第九届全国生物信息学与系统生物学学术大会

The 9th National Conference on Bioinformatics and Systems Biology of China

摘要集

ABSTRACT

2020年9月26日-29日　上海·圣诺亚皇冠假日酒店

目 录

主题 S4: 生物网络与计算系统生物学

主题 S5: 生物大分子结构模拟、预测与药物分子设计

主题 S6: 表观遗传与转录调控组学

主题 S7: 生物大数据的审编与整合

主题 S8: 微生物组学分析方法与应用

主题 S9: 复杂疾病的系统生物学

主题 S10: 生物信息学与植物科学、农业科学

主题 S11: 人工智能与生物信息学

主题 S12: 非编码 RNA 的识别与应用

特色专题 T1: 计算蛋白质组学

特色专题 T2: 计算合成生物学

特色专题 T3: 表型组学

特色专题 T4: 脑科学与生物信息学

从新冠病毒、精准医学到核酸药物

陈润生

中国科学院生物物理研究所

摘要：新型冠状病毒是一种生物结构简单、基因组仅含有不到三万个碱基的病毒，但是给人类的健康、经济造成了巨大的影响。针对新冠病毒，目前有很多问题尚未解决，例如病毒的来源无法简单地从测序的结果得出，病毒在不同物种之间、人与人之间的传播条件没有清晰的标准，感染了新冠病毒的患者表现出的症状有很大差异等。随着研究的进展，我们能够更加快速、准确地开展防控和治疗工作。

大数据方法的应用，使得医学诊断愈发有效、准确，生物医学领域也进入以大数据为特征的精准医学时代。组学数据可以对疾病进行准确的预判，同时也在药物研发和临床治疗方面发挥重要的作用。精准医学的发展将会带动相关产业的发展，同时也会影响到国家医疗体系的政策法规、安全保障制度、药物管理体系等，已经成为引领生命医学发展潮流的战略制高点。

核酸药物的研发在近几年成为药物领域快速发展的前沿方向。核酸药物具有风险小、成本低的优点，mRNA 疫苗是最有希望能够应用于治疗新冠病毒的药物之一。长非编码核酸药物具有极大的潜力，可能会对包括肿瘤在内的疾病治疗提供全新的思路，目前在肿瘤、传染病、衰老等方向均具有极大的应用潜力。

报告人 Email：chenrs@ibp.ac.cn

网络标志物的疾病诊断和动态网络标志物的疾病预警

陈洛南

中国科学院生物化学与细胞生物学研究所

摘要：网络标志物(疾病诊断和风险评估):生物标志物是医学检测的最基本工具，但传统的分子标志物一般来说随着时间和条件变化，其稳定性和准确性不能满足当前医疗检测的需求。而从系统的观点，网络是表征生物系统状态的稳定标志。由观测的分子数据，我们提出的单样本网络构建方法，可由单个样本构建分子网络标志物，因此实现由网络或网络熵诊断或预后疾病，进一步进行人体健康风险定量评估。不同于分子的表达量（平均值或一阶统计量）等，网络标志物(NB: network biomarker)是由一群分子的关联性（二阶统计量）来表征生物系统状态，所以网络标志物相对于单分子标记物的维度和量纲都不同等特点，也使得网络标志物比传统的单分子标志物的稳定性和准确性高等优点。

动态网络标志物（疾病预警和状态评估):动态网络标志物(DNB: dynamic network biomarker)方法首次提出由网络波动和关联特征预测疾病临界状态，由此可定量地检测未病状态或疾病预警，从而为实现未病预测和健康状态定量检测提供切实可行的技术手段。特别是 DNB 理论为健康人群的健康状态定量评估提供了切实可行的理论基础与方法。不同于现行医学方法所鉴定疾病后的“坏”分子（如致病基因）或网络，动态网络标志物方法能用于鉴定疾病前的“好”DNB 分子或网络。在临界状态到达前干预或提升该 DNB 分子功能，可以显著延迟健康状态恶化的临界状态的到来，从而可极大改善人类健康及提高生存质量。DNB 方法也可应用于流行性疾病等的预警。

报告人 Email：lnchen@sibs.ac.cn

多层次信息整合的分子溯源揭示 SARS-CoV2 传播过程

李亦学，张国庆，王泽峰，赵国屏，凌鋆超，王振，曹瑞芳，郑广勇

中国科学院上海营养与健康研究所

摘要：COVID-19 于 2019 年 12 月首次报道，是由新的 β-冠状病毒 SARS-CoV-2 引起的，该病毒被认为是天然来源的，并可能在中国武汉的海鲜市场（华南）扩增。尽管对病因学的基因组学进行了深入研究，但与 1 月份在该市流行的早期爆发，随后到 2 月下旬在中国的广泛传播以及 3 月份的全球爆发有关的主要遗传和流行病学事件在很大程度上仍不清楚。通过收集整理 COVID-19 全球爆发之前到爆发之后的两阶段的 SARS-CoV-2 严格质控的基因组开放数据，我们基于简约信息位点（PISs）构建了动态的系统进化网络，同时对获得的系统发育图进行拓扑分割，进而将挖掘大量的互联网信息而获得的流行病学元数据耦合到分割后的拓扑网络中。由于我们构建病毒演化系统发育树的过程更加符合病毒传播的真实场景，在这种知识增强和拓扑重建的系统进化网络的帮助下，从基于证据和由多种 PIS 定义的病毒亚型的区域分布出发，第一，我们可以从拓扑分割的 clade 内部和外部的病毒感染者的关系链出发再现病毒传播链的结构特征，从而能够从世界各个角度或水平广泛追踪 3 月中旬之前的 SARS-CoV-2 的爆发和传播途径；第二，我们揭示出了 SARS-CoV-2 传播随时间发生的 “遗传多样性”由高到低，再由低到高的阶段性分子演化特征，伴随着 COVID-19 致死率的由强到弱的变化。特别的，SARS-CoV-2 在中国武汉爆发时所呈现的丰富的“遗传多样性”表明这之前也应该存在有一个足够长的病毒“遗传多样性”由低到高演化的时间周期。与以前的基因组研究相反，分子钟测年表明与华南市场有关的菌株在时间上处于武汉早期爆发时段，但是并不是源头，这个结果也与我们的推论吻合。我们的研究结果还表明，在全球传播过程中，特别是在某些关键的时间和地理分支点，产生了广泛的奠基者效应，计算发现相应的保守蛋白经历了衰减的净化选择也支持这一点。此外，我们还通过频发突变位点的识别，揭示了 SARS-CoV-2 全球爆发的地理因素和“多样性分布”相关的平行进化和趋同进化现象。我们的研究提供了全球预防和控制 COVID-19 的关键信息。

报告人 Email：yxli@sibs.ac.cn

Splicing circRNAs from the inside out

Li Yang(杨力)

CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China

摘要：A fundamental feature of eukaryotic protein-coding and long non-protein coding genes is that they are in pieces. It is crucial that intragenic regions (introns) are spliced out of the precursor RNA and expressed regions (exons) are ligated together to form mature RNA, which is generally characteristic with 5'-cap and 3'-poly(A) tail. Surprisingly, recent transcriptome-wide analyses with specific computational pipelines have suggested that a variety of novel types of lncRNAs can be spliced out of genes, leading to their formation without 5'-cap and/or 3'-poly(A) tail. Strikingly, many of them are generated from back-spliced exons in the middle of genes to eventually form circRNAs. In the last ten years, by applying both computational and experimental methods, we and others have contributed to profile thousands of circRNAs among tissues and across species, to reveal their biogenesis mechanisms, to examine their unique secondary structures, and to uncover their biological significance. However, as sequences of individual circRNAs fully overlap with those of their cognate linear RNA isoforms, limited tools have been available to distinguish circRNAs from their cognate linear RNAs processed from the same pre-RNAs. This impedes precise quantitation and function annotation of circRNAs. In this talk, I will summarize our current understanding on circRNA biogenesis and function, and importantly, further discuss our efforts that aim to develop new methods for circRNA study.

关键词：circular RNA, splicing, back-splicing, lncRNAs, pre-RNA, transcriptome

报告人 Email: liyang@picb.ac.cn

单细胞生物信息学及其在肿瘤免疫微环境的应用

张泽民

北京大学

摘要：单细胞转录组测序技术的飞速发展，为在细胞分辨率刻画和解析生殖、发育、免疫、神经、肿瘤等生命过程提供了强有力的技术手段。大量的新型单细胞数据的产生也为生物信息学带来了挑战和机遇。我们在处理单细胞数据的聚类、分型、定义、整合、可视化，以及在基于单细胞数据分析细胞的相互作用和动态关系方面做了方法学研究，然后利用这些生物信息方法解读大量和肿瘤免疫微环境的单细胞转录本数据。我们分析了肝癌、结直肠癌、肺癌、和多种其他癌种中的浸润性免疫细胞，发现了不同组织部位免疫细胞在组成、亚型、功能方面的不同，进一步发现了不同癌种之间的差异。我们进一步对不同组织部位的免疫细胞进行了动态跟踪，发现了某些细胞类型特有的迁移规律。进而我们发现了肿瘤免疫的新靶点、以及和临床预后相关的生物标记物，也对肿瘤微环境的系统性单细胞分析提供了新的模式。

报告人 Email：zemin@pku.edu.cn

Linking Gut Dysbiosis to Human Health: Resources, Tools and Cofounding Factors

Sicheng Wu[1], Puzi Jiang [1], Wei-Hua Chen [1](陈卫华)

[1] *College of Life Science and technology, Huazhong University of Science and Technology, Wuhan China*

摘要：Gut dysbiosis has been extensively studied and linked to various diseases. Common tasks for bioinformaticians include identifying differential features, building machine learning models and cross-validation from different studies (i.e. meta-analysis). Here I introduce recently developed resources and tools by our group to facilitate these tasks. At the end, I will demonstrate how to utilize them to build diagnostic models, perform meta-analysis and identify possible confounding factors

关键词：gut metagenomics, dysbiosis, health, machine learning, database

报告人 Email: weihuachen@hust.edu.cn

癌症基因组中的克隆和子克隆演化结构推断

寸玉鹏

中国科学院昆明植物研究所

摘要：肿瘤的发生是癌细胞基因组经历一系列复杂的遗传变异的结果。生物学家用“克隆(clone)”来描述来自同一个遗传祖先的细胞，和正常细胞一样，这些癌细胞的扩增复制也服从达尔文自然演化理论。癌细胞的克隆演化理论很好地解释了癌细胞在个体中的起源和发展，为癌细胞强大的入侵、生存，转移和抗药性能力提供了定量理论研究基础。如何从基因组数据中鉴定出克隆和亚克隆群体结构，需要精确的模型来对基因组数据进行统计推断。常用的克隆结构推断模型大都基于狄利克雷过程，这类方法计算耗时长，精度也不稳定性。基于现有方法的劣势，我们提出了一种全新的完全非参数突变聚类方法用于亚克隆群体结构的推断，即通过一个样条(spline)函数和反卷积计算来超快速地找出癌基因组中的克隆和亚克隆群体，并推出 Sclust 软件包来对癌组织样本的癌细胞纯度和基因拷贝数变异进行联合估计，然后计算每个变异的癌细胞组分用于亚克隆群体结构的重构，以及相关的克隆/亚克隆拷贝数。

报告人 Email：yp.cun@outlook.com

基因组加倍与植物适应性进化

Yuannian Jiao[1,2](焦远年)

[1] *State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China.*

[2] *University of Chinese Academy of Sciences, Beijing 100049, China.*

摘要：Ancient whole-genome duplications (WGD or polyploidy) are prevalent in plants, and some WGDs occurred during the timing of severe global environmental changes. It has been suggested that WGDs may have contributed to plant adaptation. However, it still lacks of empirical evidence from genetic level to support the hypothesis. Here, we investigated the survivors of gene duplicates from multiple ancient WGD events on the major branches of angiosperm phylogeny, and aimed to explore genetic evidence supporting the significance of polyploidy. Duplicated genes co-retained from three waves of independent WGDs (~120 million years ago (Ma), ~66 Ma and <20 Ma) were investigated in 25 selected species. Gene families functioning in low temperature and darkness were commonly retained gene duplicates after the eight independently occurred WGDs in many lineages around the Cretaceous–Paleocene (K-Pg) boundary, when the global cooling and darkness were the two main stresses. Moreover, the commonly retained duplicates could be key factors which may have contributed to the robustness of the critical stress related pathways. In addition, genome-wide transcription factors (TFs) functioning in stresses tend to retain duplicates after waves of WGDs, and the co-selected gene duplicates in many lineages may play critical roles during severe environmental stresses. Finally, our results shed new light on the significant contribution of paleopolyploidy to plant adaptation during global environmental changes in the evolutionary history of angiosperms.

关键词：whole-genome duplication, paleopolyploidy, adaptive evolution, phylogenomic, Cretaceous-Paleocene boundary, gene regulatory network

报告人 Email：jiaoyn@ibcas.ac.cn

昆虫基因组：仍需挖掘的财富

李飞[1]，杨念婉[2]，黄聪[2]，樊晓丹[3]，钱万强[4]，万方浩[4]

[1] *浙江大学昆虫科学研究所，310058，杭州*
[2] *中国农科院植保所，100193, 北京*
[3] *香港中文大学统计系，香港*
[4] *中国农科院深圳基因组研究所，518120，深圳*

摘要：昆虫是地球上物种最丰富的动物类群，人类已知的昆虫达 100 多万种，超过所有生物种类的一半。黑腹果蝇作为重要的模式生物，对其研究非常深入。然而，对其他非模式昆虫的研究，仍需进一步加强。通过收集和整理近年来的研究结果，将简要地介绍我国学者在观赏昆虫（蝴蝶）、授粉昆虫、蝗虫、农业害虫、天敌昆虫（寄生蜂和瓢虫等）、入侵有害生物等领域的基因组分析和比较基因组学研究进展。

以入侵昆虫基因组学研究为例，介绍 “1000 种入侵生物全基因组计划(IAS1000)”及其进展、入侵有害生物数据库的构建（InvsionDB）。在此基础上，通过对 37 个入侵昆虫和 6 个非入侵昆虫的比较基因组学，发现了 24 个与防御、营养、化学感受等相关的基因家族在入侵昆虫中显著扩增，利用逻辑回归模型提出了利用入侵指数（Invasion index）对有害生物的入侵性进行预测。进一步地，利用随机森林提取了入侵昆虫的基因组特征，构建了逻辑回归模型的分类器，对昆虫的入侵性进行预测，准确率达 93.2%，灵敏性和特异性均达到了 100%。外来入侵有害生物对我国的环境、农业、生态、卫生和经济均造成了巨大的影响，我国海关部门在入侵生物的检验检疫等领域投入了大量的人力物力。入侵昆虫数据库的构建及昆虫入侵性预测算法的开发，对有害生物检验检疫具有重要的推动作用。

迄今为止，已有 500 多种昆虫的基因组被测序报道，但绝大多数昆虫基因组尚没有得到深入地分析和挖掘，期待更多的生物信息同行加入昆虫基因组领域的研究。

报告人 Email：lifei15@zju.edu.cn

南极衣藻适应极端环境的进化基因组学研究

Zhenhua Zhang[1], Changfeng Qu[2,3], Kaijian Zhang[4], Yingying He[2], Xing Zhao[4], Lingxiao Yang[1], Zhou Zheng[2,3], Xiaoya Ma[1], Xixi Wang[2], Wenyu Wang[2], Kai Wang[2], Dan Li[2], Liping Zhang[2], Xin Zhang[2], Danyan Su[1], Xin Chang[1], Mengyan Zhou[4], Dan Gao[4], Wenkai Jiang[4], Frederik Leliaert[5,6], Debashish Bhattacharya[7], Olivier De Clerck[5], Bojian Zhong[1] (钟伯坚), Jinlai Miao[2,3]

[1] *College of Life Sciences, Nanjing Normal University, 210023 Nanjing, China*
[2] *First Institute of Oceanography, Ministry of Natural Resources, 266061 Qingdao, China*
[3] *Laboratory for Marine Drugs and Bioproducts of Qingdao National Laboratory for Marine Science and Technology, 266237 Qingdao, China*
[4] *Novogene Bioinformatics Institute, 100083 Beijing, China*
[5] *Biology Department, Ghent University, 9000 Ghent, Belgium*
[6] *Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium*
[7] *Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey 08901, USA*

摘要：The unicellular green alga *Chlamydomonas* sp. ICE-L thrives in polar sea ice, where it tolerates extreme low temperatures, high salinity, and broad seasonal fluctuations in light conditions. Despite the high interest in biotechnological uses of this species, little is known about the adaptations that allow it to thrive in this harsh and complex environment. Here we assembled a high-quality genome sequence of ~542 megabases and found that retrotransposon proliferation contributed to the relatively large genome size of ICE-L when compared to other chlorophytes. Genomic features that may support the extremophilic lifestyle of this sea ice alga include massively expanded gene families involved in unsaturated fatty acid biosynthesis, DNA repair, photoprotection, ionic homeostasis, osmotic homeostasis, and reactive oxygen species detoxification. The acquisition of multiple ice binding proteins through putative horizontal gene transfer likely contributed to the origin of the psychrophilic lifestyle in ICE-L. Additional innovations include the significant up-regulation, under abiotic stress, of several expanded ICE-L gene families, likely reflecting adaptive changes among diverse metabolic processes. Our analyses of the genome, transcriptome, and functional assays advance general understanding of the Antarctic green algae, and offer potential explanations for how green plants adapt to extreme environments.

关键词：adaptive evolution, extreme Antarctic environments, comparative genomics, sea ice green algae, de novo genome

报告人 Email：bjzhong@gmail.com

Anti-masculinization Induced by Aromatase Inhibitors in Adult Female Zebrafish

Lu Chen[1], Li Wang[1], Qiwei Cheng[1], Yi-Xuan Tu[1], Zhuang Yang[1], Run-Ze Li[1], Zhi-Hui Luo[1], Zhen-Xia Chen[1](陈振夏)

[1] *College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, PR China*

摘要： Early sex differentiation genes of zebrafish remain an unsolved mystery due to the difficulty to distinguish the sex of juvenile zebrafish. However, aromatase inhibitors (AIs) could direct juvenile zebrafish sex differentiation to male and even induce ovary-to-testis reversal in adult zebrafish. In order to determine the transcriptomic changes of sex differentiation in juvenile zebrafish and early sex reversal in adult zebrafish, we sequenced the transcriptomes of juvenile and adult zebrafish treated with AI exemestane (EM) for 32 days, when juvenile zebrafish sex differentiation finished. EM treatment in females upregulated the expression of genes involved in estrogen metabolic process, female gamete generation and oogenesis, including gsdf, macf1a and paqr5a, while down-regulated the expression of vitellogenin (vtg) genes, including vtg6, vtg2, vtg4, and vtg7 due to the lower level of Estradiol (E2). Furthermore, EM-juveniles showed upregulation in genes related to cell death and apoptosis, such as bcl2l16 and anax1c, while the control-juveniles exhibited up-regulation of genes involved in positive regulation of reproductive process and oocyte differentiation such as zar1 and zpcx. Moreover, EM-females showed higher enrichment than control females in genes involved in VEGF signaling pathway, glycosaminoglycan degradation, hedgehog signaling pathway, GnRH signaling pathway and steroid hormone biosynthesis. Our study shows anti-masculinization in EM-treated adult females but not in EM-treated juveniles. This may be responsible for the lower sex plasticity in adults than juveniles.

关键词： Sex differentiation, Zebrafish, RNA-Seq, Adult, Aromatase inhibitor, Sex reversal

报告人 Email: zhen-xia.chen@mail.hzau.edu.cn

Mako: a Graph-based Pattern Growth Approach to Detect Complex Structural Variants

Jiadong Lin[1,2,3,4], Xiaofei Yang[2,5], Tun Xu[1], Qihui Zhu[6], Eliza Cerveira[6], Mallory Ryan[6], Charles Lee[6,7], Li Guo[2,8], Walter Kosters[4], Chengsheng Zhang[6,7], the Human Genome Structural Variation Consortium, Kai Ye[1,2,3,8](叶凯)

[1] *School of Automation Science and Engineering, Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China.*

[2] *MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China.*

[3] *Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061 China.*

[4] *Leiden Institute of Advanced Computer Science, Faculty of Science, Leiden University, Leiden, Netherland.*

[5] *The School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China.*

[6] *The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032.*

[7] *Precision Medicine Center, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061 China.*

[8] *The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, 710049 China.*

摘要： Complex structural variants (CSVs) are genomic alterations that have more than two breakpoints and are considered as simultaneous occurrence of simple structural variants. However, detecting the compounded mutational signals of CSVs is challenging through a commonly used model-match strategy. As a result, there has been little progress for CSV discovery compared with simple structural variants. We systematically analyzed the multi-breakpoint connection feature of CSVs, and proposed Mako, utilizing a bottom-up guided model- free strategy, to detect CSVs from paired-end short-read sequencing. Specifically, we implemented a graph-based pattern growth approach, where the graph depicts potential breakpoint connections and pattern growth enables CSV detection without predefined models. Comprehensive evaluations on both simulated and real datasets revealed that Mako outperformed other CSV discovery algorithms. Meanwhile, Mako CSV subgraph effectively characterized the breakpoint connections of a CSV event. Additionally, Mako uncovered a total of 15 CSV types from HG00514, HG00733 and NA19240, including two novel types of adjacent segments swap and tandem dispersed duplication. Further analysis of these CSVs also revealed a novel role of sequence homology in the formation of different CSVs.

关键词： structural variant, pattern growth, graph, model- free strategy, bottom-up

报告人 Email: kaiye@xjtu.edu.cn

Whole Genome Analyses of a Healthy Chinese Population

Zhenglin Du[1](杜政霖), Liang Ma[1], Hongzhu Qu[1], Wei Chen[1], Bing Zhang[1], Na Yuan[1], Xi Lu[1], Fan Liu[1,2], Xiangdong Fang[1,2], Hua Chen[1,2], Xin Liu[1,2], Jingfa Xiao[1,2], and Changqing Zeng[1,2]

[1] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

摘要：Unraveling the genetic mechanism of diseases and physiological traits requires comprehensive sequencing analysis on large amount of population samples. As a main part of the Chinese Academy of Sciences Precision Medicine Initiative (CASPMI) project, a healthy Chinese cohort was established including ~1000 participants from nine ethnic groups of 30 provinces or autonomous regions in China. By whole-genome sequencing at a depth of ~30X and variation identification, a genomic variation map of Chinese populations was generated, which contained 24.85 million single-nucleotide variants and 3.85 million small indels.

To further characterize population-specific variants in the Chinese cohort, the frequencies of the SNPs and indels from our samples were compared with those of different populations from the 1000 genomes project. Among these identified population-specific variants, 40 SNPs in 38 genes were correlated with various metabolism-related traits or diseases, according to the GWAS Catalog records, and the SNP rs1549293 in *KAT8* was found significantly associated with the waist circumference in the males of the CASPMI cohort. Furthermore, significant genetic diversity was observed between northerners and southerners in the cohort. Especially, the SNP rs1801133, which is in the locus of *MTHFR* (methylenetetrahydrofolate reductase) on chromosome 1, shows visible differentiation. The frequency of allele A (*MTHFR* 667T) was 57.8% in northerners and 37.9% in southerners. By mapping the frequencies in our population and other reported frequencies to their geographic locations, there exists an adaptation zone between 35–45 degree North on Afro-Eurasia continents (**Figure 1**), where the frequency of 667T was maintained high and decreases both northward and southward from this belt.
The population-specific data reported here will provide a useful resource for population studies and support future investigations aimed at providing precision medicine and individualized healthcare.

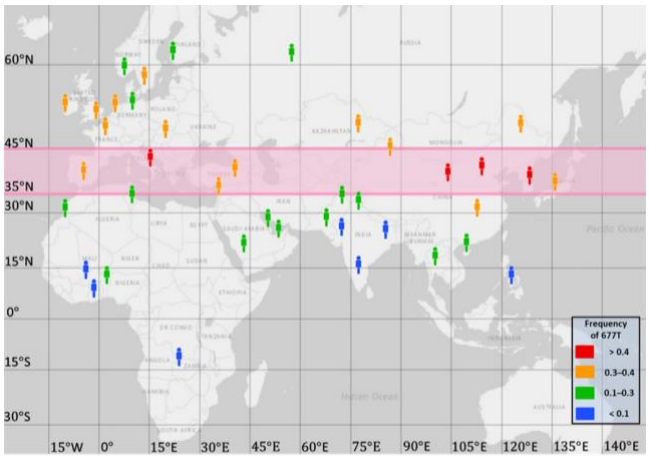

Figure 1 Frequency distribution of *MTHFR* 667T (rs1801133)

关键词：Variation map; Phenotype association; Large population

报告人 Email：duzhl@big.ac.cn

Computational Methods for Scalable Embedding and Network Reconstruction using Single-cell ATAC-seq Data

张世华

中国科学院数学与系统科学研究院

摘要： With the rapid development of single-cell ATAC-seq technology, it has become possible to profile the chromatin accessibility of massive individual cells. However, it remains challenging to characterize their regulatory heterogeneity due to the high-dimensional, sparse and near-binary nature of data. First, we developed a network diffusion method for scalable embedding of massive single-cell ATAC-seq data (named as scAND). scAND can take information from similar cells to alleviate the sparsity and improve cell type identification. Extensive tests and comparison with existing methods using synthetic and real data as benchmarks demonstrated its distinct superiorities in terms of clustering accuracy, robustness, scalability and data integration. Second, we adopted a computational method to **j**ointly **r**econstruct *cis*-regulatory **i**nteraction **m**aps (JRIM) of multiple cell populations based on patterns of co-accessibility in single-cell data. Reconstructed common interactions among 13 tissues indeed relate to basic biological functions, and individual *cis*-regulatory networks show strong tissue specificity and functional relevance. More importantly, tissue-specific regulatory interactions are mediated by coordination of histone modifications and tissue-related TFs, and many of them reveal novel regulatory mechanisms.

报告人 Email： zsh@amss.ac.cn

Mapping Mammalian Cell Landscapes by Single cell mRNA-seq

Xiaoping Han[1], & Guoji Guo[1](郭国骥)

[1] *Center for Stem Cell and Regenerative Medicine, Zhejiang University School of Medicine, Hangzhou 310058, China*

摘要：Single-cell analysis is a valuable tool to dissect cellular heterogeneity in complex systems. We used single-cell RNA sequencing to determine the cell-type composition of all major mouse and human organs to construct cell landscapes for the mammalian systems. We revealed single-cell hierarchies for many tissues that have not been well characterised. We established a cell mapping pipeline that helps to define mammalian cell identity. Finally, we performed a single-cell comparative analysis of landscapes from both human and mouse to reveal the conserved genetic networks. In the mammalian systems, stem and progenitor cells exhibt strong transcriptomic stochasticity, while the differentiated cells are more distinct.

报告人 Email：ggj@zju.edu.cn

单细胞转录谱功能通路活性挖掘在寻找癌细胞起源的应用研究

苏建忠

温州医科大学

摘要：单细胞转录组测序技术可以从单个细胞的水平分析细胞间的异质性，研究肿瘤微环境，寻找癌细胞的起源等，在传统分析流程中对细胞分群进行功能性解释一直是单细胞转录组解析的一大挑战。我们开发对单细胞进行转录组功能通路注释新方法 scTPA，提供在通路水平上对单细胞转录组数据预处理，功能通路活性打分，聚类和细胞类型特异的激活通路挖掘及可视化等系统研究。并应用到视网膜母细胞瘤和脑癌中来挖掘癌细胞的起源。

报告人 Email：sujz@wmu.edu.cn

基于单细胞转录组测序数据从头重构细胞空间关系

任仙文

北京大学

摘要：单细胞转录组测序(scRNA-seq)通过提供前所未有的细胞和分子通量，彻底改变了转录组研究，但在组织分离过程中，单个细胞的空间信息会丢失。虽然基于成像的技术如原位测序显示了巨大的前景，但目前的技术挑战限制了它们的广泛应用。在这里，我们假设细胞空间关系是由细胞表型内在编码的，并且通过配体和受体的相互作用发生自组织形成。因此，我们提出了名为 CSOmap 的计算工具，只基于单细胞转录组测序数据从头重构细胞的空间关系。在人和小鼠的十多个器官的五种单细胞转录组测序平台上，我们对 CSOmap 进行了验证，证明了其有效性。特别地，CSOmap 可以在计算机水平上模拟基因或细胞类型的干扰实验，来检验其对细胞空间关系的影响，从而获得新的生物学洞见。我们把 CSOmap 用于研究肿瘤微环境，做出来一系列新的发现。CSOmap 是一个可广泛适用于不同单细胞转录组测序数据的计算工具，对解析细胞空间关系具有重要作用。

报告人 Email：renxwise@pku.edu.cn

整合单细胞及群体细胞多组学数据的统一数学框架

曾婉雯

南开大学

摘要：单细胞测序技术的飞速发展，使不同细胞类型得以精细区分，使得科学家们在单细胞水平进行分子机制研究成为可能。单细胞测序与群体测序不同在于群体测序所提取的 RNA（或 DNA）源于样本中的多个细胞，所以群体测序的结果不可避免的会受到不同细胞间异质性的影响，而单细胞测序则是针对单个细胞的基因组进行测序，能够更好地帮助我们认识细胞与细胞之间的差异。对同一个样本，如果既有单细胞测序数据，也有群体测序数据，由于这些不同的数据都有一样的亚群信息，他们之间的分析应该能够相辅相成，使得分析结果更加准确。对单细胞数据进行聚类得到细胞亚群的信息，并将群体细胞数据解卷积得到对应细胞亚群的数据，将具有相当重要的科学意义。特别地，如果能够得到群体数据中准确的调控网络数据，并将其解卷积成亚群特异的调控网络数据，会大大促进对单细胞异质性的理解。

作者提出了提出了整合单细胞多组学数据和细胞群体组学数据的统一数学框架 DC3，将基因调控网络建模推进到单细胞层面，可以使用以下常见类型的群体数据和单细胞数据的组合作为输入：1）scRNA-seq，scATAC-seq 和 scHi-C; 2）scRNA-seq，scATAC-seq 和群体 HiChIP; 3）scRNA-seq，群体 ATAC-seq，群体 HiChIP; 4）群体 RNA-seq，scATAC-seq，群体 HiChIP。通过输入单细胞水平数据及群体细胞水平数据，DC3 对单细胞水平的数据进行聚类得到对应的细胞亚群水平的数据，并对群体细胞数据解卷积到对应细胞亚群水平的数据。大量的模拟实验表示，DC3 可以成功地将群体测序数据解卷积为细胞亚群特异的数据。同时，亚群特异数据反过来改善单细胞水平数据的联合聚类结果。DC3 克服单细胞层面样本难以匹配的困难，能够对单细胞多组学数据进行更准确的聚类。同时突破三维基因组学数据难以在单细胞层次观测的瓶颈，实现对群体数据进行解卷积分解，计算结果得到实验验证。

报告人 Email：wwzeng@nankai.edu.cn

Deep soft K-means clustering with self-training for single-cell RNA sequence data

Liang Chen[1], Weinan Wang[1], Yuyao Zhai[2] and Minghua Deng[1,3](邓明华)

[1] *School of Mathematical Sciences, Peking University, Beijing 100871, China*
[2] *Northeast Normal University, Changchun 130024, China*
[3] *Center for Quantitative Biology, Peking University, Beijing 100871, China*

摘要： Single-cell RNA sequencing (scRNA-seq) allows researchers to study cell heterogeneity at the cellular level. A crucial step in analyzing scRNA-seq data is to cluster cells into subpopulations to facilitate subsequent downstream analysis. However, frequent dropout events and increasing size of scRNA-seq data make clustering such high-dimensional, sparse and massive transcriptional expression profiles challenging. Although some existing deep learning based clustering algorithms for single cells combine dimensionality reduction with clustering, they either ignore the distance and affinity constraints between similar cells or make some additional latent space assumptions like mixture Gaussian distribution, failing to learn cluster-friendly low-dimensional space. Therefore, in this paper, we combine the deep learning technique with the use of a denoising autoencoder to characterize scRNA-seq data while propose a soft self-training K-means algorithm to cluster the cell population in the learned latent space. The self-training procedure can effectively aggregate the similar cells and pursue more cluster-friendly latent space. Our method, called 'scziDesk', alternately performs data compression, data reconstruction and soft clustering iteratively, and the results exhibit excellent compatibility and robustness in both simulated and real data. Moreover, our proposed method has perfect scalability in line with cell size on large scale datasets.

关键词： single cell RNA-seq, self-training k-means algorithm, Denoising autoencoder, zero-inflated negative binomial, KL divergence.

报告人 Email: dengmh@pku.edu.cn

Integrative Analyses of Single-cell Transcriptome and Regulome using MAESTRO

Chenfei Wang[1,2](王晨飞), Dongqing Sun[3], Xin Huang[4], Changxin Wan[3], Ziyi Li[3], Ya Han[3], Qian Qin[3], Jingyu Fan[3], Xintao Qiu[2,5], Yingtian Xie[2,5], Clifford A. Meyer[1,2], Myles Brown[2,5], Ming Tang[1,2], Henry Long[2,5], Tao Liu[6], X. Shirley Liu[1,2]

[1] *Department of Data Science, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA, 02215, USA*

[2] *Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA*

[3] *Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Science and Technology, Tongji University, Shanghai 200433, China.*

[4] *Beijing Institute of Radiation Medicine, Beijing 100850, China.*

[5] *Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA.*

[6] *Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA.*

摘要： We present Model-based AnalysEs of Transcriptome and RegulOme (MAESTRO), a comprehensive open-source computational workflow (http://github.com/liulab-dfci/MAESTRO) for the integrative analyses of single-cell RNA-seq (scRNA-seq) and ATAC- seq (scATAC-seq) data from multiple platforms. MAESTRO provides functions for pre-processing, alignment, quality control, expression and chromatin accessibility quantification, clustering, differential analysis, and annotation. By modeling gene regulatory potential from chromatin accessibilities at the single-cell level, MAESTRO outperforms the existing methods for integrating the cell clusters between scRNA- seq and scATAC-seq. Furthermore, MAESTRO supports automatic cell-type annotation using predefined cell type marker genes and identifies driver regulators from differential scRNA-seq genes and scATAC-seq peaks.

关键词： Single-cell RNA-seq, Single-cell ATAC-seq, Computational workflow, Integrate scRNA-seq and scATAC-seq, Cell-type annotation, Predict transcriptional regulators

报告人 Email: 08chenfeiwang@tongji.edu.cn

CytoTalk: *De novo* construction of signal transduction networks using single-cell RNA-Seq data

Yuxuan Hu [1](胡宇轩), Tao Peng [2,3], Lin Gao [1], Kai Tan [2,3,4,5]

[1]*School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China*
[2]*Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA*
[3]*Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA*
[4]*Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA 19104, USA*
[5]*Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

摘要：Single-cell technology has opened the door for studying signal transduction in a complex tissue at unprecedented resolution. However, there is a lack of analytical methods for *de novo* construction of signal transduction pathways using single-cell omics data. Here we present CytoTalk, a computational method for *de novo* constructing cell type-specific signal transduction networks using single-cell RNA-Seq data. CytoTalk first constructs intracellular and intercellular gene-gene interaction networks using an information-theoretic measure between two cell types. Candidate signal transduction pathways in the integrated network are identified using the prize-collecting Steiner forest algorithm. We applied CytoTalk to a single-cell RNA-Seq data set on mouse visual cortex and evaluated predictions using high-throughput spatial transcriptomics data generated from the same tissue. Compared to published methods, genes in our inferred signaling pathways have significantly higher spatial expression correlation only in cells that are spatially closer to each other, suggesting improved accuracy of CytoTalk. Furthermore, using single-cell RNA-Seq data with receptor gene perturbation, we found that predicted pathways are enriched for differentially expressed genes between the receptor knockout and wild type cells, further validating the accuracy of CytoTalk. In summary, CytoTalk enables *de novo* construction of signal transduction pathways and facilitates comparative analysis of these pathways across tissues and conditions.

关键词： cell-cell communication, signaling networks, single-cell RNA-Seq data, spatial transcriptomics

报告人 Email: yuxuan_hu_xd@163.com

基于电子病历数据的疾病严重程度分型预警: HNC-LL Model for COVID-19

Lu-shan Xiao[1,2], Wen-Feng Zhang[3], Meng-chun Gong[4], Yan-pei Zhang[2], Li-ya Chen[1], Hong-bo Zhu[2,5], Chen-yi Hu[2], Pei Kang[1], Li Liu[1,2](刘莉), Hong Zhu[4]

[1] *Department of Medical Quality Management, Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China*

[2] *Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China*

[3] *Department of Infectious Diseases, The First Affiliated Hospital, Nanchang University, Nanchang, 330006, China*

[4] *Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China*

[5] *Department of Oncology, the First Affiliated Hospital of University of South China, Hengyang 421001, China*

摘要： *Background:* Information regarding risk factors associated with severe coronavirus disease (COVID-19) is limited. This study aimed to develop a model for predicting COVID-19 severity.

Methods: Overall, 690 patients with confirmed COVID-19 were recruited between 1 January and 18 March 2020 from hospitals in Honghu and Nanchang; finally, 442 patients were assessed. Data were categorised into the training and test sets to develop and validate the model, respectively.

Findings: A predictive HNC-LL (**H**ypertension, **N**eutrophil count, **C**-reactive protein, **L**ymphocyte count, **L**actate dehydrogenase) score was established using multivariate logistic regression analysis. The HNC-LL score accurately predicted disease severity in the Honghu training cohort (area under the curve [AUC]=0.861, 95% confidence interval [CI]: 0.800–0.922; $P<0.001$); Honghu internal validation cohort (AUC=0.871, 95% CI: 0.769–0.972; $P<0.001$); and Nanchang external validation cohort (AUC=0.826, 95% CI: 0.746–0.907; $P<0.001$) and outperformed other models, including CURB-65 (confusion, uraemia, respiratory rate, BP, age ≥65 years) score model, MuLBSTA (multilobular infiltration, hypo-lymphocytosis, bacterial coinfection, smoking history, hypertension, and age) score model, and neutrophil-to-lymphocyte ratio model. The clinical significance of HNC-LL in accurately predicting the risk of future development of severe COVID-19 was confirmed.

Interpretation: We developed an accurate tool for predicting disease severity among COVID-19 patients. This model can potentially be used to identify patients at risks of developing severe disease in the early stage and therefore guide treatment decisions.

Funding: This work was supported by the National Nature Science Foundation of China (grant no. 81972897), Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2015) and National Key Research & Development Program of China (Project No. 2018YFC0116901).

关键词： COVID-19, SARS-CoV-2, severity, prediction, HNC-LL

报告人 Email： liuli@i.smu.edu.cn

基于组学大数据的新冠病毒基因功能预测与潜在药物筛选

Hui Zhi[1], Xu Pan[1], Xin Li[1], Shangwei Ning[1](宁尚伟)

[1] *College of Bioinformatics, Harbin Medical University, Harbin, China*

摘要：The rapid spread of the coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 has aroused worldwide concern. However, to date, our knowledge of the function of viral genes remains limited. To systematically investigate the viral genome transcription, we first obtained the RNA-seq datasets of SARS-CoV-2 infected cells from GSE1475074. We aligned the raw reads to the GENCODE GRCh38 human genome assembly using Bowtie2 and normalized the counts of protein coding genes (PCGs) and long non-coding RNAs (lncRNAs) to TPM (Transcripts Per Million) using RSEM6. Next, the un-aligned reads were aligned to the SARS-CoV-2 reference file (NCBI Reference Sequence: NC_045512.2) using Bowtie2. The viral gene quantification referred to the viGEN pipeline. Compared with other viral genes, the nucleocapsid (N) gene expression level was the highest and the ORF7b was the lowest. Moreover, we also identified the differentially expressed PCGs/lncRNAs in 5 SARS-CoV-2-infected- and mock-infected control experiments using the Mann-Whitney U test. There were 1,078 differentially expressed PCGs, and 68 differentially expressed lncRNAs with p-vlaue<0.05 in more than 3 control experiments. Next, we identified the SARS-CoV-2 co-expressed genes (SCOGs). 762 SCOGs were found in the SARS-CoV-2-infected cells. Moreover, we obtained the drug-SCOGs interactions from DGIdb10, and TTD11 and constructed the drug-SCOGs interaction network (149 drugs, 52 SCOGs). There were several potentially antiviral drug-target gene interactions revealed in the network. Among them, JQ1 was a potent inhibitor of the BRD2 (Bromodomain Containing 2) that could co-express with the SARA-CoV-2 E gene. NMS-873, a small molecule inhibitor of VCP, was found to inhibit SARS-CoV-2 replication at low nanomolar concentrations. Next, we analyzed the competitive endogenous RNA (ceRNA) in host cells to better characterize the host environment for SARS-CoV-2 infection. we obtained the miRNA, mRNA, and lncRNA expression profiles of SARS-CoV-2-infected Calu3 cells from GSE14872915. The ceRNA network contained 3 miRNAs (miR-4745-3p, miR-3150b-3p, and miR-411-5p), 2 lncRNAs (AL365203.2, AC135983.2), and 38 mRNAs. Besides, these non-coding RNAs also show antiviral and immune activity. In summary, our study annotated the SARS-CoV-2 functional genomics and identified the potential drugs for SCOGs. These results provided novel insights into the understanding of SARS-CoV-2 functional genomics and potential host-targeting antiviral strategies for SARS-CoV-2 infection.

关键词：COVID-19, PCGs/lncRNAs, co-expression, drug-SCOGs, ceRNA

报告人 Email: ningsw@ems.hrbmu.edu.cn

新型冠状病毒信息资源整合及变异分析

Shuhui Song[1,2,3,4](宋述慧), Lina Ma[1,2,3], Dong Zou[1,2,3], Dongmei Tian[1,2], Cuiping Li[1,2], Junwei Zhu[1,2], Zhang Zhang[1,2,3,4], Wenming Zhao[1,2,3,4], Yongbiao Xue[1,2,4], Yiming Bao[1,2,3,4]

[1] China National Center for Bioinformation, Beijing 100101, China
[2] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[3] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China

摘要：On 22 January 2020, the China National Center for Bioinformation (CNCB) / National Genomics Data Center (NGDC) created the 2019 Novel Coronavirus Resource (2019nCoVR,https://bigd.big.ac.cn/ncov/), an open-accessed SARS-CoV-2 information resource. 2019nCoVR features comprehensive integration of sequence and clinical information for all publicly available SARS-CoV-2 isolates, which are manually curated with value-added annotations and quality evaluated by our in-house automated pipeline. Of particular note, 2019nCoVR performs systematic analyses to obtain a dynamic landscape of SARS-CoV-2 genomic variations at a global scale. It provides all identified variants and detailed statistics for each virus isolate, and congregates the quality score, functional annotation, and population frequency for each variant. It also generates visualization of the spatiotemporal change for each variant and yields historical viral haplotype network maps for the course of the outbreak from all complete and high-quality genomes. Moreover, 2019nCoVR provides a full collection of literatures on COVID-19, including published papers from PubMed as well as preprints from services such as bioRxiv and medRxiv through Europe PMC. Furthermore, by linking with relevant databases in CNCB/NGDC, 2019nCoVR offers data submission services for raw sequence reads and assembled genomes, and data sharing with National Center for Biotechnology Information. Collectively, all SARS-CoV-2 genome sequences, variants, haplotypes and literatures are updated daily to provide timely information, making 2019nCoVR a valuable resource for the global research community.

关键词：2019nCoVR; SARS-CoV-2; Genome assemblies; Genomic variation; Haplotype

报告人 Email: songshh@big.ac.cn

Data-driven Hybrid Surveillance System for COVID-19: the Honghu System

Gong Mengchun[1](弓孟春), Liu Li[2], Sun Xin[3], Yang Yue[2], Shuang Wang[4], Zhu Hong[1,2]

[1] *Institute of Health Management, Southern Medical University, Guangzhou, China*
[2] *Nanfang Hospital, Southern Medical University, Guangzhou, China*
[3] *Chinese Evidence-based Medicine Center, West China Hospital, Sichuan University, Chengdu, China*
[4] *Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu, China*

摘要：*Background and Objective:* COVID-19 has been an unprecedented challenge to the global healthcare system. Tools that can improve the focus of surveillance efforts and clinical decision support are of paramount importance. Therefore, new medical informatics technologies are needed to enable effective control of the pandemic.

Methods: The Honghu Hybrid System (HHS) for COVID-19 collected, integrated, standardized and analyzed data from multiple sources, including the case reporting system, diagnostic labs, electronic medical records and social media on mobile devices.

Results: HHS was developed and successfully deployed within 72 hours in the city of Honghu in Hubei Province, China. Syndromic surveillance component in HHS covered over 95% of the population of over 900,000 people and provided near real-time evidence for the control of epidemic emergencies. Clinical decision support component in HHS was also provided to improve patient care and prioritize the limited medical resources.

Conclusions: The facilitating factors and challenges are discussed to provide useful insights to other cities to build up suitable solutions based on big-data technologies. The HHS for COVID-19 proved to be feasible, sustainable and effective and can be migrated.

关键词：COVID-19, Big Data, Syndromic Surveillance, Clinical Decision Support, Stakeholders involvement

报告人 Email：gmc@nrdrs.org

Single-cell Analysis of Two Severe COVID-19 Patients Reveals a Monocyte-associated and Tocilizumab-responding Cytokine Storm

Chuang Guo[1], Bin Li[1], Huan Ma[1], Xiaofang Wang[2], Pengfei Cai[1], Qiaoni Yu[1], Lin Zhu[1], Liying Jin[1], Chen Jiang[1], Jingwen Fang[3], Qian Liu[1], Dandan Zong[1], Wen Zhang[1], Yichen Lu[1], Kun Li[1], Xuyuan Gao[1], Binqing Fu[1,4], Lianxin Liu[2], Xiaoling Ma[5], Jianping Weng[6], Haiming Wei[1,4], Tengchuan Jin[1,4], Jun Lin[1,4], Kun Qu[1,4,7](瞿昆)

[1] Department of oncology, The First Affiliated Hospital of USTC, Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230021, China.

[2] Department of Hepatobiliary Surgery, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230021, China.

[3] HanGene Biotech, Xiaoshan Innovation Polis, Hangzhou, Zhejiang, 311200, China.

[4] CAS Center for Excellence in Molecular Cell Sciences, the CAS Key Laboratory of Innate Immunity and Chronic Disease, University of Science and Technology of China, Hefei, Anhui, 230027, China.

[5] Department of Laboratory Medicine, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, China.

[6] Department of Endocrinology and Metabolism, The First Affiliated Hospital of USTC, Division of Life Sciences of Medicine, University of Science and Technology of China, Hefei 230026, China.

[7] School of Data Science, University of Science and Technology of China, Hefei 230026, China.

摘要：Several studies show that the immunosuppressive drugs targeting the interleukin-6 (IL-6) receptor, including tocilizumab, ameliorate lethal inflammatory responses in COVID-19 patients infected with SARS-CoV-2. Here, by employing single-cell analysis of the immune cell composition of two severe-stage COVID-19 patients prior to and following tocilizumab-induced remission, we identify a monocyte subpopulation that contributes to the inflammatory cytokine storms. Furthermore, although tocilizumab treatment attenuates the inflammation, immune cells, including plasma B cells and $CD8^+$ T cells, still exhibit robust humoral and cellular antiviral immune responses. Thus, in addition to providing a high-dimensional dataset on the immune cell distribution at multiple stages of the COVID-19, our work also provides insights into the therapeutic effects of tocilizumab, and identifies potential target cell populations for treating COVID-19-related cytokine storms.

关键词：Coronavirus disease 2019 (COVID-19); Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); Tocilizumab; Single-cell RNA sequencing (scRNA-seq); Inflammatory storm

报告人 Email：qukun@ustc.edu.cn

Compositional Diversity and Evolutionary Pattern of Genomes of SARS-CoV-2 and Related Coronaviruses

Aiping Wu[1](吴爱平)

[1] *Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, Suzhou Institute of Systems Medicine, Suzhou, Jiangsu 215123, China*

摘要： The 2019 novel coronavirus (SARS-CoV-2) has spread more rapidly than any other betacoronavirus including SARS-CoV and MERS-CoV. However, the mechanisms responsible for the infection and molecular evolution of this virus remained unclear. Here, the compositional diversity and evolutionary pattern of genomes of SARS-CoV-2 and its related coronaviruses had been investigated to promote the understanding of viral adaption and evolution of SARS-CoV-2.

Firstly, an in-depth annotation of SARS-CoV-2 genome has revealed the differences between SARS-CoV-2 and SARS or SARS-like coronaviruses[1]. A systematic comparison identified 380 amino acid substitutions between these coronaviruses, which may have caused functional and pathogenic divergence of SARS-CoV-2.

Then, we collected and analyzed 305 genomic sequences of SARS-CoV-2 including 11 novel genomes from patients in China at the early-stage of outbreak. Through comprehensive analysis of the available genome sequences of SARS-CoV-2 strains, we have tracked multiple inheritable SNPs and determined the evolution of SARS-CoV-2 relative to other coronaviruses. Systematic analysis of 305 genomic sequences of SARS-CoV-2 revealed co-circulation of two genetic subgroups with distinct SNPs markers, which can be used to trace the SARS-CoV-2 spreading pathways to different regions and countries. Although SARS-CoV-2, human and bat SARS-CoV share high homologous in overall genome structures, they evolved into two distinct groups with different receptor entry specificities through potential recombination in the receptor binding regions. In addition, SARS-CoV-2 has a unique four amino acid insertion between S1 and S2 domains of the spike protein, which created a potential furin or TMPRSS2 cleavage site.

Furthermore, we had carried out a comprehensive study for the compositional diversity and evolutionary patterns of accessory proteins to understand the host adaptation and epidemic variation of all coronaviruses. We developed a standardized genome annotation tool for coronavirus named CoroAnnoter, by combining open reading frame (ORF) prediction, transcription regulatory sequence (TRS) recognition and homologous alignment. With CoroAnnoter, we annotated 39 representatively reference coronavirus strains to form a compositional profile for all accessary proteins. Huge variations were observed in the number of accessory proteins from 1 to 10 for different coronaviruses, in which SARS-CoV-2 and SARS-CoV have the most ones as of 9 and 10. The genomic distribution of accessory proteins have significant intra-genus conservation and inter-genus diversity, which could be grouped into 1, 4, 2 and 1 types for α-, β-, γ-, and δ-coronaviruses, respectively. Evolutionary analysis suggested that accessory proteins are more conservative in front of the proteins E and M (E-M), while more diverse behind the E-M proteins. Furthermore,

comparison of virus-host interaction networks of SARS-CoV-2 and SARS-CoV accessory proteins showed that they share multiple antiviral signaling pathways including apoptotic process, viral life cycle and response to oxidative stress. Our study not only provides a tool for coronavirus genome annotation, but also builds a comprehensive profile for coronavirus accessory proteins covering their composition, classification, evolutionary pattern and host interaction.

In summary, our studies provided a comprehensive insight into the diversity, evolution and spread of the SARS-CoV-2.

关键词：SARS-CoV-2, Coronavirus, Diversity, Evolution, Mutation

报告人 E-mail: wap@ism.cams.cn

Virus Knowledge Mining from Literatures to Identify the Key Factors in the Virus and Host Cell Interaction

Zeyu Zhang[1](张泽宇), Ying Wang[1,2], Xiaoyan Zhang[1]

[1] *Research Center for Translational Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China*

[2] *Department of Laboratory Medicine, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai, 200438, China*

摘要：*Aim*：COVID-19 is a highly infectious and highly pathogenic disease caused by the novel coronavirus (SARS-CoV-2), which seriously endangers human public health. At present, the pathogenic mechanism of COVID-19 is not very clear, and its diagnosis and treatment are difficult. With the accumulation of virus studies, the explosive growth of the literature has provided abundant resources for the construction of virus knowledge graphs. It is urgent to construct virus knowledge mining from literatures for identifying key factors in the virus and host cell interaction and drugs which can help early diagnosis and treatment.

Methods：Natural language processing and knowledge representation were applied to construct a virus knowledge graph and systematically analyze the key immune factors in the process of virus and cell interaction. We collected six named entity recognition datasets as our corpus, including 19,575 disease data, 95,253 drug data, and 56,163 gene data. We combined datasets mentioned above into one training set and validation set for each entity, respectively. We applied BioBERT to train the NER models on these datasets. Through multiple downstream training and fine-tuning, we generated multiple models and tested the performance on the validation set, and selected the model that performs best on BioBERT-Base and BioBERT-Large.

Results：We have established a comprehensive database of human disease-related virus mutations, integrations and cis-effects. Moreover, by selecting the best performing model (0.8479 f1-score for disease, 0.8899 f1-score for drug and 0.7153 f1-score for gene) for the three entities based on pretrained BioBERT model, we randomly select a small sample to test the performance of virus literature migration learning. We then used the named entity recognition and annotation tool Brat to annotate virus-related literature, applied the IOBES strategy to annotate drugs, diseases, biomarkers/targets (genes, proteins, cytokines, etc.) and viruses, set labels for different entities, and completed the design of the annotation strategy for relation extraction. After extracting drugs, virus-related diseases, biomarkers (genes, proteins, cytokines, etc.), and defining relationships among these entities, it is feasible to draw a virus knowledge map based on the relationship matrix.

Conclusion：Our established database, models and platforms will contribute to identify the key factors in the virus and host cell interaction and screening the COVID-19 path in the viral knowledge graph. The research provides a knowledge and technical platform which may contribute to the clinical diagnosis and treatment of COVID-19.

关键词：COVID-19, virus, knowledge mining, natural language processing, viral knowledge graph, virus-host interaction

报告人 Email: zhzyvv@gmail.com

The Support of Genetic Evidence for Cardiovascular Risk Induced by Antineoplastic Drugs

Hui Cui[1,2,3], Shengkai Zuo[3], Zipeng Liu[4], Huanhuan Liu[3], Jianhua Wang[3], Tianyi You[3], Zhanye Zheng[3], Yao Zhou[3], Xinyi Qian[3], Hongcheng Yao[4], Lu Xie[5], Tong Liu[6], Pak Chung Sham[4], Ying Yu[1,2,3], Mulin Jun Li[3,7](李俊)

[1] School of Life Science and Technology, ShanghaiTech University, Shanghai, China.

[2] Key Laboratory of Food Safety Research, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute for Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China.

[3] Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China.

[4] Centre of Genomics Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

[5] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai, China

[6] Tianjin Key Laboratory of Ionic-Molecular Function of Cardiovascular Disease, Department of Cardiology, Tianjin Institute of Cardiology, Second Hospital of Tianjin Medical University, Tianjin, China.

[7] Department of Epidemiology and Biostatistics, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China.

摘要：Cardiovascular dysfunction is one of the most common complications of long-term cancer treatment. Growing evidence has shown that antineoplastic drugs can increase cardiovascular risk during cancer therapy, seriously affecting patient survival. However, little is known about the genetic factors associated with the cardiovascular risk of antineoplastic drugs. For the first time, we established a compendium of genetic evidence that supports cardiovascular risk induced by antineoplastic drugs. Most of this genetic evidence is attributed to causal alleles altering the expression of cardiovascular disease genes. We found that antineoplastic drugs predicted to induce cardiovascular risk are significantly enriched in drugs associated with cardiovascular adverse reactions, including many first-line cancer treatments. Functional experiments validated that retinoid X receptor agonists can reduce triglyceride lipolysis, thus modulating cardiovascular risk. Our results establish a link between the causal allele of cardiovascular disease genes and the direction of pharmacological modulation, which could facilitate cancer drug discovery and clinical trial design.

关键词：genetic variants, anti-tumor drug, cardiovascular dysfunction, GWAS, drug side effect

报告人 Email: mulinli@connect.hku.hk

机器智能赋能新药研发

曾坚阳

清华大学

摘要：近年来，大规模基因组学、生物化学和药理学数据不断累积，为机器智能技术在新药研发和药物重定位中的应用提供了新的契机。在本次报告中，我将围绕针对新冠肺炎的老药新用，介绍本课题组最近在 AI+药物发现领域的研究进展。我们团队开发了一套系统性药物重定位框架，整合机器学习和统计分析等方法，集成并挖掘大规模知识图谱、文献和转录组数据，从老药中寻找治疗新冠肺炎的潜在候选药物。目前相关预测结果已经成功获得细胞水平的实验验证。

报告人 Email：zengjy321@tsinghua.edu.cn

Protein-protein Contact Prediction for Integrative Protein Docking

Yumeng Yan[1], Sheng-You Huang[1](黄胜友)

[1] *Institute of Biophysics, School of Physics, Huazhong University of Science and Technology, Wuhan 430074, China*

摘要：Protein-protein interactions play a fundamental role in all cellular processes. Therefore, determining the structure of protein-protein complexes is crucial to understand their molecular mechanisms and develop drugs targeting the protein-protein interactions. Recently, deep learning has led to a revolutionary advancement in protein contact prediction, giving an unusual high accuracy in protein structure prediction. However, due to the limited number of homologous protein-protein complexes in the protein data bank (PDB) and the pairing challenge of multiple sequence alignments (MSA), the accuracy for inter-protein contact prediction is still low, compared to that for intra-protein contact prediction. Given that homo-oligomer proteins constitute more than one-third of protein complexes, we have proposed a deep learning framework to predict inter-protein residue-residue contacts across homo-oligomeric protein interfaces by integrating evolutionary coupling, sequence conservation, distance map, and physic-chemical information of monomers, named as DeepHomo. Our DeepHomo was extensively tested on 300 diverse homo-oligomer complexes from the PDB and 28 realistic targets from recent CASP-CAPRI challenges, and compared with state-of-the-art direct-coupling analysis (DCA) and machine learning (ML)-based approaches. It was shown that DeepHomo far outperformed the existing DCA and ML-based methods and achieved a high accuracy of >60% for the top predicted contact. By integrating the predicted contacts, our HSYMDOCK docking algorithm obtained correct complex structures for 67.9% of 28 realistic homo-oligomer proteins from CASP-CAPRI challenges within top five predictions, compared to 42,9% for *ab initio* docking. These results demonstrated the high accuracy of DeepHomo for inter-protein contact prediction and its important role in protein complex structure prediction.

关键词：Protein-protein interaction, Deep learning, Residue-residue contact, Homo-oligomer, Direct coupling analysis

报告人 Email: huangsy@hust.edu.cn

大脑网络结构中的最大熵原理

李松挺

上海交通大学

摘要：大脑网络通常具有大量的短程连接和少量的长程连接，且连接结构较为错综复杂。实验测量发现，不同生物的脑网络连接长度分布有较大差异，但其背后的原理并不清楚。我们通过分析发现，从无脊椎动物到脊椎动物等多种生物的脑网络连接长度分布均符合在几何空间约束和生物材料约束前提下的最大熵原理，并进一步提出最大熵原理优化的生物学实现过程以及相应的脑网络结构生成模型，可较为准确地重构不同物种脑网络结构的统计特性。该工作或表明大脑网络通过演化使其具有高熵的结构多样性特点，以支持其高效的信息处理功能。

报告人 Email：songting@sjtu.edu.cn

解释遗传变异的调控网络建模

Yong Wang [1](王勇)

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing100190, China*

摘要：Interpreting genetic variant (including SNP and structural variant) is the key to precision health. Most of these variants will affect disease risk, response to drugs or other traits such as height in a tissue or condition specific way. How can we figure out which variants affect the function and regulation of genes in which condition? We propose to use gene regulatory network to integrating omics data and interpret genetic variants. Particularly, we will discuss the models and algorithms to organize, analyze, model, and integrate the genetic variant, DNA accessibility data, transcriptional data, and functional genomic regions together. We believe that the integrative paradigm on chromatin and expression levels will eventually help us to understand the information flow in cell and will influence research directions across many fields.

关键词：Gene regulatory network, Genetic variant, Model, Data integration

报告人 E-mail: ywang@amss.ac.cn

Modelling and Analysis of Non-Markovian Biochemical Reaction Networks

张家军

中山大学

摘要： Modeling intracellular processes has long relied on the Markovian assumption. However, as soon as a reactant interacts with its environment, molecular memory definitely exists and its effects cannot be neglected. Since the Markov theory cannot translate directly to modeling and analysis of non-Markovian processes, this leads to many significant challenges. We develop a formulation, namely the stationary generalized chemical-master equation, to model intracellular processes with molecular memory. This formulation converts a non-Markovian question to a Markovian one while keeping the stationary probabilistic behavior unchanged. Both a stationary generalized Fokker–Planck equation and a generalized linear noise approximation are further developed for the fast evaluation of fluctuations. These formulations can have broad applications and may help us discover new biological knowledge.

报告人 Email： zhjiajun@mail.sysu.edu.cn

DiverRWH: Discovering Personalized Cancer Driver Genes by Hypergraph Model

Chenye Wang[1], Junhan Shi[1], Yusen Zhang[1], Naiqian Zhang[1](张乃千)

[1] *Shandong University at Weihai*

摘要：Cancer is a complex genetic disease primarily caused by the accumulation of genetic alteration. Identifying molecular cancer drivers is critical for precision oncology. Although rapid progress has been made in computational approaches for prioritizing cancer driver, current researches mainly focus on cohort-lever driver gene identification. It is acknowledged that even individual tumor samples belonged to the same type exhibit extensive mutation heterogeneity and have diverse genomic alterations. Most of the existing methods lack the ability of discovering patient-specify driver genes. We present a hypergraph model which integrates somatic mutation data and molecular interaction data effectively. By ingenious design, the hypergraph model can store the mutation information in samples accurately and capture the relationship structure inherent in them. After that, we develop random walk algorithm in hypergraph called DiverRWH to identify cancer driver genes. By setting different initial values, DiverRWH can not only predict cohort-lever cancer driver mutations, but more importantly it can also identify personalized driver mutation profiles. Applications to TCGA datasets demonstrated that DiverRWH significantly outperforms the existing state-of-art methods in terms of predictive accuracy, sensitivity, and specificity. In addition, DiverRWH is also highly robust to data perturbation. We believe DiverRWH complements existing driver gene identification methods which would inform potential personalized therapies targeted against the products of these aberrant genomic alterations.

关键词： cancer, driver gene, hypergraph model, random walk

报告人 E-mail: nqzhang@email.sdu.edu.cn

A DNA Methylation State Transition Model Reveals the Programmed Epigenetic Heterogeneity in Human Pre-Implantation Embryos

Chengchen Zhao[1](赵程辰), Naiqian Zhang[2], Yalin Zhang[1,3,$], Nuermaimaiti Tuersunjiang[1], Shaorong Gao[1,3], Wenqiang Liu[1,3], Yong Zhang[1]

[1] Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, Frontier Science Center for Stem Cell Research, School of Life Science and Technology, Tongji University, Shanghai 200092, China

[2] School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China

[3] Clinical and Translational Research Center of Shanghai First Maternity and Infant Hospital, Tongji University, Shanghai 200092, China

摘要：Expression and epigenetic heterogeneity emerge before the first cell fate determination during mammalian early embryogenesis, but the programs causing such determinate heterogeneity are largely unexplored. Here, we present MethylTransition, a novel DNA methylation state transition model, for characterizing methylation changes during one or a few cell cycles at single-cell resolution. MethylTransition involves the creation of a transition matrix comprising by 3 parameters that represent the probabilities of DNA methylation-modifying activities in order to link the methylation states before and after a cell cycle. We applied MethylTransition to single-cell DNA methylome data during human pre-implantation embryogenesis and elucidated that the DNA methylation heterogeneity that emerges at promoters during this process is largely an intrinsic output of a program with unique probabilities of DNA methylation-modifying activities. Moreover, we experimentally validated the effect of initial DNA methylation on expression heterogeneity in pre-implantation mouse embryos. Our study reveals the programmed DNA methylation heterogeneity during human pre-implantation embryogenesis though a novel mathematic model and provides valuable clues for identifying driving factors of the first cell fate determination during this process.

关键词：DNA methylation, heterogeneity, first cell fate determination

报告人 Email: cczhao@tongji.edu.cn

Robustness and Lethality in Multilayer Biological Molecular Networks

Xueming Liu[1](刘雪明), Enrico Maiorino[2], Arda Halu[2], Kimberly Glass[2], Rashmi B. Prasad[3], Joseph Loscalzo[2], Jianxi Gao[4], Amitabh Sharma[2]

[1] *Key Laboratory of Imaging Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China*

[2] *Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*

[3] *Genomics Diabetes and Endocrinology, Lund University Diabetes Centre, CRC, Malmo 21429*

[4] *Department of Computer Science, Rensselaer Polytechnic Institute,Troy, New York, 12180, USA*

摘要：Robustness is a prominent feature of most biological systems. In a cell, the structure of the interactions between genes, proteins, and metabolites has a crucial role in maintaining the cell's functionality and viability in the presence of external perturbations and noise. Despite advances in characterizing the robustness of biological systems, most of the current efforts have been focused on studying homogeneous molecular networks in isolation, such as protein-protein or gene regulatory networks, neglecting the interactions among different molecular species. Here we propose a comprehensive framework for understanding how the interactions between genes, proteins, and metabolites contribute to the determinants of robustness in a heterogeneous biological network. We integrate heterogeneous sources of data to construct a multilayer interaction network composed of a gene regulatory layer, a protein-protein interaction layer, and a metabolic layer. We design a simulated perturbation process to characterize the contribution of each gene to the overall system's robustness, defined as its influence over the global network. We find that highly influential genes are enriched in essential and cancer genes, confirming the central role of these genes in critical cellular processes. Furthermore, we show that the proposed mechanism predicts a higher vulnerability of the metabolic layer to perturbations applied to genes associated with metabolic diseases. By comparing the robustness of the network to multiple randomized network models, we find that the real network is comparably or more robust than expected in the random realizations. Finally, we analytically derive the expected robustness of multilayer biological networks starting from the degree distributions within or between layers. These results provide new insights into the non-trivial dynamics occurring in the cell after a genetic perturbation is applied, confirming the importance of including the coupling between different layers of interaction in models of complex biological systems.

关键词：Multilayer networks, robustness, gene regulatory network, protein interaction network, metabolic network

报告人 E-mail: xm_liu@hust.edu.cn

FALCON-ProFOLD: 基于人工智能技术的蛋白质结构“从头预测”算法

卜东波

摘要： Protein functions rely largely on the final details of their tertiary structures, and the structures could be accurately reconstructed using inter-residue distances. Residue co-evolution has become the primary principle for estimating inter-residue distances since the residues in close proximity tend to co-mutate during protein evolutionary history. Widely-used approaches infer the residue co-evolution using an indirect strategy, i.e., they extract from the multiple sequence alignment (MSA) of query protein some handcrafted features, say, co-variance matrix, rather than using the raw information carried by MSA. This indirect strategy, however, will transform two different MSAs into the same co-variance matrix, thus leading to considerable information loss at the very beginning. Here, we report a deep neu- ral network framework (called CopulaNet) to learn residue co-evolution directly from MSA without any handcrafted features. The CopulaNet consists of two key elements: an encoder to model structural-context specific mutation for each residue, and an aggregator to aggregate these embeddings for modeling copula among residues and thereafter infer residue co-mutations. Using the CASP13 (the 13th Critical Assessment of Protein Structure Prediction) target proteins as representatives, we demonstrated the successful application of CopulaNet for estimating inter-residue distances and thereafter reconstructing protein tertiary structure with improved accuracy and efficiency. On the 31 free modeling target proteins, our approach generated high-quality structures with an average TM-score of 0.658, and thus outperformed the state-of-the-art approaches AlphaFold (0.582) and trRosetta (0.580).

报告人 Email： dbu@ict.ac.cn

Predicting the Real-valued Inter-residue Distances for Proteins

Wenze Ding and Haipeng Gong(龚海鹏)

MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China.

摘要： Predicting protein structure from the amino acid sequence has been a challenge with theoretical and practical significance in biophysics. Despite the recent progresses elicited by improved inter-residue contact prediction, contact-based structure prediction has gradually reached the performance ceiling. New methods have been proposed to predict the inter-residue distance, but unanimously by simplifying the real-valued distance prediction into a multiclass classification problem. Here we show a lightweight regression-based distance prediction method, which adopts the generative adversarial network to capture the delicate geometric relationship between residue pairs and thus could predict the continuous, real-valued inter-residue distance rapidly and satisfactorily. The predicted residue distance map allows quick structure modeling by the CNS suite, and the constructed models approach the same level of quality as the other state-of-the-art protein structure prediction methods when tested on CASP13 targets. Moreover, this method can be used directly for the structure prediction of membrane proteins without transfer learning.

关键词： protein inter-residue distance, deep learning, real-valued distance prediction, protein structure prediction, generative adversarial network.

报告人 Email: hgong@tsinghua.edu.cn

固有无序蛋白的模糊相互作用：在分子模拟的显微镜下

王文宁

复旦大学化学系，上海市淞沪路 2205 号，200438

摘要：Protein interactions involving intrinsically disordered proteins (IDPs) greatly extend the range of binding mechanisms available to proteins. In some cases, known as coupled folding and binding model, IDPs undergo disorder-to-order transitions to form a protein complex with a well-defined structure. In many other cases, IDPs retain structural plasticity in the final complexes, which have been defined as the fuzzy complexes. While a large number of fuzzy complexes have been characterized with variety of fuzzy patterns, many of the interactions are between an IDP and a structured protein. Thus, whether two IDPs can interact directly to form a fuzzy complex without disorder-to-order transition remains an open question. In this study, MD simulation, NMR, and smFRET are combined to characterize the interaction between two IDPs, the C-terminal domain (CTD) of protein 4.1G and the nuclear mitotic apparatus (NuMA) protein. It is revealed that CTD and NuMA form a fuzzy complex with remaining structural disorder. Multiple binding sites on both proteins were identified by molecular dynamics and mutagenesis studies. This study provides an atomic scenario in which two IDPs bearing multiple binding sites interact with each other in dynamic equilibrium, which we name as the extreme fuzziness. Extreme fuzziness completes the full spectrum of protein-protein interaction modes, suggesting that a more generalized model beyond existing binding mechanisms is required.

报告人 Email：wnwang@fudan.edu.cn

Push to Open: The Gating Mechanism of the Tethered Mechanosensitive Ion Channel NompC

Chen Song(宋晨)

Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

摘要：NompC is one of the earliest identified mechanosensitive ion channels responsible for the sensation of touch and balance in Drosophila melanogaster, and serves as an ideal model for studying tethered mechanosensitive ion channels. We performed extensive molecular dynamics simulations on a recently resolved Cryo-EM NompC structure, and observed the atomistic and dynamic details of the channel gating under the compression of the intracellular domain. When pushing the intracellular domain toward the membrane, the ankyrin repeat region is compressed and passes the mechanical force to the linker helices like a spring, the force constant of which is ~3.3 pN/nm. The TRP domain undergoes a clockwise rotation (intracellular view) and a slight tilt that leads to the opening of the channel. We identified the key residues responsible for the force transfer, supported by electrophysiology experiments. We think this push-to-open mechanism might be a universal gating mechanism of similar tethered mechanosensitive ion channels, enabling cells to feel and respond to compression or shrinking.

报告人 Email：c.song@pku.edu.cn

Development and Application of Enhanced Sampling Method of Biomolecular Conformations Based on Feature Space of Structures

Hao Dong(董昊)

Kuang Yaming Honors School, Nanjing University, Nanjing 210023, P. R. China

摘要：Conformational change of proteins, especially the transition between functional states, is generally associated with their biological processes. However, conformational transitions in proteins are usually not easy to be well characterized by experimental protocols, mainly because of their inadequate temporal and spatial resolution. Recently, we proposed an enhanced conformational sampling protocol within the framework of the feature space of protein structures, and developed a DAta-Driven Accelerated (DA2) sampling method and a two-ended DA2 (teDA2) sampling method: DA2 was designed to search new functional states of a biomolecule from a known structure, and teDA2 was designed to identify the possible paths between two available states of a biomolecule. Both methods drive the conformational change of biomolecules in an adaptively updated feature space of biomolecular structures without introducing bias. DA2 was used to predict a closed conformation of N-terminal calmodulin (nCaM) from the open one. The identified structure resembles the crystal structure of nCaM in its closed state, with a root-mean-square deviation between the two of only 1.8 Å. The teDA2 was applied to explore the conformational transition of adenylate kinase (ADK), a model system with well-characterized open and closed state structures. Our calculations identified three different mechanisms and the associated multiple pathways for domain motion of ADK. As a reliable and efficient enhanced conformational sampling protocol, DA2 and teDA2 could be employed to study the dynamics between different functional states of a broad spectrum of proteins and biomolecular machines.

报告人Email: donghao@nju.edu.cn

Accurate and fast *de novo* protein structure prediction through deep learning and Rosetta

Jianyi Yang(杨建益)

School of Mathematical Sciences, Nankai University, Tianjin, 300071.

摘要：The prediction of interresidue contacts and distances from co-evolutionary data using deep learning has considerably advanced protein structure prediction. Here we build on these advances by developing a deep residual network for predicting inter-residue orientations in addition to distances, and a Rosetta constrained energy minimization protocol for rapidly and accurately generating structure models guided by these restraints. In benchmark tests on CASP13 and CAMEO derived sets, the method outperforms all previously described structure prediction methods. Although trained entirely on native proteins, the network consistently assigns higher probability to *de novo* designed proteins, identifying the key fold determining residues and providing an independent quantitative measure of the "ideality" of a protein structure. The method promises to be useful for a broad range of protein structure prediction and design problems. A web server and a standalone package for the proposed method are freely available at: https://yanglab.nankai.edu.cn/trRosetta/.

关键词：Protein structure prediction, deep learning, Rosetta

报告人Email: yangjy@nankai.edu.cn

H-RACS: a Handy Tool to Rank Anti-cancer Synergistic Drugs

Xinmiao Yan [1], Yiyan Yang [1], Zikun Chen [1], Zuojing Yin [1], Zeliang Deng [1], Tianyi Qiu [2], Kailin Tang [1], Zhiwei Cao [1](曹志伟)

[1]*Department of Gastroenterology, Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China*
[2]*Shanghai Public Health Clinical Center, Fudan University, Shanghai 200032, China*

摘要：Though promising, identifying synergistic combinations from a large pool of candidate drugs remains challenging for cancer treatment. Due to unclear mechanism and limited confirmed cases, only a few computational algorithms are able to predict drug synergy. Yet they normally require the drug-cell treatment results as an essential input, thus exclude the possibility to pre-screen those unexplored drugs without cell treatment profiling. Based on the largest dataset of 33,574 combinational scenarios, we proposed a handy webserver, H-RACS, to overcome the above problems. Being loaded with chemical structures and target information, H-RACS can recommend potential synergistic pairs between candidate drugs on 928 cell lines of 24 prevalent cancer types. A high model performance was achieved with AUC of 0.89 on independent combinational scenarios. On the second independent validation of DREAM dataset, H-RACS obtained precision of 67% among its top 5% ranking list. When being tested on new combinations and new cell lines, H-RACS showed strong extendibility with AUC of 0.84 and 0.81 respectively. As the first online server freely accessible at http://www.badd-cao.net/h-racs, H-RACS may promote the pre-screening of synergistic combinations for new chemical drugs on unexplored cancers.

关键词：anti-cancer; synergistic combination; drug synergy; web server; bioinformatics

报告人 Email: zwcao@tongji.edu.cn

Intrinsically Disordered Protein Specific Force Field CHARMM36IDPSFF

Hao Liu[1], Dong Song[1], and Hai-Feng Chen[1](陈海峰)

[1] *State Key Laboratory of Microbial Metabolism, Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, 200240, China*

摘要： Intrinsically disordered proteins (IDPs) are closely related to various human diseases. Because IDPs lack certain tertiary structure, it is difficult to use X-ray and NMR methods to measure their structures. Therefore, molecular dynamics simulation is a useful tool to study the conformer distribution of IDPs. However, most generic protein force fields were found to be insufficient in simulations of IDPs. Here we report our development for the CHARMM community. Our residue-specific IDP force field (CHARMM36IDPSFF) was developed based on the base generic force field with CMAP corrections of for all 20 naturally occurring amino acids. Multiple and extended tests show that the simulated chemical shifts with the newly developed force field are in quantitative agreement with NMR experiment and are more accurate than the base generic force field of Charmm36m. Comparison of J-couplings with previous work shows that CHARMM36IDPSFF and its corresponding base generic force field have their own advantages. In addition, CHARMM36IDPSFF simulations also agree with experiment for SAXS profiles and radii of gyration of IDPs. Detailed analysis shows that CHARMM36IDPSFF can sample more diverse and disordered conformers. These findings confirm that the newly developed force field can improve the balance of accuracy and efficiency for the conformer sampling of IDPs. This newly force field can be used to explore the sequence-disorder-function paradigm of IDPs.

报告人 Email： haifengchen@sjtu.edu.cn

三维基因组与疾病

李程

北京大学

摘要：近年来，三维基因组学技术快速发展并广泛应用于生物医学问题，从三维空间和时间尺度上研究细胞核内染色质高级结构的组织、功能和动态。本报告将介绍三维基因组学实验、分析和在生物学问题中的应用。1.基于对癌症基因组中变异出现的原因和后果的研究兴趣，通过 Hi-C 实验和分析流程，研究多发性骨髓瘤细胞中非整倍体变异对三维基因组和表达谱的影响。2.开发数据建模新算法，预测转录因子依赖于染色质三维结构的空间分布规律。3.结合成像实验，研究 lamin 缺失对三维基因组结构的影响。

报告人 Email：cheng_li@pku.edu.cn

Mapping the functional m6Am methylome in the mammalian transcriptome

Hanxiao Sun[1], Meiling Zhang[1], June Liu1, Kai Li[2], Chengqi Yi[1,3,4](伊成器)

[1] *State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China*

[2] *Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

[3] *Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China*

[4] *Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

摘要：Reversible and dynamic RNA methylation has greatly changed our views of gene expression regulation. In addition to the well-studied m6A modification, there exists another reversible epitranscriptomic mark commonly found adjacent to the mRNA cap, called m6Am. However, despite its initial discovery over 40 years ago, the functional consequences and molecular mechanisms of m6Am remain largely unexplored. Combining chemical biology, cell biology, high-throughput sequencing and genetic approaches, we aim to understand m6Am-mediated regulation of gene expression. In our previous studies, we have spearheaded the discovery of several novel chemical modifications in mRNA, revealed their transcriptome-wide distribution pattern, and characterized multiple disease-related modification enzymes. Building on our expertise in the field of epitranscriptomics, my group is one of the first two labs to identify the long-sought m6Am methyltransferase last year. Our future work will identify novel modification enzymes and effector proteins of m6Am, elucidate its roles in physiological and pathological conditions, and screen for small molecule inhibitors to manipulate m6Am in vivo. This cutting-edge and comprehensive research will open up a new direction in the field of RNA biology.

关键词：m6Am, RNA methylation, high-throughput sequencing, Mapping, Epitranscriptomics

报告人 Email: chengqi.yi@pku.edu.cn

三维基因组 CTCF 拓扑绝缘子作用机制研究

吴强

上海交通大学

摘要：CTCF is a key insulator-binding protein, and mammalian genomes contain numerous CTCF sites, many of which are organized in tandem. Using CRISPR DNA-fragment editing, in conjunction with chromosome conformation capture, we find that CTCF sites, if located between enhancers and promoters in the protocadherin (*Pcdh*) and *β-globin* clusters, function as an enhancer-blocking chromatin insulator by forming distinct directional chromatin loops, regardless whether enhancers contain CTCF sites or not. Moreover, computational simulation in silico and genetic deletions in vivo as well as dCas9 blocking in vitro revealed balanced promoter usage in cell populations and stochastic monoallelic expression in single cells by large arrays of tandem CTCF sites in the *Pcdh* and immunoglobulin heavy chain (*Igh*) clusters. Furthermore, topological CTCF insulators promote, counter-intuitively, long-range chromatin interactions with distal directional CTCF sites, consistent with the cohesin “loop extrusion” model. Finally, gene expression levels are negatively correlated with CTCF chromatin insulators located between enhancers and promoters on a genome-wide scale. Thus, single CTCF insulators ensure proper enhancer insulation and promoter activation while tandem CTCF topological insulators determine balanced spatial contacts and promoter choice. These findings have interesting implications on the role of topological chromatin insulators in 3D genome folding and developmental gene regulation.

报告人 Email：qwu123@gmail.com

人类衰老速率的异质性

Jing-Dong J. Han(韩敬东)

Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, 200031, China.

摘要： We have always been interested in finding quantitative aging biomarkers to accurately assess the aging status by focusing on epigenetic changes (Cheng et al., 2018; Han et al., 2012; Jin et al., 2011). Recently by analyzing the 3D facial images, we generated the first comprehensive mapping of the aging human facial phenome. We found quantitative facial features, such as eye slopes, highly associated with age. We constructed a robust age predictor and found that on average people of the same chronological age differ by +/-6 years in facial age, with the deviations increasing after age 40. The predictor is as accurate as the most accurate to-date physiological age predictor – the one based on blood cell DNA methylation sites. Using this predictor we identified slow- and fast-agers that are significantly supported by health indicators (Chen et al., 2015). We further profiled blood cell mRNA and lncRNA expression by RNA-seq of this cohort and computationally predict their regulatory networks and their contributions to the variation in aging rate among different individuals, and those that are modifiable by their lifestyles. By extending the study to a large Northern Chinese cohort of 10,000 people we can now use deep learning AI approaches to precisely estimate aging status based on 3D facial images and their associations with individuals' health and medical history.

报告人 Email： jackie.han@pku.edu.cn

非编码区遗传变异功能解析的多组学方法

Xin Li[1](李昕), Jiangxue Li [1], Danyue Dong [1], Zhenguo Wang [1], Zengming Wang [1], Liandong Zhang [1]

[1] *Shanghai Institute of Nutrition and Health, CAS-MPG Partner Institute for Computational Biology*

摘要：Most verified disease-causing mutations have been limited to protein coding regions, with the function of non-coding variation remains largely unknown. The efficient interpretation of non-coding variants among human populations holds a substantial potential for furthering our understanding of disease etiology and finding new therapeutic targets. However, investigation of noncoding variation has been significantly barred by the lack of explicit genetic code, the requirement of extra-large sample sizes and multi omics analysis. Therefore, datasets with large sample sizes and multi-omics may provide us with the unique opportunity to unveil a comprehensive landscape of the functional impact of rare non-coding variants.

We will train a novel machine learning model taking into account genetic factors, which may greatly improve the efficacy of the fine-mapping method derived from extracted sequence and other biological features in the human genome, transcriptome, epigenome, proteome and metabolome that may impact transcription factor binding from sequence annotation. This will help us build a detailed regulatory cascade of rare genetic variants and answer the question how rare genetic variation influence human disease. The model will be trained using this cohort and parameters can be derived for biological interpretation. Utilizing solved and unresolved cases in the cohort, we seek to train the model over a variety of genomic features, aiming at identifying potential rare non-coding variants that are more prone to regulatory modification and pathogenesis. A successful outcome of this study will build a detailed catalog of all types of rare non-coding variants and their regulatory effects, forming the basis for prioritizing rare variants for further phenotypic characterization.

关键词：rare variation, non-coding region, multi-omics, machine learning model, regulatory effects

报告人 Email: lixin@picb.ac.cn

G-quadruplexes May Regulate Gene Transcription by Affecting the Three-dimensional Chromatin Structure

Xiao Sun[1](孙啸), Yue Hou[1,2], Rongxin Zhang[1], Yu Gu[1], Zhaohui Qin[3]

[1] *State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*

[2] *Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*

[3] *Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA USA*

摘要：G-quadruplex is a kind of special secondary structure of nucleic acid molecules. Several lines of evidence indicate that G-quadruplex structures in genomic DNA have functional effects on telomere maintenance, DNA replication, and especially gene transcriptional regulation. Currently we know less about the underlying mechanism of gene transcription of G-quadruplexes. We believe that G-quadruplexes can play a role in biological processes by affecting the chromatin structure. In this study, we investigated the relationship between G-quadruplexes and the global conformation of the chromatin. We found that G-quadruplexes were significantly enriched at boundaries of topological associated domains (TADs). Architectural protein occupancy, which plays critical roles in the formation of TADs, was highly correlated with the content of G-quadruplexes at TAD boundaries. Moreover, adjacent boundaries containing G-quadruplexes frequently interacted with each other because of the high enrichment of architectural protein binding sites. Furthermore, G-quadruplex motifs on different strands were associated with the orientation of CTCF binding sites, suggesting the potential function for G-quadruplexes in loop extrusion. It has been known that G-quadruplexes in promoters have effect on gene transcriptional regulations. But it is unclear whether G-quadruplexes can regulate gene transcription by affecting the remote interaction between enhancers and promoters. Actually, we found that G-quadruplexes can mediate some enhancer–promoter interaction pairs, and the interaction frequency of these pairs was significantly higher than that of other enhancer–promoter pairs. We also found that the G-quadruplexes on enhancers were related to high expression of elncRNAs, reinforcing the enhancer-promoter interaction and leading to high expression level of target genes. Intriguingly, we also found that more than 99% of G-quadruplexes overlapped with transcriptional factor binding sites, suggesting G-quadruplexes may recruit transcriptional factors and mediate the chromatin interaction.

关键词：G-quadruplex, Chromatin, Topological Associated Domain, Transcription

报告人Email：xsun@seu.edu.cn

Dux-mediated Corrections of Aberrant H3k9ac during 2-cell Genome Activation Optimizes Efficiency of Somatic Cell Nuclear Transfer

Guang Yang[1,2], Linfeng Zhang[1], Wenqiang Liu[1], Shaorong Gao[1], Jiayu Chen[1], Cizhong Jiang[2](江赐忠)

[1] Clinical and Translation Research Center of Shanghai First Maternity & Infant Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[2] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Orthopaedic Department of Tongji Hospital, Tongji University, Shanghai 200065, China

摘要：Differentiated somatic cells can be reprogrammed to totipotent embryos through somatic cell nuclear transfer (SCNT) with low efficiency. The histone deacetylase inhibitor Trichostatin A (TSA) has been found to improve SCNT efficiency, but the underlying mechanism remains undetermined. Here, we examined genome-wide H3K9ac during SCNT embryo development and found that aberrant H3K9ac regions resulted in reduced 2-cell genome activation. TSA treatment largely corrects aberrant acetylation in SCNT embryos with an efficiency that is dictated by the native epigenetic environment. We further identified that the overexpression of *Dux* greatly improves SCNT efficiency by correcting the aberrant H3K9ac signal at its target sites, ensuring appropriate 2-cell genome activation. Intriguingly, the improvement in development mediated by TSA and *Kdm4b* is impeded by *Dux* knockout in SCNT embryos. Together, our study reveals that reprogramming of H3K9ac is important for optimal SCNT efficiency, and for the first time identifies *Dux* as a crucial transcription factor in this process.

关键字：*Dux*, H3K9ac, reprogramming, epigenetic, somatic cell nuclear transfer

报告人 Email：czjiang@tongji.edu.cn

Transcriptome Analysis Reveals a Silica-induced Immune Response, Fibrosis Character and Alternative Splicing Profile in a Silicosis Rat Model

Hongyu Zhao, Da Lv, Zhiyan Jiang, Jingyan Wang, Lu Cai[1,2](蔡禄)

School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, 014010, China.

摘要：Silicosis is a type of pneumoconiosis caused by the inhalation of silica dust. It is characterized by inflammation and fibrosis of the lung. Although many studies have reported that crystalline silica-inhalation into the lung initiates the immune response, activating effector cells and triggering the inflammatory cascade with subsequent elaboration of the extracellular matrix and fibrosis, the mechanism of silicosis pathogenesis remains unclear. We established a silica inhalation-induced silicosis rat model validated by histological and cytokine analyses.

RNA-seq and bioinformatic analyses showed that 600 genes were upregulated and 537 genes were downregulated in the silica-treated group. GO enrichment analysis indicated that the pathological process of lung tissue stressed by silica is mainly involved with immune-associated processes, response to stimulus, organism remodeling-associated process, extracellular matrix remodeling, cell adhesion and migration process, signaling pathway, cell activation, chemotaxis, regulation process, and development process.

53 KEGG pathways were enriched. The first category of enriched pathways is involved in the immune response. Here, we directly identified several immune-associated pathways, such as “B cell receptor signaling pathway”, “natural killer cell mediated cytotoxicity” and “complement and coagulation cascades”. Thus, a strong inflammatory reaction occurred in response to inhalation exposure of silica. Phagocytosis is one of the main immune responses to silica particle exposure. The pathways of “phagosome” and “Fc gamma R-mediated phagocytosis” were significantly enriched. The second category of pathways is associated with the fibrosis process. Along with persistent inflammation, silica triggers a rapid-onset fibrotic response with deposition of extracellular matrix (ECM). We identified 4 significant KEGG pathways associated with pulmonary fibrotic processes: 1) "ECM-receptor interaction": The ECM consists of a complex mixture of structural and functional macromolecules, and plays an important role in the morphogenesis of tissues and organs; 2) another fibrosis-associated pathway: At the cell-extracellular matrix contact points, specialized structures are formed, termed focal adhesions; 3) "cell adhesion molecules (CAMs)": CAMs on the cell surface can mediate the interaction between cells and ECM; 4) "protein digestion and absorption".

Since alternative splicing of pre-mRNAs is also essential for the regulation of gene expression, we identified several alternative pre-mRNA splicing events in the fibrotic process. We identified 22,010 and 34,536 AS events based on Hisat2 and STAR, respectively. In our transcriptome data, except for the 3,472 and 3,489 known AS events, 18,538 and 31,047 AS events, involving 7,702

and 9,173 genes, were identified for the first time using Hisats2 and STAR, respectively. Skipped exon (SE) was the most frequent alternative splicing events, accounting for more than 80%, whereas retained intron (RI), alternative 5′ splice site (A5SS) and alternative 3′ splice site (A3SS) accounted for only approximately 1% of alternative splicing events in each group. Under an FDR cutoff of 0.05 and ΔPSI of 5%, we analyzed the differential alternative splicing events (DASEs) between the silica-treated and saline-treated groups. Based on the mapping results of the Hisat2 method, 33 DASEs, including 30 SEs, 2 MXEs and 1 A3SS, were identified. Total 43 DASEs, including 30 SEs, 12 MXEs and 1 A5SS, were identified using STAR software. Most of these DASEs were novel and had not been annotated in the Ensembl database so far.

This study will provide a foundation to understand the molecular mechanism of the pulmonary fibrosis caused by silica.

关键词： Silicosis, Immune response, Pulmonary fibrosis, Alternative splicing

报告人 Email： nmcailu@163.com

精准医学知识图谱构建

Fan Zhong [1](钟凡), Xing Wang [1], Huijiang Zou [2], Wei Wang [2], Lei Liu [1]

[1] *Department of Systems Biology for Medicine, Shanghai Medical College, Fudan University, 138 Yixueyuan Road, Shanghai 200032, China*

[2] *School of Computer Sciences, Fudan University, 825 Zhangheng Road, Shanghai 201203, China*

摘要： The precision medicine aims at providing personalized therapies and diagnostics. It requires us to obtain most, if not all, of the specific and accurate biomedical information from individual genotyping and profiling data. The prerequisite of that is a comprehensive knowledgebase containing extensive heterogeneous biomedical knowledge. The more promoting attempt is to reorganize the extracted knowledge into a rational structure other than just islets, so as to help us to raise hypotheses or make decisions. Here, we present **P**recision **M**edicine knowledgeb**a**se **a**pplication **p**latform (PMapp), a knowledgebase that supports functions of retrieval by all aspects of biomedical keywords, pathways/networks visualization, biomedical literature mining, analytic workflow for omics-derived gene lists, and especially distinctive knowledge association by knowledge graph reasoning. Compared to many other databases biased towards particular domains, PMapp organizes knowledge of human genes and gene products, variations, small molecules, drugs, interactions, pathways, diseases and phenotypes as heterogeneous biomedical entities with interrelationships among them in a knowledge graph. Using a semantic model called Precision Medicine Ontology (PMO) as the conceptual support for providing, accessing and structuring information, the infrastructure of PMapp expands the traditional indexing and searching schema from keyword-based to knowledge-based. The knowledge in PMapp were not only from over sixty public databases, but also from our self-developed deep learning based biomedical literature mining tool BioIE (Biomedical Information Extraction). In summary, PMapp has integrated over 1.2 million biomedical entities including genes and their products, variations, drugs, small molecules, interactions, pathways, diseases and phenotypes in a unified biomedical ontology framework which standardize these biomedical entities, and about 2.1 million relationships with about 56.9 million properties among them. The knowledgebase is free and available for all users at http://www.pmap.org.cn, and there is no login requirement.

What is worth mentioning, is that the search engine in Precision Medicine Knowledge Graph Generator (PMKGG) module provides the distinctive function of knowledge association. PMKGG can generate three kinds of high-level knowledge graphs that (1) follows some desired logical connections from any genotype or drug entity query; (2) links any two biomedical entities with the shortest paths; (3) outputs the adjacent nodes of a queried biomedical entities and extended paths by stepwise selections. PMKGG acts as a mind map builder to guide thoughts of analysis, and can transmit results to the annotations and pathway/network mapping system seamlessly for further in-depth analyses. The generated framework map will guide scientific research analysis, assist hypothesis generation and decision-making, and provide broader and deeper understanding of

biomedical problems. There are three kinds of inputs in PMKGG: 1) a keyword to a category, such as from the gene “KRAS” or the drug “Stepronin” to the category “disease”; 2) any two keywords such as “Chloroquine” and “pneumonia” with the shortest path; 3) to type a keyword such as “SPP1”, “breast cancer” or “Roxithromycin” to obtain its adjacent nodes. The output of PMKGG is a knowledge graph following some desired logical connection patterns (for the inputs 1 and 2 above), or the adjacent nodes of queried biomedical entities and extended paths by stepwise selections (for the input 3 above).

关键词：precision medicine, knowledge graph, knowledgebase, intelligent query, mind map

报告人 Email: zhongfan@fudan.edu.cn

精准医学本体和语义网络构建与应用

李姣

中国医学科学院医学信息研究所

摘要：本体（Ontology）在机器可理解的知识表示中发挥重要的作用。本研究面向精准医学知识库构建，研发了精准医学本体（PMO, Precision Medicine Ontology），标准化定义了疾病、基因、药物、表型，变异，通路等概念以及概念间的语义关系网络，提供与基因本体 GO，NCBI Gene，表型本体 HPO， 疾病本体 DO 等国际本体的映射，确保与国际相关本体的互操作性。基于精准医学本体，课题组开展了药物基因组学数据整合和药效预测研究，以及跨语种的医学实体对齐技术研发和开放语义关系识别研究。

报告人 Email：li.jiao@imicams.ac.cn

整合生物大数据解析长链非编码调控与功能

Juan Xu[1](徐娟), Dezhong Lv[1], Kang Xu[1], Xinhui Li[1], Xiyun Jin[1], Tingting Shao, Xia Li[1]

[1] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China.*

摘要：Long non-coding RNAs (lncRNAs) control various crucial biological functions to maintain morphology and function of tissues. Their precise regulatory effectiveness is closely associated with spatial expression patterns across tissues, whose dysfunction often influences disease development and progression. Thus, integration of biological data, such as spatial characterization of lncRNA expression and gene regulation data, would improve our understanding of lncRNA functions in complex diseases.

The spatial expression atlas for lncRNAs across different human tissues and adult or pediatric cancer types is still limited. To fill this gap, we first constructed a user-friendly resource LncSpA (LncRNA Spatial Atlas of expression), available at http://bio-bigdata.hrbmu.edu.cn/LncSpA/. LncSpA provides comprehensive information about spatial expression for lncRNAs. Various types of visualization and tables were used to characterize TE lncRNAs, predicted functions based on co-expression mRNAs, diseases association and the potential as diagnostic, or prognostic markers. LncSpA not only facilitates computational investigators to perform integrative analysis of TE lncRNAs in interesting tissues, but also enables experimental scientists to analyze their own data in the context of other related public data.

Moreover, lncRNAs are emerging as critical regulatory elements and play fundamental roles in the biology of various cancers. To identify the potential functional lncRNAs in cancer, we systematically investigated the expression pattern of lncRNAs. Particularly, we found that the expression of *MIR22HG* was significantly decreased in colorectal cancer (CRC), which was mainly driven by copy number deletion. *MIR22HG* functions through regulating the TGFβ pathway via competing interacting with *SMAD2*. Overexpression of MIR22HG can enhance the response of immunotherapy by increasing the CD8 T cell infiltration, suggesting a rational combinational therapy strategy in CRC.

Taken together, all these results suggest that integration of biological data is critical for understanding lncRNA function and to advance identification of lncRNA-based immunotherapy targets.

关键词：LncRNAs, Biological data, Biological function, Regulation, Network biology

报告人 Email: xujuanbiocc@ems.hrbmu.edu.cn

基于组学数据和临床信息的整合分析及其肿瘤标志物挖掘

Yan Zhang[1](张岩), Yihan Wang[2]

[1] *School of Life Science and Technology, Harbin Institute of Technology*

[2] *College of Bioinformatics Science and Technology, Harbin Medical University*

摘要： In breast cancer, highly intratumor epigenetic heterogeneity can lead to drug-resistant, metastasis and poor prognosis of tumors, which increases the complexity in the diagnosis and treatment of cancer. However, most studies are limited to average DNA methylation level of individual CpGs and ignore heterogeneous DNA methylation patterns of cell subpopulations within the tumor. Thus, quantifying the variability in DNA methylation pattern in sequencing reads is valuable for understanding differences within the tumor. To better understand the complex DNA methylation patterns, we focus on epigenetic alleles (epialles) which can reveal the dynamics of DNA methylation status and help to discover the characteristics of different cell subpopulations at the cellular level. Compared with single CpG site, the analysis of epialleles can reveal the dynamics of methylation status and is suitable for small sample size.

Here, we performed RRBS for multiple regions within a tumor from one breast cancer patient to shed light on intratumoral epigenetic heterogeneity. We proposed a method “epialleJS” based on Jensen-Shannon divergence (JSD) to identify differential epialleles between tumor core and tumor periphery (CPDEs) and characterized tumor subpopulations which defined by distinct epiallelic patterns. The epiallelic patterns of tumor core were more disordered, suggesting a higher epigenetic heterogeneity. More than 70% of CPDEs have higher epipolymorphism in tumor core than tumor periphery, and these CPDEs have lower average methylation level in tumor core. We also found the genes with higher epigenetic heterogeneity also had higher transcriptional heterogeneity.

Moreover, we analyzed the function of intratumoral heterogeneous epialleles in breast cancer. Using GREAT software to perform GO and KEGG functional enrichment analysis, we found that the CPDEs were involved in cancer-related functions and pathways such as hypoxic response, angiogenesis, cell cycle, glucose metabolism, Notch signaling pathway and Wnt signaling pathway, which revealed that intratumor heterogeneity was associated with hypoxic processes and inferring that tumor core was hypoxic with respect to tumor periphery. Further verification was performed by integrating 450K methylation and gene expression data in TCGA and gene expression data in GEO. By screening hypoxia related differentially expressed genes and combining the hypoxic related genes collected from literatures, the gene expression profiles of TCGA breast cancer patients were clustered to determine the hypoxic state of the patients. Then the CPDEs were divided into two groups based on the methylation levels of tumor center and tumor periphery. Two types of CPDEs were used to cluster the methylation patterns of breast cancer patients in TCGA, respectively, and the patients were divided into high, medium, and low methylation groups. Then combined the hypoxic status of the patients to explore the relationship between two types of CPDEs and hypoxia. As a result, it was found that the tumor core was more prone to hypoxia, and that only the heterogeneity produced by the CPDEs which had lower methylation level in tumor core than tumor periphery was related to the hypoxic microenvironment within the tumor. Based on the results of

the random forest model, we found that these CPDEs can more accurately predict the hypoxic status of tumors of breast cancer patients in TCGA. At the same time, we performed survival analysis and found that five hypoxia-related DNA methylation markers were significantly associated with progression-free survival in breast cancer patients, including a CpG site cg15190451 in gene *SLC16A5*. Finally, immunohistochemical analysis also confirmed that the expression of *SLC16A5* was associated with clinicopathological characteristics and survival of breast cancer patients.

Collectively, our study systematically analyzed the intratumoral epigenetic pattern discrepancies of cell subpopulations in breast cancer, which is propitious to explain the causes and mechanisms of heterogeneity to provide precise personalized treatment protocols and evaluate disease progression for breast cancer patients.

关键词： DNA methylation, breast cancer, intratumoral heterogeneous, hypoxia, clinical information

报告人 Email: zhangtyo@hit.edu.cn

国家生物信息中心数据资源

Zhang Zhang(章张), Wenming Zhao, Jingfa Xiao, Yongbiao Xue, Yiming Bao

China National Center for Bioinformation and Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

摘要：Genome data are increasing dramatically as the result of new technologies. Often, these data are required to be deposited into international databases such as DDBJ, EBI and NCBI, in order to obtain accession numbers needed for publication. This could be challenging sometimes for researchers in China because of large data size, slow data transfer due to limited international internet bandwidth, and language barrier and technical issues in communication. To alleviate these problems, the BIG Data Center (BIGD, https://bigd.big.ac.cn) was launched in 2016 at Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS). During the past few years, BIGD has grown and expanded considerably and became one of the major global centers. In 2019, the National Genomics Data Center (NGDC) was created based on BIGD. Later in the same year, BIG was given the title of China National Center for Bioinformation (CNCB). CNCB will be built on the well-established NGDC multi-omics databases such as Genome Sequence Archive (GSA), Genome Variation Map (GVM), Genome Warehouse (GWH) and 2019 Novel Coronavirus Resource (2019nCoVR), together with specialized resources from many institutions under CAS and other ministries. CNCB is dedicated to providing freely accessible data repositories and a variety of data resources in support of worldwide research activities.

报告人 Email：zhangzhang@big.ac.cn

Ultrafast and Scalable Variant Annotation and Prioritization with Big functional genomics data

Dandan Huang[1,2](黄丹丹), Xianfu Yi[3], Yao Zhou[1], Hongcheng Yao[4], Hang Xu1[4], Jianhua Wang[1], Shijie Zhang[1]

[1] Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China.

[2] Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China.

[3] School of Biomedical Engineering, Tianjin Medical University, Tianjin, China.

[4] School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

摘要： The advances of large-scale genomics studies have enabled compilation of cell type-specific, genome-wide DNA functional elements at high resolution. With the growing volume of functional annotation data and sequencing variants, existing variant annotation algorithms lack the efficiency and scalability to process big genomic data, particularly when annotating whole genome sequencing variants against a huge database with billions of genomic features. Here, we develop VarNote to rapidly annotate genome-scale variants in large and complex functional annotation resources. Equipped with a novel index system and a parallel random-sweep searching algorithm, VarNote shows substantial performance improvements (two to three orders of magnitude) over existing algorithms at different scales. It supports both region-based and allele-specific annotations, and introduces advanced functions for the flexible extraction of annotations. By integrating massive base-wise and context-dependent annotations in the VarNote framework, we introduce three efficient and accurate pipelines to prioritize the causal regulatory variants for common diseases, Mendelian disorders and cancers.

关键词： variant annotation, big genomics data, algorithms, GWAS, WGS

报告人 Email: 116327382@qq.com

BIGSearch: a Cross-database Search System for Multi-dimensional Biological Big Data.

Dong Zou[1,2,3](邹东), Zhang Zhang [1,2,3]

[1] *China National Center for Bioinformation, Beijing 100101, China*
[2] *National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

摘要：With the rapid development of high-throughput sequencing technologies and the dramatic decrease of sequencing cost, biological omics data are generated at increasingly explosive rates, accordingly posing considerable challenges in biological big data sharing and search. Database serves as an important way in supporting data openness, sharing, query, and retrieval. Globally, a number of new databases are developed every year. However, such valuable big data are distributed among databases yet without standardized data index, thus making it very difficult to achieve efficient big data search, not mention to search in a cross-database manner. Towards this end, here we present BIG Search, a distributed and scalable full-text search system. It features cross-domain search and facilitates users to gain access to a wide range of biological data almost in real-time. This search system will provide one-stop cross-database search services for worldwide researchers, with the aim to promote big-data-driven scientific research and innovative development.

关键词：biological big data, search engine, cross-database

报告人 Email：zoud@big.ac.cn

Recent Advances in Large-scale Biomedical Semantic Indexing

Ronghui You[1,2], Suyang Dai[1,2], Yuxuan Liu [1,2], Hiroshi Mamitsuka[3], Shanfeng Zhu[2,4](朱山风)

[1]*School of Computer Science, Fudan University, Shanghai, China*
[2]*Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China*
[3]*Institute for Chemical Research, Kyoto University, Kyoto, Japan*
[4]*Institute for Science and Technology of Brain-Inspired Intelligence, Fudan University, Shanghai, China*

摘要： With the rapid increase of biomedical articles, large-scale automatic semantic indexing has become increasingly important. For example, Medical Subject Headings (MeSHs) are used by the National Library of Medicine (NLM) to index almost all 30 million citations in MEDLINE. This greatly facilitates the applications of biomedical information retrieval and text mining. However, the automatic MeSH indexing still faces challenges as results of 1) a large number of labels (around 30000); 2) deep semantic information of biomedical text; and 3) limited information in title and abstract. The BioASQ challenge provides a realistic and practical benchmark to advance the design of effective algorithms for large-scale MeSH indexing. Over the last few years, we have developed a series of SOTA machine learning-based methods to address these challenges in large-scale MeSH indexing, such as MeSHLabeler, DeepMeSH, AttentionXML and FullMeSH. Specifically, AttentionXML has achieved the first place in the BioASQ2020 challenge. It obtained an MiF of 0.707, which is around 8.5% higher than that of NLM's official tool, Medical Text Indexer (MTI).

关键词： Biomedical Text Mining, BioCuration, Deep Learning, Large-scale Multi-label learning, MeSH Indexing

报告人 Email: zhusf@fudan.edu.cn

人体微生物与新生儿健康

Fangqing Zhao(赵方庆)

Beijing Institutes of Life Science, Chinese Academy of Sciences

摘要：Early colonization and development of the gut microbiota not only has a great impact on childhood but also influences their health when children grow up. Unlike that in adults, microbial community in infants, especially the first few years after birth, exhibits a less stable and more fragile pattern. Even a slight disturbance (e.g. diet, antibiotics) may break the balance of intestinal microbial community structure and change it to another dramatically different state. Such high plasticity and dependence on external environments may provide an opportunity to use external intervention during the early stage to improve children's health in the future. The simple composition of neonatal microbiota also helps us unveil the establishment of the symbiotic relationship between host and microbes. Integrated with clinical and geographical factors, the assembly, succession and maturation of neonatal microbiota during the early life were analyzed. Our findings provide comprehensive insights into the initial colonization and development of human microbiota through enterotype analyses and ecological modelling, which will undoubtedly improve our understanding of human microbiota and its impact on infant health.

报告人 Email：zhfq@biols.ac.cn

Comprehensive Analysis and Genome-wide Prediction of DNA Replication Origins in *Saccharomyces cerevisiae*

Dan Wang[1,2,3], Fei-Liao Lai[1,2,3], Mei-Jing Dong[1,2,3], Tao Liu[1,2,3], Hao Luo[1,2,3] and Feng Gao[1,2,3](高峰)

[1] *Department of Physics, School of Science, Tianjin University, Tianjin, China*

[2] *Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China*

[3] *SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin, China*

摘要：*Background:* DNA replication is a fundamental process in all organisms; this event initiates at sites termed origins of replication (ORIs). For eukaryotic organisms, replication origins are best characterized in the unicellular eukaryote *Saccharomyces cerevisiae*, which is one of the most established model organisms and considered as an ideal model for the study of human diseases. With the rapid increase of the complete *S. cerevisiae* genome sequences, the population genomic study is urgently needed to gain a more thorough understanding of the sequence conservation and evolutionary relationships of ORIs in budding yeasts. Notably, the development of algorithms and bioinformatic tools for identification of autonomously replicating sequences (ARSs) in *S. cerevisiae* genomes would provide high-efficient and low-cost methods to facilitate the reveal of the eukaryotic DNA replication mechanisms.

Results: We have presented a systematic study of conservation analysis for the replication origins of *S. cerevisiae* from both genome-wide and population genomics perspectives. Our results indicate that most of ARSs are unique across the whole genome of reference *S. cerevisiae* and those with high sequence similarity are prone to locate in subtelomeres. In the budding yeast population, most of the homologous ARSs are not only conserved in genomic sequence but also relatively conserved in chromosomal position. The non-conserved ARSs tend to distribute in the subtelomeric regions. Besides that, the genes adjacent to replication origins among the *S. cerevisiae* population have been extracted. We found that the genes adjacent to conserved ARSs are significantly enriched in DNA binding, enzyme activity, transportation and energy, whereas for the genes adjacent to non-conserved ARSs are significantly enriched in response to environmental stress, metabolites biosynthetic process and biosynthesis of antibiotics. By utilizing the Z-curve methodology, we have developed a novel pipeline, Ori-Finder 3, for the computational prediction of replication origins in *S. cerevisiae* at the genome-wide level based solely on DNA sequences. The ARS exhibiting both an AT-rich stretch and ACS element can be predicted at the single-nucleotide level. For the identified ARSs in the *S. cerevisiae* reference genome, 83 and 60% of the top 100 and top 300 predictions matched the known ARS records, respectively. Based on Ori-Finder 3, we subsequently built a database of the predicted ARSs identified in more than a hundred *S. cerevisiae* genomes. Consequently, a user-friendly web server including the ARS prediction pipeline and the predicted ARSs database has been developed, which can be freely available at http://tubic.tju.edu.cn/Ori-Finder3.

关键词：*Saccharomyces cerevisiae*; Replication origin; Z-curve; Population genomics; Autonomously replicating sequence (ARS)

报告人 Email: fgao@tju.edu.cn

病毒进化与病毒圈

崔杰

中国科学院上海巴斯德研究所

摘要：病毒圈指特定生态系统、生物群体或个体所包含或携带的全部病毒的总和，既包括人们已知的各种类病毒，也包含大量的未鉴定的新病毒。由于样品采集的限制以及检测方法的不完善，人们对于病毒圈的认识非常有限且具有一定的偏差。随着高通量测序技术的兴起，以及其在病原生物学领域的广泛应用，人们得以发现更多的全新病毒；更重要的是，这些新病毒的发现挑战了传统的病毒分类学。病毒圈的研究有助于理解病毒的宏观进化和基因组的多样性，了解病毒和宿主的共进化关系。病毒圈的方法学开发为人们了解病毒的起源与传播、重组和突变提供了依据，是描绘病毒生态圈的理论基础。

报告人 Email：jcui@ips.ac.cn

呼吸道微生物组研究进展与方法

Mingkun Li(李明锟)

Beijing Institute for Genomics, Chinese Academy of Sciences
China National Center for Bioinformation

摘要：Investigation of the lung microbiota is a relatively young field; however, there has been remarkable progress in understanding the composition and function of the lung microbiota in the last few years. Alterations of the lung microbiota have been observed in many respiratory diseases, including chronic obstructive pulmonary disease (COPD), asthma, and cystic fibrosis, but associations with clinical features and interactions with host genes are largely unknown. Meanwhile, technologies developed for lung microbiota have the potential to identify the pathogen that causes an infection in the respiratory tract. For instance, SARS-CoV-2 was first identified in the metatranscriptome data of the bronchoalveolar lavage fluid (BALF).

Our lab has conducted lung microbiota analysis on over 2000 samples, including the oropharyngeal swab, sputum, and BALF which were collected from pneumonia, COPD, COVID-19 patients, and healthy controls, to disentangle the association between the lung microbiota and disease progression. Meanwhile, we were also working on the development and optimization of the methods and protocols to manipulate different types of specimens as well as new algorithms to analyze the data. For example, a pretreatment method was developed to remove the overwhelming host DNA before library preparation, and Tn5 transposase was used to shorten the time of library preparation. In addition, a computational workflow was developed to analyze the microbiota composition at the species level.

SARS-CoV-2 first targets the cell in the respiratory tract and could potentially influence the microenvironment, which in turn changes the lung microbiota. Meanwhile, lung microbiota was proposed to alter the susceptibility of influenza, whether the lung microbiota could interact with the SARS-CoV-2 is unclear. Recently, we have analyzed the dynamics of lung microbiota in 192 COVID-19 patients with severe symptoms. We found that the microbiota dynamics were significantly different between patients with different clinical outcomes (deceased or recovered), and multiple species were identified as potential prognostic biomarkers in our study.
In this study, we want to review the recent research work and progress in the technique in the field of lung microbiota, and introduced the work conducted in our lab.

报告人 Email: limk@big.ac.cn

Computational Method Study for Phages and Plasmids from Metagenomic Sequences

Zhencheng Fang, Jie Tan, Shufang Wu and Huaiqiu Zhu(朱怀球)

Department of Biomedical Engineering, and Center for Quantitative Biology, Peking University, Beijing 100871, China.

摘要：Phage and plasmid are the main components of mobile genetic elements (MGEs) and key players in horizontal gene transfer (HGT). In sequenced metagenomic data, fragments from phage and plasmid generally co-exist with bacterial chromosome-derived fragments. By far, we lack a series of non-chromosome specific bioinformatics tools to perform sequence identification and annotation, which prevents biologists from effectively analysing the regulation of a complex microbial community. Herein, based on deep learning technique, we present two novel algorithms, namely PPR-Meta and PlasGUN, for sequence identification and gene prediction of phage and plasmid sequence in metagenomic data. PPR-Meta is the first tool that can simultaneously identify phage and plasmid sequences in metagenomic data. We designed a deep neural network structure named Bi-path Convolutional Neural Network (BiPathCNN). The “base path” of BiPathCNN takes the base sequence of the DNA as input, which is beneficial to extract the sequence signatures of the non-coding region, while the “codon path” of BiPathCNN takes the codon sequence of the DNA as input, which is beneficial to extract the sequence signatures of the coding region. Testing over a benchmark dataset, PPR-Meta performed much better on phage and plasmid identification respectively than similar bioinformatic tools. We used PPR-Meta to analyse the percentages of phages, bacterial chromosomes, and plasmids in microbial communities from the human digestive tract. The results showed that in the inner end of the digestive tract, the percentages of the phages and plasmids are relatively lower, while higher in the outer end. Since the phage and plasmid are the main mediators of horizontal gene transfer, such a phenomenon indicated that the frequency of horizontal gene transfer at the outer end of the digestive tract may be higher. This phenomenon also indicated that horizontal gene transfer seemed to be a way for microbial communities in the outer end of the digestive tract to adapt to the external environment. Correspondingly, PlasGUN is the first tool of gene prediction for plasmid metagenomic short read data. We designed a deep neural network structure named multiple input Convolutional Neural Network (miCNN), which extracts the sequence signatures of each plasmid candidate ORF (open reading frame) from multiple dimensions. Considering the complexity of the ORF structure, we toke the codon sequence of the ORF, upstream and downstream sequence of the start codons, ORF length and ORF integrity type as the input signals of PlasGUN. Testing over the benchmark dataset, PlasGUN demonstrates much better performance on plasmid short read data than traditional gene prediction tools. To further verify the effectiveness of PlasGUN algorithm, we tried to construct a gene prediction tool named VirGUN for phage short read sequences using the strategy similar to PlasGUN. Testing over a benchmark dataset, VirGUN can also achieve the best performance among related tools. This shows that the algorithm of PlasGUN is suitable for gene prediction of extrachromosomal elements like

the phage and plasmid. We expect that PPR-Meta and PlasGUN will be powerful bioinformatic tools to analyze the phage and plasmid in the metagenome for biologists, which will facilitate biologists' understanding of the complexity of microbial communities and their regulation mechanisms. PPR-Meta and PlasGUN are freely available via https://github.com/zhenchengfang/PPR-Meta and https://github.com/zhenchengfang/PlasGUN.

关键词：Microbiome, Metagenome, Phage, Plasmid, Deep learning

报告人 Email: hqzhu@pku.edu.cn

Identification and Characterization of Bacterial Toxin-antitoxin Loci

Hong-Yu Ou(欧竑宇)

State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade joint innovation center on Antibacterial Resistances, School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, China

摘要：Bacterial Toxin–antitoxin (TA) systems are highly abundant in most free-living bacteria (1). The TA systems are involved in multiple life activities of bacteria, such as nutrition starvation, programmed cell death, protection from bacteriophage, and the antimicrobial resistance. TA system consists of a stable toxin protein and a labile cognate antitoxin encoded by a bicistronic locus. Recently, we have developed an open-access database, TADB2.0, which provides comprehensive information about bacterial type II toxin-antitoxin (TA) loci. With the aid of text mining and manual curation, it recorded 6193 type II TA loci in 870 replicons of bacteria and archaea, including 105 experimentally validated TA loci. Besides, the newly developed tool TAfinder combines the homolog searches and the operon structure detection, allowing the prediction for type II TA pairs in bacterial genome sequences. After the examination of 2786 species of prokaryotes with the publicly available complete genome sequences, TAfinder prediction results showed that 66% of species harbored 1–20 type II TA loci in individual strains while 20% of species carried more than 20 loci. Remarkably, the newly characterized acetyltransferase-type TA loci are widely distributed in *Enterobacteriaceae* bacteria, including *Escherichia coli*, *Salmonella enterica*, and *Klebsiella pneumoniae*.

Gcn5-related N-acetyltransferase (GNAT) toxins inhibit translation by acetylation of aminoacyl-tRNAs and are counteracted by cognate ribbon-helix-helix (RHH) antitoxins, but the toxicity neutralization mechanisms need to be clarified. By using TAfinder, we identified a GNAT-RHH TA locus *kacAT* in the chromosome of *K. pneumoniae* HS11286, an ST11 strain resistant to multiple antibiotics. And we found that the overexpression of KacT halted cell growth while co-expression of KacA neutralized KacT toxicity via forming a heterohexamer complex. Besides, this complex interacted with the cognate promoter DNA, resulting in negative auto-regulation of *kacAT* transcription. The crystal structures of the KacA-KacT-operator DNA complex revealed the formation of a unique heterohexamer, KacT-KacA2-KacA2-KacT. The direct interaction of KacA and KacT involves a unique W-shaped structure with the two KacT molecules at opposite ends. Our subsequent mutagenesis confirmed that the inhibition of KacT is achieved by the binding of four KacA proteins that preclude the formation of an active KacT dimer. And we presented an experimentally supported molecular model proposing that the KacT:KacA ratio controls *kacAT* transcription by conditional cooperativity. Our results showed how the RHH antitoxin neutralizes the GNAT toxin in a unique way and how the GNAT-RHH complex auto-regulates its transcription.

关键词：bacterial toxin and antitoxin system, biological database, prediction tool, GNAT-RHH toxin-antitoxin loci, *Klebsiella pneumoniae*

报告人 E-mail: hyou@sjtu.edu.cn

CVTree: Whole-Genome-Based and Alignment-Free Phylogeny/Taxonomy of Prokaryotes

Guanghong Zuo[1](左光宏)

[1] *T-life Research Center and Department of Physics, Fudan University, Shanghai, 200433, China*

摘要：It has been estimated that there are 10^{30} living Archaea and Bacteria cells on Earth, making almost half of the biomass. Yet our knowledge on prokaryotes is rather limited. Prokaryote phylogeny and taxonomy were mainly based on morphological characters until Carl Woese and coworkers suggested to use the small subunit rRNA sequence as molecular clock for Archaea and Bacteria in the late 1970s. The 16S rRNA analysis has been so successful that the second edition of the Bergey's Manual of Systematic Bacteriology follows "a phylogenetic framework based on analysis of the nucleotide sequence of the small ribosomal subunit RNA, rather than a phenotypic structure". However, it is just this "congruency" of prokaryote phylogeny and taxonomy on the basis of 16S rRNA analysis makes independent verification of the latter an urgent task. The verification may be considered independent as long as both the input data and the methodology are distinct from the small subunit rRNA sequence analysis. As complete genomes contain much more phylogenetic information than a single gene or a set of proteins, whole-genome based approaches are destined to play a central role in phylogeny. Furthermore, the significance of using whole genomes goes far beyond verification of the 16S rRNA analysis.

CVTree constructs whole-genome based phylogenetic trees without sequence alignment by using a Composition Vector (CV) approach. It was first developed to infer evolutionary relatedness of Bacteria and Archaea and then successfully applied to viruses, chloroplasts, and fungi. In this talk, I will introduce the methodology and describe the newest CVTree4 web server (http://tlife.fudan.edu.cn/cvtree4). Prokaryote phylogeny is verified by direct comparison with taxonomy at all ranks from phyla down to species. An interactive tree-viewer allows searching, collapsing, and expanding the tree branches. An automatic generated list of taxa, which are monophyletic or not, helps to comprehend the overall agreement of phylogeny with taxonomy and hints on possible taxonomic revisions. By using CVTree, and together with 16S rRNA-based phylogenic tree (obtained from "The All-Species Living Tree" Project and rearranged in LVTreeViewer: http://tlife.fudan.edu.cn/lvtree) and taxonomic system, we studied the phylogeny and taxonomy of prokaryotes in detail. This will help to put phylogeny and taxonomy of the prokaryotes on a solid ground. And we hope CVTree will become a convenient tool in the hands of microbiologists.

关键词：Phylogeny, Taxonomy, Prokaryotes, Composition Vector

报告人 Email: ghzuo@fudan.edu.cn

Altered Gut Microbiota in Parkinson's Disease Patients/Healthy Spouses and its Association with Clinical Features

Fan Zhang[1], Liya Yue[2], Xing Fang[1], Gengchao Wang[2,3], Cuidan Li[2], Xiaodong Sun[1], Xinmiao Jia[4], Jingjing Yang[1], Jinhui Song[1], Yu Zhang[1], Chongye Guo[2], Guannan Ma[2], Ming Sang[1], Fei Chen[2,3](陈非), Puqing Wang[1]

[1] *Department of Neurology, Xiangyang No.1 People's Hospital, Hubei University of Medicine, Xiangyang, Hubei, China*

[2] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China National Center for Bioinformation, Beijing, China*

[3] *University of Chinese Academy of Sciences, Beijing, China*

[4] *Department of Medical Research Center, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing, China*

摘要：*Introduction*: Increasing evidence shows that gut microbiota dysbiosis may play important roles in the occurrence and progression of Parkinson's disease (PD), but the findings are inconsistent. Besides, the effect of family environment on gut microbiota dysbiosis remains unclear.

Methods: We characterized the gut microbial compositions of 63 PD patients, 63 healthy spouses (HS) and 74 healthy people (HP) using 16S rRNA sequencing. Clinical phenotypes and microbial composition were analyzed comprehensively.

Results: There were markedly different microbial compositions among PD, HS and HP samples by alpha/beta diversity. We also found differential microbial compositions among Hoehn & Yahr stage/disease duration. Eight inflammation-associated microbial genera shared a continuously increase trend with increased Hoehn & Yahr stage and disease duration, indicating characteristic bacteria associated with deterioration in PD. Additionally, seven bacterial markers were identified for accurately differentiating PD patients from the controls (area under the curve [AUC]: 0.856).

Conclusions: Our study shows altered gut microbiota in PD patients. Importantly, inflammation-associated microbial genera may play roles in PD progression. Differential microbial compositions in HS and HP samples demonstrate that the gut microbiota are also affected by family environment. Disease-associated metagenomics studies should consider the family environmental factor. Our research provides an important reference and improves the understanding of gut microbiota in PD patients.

关键词：Gut microbiota, Parkinson's disease, Hoehn & Yahr stage

报告人 Email: chenfei@big.ac.cn

重复序列扩增疾病与抗新冠病毒药物预测

史庆[1]，刘鑫[1]，陈斌[1]，王秀杰[1]

[1] *中国科学院遗传与发育生物学研究所　北京 100101*

摘要：许多神经肌肉系统疾病（如脆性 X 染色体综合症、亨廷顿舞蹈症、脊髓小脑性共济失调等）的发生都与某些基因内短序列的重复扩增有关，重复序列的扩增也会影响蛋白质的相变特征。目前已知能通过重复序列扩增导致疾病的基因还很少，并且分散于不同的文献报道中。为建立统一的重复序列扩增疾病的数据资源，以方便相关研究，我们构建了重复序列扩增疾病数据库—DRED (DRED, Database of genes related to Repeat Expansion Diseases, http://omicslab.genetics.ac.cn/dred/index.php)。DRED 数据库不仅收集了已经发表的所有可以通过序列扩增导致疾病的基因，也利用机器学习的方法，对可能通过序列扩增导致疾病的基因进行了预测，从而为研究其功能提供便利。

在新冠疫情爆发的早期，我们针对 COVID-19 的 M^{Pro} 水解酶，预测了已有临床药物对 M^{Pro} 水解酶的抑制能力，进而筛选了可能的抗 COVID-19 药物。我们也与中国中医科学院合作，对抗新冠病毒方剂“化湿败毒方”的功效原理进行了分析。

报告人 Email：xjwang@genetics.ac.cn

复杂疾病生物标志物寻找的原理与应用

Bairong Shen(沈百荣)

Institutes for Systems Genetics, West China Hospital，Sichuan University, Chengdu 610041， China

摘要：Biomarker discovery for complex diseases are often based on experiments or disease specific modeling. We here take cancer microRNA biomarker as the case study and aim to identify the key and vulnerability parts in the systems based on the computational network structural and functional analysis, combining with evidence-based pattern recognition. We discovered several principles for the microRNA biomarker discovery with characterizing of the structural, functional and evolutional properties of the candidates, as well as their genotyping-phenotyping relationships. These principles are validated with many different diseases such as cancers, cardiovascular disease and neurodegenerative diseases and others. We therefore developed software tools for the application of the models for novel biomarker discovery in different complex diseases. We now extended the model to the study of other ncRNAs and the ceRNA biomarkers.

关键词：Biomarker discovery, Complex disease, microRNA, network characterization

报告人 Email: bairong.shen@scu.edu.cn

环形 RNA 编码蛋白潜能的生物信息学研究

宋晓峰

南京航空航天大学自动化学院

摘要：真核细胞在 DNA 转录后通过反向剪接机制形成的环形 RNA 具有重要生物学功能。已有研究表明部分环形 RNA 具有编码蛋白的能力，但调控其翻译活性的分子机制尚不清楚。有研究证实人工合成的含有“内部核糖体进入位点”（Internal Ribosome Entry Site，IRES）元件和开放阅读框的环形 RNA 可以在体外翻译产生蛋白质，有研究也发现部分内源性环形 RNA 翻译多肽分子。随着第二代测序技术、Ribo-seq、ChIRP 技术、翻译组学技术、蛋白质谱技术等各种高通量实验技术的发展，以及国际上多种环形 RNA 数据库的建立，我们能够从系统生物学与生物信息学角度，深入探讨环形 RNA 分子在其编码蛋白潜能上的序列与结构特征。

报告人 Email：xfsong@nuaa.edu.cn

Bioinformatics Methods and Applications for T-cell Immunology and Immune Checkpoint Therapy

Ya-Ru Miao, Qiong Zhang, Si-Yi Chen, Fei-Fei Hu, Chun-Jie Liu, An-Yuan Guo(郭安源)

Department of Bioinformatics and Systems Biology, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

摘要：T-cell immunology plays a pivotal role in human immune respone. The distribution and abundance of immune cells, particularly T-cell subsets, are very important in cancer immunology and therapy. T cells has many subsets with specific function and current methods are limited in estimating them, thus, a method for predicting comprehensive T-cell subsets is urgent need in cancer immunology research. Here we introduce Immune Cell Abundance Identifier (ImmuCellAI), a novel gene set signature-based method, for precisely estimating the abundance of 24 immune cell types including 18 T-cell subsets, from gene expression data. Performance evaluation on both our sequencing data with flow cytometry results and public expression data indicated that ImmuCellAI can estimate the abundance of immune cells with superior accuracy than other methods especially on many T-cell subsets. Application of ImmuCellAI to immunotherapy datasets revealed that the abundance of dendritic cells (DC), cytotoxic T, and gamma delta T cells was significantly higher both in comparisons of on-treatment vs. pre-treatment and responders vs. non-responders. Meanwhile, we built an ImmuCellAI result-based model for predicting the immunotherapy response with high accuracy (AUC 0.80~0.91). These results demonstrated the powerful and unique function of ImmuCellAI in tumor immune infiltration estimation and immunotherapy response prediction. The ImmuCellAI online server is freely available at http://bioinfo.life.hust.edu.cn/web/ImmuCellAI/.

T cell receptors (TCR) function to recognize antigens and play vital roles in T-cell immunology. Surveying TCR repertoires by characterizing complementarity-determining region 3 (CDR3) is a key issue. Due to the high diversity of CDR3 and technological limitation, accurate characterization of CDR3 repertoires remains a great challenge. We propose a computational method named CATT for ultrasensitive and precise TCR CDR3 sequences detection. CATT can be applied on TCR sequencing (TCR-Seq), RNA-Seq, and single-cell TCR(RNA)-Seq data to characterize CDR3 repertoires. CATT integrated de Bruijn graph based micro-assembly algorithm, data-driven error correction model, and Bayesian inference algorithm, to self-adaptively and ultra-sensitively characterize CDR3 repertoires with high performance. Benchmark results of datasets from in silico and experimental data demonstrated that CATT showed superior recall and precision compared with existing tools, especially for data with short read length and small size, and single-cell sequencing data. Thus, CATT will be a useful tool for TCR analysis in researches of cancer and immunology. CATT is freely available at http://bioinfo.life.hust.edu.cn/CATT.

Immune checkpoint genes (ICGs) play critical roles in circumventing self-reactivity and

represent a novel target to develop treatments for cancers. However, a comprehensive analysis for the expression profile of ICGs at a pan-cancer level and their correlation with patient response to immune checkpoint blockade (ICB) based therapy is still lacking. In this study, we defined three expression patterns of ICGs using a comprehensive survey of RNA-seq data of tumor and immune cells from the FANTOM5 project. The correlation between the expression patterns of ICGs and patient survival and response to ICB therapy was investigated. The expression patterns of ICGs were robust across cancers, and upregulation of ICGs was positively correlated with high lymphocyte infiltration and good prognosis. Furthermore, we built a model (ICGe) to predict the response of patients to ICB therapy using five features of ICG expression. A validation scenario of six independent datasets containing data of 261 patients with CTLA-4 and PD-1 blockade immunotherapies demonstrated that ICGe achieved AUCs of 0.64–0.82 and showed a robust performance and outperformed other mRNA-based predictors. In conclusion, this work revealed expression patterns of ICGs and underlying correlations between ICGs and response to ICB, which helps understand the mechanisms of ICGs in ICB signal pathways and other anticancer treatments.

关键词：immune cell, T-cell subsets, cancer, immunotherapy, T cell receptor

报告人 Email: guoay@hust.edu.cn

Systematically Analyzing the Regulation of Immune Pathways Identifies Potential Oncogenic Biomarkers

Yongsheng Li[1,2](李永生), Tiantongfei Jiang[1], Weiwei Zhou[1], Haozhe Zou[1], Qi Wang[1], Xia Li[1,2]

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China.

[2] Key Laboratory of Tropical Translational Medicine of Ministry of Education, Hainan Medical University, Haikou 571199, China.

摘要：Genetic and epigenetic alterations in immune-related pathways are common hallmarks of cancer. However, a comprehensive understanding of immune networks and how immune regulators impact network structure and functional output across cancer types is instrumental.

Herein we systematically interrogated somatic mutations and long noncoding RNAs (lncRNAs) regulation on immune-related pathways. A network-based computational model (NIPPER) with scoring systems was proposed to prioritize critical genes and mutations eliciting differential HLA binding affinity and alternate responses to immunotherapy. Mutations genes involved in immune pathways are enriched in essential protein domains and often alter tumor infiltration by immune cells, affecting T cell receptor repertoire and B cell clonal expansion. Interactome network propagation framework integrated with drug associated gene signatures can be used to identify potential immunomodulatory drug candidates.

Moreover, we developed ImmLnc (http://bio-bigdata.hrbmu.edu.cn/ImmLnc), for identifying lncRNA regulators of immune-related pathways. The immune-related lncRNAs are likely to show expression perturbation in cancer and are significantly correlated with immune cell infiltration. ImmLnc can help prioritize cancer-related lncRNAs and further identify three molecular subtypes (proliferative, intermediate, and immunological) of non-small cell lung cancer. These subtypes are characterized by differences in mutation burden, immune cell infiltration, expression of immunomodulatory genes, response to chemotherapy, and prognosis.

Taken together, our systems-level analyses of the regulation to immune-related pathways help interpret the heterogeneous immune responses among patients. Our results identified potential oncogenic biomarkers and serve as an important resource for future functional studies and targeted therapeutics.

关键词：immune pathways, mutations, lncRNAs, biomarkers, network biology

报告人 Email：liyongsheng@ems.hrbmu.edu.cn

Reference Gene Selection for Quantitative Gene Expression Analysis in Platelet of Tumor

Guishu Yang[1,2], Dongsheng Wang[2], Ruiling Zu[2], Yulin Liao[2], Kaijiong Zhang[2], Ying Lin[3]& Huaichao Luo[2](罗怀超)

[1] Department of Clinical Laboratory, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, China,610041;

[2] School of Clinical Medicine, Southwest Medical University, Luzhou, Sichuan, China,646000;

[3] Department of Laboratory Medicine, Sichuan Academy of Medical Sciences, Sichuan Provincial People's Hospital, School of Clinical Medicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, China,610072.

摘要：The main purpose of this study is to explore the expression of ACTB, B2M and GAPDH, the conventional internal reference genes, in platelets of different types of tumors and healthy people, and to find the most suitable reference gene for normalizing gene expression data after evaluating the stability of candidate genes. Which offers a solid foundation for further studies of the molecular biology of platelet of tumor. According to the final diagnosis basis, the whole blood samples of peripheral veins from patients with lung cancer, breast cancer, liver cancer, colon cancer, which had been diagnosed in our hospital were collected. Ten cases of each tumor and ten healthy individuals were selected. Separated by centrifugation, the platelets in the whole blood were extracted. The RNA in the platelets was extracted by using Trizol Reagent and reversed transcribed into cDNA. Then the expression levels of the three candidate reference genes in the platelets were detected by the technology of Quantitative RT-PCR, five algorithms, including GeNorm, NormFinder, Bestkeeper, Delta CT and Comprehensive Ranking, was carried out to evaluate the stability of candidate genes.

The results show that the expression stability of B2M and GAPDH are similar, and both are more stable to the expression of ACTB by using the GeNorm algorithm analysis. While the results of the other four algorithms (NormFinder, Bestkeeper, Delta CT, and Comprehensive Ranking) were shown that in 50 platelet samples, the expression of GAPDH is the most stable, which indicates GAPDH is more suitable as a reference gene for platelet RNA detection in tumor. Based on the evaluation results of the stability of the three candidate reference genes, GAPDH has higher expression stability and is more suitable as a standard internal reference for analysis of gene expression in platelets in tumor-related research.

关键词：Reference Gene, Platelet, tumor

报告人 Email：luo1987cc@163.com.

An Integrative Pharmacogenomics Analysis Identifies CK2 Alpha as a Promising Therapeutic Target in KRAS($G_{12}C$) Mutant Lung Cancer

Haiyun Wang[1](王海芸), Qi Lv[1], Yue Xu[1], Zhaoqing Cai[1], Jie Zheng[1], Xiaojie Cheng[1], Yao Dai[1], Pasi A. Jänne[2,3], Chiara Ambrogio[2], Jens Köhler[2]

[1] *School of Life Sciences and Technology, Tongji University, Shanghai 200092, China*
[2] *Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA*
[3] *Belfer Center for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA*

摘要：KRAS mutations are the most frequent oncogenic aberration in lung adenocarcinoma. KRAS mutant isoforms differentially shape the biology of tumours and influence drug responses. This heterogeneity challenges the development of effective targeted therapies for KRAS-driven lung cancer. Here, we systematically investigated MEK/ERK inhibitors sensitivity for different KRAS mutant isoforms. Then we developed an integrative pharmacogenomics analysis to identify potential targets in lung cancer with KRAS(G12C) mutation, the most frequent aberration in patients with primary or metastatic KRAS mutant non-small cell lung cancer. We validated our predictive in silico results with in vitro models using gene knockdown, pharmacological target inhibition and reporter assays. CSNK2A1 knockdown reduces cell proliferation, inhibits Wnt/β-catenin signalling and increases the anti-proliferative effect of MEK inhibition selectively in KRAS(G12C) mutant lung cancer cells. The specific CK2-inhibitor phenocopies the CSNK2A1 knockdown effect and sensitizes KRAS(G12C) mutant cells to MEK inhibition.

关键词：Pharmacogenomic profiles, KRAS mutations, Lung adenocarcinoma, CSNK2A1

报告人 Email：wanghaiyun@tongji.edu.cn

复杂疾病相关 lncRNA 竞争三元组机制解析与精准医疗

Peng Wang [1](王鹏), Yingfeng Luo [2], Jianfeng Huang [1], Shenghan Gao [2], Songnian Hu [2], Yeyuan Chen [1]

[1] Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences & Ministry of Agriculture Key Laboratory of Crop Gene Resources and Germplasm Enhancement in Southern China, No. 4 Xueyuan Road, Haikou 571100, Hainan, China

[2] State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, 1-3 West Beichen Road, Beijing 100101, China

摘要：Mango is one of the most important tropical fruits. It belongs to the family Anacardiaceae, a majority of whose members produce family-specific urushiols and related phenols, which can induce contact dermatitis. We generate a chromosome-scale genome assembly of mango, providing a reference genome for the Anacardiaceae family. Our results indicate the occurrence of a recent whole-genome duplication event in mango. Duplicated genes preferentially retained include photosynthetic, photorespiration, and lipid metabolic genes that may have provided adaptive advantages to sharp historical decreases in atmospheric carbon dioxide and global temperatures. A notable example of an extended gene family is the chalcone synthase (CHS) family of genes, and particular genes in this family show universally higher expression in peels than in flesh, likely for the biosynthesis of urushiols and related phenols. Genome resequencing reveals two distinct groups of mango varieties, with commercial varieties clustered with India germplasm demonstrating allelic admixture, and indigenous varieties from Southeast Asia in the second group. Landraces indigenous in China formed distinct clades, and some showed admixture in genomes. In conclusion, analysis of chromosome-scale mango genome sequences reveals photosynthesis and lipid metabolism are preferentially retained after a recent WGD event, and expansion of CHS genes is likely associated with urushiol biosynthesis in mango. Genome resequencing clarifies two groups of mango varieties, discovers allelic admixture in commercial varieties, and shows distinct genetic background of landraces.

关键词：mango, genome, domestication, urushiol

报告人 Email：pwang521@163.com

ePmiRNA_finder: Identification of Extracellular Plant miRNAs in Human and Animals

Longjiang Fan(樊龙江)

Institute of Crop Sciences & Institute of Bioinformatics, Zhejiang University, Hangzhou 310058, China

摘要： Extracellular microRNAs (miRNAs) have been proposed to have a potential cross-kingdom gene regulatory action. Among these, plant derived miRNAs of dietary origin have been reported to survive the harsh conditions of the human digestive system, enter the circulatory system, and regulate gene expression and metabolic function. Definitive evidence in support of the presence and scope of plant miRNAs in human serum and cells and the speculative conclusions on their regulatory role has been difficult to obtain due to limited sample sizes and low statistical power during analysis. We have developed a bioinformatics pipeline (ePmiRNA_finder) and applied it to analyze 421 small RNA sequencing data sets from 10 types of human body fluids and tissues and comparative samples from carnivores or herbivores. Following strident rule for classification, a total of 35 miRNAs were identified that map to edible plants were found in at least one human blood sample and were at levels significantly different in cow or microbiome, suggesting that plant miRNA abundance correlates with dietary intake without attribution from contaminations. Plant miRNA profiles were characterized also to be body fluid/tissue-specific, especially the abundant identification in the brain and the breast milk samples, which support the capable transportation through the barriers and selective absorption of dietary miRNAs. Our study provides supportive evidence for the dietary intake of plant miRNAs and their cross-kingdom action within human circulating system.

报告人 Email: fanlj@zju.edu.cn

水稻基因组序列变异的功能注释

Hu Zhao, Weibo Xie(谢为博)

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

摘要：Interpreting the functional impacts of genetic variants is an important challenge for functional genomic studies in crops and next-generation breeding. Currently, studies in rice have mainly focused on the identification of genetic variants, while the functional annotation of variants has not yet been carried out. Here we curate haplotype information of 17,397,026 genomic variations from sequencing data of 4,507 rice accessions and quantitatively evaluate the effects of missense mutations in coding regions in each haplotype based on the conservation of amino acid residues. We also generate high-quality chromatin accessibility (CA) data from six representative rice tissues and train deep convolutional neural networks (AUROC: 0.93-0.95) using these data to predict the impacts of variants for CA in non-coding regions. We characterize the functional properties and tissue specificity of the effects of genetic variants. We finally demonstrate how the functional annotation data of genetic variants can be used to identify the causal variations in mapping populations. The annotation data would be a valuable resource for prioritizing genetic variants in rice and can be freely queried in RiceVarMap v2.0 (http://ricevarmap.ncpgr.cn).

关键词：Rice, Genetic variant, Functional annotation, Chromatin accessibility, Deep learning

报告人 Email: weibo.xie@mail.hzau.edu.cn

植物转录组数据大规模整合与挖掘

Siyuan Chen, Zhixu Qiu, Mingui Song, Chuang Ma (马闯)

State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest Agriculture and Forestry University, Yangling, Shaanxi 712100, China

摘要：High-throughput sequencing of RNA (RNA-Seq) has entered almost all research laboratories, and become a key research tool to perform genome-wide transcriptome profiling in both model and non-model species. The constant improvement of RNA-Seq technologies coupled with sharp decreases in sequencing costs and data generation timelines, now enables investigators to perform large-scale sequencing-based projects for hundreds of thousands of samples from different cells, tissues, organs, experimental conditions, individuals and species. The large-scale transcriptome sequencing brings considerable challenges for comprehensive data analysis and knowledge discovery. Towards addressing these challenges, we have developed several bioinformatics methods and web-based platforms (available at https://github.com/cma2015) that aim to facilitate large-scale integration and intelligent analysis of plant transcriptomes. In the following, we present two recently developed methods for the utilization of unmapped reads and for the mining of high-dimensional gene expression matrix.

The commonly used RNA-Seq workflows usually yield millions of unmapped and thus uncharacterized reads. We examined ~4.71 billion reads from 117 maize B73 samples, identified ~50 million unmapped reads, and using a newly developed bioinformatics pipeline *de novo* assembled 5,419 maize transcripts. Of these, 635 had strong evidence support at the genome and/or transcript level. 83 of these 635 assembled transcripts were classified as protein-coding RNAs, and 40 exhibited tissue specificity. Moreover, we found that a number of assembled transcripts, including those encoding translation factors, were differentially expressed in the drought-stress response in ear and leaf tissues of maize. Our unmapped reads analysis provides a comprehensive resource of unmapped maize transcripts and reveals their associations with drought stress, providing potentially important new genes for the investigation of drought stress-related mechanisms.

Large-scale RNA-Seq experiments produce high-dimensional gene expression data, which can not be effectively analyzed with traditional approaches like differential expression analysis and gene co-expression analysis. We explored the gene expression data from 940 maize B73 samples with an unsupervised machine learning algorithm---matrix fractionalization (MF). Our MF-based analysis enriched the biological knowledge of maize seed through the identification of seed specifically expressed genes in maize, the exploration of the temporal effect on gene expression from temporal transcriptomes, and the identification of compartment-specific genes and biological processes from spatial transcriptomes.

关键词：Machine learning, Matrix fractionalization, RNA sequencing, Transcriptome, Unmapped reads

报告人 Email: chuangma2006@gmail.com

遗传与表观遗传互作决定普通小麦亚基因组分化的分子机制研究

Meiyue Wang[1,2], Zijuan Li[1,2], Yuyun Zhang[1,2], Yijing Zhang[1,2](张一婧)

[1] *National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 300 Fenglin Road, Shanghai 200032, China*

[2] *University of the Chinese Academy of Sciences, Beijing, 100049, China*

摘要：The widely cultivated wheat has a large allohexaploid genome. Subgenome-divergent regulation contributed to the genome plasticity and success of polyploid wheat in domestication. However, the specificity encoded in wheat genome determining the subgenome-divergent spatio-temporal regulation has been largely unexplored. The considerable size and complexity of the genome are major obstacles to dissecting the regulatory specificity. Here, we compared the epigenomes and transcriptomes from a large spectrum of samples under diverse developmental and environmental conditions. A total of 223,976 distal regulatory elements (REs) were specifically linked to their target promoters with orchestrated epigenomic activity. We detected distinct epigenetic architectures of REs representing different levels of subgenome divergence. Furthermore, through employing quantitative epigenomic approaches, we detected key responsive cis- and trans-acting factors validated by DNA Affinity Purification and sequencing (DAP-seq), and demonstrated the coordinated interplay between RE sequence contexts, epigenetic factors, and transcription factors in determining subgenome divergence. Altogether, this study provides a wealth of resources for elucidating the RE regulomics and subgenome-divergent regulation in hexaploid wheat, and gives new clues for interpreting genetic and epigenetic interplay in regulating the benefits of polyploid wheat.

关键词：common wheat, epigenomic atlas, subgenome divergence, regulatory elements, enhancer, development, stress response

报告人 Email：zhangyijing@cemps.ac.cn

ChIP-Hub: an Integrative Platform for Exploring Plant Regulome

Liang-Yu Fu[1,2], Zhigui Wu[1], Peijing Zhang[3], Ming Chen[3], Kerstin Kaufmann[2] and Dijun Chen[1,2](陈迪俊)

[1] *State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China*

[2] *Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, 10115 Berlin, Germany*

[3] *Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, 10115 Berlin, Germany*

摘要：ChIP-seq and complementary assays are powerful methods to measure protein-DNA binding events and chemical modifications of histone proteins at genome-wide level. In recent years, an explosion of regulome data are being generated in plants. However, reuse and comparison of data generated by different laboratories is not straightforward, hampering data integration to generate novel hypotheses for further investigation. Here, we adapt the standard provided by the ENCODE consortium to set up computational pipelines and systematically reanalyze public ChIP-seq (as well as DAP-seq) data in plants. We further present an integrative web-based platform (ChIP-Hub) for exploring the comprehensive reanalysis of more than 7000 ChIP-seq datasets in >40 plant species. The resources are bundled in a well-accessible database that also allows visualization and meta-analyses. This resource will allow experimental biologists from various fields to comprehensively use all available epigenomic information to get novel insights into their specific questions. ChIP-Hub is available at http://www.chiphub.org/.

关键词：ChIP-Hub, Regulome, Epigenome, Plants, Big Data

报告人 Email: dijunchen@nju.edu.cn

Incorporation of Parental Phenotypic Data into Multi-omic Models Improves Prediction of Yield-related Traits in Hybrid Rice

Yang Xu[1](徐扬), Yue Zhao[1], Shizhong Xu[2], Chenwu Xu[1,*]

[1] *Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology/Co-Innovation Center for Modern Production Technology of Grain Crops, Key Laboratory of Plant Functional Genomics of Ministry of Education/ Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding, College of Agriculture, Yangzhou University, Yangzhou 225009, China*

[2] *Department of Botany and Plant Sciences, University of California, Riverside, CA, 92507, USA*

摘要： Hybrid breeding has been shown to effectively increase rice productivity. However, identifying desirable hybrids out of numerous potential combinations is a daunting challenge. Genomic selection holds great promise for accelerating hybrid breeding by enabling early selection before phenotypes are measured. With the recent advances in multi-omic technologies, hybrid prediction based on transcriptomic and metabolomic data has received increasing attention. However, the current omic-based hybrid prediction has ignored parental phenotypic information, which is of fundamental importance in plant breeding. In this study, we integrated parental phenotypic information into various multi-omic prediction models applied in hybrid breeding of rice and compared the predictabilities of 15 combinations from four sets of predictors from the parents, i.e., genome, transcriptome, metabolome and phenome. The predictability for each combination was evaluated using the best linear unbiased prediction and a modified fast HAT method. We found significant interactions between predictors and traits in predictability, but joint prediction with various combinations of the predictors significantly improved predictability relative to prediction of any single source omic data for each trait investigated. Incorporation of parental phenotypic data into various omic predictors increased the predictability, averagely by 13.6%, 54.5%, 19.9%, and 8.3%, for grain yield, number of tillers per plant, number of grains per panicle, and 1,000 grain weight, respectively. Among nine models of incorporating parental traits, the AD-All model was the most effective one. This novel strategy of incorporating parental phenotypic data into multi-omic prediction is expected to improve hybrid breeding progress, especially with the development of high-throughput phenotyping technologies.

关键词： genomic selection, hybrid rice, multi-omic data, parental traits, best linear unbiased prediction

报告人 Email： yangx@yzu.edu.cn

The Detection of Differentially Expressed Gene and Atlas Construction of Pre-mRNA Alternative Splicing during Seed Germination of *Arabidopsis thaliana*

Yongqiang Xing[1,2](邢永强), YanXin Wang[1], Lu Cai[1,2]

[1] *School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010*

[2] *The Inner Mongolia Key Laboratory of Functional Genome Bioinformatics, Inner Mongolia University of Science and Technology, Baotou 014010*

摘要：Seed germination is the initial phase of life cycle for plants and plays a pivotal role in regeneration of plant population, continuation of plant species and crop yield. However, seed germination is an extremely complex biological process and its regulatory mechanisms remain unclear. It is well known that gene expression and pre-mRNA alternative splicing are vital to the central dogma. In *Arabidopsis* genome, ~60% multi-exon genes are alternatively spliced. Alternative splicing has been shown to be involved in the response to environmental cues, including abiotic and biotic stresses, in the regulation of crucial developmental processes such as flowering, and in circadian timekeeping. The present paper studies the effect of gene expression and alternative splicing on seed germination in *Arabidopsis*.

We constructed RNA-seq datasets from Dry seeds, Germinating seeds 1-3, Young seeds and Seeds in *Arabidopsis* and carried out quality control for the raw reads by using FastQC and Trimmomatic. StringTie estimated transcript and gene expression from a mapping to the *Arabidopsis* genome obtained from Hisat2. We next created the gene expression matrix for *Arabidopsis*, which includes 37336 genes along the rows and 12 samples along the columns, and employed DESeq2 to systematically identify differentially expressed genes between different stages during seed germination. To elucidate the biological functions of differentially expressed genes, we performed a functional enrichment analysis of these genes using ClusterProfiler tool. The results showed that many genes involved in metabolic processes, such as glycolysis and DNA synthesis, were activated during the seed germination (see Fig.1). It ensures the reactivation of metabolism and basic cellular activities of seeds. Some genes involved in environmental responses, such as oxidative stress, toxic substance, heat, and water, were inhibited. This is essential for seeds to be successfully activated from a dormant state. Furthermore, rMATS was utilized to detect the pre-mRNA alternative splicing events and differential alternative splicing events under different conditions. We constructed the atlas of pre-mRNA alternative splicing at different stages of seed germination and analyzed the distribution characteristics of alternative splicing patterns across every stage. Besides, we detected the differential alternative splicing events between different germination stages and analyzed the biological functions of differential alternative splicing events during seed germination using ClusterProfiler. A total of 18 genes, in which differential retained intron events were identified between Dry Seeds and Germinating seeds 3, were enriched on splicing process. By visualization of alternative splicing events in *AT4G25500*, which was done by using the rmats2sashimiplot tool, the inclusion levels and functions of differential alternative

splicing between different stages of seed germination were clearly illuminated (see Fig.2).

Overall, we investigated the dynamic changes of gene expressed levels and alternative splicing events during seed germination of *Arabidopsis* based on bioinformatics methods and dissected omics features and regulated mechanisms during seed germination from the perspective of gene expression and alternative splicing of pre-mRNA. This study is expected to be helpful for elucidating the regulated mechanisms of seed germination and to provide a guideline for relevant experimental works.

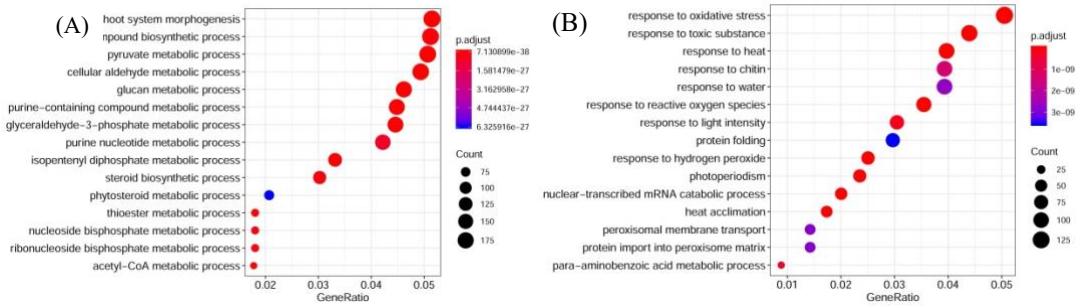

Figure 1 GO-BP enrichment analysis of differentially expressed genes between Dry seeds and Germinating seeds. (A) The top 15 enriched GO-BP terms ranked by *p-adjust* of up-regulated genes; (B) The top 15 enriched GO-BP terms ranked by *p-adjust* of down-regulated genes. GeneRatio of *x*-axis denotes the number of enriched genes associated with the given GO term divided by the total number of input genes. The size of the dot denotes gene number. The color denotes *p-adjust* measuring the differential level of gene expression.

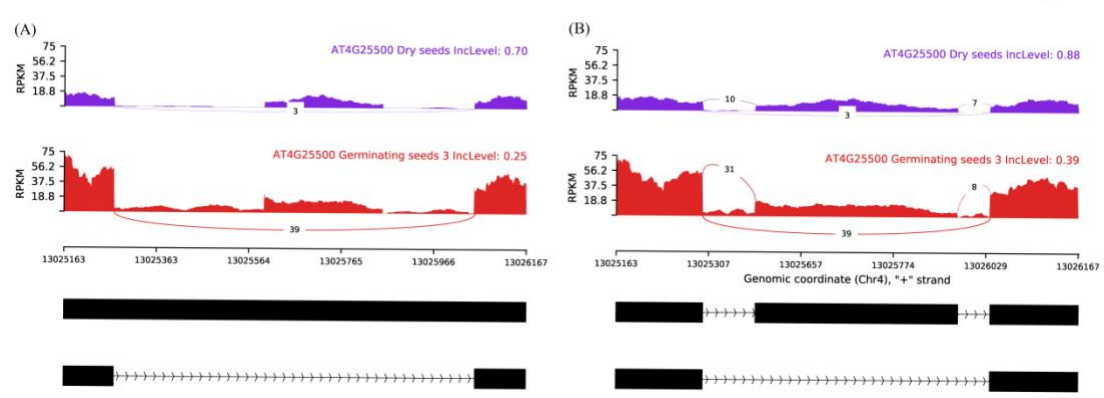

Figure 2 Differentially alternative splicing events of *AT4G25500* between Dry Seeds and Germinating seeds 3. (a) retained intron; (b) skipping exon

Funding: This work was supported by grants from the National Natural Science Foundation of China (61662055), the Natural Science Foundation of Inner Mongolia (2018MS03024), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-20-B05), and Inner Mongolia University of Science and Technology Innovation Fund for Excellent Young Scholars (2019YQL01).

关键词： *Arabidopsis thaliana*, Seed Germination, Differentially Expressed Genes, Alternative Splicing, Enrichment Analysis

报告人 Email： xingyongqiang1984@163.com

Biased Gene Retention during Diploidization in *Brassica* linked to Three-dimensional Genome Organization

Ting Xie [1](谢婷), Fu-Gui Zhang[1], Hong-Yu Zhang[2], Xiao-Tao Wang[3], Ji-Hong Hu[1] and Xiao-Ming Wu[1]

[1] *Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture and Rural Affairs, Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Wuhan 430062, China*

[2] *Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China*

[3] *Department of Biochemistry and Molecular Biology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA*

摘要： The non-random three-dimensional (3D) organization of the genome in the nucleus is critical to gene regulation and genome function. Using high-throughput chromatin conformation capture, we generated chromatin interaction maps for *Brassica rapa* and *Brassica oleracea* at a high resolution and characterized the conservation and divergence of chromatin organization in these two species. Large-scale chromatin structures, including A/B compartments and topologically associating domains, are notably conserved between *B. rapa* and *B. oleracea*, yet their KNOT structures are highly divergent. We found that genes retained in less fractionated subgenomes exhibited stronger interaction strengths, and diploidization-resistant duplicates retained in pairs or triplets are more likely to be colocalized in both *B. rapa* and *B. oleracea*. These observations suggest that spatial constraint in duplicated genes is correlated to their biased retention in the diploidization process. In addition, we found strong similarities in the epigenetic modification and Gene Ontology terms of colocalized paralogues, which were largely conserved across *B. rapa* and *B. oleracea*, indicating functional constraints on their 3D positioning in the nucleus. This study presents an investigation of the spatial organization of genomes in *Brassica* and provides insights on the role of 3D organization in the genome evolution of this genus.

关 键 词： *Brassica*, high-throughput chromatin conformation capture (Hi-C), 3D genome organization, gene retention, spatial colocalization

报告人 Email: xieting@caas.cn

Deep Learning for Motif Mining in DNA/RNA Sequences

黄德双

同济大学电子信息与工程学院

摘要：Recent biological studies have shown that binding-site motif mining plays a crucial role in the transcription and translation phases of gene expression, so the study of motif will help to understand the complex biomolecular system and explain disease pathogenesis. How to carry out an in-depth research on motifs through computational methods has always been one of the core issues in the modeling of life system gene regulation processes. In this report, we will systematically study and explore motif prediction of biological sequences in combination with the popular emerging technology “Deep Neural Networks”. Firstly, we will briefly introduce the development of deep neural networks and the research status of biological sequence motif mining. Secondly, we will discuss the existing shortcomings of deep-learning based motif prediction, and correspondingly propose a variety of improved motif mining methods, such as high-order convolutional neural network which employs a high-order encoding method to build high-order dependencies among DNA nucleotides, weakly-supervised convolutional neural network which combines multiple-instance learning with convolutional neural network, hybrid deep neural network which integrates DNA sequences and shape features. Finally, we will discuss the future researches from multiple perspectives.

报告人 Email：dshuang@tongji.edu.cn

Computational Prediction of RNA Tertiary Structures using Machine Learning Methods

Bin Huang, Yuanyang Du, Shuai Zhang, Wenfei Li, Jun Wang, Jian Zhang (张建)

National Laboratory of Solid State Microstructures, School of Physics, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210093
Institute for Brain Sciences, Kuang Yaming Honors School, Nanjing University, Nanjing 210093

摘要：RNAs play crucial and versatile roles in biological processes. Computational prediction approaches can help to understand RNA structures and their stabilizing factors, thus providing information on their functions, and facilitating the design of new RNAs. Machine learning (ML) techniques have made tremendous progress in many fields in the past few years. Although their usage in protein-related fields has a long history, the use of ML methods in predicting RNA tertiary structures is new and rare. Here, we introduce our recent work of using ML methods on RNA structure predictions and discuss the advantages and limitation, the difficulties and potentials of these approaches when applied in the field.

关键词：RNA structure prediction; RNA scoring function; Knowledge-based potentials; Machine learning; Convolutional neural networks

报告人 Email：jzhang@nju.edu.cn

Artificial Intelligence Biology: from PTM to COVID-19

Wanshan Ning[1], Chenwei Wang[1], Shaofeng Lin[1], Yu Xue[1](薛宇)

[1] *Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*

摘要：After introduction of machine learning in sequence analysis by Drs. Zhirong Sun and Yanda Li in 1999, application of artificial intelligence in biology has emerged to be a particularly active field in Bioinformatics. Recent advances in deep learning have provided a great opportunity to accurately infer the complex causality from big biological data, and boomed the establishment of a new interdisciplinary field named artificial intelligence biology (AIBIO). Here, we reported our progresses in AIBIO during the past two years. To predict post-translational modification (PTM) sites from protein sequences, we encoded PTM peptides into numerical data using a variety types of sequence and structure features. We compared conventional machine learning algorithms, such as penalized logistic regression (PLR), support vector machine (SVM), and random forest (RF), to an extensively used framework of deep neural network (DNN), and found that deep learning and conventional machine learning algorithms exhibited strikingly different advantages for representing distinct features. Thus, we implemented a new hybrid learning architecture by integrating DNN and PLR into a single framework, and developed a new tool named HybridSucc (http://hybridsucc.biocuckoo.org/), which achieved a ≥ 17.84% improvement of the area under curve (AUC) value (0.885 vs. 0.751) for the prediction of lysine succinylation sites than other existing methods. Furthermore, we designed a new method named number-to-image transformation (NIT) to transform numerical sequence data into imaging data, and implemented a hybrid learning framework consisting of parallel convolutional neural networks (pCNNs) and PLR algorithm. Using this architecture, we developed a new tool named GPS-Palm for predicting *S*-palmitoylation sites, with a >31.3% improvement AUC value against other existing tools (http://gpspalm.biocuckoo.cn/). Recently, we developed an engineering framework of Hybrid-learning for UnbiaSed predicTion of COVID-19 patients (HUST-19) to predict morbidity and mortality outcomes. We found the integration of computed tomography (CT) images and clinical features (CFs) achieved a striking accuracy for predicting different types of COVID-19 patients. Taken together, we anticipate our hybrid learning architecture can be a useful framework to integrate heterogenous big data, and believe AIBIO will be more attractive for both bioinformaticians and biologists.

关键词：artificial intelligence, AIBIO, machine learning, post-translational modification, COVID-19

报告人 Email: xueyu@hust.edu.cn

新的集成分类、降维策略与生物信息应用

邹权

电子科技大学

摘要：对多个确信度不高的结果进行汇集，往往通过多数战胜少数的策略可以得到优质的结果。但简单的计分、投票等规则无法被证明一定能够得到更优的结果。本次报告分别就特征排序、不平衡分类提出集成策略，并从不同的角度证明其优越性。最后将相关的继承策略应用于肿瘤侵袭性、蛋白质耐热性预测等方面，取得了较好的预测效果。

报告人 Email：zouquan@nclab.net

基于深度学习的微生物相关预测研究

骆嘉伟

湖南大学

摘要：微生物菌落系统的稳定和平衡是人体健康的保证，大量研究表明菌落系统的动态变化可能引起人体各种疾病的发生，因此微生物成为了干预治疗的最新靶标。充分利用海量的微生物数据，挖掘蕴含的有医学价值的信息，进而揭露微生物的功能机制，将有利于推动精准医疗、个性化诊疗的发展。近几年来，深度学习技术得到了快速发展，已经广泛应用于包括计算机视觉、语音识别、自然语言处理、生物信息学等诸多领域。如何有效整合微生物数据并合理地利用深度学习技术分析微生物生物学功能，是当前国内外一个极具挑战性的问题。我们将在分析微生物计算研究发展现状的基础上，与大家分享交流目前深度学习在识别微生物与药物、疾病的关联关系方面的相关研究工作。

报告人 Email：luojiawei@hnu.edu.cn

Functional Multimer Protein-protein Interaction Complex Structure Prediction by Machine Learning Approaches

Xinqi Gong[1,2](龚新奇)

[1] *Institute for Mathematical Sciences, Renmin University of China*

[2] *Beijing Advanced Center for Structural Biology, Tsinghua University*

摘要：On the basis of previous studies of protein-protein molecular docking, and in combination with the more multimer protein-protein complex structures that have been resolved due to the revolution of cryoelectron microscopy in recent years, my team has been focusing on the computation and prediction of functional multimer protein-protein interactions in recent years. We have mined the geometric features of multimer protein interactions, discovered new rules of amino acid pairing at the interface, and developed machine learning algorithms of convolutional neuronal network and graph neuronal network, for predicting the interface amino acid pairing and molecule packing of three, and more body protein interactions. We compared the differences of different interfaces in the multimer protein complexes, and found that some interfaces are important hot spots, which are easier to form and play more prominent roles than others. Additionally, we have developed a new method to distinguish which two proteins will interact or not interact among many proteins. Furthermore, we are integrating our multimer protein interaction prediction into large proteomics network to discover new biological understanding of the predicted interactions. Our machine learning approaches for multimer protein-protein interaction prediction will help to improve our knowledge about structure based life science and molecular/drug design.

关键词：multimer, protein-protein interaction, functional complex structure, machine learning algorithms, Convolutional Neuronal Network.

报告人 Email: xinqigong@ruc.edu.cn

Querying Heterogeneous Single-cell Transcriptomics Datasets via Adversarial Learning

Zhi-Jie Cao[1](曹智杰), Lin Wei[1], Shen Lu[1], De-Chang Yang[1], Ge Gao[1]

[1] *Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), and State Key Laboratory of Protein and Plant Gene Research at School of Life Sciences, Peking University, 100871 Beijing, China*

摘要：Single-cell transcriptomic profiling provides valuable insights for important biological questions like cellular function and gene expression regulation. Recent technological innovations lead to a rapid accumulation of single-cell transcriptomic data, which further enables data-driven cell annotation. However, multiple confounding effects like intra-/inter-dataset batch effect raise serious challenges for effective and efficient cross-dataset querying and integration. Herein, we designed Cell BLAST, a single-cell transcriptomics data querying method based on deep generative modeling. By introducing adversarial domain adaptation to correct for batch effect, as well as a posterior-based cell-to-cell similarity metric, we significantly improve the accuracy of cell querying over existing tools. Combined with a well-curated reference database ACA and a user-friendly Web server (https://cblast.gao-lab.org), Cell BLAST provides the one-stop solution for real-world scRNA-seq cell querying and annotation.

关键词：Deep generative modeling, scRNA-seq, cell querying, database

报告人 Email：gaog@mail.cbi.pku.edu.cn

INSCT: Integrating Millions of Single Cells using Batch-aware Triplet Neural Networks

Yin-Ying Wang[1](王银鹰), Lukas M. Simon[1,2], Zhongming Zhao[1,3,4]

[1] *Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[2] *Baylor College of Medicine, Therapeutic Innovation Center, Houston, TX, 77030, USA*
[3] *Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[4] *MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA*

摘要：Due to the minute amount of starting material, scRNA-seq data is prone to batch effects, which obstruct analysis of data generated by different laboratories, platforms, or experiments. Thus, the integration of heterogeneous, large-scale scRNA-seq data is one of the most critical challenges for large atlasing projects, such as the Human Cell Atlas. However, most of the existing methods are time-consuming and require large computational resources for the integration of increasingly frequent multi-million transcriptome datasets. To solve this issue, we developed a novel deep learning algorithm to overcome batch effects using batch-aware triplet neural networks, called INSCT ("Insight"). Using simulated and real data, we demonstrate that INSCT generates an embedding space which accurately integrates cells across experiments, platforms and species. Our benchmark comparisons with current state-of-the-art scRNA-seq integration methods revealed that INSCT outperforms competing methods in scalability while achieving comparable accuracies. Moreover, using INSCT in semi-supervised mode enables users to classify unlabeled cells by projecting them into a reference collection of annotated cells. To demonstrate scalability, we applied INSCT to integrate more than 2.6 million transcriptomes from four independent studies of mouse brains in less than 1.5 hours using less than 25 gigabytes of memory. This feature empowers researchers to perform atlasing scale data integration in a typical desktop computer environment. INSCT, including interactive usage tutorial, is freely available at https://github.com/lkmklsmn/insct.

报告人 Email：yingxiao8958@gmail.com

Higher-order Structure and Function of Noncoding RNA

薛愿超

中国科学院生物物理研究所

摘要：The human genome project revealed that at least 98% of the human genome does not encode proteins. This so-called noncoding part is pervasively transcribed and generated a large number of noncoding RNAs. Compared with mRNA, noncoding RNAs usually form intricate tertiary structures via intramolecular base-pairings and RNA-binding proteins. The highly structured noncoding RNAs further interact with other RNA molecules via intermolecular RNA-RNA interactions to regulate diverse essential biological processes. However, the in situ higher-order conformation and dynamic regulation of noncoding RNAs remain largely unknown.

By coupling proximity ligation mediated by RNA-binding proteins with deep sequencing, we recently developed an RNA in situ conformation sequencing (RIC-seq) technology for the global profiling of intra- and intermolecular RNA–RNA interactions. This technique not only recapitulates known RNA secondary structures and tertiary interactions but also facilitates the generation of 3D interaction maps of RNA in human cells. Using these maps, we identify noncoding RNA targets globally and discern RNA topological domains and trans-interacting hubs. We find that the functional connectivity of enhancers and promoters can be assigned using their pairwise-interacting RNAs. Unexpectedly, the super-enhancer long noncoding RNA CCAT1-5L can interact with MYC promoter- and enhancer-RNAs to regulate MYC oncogene's transcription by modulating long-range chromatin looping. I will share our current understanding of noncoding RNA structures and their functions in this meeting.

报告人 Email：ycxue@ibp.ac.cn

MicroRNA 集合分析：概念、方法与应用

崔庆华

北京大学

摘要：microRNA (miRNA)是一类重要的小的非编码 RNA 分子，在很多重要的分子通路扮演关键的调控角色，是疾病诊疗的新型分子。生物信息学在研究和探索 miRNA 在疾病中的作用发挥了重要作用。本报告将汇报 miRNA 集合分析的历史，从概念的提出，到生物信息方法与技术的建立，乃至在复杂疾病中的转化应用研究，从疾病相关 miRNA 的模式识别，到 miRNA 和疾病调控规律的模式发现，以及疾病诊断、预后新型标志物的挖掘，最后对未来做出展望。

报告人 Email：cuiqinghua@bjmu.edu.cn

Understanding Human Evolution in the Genomic Framework of Rhesus Monkey

Ni A. An[1], Wanqiu Ding[1], Qi Peng[1], Jie Zhang[1], Xiangshang Li[1], Chuan-Yun Li[1](李川昀)

[1] *Institute of Molecular Medicine, Peking University, Beijing, China*

摘要：We focus on the fundamental questions of human evolution, such as "what makes us uniquely human", from the perspectives of human-specific genes and regulatory events. Taking the dual-advantages of rhesus macaque as a central model animal and a species closely related to human, we identified 43 human-specific protein-coding genes and proposed a novel model for new gene origination, which states that protein-coding genes may generally emerge out of lncRNAs. Moreover, we clarified the mechanisms and evolution of primate regulations across multiple regulatory levels, such as alternative splicing, RNA editing and nucleosome occupancy, and identified a catalog of more than ten thousand human-biased regulatory events. Among these new genes and regulations, we found that a primate-specific, *de novo*-originated miRNA expands brain stem cells and induces *de novo* cortical folding that is particularly distinct in primates. Strikingly, ectopic expression of this microRNA in mice ultimately causes *de novo* cortical folding in the otherwise lissencephalic brain. Besides this primate-specific microRNA, we also found two human-specific *de novo* genes play essential roles in neocortex expanding. These findings provide novel clues to the understanding of human brain evolution.

报告人 Email：chuanyunli@pku.edu.cn

RNA Systems Biology Powered by Big Data and Machine Intelligence

张强锋

清华大学

摘要： Interactions with RNA-binding proteins (RBPs) are crucial for RNA regulation and function. While both RNA sequence and structure are critical determinants, RNA structure is dependent on cellular environment and especially important in regulating RBP binding changes in different conditions. However, how distinct it contributes to RBP binding in vivo remains poorly understood. In our study, we determined a rich dataset of transcriptome-wide RNA secondary structure profiles in multiple cell types by using icSHAPE, a cutting-edge technology for RNA structure probing. We observed a high level of association between RNA structural variable regions and variable RBP binding sites across cell types, suggesting RNA structure plays a critical role in determining RBP binding changes in specific cellular conditions. We thus developed a deep neural network, PrismNet, that includes in vivo RNA structure to model the RNA sequence and structural preferences of protein-RNA interactions in cellular conditions. PrismNet accurately predict RBP binding and the impact of genetic variants on RNA structure and RBP binding. The predicted binding sites are more conserved with harbored mutations tend to be more deleterious than non-binding sites. Remarkably, compared to genetic variants with stable RNA structures, structure-changing variants (riboSNitches) within the binding sites are more frequently associated with human disease.

报告人 Email： qczhang@tsinghua.edu.cn

细胞核仁中 lncRNA 的关键角色：从数据挖掘到分子机制

杨雪瑞

清华大学

摘要：A vast number of long noncoding RNA (lncRNA) species have been annotated during the past years. However, identification of physiologically relevant lncRNAs and in-depth investigation of lncRNA molecular functions remain challenging due to lack of prior knowledge and insights for generating testable hypotheses. We developed a multi-omics data-mining pipeline to search for the lncRNAs that take parts in shaping the largely shifted gene expression regulation programs in cancers. Our results prioritized the lncRNAs according to their regulatory potentials on the transcriptional regulation circuitry, in a cancer type-specific or pan-cancer manner. For example, guided by the lncRNA function survey in liver cancer, we have uncovered the mutual functional dependency between a previously untouched human long non-coding RNA, which we renamed LETN, and a key nucleolar protein, NPM1. Specifically, LETN plays a critical role in facilitating the formation of NPM1 pentamers, which are essential building blocks of the nucleolar granular component and control the nucleolar functions. Repression of LETN or NPM1 led to identical and strong shift of the nucleolar morphology and arrest of the nucleolar functions, which led to proliferation inhibition of human cancer cells and neural progenitor cells. Interestingly, this inter-dependency between LETN and NPM1 is associated with the evolutionarily new variations of NPM1 and the coincidental emergence of LETN in higher primates. Therefore, this human-specific protein-lncRNA axis renders an additional yet critical layer of regulation with high physiological relevance in both cancerous and normal developmental processes that require hyperactive nucleoli.

报告人 Email：yangxuerui@tsinghua.edu.cn

Multi-omics Annotation of Human Long Non-coding RNAs

Qianpeng Li[1,2,3,4], Zhao Li[1,2,3,4], Changrui Feng[1,2,3,4], Shuai Jiang[1,2,3], Zhang Zhang[1,2,3,4], Lina Ma[1,2,3](马利娜)

[1] *China National Center for Bioinformation, Beijing 100101, China*

[2] *National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

[4] *University of Chinese Academy of Sciences, Beijing 100101, China*

摘要：Long non-coding RNAs (lncRNAs), a class of non-coding transcripts longer than 200 nucleotides, have emerged as key regulators of multiple essential biological processes involved in physiology and pathology. A high-quality and comprehensive lncRNA annotation is essential for subsequent functional investigation of human genome. However, the relatively low and tissue-specific expression and low conservation of lncRNAs as well as the enormous quantity have challenged the annotation of lncRNAs. The immense amount of omics data is generated at an unprecedented rate by high-throughput sequencing technologies, providing possibilities in the large-scale collection and annotation of lncRNAs. Here, we underline the specialized omics resources dedicated to lncRNA research and present the integrative analysis pipeline of various omics data including expression, DNA methylation, genome variation and miRNA interaction. Moreover, we highlight the multi-omics integrative analysis as a powerful strategy to efficiently discover and characterize the functional lncRNAs and elucidate their potential molecular mechanisms.

关键词：long non-coding RNA, multi-omics data, integrative annotation

报告人 Email：malina@big.ac.cn

ncRFP: a Novel End-to-end Method for Non-coding RNA Family Prediction Based on Deep Learning

Linyu Wang[1,2], Shaoge Zheng[1,2], Hao Zhang[1,2], Zhiyang Qiu[1,2], Xiaodan Zhong[1,2], Haiming Liu[1,2], Yuanning Liu[1,2](刘元宁)

[1] *College of Computer Science and Technology, Jilin University*
[2] *China and Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University*

摘要：Evidence has accumulated enough to prove non-coding RNAs(ncRNAs) play important roles in cellular biological processes and diseases pathogenesis. High throughput techniques have produced a large number of ncRNAs whose functions remain unknown. Since the accurate identification of ncRNA family is helpful to the research on their functions, it is of necessity and urgency to predict the family of each ncRNA. Although several methods are applicable to prediction, their complex procedures or inaccurate performance remain major problems confronting us. The main idea of those methods is first to predict the secondary structure, and then identify ncRNAs according to properties of the secondary structure. Unfortunately, the inaccuracy of RNA secondary structure prediction tools, which may lead to the low accuracy. Acknowledge the crucial roles of ncRNAs in cellular processes and weaknesses of those methods, it is required to develop a new method so as to simplify and precisely predict the family of each ncRNA. In this paper, a novel method 'ncRFP' is proposed to complete the prediction task based on Deep Learning. Instead of predicting the secondary structure as other methods do, ncRFP directly predicts the family of ncRNA by automatically extracting features from ncRNA sequences.

关键词：ncRNAs, Deep Learning, secondary structure, ncRFP, ncRNA sequences

报告人 Email: liuyn@jlu.edu.cn

Prediction of Plant Long Non-coding RNA and Its Function Analysis

YAN Lingjuan[1](闫玲娟), CHEN Yingli[1], FAN Zhiyu[1], YAN Dongxue[1]

[1] *School of Physics and Technology, Inner Mongolia University, Hohhot 010021, China*

摘要：Long non-coding RNA (lncRNA) is a type of RNA transcript defined as having a length greater than 200nt and no protein coding ability. Previously thought to be dark matters of the genome, lncRNA have been gradually recognized as crucial gene regulators. Compared with human and animals, the transcriptomic identification of lncRNA in plants is still in its infancy. So far, the accurate identification of lncRNA is still one of the main problems in the field of plant research.

We used bioinformatics methods to predict plant lncRNA, it mainly includes the following aspects, constructed a new plant lncRNA and mRNA data set, analyzed the sequence and structural features of the plant lncRNA in the data set. and extracted the k-mer frequency information, secondary structure, open reading frame and geometric flexibility information of the sequence. In addition, based on the SVM algorithm, with a Jackknife test on plant lncRNA is predicted, and calculated the fusion of various features of plant lncRNA predicted results, the accuracy reached 96.14%.

By analyzing the sequence characteristics of plant lncRNA, it was found that plant lncRNA was rich in A and U, while mRNA was rich in C and G, and the frequency of AA/AU/ UA /UU of plant lncRNA was significantly higher than that of mRNA. The 4-mer sequence features of lncRNA were extracted as the input vectors of SVM to identify plant lncRNA, the accuracy reached 93.36%. Geometric flexible information feature of the sequence were extracted to prediction, there are two parameters ω and λ, when λ is 5 and ω is 0.1, the result is relatively good up to 85.56%. Considering the limitations of the traditional plant lncRNA identification based on a single feature. Therefore, several features with better information parameters were fused to predict, and the results were found to be improved to a certain extent, found that the relative length of the longest open reading frame and the 4-mer component information are relatively important feature information in the prediction of plant lncRNA.

Although more and more lncRNA transcripts have been identified, the biological function of most lncRNA remains unclear. Because sequence conservation are accepted as indicators of biological function, the analysis of the conserved fragments of plant lncRNA sequences can provide valuable sequences for the functional study of plant lncRNA.

关键词：Plant lncRNA; Feature extraction; Support Vector Machine; conservation

报告人 Email: 1763491452@qq.com

人类新蛋白质组的发现

Shaohua Lu, Jing Zhang, Xinlei Lian, Li Sun, Kun Meng, Yang Chen, Zhenghua Sun, Xingfeng Yin, Yaxing Li, Jing Zhao, Tong Wang(王通), Gong Zhang and Qing-Yu He

Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, Guangzhou 510632, China

摘要：It has been a long debate whether the 98% 'non-coding' fraction of human genome can encode functional proteins besides short peptides. With full-length translating mRNA sequencing and ribosome profiling, we found that up to 3330 long non-coding RNAs (lncRNAs) were bound to ribosomes with active translation elongation. With shotgun proteomics, 308 lncRNA-encoded new proteins were detected. A total of 207 unique peptides of these new proteins were verified by multiple reaction monitoring (MRM) and/or parallel reaction monitoring (PRM); and 10 new proteins were verified by immunoblotting. We found that these new proteins deviated from the canonical proteins with various physical and chemical properties, and emerged mostly in primates during evolution. We further deduced the protein functions by the assays of translation efficiency, RNA folding and intracellular localizations. As the new protein UBAP1-AST6 is localized in the nucleoli and is preferentially expressed by lung cancer cell lines, we biologically verified that it has a function associated with cell proliferation. In sum, we experimentally evidenced a hidden human functional proteome encoded by purported lncRNAs, suggesting a resource for annotating new human proteins.

报告人 Email: tongwang@jnu.edu.cn

Multi-omics Analysis Defines Biomarkers for Renal Aging

Meiqi Yi, Yingying Ma, Songbiao Zhu, Yuling Chen, Haiteng Deng(邓海腾)

MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systematic Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

摘要：Kidney aging is one of four aging types in elderly population. Characterization of molecular signatures in kidney aging provides valuable information for understanding molecular mechanisms underlying aging as well as for identification of aging biomarkers. In the present study, we profiled age-associated changes in proteome, glutathionylome and epigenome in mouse kidneys. We defined 27 proteins as biomarkers for renal aging. We further revealed that peroxisomal biogenesis were significantly decreased in aged mice, suggesting that peroxisome deterioration was a hallmark for renal aging. Downregulation of catalase and GRX1 resulted in significant increase in protein glutathionylation in aged mice as revealed by glutathionylome analysis. We further uncovered that nicotinamide mononucleotide (NMN) supplementation increased peroxisome biogenesis and restored the proteome homeostasis. Histone modification profiling revealed epigenetic signatures associated with human kidney aging. Our data provide a valuable resource for understanding the age-associated changes in kidneys.

关键词：renal aging; multi-omics; biomarkers; glutathionylation; NMN

报告人 Email: dht@mail.tsinghua.edu.cn

Using Structural Analysis to Explore the Role of HBV Mutations in Immune Escape from Liver Cancer in Chinese, European and American Populations

Shuanglin Gu[1], Lv Li[1], Xue Lin[2], Xingyu Li[1], Juncheng Dai[2], Jianqiong Zhang[1], Ren Kong[3], Wei Xie[1], Jian Li[1](李健)

[1] *Key Laboratory of DGHD, MOE, Institute of Life Sciences, Southeast University, Nanjing 210096, China*

[2] *Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China*

[3] *Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China*

摘要：HBV infection is an important problem threatening human health. After HBV virus invades human body, it may assemble a complete virus particle in the cytoplasm to trigger the immune reaction, especially the interaction between the HBV virus and the host that mediated by $CD8^+$ T cell. We collected the sequences of HBV from the HBVdb database, then screened candidate mutation sites in Chinese, European and American populations based on conservation and physicochemical properties. After that we constructed the three-dimensional structure of MHC I-peptide complexes, performed molecular docking, run molecular dynamics to compare the binding free energy, stability, and affinity of MHC I-peptide complexes with the aim to estimate the effect of peptide mutation. We analyzed the specific HBV virus subtypes of the Chinese, the European and American population and candidate mutation sites was used to predict the mutant antigen peptide. Finally, based on physical and chemical properties and antigen peptide prediction scores, 21 HBV mutation sites were selected. Then combined with specific HLA subtypes, 11 mutations were found to have a significant negative impact on affinity, stability and binding free energy. Overall our work found important potential mutations, which provide an evaluation of HBV mutations and a clue of it in immunotherapy.

关键词：HBV, stability, affinity, immune escape, molecular dynamics

报告人 Email：jianli2014@seu.edu.cn

A Completely Designer Chromosome Arm Functions in Yeast

Junbiao Dai(戴俊彪)

Guangdong Provincial Key Laboratory of Synthetic Genomics, Shenzhen Key Laboratory of Synthetic Genomics and Center for Synthetic Genomics, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

摘要：Facilitated by advances in technologies for DNA synthesis and assembly, the entire genome of an organism, from viruses to bacteria, has now become the target of redesign and reprogramming. The synthesis of first eukaryote, *Saccharomyces cerevisiae*, genome (Sc2.0) is near completion, tackled by an international consortium. Sets of design features were grafted into the synthetic genomes for future application including, for example, the integration of water markers or PCRtags and elimination of one or several codons. However, it is still questionable if we can construct a functional synthetic genome with no or minimal homologs to that of the native one. In this talk, I will first briefly summarize findings during the construction of synthetic chromosome XII (synXII) as part of the Sc2.0 project. Next, I will introduce how can we simplify the left arm of synXII by combining SCRaMbLE and targeted chromosome deletion. Finally, we showed that a neochromosome could be constructed using completely refactored sequences, which is able to functionally replace the native chromosome arm.

报告人 Email：junbiao.dai@siat.ac.cn

基因调控元件的人工智能设计

汪小我

清华大学

摘要：近年来，基因的编辑与合成等生物技术取得突破，使得我们从最底层设计和构建生物系统成为可能。将细胞作为控制对象，将极大地扩展智能控制研究的内涵。人工智能技术与合成生物学的交叉具有颠覆性，未来将可能对促进代谢工程、分子育种、基因治疗等领域的发展产生深远影响。如何发展智能技术实现人工生物系统的优化设计和精确控制是我们面临的挑战与机遇。

我们近期尝试用深度学习模型来设计和产生全新的基因启动子。过去，人工元件的获取主要通过对自然元件的简单改造，例如通过对天然序列的随机突变、功能片段拼接组合等方法，结合定向进化等实验来筛选获得新的元器件。这些方法一方面成功率低，另一方面通常只能获得与天然序列非常相似的元件，难以发现全新的调控元件。我们尝试将人工智能对抗学习与合成生物实验闭环耦合，采用数据与知识双驱动的策略解析基因转录调控复杂编码模式，进而用机器学习方法优化设计获得大量全新的调控元件，完成从物理-虚拟-物理世界的迭代映射过程。从实践上探索了用 AI 设计基因调控元件的可行性，用人工智能优化设计大大提高生物实验筛选效率。对推动工程生物系统更加高效、安全、可控的智能化设计与构建具有参考意义。

报告人 Email：xwwang@tsinghua.edu.cn

可预测组装调控元件的设计原则

娄春波

中科院深圳先进技术研究院

摘要：在微生物生命系统中，除了各种代谢途径所需的酶以外，还存在各种层次调控和反馈控制系统。这些层次包括 DNA 复制、转录、翻译、转录/翻译后修饰、降解等等。这些调控系统可以维护生命体的各种代谢物和无机物的体内平衡；也可以控制单个细胞分化成复杂的多细胞个体。但是现有调控元件多数是保留自然界进化过程中“补鞋匠”痕迹，不能满足人工设计生命系统的需要。因此，如何设计优质调控元件成为合成生物学健康发展的重要障碍之一。在本报告中，我将在基因元件模块化设计原理、正交化拓展以及相关元件深度挖掘等问题，展开一系列探讨。

报告人 Email：louchunbo@gmail.com

整合临床表型和基因组变异信息筛选罕见遗传病

Weidong Tian(田卫东)

Fudan University

摘要：Whole exome sequencing (WES) has become widely used in clinical practice of the diagnosis of the causal genes of Mendelian diseases. In order to prioritize disease-causing variants called from WES data, inspection of a patient's clinical phenotypes that are usually translated into the Human Phenotype Ontology (HPO) terms is necessary. In this study, we introduce a probabilistic approach for prioritizing diseases from a patient's phenotypes (HPO terms). We further develop a tool called PhenoPro that prioritizes the causal gene of Mendelian disease given both the HPO terms and the variant data of a patient. PhenoPro shows significant improvements over previously developed tools in both simulated and real patient data. To make PhenoPro fully automated, we also include a natural language processing (NLP) component that automatically assigns HPO terms based on a patient's clinical report, and demonstrate that NLP is as effective as manual HPO assignment by using real clinical reports. As such, PhenoPro is of great use as a pre-screening tool to assist in the diagnosis of Mendelian disease genes.

报告人 Email：weidong.tian@fudan.edu.cn

基于网络数量性状位点的基因型-表型关联模型

Kai Yuan[1], Tao Zeng[1,2,3](曾涛), Luonan Chen[1,3,4,5]

[1] *Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

[2] *Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences.*

[3] *Institute of Brain-Intelligence Technology, Zhangjiang Laboratory, Shanghai 201210, China*

[4] *School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China*

[5] *Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.*

摘要：Many of the conventional eQTL methods are using network concept or model to interpret the biological or biomedical significance of their discovery, however, they don't consider the high-order associations between SNP and gene-pair/edge groups (networks). Thus, many network-associated phenotypic determinates would be disregarded. To address such issue, we propose network QTL (nQTL) to identify the cascade association of genotype -> network -> phenotype rather than traditional genotype -> expression -> phenotype. We have implemented a nQTL framework to study the associations from genotypes to networks and further to phenotypes at a system level on the basis of single-sample network theory and method, which provides gene-pairs' correlation data transformation, edge trait detection, edge/network signatures identification, and cascade association reconstruction. Both the simulation studies and real data case studies demonstrate the efficiency of nQTL compared to traditional eQTL. Especially, the case study of nQTL on both healthy human bulk and scRNA-seq data has identified immune associated edge traits, network signatures and post-targeted phenotypes. Collectively, all results support that nQTL can reveal significant functional impacts of particular genotype on biology and medicine.

报告人 Email：zengtao@sibs.ac.cn

Hierarchical Segmentation based 3D Facial Genetics Analysis

Jairui Li[1,2,3](李嘉睿), Siyuan Du[3], Manfei Zhang[3,5], Sijie Wu[3,4,5], Wei Qian[3,5], Sijia Wang[3,5]

[1] *Medical Imaging Research Center, MIRC, University Hospitals Leuven, Leuven, Belgium*

[2] *Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium*

[3] *CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

[4] *State Key Laboratory of Genetic Engineering and Ministry of Education, Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200438, China*

[5] *Human Phenome Institute, Fudan University, 825 Zhangheng Road, Shanghai 201203, China*

摘要：Phenomics investigates high-dimensional phenotypic data on an organism-wide scale, which requires special statistical models and efficient algorithms. The human face is a typical complex and high dimensional phenotype. Methods and algorithms developed for the human facial shape are applicable to phenomic studies using high-dimensional data. As illustrated by population and sex differences, family resemblances, and identical twins, the human face is highly heritable, and the genetic architecture of human facial shape has received increasing attention. In the recent decade, 3D imaging techniques have been widely used in craniofacial genetic studies to characterize facial morphology. Like the increasing dimensionality used in phenomic studies, many facial studies evolve from sparse configuration of facial shape using limited number of landmarks to spatially dense landmarking to investigate the human facial shape variation.

The high dimensionality of human facial shape requires special statistical models and efficient algorithms for many genetic analyses. The hierarchical organized facial phenotyping was succeeded in identifying hundreds of independent signals associated with the global and local facial morphometrical features. The hierarchical facial modules were defined by a hierarchical spectral clustering algorithm using the phenotypic correlation matrix between landmarks as the similarity matrix. On each module, we used canonical correlation analysis (CCA) to define the linear combination of the facial segment PCs that are mostly correlated with each SNP. The correlation is tested for significance based on Rao's exact F-test. Despite the phenotypic correlation, we also used the genetic correlation and environment correlation to define the hierarchical facial modules. Utilizing the hierarchical facial segmentation, our GWAS on a Han Chinese cohort identified 246 genome-wide significant loci associated with the facial shape variation. The hierarchical facial spectral clustering was performed on a combined matrix of phenotypic similarities and Euclidean

distances between landmarks to prevent the isolated points and discrete regions.

Depending on the cohort, the facial mask, and the similarity matrix used, data-driven hierarchical facial phenotyping produces different facial segments across studies. The difference between the facial segmentations leads to unreplicable results. For example, there is a segment highly overlapped with the zygoma area in Chinese cohort, while there is no corresponding segment can be found in the European cohort. In addition, there are still many associated variants found in the studies using distance-based traits which cannot be replicated in the facial segment-based studies. Finally, unsupervised data-driven facial segmentation could produce facial segments with isolated landmarks or discrete parts without any biological meaning. It is possible to define the shape by pairwise distances between all the landmarks and construct a PCA space on these distances. The pairwise distances between landmarks are affine-invariant and thus invariant to generalized Procrustes analysis. In the literature, only the distances between the spatially sparse landmarks on either 2D or 3D facial shapes were studied. The facial phenotyping based on the distances between the spatially dense landmarks is a plausible future direction for 3D facial phenotyping to investigate. On the other hand, to utilize related individuals, it is possible to use unrelated samples to define the CCA direction for an SNP and then project all the samples (including related individuals) onto that direction and perform a univariate LMM fixed effect test.

关键词： Phenomics, Craniofacial Shape, Hierarchical Segmentation, GWASs

报告人 Email: lijiarui@picb.ac.cn

大数据时代的脑科学：张江国际脑库介绍

赵兴明

复旦大学

摘要：近年来，伴随着测序、光遗传、钙成像和脑影像等各类技术的快速发展，围绕脑科学的不同尺度和模态的海量数据开始不断涌现。在数据科学时代，数据正成为理解大脑，阐述脑疾病发生机制和发展类脑智能的关键，也逐渐成为神经科学理论的基石。在此背景下，复旦大学正在建设全维度脑科学数据平台-张江国际脑库，以期为理解脑功能机制和脑疾病发生机理提供帮助。在该报告中，我将介绍张江国际脑库的建设情况和相关研究进展。

报告人 Email：xmzhao@fudan.edu.cn

Human Brain Evolution: Insights from Transcriptome Analysis

诸颖

复旦大学

摘要：The human brain is about three times as large as those of our closest living relatives, the nonhuman African great apes. However, the increased size and neural cell counts alone fail to explain the human-specific aspects of behavior, cognition and disorders, and the precise molecular mechanisms underlying shared and unique features of the developing human nervous system have been only minimally characterized. Our integrative analysis of tissue and single-cell transcriptomic data revealed diverse molecular and cellular features of the phylogenetic reorganization of the human brain across multiple levels, with relevance for brain function and diseases.

报告人 Email：ying_zhu@fudan.edu.cn

Germline OGDHL Variant Could be a Major Driving Genetic Factor to Cause Chinese Familial Major Depression Disorder

Jun Du[1], Pan You[2], Wen-Qiang Wang[2], Zhi-Liang Ji[1](纪志梁)

[1] *State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, Fujian, P R China*

[2] *Xiamen Xianyue Hospital, Xianyue Road 387-399, Xiamen 361012, Fujian, P R China*

摘要：Depression disorder is a severe psychiatric and social problem that affects more than 4% of global population. Major depression disorders (MDDs) exhibit significant hereditary properties; however, the exact driving genetic force largely remains unclear. In the study, we recruited a three-generation Chinese pedigree in which 5 out of 17 members suffered long-term MDDs. We conducted the whole exome sequencing (WES) to portray the genetic mutation profiles for the family, upon which we identified a list of susceptible genetic variations highly associated with the MDD onset. In particular, a non-synonymous single nucleotide variation (SNV) on the oxoglutarate dehydrogenase like gene (ODGHL) may cause the significant protein structure change, abnormal decrease of glutamate, and eventually the low mood. Further brain image analyses unveiled that the ODGHL variant might also answer for the significant volume reduce of the amygdala. In summary, this work may improve insights into the underlying biological processes of MDDs. It also provides valuable clues in planning relevant therapeutic interventions for MDD patients.

关键词：Major Depression Disorder; Whole exome sequencing; Brain image

报告人 Email: appo@xmu.edu.cn

会议主办：

中国生物信息学学会（筹）

会议协办：

清华大学生物信息学教育部重点实验室
中国细胞生物学学会功能基因组信息学与系统生物学分会

会议承办：

同济大学、复旦大学