北京生科院开发出菌群大数据挖掘及标识物识别新技术

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/18493.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

近年来，肠道菌群与人体健康和疾病的关系受到关注，该领域相关研究持续增长，积累了大量的数据信息。这些菌群数据类型复杂、数量庞大，但蕴含着研究和应用价值。如何从海量零散的数据中挖掘出与人类健康和疾病紧密相关的信息，这成为肠道菌群领域关注的首要问题。一方面，肠道微生物的种类和丰度与基因表达阵列不同，更易受到饮食、地域等各种因素的影响，不同人群之间肠道菌群组成可能存在较大差异，这使得直接基于菌群丰度整合数据并进行生物标识物挖掘的方法必然产生偏差；另一方面，微生物丰度矩阵过于稀疏，常规计算方法也很难基于此稀疏矩阵对不同批次的丰度进行校正。因此，建立更为高效的算法模型来解读复杂菌群数据的内在规律，显得迫切和重要。

中国科学院北京生命科学研究院赵方庆团队开发了一种高效的菌群大数据整合与标识物识别的新算法——NetMoss。5月23日，相关研究成果以*Large-scale microbiome data integration enables robust biomarker identification*为题，在线发表在*Nature Computational Science*上。该算法针对不同菌群数据集的高异质性特点，利用微生物互作网络对不同来源的数据进行有效整合，通过比较不同状态下微生物网络的扰动情况，量化不同网络模块间的拓扑结构差异，从而实现对疾病标识物的识别。与既往方法相比，NetMoss可以更高效地对不同批次的微生物组大数据进行无偏整合，挖掘与疾病相关的菌群标识物，并识别出驱动多种疾病发生的菌群失调共变模式。

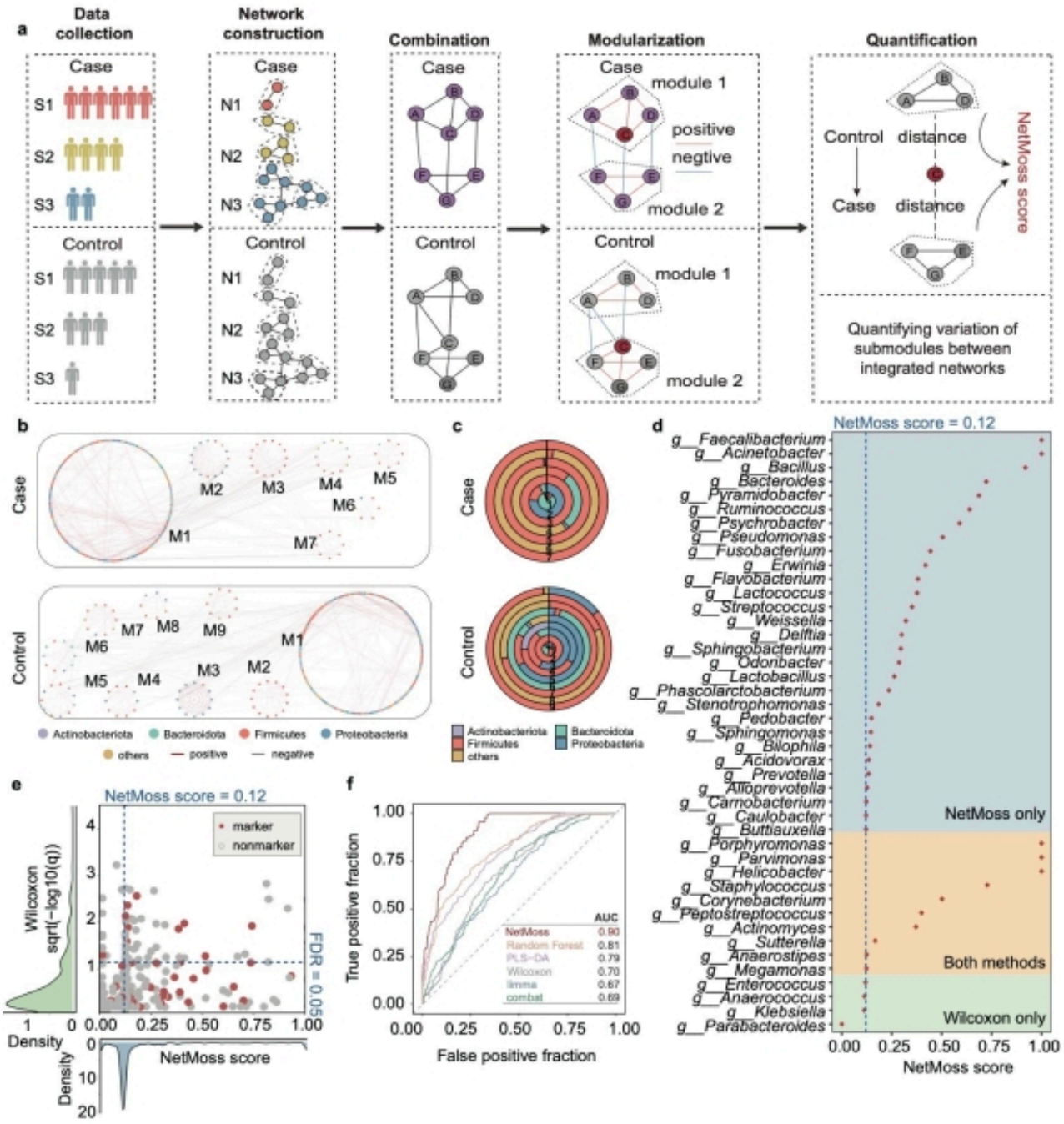
科研人员收集整理了11,377例包括疾病与健康对照的肠道菌群测序样本，覆盖78项研究、37种疾病、13个国家或地区。针对这些不同人群的多种数据集，研究发现目前常用的计算方法，极难去除实验和测序过程中导致的批次效应。为了有效地进行后续分析、避免偏倚性，研究开发了一种高效的数据整合和生物标识物挖掘的计算模型（如图）。该模型以微生物互作网络作为理论基础，在每个子数据集中单独构建微生物互作网络。而后，依据微生物网络的结构特点，为每个网络赋予不同权重再加以整合。通过此方式，每个子数据集中最原始的生物互作信息都可以得到有效保留，减少了不同批次对最终整合数据的影响。对疾病相关的生物标识物的识别，则是基于整合后的微生物网络进行。根据网络中不同微生物的互作关系将网络划分出不同的模块，通过量化疾病和健康网络中不同模块的拓扑结构差异，找出对外界影响扰动最敏感的细菌，对这些细菌在网络中的扰动程度进行打分，从而识别出与疾病发生发展密切相关的细菌。研究将该计算模型应用于模拟和真实数据集，发现该模型具有很高的准确性和鲁棒性，无论是在整合后的数据集还是在单一数据集中，其对疾病相关标识物的识别效率均高于其他方法。研究进一步发现，大部分疾病标识物不只单单导致一种疾病的发生，而是与多种疾病存在显著关联；这些相似的菌群失调现象

可能为不同疾病的共性致病机理提供重要线索。

该研究首次提出基于微生物互作网络的新算法，实现对大规模菌群数据的高效整合与疾病相关标识物的精准识别。基于庞大的微生物互作网络，可以挖掘出一些丰度尚未发生显著改变，但在生态互作网络中已被扰动的关键微生物类群。关注这些在生态互作中发生改变的细菌，为疾病的机制研究提供了新线索和关键靶标。通过研究它们在不同系统之间的作用，将有助于理解菌群与宿主互作的本质，也可以更好地指导我们对多种疾病进行预防和治疗。将肠道菌群及它们内部的互作关系作为整体来看待，能够捕捉到更多以往由于孤立研究而被忽视的信息，这种全面系统的整合分析的思维模式也为其他方向的研究提供了重要启示。

研究工作得到国家杰出青年科学基金、国家重点研发计划和中科院战略性先导科技专项。赵方庆团队致力于建立高效的算法模型和实验技术，探索人体微生物与非编码RNA的结构组成与变化规律，以期解析它们与人类健康和疾病的关系，相关成果发表在*Cell*（2020）、*Gut*（2022/2020/2018）、*Nature Biotechnology*（2021）、*Nature Computational Science*（2022）、*Nature Communications*（2022a/2022b/2021/2020/2017/2016）、*Genome Biology*（2021/2020/2016）、*ISME J*（2019）等上，这丰富了我们对人体微生物与非编码RNA多样性、结构组成与功能的认识，并为相关数据挖掘及功能机制研究提供了重要的方法学工具。

论文链接



NetMoss算法及CRC相关菌群标识物的挖掘

研究团队单位：北京生命科学研究院

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发