
深圳先进院提出微生物组数据挖掘新方法tmap

作者：writer 来源：中国科学院

本文原地址：<https://www.iikx.com/news/progress/8057.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！

近日，中国科学院深圳先进技术研究院合成生物研究所赵国屏课题组（周豪魁为通讯作者，廖天华为第一作者）在微生物组学数据挖掘领域取得新突破，相关论文 *tmap: an integrative framework based on topological data analysis for population-scale microbiome stratification and association studies*（《tmap 微生物组大数据挖掘方法开发——基于拓扑学数据分析的人群分层及宿主表型关联分析》）在线发表在期刊 *Genome Biology* 上。

随着高通量多组学技术的发展，传统的数据分析方法难以在日益复杂、规模庞大的微生物组数据中挖掘出有效可靠的信息，在数据集庞大、样本之间存在各种复杂差异性，微生物组和宿主表型特征之间存在的细微关联可能是研究目的所在，但是极易被掩盖掉。这是对数据分析工作的挑战，也对开发创新性分析方法提出迫切需求。

本课题组结合mapper算法和网络富集分析，开发基于拓扑学数据分析(topological data analysis, TDA)的 *tmap* 软件，提供了整合大规模人群微生物组数据的分析框架。 *tmap*

将不同国家地区、不同疾病类型和不同宿主表型的人群微生物组大数据加以整合，融合不同来源的微生物组及其宿主表型的医疗健康大数据，是目前微生物组数据共享和整合的创新性方法。这对大规模人群队列的数据分析而言，不仅归纳总结了数据隐含的规律，也对科学假设的提出，科学问题的提炼有很大帮助。

本研究中 *tmap*

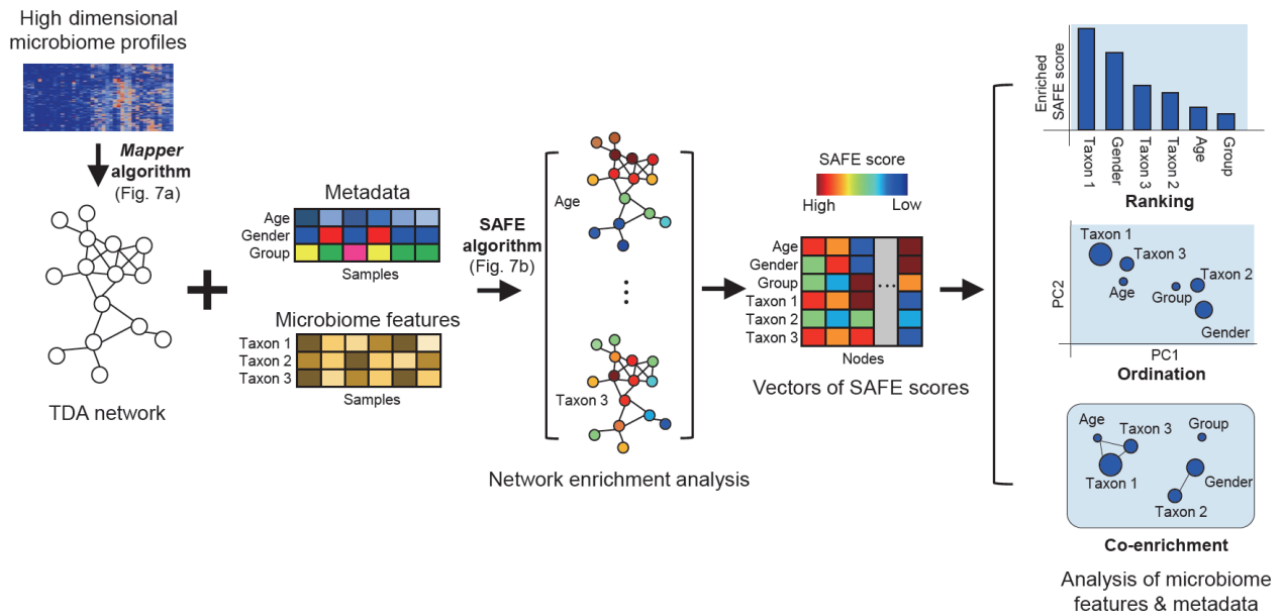
的大致工作流程是将高维微生物组数据转化为拓扑学网络结构，并用于微生物组人群分层及相关性分析。通过Mapper算法及网络分析， *tmap* 利用样本的微生物组数据构建TDA网络结构，并通过SAFE score量化了样本表型或者物种在一簇样本中的富集程度。基于SAFE score可以实现以下3个分析目的：一是Ranking，即对每个表型或物种的SAFE score之和排序，其含义与linear regression中的effective size类似，但是对线性及非线性的关联都可以较好地识别；二是Ordination，对SAFE score做PCA，目的为展示一个整体富集规律（overall enrichment pattern）不仅可以识别对微生物组变化贡献最大的物种或表型，也可以了解哪些表型或物种的联系比较紧密；三是Co-enrichment network，这里的网络基于物种及表型关联的显著性构建，用于探索其生物学意义上的关联或者混杂因素（confounder），需要以研究目的和意义为导向。

另外，本研究通过对4个数据集（包括1个合成数据和3个真实世界数据）对*tmap*方法加以验证，结果表明与传统的线性方法（如*envfit*, *adonis*, *ADONISM*）相比，*tmap*都表现更好，对非线性关联及潜在规律都有较好的识别。以比利时队列FGFP为例：2016年发表在*Science*，研究内容为健康正常人群的肠道微生物组。研究纳入了1106个粪便16S rRNA样本，及样本的69个宿主表型特征，包括了7个metadata类别：社会人口学特征，生活方式，血液指标，健康状况，饮食习惯，肠道健康状况及病史用药情况。文章表明，仅有7.63%的微生物组改变与宿主表型的变化有关。*tmap*提高了识别并解读population-scale的微生物组-宿主表型关联的能力。如下图所示，*tmap*不仅可以较好地复现*envfit*这种基于线性回归的方法，也在对非线性变化的捕捉上，比*envfit*更加灵敏。通过对FGFP数据转化的SAFE score分析，可以得到许多原始文章中没有挖掘到的规律和线索。如下图所示，比如两种药物虽然与同一种疾病相关联，但其对微生物组的影响是不同的，具体表现为在网络图上的富集部位不一样。这都是传统线性分析难以得到的。

总的来说，*tmap*有助于将不同国家地区、不同疾病类型、不同宿主表型的人群微生物组大数据加以整合，对人群做精细分层，为特异性的干预性治疗、数据来源多样化的meta分析等研究提供了有效线索。有效融合不同来源的微生物组及其宿主的医疗健康大数据，是实现微生物组数据共享和整合的创新性方向。其次，*tmap*不仅可以胜任传统线性研究对线性关系的刻画，也对非线性关联灵敏，优于传统分析方法。且基于微生物组网络地图，可以挖掘基于微生物组变化背景的宿主表型关联，使表型之间的关系不会与微生物组变化割裂开来。最后，了解不同宿主表型在网络图上的分布，捕捉变量之间联系及变量自身的重要性排序，对大规模人群队列的数据分析，不仅归纳总结了现有数据规律，也对科学假设的提出，科学问题的提炼有很大帮助。

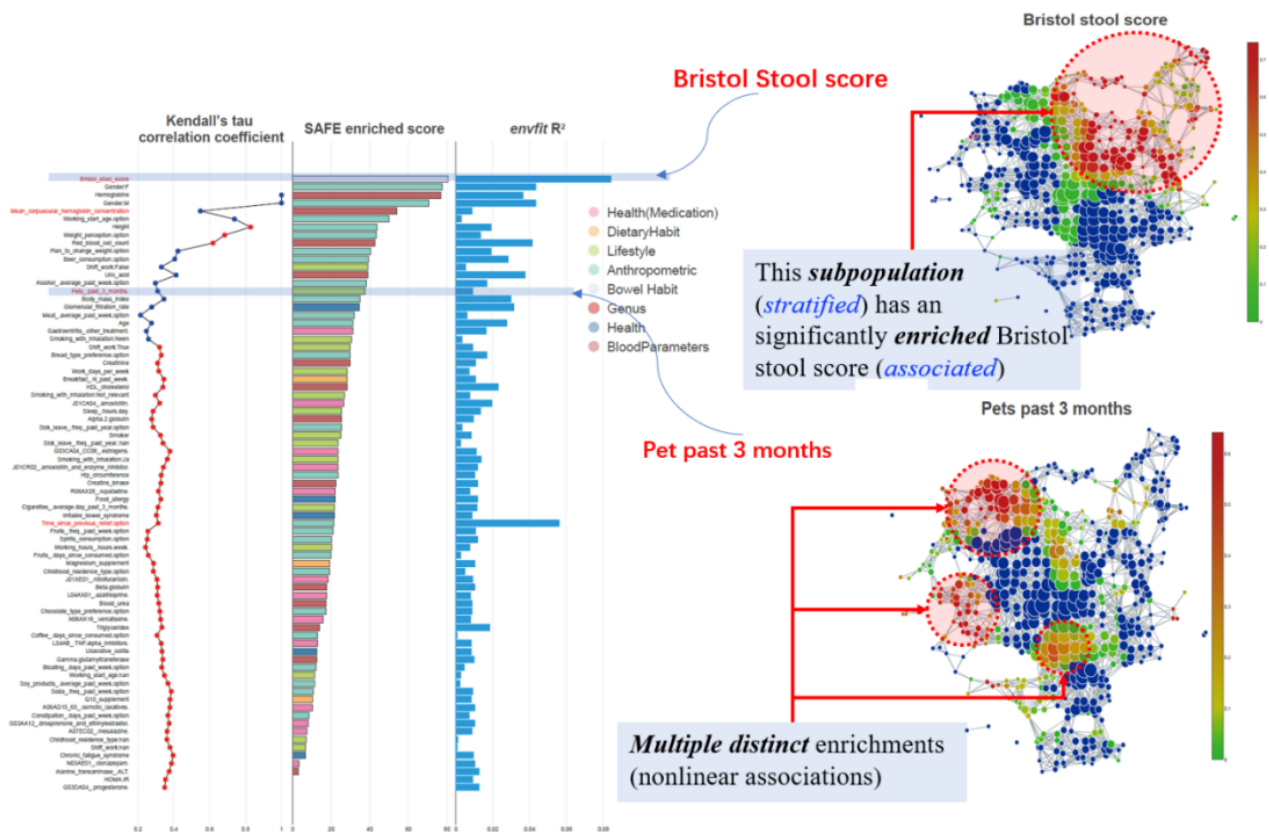
本研究工作得到深圳市科技计划基金支持。

[论文链接](#)



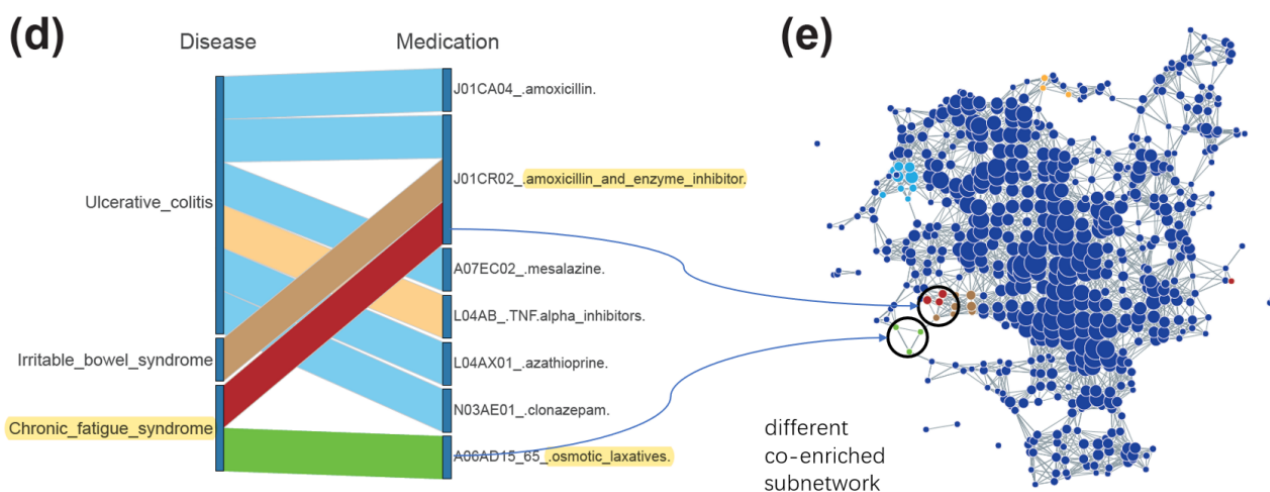
tmap

的大致工作流程：将高维微生物组数据转化为拓扑学网络结构，并用于微生物组人群分层及相关性分析。



以比利时队列为例，左图为关联

微生物组特征与宿主表型，对宿主表型进行排序。并与传统*envfit*方法比较。右图表现了*tmap*对线性及非线性关联均有良好的捕捉。



以比利时队列为例，对疾病和用药的共富集分析。左图为关联微生物组特征与宿主表型，对宿主表型进行排序。并与传统*envfit*方法比较。右图表现了*tmap*对线性及非线性关联均有良好的捕捉。

研究团队单位：深圳先进技术研究院

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](https://www.iikx.com)转发