

**基于大数据引擎的全球微生物组转化网络**

A Scale-Free, Fully Connected Global Transition Network Underlies Known Microbiome Diversity

原文链接：<https://journals.asm.org/doi/10.1128/mSystems.00394-21>

第一作者：Gongchao Jing, Yufeng Zhang

通讯作者：**Xiaoquan Su, Jian Xu**

主要单位：青岛大学计算机科学与技术学院，中国科学院青岛生物能源与过程研究所单细胞中心

**摘要**

在自然界的各种生态系统中，微生物以群落的形式广泛存在并相互作用，从而深刻地塑造着地球生物圈的功能。然而，菌群多样性究竟是怎样形成和演变的？本文提出了一种基于大数据搜索的理论模型，构建了首个全球性的微生物组相互转化网络(Microbiome Transition Network)作为解决方案。该全局性网络基于微生物组结构的相似性高低，节点分别是177,022例菌群样本，通过分析样本我们发现尽管微生物组的结构与其各自迥异的生态系统有着显著的关联性，但每个菌群与任一其他菌群之间平均只需要“七”步，便可实现相互转化。因此，微生物组在全球范围上具有其内在的同源性。该网络是无标度的，表明微生物组转化具有高度的稳定性。进而，通过跟踪该网络中的最小生成树，研究人员得出微生物组相互转化的全球路线图，该路线图刻画了在不同生态系统内部与之间，每个微生物组最可能的演化途径和互作过程。这种基于搜索的全球微生物组网络仅数小时内就能在一个计算节点上重建，为追踪现有或新微生物组的起源和进化提供了一个易于扩展的参考。

**关键词**

微生物组转变(Microbiome Transition),无标度(Scale-free),网络(Network),数据挖掘(Data Mining),Beta多样性(Beta Diversity)

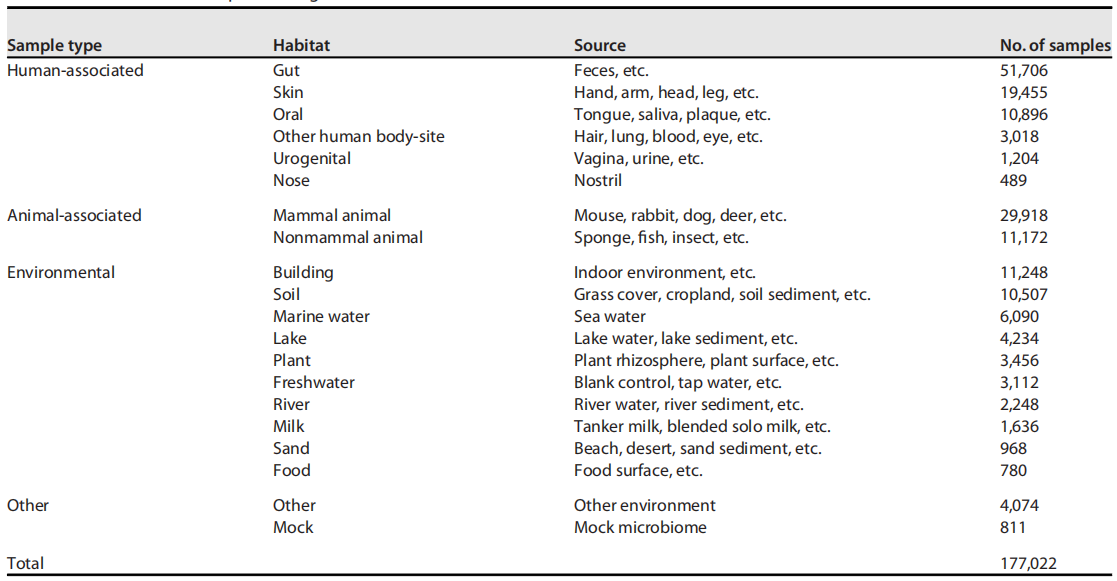
**背景**

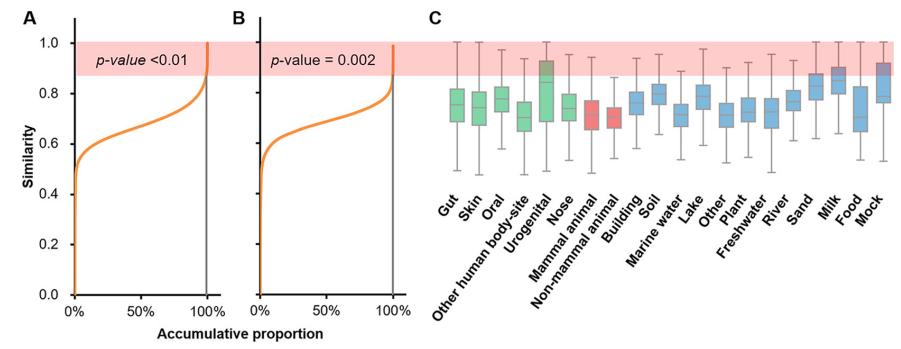
微生物组组成是自然界所有微生物组的基本特征，它受很多种环境因素的影响，例如栖息地、地理位置、温度、氧气水平甚至日长(1,2)。然而，微生物组之间“群落对群落”的水平组成变化是否以及如何与全球微生物组多样性的起源和进化有关(3-5)，目前尚不清楚。例如，来自不同环境的微生物组是分别出现和发展的，还是全球微生物组同源开始，然后通过组成扩散和动态传播到其他栖息地的(6)？近年来，大量的微生物组样本(如人类微生物组计划(7)、地球微生物组计划(1)、塔拉海洋计划(8)等)产生并积累(4)；但是，此类数据的数量相当大且复杂，使得在全球范围内对微生物组进行聚类和建模受到了阻碍(例如，100,000个微生物组的距离矩阵包含约个元素)。

**结果**

**微生物组转化模型和基于搜索的网络。**我们通过微生物组转化模型描述了微生物群落之间的组成动态和变化。在该模型中，微生物群落本质上是微生物物种的组合，并且可以通过传播、融合等方式添加、移除物种或者改变其丰度，来将群落结构转化(Transition)为另一种形式(9,10)(补充材料中的**图S1**)。理论上来说，相似度越高的两个群落，彼此之间转化需要改变的成分越少，转化的可能性就越大；但是，两个菌群究竟达到何种程度的相似才能发生微生物转化，目前尚不清楚。因此，我们基于微生物组搜索引擎（Microbiome Search Engine; MSE(11)）数据库（总共包含177,022个样本，详见“微生物组样本收集”部分，**表1**)，使用Meta-Storms算法(12,13)计算所有微生物组样本之间的相似度，从而推测其中相似度具有显著性（Permutation test *P* < 0.01；**图1**）的样本对之间可能存在“直接转化(Direction transition)”。因此，我们将*P*<0.01所对应的Meta-Storms相似度值0.868定义为微生物组之间直接转化的阈值。通过进一步分析每个栖息地的成对相似度，我们发现0.868的阈值相似度不仅在栖息地之间的相似度分布中显著较高(*P*值=0.0022)(**图1B**)；还高于大多数栖息地内相似度的上限(20个中的17个；**图1C**)。因此，0.868的相似度阈值对于定义生态系统之间的微生物组转化已经足够严格了。

**表1. 不同来源栖息地的微生物组样本分布**

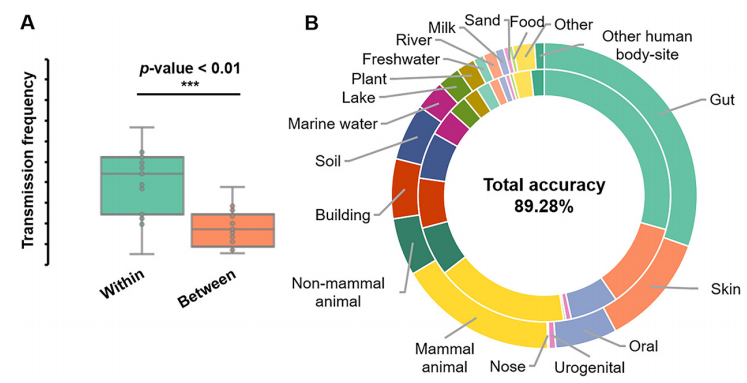




**图1. n = 177,022个微生物组样本相似度的分布。**(**A**) 所有样本中相似度的显著性*P* < 0.01所对应的相似度阈值为0.868(阴影下)。(B) 在不同栖息地之间，0.868的相似度阈值使得*P* = 0.0022。(C) 对于大多数栖息地，阈值高于栖息地内相似度的上限。三个面板使用了同样的y轴。P值通过Permutation检验计算得出。

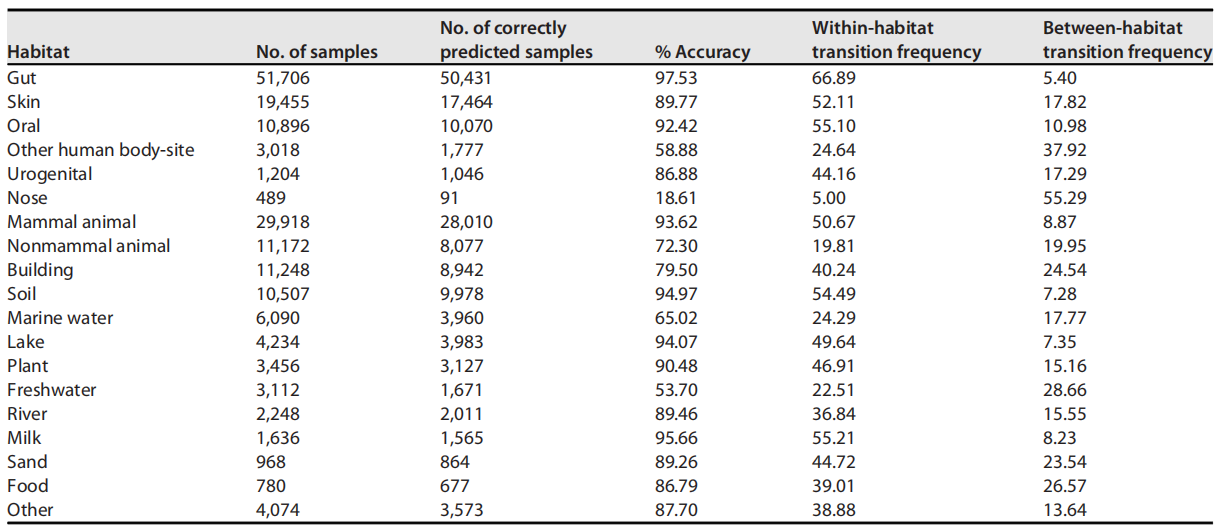
接着，对于所有177,022个微生物组，我们将每一个样本作为输入，使用MSE对其他所有样本进行搜索，找到相似度高于0.868的最佳匹配项。基于搜索结果，我们构建了一个转化网络，该网络由177,022个节点(样本)和11,175,742条边组成，其中每个节点都是一个微生物组，每条边代表一个直接转化(**方程2;图S2**)。我们注意到，在该网络中，一对具有低相似度的样本可以通过多条边连接(即，经由一系列跨中间转移样本的直接转化)，这样的样本对被称为“间接转化”(Indirect transition;**方程3**)。

**转化网络能够在全球范围内预测微生物组的来源。**在全球范围内，微生物组结构的相似是否意味着其来源的生态系统特征的相似(8,14)？为了定量的研究这个问题，我们比较了转化网络中栖息地内(来自同一栖息地的样本对的转化)和栖息地间(两个不通栖息地之间的样本对的转化)的直接转化频率。其中，每个栖息地的直接转化频率由该栖息地每个样本的平均直接转化次数来计算。我们注意到，同一栖息地的样本之间存在更频繁的直接转化(**图2A**；双尾配对t检验；*P* < 0.01)。因此，微生物组的来源环境决定了微生物的组成。接着我们使用转化网络来预测每个样本的栖息地(参见材料和方法)。通过留一交叉法验证(LOOCV)，89.28%样本的来源信息可以被正确地预测(**图2B；表2**)。因此，在全球范围内，微生物组结构与其环境特征密切相关。



**图2. 全球微生物组网络预测微生物组栖息地。**(A) 栖息地内直接转化的频率明显高于栖息地间的频率。P值通过双侧t检验计算。(B) 微生物组网络正确预测了89.28%样本的栖息地。内环代表真实栖息地的比例，外环代表预测栖息地的比例。

**表2. 基于微生物组网络的栖息地预测**

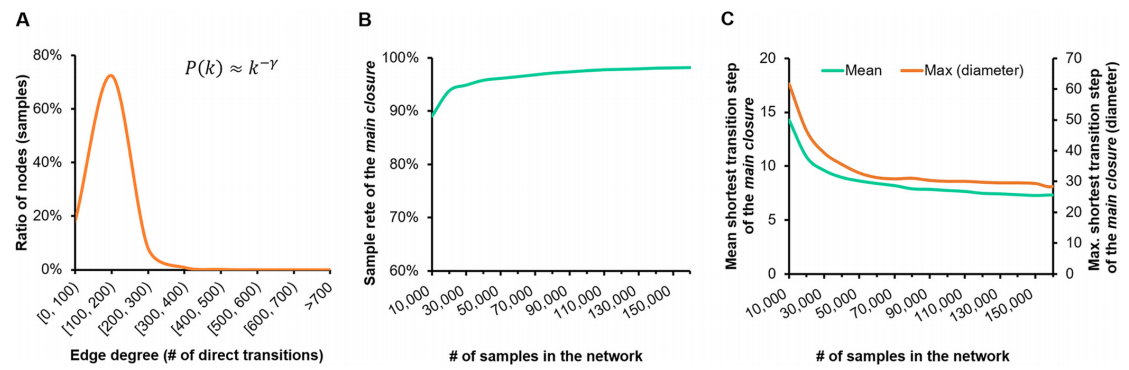


与此同时，“错配”占所有样本的10.72%(176,211个中的18,894个)。这种错配很有趣，他们可能是由栖息地之间微生物组的频繁接触和交换引起的。人体部位之间的错配是最常观察到的错配(占所有样本的1.86%)，由日常接触和微生物成分交换引起(15)。非哺乳动物(如海绵)与海水之间的匹配是第二常见的错配(1.76%)。由共享室内环境(其中微生物组主要来自人类)(5，16)，“人的皮肤”、“动物(宠物)”和“建筑(室内环境)”之间的错配占样本的1.68%。此外，0.81%的错配发生在人-肠道和哺乳动物-动物-肠道之间，这可以通过人类与其他哺乳动物之间系统发育的密切关系，以及哺乳动物与其肠道微生物组的共同进化来解释(17)。我们还观察到一种错配(0.14%)是湖水样本被预测为河水(湖的源流)，反之亦然。因此反映出，尽管全球范围内的微生物组结构主要是由它们的栖息地决定的，但微生物组结构可以通过来自不同环境的微生物组的接触和交换来改变。

**全球微生物组通过转化网络彼此相连。**全球微生物组的β多样性可能通过两种情况进化：第一，“异源”，即来自不同环境的微生物组分别产生和发展(**图S3A**)；第二，“同源”，微生物组以同源性开始，然后分散到其他栖息地(例如，通过成分转化、交换或融合)(**图S3B**)。为了区分这两种情况，我们使用传递闭包算法来检查这个转化网络的连通性(参见材料和方法)。闭包是指一组节点(微生物组)，其中每个微生物组可以通过直接或有限的间接转化来连接到其他任何节点。因此，处于封闭状态意味着可能会通过成分交换可以在样本之间进行转化。通过传递闭包算法遍历网络中的所有节点，我们发现98.31%的样本(177,022个中的174,032个)可以聚类成单个闭包(也为称“主闭包”)。而在上一节的结果中，89.28%的样本结构与其栖息地直接相关，彼此之间差异很大(**图2B**)，在这种情况下，大部分样本仍然能够进入单个闭包，说明微生物组异源的可能性非常低(概率 < 1.5e-05；由方程5估计)，所以我们认为全球微生物组具有同源性，并且是通过相互转化来形成新的微生物组(**图S3B**)。值得注意的是，仍有1.69%(177,022中的2,990个)样本未包含在主闭包中，这主要是由于统计不准确(1.47%的样本显示的相似度水平仅略低于被招募到主闭包的阈值；P值介于0.01和0.05之间)或管理错误(例如，MSE数据库中，0.16%的样本被标记为微生物组，但实际上是纯培养物或18S/ITS扩增子样本)。因此，同源假说最大可能性地解释了微生物组结构的起源和进化。

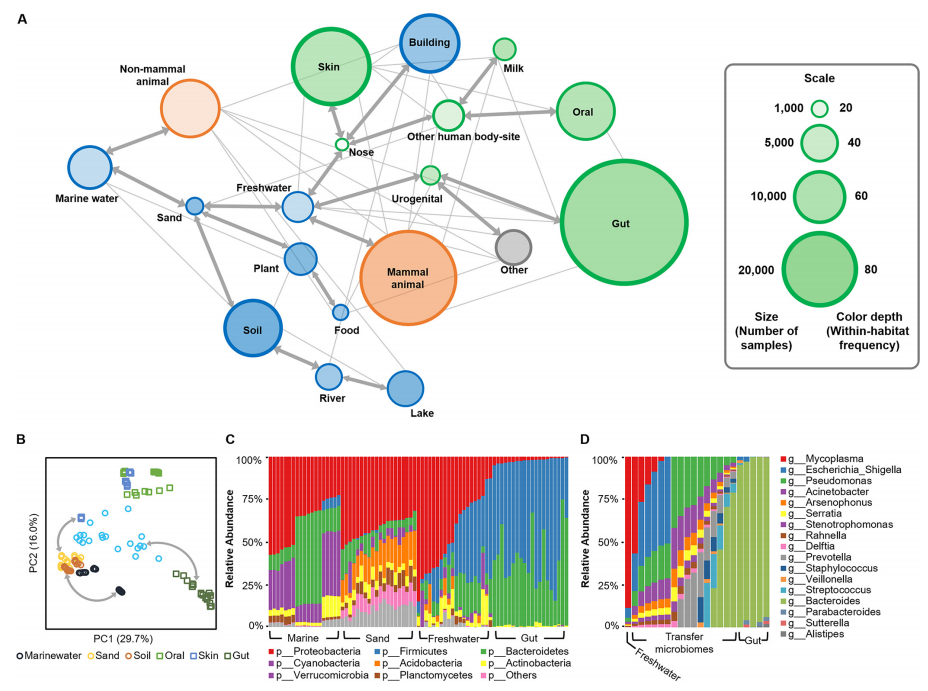
为了确定全球微生物组网络的规模，我们使用Dijkstra算法(18)(参见材料和方法)计算了主闭包中所有样本对的成对最短转化步长。有趣的是，就像社交网络的“小世界”原则(19)一样，微生物组转化网络遵循“7度空间”模式(**图S4A**)。具体来说，主闭包中的任何两个微生物组，即使它们来自不同的栖息地并且表现出低相似度，每个菌群与任一其他菌群之间也可以平均只需要7步的直接转化(20) (**图S4B**)。这种模式强调了高度的连通性，因此，来自全球范围内不同栖息地的微生物组之间有着惊人的密切相互作用。

**全球微生物组网络是无标度的，且稳定性很强。**值得注意的是，在这个全球转化网络中，对于每个节点的度k(即与之直接相连的邻居的数量)遵循泊松分布(**图3A**)，其中和之间的Pearson r = －0.836，即，表明网络是无标度的(21,22)。无标度网络的一个关键特征是拓扑的稳定性，即使移除其中相当数量的节点，仍然不会改变其结构。为了测试其稳定性，我们移除了不同数量的随机选择样本及其相关边，然后评估了剩余网络中主闭包的大小和连通性(剩余节点的百分比范围在5%到95%之间；每个这样的去除程序重复10次)(参见材料和方法)。当网络中的节点数达到80,000(占节点总数的45%)时，97.19%的样本(原为98.31%)(**图3B**)的主闭包连通率曲线已经呈现出平坦的趋势，此外，平均转化步长和最大转化步长(直径)分别收敛到8和33(原为7和32)(**图3C**)。因此，这些参数非常稳定，不依赖于网络中总样本数的增加。这些发现表明了全球范围内生态系统之间微生物组多样性模式的鲁棒性和稳定性。



**图3. 全球微生物组网络的鲁棒性。**(**A**) 网络的节点度(连接邻居的数量)遵循泊松分布，表明网络是无标度的。(**B**) 随机节点移除对主闭包中样本覆盖度的影响。(**C**) 随机节点移除对平均最短转化步长和最大转化步长(直径)的影响。

**微生物组转化路线图模拟了全球微生物多样性在多个生态系统之间的发展。**微生物组的组成因栖息地不同而迥异，同时，微生物组在全球范围内也有完全连通性，所以，全球转化网络能够在宏观尺度上重现微生物多样性是如何在不同栖息地之间传播的。这种“微生物扩散”路线图可以通过一个子网络来模拟，该子网络具有两个特征，一、覆盖并连接所有样本，二、没有循环或冗余路线。因此，我们通过使用Kruskal算法(23)(参见材料和方法)解析主闭包的最小生成树(Minimum Spanning Tree; MST)来得出这样一个路线图(**图4A**)。作为具有最高整体转化概率(相似度)的全局最优值，MST最大限度地捕捉了所有19个栖息地(不包括Mock样本)之间全球微生物多样性的转化模式。例如，海洋微生物组最有可能与其他两种环境交换，一种是地理上靠近海岸的沙子，另一种是非哺乳动物，如鱼。这些观察结果还表明，沙子和淡水微生物组是土壤、植物和人类相关栖息地如肠道、口腔、皮肤和人类生活环境的“门户”。

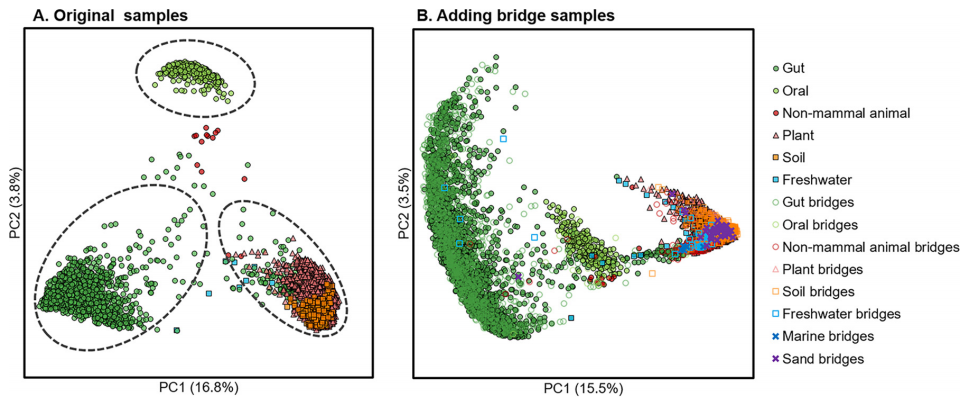


**图4. 全球微生物组转化路线图。**(**A**) 粗线代表最大整体相似度的路线图，其中箭头表示这种转化是双向的。每个栖息地的样本数量由节点大小来表示，栖息地内转化频率由节点颜色深度表示。细线显示栖息地之间的频繁转化。(**B**) 140个微生物组的主坐标分析(PCoA)与全球路线图是同构的。(**C**) 从海洋到肠道环境的微生物组转化途径的门级组成转变。(**D**) 从淡水微生物组到肠道样本的转化案例的属级组成转变。

我们从六个栖息地中随机选择了140个样本，对其进行主坐标分析(PCoA)，得到了与路线图同构的结果，表明了该路线图的合理性(**图4B**)。此外，我们从Phylum层次对海洋-肠道转化路线（该子集中最长的转化路线之一）进行了进一步的详细描述(**图4C**)。当放大这条路线时，可以发现该路线中，淡水微生物组转化为肠道样本的一系列结构变化。从实际的淡水样本(24)开始，在每个步骤中，去除/减少富含淡水的生物体(25)(例如*Mycoplasma*和*Escherichia*)，并添加富含肠道的生物体(15)(例如*Bacteroides*和*Parabacteroides*)。尽管单个步骤可能只对微生物组结构造成轻微的改变，但经过多次迭代后，该样本可以通过一系列转移样本(**图4D**)顺利转移到网络中的肠道微生物组(26)。

**微生物组在时间尺度和地理位置尺度上的转化。**为了测试全球微生物组转化网络是否可以对微生物组的动态变化进行追踪，我们使用一个时间队列来描述人类微生物组随时间的转化。在该数据集中，从两个个体(I,男性；II,女性)的三个身体部位(肠道、口腔和皮肤)在超过396个时间点(27)中收集了1,963个样本。我们基于搜索网络的分析显示，每个身体部位的微生物组组成随着时间的推移呈现出显著的变化(**图5A到C；图S5**)，而皮肤和口腔微生物组则通过直接转化聚集到同一闭合中(**图5D到F**)。这些表明，在不同的时间点，每个部位以及皮肤和口腔部位之间都在进行微生物组的转化。此外，肠道样本都与皮肤及口腔样本有很大差异，这与全球微生物组转化图(**图4A**)一致，并且在该图中，肠道微生物组与皮肤和口腔微生物组处于不同的路径中。因此，尽管口腔和肠道都是消化道，并且可能发生微生物从口腔到肠道的转移(28)，但口腔微生物组可能更多的来自皮肤微生物组(或更显著地由皮肤微生物组形成，反之亦然)，而不是肠道微生物组。这一看似违反直觉的发现实际上可以通过皮肤和口腔的共享来得到更多有氧和更少酸性的环境的支持(pH值和氧气水平对微生物组结构有很大的影响)(1)。

另一方面，为了验证来自不同地理位置的微生物组的连通性，我们通过单个数据集(29)构建了一个基于搜索的网络，该数据集包含从六种生态环境(人类肠道、人类口腔、非哺乳动物动物、植物、土壤和淡水)收集的3,850个样本。产生这个数据集的原始研究表示，来自人类肠道和植物根部的微生物群落之间鲜有共同的菌群组成(29)。我们的分析与该结论一致，基于该数据集的网络分析发现，样本被分布到三个独立的闭包中(**图6A**)。然而，一旦从MSE数据库中添加了额外的1,635个连接不同闭包的样本，就会出现一个新的闭包，它整合了原始的三个独立闭包，并能够囊括高达97.74%的样本。这是因为新添加的样本充当了“转移节点”(材料和方法中的**方程3**；**图6B**)。值得注意的是，在这些“转移”微生物组中，大多数(96.89%)来自与原始数据集相同的栖息地，其他主要来自沙子和海洋，而且我们发现它们也是我们的全球微生物组转化路线图中非哺乳动物动物、植物和土壤微生物组之间的转移节点(**图4**)。这个例子表明，尽管来自不同环境和孤立地理位置的微生物组具有非常不同的结构，但只要充分调查和覆盖全球β多样性，它们仍然可以在微生物组转化网络中的单个闭包内连接，即相互进化。这些结果直接挑战了多宿主微生物群组成独立进化的结论(29)，尤其是当讨论相似度(即，β多样性)、相互作用或微生物组之间的其他类型的关系时，强调了在我们全球微生物组网络的背景下推导或验证“本地”数据集的重要性。



**图6. 微生物组在栖息地和地理位置之间的转化。**(**A**) 来自6个栖息地的3,850个样本被纳入在三个孤立的转化封闭包中，其中样本比例分别为43.22%、7.90%和45.53%。(**B**) 从MSE数据库中添加额外1,635个桥梁样本后，三个闭包通过直接转化合并为一个闭包，这个闭包包含97.74%的样本。

**讨论**

微生物群一直在与我们的星球共同进化并塑造我们的星球，但由于化石的缺乏以及整合和挖掘如此大规模、高度复杂的数据(14)所涉及的的方法上的挑战，它们在全球范围内的起源和进化仍然难以捉摸。尽管“物种对物种”的相互作用已经通过对不同栖息地微生物组的共现分析绘制出来(30,31,38)，但在全球范围从“群落对群落”层面对庞大的微生物组数据空间的研究才刚刚开始(32-34)。在这里，我们提出了一种基于网络的分析框架，用于描述和模拟全球微生物组β多样性在多个栖息地的变化和扩散。得益于微生物组搜索引擎(11)极高的搜索速度，我们引入了一个全球微生物组网络，包含177,022个微生物组样本和113亿个序列。通过遍历这样的网络，我们展示了微生物组结构通过遵循“小世界”原则的显著相似度在全球范围内连接。这一努力揭示了全球微生物组多样性的内在同源性，并支持地球上所有微生物组的单系起源。此外，我们绘制了第一个全球微生物组转化路线图，来说明可以解释全球微生物组进化过程的潜在和最可能的路径。

由于微生物组测序数据呈指数级增长，目前主要依赖于-复杂度成对关系(n是样本数)的β多样性分析方法，例如主坐标分析(PCoA)和聚类分析，尤其是在计算资源有限的情况下变得越来越不切实际。在这里，我们通过一个基于搜索的网络来应对这一挑战，该网络建立在“邻居”的基础上，即每个样本具有最高相似度的邻居；这种策略将计算复杂度降低到(c是常数，即邻居的数量)，从而能够在3小时内在单个计算节点上破译超过100,000个微生物组的成对相似度。因此，全球微生物组转化路线图将定期更新，因为本地微生物组数据集可以很容易地根据他们的共享节点与这个全球路线图保持一致，所以可以为解释或验证现有或未来对微生物组间相似性、关联性或相互作用的观察提供参考。此外，这种基于网络的分析框架可以扩展到鸟枪法宏基因组数据集，为追踪或预测微生物组进化提供了新的视角，即使在全球范围内也具有良好的分辨率。

**方法和材料**

**微生物组样本的收集。**我们使用了微生物组搜索引擎数据库([http://mes.ac.cn)中的所有微生物组样本。从包括20个栖息地(补充材料中的**表S1**)的572个研究/](http://mes.ac.cn)中的所有微生物组样本。从包括20个栖息地的572个研究/)项目中收集样本。通过使用Parallel-META 3(36)(3.4.4版)软件生成OTU，并采用Greengenes(35)全长16S rRNA数据库(13-8版)在97%的相似度上进行注释。16S rRNA基因拷贝数的变化基于IMG/M数据库(37)进行了修正。我们为每个样本设置了500的最小序列号和80%的最小16S rRNA比对率，以确保高质量的参考数据集。最后，n=177,022个样本和11,302,841,991条序列通过了质量控制和筛选(**表S1**)。

**计算成对微生物组相似度矩阵来定义直接转化。**在Parallel-META 3的软件包中使用Meta-Storms算法(12,13)，完全排列所有n = 177,022个样本的成对相似度矩阵(总共 = 15,668,305,731次)来检查微生物组系统发育相似度的分布。通过在相似度排列中设置一个截止P值小于0.01(排名前1%)，我们得到了Meta-Storms相似度0.868并作为显著高值的统计阈值来定义直接转化(这个阈值也是称为)。因此，转换模型可以用以下形式描述

(1)

其中和是两个任意微生物组，sim(,)表示它们的Meta-Storms相似度。

**基于搜索的微生物组网络。**基于搜索的微生物组网络是使用微生物组搜索引擎(MSE)(11)构建的。对于每个样本，我们在所有其他样本中搜索前100个匹配项，并将其与相似度高于直接转化阈值( = 0.868)的匹配样本连接起来.通过对所有样本迭代这样的搜索，我们构建了一个全局网络G。

(2)

在这个网络中，一个节点(例如，**方程2**中的a或b)是一个单一的微生物组，而连接节点的边(例如，**方程2**中的边[a，b])是直接转化(**图S2**)。

最后，在网络中有177,022个节点(样本)和总共11,175,742条边(直接转化)。在这个网络中，一对相似度低的样本可以通过多条边的路径连接，即间接转化：

(3)

这里被定义为从a到b的间接转化的基础的“转移样本”。

**使用微生物组网络预测栖息地。**在网络G中，我们通过每个微生物组的前n = 10个相邻样本和相似度来预测其来源栖息地。对于网络中的任意微生物组样本a，与其前10个邻居的相似度是，而n个邻居来自个不同的栖息地，如，则微生物组a的预测栖息地的概率计算公式为

(4)

这里，表示邻居j的栖息地是(1≤k≤m)。然后将具有最高概率P的预测栖息地作为样本a的预测结果。

**网络中所有样本之间的转化概率。**通过方程4，我们计算出在全球范围内，微生物群组成因栖息地而异，并且同一栖息地之间的转化概率为89.28%(**图2B**)。为了计算转化网络中连接所有栖息地的整体概率，我们可以从连接概率为的任意两个栖息地开始。当网络中再增加一个栖息地时，三个栖息地之间的转化概率可以计算为，其中89.28%的平方代表增加的栖息地与前两个栖息地之间没有直接转化的概率。那么我们可以扩展这样一个过程来估计连接转化网络中n个栖息地的概率为

(5)

**微生物组网络的传递闭包算法。**微生物组转换网络中，闭包是完全连接的节点(微生物组)的子集，因此每个微生物组都可以通过直接或间接转化(具有有限的转移节点)与任何其他样本相连。闭包可以由网络中的任意节点初始化，然后通过添加更多与该闭包直接连接的外部节点进行扩展(**图S6**)。如果两个或多个闭包通过任何边连接，这些闭包也可以合并为一个闭包。通过在网络G中所有节点之间的遍历，我们得到了一个包含98.31%样本的主闭包C。

**微生物组网络的大小。**在主闭包C中，两个间接连接的节点(微生物组)之间总是存在多条路线。我们将两个直接相连节点之间的边数计为1，因此间接路线的长度就是这条路线上的转移节点11的数量(**图S4**)。我们使用Python包igraph(在Python3.6.1中运行的0.7.1)的Dijkstra算法(19)来找到主闭包C中所有间接连接节点对之间的成对最短转化步骤(转移节点数最少)。因此，最短路径中的最大步数就是闭包的直径。直径意味着在这个闭包中，任何两个微生物组都可以通过一条比直径小的几步的路径相互连接。

**微生物组网络路线图的最小生成树。**在转化闭包中，生成树是连接所有节点(微生物组)而没有循环的子网络。对于两个直接连接的样本a和b，我们将它们的距离定义为

(6)

最小生成树(MST)可以被认为是具有最高整体转化概率的样本的全球转化路径，因为它连接了所有具有最短总距离的样本。在主闭包C中，我们使用Kruskal算法(23)来计算二级MST，以反映全球范围内不同栖息地之间的转化。

第一层MST是“样本级分辨率”的，我们在此基础上制作了关于“栖息地级分辨率”的第二层MST。最初我们计算了主闭包C的第一层MST，然后生成了基于栖息地的网络(**方程2**)，其中每个节点代表一个栖息地，两个栖息地之间的距离和是MST中连接两个栖息地的所有边的平均距离。然后我们计算了二层MST()，它展示了跨多个栖息地的全球微生物组转化路线图。

(7)

路线图(MST[])的重要性通过原始网络主闭包C中拓扑等价子网络的排列检验进行评估。具体来说，在排列中，对于路线图中连接两个栖息地(例如，栖息地和)的每条边，我们还分别从这两个栖息地中(例如和)随机选择了一个连接两个样本(例如样本a和样本b)的边。由于我们迭代了10,000次排列，如果路线图的总距离小于排列网络的99%（也意味着总概率在前1%，P值＜0.01），我们可以认为路线图MST()在主闭包C中很重要。

**从参考数据库到连接分离闭包的基于搜索的样本选择。**从参考数据库中选择转移样本来连接两个分离的闭包，我们对照参考存储库搜索每个闭包的所有样本，寻找高于直接转化阈值（Td.t=0.868）的最佳匹配，并且两个闭包之间的重叠匹配是连接两个闭包的转移微生物组。如果匹配中没有重叠，那么我们通过添加它们的匹配来扩展每个闭包并重复搜索过程，直到找到任意的转移样本。另一方面，一旦无法通过数据库搜索进一步扩展闭包，而且仍然找不到可用的转移样本，这意味着参考数据库中没有样本能够作为转移节点，通过直接转化将两个分离的闭包联系起来。

**数据和材料的可用性。**这里的关键生物信息学工具，微生物组搜索引擎(MSE)，可以通过http://mse.ac.cn作为在线服务免费访问。此外，对于自定义微生物组数据库的独立搜索，在GitHub(https://github.com/qibebt-bioinfo/meta -storms)上提供了MSE的内核代码和教程。这项工作中使用的所有数据和分析脚本都可以在GitHub(https://github.com/qibebtbioinfo/microbiomenetwork)上获得，以确保可重复性。

**参考文献**

1.→Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vazquez-Baeza Y, Gonzalez A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Leflfler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi .org/10.1038/nature24621.

2.→Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108:4516–4522. https://doi.org/10.1073/pnas.1000080107.

3.→Mihaljevic JR. 2012. Linking metacommunity theory and symbiont evolutionary ecology. Trends Ecol Evol 27:323–329. https://doi.org/10.1016/j .tree.2012.01.011.

4.→Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. 2016. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. PLoS Biol 14:e2000225. https:// doi.org/10.1371/journal.pbio.2000225.

5.→Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun 8:14319. https://doi.org/10.1038/ ncomms14319.

6.→Garud NR, Pollard KS. 2020. Population genetics in the human microbiome. Trends Genet 36:53-67. https://doi.org/10.1016/j.tig.2019.10.010.

7.→Integrative HMP (iHMP) Research Network Consortium. 2019. The Integrative Human Microbiome Project. Nature 569:641–648. https://doi.org/10 .1038/s41586-019-1238-8.

8.→Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, RoyoLlonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Tara Oceans coordinators. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359. https://doi.org/10.1126/science.1261359.

9.→Charubin K, Modla S, Caplan JL, Papoutsakis ET. 2020. Interspecies microbial fusion and large-scale exchange of cytoplasmic proteins and RNA in a syntrophic Clostridium coculture. mBio 11. https://doi.org/10.1128/mBio .02030-20.

10. →Sarkar A, Harty S, Johnson KVA, Moeller AH, Archie EA, Schell LD, Carmody RN, Clutton-Brock TH, Dunbar RIM, Burnet PWJ. 2020. Microbial transmission in animal social networks and the social microbiome. Nat Ecol Evol 4:1020–1035. <https://doi.org/10.1038/s41559-020-1220-8>.

11.→Jing G, Liu L, Wang Z, Zhang Y, Qian L, Gao C, Zhang M, Li M, Zhang Z, Liu X, Xu J, Su X. 2021. Microbiome Search Engine 2: a platform for taxonomic and functional search of global microbiomes on the whole-microbiome level. mSystems 6:e00943-20. https://doi.org/10.1128/mSystems.00943 -20.

12.→Su X, Wang X, Jing G, Ning K. 2014. GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU. Bioinformatics 30:1031–1033. https://doi.org/10 .1093/bioinformatics/btt736.

13.→Su X, Xu J, Ning K. 2012. Meta-Storms: effificient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics 28:2493–2501. doi:10.1093/bioinformatics/bts470. https://doi.org/10.1093/bioinformatics/bts470.

14.→Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. 2016. Microbiome data science: understanding our microbial planet. Trends Microbiol 24:425–427. https://doi.org/10.1016/j.tim.2016.02.011.

15.→Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla A, Creasy HH, Earl AM, Fitzgerald M, Fulton RS, Human Microbiome Project Consortium, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10 .1038/nature11234.

16.→Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK, VázquezBaeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 345:1048–1052. https://doi.org/10.1126/science.1254529.

17.→Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, Henrissat B, Knight R, Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332:970–974. https://doi.org/10.1126/science.1198719.

18.→Dijkstra EW. 1959. A note on two problems in connexion with graphs. Numer Math 1:269–271. https://doi.org/10.1007/BF01386390.

19.→Milgram S. 1967. The small world problem. Psychology Today 2:60–67.

20.→Leskovec J, Horvitz E. 2008. Planetary-scale views on a large instant-messaging network; abstr, Proceedings of the 17th international conference on World Wide Web, Beijing, China, Association for Computing Machinery. https://doi.org/10.1145/1367497.1367620.

21.→Barabasi AL. 2009. Scale-free networks: a decade and beyond. Science 325:412–413. https://doi.org/10.1126/science.1173299.

22.→Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. The large-scale organization of metabolic networks. Nature 407:651–654. https://doi.org/ 10.1038/35036627.

23.→Kruskal JB. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Amer Math Soc 7:48–50. https://doi .org/10.1090/S0002-9939-1956-0078686-7.

24.→Dunn RR, Fierer N, Henley JB, Leff JW, Menninger HL. 2013. Home life: factors structuring the bacterial diversity found within and between homes. PLoS One 8:e64133. <https://doi.org/10.1371/journal.pone.0064133>.

25.→Lee CS, Kim M, Lee C, Yu Z, Lee J. 2016. The microbiota of recreational freshwaters and the implications for environmental and public health. Front Microbiol 7:1826. https://doi.org/10.3389/fmicb.2016.01826.

26.→Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. 2016. Preservation methods differ in fecal microbiome stability, affecting suitability for fifield studies. mSystems 1:e00021-16. https://doi.org/10 .1128/mSystems.00021-16.

27.→Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. 2011. Moving pictures of the human microbiome. Genome Biol 12:R50. https://doi.org/10 .1186/gb-2011-12-5-r50.

28.→Schmidt TS, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, Wirbel J, Maistrenko OM, Alves RJ, Bergsten E, de Beaufort C, Sobhani I, HeintzBuschart A, Sunagawa S, Zeller G, Wilmes P, Bork P. 2019. Extensive transmission of microbes along the gastrointestinal tract. Elife 8:e42693.

https://doi.org/10.7554/eLife.42693.

29.→Hacquard S, Garrido-Oter R, Gonzalez A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, Schulze-Lefert P. 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host Microbe 17:603–616. https://doi.org/10.1016/j.chom.2015.04.009.

30.→Ma B, Wang Y, Ye S, Liu S, Stirling E, Gilbert JA, Faust K, Knight R, Jansson JK, Cardona C, Rottjers L, Xu J. 2020. Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome 8:82. https://doi.org/10.1186/s40168-020-00857-2.

31.→Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. 2012. Microbial co-occurrence relationships in thehuman microbiome. PLoS Comput Biol 8:e1002606. https://doi.org/10 .1371/journal.pcbi.1002606.

32.→Su X, Jing G, McDonald D, Wang H, Wang Z, Gonzalez A, Sun Z, Huang S, Navas J, Knight R, Xu J. 2018. Identifying and predicting novelty in microbiome studies. mBio 9:e02099-18. https://doi.org/10.1128/mBio.02099-18.

33.→Su X, Jing G, Sun Z, Liu L, Xu Z, McDonald D, Wang Z, Wang H, Gonzalez A, Zhang Y, Huang S, Huttley G, Knight R, Xu J. 2020. Multiple-disease detection and classifification across cohorts via microbiome search. mSystems 5:e00150-20. https://doi.org/10.1128/mSystems.00150-20.

34.→Zhu M, Kang K, Ning K. 2020. Meta-Prism: ultra-fast and highly accurate microbial community structure search utilizing dual indexing and parallel computation. Brief Bioinform 22:557–567. https://doi.org/10.1093/bib/ bbaa009.

35.→McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. https://doi.org/10.1038/ismej.2011.139.

36.→Jing G, Sun Z, Wang H, Gong Y, Huang S, Ning K, Xu J, Su X. 2017. Parallel-META 3: comprehensive taxonomical and functional analysis platform for effificient comparison of microbial communities. Sci Rep 7:40371. https:// doi.org/10.1038/srep40371.

37.→Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen I-MA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. 2008. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36:D534–D538. https://

doi.org/10.1093/nar/gkm869.

38. Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 12:315–330. https://doi.org/10.1007/s13238-020-00724-8.