Temporal development of the gut microbiome in early childhood from the TEDDY study

TEDDY研究计划中的幼儿肠道微生物组随时间的发育

研究总结

从婴儿期到儿童期，微生物组的发展取决于一系列因素，**微生物-免疫串扰（crosstalk）**被认为与之后的疾病病理生物学有关，包括：持续性胰岛自身免疫病（persistent islet autoimmunity）和1型糖尿病。然而，据我们所知，还没有研究对大规模，多中心（multi-center）人群生命早期的微生物组进行过广泛的描述。在这里，我们通过**16S rRNA基因测序**( n = 12005 )和**宏基因组测序**( n = 10867 )分析了**903名3个月至46个月**大的儿童纵向的**粪便样本**，这是青少年糖尿病的环境决定因素研究（The Environmental Determinants of Diabetes in the Young，TEDDY）的一部分。我们的结果表明正在发育的肠道微生物组经历了微生物组发展的**三个不同阶段：发育阶段( 3 - 14个月)、过渡阶段( 15 - 30个月)和稳定阶段( 31 - 46个月)**。**母乳喂养，无论是专一地还是部分母乳的喂养，都是与微生物组结构相关的最重要因素。**母乳喂养与较高水平的*Bifidobacterium* species (*B. breve* 和 *B. bifidum*)有关，停止母乳喂养导致肠道微生物组更快地成熟，这以Firmicutes门为标志。在发育阶段，出生模式也与微生物组显著相关，阴道分娩的婴儿肠道微生物由较高水平的*Bacteroides* species (特别是 *B. fragilis*)驱动。无论出生方式是什么，*Bacteroides*都与更高水平的肠道多样性和更快速地成熟有关。包括地理位置和家庭暴露在内的环境因素(如兄弟姐妹和毛茸茸的宠物)也是重要的变量。一个巢式病例对照（nested case-control）分析揭示了**微生物分类学与胰岛自身免疫病或1型糖尿病发展之间的微妙联系**。这些数据决定了生命早期中微生物组的结构和功能，并为进行微生物-免疫串扰（crosstalk）对健康长期的影响机制研究奠定了基础。

在这项研究中，对来自**三个欧洲国家**(德国、瑞典和芬兰)和**三个美国州**(科罗拉多州、佐治亚州和华盛顿州)的**903名儿童的12500份粪便样本**进行了分析。这些儿童代表血清转化（seroconverted）为胰岛细胞自身抗体阳性（islet cell autoantibody positivity）或发展1型糖尿病( developed T1D )的人和对应的控制组。作为青少年1型糖尿病的环境决定因素研究（TEDDY）的一部分，这些儿童大约从3个月大起，平均每月收集粪便样本。在稀释（rarefaction）样本并将样本限制在**3 - 46个月**之后，我们分析了**微生物组**( 16S rRNA基因测序，903名儿童的12005个样本；宏基因组测序，783名儿童的10867个样本)和来自纵向的粪便样本**功能宏基因组**(仅宏基因组测序数据)。

对于这群有胰岛自身免疫(islet autoimmunity， IA )或T1D风险的儿童，我们打算( 1 )明确描述3 - 46个月大儿童纵向的肠道微生物组的发展特征；( 2 )确定在生命早期，对于发育中的细菌群落有特定影响的母体遗传（maternal）和产后的影响因素；( 3 )使用巢式病例（nested case–control）对照分析来研究微生物组预测胰岛自身免疫病或T1D发展的潜力。

补充注释1（Supplementary Note 1）和扩展数据图1（Extended Data Fig. 1）提供了细菌分类和功能性通路的概述。Dirichlet multinomial mixtures（DMM ）模型被应用在16S rRNA基因测序数据(图1 )和宏基因组测序数据。所有3到46个月大的样本都包括在内，16S rRNA基因测序图谱（profiles）形成了10个簇（clusters）(基于lowest Laplace approximation) 。在每个簇中，细菌的丰度和多样性都增加了。使用前五个门的线性混合效应模型（linear mixed-effects modelling）和香农多样性指数（Shannon’s diversity index），我们确定了微生物组发展的**三个不同阶段：发育阶段( 3 - 14个月)、过渡阶段( 15 - 30个月)和稳定阶段(≥31个月)**。在发育阶段，所有五个门和香农多样性指数（Shannon’s diversity index）都发生了显著变化；在过渡阶段，两个门(*Proteobacteria*和 *Bacteroidetes*)和香农多样性指数（Shannon’s diversity index）发生了显著变化；在稳定阶段，所有门和香农多样性指数（Shannon’s diversity index）都没有显著变化(图1c )。在最初的发育阶段，*Bifidobacterium*占主导地位，其中20 %的个体由Cluster 1转变为由Cluster 3占主导地位 (这两个簇都是*Bifidobacterium*占主导地位)。随着婴儿年龄的增长，在过渡时期，他们粪便中的微生物组在15 - 30个月，多样性增加，变成4-8簇（clusters）。第31个月开始，观察到微生物组的稳定，其中婴儿的样本在连续的时间点保持在同一个簇中。在稳定阶段，Cluster 8 到Cluster 10是最主要的，这些簇的特点是高度的α多样性和厚壁菌门内的属的优势。三个微生物群阶段和分类群的变化与其他组群一致，并得到宏基因组测序数据的支持。

接下来，我们试图从16S rRNA基因测序数据(属级)、宏基因组测序分类群(物种级)和功能代谢能力((Kyoto encyclopedia of genes和genomes (KEGG) modules)中确定与微生物组相关的重要因素。为了进行统计分析，将样本分成（stratify）离散时间点(第3到6个月、第7到10个月、第11到14个月、第15到18个月、第19到22个月、第23到26个月、第27到30个月和第31到40个月)来分析协变量（covariate），并且只包括每个婴儿的第一个样本。扩展数据表1（Extended Data Table 1）显示了每个协变量（covariate）的基本分组信息。几个协变量与3至18月之间的属水平和种水平细菌群落分布显著相关，特别是在第3到6月的第一时间点。相反，细菌代谢潜能仅与从出生3到14个月的母乳喂养相关。

**母乳喂养解释了从出生第3个月至14个月最大的变量（variance），**之后只有10 %的婴儿接受母乳喂养。母乳喂养对微生物组的发育有着类似的影响，不管它是完全母乳喂养，还是与配方奶粉和固体食物一起混合喂养的。在属一级，母乳的摄入与*Bifidobacterium*的相关性在每一个时期都是最大的。在物种水平上，母乳喂养与121种不同的细菌物种显著相关，母乳喂养的婴儿中*B. bifidum, B. breve, B. dentium, Lactobacillus rhamnosus*和*Staphylococcus epidermidis*的水平较高，*Escherichia coli, Tyzzerella nexilis, Eggerthella lenta, Ruminococcus torques*和*Roseburia intestinalis*的水平较低。*Bifidobacterium*属的物种和*Lactobacillus*属的物种存在于母乳中，*Staphylococcus*属的物种定居在乳晕皮肤上，因此这些物种可以从母亲直接转移到婴儿身上。与其他*Bifidobacterium*属物种相比，*B. longum*与母乳喂养没有显著的相关性但是保持较高的相对丰度。在Vatanen等人的文章原稿中，发现大部分*B. longum*菌株含有来自人乳寡糖( human milk oligosaccharide，HMO )基因簇的基因，而在停止母乳喂养后，大部分*B. longum*菌株不再携带这些基因。这可能反映了*B. longum* 的亚种 *infantis* 和*longum* 的各自使用哺乳动物来源和植物来源的寡糖的能力。停止母乳喂养后，*B. bifidum*也持续存在，这种细菌能够将HMO转化为粘蛋白降解物（mucin degradation）。Vatanen等人的实验表明，构成DMM Clusters 1 – 3的*B. breve, B. longum和B. bifidum*具有不同的糖利用特征，这表明婴儿之间可以利用的不同的营养物质可以促进特定*Bifidobacterium*物种的定殖。

随着婴儿年龄的增长，固体食物在饮食中的比例也增加。在当前的研究中，一些母乳喂养的婴儿和不再母乳喂养的婴儿之间的香农多样性指数（Shannon diversity index）随着时间的推移开始收敛（converge），这可能是由于饮食中母乳的比例降低，因而*Bifidobacterium*的优势减弱。由于*Bifidobacterium*在母乳喂养的婴儿中占主导地位，因此在所有阶段，部分母乳喂养的婴儿与不在接受母乳喂养的婴儿相比多样性显著更低。为了进一步探索微生物的成熟，我们使用了微生态年龄（microbiota age）和按年龄划分的Z分数( microbiota-by-age Z-scores，MAZ )，包含了20个OTU 的模型解释了72 %的差异（variance）(模型包括所有OTU时，这个比例为74 % )。相比之下，在发育和过渡阶段接受母乳喂养的婴儿中，微生物年龄和MAZ分数显著更低，但在稳定阶段趋于一致。停止母乳喂养之后，110个细菌物种( 89个来自Firmicutes门)从第3到14个月显著增加。最近，人们注意到在接受部分母乳喂养时，Firmicutes细菌会受到抑制。这些数据一起支持了现有的报道，**即肠道微生物群的成熟是由停止母乳喂养驱动的，而不是固体食物的引入驱动，其特征就是Firmicutes的增加。**

**母乳喂养是唯一与代谢潜能显著相关的协变量（covariate）。**绘制前三个时间点( 3 - 14个月)所有重要模块的图，显示了基于母乳喂养的清晰聚类，无论时间点如何，代谢能力都具有可比性。与母乳喂养婴儿最显著相关的模块来自“碳水化合物和脂肪代谢（carbohydrate and lipid metabolism）”途径，包括“脂肪酸生物合成（fatty acid biosynthesis）”和“β-氧化，酰基辅酶a合成（beta-oxidation, acyl-CoA synthesis）”。这与以前的研究一致，以前的研究发现母乳喂养的婴儿在婴儿期脂肪酸生物合成相关的基因增加了。相反，没有接受母乳的婴儿显示代谢能力的快速转换（turnover），“二羧酸酯-羟基丁酸酯循环（dicarboxylate-hydroxybutyrate cycle）”和“还原乙酰辅酶a（reductive acetyl-CoA）”途径增加。与维生素B7 (“核苷酸和氨基酸代谢（nucleotide and amino acid metabolism）”途径)相关的模块，在非母乳喂养婴儿在14个月以前的所有时间段都增加了，这是一种与成年人微生物组相关的功能。

通过将与已知KEGG同源（orthologues）有重叠，并且含有基因组坐标（genomic coordinates）的reads映射到KEGG模块上，我们就能直接确定每一个同源基因（gene orthology）以及模块是从哪一个分类单元（taxa）衍生出来的。根据区分母乳喂养状况的主要物种绘制了每个重要模块的每条路径。在母乳喂养的婴儿中，*B. breve*占据了生命早期最大数量的重要模块，并在6个月后被*B. bifidum*所取代。在非母乳喂养的婴儿中，*E. coli*主占第3至14个月的重要模块。这为证明在停止母乳喂养后，肠道微生物组在分类和功能两个层面上迅速成熟提供了进一步的证据。

TEDDY研究的动力是基于特定的1 : 1巢式病例对照（nested case-control）设计，从两个嵌套病例对照研究( IA（autoimmunity ）或T1D )中，使用风险集抽样（risk of sampling），检测与IA和T1D发展相关的微生物群。分析队列由一个每个病例对照对的样本数相等的子集组成。截至2012年5月31日，IA队列包括632名儿童和6194份粪便样本，T1D队列包括196名儿童和1540份粪便样本。IA组和T1D时间序列上的的α多样性(richness和Shannon多样性)、微生态年龄（microbiota age）和MAZ分数在病例（cases）和匹配对照（matched control）之间是可比较的。在IA病例（cases）中，来自16S rRNA基因测序的前50个最丰富的属的相对丰度仅显示出细微的组成差异，未分类的Erysipelotrichaceae的相对丰度更高。在T1D和对照队列中，5个细菌属与T1D发病相关，其中*Parabacteroides*最为显著。T1D病例中有11个细菌属较低，包括4个未分类的Ruminococcaceae, *Lactococcus, Streptococcus* 和*Akkermansia*。

条件逻辑回归模型（Conditional logistic regression models）显示IA的唯一状态数量或每个受试者（subject）不同状态之间的转换数之间没有显著关联。T1D也缺乏关联性，例外的是T1D发病前6 - 12个月，病例显示较少的唯一状态。值得注意的是，T1D发病前6 - 12个月的组中，任何一个时间点的样本数都是最低的，因此，统计上显著的结果应该谨慎解释。总的来说，群落动力学的条件逻辑回归模型（Conditional logistic regression models）表明，**微生物组稳定性与IA或T1D的发病没有密切关系**。

补充注释3（Supplementary Note 3）进一步分析了在几个时间点都具有显著性差异，在16S rRNA基因测序和宏基因组分析中都具有显著性差异的协变量。简而言之，**出生模式（birth mode）与生命第一年的微生物组发展有着显著的关联，阴道分娩的婴儿*Bacteroides* 属的物种水平较高。**这在不同的母乳接触人群（exposure groups）和在不同地理位置都大体一致。尽管核心微生物组是一致的，多样性、微生物群年龄和MAZ分数也有可以比较的轨迹，但是在第3至22个月地理位置之间的差异也存在。家庭暴露（household exposure）(例如，与兄弟姐妹和毛茸茸的宠物生活在一起)也与早期生命中微生物组特征的差异有关，在生命早期，与兄弟姐妹和/或毛茸茸的宠物生活在一起的婴儿微生物组的成熟速度更快。

TEDDY研究通过定期取样，对广泛的元数据（metadata）以及扩增子测序和宏基因组测序进行分析，对903名3至46个月大的婴儿的肠道微生物发展进行了有力的分析。我们表明，**生命的第一年是微生物组发展的关键阶段，母乳的摄入是这一时期影响微生物群发展的主要因素。出生模式、地理位置、家庭兄弟姐妹和毛茸茸的宠物也与这一时期的微生物组有关。**我们认为生命的第一年是发育时期，生命的第二年是过渡时期，从生命的第三年起，微生物组进入稳定时期。当研究人员在第3个月之前或第46个月之后采集样本时，这些精确的年龄可能会发生变化。

目前的队列（cohort）大部分是白人，非西班牙裔，T1D遗传风险高的婴儿群体，其中一些人患有自身免疫或糖尿病。病例和对照之间的时间序列上的α-多样性和群落动态是可比较的（comparable），这与其他队列中报道的发现相反，这可能是由于TEDDY研究队列（cohort）中受试者和样本数量的增加。我们发现病例( IA和/或T1D )和匹配对照之间的细菌相对丰度有细微的变化。**T1D病例显示*Streptococcus* 物种水平和*Lactococcus*物种较高**，这与Vatanen等人在论文中的发现一致。根据以前的研究，在目前的研究中**，匹配对照中*Akkermansia*的丰度也更高，这可能表明肠道完整性增加**。

总体上，微生物组的发展和显著的协变量与西方化人群以前的报告一致，尽管需要谨慎地将具有T1D风险的TEDDY研究中的儿童群体的研究结果推断到更广泛的人群。然而，当前研究中报告的显著协变量已经独立地与后期疾病的风险相关联，如肥胖、哮喘和过敏。目前的研究提供了几个婴儿时期微生物群发育的可测试假设，确定生命早期微生物组变化的潜在机制以及随后对免疫发育和功能的影响仍然很重要。随着对生命早期关键阶段及其对健康和疾病的影响更全面地理解，生活方式和治疗方法可以被定制来支持最佳的微生物-免疫稳态。

配图：

图.1 16S rRNA 基因测序数据的DMM聚类（clutering）

基于lowest Laplace approximation，所有的数据形成了10不同的聚类（cluters）

A 热图展示了每个DMM聚类的25个最占优势的属的相对丰度。方括号中的分类群名称需要正式的分类修订。

B 箱线图展示每个DMM聚类的α-多样性（丰度（richness）和香农多样性指数（Shannon’s diversity））。中线表示中位数，分箱包含了25%和75%的数据。分箱上下的竖线表示最极端的数据点。在线条之外的点则表示了异常值

C 转变模型（transition model）表示了每个DMM聚类在3到46个月的每个时间点的发育。虚线分箱表示微生物组的三个发育时期（发育，过渡和稳定）。旁边的实心正方形表示基于每个时期的门和香农多样性指数（Shannon’s diversity）( H’)的显著变化。在发育时期，所有的门和H’ 的变化都是显著的；在过渡时期，两个门和H’的变化是显著的。在稳定时期，没有变化是显著的。节点和边界（edge）的大小由总体数量决定。节点的颜色由DMM聚类的编号决定，边界（edge）的大小由转变频率（transition frequency）决定。频率低于4%的转变没有展示。

图.2 所有数据类型中由EnvFit建模的22个微生物组协变量的显著性和解释方差

水平栏显示由EnvFit确定的模型中每个协变量解释的方差( r2 )。每个协变量中的组详见扩展数据表1（Extended Data Table 1）。协变量基于总体元数据组（metadata）进行着色。显著协变量（错误发现率（false rate rate, FDR, P < 0.05））以粗体表示。星号表示每个时间点的显著协变量。BMI表示体重指数；wtgain表示体重增加。

a、基于16S rRNA基因测序数据的属级微生物组图谱（profiles）( n = 4，069 )。

b、基于宏基因组测序的物种级微生物组图谱（profiles）( n = 3，843 )。

c、基于宏基因组测序的模块级功能宏基因组能力图谱（profiles）( n = 3，843 )。

图.3 在所有数据集中，出生后第一年母乳喂养状态是最显著的微生物组协变量

母乳喂养状态与与前三个时间点( 3 - 14个月，n = 2，257；补充表1（Supplementary Table 1） )的微生物组图谱（profiles）显著相关。曲线显示每个类别数据的局部加权散点图的平滑曲线(LOESS)，阴影区域显示基于拟合排列（permutation-based）的95 %置信区间。

A、非度量多维标度( Non-metric multidimensional scaling, NMDS )排序图显示了每个母乳喂养状态组的平均几何中心（mean centroid）。曲线图仅包括在给定时间点内从患者取得的第一个样本；第3至6、7至10和11至14个月。平均几何中心（mean centeoid）大小是基于样本量的，条形代表95 %的置信区间。

b、显示第3至40个月母乳摄入量的曲线图，与同期六种最丰富的*Bifidobacterium*的相对丰度相比( n = 11，717 )。

c、第3至40个的纵向香农多样性指数（Shannon diversity index）( n = 11717 )。

d、基于取样时微生物组年龄和婴儿年龄的MAZ分数的微生物组成熟的纵向发展( n = 11717 )。

e、热图，显示前三个时间点中每一个时间点由MaAsLin分析确定的所有重要模块的平均丰度。还给出了每个模块的相应路径。BM表示母乳。

f、堆叠条形图，显示了在路径级别入库的每个重要模块的丰度。每种细菌的丰度图，五种最重要的细菌分别与母乳喂养和非母乳喂养的婴儿有关。

拓展阅读：

1 Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030126>

2 the human gut microbiome in early-onset type 1 diabetes from the teddy study

<https://www.nature.com/articles/s41586-018-0620-2>

3 Persistent gut microbiota immaturity in malnourished Bangladeshi children.

<https://www.nature.com/articles/nature13421>

4 分析过程使用的代码

https://github.com/StewartLab/ Stewart\_TEDDY\_Microbiome\_Analysis

研究方法

研究人群（Study population）

TEDDY研究由六个临床研究中心组成：三个在美国(科罗拉多州、佐治亚州/佛罗里达州和华盛顿)，三个在欧洲(芬兰、德国和瑞典)。根据自身抗体的阳性情况（anto-antibody positivity），被招募的儿童被前瞻性地随访3个月大至15岁，四岁前每3个月进行一次研究访问，此后每3个月或6个月进行一次随访。截至2012年5月31日，收集了粪便样本和相关元数据（metadata）。从出生3到48个月，每个月收集粪便样本，在10岁以前每三个月收集一次，之后每两年收集一次，放入临床中心提供的三个塑料粪便容器中。4岁后抗体阴性的儿童一年提交4次报告，尽管4岁后他们的随访改为一年两次。如果他们住在美国，父母在环境温度或+ 4°C温度下将粪便容器放入适当的装运箱中，保证在24小时内送到NIDDK仓库。如果他们住在欧洲，会送到他们的附属临床中心。欧洲临床中心储存粪便样本，每月将大量冷冻粪便运送到NIDDK仓库。研究人群的选择(病例和匹配对照)是基于他们患T1D的高风险，基于他们的HLA基因型，10 %基于的家族史。详细的研究设计和方法之前已经发表过了。病例和对照儿童的匹配因素是地理位置、性别和T1D家族史。

元数据是使用经过验证的问卷收集的，这些问卷已经发表或由专家广泛审查过。在为期三个月的临床访问中，通过问卷调查收集了关于母亲、怀孕和出生的信息，包括出生模式(阴道分娩和剖腹产)、婴儿的5分钟Apgar评分、怀孕期间的并发症、母亲糖尿病( T1D、2型糖尿病( T2D )或孕期糖尿病（gestational diabetes）)、妊娠期和孕期母亲用药(胰岛素（insulin）、二甲双胍（metformin）、格列本脲（glyburide）、抗高血压药（antihypertensives）)的信息。TEDDY研究为父母提供了许多工具，如“TEDDY book”，帮助他们实时收集孩子生活中的所有事件，以确保偏差和错误最小化。每次随访时，研究人员都会和主要看护人一起阅读TEDDY book，并使用标准化的研究表格提取相关信息。从3个月大开始，训练有素的工作人员每三个月定期访问一次，收集数据，并通过电子传输的表格(网络表格（web forms）或远程表格(teleforms))直接输入。web应用程序中使用前端约束（Front-end constraints）来防止无效数据的输入。TEDDY的报错和确认系统(The TEDDY Error Reporting and Verification System ,ERVS )由一组对数据进行自动质量控制、报告和解决错误的程序、一个用于存储错误数据的集成数据库以及一组生成用监控数据清理工作的报告的程序组成。该系统的细节已经发表。鉴于TEDDY研究设计的前瞻性，信息偏差和召回偏差被大大减少。由于儿童在加入研究时没有事件结果，因此没有理由在研究参与者群体之间收集的信息的准确性方面存在任何系统性差异。

TEDDY研究得到了科罗拉多州科罗拉多多机构审查委员会、佐治亚州医学院佐治亚州人类保障委员会( 2004 - 2010 )、佐治亚州健康科学大学人类保障委员会( 2011 - 2012 )、佐治亚州摄政大学机构审查委员会( 2013 - 2015 )、奥古斯塔大学机构审查委员会( 2015年至今)、佛罗里达州佛罗里达大学佛罗里达健康中心机构审查委员会、华盛顿州机构审查委员会( 2004 - 2012年)的美国地方机构审查委员会和芬兰西南医院区伦理委员会、德国巴伐利亚医学协会伦理委员会、瑞典隆德地区伦理委员会第2节( 2004 - 2012年)和隆德大学持续伦理审查委员会( 2013年至今)的欧洲伦理委员会的批准。 所有父母或监护人在参与基因筛查和登记之前都提供了书面知情同意。这项研究是按照所有相关的道德规范进行的。

对于匹配的IA病例-对照研究，使用discrete Cox’s proportional hazards regression计算了a priori power calculations，估计了80 %的power，α= 0.01，双端检验检测出5 %流行率的暴露的比值，而20 %流行率的暴露的比值比。实验不是随机的，研究人员在实验和结果评估期间也不会对分配blind。

A priori power calculations using discrete Cox’s proportional hazards regression for the matched IA case–control study estimated 80% power, α = 0.01, two- sided test to detect an odds ratio > 3 for an exposure with 5% prevalence to an odds ratio > 1.8 for an exposure with 20% prevalence. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

16S rRNA 基因测序（16S rRNA gene sequencing）

16S rRNA基因测序方法是根据NIH -人类微生物组计划（NIH-Human Microbiome Project）和地球微生物组计划（Earth Microbiome Project）开发的方法改编的。按照制造商的说明，使用PowerMag Microbiome DNA isolation kit提取细菌DNA。16S rRNA基因的V4区通过PCR扩增，并在MiSeq平台( Illumina )上使用2 × 250 bp paired-end read protocol进行测序。Reads的pairs被解复用（demultiplexed），使用USEARCH v7.0.1090讲reads进行合并。合并允许0个错误和最小50个碱基的重叠，合并的reads在第一个碱基处被trimmed，q ≤ 5。质量过滤被应用于最终的合并reads，包含超过0.5 %预期错误的reads被丢弃。使用UPARSE算法，序列以97 %的相似性阈值（cutoff）逐步聚类成OTU。使用USEARCH v7.0.1090和UCHIME v4.2移除嵌合体（Chimeras）。为了确定分类，使用USEARCH v7.0.1090将OTU映射（mapped）到仅包含16S V4区域的SILVA数据库。通过将合并的reads映射（mapped）到UPARSE OTUs来计算丰度。根据前两个步骤中生成的输出文件，使用一个定制脚本构建了一个稀释后（rarefied）的OTU表，用于分类相对丰度、α多样性和β多样性(包括UniFrac )的下游分析。从16S rRNA基因测序中总共获得了114313601个reads(每个样品的中位数为8442个reads)，每个样品被稀释（rarefied）到3000个读数。严格的合并参数导致OTU数量相对较少，宏基因组学的物种数量比16S rRNA基因测序的OTU数量高四倍左右。

宏基因组鸟枪测序（Metagenomic shotgun sequencing）

从每个样本构建的单个文库被合并（pooled）并加载（loaded）到HiSeq 2000平台( Illumina )上，并使用e2 × 100 bp paired-end read protocol进行测序。质量过滤、trimming和demultiplexing的过程是由内部流程完成的，该流程是通过公开可获取的工具开发的，例如Casava v1.8.2 ( Illumina )用于生成fastq，Trim Galore v0.2.8 (http://www.bioinformatics. babraham.ac.uk/projects/trim\_galore/)和cutadapt v1.9dev2用于adaptors和quality trimming，PRINSEQ v0.20.5用于样本的去重复和低复杂度的过滤。此外，Bowtie2 v2.2.3用于将reads映射（map）到包含NCBI全基因组测序( whole-genome sequencing，WGS )的档案中细菌、病毒、人类和载体的完整基因组和集合的数据库(截至2015年3月)。从后续分析中删除最高identity 匹配（match）不是细菌的reads。编辑距离( Levenshtein distance)用于确定与参考基因组的比对分数。对于细菌reads，仅考虑前25个最高得分的比对结果，每个read选择最高得分匹配。在多个相同的最高得分命中（hits）的情况下，确定最低的共同祖先（the lowest common ancestor）。

将与已知KEGG同源（orthologues）的reads列表化，并逐步（step wise）计算KEGG模块，如果每个检测物种和宏基因组存在65 %的反应步骤，则确定该模块是完整的。通过MinPath计算由存在的基因同源词产生的的最小集合（minimum set），为每个分类群和宏基因组构建了通路。从宏基因组测序中总共获得了19967936136个reads(每个样品的中位数为1606240个reads)，随后的分析中，每个样品被稀释（rarefied）到100000个读数。

统计分析（Statistical analysis）

分析分两部分进行: ( 1 )表征微生物组的时间纵向上的成熟；( 2 )确定影响微生物组发育的显著协变量。对于分析的两个部分，使用16S rRNA基因测序数据在OTU水平上为计算α多样性(丰富度和香农多样性)，使用宏基因组数据在物种水平上计算。在R中用ggplot软件包使用LOESS回归方法为α多样性和分类丰度（taxonomic abundance）建模，并且以95 %的置信区间绘图。

DMM聚类（DMM clustering）

分析的第一部分确定了微生物发展的关键阶段，包括DMM的使用。DMM根据微生物组结构bins样品。基于lowest Laplace approximation score来确定合适的聚类数。对于这一具体分析，包括了46个月的样本，而所有其他分析包括了40个月的样本。这里包括额外的样品，可以更准确地测定微生物关键阶段。

分析的第二部分试图确定在离散时间点形成微生物组图谱的显著协变量，并根据40个月前的样本进一步确定显著改变的分类群。统计分析框架考虑了数据集的纵向性质，并考虑了协变量的动态性质。由于一些协变量可能会在开始日期之前影响微生物组(例如抗生素处方的基本指示（underlying indication）)，并且一些协变量会在未知的时间范围内改变微生物组(例如，抗生素破坏的微生物群在治疗后几个月内可能会继续改变)，协变量被分类为‘before’, ‘during’和‘after’。如果婴儿的协变量为负，所有样本都将被归类为“never”。在一个协变量可能有几个开始起作用的情况下(例如，多重（multiple）抗生素开始和结束时间点)。在第一次开始后，对于剩余的样本，协变量被归类为“after”，除非发生了另一个事件，在这种情况下，根据开始和停止日期，在适当的时候将应用“during”。分析是在特定时间窗口进行的，包括在3至6月、7至10月、11至14月、15至18月、19至22月、23 至26月、27至 30月和31至40月收集的样本。每个时间窗口中只包括从给定儿童收集的第一个样本，以考虑重复测量。

EnvFit分析以确定显著的协变量（MaAsLin analysis to determine significant taxa associated with each covariate）

每个协变量的效应大小和显著性是使用“vegan”( https : / / cran.r-project.org/web/packages/vegan/index.html )中的“envfit”函数来确定的，该函数比较了各组centroids相对于总变量的差异。（Ordination）是使用基于Bray–Curtis dissimilarity的NMDS进行的。显著值（significance value）是基于10000个排列（permutations）来确定的。使用FDR调整（FDR adjustment），对来自envfit的所有p值进行多次比较调整。总共有22个已知与新生儿、婴儿和儿童肠道微生物发育相关的协变量被纳入EnvFit分析，每个变量中使用的分组显示在扩展数据表1（Extended Data Table 1）中。具体来说，我们测试了母体因素，包括糖尿病(孕期糖尿病（gestational diabetes、T1D、T2D或无（none）) 、糖尿病药物(胰岛素（insulin）、二甲双胍（metformin）、格列本脲（glyburide）、抗高血压药（antihypertensives）) 、BMI、妊娠体重增加类别(过量或非过量) 、先兆子痫（preeclampsia）、母亲益生菌消耗量，以及后代的因素，如早产、55、出生模式、性别、母乳喂养和/或配方奶粉喂养。

MaAsLin分析，以确定与每个协变量相关的重要分类群（MaAsLin analysis to determine significant taxa associated with each covariate）

MaAsLin被用于协变量的调整，当确定分类群( 16S rRNA基因测序的属水平和宏基因组测序的物种水平)对特定变量的影响时，同时考虑了潜在的混淆协变量。简而言之，这种微生物数据的多元线性建模系统从一组协变量中进行选择，用来与微生物分类单元或通路的丰度相关联。然后，使用variance-stabilizing arcsin square root transform的mixed-effects linear models，从这个简化的集合中确定推定关联的显著性。所有关联的标称（nominal）p值随后使用Benjamini–Hochberg FDR方法进行调整。这里，报告了校正q < 0.25的微生物特征。Enfit分析中测试的所有22个协变量都包括在调整中，不管enfit的显著性如何。受试者年龄也包括在内，以调整每三个月时间窗口中分类群的潜在年龄驱动变化，IA和T1D结果也包括在内，以调整队列的巢式病例控制性质。应用默认MaAsLin参数(元数据（metadata）中样本NA的最大百分比为10 %，最小百分比为0.01 %，P < 0.05，q< 0.25 )。使用FDR对所有p值进行多次比较调整。

微生物态maturation modelling和linear mixed-effects分析（Microbiota maturation modelling and linear mixed-effects analysis）

随机森林回归模型（random forest regression model）如前所述使用“ran- domForest”R包执行。简而言之，该模型是在随机选择的150名足月、阴道分娩、母乳喂养的婴儿身上训练的，这些婴儿在最终数据集中至少包含10个样本。该模型是使用默认参数建立的:生长10000棵树，在每次分割时随机抽样n/3的OTU，其中n代表OTU的数量。通过应用具有十倍交叉验证（tenfold cross-validation）的“rfcv”进一步完善了该模型，根据增加均方误差的百分比来训练最终模型，结果包括了20个OTUs。这20个OTU解释了模型总方差的72 % (相比之下，包括所有OTU的总方差为75 % )。该模型预测的受试者年龄被称为微生物年龄，并被进一步用于使用先前描述的公式确定MAZ分数25。使用R“lme4”包中 “lmer”命令，基于linear mixed-effects，计算α多样性、微生物年龄和MAZ分数的显著差异。我们包括对每个（individual）儿童的随机斜率（random slopes）和截距（intercept），并将出生模式、年龄、*Bacteroides*阳性或阴性、主要饮食、地理位置、兄弟姐妹的存在以及家庭宠物的存在作为固定影响进行评估。为了执行这些分段纵向模型，我们将样本分为三个发育阶段(小于14个月、15至30个月和大于31个月)。由于完全母乳喂养和从未母乳喂养组中的样本数量相对较少，母乳状况分析是基于“部分母乳喂养（some breast milk）”或“母乳之后（after breast milk）”进行的，发现这些组别分别与完全母乳喂养和从未母乳喂养聚类在一起。

IA和T1D巢式病例-对照稳定性分析数据集的确定（Determination of the datasets for IA and T1D nested case–control stability analyses）

持续的确认的IA的发展情况每三个月评估一次。持续的自身免疫是指在两次或更多次连续对方中确认胰岛自身抗体的存在。持续自身免疫的日期被定义为两个连续样本中的第一个样本的抽取日期，这两个样本被认为儿童证实了持续的特定自身抗体(或任何自身抗体)的阳性。T1D是根据美国糖尿病协会诊断标准定义的。创建了具有相同病例数和对照样本数的数据集，以对汇总度量变量进行conditional logistic regression。平均来说，病例往往比对照有更多的样本，因此有更多的转变和观察到的状态，这导致了我们的指标和疾病结果之间的虚假（spurious）关联。为此，我们创建了一个数据集，其中病例和对照样本根据生命中最近的样本匹配成对病例(不匹配的样本或20 %以外的样本从分析中被省略)。这导致了316例IA病例和316对对照的分析队列(每组中n = 3097个粪便样本)，98例T1D病例和98对对照(每组中n = 1270个粪便样本)。为了一致性，我们将这些数据集用于所有匹配的病例对照分析。IA和T1D分析仅基于16S rRNA基因测序数据，对宏基因组测序数据的分析(即物种级别的分类概况和功能能力)在另一篇文章中。

在病例-匹配对照数据集中与IA发病相关的分类学和代谢特征分析（Taxonomic and metabolic profiling relative to IA onset in the matched case–control dataset）

16S rRNA基因测序数据用于确定α多样性( OTUs数量(丰富度)和香农多样性指数)、微生物群年龄和MAZ分数之间的差异。使用R中的linear mixed-effects models模型计算α多样性、微生物群年龄和MAZ分数的显著差异，使用位于“lme4”包中的“lmer”命令。为了执行这些分段纵向模型，我们将样本分为三个发展阶段(小于14个月，15至30个月，大于31个月)。疾病诊断前，对来自样本的前50个最主要细菌属进行病例-匹配对照对的conditional logistic regression。几率比以95 %的置信区间计算，并根据潜在的混杂变量进行调整，包括样本收集时的年龄、HLA基因型、出生模式和母乳喂养时间。各属的丰度信息转换为log2的输入模型。为了避免0，添加了0.01的值。Benjamini–Hochberg procedure用于校正多次比较，校正后的p＜0.05被认为是有意义的。

基于IA或T1D病例和对照之间DMM簇的微生物组不稳定性评估（Assessment of microbiome instability based on DMM clusters between IA or T1D cases and controls）

对于每一个受试者，计算每个婴儿显示出的完全取样的clusters总数以及不同集群之间从一个样本到下一个样本的转换次数，以提供微生物稳定性随时间变化的汇总。这些汇总指标随后被用于conditional logistic regression，以评估微生物稳定性与IA和T1D的关系。Odds ratios以95 %的置信区间计算，并根据潜在的混杂变量进行调整，包括HLA基因型、出生模式、母乳喂养时间、抗生素疗程数和感染事件数。