成年单卵双胞胎的病毒组多样性与肠道微生物组多样性相关

Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins

图形摘要：

病毒组仍然是微生物组中相对未被探索的组分。Moreno-Gallego和Chou等人，在没有考虑宿主遗传变量的情况下，研究了同卵双胞胎的病毒组，探究微生物多样性如何与病毒组多样性相关。按微生物组一致性的高或低对同卵双胞胎进行分类，揭示了相关的病毒组相关性。

研究总结：

病毒组是人类肠道微生物子中最可变的组分之一。在双胞胎中，病毒组是相似的，但成年人却不一样，这表明随着双胞胎年龄的增长，他们的环境和微生物组发生分歧（diverge），病毒组也一样发生分歧（diverge）。微生物组在多大程度上驱动了巨大的病毒组多样性尚不清楚。在这个研究中，我们研究了21对成年同卵双胞胎中微生物组和病毒组多样性之间的关系，这些双胞胎按微生物组一致性的高或低来分类。源自类病毒颗粒（virus-like particles）的病毒组对于每个个体都是独一无二的，*Caudovirales* 和 *Microviridae*占优势，并显示出包含crAssphage的小核心（small core）。与微生物组具有一致性的双胞胎相比，在不具有一致性的双胞胎中显示出，微生物组越丰富，病毒组就越丰富。这些模式是由噬菌体而不是真核病毒驱动的。总的来说，**这些观察支持微生物组在病毒组模式中的强大作用。**

研究介绍：

人类肠道微生物组由大量的细菌细胞组成，连同少数古菌和真核细胞，共同形成一个密度非常高的微生态系统(每克粪便1011 - 1012个细胞) 。**微生物组的细胞和病毒组的成分(每克粪便中109 - 1012个类病毒颗粒（virus-like particles）的比例大致相等。**病毒组主要由噬菌体和溶源噬菌体组成，也包括更罕见的真核病毒和内源性逆转录病毒。目前，大多数噬菌体在数据库中没有匹配，它们的宿主仍有待阐明。将噬菌体与宿主相匹配是一项挑战：例如，最常见的人类肠道噬菌体——crAssphage——的宿主最近才被鉴定为*Bacteroides*的物种。除了对宿主的鉴定，还存在其他问题，比如塑造病毒组最重要的因素，以及微生物组的细胞部分对病毒组的预测程度。

普遍认为噬菌体及其宿主在时间序列上的种群动态可能是相关联的。事实上，研究者描述过水生系统中的病毒及其细菌宿主的群体振荡，这些研究表明病毒在调节细菌种群（population）方面起关键性作用。**但是这种捕食者-被捕食者模式（predator-prey）对于人类肠道病毒组和微生物组中并不典型。**(为了清楚起见，从这里开始，我们使用“微生物组（microbiome）”来指代微生物组的细胞部分（cellular fraction of the microbiome），例如，主要是细菌细胞。)尽管如此，病毒组和微生物组确实显示出相似的跨寄主的多样性模式，例如人与人之间的（interpersonal）高度的差异和相对长期的稳定性。与无关（unrelated）个体相比，有联系（related）的个体的微生物组更相似，这可能是由于共同的饮食习惯推动了微生物组之间的相似性。相应地，饮食与病毒组多样性相关，很可能这是通过饮食对微生物组的影响。**在婴儿中，同卵双胞胎之间的病毒组比无关（unrelated）个体之间的病毒组更相似。在成年双胞胎中没有观察到这种模式，可能是因为它们的微生物组已经发生了分歧（divergence）。**由于宿主相关性等混杂因素，微生物组本身在多大程度上驱动宿主间病毒组的多样性模式难以评估。

在这个研究中，我们关注成年同卵双胞胎的肠道微生物组，以进一步探索微生物组与病毒组多样性之间的关系。通过研究同卵双胞胎的病毒组，我们控制了宿主的遗传相关性。尽管与双卵双胞胎或不相关的个体相比，同卵双胞胎通常具有更相似的微生物组，但是同卵双胞胎仍然可以显示出大量双胞胎内的微生物组多样性。我们以前从TwinsUK队列（cohort）中收集了双胞胎的粪便微生物，基于这一信息，我们选择了具有高度一致性或高度不一致性的微生物组的双胞胎。我们从类病毒颗粒（virus-like particles）中产生病毒，这些类病毒颗粒（virus-like particles）是从与微生物组来源相同的样品中获得的。结果表明，**微生物组多样性和病毒组多样性的指标呈正相关关系。**

研究结果

选择微生物组一致或者不一致的同卵双胞胎

Selection of Microbiome-Concordant and -Discordant Monozygotic Twin Pairs

我们选择了具有相似体重指数( BMI )的双胞胎，基于之前获得的16S rRNA基因测序数据，它们的微生物组样本间多样性( β-多样性)要么一致，要么不一致。这项研究中的成年同卵双胞胎不共享一个家庭（household），我们假设双胞胎之间的其他环境变量相似。我们基于三个β-多样性距离确定同卵双胞胎微生物组之间的一致或不一致程度，这三个距离是Bray-Curtis、加权（weighted）UniFrac和未加权（unweighted）UniFrac。与预期一致，**β-多样性是相关的**。基于成对距离(pairwise distance)的分布，我们从所有三种分布的边界中选择了21对同卵双胞胎，同时保持年龄和体重指数在整个组中的平衡。在所选的21对双胞胎中，微生物组具有一致性的同卵双胞胎之间的微生物组相比不具有一致性的更相似。不具有微生物组一致性的同卵双胞胎的微生物组在所有的分类学水平上都有不同的组成，特别是在门的水平上，Firmicutes和Bacteroidetes是两个主要的门，对同卵双胞胎之间的差异贡献最大。

类病毒颗粒（VLPs）的鸟枪法宏基因组测序

Shotgun Metagenomes of VLPs

我们从用于16S rRNA基因多样性分析的相同粪便样品中分离出类病毒颗粒（virus-like particles）。从类病毒颗粒（virus-like particles）中提取的DNA用于全基因组扩增，然后进行鸟枪法宏基因组测序。选择第一个库（“large-insert-size library”），平均插入大小为500 bp (总共34325116对reads；质量控制后每个样品817265 ± 249550对reads)，用于病毒contig的重新组装（de novo assembly）。平均插入大小为300 bp的较小片段在第二个文库(“small-insert-size library”)中纯化并测序。最终的pair-end reads合并为25324163条较长的过滤过质量的reads，以提高映射（mapping）精确度。

识别推定的细菌污染

Identification of Putative Bacterial Contaminants

从类病毒颗粒（virus-like particles）制备和测序的病毒可能被细菌DNA污染。然而，鉴于噬菌体是水平基因转移的主要媒介，在溶源噬菌体状态下，温和病毒通常占细菌基因组的10 %，去除潜在的细菌污染也可能去除病毒reads。为了评估细菌DNA的污染，我们把病毒reads映射（mapped）到一组8163个完全组装的细菌基因组上。我们的策略，包括评估每个基因组整个长度上的覆盖率(in bins of 100kb)，那些中值覆盖率大于100的基因组被认为是污染。映射到短区域的reads被认为是溶源噬菌体或水平转移的基因而被保留下来。映射（map）到基因组的reads被确定为潜在污染物，去除它们之后再进行进一步的分析。

我们识别了**65个细菌基因组是潜在污染**，每个样品的reads中1 % ± 1.125 %映射（mapping）到细菌基因组。**大多数( 37 / 68 )属于Firmicutes门**；在物种水平上，在至少50 %的样品中检测到*Bacteroides dorei, B. vulgatus, Ruminococcus bromii, Faecalibacterium prausnitzii, B. xylanisolvens, Odoribacter splanchnicus*和*B. caecimuris*。如果微生物组中最丰富的细菌种类是最可能的污染源，那么它们作为污染物的相对丰度应该与它们在微生物组中的相对丰度相对应。然而，我们观察到污染的DNA和微生物组中代表分类群的相对丰度之间没有显著的相关性。

功能图谱支持病毒在类病毒颗粒纯化中的富集

Functional Profiles Support Viral Enrichment in VLP Purification

为了评估病毒组的功能性内容，我们使用综合基因目录(Integrated Gene Catalog，IGC )的KEGG对“short-insert-size”文库的原始reads进行了注释。根据以前的研究，大部分宏基因组类病毒颗粒（virus-like particles）的reads( 85.43 % ± 5.74 % )被映射(mapped)到具有未知功能的基因。

为了进一步验证序列来源于类病毒颗粒（virus-like particles）而不是微生物组，我们生成并且比较了四个补充个体的类病毒颗粒（virus-like particles）和大量粪便样品的宏基因组数据。与预期一致，**来自相同样品的病毒组和微生物组的功能分布是不同的。**映射（map）到注释基因的病毒组reads在两个类别中富集：遗传信息过程（Genetic Information Process）( 48.87 % ± 12.12 % )和核苷酸代谢（Nucleotide Metabolism）( 17.59 % ± 8.81 % )，相比之下微生物宏基因组中分别为24.31 % ± 1.28 %和5.47 % ± 0.4 %。基本上，**细菌宏基因组中存在的大多数其他功能类别在病毒组中都不存在。**此外，相对于微生物组，病毒组的功能注释显示样本间的更高的变异，更低的组内相关系数（intraclass correlation coefficient）。

病毒组是个体独有的

Viromes Are Unique to Individuals

我们从“large-insert-size library”中组装了reads，总共产生了大于500 nt 的107307条contigs (最大值：79863 nt；平均值，118；最小值：1741元)。为了评估病毒组的结构和组成，建立了一个比对到去重复的contigs的reads补充矩阵（a matrix of the recruitment of reads against dereplicated contigs was built）。这个补充的矩阵包括14584条，既长(＞1300 nt )，覆盖面又广( ＞ 5X )的contig，他们被称为“病毒类型（virotypes）”。对补充矩阵的分析表明，**每个个体都有一组独特的病毒类型（virotypes）**： 3415种病毒类型只存在于一个个体中(占总数的23.41 % )， 413种病毒类型存在于至少50 %的个体中 ( 2.83 % )，只有18种病毒类型存在于所有个体中 ( 0.1 % )。

微生物组具有一致性的双胞胎共享病毒类型

Twins with Concordant Microbiomes Share Virotypes

我们检查了双胞胎之间共享的病毒类型（virotypes），并观察到同卵双胞胎共享的病毒类型（virotypes）并不比无关的个体多。然后，我们分别评估了微生物组具有一致和不具有一致性的双胞胎：微生物组不具有一致性的双胞胎与无关个体相比不共享更多的病毒类型( p = 0.254 )，而**微生物组具有一致性的双胞胎确实比无关个体共享更多的病毒类型( p = 0.048 )**。此外，**我们还发现微生物组具有一致性的双胞胎比微生物组不具有一致性的双胞胎共享更多的病毒类型**( p = 0.015 )。

在肠道病毒组中噬菌体占优势

Bacteriophage Dominance of the Gut Virome

为了描述病毒组的分类组成，我们试图使用投票系统方法（a voting system approach that）对所有66446个去重复并且有良好覆盖率的contig进行注释，该方法利用了组装后的contig及其编码蛋白中的信息。此外，我们对两个高度丰富的肠道相关噬菌体家族（families）进行了定制注释: ( 1 )crAssphage；( 2 ) *Microviridae*科。为此，我们使用隐马尔可夫模型( Hidden Markov Models ，HMMs )来搜索crAssphage (双链DNA，dsDNA )病毒和*Microviridae*科(单链DNA，ssDNA )病毒contigs。

HMMs使我们能够识别远源的同源物（distant homologs），然后我们将它们与已知的参考序列整合到系统发育树中，以确认注释并更好地解析分类。**我们注释了108个contigs( 19个crAssphage，90个*Microviridae*科)**，验证了68个contigs在科水平的分配（assignments），并为97个contigs分配了一个亚科。对于*Microviridae*科，只有11个contigs以前有分类分配（taxonomic assignment），都属于*Gokushovirinae*，我们确认了这些分配，另外有23个contigs为*Gokushovirinae*，54个contigs为*Alpavirinae*，1个contigs为*Pichovirinae*。对于crAssphage，11个contigs与原始的crAssphage聚集在一起，3个contigs与Chlamydia噬菌体参考基因组聚集在一起，5个contigs与IAS病毒聚集在一起。

在校对了投票系统（voting system）注释和HMM注释之后，总共有12751个contigs( 29.62 % )被分类分配（taxonomically assigned）。**病毒组以噬菌体为主**，只有6.42 %的contigs被注释为真核病毒。与预期一致，**大多数contigs( 96.98 % )是dsDNA病毒**，而只有2.43 %的contigs被标注为ssDNA病毒。*Caudovirales*是最丰富的目，其三个主要科分别为: *Myoviridae* ( 20.22 % ± 4.83 % )、*Podoviridae* ( 10.54 % ± 3.27 % )和*Siphoviridae* ( 35.25 % ± 7.19 % )。crAssphage家族平均构成重叠群的13.26 % ( ± 12.24 % )，在一个病毒组中最大的分布可以达到55.80%，*Microviridae*占病毒群的3.87 % ± 2.57 %。有趣的是，我们观察到*Phycodnaviridae*超过平均丰度的1 % ( 1.77 % ± 1.12 % )，以及与nucleocytoplasmic large DNA viruses (NCLDV)相关的contigs的平均相对分布为3.99 % ± 2.22 %。在所有样品中都存在的18个contigs包括10个标记为crAssphage的contigs，2个标记为“未分类的*Myoviridae*”，2个标记为“未分类的*Caudovirales*”，1个标记为*Microviridae*，3个未分类。在每个样本定义的分类概况中，我们在所有分类水平上寻找具有微生物组一致和不具有微生物组一致性的双胞胎的病毒组组成的差异。**任何分类群在目和科水平上没有显著差异，包括crAssphage和*Microviridae*科。**

我们使用CRISPR映射（mapping）和微生物与噬菌体数据库(microbe-versue-phage，MVP)预测病毒类型（virotypes）和分类特征contigs（taxonomically characterized）的细菌寄主。由于注释噬菌体的宿主，我们没有获得任何注释为真核病毒的contig信息。这些方法使我们能够识别910个contig推定的（putative）宿主。在这910个contigs中，只有一个先前被注释为crAssphage，与预期一致，其宿主被推断为*Bacteroidetes*的物种。我们总共鉴定了1280个推定的（putative）细菌宿主菌株，包括来自多个门87个属的187个物种：其中大多数来自Firmicutes门（92），其次是Bacteroidetes门( 41 )和Proteobacteria门 ( 38 )。每个重contigs的宿主数量中值为1（IQR = 1–2 )，而在菌株水平上，每个宿主的噬菌体数量中值为2 (IQR= 1–3 )。

病毒多样性与微生物多样性相关

Virome Diversity Correlates with Microbiome Diversity

为了评估病毒组和微生物组多样性之间的关系，我们使用从测序数据中获取的三个不同层次的信息检查了病毒组的样本内多样性( α-多样性)和β-多样性: ( 1 )病毒类型（virotypes），( 2 )分类注释的contigs，( 3 )从短reads中注释的基因。

α-多样性

微生物组和病毒组的α-多样性在用于测试相关性的三层信息中的两层中呈正相关(病毒类型（virotypes）和分类注释的contigs)。我们使用带注释的contigs来查询病毒亚组（subgroups）(真核ssDNA、真核dsDNA、细菌ssDNA和细菌dsDNA)中的α-多样性。我们的结果表明真核病毒的多样性与微生物组α -多样性无关。相反，**ssDNA或dsDNA的噬菌体与微生物组α-多样性呈正相关关系。**

β-多样性

我们观察到，当使用Hellinger距离时，与不具有一致性的微生物组的双胞胎相比，具有微生物组一致性的双胞胎，病毒组β-多样性较低；平均binary Jaccard 距离和Bray-Curtis dissimilarity也显示出相同的趋势。与我们观察到的α-多样性相似，**无论使用哪一层信息，具有一致性的微生物组的同卵双胞胎中病毒组的平均Hellinger距离明显低于不具有一致性的微生物组的同卵双胞胎。**把具有微生物组一致性的双胞胎或不具有一致性的微生物组的双胞胎按性别分开时，我们没有观察到β-多样性的显著差异。尽管如此，任何关于性别影响的推断都是有局限的，因为每个群体的个体数量都减半了。此外，当使用注释contigs时，在微生物组和病毒组β-多样性之间观察到显著正相关。这种关系是由噬菌体而不是真核病毒驱动的。

最后，我们比较了相关(同卵双胞胎)和无关个体之间的病毒组和微生物组的成对距离（pairwise distance）。成对距离（pairwise distance）矩阵显示病毒组和微生物组β-多样性测量值之间的正相关，不仅在双胞胎之间，在所有个体之间都是如此。这些结果表明，**不管宿主之间的遗传相关性如何，个体的微生物组越相似，病毒组也会越相似。**

研究讨论

**同卵双胞胎，像其他兄弟姐妹一样，通常比无关个体拥有更相似的肠道微生物组。**此外，**同卵双胞胎总体上比异卵双胞胎具有更相似的微生物组**，尽管在整个微生物组水平上，这种影响很小，这主要是由一小组可遗传的微生物驱动的。然而，在同卵双胞胎群体中，微生物组中双胞胎内部差异可能和异卵双胞胎的一样大。我们利用同卵双胞胎β-多样性的巨大差异，选择了肠道微生物组高度一致或不一致的同卵双胞胎。**我们对它们的病毒环境的分析表明，尽管个体间肠道病毒环境差异很大，而且不管宿主的相关性如何，他们的微生物环境越不相似，他们的病毒环境就越不相似。这种模式是由病毒组的噬菌体驱动的。**

通过从微生物组的β-多样性分布中选择同卵双胞胎，我们去除了宿主遗传相关性作为可能影响病毒组的变量。以前对婴儿双胞胎病毒组和微生物组的研究表明，同卵双胞胎的微生物组和病毒组比无关个体的更相似，这表明共享的宿主基因型和/或环境是关键。相反，一项对成年双胞胎病毒组的研究表明，成年同卵双胞胎没有比不相关的个体拥有更相似的病毒组；然而，根据目前的研究结果，这可能是一个效力（power）问题。事实上，在我们的数据集中，**我们观察到无论双胞胎的微生物组一致还是不一致，同卵双胞胎比无关个体都具有更多相似的病毒组(病毒类型（virotypes）和分类（taxonomy）)。**

之前的研究表明，与成年双胞胎相比，年轻双胞胎的病毒组相似性更高，这与婴儿双胞胎共享更大的环境有关，特别是在饮食方面。Minot等人的研究也表明，相同饮食的个体比不同饮食的个体具有更相似的肠道病毒组。众所周知，饮食是日常微生物组波动的强大驱动力，所以饮食对病毒组的影响可能是由微生物组介导的。然而，我们没有控制饮食，所以我们观察到的微生物组不一致可能是由于双胞胎在取样时饮食不同造成的。**不管影响微生物组一致性的差异的基础是什么，它都与病毒组的一致性密切相关。**

以前没有直接讨论过成年人的病毒组丰度（richness）和微生物组丰度（richness）之间的关系。我们观察到，使用描述病毒多样性的三层信息中的两层，微生物组和病毒组的α-多样性正相关。具体来说，这种模式在病毒类型（virotypes）和分类学（taxonomy）上被观察到，但在基因（gene）上没有。然而，由于观察到病毒基因仅在两类中得到富集，即遗传信息处理（Genetic Information Processing）和核苷酸代谢（Nucleotide Metabolism），我们预计受试者之间病毒组基因的多样性不会有差异。**分类注释层（taxonomic annotation layer）的信息表明，是病毒的噬菌体组分而不是真核病毒推动了这种α-多样性相关的模式。**

病毒组和微生物组多样性之间的正相关关系表明，宿主的可获取性（availability）越强，病毒的多样性就越高。这些观察与“piggyback the winner”模型一致，该模型假设在密度大的环境中，噬菌体选择进入溶源循环，并与其宿主一起复制。事实上，对人类肠道病毒组的纵向研究已经报道了与溶源性相关的基因，温带型的contigs随时间（over time）的低突变率，以及病毒组的长期稳定性，表明了对溶源性周期的偏好。然而，噬菌体捕食行为（phage predation）被认为是维持高度多样性和高效生态系统的一个重要因素，并可能在快速变化的生态系统中扮演维持多样性的角色，如人类肠道中。对病毒组-微生物组相互作用的短期时间序列（time series）分析，以及对噬菌体繁殖中溶源-裂解转变（lysogenic-lytic switch）的更好理解，将有助于解释在人类肠道病毒组中观察到的模式。

这里描述的病毒组的组成与以前报道的成人粪便病毒组的组成有相似性。从经过注释的部分来看，*Caudovirales*目及其科*Siphoviridae、Myoviridae*和*Podoviridae*以及crAssphage是所有样本中的优势噬菌体。Manrique等人将婴儿肠道的噬菌体定殖过程总结如下：真核病毒首先占据新生儿肠道，随后是*Caudovirales*，在2.5岁前*Microviridae*开始成为优势噬菌体。的确，**在我们的样本集中观察到了丰富的*Microviridae*，但是*Caudovirales*是优势群体。**年龄与这里研究的成人受试者的多样性模式无关。

尽管这里描述的每个病毒组具有高度的多样性和独特性，但我们在受试者中发现了一组病毒：所有样本中都存在18个contigs。这些contigs中有一半以上被注释为crAssphage，这与最近这种噬菌体广泛存在的报道一致。我们的数据集中其他共有的病毒类型（virotypes）被分类为*Myoviridae*和*Microviridae*。我们还发现了映射（mapping）到以nucleocytoplasmic large DNA viruses (NCLDVs)、*Phycodnaviridae*和*Mimiviridae*为代表的科水平的contigs。这些类型的病毒越来越多地被报道为人类肠道病毒组的成员。一组核心噬菌体由九个代表组成，包括crAssphage。广泛共享的病毒类型（virotypes）可能表示了，个体之间特定宿主的广泛共享，或者这些病毒在人类微生物组中具有广泛的宿主范围。

我们使用HMMs来注释病毒contigs，这使得我们能够深入探讨病毒组的分类学的内容。除了与公共数据库的比较所揭示的内容之外，我们并确认了这些注释。因为每种类型的病毒(例如，科水平)都需要自己的HMM，所以我们将这种方法应用于几个关键组（key groups）。当应用于crAssphage时，HMM检索的contigs仅与来自粪便病毒的序列聚在一组，而不与来自其他环境(例如陆地或海洋)的序列。这表明，尽管crAssphage是一个多样化的噬菌体群体，但其在人类肠道中的多样性仅限于crAssphage参考基因组，IAS病毒参考基因组或者*Chlamydia*噬菌体相关的序列。我们还将HHM应用于*Microviridae*科，这是一种ssDNA噬菌体。我们能够确认*Gokushovirinae*和Alpavirinae亚科的不同成员的存在。尽管有证据表明，所描述的Alpavirinae基因组构成了*Microviridae*科的第三组，它们对应于溶源噬菌体，这使得很难将它们整合到国际病毒分类委员会(International Committee on Taxonomy of Viruses， ICTV )的分类中；因此，在应用HMM图谱（profiles）前，没有任何contigs被标注为Alpavirinae。

对于每一个病毒分类组（taxonomic group），都有一组相应的细菌宿主。从我们用来选择双胞胎的16S rRNA基因多样性数据中，很明显是哪一个门水平的细菌对具有一致性的微生物组双胞胎和不具有一致性的微生物组双胞胎的微生物组的差异贡献最大。但是与细菌不同的是，我们无法根据病毒中的目或者科来辨别出这样清晰的模式。事实上，大部分噬菌体多样性只归入一个目*Caudovirales*，及其三个科：*Myoviridae*, *Podoviridae*和*Siphoviridae*。这些科的病毒可以感染不相关的宿主。因此，**我们没有必要期望特定的病毒目或病毒科显示在细菌门水平上观察到的这种模式（很明显是哪一个门水平的细菌对具有一致性的微生物组双胞胎和不具有一致性的微生物组双胞胎的微生物组的差异贡献最大）。**

最后，我们注意到了一个有趣的模式，即在选定的细菌种类中，病毒组中完全覆盖细菌基因组。由于这些推定的污染物不是微生物组中最丰富的成员，它们不太可能代表大量（bulk）DNA的随机污染。目前还不清楚为什么某些细菌基因组显示如此高的覆盖率。一种可能性是，我们正在观察具有转座功能的噬菌体寄主的物种范围。噬菌体，比如Mu噬菌体可以随机整合到宿主基因组中，通过连续几轮复制转座扩增，然后可以包装宿主基因组的任何部分。有趣的是，在此检测到的几种污染也被报道为其他人类肠道病毒研究中的污染，这可能表明具有转座功能的噬菌体的宿主特异性。另一种解释包括囊泡产生、基因转移剂和/或广义转导过程。对不同病毒数据库中发现的完整细菌基因组的进一步比较，可能有助于揭示它们的来源，特别是在多项研究中回收的相同细菌物种。

配图

图1. 双胞胎的微生物组不一致

( A )来自之前研究中的354对单卵双胞胎微生物组的β-多样性。每个点代表一对双胞胎的β-多样性，使用加权（weighted）UniFrac (x轴)、未加权（unweighted）UniFrac (z轴)和Bray-Curtis (y轴)方法测量的。平面是最小二乘法（the least squared）拟合的Bray-Curtis ~ Weighted UniFrac + Unweighted UniFrac平面。从两个边缘（edges）中选择具有微生物组具有一致性(蓝色)和微生物组不具有一致性(橙色)的同卵双胞胎子集。黑点表示用于病毒组和整个粪便宏基因组比较的样本。

( B )比较21对同卵双胞胎在门水平上的分类学概况(相对丰度)，1 – 9组的微生物组具有一致性，10 – 21组的微生物不具有不一致性。

( C )微生物组具有一致性(蓝色点，样本量n = 9 )和不具有一致性(橙色点，样本量n = 12 )同卵双胞胎的主要门水平的相对丰度的差异。使用了Mann-Whitney’s U test. \*\*\*表示p < 0.0005, \*表示p = 0.055。

图2. 类病毒颗粒（VLP）在准备过程中的细菌污染

( A )上图单个样品（4A）在去除被认定为污染的read之前的类病毒颗粒（VLP）映射到细菌基因组的热图，下图是去除之后的。细菌基因组用竖条表示，按长度排序，分成了100000bp的bins。中值覆盖率超过了100的reads被认为是污染。左边的色标显示了bin覆盖率，散点图显示了每个基因组bin覆盖率的中位数。

( B )从所有类病毒颗粒（VLP）提取物中识别出来的65个污染基于NCBI的进化树图。右图: 类病毒（VLP）提取物中细菌基因组的丰度与微生物组中16S rRNA基因图谱之间的Spearman相关系数（rank correlation co-efficient (rho)）。左图:所有个体中添加了细菌基因组后的总丰度。 Left: total abundance of each bacterial genome added across all individuals.

图3. 粪便宏基因组和病毒组的功能基因（gene content）比较

( A )粪便宏基因组和病毒组中KEGG类别的相对丰度，包括综合基因目录(Integrated Gene Catalog，IGC)基因的所有命中（hits），不考虑注释。

( B )粪便宏基因组和病毒组在KEGG注释第二季类别中相对丰度的热图，不包括带有未知注释的综合基因目录(Integrated Gene Catalog，IGC)基因。色标显示相对丰度的平方根。A.V.表示附加的病毒组；A.M.表示附加的微生物组(全基因组提取)。A.M.的Intra-class coefficient(ICC) 值为0.99，A.V.的ICC值为0.85。具有微生物一致性的同卵双胞胎的ICC值0.69，不具有一致性的ICC值为0.68。

图4.病毒组组成

21对同卵双胞胎的微生物组在科水平的分类特征比较，1 - 9 组是具有一致性的微生物组，10 – 21组是不具有一致性的微生物组。

( A )同卵双胞胎的病毒组在科水平的组成。

( B )微生物组具有一致性(蓝色点，样本量n = 9 )和不具有一致性的(橙色点，样本量n = 12 )同卵双胞胎在每个科水平的相对丰度差异。

图5. 噬菌体多样性与微生物组多样性相关，但与真核病毒多样性不相关

( A )病毒组香农α-多样性（Shannon α-diversity）与微生物组香农α-多样性（（Shannon α-diversity））的相关性( 样本量n = 42 )。绘制了线性回归95 %置信区间的最佳拟合线。病毒类型（Virotypes）：皮尔逊相关系数（Pearson correlation coefficient）= 0.406，m= 0.3，p = 0.007，R2 = 0.165。分类学：皮尔逊相关系数（Pearson correlation coefficient）= 0.389，m = 0.25，p = 0.010，R2 = 0.151。基因：皮尔逊相关系数（Pearson correlation coefficient），m = 0.11，p = 0.506，R2 = 0.011。

( B )根据注释为真核ssDNA病毒、ssDNA噬菌体、真核dsDNA病毒和dsDNA噬菌体的contigs计算的病毒组香农α-多样性（Shannon α-diversity）与微生物组香农多样性（Shannon α-diversity）的相关性( 样本量n = 42 )。绘制了线性回归95 %置信区间的最佳拟合线。真核病毒: 皮尔逊相关系数（Pearson correlation coefficient）= 0.027，m= 0.034，p = 0.863，R2= 0.000751。sDNA噬菌体: Pearson相关系数= 0.394，m= 0.35，p = 0.009，R2 = 0.155。dsDNA真核病毒: 皮尔逊相关系数（Pearson correlation coefficient）= 0.143，m= 0.15，p = 0.368，R2 = 0.020。dsDNA噬菌体: 皮尔逊相关系数（Pearson correlation coefficient）= 0.400，m = 0.25，p = 0.008，R2 = 0.16。

图6. 病毒组β-多样性模式反映了微生物组的β-多样性

根据分析的三个不同的信息层(病毒类型（virotypes）、基因（genes）和分类学（taxonomy）)，箱线图显示了微生物组和病毒组的Hellinger距离分布，对于微生物组具有一致性的同卵双胞胎(蓝色，样本量n = 9 )，微生物组不具有一致性的同卵双胞胎(橙色，样本量n = 12 )，微生物组具有一致的同卵双胞胎中的无关样本(蓝色，样本量n = 144 )，以及微生物组不具有一致性的同卵双胞胎中的无关样本(橙色，样本量n = 264 )。均值之间的显著差异(Mann-Whitney’s U test, p < 0.020) 用不同的字母表示。

研究方法

粪便样品

本研究中使用的粪便样本是作为之前的研究的一部分而获得的。从16S rRNA基因多样性中，先前共测量了354对同卵双胞胎，他们的粪便样本是在2013年1月28日至2014年7月14日期间收集的。我们基于双胞胎中的三个微生物组β-多样性距离，选择了9对微生物组具有一致性和12对微生物组不具有一致性的同卵双胞胎。这三种距离是，unweighted UniFrac, weighted UniFrac和Bray-Curtis。在微生物组具有一致性和不具有一致性组内都选择了能够在性别、年龄、BMI和BMI差异方面（BMI difference）保持平衡的双胞胎。一致组中的双胞胎年龄在23至77岁之间，包括5对男性和4对女性，微生物组不具有一致性的双胞胎年龄在29至81岁之间，包括5对男性和7对女性。所有涉及使用这些先前收集的样本的工作都得到康奈尔大学IRB的批准。

从人类粪便样品中分离类病毒样颗粒（Virus-like Particles, VLPs）

类病毒颗粒（Virus-like Particles, VLPs）的分离基于之前描述过的方法。对于类病毒样颗粒（Virus-like Particles, VLPs）分离，之前使用0.02 μm滤膜( Whatman )过滤，0.5 g粪便样品在15 mL PBS中涡旋5 - 10分钟，使之重新悬浮。将匀浆在4500 \* g下离心30分钟，上清液通过0.22 μm polyethersulfone（PES ) Express Plus Millipore Stericup ( 150 ml )过滤，以去除细胞碎片和细菌大小的颗粒。然后滤液在Millipore Amicon Ultra-15 Centrifugal Filter Unit上浓缩100K至1毫升（The filtrate was then concentrated on a Millipore Amicon Ultra-15 Centrifugal Filter Unit 100K to ~1 ml）。将浓缩物转移到5 Prime Phase Lock Gel中，并在室温下用200 ml氯仿（chloroform）温育10分钟。在15000 \* g下离心1分钟后，将水层转移到新的微量离心管中，并在37℃用Invitrogen TURBO DNase (14 U )、Promega RNase One (20 U )和1 ml Benzonase Nuclease (E1014 Sigma Benzonase Nuclease) 处理3小时。温育后，向每个样品中加入0.04 体积（volumes）0.5 M EDTA。在进一步处理之前样品被储存在- 80℃环境下。

病毒DNA鸟枪法测序

病毒DNA是用从Invitrogen购买的PureLink Viral RNA/DNA Mini Kit提取的。然后使用从Sigma-Aldrich购买的GenomePlex Complete Whole Genome Amplification (WGA2) Kit扩增每个病毒DNA样品。该步骤包括两个空白对照，但是产量非常低，无法构建文库。然后用Covaris S2 Adaptive Focused Acoustic Disruptor with对扩增产物进行片段化，参数设置如下：duty cycle为10 %，cycle per burst为200，intensity为4 ，duration时长 60 s。每个病毒测序文库按照Illumina TruSeq DNA Preparation Protocol制备，每个样品有一个独特的条形码。所有条形码库汇集在一起。其中一半由BluePippin根据大小选择用来富集更长的插入物( 425 - 875 bp，包括adapters)。在Illumina HiSeq 2500仪器上，两个文库（pools），即“large-insert-size library”和“short-insert-size library”，在Cornell Biotechnology Resource Center Genomics Facility的Illumina HiSeq 2500上，以250 bp paired-end chemistry的Rapid Run Mode模式运行，在独立lane上进行测序。

全部粪便样品的宏基因组鸟枪法测序

基因组DNA是使用PowerSoil - htp DNA isolation kit从每份样品中的大约100毫克样品中提取的。然后，每个测序文库都是按照Illumina TruSeq DNA Preparation Protocol，用500 ng DNA，使用gel-free方法，经过14次PCR循环制备的。在Cornell Biotechnology Resource Center Genomics Facility的Illumina HiSeq 2500上，以2x150 bp paired-end chemistry的Rapid Run Mode模式运行，在独立lane上进行测序。

细菌污染的评估

2017年2月21日，从NCBI FTP上检索了8163个完整的细菌基因组。使用Bowtie 2 v.2.2.8映射（mapped）每个样本的reads到细菌参考基因组上，参数如下: -local-maxins 800 -k = 3。使用view和depth Samtools commands v.1.5，计算每个碱基的基因组覆盖率时，只考虑了映射（mapping）质量超过20的reads。接下来，平均基因组覆盖率为100K bp bins。我们观察到均匀覆盖的基因组的median bin coverage至少是100；那些median bin coverage大于100的基因组被认为是污染。映射到这些基因组的reads被删除。细菌基因组中可能有一个或多个溶源噬菌体；这些溶源噬菌体的爆发（bursting）事件可能会发生，产生几个类病毒样颗粒（Virus-like Particles, VLPs）。作为一项保守措施，为了避免源自溶源噬菌体而非细菌基因组本身的reads丢失，bins with a coverage超过细菌mean coverage三个标准差的bins也被识别并归类为类溶源噬菌体（prophages-like）区域。映射（mapping）到潜在污染基因组的read被标记为“污染”，并从进一步的分析中移除，而映射（mapping）到high coverage bins的reads映射被标记为“可能的溶源噬菌体”。

使用内部Python脚本构建每个样品中每个潜在污染的丰度的矩阵，并通过RPKM标准化（normalized）。与此同时，来自古德里奇等人的数据，每个OTU的相对丰度被回收，并在物种层面使用summarize\_taxa.py qiime脚本进行汇总。针对两组物种，计算了污染物相对丰度与其对应的16S rRNAs数据之间的Spearman rank order correlation between相关性。

功能图谱（profiles）

经过joined和trimmed的来自“short-insert-size library”的reads被映射（mapped）到整合基因目录( Integrated Gene Catalogs，IGC )上，这是人类肠道微生物组中参考基因的整合目录。使用的方法是BLASTX DIAMONAL v.0.7.5，最大e-value值的截止值为0.001，要报告的最大目标序列数目设置为25。

映射到整合基因目录( Integrated Gene Catalogs，IGC )后，使用内部Python脚本生成了丰度矩阵。然后根据整合基因目录( Integrated Gene Catalogs，IGC )提供的每一个基因，使用KEGG对矩阵进行注释。注释的丰度矩阵被稀释(rarefied，subsampling without replacement)为每个样本2000000 read命中（hits）。然后使用QIIME 1.9的命令summarize\_taxa\_through\_plots.py生成KEGG功能图谱。使用R包Psych计算每组(附加的微生物组、附加的病毒组、具有一致性的微生物组样品病毒组和不具有一致性的微生物组样品病毒组)功能图谱的组内相关系数（Intraclass Correlation Coefficient）。

从头（De novo）组装

经过trimming步骤的reads还留着pair（正向和反向），使用针对于短reads的Integrated metagenomic assembly pipeline (InteMAP)流程组装，插入（insert）尺寸为325 bp ± 100 bp。每个样品单独组装。组装第一次运行后，所有干净的reads都使用Bowtie 2 v.2.2.8映射到组装的contigs，参数如下: -local-maxins 800。然后至少一次比对具有一致性的reads都提交到Integrated metagenomic assembly pipeline (InteMAP)，以供第二次组装。使用内部Perl脚本，将来自所有样本的大于500 bp的contigs汇集在一起，并用all versus all方法进行比较。从这一分析中，有可能识别出潜在的环状基因组，并在另一个contigs中讲超过90 %长度的contigs去重复。

根据Roux等人对覆盖度（coverage）和长度（length）的建议，对去重复后的宏基因组集合（assemblies）进行reads的补充，用于构建丰度矩阵。使用Rsubread v.1.28.0将reads映射（mapped）到去重复的contigs。使用内部Python脚本将映射（mapping）的输出解析成丰度矩阵，该矩阵通过每个样品中每一百万条测序reads的contig长度的每一千bp的reads（reads per kilobase of contig length per million sequenced reads per sample，RPKM )进行标准化，并转换为Log10(x+1 )，x为标准化丰度。标准化覆盖率低于5x的contig被除去。最后，应用contig长度过滤获得病毒类型（virotypes）。绘制contigs数量作为长度的函数的衰减曲线，长度阈值就是衰减曲线的转折点1300bp。

HMM注释

建立了独立的HMM图谱，以识别类-crAss（crAss-like） contigs和*Microviridae* contigs。为了构建HMM 类- crAss（crAss-like）图谱（profiles），我们从NCBI网站上下载了类-crAss（crAss-like）家族的主要衣壳蛋白( Major Capsid Protein，MCP )序列。多序列比对(Multiple sequence alignments，MSA )是使用MUSCLE v.3.8.31进行的，并使用UGENE v.1.31.0进行检查。间隙（gaps）超过30 %的位置（positions）被移除。最后，HMM – crAsslike图谱（profile）是使用软件包HMMER v.3.1b2 的hmmbuild构建的。对于*Microviridae*科的例子，还使用了Alves等人开发的病毒蛋白1 (viral protein 1，VP1 )的所有HMM -图谱（HMM-profile）。

使用hmmsearch查询组装好的contig的预测蛋白质，以匹配HMM -图谱（HMM-profile）。e-value值低于1x10-5的蛋白质被认为是真正的同源物，其他的标准还包括，参考蛋白质的大小在参考蛋白质的大小范围内( crAsslike MCP : 450 - 510残基（residues）；Microviridae: 450 - 800个残基（residues）)，至少50 %的覆盖率（coverage）和至少40 %的同一性（identity）。覆盖率和同一性通过BLASTp，用真实同源物查询参考序列来确定。

通过上述步骤的真正同源物被用于系统发育分析。参考序列和同源序列使用MUSCLE v.3.8.31进行比对，使用UGENE v.1.31.0去除至少30 %缺口（gaps）的位点（sites）。使用RAxML v.8.2.4进行最大似然( ML )系统发育分析，使用prottest v.3.4.2获得了最佳进化模型，ML树中节点的支持是通过具有100次伪复制（pseudoreplicates）的自举（bootstrap）获得的。

分类学图谱（Taxonomic Profile）

为了推断组装的类病毒颗粒（Virus-like Particles, VLPs）的分类归属（taxonomic affiliation），使用GeneMarkS v.4.32对所有大于500 bp的contigs预测基因。预测基因的氨基酸序列随后被用于BLASTp搜索中，使用DIAMOND v.0.7.5对NR NCBI病毒数据库进行搜索最大e-value截止值（cutoff）为0.001，要报告的最大目标序列数目设置为25。使用BLASTp结果，每个基因的分类由MEtaGenome Analyzer（MEGAN5 ）v.5.11.3中的lowest-common-ancestor algorithm进行分配。参数如下: Min Support: 1, Min Score: 40.0, Max Expected: 0.01, Top Percent: 10.0, Min-Complexity filter: 0.44。独立地，使用CENTRIFUGE v.1.0.4查询NT NCBI病毒基因组数据库获得每个contigs的分类结果。然后使用投票系统（voting system）分配每个contigs的最终分类注释，其中每个蛋白质的分类注释和contigs的CENTRIFUGE注释被视为投票。通过对contigs的所有可能投票，构建了一个N元树（N-ary tree），每个节点（node）的权重（weight）是包括该节点的投票数。contigs的分类注释将是遍历（traversing）通过权重最高节点（heaviest nodes）的树的结果，但有一个考虑：如果节点的所有子节点具有相同的权重，遍历（traversing）必须停止。分类概况被认为是补充矩阵的一个子集，包含所有由投票系统（voting system）注释或通过HMM概况标注的contigs (见上文)。

噬菌体-寄主相互作用的预测

使用PilerCR v.1.06，从用于评估细菌污染的8163种细菌基因组中识别出有规则间隔的短回文重复序列（Clustered Regularly Interspaced Short Palindromic Repeats，(CRISPRs)。间隔物（spacers）的预期大小为20 bp和72 bp，被用作对病毒类型和分类标注重叠群的查询，使用BLASTn (v.2.6.0+ )和短查询参数。覆盖至少90 %间隔区且e值&lt; 0.001的匹配被认为是CRISPR间隔区-病毒关联。此外，病毒类型和分类宣布的重叠群被映射到MVP数据库中病毒群的代表基因组上。使用LAST-959。由于MVP中的病毒簇包含至少95 %同一性的序列，其长度的至少80 %，因此只保留满足这些限制的匹配。重叠群的宿主是由其匹配的病毒群确定的。

多样性指数

样本内的香农多样性指数( a -多样性)和同卵双胞胎内的Hellinger距离( B-多样性)是使用生成的所有三个丰度矩阵(函数、分类和读取补充矩阵)的Vegan R包的多样性和vegdist函数计算的。使用皮尔逊相关系数（Pearson correlation coefficient）测量病毒组α-多样性和微生物组α -多样性之间的相关性。病毒组β-多样性和微生物组β-多样性之间的相关性是通过使用皮尔逊相关系数（Pearson correlation coefficient）的曼特尔检验（Mantel test using the Pearson correlation coefficient）来计算的。此外，对微生物组具有一致性的同卵双胞胎之间的β-多样性与微生物组不具有一致性的同卵双胞胎之间的β-多样性进行了比较；p值使用Mann-Whitney U test来计算。

统计分析

p值是使用使用Python “scipy”包的Mann-Whitney U测试或Mantel测试获得的；使用Python “scipy” 包将相关系数测量皮尔逊相关系数（Pearson correlation coefficient）；α和β多样性是用R的 “vegan” 包计算的；使用R包“psych”计算类别内系数（Intra-class coefficien）；使用RAxML进行最大似然系统发育分析。

拓展阅读：

1 数据分析过程中使用的脚本

<https://github.com/leylabmpi/TwinsUK_virome>

2 Human Genetics Shape the Gut Microbiome

<https://www.cell.com/fulltext/S0092-8674(14)01241-0>

3 Environment dominates over host genetics in shaping human gut microbiota

<https://www.nature.com/articles/nature25973?platform=hootsuite>

4 Early life dynamics of the human gut virome and bacterial microbiome in infants

<https://www.nature.com/articles/nm.3950>

5 The Human Gut Phage Community and Its Implications for Health and Disease

<https://www.mdpi.com/1999-4915/9/6/141>

6 Healthy human gut phageome

<https://www.pnas.org/content/113/37/10400>