Study notes for – A unified catalog of 204,938 reference genomes from the human gut microbiome

标题：Nature子刊：高达20万个微生物的人类肠道微生物组参考基因组集

<https://www.mr-gut.cn/papers/read/1059900495>

热心肠日报导读：

* 人类胃肠道基因组UHGG集中有20多万种人类肠道微生物基因组；
* 超过60％的肠道基因组无法分配给现有物种，表明大多数UHGG物种在当前参考数据库中缺乏代表性；
* 虽然UHGP总体上包含了更多的蛋白质簇，但大多数新添加的蛋白质在个体样本中的丰度/流行率较低；
* 肠道菌群可能包含许多具有重要代谢活性的物种，尚待在实验室条件下进行培养和功能鉴定；
* 各大洲之间有很高的菌株变异性，并且仍有相当程度的多样性有待发现。

推荐语：人类肠道微生物组与人类健康和疾病有关的重要表型密切相关。 缺乏足够微生物多样性的完整参考基因组，妨碍了对微生物组中单个物种的作用及其功能和相互作用的理解。因此，建立微生物参考基因组和基因的整合数据库，是准确表征肠道微生物物种分类和功能的重要步骤。本文生成了代表人类肠道微生物组的20万个基因组和1.71亿个蛋白质序列的基因集，其中的4,644个物种中，有71％缺乏培养的菌株，表明大多数微生物仍有待实验表征。本研究获得的高质量参考基因组集可极大提高宏基因组研究的分辨率和准确性，对人类肠道微生物组与表型的关联分析提供了数据基础。遥想十年前Science和Nature发表的微生物组里程碑研究中鉴定出的178个微生物基因组和330万个微生物基因，这十年间科学家们在微生物组参考基因集方面的进展速度可谓是呈指数级增长。

**引言**

在2019年，在Nature， Nature Biotechnology(NBT) 和Cell 期刊相继发表的几篇大规模的基于可培养或宏基因组组装的微生物基因组集（genome catalog）- 参考集如下，以前所未有的规模，在基因组水平上扩展了我们对人体肠道微生物群落及其功能多样性的认识，同时也让我们认识到不可培养微生物所占比例之大并且其功能和代谢特点显著相异于可培养微生物。这篇在2020年7月下旬发表在NBT上的文章是在综合归纳和总结之前多国肠道微生物研究的数据基础上，提供了更加全面的微生物基因组集及基于基因组信息的非冗余蛋白集。这在很大程度上提高了肠道微生物宏基因组研究的比对比率，并且使得分类注释可以深入到菌株水平。

1. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. 37, 179–185 (2019).

2. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. 37, 186–192 (2019).

3. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. Nature 568, 499–504 (2019).

4. Nayfach, S. et al. New insights from uncultivated genomes of the global human gut microbiome. Nature 568, 505–510 (2019).

5. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000

genomes from metagenomes spanning age, geography, and lifestyle. Cell 176, 649–662.e20 (2019).

**摘要**

全面且高质量的参考基因组对于人体肠道微生物群落组成的物种分类和功能分析是必要的。本文中我们提供了一个人体肠道基因组整合集（unified human gastrointestinal genome, UHGG），包含20万4938个非冗余基因组，隶属于4644个原核生物种。这些基因组编码大于1.70亿的蛋白质序列，被汇总归结在人类肠道微生物蛋白整合集（unified human gastrointestinal protein, UHGP）。相比于之前的IGC (integrated Gene Catalog)，这个UHGP涵盖的数量是之前的两倍多。通过比对当前的主流功能注释数据库，在UHGG当中有70%找不到其对应的可培养微生物代表，有大约40%的UHGP的功能无法被注释。种内基因组变异性分析结果表明：修饰基因（accessory genes）和单核苷酸多态性（single nucleotide variants, SNVs）广泛存在，某些SNVs呈现人口地理分布的特异性。总之，这个人体肠道微生物基因组和蛋白质组的整合集将会使研究者更加高效的将微生物基因型及其表型关联在一起进行分析。

**背景介绍**

自人类微生物计划启动（2008年），研究者相继发现百上千的新细菌基因组。在2014年李俊桦等通过整合源于欧洲，美国和中国多达1200个样品的人体肠道微生物宏基因组测序结果，提供了一个更加全面的人体微生物参考基因集。这个基因集被广泛用于人体肠道微生物与人体疾病（比如，二型糖尿病，肥胖症等）的相关性研究中。但是，基于**非冗余基因集的研究存在其自身的缺陷性**，例如, 我们**无法将这些比对到的非冗余基因对应到其所在的微生物基因组**，这就使得我们无法从基因组的整体水平上去认识某些富集或者缺失的基因以更好的认识其功能。

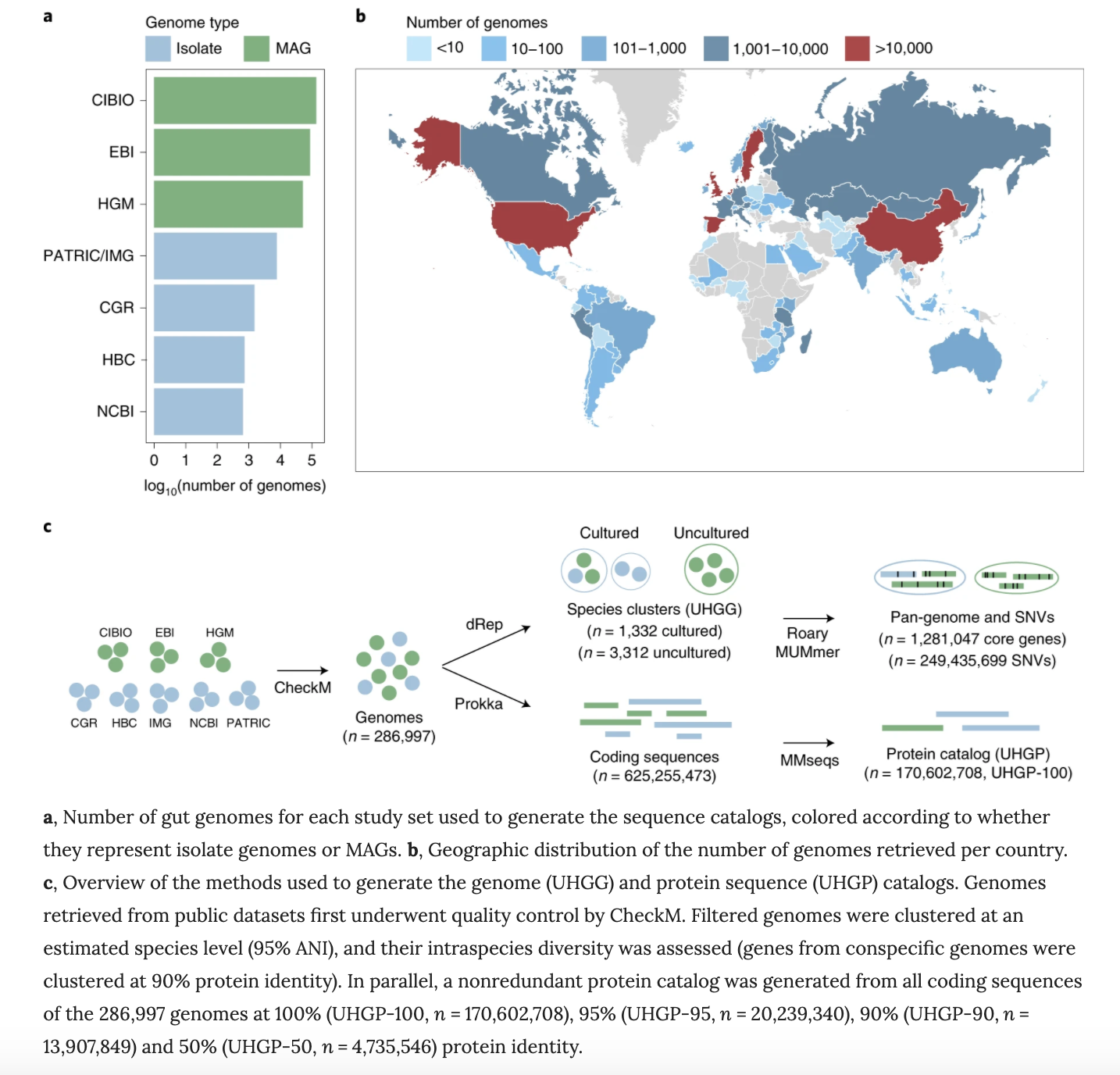
基于分离培养的肠道微生物组的基因组和相应的功能试验验证分析，不断扩充了我们对肠道微生物功能及其对人体群落结构稳定性以及宿主健康的影响机制的理解。与此同时，基于宏基因组序列拼接（assembly）和分箱（binning）技术所获得的宏基因组组装基因组（metagenome-assembled genomes, MAGs）也以前所未有的速度迅速积累。但是，MAGs拼接也面临着错误拼接和分箱的可能性，因此，对于MAGs基因组比较分析解释时需要特别谨慎。以上基于分离培养和宏基因组测序拼接所积累的大量的微生物基因组，在很大程度上扩大了人类肠道微生物可知物种库，本文就综合了之前的研究结果并建立一个综合的肠道微生物基因组集合（UHGG）及其蛋白质组集（UHGP）。

**结果**

**UHGG涵盖范围及其质量描述（见下图）**

图1. 人类肠道微生物组的整合序列集。

Fig. 1 | The unified sequence catalog of the human gut microbiome.



a，用于产生序列集的每个研究集的肠道基因组数量，根据它们代表分离的基因组(蓝色)还是MAGs(绿色)进行着色。 b，每个国家/地区检索到的基因组数量的地理分布。 c，用于生成基因组（UHGG）和蛋白质序列（UHGP）集的方法概述。 从公共数据集中检索的基因组首先由CheckM进行质量控制。 过滤后的基因组采用dRep以估计的物种水平（95％ANI）聚类，并评估其种内多样性（来自同种基因组的基因以90％的蛋白质同一性聚类）。 同时，从286,997个基因组的所有编码序列中生成了非冗余蛋白集，分别为100％（UHGP-100，n = 170,602,708），95％（UHGP-95，n = 20,239,340），90％（UHGP-90，n = 13,907,849）和50％（UHGP-50，n = 4,735,546）蛋白质同一性。

图注. 是文章的重点。重点看图。

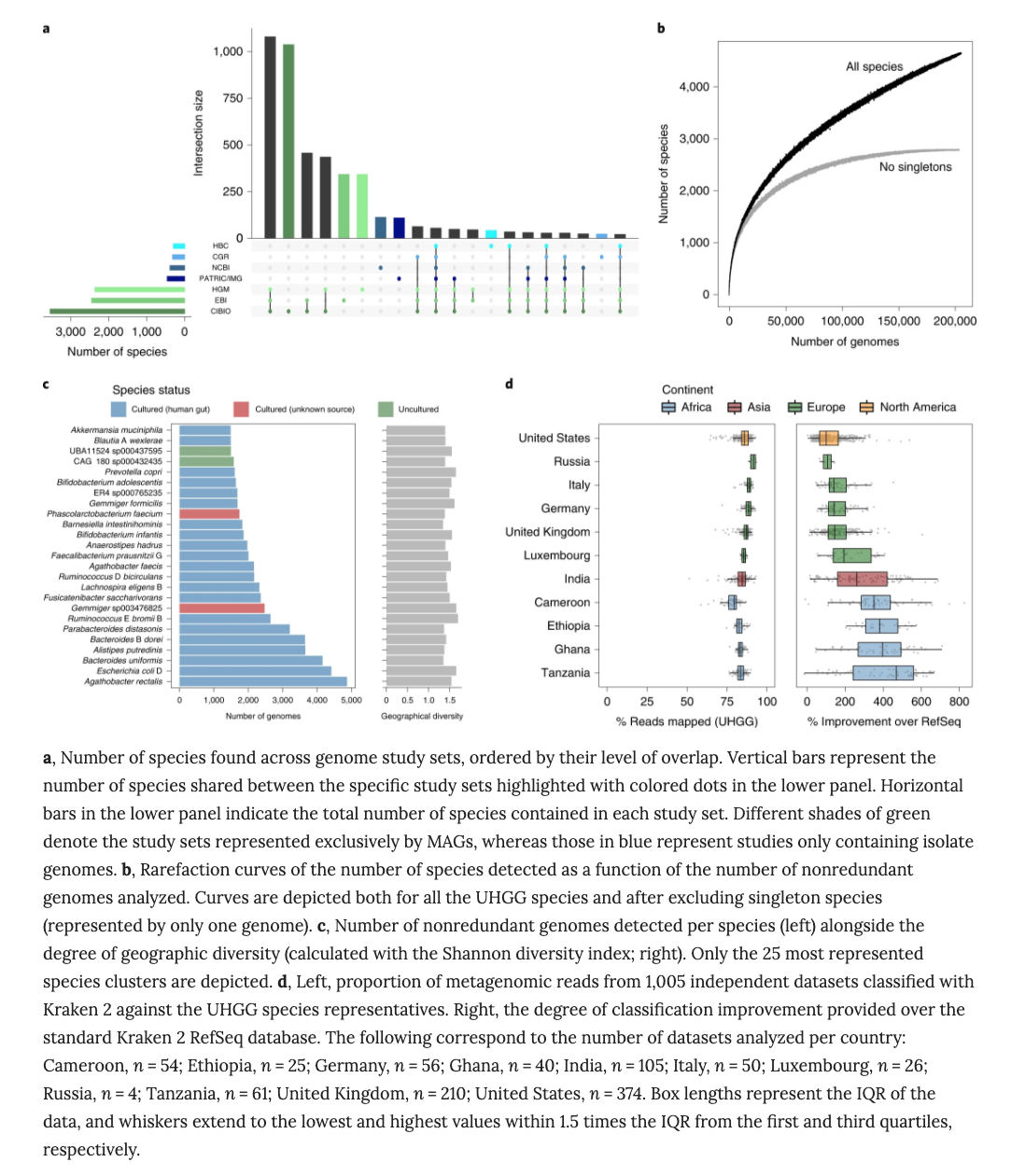
本研究中涵盖了2019年之前几个大规模微生物基因组数据：其中分离菌株基因组来自于Foster et.al. (2019) 等HBC（Human Gastrointestinal Bacteria Culture Collection）包含的基因组，Zou et.al. （2019）等CGR（Culturable Genome Reference）包含的基因组以及其他数据库中的可培养微生物基因组（包含NCBI，PATRIC和IMG）；另外，MAGs基因组主要来自于Pasolli et.al. （2019）CIBIO，Almeida et.al. （2019）EBI和Nayfach et.al. （2019）HGM研究中所获得的微生物基因组。对以上所有基因组的筛选条件包含有三个：**1）>50%完整度；2）<5%污染度；3）质量分数（完整度-5\*污染度）>50**。通过这个条件筛选之后，一共获得了286997个高质量基因组，其在各数据库中的分布如图1a：

为了建立物种水平上的微生物集，以上286997基因组被聚类成物种，这里物种是基于两个基因组间核酸序列一致性定义的 - **在大于30%的可比对序列长度上具有>95%的平均核酸一致性即被定义为一个物种**。每个物种的代表基因组是从其聚类堆内选择的质量最好的一个作为代表，并且这些代表基因组综合起来建立了UHGG集合。基于此定义，以上286997个基因组被聚类到4644个原核生物种，其中4616属于细菌而28个为古菌，其中3207的代表基因组的完整度>90，污染度<5%，并且其中573个基因组含有5S，16S，23S以及至少18个tRNAs（这个是Genomic Standards Consortium定义的高质量MAG的标准）。GTD-Tk（Genome Taxonomy Database Toolkit）被用来注释这些物种代表基因组的物种信息。结果表明，有大于60%的肠道微生物基因组无法被注释到物种，也就是说有大量的UHGG在对应的物种数据库中是缺乏代表物种的。与此同时，作者也使用CMseq分析了所有MAGs的菌株异质性（strain heterogeneity=proportion of polymorphic porsitions），其中位数是0.06%，这个数值低于0.5%，意味着这个研究中得到的MAGs是数据高质量的MAGs伴随着很微弱的菌株异质性。

**UHGG与之前研究结果及其数据库的比对（见下图）**

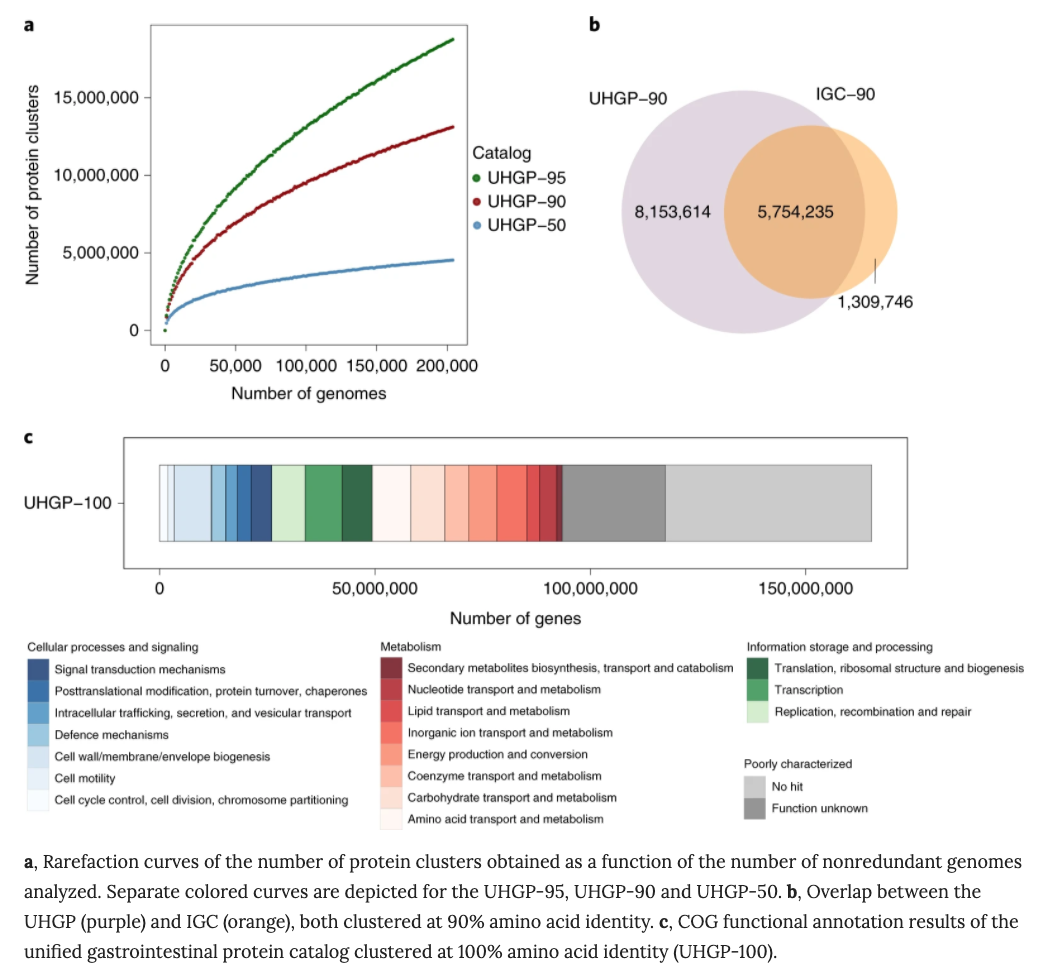
几个大规模的MAGs研究，虽然使用了不同的序列拼接，分箱及其提炼优化方法，其生成的MAGs存在很大比例的重合度，其中1081物种分别同时在CIBIO，EBI和HGM获得。通过比较这几个研究中均用到的样品MAGs发现，79%-86%的MAGs在所有的研究中都有检测到。另外，通过比较同一个样品在不同研究中所拼接获得的属于同一个物种的MAGs发现，对应物种的ANI和AF的中位数分别为99% 和92.1%，其中高质量（90%完整度）基因组其对应的AF为94.5%，然而中间质量的基因组，这个AF只有86.6% （Extended Data Fig. 3c）。对应的物种饱和曲线表明，不可培养微生物种在此研究中并未达到饱和，但是这个不饱和很有可能是稀有种所引起的，因为当单基因组单独代表的物种去掉时，物种曲线接近饱和。不同于MAGs在不同研究中的高度一致性，对应在HBC，CGR和NCBI中的可培养微生物基因组的重叠性很小，大部分（70%）是研究或数据库特有的。大约81%（3705）UHGG肠道微生物中在现有的肠道微生物可培养数据库中找不到对应的代表性物种。当作者将比对数据库扩展到NCBI RefSeq所有的可培养菌时，另外的438个物种找到了对应的物种信息，即使这样依旧有3312（71%）的UHGG物种在现阶段缺乏可培养种代表。其中大约有66%和31%的细菌和真菌没有对应的可培养微生物代表。几个最大的代表性分支为4C28d-15目（包含167个物种），RF39目（包含139个物种）和CAG-272目（包含67个物种）。

另外，通过比较对应物种所包含的基因组数目发现，包含最多细菌基因组的前25个物种中仅有两个物种是没有可培养菌株的，排名前三的细菌物种是Agathobacter rectalis, Escherichia coli D和Bacteroides uniformis，而对应的古菌为Methanobrevibacter A smithii。本文中也计算了这些UHGG微生物种的地理分布Shannon多样性，结果表明其中最大的物种不受到样品地理位置的限制，而是在不同地理位置分布的样品中呈现相似的高丰富度。为了更好的评估本文UHGG在宏基因组物种注释中的独特优势，作者选取了1005个宏基因组数据并将序列使用Kraken2比对到UHGG以及RefSeq数据库，结果表明前者足足提高了155%的比对比例相比也后者而言，其可比对比例高达85.9%（中位数）。并且UHGG作为参考基因组的比对优势在非西方国家的宏基因组样品中更为明显。



**UHGP蛋白集合的涵盖范围及其组成特性（见下图）**

首先UHGP是通过注释所有的UHGG所包含的286997个基因组而获得的蛋白组集合，这个不同于之前流行的非冗余基因集概念，因为后者没有对应非冗余基因的基因组信息而UHGP蛋白集合中对应的蛋白条目可以找到其所在的基因组来源。UHGP包含625255473个完整蛋白序列，随后这些条目在50%，90%和95%蛋白序列一致性标准基础上被聚类并命名为UHGP50，UHGP90，UHGP95和UHGP100。饱和曲线显示，在90%和95%一致性水平上，非冗余蛋白条目的数量呈现持续上升而未能饱和，相比而言在50%相似度水平上，对应的蛋白条目随着样品数量的增加达到饱和。

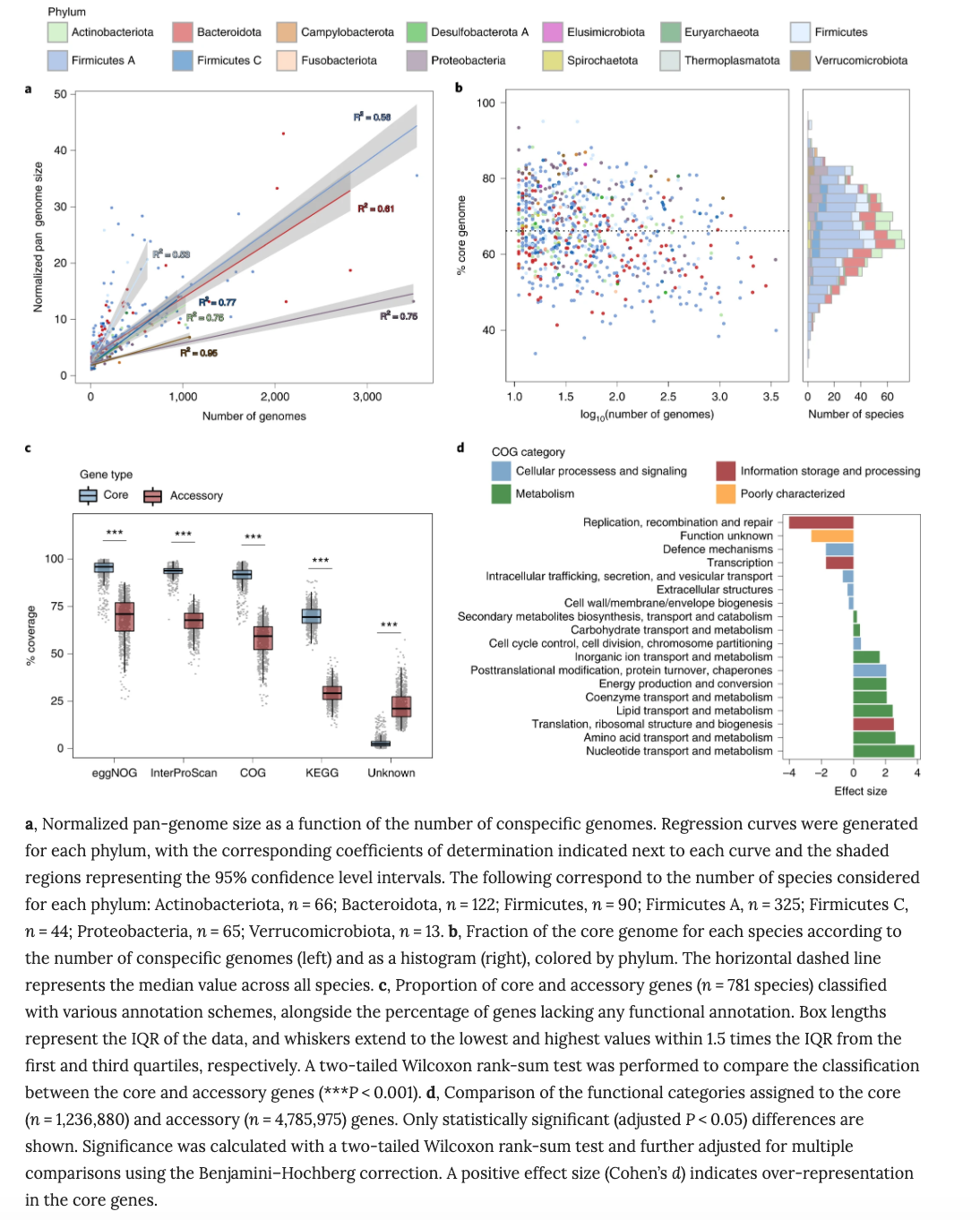
为了评估这个蛋白集合作为参考集其比对能力，文中以之前的IGC作为参考并将其非冗余基因组以同样的90%蛋白序列一致性进行去冗余，结果得到包含有7063981条目的IGC-90蛋白集合。结合UHGP-90的13.9百万的蛋白条目，两者一起共收录15.2百万蛋白簇，其中5.8百万共存于UHGP-90和IGC-90集合里。相比于IGC-90而言，UHGP-90收录的条目数量同比增加了115%。然而基于1005个样品宏基因组的比对结果显示，UHGP-90并没有在很大程度上增加对应测序系列的比对率（mapping ratio），对应的提高比例只有5%。这个结果暗示，极有可能UHPG中很多的蛋白条目隶属于样品中的低丰度蛋白簇。为了降低因MAG污染而带来的虚高的蛋白质组集合，最后的UHGP只选用了那些在同物种基因组中至少出现在两个不同的基因组中的一类，基于这个筛选条件所得到的UHGP-95，UHGP-90和UHGP-50分别包含10798224，8082122和3088278个蛋白条目。

**UHGP功能组成，（见下图）**

通过比对当下的主要功能注释数据库（eggNOG，InterPro，COG和KEGG），UHGP-100中存在27.3%的条目在以上数据库中无法匹配到的对应的功能。基于COG的注释结果显示，其中高代表性的功能类群主要参与氨基酸转运和代谢，细胞壁/细胞膜/envelope 生物发生以及转录。另外为了比较微生物物种间功能组成差异，作者们将同隶属于同一个物种的基因组的蛋白序列在90%的氨基酸序列一致性的标准下进行了去冗余，对应物种的pan-genome被坚定到363个KEGG模块（module）。其中保守性的模块的功能主要参与如下过程或者功能：核糖体结果，糖酵解，肌酐一磷酸生物合成，糖原异生以及蔓草酸酯信号路径。另外，本文发现某些细菌门对应的pan-genome可注释的功能比列相对较小，比如Myxococcota, Bdellovibrionota, Thermoplasmatota, Patescibacteria和Verrucomicrobiota。与此同时，某些细菌类群对应的碳水化合物先关的活性酶的编码基因含量相对较高，比如：Fibrobacterota，Bacteroidota，Firmicutes I，Verrucomicrobiota和Patescibacteria.

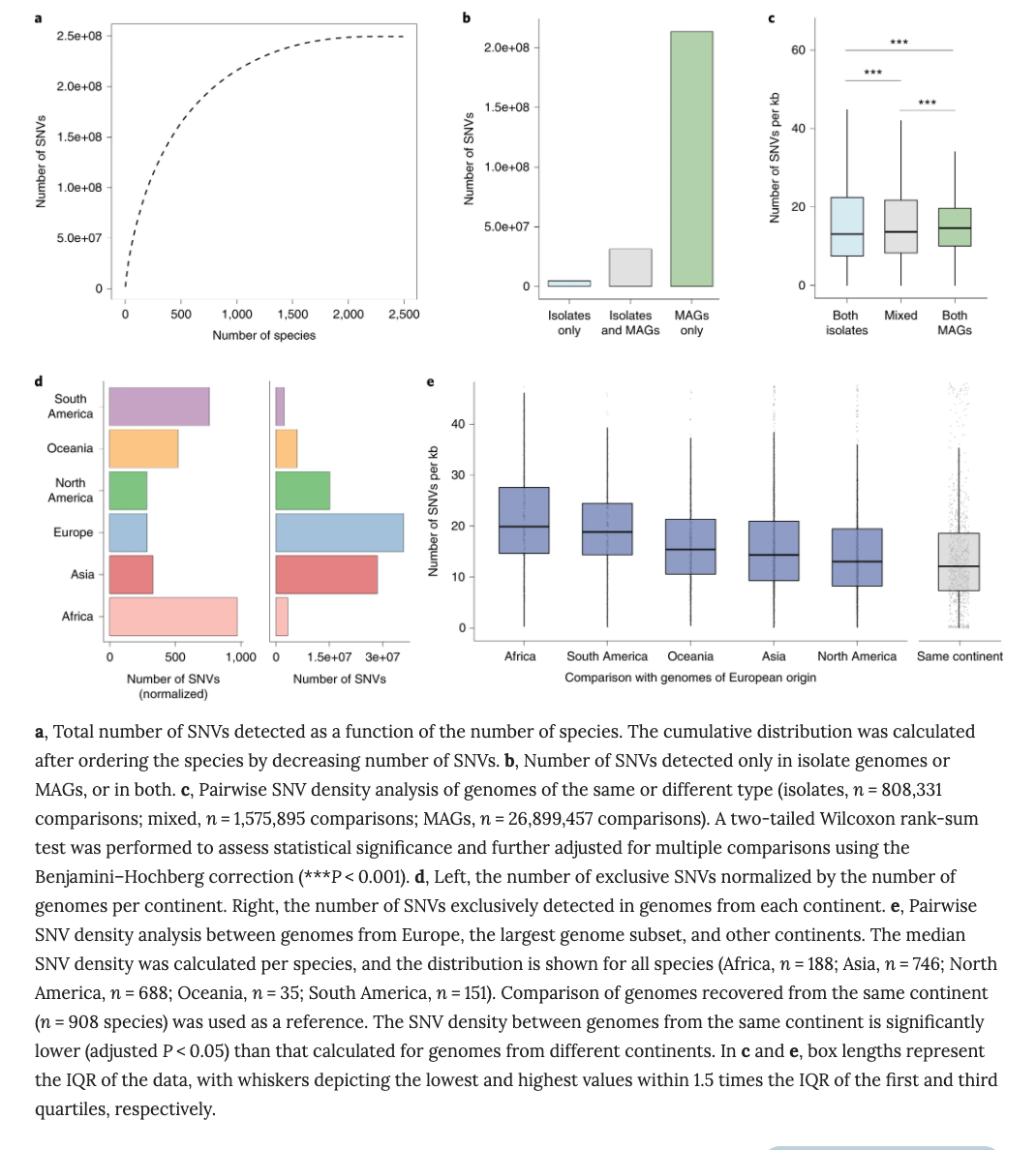
为了比较种内核心基因以及修饰基因组成特点，只有那些完整度>90同时对应物种的基因组总数大于10的物种（共781）及其基因组被用于比较分析。每个物种的基因频数分布呈现双峰模式，总体上来讲大部分的基因或属于核心基因（出现在其对应物种>90%的基因组中））亦或属于修饰基因（出现在其对应物种<10%的基因组中）。另外不同微生物种pan-genome大小随着其基因组个数的增加速率不同，其中Firmicute对应的增长速率较大（但是个人认为这个趋势并没有特别明显，线性关系有点牵强）。但是有一点比较有意思的是，物种对应的核心基因占总基因的比例存在很大的差异，即使当同一个细菌门的两个物种包含基因组个数都高达1000，其核心基因比例依旧差异很大。

通过对核心基因和修饰基因进行功能注释发现，对应核心基因的可注释比例（96% eggNOG，94% InterPro，92% COG和69% KEGG）明显高于修饰基因，后者平均有21%基因无法被以上数据库注释。基于COG数据库的功能注释结果显示，核心基因编码蛋白功能往往与关键的碳，氨基酸以及脂肪的代谢途径及其它看家功能相关。相比而言，修饰基因编码的蛋白往往与微生物的防御机制相关，包含广谱耐药性机制相关的ABC transporter efflux pumps以及更加具有目标性的对噬菌体病毒的防御机制，比如，CRISPRA-cas和restriction modification systems。



**种间单核苷酸多态性（见下图）**

另外本研究也分析了物种内不同基因组的单核苷酸多态性（single-nucleotide variants, SNVs），结果显示85%的SNVs是MAGs特有的，而可培养微生物的基因组特有的种内SNVs仅有2.2%。另外发现，非洲肠道微生物样品种的SNVs比例相比于其所包含的总基因组数量而言比例远远高于欧洲，美国以及亚洲样品。



**相关领域研究者的福利**

**数据可用性**

UHGG的基因组已经被存放在European Nucleotide Archive数据库的ERP116715研究序列号之下。本文中生成的UHGG，UHGP以及SNV，另外其对应的功能注释，泛基因组结果以及基于UHGG而生成的Kraken2格式的参考比对数据集等MGnity ftp网站（<http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/>

）均可见。 另外UHGG基因组集合和UHGP蛋白质组集合已经被并入到MGnity网站。另外，此研究还生成UHGG对应的Bitsliced Genomic Signature Index (BIGSI)，使得用户可以通过交互式界面从数据库中搜索到搜寻序列（query sequence <5kb）对应在数据库中的相似序列的相关信息。为了保持数据库能及时的并入最新研究的数据，MAGs也会从ENA（European Nucleotide Archive）数据库中提取相应的MAGs数据并及时并入数据库中以补充或者取代低质量的代表序列。