第 3 章

宏基因组样本收集

摘　要：高质量的样本是进行宏基因组学研究的先决条件。早些年，辛勤的学生每天早上都要在医院里等待，以确保采集到的每一勺粪便样本都能保存在冰冻的无菌管中，以最快速度冷冻于 –80℃的冰箱，随后装上大量干冰分批运输。本章涉及对样本采集步骤的建议，从样本收集、测序、统计分析时需要考虑的因素。除了宏基因组测序外，还需要多种方法来评估低生物量样本中的微生物数量，并将微生物与潜在功能联系起来。污染控制和身体多采样位置的研究设计的原则同样适用于各种微生物组样本。

关键词：宏基因组鸟枪法测序，扩增子测序，核糖体 rRNA 基因，微生物组，低生物量标本，胎盘微生物组，脂肪组织，宏基因组全关联研究，标本储存

3.1　样本中非微生物的部分，会影响DNA提取和测序量

根据布里斯托（Bristol’s stool score，BSS）大便分类法，可根据形状对粪便样本打分，这是对样本中水分含量、肠道经过时间以及样本中微生物数量的有效估计 [1-3]。被试者可根据问卷备注给出粪便样本 BSS 打分。问卷得出的结果和粪便宏基因组样本中得到的次级胆汁酸代谢酶数量相关，而这在机制上是说得通的 [3]。

自动采样马桶可以自动记录诸如 BSS 得分、尿样的体积和流速，但还不能自动采集样本 [4]。

每克宏基因组样本中的微生物数量，受到食物残渣的影响，这对于诸如大熊猫这类食草动物尤其明显（图 3.1A）。一项基于英国健康成人饮食的粪便研究，在剧烈震荡和洗涤剂处理后，发现样本中 55% 为细菌、17% 为纤维、24% 为可溶性物质。随着微生物群落的不同，样本中细菌所占的比例会有所差异。

可折断的尼龙拭子，能被用于粪便和口腔样本采样（图 3.1B）。皮肤或鼻腔采样常使用拭子，采样前通常会用生理盐水或缓冲液浸润，但目前尚不清楚短暂接触矿物质会对微生物群落有多少影响。对于生物量较低的样本，需要的采样量会更高。为避免样本在运输过程中漏液，需谨记拧紧瓶盖。如果样本是受试者自采集的，尽量配有清晰的图片及视频说明采样流程，并对样本拍照。不同队列间受试者的配合程度不

同，需要尽早开展质控环节。

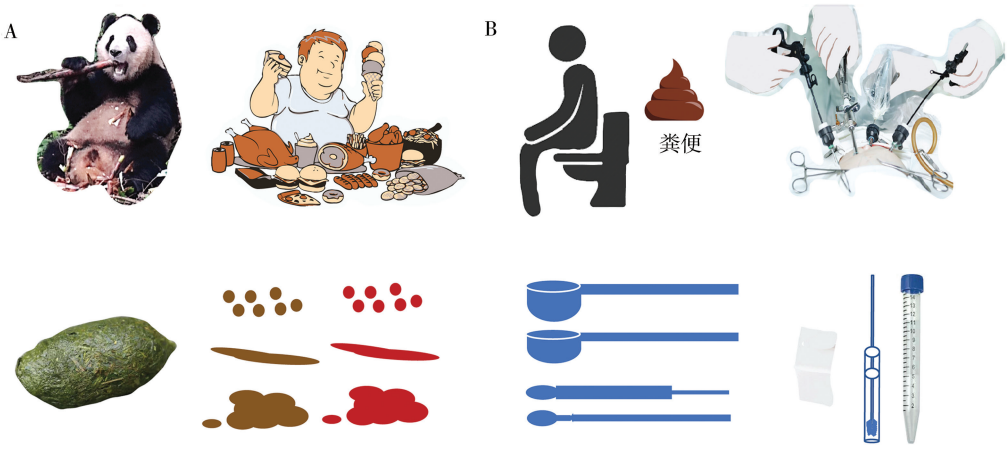

图 3.1　在采集宏基因组样本时，需获得足量 DNA

A. 宏基因组样本可能含有除微生物之外的其他物质。人类粪便按照简化的BSS打分描述，最硬的BSS得分为1，中间硬度的为4，水分含量最多的为7。B. 采集样本时，固体样品可用勺子，表面可用拭子，液体可用管子。腹腔镜或者其他新技术可以在手术室采集微生物组样本，降低切口污染的概率。皮肤和其他表面采样前都要经过净化。棉签或刷子可以在保护管之前到达取样地点。

若受试者患有痔疮或炎症性肠炎等胃肠道疾病，粪便样本中会包含血液，导致宏基因组样本中人类基因组的比例远超 1%（表 3.1）。除了粪便和龈上菌斑，大部分人类宏基因组样本中都包含更高比例的来自人类基因组的序列，甚至在某些组织的样本中高达 99%。在签署知情同意书后，宏基因组获得的人类低深度全基因组测序数据，可以与微生物数据共同分析，但需要注意不同来源组织与来自血液的人基因组数据存在差异。

表 3.1　来自不同采样部位的人类宏基因组样本中，采用短序列测序时，包含的人源 DNA 序列占比

采样区域	采样点	人源基因比例	参考文献
肠道	粪便	1%	[5,6]
肠道	粪便（克罗恩病）	20% 或以上	[7]
口腔	颊黏膜	82% ~ 90%	[5,8,9]
口腔	龈上菌斑	40%，5.6%	[5,10]
口腔	龈下菌斑	79%	[5]
口腔	舌苔	30%	[9,11]
口腔	唾液	77% ~ 91%	[5,11]
皮肤	干燥（如前臂）	36%	[12]
皮肤	潮湿（如肘窝）	44%	[12]
皮肤	皮脂（如耳后皱褶）	59% ~ 73%	[5,12]
泌尿生殖道	阴道	90% ~ 98%	[5,13]
泌尿生殖道	子宫颈口	98%	[14]
泌尿生殖道	腹水	99.8%	[15]

续表

采样区域	采样点	人源基因比例	参考文献
泌尿生殖道	胎盘	＞ 99%	[16]
呼吸道	前鼻孔	96%	[5]

注：这并不是一个详尽的列表。当研究人员和临床医生决定宏基因组NGS的测序数量（例如，10 GB的PE100的读数），或选择扩增子测序、原位杂交等方法时，该表可提供参考。

通过低速离心去除宿主细胞的方式会导致某些细菌物种的消失，这在支气管肺泡灌洗（bronchoalveolar lavage）样本中已被论证[17]。通过分子生物学或者化学手段去除宿主细胞，也会影响微生物的组成[15-18]，但这可在未来进行优化。目前我们推荐进行无偏测序，之后使用生物信息学手段去除宿主。

根据研究问题选择适当的技术组合（如胰腺癌中的细菌，图 1-9）。对于采集完备的样本，若仅仅因为数据量过大而不去穷尽其分析潜力则很可惜。毕竟，很多微生物的演化和互作发生在局部。

3.2　对于粪便及生物量较低的宏基因组，要注意减少采样过程中每一步的污染

正如在第 1 章简要提及的，对于生物量较低的宏基因组，要注意采样过程中每一步的污染风险，例如对胎盘样本是否包含微生物存疑（图 1.8），但对于肿瘤样本接受度就更高（图 1.9）。研究生物量较低的宏基因组样本，有助于建立在采样过程中每一步注意事项的清晰认识（图 3.2）。另外，证明某一部位缺少微生物，比证明微生物的存在更难。毕竟大自然中无论怎样极端的环境，细菌、真菌和古菌总能占有一席之地。

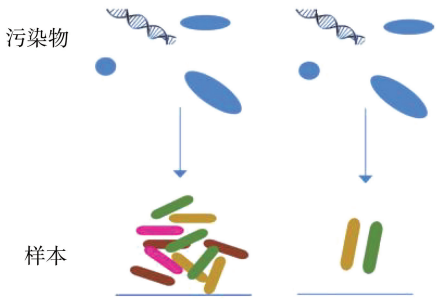

图 3.2　采样过程中需注意事项

从采样到测序的任何步骤都可能引入污染的DNA或微生物。只是生物量低的样品更容易被污染物淹没

对于粪便、口腔和阴道样本，通常拭子中包含超过 10^{10} CFU 微生物细胞。庞大的细胞数使我们在样本采集、DNA 提取、文库制备及测序过程中，无需为避免污染

而过度谨慎担忧。而对于生物量较低的样本（图 3.2 和图 3.3），需要经常检验上述步骤所用试剂是否包含活的或死的微生物[25]。例如，使用扩增子测序（图 3.4）。关于样本污染，经常被引的一篇文献使用连续稀释的沙门氏菌，在五次连续稀释之后，即包含约 1000 个沙门菌细胞时，16S 测得的其他菌看上去更多了[26]。尽管实验关注的是试剂污染，作者并没有报道有关稀释过程是否引入了污染的信息（PCR 扩增是在一个超净工作台里用高压灭菌微量离心管和滤芯移液吸头进行的），也没有报道有采用什么措施避免样本被操作人员污染，这在古 DNA 领域中众所周知。在进行扩增子测序时，还需要考虑批次效应带来的影响[27]。

哪怕是医院的房间，甚至包括空间站，在使用后也包含微生物，它们可能来自患者和员工[28-32]。医院的通风系统中，可能检出细菌或真菌（图 3.3），当其与我们关注的样本存在差异时，仍可以认为样本中包含微生物，并寻找进一步的证据。对于作为阴性对照的培养皿上的活菌群以及 qPCR，上述结论也适用。

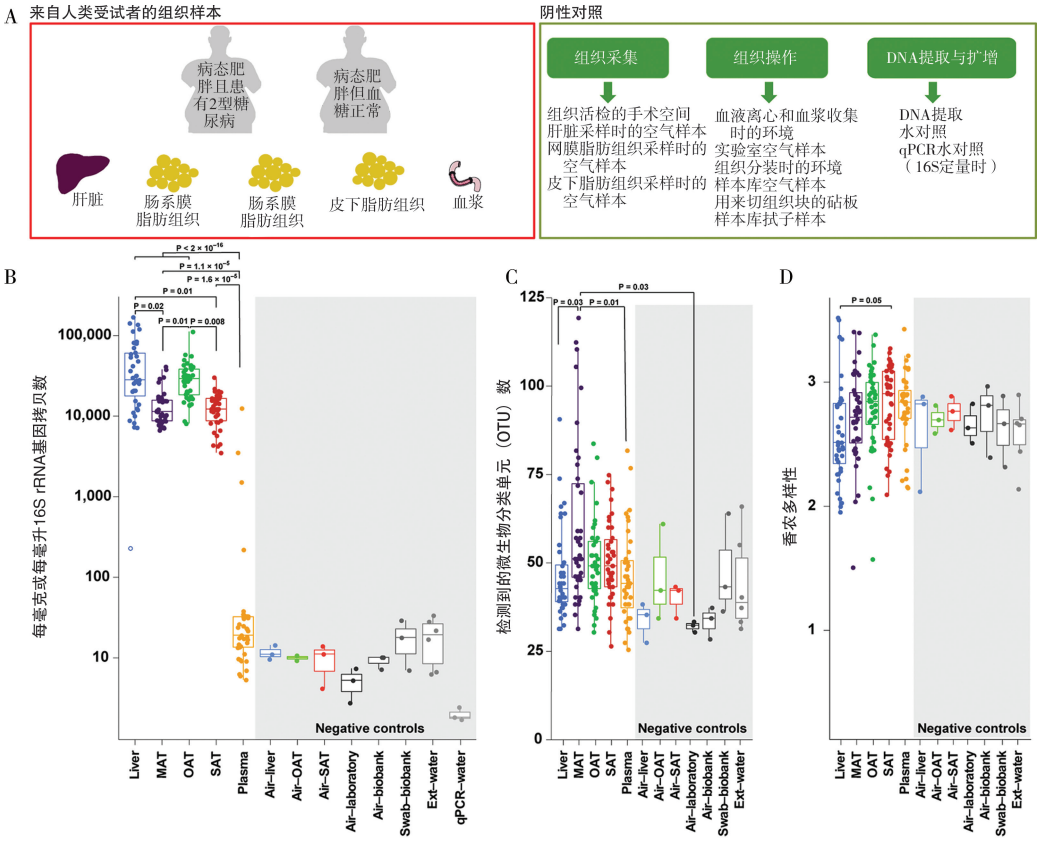

图 3.3　一个低生物量样本研究设计的例子

包括阴性对照以及不同的生理条件下的关联研究。A. 采集病态肥胖者，2型糖尿病T2D（n=20）和正常血糖（n=20）的肝脏、3个不同的脂肪组织（OAT、MAT、SAT）和血浆样本。采用优化的血浆和组织细菌DNA检测条件进行DNA提取和扩增。在分析的主要步骤：组织收集、组织操作、DNA提取和扩增，采用一套全面

的阴性质控样本，来检测环境带来的样品污染。在组织收集过程中，输卵管在整个手术过程中（空气-肝脏、空气-AOAT和空气-SAT）一直开放在手术区附近。来自组织操作的污染由另一组管子予以避免，在整个血液离心和等离子体收集（空气实验室）以及组织外引用（空气生物库）过程中，这些管子一直开放在操作者旁边。用于检验组织的砧板在组织操作（拭子生物库）之前被取样。采用定量PCR技术，使用纯水样本对DNA提取（ext-water）过程中的标本、试剂和（或）环境污染进行质控，并对组织16S RNA 进行扩增定量分析。在逐个病例对阴性对照进行全面验证后，16S定量和测序数据被用于发现与T2D相关的组织特异性细菌特征。（B～D）身体各部位的细菌数量。B. 16S rRNA基因计数。C. 观察到的OTU分类。D. 肝脏中的香农指数、三种不同的脂肪组织（OAT、MAT和SAT）和肥胖者的血浆。在分析的主要步骤：组织采集（空气-肝脏、空气-OAT、空气-SAT）、组织操作（空气-实验室、空气-生物库和棉签生物库）和DNA提取或扩增（纯水、qPCR-水）中，检测阴性对照以控制环境样本污染。在图（B～D）中，各组用Kruskal-Wallis单因素方差分析进行比较，然后用Dunn的成对检验进行比较。方框图描绘了第一和第三个四分位数，中位数由方框内的垂直线表示；线段分别从第一和第三个四分位数延伸到最高和最低观测值，不超过1.5×IQR。来源：Nat Metab, 2020, 2:233-242. https://doi.org/10.1038/s42255-020-0178-9.

对于包含大量人源 DNA 的样本，扩增过程可能会非特异性地扩增大量人源序列（如肾结石样本）[33]，除了最优化 PCR 条件，这样的样本还需要根据 DNA 片段长度进行纯化，或在控制污染的前提下，进行目标区域测序。

对于来自同一人的多份样本，包括采集自不同部位的样本，也能够识别出个人特有的模式，从而有助于排除采集到的微生物是随机污染，或是来自于试剂的系统性污染 [15,34,35]。当研究没有怀孕也没有炎症的女性生殖道样本时，笔者合作的临床专家吴瑞芳教授确保在皮肤上进行亚厘米的切割后，她的团队从道格拉斯窝中的腹腔液（盆腔积液）开始，将腹腔镜转移到输卵管（如输卵管有梗阻的患者），然后是子宫内膜（子宫肌瘤、子宫内膜异位症或子宫腺肌病），根据生物量由低到高，逐个进行采样，阴道和宫颈样本在首次就诊当天采集。在操作时有保护性套管，到达采样部位时才伸出采样。尽管乳酸杆菌占优势，但是阴道和宫颈样本与几天后在手术室为同一位志愿者收集的上生殖道样本的微生物分布有很好的一致性。腹腔液 pH 相比更加接近中性，也可能是低生物量但多样性更高微生物的来源（第 2 章）[34.35]。

目前，胎盘中是否存在微生物，仍然存在争议（图 1.8）。根据对保守的 16S RNA 区域进行原位杂交的结果，Kjersti M. Aagaard 博士报告说，小团细菌主要位于绒毛薄壁组织或合胞体滋胚层，而在绒毛膜和母体绒毛间隙中较少见 [36]。图 3.4 展示了核糖体 RNA（rRNA）基因簇（rDNA）的示意图。最近的一项研究从胎盘绒毛末端采集标本，要求能使用 16S rRNA 基因扩增子测序（V1 ～ V2 区，图 3.4）和宏基因组短序列测序都能检测到细菌的存在，该研究否认了在同一产妇的阴道样本中也包含的细菌 [25]。这样各种删结果之后，只有新生儿病原性无乳链球菌（B 型链球菌），被认定为胎盘微生物 [14]。

来自血液的污染一直是一个难以回避的问题（图 3.3）。在胎盘研究中，由于测序量极低［平均 2650 万条读段（reads），其中超过 99% 是人源的，如果 99.9% 都来自人源，那么每个样本平均只剩下 2.65 万条读段，相当于每个细菌的

基因组覆盖度不足 1%］，导致从 KEGG 第二层看到的功能分布，存在样本间的波动[16]。对于宫颈口样本，我们发现优势菌种与人源序列百分比呈现相关性[9,14]。然而，Aagaard 博士的首篇研究中确定的一些分类单元，可能仍然真正地存在于胎盘中。这项宏基因组学研究中，胎盘微生物组显示出一种丰度较高的大肠埃希菌存在，该菌已知在新生儿粪便中出现[37-41]，同时还包含一些可能来自肠道或口腔的细菌，如拟杆菌属、副血链球菌、普雷沃黑色素细菌，以及一些可能来自生殖道的细菌，如痤疮丙酸杆菌、惰性乳杆菌、卷曲乳杆菌等，这些细菌都在包含众多控制样本的实验组中出现。值得注意的是，痤疮丙酸杆菌和副血链球菌也是婴儿口腔和肠道微生物组的成员[16,38,40,42]。如果对奈瑟菌的物种分配是正确的，尽管测序覆盖率低[16]（更多关于分类学的内容见第 5 章），它将为幼儿鼻咽部携带的细菌提供一个潜在的来源。16S rRNA 扩增子测序[36]检测到了更多的分类群，这样的测序方式不会包含人类序列。

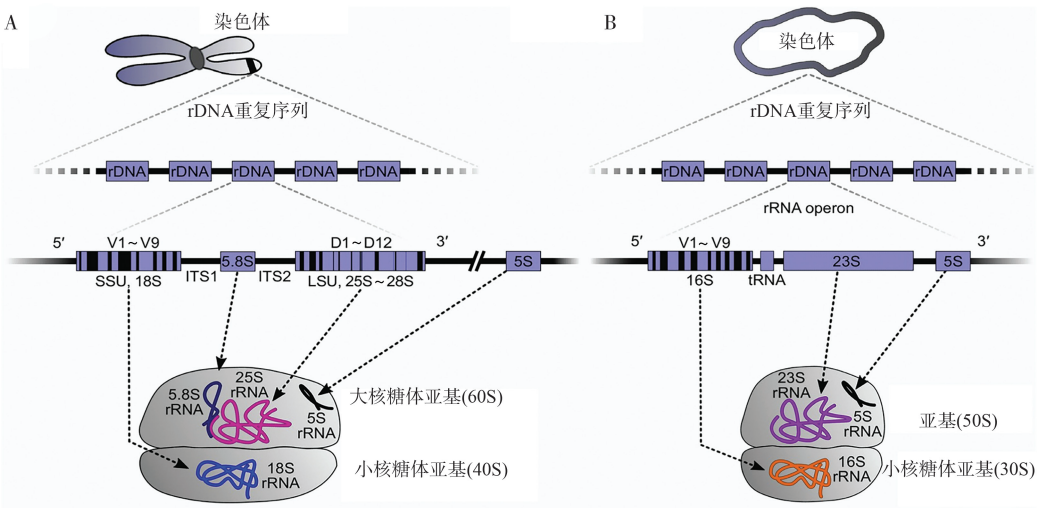

图 3.4　核糖体 RNA（rRNA）基因簇（或 rDNA）的示意图

真核生物和原核生物RNA位点的可变区通常用于表征微生物的分类群，并通过扩增子测序和分析来找出它们之间的系统发育关系。在大多数真菌中，rRNA基因簇包括小核糖体亚基（SSU，18S）和大核糖体亚基（LSU，25～28S），其中内转录间隔区（ITS1和ITS2）位于5.8S的侧翼。在细菌中，rRNA操纵子包括SSU（16S）、LSU（23S）和5S位点。黑色的垂直线以串行顺序显示了SSU（V1～V9）和LSU（D1～D12）中的可变区域，最适合通过微生物群落剖面进行生物多样性评估。来源：Trends Microbiol, 2021, 29:19-27. https://doi.org/10.1016/j.tim.2020.05.019.

根据对 T2D 患者脂肪组织细菌及其异常菌群的有趣研究，血浆和阴性对照组细菌多样性并不低于脂肪组织样本，但 16S rRNA 基因总拷贝数高于阴性对照组（图 3.3）[43,44]。每毫克组织[44]中含有的细菌数量，看起来比每微克（μm）单位 DNA[43] 的细菌数量更多，因为后者包括人类基因组（3.2 GB 的 2 个拷贝）。白色脂肪细胞的直径从小于 30 μm 到大于 300 μm 不等[45]。回到第一章关于人体细胞与细菌

比例的问题，我们可以尝试一些估算。如果这些都是直径为 20 μm 的微小脂肪细胞，那么根据图 3.3 中 16S RNA 基因拷贝数 [44]，单位组织重量或体积中，细菌和人体细胞的比率大约为 1：10。然而，如果脂肪细胞直径为中位数，约为 100 μm，细菌细胞与人类细胞的比例就会反过来，成为 10：1 左右。

当开始研究大脑和肺部（图 3.5、图 3.6）或其他组织时，我们是否按照相同的标准，收集了所有相关的样本？好消息是对于支气管肺泡灌洗液，在有保护性插管时，从口腔和从鼻腔采集的样本之间，没有明显的差异 [46]。但是，相关的样本应该在个体间进行两两比较，而不是显示在一个粗糙的 PCA（主成分分析）中。这样的样本中会有多少微生物细胞？对于病毒和真菌，我们是否需要其他信息来更好地了解其当地栖息地和形态学特征？

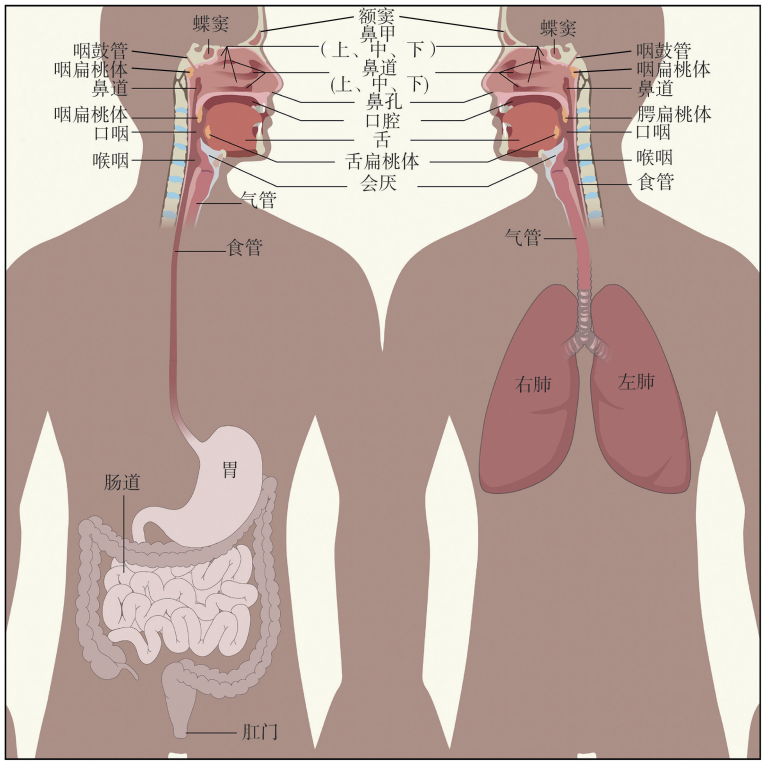

图 3.5　微生物在人体内的分布情况

微生物可能从鼻子、口腔或咽喉扩散到远端身体部位的示意图，以及不同部位之间的相互联系。例如，咽扁桃体（又称腺样体），是口腔/鼻咽部淋巴组织的一个主要部位，可能成为中耳感染的感染菌的储备来源，是因为它们可以通过咽鼓管传播。来源：Cell Host Microbe, 2017, 21:421-432. https://doi.org/10.1016/j.chom.2017.03.011.

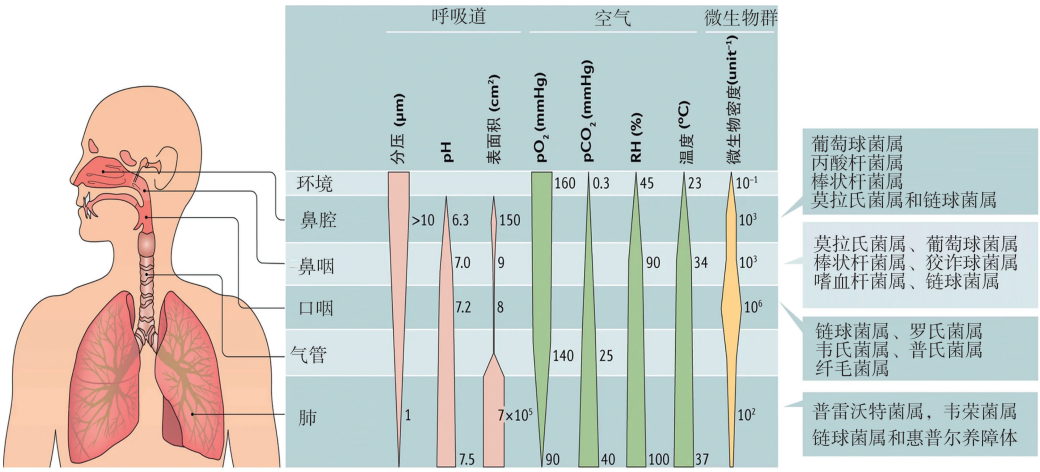

图 3.6　呼吸道的生理和微生物梯度

呼吸道从鼻腔、鼻咽、口咽、气管到肺部，具有不同的生理和微生物的梯度特征。pH随呼吸道逐渐升高，而相对湿度和温度的升高主要发生在鼻腔。此外，氧气和二氧化碳的分压呈相反的梯度，这取决于环境空气条件和肺表面的气体交换。呼吸道还受到来自环境的颗粒的影响，其中包括细菌和病毒。颗粒的沉积位置与其直径有关，大于10 μm的颗粒主要沉积在呼吸道上部，小于1 μm的颗粒可以进入肺部。这些颗粒通常含有细菌和病毒，其直径一般大于0.4 μm。呼吸道的生理参数决定了微生物在不同生态位的选择性生长的条件，从而形成了呼吸道不同位置的微生物群落。测量细菌密度的单位根据生态位不同而有所差异，可用以下方式表示：环境中的细菌密度可以用每立方厘米空气中的细菌数来表示，而鼻腔、鼻咽、口咽和肺部的细菌密度则可以用每个鼻拭子、每毫升口腔清洗液或每毫升支气管肺泡灌洗液中的细菌数来估计。来源：Nat Rev Microbiol, 2017, 15:259-70. https://doi.org/10.1038/nrmicro.2017.14.

3.3　采样后，防止微生物增殖的试剂

在宏基因组研究的早期，粪便样本中兼性厌氧性大肠埃希菌的相对丰度有时超过30%，研究人员怀疑是由粪便样本长时间暴露在室温下造成的。但是这一现象也可能反映了一些疾病状态下的肠道菌群的真实情况，如大肠癌、克罗恩病、IgA 缺乏、2 型糖尿病等[7,47-50]。

为了避免这类情况，研究人员不再需要让志愿者们在家里的冰箱里暂存粪便，或者每天在诊所里放干冰。冷冻过程也会影响宏基因组样品的组成，例如，由于水结晶过程中 pH 和其他浓度的变化，宏基因组样品中的组分可能会影响冷冻效率。目前，有一些商业试剂可使微生物组样品在室温下保存 2 ~ 4 周（如 DNA Genotek Inc., Mawi DNA Technologies, MGI Tech 提供的产品）。这些试剂的保存时间通常远远超过快递公司的运送时间。过滤纸也可用于粪便和宫颈样本，取样后风干，然后密封，但在滤纸风干、裁剪等过程中如何最大限度地减少污染，并没有形成共识。我们必须确保滤纸上的 DNA 数量足够进行高通量宏基因组测序，而不仅仅是使用 16S rRNA

基因扩增子测序。

对于宏转录组的研究，目前还没有发表足够的文献，这不仅需要高质量的保存 RNA，还要求后续（无论多么不完全）去除核糖体 RNA[51,52]。宏蛋白质组学也在兴起，我们目前只尝试了新鲜或冷冻的样品。

稳定剂只抑制细菌生长和降解，不杀灭细菌。因此，仍有一些微生物可以在培养皿上生长。另外，使用质谱法分析相同样本的代谢组，通常需要将（尼龙）拭子浸泡在与存储微生物组不同的试剂中（如用于皮肤拭子代谢组检测的乙醇与水 50∶50 混合液 [53]），但是也有一些商业产品试图兼顾这两种用途。

思考题 3.1

（1）对于一个你感兴趣的微生物样本，你知道它在身体的哪个部位，以及那里大概有多少微生物细胞和物种吗？当患上某种疾病时，你认为这些数字将如何改变？

（2）如果你想构建一个宏基因组文库，你需要准备多少 DNA（例如 0.5 μg）？你会使用拭子还是塑料采样器来采集样本？你能否立刻处理或冷冻样本（之后用于其他组学分析）？如果你使用一个商业试剂，在室温下保存样本数周，并通过商业快递运输，你认为这会不会对微生物群落中的某些菌群产生较大的影响？

（3）你是否有一个标准的混合后的真实样本，用来与真实的、冷冻保存的和保存液保存的样本之间的微生物群落进行对比？

3.4　对于宏基因组样本的DNA提取方法

宏基因组样本的 DNA 提取比哺乳动物细胞或单一微生物种类的 DNA 提取更具挑战性，因为它涉及破坏多种不同类型的细胞壁（图 3.7）[54-56]，即使样本中没有过多的植物纤维或原生生物（图 3.1）。为了提高 DNA 提取的效率和质量，可以采用物理和化学的方法，同时使用对照样品进行平行测序。还可以掺入与样本中序列无关的质控品用于定量。如果需要进行纳米孔测序，要求片段长度超过 20kb，那么就需要采用更加温和的提取方法。

珠磨法是一种常用的手段，它使用坚硬的四方锆多晶体珠子来破碎细胞壁，比玻璃珠更有效。珠子的大小也要根据样品中的细菌和真菌的特点进行选择。较小的珠子可以在单位时间内产生更多的碰撞，但可能无法完全破坏一些真菌的细胞壁。

与扩增子测序不同，高通量宏基因组测序不涉及 PCR 步骤，因此对 DNA 的纯度要求不那么严格。例如，从粪便样本中提取的 DNA 可能仍然会有点发黄，但这不会影响测序的结果。

此外，还可以使用自动化平台（如 96 孔板）来提高样品处理的质量和一致性，

减少随机污染的风险，并节省人力和时间。这种自动化还可以更好地保护工作人员免受任何临床样本污染。

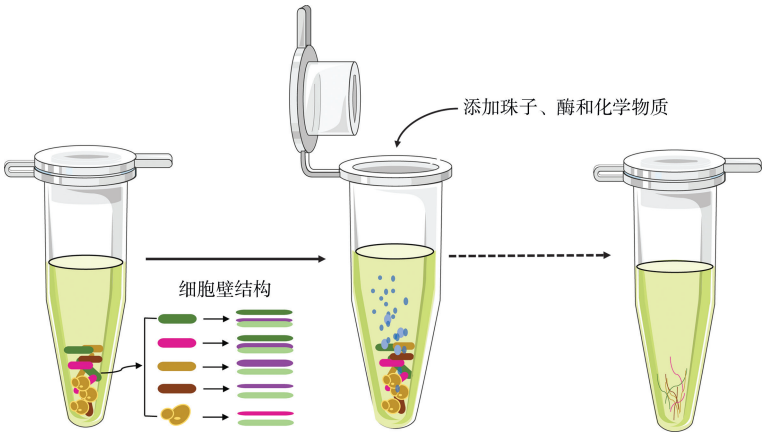

图 3.7　复杂微生物群落的 DNA 提取

粗线条代表革兰氏阴性菌、革兰氏阳性菌和真菌的不同细胞壁结构。

对于构建测序文库，提取的 DNA 片段要经过超声打断。使用诸如 Tn5 酶进行酶纯化，通量更高，也更适合自动化。

3.5　测序量

理论上，只要一个测序读取能够唯一地比对到一个分类单元，就可以检测出特定微生物的存在（详见第 5 章关于分类的内容）的存在。宏基因组鸟枪法测序不需要 PCR 扩增步骤，因此其错误率可以忽略不计。对于一个包含 1 亿个读数的宏基因组样本（如读长 PE100，即 $100 \times 10^6 \times 100 = 10$ GB 数据），直接检测到的一个分类群或基因的最低可能相对丰度是 10^{-8}。对于含 10^{11} CFU 微生物细胞的样品，检出限应该达到单细胞，可以用连续稀释的对照 [22,60] 进行验证。对于低丰度类群，测序量仍然是一个限制因素。在这种情况下，可以考虑在同一个体其他部位样本或其他个体的样本中寻找目标微生物，或者首先尝试对样本进行培养。

DNA 作为一种长链大分子，遵循长链一端固定时或弯或直形态分布的物理规律，桥式 PCR 测序平台倾向于对高 GC 区域（如双歧杆菌基因组的 GC 含量约为 60%）进行过度测序，并且可能需要对过度测序导致的丰度偏差进行数值校正 [57,61,62]。

对于细菌的 16S RNA 扩增子测序，以及真菌的 ITS（内转录间隔区）基因扩增子测序（图 3.4），即使在 DNA 浓度低于紫外线检测（例如，在一些尿样中）的下限时，仍可以检测到微生物。不同的高变区（如 V4 ~ V5，V1 ~ V2）具有不同的分类学分辨率。全长扩增子测序，即包含了完整 16S 和 18S ITS 序列的测序，是一种可靠的