血液对比唾液：分析样本类型对人类全基因组测序的变异调用置信度的影响

Mike Tayeb[1]、Ana Mijalkovic Lazic[2]、Milena Kovacevic[2]、Milos Popovic[2]、Sebastian Wernicke[2]、Christina Dillane[1]
Aaron Del Duca[1]和Rafal M. Iwasiow[1]

1 DNA Genotek Inc, Ottawa, Ontario
2 Seven Bridges Genomics Inc., Cambridge, Massachusetts

前言

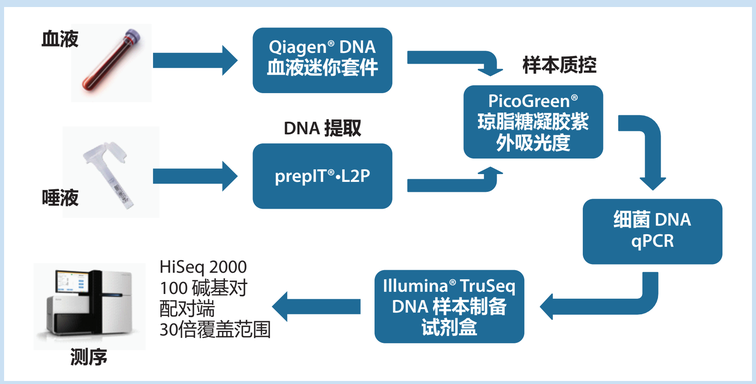
使用Oragene®自行采集套件采集的唾液不仅是血液的非侵入性替代品，而且是大量高质量基因组DNA的来源。Oragene提高供者人数和依从性，使大型人口研究成为可能，其实用性已在一千多篇经同行评审的文章中得到充分证明。但是，在现有文献中，有关全基因组测序中唾液DNA性能的数据很少。

在这项研究中，我们进行了系统性多样本分析，分析了样本类型（血液对比唾液）对变异调用置信度的影响以及唾液中细菌DNA对序列比对的影响。

材料与方法

样本采集：使用K-EDTA管和Oragene自行采集套件分别从两个家庭的每名成员身上采集血液和唾液样本。选择这些特定研究参加者的原因是唾液样本中细菌DNA的含量（由16S qPCR确定）从低于平均值到显著高于平均值。从家庭1和家庭2分别获取四对和三对血液/唾液样本。

样本制备和测序：使用标准样本制备方案提取和定量DNA，并准备TruSeq文库用于Illumina HiSeq 2000测序。制备了来自家族2的样本并测序了两份重复样品以提供技术性重复样品。对制备的所有20个文库进行测序，目标覆盖范围为30×。

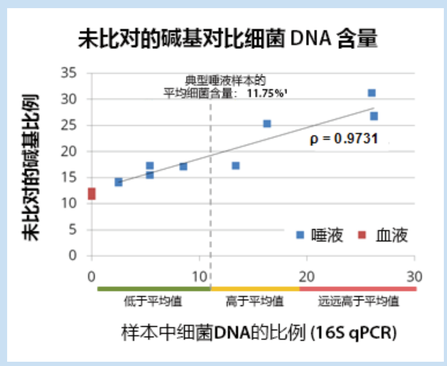


数据分析：使用符合Broad Institute最佳实践建议的BWA + GATK管道，从Seven Bridges平台上读取的序列中调用变异进行生物信息学分析。读取的序列与hg19/B37参考序列比对，所有被调用的变异均使用根据Broad Institute建议设置的硬过滤器进行过滤。

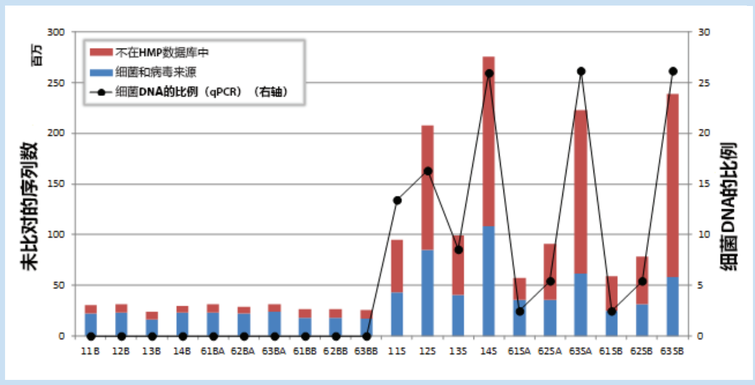
为了确定血液和唾液样本中未比对的序列是否属于细菌来源，使用BWA MEM 0.7.4将它们与人类微生物组计划（HMP）[1]数据库中包含的序列进行比对。

结果

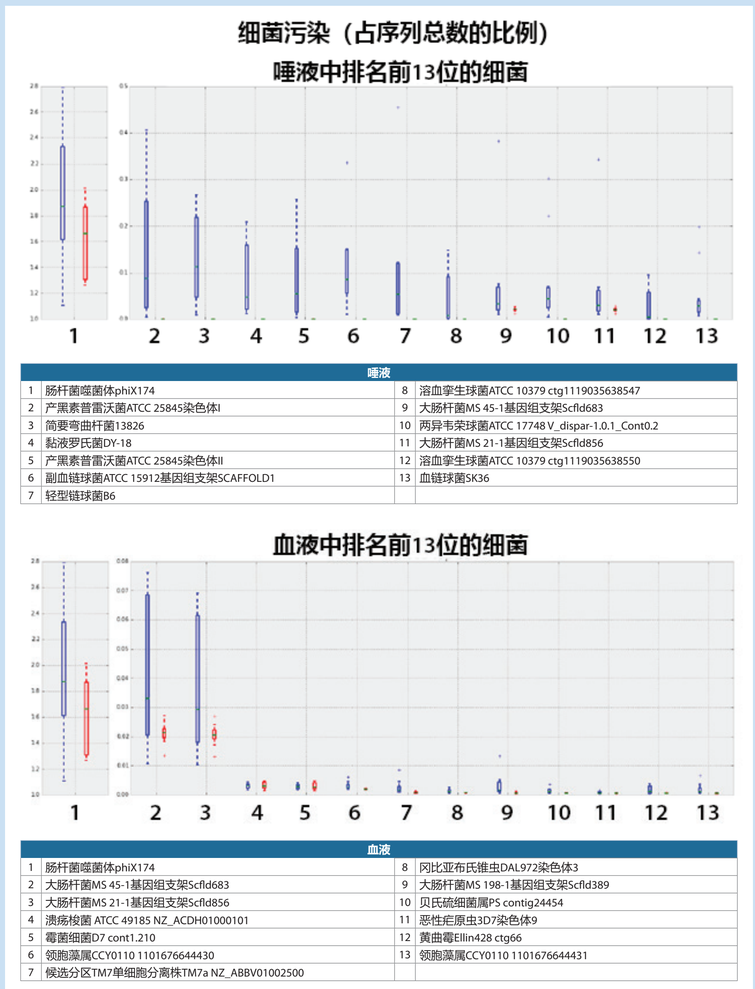
样本中细菌DNA含量与对应于hg19参考序列的碱基（读取的序列）数密切相关，Pearson相关系数为0.9731。这表明样本中细菌DNA含量对测序覆盖范围具有线性影响。



将未对应于hg19参考序列的序列与HMP数据库比对。唾液中平均37%的未对应序列是细菌或病毒来源，而血液中此比例更高，平均为72%。组装未与hg19或HMP对应的序列后，显示出重叠片段，类似于先前在肠道和土壤样本中鉴定的细菌与一些未知生物。存在来自此类生物的序列表明HMP数据库不完整。



为了量化每个样本中不同细菌的数量，将对应于每个细菌基因组的序列数量表达为占样本中序列总数的比例。下图显示了在唾液和血液中发现的前13种病毒和细菌的序列数量。

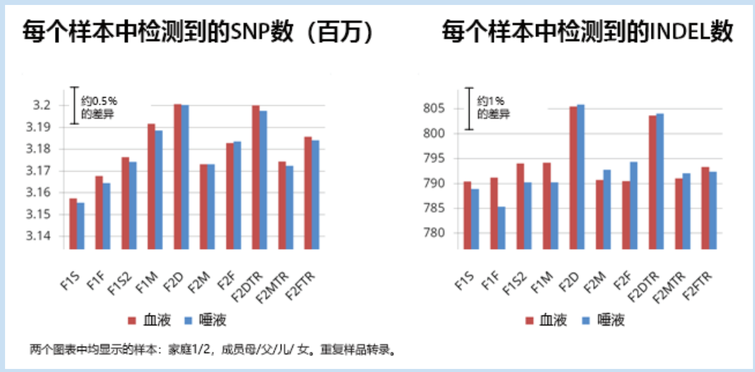


唾液			
1	肠杆菌噬菌体phiX174	8	溶血挛生球菌ATCC 10379 ctg1119035638547
2	产黑素普雷沃菌ATCC 25845染色体I	9	大肠杆菌MS 45-1基因组支架Scfld683
3	简要弯曲杆菌13826	10	两异韦荣球菌ATCC 17748 V_dispar-1.0.1_Cont0.2
4	黏液罗氏菌DY-18	11	大肠杆菌MS 21-1基因组支架Scfld856
5	产黑素普雷沃菌ATCC 25845染色体II	12	溶血挛生球菌ATCC 10379 ctg1119035638550
6	副血链球菌ATCC 15912基因组支架SCAFFOLD1	13	血链球菌SK36
7	轻型链球菌B6		

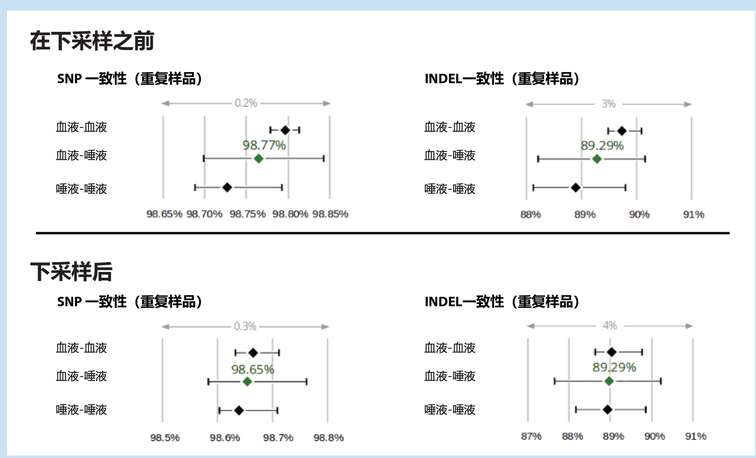
血液			
1	肠杆菌噬菌体phiX174	8	冈比亚布氏锥虫DAL972染色体3
2	大肠杆菌MS 45-1基因组支架Scfld683	9	大肠杆菌MS 198-1基因组支架Scfld389
3	大肠杆菌MS 21-1基因组支架Scfld856	10	贝氏硫细菌属PS contig24454
4	溃疡梭菌 ATCC 49185 NZ_ACDH01000101	11	恶性疟原虫3D7染色体9
5	霉菌细菌D7 cont1.210	12	黄曲霉Ellin428 ctg66
6	领胞藻属CCY0110 1101676644430	13	领胞藻属CCY0110 1101676644431
7	候选分区TM7单细胞分离株TM7a NZ_ABBV01002500		

序列的最显著贡献因素（血液和唾液中分别为2.0%和2.8%）是肠杆菌噬菌体PhiX174。在准备测序期间，将该病毒作为Illumina制备方案的一部分添加到每个样本中，以改善校准和质量控制。[2]存在的其他细菌/病毒序列的数量要低得多（唾液<0.5%，血液<0.08%），唾液中发现的大多数物种都是已知的口腔常驻生物体。血液中存在的细菌序列（如大肠杆菌序列）可能是由于样本或文库制备过程中发生了污染，唾液样本中也存在类似的污染。

从血液和唾液样本中调用的[变单核苷酸多态性（SNP）和插入/缺失INDEL）总数没有观察到显著差异。SNP和INDEL计数的平均差异分别为0.06%和0.30%。



血液重复样品、唾液重复样品和血液-唾液对之间的SNP和INDEL的一致性通常非常高，但是可以观察到血液和唾液之间存在细微的系统性差异。为了确定一致性差异是否是由于覆盖范围差异引起的，血液样本被下采样到与唾液样本相同的覆盖范围。考虑到覆盖范围差异后，重复样品的平均SNP和INDEL一致性分别在0.05%和0.25%之间。



为了检查是否存在富含细菌序列的人类基因组区域，还将对应于人类参考序列的序列与HMP参考序列比对。所有未对应于hg19和HMP的序列均被丢弃，并以100碱基对窗口计算每个碱基的移动平均覆盖范围。如果观察到20×的平均覆盖范围，则该区域被分类为富集区域。对这些区域检查了以下项目，以确定潜在的细菌污染：

- 异常高的不匹配率（序列中不匹配数/区域中总碱基数）
- 仅在唾液样本中检测到富含HMP的区域
- 高比对比例，而图谱质量极差
- 血液和唾液之间的一致性异常低

对上述一个或多个类别的区域进行手动检查，显示这些区域都没有细菌序列污染的迹象。虽然这并不是没有细菌序列污染的确凿证据，但仍然说明细菌序列没有累积到足以影响整体突变调用质量的程度。

结论

唾液样本中细菌DNA的量和未对应于人类参考序列的序列数密切相关。但是，细菌DNA导致的覆盖范围损失相对较小，样本中每5%细菌DNA的覆盖范围下降率约3%。

血液中大多数（72%）未比对的序列与HMP数据库比对，表明这些序列的来源确实是细菌或病毒。在唾液中，该指标较低（37%），但是其他许多未比对的序列显示与HMP中未发现的其他细菌/病毒物种相似，表明口腔中存在其他可能来自环境的物种。

尽管由于唾液样本中存在细菌DNA导致覆盖范围降低，但调用的SNP和INDEL数量没有显著差异。当血液数据被下采样到与唾液相同的覆盖范围时，基本上消除了重复样品和唾液/血液对之间的一致性差异，表明覆盖范围差异是目前已知的导致样本类型之间一致性差异的最重要原因。

最后，仔细查看了基因组中富含HMP的区域，发现细菌序列很可能不会累积到足以影响突变调用的程度。

参考文献

1. NIH Human Microbiome Project. http://hmpdacc.org/HMREFG
2. Using a PhiX Control for HiSeq Sequencing Runs. Illumina Inc. March 2013. http://res.illumina.com/documents/products/technotes/technote_phixcontrolv3.pdf

DNA genotek

SevenBridges genomics