DOI: 10.13345/j.cjb.190281

Dec. 25, 2019, 35(12): 2284−2294
©2019 Chin J Biotech, All rights reserved

·生物工程与大健康·

汪小我 清华大学自动化系长聘副教授，博士生导师。主要研究方向为模式识别、生物信息学、合成生物学。在 *Proc Natl Acad Sci USA*、*Bioinformatics* 等学术期刊发表论文 40 余篇，被 SCI 他引 3 000 余次。曾获全国优秀博士学位论文奖、国家自然科学基金优秀青年基金、教育部新世纪优秀人才、中国自动化学会青年科学家奖等。目前担任中国生物工程学会青年工作委员会主任、中国人工智能学会生物信息学与人工生命专委会副主任、中国自动化学会青工委常委、中国计算机学会生物信息学专委会常委等。

血浆游离 DNA 全基因组甲基化测序的实用稳定性评估

方欢，钟碧溪，魏磊，张祥林，张威，汪小我

清华大学 自动化系 合成与系统生物学研究中心 生物信息学教育部重点实验室 北京信息科学与技术国家研究中心生物信息学研究部，北京 100084

方欢, 钟碧溪, 魏磊, 等. 血浆游离 DNA 全基因组甲基化测序的实用稳定性评估. 生物工程学报, 2019, 35(12): 2284–2294.
Fang H, Zhong BX, Wei L, et al. Practical stability of whole-genome bisulfite sequencing using plasma cell-free DNA. Chin J Biotech, 2019, 35(12): 2284–2294.

摘 要: 随着液体活检技术的发展，血浆游离 DNA 成为当前的研究热点之一。血浆游离 DNA 的全基因组甲基化测序被认为在癌症检测等医学应用拥有巨大潜力，但目前尚缺乏针对该实验流程的实用稳定性评估。文中利用两名志愿者在不同时间采样的血浆游离 DNA，在不同实验平台分别进行 DNA 甲基化的重亚硫酸盐转化前建库 (Pre-BS)、转化后建库 (Post-BS) 和常规 DNA 建库，获取多因素影响下的测序数据样本。在此基础上，建立了一套血浆游离 DNA 测序数据分析的质量控制参考流程，综合评估了血液采集提取、游离 DNA 建库测序过程的实用稳定性，为血浆游离 DNA 全基因组甲基化测序应用于临床液体活检提供实用性的基础参考。

关键词: 血浆游离 DNA，全基因组，DNA 甲基化，片段化模式，微量 DNA 建库

Received: June 26, 2019; **Accepted:** August 30, 2019
Supported by: National Natural Science Foundation of China (No. 61721003).
Corresponding author: Xiaowo Wang. Tel: +86-10-62794294-808; Fax: +86-10-62783552; E-mail: xwwang@tsinghua.edu.cn
国家自然科学基金 (No. 61721003) 资助。

Practical stability of whole-genome bisulfite sequencing using plasma cell-free DNA

Huan Fang, Bixi Zhong, Lei Wei, Xianglin Zhang, Wei Zhang, and Xiaowo Wang

Bioinformatics Division, BNRIST, Ministry of Education Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, *China*

Abstract: With the development of liquid biopsy technology, plasma cell-free DNA (cfDNA) becomes one of the research hotspots. Whole-genome bisulfite sequencing of plasma cell-free DNA has shown great potential medical applications such as cancer detection. However, the practical stability evaluation is still lacking. In this study, plasma cell-free DNA samples from two volunteers at different time were collected and prepared for sequencing in multiple laboratories. The library preparation strategies include pre-bisulfite, post-bisulfite and regular whole-genome sequencing. We established a set of quality control references for plasma cell-free DNA sequencing data and evaluated practical stability of blood collection, DNA extraction, and library preparation and sequencing depth. This work provided a technical practice guide for the application of plasma cfDNA methylation sequencing for clinical applications.

Keywords: plasma cell-free DNA, whole genome, DNA methylation, fragmentation pattern, low input library

细胞游离 DNA (Cell-free DNA，cfDNA) 广泛存在于人体的血液、唾液、尿液、粪便、脑脊液及其他体液中[1]，是无创液体活检技术中的重要标记物之一[2-3]。血浆游离 DNA 由全身各组织细胞的基因组经过细胞凋亡、细胞坏死、主动分泌等断裂释放到血液中[4]，携带了其来源细胞的基因组变异、DNA 甲基化、核小体排布等多方面信息[5]，可应用于孕妇的无创产前检测、器官移植受体的术后排异评估和癌症的溯源检测等领域[6-8]，具有极高的研究价值和应用潜力。

血浆游离 DNA 的检测方式多种多样，按照检测位点数量来分，包括基于 PCR 的单位点检测[9-10]、基于杂交捕获的基因芯片和靶向测序[11-13]、基于高通量测序的全基因组检测[6,14]等；按照检测的信息来分，包括检测突变的有无[15]、染色体拷贝数的多少[16]、DNA 甲基化的程度[17]、病原体的感染[18]等。随着二代测序技术的发展，全基因组测序、多维度信息整合是液体活检未来的发展趋势[2]。血浆游离 DNA 的全基因组甲基化测序数据除了用以获取 DNA 甲基化程度外，还蕴含着部分点突变、拷贝数变异、片段化模式的信息，具有极大的应用潜力。然而，目前尚缺乏对血浆游离 DNA 全基因组甲基化数据中非甲基化信息的详细论证与研究。

目前最常见的血浆游离 DNA 的全基因组甲基化检测是基于全基因组重亚硫酸盐测序 (Whole genome bisulfite sequencing，WGBS)。其基本原理为：在重亚硫酸盐的作用下，DNA 序列中非甲基化的胞嘧啶 C 会转化为尿嘧啶 U，经过 PCR 扩增后变成胸腺嘧啶 T，而甲基化的 C 则不变，将测序片段与人的基因组比对后，即可定量 C 位点的甲基化状态。这一过程主要分为建库测序实验和数据分析两部分，其中，甲基化建库方式分为转化前建库 (Pre-BS) 和转化后建库 (Post-BS) 两种[19]。在重亚硫酸盐转化过程中 DNA 有可能被打断。Pre-BS 建库在转化之前连接测序接头，部分 DNA 被打断丢失，但测序获得的序列均为真实的原始片段；Post-BS 建库在转化之后连接测序接头，避免了 DNA 因断裂失去接头而损失，但人工引入了被打断的片段。因此，两种建库方式在检出甲基化和片段化模式中各有利弊，Pre-BS 保留了片段长度信息，Post-BS 所需 DNA 的起始量更低。目前，

测序公司对于 Pre-BS 建库的送样要求为 1 μg，对于 Post-BS 建库的送样要求则至少为 20 ng。而在研究型实验室中，在使用更高成本的试剂的前提下，Pre-BS 的起始量可低至 10 ng，Post-BS 的起始量则仅需 0.5 ng。按照血浆游离 DNA 提取浓度为 20 ng/mL 计算，仅需抽取 5 mL 血液，获取 2–3 mL 血浆即可满足公司的 Post-BS 建库和实验室的 Pre-BS 建库需求。

本研究从血浆游离 DNA 全基因组甲基化测序的实用性和稳定性角度出发，探究血浆游离 DNA 的采集提取过程的稳定性、不同甲基化建库方式的异同点与质量控制评价、甲基化测序中的片段化模式信息，为血浆游离 DNA 全基因组甲基化测序应用于癌症检测、无创产前检测等液体活检领域提供实用性的基础参考。

1 材料与方法

1.1 血浆游离 DNA 测序数据的获取

1.1.1 血液采集及血浆分离

为了比较不同人、不同采血时间、是否为冻存血样等因素对血浆游离 DNA 采集的影响，研究采集了两名志愿者 (P1 为男性，P2 为女性) 在两个时间点 (第一次采血时间记为 w0，6 周后第二次采血时间记为 w6) 的血样。血液采集方式使用标准采血流程，使用 10 mL EDTA 抗凝管低温运输，并在采血后的 2 h 内分离血浆。血浆分离步骤为：将混有 EDTA 的血液置于冷冻高速离心机中，4 ℃、1 600×*g* 离心 10 min；在超净台中将上层血浆转移至离心管中，再次 4 ℃、16 000×*g* 离心 10 min；将上清分装至 1.5 mL EP 管中，用于后续提取血浆游离 DNA，或−80 ℃冷冻保存。

1.1.2 血浆游离 DNA 的提取

不同于组织或细胞的基因组 DNA，血浆游离 DNA 在血液中的含量极低，不能使用基因组 DNA 的试剂盒提取血浆游离 DNA。通过文献调研与比较，我们选择了 QIAamp Circulating Nucleic Acid 试剂盒提取血浆游离 DNA。同时，将相同的血浆样品通过干冰运输分别送至市面上主流的两家测序服务提供商 (公司 A、公司 B)，进行血浆游离 DNA 的提取服务，每家公司重复一次。通过 Qubit 荧光剂测量血浆游离 DNA 的浓度，并计算每毫升血浆中的得率；通过安捷伦 2100 生物分析仪测量血浆游离 DNA 的片段大小，合格的样品主峰应该在 170 bp 左右，且无大片段的基因组污染；通过重复试验间的浓度比较，确定提取的稳定性。

1.1.3 血浆游离 DNA 的建库和测序

对于血浆游离 DNA，市场上主流测序公司仅提供微量全基因组甲基化建库服务 (Post-BS 建库方式)，我们在实验室进行了全基因组甲基化的微量 Pre-BS 建库，并将两种建库方式的结果进行比较。同时，为比较 A 和 B 两家公司提取血浆游离 DNA 的得率和稳定性，选择公司 A 对两名志愿者的 3 次血样进行提取建库和测序，公司 B 对一次血样进行提取建库和测序，每个血样设置一次重复。公司 A 与 B 一共构建完成 8 个甲基化文库。此外，为了探究不同甲基化建库方式对 DNA 甲基化、片段长度的影响，我们对同样的血浆样品进行了非甲基化的常规全基因组建库测序作为参照。本研究使用的测序平台包括 Illumina 的 NovaSeq 6000 和 Hiseq X Ten，测序读长均为双端 150 bp。本研究中关于血浆游离 DNA 样本的血液来源志愿者、采血时间、建库方式、测序平台、建库实验室等详细信息见表 1。

1.2 血浆游离 DNA 数据的质量控制

针对血浆游离 DNA 测序数据的特点，本研究开发了一套血浆游离 DNA 预处理与质量控制的流程，包括检验测序质量、去除接头序列、比对基因组、去除 PCR 重复、计算转化效率、片段长度分布、覆盖度和深度等。

表 1 研究涉及血浆游离 DNA 的样本信息

Table 1 Information of plasma cfDNA samples in this study

Sample name	Blood source	Blood drawing time	Library preparation strategy	Sequencing strategy	Library preparation laboratory
SP1	P1	w0	Post-BS	NovaSeq 6000, PE150	Company A
SP2	P1	w0	Post-BS	NovaSeq 6000, PE150	Company A
SP3	P2	w6	Post-BS	NovaSeq 6000, PE150	Company A
SP4	P2	w6	Post-BS	NovaSeq 6000, PE150	Company A
SP5	P1	w6	Post-BS	NovaSeq 6000, PE150	Company A
SP6	P1	w6	Post-BS	NovaSeq 6000, PE150	Company A
SP7	P1	w6	WGS	NovaSeq 6000, PE150	Company A
SP8	P1	w6	Post-BS	Hiseq X Ten, PE150	Company B
SP9	P1	w6	Post-BS	Hiseq X Ten, PE150	Company B
SP10	P1	w6	Pre-BS	NovaSeq 6000, PE150	Our lab
SP11	P1	w6	Pre-BS	NovaSeq 6000, PE150	Our lab

1.2.1 cfDNA 全基因组甲基化数据的预处理

对于 SP7 之外的全基因组甲基化数据，进行如下预处理步骤：首先，使用 FastQC 软件对数据的测序质量、碱基分布、序列重复次数、接头污染情况等进行初步统计。由于基因组大部分 C 位点无甲基化修饰，测序为 T 碱基，因此在甲基化文库序列中，T 的比例最高，C 的比例最低，造成碱基不均衡现象。然后，由于血浆游离 DNA 的片段长度较短，当片段长度短于测序读长时，双端测序读段的 3′端包含部分测序接头序列。我们通过 Cutadapt 软件截去接头序列。下一步，使用 BS-Seeker2 软件[20]将甲基化测序数据比对到人的 hg19 基因组上，匹配模式为局部比对。最后，使用 Picard 软件去除比对到同一位置的重复片段，计算全基因组每一 CpG 位点上比对的片段数目和甲基化片段数目。

1.2.2 cfDNA 全基因组测序数据的预处理

由于本研究涉及的血浆游离 DNA 全基因组甲基化测序受到多因素影响，而采集的样本有限，因此，设置 cfDNA 全基因组测序数据 (Whole genome sequencing, WGS) 作为基础对照，并搜集公共数据集中健康人的血浆游离 DNA 全基因组数据，弥补样本数量的不足，增强结论的可信度。本研究搜集的公共数据集包括 GEO 数据库中的 GSE71378 数据集 (样本编号 BH01，美国多个健康人的血浆游离 DNA 混合测序数据)[21]以及 EGA 数据库中的 EGAS00001001024 数据集 (样本编号 C309–314，中国香港健康人血浆游离 DNA 测序数据)[22]。与甲基化测序数据相比，血浆游离 DNA 的全基因组数据处理流程类似，使用 esATAC 包[23]去除接头序列和比对基因组，使用 Picard 去除 PCR 重复扩增的片段。

1.2.3 cfDNA 文库的质量控制标准

在测序数据的预处理过程中，分别统计测序片段数、成功比对片段数、去重后片段数，计算比对率、PCR 重复率和测序有效率，其中 PCR 重复率越低越好，比对率和测序有效率越高越好。对有效片段统计其长度分布，观测主峰位置是否符合血浆游离 DNA 的长度特点。再基于片段的实际长度，用片段覆盖到的基因组区域除以基因组总长度计算覆盖度，使用片段总碱基数除以覆盖基因组范围得到平均测序深度。我们以上述参数为指标，比较不同建库方式、不同实验室建库获得的血浆游离 DNA 甲基化测序数据的质量。

PCR 重复率与测序深度息息相关，对同一文库来说，测序较浅时，我们发现重复片段的概率较

低，PCR 重复率低；随着测序加深，PCR 重复率逐渐升高。对原始测序数据进行降采样，以 5 M 测序片段为步长设计采样点，每个采样点重复两次，分别计算重复率，取两次重复的平均值观测 PCR 重复率与测序深度的关系。此外，自建库的 SP10 和 SP11 测序过程分为两步，首先测序 5G 原始数据，根据上述标准判断文库质量，然后根据文库复杂度进行加测，由此得到的两次独立上机的数据可用于探索同一文库在两次测序中的结果异同。后续分析比较使用两次测序合并的数据。

1.3 cfDNA 甲基化数据的片段化模式评估

借鉴文献中的短片段比例作为片段化模式特征[14]，分别计算 SP1–11 和美国人血浆游离 DNA 混合测序数据 BH01、中国香港人血浆游离 DNA 测序数据 C309–314 的全基因组片段化模式图谱。具体步骤如下：首先，将基因组划分为不重叠的 5 M 宽的窗口，计算每个窗口内长度为 100–150 bp 的短片段数量和长度为 150–220 bp 的长片段数量，短片段占比为短片段数量除以长短片段总数。为了避免性别对片段化模式特征的影响，将 X 和 Y 性染色体排除在外。为了使不同样本间可比，对短片段占比进行标准化，通过减去均值使片段化模式特征分布于 0 附近，再用于后续聚类分析。

为了定量刻画不同血浆游离 DNA 文库的片段化模式差异，以全基因组的短片段占比为特征向量，计算两两样本间的皮尔森相关系数，再采用非加权配对算术平均法进行聚类，得到血浆游离 DNA 甲基化数据的片段化模式关系。

2 结果与分析

2.1 血浆游离 DNA 的提取稳定性

为了探究人群差异、采血时间、血浆冻存时间、起始血浆体积等因素对血浆游离 DNA 提取的影响，本研究采集了两名志愿者在两个时间点的血样，将血浆冻存不同时间后使用同一流程进行多次提取，血浆游离 DNA 的提取浓度见表 2。结果表明，两名志愿者的血浆游离 DNA 浓度存在显著差异；同一个人在不同时间采集的血浆游离 DNA 趋于稳定，采血时间对血浆游离 DNA 浓度的影响不及不同人的影响大；冻存时间与游离 DNA 浓度并不是正比关系，而是呈现游离 DNA 浓度随着冻存时间加长而先增大后减小的趋势。这可能是因为：在新鲜血浆中，细胞的基因组 DNA 较为完整，容易在提取过程中去除；当血浆冻存一段时间后，基因组 DNA 逐渐降解，混入游离 DNA 中一起提取，从而增加了血浆游离 DNA 的浓度；当血浆经过长期冷冻后，DNA 降解消失现象增多，血浆游离 DNA 的浓度逐渐减少。通过使用安捷伦 2100 生物分析仪对冻存血浆中游离 DNA 的长度分布进行检测，证实了冻存血浆中大片段基因组污染增多的观

表 2 多因素对血浆游离 DNA 提取浓度的影响

Table 2 Impact of multiple factors to cfDNA extraction concentration

Blood source	Blood drawing time	Freezing time (d)	Input plasma volume (mL)	Plasma concentration (ng/mL)
P1	w0	0	4	17.13
P1	w0	42	3	38.00
P1	w0	125	11	30.00
P1	w6	0	3	22.93
P1	w6	0	3	24.93
P1	w6	0	3	25.20
P1	**w6**	**0**	**3**	**24.27**
P1	**w6**	**90**	**2**	**22.25**
P2	w0	0	4	10.00
P2	w0	42	3	30.67
P2	w0	125	15	15.51
P2	w6	0	3	14.93
P2	w6	0	3	11.33
P2	w6	0	3	14.47
P2	w6	0	3	12.47
P2	w6	90	3	16.89

The lines in bold corresponding to SP10 and SP11 in table 3.

点。实际上，DNA 浓度不仅与血浆中片段的摩尔量有关，而且与片段长度直接相关。因此，当存在基因组污染时，使用质量浓度定量血浆游离 DNA 是不准确的，即使大片段的摩尔量很少，百倍的片段长度也会使测得的游离 DNA 浓度虚高。可能更精确定量血浆游离 DNA 的方案有：方案 1 是通过片段选择去除长片段后检测游离 DNA 的质量浓度，排除长片段的干扰；方案 2 是使用摩尔浓度替代质量浓度，通过 qPCR 定量血浆中游离 DNA 的摩尔数量，避免片段长度对定量准确性的影响。

为了比较不同实验室提取血浆游离 DNA 的浓度与稳定性差异，将两名志愿者的多个血浆样本分别送至公司 A、B 进行血浆游离 DNA 提取，最终提取浓度与本实验室的两样本提取结果汇总见表 3。从表中可以观察到不同实验室是影响血浆游离 DNA 提取浓度的关键因素，公司 A 提取的 SP1–7 浓度较低，公司 B 提取的 SP8–9 浓度较高，但重复实验间的一致性较差。此外，在公司 A 提取的样本中，来自 P2 个体的样本 (SP3–4) 的 cfDNA 浓度低于来自 P1 个体的样本 (SP1–2、SP5–7)，与前面本实验室提取的结果一致。最后，使用安捷伦 2 100 生物分析仪检测血浆游离 DNA 的片段长度分布，确认 SP1–11 无基因组污染，提取结果达到了建库要求，可以进行后续建库和测序。

表 3 研究中血浆游离 DNA 的提取浓度

Table 3 Extraction concentration of plasma cell-free DNA in this study

Sample name	Input plasma volume (mL)	Total cfDNA mass (ng)	Plasma concentration (ng/mL)
SP1	1	19.12	19.12
SP2	1	19.12	19.12
SP3	2	13.35	6.68
SP4	2	11.44	5.72
SP5	2	16.78	8.39
SP6	2	17.55	8.78
SP7	2	15.64	7.82
SP8	1	65.00	65.00
SP9	1	25.00	25.00
SP10	3	72.80	24.27
SP11	2	44.50	22.25

SP1 and SP2 are different libraries from the same extracted DNA.

2.2 血浆游离 DNA 的建库一致性

2.2.1 血浆游离 DNA 数据的质量控制

本研究涉及 Pre-BS 和 Post-BS 两种甲基化建库方式。10 个血浆游离 DNA 甲基化文库的 Qubit 浓度和 2 100 片段分布都符合上机测序要求，血浆游离 DNA 甲基化测序数据的预处理和质量控制结果见表 4。公司 A 建库的 SP1–6 的局部比对率不超过 40%，远低于公司 B 建库的 SP8–9。当使用全局匹配模式进行比对时，SP1–6 的成功比对率只有 25%左右。在排除了接头序列污染等流程不当的因素后，我们在未成功比对的片段中观测到了部分 PCR 异源双链核酸分子，确认造成比对率低的原因发生于公司 A 的甲基化建库过程中。由于比对率的不足，造成与 SP8–9 相似测序深度的 SP1–4、SP6 的总有效片段数量较少，基因组覆盖度偏低，平均有效深度仅 (1–2)×；对于测序量是 SP8 或 SP9 三倍的 SP5，其获得的有效片段数、平均深度与 SP8、SP9 持平。综合来看，公司 B 的血浆游离 DNA 甲基化建库的数据有效率比公司 A 更高。考虑 Pre-BS 与 Post-BS 两种建库方式的特点，可见 SP10–11 的 PCR 重复率略高于 SP8–9，这可能是因为 Pre-BS 建库打断损失了部分血浆游离 DNA，文库复杂度降低导致 PCR 重复片段增多。

2.2.2 血浆游离 DNA 数据的文库复杂度评价

在表 4 中，可见 PCR 重复率与测序深度高度相关，测序最深的 SP5 和 SP10 是重复率最高的两个样本。为了更精确地定量比较血浆游离 DNA 甲基化文库的复杂度差异，我们对有效片段数较多的 SP5、SP8–11 的原始甲基化测序数据作了降采样分析，结果如图 1 所示。随着测序数据的采样下降，

表 4 血浆游离 DNA 甲基化数据的质量控制

Table 4 Quality control of cfDNA methylation data

Sample name	Raw fragments (M)	Local alignment rate (%)	Duplication rate (%)	Effective fragments (M)	Genome coverage (%)	Sequencing depth (×)
SP1	84.85	38.67	28.84	23.35	49.18	2.33
SP2	87.99	39.17	24.53	26.01	55.54	2.36
SP3	72.30	34.30	25.91	18.37	46.07	1.96
SP4	82.18	33.97	26.48	20.52	53.07	1.98
SP5	230.78	38.46	40.90	52.46	70.01	3.70
SP6	62.16	39.04	28.31	17.39	44.22	1.95
SP8	77.81	77.10	17.05	49.77	85.62	3.63
SP9	77.40	77.72	16.38	50.30	85.24	3.68
SP10	218.31	74.55	42.92	91.27	83.26	6.52
SP11	69.19	76.27	23.05	40.60	72.13	3.30

PCR 重复率下降，符合正相关关系。对比同一降采样深度下不同文库的重复率，发现公司 A 的 SP5 文库重复率最高；公司 B 的 SP8、SP9 文库重复率最低；SP10、SP11 两个 Pre-BS 文库理论上应该重复率较高，但实际上介于公司 A 和公司 B 的 Post-BS 文库之间。重复率的高低直接关系到测序成本和文库的极限测序量，是建库水平的一个直观反映。有趣的一点是，同一文库在两次独立测序中得到的重复率曲线不完全相同。具体来说，图 1 中 test_SP10 与 test_SP11 的重复率高于 SP10 和 SP11，当预期测序深度较低时，所得数据的重复率高于高深度数据的降采样重复率。这可能是因为：血浆游离 DNA 甲基化文库的复杂度有限，不同预期测序量下的上机文库量不同，造成文库的复杂度、重复率差异。

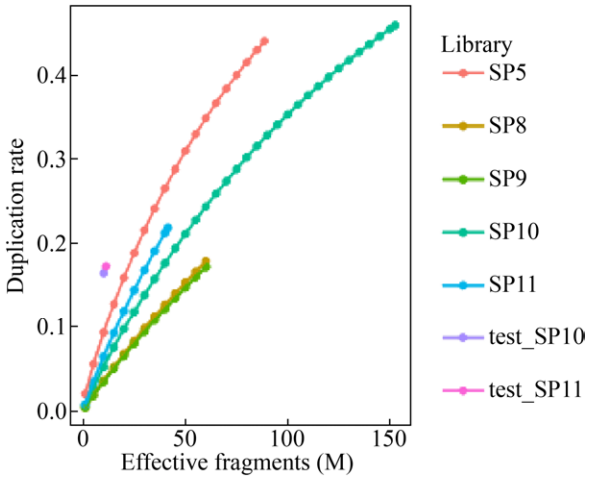

图 1 测序深度与文库重复率的关系

Fig. 1 Relationship between sequencing depth and library duplication rate.

2.2.3 血浆游离 DNA 数据的片段长度分布

除了成功比对率、重复率、覆盖度深度等质量控制指标之外，血浆游离 DNA 甲基化文库的评价标准还包括片段长度分布。根据成功比对的片段绘制血浆游离 DNA 甲基化文库的长度分布曲线，同时对比非甲基化建库的 SP7 和多个健康人混合数据 BH01，所得结果如图 2 所示。图 2 中虚线为 170 bp，BH01 和 SP7–11 文库的主峰均在 170 bp 附近，而公司 A 建库的 SP1–6 主峰明显偏离 170 bp。此外，更为严重的问题是公司 A 实验的 Post-BS 文库和 SP7 都存在长度分布截断的现象。事实上，实验中的片段选择步骤难以做到完全截断，不会是垂直的长度分布曲线。经过反复排查测序数据的预处理流程，采取局部比对基因组的策略，问题依然存在。考虑到片段长度截断发生在 150 bp 以内，有可能是因为测序公司在生成原始数据时，默认过滤了含有部分接头序列的片段。对于基因组 DNA 的常规建库来说，文库大小 (200–400 bp) 远高于测序的

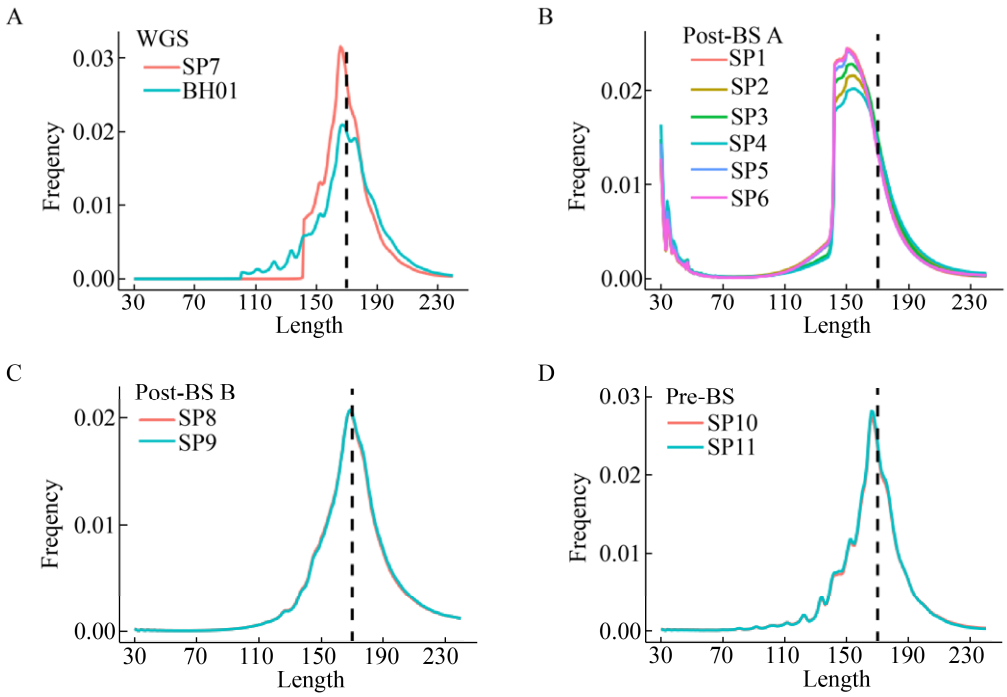

图 2　血浆游离 DNA 文库的长度分布 (A：WGS 文库的长度分布；B：公司 A 构建的 Post-BS 文库的长度分布；C：公司 B 构建的 Post-BS 文库的长度分布；D：Pre-BS 甲基化文库的长度分布)

Fig. 2　Length distribution of plasma cfDNA library. (A) Length distribution of WGS library. (B) Length distribution of Post-BS library accomplished by company A. (C) Length distribution of Post-BS library accomplished by company B. (D) Length distribution of Pre-BS library.

读长 (150 bp)，因此舍弃短片段不会影响测序数据的分析，但是对于血浆游离 DNA 来说，过滤短片段会对后续分析的影响较大。例如，癌症病人血浆游离 DNA 比健康人更短[24]，使用长度为 90–150 bp 的血浆游离 DNA 可以更灵敏地区分癌症病人和健康人[25]。换句话说，血浆游离 DNA 的短片段中，癌症来源片段的比例更高，舍弃短片段将直接影响血浆游离 DNA 的应用。

对比 Pre-BS 和 Post-BS 建库的片段长度区别，可见 SP10 与 SP11 这两个 Pre-BS 文库的片段长度分布呈现 10 bp 左右的周期，与非甲基化建库 (WGS) 的公共数据 BH01 中的周期一致，而两家公司的 Post-BS 数据中，血浆游离 DNA 片段长度分布比较平滑。由于 Pre-BS 与 Post-BS 两种建库方式是分别在不同实验室完成的，所以不能排除实验室差异的因素造成片段长度分布的差异。但至少，这一现象证实了 Pre-BS 可以保留部分片段长度信息，而公司 A、B 的 Post-BS 建库对精细的片段长度信息损失较多。

2.3　血浆游离 DNA 甲基化数据的长度信息

为了进一步阐述血浆游离 DNA 甲基化文库中的片段长度信息，研究参考了文献中的血浆游离 DNA 长度特征[14]，使用全基因组 5 M 窗口内的短片段比例作为特征，绘制甲基化与常规文库的全基因组片段化模式图谱，所得结果见图 3。

由于在考察血浆游离 DNA 的长度特征时，全基因组测序比全基因组甲基化测序更能反映真实情况，因此，本研究将全基因组甲基化数据中的片段长度特征与全基因组测序数据进行比较，以期评估血浆游离 DNA 甲基化数据中的片段长度信息。

此外，为了反映片段长度在人群中的差异，还搜集了两个公共数据集中的健康人血浆游离 DNA 全基因组测序数据 (美国多个健康人混合血浆测序数据 BH01、中国香港健康人血浆测序数据 C309–314)。从图 3 可以看出，SP10–11 中 Pre-BS 甲基化数据得到的片段化模式特征在全基因组的方差较大，信息量更大，其分布模式与 WGS 数据 BH01、SP7 相似，而与 SP1–6 与 SP8–9 中 Post-BS 数据存在显著差异。这一现象再次印证了 Pre-BS 保留片段长度信息的特点。

我们根据血浆游离 DNA 的全基因组片段化模式图谱，计算样本间的相关关系，聚类分析推断样本间相似性，聚类结果如图 4 所示。SP10–11 中 Pre-BS 数据与 SP7、BH01 相关性高；公司 B 建库的 Post-BS 数据和 C309–314 聚在一起；而公司 A 建库的 Post-BS 数据单独聚在一起。从聚类结果可以观察到血浆游离 DNA 的 Pre-BS 数据可以保留 WGS 的片段化模式特征；公司 B 建库 Post-BS 数据也捕获了部分片段化模式特征，对应到图 3 的基因组图谱中，例如 1 号、10 号染色体的图谱模式印证了 Post-BS B 与 C309–314 更相似。总体来说，不管是 Pre-BS 还是 Post-BS 甲基化建库，都含有

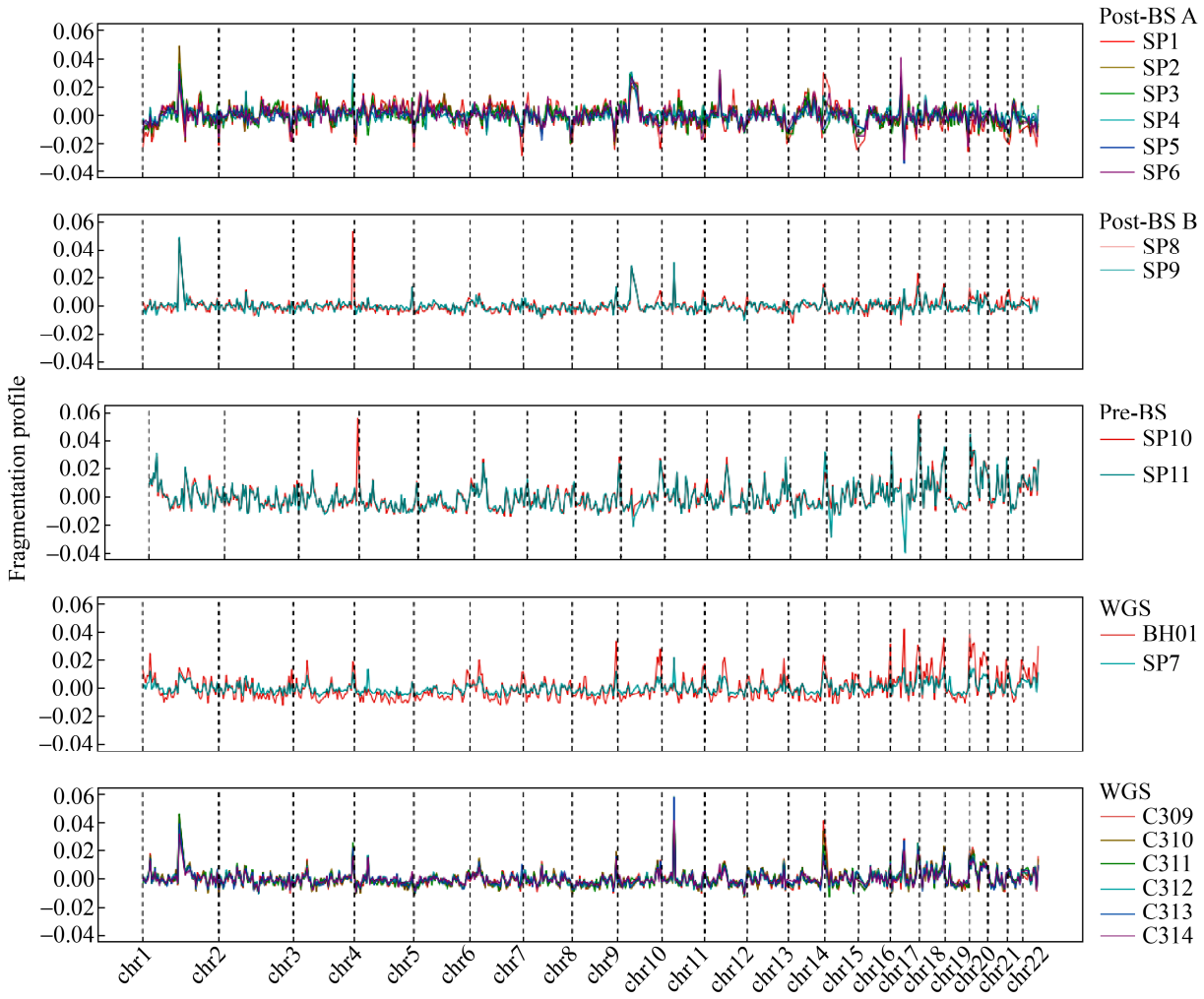

图 3　血浆游离 DNA 全基因组数据中的片段化模式

Fig. 3　Fragmentation pattern in whole-genome plasma cfDNA data.

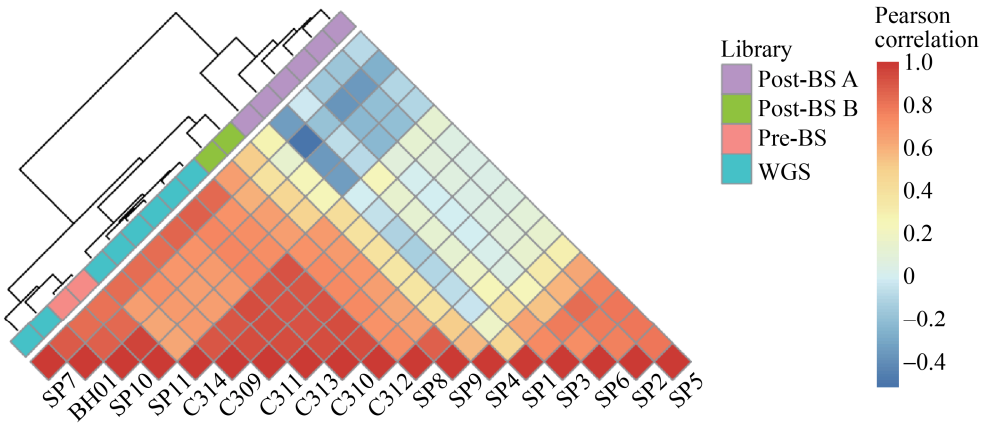

图 4　血浆游离 DNA 文库间的片段化模式相关性

Fig. 4　Correlations of fragmentation pattern among different cfDNA libraries.

血浆游离 DNA 的片段化模式信息，不同甲基化建库方式的数据差异与不同实验室的 WGS 数据差异相当，说明甲基化与非甲基化建库方式对片段化模式的影响甚至不及人群因素的影响大。但是，对于公司 A 产出的 Post-BS 数据，其片段化模式与 WGS 数据相关性差，这一点有可能与公司 A 的短片段截断有关。

3　讨论

血浆游离 DNA 在液体活检领域有着广泛的应用，在癌症检测领域，血浆游离 DNA 中部分位点的检测试剂盒已经得到国家食品药品监督管理总局的批准，应用于人群的癌症筛查中。随着新一代测序技术的不断发展，血浆游离 DNA 的全基因组甲基化测序在早期癌症的检测与肿瘤溯源方面具有良好的应用前景。

本研究通过对不同人、不同采血时间、不同冻存时间、不同建库方式、不同测序平台、不同公司等变量的分析，探究了血浆游离 DNA 的采集、提取、建库、测序、信息分析过程，综合评估实验的稳定性和甲基化数据的实用性。研究结果表明：针对血浆游离 DNA 的采集提取稳定性，两名健康异性志愿者的血浆游离 DNA 浓度在 5–65 ng/mL 范围内，两人的血浆游离 DNA 浓度有显著差异，而不同采血时间的浓度较为稳定；针对血浆游离 DNA 的建库测序稳定性，我们发现不同测序公司间的建库质量存在较大差异，并存在短序列读段的长度截断等干扰因素；针对 Pre-BS 与 Post-BS 甲基化建库的实用性，结果表明抽取 5 mL 血液即可满足 Pre-BS 与 Post-BS 的建库需求，同时 Pre-BS 保留了更精细的片段长度信息；针对血浆游离 DNA 甲基化数据中的片段化模式，我们发现 Pre-BS 甲基化测序数据中能提取到与 WGS 相似的片段化模式特征。这些实用性与稳定性的评估结果将为血浆游离 DNA 全基因组甲基化测序应用于液体活检领域提供有力的参考和支撑。

REFERENCES

[1] Chan AK, Chiu RW, Lo YM, et al. Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. Ann Clin Biochem, 2003, 40(2): 122–130.

[2] Heitzer E, Haque IS, Roberts CES, et al. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet, 2019, 20(2): 71–88.

[3] Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev

Cancer, 2011, 11(6): 426–437.
[4] Lui YY, Chik KW, Chiu RW, et al. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clin Chem, 2002, 48(3): 421–427.
[5] Wan JCM, Massie C, Garcia-Corbacho J, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer, 2017, 17(4): 223–238.
[6] Sun K, Jiang PY, Chan KCA, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci USA, 2015, 112(40): E5503–E5512.
[7] De Rubis G, Rajeev Krishnan S, Bebawy M. Liquid biopsies in cancer diagnosis, monitoring, and prognosis. Trends Pharmacol Sci, 2019, 40(3): 172–186.
[8] Li WY, Li QJ, Kang SL, et al. Cancer Detector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res, 2018, 46(15): e89.
[9] Ooki A, Maleki Z, Tsay JJ, et al. A panel of novel detection and prognostic methylated DNA markers in primary non-small cell lung cancer and serum DNA. Clin Cancer Res, 2017, 23(22): 7141–7152.
[10] Cohen JD, Li L, Wang YX, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 2018, 359(6378): 926–930.
[11] Liu L, Toung JM, Jassowicz AF, et al. Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification. Ann Oncol, 2018, 29(6): 1445–1453.
[12] Jamal-Hanjani M, Wilson GA, Horswell S, et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. Ann Oncol, 2016, 27(5): 862–867.
[13] Guo SC, Diep D, Plongthongkum N, et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet, 2017, 49(4): 635–642.
[14] Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 2019, 570(7761): 385–389.
[15] Qu CF, Wang YT, Wang P, et al. Detection of early-stage hepatocellular carcinoma in asymptomatic HBsAg- seropositive individuals by liquid biopsy. Proc Natl Acad Sci USA, 2019, 116(13): 6308–6312.
[16] Chan KCA, Jiang PY, Chan CWM, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci USA, 2013, 110(47): 18761–18768.
[17] Feng H, Jin P, Wu H. Disease prediction by cell-free DNA methylation. Brief Bioinform, 2019, 20(2): 585–597.
[18] Lam WKJ, Jiang PY, Chan KCA, et al. Sequencing-based counting and size profiling of plasma Epstein-Barr virus DNA enhance population screening of nasopharyngeal carcinoma. Proc Natl Acad Sci USA, 2018, 115(22): E5115–E5124.
[19] Olova N, Krueger F, Andrews S, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol, 2018, 19(1): 33.
[20] Guo WL, Fiziev P, Yan WH, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. Bmc Genomics, 2013, 14: 774.
[21] Snyder MW, Kircher M, Hill AJ, et al. Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. Cell, 2016, 164(1/2): 57–68.
[22] Jiang PY, Chan CWM, Chan KCA, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci USA, 2015, 112(11): E1317–E1325.
[23] Wei Z, Zhang W, Fang H, et al. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. Bioinformatics, 2018, 34(15): 2664–2665.
[24] Underhill HR, Kitzman JO, Hellwig S, et al. Fragment length of circulating tumor DNA. PLoS Genet, 2016, 12(7): e1006162.
[25] Mouliere F, Chandrananda D, Piskorz AM, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med, 2018, 10(466): eaat4921.

(本文责编 陈宏宇)