单管长片段测序(stLFR)技术的系统性评估

华大智造基于DNBSEQ测序平台的stLFR技术赋能长读长测序

本研究提出基于将相同的条形码序列添加到原始长DNA分子的亚片段中共标记亚片段的单管长片段测序(single-tube long fragment read，stLFR)技术,以短读序列获取长片段DNA分子信息，并利用基因组信息已知的人类DNA标准品NA12878系统性的在华大智造DNBSEQ测序平台上验证了stLFR测序技术在长片段DNA测序过程中的性能。

具体研究成果已于2019年发表于*Genome Research*杂志题为“Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly”[1]。

推荐应用：DNA长读长测序

推荐机型：DNBSEQ-T7RS, MGISEQ-2000RS

- **单倍体组装**

  杂合位点phasing rate高达99.7%，单倍体组装区块最大N50值可达10 Mb。

- **小变异检测**

  高SNP和InDel变异检测精准度和敏感度高，该技术类似于单分子测序但是碱基错误率低，检测小变异质量高。

- **结构变异（Structure Variant，SV）检测**

  可有效检出倒位、易位、删除、缺失、插入等大于20kb的结构变异。

- **自定义数据挖掘**

  可以分析常规WGS难以处理的区域例如高同源区域、高重复区域等。

- **起始数据量低**

  起始量低至1 ng即可获得100 ng DNA起始的常规文库相当的基因组覆盖均匀性表现。

背景介绍

目前绝大多数生物体的全基因组序列缺乏同源染色体上连续块状的单碱基到多碱基变异的单倍体信息，并且大多数基因组中存在短读测序无法解析大结构变异以及其他的区域。早期研究对于这些信息不够重视，但是对于全面了解基因组如何帮助表现表型来说这些信息是十分重要的。

因此，为了解决这一问题，本研究提出长DNA分子测序技术单管长片段测序技术(stLFR)，基于在同一来源的长DNA分子亚片段上标记cobarcode方法[2]，在单管中对提取好的长DNA分子在Tn5转座酶的作用下随机插入转座子序列，转座子连接方法有两种分别为双转座子连接法和3′端分支连接法，之后通过接头序列将转座子与带有多拷贝的分子标签磁珠载体结合，之后对捕获出来的长DNA分子进行打断，PCR扩增和环化建库。测序并拼接后的DNA分子的长度范围可达20-300Kb，克服短read测序无法获得的这句话读不通。利用stLFR技术对基因组信息已知的人类DNA标准品NA12878进行文库构建，并利用DNBSEQ测序平台进行高通量测序，结果显示了高质量的变异检测和最大定相区块的N50长度达到34Mb，还揭示了NA12878基因组内的结构变异(Structure Variant，SV)情况。充分的信息表明：stLFR 是一种可以帮助我们对许多未了解的基因组区域进行有效探索和发掘的长片段测序技术。

研究描述

以深圳华大生命科学研究院为主体的某研究团队为解决短读长测序在解读全基因组信息方面存在的缺陷而开发了一种全新的技术——stLFR。该长片段测序技术基于无分隔共标记理念和高通量短读长测序技术开发且其优势明显。该团队在完成技术开发后还利用DNA标准品NA12878对技术的各项指标参数进行了验证[1]。

实验方法

文库制备

首先对人类基因组NA12878样本经过实验手段利用RecoverEase DNA分离试剂盒从细胞系中获取长DNA片段。基于分裂和池连接的策略使用三组双链条形码DNA分子构建barcode磁珠，在带有5′双生物素连接子的链霉亲和素上连接相同的接头序列，采用集成DNA技术构建了3组1536个含有重叠序列区域的barcode寡核苷酸。将寡核苷酸连接到磁珠上并收集使用。带有两个转座子的stLFR测序在Tn5偶联转座子（包含杂交的单链和被酶识别的双链序列）的作用下将两个不同的转座子插入到基因组DNA中，反应结束后与磁珠在杂交缓冲液中结合，用磁铁收集磁珠到管的侧面并洗涤，去除多余的barcode寡核苷酸。带有3′端接头序列的stLFR在杂交插入时使用一个转座子，在捕获和barcode连接完成后将磁珠收集到管侧面，使用外切酶I和外切酶III消化接头序列，之后收集磁珠捕获得到的长DNA分子。带有两个转座子stLFR测序方法中插入长DNA片段的的是两种不同的转座子。然而，这种方法导致每个长DNA分子的覆盖率降低50%，因为将两个不同的转座子彼此相邻插入作为PCR引物。为了实现每个长DNA片段的最高覆盖率，在插入步骤中使用了单个转座子，并通过连接一个额外的3′端引物，称为3′分支连接[3]。基于这两种转座子连接方法收集得到长DNA片段进行PCR扩增建库，最后进行高通量测序(图1)。基于此方法华大智造研发出MGIEasy stLFR文库制备试剂盒套装(16RXN货号:1000005622)进行扩增建库。

测序

基于BGISEQ-500高通量测序仪，对构建好的stLFR文库进行双端100pb(PE100)测序。目前华大智造对测序仪进一步研发升级为通量更大，测序更快速的MGISEQ-2000和DNBSEQ-T7RS测序仪，为长DNA分子长连续信息提取提供可靠准确的测序方法。

数据处理

本研究使用第三方公开工具对原始read数据进行序列比对和变异检测，与GIAB[4]变异集相比较可以确定变异检测的真阳性(TP)、假阳性(FP)和假阴性(FN)率。为了进一步提高 stLFR测序方法的变异信息的假阳性率，我们开发了一个基于XGboost[5]的用于变异过滤的二元分类模型用于变异过滤。获得比对结果bam文件和变异信息数据之后将这两个数据作为输入，利用HapCUT2[6]对单倍体进行组装。通过计算基因组区域之间的cobarcode[7]来检测SV，使用Jaccard系数计算cobarcode之间的比例以识别结构变异。利用Long Ranger对至少读取10次的barcode进行优化到~470万个barcode列表。转换stLFR FASTQ格式作为Supernova输入，生成伪hap组装输出，并将最小长度为10Kb的scaffolds与GRCh38进行比对。

样本采集	文库制备和测序	生信分析	结果分析
人类基因组 NA12878样本	 MGIEasy stLFR文库制备试剂盒  DNBSEQ测序平台	可用 MGI-tech-bioinformatics /stLFR_v1	stLFR方法系统性评估

仅供研究使用，不适用于临床诊断

结果

stLFR测序建库流程

stLFR测序首先是在长DNA片段中以固定的间隔通过Tn5转座酶插入杂交序列。该转座酶包含用于杂交的单链区域和可以被酶识别的双链序列。转座酶在转座完成后仍然与长DNA片段结合保持长DNA片段不受损伤。之后将插入杂交序列的长DNA与包含40万个捕获序列的barcode磁珠结合捕获(每个磁珠拥有唯一共同的barcode),进行滚动运动将长DNA分子包裹在磁珠周围。接下来收集磁珠并通过连接杂交序列和捕获接头之间的缺口将个性化barcode序列转录到每个长DNA分子中。之后去除转座酶，消化Oligo之后经过PCR扩增环化进入下一代测序流程，通过华大智造自主研发专利技术DNB纳米球滚环扩增，包埋到阵列式芯片上，最后通过BGISEQ-500进行高通量PE100测序（图1A）。

测序结束后通过自定义方法提取barcode信息，通过唯一barcode映射read结果显示，大多数具有相同barcode的read数据都聚集再基因组中的一个区域，该区域对应文库制备时期使用的长DNA分子长度(图1B)。

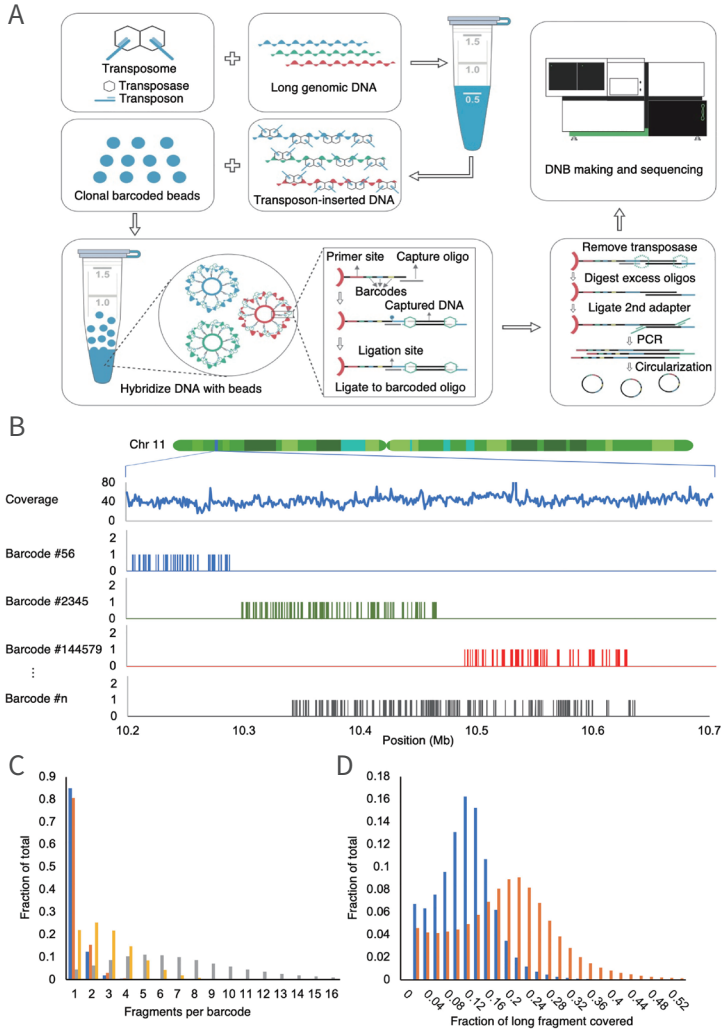

图1. A表示stLFR建库流程；B表示barcode映射的read在11号染色体一小片段上成簇聚集；C表示4个文库里每个barcode映射的片段数量；D表示为stLFR-1文库的覆盖每个原始长DNA片段的非重叠序列读数（蓝色）和捕获的亚片段（橙色）的比例。

stLFR测序 read覆盖度和变异检测

我们使用DNA的1ng（stLFR-1和stLFR-2）和10 ng (stLFR-3和stLFR-4)以及stLFR-1-3采用3′分支连接法，对stLFR-4采用双转座子法的4种起始量和不同转座子连接方法与其他变异检测方法的变异检测结果比较。stLFR-1和stLFR-2总碱基覆盖率分别为336GB和660Gb，stLFR-3和stLFR-4分别达到117和126Gb的更适度水平(表1)。非重复覆盖率从34×到58×不等，每个条形码的长DNA分子数量从1.2到6.8(图1C)。观察到每个长DNA分子的最高平均非重复覆盖率为10.7%-12.1%，每个长DNA分子捕获亚片段的最高平均非重复碱基覆盖率为17.9%-18.4%(图1D)。对测序后数据进行分析发现stLFR在InDel的检测上都比同类型方法的表现要好很多。在SNP检测上，stLFR文库也会略优于其他方法结果。过滤后SNP的与过滤前FP率相似，但是FN率高2到3倍。(表1)。

	stLFR-1	stLFR-2	stLFR-3	stLFR-4	10X Genomics	IlluminaBeads Haplotyping	BGISEQ-500SD	BGISEQ-500 PCR-free SD	BGISEQ-500SD
Library statistics									
Total bases sequence(Gb)	336	660	117	126	128	99	81	129	132
SNP									
TP	3194945	3197686	3193507	3175921	3202498	-	3194780	3201626	3201452
FP	9125	9544	7144	9544	7144	-	5192	6372	4800
FN	15312	12571	16750	12571	16750	-	15477	8630	8805
Precision	0.997	0.997	0.998	0.995	0.971	0.997	0.998	0.998	0.999
Sensitivity	0.995	0.996	0.995	0.989	0.998	0.952	0.995	0.997	0.997
TP(Filter)	3193269	3194955	3192891	3174874	3200472		3194254	3200960	3201273
FP(Filter)	4491	4606	4814	8982	18615		4271	4153	3111
FN(Filter)	16988	15302	17366	35382	9785		16003	9396	8984
Precision (Filter)	0.999	0.999	0.999	0.997	0.994		0.999	0.999	0.999
Sensitivity (Filter)	0.995	o.995	0.995	0.989	0.997		0.995	0.997	0.997
INDEL									
TP	460144	464451	459979	440718	415613	-	463273	465316	467612
FP	32437	30487	17375	22075	235331	-	10541	17136	19514
FN	21120	16816	21288	40547	65656	-	17993	15951	13655
Precision	0.934	0.938	0.964	0.952	0.636	0.932	0.978	0.965	0.96
Sensitivity	0.956	0.965	0.957	0.916	0.864	0.832	0.963	0.967	0.972

表1. stLFR与其他方法变异检测结果比较

stLFR测序定相分析

使用HapCUT2方法进行变异定相，大多数杂合子SNP的99%被定向在区块内，二倍体基因组上的定相信息可以利用带有cobarcode的短读序列去获取，从而可解析基因调控和编码区变异组合。结果显示stLFR-1的read覆盖率高，在40X深度下，stLFR-1文库数据定相区块N50值最高可达34 Mb，能被定相的杂合位点比例高达99.7%（图2）。

结构变异(SV)检测

本研究为了揭示stLFR测序对多种结构变异进行准确检测的能力，根据已知的SV位点对stLFR-1和stLFR-4进行检测，发现即使覆盖率较低仍能检测到删除位点(图3A)。检测到NA12878中8号染色体的150Kb的杂合缺失(图3B和3C)。使用5号染色体和12号染色体之间具有已知易位的患者的细胞系[8](图3D)和已倒位变异的GM20759细胞系[9]样本进行验证检测其他SV的能力(图3E)，显示均能检测到与已知一致的结构变异。

stLFR重头组装

1ng的stLFR-1文库中85%的片段（基因组中小于300kb的区域）由唯一的barcode共标签，这有助于我们对于重头组装的简化和改进。为了检测stLFR进行从头组装的能力，我们使用stLFR-1和stLFR-2文库进行从头组装。将组装的contig与人类参考基因组GRCh38染色体进行比对，结果显示高度一致性（图4）。

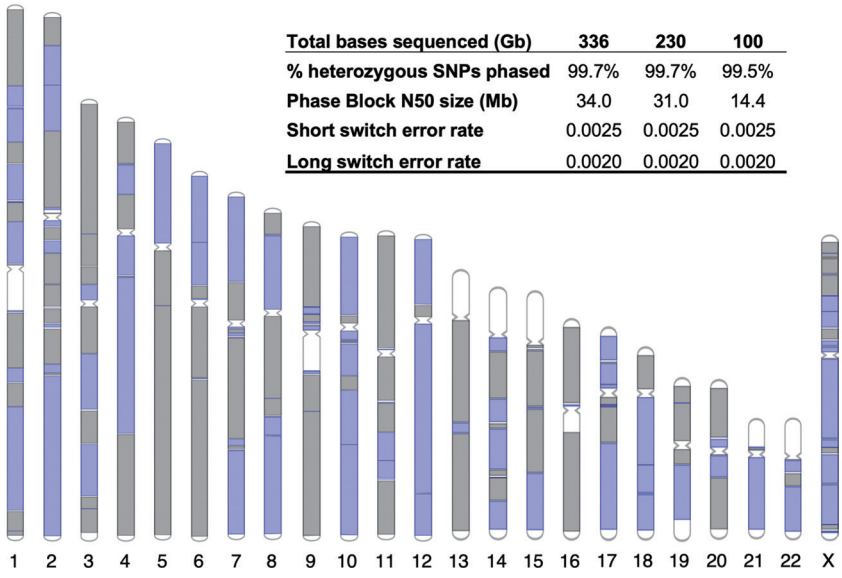

图2. stLFR-1定相数据统计

A

Location	Zygosity	Size (kb)	Found by stLFR
Chr1:189,704,509-189,783,359	Het	78.9	Yes
Chr3:65,189,000-65,213,999	Het	25.0	Yes
Chr3:162,512,134-162,626,335	Hom	114.2	Yes
Chr4:116,167,000-116,176,999	Het	10.0	Yes
Chr4:187,094,000-187,097,999	Het	4.0	Yes
Chr5:104,432,113-104,503,673	Hom	71.6	Yes
Chr6:78,967,194-79,036,419	Het	69.2	Yes
Chr7:110,182,000-110,187,999	Het	6.0	Yes
Chr8:39,232,074-39,387,229	Het	155.2	Yes
Chr16:62,545,000-62,549,999	Het	5.0	Yes

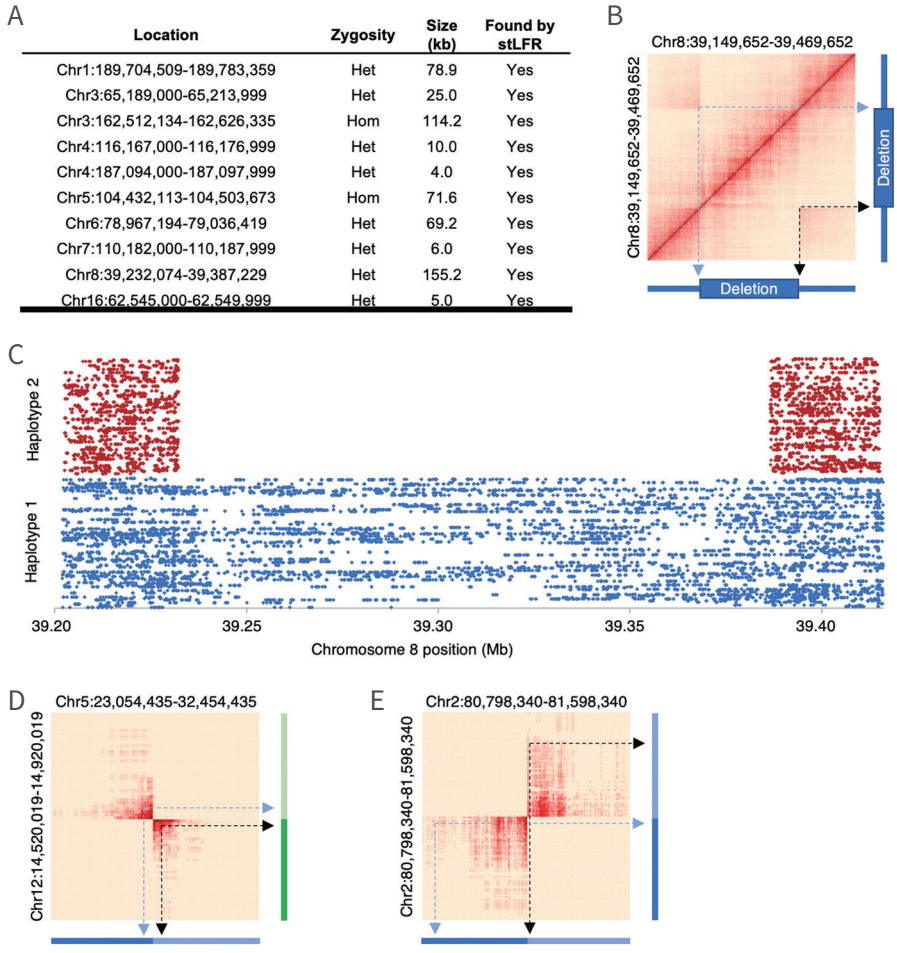

图3. A表示stLFR根据已知位点进行检测验证stLFR检测SV能力；B、C表示8号染色体片段缺失；D表示已知易位患者细胞系中5号染色体上检测倒位和已知一致E表示GM20759检测得到的倒位和已知倒位一致[9]。

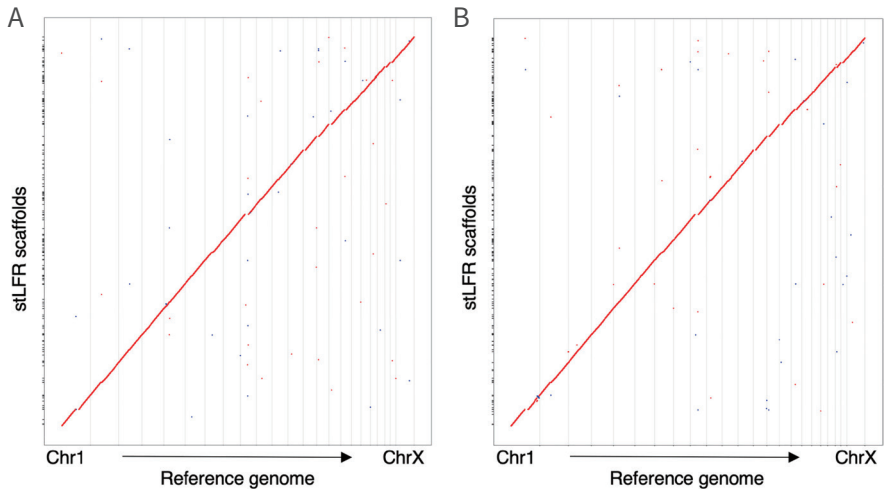

图4. A、B分别表示stLFR-1文库和stLFR-2从头组装的contig与GRCh38比对

总结

本研究利用NA12878细胞系验证了单管长片段测序(stLFR)技术的适用性，这些分析都是使用单个stLFR测序文库进行的，它们的构建并没有显着增加全基因组测序(WGS)文库制备的时间或成本。stLFR技术可实现高质量测序、定相、SV检测、二倍体从头基因组组装和其他长 DNA 测序应用。

基于该技术开发出的MGIEasy stLFR文库制备试剂盒可完美搭配华大智造自主研发的DNBSEQ测序平台开展长度长测序研究。该测序平台基于独有的DNBSEQ™技术，具有高准确性、低重复序列率、低标签跳跃的优势。

基因测序仪MGISEQ-2000

参考文献

1. Wang, O., *et al.*, Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res*, 2019. **29**(5): p. 798-808.
2. Peters, B.A., J. Liu, and R. Drmanac, Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. *Front Genet*, 2014. **5**: p. 466.
3. Wang, L., *et al.*, 3' Branch ligation: a novel method to ligate non-complementary DNA to recessed or internal 3'OH ends in DNA or RNA. *DNA Res*, 2019. **26**(1): p. 45-53.
4. Zook, J.M., *et al.*, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*, 2014. **32**(3): p. 246-51.
5. Chen, T. and C. Guestrin, XGBoost, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785-794.
6. Edge, P., V. Bafna, and V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*, 2017. **27**(5): p. 801-812.
7. Zhang, F., *et al.*, Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol*, 2017. **35**(9): p. 852-857.
8. Dong, Z., *et al.*, Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med*, 2016. **18**(9): p. 940-8.
9. Dong, Z., *et al.*, Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the future of clinical cytogenetics. *Genet Med*, 2018. **20**(7): p. 697-707.

推荐订购信息

产品类型	产品名称	产品货号
仪器	基因测序仪MGISEQ-2000RS	900-000035-00
	基因测序仪DNBSEQ-T7RS	900-000236-00
	MGISP-960RS 自动化样本制备系统	900-000100-00
软件	MegaBOLT生信分析加速器（工作站式服务器）	970-000085-00
	ZTRON Pro 一体机（T500）	900-000444-00
	MGI-tech-bioinformatics/stLFR_v1	https://github.com/MGI-tech-bioinformatics/stLFR_v1
建库试剂	MGIEasy stLFR文库制备试剂盒（16RXN）	940-000193-00
测序试剂	MGISEQ-2000RS 高通量测序试剂套装（stLFR FCL PE100）	1000016984
	DNBSEQ-T7RS 高通量测序试剂套装(stLFR FCL PE100)	1000019251

深圳华大智造科技股份有限公司

深圳市盐田区北山工业区综合楼11栋

4000-688-114

www.mgi-tech.com

MGI-service@mgi-tech.com

股票简称：华大智造

股票代码：688114

版本：2023年6月版

撰稿：黎金兰 黄艳玲

责任编辑：王其伟

审稿：江遥