stLFR 长片段技术解密WGS盲区

——基于DNBSEQ平台的stLFR数据在WGS盲区的覆盖率＞70%

亮点

起始量低

建库起始量低至1.5ng

变异检测全

SNP/Indel/CNV/SV均可检测，SV检测结果准，有效检出倒位、异位等结构变异

盲区覆盖度高

可分析常规WGS无法有效覆盖的高同源/高重复区域

单倍型分型准确率高

可分型，杂合位点分型比例高达99.9%

背景

stLFR(single tube Long Fragment Read)是由华大智造研发的无分隔共标签长片段读取技术（图1），搭配DNBSEQ™系列测序平台，可通过短读长测序获取长片段DNA信息，从而实现一次测序获得高准确度的SNP/InDel/CNV/SV变异检测结果，同时可有效覆盖常规WGS无法有效覆盖的高同源/高重复等区域，为大规模WGS提供新的产品！

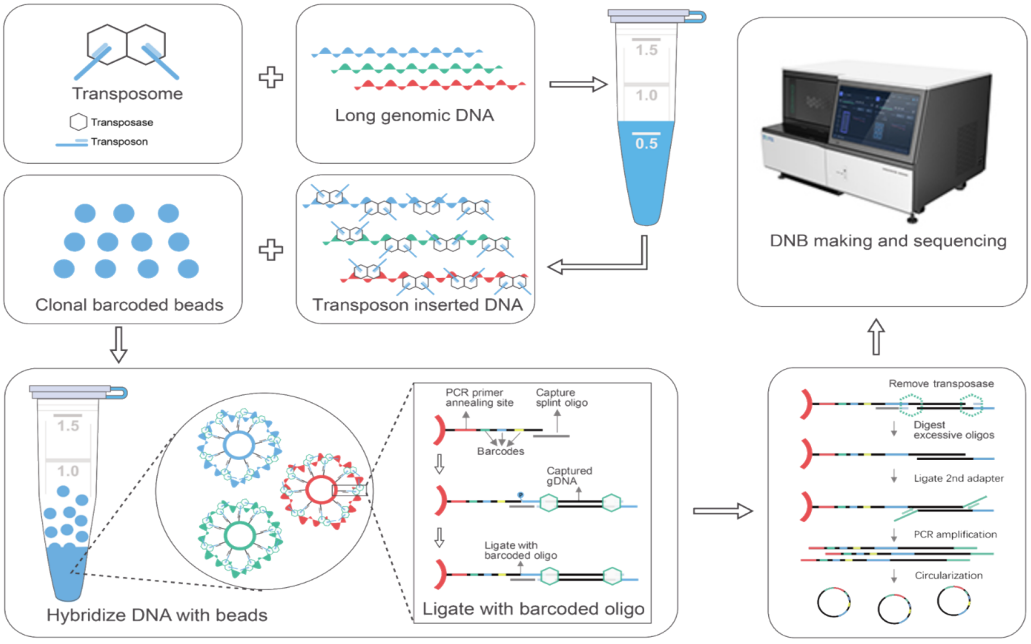

图1　stLFR 文库构建流程示意图

以中国食品药品检定研究院在2020年6月发表的学术文章为例：针对同一个体的样本在DNBSEQ/illumina/PacBio CCS/stLFR等多个测序平台的检测结果的平行比较，结果表明：stLFR数据在变异检测/盲区覆盖度/疾病相关基因的检测/单倍体分型等方面的结果表现明显优于常规短读长测序，结果可达到PB CCS同等效果。

文章详情

样本来源

来自中国北京的汉族男性志愿者的HJ细胞系标准品（XHEC-C-2019-086, HJ）。

测序策略

表1 不同平台的测序读长及测序深度

数据类型	测序平台	读长	测序深度
短读长数据	BGISEQ-500	PE100 bp	86.58 X
	DNBSEQ-G400	PE100 bp	86.58 X
	NextSeq-CN500	PE150 bp	60.07 X
	NextSeq550Dx	PE150 bp	60.07 X
	NovaSeq6000	PE150 bp	60.07 X
长读长数据	stLFR(DNBSEQ-G400)	11.7 kb	83.02 X
	PacBio Sequel II HIFI (CCS)	12.1 kb	24.4 X

注：stLFR 读长11.7Kb是指long fragment 长度，测序读长是PE100+42；
DNBSEQ-G400：即华大智造基因测序仪MGISEQ-2000在部分海外国家和地区的名称。

1不同平台数据分析流程介绍：

1） 短读长数据先后经过SOAPnuke进行低质量过滤、BWA比对，最后通过GATK HaplotypeCaller进行SNV与InDel变异检测（如图2绿色色块和蓝色色块所示）；

2） stLFR数据通过Long Ranger WGS模式进行分析及变异检测，使用HapCUT2进行单倍体定相分析（如图2红色色块所示）；

3） PacBio CCS数据使用pbmm2软件与hs37d5参考基因组进行比对，并通过GATK HaplotypeCaller进行变异检测，使用WhatsHap软件进行单倍体定相分析（如图2黄色色块所示）。

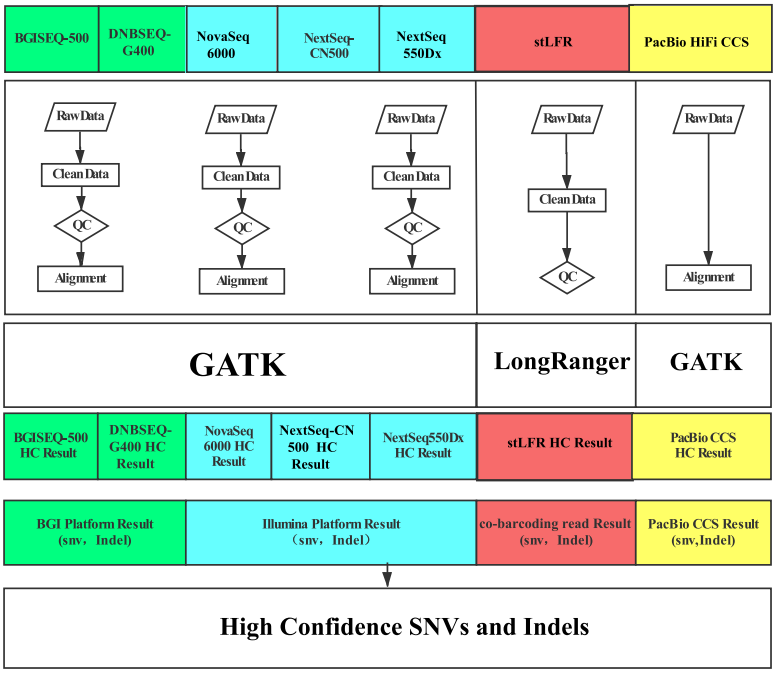

图2 不同类型的数据的信息分析流程图

结果

1. stLFR 数据在SNV和Indel的检测能力优于常规短读长数据

从SNV和InDel的比较结果来看，stLFR 与CCS的检测结果优于短读长平台。从下图可以看出，stLFR WGS 特有的SNV和Indel 数量超过20万个，数量远超常规WGS。

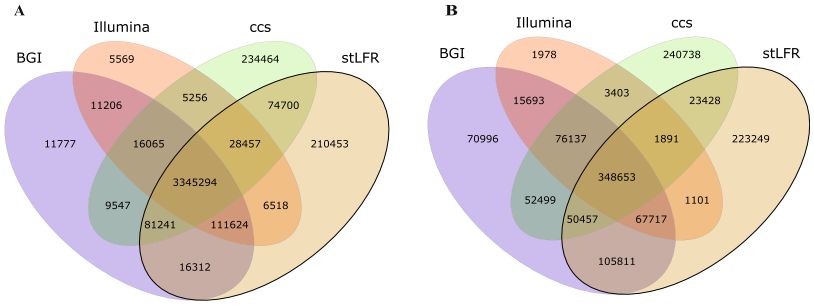

图3 各个平台检测到的SNV和InDel 变异结果展示

注：BGI 即指DNBSEQ平台；

2. stLFR 数据在盲区的覆盖度和常规短读长数据相比提升35倍

在30X测序深度的情况下，各个平台对于基因组中大约44.41Mb的盲区覆盖度的比较结果中，常规WGS数据覆盖度低于2%，无法有效覆盖盲区，但stLFR与CCS覆盖率分别是73.3%及68.53%，stLFR WGS数据表现非常优秀（表2）。

表2 CCS 和stLFR 在盲区的覆盖度

	CCS	stLFR
Length(bp)	30403236	32519955
Percentage(%)	68.53	73.3

备注：盲区是指高度同源区域/高重复区域/高GC区域/以及reference 组装不完整的区域。

3. stLFR数据单倍体分型结果优于CCS数据

常规短读长数据无法进行单倍体定相分析，但PacBio CCS数据和stLFR数据可以，分型结果见表3。PacBio CCS数据和stLFR数据对于杂合位点的分型比例分别为99.63%和99.91%。从分型结果来看，stLFR 数据表现比PacBio CCS 数据表现更胜一筹！

表3 PacBio CCS 和stLFR数据分型结果

	CCS			stLFR		
Chr	Heterozygous	Phased SNV	Phased rate(%)	Heterozygous	Phased SNV	Phased rate(%)
1	169,906	169,174	99.57	172,790	172,682	99.94
2	167,518	166,806	99.57	103,486	103,435	99.95
3	143,618	142,968	99.55	101,126	101,070	99.94
4	151,585	151,033	99.64	102,317	102,274	99.96
5	128,296	127,772	99.59	75,908	75,874	99.96
6	131,798	131,325	99.64	70,091	70,064	99.96
7	123,689	123,253	99.65	68,411	68,379	99.95
8	120,782	120,391	99.68	72,564	72,536	99.96
9	94,946	94,653	99.69	53,789	53,762	99.95
10	99,256	98,894	99.64	60,279	60,254	99.96
11	100,822	100,465	99.65	49,744	49,724	99.96
12	101,519	101,168	99.65	172,818	172,720	99.94
13	75,515	75,282	99.69	43,553	43,531	99.95
14	68,223	67,954	99.61	37,388	37,370	99.95
15	67,759	67,549	99.69	34,105	34,089	99.95
16	69,062	68,823	99.65	145,073	145,017	99.96
17	54,620	54,358	99.52	151,677	151,603	99.95
18	59,025	58,847	99.7	130,925	130,865	99.95
19	48,314	48,195	99.75	135,438	135,376	99.95
20	42,939	42,750	99.56	126,619	126,552	99.95
21	39,076	39,010	99.83	122,931	122,888	99.97
22	31,029	30,979	99.84	111,751	111,697	99.95
X	—	—	—	4,418	3,636	82.3
Y	—	—	—	4,444	4,354	97.97
Genome	2,089,297	2,081,649	99.63	2,151,645	2,149,752	99.91

4. stLFR数据检测疾病相关基因能力优于常规短读长数据

在盲区中的基因及其变异，有很多和人类疾病相关，举例如下：

NBPF4：神经母细胞瘤断点基因家族（NBPF）的一员，该家族由几十个相近重复的基因组成，主要位于人类1号染色体的重复片段中。各平台的检测结果显示，只有stLFR和PB CCS测序完整检出。

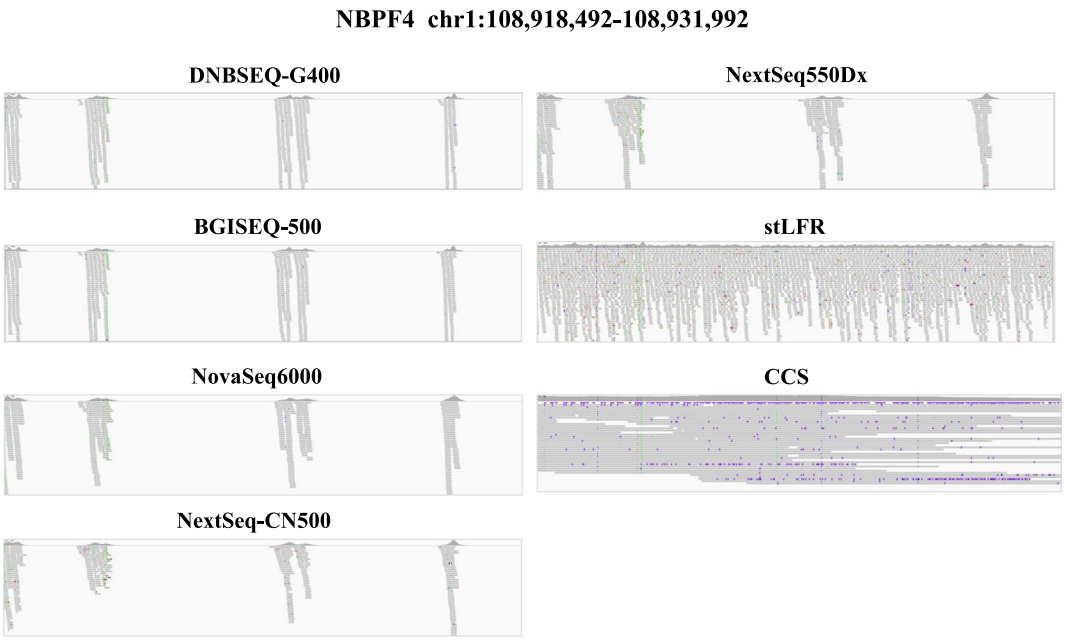

图4 不同平台在NBPF4基因区的覆盖度和测序深度的IGV结果图

NAIP: 是5号染色体长臂1区3亚区的一段500kb的反向重复区域的一部分，包含至少4个重复元件和基因，很容易发生重排和缺失。序列的高复杂度使得这段区域很难被准确检测。这个基因被认为是其临近基因SMN1（与脊髓性肌萎缩症相关）的突变基因，短读长平台检测到的NAIP区域的变异都是一些小的变异，但在stLFR和CCS 平台几乎所有NAIP的变异都检测到。。

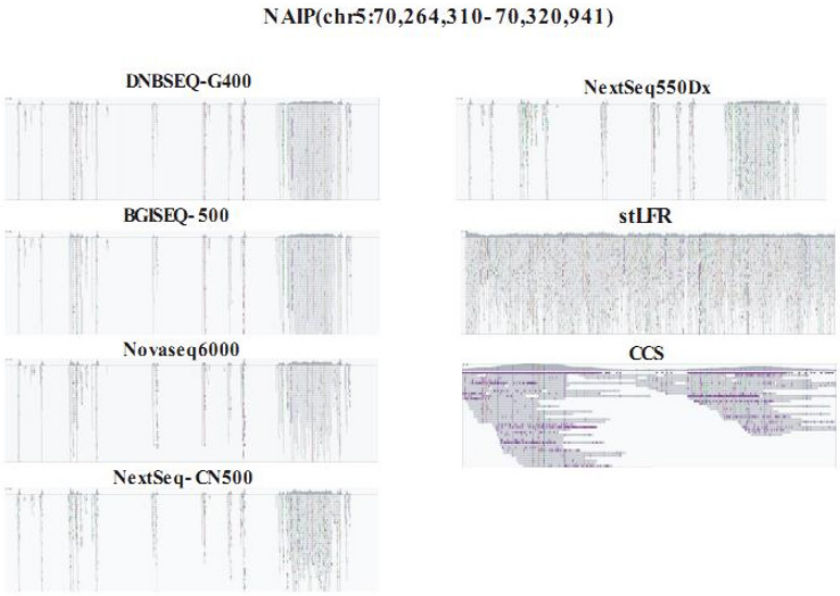

图5 不同平台在NAIP基因区的覆盖度和测序深度的IGV结果图

总结

在本次介绍的案例中，研究团队比较了不同测序平台平行比较的结果，发现stLFR结合DNBSEQ测序在变异检测/盲区覆盖度/疾病相关基因的检测/单倍体分型等方面的结果表现明显优于常规短读长测序，甚至媲美于PB CCS结果。

应用推荐：

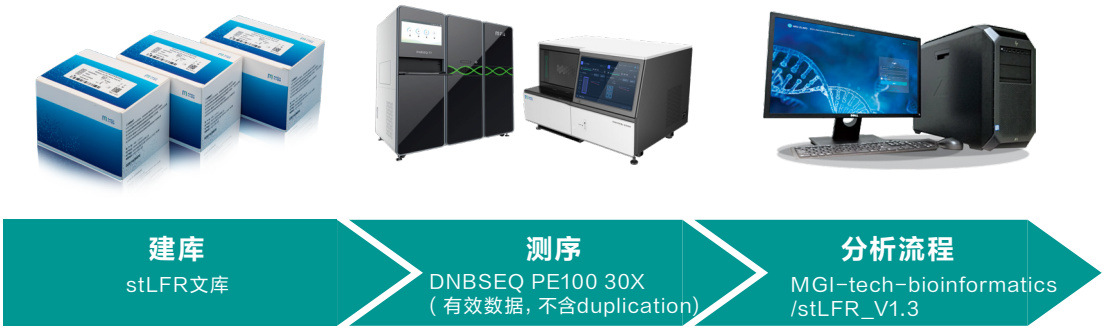

图6 基于stLFR 建库的WGS 产品组合推荐策略

参考文献

1. Wang O, Chin R, Cheng X, 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[J]. Genome research, 2019, 29(5): 798-808.
2. Huang C, Shao L, Qu S, et al. An integrated Asian human SNV and indel benchmark established using multiple sequencing methods[J]. Scientific reports, 2020, 10(1): 1-11.

附录：其它stLFR 相关的已发表文章

发表时间	发表杂志	IF	文献标题	研究对象	应用分类
2019.04	Genome Research	10.989	Efficient and uniqueco-barcoding of second-generation sequencing reads from long DNA moleculesenabling cost effective and accurate sequencing, haplotyping, and de novoassembly[J]	NA12878	方法学
2020.03	Peer J	2.457	IterCluster: a barcode clustering algorithm for long fragment read analysis	–	方法学
2020.06	Scientific Report	4.149	An integrated Asian human SnV and indel benchmark established using multiple sequencing methods	HJ	人重
2020.05	Scientific Report	4.149	The genome of Mekong tiger perch (Datnioides undec-imradiatus) provides insights into the phylogenetic position of Lobotiformes	曼谷拟松鲷（泰北虎鱼）	组装
2020.08	GigaScience	7.551	Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K)	万种鱼类基因组计划	组装
2020.10	Protein Cell	10.164	Genomic and transcriptomic analysis unveils population evolution and development of pesticide resistance in fall armyworm Spodoptera frugiperda	草地贪夜蛾	组装

订购信息

分类	产品名称	规格	货号
建库试剂	MGIEasy stLFR文库制备试剂盒	16 RXN	1000005622
建库试剂	stLFR文库制备试剂盒V2.0	48 RXN	1000021745
测序试剂	MGISEQ-2000RS 高通量测序试剂套装（stLFR FCL PE100）	/	1000011545
测序试剂	DNBSEQ-T7RS 高通量测序试剂套装（stLFR FCL PE100）	/	1000019251
测序平台	基因测序仪MGISEQ-2000RS	台	90000003500

深圳华大智造科技股份有限公司

MGI-service@mgi-tech.com | www.mgi-tech.com | 4000-688-114 | 深圳市盐田区北山工业区综合楼及11栋2楼