stLFR技术准确检测染色体结构变异

——精准确定染色体平衡易位断点位置

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

染色体结构变异（Structural Variants，简称SVs）指的是基因组中大于50 bp的变异，包括缺失、插入、重复、倒位和易位。SV会比单核苷酸多态性（SNP）或小的插入缺失（INDEL）带来更多的基因组序列差异,某些SV还会造成某些特定的疾病。尽管SV非常重要，但对其进行准确的检测仍是很大的挑战。SV变异涉及的范围大，而且还存在非常复杂的变异（倒位、平衡易位等），传统的短读长的测序方法很难检测到相应的变异，更难以精确断点位置。长读长测序技术也限于测序准确性和成本，制约其在SV检测中的应用。为了弥补短读长测序的长片段信息的不足以及平衡测序成本，开发出的共标签长片段读取技术，有助于发现基因组中的SV。

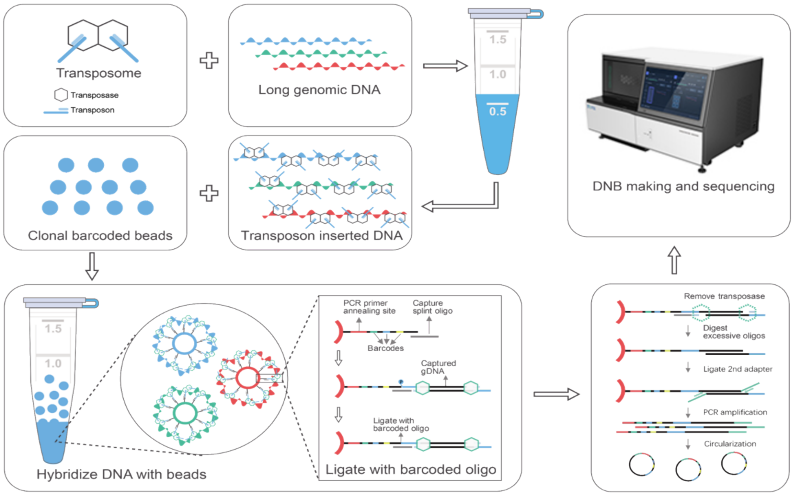


图1 stLFR 文库构建流程示意图

stLFR(single tube Long Fragment Read)是由华大智造研发的无分割共标签长片段读取技术（图1），搭配DNBSEQTM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，以及单倍体分型结果。因为可精确断点位置，不仅可确认断点位置是否有功能基因，而且可用于PGT-SR的检测，解决临床上的难点。基于stLFR的共标签长片段信息，利用特别为stLFR数据特征开发的SV分析软件stLFRsv能识别co-barcoded reads之间的异常缺口，检测出复杂结构变异的断点。

案例介绍

1. 使用HG002样本对stLFR测序数据和stLFRsv软件检测SV的能力进行评估，确定stLFR在检测结构变异具有优势。

2. 使用5例已知核型结果的平衡易位样本评估stLFR检测复杂SV的能力，确定stLFR可以准确检测平衡易位，并且能精确到断点上下游10K以内的位置，sanger验证结果确定了stLFR检测断点位置的准确。

方法

1. 使用stLFR技术对HG002细胞系样本进行建库测序，测序深度为100X。用stLFRsv进行SV的检测，并且将数据向下选取50X和30X的数据进行检测。同时使用其他4种SV分析软件（LongRanger、NAIBR、smoove和GROC-SVs）对stLFR测序数据进行SV检测，stLFR数据分析得到的结果会与GIAB v0.6.2结构变异集（GIAB v0.6.2结构变异集是使用HG002细胞系样本建立的一个结构变异标准集，包括7,172个插入和缺失）进行比较。此外，stLFR数据分析的SV结果会与100X Nanopore长读长数据的SV检测结果进行比较。

2. 筛选并收集已进行核型检测的5例平衡易位血液样本，使用透析法提取长度大于40kb的长片段DNA，使用stLFR技术进行建库，然后使用MGISEQ-2000测序仪进行测序，得到的测序数据使用stLFRsv分析软件进行分析，再将检测的平衡易位的结果与核型结果进行比较之后，确定检测出的平衡易位的断点位置，根据断点位置进行PCR引物的设计，设计的引物的正反向引物分别位于发生易位的2条染色体上，可以扩增出衍生染色体的断点附近的序列。PCR扩增产物进行sanger测序，得到的序列信息与检测到的断点信息进行验证，确定检测到断点位置的附近的序列与sanger测序得到的序列是否一致。（图2）。

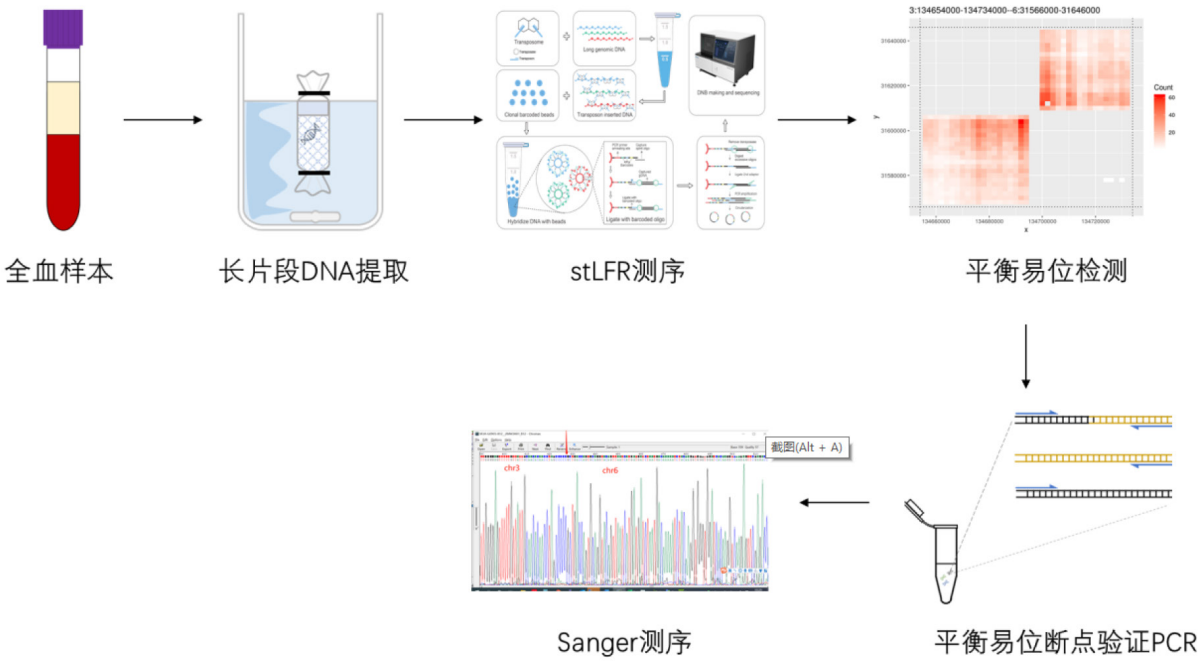


图2　平衡易位样本检测实验流程

结果

1. 100X stLFR数据与Nanopore 数据的SV分析结果和GIAB结构变异集的比较

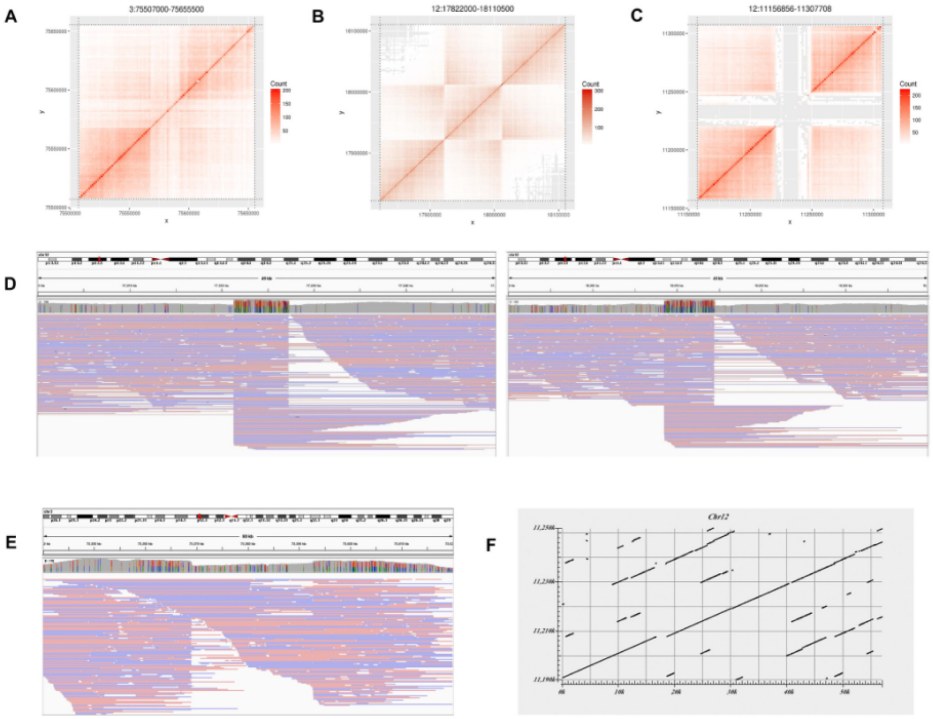
在缺失片段长度在50-1K，1K-10K和 >30K时，长读长的召回率都高于stLFR，但精确率要低于stLFR。但是在10K-30K时，stLFR测序数据加上stLFRsv软件的召回率和精确率都高于长读长，stLFRsv软件的召回率也高于其他SV分析软件。当缺失片段长度>30K时，stLFR数据使用stLFRsv可以保证较高的召回率和一定的精确率。但是stLFR测序数据和stLFRsv的组合在 50-10K的片段范围表现不好。总的来说，stLFR数据加上stLFRsv可以最大可能的准确检测出10K以上的缺失（数据见表1）。

表1 不同大小缺失与GIAB结构变异集的比较评估结果

		100 x long reads			100 x co-barcoded reads			
	Mapping	Sniffles Minimap2	Long Ranger lariat	NAIBR bwamem2	stLFRsv bwamem2	smoove bwamem2	stLFRsvb + smoove bwamem2	GROC-SVs bwamem2
50-1 k	Benchmark				4719			
	Total call	9453	3583	2	0	972	972	0
	TP	4168	2304	2	0	724	724	0
	FP	5285	1279	0	0	248	248	0
	FN	551	2415	4717	4719	3995	3995	4719
	Precision	44.09%	64.30%	100.00%	-	74.49%	74.49%	-
	Recall	88.32%	48.82%	0.04%	-	15.34%	15.34%	-
1 k-10 k	Benchmark				577			
	Total call	902	489	155	13	554	556	0
	TP	533	391	125	12	434	436	0
	FP	369	98	30	1	120	120	0
	FN	44	186	452	565	143	141	577
	Precision	59.09%	79.96%	80.65%	92.31%	78.34%	78.42%	-
	Recall	92.37%	67.76%	21.66%	2.08%	75.22%	75.56%	-
10 k-30 k	Benchmark				31			
	Total call	60	27	31	56	35	56	9
	TP	28	19	24	30	22	30	7
	FP	32	8	7	26	13	26	2
	FN	3	12	7	1	9	1	24
	Precision	46.67%	70.37%	77.42%	53.57%	62.86%	53.57%	77.78%
	Recall	90.32%	61.29%	77.42%	96.77%	70.97%	96.77%	22.58%
>30 k	Benchmark				9			
	Total call	55	14	28	23	36	23	13
	TP	9	6	8	7	7	7	4
	FP	46	8	20	16	29	16	9
	FN	0	3	1	2	2	2	5
	Precision	16.36%	42.86%	28.57%	30.43%	19.44%	30.43%	30.77%
	Recall	100.00%	66.67%	88.89%	77.78%	77.78%	77.78%	44.44%

使用stLFRsv检测出的部分与变异集不一致的缺失，可以与长读长测序结果可以相互验证（图3A，3E，3C，3F，3G），只有三个大于10kbp的缺失没有检测到。说明stLFR检测大片段缺失具有优势。

除缺失以外，还检测出55个倒位，重复和易位，其中的大多数存在多个基因组中，可能是由于参考序列问题或重复区域造成的结果。部分倒位可能由于参考序列人群的差异，这部分结果可以与长读长测序相互验证（图3B，3D，3H）。



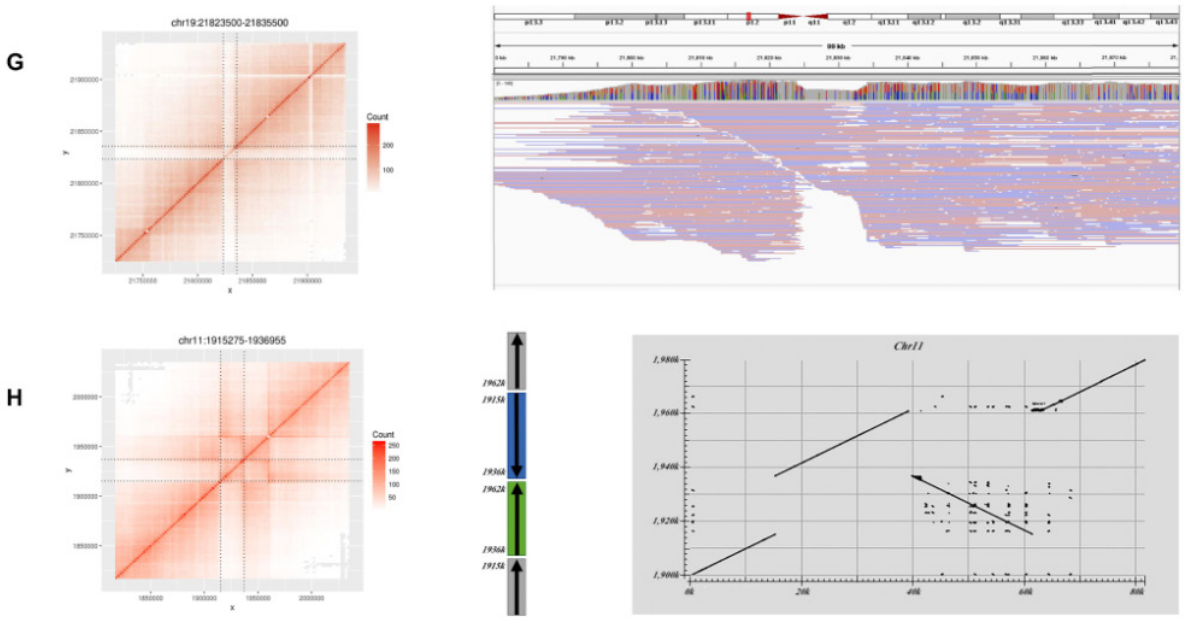


图3 未能与GIAB结构变异集匹配的大片段变异

（A）3号染色体上缺失位置的barcode热图。

（B）12号染色体上倒位的barcode热图。

（C）12号染色体上缺失位置的barcode热图。

（D）长读长测序数据比对结果支持（B）图中12号染色体上的倒位。

（E）长读长测序数据比对结果支持（A）图中3号染色体上的缺失。

（F）（C）图中12号染色体上缺失位置组装比对到参考基因组。

（G）19号染色体上缺失位置的barcode热图与长读长测序数据比对结果IGV展示图。

（H）11号染色体上倒位的barcode热图及组装比对结果。

向下取不同深度的数据（30X，50X）进行分析比较结果显示10K–30K结构变异检测的召回率和精确率会随深度降低而降低，30K以上在50X深度数据量时与100X基本没有差别。

表2 不同深度stLFR数据检测到的缺失与GIAB结构变异集评估结果

		stLFRsv	stLFRsv	stLFRsv	stLFRsv+smoove	stLFRsv+smoove	stLFRsv+smoove
	Coverage	30X	50X	100X	30X	50X	100X
50-1 k	Benchmark				4,719		
	Total call	0	0	0	359	649	972
	TP	0	0	0	276	490	724
	FP	0	0	0	83	159	248
	FN	4719	4719	4719	4443	4229	3995
	Precision	-	-	-	0.7688	0.7550	0.7449
	Recall	-	-	-	0.0585	0.1038	0.1534
1 k-10 k	Benchmark				577		
	Total call	5	11	13	403	506	556
	TP	4	10	12	325	408	436
	FP	1	1	1	78	98	120
	FN	573	567	565	252	169	141
	Precision	0.8000	0.9091	0.9231	0.8065	0.8063	0.7842
	Recall	0.0069	0.0173	0.0208	0.5633	0.7071	0.7556
10 k-30 k	Benchmark				31		
	Total call	59	57	56	58	57	56
	TP	18	26	30	18	26	30
	FP	41	31	26	40	31	26
	FN	13	5	1	13	5	1
	Precision	0.3051	0.4561	0.5357	0.3103	0.4561	0.5357
	Recall	0.5806	0.8387	0.9677	0.5806	0.8387	0.9677
>30 k	Benchmark				9		
	Total call	12	20	23	12	19	23
	TP	3	7	7	3	7	7
	FP	9	13	16	9	12	16
	FN	6	2	2	6	2	2
	Precision	0.2500	0.3500	0.3043	0.2500	0.3684	0.3043
	Recall	0.3333	0.7778	0.7778	0.3333	0.7778	0.7778

通过以上的比较结果，我们确定stLFR在检测结构变异，尤其是在比较大的的结构变异方面具有优势，并且还能检测复杂的结构变异，更进一步可以检测出复杂结构变异的断点位置。平衡易位作为一种复杂结构变异，会导致子代染色体异常，会引起复发性自然流产，在人群中的发生率是0.1%~0.2%，反复流产患者中的发生率3%~5%。为验证stLFR检测复杂结构变异断点位置的能力，我们收集平衡易位样本进行实验。

2. 5例平衡易位样本stLFR检测结果

5例平衡易位样本血液样本经过长片段DNA提取、stLFR建库和测序之后，得到stLFR测序数据的深度在23X~33X。为确定stLFR平衡易位染色体断点检测的准确性，针对断点设计引物进行sanger测序验证，确定断点位置处于stLFR检测到的范围内。

表3 5例样本的测序数据基本统计结果

Sample name	Raw reads	Clean reads	Clean data rate (%)	Barcode number	Mapped reads	Mapping rate	Duplicate rate	Average sequencing depth	Coverage (%)	Coverage(20X)(%)
M3	1677138338	1532563606	78.57	55920527	1521616483	0.9929	0.3329	32.49	0.9978	0.797
M6	1554820452	1213495850	78.05	68500575	1212308358	0.999	0.3715	25.76	0.9974	0.7147
M10	1384952028	1123175560	81.1	61774033	1122016048	0.999	0.386	23.31	0.9972	0.6201
F9	1542627832	1208865290	78.36	67996341	1208085035	0.9994	0.3263	27.54	0.991	0.7937
F10	1575274966	1264116004	80.25	69500194	1263345784	0.9994	0.3333	28.51	0.991	0.821

使用stLFRsv软件进行分析，能检测到相应的平衡易位，分析得到的断点范围在4K~10K以内。分析得到的平衡易位的断点位置转换成染色体区带信息后，与核型位置有1个带的差别。

表4 stLFR平衡易位检测结果与sanger测序验证结果

样品名称 sample name	核型结果	stLFR检测结果			sanger验证结果	sanger测序结果
		stLFR SV 检出位置	是否与核型结果一致	核型表示	sanger验证位置	
M3	46,XY,t(3;6)(q23;p21.3)	chr3:134694000-134702000	N，位置差别1个带	46,XY,t(3;6)(q22;p21.3)	chr3:134700052 chr6:31611266	chr3断点位置有5133bp的del
		chr6:31606000-31612000	Y		chr3:134694918 chr6:31606565	chr6断点位置有4700bp的del
M6	46,XY,t(4;8)(q21;q21)	chr4:85110000	Y	46,XY,t(4;8)(q21;q12)	chr4:85107995 chr8:62522642	chr4断点位置有6bp的del
		chr8:62520000	N，位置差别1个区		chr4:85108002 chr8:62522618	chr8断点位置有23bp的del
M10	46,XY,t(3;7)(p21;p13)	chr3:53884000-53888000	Y	46,XY,t(3;7)(p21;p14.1)	chr3:53886571 chr7:39765756	
		chr7:39764000-39768000	N，位置差别1个带		chr3:53886572 chr7:39765750	chr7断点位置有5bp的del
F9	46,XX,t(3;13)(p13;q32)	chr3:84874000-84884000	N，位置差别1个带	46,XX,t(3;13)(p12.1;q22.1)	chr3:84881365 chr13:74627989	chr3断点位置有4837bp的del
		chr13:74628000-74632000	N，位置差别1个区		chr3:84876528 chr13:74629483	chr13断点位置有1494bp的del
F10	46,XX,t(4;12)(p14;p11.2)	chr4:43116000-43122000	N，位置差别1个带	46,XX,t(4;12)(p13;p12.3)	chr4:43118167 chr12:19982536	chr4断点位置有5bp的del
		chr12:19980000-19984000	N，位置差别1个带		chr4:43118162 chr12:19982536	

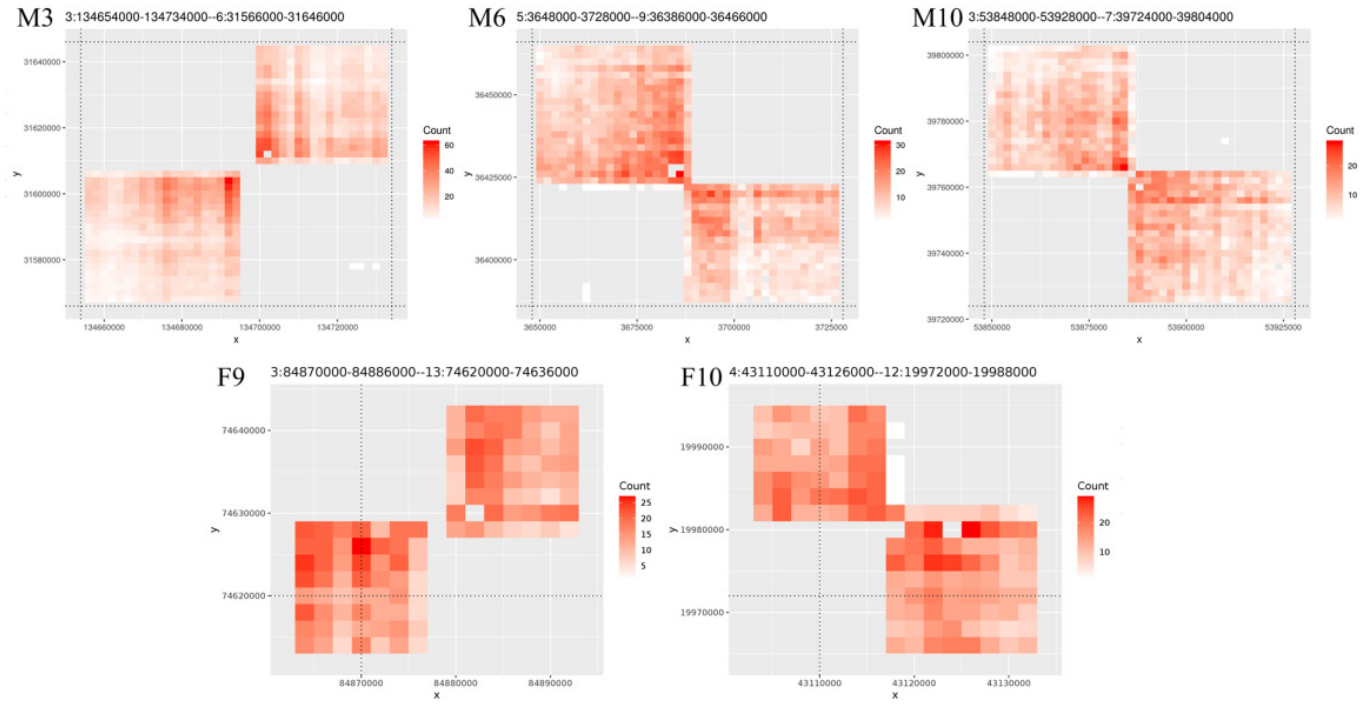


图4　5例样本的平衡易位位置的barcode的热图

设计PCR引物，扩增出衍生染色体的断点上下游的片段，进行sanger测序后得到的序列，确定了精确的断点位置，并且发现断点附近存在不同大小的缺失（表4，图5）。验证到的断点位置都在stLFR检测结果的范围内，说明stLFR能精准确定染色体平衡易位断点位置。

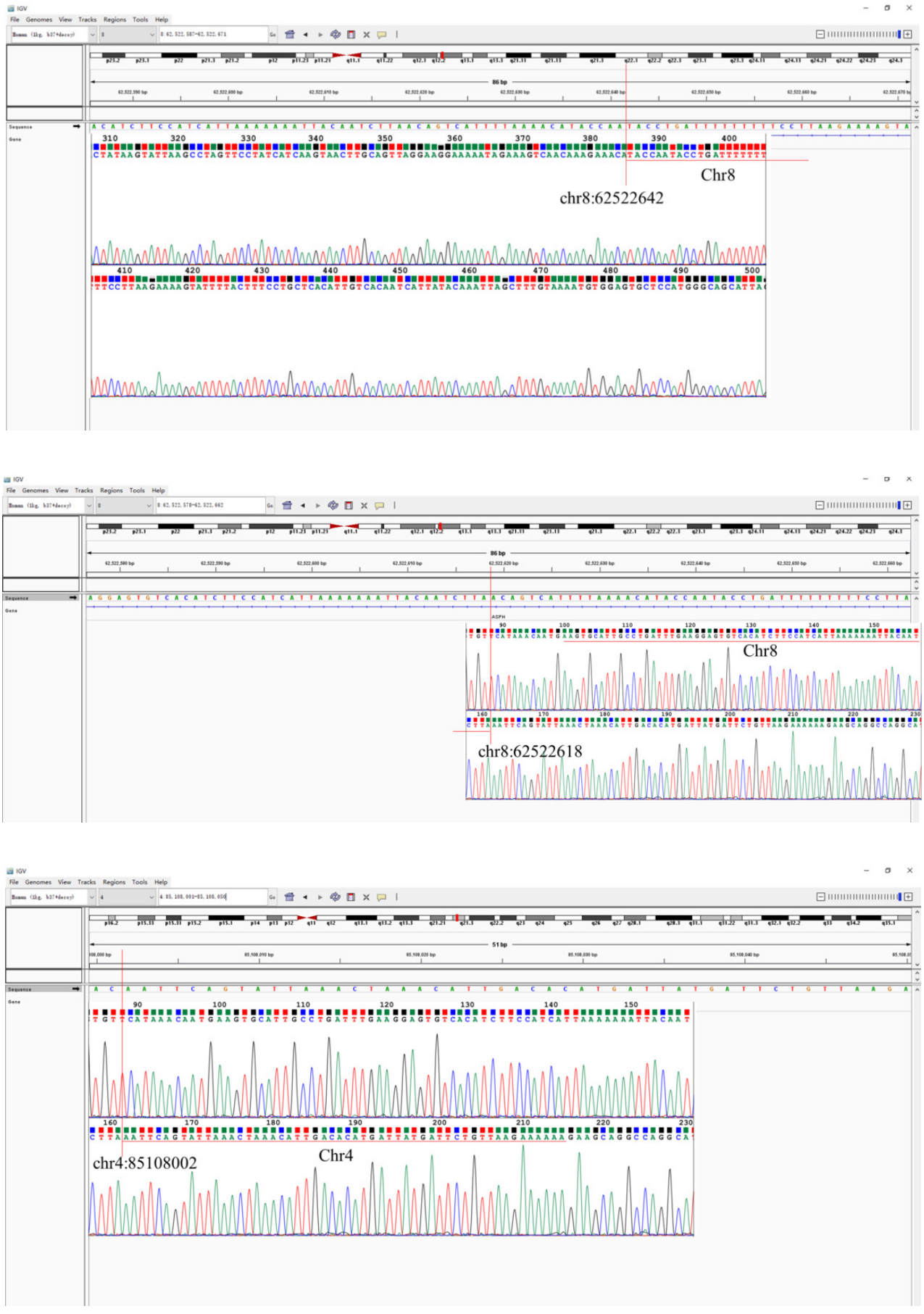


图5　M6样本sanger验证结果与对应的断点位置IGV展示图

总结

应用方案推荐：

分类	推荐方案
建库	stLFR文库
测序	DNBSEQ PE100 30X(有效数据，不含duplication)
分析流程	MGI-tech-bioinformatics/stLFR_V1.3

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. Guo J, Shi C , Chen X , et al. stLFRsv: A Germline Structural Variant Analysis Pipeline Using Co-barcoded Reads[J]. Frontiers in Genetics, 2021, 12:636239.

深圳华大智造科技股份有限公司
MGI-service@mgi-tech.com | www.mgi-tech.com