

华大智造SE400长读长测序助力地中海贫血患者中基因簇缺失的发现

2021年，广州AmCare基因组实验室与广州多家医院合作，在 *International Journal of Laboratory Hematology* 杂志上发表了题为 “Identification of thalassemia gene cluster deletion by long-read whole-genome sequencing (LR-WGS)” 的研究论文。该研究通过使用华大智造推出的MGISEQ-2000 SE400长读长模式对7例地中海贫血患者进行WGS分析发现：基于MGISEQ-2000测序仪的长读长SE400 WGS是地中海贫血致病突变检测的有力工具，可实现断裂点的有效检测，不仅可以检测已知的缺失，还可以发现与该疾病相关的新的编码/非编码缺失¹。

推荐应用：遗传病诊断

推荐机型：MGISEQ-2000RS

• SE400助力更高的诊断率

长读长SE400 全基因组测序在检测高度重复序列、结构变异以及 Indel变异等方面有显著优势，可显著提高地中海贫血致病突变的检出概率。

• 测序数据产出高效且质量高

DNBSEQ测序技术具有高准确性，低重复序列率以及低标签跳跃等重要特性。

• 建库全流程可高度自动化

该建库流程可借助华大智造自研的自动化系统实现高效、快捷的实验流程。



背景介绍

目前，基于高通量的靶向测序或全外显子组测序(WES)已被广泛应用于遗传病的临床诊断，阳性检出率为20%-50%。外显子组测序由于技术限制不能可靠地诊断一些基因突变，例如包括假基因在内的具有高度同源性的序列。通过单基因克隆获得的假基因于1997年被首次定义为基因组中与编码基因序列相似但不正常表达的基因序列。目前，人类基因组中大约有2万个假基因，而现已知共98个致病基因存在于假基因或具有高度同源性的序列中，例如珠蛋白基因簇。

高通量测序的读长通常为100-150bp，并不能区分源于基因与假基因的遗传突变。而长读长高通量测序因可直接检测重复序列故而可以将高度同源的基因与假基因区分开来。与WES相比，全基因组测序(WGS)可以通过长读长测序模式在结构变异(SVs)和Indel变异检测以及测序深度上展示出更好的性能从而有助于发现与罕见疾病有关的治疗变异²⁻⁴。

地中海贫血是一种常染色体隐性遗传的血液病，类型复杂，发病率和死亡率都很高^{5,6}。尤其是在中国南方，地中海贫血每年仍影响超过20000名新生儿⁷。PCR技术的不断完善使得分子技术成为地中海贫血诊断的金标准，但相关检测方法都或多或少的具有一定的缺陷，如只能检测有限的突变类型；不能自发的检测缺失或非缺失突变。故该领域急需要通过新的方法检测基因断裂点来填补这些检测方法的空白。

研究描述

本研究选择了7例地中海贫血缺失患者使用MGISEQ-2000的SE400长读长模式进行WGS测序，确定了基因组缺失断裂点，并发现7例患者均携带珠蛋白基因簇的杂合性缺失，其中3例突变位于 α 基因簇，3例突变位于 β 基因簇，1例在两个珠蛋白簇中都有缺失。这一发现将有助于诊断带有假基因或高度重复序列的遗传病并提供产前诊断。

实验方法

样本收集

本研究共收集7例地中海贫血并且血液学特征提示 α 或 β 基因缺失的患者样本且获得患者知情同意。Sanger测序排除了HBA1、HBA2或HBB基因中存在点突变的可能性。多重连接探针扩增(MLPA)分析表明，7例受试者均在 α -珠蛋白基因簇或 β -珠蛋白基因簇中存在杂合性大片段缺失，但缺失断裂点尚未确定。

DNA提取、文库制备和全基因组测序

使用SolPure血液DNA提取试剂盒从患者的外周血样本中提取基因组DNA，随后利用华大智造相关产品完成文库制备与测序工作。文库构建操作步骤大至

如下：先将基因组DNA打断后进行末端修复和接头连接，后进行PCR扩增，纯化后的PCR产物(片段长度为400~600bp)通过单链环化并制备成DNA纳米球后，在MGISEQ-2000基因测序仪上完成单端400个碱基的长读长测序(SE400)。值得关注的是，样本量较大时可使用华大智造提供的自动化提取、建库技术，可极大的节约人力并提高效率。

数据分析

MGISEQ-2000测序平台上的原始测序数据记录在FASTQ文件(Fq)中。使用BWA(版本：0.7.5)软件中的MEM算法将测序数据比对到人类参考基因组GRCh37/hg19上。然后，使用SAMTOOLS(版本：0.1.18)和Picard工具(版本：1.93, <https://broadinstitute.github.io/picard/>)对BAM文件进行分类、增加索引和消除重复数据。接下来，使用GATK(版本：3.1-1)提供的HaplotypeCaller工具来检测SNVs和INDELs，并使用ANNOVAR来注释SNVs和INDELs(参考信息来自数据库SNP、1000基因组和其他已发表数据库的信息)。此外，采用LUMPY来计算结构变异(SVs)。通过统计能比对上人类参考基因组的softClip读数总数以找到潜在可靠的断裂点，然后检索经softClip切割的序列从而找到断点的另一端。

样本采集	文库制备和测序	生信分析	结果分析
7例地贫样本收集	 MGIEasy 通用DNA文库制备试剂套装  MGISEQ-2000 基因测序仪	BWA SAMTOOLS Picard GATK ANNOVAR LUMPY	地中海贫血患者中SNVs、SVs和InDels等突变的检测

结果分析

地中海贫血患者的SE400 WGS分析

选择广州市妇女儿童医学中心编码的地中海贫血患者7例，将其外周血标本送入独立实验室(中国广州AMCARE基因组实验室)进行WGS和数据分析。MGISEQ-2000测序仪产生了118.89 GB的WGS

数据。结果发现所有7例患者都携带杂合性缺失，其中3例在 α 基因簇中，3例在 β 基因簇中，1例在两个珠蛋白簇中。其中，两名患者的DNA断裂点位置相同并导致了19 kb的基因组缺失；另有两名患者的断裂点相同且基因缺失长度为27 kb(表1和图1)。

Patient	Variant type	Hr	Deletion region	Deletion length	Copy number ^a
31941-ZZC	HBA1, HBA2 deletions	16	157 009-330 001	172 993	1
31942-HZX	HBA1, HBA2 deletions	16	215 396-234 699	19 304	1
31943-PHY	HBB deletions	11	5 222 878-5 250 289	27 412	1
31944-GXM	HBB deletions	11	5 236 361-5 257 771	21 411	1
31945-ZJL	HBB deletions	11	5 191 121-5 270 050	78 930	1
31946-LY	HBA1, HBA2, HBB deletions	11	5 222 878-5 250 288	27 411	1
		16	215 391-234 699	19 309	1
48197-YCT	HBA1 HBA2 deletions	16	220 862-231 981	11 120	1

表1. WGS分析鉴定HBA1/2和HBB突变

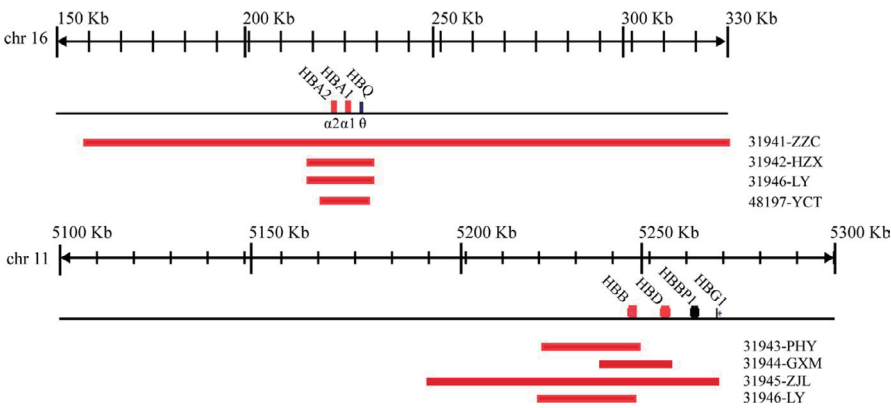


图1. 7例珠蛋白基因缺失总结。黑线表示正常区域序列，红色填充框表示chr11和chr16上缺失的序列。

7位患者在11号和16号染色体(GRCh37/hg19)上的DNA结构变异

1)患者31941-ZZC，32岁男性，表现出典型的α-地中海贫血特征。WGS分析检测到α-珠蛋白基因簇在chr16: 57009-330 001处有172kb的罕见缺失，缺失了HBA1和HBA2的所有编码外显子(图1，表1和表2)。

2)患者31942-HZX和31946-LY在chr16: 215396-234699有相同的DNA断裂点，HBA1和HBA2的量减少到正常对照组的一半(图1和表2)。通过设计一对能覆盖缺失区域的引物扩增出一个长度为1344 bp的片段(图3D)发现了基因组缺失范围为chr16: 215 256-235 908。此外，我们还发现患者31946-LY在chr: 115222878-5250288处存在包括HBB的所有编码外显子在内的β-珠蛋白基因缺失(图1，表1和表2)，并经PCR验证(图2D)。患者31943-PHY和31946-LY都有高血红蛋白F(Hb F)，用MLPA检测了β-珠蛋白簇，发现有缺失。

3)患者31943-PHY，女性，28岁，有持续性小红细胞和低色素性贫血，Hb F升高(19.3%)，这是遗传性胎儿血红蛋白持续存在症(HPFH)的典型特征(表1和表2)。

传性胎儿血红蛋白持续存在症(HPFH)的典型特征(表1和表2)。WGS检测到β-珠蛋白基因的杂合性缺失(图1)，后经PCR证实(图2D)。

4)患者31944-GXM，有小细胞性贫血和Hb F型(21%) HPFH(表1和表2)。WGS检测到β-珠蛋白基因在chr11: 5236361-5257771中存在21.4 kb的杂合性缺失(表2和图1)，并用PCR进行了验证(图2D)。

5)患者31945-ZJL，有小细胞性贫血和高脂血症的特征(表1和表2)。WGS检测到β-珠蛋白基因在chr11: 5 191 121-5 270 050中存在78.9 kb的杂合缺失(表2和图1)，经并用PCR进行了验证(图2D)。

6)对于48197-YCT患者，MLPA检测到α-珠蛋白基因簇的缺失(缺失探针涉及16P13.3: 160、313-196、305之间的18个探针，包括HBA2-、HBA2-、HBA1-和HBQ相关区域的上游)。然而，断点的确切位置尚不清楚。WGS检测到16号染色体的α-珠蛋白基因簇在220861至234981之间有11 kb的缺失(表2，图2A，B)。利用这一WGS信息，进行了PCR扩增并通过Sanger测序验证了断点片段(图2C)。

Sample ID	Gender	Age	Hb (g/L)	MCV (fL)	MCH (pg)	HbA (%)	HbA2 (%)	HbF (%)
31941-ZZC	Male	32	142	66.6	21.3	97.5	2.5	0
31942-HZX	Male	30	139	70.2	23.6	97.6	2.4	0
31943-PHY	Female	28	120	71.7	23.3	76	4.7	19.3
31944-GXM	Female	31	123	76.5	24.9	77.2	1.8	21
31945-ZJL	Male	36	136	68.4	21	84.3	2.4	13.3
31946-LY	Female	25	123	70	22.2	80.5	4	15.5
48197-YCT	Female	28	103	67	22	97.4	2.6	0

表2. 7例地中海贫血患者的血液学资料

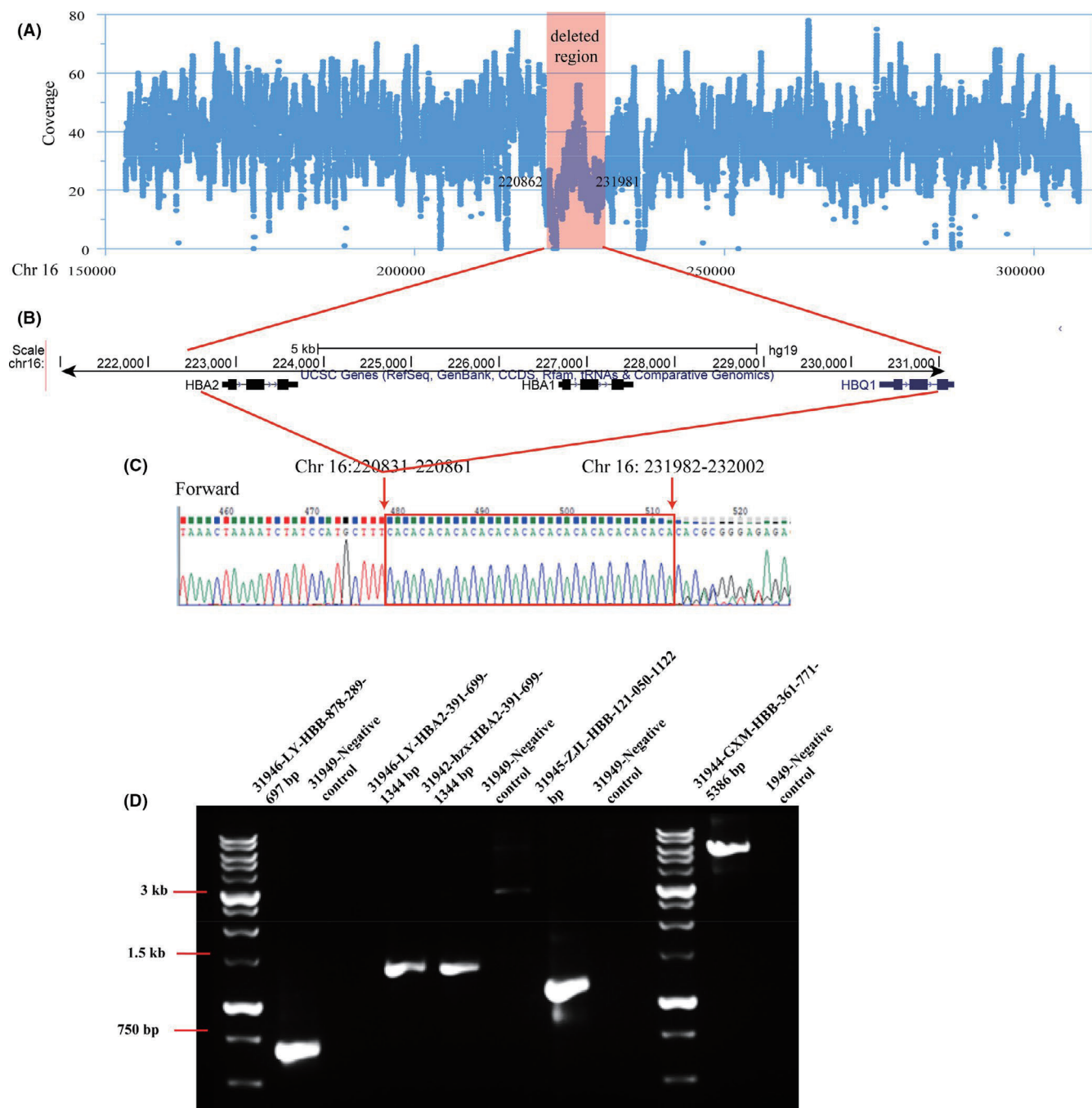


图2. 7例血红蛋白缺失的分子分析总结。A: 红色矩形: WGS的报道成功地定义了16号染色体上的一个结构变异: 220 862-231 981(GRCh37/hg19)。表格显示低覆盖率(<15人基因组)为缺失区。B: 展示了包含 16 号染色体上 HBA1/2 簇的正常区域。C: Sanger测序证实了48197-YCT患者的chr16: 220862-231 981中的断裂点。D: PCR证实了这七名患者的DNA断裂点。

总结

本研究选择7例地中海贫血患者为研究对象，进行了400bp长读长的全基因组测序，以准确检测地中海贫血基因缺失。

结果显示基于华大智造自主研发的SE400 WGS是地中海贫血断裂点检测的有力工具，不仅可以检测已知的缺失，还可以发现与该疾病相关的新的编码/非编码缺失，从而将研究范围扩大到非编码区，进而提高了诊断效率。此研究成果表明这是一款具有极大应用前景的检测工具。



基因测序仪MGISEQ-2000

参考文献

1. Jiang F, Lyu GZ, Zhang VW, Li DZ. Identification of thalassemia gene cluster deletion by long-read whole-genome sequencing (LR-WGS). *Int J Lab Hematol*. 2021 Aug;43(4):859-865.
2. Verdura E, Schluter A, Fernandez-Alvarez G, et al. A deep intronic splice variant advises reevaluation of presumably dominant SPG7 cases. *Ann Clin Transl Neurol*. 2020;7(1):105-111.
3. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*. 2014;511(7509):344-347.
4. Lindstrand A, Eisfeldt J, Pettersson M, et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Med*. 2019;11(1):68.
5. Taher AT, Weatherall DJ, Cappellini MD. Thalassemia. *Lancet (London, England)*. 2018;391(10116):155-167.
6. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. *Bull World Health Organ*. 2001;79(8):704-712.
7. Lai K, Huang G, Su L, He Y. The prevalence of thalassemia in mainland China: evidence from epidemiological surveys. *Sci Rep*. 2017;7(1):920-930.

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	900-0000350-00
	MGISP-100RS自动化样本制备系统	900-000070-00
	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
建库试剂	MGIEasy 通用DNA文库制备试剂套装(16 RXN)	1000006985
测序试剂	MGISEQ-2000RS 高通量测序试剂套装(FCL SE400)	1000013857

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



仅供研究使用

版权声明：本手册版权属于深圳华大智造科技股份有限公司所有,未经本公司书面许可,任何其他个人或组织不得以任何形式将本手册中的各项内容进行复制拷贝、编辑或翻译为其他语言。本手册中所有商标或标识均属于深圳华大智造科技股份有限公司及其提供者所有。

版本：2023年10月版

撰稿：霍明星 马兢璘

责任编辑：王其伟

审稿：江遥