

三代测序与靶向捕获技术联用进行高分辨 *HLA* 基因分型及 MHC 区域单倍体型精细鉴定

陈佳¹, 舒明月¹, 里进², 付爱思¹, 杨帆³, 王邹³, 李一荣²,
邓子新¹, 刘天罡¹

1. 武汉大学药学院, 组合生物合成与新药发现教育部重点实验室, 武汉 430071

2. 武汉大学中南医院检验医学中心, 武汉 430071

3. 武汉生物技术创新公共技术服务平台, 武汉 430071

摘要: 人类白细胞抗原(human leukocyte antigen, *HLA*)的高分辨率、精准分型对于组织配型以及 *HLA* 相关疾病研究具有重要意义。本研究以 12 位原发性肝细胞癌病人的外周血为供试样本, 分析二、三代测序数据用于高分辨率 *HLA* 分型的优劣势, 同时结合探针捕获与三代测序技术对 YH、HeLa 标准细胞系以及一个原发性肝细胞癌病人的主要组织相容性复合体(major histocompatibility complex, MHC)区域进行靶向分析, 探究长读长测序技术对于整个 MHC 区域精细分析的潜力。研究表明: (1)二、三代测序技术均能实现 6~8 位高分辨 *HLA* 分型, 且两者分型结果一致。但是三代数据的覆盖均一度显著优于二代, 不会出现明显的“断层”现象; (2)超长的三代数据可直接跨越整个扩增子, 对于基因单倍体型的判定(phasing)具有明显优势。样本中 92.79% 的 *HLA* 基因能够得到准确的单倍体分型结果, 远高于二代的 75.65%; (3)长读长的三代测序数据不但能实现对 MHC 区域的更好组装, 还具有对整个 MHC 共计 3.6 Mb 区域进行 phasing 的能力, 而这将有助于明确各个突变位点、等位基因、非编码区等基因原件在每个 MHC 单倍体型上的定位与相互连锁信息, 为免疫等相关疾病的研究提供理论依据。

关键词: *HLA*; MHC; 三代测序; *HLA* 单倍体分型; NimbleGen 探针捕获技术

收稿日期: 2018-11-26; 修回日期: 2019-01-25

基金项目: “万人计划”青年拔尖人才项目资助[Supported by the Young Talents Program of National High-level Personnel of Special Support Program (The “Ten Thousand Talent Program”)]

作者简介: 陈佳, 在读硕士研究生, 专业方向: 微生物与生化制药。E-mail: chenjjia19940216@whu.edu.cn

通讯作者: 李一荣, 主任医师, 研究方向: 临床分子免疫学诊断。E-mail: liyirong838@163.com

邓子新, 教授, 研究方向: 合成生物学。E-mail: zxdeng@whu.edu.cn

刘天罡, 教授, 研究方向: 合成生物学。E-mail: liutg@whu.edu.cn

DOI: 10.16288/j.ycz.18-282

网络出版时间: 2019/2/25 17:19:18

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20190225.1719.008.html>

The third-generation sequencing combined with targeted capture technology for high-resolution *HLA* typing and MHC region haplotype identification

Jia Chen¹, Mingyue Shu¹, Jin Li², Aisi Fu¹, Fan Yang³, Zou Wang³, Yirong Li², Zixin Deng¹, Tiangang Liu¹

1. Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education, School of Pharmaceutical Sciences, Wuhan University, Wuhan 430071, China

2. Department of Laboratory, Zhongnan Hospital of Wuhan University, Wuhan 430071, China.

3. Public Technology Service Platform, Wuhan Institute of Biotechnology, Wuhan 430071, China

Abstract: The high-resolution and accurate typing of human leukocyte antigen (HLA) is of great significance for the study of tissue matching in organ transplantation and the correlation between *HLA* and disease. In this study, the peripheral blood of 12 patients with primary hepatocellular carcinoma was used to compare the advantages and disadvantages of the next- and third-generation sequencing technology for high-resolution *HLA* typing. In addition, probe capture technology was used to capture the MHC region of YH and HeLa standard cell lines, and a primary hepatocellular carcinoma patient. The captured products were sequenced using PacBio platform to assess the potential of ultra-long reads sequencing technology for analysis of the entire MHC region. Our results showed that: (1) the next- and third-generation sequencing technology can both achieve 6-8 digit high resolution in *HLA* typing. However, the coverage of the third-generation is significantly better than the next-generation sequencing technology. (2) The ultra-long reads of the third generation sequencing can directly span the entire amplicon region, which has obvious advantages for haplotype phasing, with 92.79% of the *HLA* genes having accurate phasing results, which is much higher than the 75.65% from the next-generation data. (3) The long-reads from the third generating sequencing can not only be used to assemble the MHC region but also the ability to phase the entire MHC region of 3.6 Mb, thereby helping to clarify the localization information of the mutation sites, alleles and non-coding regions on each MHC haplotype, and providing a theoretical basis for the study of immune and other related diseases.

Keywords: *HLA*; MHC; the third-generation sequencing; *HLA* haplotype phasing; NimbleGen probe capture technology

人类白细胞抗原(human leukocyte antigen, *HLA*)基因位于人体第 6 号染色体的短臂, 受控于人类主要组织相容性复合体(major histocompatibility complex, MHC)的基因簇^[1], 全长约 3.6 Mb, 是目前所知人体最复杂的遗传多态性系统^[2,3]。

前期研究表明, *HLA* 基因的变异与传染病^[4]、药物过敏反应^[5]、自身免疫疾病^[6]、器官移植反应^[7]以及恶性肿瘤^[1]等均有关联。此外, 近期研究也表明 *HLA* 特定等位基因多态性与原发性肝细胞癌的发生相关^[8]。因此, 准确的 *HLA* 分型技术对于组织

配型以及研究 *HLA* 与疾病相关性具有重要意义。

由于 MHC 区域存在高度的多态性和广泛的连锁不平衡, 因此研究人员对该区域所涉及到的分子机制的研究受到一定的限制。传统基于聚合酶链式反应等分型方法存在分辨率低、无法获得单倍体型结果以及新等位基因信息等诸多问题^[9,10]。虽然以往 Sanger 测序技术被视为是 *HLA* 分型的金标准^[11], 但其因通量低而逐步被各种第二代测序(next-generation sequencing, NGS)平台所取代^[9,12,13]。但是, NGS 对于个体所具备的两套同源染色体的独特核苷酸信息,

即“对单倍体型的判定”(phasing)的解析依然存在困难,而已有文献表明准确的单倍体分型能更好解读基因与表型(包括疾病)之间的关系^[14],尤其是对于 *HLA-DP* 等较大的 *HLA* 基因,NGS 因其读长短而很难准确获得单倍体型结果^[15]。基于单分子实时测序技术(single molecule real-time, SMRT)的第三代测序仪能产生平均长度在 10 kb 以上的数据,这不但有利于基因精确分型,还能实现对基因组复杂区域的组装以及对某个基因内及等位基因间差异的细致解析。目前利用该技术进行基因组完成图组装或用于多倍体基因组中单倍体型的解析研究已有所报道^[16,17]。

三代测序技术预期能实现对 *HLA* 基因及 MHC 区域更精细的分析,从而准确地确定每个基因甚至整个 MHC 区域的单倍体分型结果,并且可有效挖掘包括单核苷酸多态性(single nucleotide polymorphism, SNPs)在内的一系列遗传信息,这将极大地推进各类与人体免疫相关研究的发展。本研究以 12 位原发性肝细胞癌病人的外周血为供试样本,分析二、三代测序数据用于高分辨率 *HLA* 分型的优劣势,同时结合探针捕获与三代测序技术对 MHC 区域进行靶向分析,探究长读长测序技术对于整个 MHC 区域精细分析的潜力。

1 材料与方法

1.1 研究对象

本研究收集的 12 位原发性肝细胞癌(hepatic-cellular carcinoma, HCC)患者外周血样本均由武汉大学中南医院提供,所有患者均签署了知情同意书。YH 标准细胞系由中国国家基因库提供,HeLa 细胞采购自美国菌种保藏中心。

1.2 *HLA* 基因的测序分型

外周血样本总 DNA 提取使用 Hipure Blood DNA Mini Kit (广州美基生物科技有限公司)完成,YH、HeLa 细胞系总 DNA 的提取采用酚-氯仿提取法。*HLA* 基因扩增使用 GENDX NGSgo-AmpX 试剂盒以及 QIAGEN Long range PCR 试剂盒(QIAGEN 公司,德国),扩增产物使用 Qubit 2.0 定量后等摩尔混合,不同样本的 *HLA* 扩增子混合物分别采用

Illumina Miseq 以及 PacBio RSII 的标准混样建库流程进行文库制备及测序。PacBio RSII 原始数据使用 SMRT Portal 中的 RS_ReadsOfInsert 方法进行质量过滤,得到环形一致性序列(circular consensus sequences, CCS)的数据,将过滤的质量值(即 minimum predicted accuracy 参数)分别设为 0.80、0.85、0.90、0.95 和 0.99,分别得到 CCS0.80、CCS0.85、CCS0.90、CCS0.95 和 CCS0.99 的高质量数据。

HLA 分型分析中使用的各种开源或商业分型软件均采用从官方网站或者商业公司处获得的最新版本,其中,NGSengine、HLAssign、HLA-reporter、Omixon、HLAminer、HLA-VBseq 和 OptiType 用于二代数据分型,NGSengine 和 HLAminer 用于三代不同数据类型以及不同质量值 CCS 的数据分型。

1.3 MHC 区域捕获测序以及数据分析

使用 Roche NimbleGen MHC 探针捕获试剂盒对 MHC 相关区域(包括传统 MHC 区域和约 1.0 Mb 的 MHC 周边区域)进行捕获,实验过程根据三代测序长读长的特点对 DNA 打断、DNA 纯化体系以及 DNA 杂交时间等操作进行了优化,采用 PacBio RSII 的标准流程进行文库制备及测序。使用 FALCON 软件进行数据组装并且使用 SMRT Portal 中 RS_Resequencing 标准流程将原始测序数据比对到对应参考基因组 MHC 参考序列,以计算数据覆盖度并根据原始覆盖度(该流程的默认参数)统计 SNPs 在 MHC 区域的分布(未对覆盖度进行筛选,95%以上的覆盖度为 100×)。

MHC 区域单体型分析分别采用了 FALCON-Unzip 软件以及 targeted-phasing-consensus 脚本(<https://github.com/PacificBiosciences/targeted-phasing-consensus>)两套方法,并使用 MUMmer 软件将得到的单倍体分型结果与其对应的 *HLA* 基因分型结果进行比对,评估上述两套方法对 MHC 区域单倍体分型结果的差异。

2 结果与分析

2.1 基于二代测序数据对 *HLA* 基因分型

利用包括全基因组、全外显子组、转录组及 *HLA*

基因扩增子数据在内的多种数据类型进行 *HLA* 分型的各种学术和商业软件已被广泛使用(表 1)。为评估分型软件的准确性以便从中选择最佳的分型软件用于与后续三代数据分型结果的比较,比较了 7 种不同的 *HLA* 分型软件对基于全长扩增子二代测序数据的分析性能。为了确保分型结果的可靠性,不论是 Illumina Miseq 还是 PacBio RSII 均保证了足够的数据量。由于二代数据存在一定偏好,因此其覆盖度会存在不均一的情况,不同基因的覆盖度有一定的差异(图 1),但是总体而言,其 95% 以上的区域覆盖度会大于 200 \times 。而三代数据的覆盖度比二代数据更高(图 1)。通过比较,7 种软件对二代数据的分型结果展现出较大差异。从总体上来看,NGSengine 对 class 和 class 基因的分型结果敏感度都较高,结果与血清学鉴定结果吻合。而其他 6 种分型软件仅对 class 类基因的分型均较为准确,但对 class 类基因则不敏感,部分软件甚至不能给出分型结果。例如,HLAssign 和 HLAReporter 只有 2 个(2/12)和 0 个样本(0/12)预测到了 *DPA1* 基因分型。而 HLAminer 和 Omixon 对 class 基因的分型结果的判定与 NGSengine/血清学结果相比有很大的差异。由于 HCC27 样本的 *HLA* 基因分型结果经过了血清学结果及 MHC 区域捕获测序结果的双重验证,故以该样本作为示例(表 2),NGSengine 对 class 和 class 基因的分型结果与血清学鉴定结果以及 MHC 区域捕获测序结果高度吻合。但是 HLAAssign 和

HLAReporter 无法预测到 *DPA1* 的基因分型,而 HLAminer 和 Omixon 对 DPB1、DQB1、DRB1 等 class 基因的分型结果判定错误率较高(表 2)。

此外,OptiType 与 NGSengine 的分型结果很相似,但是其分辨率只有 4 位(表 2)。有文献表明,位于外显子外部的单核苷酸变异可能在疾病的发病机制中起关键作用^[27]。因此,高分辨率的 *HLA* 分型软件具有更高的应用价值。在测试的 7 种分型软件中,NGSengine 分型最准确,分辨率最高。后续将以 NGSengine 产生的结果为参照,评估三代测序数据分析结果。

2.2 基于三代测序数据对 *HLA* 基因分型

采用两种已公开能够使用三代测序数据的分型软件——HLAminer 和 NGSengine 对基于三代测序数据的 *HLA* 基因进行分型。此外,PacBio 三代测序数据分为 subreads 和环形一致性序列(circular consensus sequences, CCS)两类:subreads 是去除接头序列和低质量部分所得到的未经矫正的数据,而 CCS 是来自于同一个 DNA 分子经过环状反复测序产生的多条 subreads 相互矫正后得到的高准确性数据。理论上,CCS 数据的准确性越高越有利于获得准确的分型结果,但是矫正准确性设置越高最终所获得的 CCS 数据量也会减少,存在不能满足分型最低数据量需求的风险。因此,本研究测试了不同的数据类型和 CCS 准确性对分型结果的影响。

表 1 7 种 *HLA* 分型软件比较

Table 1 Comparison of *HLA* typing software

软件	原理	分辨率	数据类型	是否测试	参考文献
NGSengine		8-digit	Amplicon	Y	—
HLAminer	比对/组装	4-digit	WGS/WES/RNA-seq/amplicon	Y	[18]
ATHLATES	组装	4-digit	WGS/WES/amplicon	N	[19]
HLAFOREST	比对	4-digit	RNA-seq	N	[20]
OptiType	比对	4-digit	WGS/WES/RNA-seq	Y	[21]
Omixon	比对	6-digit	WGS/WES/amplicon	Y	[22]
HLAReporter	比对和组装	4-digit	WGS/WES	Y	[23]
HLA-VBSeq	比对	8-digit	WGS	Y	[24]
HLAssign	比对	6-digit	Sequence capture	Y	[25]
SEQ2HLA	比对	4-digit	RNA-seq	N	[26]

WGS: 全基因组数据; WES: 全外显子组数据; RNA-seq: 转录组数据; Amplicon: 基因扩增子数据; Y: 对应软件已测试; N: 对应软件未测试。

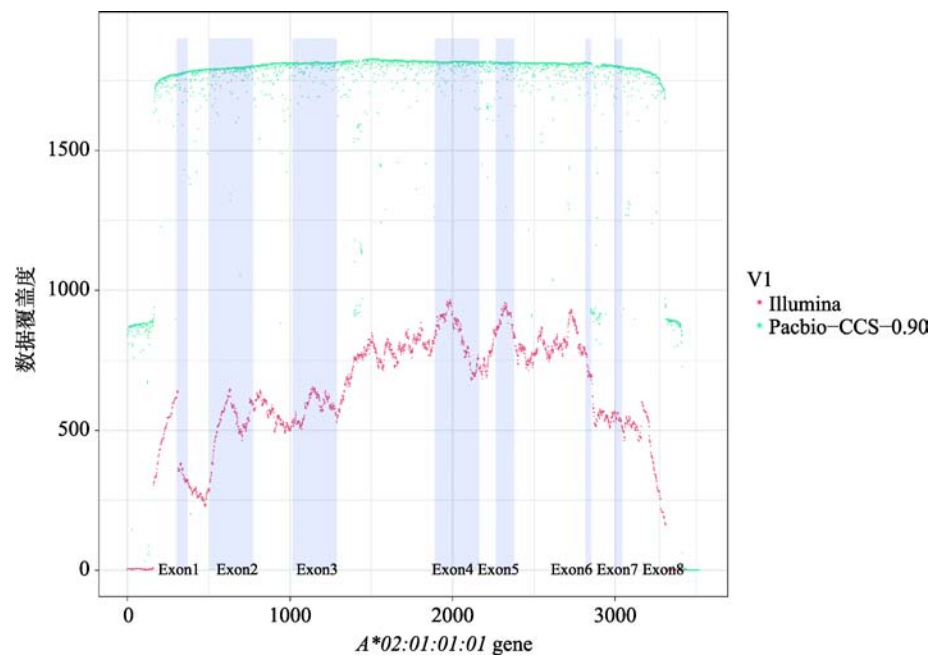


图 1 HCC27 样本 *HLA-A* 基因 Illumina 和 PacBio-CCS-0.90 数据覆盖度

Fig. 1 Comparison of coverage based on Illumina and PacBio-CCS-0.90 data from *HLA-A* of HCC27

分型结果	软件名称						
	NGSengine	HLAminer	HLAssign	HLA-reporter	Omixon	HLA-VBseq	Opti Type
HLA-A	A*02:01:01:01	A*02:01P	A*02:01:01	A*02:01:01G	A*02:01:01	A*02:01:01:01	A*02:01
	A*31:01:02:01	A*31:01P	A*31:01:02	A*31:01:02G	A*31:01:02	A*31:01:02	A*31:01
HLA-B	B*40:01:02	B*40:01:02:04	B*40:01:02	B*40:01:01G	B*40:01:02	B*40:01:02	B*40:01
	B*46:01:01	B*46:01P	B*46:01:01	B*46:01:01G	B*46:01:01	B*46:01:01	B*46:01
HLA-C	C*01:02:01	C*01:02P	C*01:02:01	C*01:02:01G	C*01:02:01	C*01:02:01	C*01:02
	C*07:02:01:01	C*15:102	C*07:02:01	C*07:02:01G	C*07:02:01	C*07:02:01:01	C*07:02
DPA1	DPA1*02:02:02	DPA1*02:01P	—	—	DPA1*02:02:02	—	—
	DPA1*04:01	DPA1*01:03P	—	—	DPA1*04:01	—	—
DPB1	DPB1*05:01:01	DPB1*40:01	DPB1*05:01:01	DPB1*05:01:01G	DPB1*135:01	—	—
	DPB1*13:01:01	DPB1*19:01P	DPB1*13:01:01	DPB1*13:01:01G	DPB1*519:01	—	—
DQA1	DQA1*01:03:01:04	DQA1*01:02P	DQA1*01:03:01	DQA1*01:03:01G	DQA1*01:03:01	DQA1*01:03:01:01	—
	DQA1*03:02	DQA1*02:01P	DQA1*03:02	DQA1*03:01:01G	DQA1*03:02	DQA1*03:02	—
DQB1	DQB1*03:03:02:02	DQB1*03:05P	—	DQB1*03:03:02G	DQB1*03:03:02	DQB1*03:03:02:03	—
	DQB1*06:01:01	DQB1*04:32	—	DQB1*06:01:01G	DQB1*06:01:15	DQB1*06:01:01	—
DRB1	DRB1*08:03:02	DRB1*07:01P	—	DRB1*08:03:02	DRB1*08:03:02	DRB1*08:03:02	—
	DRB1*09:01:02	DRB1*13:02P	—	DRB1*09:01:02	DRB1*09:01:02	DRB1*09:21	—
DRB4	DRB4*01:03:01:01	DRB4*01:01P	DRB4*01:03:01	—	DRB4*01:03:01N	—	—
	—	—	—	—	DRB4*01:10	—	—

加粗标记表明与 NGSengine 不一致的分型结果 ; “ — ” 表明无法得到该基因的分型结果。

首先,HLAminer 软件三代与二代数据的分型结果间存在差异,而 NGSengine 软件三代与二代数据的分型结果保持一致(表 3)。此外,当 subreads 的数据量足够时,基于 subreads 数据的分型结果与 CCS 数据基本一致,但前者往往会多出一个代表错配信息的后缀。以 HCC5 样本的 *DRB5* 基因为例,其基于 subreads 的分型结果为 *DRB5*02:02₃*,而 CCS 对应的分型结果为 *DRB5*02:02*,这表明 subreads 的结果在外显子区域存在 3 个错配碱基。从单碱基准确

性角度考虑,使用 CCS 数据对 *HLA* 基因分型应该是更好的选择。

随着 CCS 数据准确性的不断提高,其数据量会显著下降(图 2)。以 HCC27 样本为例,数据量下降最大的两个断层分别在 subreads 到 CCS0.80,以及 CCS0.95 到 CCS0.99 两处,其数据量分别降低了 78.86%和 89.84%。虽然 CCS0.99 准确性最高,但其 reads 数量从 subreads 的 3000 条降至 50 条,无法满足后续的分型需求。另一方面,数据量远多于

表 3 NGSengine 和 HLAminer 的分型结果分辨率及与 Illumina 分型结果一致性统计
Table 3 Resolution of the typing results of NGSengine and HLAminer and their consistency with Illumina typing results

软件	分型分辨率 6 位以上的基因个数	分型分辨率 8 位以上的基因个数	与 NGSengine-Illumina 分型结果一致的基因个数			
			CCS0.80	CCS0.85	CCS0.90	CCS0.95
NGSengine	103/114	49/114	99/114	99/114	99/114	99/114
HLAminer	4/114	2/114	28/114	30/114	36/114	43/114

114 个基因中有 99 个完全一致,15 个不一致,分型结果体现在第 8 位分型结果的差异,这类差异可通过软件优化与参数调整进一步减少。

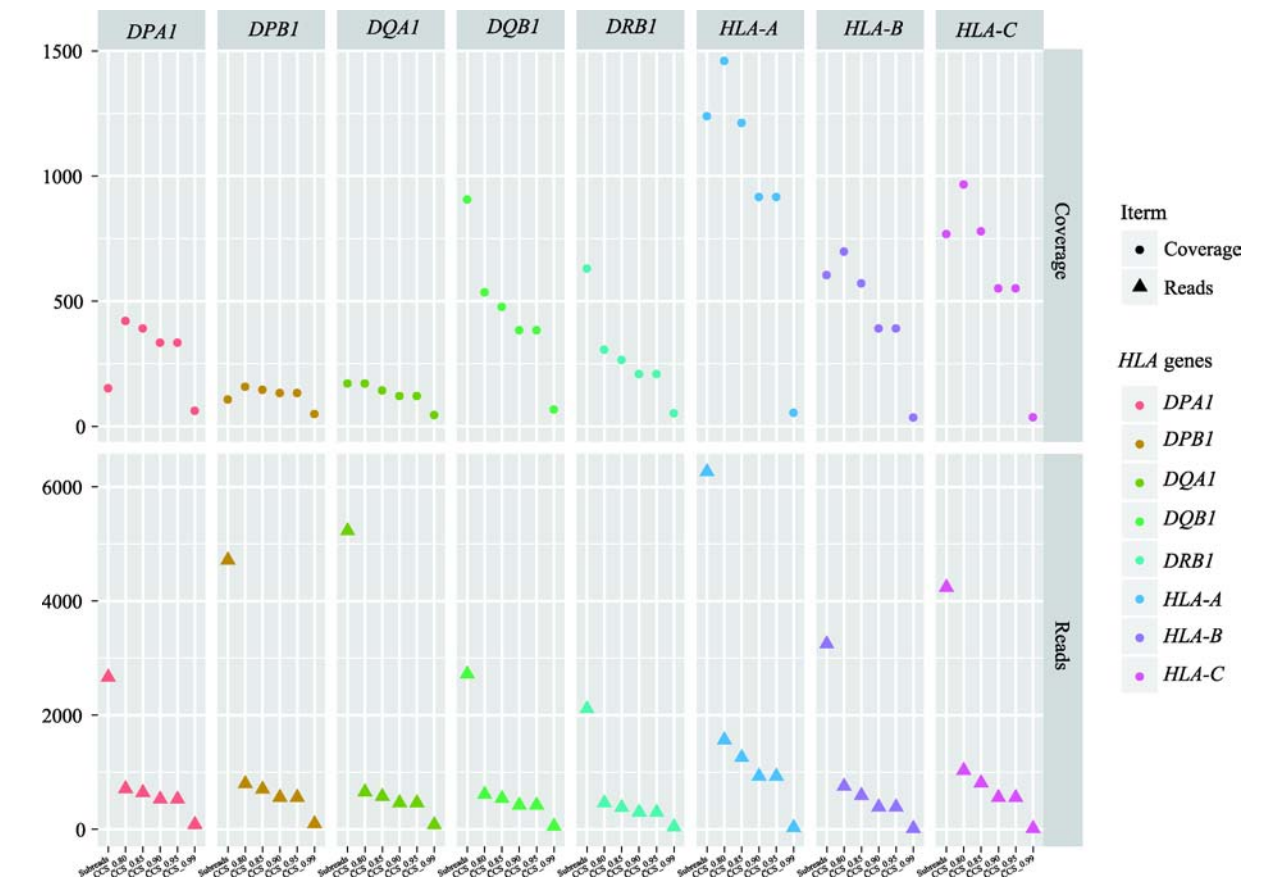


图 2 HCC27 样本在基于不同数据类型和 CCS 准确性下各 HLA 基因的覆盖度和数据量
Fig. 2 Coverage and reads of HLA genes in HCC27 based on different data types and CCS accuracy

CCS 数据的 subreads 并没有获得最好的覆盖度,除了 *DQB1* 和 *DRB1* 其他基因的 subreads 覆盖度甚至还低于 CCS0.80 (图 2),这可能是由于 subreads 错误率较高导致大量结果被过滤。因此,考虑到准确性和数据量之间的平衡,本研究最终选用 CCS0.90 的数据用于最终分型结果的比较。

虽然基于二代和三代 CCS0.90 数据的分型结果一致(表 3),但是两种数据的基因覆盖度却存在较大差异(图 1)。Illumina 数据存在一定偏好,在某些区段上会出现明显的“断层”现象,尤其在基因两端区域。相比之下,CCS0.90 数据的覆盖度更均匀,对于一些复杂的或全新的 *HLA* 基因具有更强的分型能力和更低的错误风险。

2.3 基于三代测序数据对 *HLA* 基因进行单倍体分型

Phasing regions 是用以评估单倍体分型效果的重要指标,主要代表目的基因中能准确分型单倍体区域的数目。分析结果显示基于二、三代数据进行 *HLA* 基因单倍体分型的结果间存在一定的差异(表 4)。基于三代数据,有 92.79% (103/111)的基因可以得到一条完整的单倍体结果,而这一比例在 NGS 数据里仅占 75.65% (87/115)。与此同时,同一个单倍体被定相到 3 个以上区域的比例中,NGS 占比 13.91% (16/115),而三代测序数据对应的占比仅为 3.6% (4/111) (表 4)。因此,三代测序更有利于提高单倍体分型的准确性,减少不确定性。

2.4 利用 PacBio 数据组装 MHC 区域以及 SNP 在 MHC 区域上的分布

MHC 捕获探针设计区域大小为 4 970 458 bp

表 4 二代(Illumina)和三代(PacBio)测序数据中所有 *HLA* 基因 phasing regions 个数的差异

Table 4 Differences in the number of phasing regions of all *HLA* genes between the Next(Illumina) and Third-(PacBio)generation sequencing technology

测序平台	单倍体分型区域数目		
	1	2	>3
Illumina	87	12	16
PacBio	103	4	4

(Chr.6: 28 477 797~33 448 354 bp),使用 FALCON 软件对 YH 标准细胞系进行组装,得到的最佳组装结果为:MHC 组装大小为 4.46 Mb,Contig N50 为 85 kb,Contig 总数 154 个。将组装的 Contig 比对到 YH 基因组的参考序列上(图 3),可以完整覆盖其 MHC 的参考序列。数据总覆盖度 97.86% (4 864 179/4 970 458 bp),跟以往使用二代数据结果获得的覆盖度为 97.29% 的结果相比有所提升^[28]。

随后,使用根据 YH 细胞系组装优化参数对 HCC27 样本 MHC 区域进行组装,得到的最佳组装效果为:MHC 组装大小为 4.79 Mb,Contig N50 为 90 kb,Contig 总数 223 个。数据覆盖度 99.8% (4 960 480 bp/4 970 458 bp)。0×以下的覆盖度比例为 0.21% (10 363/4 970 458 bp),30×以下的覆盖度比例为 2.23% (110 974/4 970 458 bp),意味着有 97.77% 的序列覆盖度达到 30×以上,组装效果进一步提升。

此外,本研究统计了 HCC27 样本、YH 和 HeLa 细胞系 MHC 区域 SNP 的分布情况(图 4),发现 *HLA* 基因区域的 SNP 频率显著升高,该结果与之前报道的结果一致^[28],这也是 *HLA* 基因多态性高的重要表现。

2.5 关于 HCC27 样本 MHC 区域的单倍体分型

FALCON+FALCON-Unzip 以及 targeted-phasing-consensus 采用了两种不同的单倍体分型原理。前者是基于数据的从头组装,是无参考序列的单倍体分析方法。后者的分析思路是基于数据比对,将原始的测序数据比对到参考序列上,然后根据比对的结果得到两条单倍体型结果,即 consensus0 和 consensus1。两种方法所得结果与本文 2.2 部分的结果基本一致(表 5)。为进一步验证上述 HCC27 样本 *HLA* 基因单倍体分型结果的准确性,对其产生的两条单倍体型序列中所含有 *HLA-A* 序列与各外显子区域内 *HLA-A* 基因进行比对(图 5, A 和 B),两者在单碱基水平上均保持一致。这表明基于长读长的三代测序数据,两种方法均可以对 MHC 区域上各 *HLA* 基因进行较为准确的单倍体型分析。此外,基于捕获的测序结果,还可以获得扩增测序难以得到的内含子信息。

另一方面,FALCON+FALCON-Unzip 会出现同

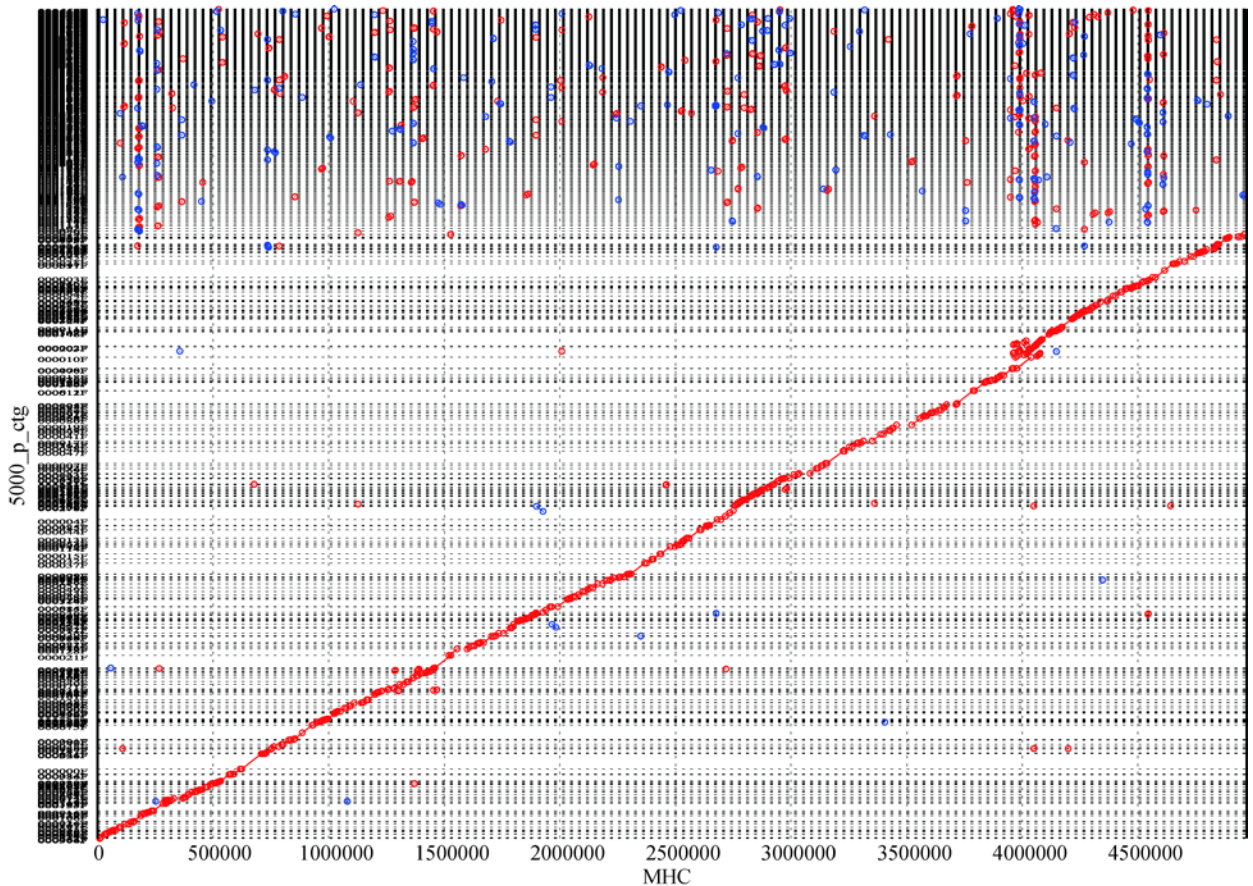


图 3 YH 细胞系 MHC 区域 FALCOM 组装序列与人类 MHC 的参考序列的比对结果

Fig. 3 Alignment of the FALCOM assembly sequence of the MHC region of the YH cell line with the reference sequence of human MHC

一基因被比对到不同 contig 上的情况(表 5),表明该组装方法对于 MHC 这类的多态性极高的区域可能会存在一定的单倍体分型错误。而基于比对方法得到的 consensus 序列,由于捕获测序长度限制,也可能导致在进行单倍体分型时,同一个位点的 SNP 信息难以准确定位到不同的 consensus 序列,从而导致模棱两可的结果。

因此,本研究整合了上述两种单倍体分型方法的所有信息以及各 HLA 基因扩增子的分型信息,对 HCC27 样本 MHC 区域的单倍体上的 HLA 基因的分布进行了校正和预测(图 6),两套 HLA 基因分别被定位到 consensus0 和 consensus1 两个单倍体型上。其中,没有用虚线标注的,即 *DPA1* 和 *DPB1* 等位基因,还通过从头组装的方法(FALCON+FALCON-Unzip 结果)验证了这两个等位基因的确位于一条 contig 上。基于此,可以大致了解来自双亲的两

套 HLA 等位基因、以及基因间其他功能原件在 MHC 上的确切位置与连锁关系,这对更深入研究基因与表型(包括疾病)之间的关系具有重要意义。

3 讨论

本研究评估了 7 种可使用二代测序数据和 2 种可使用三代数据的 HLA 分型软件。结果显示,二代数据的分型结果差异很大,其主要归因于各软件最适合输入数据类型、分析原理、数据库的差异(表 1)。而不论是基于二代还是三代数据,NGSengine 均能产生准确,分辨率高的分型结果,这说明分型实验设计与分析流程匹配的重要性。CCS 数据有助于提高分型结果的单碱基准确性,基于 CCS 分型结果其外显子错配信息会大幅减少,需要进行单碱基级别分析的研究可采用 CCS0.90 的数据进行分析。不同

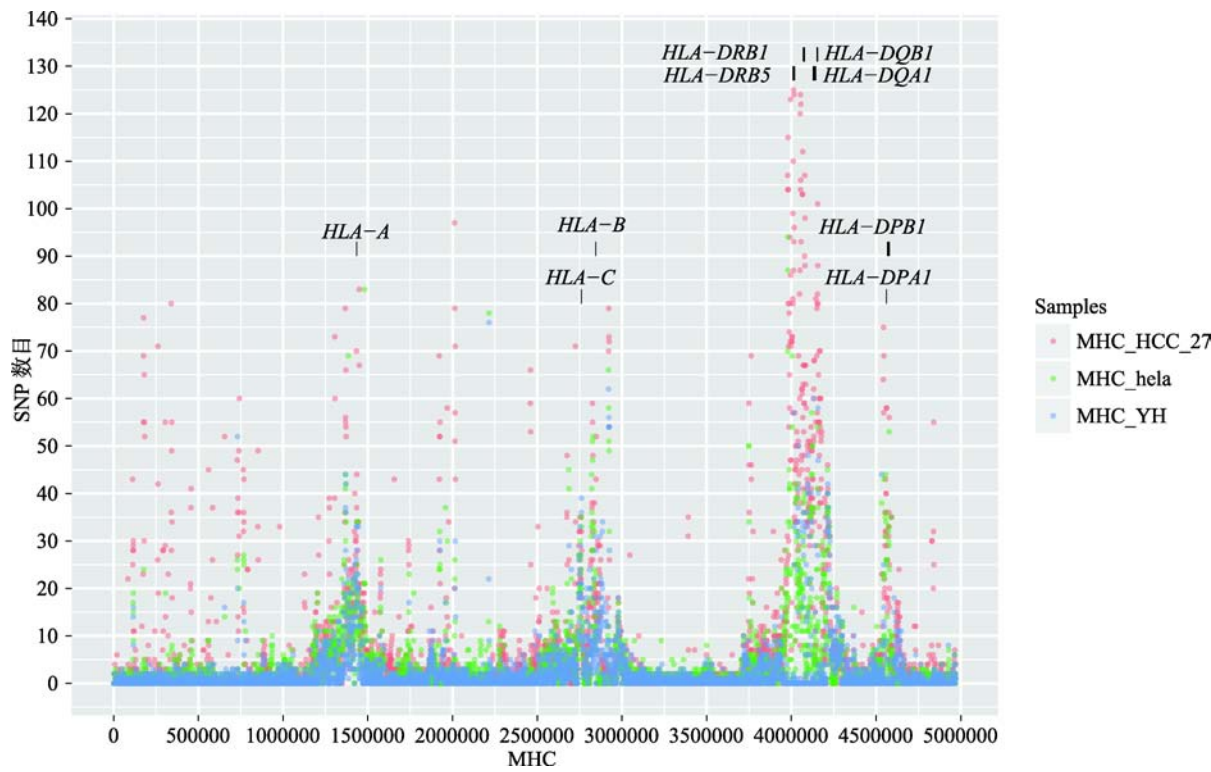


图 4 HCC27、YH 和 HeLa 标准细胞系 SNP 在 MHC 区域的分布图
Fig. 4 Distribution of SNPs in the MHC region of HCC27,YH and HeLa standard cell lines

表 5 HCC27 基于 FALCON+FALCON-Unzip 以及 Targeted-phasing-consensus 的单倍体分型结果
Table 5 The phasing results of HCC27 based on the FALCON+FALCON-Unzip and Targeted-phasing-consensus methods

FALCON+FALCON-Unzip		Targeted-phasing-consensus	
contig_Tags	单倍体分型结果	consensus_Tags	单倍体分型结果
000009F arrow	<i>A*02:01:01:01_gen</i>	consensus1	<i>A*02:01:01:01_gen</i>
000009F_001 arrow	<i>A*02:01:01:01_gen</i>		
000009F_001 arrow	<i>A*31:01:02:01_gen</i>	consensus0	<i>A*31:01:02:01_gen</i>
000018F arrow	<i>B*46:01:01_gen</i>	consensus1	<i>B*46:01:01_gen</i>
000018F_002 arrow	<i>B*46:01:01_gen</i>		
000018F_002 arrow	<i>B*40:01:02:01_gen</i>	consensus0	<i>B*40:01:02:01_gen</i>
000018F arrow	<i>C*01:02:01_gen</i>	consensus1	<i>C*01:02:01_gen</i>
000018F_002 arrow	<i>C*01:02:01_gen</i>		
000018F_002 arrow	<i>C*07:02:01:01_gen</i>	consensus0	<i>C*07:02:01:01_gen</i>
000027F_001 arrow	<i>DPA1*02:02:02_gen</i>	consensus0	<i>DPA1*02:02:02_gen</i>
000027F arrow	<i>DPA1*04:01_nuc</i>	consensus1	<i>DPA1*04:01_nuc</i>
000027F_001 arrow	<i>DPB1*05:01:01_nuc</i>	consensus0	<i>DPB1*05:01:01_nuc</i>
000027F arrow	<i>DPB1*13:01:01_nuc</i>	consensus1	<i>DPB1*13:01:01_nuc</i>
000017F arrow	<i>DQA1*01:03:01:04_gen</i>	consensus0	<i>DQA1*01:03:01:04_gen</i>
000017F arrow	<i>DQA1*03:02_gen</i>	consensus1	<i>DQA1*03:02_gen</i>
000017F arrow	<i>DQB1*03:03:02:02_gen</i>	consensus1	<i>DQB1*03:03:02:02_gen</i>

contig_Tags 和 consensus_Tags 表明单倍体分型结果对应的 contig 和 consensus 编号；加粗标记表示同一单倍体分型被比对到不同的 contig 上。

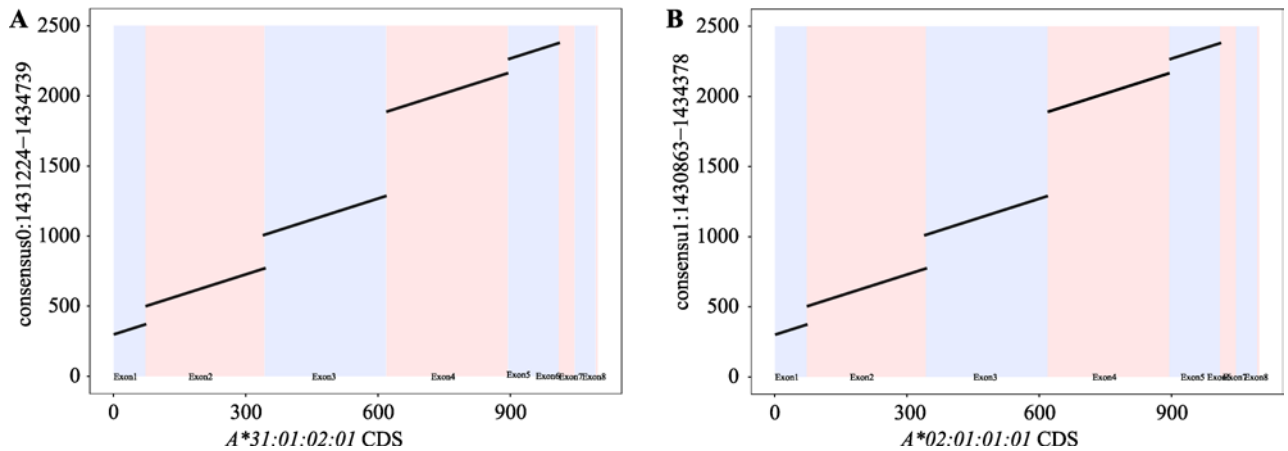


图 5 HCC27 样本 consensus0 和 consensus1 与 *HLA-A* 基因的比对结果

Fig. 5 Sequence alignment of consensus0 and consensus1 and *HLA-A* gene

A: HCC27 样本 consensus0 与 A*31:01:02:01 的序列比对结果; B: HCC27 样本 consensus1 与 A*02:01:01:01 的序列比对结果。

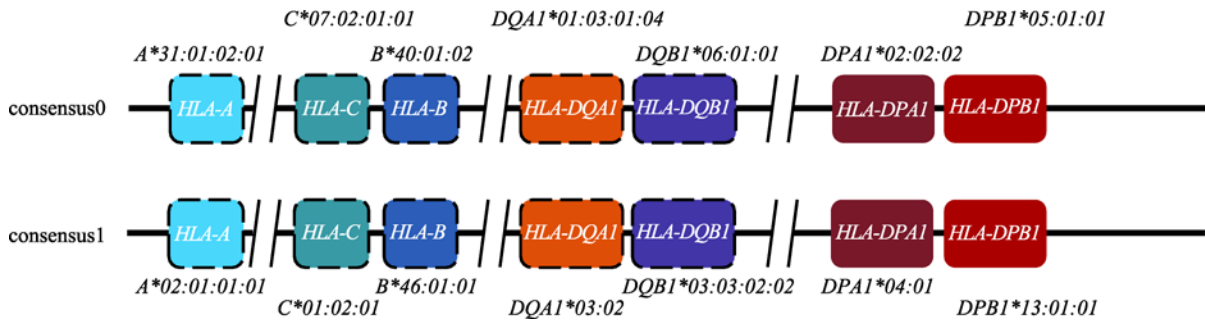


图 6 HCC27 MHC 区域的单倍体分型结果预测

Fig. 6 Proposed phasing results of MHC region of HCC27

consensus0 和 consensus1 表示利用 targeted-phasing-consensus 脚本得到的两个单倍体型,不同颜色表示不同的等位 *HLA* 基因。未用虚线标注的,表明其结果可通过从头组装的结果(FALCON+FALCON-Unzip 结果)得到验证,虚线边框标注的表示通过从头组装结果无法得到验证信息。

于二代数据需要通过多条数据组装/计算等分析手段对基因上各个位点的相位进行单倍体分型。单条三代测序数据可跨越较长的区域,基因分型与单倍体分型过程不涉及数据组装,减少了因组装而导致的错误。虽然在本研究中二、三代的分型结果基本一致,但是三代的覆盖均一度与单倍体分型结果均优于二代(图 1,表 4),可以大幅提高单倍体分型的准确性,减少模棱两可的分型结果,更适用于 *HLA* 基因的分型与单倍体分析。

YH、HeLa 和 HCC 样本的 MHC 区域捕获和三代测序的结果表明,三代数据的组装结果优于以往文献使用二代测序的结果,且结果准确性可以达到单碱基水平。此外,FALCON+FALCON-Unzip 软件由于是基于三代测序原始数据的无参考基因组单倍

体分型方式,可能出现同一基因被比对到不同 Contig 上的情况(表 5),从而导致组装出错。而 targeted-phasing-consensus 方法虽然是基于参考基因组序列的单倍体分型方法,但由于受到捕获产物测序数据长度的限制,同一个位点的 SNP 位点难以准确定位到不同的 consensus 序列,同样可能导致模棱两可的结果。因此,本研究将上述两种基于组装和比对的单倍体分型方法所得到的单倍体型信息以及本文 2.2 部分的 *HLA* 基因扩增子分型结果进行整合,对 HCC27 样本 MHC 区域的单倍体上的 *HLA* 基因的分布进行了校正和预测,通过从头组装的方法对预测结果进行验证,发现 *DPA1* 和 *DPB1* 等位基因的确位于同一 contig 上。基于该方法,可以从整体上了解来自于双亲的两套 *HLA* 等位基因、以及基因间其他

功能原件在 MHC 上的位置与连锁关系, 这将有助于对 MHC 这类结构复杂的基因区域进行系统研究并极大的推进各类相关疾病的相关性分析。

参考文献(References):

- [1] TANG MZ, CAI YL, ZHENG YM, ZENG Y. Association between human leukocyte antigen and nasopharyngeal carcinoma. *Hereditas(Beijing)*, 2012, 34(12): 1505–1512. 汤敏中, 蔡永林, 郑裕明, 曾毅. 人类白细胞抗原与鼻咽癌的相关性. *遗传*, 2012, 34(12): 1505–1512. [DOI]
- [2] YANG Zhao-Qing, CHU Jia-You. The research progress of human genetic diversity in China. *Hereditas(Beijing)*, 2012, 34(11): 1351–1364. 杨昭庆, 褚嘉祐. 中国人类遗传多样性研究进展. *遗传*, 2012, 34(11): 1351–1364. [DOI]
- [3] XU Jun-Pin, DENG Zhi-Hui, JU Gong-Yan, GAO Su-Jing, WANG Da-Meng, HE Liu-Mei, WEI Tian-Chi. Cloning and sequencing *HLA-A* and *-B* genomic DNA and analyzing polymorphism in regulatory regions in Chinese Han individuals. *Hereditas(Beijing)*, 2010, 32(7): 685–693. 徐筠婷, 邓志辉, 邹红岩, 高素青, 王大明, 何柳媚, 魏天莉. 中国汉族个体 *HLA-A*、*-B* 基因全长序列的测定及调控区多态性. *遗传*, 2010, 32(7): 685–693. [DOI]
- [4] Kløverpris HN, Adland E, Koyanagi M, Stryhn A, Harndahl M, Matthews PC, Shapiro R, Walker BD, Ndung'u T, Brander C, Takiguchi M, Buus S, Goulder P. HIV subtype influences *HLA-B*07:02*-associated HIV disease outcome. *AIDS Res Hum Retroviruses*, 2014, 30(5): 468–475. [DOI]
- [5] Mallal S, Nolan D, Witt C, Masel G, Martin A, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, James I, Christiansen FT. Association between presence of *HLA-B*5701*, *HLA-DR7*, and *HLA-DQ3* and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet*, 2002, 359(9308): 727–732. [DOI]
- [6] Galeazzi M, Sebastiani GD, Passiu G, Angelini G, Delfino L, Asherson RA, Khamashta MA, Hughes GR. *HLA-DP* genotyping in patients with systemic lupus erythematosus: correlations with autoantibody subsets. *J Rheumatol*, 1992, 19(1): 42–46. [DOI]
- [7] Sasazuki T, Fuji T, Morishima Y, Kinukawa N, Kashiwabara H, Inoko H, Yoshida T, Kimura A, Akaza T, Kamikawaji N, Koderia Y, Takaku F. Effect of matching of class I *HLA* alleles on clinical outcome after transplantation of hematopoietic stem cells from an unrelated donor. Japan Marrow Donor Program. *N Engl J Med*, 1998, 339(17): 1177–1185. [DOI]
- [8] Donaldson PT, Ho S, Williams R, Johnson PJ. *HLA* class II alleles in Chinese patients with hepatocellular carcinoma. *Liver*, 2001, 21(2): 143–148. [DOI]
- [9] Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D, Monos D. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Hum Immunol*, 2010, 71(10): 1033–1042. [DOI]
- [10] Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, Hayashi Y, Paumen M, Katsuyama Y, Mitsunaga S, Ota M, Kulski JK, Inoko H. Super high resolution for single molecule-sequence-based typing of classical *HLA* loci at the 8-digit level using next generation sequencers. *Tissue Antigen*, 2012, 80(4): 305–316. [DOI]
- [11] Latham K, Little AM, Madrigal JA. An overview of *HLA* typing for hematopoietic stem cell transplantation. *Methods Mol Biol*, 2014, 1109: 73. [DOI]
- [12] Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the *HLA* genes by next-generation sequencing. *BMC Genomics*, 2013, 14: 355. [DOI]
- [13] Barone JC, Saito K, Beutner K, Campo M, Dong W, Goswami CP, Johnson ES, Wang ZX, Hsu S. *HLA*-genotyping of clinical specimens using Ion Torrent-based NGS. *Hum Immunol*, 2015, 76(12): 903–909. [DOI]
- [14] Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. *Nat Rev Genet*, 2011, 12(3): 215–223. [DOI]
- [15] Nelson WC, Pyo CW, Vogan D, Wang R, Pyon YS, Hennessey C, Smith A, Pereira S, Ishitani A, Geraghty DE. An integrated genotyping approach for *HLA* and other complex genetic systems. *Hum Immunol*, 2015, 76(12): 928–938. [DOI]
- [16] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, 2013, 10(6): 563–569. [DOI]
- [17] Bowman B, Ranade S, Harting J, Lleras R. A novel analytical pipeline for de novo haplotype phasing and amplicon analysis using SMRT® sequencing technology. *J Biochem Technol*, 2014, 25(Suppl.): S17–S18. [DOI]
- [18] Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA. Derivation of *HLA* types from shotgun

- sequence datasets. *Genome Med*, 2012, 4(12): 95. [DOI]
- [19] Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, Pfeifer JD. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res*, 2013, 41(14): e142. [DOI]
- [20] Kim HJ, Pourmand N. HLA typing from RNA-seq Data Using Hierarchical Read Weighting. *PLoS One*, 2013, 8(6): e67885. [DOI]
- [21] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 2014, 30(23): 3310–3316. [DOI]
- [22] Major E, Rigó K, Hague T, Bérces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLoS One*, 2013, 8(11): e78410. [DOI]
- [23] Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, Sham PC, Lau YL, Yang W. HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med*, 2015, 7(1): 25. [DOI]
- [24] Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, Yamaguchi-Kabata Y, Yasuda J, Nagasaki M. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics*, 2015, 16(S2): S7. [DOI]
- [25] Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, Hov JR, Sauer S, Schimpler M, Ziemann M, Görg S, Jacob F, Karlsen TH, Franke A. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res*, 2015, 43(11): e70. [DOI]
- [26] Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci O, Diken M, Castle JC, Sahin U. HLA typing from RNA-Seq sequence reads. *Genome Med*, 2012, 4(12): 102. [DOI]
- [27] Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. *J Hum Genet*, 2015, 60(11): 665–673. [DOI]
- [28] Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, Xu Y, Liang D, Gao P, Sun Y, Gifford B, D'Ascenzo M, Liu X, Tellier LC, Yang F, Tong X, Chen D, Zheng J, Li W, Richmond T, Xu X, Wang J, Li Y. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. *PLoS One*, 2013, 8(7): e69388. [DOI]

(责任编辑: 方向东)