

鸡 T 细胞受体 γ 链基因位点重新测序和组装

马俊平^{1,2}, 杨犀^{1,2}, 律娜¹, 刘飞¹, 陈燕^{1,3}, 朱宝利¹

1. 中国科学院微生物研究所, 北京 100101;
2. 中国科学院大学, 北京 100049;
3. 中国农业科学院北京畜牧兽医研究所 ASTIP-IAS03, 北京 100193

摘要: 动物 T 细胞受体(T cell receptor, *TCR*)基因由多个不同的高度同源的基因家族组成, 通过全基因组测序很难获得准确的基因序列和排列位置。文章通过在 NCBI 中发布的鸡 *TCR* 的 γ 链(*TCR γ* 或 *TRG*)基因片段序列定位了鸡 *TRG* 基因所在区域, 并确定了与鸡 *TRG* 基因位点对应的细菌人工染色体(BAC)克隆(CH261-174P24)。对该克隆进行高通量的重新测序和组装后, 得到含有 10 个 scaffolds 的基因组草图, 较完整地覆盖了鸡 *TRG* 基因位点及两侧区域。通过 PCR 扩增和测序证明了 scaffold 内部结构的正确性, 校正了鸡参考基因组 *TRG* 基因位点一个可变基因和一个缺口序列(gap)附近各一处错误序列, 以及可变基因区多处序列错误。文章通过校正鸡参考基因组 *TRG* 基因位点的序列, 为鸡 *TRA/D* 和 *TRB* 基因位点的基因组序列分析提供了新方法。

关键词: 鸡; T 细胞受体(*TCR*); 伽马链(*TCR γ* 或 *TRG*)基因位点; 测序; 组装

Re-sequencing and assembly of chicken T cell receptor gamma locus

Junping Ma^{1,2}, Xi Yang^{1,2}, Na Lv¹, Fei Liu¹, Yan Chen^{1,3}, Baoli Zhu¹

1. Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, ASTIP-IAS03, Beijing 100193, China

Abstract: The genomic organization of the animal T cell receptor (*TCR*) loci is characterized by different gene families with high homology, and it is quite difficult to obtain accurate gene sequences and arrangements of these gene families. In this study, we identified the location of chicken *TCR* gamma chain (*TCR γ* or *TRG*) genes by comparing those *TRG* gene sequences with the chicken reference genome, and the corresponding bacterial artificial chromosome (BAC) clone, CH261-174P24, was chosen for further high-throughput DNA re-sequencing and assembly. As a result, a draft genome assembly containing ten scaffolds was obtained, which almost covered the chicken *TRG* gene locus and the flanking regions. Subsequently, the internal structure of these scaffolds was confirmed by PCR amplification and Sanger sequencing. Our analysis corrected two errors in the sequence—one near a *TRG* variable gene and one close to a gap, respectively, and several errors in the *TRG* variable genes in the chicken reference

收稿日期: 2014-11-04; 修回日期: 2015-03-09

基金项目: 国家重点基础研究发展规划项目(编号: 2015CB554204)资助

作者简介: 马俊平, 硕士研究生, 专业方向: 遗传学和基因组学。E-mail: majp12@im.ac.cn

通讯作者: 朱宝利, 博士, 研究员, 研究方向: 微生物与免疫基因组学。E-mail: zhubaoli@im.ac.cn

陈燕, 博士, 助理研究员, 研究方向: 动物遗传育种。E-mail: chenyan0204@163.com

DOI: 10.16288/j.ycz.14-382

网络出版时间: 2015-3-16 10:49:31

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20150316.1049.001.html>

genome. In conclusion, our work has partially corrected the erroneously assembled sequences of the *TRG* gene locus in the chicken reference genome and thus provides a new method for genome sequence analysis of chicken *TRA/D* and *TRB* gene loci.

Keywords: chicken; T cell receptor (*TCR*); gamma gene locus; re-sequencing; assembling

T 细胞受体(T cell receptor, *TCR*)是 T 淋巴细胞表面特异性识别抗原的受体,在机体的免疫应答和免疫调节中发挥着重要作用。对 *TCR* 基因结构与功能的研究,是深入探索机体免疫应答机理的重要内容之一。

对人(*Homo sapiens*)和小鼠(*Mus musculus*)*TCR* 的研究发现^[1~4],*TCR* 包括 α 、 β 、 γ 和 δ 4 种肽链,以两种形式存在,它们是由二硫键连接形成的 α/β 或 γ/δ 异二聚体蛋白链。在胚系中,编码 *TCR* 的基因序列是由多个可变的基因区域组成。*TCR β* 和 *TCR δ* 基因位点包括可变基因区(Variable gene region, V 基因区)、多样基因区(Diversity gene region, D 基因区)、结合基因区(Joining gene region, J 基因区)和恒定基因区(Constant gene region, C 基因区)等,*TCR α* 和 *TCR γ* 基因位点由 V 基因区、J 基因区和 C 基因区组成。但这些基因之间有较高同源性,且基因个数和排列顺序等较难确定,尤其是 V 基因区的基因数目众多且以基因簇的形式存在。这些都成为 *TCR* 研究的难点,使得 *TCR* 基因在染色体上很难进行准确定位。20 世纪末期,随着以 Sanger 双脱氧链终止法为代表的第一代测序技术在研究领域的广泛应用,人和小鼠 *TCR* 基因位点的序列和结构才逐渐被解析清楚,其功能也得到了注释。

但到目前为止,鸡 *TCR* 基因的相关研究开展得较少,其序列和结构并不完全清晰。20 世纪 90 年代以来,Cooper 等相继发现了鸡 *TCR* 的 α 、 β 、 γ 、 δ 等 4 种亚型,并对胸腺、肝脏等器官中的部分 *TCR* 基因序列进行了测定和分析^[5~13]。2004 年,国际鸡基因组测序中心在《自然》杂志上公布了红原鸡(*Gallus gallus*)的参考基因组草图^[14],采用鸟枪法对红原鸡的全基因组进行了测序和组装,并分别于 2006 年和 2011 年进行过两次补充。即便如此,NCBI 数据库中鸡参考基因组上 *TCR* 基因位点仍然存在很多缺口序列。

本研究拟通过对鸡 *TCR γ* (*TRG*)基因片段序列的分析,在红原鸡 BAC 克隆文库(CHORI-261)中挑选

包含鸡 *TRG* 基因的 BAC 克隆,采用 Illumina 公司的 Miseq 高通量测序系统对这些克隆进行重新测序和组装,以期获得鸡 *TRG* 基因位点的草图,并对鸡参考基因组上 *TRG* 基因位点的错误信息进行校正。

1 材料和方法

1.1 鸡 *TCR γ* 链(*TRG* 链)编码序列的定位

从 NCBI 数据库中下载鸡 *TRG* 链的编码序列。通过参考文献中报道的鸡 *TRG* 基因的一些特征^[8~11],分别提取 V 区、J 区和 C 区基因片段。然后与鸡参考基因组(WUGSC2.1/galGal3)进行比对。根据基因片段密集区域判断鸡 *TRG* 基因位点区间。

1.2 BAC 克隆的确定

根据加州大学圣克鲁兹分校基因组网站上(UCSC Genome Browser)鸡参考基因组注释的 BAC 克隆信息(galGal3),参照上一步定位的鸡 *TRG* 基因位点区间,在鸡 *TRG* 基因位点上下游约 50~100 kb 范围内挑选出载有鸡全部 *TRG* 基因的 1 个 BAC 克隆,编号为 CH261-174P24。

该克隆 CHORI-261-174P24(以下简称克隆 174P24)由奥克兰儿童医院及研究中心(Children's Hospital Oakland research Institute, CHORI)构建(网站: <http://bacpac.chori.org/library.php?id=120>),所用载体为质粒 pTARBAC2.1,此 BAC 克隆长度约为 218 kb。

1.3 BAC DNA 的提取

克隆 174P24 以半固体穿刺法保存,经过在含有氯霉素(终浓度为 12.5 ug/mL)的固体 LB 平板上划线和活化培养,得到单克隆。挑取单克隆于液体 LB 培养基中,采用氯霉素做诱导,进行过夜培养。利用 QIAGEN 公司的质粒中量提取试剂盒(QIAGEN Plasmid Midi Kit)提取 BAC DNA。

1.4 BAC DNA 的初步鉴定

1.4.1 酶切鉴定

从 NCBI 网站下载克隆 174P24 的参考序列(鸡

参考基因组版本为 ICGSC *Gallus_gallus*-4.0/galGal4)。运用 NEBcutter2.0 软件预测参考序列的限制性内切酶酶切位点。在预测结果里筛选所用限制性内切酶, 筛选标准为: (1)在参考序列中酶切位点较少, 以便于识别产物条带; (2)需为常用限制性内切酶, 以便购买和实验操作。经过筛选, 限制性内切酶 *Not* 适合用于鉴定。经内切酶 *Not* 充分消化后, 产物进行脉冲场凝胶电泳。根据酶切产物的条带初步判断所选克隆。

1.4.2 PCR 鉴定

依据克隆 174P24 的已知末端序列, 运用 Primer5.0 软件设计引物。以克隆 174P24 的 DNA 为模板, 进行 PCR 扩增。产物由北京六合华大基因科技股份有限公司测序。测序序列再与该克隆末端序列比对, 证明所选克隆的正确性。

1.5 BAC 克隆的高通量测序

克隆 174P24 的 DNA 经纯化后, 采用美国柏尔生物技术公司的快速 DNA 测序试剂盒(Bioo Scientific NEXTflex™ Rapid DNA-Seq Kit)构建 500~700 bp 的测序文库。由北京诺禾致源生物信息科技有限公司进行双端 250 bp 的 MiSeq 测序。

1.6 BAC 克隆的组装、分析和证明

运用 HTQC 软件^[15]和 Bowtie 软件对测序的原始 reads 进行预处理。获得的高质量序列对, 使用 ABySS 软件^[16]和 Phrap 软件^[17]进行组装。组装得到的 scaffolds 使用 NCBI BLAST 软件与鸡参考基因组(galGal4)进行比对, 并找出同源区域。比对结果及找到的同源区域使用自定义脚本进行提取与可视化。

针对 scaffolds 内部分段的连接处, 运用 Primer5.0

软件分别设计引物(表 1)。以克隆 174P24 的 DNA 为模板, 通过 PCR 扩增、测序和 BLAST 比对进行证明。

2 结果与分析

2.1 鸡 TCR γ 链编码序列的定位

从 NCBI 数据库中下载鸡 TRG 链的编码序列, 共有 203 条。从中提取 V 基因、J 基因、C 基因集合与鸡参考基因组(galGal3)的比对结果(图 1)显示, 鸡 TRG 基因定位于 2 号染色体的约 49 150~49 230 kb 区间内, 其中 C 基因区、V 基因区分别处于 5'端和 3'端, J 基因区位于两者中间。TRG V 基因平均间距约为 1.5 kb。从现有鸡 TRG 链编码序列在鸡参考基因组上的比对位置发现, 鸡 TRG 基因位点只含有一个 TRG C 基因, 包含 3 个外显子。

2.2 BAC 克隆的确定和初步鉴定

参照 UCSC 数据库中鸡参考基因组(galGal3)上注释的 BAC 克隆信息, 根据编码序列比对结果定位的鸡 TRG 基因位点在染色体上的位置, 在鸡 TRG 基因位点上下游约 50~100 kb 范围内选取了载有鸡全部 TRG 基因的一个克隆, 编号为 CH261-174P24 (图 1)。该克隆位于鸡参考基因组(galGal3)2 号染色体的约 49 109~49 313 kb 区间内。

利用内切酶 *Not* 充分消化克隆 174P24 的质粒 DNA, 并对产物进行脉冲场凝胶电泳。克隆 174P24 经酶切后在 10 kb 左右有载体 DNA 条带; 该克隆的插入片段中仅有一个 *Not* 酶切位点, 约 220 kb 的插入片段被酶切为约 190 kb 和约 30 kb 的两条带。酶切结果表明, 该克隆含有插入片段, 且大小与克隆 174P24 相符。此外, PCR 结果表明, 克隆 174P24

表 1 克隆 174P24 组装 scaffolds 内部连接处证明的引物信息

scaffold	引物			复性温度()	产物长度(bp)
	位置	正向(F)	反向(R)		
k192/174p24.scaffold1	1A-B	AACCTGGGAAATAAGCAA	CGTTAGATGGGCAAAGAT	52	1010
	1B+	TTTCTTTCAGGCTTGGTG	AGTGTCCCAGCGTTTAT	52	1233
k192/174p24.scaffold10	10A-B	AATCCGAATCATAGGAAACT	TCATTGGGAATCACAGAAA	52	1267
	10B-C	CATCTGGCACAGTAGTAGGT	ACAGGTGGTAAGGTTGAGT	56	2595
k192/174p24.scaffold9	9A-C	AGAGTCAGGGAGCAAGCG	TTGATTCGCGCCATTAG	52	1322
k192/174p24.scaffold3	3A-B	CAATGGCTGGATGAGAAA	AAACTAGAGTTGAAAGTGGC	52	1143
k192/174p24.scaffold2	2A-B	CTCAAGCAGAAGGGATGT	GAGGAAGGCTAGGCAAAC	52	1533

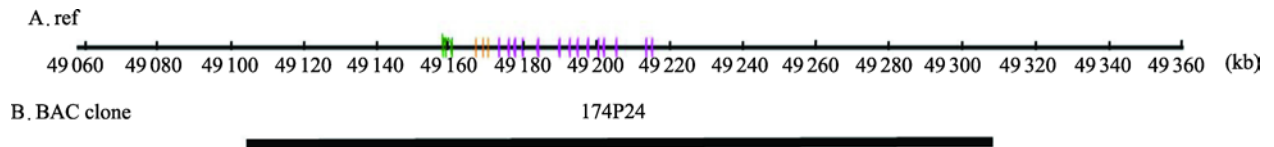


图 1 鸡 TRG 链编码序列比对到鸡参考基因组的示意图及对应的 BAC 克隆位置示意图

A: 鸡 TRG 链已知 V 基因(紫色竖线)、J 基因(黄色竖线)、C 基因(绿色竖线)集合比对到鸡参考基因组的示意图; B: 鸡 TRG 基因对应的 BAC 克隆及其在鸡参考基因组上的位置。

经过扩增得到约 400 bp 的单一产物, 对产物的测序结果与已知末端序列完全一致。推测鸡 TRG 基因位点位于克隆 174P24 中。

2.3 BAC 克隆的高通量测序

克隆 174P24 的测序数据量约为 237 Mb, 获得了 473 548 条 reads。其中, 宿主菌基因组的污染低于 1/1000; 两侧 read 质量均在 29 以上, 即测序准确性在 99% 以上; 测序深度为 1331 倍。

2.4 BAC 克隆的组装、分析和证明

2.4.1 BAC 克隆的组装和比对分析

本研究中鸡 TRG 链克隆 174P24 的测序数据已提交至 NCBI/SRA 数据库(登录号: SRR1615248), 该克隆组装的 scaffolds 信息已提交至 GenBank 数据库(登录号: KP058483)。本文运用 ABySS 软件和 Phrap 软件对该克隆进行组装。组装过程中经过多次尝试(k-mer 值: 96~192)和优化, 发现当 k-mer 值取 192 时组装获得的 scaffolds 最为完整。利用 NCBI BLAST 软件对得到的 scaffolds 与鸡参考基因组(galGal4)进行比对, 并找出同源区域, 结果使用自定义脚本进

行提取和可视化, 组装和比对结果见图 2。

克隆 174P24 参考序列的 GC 含量分布图显示, 在该克隆约 178 kb 区间内, GC 含量普遍低于 50%, 总 GC 含量为 39.13%; 在鸡 TRG 基因集中分布的约 49 670~49 730 kb(约 60 kb)区间, GC 含量为 42.29%, 较两侧序列高。参考序列 6 处 gap 中的 5 处也集中在这个区间内。从该克隆测序 reads 的丰度分布图可以看出, 与两侧序列比较, TRG 基因片段密集的约 60 kb 区间内的 reads 丰度稍低且分布不均匀。

经过组装本文得到了鸡 TRG 链克隆 174P24 的基因组草图, 该草图包含 10 个长度介于 1677~79 832 bp 之间的 scaffolds, 与参考序列的比对结果见图 2 和表 2。这 10 个 scaffolds 的总长度为 18 6931 bp, 与参考序列长度(约 178 kb)接近, 平均长度约为 18 kb。经过与参考序列的比较发现, scaffold7 长度为 62 793 bp, 位于 TRG 基因位点上游, scaffold2 长度为 79 832 bp, 位于 TRG 基因位点下游, 二者总长度共占 10 个 scaffolds 序列总长度的 78%。这两个 scaffolds 组装得比较完整, 并且位置与参考序列对应。而在 TRG 基因密集的 60 kb 区间内有以下 8 个

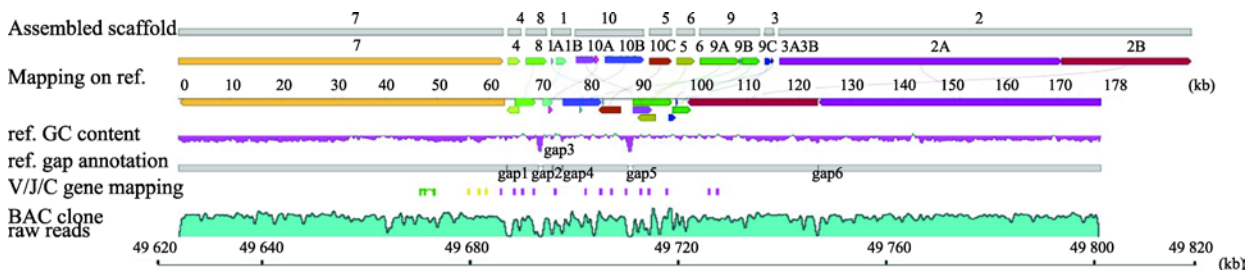


图 2 鸡 TRG 链克隆 174P24 的组装草图和与参考序列的比对结果

1: Assembled scaffold.(测序 reads 组装的 scaffold(500 bp 以上), 上方数字表示 scaffold 编号); 2: Mapping on ref.(scaffold 与参考序列的比对情况); 3: ref. GC content 反应参考序列的 GC 含量分布, 水平线代表 GC 含量为 50%; 4: ref. annotation 展示参考序列的 gap 注释情况(gap1~gap6: 100 bp、848 bp、100 bp、150 bp、958 bp、127 bp); 5: V/J/C gene mapping(从鸡 TRG 链 mRNA 序列提取的 V、J、C 基因片段比对到参考序列的位置, 其中绿色代表 C 基因, 黄色代表 J 基因, 紫色代表 V 基因); 6: BAC clone raw reads 展示测序 reads 的丰度分布。

表 2 鸡 *TRG* 基因位点克隆 174P24 组装的 scaffold 信息

scaffold	长度(bp)	GC 含量(%)	位置	方向 ^a	scaffold(bp)		参考序列 ^b (bp)	
					起始	终止	起始	终止
k192/174p24.scaffold7	62 793	36.77	7	-1	1	62 793	287	63 072
k192/174p24.scaffold4	2432	45.00	4	-1	1	2216	63 582	65 829
k192/174p24.scaffold8	3973	45.49	8	1	1	3972	65 087	69 054
k192/174p24.scaffold1	3680	49.78	1A	1	1	272	82 055	82 326
			1B	1	925	2786	70 486	72 347
k192/174p24.scaffold10	13 242	47.21	10A	1	104	3712	87 921	91 529
			10B	1	3643	4425	71 565	72 347
			10C	1	5686	13 151	74 359	81 820
k192/174p24.scaffold5	4257	48.82	5	-1	1	4257	81 296	85 552
k192/174p24.scaffold6	3430	46.15	6	-1	1	3430	88 847	92 273
k192/174p24.scaffold9	11 615	45.82	9A	1	96	7618	87 921	95 452
			9B	1	7669	8097	77 614	78 043
			9C	1	8148	11 615	95 606	99 078
k192/174p24.scaffold3	1677	48.60	3A	1	1	1298	94 856	96 156
			3B	1	1297	1677	96 230	96 610
k192/174p24.scaffold2	79 832	38.97	2A	-1	244	54 703	123 924	178 384
			2B	-1	54705	79 832	98 493	123 628

注：a 方向：“1”表示与参考序列方向相同，“-1”表示与参考序列方向相反；b 参考序列长度按 BAC 克隆长度“0~178384”标注。

长度在 1677~13 242 bp 间的 scaffolds 包括 scaffold1、scaffold3、scaffold4、scaffold5、scaffold6、scaffold8、scaffold9 和 scaffold10 等(图 2 和表 2)，其平均长度约为 5.5 kb。从图 2 可以看出，以上 8 个 scaffolds 在参考序列上出现多处重复比对现象(两个及以上 scaffolds 的内部区段比对到参考序列的相同或相近位置)以及同一个 scaffold 分段比对到参考序列不同位置的现象(图 2 和表 2)。例如，scaffold6 的全部、scaffold9 和 scaffold10 的左端均比对到参考序列上 90 kb 附近，scaffold10 分 3 段比对到参考序列上 70 kb、80 kb、90 kb 左右的 3 个不同区段。同时，本文统计了组装 scaffolds 的 GC 含量：总 GC 含量为 40.12%，*TRG* 基因位点的 8 个 scaffolds 的 GC 含量均高于平均值，且远高于两侧 scaffold7 和 scaffold2 的 GC 含量(36.77%、39.77%)(表 2)。

2.4.2 scaffolds 内部连接的证明

克隆 174P24 组装的 scaffolds 与鸡参考基因组的比对结果表明，scaffold1、scaffold10、scaffold9、scaffold3、scaffold2 等 5 个 scaffolds 的序列分 2~3 段分别比对到参考序列上相近或相距约 10 kb 的不

同位置。针对这 5 个 scaffolds 内部区段连接处的 PCR 结果表明，本研究组装的这 5 个 scaffolds 内部区段的顺序、连接方向和序列正确：scaffold2 内部 PCR 证明了参考序列上 gap6 和 gap6 两侧共约 296 bp 的序列(克隆 174P24 参考序列 123 628~123 924 bp)不存在，scaffold9 内部 PCR 结果表明 9A 和 9C 比对到参考序列上两个区段之间约 154 bp(95 452 bp~95 606 bp)的间隔序列应该为 530 bp，且 9A、9B 和 9C 3 个区段相连；scaffold1、scaffold10 和 scaffold3 的内部分段按图 2 中组装的顺序和方向相连，且序列与组装结果一致。

3 讨论

目前，多数物种 *TCR* 基因位点的序列信息不完整、各基因片段的排列位置不清楚，对人和小鼠 *TCR* 近 30 多年的研究才得到了基本完整的 *TCR* 基因序列和基因片段的排列位置。*TCR* 基因位点内的基因数目众多、排列位置难以确定，且每个家族内的基因片段同源性均很高，这些都导致 *TCR* 基因位点的测序和组装难度远远高于基因组上其他区段。以人 *TRB* 基因位点为例，位于 7 号染色体上约 620 kb 的

区间内集中了 68 个 $TCR\beta$ V 基因片段、2 个 $TCR\beta$ D 基因片段、14 个 $TCR\beta$ J 基因片段和 2 个 $TCR\beta$ C 基因片段; $TCR\beta$ V 区排列在染色体的 5'端, $TCR\beta$ D 区、 $TCR\beta$ J 区和 $TCR\beta$ C 区以 D-J-C 为重复单元位于染色体的 3'端, 两个重复单元的 J 基因片段个数分别为 6 个和 8 个, 整体结构比较复杂^[4,18-21]。与人和小鼠相似, 鸡 TCR 基因位点的 V 基因片段数目众多、彼此间同源性很高, 这使得测序和组装难度也很大。自 21 世纪以来, 鸡 TRG 基因的研究报道较少。目前为止, 已有研究结果表明, 鸡 TRG 基因包括 L 区、V 区、J 区和 C 区; 在 T 细胞分化成熟的过程中, 在特异性重组酶作用下, TRG 胚系基因发生重排, 由分割的、无转录活性的 V 区、J 区和 C 区基因片段连接形成一个完整的、有转录功能的 TRG 重排基因, 它编码约含 300 个氨基酸的多肽^[6,12]。但目前由全基因组鸟枪法测序得到的鸡参考基因组中 TRG 基因位点的序列不完整也不完全准确, 并且基因个数和排列位置等数据依然缺乏。

为了获得更加完整和准确的鸡 TRG 基因位点的序列, 本研究利用 Illumina 公司的 MiSeq 高通量测序平台对鸡 TRG 链的克隆 174P24 进行重新测序和组装。与常规的鸟枪法测序相比, 单个 BAC 克隆的测序和组装有以下优势: (1) BAC 克隆仅插入约 200 kb 的目的片段, 并且插入片段经指纹图谱法在染色体上进行过准确定位。与全基因组测序相比, BAC 克隆测序的基因序列短, 且目的性强, 准确性会更高; (2) 排除 BAC 克隆区间外染色体上同源序列的干扰, 降低了组装难度。综上, BAC 克隆的测序和组装对基因片段数目众多、排列位置难以确定的 TCR 基因位点更具有优势。

本研究对鸡 TRG 链克隆 174P24 的测序和组装分析, 既获得了新的信息, 同时也发现了一些问题。组装 scaffolds 与鸡参考基因组的比对结果显示, 鸡 TRG 基因位点定位于克隆 174P24 中间约 60 kb 的区间内, V 区、J 区和 C 区等基因片段均以基因簇的形式分布。在这个区间内, 总 GC 含量(42.29%)略高于两侧区域(39.13%), 并且分布不均一。而 Illumina 测序在高 GC 和高 AT 区域产生的 reads 数偏少, 这可能也是导致 TRG 基因位点区间测序 reads 丰度比两侧区域偏低的原因。本文运用 Celera 软件、ABYSS 软件和 Phrap 软件等进行过多次尝试, 并把在参考

基因组上定位到克隆 174P24 区间内的 fosmid 文库的两侧末端序列用作辅助定位。结果发现, 运用 ABYSS 软件拼接获得叠连群, 进一步采用 Phrap 软件进行融合, 获得的 scaffolds 最为完整; 本研究中 fosmid 文库末端序列对组装的帮助并不大。

经过多次优化后, 本文获得了克隆 174P24 的基因组草图, 该草图包含 10 个平均长度约为 18 kb 的 scaffolds。与参考序列的比对结果显示, 本文得到的 10 个 scaffolds 较完整地覆盖了鸡 TRG 基因位点及旁侧区域。旁侧区域包括 scaffold7 和 scaffold2, 其位置和序列与参考序列对应。而 TRG 基因位点包含在 8 个平均长度约为 5.5 kb 的 scaffolds 里, 但它们的位置和序列与参考序列并不完全对应。这种不完全对应主要指多个 scaffolds 比对到参考序列的相近位置及一个 scaffold 分段对应到参考序列的不同位置。针对这些 scaffolds 的内部连接处的 PCR 结果显示, 本研究组装的 scaffold1、scaffold2、scaffold3、scaffold9、scaffold10 等 5 个 scaffolds 内部结构正确。从图 2 中可以看出, scaffold2 分两段比对到参考基因组约 100 kb~178 kb 的位置, 且两区段连接处与参考序列的 gap6(长度为 127 bp)相互对应。但该 scaffold 内部 PCR 结果表明, 这个 gap 和其两侧共约 296 bp 的序列(克隆 174P24 参考序列 123 628~123 924 bp 间)不存在。这表明由全基因组高通量测序组装得到的参考序列并不完全准确, 组装时这 178 kb 的区间掺入了染色体上其他区间的序列, 从而出现 PCR 扩增时参考序列并不存在或并不完全准确的情况。而 scaffold9 的 9A、9B 和 9C 3 段分别比对到参考序列上约 80 kb、90 kb 和 100 kb 的 3 个位置, 但 PCR 结果表明这 3 个区段按图 2 中我们组装的顺序和方向连接形成一个完整的 scaffold, 而且 9A 和 9C 对应参考序列上相距约 154 bp(95 452~95 606 bp)的间隔处应该为 530 bp, 这校正和补充了参考序列上不完全准确的序列信息。同样, scaffold10 分成 3 段比对到参考序列上约 70 kb、80 kb 和 90 kb 附近的 3 个区域, 而且前后顺序与组装结果不同。但该 scaffold 内部 PCR 结果表明, 这 3 个区段连接在一起, 而且顺序和方向与图中 scaffold10 一致。另外, 经过 PCR 证明, scaffold1 和 scaffold3 的内部区段按图 2 中组装的顺序和方向相互连接在一起。这表明鸡参考基因组 TRG 基因位点上 V 基因区(70~100 kb)

内的序列信息不完全准确。这种情况可能是由于重复序列的存在、基因结构的复杂性和全基因组测序导致组装时出现了偏差。以上 PCR 结果表明单个 BAC 克隆的测序和组装能得到更加准确的序列信息, 可以对现有参考基因组的序列信息进行校正。此外, 这 10 个 scaffolds 间的顺序和方向虽然根据参考序列进行了初步排序, 但还需要进一步的 PCR 证明。本研究虽然多次尝试过 PCR 扩增, 但经常出现无扩增条带、非特异性扩增、测序时无反应及非特异结合等现象, 而不能获得有用的序列信息。

综上所述, 本研究通过对鸡 *TRG* 基因位点 BAC 克隆的重新测序和组装分析, 得到了鸡 *TRG* 链克隆 174P24 的基因组草图, 并校正了鸡参考基因组的多处序列错误, 为鸡 *TRA/D*、*TRB* 基因位点及其他物种 *TCR* 基因位点等相关方面的研究提供科研思路 and 理论依据, 为基因组复杂区域的研究提供参考和建议。

参考文献

- [1] Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AFA, Wang K, Koop BF, Hood L. Comparative genomics of the human and mouse T cell receptor loci. *Immunity*, 2001, 15(3): 337–349. [\[DOI\]](#)
- [2] Clark SP, Arden B, Kabelitz D, Mak TW. Comparison of human and mouse T-cell receptor variable gene segment subfamilies. *Immunogenetics*, 1995, 42(6): 531–540. [\[DOI\]](#)
- [3] Mackelprang R, Carlson CS, Subrahmanyam L, Livingston RJ, Eberle MA, Nickerson DA. Sequence variation in the human T-cell receptor loci. *Immunol Rev*, 2002, 190(1): 26–39. [\[DOI\]](#)
- [4] 李扬秋. T 细胞受体的研究和应用. 北京: 人民卫生出版社, 2009: 53–55. [\[DOI\]](#)
- [5] Chen CL, Ager LL, Gartland GL, Cooper MD. Identification of a T3/T cell receptor complex in chickens. *J Exp Med*, 1986, 164(1): 375–380. [\[DOI\]](#)
- [6] Sowder JT, Chen CL, Ager LL, Chan MM, Cooper MD. A large subpopulation of avian T cells express a homologue of the mammalian T gamma/delta receptor. *J Exp Med*, 1988, 167(2): 315–322. [\[DOI\]](#)
- [7] Vainio O, Lassila O, Cihak J, Löscher U, Houssaint E. Tissue distribution and appearance in ontogeny of β T cell receptor (TCR2) in chicken. *Cell Immunol*, 1990, 125(1): 254–260. [\[DOI\]](#)
- [8] Tjoelker LW, Carlson LM, Lee K, Lahti J, McCormack WT, Leiden JM, Chen CL, Cooper MD, Thompson CB. Evolutionary conservation of antigen recognition: the chicken T-cell receptor beta chain. *Proc Natl Acad Sci USA*, 1990, 87(20): 7856–7860. [\[DOI\]](#)
- [9] McCormack WT, Tjoelker LW, Stella G, Postema CE, Thompson CB. Chicken T-cell receptor beta-chain diversity: an evolutionarily conserved D beta-encoded glycine turn within the hypervariable CDR3 domain. *Proc Natl Acad Sci USA*, 1991, 88(17): 7699–7703. [\[DOI\]](#)
- [10] Cihak J, Hoffmann-Fezer G, Ziegler-Heibrock HW, Stein H, Kaspers B, Chen CH, Cooper MD, Löscher U. T cells expressing the V beta 1 T-cell receptor are required for IgA production in the chicken. *Proc Natl Acad Sci USA*, 1991, 88(23): 10951–10955. [\[DOI\]](#)
- [11] Göbel TW, Chen CL, Lahti J, Kubota T, Kuo CL, Aebersold R, Hood L, Cooper MD. Identification of T-cell receptor alpha-chain genes in the chicken. *Proc Natl Acad Sci USA*, 1994, 91(3): 1094–1098. [\[DOI\]](#)
- [12] Six A, Rast JP, McCormack WT, Dunon D, Courtois D, Li Y, Chen CH, Cooper MD. Characterization of avian T-cell receptor genes. *Proc Natl Acad Sci USA*, 1996, 93(26): 15329–15334. [\[DOI\]](#)
- [13] Kubota T, Wang JY, Göbel TWF, Hockett RD, Cooper MD, Chen CH. Characterization of an avian (*Gallus gallus domesticus*) TCR gene locus. *J Immunol*, 1999, 163(7): 3858–3866. [\[DOI\]](#)
- [14] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 2004, 432(7018): 695–716. [\[DOI\]](#)
- [15] Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, Zhao FQ, Zhu BL. HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, 2013, 14: 33. [\[DOI\]](#)
- [16] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res*, 2009, 19(6): 1117–1123. [\[DOI\]](#)
- [17] Bonfield JK, Staden R. The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res*, 1995, 23(8): 1406–1410. [\[DOI\]](#)
- [18] Tillinghast JP, Behlke MA, Loh DY. Structure and diversity of the human T-cell receptor beta-chain variable region genes. *Science*, 1986, 233(4766): 879–883. [\[DOI\]](#)
- [19] Rowen L, Koop BF, Hood L. The complete 685-kilobase DNA sequence of the human T cell receptor locus. *Science*, 1996, 272(5269): 1755–1762. [\[DOI\]](#)
- [20] Ikuta K, Ogura T, Shimizu A, Honjo T. A joining-diversity-joining complex generated by inversion mechanism and a variable-diversity complex in the β -chain gene of the human T-cell receptor. *Nucl Acids Res*, 1986, 14(12): 4899–4909. [\[DOI\]](#)
- [21] Tunnacliffe A, Kefford R, Milstein C, Forster A, Rabbitts TH. Sequence and evolution of the human T-cell antigen receptor beta-chain genes. *Proc Natl Acad Sci USA*, 1985, 82(15): 5068–5072. [\[DOI\]](#)