Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Improving the genome assembly of rabbits with long-read sequencing

Yiqin Bai [a], Weili Lin [b], Jie Xu [c], Jun Song [c], Dongshan Yang [c], Y. Eugene Chen [c], Lin Li [a,d], Yixue Li [b,d,e,*], Zhen Wang [b,*], Jifeng Zhang [c,*]

[a] State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China
[b] Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[c] Center for Advanced Models for Translational Sciences and Therapeutics, University of Michigan Medical Center, Ann Arbor, MI, USA
[d] School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[e] Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, China

ARTICLE INFO

Keywords:
Rabbit genomes
Reference assembly
Long-read sequencing
Gap closing

ABSTRACT

The European rabbit (*Oryctolagus cuniculus*) is important as a biomedical model given its unique features in immunity and metabolism. The current reference genome OryCun2.0 established with whole-genome shotgun sequencing was quite fragmented and had not been updated for ten years. In this work, we provided a new rabbit genome assembly UM_NZW_1.0 to improve OryCun2.0 by leveraging the contig lengths based on long-read sequencing and a wealth of available Illumina paired-end sequence data. UM_NZW_1.0 showed a remarkable increase of continuity compared with OryCun2.0, with 5 times longer contig N50 and approximately 75% gaps closed. Many of the closed gaps were overlapped with protein-coding genes or transcriptional features, resulting in an enhancement of gene annotations. In particular, UM_NZW_1.0 presented a more complete landscape of the MHC region and the IGH locus, therefore provided a valuable resource for future researches on rabbits.

1. Introduction

European rabbits (*Oryctolagus cuniculus*) are actively used as animal models for human disease research. Rabbits have a longer life span than that of rodents, suitable for long-term observational and pre-clinical studies, and they are more genetically related to humans than other rodents [1]. Rabbits can carry some pathogens, which help them a more appropriate animal model than rodents in mimicking human host-pathogen interaction and infection [2,3]. Rabbits are frequently used as a laboratory model for several non-infectious conditions as well, including cancer, arthritis, eye disease, cardiovascular disease, lipid metabolism and Alzheimer's disease [4]. Rabbits have a unique feature of lipoprotein metabolism which is like humans but unlike rodents, and are sensitive to a cholesterol diet [5]. Additionally, rabbits are of an intermediate size between rodents and other larger and more costly animal models, easy for handling and able to be reared in a small space. Thus, rabbits can provide more cells and tissues and are more suitable for the industrial production of antibodies [6].

Understanding the genetic diversity and evolution of species is based on accurate sequences and a high-quality assembly of genomes. The previously published genome assembly of *Oryctolagus cuniculus* was generated by whole-genome shotgun sequencing. Initially, rabbit genome assembly OryCun1.0 was constructed using paired-end reads from 4 kb plasmids and 40 kb fosmids at coverage of 2× from a single female rabbit [7]. Based on the sequencing data from OryCun1.0 and additional paired-end reads from 4 kb plasmids, 10 kb plasmids, 40 kb fosmids, and bacterial artificial chromosomes, OryCun2.0 was generated at a coverage of 6–7× [7]. The assembly size was 2.6 Gb, consisting of 3318 scaffolds and 84,024 contigs. OryCun2.0 was anchored to chromosomes with a cytogenetically anchored microsatellite map [7]. Because traditional Sanger sequencing was applied, the accuracy of OryCun2.0 was high and many important inferences could be made. It also served as the reference genome for reads mapping in many whole-genome and transcriptome studies. However, OryCun2.0 contains more than 80 thousand gaps which may lead to incomplete gene annotations or misassembled scaffolds. Unfortunately, in January 2005, all rabbits of this strain were lost in a fire [8], and the assembly remains not updated for more than 10 years.

Recently, the emergence of the third generation sequencing contributes to obtain longer reads with random base errors that can be

[*] Corresponding authors.
E-mail addresses: yxli@sibs.ac.cn (Y. Li), zwang01@sibs.ac.cn (Z. Wang), jifengz@umich.edu (J. Zhang).

https://doi.org/10.1016/j.ygeno.2021.05.031
Received 9 September 2020; Received in revised form 21 May 2021; Accepted 25 May 2021
Available online 27 May 2021

corrected. Pacific Biosciences (PacBio) is one of the platforms based on Single Molecule, Real-Time (SMRT) sequencing technology. Reads obtained are tens of kilobases in length. Although the base error rate of the single-molecule reads is high (~15%), consensus sequences generated at 15-fold coverage show a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates [9]. The recent development and application of the PacBio sequencing technologies have shown considerable promise in improving genome assemblies of some species such as chimpanzee, orangutan and human [10,11]. In this study, we applied the PacBio sequencing to improve the genome assembly of the rabbit. We obtained a more continuous assembly and showed that about 75% of the traceable gaps in OryCun2.0 could be closed.

2. Methods and materials

2.1. Genome sequencing and assembly

Frozen liver tissue from a male New Zealand white (NZW) rabbit was subject to whole-genome sequencing on a PacBio RS II instrument. All animal experiments were performed with the approval of the animal ethics committee of the University of Michigan. Purified DNA was fragmented and end-repaired. Adaptors were ligated to the DNA fragments. Prepared libraries were size-selected to isolate molecules of ~20 kb. Sequencing primers were annealed and complexes with polymerase were generated and the complexes were sequenced. HGAP4 [12,13] was used for the contig assembly UM_NZW_0.0. The assembly process proceeded in two rounds. The first round of assembly involved the selection of seed reads or the longest reads in the dataset. All shorter reads were aligned to the seed reads, in order to generate consensus sequences with high accuracy. In the next round, the error corrected reads were aligned to each other and assembled into genomic contigs.

2.2. Intra-scaffold gap closing

UM_NZW_0.0 was used to close the intra-scaffold gaps of OryCun2.0 with LR_Gapcloser [14]. "-a 0.75" option was applied and 3 iterations were run as the default.

2.3. Scaffold extending

The gap-closed OryCun2.0 and UM_NZW_0.0 were aligned using MUMmer [15] v4.0.0.beta2 with parameters "-L 1500". Output was processed with delta-filter to remove alignments due to repeats and duplicates, and "-l 3000" was used for filtering out alignments between non-syntenic repeats. Based on the filtered alignments, quickmerge [16] was conducted with "-hco 5.0 -c 1.5 -l 500000" to merge the two assemblies.

2.4. Assembly polishing

Ten Illumina paired-end whole-genome sequencing samples from the database RabGTD [17] were mapped to the extended assembly with BWA-MEM [18]. Two rounds of Pilon [19] were conducted to polish the assembly with alignments mapped in proper pairs. The polished assembly was named UM_NZW_1.0. Freebayes [20] was adopted to call the variants. Base coverage was obtained using samtools [21] depth.

2.5. Relocating gaps

We identified the inter-scaffold gaps with MUMmer [15], and identified intra-scaffold gaps with further analysis using BLAST [22]. UM_NZW_1.0 and OryCun2.0 were aligned with MUMmer4 [15] choosing the option "-l 100 -c 500 -L 1500 –maxmatch". The alignments were filtered with "delta-filter −1" and converted to coordinates with "show-coords -rclTH". BLAST [22] was adopted to relocate the potential intra-scaffold gap segments with the up- and downstream 2000-bp sequences. Unique BLAST results with the highest scores were kept to extract corresponding sequences of these gap regions in UM_NZW_1.0.

2.6. Assembly comparison

To compare the continuity of assemblies, we used QUAST [23] with the option "–eukaryote –scaffolds" and no reference assembly. To compare the accuracy of assemblies, FRC_align [24] was used with the ten Illumina whole-genome sequencing samples from RabGTD [17] and the options "–genome-size 2737000000 –CEstats-PE-min -10 –CEstats-PE-max 10". A standalone BUSCO [25] analysis using the mammalia_odb9 database was adopted to compare the completeness of the assemblies.

2.7. Annotation

RepeatMasker [26] version open-4.0.6 was run to annotated the interspersed repeats and low complexity DNA sequences in UM_NZW_1.0 with "-species rabbit". Three approaches were adopted for gene annotation. 1) The RefSeq transcripts of OryCun2.0 were mapped to UM_NZW_1.0 using minimap2 [27] with the option "-ax splice". A GTF file marking the locations of all transcripts and exons in UM_NZW_1.0 was then extracted from the alignment file. 2) An Illumina short-read RNA-Seq dataset of ten rabbit issues from SRA PRJNA78323 [7] were aligned to UM_NZW_1.0 with STAR [28]. Cufflinks [29] (2.2.1) was used to generate potential transcripts and exons. 3) As for *de novo* gene annotation, AUGUSTUS [30] (3.2.3) was used with the setting "–species human –protein=on –introns=on –start=on –stop=on –cds=on –codingseq=on". Eventually, EVM [31] (1.1.1) was applied to integrate these annotation results.

The transcription in gap regions was validated based on the RNA-Seq dataset of ten tissues. The depth for each base was obtained with samtools [21] depth.

2.8. Identifying MHC region

The MHC region on Chromosome 12 of UM_NZW_1.0 was reannotated with GeneWise [32] using homologous protein sequences from human and mouse. Protein sequences of human and mouse were downloaded from RefSeq, and initial alignment was performed using genblasta [33] "-e 1e-5". Only target regions with alignment coverage greater than 0.7 were preserved for the GeneWise analysis. The MHC subregions I-III were delineated according to humans [34].

2.9. Immunoglobulin gene mapping

Rabbit nucleotide sequences of the V, J and D genes of IGH/IGK/IGL were obtained from "F + ORF+all P" sets in the IMGT [35]/GENE-DB reference directory, and the nucleotide sequences of the C genes were fetched from the IMGT/GENE-DB query page. The sequences were aligned to UM_NZW_1.0 with BLAST [22], of which D genes were mapped using "-task blastn-short" due to their short sequence length. For each D, J and C gene, only the BLAST result with the highest score was kept. As for V genes, all results with alignment coverage greater than 0.9 were preserved. In the case of overlapping target regions, the blast result with the smallest e-value was chosen. We specifically assembled the scaffolds containing IGH genes by using quickmerge [16] with a looser cutoff "-l 3000". The final results were manually curated by comparing with BAC clone sequences [36].

3. Results and discussion

3.1. Sequencing data set

We performed whole-genome sequencing with frozen tissue from a

New Zealand white (NZW) rabbit using the PacBio technology. In total, 18 PacBio Sequel SMRT cells generated 12.12 million subreads containing 116.56 Gb with a subread N50 of 16.79 kb (Fig. 1a, b, c). 81.17% of the subreads had at least one alignment to the current rabbit reference genome (OryCun2.0) and the mean sequencing depth was 40 × (Fig. 1d). 96.85% of the alignments had an identity rate over 70% and the mean identity rate was 82.06%(Fig. 1e). No length bias was observed in the error rate (Fig. 1f).

3.2. Genome assembly and improvement

We carried out a *de novo* assembly of the PacBio data set with HGAP4 [12,13] and obtained the new assembly UM_NZW_0.0. This assembly comprised 2.62 Gb, including 11,725 contigs with a contig N50 size of 506.82 kb, which was much longer than that of the OryCun2.0 contigs (64.6 kb) [7]. The GC content was 43.88%, which was similar to that of OryCun2.0 (43.75%). The average contig length reached 223.60 kb, and the maximum contig length was 4318.74 kb. Using MUMmer4 [15] for whole-genome alignment, 98.98% contigs of UM_NZW_0.0 were aligned to OryCun2.0, where 99.17% alignments were 1-to-1.

Over 80 thousand gaps remain unresolved in OryCun2.0. Using contig sequences from UM_NZW_0.0, we closed part of these intra-scaffold gaps with LR_Gapcloser [14]. Then the gap-closed version of OryCun2.0 and UM_NZW_0.0 were merged using quickmerge [16] to close potential inter-scaffold gaps. To ensure the base accuracy of the assembly, we conducted two rounds of Pilon [19] with Illumina whole-genome sequencing data totaling 130× from 10 additional NZW rabbits, resulting a chimeric assembly UM_NZW_1.0 of 2.84 Gb. After the two rounds of polishing the number of variants called from the genome tended to be stable, suggesting this assembly combined high frequency variants from the 10 NZW rabbits (Table 1).

Table 1
Variants called after polishing in UM_NZW_1.0. Only sites with ≥ 3× coverage and SNPs and INDELs of which frequency > 0.5 are considered.

	No. variants	Total sites	Identity
UM_NZW_1.0 (without Pilon)	9,043,297	2,760,894,201	99.67%
UM_NZW_1.0 (Pilon × 1)	4,572,897	2,760,579,762	99.83%
UM_NZW_1.0 (Pilon × 2)	4,354,430	2,760,899,422	99.84%

3.3. Assembly evaluation

UM_NZW_1.0 consisted of 3159 scaffolds, including 21,334 contigs, among which the largest scaffold reached 208.59 Mb. Although the scaffold N50 was 111.64 Mb and similar to OryCun2.0, the contig N50 was 337.73 kb, significantly greater than 64.64 kb in OryCun2.0. Because only contigs consistent between UM_NZW_0.0 and OryCun2.0 were incorporated, the contig N50 of UM_NZW_1.0 was lower than that of UM_NZW_0.0. The GC content of UM_NZW_1.0 was 43.83%, comparable with 43.75% of OryCun2.0. Total interspersed repeats were annotated to be 41.13%, slightly higher than 39.97% in OryCun2.0 (Table 2).

We then compared the accuracy of the assemblies OryCun2.0 and UM_NZW_1.0 with the feature response curve (FRC) [24], which is a metric designed to captures the trade-offs between quality and contig size. Features are areas on the assembly that show indications of assembly errors based on the alignment of short sequencing reads [37]. A steeper curve indicated an assembly of higher quality. The FRCs for UM_NZW_1.0 and OryCun2.0 diverged at a higher feature threshold, with UM_NZW_1.0 being steeper, indicating UM_NZW_1.0 achieved a better performance than OryCun2.0 (Fig. 2a).

To compare the completeness of OryCun2.0 and UM_NZW_1.0, we used benchmarking universal single-copy orthologs (BUSCO) [25] analyses against the dataset *mammalia_odb*. BUSCO analysis suggested that the new assembly was more complete, with 92.3% complete BUSCOs for UM_NZW_1.0, a marked improvement compared with the 88.3% in OryCun2.0 (Fig. 2b).

3.4. Annotation of closed gaps

With all the 80,789 intra-scaffold gaps in OryCun2.0, 74,055 could be unambiguously located on UM_NZW_1.0 and were used for further

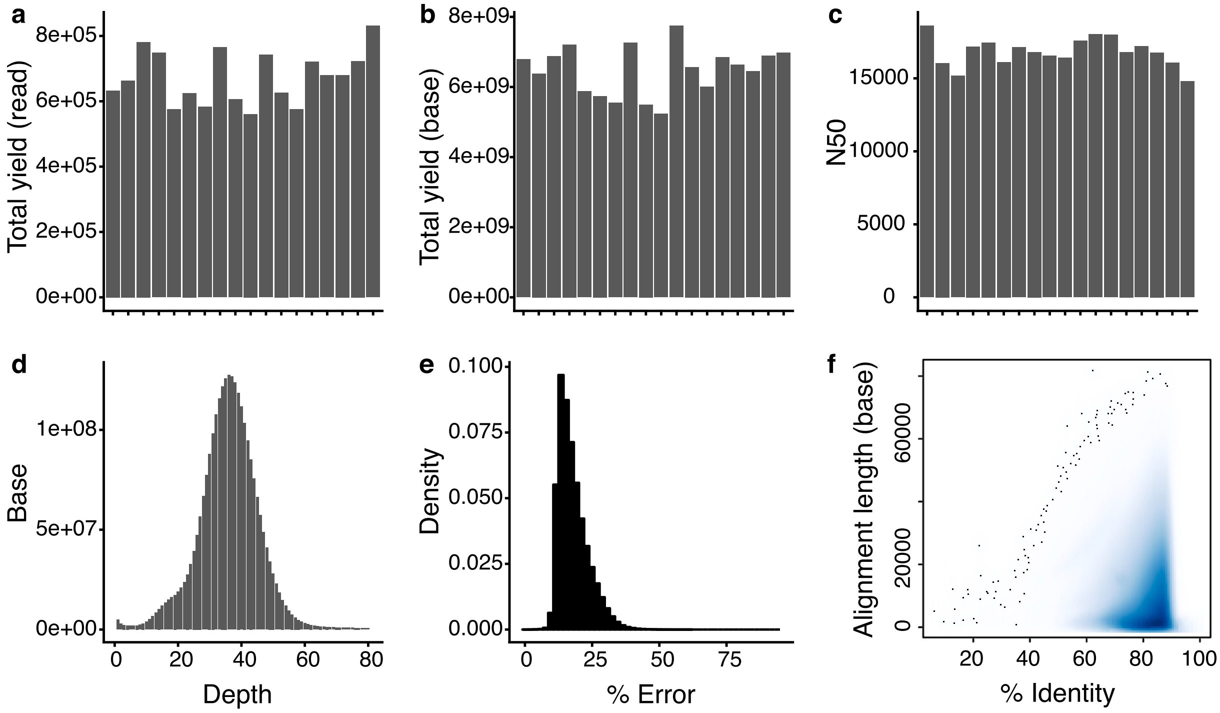

Fig. 1. Summary of data set. (a) Total yield read, (b) total yield base, and (c) read length N50s per flow cell. (d) Sequencing depth of each position. (e) Alignment identity to OryCun2.0. (f) Alignment identity compared to alignment length. No length bias was observed.

Table 2
Metrics of UM_NZW_1.0 compared with OryCun2.0.

	No. scaffold	No. contig	Scaffold N50	Contig N50	GC %	Repeat %
UM_NZW_1.0	3159	21,334	111,644,748	337,730	43.83	41.13
OryCun2.0	3241	84,030	111,795,807	64,648	43.75	39.97

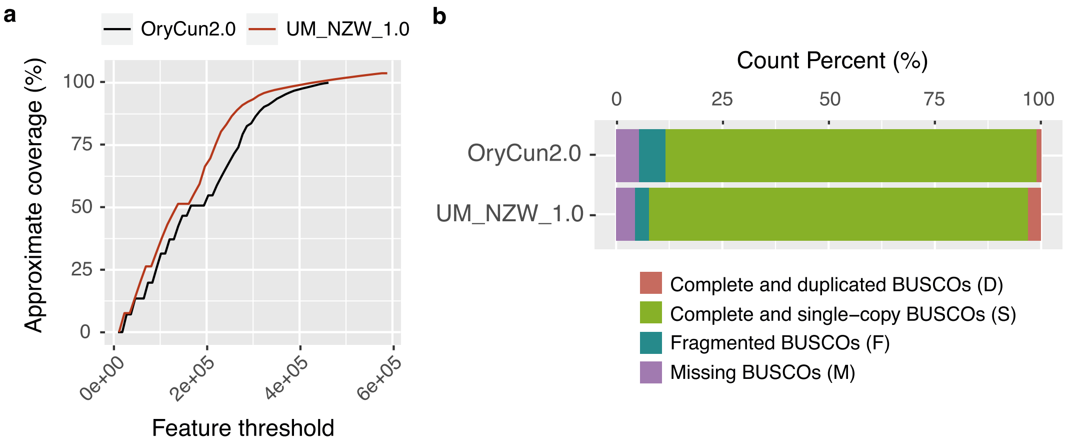

Fig. 2. Quality of UM_NZW_1.0 compared with OryCun2.0. (a) Feature response curves for UM_NZW_1.0 and OryCun2.0. The FRCs were generated using FRC_align and plotted in R. (b) Comparison of the completeness of gene annotation in UM_NZW_1.0 and OryCun2.0, as a percentage of 4104 mammalian genes from BUSCO.

analysis. 75.3% (55,829/74,055) of these gaps were completely closed. The observed length of closed gap regions showed good concordance with the corresponding estimated length in OryCun2.0. For 53,840 spanned gaps with a specified size in OryCun2.0, the observed sizes were in agreement with the estimated sizes (correlation = 0.99, mean ratio between observed and estimated size = 1.08). Among them 41,075 were completely closed (Fig. 3a). Although a few gaps still left, 12.4% of them showed less than half of the gap length estimated in OryCun2.0 (Fig. 3b). Therefore, we considered these gaps as partially closed. There were also 20,215 unspanned gaps with unknown size in OryCun2.0 that were closed with a mean observed size of 118.9 bp and the maximum observed size of 6601 bp. The total length of 74,055 gap regions was 106.7 Mb. Among the closed gaps of 35.2 Mb the proportion of repeat sequences reached 50.9%, higher than that of genome average (Fig. 3c). Most of the closed gaps coincided with the presence of LINE elements, which were long repetitive elements that were difficult to span with short reads.

Besides these intra-scaffold gaps, there were also 292 regions of 23.0 Mb that linked multiple scaffolds of OryCun2.0 into larger scaffolds in UM_NZW_1.0, which we here referred as inter-scaffold gaps. Except for 4 partially closed regions, other 288 inter-scaffold gaps were completely closed in UM_NZW_1.0. 36.9% of these closed inter-scaffold gaps were annotated as repeat sequences(Fig. 3c).

To evaluate the functional potentials of closed gap regions in UM_NZW_1.0, we collected a public RNA-Seq dataset including 10 different tissues and aligned them to UM_NZW_1.0. The mean ratio of uniquely mapped reads was about 84.2%. 24.6% of the intra-scaffold gap regions consisted of RNA-Seq reads with an average depth of 15× and the mean base coverage in these gap regions was 25.9%(Fig. 3d). As for inter-scaffold gap regions, 83.2% were mapped with RNA-Seq reads, and the mean base coverage was 20.7% with an average depth of 44 × (Fig. 3e). These facts indicated that the closed gap regions contained a lot of transcriptional signals which were missing in OryCun2.0.

According to the RefSeq annotation of OryCun2.0, 21,605 closed gaps were overlapped with 7911 genes and 43 closed gaps were overlapped with 45 exons (Fig. 3f). We re-annotated these genes with EVM [31] and manually checked the gaps in exons. For example, THBD (thrombomodulin) was partial on OryCun2.0, and it was a validated gene (NM_001082144.1) with the longest overlapped region between an exon and a closed gap in UM_NZW_1.0. We aligned the gene sequences from OryCun2.0, UM_NZW_1.0 and the human and mouse genome assembly using Clustal Omega [38] (Fig. 3g). The closed gap overlapped with the exon was 932 bp and was similar to the homologous segment. Similar gap closure on an exon was also found in APOE (Fig. 3h), which is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. APOE was partial in OryCun2.0 with a gap estimated to be of 860 bp at the end of its last exon. In UM_NZW_1.0, the gap was closed and annotated to be an exon, which was also similar to the homologous segment in human and mouse, offering an intact sequence of APOE in the rabbit.

3.5. *MHC region*

The European rabbit has been used as a laboratory animal model in immunology. MHC region is considered to play an essential role in the vertebrate adaptive immune system, and is responsible for the recognition and presentation of antigens. The MHC region is difficult to resolve using short reads due to its repetitive and highly polymorphic nature, and recent efforts to apply long-read sequencing to this problem have shown promise. Therefore, the continuity and completeness of the MHC region are usually adopted to illustrate the benefits of long-read sequencing. Rabbit MHC was mapped to Chromosome 12 [39]. Considering a standardized nomenclature for the rabbit MHC Class I and Class II genes has not yet been established [40], we located the rabbit classical MHC genes on Chromosome 12 of UM_NZW_1.0 according to the homologous genes of human and mouse as delineated by previous studies [34,41] (Fig. 4a). The rabbit classical MHC region was estimated to extend 3.1 Mb, comparable with 3.4 Mb in human. The MHC subregion I and II came after subregion III, which was different from the MHC subregion organizations in both human and mouse and consistent with the previous conclusion [39]. The interval region between subregion I and III extending about 1 Mb was homologous to the extended subregion I in humans, including MOG, olfactory receptor family genes, butyrophilin family genes and genes encoding zinc finger protein.

94 out of 131 gaps left in the MHC region of OryCun2.0 were closed in UM_NZW_1.0, including 36 in subregion I, 6 in subregion II, 26 in subregion III, 25 between subregion I and III and 1 between subregion II and I. The mean ratio between observed and expected length of the

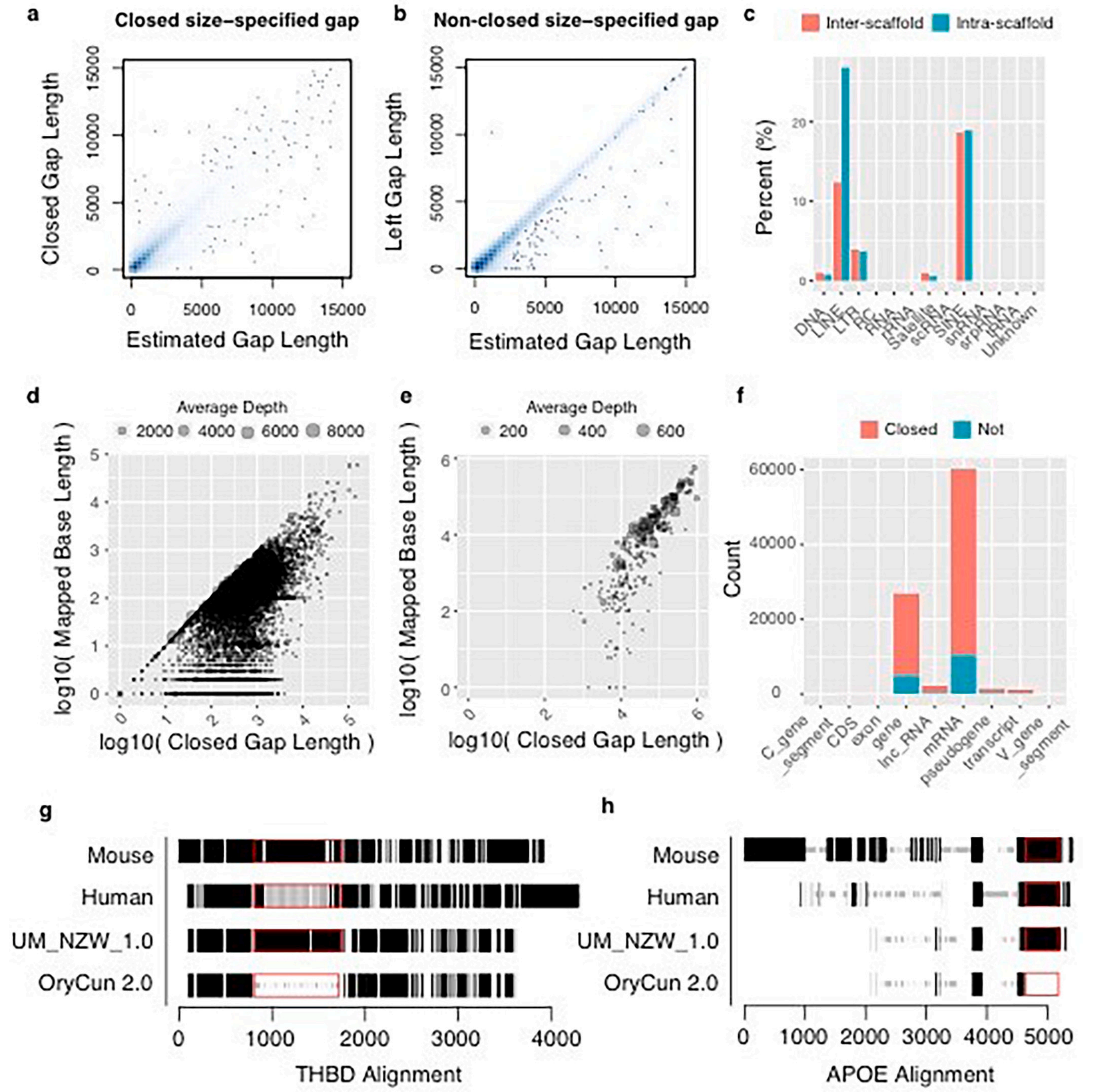

Fig. 3. Statistics and annotations in gap regions. (a) Estimated gap length in OryCun2.0 compared with the closed gap length of spanned gaps that were closed in UM_NZW_1.0. (b) Estimated gap length in OryCun2.0 compared with the left gap length of non-closed spanned gaps. (c) Ratio of different types of repeat elements in closed gap sequences. (d,e) RNA-Seq reads mapped to intra-scaffold(d) and inter-scaffold(e) gap regions. Positions of 0 depth were not shown. (f) RefSeq annotation elements overlapped with gap regions. (g,h) Rabbit gene THBD(g)/APOE(h) in UM_NZW_1.0 was aligned with the gene in OryCun2.0 and its homologous gene in human and mouse. Gray lines denoted the aligned bases in the gene region and black regions denoted the bases on exons. Red blocks denoted the gap region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

closed spanned gap regions was 0.95 (Fig. 4b). 13 of these gaps were overlapped with exons in UM_NZW_1.0. In particular, RLA-DMA (NM_001190432.1), a member of MHC Class II genes, was a validated gene with the longest overlapped region between an exon and a closed gap. The closed gap segment was estimated as 819 bp and 131 bp was overlapped with an exon (Fig. 4c). The completed exon was similar to that in the human orthologous gene.

3.6. *Immunoglobulin gene locus*

Immunoglobulins play an essential role in the adaptive immune system. Complete knowledge of gene sequences encoding the kappa (IGK), lambda (IGL) and heavy chains (IGH) will lead to a better understanding of how and why rabbits produce antibodies of high specificity and affinity.

The IGK genes were located on Chromosome 2 of UM_NZW_1.0,

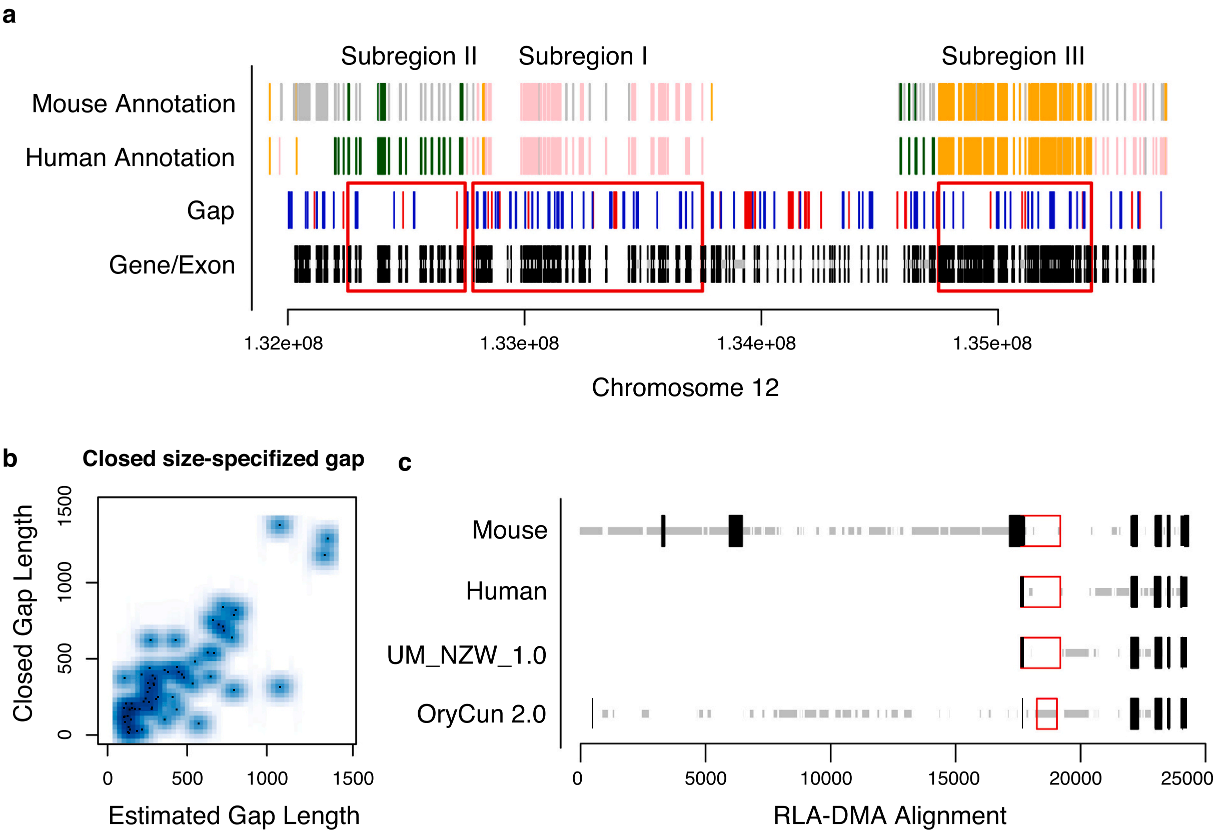

Fig. 4. Structure, gap closure ratio and homology alignment in MHC. (a) MHC region locus in UM_NZW_1.0. Gray lines denoted the transcript regions and black regions denoted the exons. Closed and non-closed gaps were denoted as blue and red respectively. Human and mouse proteins were used to annotate genes in this region with GeneWise. And the genes were classified into MHC subregion I/II/III according to previous classification in human and mouse. Genes annotated with proteins from MHC subregion I/II/III were marked as pink/green/orange respectively, and gray denoted that no classification information was mapped. Red blocks showed the estimated regions of MHC subregion I/II/III in rabbits. (b) Estimated gap length in OryCun2.0 compared with the closed length of spanned gaps that were closed and located in the MHC region. (c) Reannotated RLA-DMA in UM_NZW_1.0 was aligned with RLA-DMA in OryCun2.0 and its homologous gene in human and mouse. Gray lines denoted the aligned bases in the gene region and black regions denoted the bases on exons. Red blocks denoted the gap region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

including 89 IGKV, 8 IGKJ and 2 IGKC genes (Fig. 5a). The genomic organizations of the IGK genes was in agreement with the previous report [40]. 13 out of 18 gaps in this region were closed, though none of them overlapped with coding sequences. Also similar to the previous study [40], the IGL genes were aligned to Chromosome 21 of UM_NZW_1.0, with 43 IGLV, 4 IGLJ and 4 IGLC genes. 2 gaps overlapped with IGLV sequences were closed (Fig. 5b). The complete structure of IGH remained unsolved in the previous study, and the locus remained quite fragmented in OryCun2.0. By merging these scaffolds with contigs assembled from long-read sequencing, we found most of the IGH genes could be mapped to an extended scaffold (Fig. 5c), which bridged original scaffolds NW_003160813.1, NW_003159763.1 and NW_003160725.1 in OryCun2.0. Two of the junctions of the scaffolds could be validated by independent BAC clone sequences of rabbit IGH locus (AY386696.1 and AY386697.1) [36]. The scaffold revealed a complete genomic structure of V-D-J-C genes spanning 1.5 Mb, including 54 IGHV genes, 8 IGHD genes, 6 IGHJ genes and 6 IGHC genes (IGHA10, IGHA3, IGHA4, IGHE, IGHG, IGHM). Another 5 IGHC genes (IGHA13, IGHA14, IGHA9, IGHA8 and IGHA12) were found on a separated scaffold mainly assembled from long-read sequencing. The scaffold contained NW_003161178.1 in OryCun2.0, on which only IGHA13 was reported [8], and part of NW_003160813.1 and NW_003161893.1. These two major scaffolds along with other 6 scaffolds containing 90 IGHV genes made the IGH presentation the most intact version so far.

4. Conclusion

We provided a new rabbit genome assembly UM_NZW_1.0 by leveraging the contig lengths provided by long reads, a wealth of available data of Illumina paired-end sequence, and the current reference assembly OryCun2.0. The new genome assembly UM_NZW_1.0 was more continuous than OryCun2.0, with its contig N50 of 337,730 bp and 18,178 gaps. FRC analysis suggested that the new assembly remained qualified besides its improved continuity. The public RNA-Seq data were aligned to UM_NZW_1.0, suggesting many of the closed gaps comprising transcriptional signals. The improvement in the number of complete BUSCO genes also indicated that the new assembly UM_NZW_1.0 provided better gene annotations. Collectively, these metrics demonstrated that UM_NZW_1.0 greatly enhanced the continuity and completeness of the rabbit genome while maintained high accuracy.

The increased read length obtained with PacBio sequencing technology enabled us to resolve complex genomic structures like the MHC region and immunoglobulin loci. We located the MHC genes on UM_NZW_1.0 with homologous annotation and identified the classical MHC region expanding about 3.1 Mb. Most gaps left within these regions were closed. Furthermore, we provided a more intact genomic structure of immunoglobulin genes, especially for IGH. Previous studies presented quite fragmental segments containing IGH genes [8], while our study offered merged scaffolds consisting of several unlocated segments, on which the V-D-J-C arrangement of IGH genes could be resolved. The enhanced immune gene content will contribute to better utilization of rabbits in biomedical research and antibody production.

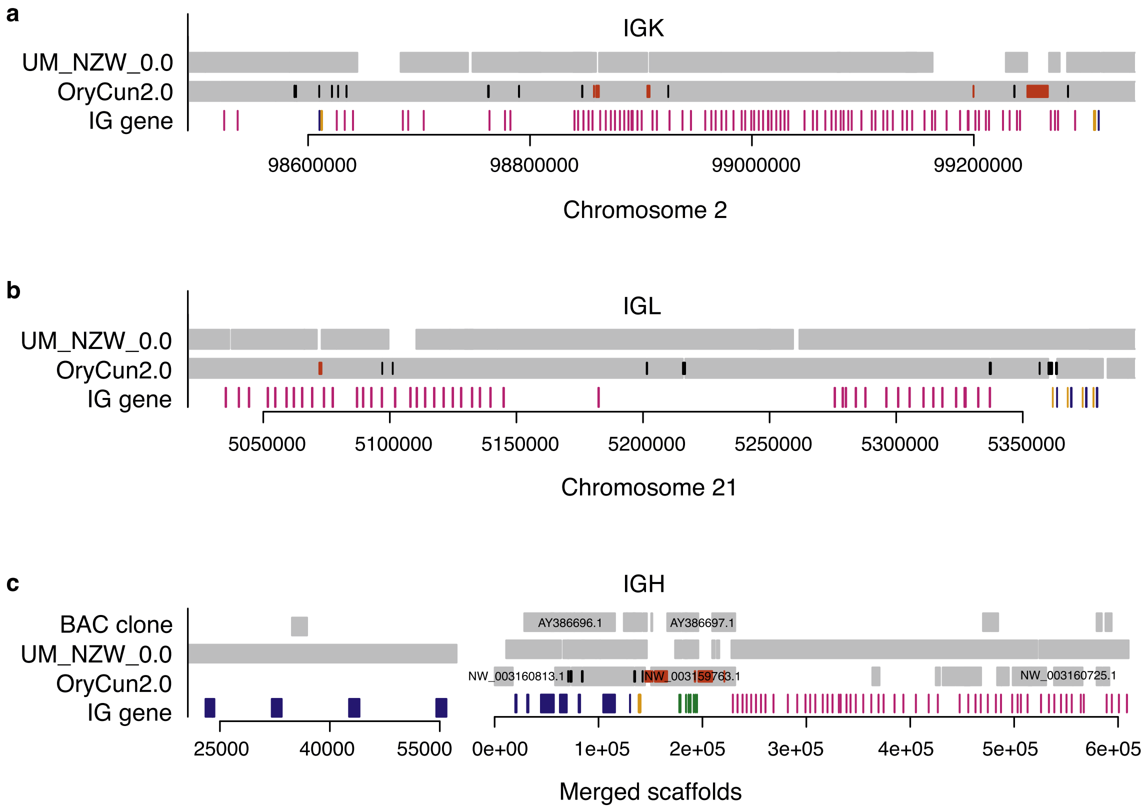

Fig. 5. Identification of immunoglobulin genes. Immunoglobulin gene locus of IGK(a), IGL(b) and IGH(c). V/D/J/C genes were denoted as pink/green/orange/blue blocks. Scaffolds in OryCun2.0 and UM_NZW_0.0 were aligned with MUMmer4. Closed gaps were denoted as black in OryCun2.0 scaffolds, while gaps not closed were denoted as red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Data availability

This assembly has been deposited at NCBI GenBank under the accession VIYN00000000. The raw sequencing data have been submitted to the NCBI SRA under accession number PRJNA274594.

Acknowledgements

This work was supported the National Key R&D Program of China [2016YFC0901704, 2017YFC1201200, 2017YFA0505500], the National Natural Science Foundation of China [32070570], the Chinese Academy of Sciences [KFJ-STS-QYZD-126, ZDBS-SSW-DQC-02] and the Youth Innovation Promotion Association CAS [2017325].

References

[1] D. Graur, L. Duret, M. Gouy, Phylogenetic position of the order lagomorpha (Rabbits, Hares and Allies), Nature. 379 (6563) (1996) 333–335.

[2] J. James, L. Martin, M. Krenz, C. Quatman, F. Jones, R. Klevitsky, et al., Forced expression of α-myosin heavy chain in the rabbit ventricle results in cardioprotection under cardiomyopathic conditions, Circulation. 111 (18) (2005) 2339–2346.

[3] J. Zschaler, D. Schlorke, J. Arnhold, Differences in innate immune response between man and mouse, Crit. Rev. Immunol. 34 (2014) 5.

[4] P.J. Esteves, J. Abrantes, H.-M. Baldauf, L. BenMohamed, Y. Chen, N. Christensen, et al., The wide utility of rabbits as models of human diseases, Exp. Mol. Med. 50 (5) (2018) 66.

[5] J. Fan, S. Kitajima, T. Watanabe, J. Xu, J. Zhang, E. Liu, et al., Rabbit models for the study of human atherosclerosis: from pathophysiological mechanisms to translational medicine, Pharmacol. Ther. 146 (2015) 104–119.

[6] J. Weber, H. Peng, C. Rader, From rabbit antibody repertoires to rabbit monoclonal antibodies, Exp. Mol. Med. 49 (3) (2017) e305, https://doi.org/10.1038/emm.2017.23.

[7] M. Carneiro, C.-J. Rubin, F. Di Palma, F.W. Albert, J. Alföldi, A.M. Barrio, et al., Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication, Science. 345 (6200) (2014) 1074–1079.

[8] E.M. Gertz, A.A. Schäffer, R. Agarwala, A. Bonnet-Garnier, C. Rogel-Gaillard, H. Hayes, et al., Accuracy and coverage assessment of Oryctolagus cuniculus (rabbit) genes encoding immunoglobulins in the whole genome sequence assembly (OryCun2. 0) and localization of the IGH locus to chromosome 20, Immunogenetics. 65 (10) (2013) 749–762.

[9] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, et al., Real-time DNA sequencing from single polymerase molecules, Science. 323 (5910) (2009) 133–138.

[10] D. Gordon, J. Huddleston, M. Chaisson, C.M. Hill, Z.N. Kronenberg, K.M. Munson, et al., Long-read sequence assembly of the gorilla genome, Science. 352 (2016) 6281.

[11] Z.N. Kronenberg, I.T. Fiddes, D. Gordon, S.C. Murali, S. Cantsilieris, O.S. Meyerson, et al., High-resolution comparative analysis of great ape genomes, Science. 360 (2018) 6393.

[12] C. Chin, P. Peluso, F.J. Sedlazeck, M. Nattestad, G.T. Concepcion, A. Clum, et al., Phased diploid genome assembly with single-molecule real-time sequencing, Nat. Methods 13 (12) (2016) 1050–1054.

[13] C. Chin, D. Alexander, P. Marks, A. Klammer, J.P. Drake, C. Heiner, et al., Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Nat. Methods 10 (6) (2013) 563–569.

[14] G.-C. Xu, T.-J. Xu, R. Zhu, Y. Zhang, S.-Q. Li, H.-W. Wang, et al., LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly, GigaScience 8 (1) (2019), giy157.

[15] G. Marçais, A.L. Delcher, A.M. Phillippy, R. Coston, S.L. Salzberg, A. Zimin, MUMmer4: a fast and versatile genome alignment system, PLoS Comput. Biol. 14 (1) (2018), e1005944.

[16] M. Chakraborty, J.G. Baldwin-Brown, A.D. Long, J. Emerson, Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage, Nucleic Acids Res. 44 (19) (2016), e147-e.

[17] L. Zhou, Q. Xiao, J. Bi, Z. Wang, Y. Li, RabGTD: a comprehensive database of rabbit genome and transcriptome, Database. 2018 (2018).

[18] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv (2013). https://arxiv.org/pdf/1303.3997v2.pdf, 1303.3997.

[19] B.J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One 9 (11) (2014), e112963.

[20] E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, arXiv (2012). https://arxiv.org/pdf/1207.3907v2.pdf, 1207.3907.

[21] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence alignment/map format and SAMtools, Bioinformatics. 25 (16) (2009) 2078–2079.

[22] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

[23] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics. 29 (8) (2013) 1072–1075.

[24] F. Vezzi, G. Narzisi, B. Mishra, Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons, PLoS One 7 (12) (2012), e52210.

[25] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics. 31 (19) (2015) 3210–3212.

[26] A. Smit, R. Hubley, P. Green, RepeatMasker, available at, http://www.repeatmasker.org.

[27] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics. 34 (18) (2018) 3094–3100.

[28] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics. 29 (1) (2013) 15–21.

[29] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 5.

[30] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, et al., Using native and syntenically mapped cDNA alignments to improve de novo gene finding, Bioinformatics 24 (5) (1 March 2008) 637–644, https://doi.org/10.1093/bioinformatics/btn013.

[31] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments, Genome Biol. 9 (1) (2008) R7.

[32] E. Birney, M. Clamp, R. Durbin, GeneWise and genomewise, Genome Res. 14 (5) (2004) 988–995.

[33] R. She, S.C. Chu, K. Wang, J. Pei, N. Chen, genBlastA: enabling BLAST to identify homologous gene sequences, Genome Res. 19 (2008) 1.

[34] R. Horton, L. Wilming, V. Rand, R.C. Lovering, E.A. Bruford, V.K. Khodiyar, et al., Gene map of the extended human MHC, Nat. Rev. Genet. 5 (12) (2004) 889–899.

[35] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, et al., IMGT®, the international ImMunoGeneTics information system®, Nucleic Acids Res. 37 (suppl_1) (2009). D1006-D12.

[36] F. Ros, J. Puels, N. Reichenberger, W.V. Schooten, R. Buelow, J. Platzer, Sequence analysis of 0.5 Mb of the rabbit germline immunoglobulin heavy chain locus, Gene 330 (2004) 49–59.

[37] A.M. Phillippy, M.C. Schatz, M. Pop, Genome assembly forensics: finding the elusive mis-assembly, Genome Biol. 9 (3) (2008) 1–13.

[38] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (1) (2011) 539.

[39] C. Rogel-Gaillard, F. Piumi, A. Billault, N. Bourgeaux, J.-C. Save, C. Urien, et al., Construction of a rabbit bacterial artificial chromosome (BAC) library: application to the mapping of the major histocompatibility complex to position 12q1. 1, Mamm. Genome 12 (3) (2001) 253–255.

[40] A. Pinheiro, F. Neves, A.L. De Matos, J. Abrantes, W. van der Loo, R. Mage, et al., An overview of the lagomorph immune system and its genetic diversity, Immunogenetics. 68 (2) (2016) 83–107.

[41] P. Hurt, L. Walter, R. Sudbrak, S. Klages, I. Müller, T. Shiina, et al., The genomic sequence and comparative analysis of the rat major histocompatibility complex, Genome Res. 14 (4) (2004) 631–639.