The immunogenomic landscape of cattle

Ting-Ting Li[1,2,7], Tian Xia[1,2,7], Jia-Qi Wu[1,2,7], Hao Hong[1,2], Zhao-Lin Sun[3], Ming Wang[4,5], Fang-Rong Ding[4], Jing Wang[3], Shuai Jiang[1,2], Jin Li[1,2], Jie Pan[1], Guang Yang[3], Jian-Nan Feng[3], Yun-Ping Dai[4], Xue-Min Zhang[1,2,6], Tao Zhou[1,2,*] & Tao Li[1,2,6,*]

Affiliations:

[1]State Key Laboratory of Proteomics, National Center of Biomedical Analysis, 27 Tai-Ping Road, Beijing 100850, China.

[2]Nanhu Laboratory, Jiaxing, Zhejiang Province 314002, China.

[3]State Key Laboratory of Toxicology and Medical Countermeasures, Beijing Institute of Pharmacology and Toxicology, Beijing100850, China.

[4]State Key Laboratories for Agrobiotechnology, College of Biological Sciences, China Agricultural University, No.2 Yuanmingyuan Xilu, Beijing 100193, China.

[5]College of Animal Science and Technology, China Agricultural University, No.2 Yuanmingyuan Xilu, Beijing 100193, China.

[6]School of Basic Medical Sciences, Fudan University, Shanghai 200032, China.

[7]These authors contributed equally

*Correspondence: tli@ncba.ac.cn (T.L.), tzhou@ncba.ac.cn (T.Z.)

ABSTRACT

Here, we report the *de novo* assembly of a cattle genome using ultra-long-read nanopore sequencing in conjunction with other advanced technologies. The assembled genome contains only 145 contigs (N50 ~ 74.0 Mb). Compare to the current reference cattle genome ARS-UCD1.2, 154 gaps are closed, and 467 scaffolds are further placed in our

assembly. Importantly, except two remained gaps in the T-cell receptor α/δ (TRA/TRD) region, the gene loci of other TRs and immunoglobulins (IGs) are seamlessly assembled and exquisitely annotated. With the characterization of 258 IG genes and 626 TR genes that distributed in seven genomic loci, we illustrate the highest immune gene diversity in mammals to our knowledge. Moreover, the gene structures of major histocompatibility complex (MHC) are integrally depicted with properly phased haplotypes. Thus, our work not only reports a cattle genome with the most continuity and completeness, but also provide a comprehensive view of the complex immune-genome.

KEYWORDS

Cattle genome assembly, ultra-long sequencing, T cell receptor, immunoglobulin, MHC, Immune-genome

INTRODUCTION

The immune system possesses the biggest source of genetic variation, and the prodigious diversity and complexity of the immune system ensure the host to precisely distinguish non-self from self and to effectively response to the persistent but unpredictable environmental challenges (Schultze and Aschenbrenner, 2021; Sette and Crotty, 2021). At the DNA level, T cells and B cells represent the typical examples of

the genetic variations (Imkeller and Wardemann, 2018; Kumar et al., 2018). The somatic rearrangement of the V, (D) and J gene segments from the TR or B cell receptor (BCR, also known as IG) gene loci give rise to millions of different T- and B-cell receptors (Nielsen and Boyd, 2018). Each T- or B-cell, as characterized by the uniquely expressed TR or IG gene, can response to a specific antigen. The TR and IG genes, in together, encode a major part of the immune repertoire (Arunkumar and Zielinski, 2021; Chi et al., 2020). Another example is the MHC gene locus, which contains many genes that involve in the immune defenses and shows the highest diversity among population (Petersdorf and O'HUigin, 2019). Because of the structural complexity of these immunogenetic loci, a comprehensive description of these alleles remains a challenge. The complete assembly and annotation of the immunogenomic loci will provide fundamental and accurate descriptive data for immunology studies. Excitingly, using nanopore sequencing technology, human MHC gene locus was completely assembled and phased with ultra-long reads (Jain et al., 2018a).

The average cost of *de novo* assembly of a genome has dramatically decreased because of the improved next generation sequencing (NGS) technologies such as Illumina(Hu et al., 2021). More importantly, the third-generation sequencing technologies, which can produce long reads that exceeds dozens of kilobases, have led to a paradigm shift in whole-genome assembly, not only in experimental methods but also in algorithms (van Dijk et al., 2018). Pacific Biosciences (PacBio) 'single-molecule real time (SMRT)' methods can generate ~10 Kb long HiFi reads with 99%

accuracy (Ardui et al., 2018). Oxford Nanopore Technologies (ONT) recently developed an ultra-long read method that can produce reads with an average length of ~50 Kb, and the longest reads can be hundreds of kilobases or even over mega-bases (Jain *et al.*, 2018a; Nurk et al., 2022). The incredible technical progress has promoted a prosperity of genome assemblies from animals to plants. For human genome, assembly of a centromere on the Y chromosome(Jain et al., 2018b), telomere-to-telomere assembly of specific chromosome (Logsdon et al., 2021; Miga et al., 2020), and a real gapless assembly of all 22 autosomes plus X chromosome (Nurk *et al.*, 2022) have been reported recently. These advances provided detailed data and the panoramic view of all genomic variants, especially the immunogenetic diversity of humans.

As one of the most important livestock, cows made important historically contributions and are continuing contribute to the basic and applied immunology (Guzman and Montoya, 2018; Vlasova and Saif, 2021). The basis of vaccination began from the protection against smallpox by inoculation with cowpox (Pead, 2003), and CD205 was firstly identified as dendritic cell marker in cattle (Naessens and Howard, 1993). Cattle also showed privileges in studying human infectious diseases over mouse models such as tuberculosis (Waters et al., 2011) and respiratory syncytial virus (RSV) (Taylor, 2017), as human and bovine share higher similarities in immunity. Moreover, recent investigations of cattle immune responses provided novel insights into the relationships between microbes and host (Gomez et al., 2019). For example, the long third heavy chain complementary determining regions (CDRH3) in cattle has been

shown to be capable of rapid elicitation of broad-neutralizing antibodies against Human Immunodeficiency Virus (HIV) (Sok et al., 2017). However, the insufficient understanding of cattle immune system hinders the studies of this important model farm animal. A high-quality reference genome is crucial to facilitate research on cattle immunity.

Besides the current official cattle genome, ARS-UCD1.2, de novo assembly of the cattle genome has been tried with PacBio SMRT method (Rosen et al., 2020). Two genome assembly studies of water buffalo and Simmental cattle have also been reported recently (Heaton et al., 2021; Low et al., 2019). All the above assemblies showed limited genome continuity and completeness, and none of them used ONT ultra-long read method. In this study, we report a new assembly for the cattle genome with a combination of several advanced sequencing methods, in particular, the ONT ultra-long read sequencing technology. Our data significantly surpassed the continuity and accuracy of ARS-UCD1.2, and enabled the gapless assembly and refined annotation of the immunogenetic loci, including TR, IG and MHC.

RESULTS

De novo Assembly of a Cattle Genome.

To assemble a most accurate genome version of cattle, we carried out a whole genome sequencing of female cattle embryonic fibroblasts using ONT ultra-long read sequence technology in conjunction with PacBio circular consensus sequencing (HiFi), Illumina

NGS, Hi-C and BioNano optical maps. 237.8 Gb ultra-long reads were produced with an N50 length of 70.4 Kb, and the longest read was 872.5 Kb. The ONT data exhibited a great advantage in read-length compared to that of PacBio and Illumina methods (Figures S1A-C).

Using NextDenovo, we performed the *de novo* assembly of the ultra-long reads (Figure S1D). The assembly comprised only 145 primary contigs with an N50 contig size of 74.01 Mb and a total length of 2.68 Gb. The contigs were then polished and corrected with PacBio HiFi reads and Illumina reads, anchored with BioNano optical maps and Hi-C interaction matrix into a final version genome with excellent continuity (Figure S1E). The N50 scaffold size was improved to 74.72 Mb and the final genome size was 2.71 Gb, and we named our assembly as NCBA1.0 (Figure 1A). To evaluate the completeness and accuracy of the NCBA1.0, we aligned all the reads back to this new assembled genome, and 99% of the whole genome had a minimum coverage of 36 × by ultra-long reads, 4× by PacBio HiFi reads and 64× by Illumina reads (Figures 1B and S1F). The ultra-long reads showed significant low-bias on GC contents of genomic regions compared to that of Illumina reads and HiFi reads, especially at GC-poor regions (Figure S1G). This result exhibited the outstanding performance of ultra-long reads in genome assembly, as heterochromatins consists of AT-rich repeats, especially at the telomeres and acrocentric regions.

Gap Filling on the Reference Genome ARS-UCD1.2

The NCBA1.0 assembly showed tremendous sequence integrity compared to the current cattle reference genome, ARS-UCD1.2, and other cattle genomes (Figure S1H). Of the 30 chromosomes, 12 chromosomes were packaged by one single contig (Figure 1A). The q-arms of seven chromosomes were ended with a minimum of 15 Kb $(TTAGGG)_n$ telomere repeats, among which five chromosomes (chromosome 17, 20, 22, 26 and 28) were gapless centromere-to-telomere assemblies (Figure 1A). The gap-remaining regions were mainly localized at the acrocentric regions of p-arms and the chromosome X. Further, we assessed whether the remained gaps in ARS-UCD1.2 can be filled using our assembly. The reference cattle genome ARS-UCD1.2 contains 30 chromosomes and 2180 unplaced scaffolds. There are 386 gaps that denoted as Ns and 315 gaps of which were localized on chromosomes. With our assembly, 154 gaps on chromosomes were filled (Figures 1A-B). Additionally, 420 scaffolds and 47 partial scaffolds, with a total length of 24.89 Mb, can be properly placed back in the new genome (Figure S2). The scaffolds placed ranged from Kb to Mb and the largest scaffold was 4.3 Mb in chromosome X.

We further annotated the newly assembled genome. Transposable element repeats (TE) were account for 46.53% of NCBA1.0 and the total ratio of repeat sequences were 47.30% (Figure 1C). 20,288 genes were predicted with an average length of 40.4 Kb, which showed close consistency with bovine and other proximal species (Figures 1D and S3A-F). Functional annotations of the genes among databases, including Benchmarking Universal Single-Copy Orthologs (BUSCO, 95.25% overall alignment

rate), KEGG, GO, KOG, NR and Swiss-Prot, showed both high coverage and intersection ratios (Figures S4A-C). We also validated the expression of the predicted genes by RNA-Seq and confirmed the expression of 89.7% of these genes (Figure 1D). These data demonstrated the reliability of the genome annotation of NCBA1.0.

Immunoglobin Gene Loci Annotation

The assembly of a cattle genome with the most continuity and completeness allowed us to depict the detailed gene structures of the complex immunogenomic loci. We mainly looked into the IG, TR and MHC gene clusters, which localized on six different chromosomes (Figure S5). IG genes or B cell receptors are composed by three gene loci, immunoglobin heavy chain (IGH), lambda chain (IGL) and kappa chain (IGK). All IG-related gene loci were covered with gapless contigs and were well annotated in NCBA1.0, while in genome ARS-UCD1.2, IGH is missing, and IGL region remains six gaps (Figures 2A-D and S5). Detailed gene structure and functionality annotations of IG/TR loci were performed mainly following the IMGT criteria (Figures S6A and S6B).

The IGH was 616.0 Kb in size and was located in the end of the q-arm of chromosome 21 (Figures 2A, 2B and S7A). During the maturation of IG, a process called V(D)J recombination takes place. This process combines randomly selected one segment from each of the preexisting variable (V), diversity (D), and joining (J) gene clusters and give rise to the tremendous diversity of IGs on mature B cells (Nielsen and Boyd, 2018). In the NCBA1.0, the IGH locus consists of 48 IGHV genes (11 functional)

belonging to 3 IGHV subgroups, and 17 IGHD genes (all functional), 12 IGHJ genes (3 functional) and 10 IGHC genes (8 functional) (Figure 2B). A previous study assembled IHG locus by sequencing seven BAC clones and generated an IHG gene structure containing three tandem [IGHDP-IGHV3-$(IGHDv)_n$] repeats (Ma et al., 2016). In contrast, there are only two tandem repeats [IGHDP-IGHV3-$(IGHDv)_n$] in IGH locus of NCBA1.0, and the repeat regions as well as the adjacent gene loci were fully covered with multiple ultra-long reads longer than 100 Kb, demonstrating the accuracy and reliability of sequence assembly in our study (Figures S7B and S8 A-B). In addition, compared to the above work, two extra IGHV genes in the V region were identified (Figure 2B). Thus, our data suggested that the IGH locus were organized as: $(IGHV)_{46}$-$(IGHDv)_5$-$(IGHJ)_6$-IGHM1-[IGHDP-IGHV3-$(IGHDv)_n$]$_2$-$(IGHJ)_6$-IGHM2-IGHD-IGHG3-IGHG1-IGHG2-IGHE-IGHA (Figures 2A and 2B). 10% of cattle IGs contain ultra-long complementarity determining region (CDR3) that were composed of IGHV1-7 and IGHD8 (Deiss et al., 2019; Haakenson et al., 2018). We identified one unique gene locus of IGHV1-7 and one unique IGHD8 in the global IGH gene locus. We renamed these gene loci as IGHV1-6 and IGHD8-2 according to their position in NCBA1.0 (Figure 2B).

The IGL locus spanned 643.9 Kb at q-arm of reverse strand on chromosome 17. By filling the remained six gaps in the IGL locus of ARS-UCD1.2, total 125 IGLV genes (37 functional) were annotated, among which 51 IGLV genes were newly recognized (Figures 2C and S9A). We also corrected the repeat numbers of IGLJ-IGLC

clusters from nine to six (Figures 2C and S9A-B). The whole IGL genome locus was covered with an average depth of 13 by ultra-long reads longer than 100 Kb, and four ultra-long reads span over the entire IGLC region (Figure S10), giving incontrovertible evidence for the genome assembly and annotation in IGL locus. These results illustrated that the IGL genes are organized as: $(IGLV)_{125}$-$(IGLJ\text{-}IGLC)_{6}$.

The IGK is the smallest gene locus of IG and spans 214.3 Kb between 47.2 and 47.4 Mb on chromosome 11 (Figures 2D and S11). IGK consists of 28 V genes (7 functional), 5 J genes (1 functional) and 1 C gene in NCBA1.0 and 3 new V genes were found compared to the previous genome version, ARS-UCD1.2 (Figure S11). Our data suggested that the IGK locus is organized as $(IGKV)_{28}$-$(IGKJ)_{5}$-IGKC.

T Cell Receptor Gene Loci Annotation

The T cell receptors of cattle are composed by four gene subgroups, TRA, TRB, TRD and TRG, that localized at four genomic loci (Figure S5). TRA and TRD form the most complicated immunogenomic locus together that ranges over 3.3 Mb on the reverse strand of chromosome 10 (Figures 3A and S12). The entire TRD resides within the genetic region of TRA (Figures 3A and 3B). Surprisingly, the V region of TRA/TRD spanned over 3 Mb, which was more than 90% of the total TRA/TRD region. In NCBA1.0, we annotated 281 TRAV genes (148 functional) and 57 TRDV genes (46 functional), while in ARS-UCD1.2, only 183 TRAV genes (85 functional) and 39 TRDV genes (31 functional) were defined. In line with ARS-UCD1.2, the D-J-C cluster

consisted of 60 TRAJ genes, 1 TRAC gene, 9 TRDD genes, 4 TRDJ genes and 1 TRDC gene (Figure 3B). Therefore, our data suggested that the TRA(D) genomic structures were organized as $[TRA(D)V]_n$-$(TRDD)_9$-$(TRDJ)_4$-TRDC-TRDV3-$(TRAJ)_{60}$-TRAC.

The TRB genomic locus was assembled without gap in NCBA1.0. It spanned 667.3 Kb between 105.4 Mb and 106.2 Mb in chromosome 4, and consisted of 153 TRBV genes (87 functional), 3 TRBD genes (all functional), 19 TRBJ genes (15 functional) and 3 functional TRBC genes (Figures 4A and 4B). Our data closed the previously remained two gaps in ARS-UCD1.2 (Figures 4A and S13A). The TRBD, TRBJ and TRBC genes were organized into three tandem D-J-C cassettes, followed by one functional TRBV gene (TRBV30) in inverted orientation (Figure 4B). Thus, the TRB genomic structures were organized as: $(TRBV)_{152}$-$[TRBD\text{-}(TRBJ)_n\text{-}TRBC]_3$-TRBV30.

The TRG genes were localized in two separate loci on chromosome 4 which were on different strands and were 30 Mb apart from each other (Figures 4C and S13B). TRG1 spanned 229.3 Kb and comprised four tandem V-J-C cassettes, while TRG2 spanned 106.0 Kb and comprised three tandem V-J-C cassettes (Figure 4C). In total, TRG was consisted of 18 TRGV genes (17 functional), 10 TRGJ genes (8 functional) and 7 TRGC genes (all functional), and TRG genes were organized as $[(TRGV)_n\text{-}(TRGJ)_n\text{-}TRGC]_4$ for TRG1 and $[(TRGV)_n\text{-}(TRGJ)_n\text{-}TRGC]_3$ for TRG2.

In summary, the cattle genome NCBA1.0 consisted of a total number of 884 IG and TR genes (258 IG and 626 TR) that were localized in seven major loci and distributed in 710 V, 29 D, 116 J and 29 C genes (Table 1). The elaborate annotations

of NCAB1.0 greatly enriched the gene sequence diversities compared to that in IMGT database, especially in TR genes (Figure S14). Although there are still two gaps remained in the TRA/TRD region in our assembly, our work provided important data revealing the dramatic diversity of cattle immune repertoire.

Phylogenetic Analysis of V Genes

Next, we performed a phylogenetic analysis of V genes across species. Gene diversities statistics of species with detailed IG/TR annotations showed that NCBA1.0 has the highest gene number among known species, especially TR genes (Figure S15). Phylogenetic trees were constructed with all functional IG and TR V genes from human, mouse and cattle (NCBA1.0). Our data showed that all the V genes were well clustered according to their subgroups, except that TRAV genes were clustered into two separate groups (Figure 5A). This result demonstrated the evolutionary conservation of IG and TR gene subgroups among species. The gene number analysis also showed a significant deviation: both human and mouse had much more IG V genes than TR V genes, while cattle showed an opposite tendency that the TR V genes were three times more than IG V genes (Figure 5B). We also evaluated the sequence similarities among the three species. The cattle are more similar to human than mouse by both IG and TR V genes, while human is more similar to mouse by IG V genes and more similar to cattle by TR V genes (Figure 5C).

MHC Gene Annotation

The MHC plays a crucial role in determining immune responsiveness and is referred as the bovine leukocyte antigen (BoLA) in cattle. In NCBA1.0 assembly, chromosome 23, which contains the BoLA genetic region, was composed by one single contig and the BoLA ranged 3.38 Mb (Figures 6A and S16A). The MHC genes mainly contain two clusters, MHC class I and MHC class II (Petersdorf and O'HUigin, 2019), and by acquiring the seamless sequence of this gene locus, we were able to annotate the MHC genes coordinately. The cattle contain six classical MHC I genes (BoLA 1-6) with high polymorphism and ten non-classical MHC I genes (NC1-10) that show limited polymorphism (Plasil et al., 2022). Classical MHC I genes located at the 3'end of the entire BoLA region. The non-classical MHC I genes NC6-10 were adjacent to classical MHC I genes and genes NC2-5 were 600 Kb upstream away (Figure 6B). For MHC II genes, there were only DQ and DR gene pairs that were organized in adjacent sequential order (Figure 6C), unlike human MHC II that harboring an additional DP gene pair. We also annotated other genes located within the BoLA region, including C2, IL17, CTL4, TNF and TRIM families (Figure 6B and 6C).

To better understand the gene organizations of BoLA region, haplotypes were phased with ONT ultra-long reads combined with heterozygous SNPs called using Illumina data and HiFi reads. The N50 length of haplotype 1 and haplotype 2 were 198.5 Kb and 200.6 Kb, respectively (Figures 6D and 6E). Of the 3.38 Mb BoLA region, 3.17 Mb were successfully assembled into haplotypes and both haplotypes showed high

continuity (Figure S16B). The contigs of each haplotype showed delicate differences of gene structures in both MHC I and MHC II (Figure 6F), demonstrating the polymorphism and polygeny of BoLA among individuals. For example, there were two DQB genes within haplotype 1, while only one DQB gene were identified in haplotype 2. Likewise, the classical MHC I genes vary in terms of both gene types and numbers between two haplotypes (Figure 6F). These data highlighted the power of ONT ultra-long technology in resolving haplotypes of huge intricate gene clusters. Taken together, the BoLA was assembled and phased over its full length in a diploid cattle genome for the first time.

Characterization of Telomere Repeats and Satellite DNA (satDNA)

Heterochromatin, such as centromeres, normally contain long arrays of tandem repeated DNA sequences, known as satDNAs (Escudeiro et al., 2019a; Miga, 2019). Similarly, the telomeres are genomic regions at the end of chromosomes and consist of highly repeated hexanucleotide sequence, TTAGGG (Shay and Wright, 2019). The genomic structure of these repetitive regions in cattle remained elusive due to the lack of investigations of cattle genome with advanced ultra-long sequencing techniques. Thus, we further looked into the telomeres and acrocentric satDNA repeats of the cattle genome NCBA1.0. We identified 1429 ultra-long reads with minimum length of 50 Kb that contains telomere TTAGGG tandem repeats and successfully assembled telomere regions of q-arms of seven chromosomes, and the telomere arrays ranged dozens of

kilobases and the longest one spanned over 70 Kb (Figure 7A). As all autosomes of cattle genome are acrocentric (Frohlich et al., 2017), no telomere of p-arms was assembled due to the genomic structures of satDNA arrays in the centromere, which located right beside the telomeres.

We next analyzed the distributions of satDNAs in NCBA1.0. Eight types of cattle satDNAs have been reported previously (Ashari et al., 2012) (Escudeiro *et al.*, 2019a; Escudeiro et al., 2019b). In our data, FIVE types of satDNAs, SATI, SATIII, SATIV, SAT1.711a and SAT1.711b, constituted the majority of the satDNA regions (Figure 7B). We also identified thousands of copies of a 23-mer arrays, which accounted for the highest abundance of all the satDNA repeats (Figures 7B and 7C). Interestingly, one type of the satDNAs, SAT1.711b arrays, were sparsely distributed along the whole genome, while other satDNAs were located mainly within the acrocentric regions of p-arms (Figures 7B and S17). Collectively, we provided important information for telomeric and satDNA array organizations of cattle genome.

DISCUSSION

Sequencing-technology progress, especially the emerging of ONT ultra-long sequencing, has achieved great success in the complete human genome assembly, T2T-CHM13, which obtained gapless sequences for all chromosomes except Y (Nurk *et al.*, 2022). As the immune system possesses the biggest source of genetic variation, depicting the immune genomic loci was almost impossible several years ago, let alone

population genomic diversity studies of the immune system. A completely assembled genome, or at least completely assembled immune genomic loci, is the cornerstone for the in-depth understanding of the immune system of a given species. In this study by taking advantage of the ONT ultra-long sequencing method, we presented a new cattle genome assembly that exhibited remarkable improvement over existing assemblies. We illustrated the advances of our new assembly in gap filling and scaffold assignment. Importantly, we delineated the complex genomic structures of TR, IG and MHC loci, which provides fundamental immunogenomic data for further immune studies in cattle.

Our work affords the most precise roadmap of cattle immune-genome up to now. We corrected the tandem repeat regions within the IGH locus from three to two, and filled 13 gaps within the IG and TR genomic regions. 710 V genes were annotated in NCBA1.0, while only 524 V genes are currently collected in IMGT database (Table 1 and Figure S14). The cattle MHC region was also seamlessly packaged and properly phased, which is the first intact MHC assembly beside human to our knowledge (Jain *et al.*, 2018a).

Cattle showed distinct characteristics in both IG and TR immune repertoire. For IGH genes, there were only 48 V genes, which was significantly fewer than human and mouse. The relatively low diversity of IG in cattle was likely compensated by the ultralong CDRH3s that can be found in approximately 10% of the immunoglobins (Haakenson *et al.*, 2018). CDRH3s allow cattle antibodies to bind a wider range of antigens and was showed to play a key role in neutralizing HIV spike protein during

the immune response (Deiss *et al.*, 2019; Sok *et al.*, 2017; Wang et al., 2013). These cattle ultralong CDRH3s almost exclusively used the same V gene segment (IGHV1-7) that contains an eight-nucleotide duplication "TACTACTG" at its 3' end, and the same D gene segment (IGHD8-2) that is known as the longest D gene (Deiss *et al.*, 2019). Both IGHV1-7 and IGHD8-2 gene loci were clearly depicted in NCBA1.0. Based on our data, the V-D rearrangements between these two loci can be further investigated, for instance, their histone modifications and chromatin accessibility.

Cattle possess the highest TR gene diversities among all species that were annotated with detailed V(D)J gene structures (Figure S15). We identified 509 TR V genes in NCBA1.0, which was three times more than human TR V genes. The reason why cattle genome contains much more TR V genes than IG V genes remains indistinct. This is probably related with the gut-associated mucosal tissue that contacts with a great diversity of food and microbial antigens, as mucosal T cells play a central role in distinguishing dietary proteins and commensal bacteria from harmful pathogens (Chase and Kaushik, 2019). It is worth mentioning that the remarkably abundant TR repertoire in cattle may serve as a natural resource pool for the screening of specific TRs with extraordinary therapeutic activity against human diseases, such as cancer. Moreover, cattle have a large proportion of γδ T cells that showed regulatory and antigen-presenting functions (Guzman et al., 2012). Further studies of cattle T cells shall shed light on the understanding of γδ T cells, which remains elusive in humans due to their low abundance (Guzman et al., 2014) .

Limited studies in cattle centromeres and telomeres have been done (Escudeiro *et al.*, 2019b; Ilska-Warner et al., 2019). Despite the telomere length is associated with productive lifespan and fitness. In this study, we captured telomere repeats in seven chromosomes and for the first time precisely evaluated the telomere length in cattle. Evaluation of satDNAs based on the new assembly revealed sequence ranges and composition ratio of satDNAs, as well as the specific patterns of how different satDNAs were organized that have never been revealed before.

In summary, the new assembly, NCBA1.0, represented a more complete and accurate reference of cattle genome, as well as the immune genome, thereby facilitating further investigations of the immune system in cattle, and perhaps other mammals. These data can be a blueprint for the final gapless telomere-to-telomere cattle genome assembly in the near future.

ACKNOWLEDGEMENT

This work was supported by grants from the National Key Research and Development Program of China (No. 2020YFA0707702 to Tao Li and No. 2020YFA0707703 to T.Z.) and China National Natural Science Foundation (No. 81925017 to Tao Li, No. 81872153 to T.Z.).

AUTHOR CONTRIBUTIONS

T.L., T.-T.L. and T.Z. conceptualized and designed the project. T.X., Z.-L.S., M.W., and F.-R.D. prepared the cells. T.-T.L, J.-Q.W. and T.X. analyzed the data. T.-T.L, J.-Q.W., T.X., H.H., S.J., J.L., J.W., G.Y., J.-N.F., Y.-P.D. and J.P. annotated the IG, TR and MHC genes. T.L., T.-T.L. T.X. and X.-M.Z. wrote the manuscript with the help of all authors.

DECLARATION OF INTERESTS

All the authors declare no competing interests.

FIGURE TITLES and LEGENDS

Figure 1. A Global Picture of the *de novo* Cattle Genome Assembly

(A) Ideograph of cattle genome assembly NCBA1.0. Chromosomes composed of a single contig are in dark blue, satDNAs are in orange and telomeres are in purple. Closed gaps and properly placed scaffolds of ARS-UCD1.2 are depicted as red triangles and yellow strips.

(B) Read coverage of cattle genome NCBA1.0. Lane 1-3: read coverages by ONT ultra-long reads, PacBio HiFi reads and Illumine NGS reads. Lane 4-5: remaining gaps in NCBA1.0 and ARS-UCD1.2.

(C) Composition ratio of repeat elements in NCBA1.0.

(D) Gene annotations of NCBA1.0. There were 20288 genes predicted in total, and the ratio of genes with functional annotation, overlapped with BUSCO and validated with RNA-Seq were indicated.

Figure 2. The Cattle Immunoglobin Loci

(A) Genomic organization of IGH locus in KT723008, ARS-UCD1.2 and NCBA1.0. Repeated regions were drawn as blue rectangles, and ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

(B) Detailed diagram of IGH gene structure and annotation in NCBA1.0. Two repeated regions were labeled as blue rectangles (rep1-1 and rep1-2). IGHV1-6 and IGHD8-2 consists of the ultralong CDRH3 were marked with green triangles.

(C-D) Genomic organization of IGL and IGK loci in ARS-UCD1.2 and NCBA1.0. Genomic gaps in ARS-UCD1.2 were depicted beneath the genes according to their coordinates and ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

Figure 3. The Cattle TRA/TRD Loci

(A) General organization of TRA/TRD loci in ARS-UCD1.2 and NCBA1.0. Genomic gaps in ARS-UCD1.2 and NCBA1.0 were depicted beneath the genes according to their coordinates. ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

(B) The detailed genetic map of TRA/TRD loci in NCBA1.0. Labels of TRD genes starts with “D” and all TRD genes reside within the TRA genomic region.

Figure 4. The Cattle TRB and TRG Loci

(A) The genomic organization of TRB locus in ARS-UCD1.2 and NCBA1.0. Genomic gaps in ARS-UCD1.2 were depicted beneath the genes according to their coordinates. ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

(B) The detailed diagram of TRB gene structure in NCBA1.0.

(C) The genomic organization of TRG loci in NCBA1.0. ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

Figure 5. Phylogenetic Analysis of V Genes

(A) Phylogenetic tree of all functional IG and TR V genes. V genes of human, mouse and cattle were merged together and clustered well according to their biological classes.

(B) Bar plot of V gene numbers of cattle, human and mouse.

(C) Sequence similarities of functional V genes between cattle, human and mouse. Y axis indicates alignment identities of each V gene to the genes of the other two species that share the most similarity.

Figure 6. MHC Gene Locus and Haplotyping

(A) Genomic coordinates of MHC locus in chromosome 23. Chromosome 23 consists of only one contig and MHC contains two separate genomic regions.

(B-C) Detailed gene organizations of MHC class I (B) and class II (C). MHC genes

were labeled in red.

(D-F) Haplotyping of MHC genomic region. Genomic locations of two haplotigs within the MHC region (D). Length distributions of two haplotigs (E). Gene organization variation of two haplotigs (F).

Figure 7. Telomere Length and satDNA Distributions

(A) Lengths of telomere repeats in q-arms of seven chromosomes.

(B) Genome-wide distributions of cattle satDNAs. satDNAs mainly localized within the acrocentric regions except sat1.711b.

(C) Composition chart of satDNAs in NCBA1.0 assembly.

TABLES with TITLES and LEGENDS

Table 1. Gene Numbers of Each Immune Locus in NCBA 1.0 Assembly

Functional annotation for each gene of IGs and TRs was performed based on the IMGT criteria (Figure S6), and gene numbers were collected and listed in the table.

Types	IGH	IGK	IGL	TRA	TRB	TRD	TRG1	TRG2	Sum
V	48	28	125	281	153	57	14	4	710
(F/P/ORF)	(11/37/0)	(7/19/2)	(37/80/8)	(148/108/25)	(87/55/11)	(46/7/4)	(13/1/0)	(4/0/0)	(353/307/50)
D	17	NA	NA	NA	3	9	NA	NA	29
(F/P/ORF)	(17/0/0)				(3/0/0)	(6/0/3)			(26/0/3)
J	12	5	6	60	19	4	5	5	116
(F/P/ORF)	(3/1/8)	(1/0/4)	(4/0/2)	(53/2/5)	(15/1/3)	(3/0/1)	(5/0/0)	(3/0/2)	(87/4/25)
C	10	1	6	1	3	1	4	3	29
(F/P/ORF)	(8/2/0)	(1/0/0)	(3/3/0)	(1/0/0)	(3/0/0)	(1/0/0)	(4/0/0)	(3/0/0)	(24/5/0)
Sum	87	34	137	342	178	71	23	12	
(F/P/ORF)	(39/40/8)	(9/19/6)	(44/83/10)	(202/110/30)	(108/56/14)	(56/7/8)	(22/1/0)	(10/0/2)	

MATERIALS AND METHODS

Sample Collection and Genomic DNA Isolation

The holstein cattle of high genetic merit were mated to produce an elite fetus that was recovered at day 60. The cattle fibroblast cells were isolated from this fetus via disaggregation of all tissue excluding the viscera and limbs, and cultured in Dulbecco's Modified Eagle's Medium (DMEM; Gibco, Grand Island, New York, USA) supplemented with 10% fetal bovine serum (FBS; Gibco, Grand Island, New York, USA) at 37.5 °C in an atmosphere of 5% CO2 and humidified air. The genomic DNA was extracted using the QIAGEN Genomic-tips kit (QIAGEN, Valencia, CA, USA) according to the manufacturers' instructions.

Ultra-long Library Construction and Sequencing

To obtain ultra-long reads, only the large DNA fragments were recovered with BluePippin, followed with end repair and dA-tailing (NEBNext Module, MA, USA). After careful purification, the adapter ligation was performed with SQK-LSK109 ligation kit (Oxford Nanopore Technologies, Oxford, UK) and the final product was quantified by fluorometry (Qubit) to ensure >500 ng DNA was retained, and sequenced on the Oxford Nanopore PromethION platform. ONT ultra-long reads were generated by Grandomics Biosciences company and only reads with a minimum mean quality

score of 7 were kept for the following assembly.

Hi-C Library Construction and Sequencing

The Hi-C experiment was performed exactly following the *in situ* Hi-C method (Rao et al., 2014). Briefly, the cross-linked cells were lysed and digested with MboI, filled with biotin-dATP, ligated with T4 DNA ligase and reverse crosslinked. Then the biotin-labeled DNA was enriched and sequenced with Illumina sequencing platforms following the manufacturer's instructions. The read mapping, quality control and matrix building were performed with HiC-Pro (Servant et al., 2015).

De Novo Genome Assembly

The de novo assembly of ONT ultra-long reads was performed with NextDenovo (https://github.com/Nextomics/NextDenovo). The reads were first self-corrected to generate consensus sequences with NextCorrect module and then assembled into preliminary assembly with NextGraph module. To correct the preliminary assembly, the original ONT reads and PacBio CCS reads were mapped with minimap2 (Li, 2018) and corrected with Racon (Vaser et al., 2017) with default parameters for three iterations. Then the Illumina reads were used to polish the corrected assembly with NextPolish (Hu et al., 2020) for 4 iterations to generate the final polished genome assembly. The polished assembly was used as reference for the de novo assembly of BioNano data to generate scaffolds. For Hi-C data, LACHESIS (Burton et al., 2013)

was used to cluster, order and direct the scaffolds to generate the final chromosomal level genome assembly.

Assembly Evaluation

BUSCO 3.1.0 (-l mammalia_odb9 -g genome) (Manni et al., 2021) was used to evaluate the genome completeness based on included gene numbers, and CEGMA v2 (Parra et al., 2007) was used to assess the assembly based on included eukaryotic protein core families with default parameters. Sequence accuracy was assessed by total number of homozygous SNPs identified by Illumina reads mapped to the assembly. Exogenous pollution was assessed based on the distributions of GC-depth and reads coverage.

Genome Annotation

For repeated sequences, TRF (Benson, 1999) was used to identify tandem repeats and RepeatMasker (Tarailo-Graovac and Chen, 2009) was used to identify transposon-based elements. For gene structures, PASA (Haas et al., 2003)was used to predict gene coordinates based on Illumina RNA-Seq data, GeMoMa (Keilwagen et al., 2018) was used to predict gene coordinates based on protein sequences of proximal species, and GeneMark-ST (Lomsadze et al., 2005) was used to predict genes from de novo. The three gene sets were integrated into an initial gene set with EVM (Haas et al., 2008) and finally to a clean gene set by removing (http://transposonpsi.sourceforge.net/). For further gene function annotation, the protein sequences of the predicted gene set

were searched against several databases to predict their functions, including Non-Redundant Protein Database (NR), Kyoto Encyclopedia of Gene and Genome (KEGG), Eukaryotic Orthologous Groups of protein (KOG), InterProScan GO database, and Swiss-Prot.

Genome Comparison and Gap filling

Genome sequence comparison between ARS-UCD1.2 and the new assembly was performed with LASTZ (Harris, 2007) at chromosomal level. To fill the gaps of ARS-UCD1.2, 10 kb sequences up and downstream of the gap sites in chromosomes were fetched and aligned to the new assembly with minimap2 (Li, 2018). Only alignments with >90% identity were kept and the alignment results of pair of gap sequences were manually checked to ensure the gap loci were within one contig. The unplaced scaffolds were first split up at gap loci and then aligned to the new assembly with minimap2. The scaffolds were reported if > 50% alignment identity and located within one contig.

Telomeres and satDNA Annotation

To get the loci and lengths of telomeres, all short tandem repeats were identified with TRF (Benson, 1999) with in the new assembly. Then short tandem repeats of TTAGGG were identified and only these located at the end of chromosomes were kept as telomeres. To annotate satDNAs, 57 nucleotide sequences belong to 8 satDNA classes were collected from NCBI and a blast database were built containing these 57

sequences. The sequence similarities among these 57 sequences were analyzed with blast (Johnson et al., 2008). Then all short tandem repeats were identified with TRF, and their pattern sequences were cleaned up to a fasta file and then blasted against the satDNA database. The alignments were filtered with >80% sequence identity and the locations and copy numbers were merged from TRF results.

IG and TCR Gene Annotation

All cattle IG and TCR gene sequences were downloaded from IMGT database (Lefranc et al., 2015). The gene sequences were aligned to the new assembly with bowtie2 (Langmead and Salzberg, 2012), and the alignment results were merged and manually checked according to their subgroups (IGH, IGK, IGL, TRA, TRB, TRD and TRG) to make sure the gene clusters were seamlessly assembled. For each locus, all candidate variable (V), diversity (D), joining (J), and constant (C) genes were manually annotated according to the following criteria (Figure S6). Manual annotations were validated by four irrelevant people.

Phylogenetic Analysis of V Genes

Functional IG and TR V gene sequences of human and mouse were downloaded from IMGT database (Lefranc *et al.*, 2015). Functional cattle V genes were retrieved from NCBA1.0. Multiple sequence alignment was performed with Clustal Omega (Larkin et al., 2007), and the outputs were visualized using the Interactive Tree of Life software

(Letunic and Bork, 2019).

MHC Gene Annotation and Haplotyping

Totally 713 BoLA gene allele sequences were downloaded from IPD-MHC database (https://www.ebi.ac.uk/ipd/mhc/). The BoLA gene sequences were aligned to the new assembly with bowtie2 (Langmead and Salzberg, 2012), and the genomic location and order of MHC Class I/II genes were manually conformed with alignments > MAPQ 20. For haplotyping, PacBio HiFi reads that mapped to MHC region were retrieved and used for variant calling, followed with SNP genotyping with WhatsHap (Patterson et al., 2015). Then the genotyped reads were separately retrieved and assembled into haplotypes with Canu pacbio-hifi mode (Koren et al., 2017).

Statistical Analysis

Quantification methods and statistical analysis for each of the separate and integrated analyses are described and referenced in their respective Method Details subsections.

Data and Code Availability

The relevant data reported in this paper are available from the corresponding authors upon reasonable request. The related codes and figures for reproducible research are stored at GitHub (https://github.com/TintingLi/cattleGenome).

SUPPLEMENTAL INFORMATION TITLES and LEGENDS

Figure S1. Data Summary of Cattle Genomic Assembly

(A) Read length distributions of ONT ultra-long and PacBio HiFi methods. Each bar indicates a separate sequencing flow cell.

(B) Read yield distributions of ONT ultra-long and PacBio HiFi methods. Each bar indicates a separate sequencing flow cell.

(C) Length distribution of the total ONT ultra-long reads.

(D) Flow chart of the genome assembly pipeline.

(E) Continuity of ultra-long reads assembled contigs.

(F) Read coverage of NCBA1.0 by ONT ultra-long reads and HiFi reads.

(G) Influence of GC content on genome coverage of three sequencing methods.

(H) Summary of recently assembled genomes related to bovine.

Figure. S2. Properly Placed Scaffolds of ARS-UCD1.2 in NCBA1.0

Top fifty scaffolds of ARS-UCD1.2 according to their sequence length were shown. Information of all properly placed scaffolds were stored in Table S7.

Figure S3. Gene Length Distributions of Five Species

(A) Distribution of the gene lengths of five species.

(B) Distribution of the CDS lengths of five species.

(C) Distribution of the exon lengths of five species.

(D) Distribution of the exon numbers per gene of five species.

(E) Distribution of the intron lengths of five species.

(F) Distribution of the intron numbers per gene of five species.

Figure S4. Functional Annotation of Predicted Genes in NCBA1.0

(A) GO annotation of predicted genes.

(B) KEGG Ortholog annotation of predicted genes.

(C) Venn diagram of genes annotated to KOG, GO, KEGG, NR and Swiss-Prot.

Figure S5. Immune Gene Loci in NCBA1.0 Assembly

The IG, TR and MHC genomic loci dispersed in six chromosomes. The genomic coordinates for each locus were labeled and the gaps between contigs were annotated too. All immune loci were seamlessly assembled except the two gaps within the TRA/TRD region.

Figure S6. Criteria For Gene Structure and Functionality Annotation,

(A) Sequence conservation logos were created with recombination signal sequences of all functional cattle genes from IMGT by WebLogo software.

(B) Criteria for determining the functionality of an IG/TR gene. The functionality of an IG/TR gene is defined as functional (F), Open Reading Frame (ORF) or Pseudogene (P) based on the sequence analysis.

Figure S7. Ultra-long Reads Coverage of IGH Locus

(A) Global view of the IGH locus covered by ONT ultra-long reads.

(B) Enlarged tandem repeat regions of IGH locus. The repeated regions were represented as blue and red rectangles in KT723008 and NCBA1.0, respectively.

Figure S8. Dot Plots between Two IGH Genomic Sequences

(A) Dot plot of the IGH genomic sequence in NCBA1.0 assembly. The repeated regions were labeled as rep1-1 and rep1-2.

(B) Pairwise alignment between IGH in NCBA1.0 and previously reported IGH sequence (KT723008). The three tandem repeats in KT723008 were labeled as rep2-1, rep2-2 and rep2-3.

Figure S9. Detailed Annotation Map of IGL Locus

(A) Elaborate gene structures of IGL locus in NCBA1.0 assembly.

(B) Read coverage by ONT ultra-long reads in the IGL genomic region.

Figure S10. Enlarged Alignment Map of IGL J-C Cluster Region

ONT ultra-long reads that longer than 100 Kb were aligned back to the IGL locus, and four separate ONT ultra-long reads that span over the entire IJL J-C cluster region were labeled with asterisk.

Figure S11. Detailed Annotation Map of IGK Locus

Genomic coordinate and organization of IGK locus were depicted. ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn.

Figure S12. Global Genetic Map of TRA/TRD loci

Genomic coordinate and organization of TRA/TRD loci were annotated. ONT ultra-long reads that longer than 100 Kb and the mapping to the genomic region were drawn. The remaining two gaps within the TRA/D V gene region were depicted too.

Figure S13. Global Alignment Map of TRB and TRG Loci

(A) Genetic map of TRB locus in chromosome 4. TRB locus resides within contig23 and ONT ultra-long reads that mapped to the genomic region were drawn.

(B) Genetic map of TRG locus in chromosome 4. TRG contains two separate gene clusters: TRG1 and TRG2, that are 32 Mb distant away from each other. ONT ultra-long reads that mapped to the genomic region were drawn.

Figure S14. Gene Statistics of Cattle for Each Immune Locus in IMGT

Table was drawn based on cattle gene numbers of IG and TR collected from IMGT database.

Figure S15. IG and TR Statistics of Different Species in IMGT

Table was drawn based on gene numbers of IG and TR of different species collected from IMGT database.

Figure S16. Genomic Assembly and Haplotyping of MHC Locus

(A) Genomic coordinate and annotation of MHC I and MHC II. ONT ultra-long reads that mapped to the genomic region were drawn.

(B) Sequence alignments between two haplotigs and the MHC genomic region. Colors indicate the sequence identity of the alignments.

Figure S17. Distributions of satDNAs in Scaffolds

Genomic loci of satDNAs within the scaffolds of NCBA1.0 were drawn.

REFERENCES

Ardui, S., Ameur, A., Vermeesch, J.R., and Hestand, M.S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic acids research *46*, 2159-2168. 10.1093/nar/gky066.

Arunkumar, M., and Zielinski, C.E. (2021). T-Cell Receptor Repertoire Analysis with Computational Tools-An Immunologist's Perspective. Cells *10*. 10.3390/cells10123582.

Ashari, M., Busono, W., Nuryadi, and Nurgiartiningsih, A. (2012). Analysis of chromosome and karyotype in Bali cattle and Simmental-Bali (Simbal) crossbreed cattle. Pak J Biol Sci *15*, 736-741. 10.3923/pjbs.2012.736.741.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research *27*, 573-580. 10.1093/nar/27.2.573.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol *31*, 1119-1125. 10.1038/nbt.2727.

Chase, C., and Kaushik, R.S. (2019). Mucosal Immune System of Cattle: All Immune Responses Begin Here. Vet Clin North Am Food Anim Pract *35*, 431-451. 10.1016/j.cvfa.2019.08.006.

Chi, X., Li, Y., and Qiu, X. (2020). V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology *160*, 233-247. 10.1111/imm.13176.

Deiss, T.C., Vadnais, M., Wang, F., Chen, P.L., Torkamani, A., Mwangi, W., Lefranc, M.P., Criscitiello, M.F., and Smider, V.V. (2019). Immunogenetic factors driving formation of ultralong VH CDR3 in Bos taurus antibodies. Cell Mol Immunol *16*, 53-64. 10.1038/cmi.2017.117.

Escudeiro, A., Adega, F., Robinson, T.J., Heslop-Harrison, J.S., and Chaves, R. (2019a). Conservation, Divergence, and Functions of Centromeric Satellite DNA Families in the Bovidae. Genome Biol Evol *11*, 1152-1165. 10.1093/gbe/evz061.

Escudeiro, A., Ferreira, D., Mendes-da-Silva, A., Heslop-Harrison, J.S., Adega, F., and Chaves, R. (2019b). Bovine satellite DNAs – a history of the evolution of complexity and its impact in the Bovidae family. The European Zoological Journal *86*, 20-37. 10.1080/24750263.2018.1558294.

Frohlich, J., Kubickova, S., Musilova, P., Cernohorska, H., Muskova, H., Vodicka, R., and Rubes, J. (2017). Karyotype relationships among selected deer species and cattle revealed by bovine FISH probes. PloS one *12*, e0187559. 10.1371/journal.pone.0187559.

Gomez, D.E., Galvao, K.N., Rodriguez-Lecompte, J.C., and Costa, M.C. (2019). The Cattle Microbiota and the Immune System: An Evolving Field. Vet Clin North Am Food Anim Pract *35*, 485-505. 10.1016/j.cvfa.2019.08.002.

Guzman, E., Hope, J., Taylor, G., Smith, A.L., Cubillos-Zapata, C., and Charleston, B. (2014). Bovine gammadelta T cells are a major regulatory T cell subset. J Immunol *193*, 208-222.

10.4049/jimmunol.1303398.

Guzman, E., and Montoya, M. (2018). Contributions of Farm Animals to Immunology. Front Vet Sci *5*, 307. 10.3389/fvets.2018.00307.

Guzman, E., Price, S., Poulsom, H., and Hope, J. (2012). Bovine gammadelta T cells: cells with multiple functions and important roles in immunity. Vet Immunol Immunopathol *148*, 161-167. 10.1016/j.vetimm.2011.03.013.

Haakenson, J.K., Huang, R., and Smider, V.V. (2018). Diversity in the Cow Ultralong CDR H3 Antibody Repertoire. Front Immunol *9*, 1262. 10.3389/fimmu.2018.01262.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research *31*, 5654-5666. 10.1093/nar/gkg770.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology *9*, R7. 10.1186/gb-2008-9-1-r7.

Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Heaton, M.P., Smith, T.P.L., Bickhart, D.M., Vander Ley, B.L., Kuehn, L.A., Oppenheimer, J., Shafer, W.R., Schuetze, F.T., Stroud, B., McClure, J.C., et al. (2021). A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. J Hered *112*, 184-191. 10.1093/jhered/esab002.

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics *36*, 2253-2255. 10.1093/bioinformatics/btz891.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Hum Immunol *82*, 801-811. 10.1016/j.humimm.2021.02.012.

Ilska-Warner, J.J., Psifidi, A., Seeker, L.A., Wilbourn, R.V., Underwood, S.L., Fairlie, J., Whitelaw, B., Nussey, D.H., Coffey, M.P., and Banos, G. (2019). The Genetic Architecture of Bovine Telomere Length in Early Life and Association With Animal Fitness. Front Genet *10*, 1048.

10.3389/fgene.2019.01048.

Imkeller, K., and Wardemann, H. (2018). Assessing human B cell repertoire diversity and convergence. Immunological reviews *284*, 51-66. 10.1111/imr.12670.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018a). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol *36*, 338-345. 10.1038/nbt.4060.

Jain, M., Olsen, H.E., Turner, D.J., Stoddart, D., Bulazel, K.V., Paten, B., Haussler, D., Willard, H.F., Akeson, M., and Miga, K.H. (2018b). Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol *36*, 321-323. 10.1038/nbt.4109.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic acids research *36*, W5-9. 10.1093/nar/gkn201.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics *19*, 189. 10.1186/s12859-018-2203-5.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res *27*, 722-736. 10.1101/gr.215087.116.

Kumar, B.V., Connors, T.J., and Farber, D.L. (2018). Human T Cell Development, Localization, and Function throughout Life. Immunity *48*, 202-213. 10.1016/j.immuni.2018.01.007.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359. 10.1038/nmeth.1923.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948. 10.1093/bioinformatics/btm404.

Lefranc, M.P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., et al. (2015). IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic acids research *43*, D413-422.

10.1093/nar/gku1056.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research *47*, W256-W259. 10.1093/nar/gkz239.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100. 10.1093/bioinformatics/bty191.

Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2021). The structure, function and evolution of a complete human chromosome 8. Nature *593*, 101-107. 10.1038/s41586-021-03420-7.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic acids research *33*, 6494-6506. 10.1093/nar/gki937.

Low, W.Y., Tearle, R., Bickhart, D.M., Rosen, B.D., Kingan, S.B., Swale, T., Thibaud-Nissen, F., Murphy, T.D., Young, R., Lefevre, L., et al. (2019). Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nature communications *10*. ARTN 260 10.1038/s41467-018-08260-0.

Ma, L., Qin, T., Chu, D., Cheng, X., Wang, J., Wang, X., Wang, P., Han, H., Ren, L., Aitken, R., et al. (2016). Internal Duplications of DH, JH, and C Region Genes Create an Unusual IgH Gene Locus in Cattle. J Immunol *196*, 4358-4366. 10.4049/jimmunol.1600158.

Manni, M., Berkeley, M.R., Seppey, M., Simao, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol *38*, 4647-4654. 10.1093/molbev/msab199.

Miga, K.H. (2019). Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. Genes (Basel) *10*. 10.3390/genes10050352.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature. 10.1038/s41586-020-2547-7.

Naessens, J., and Howard, C.J. (1993). Leukocyte antigens of cattle and sheep. Monoclonal antibodies submitted to the Second Workshop. Vet Immunol Immunopathol *39*, 5-10.

Nielsen, S.C.A., and Boyd, S.D. (2018). Human adaptive immune receptor repertoire analysis-Past, present, and future. Immunological reviews *284*, 9-23. 10.1111/imr.12667.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44-53. 10.1126/science.abj6987.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061-1067. 10.1093/bioinformatics/btm071.

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G.W., and Schonhuth, A. (2015). WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J Comput Biol *22*, 498-509. 10.1089/cmb.2014.0157.

Pead, P.J. (2003). Benjamin Jesty: new light in the dawn of vaccination. Lancet *362*, 2104-2109. 10.1016/S0140-6736(03)15111-2.

Petersdorf, E.W., and O'HUigin, C. (2019). The MHC in the era of next-generation sequencing: Implications for bridging structure with function. Hum Immunol *80*, 67-78. 10.1016/j.humimm.2018.10.002.

Plasil, M., Futas, J., Jelinek, A., Burger, P.A., and Horin, P. (2022). Comparative Genomics of the Major Histocompatibility Complex (MHC) of Felids. Front Genet *13*, 829891. 10.3389/fgene.2022.829891.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680. 10.1016/j.cell.2014.11.021.

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C.G., Tseng, E., Rowan, T.N., Low, W.Y., Zimin, A., Couldrey, C., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience *9*. 10.1093/gigascience/giaa021.

Schultze, J.L., and Aschenbrenner, A.C. (2021). COVID-19 and the human innate immune system. Cell *184*, 1671-1692. 10.1016/j.cell.2021.02.029.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome biology *16*, 259. 10.1186/s13059-015-0831-x.

Sette, A., and Crotty, S. (2021). Adaptive immunity to SARS-CoV-2 and COVID-19. Cell *184*, 861-880. 10.1016/j.cell.2021.01.007.

Shay, J.W., and Wright, W.E. (2019). Telomeres and telomerase: three decades of progress. Nat Rev Genet *20*, 299-309. 10.1038/s41576-019-0099-1.

Sok, D., Le, K.M., Vadnais, M., Saye-Francisco, K.L., Jardine, J.G., Torres, J.L., Berndsen, Z.T., Kong, L., Stanfield, R., Ruiz, J., et al. (2017). Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. Nature *548*, 108-111. 10.1038/nature23301.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics *Chapter 4*, Unit 4 10. 10.1002/0471250953.bi0410s25.

Taylor, G. (2017). Animal models of respiratory syncytial virus infection. Vaccine *35*, 469-480. 10.1016/j.vaccine.2016.11.054.

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet *34*, 666-681. 10.1016/j.tig.2018.05.008.

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res *27*, 737-746. 10.1101/gr.214270.116.

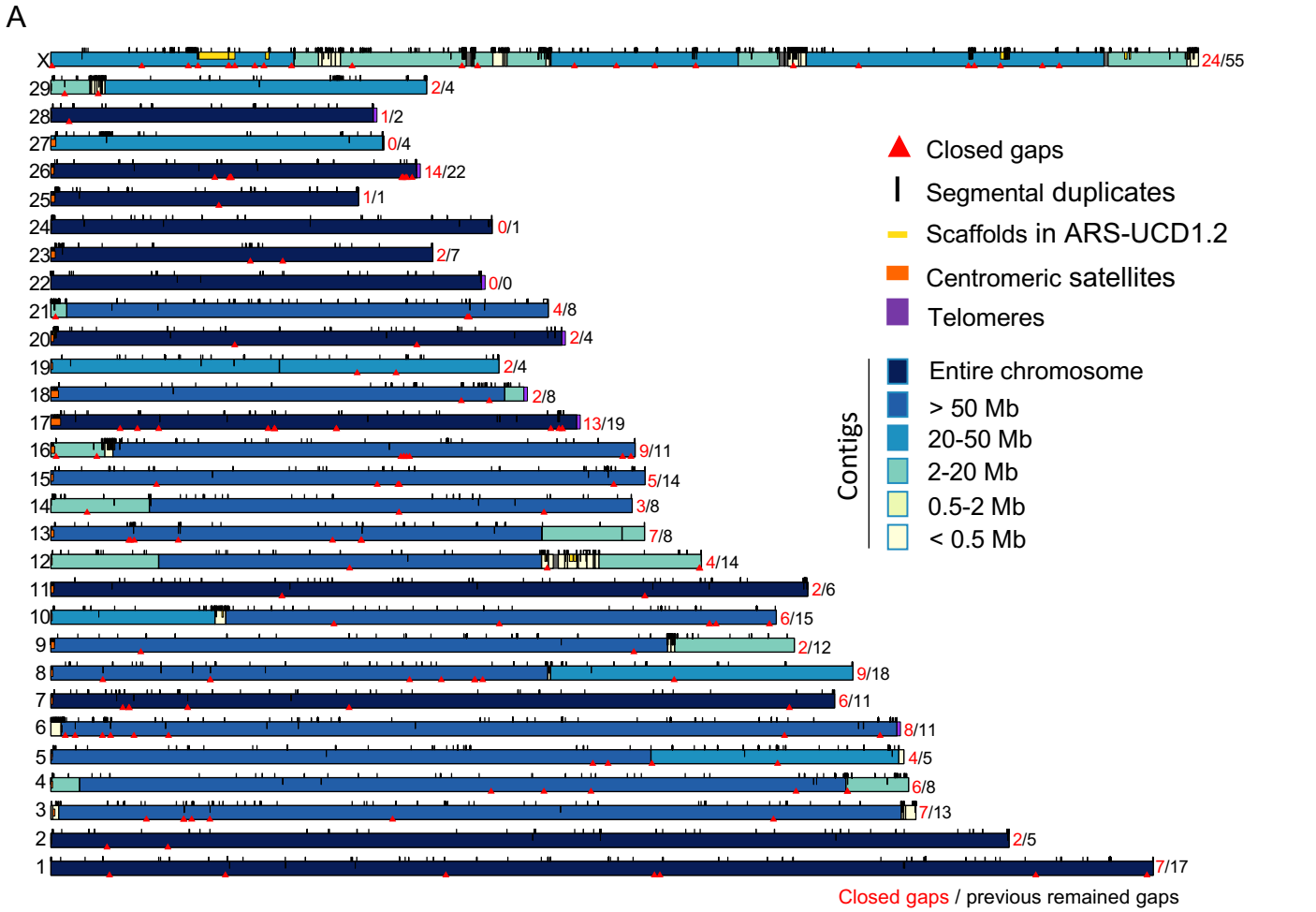
Vlasova, A.N., and Saif, L.J. (2021). Bovine Immunology: Implications for Dairy Cattle. Front Immunol *12*, 643206. 10.3389/fimmu.2021.643206.

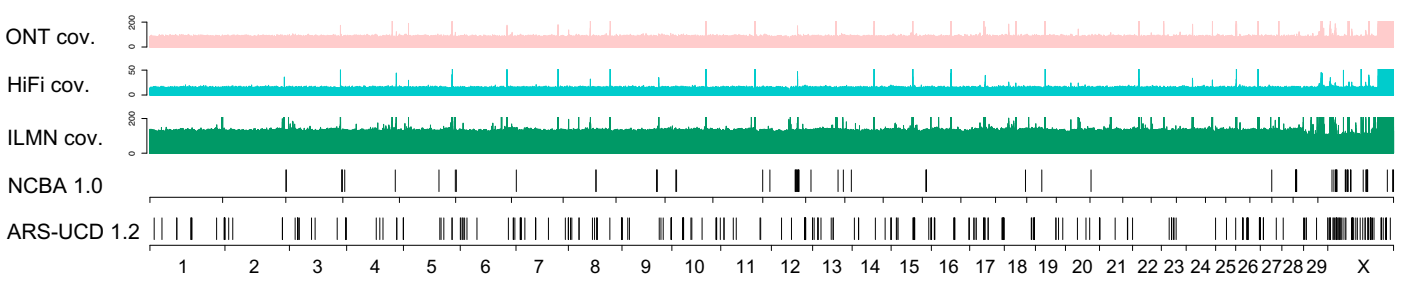
Wang, F., Ekiert, D.C., Ahmad, I., Yu, W., Zhang, Y., Bazirgan, O., Torkamani, A., Raudsepp, T., Mwangi, W., Criscitiello, M.F., et al. (2013). Reshaping antibody diversity. Cell *153*, 1379-1393. 10.1016/j.cell.2013.04.049.

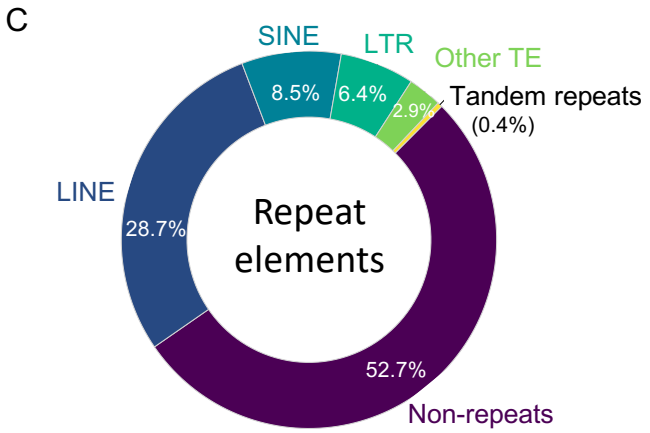
Waters, W.R., Palmer, M.V., Thacker, T.C., Davis, W.C., Sreevatsan, S., Coussens, P., Meade, K.G.,

Hope, J.C., and Estes, D.M. (2011). Tuberculosis immunity: opportunities from studies with cattle. Clin Dev Immunol *2011*, 768542. 10.1155/2011/768542.

Figure 1. A Global Picture of *de novo* Cattle Genome Assembly







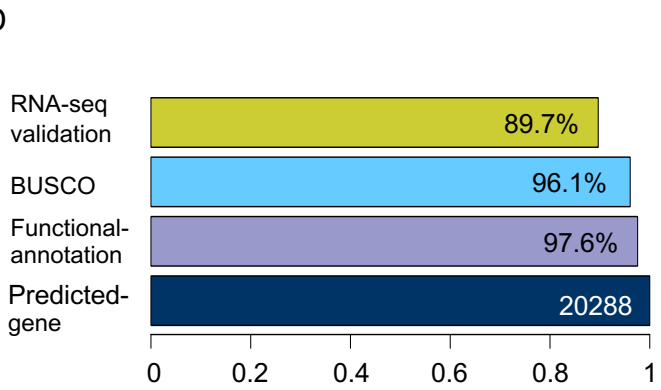


Figure 2. The Cattle Immunoglobin Loci

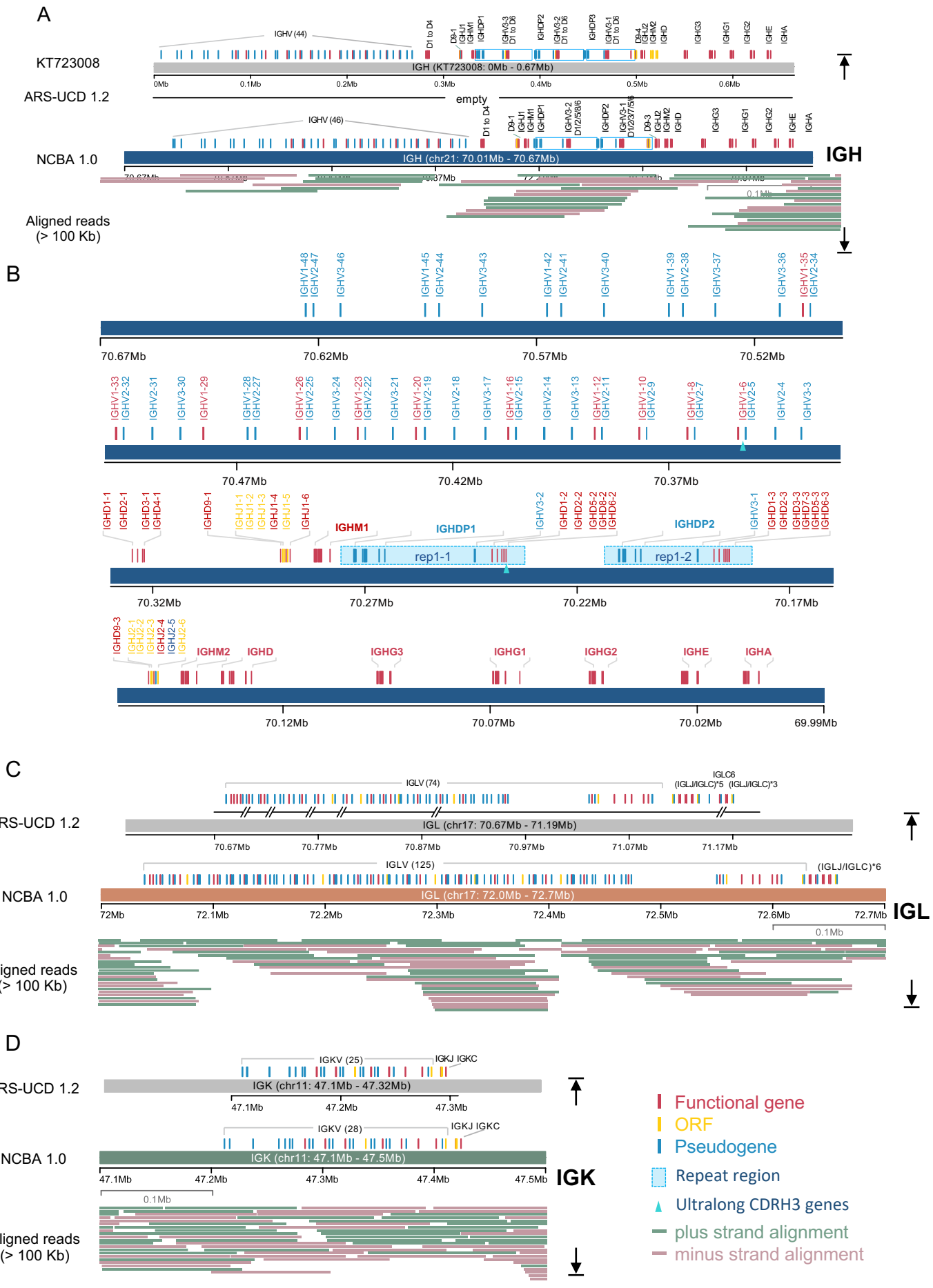


Figure 3. The Cattle TRA/TRD Loci

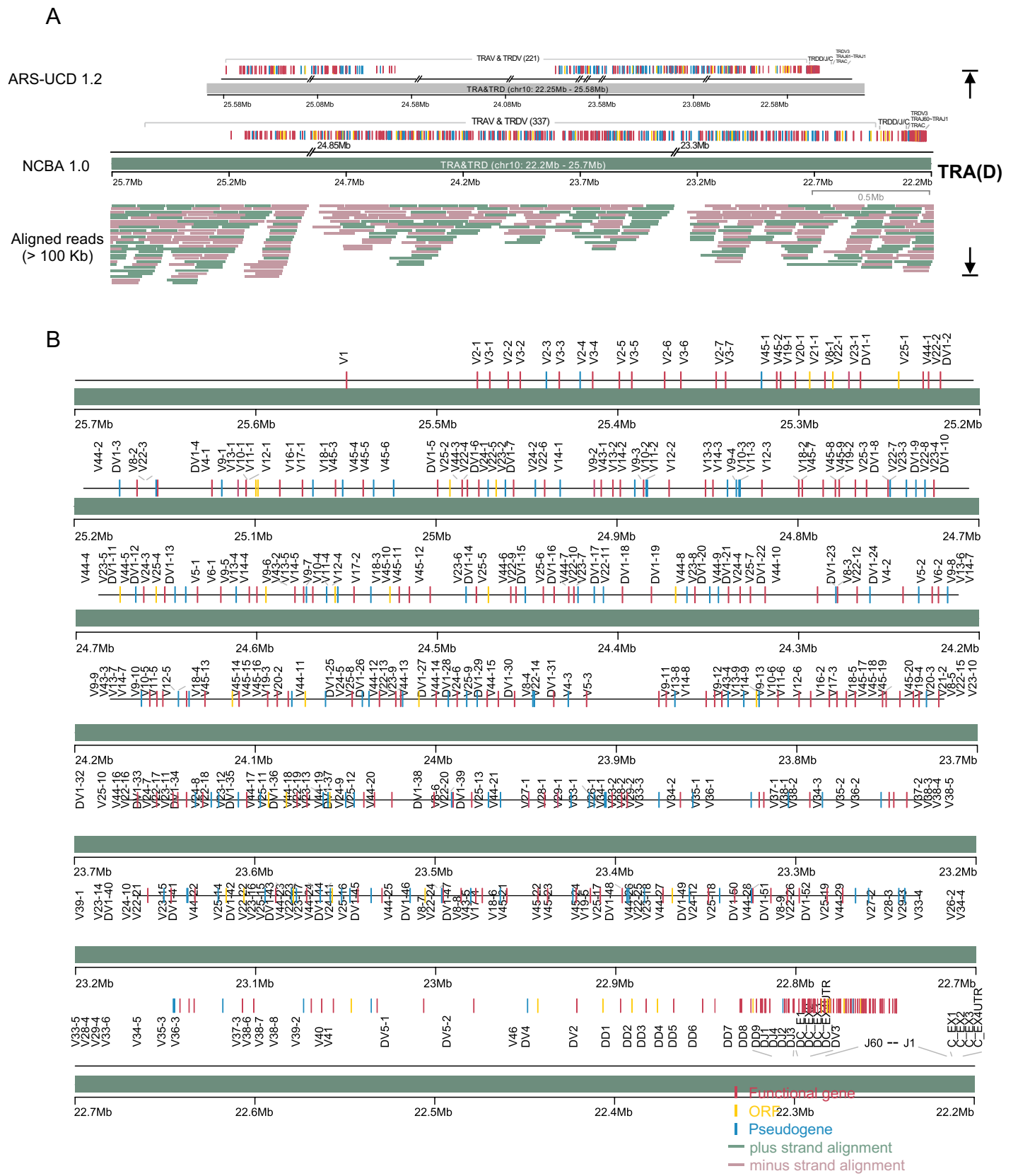


Figure 4. The Cattle TRB and TRG Loci

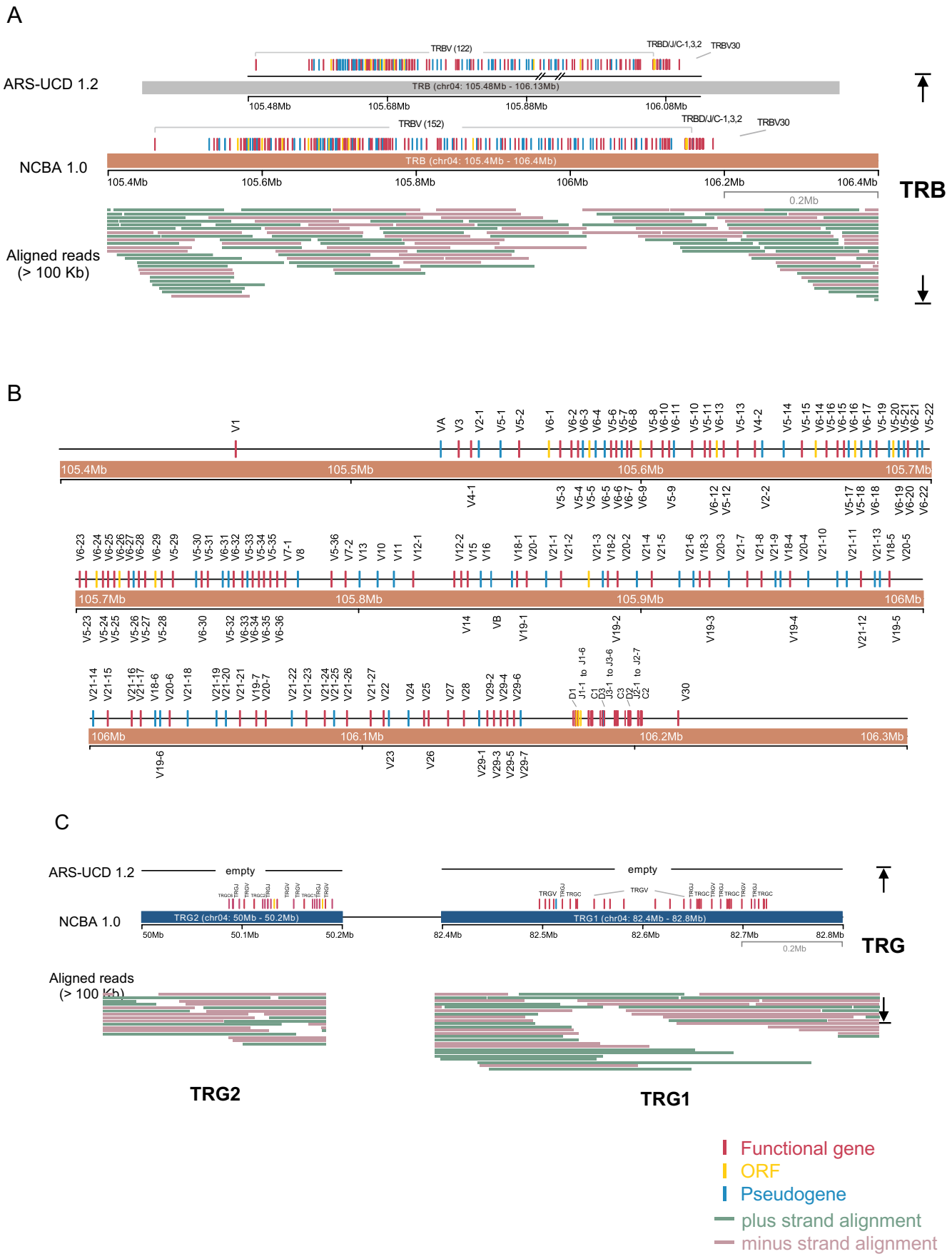


Figure 5. Phylogenetic Analysis of V Genes

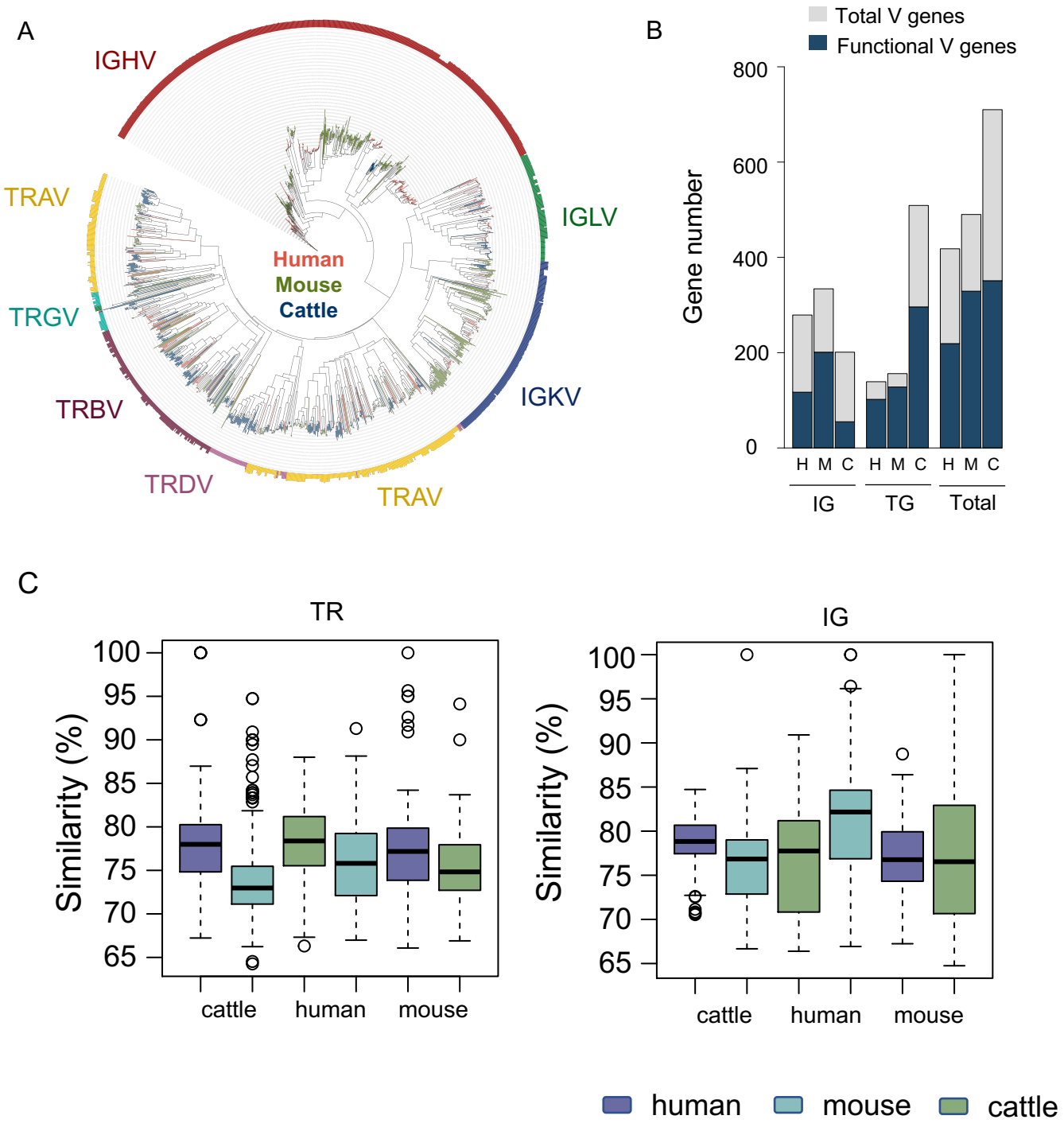


Figure 6. MHC Gene Locus and Haplotyping

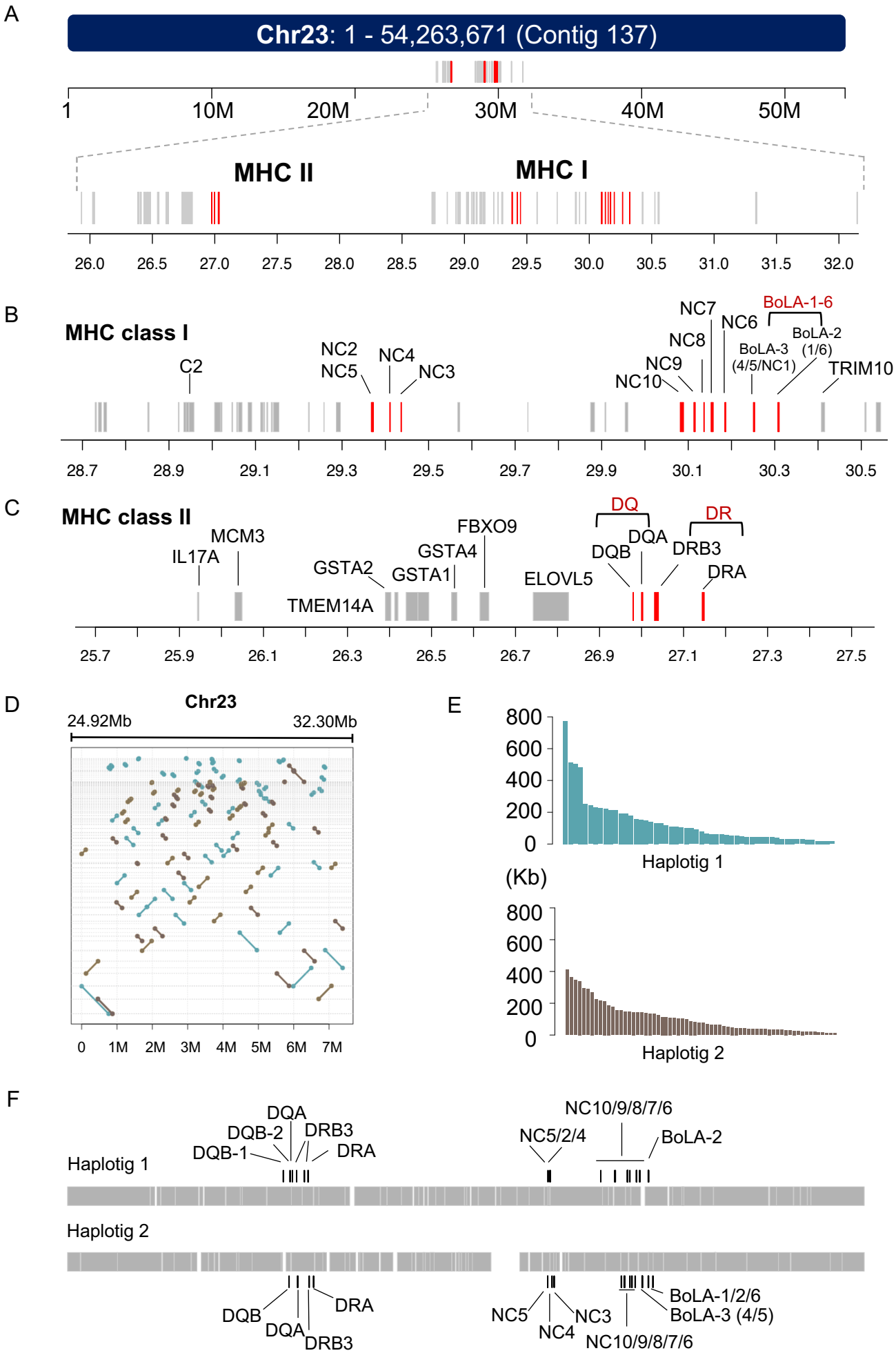


Figure 7. Telomere Length and satDNAs Distributions

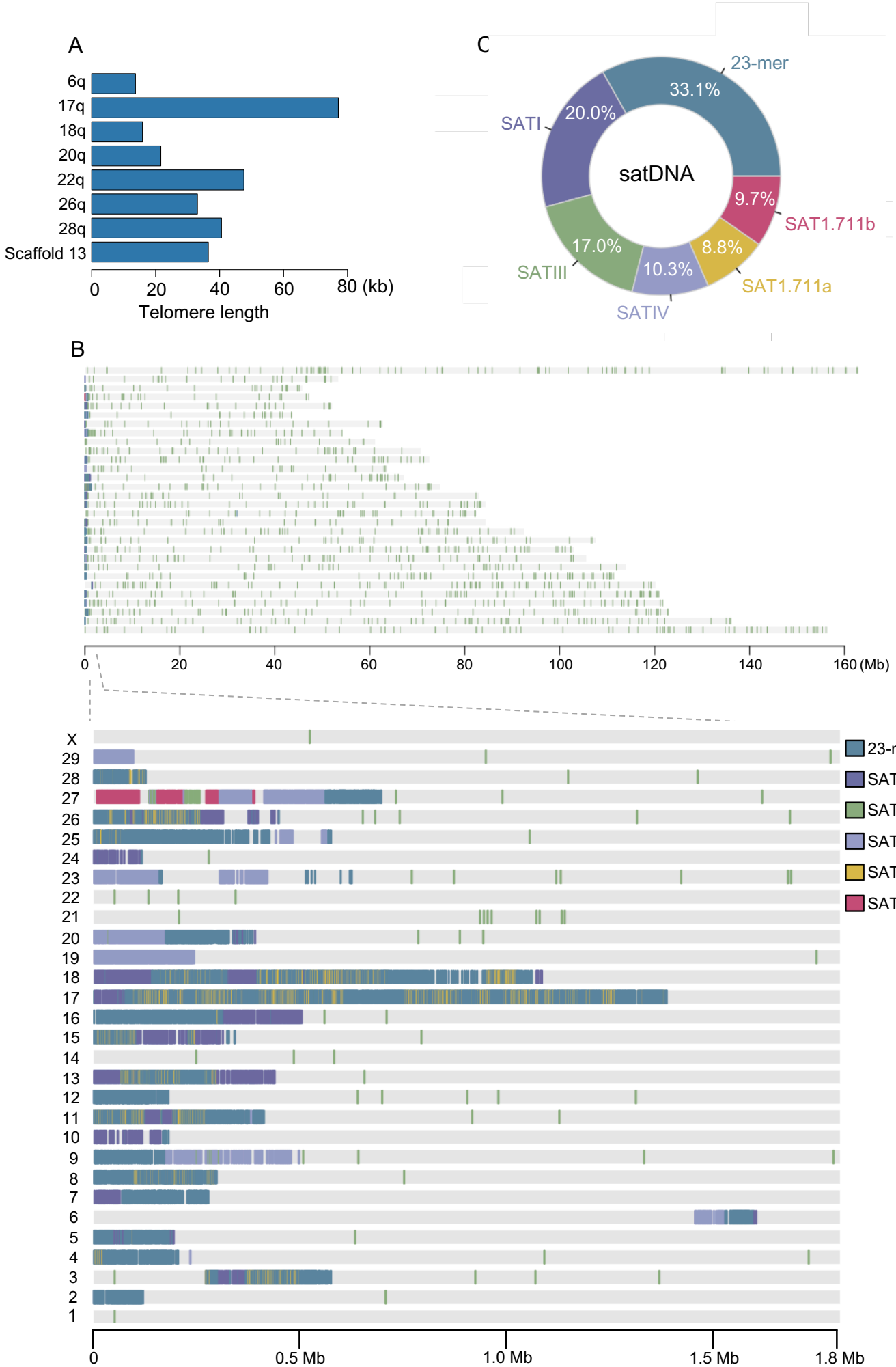
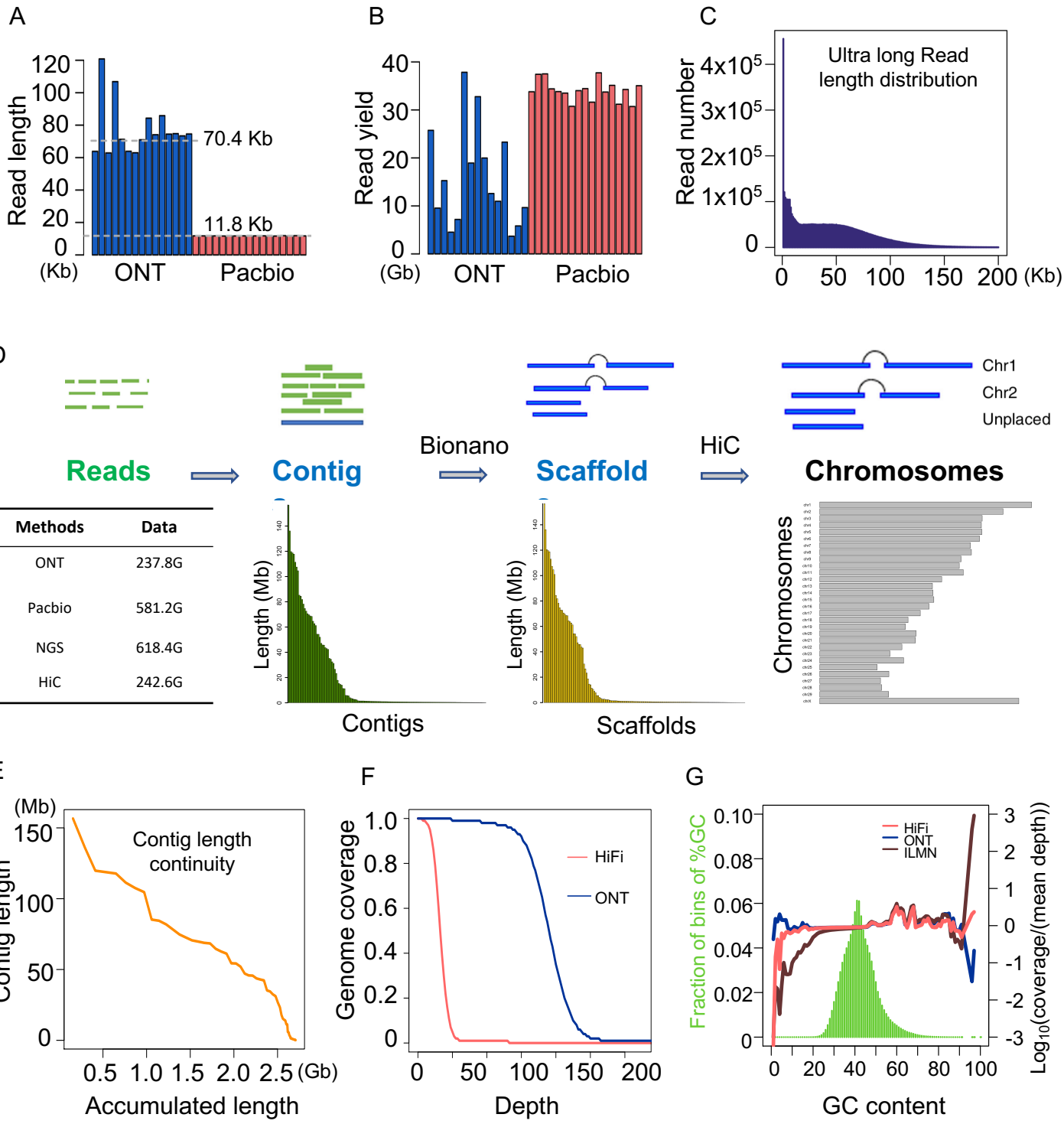


Table 1. Gene Numbers of Each Immune Locus in NCBA 1.0 Assembly

Types	IGH	IGK	IGL	TRA	TRB	TRD	TRG1	TRG2	Sum
V (F/P/ORF)	48 (11/37/0)	28 (7/19/2)	125 (37/80/8)	281 (148/108/25)	153 (87/55/11)	57 (46/7/4)	14 (13/1/0)	4 (4/0/0)	710 (353/307/50)
D (F/P/ORF)	17 (17/0/0)	NA	NA	NA	3 (3/0/0)	9 (6/0/3)	NA	NA	29 (26/0/3)
J (F/P/ORF)	12 (3/1/8)	5 (1/0/4)	6 (4/0/2)	60 (53/2/5)	19 (15/1/3)	4 (3/0/1)	5 (5/0/0)	5 (3/0/2)	116 (87/4/25)
C (F/P/ORF)	10 (8/2/0)	1 (1/0/0)	6 (3/3/0)	1 (1/0/0)	3 (3/0/0)	1 (1/0/0)	4 (4/0/0)	3 (3/0/0)	29 (24/5/0)
Sum (F/P/ORF)	87 (39/40/8)	34 (9/19/6)	137 (44/83/10)	342 (202/110/30)	178 (108/56/14)	71 (56/7/8)	23 (22/1/0)	12 (10/0/2)	

Figure S1. Data Summary of Cattle Genomic Assembly, Related to Figure 1



Methods	Data
ONT	237.8G
Pacbio	581.2G
NGS	618.4G
HiC	242.6G

H

Assembly	Sequencing methods	Contig N50 (Mb)	No. of contigs	No. of Scaffold	Genome length (Gb)
Cattle.NCBA 1.0	ONT ultra long; PacBio	74.7	145	154	2.71
ARS-UCD 1.2 (GigaScience, 2020)	PacBio	12	3077	2511	2.72
UOA_WB_1, Water buffalo (Nat Commun, 2019)	PacBio	18.8	953	509	2.65
ARS_Simm 1.0 (J Hered, 2021)	ONT	70.8	1374	1315	2.86

Figure S2. Properly Placed Scaffolds of ARS-UCD1.2 In NCBA1.0, Related to Figure 1

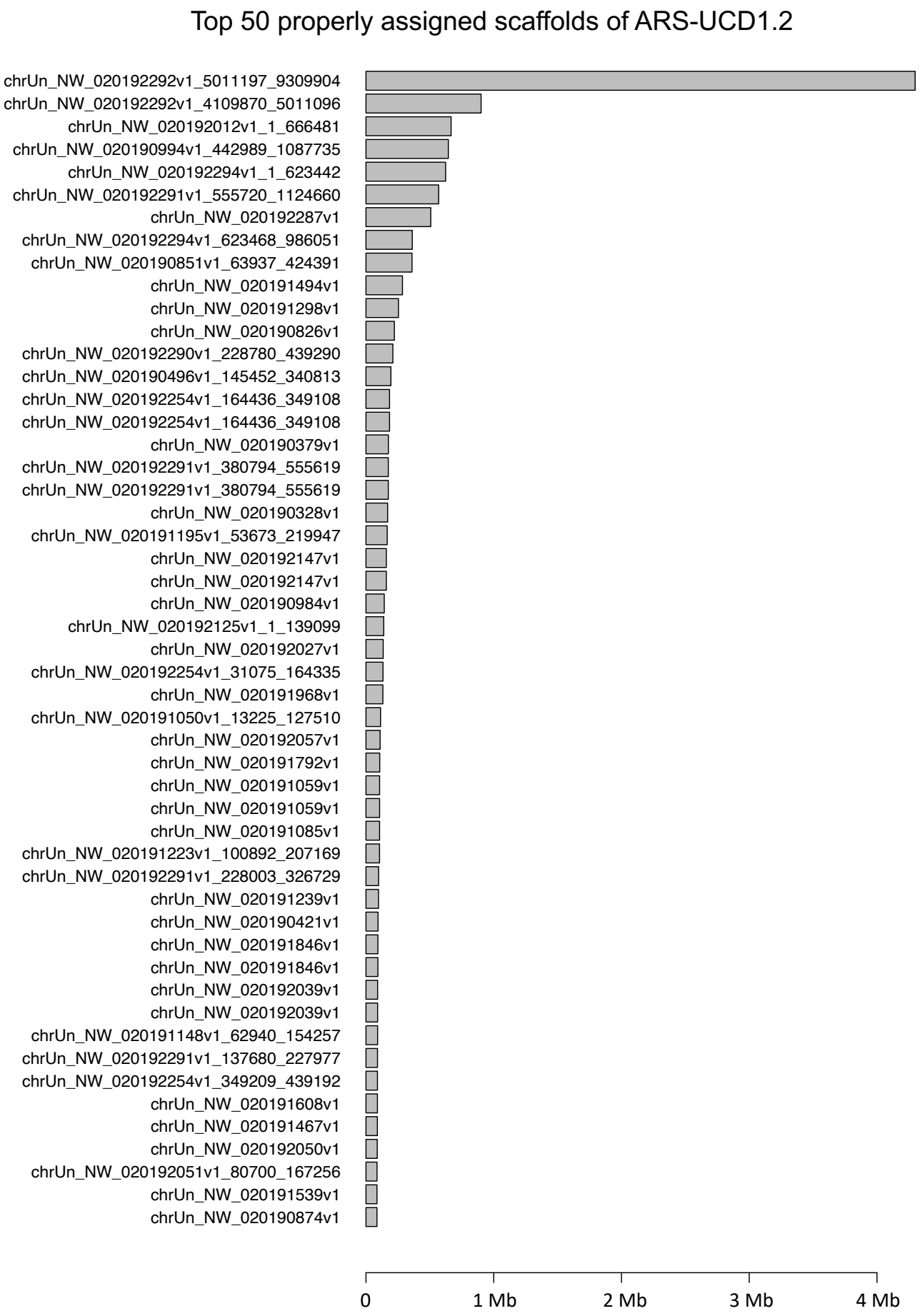


Figure S3. Gene Length Distributions of Five Species, Related to Figure 1

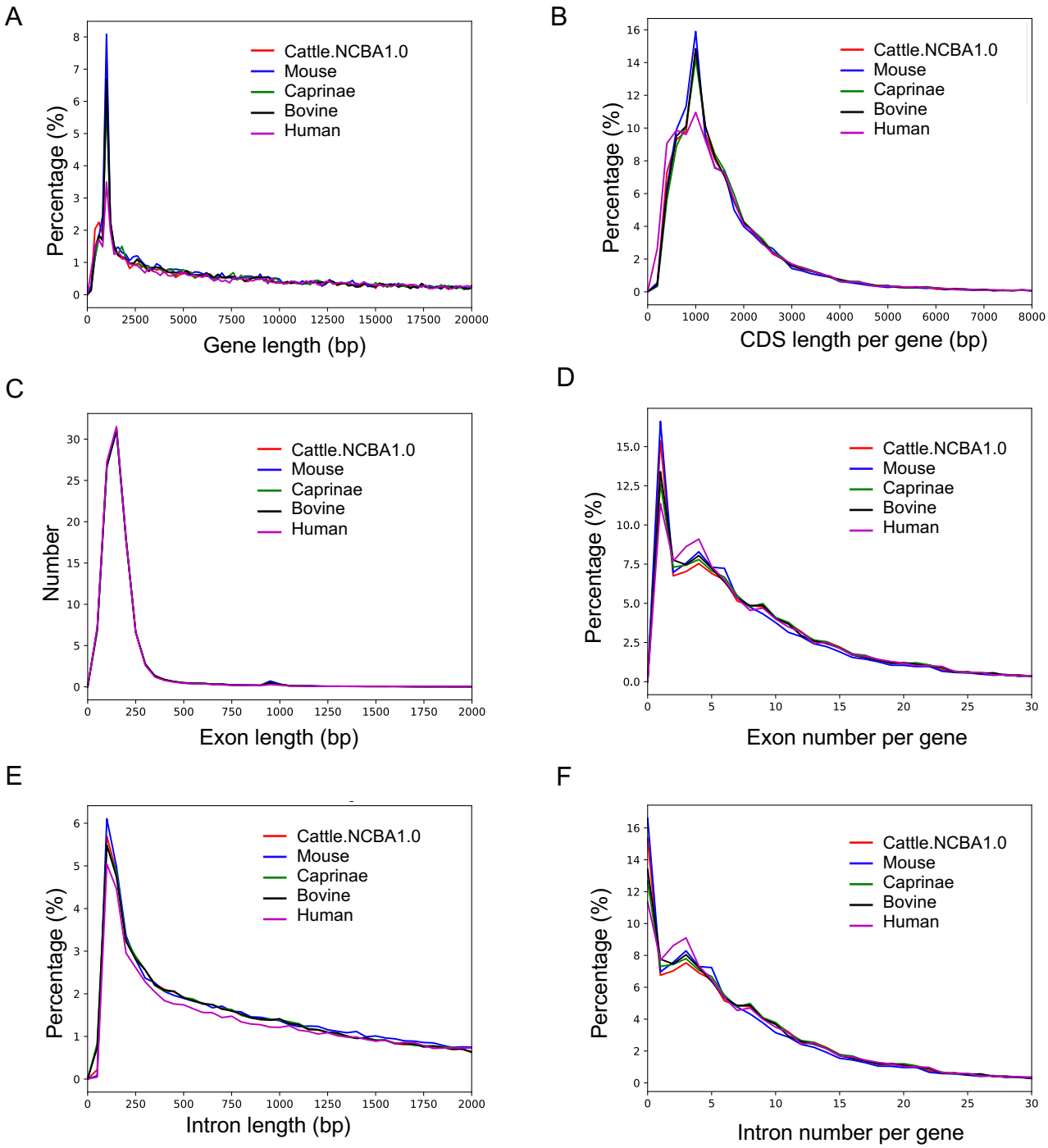
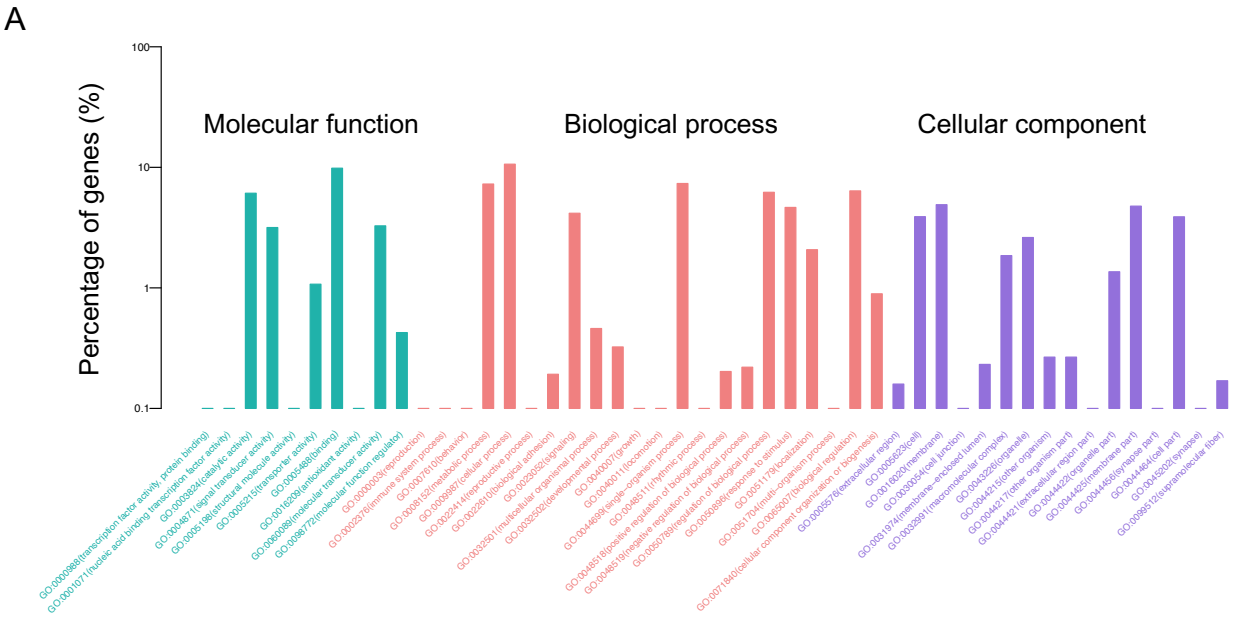
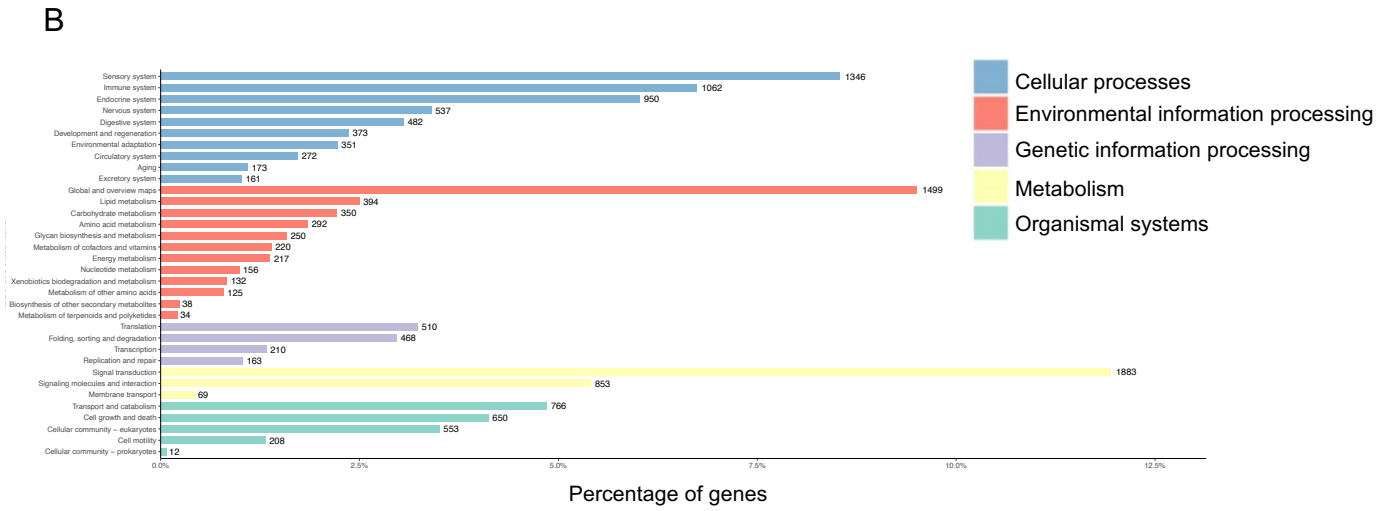


Figure S4. Functional Annotation of Predicted Genes In NCBA1.0, Related to Figure 1





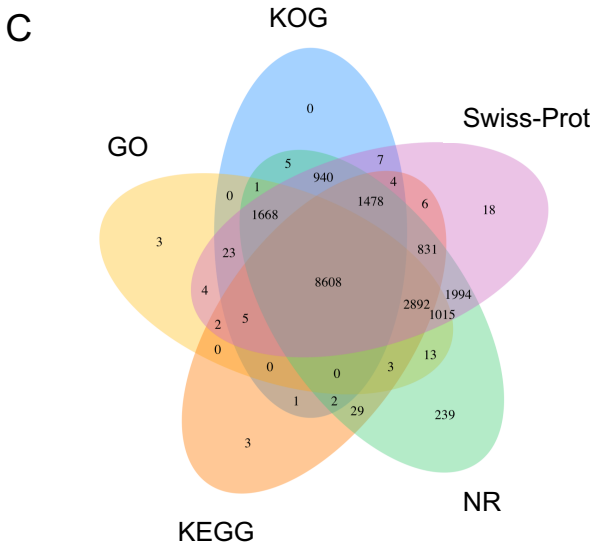


Figure S5. Immune Gene Loci in NCBA1.0 Assembly, Related to Figure 2-4

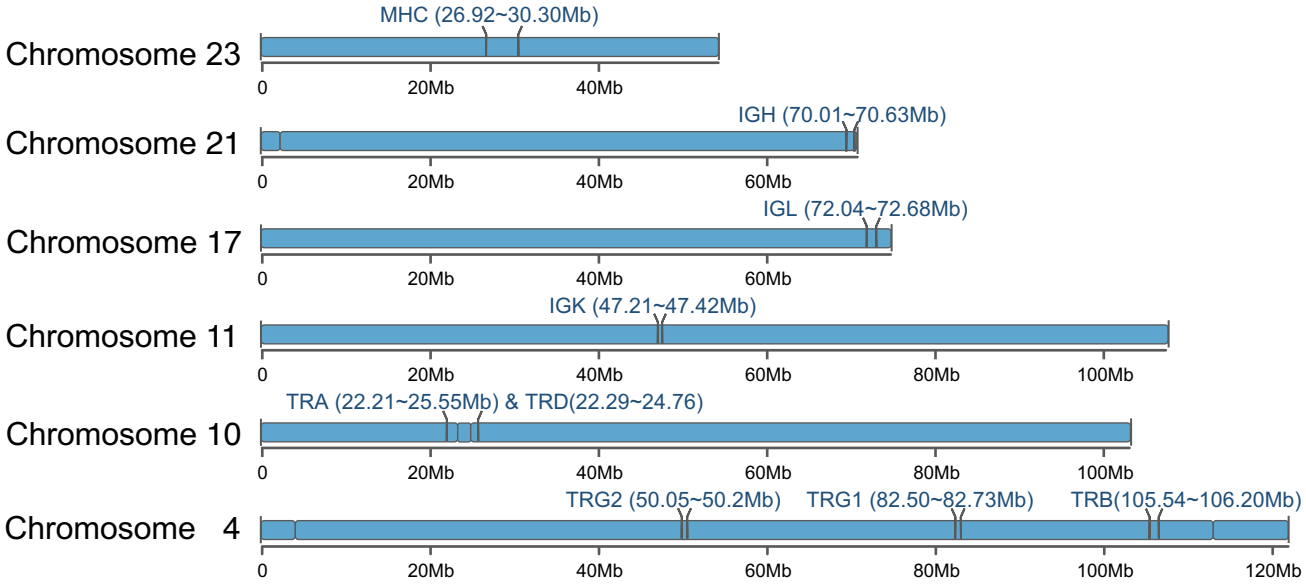
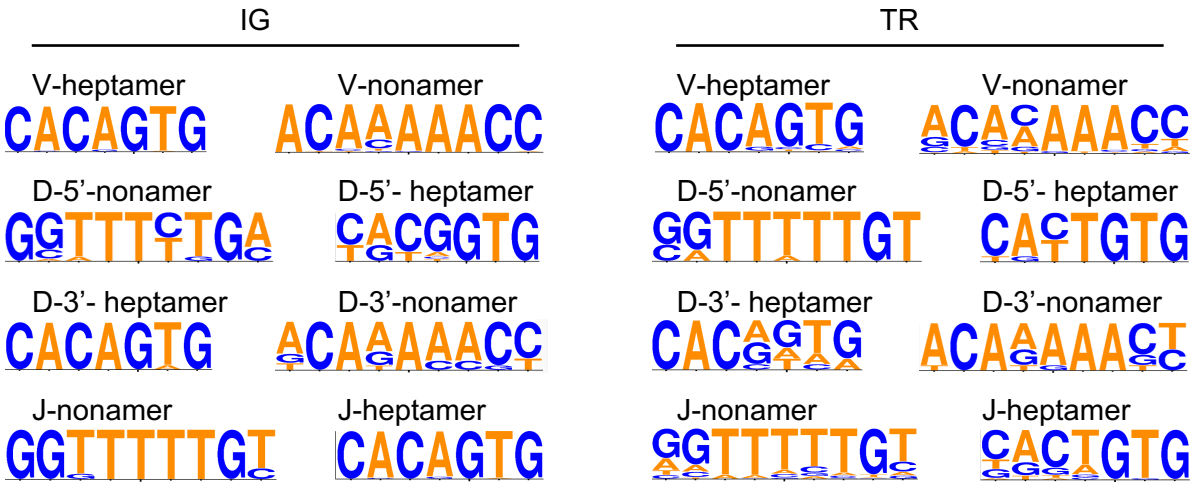


Figure S6. Criteria For Gene Structure and Functionality Annotation, Related to Figure 2-4

A



B

Functional gene	L-PART	Gene body	RS signal
V	• L-1 starts with ATG; • Length of (L1 + L2) is 3N; • Length of V-Intron < 500 nt; • No stop codon; • L-1 donor splice site is "GT"; • L-2 acceptor splice site is "AG"	• 1st-CYS in the 23rd AA, 2nd-CYS in the 104th AA, and TRP in the 41th AA (or a nearby position); • No stop codon(except the last codon in CDR3 region); no frameshift; • No long deletions in CDR1/2 region	• Canonical nonamer and heptamer with no more than 1 mismatch; • The size variation of X-Spacer length less than 3
D	NA	• No requirement	• Canonical nonamer and heptamer with no more than 1 mismatch; • The size variation of X-Spacer length less than 3
J	NA	• FGXG motif in light chain and WGXG motif in heavy chain,	• Canonical nonamer and heptamer with no more than 1 mismatch; • The size variation of X-Spacer length less than 3
C	NA	• No stop codon and no frameshift	NA

Pseudogene	L-PART	Gene body	RS signal
V	• Stop codon or frameshift; • No L-1 or L-2; • No INIT codon; • No L1 donor splice site; • No L2 acceptor splice site	• Stop codon or frameshift; • A part of V-region is missing	• No nonamer and/or heptamer
D	NA	NA	• No nonamer and/or heptamer
J	NA	NA	• No nonamer and/or heptamer
C	NA	• Stop codon or frameshift	NA

Open Reading Frame (ORF)	L-PART	Gene body	RS signal
V	• Noncanonical L1 donor splice site; • Noncanonical L2 acceptor splice site; • Unexpected V-Intron length	• Mutated 1st-CYS in the 23rd AA, mutated 2nd-CYS in the 104th AA, or mutated TRP in the 41th AA (or a nearby position); • Deletion of more than 3AA in V-region	• Noncanonical nonamer and/or heptamer; • Unexpected X-Spacer length;
D	NA	NA	• Noncanonical nonamer and/or heptamer; • Unexpected X-Spacer length;
J	NA	• Mutated FGXG motif in light chain or mutated WGXG motif in heavy chain,	• Noncanonical nonamer and/or heptamer; • Unexpected X-Spacer length;
C	NA	NA	NA

Figure S7. Ultra-long Reads Coverage of IGH Locus, Related to Figure 2

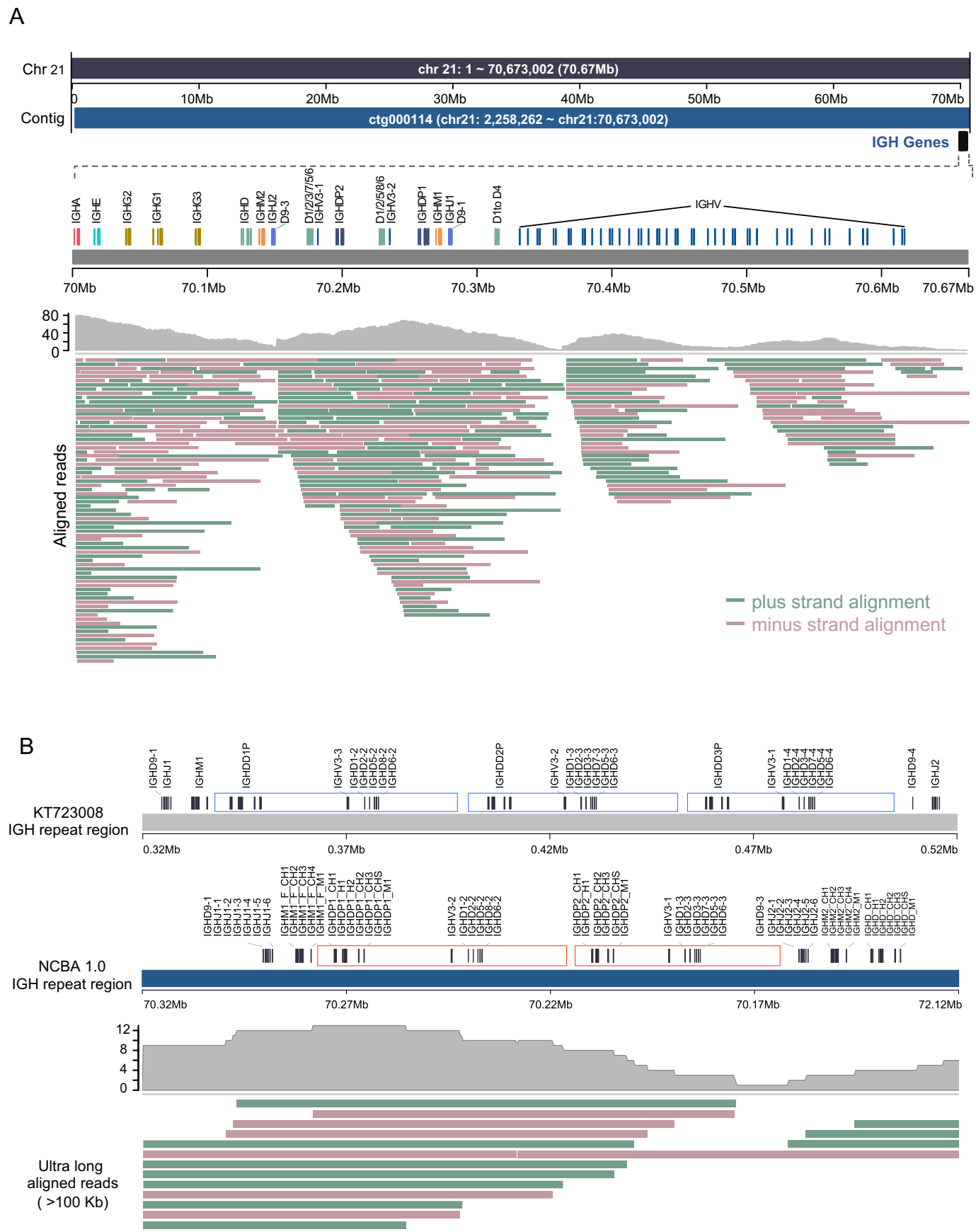


Figure S8. Dot Plots between Two IGH Genomic Sequences, Related to Figure 2

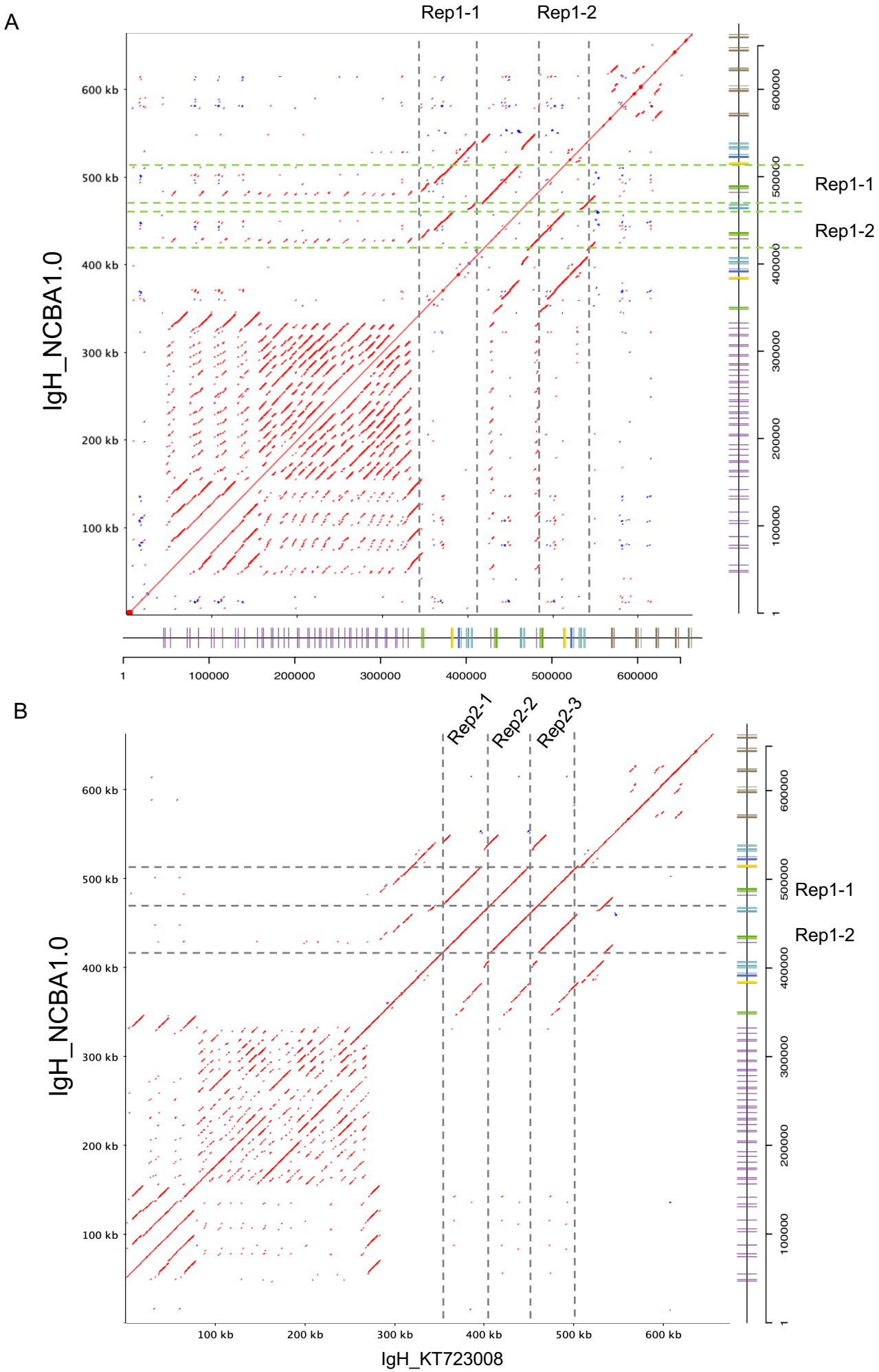


Figure S9. Detailed Annotation Map of IGL Locus, Related to Figure 2

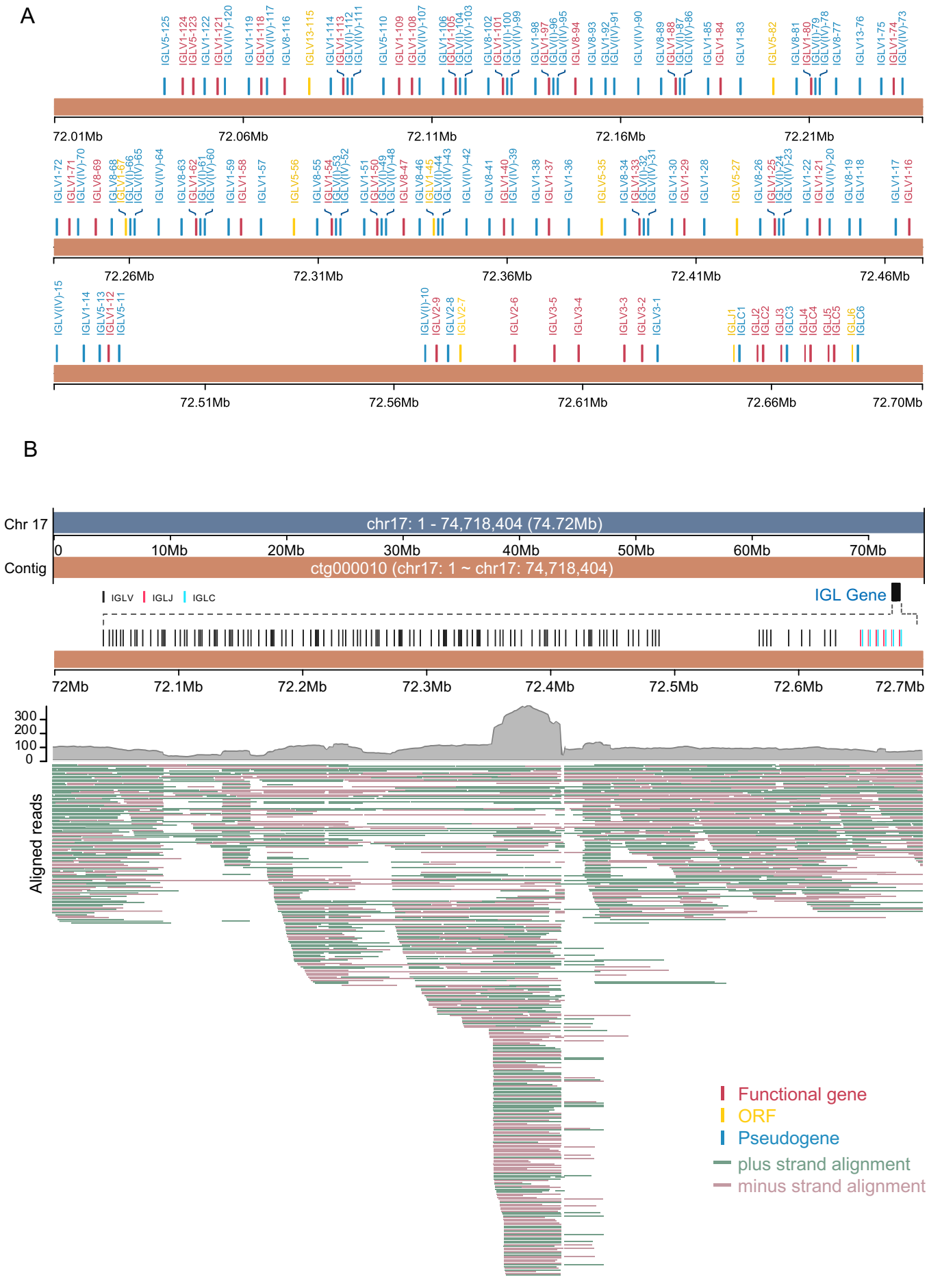


Figure S10. Enlarged Alignment Map of IGL J-C Cluster Region, Related to Figure 2

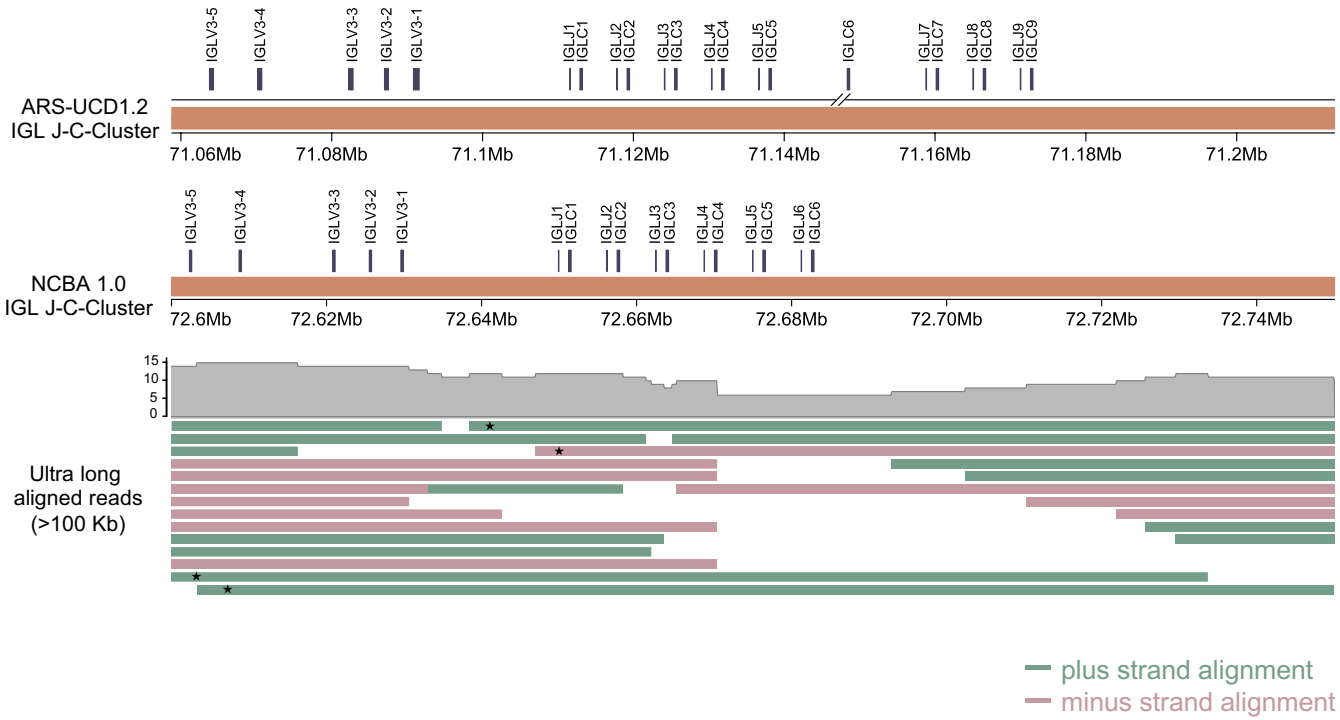


Figure S11. Detailed Annotation Map of IGK Locus, Related to Figure 2

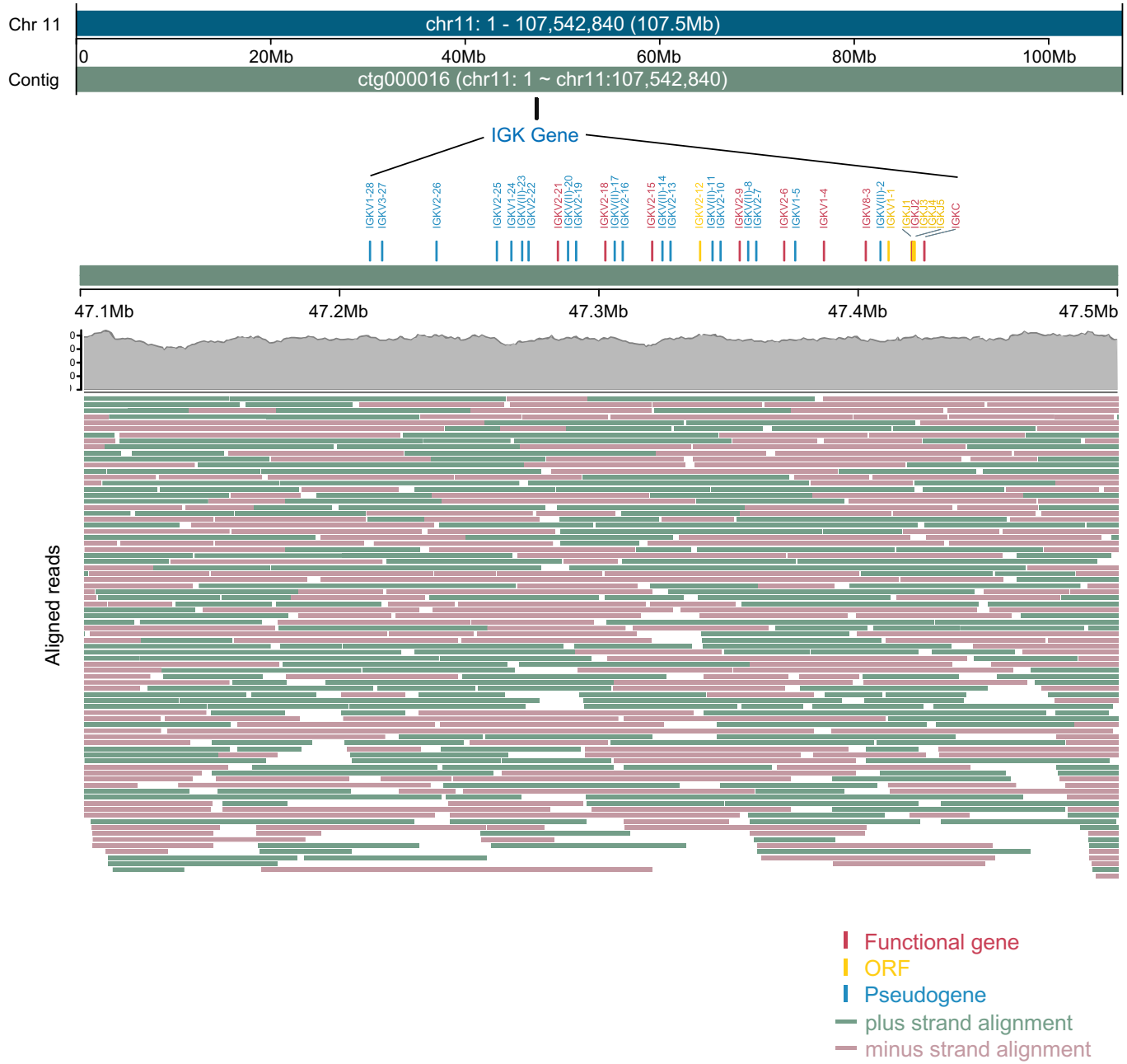


Figure S12. Global Genetic Map of TRA(D) Loci, Related to Figure 3

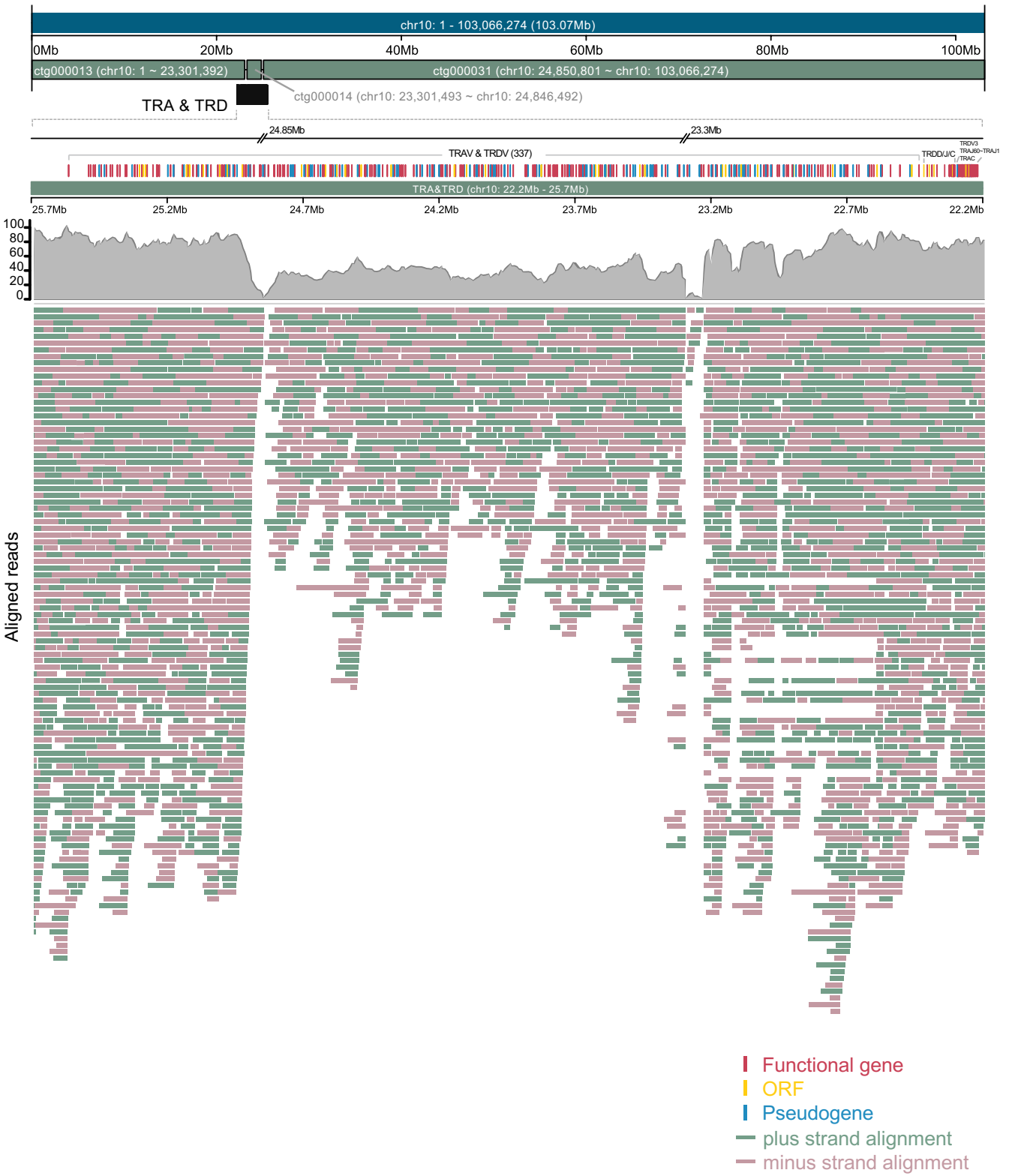
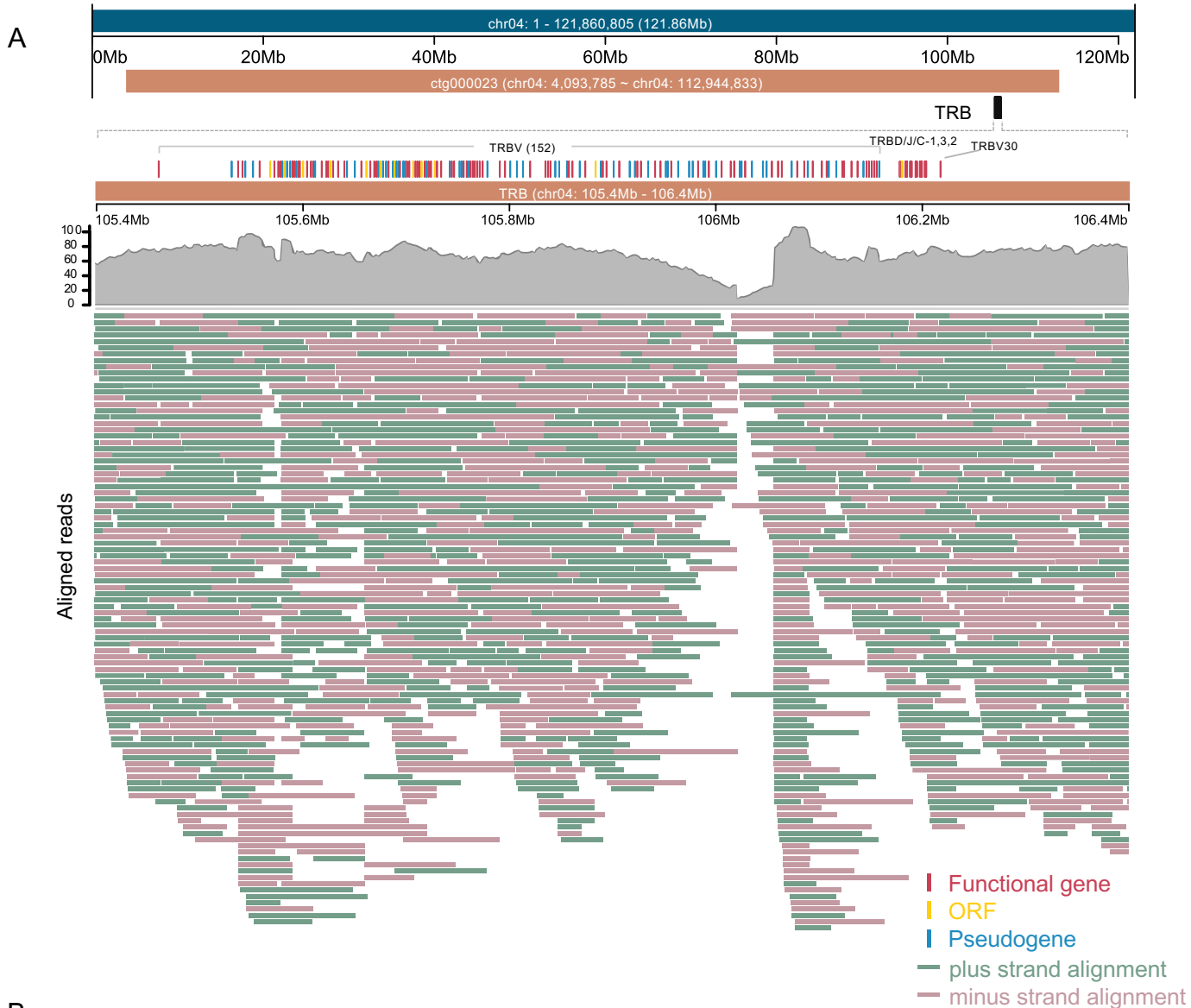


Figure S13. Global Alignment Map of TRB and TRG Loci, Related to Figure 4



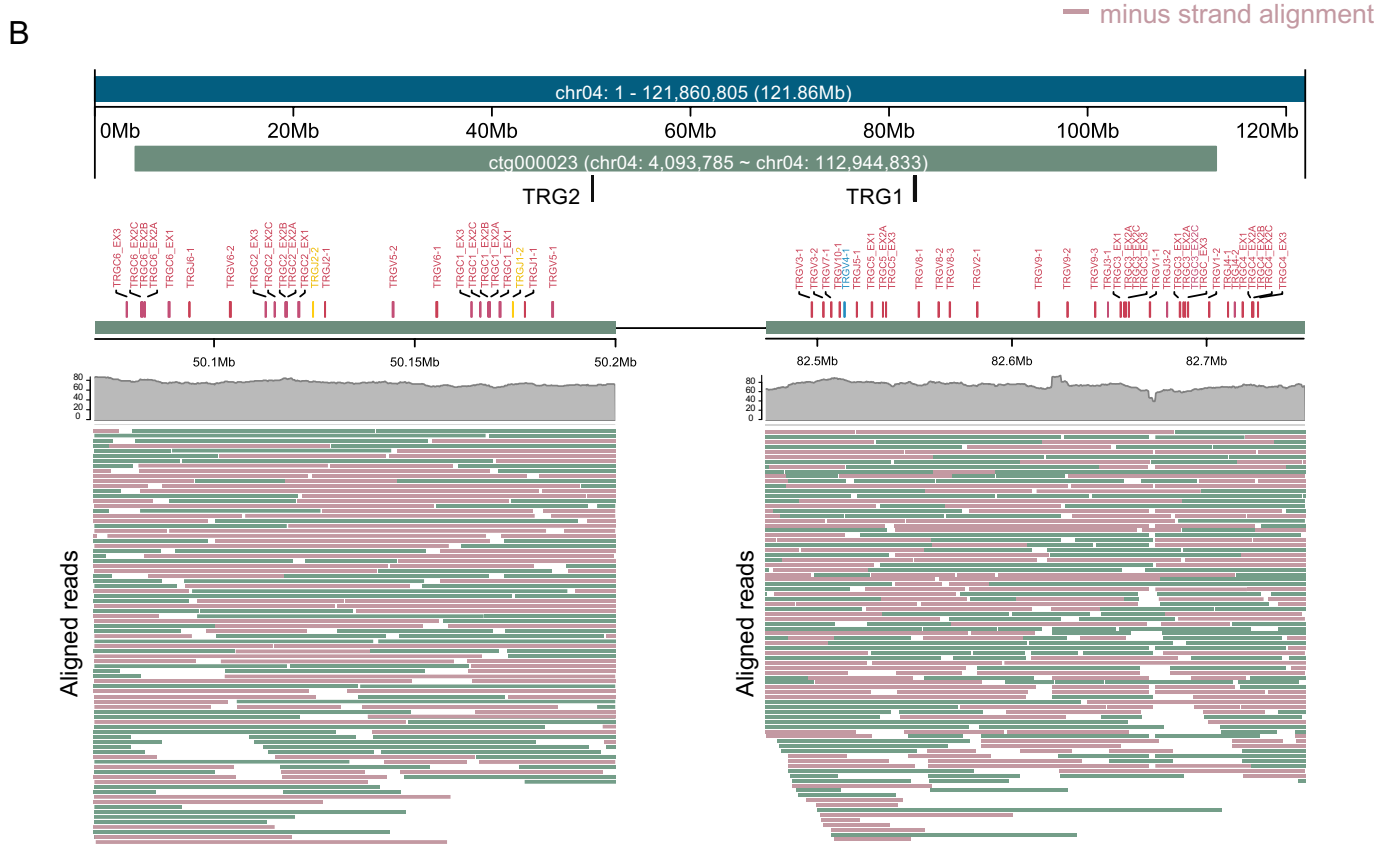


Figure S14. Gene Statistics of Cattle for Each Immune Loci in IMGT, Related to Figure 2-4

Types	IGH	IGK	IGL	TRA	TRB	TRD	TRG1	TRG2	Sum
V (F/P/ORF)	47 (12/35/0)	25 (6/17/2)	74 (22-25/ 5/44-47)	183 (79-92/ 74-87/14-21)	123 (62/52/9)	55 (45/5/5)	13 (13/0/0)	4 (4/0/0)	524
D (F/P/ORF)	23 (20-21/ 0/2-3)	NA	NA	NA	3 (3/0/0)	9 (6/0/3)	NA	NA	35
J (F/P/ORF)	12 (3/2/7)	5 (1/0/4)	8 (5/0/3)	60 (52-54/4-6/2)	19 (15/1/3)	4 (3/0/1)	4 (4/0/0)	5 (5/0/0)	117
C (F/P/ORF)	11 (8/3/0)	1 (1/0/0)	9 (4/5/0)	1 (1/0/0)	3 (3/0/0)	1 (1/0/0)	4 (3/1/0)	3 (3/0/0)	33
Sum	93	31	91	244	148	69	21	12	

Figure S15. IG and TR Statistics of Different Species in IMGT, Related to Figure 5

loci		human	mouse	cattle	NCBA1.0	sheep	dog	rabbit	chicken	cat	horse
IGH	V	123-129	152	47	48	10	89	69	94	NA	104
	D	27	17-20	23	17	4	6	11	4	NA	44
	J	9	4	12	12	6	6	6	1	NA	9
	C	11	8-9	11	10	6	8	17	3	NA	11
IGL	V	73-74	8	74	125	121	261	43	34	8	NA
	J	7-11	5	8	6	2	9	4	1	12	NA
	C	7-11	4	9	6	2	9	6	1	12	NA
IGK	V	76	174	25	28	18	25	68	NA	18	66
	J	5	5	5	5	4	5	8	NA	5	5
	C	1	1	1	1	1	1	2	NA	1	1
TRA	V	54	98	183	281	277	58	62	NA	63	NA
	J	61	60	60	60	79	59	58	NA	64	NA
	C	1	1	1	1	1	1	1	NA	1	NA
TRB	V	64-67	35	123	153	94	36	77	NA	27	NA
	D	2	2	3	3	3	2	2	NA	2	NA
	J	14	14	19	19	19	12	12	NA	12	NA
	C	2	2	3	3	3	2	2	NA	2	NA
TRD	V	8	16	55	57	70	5	4	NA	10	NA
	D	3	2	9	9	9	2	2	NA	2	NA
	J	4	2	4	4	4	4	3	NA	5	NA
	C	1	1	1	1	1	1	1	NA	1	NA
TRG	V	12-15	7	17	18	13	16	11	NA	12	NA
	J	5	4	9	10	13	16	2	NA	12	NA
	C	2	4	7	7	6	8	1	NA	6	NA

Figure S16. Genomic Assembly and Haplotyping of MHC Locus, Related to Figure 6

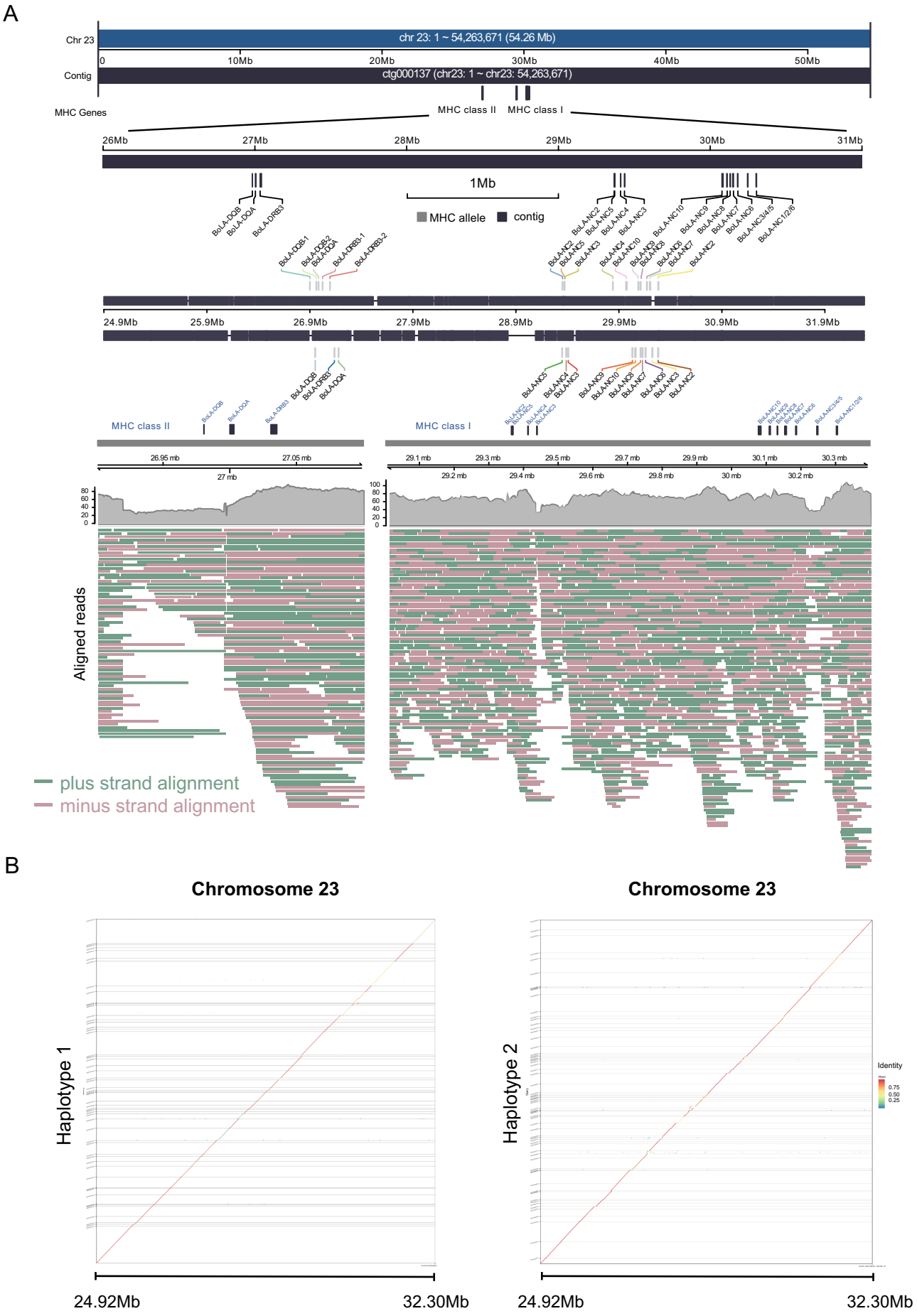


Figure S17. Distributions of satDNAs in Scaffolds, Related to Figure 7