Profiling Germline Adaptive Immune Receptor Repertoire with gAIRR Suite

Mao-Jan Lin[1,+], Yu-Chun Lin[2,+], Nae-Chyun Chen[3,+], Allen Chilun Luo[2], Sheng-Kai Lai[4], Chia-Lang Hsu[2,5,6], Jacob Shujui Hsu[2], Chien-Yu Chen[7], Wei-Shiung Yang[2,8,9,10,11], and Pei-Lung Chen[1,2,4,8,9,10,*]

[1]Department of Medical Genetics, National Taiwan University Hospital, Taipei, 10002, Taiwan
[2]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University, Taipei, 10617, Taiwan
[3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States
[4]Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei, 10617, Taiwan
[5]Graduate Institute of Oncology, School of Medicine, National Taiwan University, Taipei, 10617 Taiwan
[6]Department of Medical Research, National Taiwan University Hospital, Taipei, 10002, Taiwan
[7]Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan
[8]Division of Endocrinology and Metabolism, Department of Internal Medicine, National Taiwan University Hospital, Taipei, 10002, Taiwan
[9]Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, 10617, Taiwan
[10]Research Center for Developmental Biology and Regenerative Medicine, National Taiwan University, Taipei, 10617, Taiwan
[11]Department of Medicine, College of Medicine, National Taiwan University, Taipei, 10617, Taiwan
[+]these authors contributed equally to this work
[*]e-mail: paylong@ntu.edu.tw

ABSTRACT

The genetic profiling of germline Adaptive Immune Receptor Repertoire (AIRR), including T cell receptor (TR) and immunoglobulin (IG), is imaginably relevant to numerous immune-related conditions, but currently insurmountable due to high genetic diversity and complex recombination. In this study, we developed the gAIRR Suite comprising three modules. gAIRR-seq, a probe capture-based targeted sequencing pipeline, profiles genomic sequences of TR and IG from individual DNA samples. The computational pipelines gAIRR-call and gAIRR-annotate call alleles from gAIRR-seq reads and whole-genome assemblies. We applied gAIRR-seq and gAIRR-call to genotype TRV and TRJ alleles of Genome in a Bottle (GIAB) DNA samples with 100% accuracy. gAIRR-annotate profiled the alleles of 58 high-quality haplotype assemblies from 12 samples and further discovered 110 novel TRV alleles and 17 novel TRJ alleles. We validated a 65-kbp and a 10-kbp structural variant for HG002 on chromosomes 7 and 14, where TRD and J alleles reside. We also uncovered the disagreement of the human genome GRCh37 and GRCh38 in the TR regions; GRCh37 possesses a 270 kbp inversion and a 10 kbp deletion in chromosome 7 relative to GRCh38. The gAIRR Suite might benefit genetic study and future clinical applications for various immune-related phenotypes.

Introduction

Adaptive Immune Receptor Repertoire (AIRR)

The Adaptive Immune Receptor Repertoire (AIRR) is a collection of T-cell receptors (TR) and B-cell receptors, also known as immunoglobulins (IG), which comprise the adaptive immune system. The antigen-specific receptors are composed of three genetic regions, variable (V), diversity (D), and joining (J). The receptors' expression requires somatic V(D)J recombination, where one gene from each V, (D), and J genetic region are randomly selected and rearranged into a continuous fragment[1];[2]. Thanks to the extensive combination possibility from the V, (D), and J pool, the immune cell for recognizing foreign antigens is enormously diverse[3];[4];[5].

Although the human immune repertoire has theoretically extreme huge variability capable of binding all kinds of antigens, for each individual, the number of V(D)J genes/alleles in their genome is limited[3]. Therefore, any given individual's immunological response may be restricted according to the germline AIRR composition[3]. For example, the response to the influenza vaccine was clearly demonstrated to be modulated by *IGHV1-69* polymorphism[6],[7]. Another example is that the *IGHV3-23*03* allele binds Haemophilus influenzae type b polysaccharide more effectively than the most frequent allele,

*IGHV3-23*01*[3]. An additional example can be found that anti-HIV antibodies have biased usage of *IGHV1-69*[3]. Recently, the famous carbamazepine-induced Stevens-Johnson syndrome[8] has an additional twist[9]. In addition to the well-known HLA-B*15:02 being a disease-causing gene, drug-specific public TR was also critical for these severe cutaneous adverse reactions[9]. In a simian immunodeficiency virus infection model, the TR alleles response for the potent CD8+ T cell response was recently determined[10]. Germline AIRR alleles have started to be found associated with human diseases, including celiac disease[11], Alzheimer disease[12], rheumatic heart disease[13], and Kawasaki disease[14];[15]. However, the known examples of germline AIRR effects on human immune-related phenotypes and diseases are likely only the tip of the iceberg because determining germline AIRR for any given individual is currently insurmountable.

According to their functionality, V, D, and J genes can be further divided into a functional gene, open reading frame (ORF), or pseudogene. The functional gene's coding region has an open reading frame without stop codons, defects on splicing sites, recombination signals, or regulatory elements. A gene is classified as ORF if its coding region has an open reading frame but with alterations in the splicing sites, recombination signals, or regulatory elements. A gene whose coding region has a stop codon(s), frameshift mutation(s), or any mutation that affects the initiation codon will be defined as a pseudogene[16].

Both genomic DNA (gDNA) and messenger RNA (mRNA) can provide the AIRR information of an individual. Both of them are suitable for library construction but with advantages and limitations. mRNA provides useful dynamic information of AIRR clonotypes and expression levels, which can be extremely valuable if, and only if, the most relevant lymphocytes can be retrieved in the right tissue at the right time. mRNA-based AIRR sequencing can only cover part of the full spectrum of germline AIRR. On the other hand, gDNA includes non-rearranged and rearranged (productive and non-productive) V(D)J segments and, when applied to regular bulk gDNA, preserves the full spectrum of germline AIRR.[17]. Besides, gDNA is more stable and easier for storage compared to mRNA. Tens of thousands of gDNA samples have been collected, preserved, and readily available for germline AIRR profiling. In general, since the information of the V(D)J fragment contained in gDNA or mRNA is different, the choice of material depends on the research goal[17],[18].

Probe capture-based targeted sequencing: gAIRR-seq

Multiplex polymerase chain reaction (PCR) and probe capture-based target enrichment are two commonly used enrichment methods for gDNA and mRNA/cDNA samples. Multiplex PCR incorporates the mixture of primers specific to the known V and J genes (or C genes for mRNA/cDNA) to amplify the targeted regions. This approach has at least two concerns: the inconstancy of detecting novel alleles due to the specific primer sets and the amplification bias due to different efficiencies between primers[17],[18]. Thus, it is challenging to design PCR-based experiments that are completely immune to allele dropout.

The probe capture-based target enrichment approach begins with the fragmentation of starting materials, such as gDNA, which were then captured and enriched using biotinylated probes. After the standard next-generation sequencing (NGS) library preparation process, the enriched products are ready for sequencing to read the target regions and flanking regions[18]. The probe hybridization method can tolerate sequence mismatch to a certain degree. Hence, it has the potential to eliminate the biases in the multiplex PCR method and is more suitable for novel alleles discovery. In this study, we developed the gAIRR-seq pipeline, which integrates probe capture-based target enrichment of germline AIRR genes and NGS.

Computational tools: gAIRR-call and gAIRR-annotate

We developed a computational pipeline, gAIRR-call, to analyze the gAIRR-seq reads and call AIRR alleles (Figure 1). We designed a coarse-to-fine strategy for gAIRR-call to accurately call both known and novel AIRR alleles, as well as their flanking sequences. The pipeline starts from the coarse phase, where gAIRR-call identifies potential alleles by aligning reads to known AIRR alleles collected from the international ImMunoGeneTics (IMGT) database[19]. We use error-tolerant alignment in this phase and identify both known and novel allele candidates. In the fine phase, we re-align the reads to the candidate alleles with no errors allowed and assign reads to each allele. We then summarize each allele's coverage (read depth) and apply an adaptive read-depth filter to determine final calls. We further extend the called alleles using the SPAdes assembler[20] and phase the contigs using local variants. Finally, we report phased alleles, as well as their 200 bp flanking sequences extended from both ends of the core allele.

One major challenge for germline AIRR studies is the lack of curated annotations and no ground truth for tool development. We designed, gAIRR-annotate to annotate AIRR alleles and flanking sequences using high-quality whole-genome assemblies (Figure 1). gAIRR-annotate starts from aligning all known AIRR alleles to an assembly. We annotate all alleles that are perfectly matched. For any allele aligned with mismatches, we identify its nearest allele in the IMGT database and assign a novel allele to the associated gene. When there are multiple assemblies available for the same reference material (RM), we only annotate an allele if more than half of the assemblies report the same output. We also annotate the sequences outside core allelic regions. By default, 200 bp flanking sequences extended from both ends of an allele are annotated. We report the order of each allele in the assembly, including the novel ones, which can reveal structural variants.

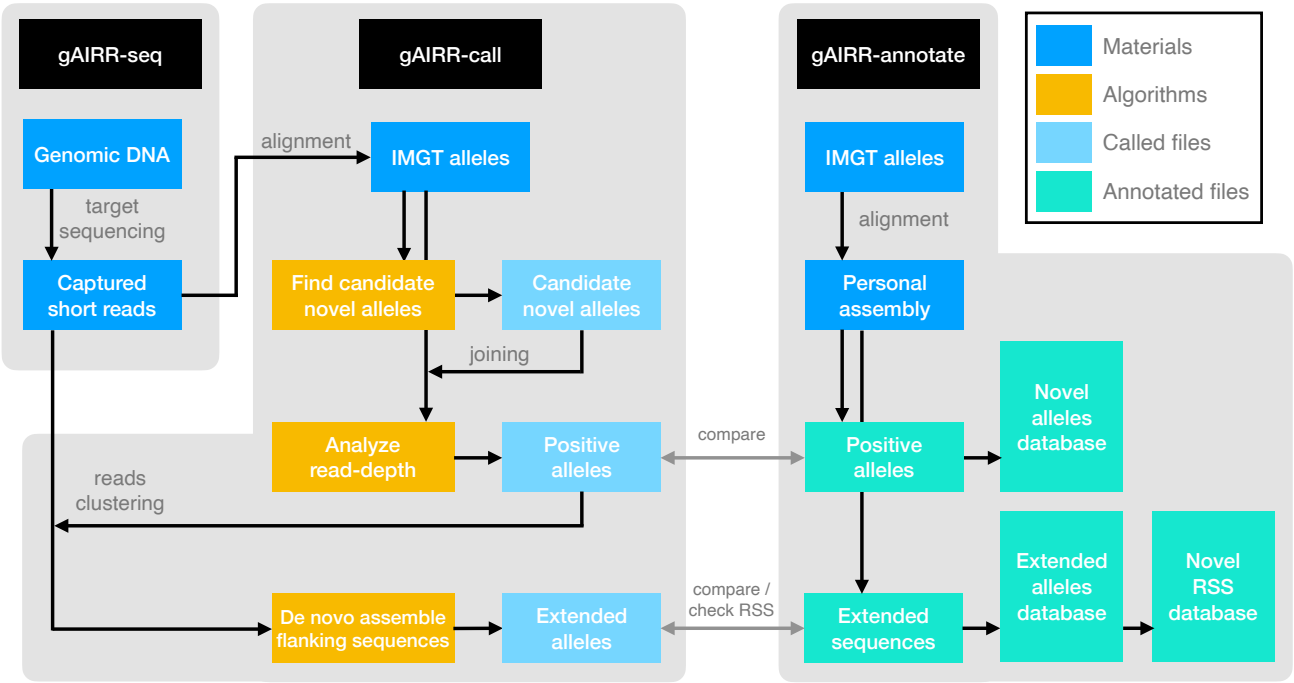


Figure 1. The gAIRR-call and gAIRR-annotate pipelines. The gray arrows show the verification methods between the two pipelines when both gAIRR-seq reads and personal assembly are available. (Methods: Comparing alleles of gAIRR-call and gAIRR-annotate)

Results

gAIRR-seq is effective and efficient in AIRR regions

Our gAIRR-seq yielded more than 280 thousand paired-end reads for each sample. We sequenced GIAB RMs HG001-7 (EBV-transformed B lymphocytes) and a primary cell sample using gAIRR-seq. There were 280-362 thousands of reads sequenced, with a mean of 310 thousands. To evaluate the sequencing quality, we aligned the gAIRR-seq reads of the GIAB RMs, HG001 and HG002, to their individual whole-genome assemblies[21]. Over 96% of the reads were successfully aligned to the assemblies. We located all TRV, TRJ, and BRV alleles in the assemblies with gAIRR-annotate to validate if the reads are in AIRR regions. Considering the fragment lengths are up to 800 bp in our sequencing method, we defined a read aligned within 800 bp from its allele boundaries to be on-target. Over 83% of the sequenced reads were on-target in HG001 and HG002 (Table S1). In core AIRR allelic regions where probes were designed, the average read depth was above 350. Despite the average read depth decreasing from central allelic regions to allelic boundaries, the sequencing depth at positions 200 bp away from the allele boundaries were typically above 200x (Figure 2). High coverage and high on-target rate make gAIRR-seq efficient and effective to enrich AIRR sequences and provide high-quality data for computational analyses.

We assessed the biases of probe preferences during library preparation using probe capture-based enrichment. The only bias we observed is an extremely long TRV pseudogene *TRAV8-5*. The gene length is 1355 bp, while the second-longest TRV gene is only 352 bp. The read-depth coverage of *TRAV8-5*, 270x, is relatively low compared to other V alleles, 405x on average. gAIRR-call tackles the uneven problem by making *TRAV8-5* a special case in adaptive threshold generation (Section Methods: The gAIRR-call pipeline). The lower read coverage of *TRAV8-5* is likely due to the high allele length-to-probe ratio. By adding more probes according to the allele length, the bias can probably be solved.

Calling known and novel germline TRV and TRJ alleles using gAIRR-call

We applied gAIRR-call using the high-coverage targeted reads from gAIRR-seq. We collected known alleles in the IMGT database and used them to identify candidate alleles. We define an allele call to be positive if the candidate allele has higher read coverage than a dataset-dependent adaptive threshold (Section Methods: The gAIRR-call pipeline). We further applied gAIRR-annotation to generate high-confidence annotations for both known and novel alleles using whole-genome assemblies (Section Results: Annotating TR alleles using high-quality assemblies). According to the annotations, gAIRR-call called TRV

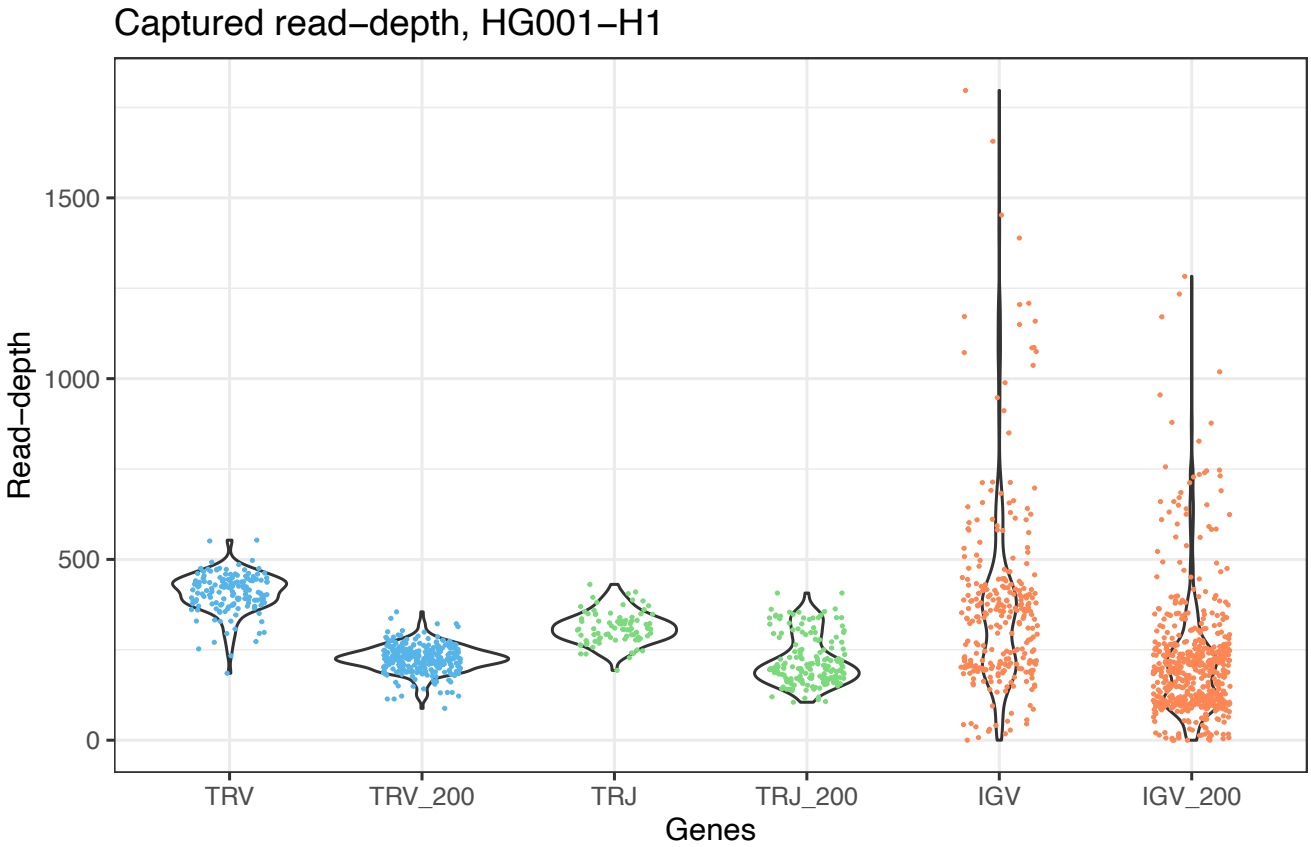


Figure 2. Read depths sequenced with gAIRR-seq in TRV, TRJ and BRV regions using data from HG001/NA12878. Columns without the "_200" suffix shows the average read-depth of a region. Columns with the "_200" suffix shows the read depth 200 bp away from the region boundaries.

and TRJ alleles with 100% accuracy, including both known and novel alleles, for both HG001 and HG002 (Figure 3).

Besides comparing novel alleles with annotations, we performed trio analysis as an additional validation. Our gAIRR-seq datasets included two trios: an Ashkenazi family (son: HG002, father: HG003 and mother: HG004) and a Chineses family (son: HG005, father: HG006 and mother: HG007). We performed gAIRR-call and collected novel alleles independently on each of the samples. We showed that there was no Mendelian violation in either family except two calls from HG002 (at *TRAJ29* and *TRBJ2-3*) (Figure 4 and Figure S1). We inspected the violated alleles and noticed they both located near known structural variant breakpoints, which were both regarded as *de novo* mutations (see Section Results: gAIRR-annotate identifies two structural variants for further discussion). Thus, we concluded that the gAIRR-seq pipeline with gAIRR-call was accurate in both known and novel alleles.

Further, we used gAIRR-call to analyze TRV alleles' flanking sequences outside the core region, which are challenging and important to the V(D)J recombination mechanism. Similar to verifying core alleles, we utilized gAIRR-annotate to provide truth sets from whole-genome assemblies. We define a call to be true-positive if it perfectly matched at least one of the haplotypes in the sample's corresponding assembly. If an annotation is not called by gAIRR-call, it's regarded as false-negative; if a gAIRR-call result doesn't have a matched annotation, it's considered to be false-positive. Among the 224 alleles called with flanking sequences from HG001, there were 7 (3%) false positives and 0 false negatives; among the 222 alleles called from HG002, there were only 2 (1%) false positives and 0 false negatives.

We additionally checked if the generated flanking sequences carried appropriate recombination signal sequences (RSS), which were important in the V(D)J recombination process for AIRR genes. All 7 GIAB RMs and the primary cell sample carried identical or nearly identical RSS with known patterns in IMGT. The finding of RSS showed the robustness of gAIRR-suite for both RMs and primary cell samples.

It is worth mentioning that there were three extended alleles not aligned perfectly to the personal assemblies HG001 and HG002, but we still considered them as true positives after manual inspection. The genes *TRBV20-1* in HG001-H1, *TRBV17*

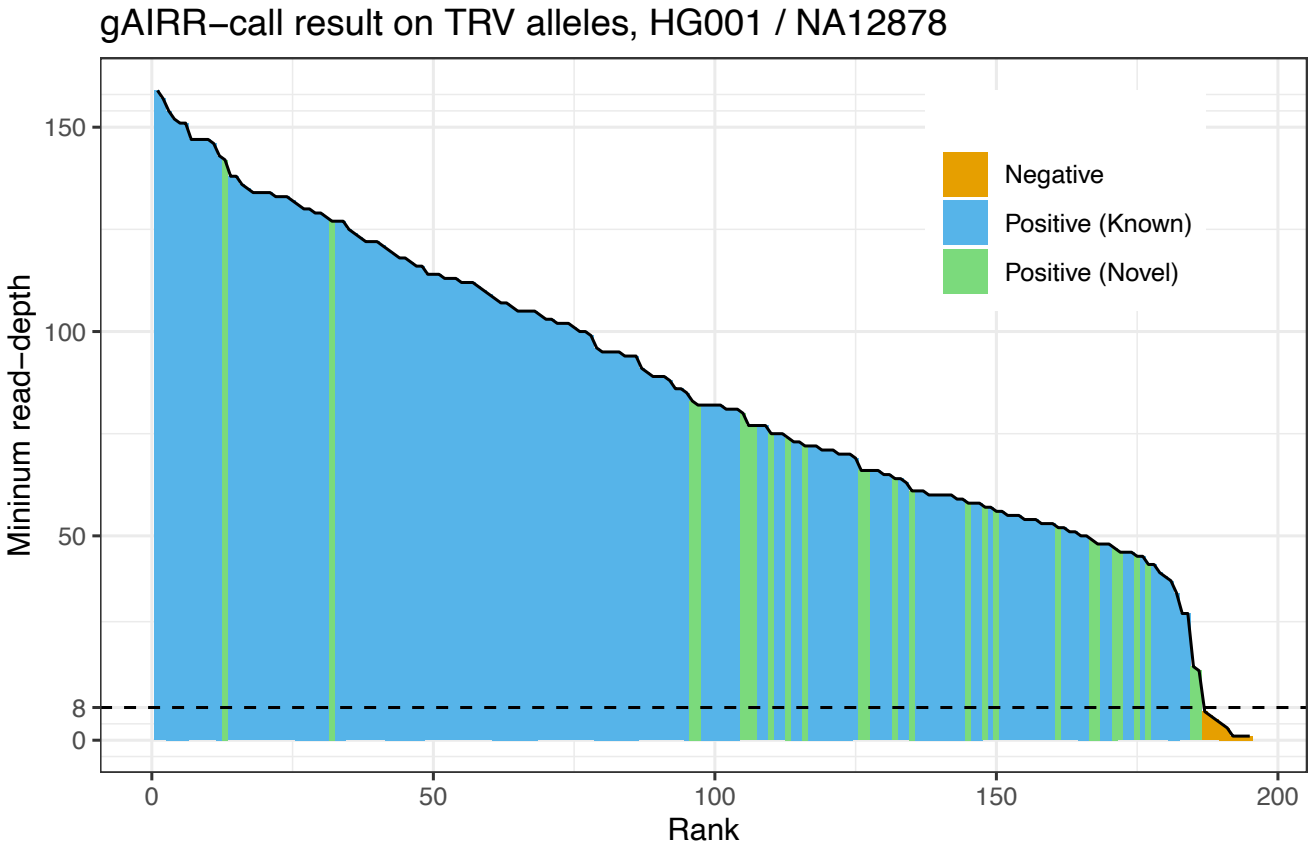


Figure 3. gAIRR-call results using HG001 data. The results are sorted by read-depth and the dash line represents the adaptive threshold in gAIRR-call. All the alleles not annotated by gAIRR-annotate, colored in orange, are below the adaptive threshold and are regarded as nocalls by gAIRR-call. True known alleles are in blue and true novel alleles are in green. In the HG001 analysis all true alleles are successfully identified by gAIRR-call. Alleles with zero minimum read-depth are excluded in the Figure.

in HG002-maternal, and *TRBV20-1* in HG002-paternal all carried one indel with respect to the personal assemblies. All three indels were in homopolymeric regions, which were known to be error-prone using PacBio sequencing. For example, in *TRBV17* the called allele had 7 consecutive guanines in the flanking region, while the assembly had 8 consecutive guanines. We examined these regions with the deep-coverage read sets sequenced with gAIRR-seq using the Integrated Genomics Viewer[22] and concluded the called alleles were likely to be correct.

Annotating TR alleles using high-quality assemblies

We collected 106 haplotypes from 6 works of literature[21;23;21;24;25;26] and annotated TR alleles using gAIRR-annotate. The dataset includes 36 individuals. We selected 12 samples to form a "HQ-12" set, where each sample could either be validated with family information or was curated with additional effort. This dataset includes HG001 (also known as NA12878, Northwest European American from Utah), HG002 (also known as NA24385, Ashkenazi Jewish), Puerto Rican trio HG00731, HG00732, and HG00733, Southern Han Chinese trio, HG00512, HG00513, and HG00514, Yoruban trio NA19238, NA19239, and NA19240, and a haploid hydatidiform moles CHM13. All samples have multiple assemblies generated from different technologies, assembly software or research groups. We showed that the numbers of TR alleles were relatively conserved — all haploids had 139-158 TRV alleles, and all but one haploid from HG002 had 5 TRD alleles and 82-86 TRJ alleles (Table S2). HG002 is known to carry two long deletions in the paternal haplotype in the TRD and TRJ region, resulting in the difference in the number of TR alleles([27]; see Section Result: gAIRR-annotate identifies two structural variants for details). The number of annotated alleles was concordant with our knowledge of the human TR germline. We also included the remaining 24 samples from HGSVC[23] in our analysis using gAIRR-annotate (Figure 4).

Our data collection included some samples with multiple assemblies available. In most cases, the alleles called using different assemblies from a single sample were concordant. The pipeline used a conservative strategy for high-precision

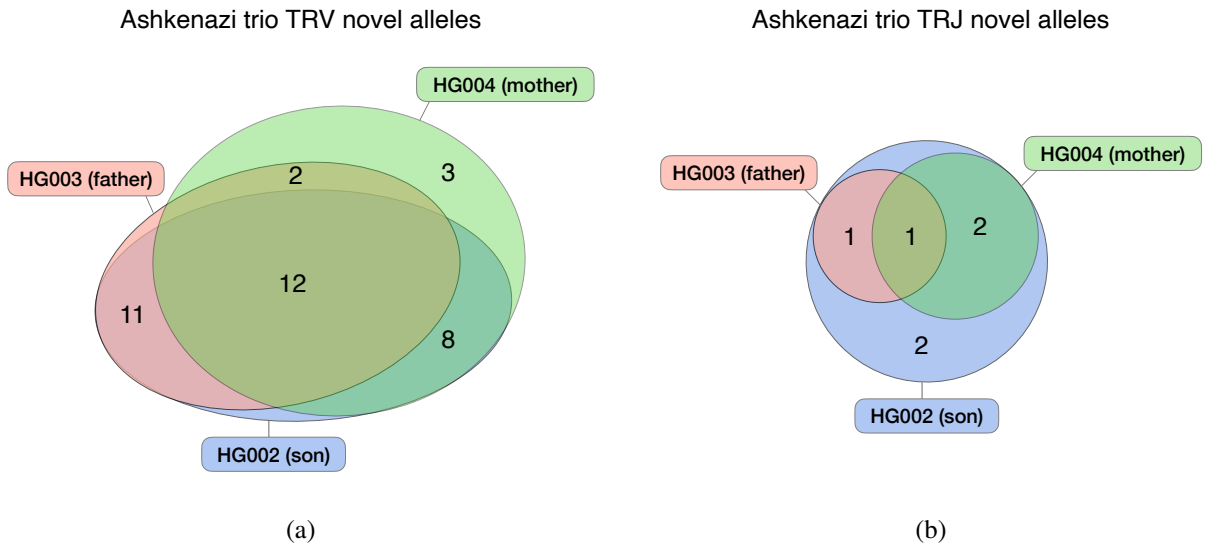


Figure 4. The (a) TRV and (b) TRJ novel allele relationship in the Ashkenazi family. HG002: son; HG003: father; HG004: mother. In TRV (a), there are no unique TRV alleles owned by HG002 and HG003. In other words, all the HG002's TRV alleles are either also owned by HG003 or HG004, and all the HG003's TRV alleles are also either owned by HG002 or HG004.

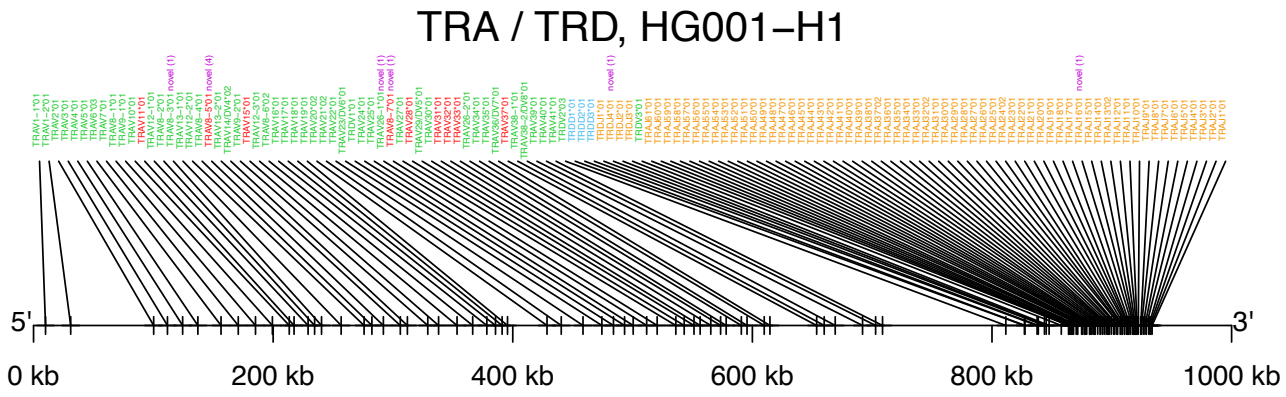


Figure 5. Positions of TRA / TRD alleles called by gAIRR-annotate. The purple text indicates that the allele is novel with its edit-distance compared to the most similar base allele inside parentheses.

annotation at conflicted loci – it took the alleles called from more than a half of the assemblies at a locus when the assemblies of a single sample generated inconsistent results. For example, we included six HG002 assemblies in the analysis, so gAIRR-annotate took an allele when there were calls from at least four assemblies (Table S3).

The novel alleles called by gAIRR-annotate showed concordance with the gene locations of existing IMGT annotations. For example, we found a novel allele between *TRAJ15*01* and *TRAJ17*01* in HG001-H1. The novel allele had only one base difference compared to *TRAJ16*01*, and according to IMGT the *TRAJ16* gene located in the middle of *TRAJ15* and *TRAJ17*. Thus, we were confident that the allele was a novel allele of a known gene. On average, there were 18 novel V alleles and

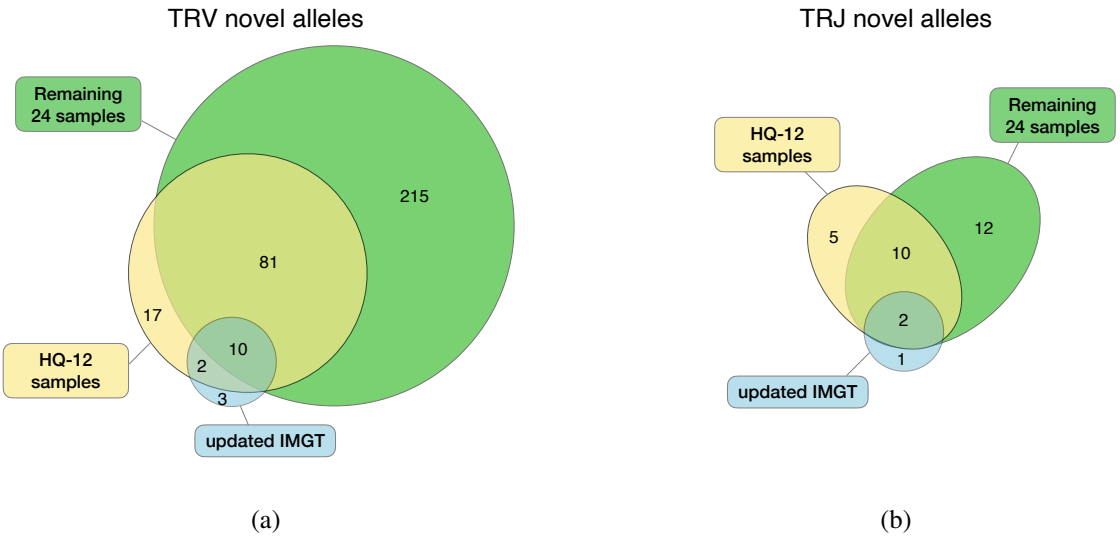


Figure 6. TRV (a) and TRJ (b) novel alleles found in HQ-12 samples (shown in yellow) and remaining 24 samples (shown in green) compared to the novel alleles updated in IMGT from v3.1.22 to v3.1.29.

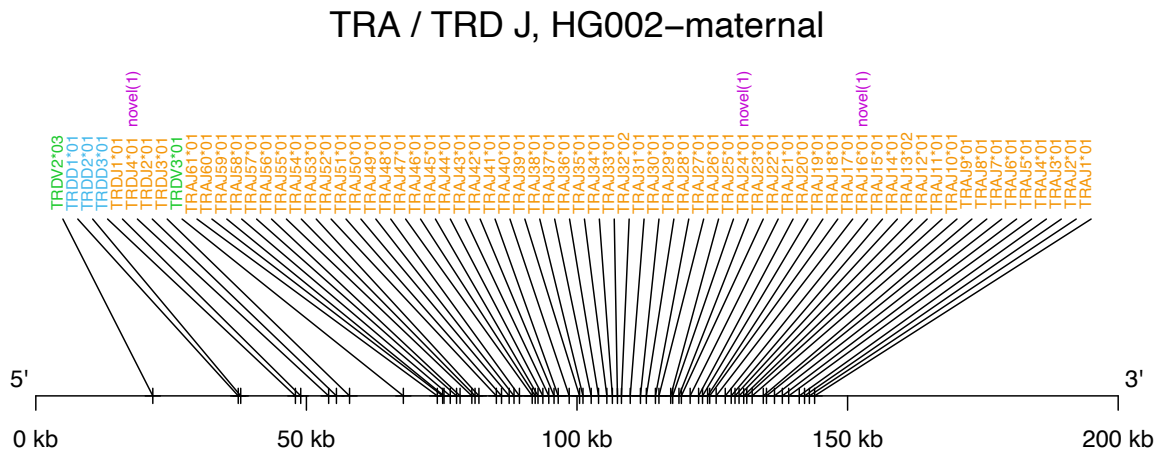
3 J alleles in each haplotype (Table S2). We examined the allele's relative positions for HG001 (Figure 5) and HG002, and the TRA and TRD alleles' relative positions were in the same pattern as the locus representation of IMGT[28] (Figure S2). It shows that the novel alleles found by gAIRR-annotate, marked by purple text, are known genes with a limited number of genetic alterations, mostly single nucleotide substitutions, to the known allele cataloged in IMGT. We also verified the structural difference found by gAIRR-annotate by comparing it with IMGT. For example, both known structurally different combinations of TRGV alleles (12 and 14 genes) in IMGT were found in HG001's haplotypes. The variations were in concordance with the IMGT locus representation. (Figure S3).

Identified novel alleles are consistent with the latest IMGT update

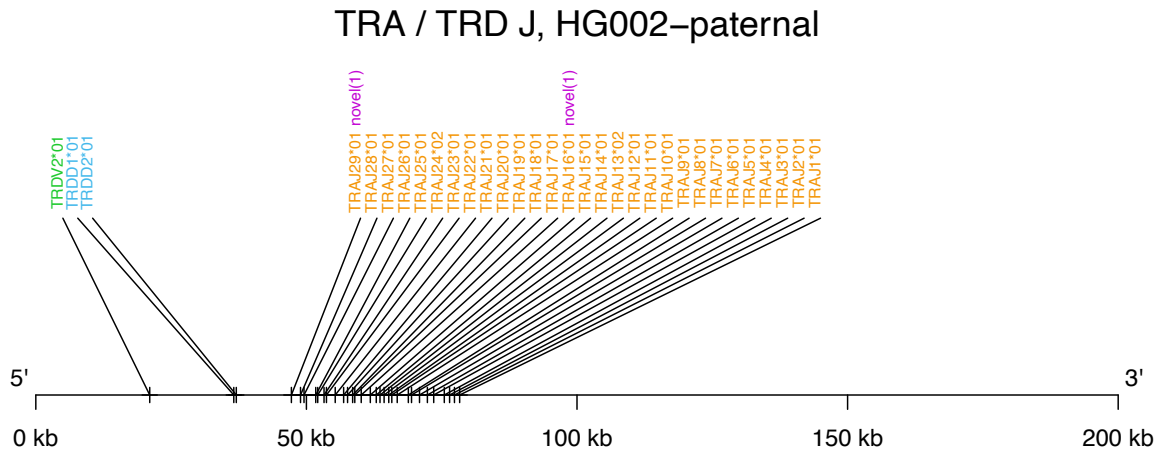
We identified novel alleles from the HQ-12 dataset and compared with the latest IMGT update. We used IMGT v3.1.22 as the baseline in this study, since it was the database we used to design gAIRR-seq probes. The latest version of the IMGT database (v3.1.29) had 15 TRV and 3 TRJ alleles added for human TR since v3.1.22. The novel alleles identified by gAIRR-annotate largely outnumbered the IMGT updates (110 vs. 15 for TRV and 17 vs. 3 for TRJ; Figures 6). For the alleles updated by IMGT, a large fraction of TRV alleles (12 out of 15) was discovered by gAIRR-annotate (Figure 6 (a)), demonstrating high accuracy of the proposed method.

The gAIRR-annotate method can also annotate extended TR alleles, including core alleles and 200-bp flanking sequences from both ends. We annotated 511 extended TRV alleles in the HQ-12 set, all of which were phased, and only 413 were known in the latest IMGT database. We further analyzed the recombination signal sequences (RSS) in the extended regions given their important participation in the V(D)J recombination mechanism. According to IMGT gene-DB, most TR RSS follow the rule that the heptamer and nonamer in the RSS are separated by either a 23-bp (191 V alleles) or a 12-bp spacer (89 J alleles). However, there are some functional TR alleles having 24-bp (10 V alleles) or 22-bp (4 V alleles) spacers and occasionally 11-bp (*TRAG20*01*), 10-bp (*TRAJ7*01*) or 13-bp (*TRDJ2*01*) spacers. There are also open reading frame (ORF) alleles that have only heptamer but no appropriate nonamer. We used gAIRR-annonate to identify known RSS patterns in the IMGT database and identify novel RSS (Table S4). We observed RSS polymorphism across different populations. For example, all 36 samples carry the IMGT *TRAJ12*01* RSS and 6 out of 36 samples also carry a novel *TRAJ12*01* RSS with a SNP at nonamer (heterozygous).

We build a novel-allele database collecting all identified novel alleles and an extended-allele database collecting alleles with 200-bp flanking sequences for both TRV and TRJ (novel-allele: Supplementary File S1 to S4; extended-allele: Supplementary File S5 to S8). We showed that we could broadly expand the size of known TCR databases using gAIRR-annotate and high-quality whole-genome assemblies. With the rapidly increasing number of high-quality assemblies utilizing accurate long-read sequencing, we are optimistic that gAIRR-annotate will help the community collect TR germline DNA information at a much greater speed than typical methods.



(a)



(b)

Figure 7. The 65 kbp structural variation in TRA / TRD J region of HG002 / NA24385. Upper: maternal haplotype, lower: paternal haplotype

gAIRR-annotate identifies two structural variants in HG002. We further confirmed the structural variants with capture-based sequencing

From Table S2, we noticed low numbers of TRJ and TRD alleles in the paternal haplotype of HG002. When visualizing alleles identified by gAIRR-annotate, we observed a 65 kbp missing region in TRA/TRD (Figure 7 (b)), containing 1 D, 1 V, and 36 J alleles from *TRDD3* to *TRAJ30* in the paternal haplotype. Similarly, we noticed a 10 kbp missing TRB region, containing 2 D alleles and 9 J alleles from *TRBD1* to *TRBJ2-2P*, in the same haplotype (Figure S4 (b)).

To further verify the 65 kbp deletion in the TRA/TRD loci of HG002, we aligned capture-based short reads of gAIRR-seq from HG002 to GRCh37 as well as gAIRR-seq reads from HG003 (HG002's father) and HG004 (HG002's mother) to perform trio analysis (Figure 8). We first examined if there were reads aligned across two deletion breakpoints, either in a chimeric form or having two paired segments separated across the region. We found 72 reads aligned across chr14:22,918,113 and chr14:22,982,924 using the gAIRR-seq dataset, providing clear evidence of a long deletion.

Further, we compared the read depth inside and outside the called deletion region. We counted the number of reads in the left and the right regions of chr14:22,982,924 and calculated the coverage drop-off rate (#left reads/#right reads). In a region

Figure 8. The Integrated Genomics Viewer visualization of the capture-based reads from HG002 (son, top), HG003 (father, middle), and HG004 (mother, bottom) aligned to GRCh37 chromosome 14. There are two red arrows indicating abrupt read-depth changes of HG002's reads at chr14:22,918,113 and chr14:22,982,924.

where reads are nearly uniformly aligned, the drop-off rate is expected to be around 1. If one of the haplotypes is missing on the left, we expect a drop-off rate around 0.5. At locus chr14:22,982,924, the drop-off rate of HG002 is 0.622, while the values for HG003 and HG004 are 1.005 and 1.138 respectively (Table S5). This verified the long deletion in the HG002 TRA/TRD region and further provided evidence that the structural variant was likely a novel mutation.

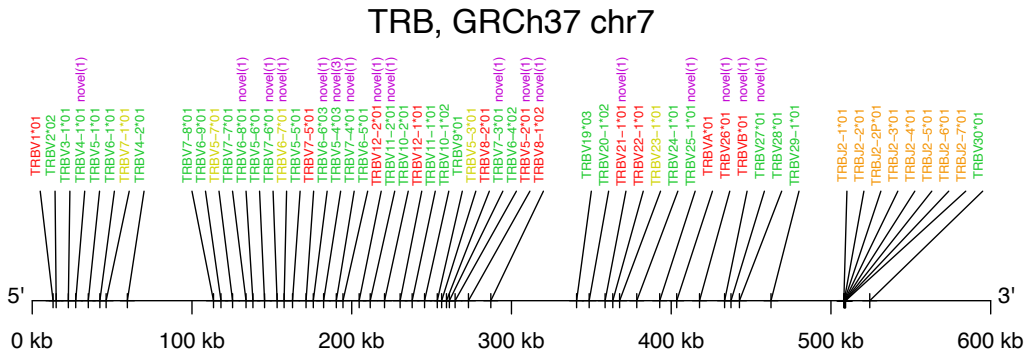
We provided two independent investigations using gAIRR-annotate with a diploid HG002 assembly[21] and using gAIRR-seq data. Both results indicated that one HG002 haplotype had a structural deletion situated in chr14 from position 22,918,114 to position 22,982,924 (GRCh37). The structure variant analysis performed by GIAB[27] also showed a deletion at chr14:22,918,114 in GRCh37, where a 13 bp fragment in HG002 replaces a 64,807 bp fragment in GRCh37. The size of deletion reported by GIAB[27] was similar to the suggestion by gAIRR-annotate and gAIRR-seq. The deletion of HG002's chromosome contributes to a loss of 32 TRAJ alleles and all 4 TRDJ alleles at the paternal haplotype.

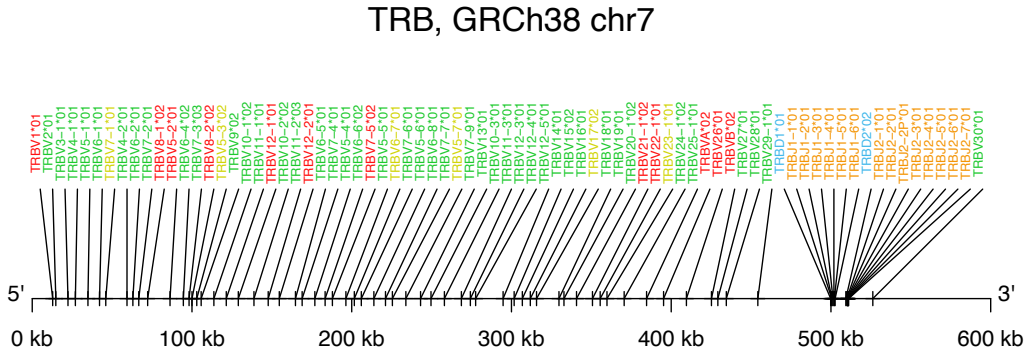
Similarly, there is a 10 kb structural deletion at chromosome 7 q34 related to TRB alleles. We verify the deletion with gAIRR-annotate, gAIRR-seq, and the report of GIAB[27]. (Supplementary Section 8 HG002's structural deletion in chromosome 7)

TR structural variations in GRCh37 and GRCh38

We performed gAIRR-annotate to analyze human reference genomes GRCh37 and GRCh38 and identified structural variants in TR regions (Figure 9). First, we identified an inversion in GRCh37 with respect to the GRCh38 primary assembly that covered TRBV alleles. In this inversion, there were 39 TRBV alleles (*TRBV6-2* to *TRBV18* following the allele order provided by IMGT) successfully annotated in GRCh38. In contrast, in GRCh37, only 27 alleles from *TRBV8-1* to *TRBV7-8* were successfully annotated, but with a reversed order. Second, we observed another long-missing sequence (about 10 kbp) in GRCh37 in the TRB region. This region covered about half of the expected TRBD and TRBJ alleles. We noticed GRCh38 didn't carry this deletion, and we annotated reasonable numbers of TRBD and TRBJ alleles as the suggestion of IMGT. When comparing annotated alleles to the IMGT database, we showed that all alleles in GRCh38 are known, and there are 18 novel alleles in GRCh37.

We compared both structural variants with the chain file provided by UCSC genome browser[29]. According to this chain





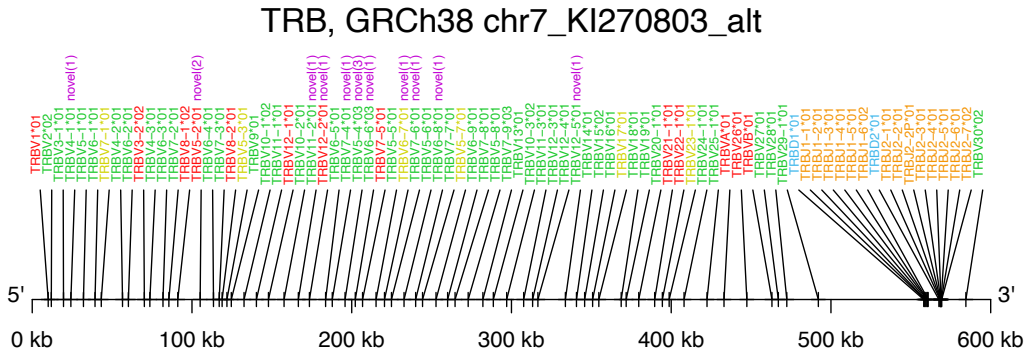


Figure 9. Subfigure from top to bottom is the gAIRR-annotate locus representation of TR beta germline alleles of GRCh37 chr7, GRCh38 chr7, GRCh38, and GRCh38 alternative contig chr7_KI270803v_alt.

file, the TRBV inversion was situated around the breakpoint at chr7:142,048,195 (GRCh37). In this region, there was a 278 kbp sequence in GRCh37 unmatched with 270 kbp one in GRCh38. The TRBD/TRBJ structural variant also had a supporting breakpoint in the chain file, at chr7:142,494,031 (GRCh37), where 18 bases in GRCh37 were unmatched with 10143 bases in GRCh38.

We further analyzed an GRCh38 alternative contig, *chr7_KI270803v1_alt*, which contained TRBV alleles (Figure 9 bottom). Eleven novel TRBV alleles in *chr7_KI270803v1_alt* were identified using gAIRR-annotate. Compared to the GRCh38 primary assembly, *chr7_KI270803v1_alt* had six additional TRBV alleles. Both TRB alleles in chr7 and *chr7_KI270803v1_alt* showed concordance with the locus representation of IMGT[28] (Figure S6). Thus, when a personal assembly is not available, we suggest using GRCh38 instead of GRCh37 for TR analysis in that GRCh38 shows concordance with the IMGT database in sequence completeness and carries better-known alleles.

Discussion

Our study aims to develop a pipeline, gAIRR-suite, that includes probe capture-based targeted sequencing and computational analyses to profile the germline AIRR. We show that gAIRR-seq can enrich the AIRR regions with high on-target rates and sufficient read depth. We then use gAIRR-call to accurately call both known and novel alleles using reads sequenced with gAIRR-seq. To utilize high-quality whole-genome assemblies, we designed gAIRR-annotate to accurately annotate AIRR genes in both core allelic and flanking regions. We used gAIRR-annotate to identify 110 novel core alleles in TRV regions and 17 novel core alleles in TRJ regions from 12 whole-genome assembled individuals. We investigated the RSS patterns in the annotated flanking sequences and observed high concordance with the IMGT database. The RSS annotations also identified alleles flanked by RSS with spacers neither 12-bp nor 23-bp, which had been mentioned by[30] and[31]. gAIRR-annotate can also help to discover structural variants. We verified two known structural variants in the TR regions in HG002. Similarly, we used gAIRR-annotate to uncover two large disagreements between reference genomes GRCh37 and GRCh38.

gAIRR-seq is advantageous in its comprehensiveness and high resolution. Compared with multiplex PCR-based methods, gAIRR-seq covers all known AIRR genes and alleles, including all TRs and IGs, in one experiment. Furthermore, novel alleles can also be identified because the probes can tolerate sequence mismatch to a certain degree. Although other technologies like iRepertoire (`https://irepertoire.com`) and Adaptive Biotechnologies (`https://www.adaptivebiotech.com`) can also adopt gDNA as the starting material for immune repertoire sequencing, they rely on the CDR3 region of rearranged V(D)J fragments to infer germline repertoires, which might result in lower resolution and allele dropout. Another advantage of probe capture-based enrichment is the cost. Each sample only costs approximately USD 170 for library preparation and target enrichment, with all IG and TR genes/alleles being captured in one experiment using easily available genomic DNA from cell lines or primary cells, such as peripheral blood mononuclear cells (PBMCs). We point out that gAIRR-seq capture probe design can be further facilitated by gAIRR-annotate thanks to the accurate collection of extended alleles. In this study, our probe design was limited to only V and J genes due to insufficient allele lengths for other genes. With a comprehensive population genomic AIRR database becoming available, we plan to include TRD, IGD and IGJ genes (all except 6 IGJ alleles are shorter than 60 bps) in future gAIRR-seq probe design.

Although the gAIRR-call is technically applicable to IG genes, we found that the accuracy for IG genes is lower than TR genes. We hypothesize that the 7 GIAB DNA samples were retrieved from EBV-transformed B lymphocytes and might suffer from reduced diversity during the establishment, growth and subculture, making them unsuitable for germline IG profiling the[32]. To circumvent this problem, we plan to use primary cells, such as PBMCs or buccal cells, for IG calling in the future. We gAIRR-seqed the genomic DNA from PBMCs of our principal investigator (Supplementary Section gAIRR-seq and gAIRR-call on primary cells) and gAIRR-called similar numbers of known and novel alleles in the TRV and TRJ regions compared to the GIAB RMs. The results are promising that we can apply the pipeline to primary cells and perform analysis in the IG genes.

Up to now, it is almost insurmountable to study the impact of germline AIRR variations on human immune-related phenotypes and diseases. The single nucleotide polymorphism (SNP) genotyping array widely used for genome-wide association study (GWAS) tags only a tiny fraction of AIRR variations[3];[13];[15]. Whole-genome sequencing (WGS) can theoretically cover AIRR variations; however, with neither a powerful bioinformatics tool (such as gAIRR-call) nor reliable reference genome(s), WGS has not yet been used to retrieve AIRR information for human genetic study due to relatively lower sequencing depth compared to enriched sequencing and high polymorphism in the AIRR regions. Whether germline AIRR variations have a broad and strong impact on human immune-related phenotypes and diseases, just like the human leukocyte antigen (HLA) variations[33];[34], is an open question awaiting more studies.

We envision that gAIRR Suite can help in at least the following scenarios. First, the discovered novel alleles can help to develop better SNP arrays and better WGS analysis of AIRR in the future. Second, the personal germline AIRR profile can be directly employed for genetic study and clinical application. Third, the personal germline AIRR genes, together with other critical immune genes (such as HLA and killer cell immunoglobulin-like receptor (KIR) genes), can be tested for di-genic or

oligo-genic effects. Last but not least, the combined analysis of germline AIRR and dynamic mRNA-based AIRR profiles might help decipher immunological responses and pathophysiological mechanisms.

Methods

The gAIRR-seq pipeline: a capture-based targeted sequencing method

Reference materials

We obtained 7 GIAB genomic DNA RMs from the Coriell Institute (https://www.coriell.org). These 7 RMs included a pilot genome NA12878/HG001 and two Personal Genome Project trios - an Ashkenazim Jewish ancestry: huAA53E0/HG002, hu6E4515/HG003, hu8E87A9/HG004; a Chinese ancestry: hu91BD69/HG005, huCA017E/HG006, hu38168C/HG007. These RMs were well-sequenced using more than 10 NGS platforms/protocols by GIAB[35]; therefore, they are appropriate RMs for benchmarking gAIRR-seq.

Probes design

We designed probes for all known V and J alleles of IG and TR based on the ImMunoGeneTics (IMGT) database Version 3.1.22 (3 April 2019) (including all functional, ORF, and pseudogenes)[19]. Each probe was a continuous 60-bp oligo (Roche NimbleGen, Madison, WI, U.S.). We designed three probes for each V allele and one probe for each J allele based on V and J alleles' length differences. (Figure S8). For the J alleles shorter than 60 bp, we padded the probes with random nucleotides "N" to 60 bp in length on both ends.

Library preparation

We quantified the input of 1000 ng gDNA and assessed the gDNA quality using a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, U.S.) before the library preparation. The gDNA was fragmented using Covaris (Coravis, Woburn, MA, U.S.), aiming at the peak length of 800 bp, which was assessed using Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, U.S.). We then added adapters and barcodes onto the ends of the fragmented DNA to generate indexed sequencing libraries using TruSeq Library Preparation Kit (Illumina, San Diego, CA, U.S.). We performed capture-based target enrichment using the SeqCap EZ Hybridization and Wash Kit (Roche NimbleGen, Madison, WI, U.S.). We then used the sequencing platform, MiSeq (Illumina, San Diego, CA, U.S.), to generate paired-end reads of 300 nucleotides.

The gAIRR-call pipeline

The gAIRR-call pipeline is composed of three steps: Finding novel alleles, calling alleles, and generating extended alleles with 200 bp flanking sequences on both ends.

Finding novel allele candidates

gAIRR-call first generates candidate novel alleles with capture-based short reads. The capture-based short reads are aligned to all the known alleles in IMGT database (v3.1.22 in this study). Then gAIRR-call checks the reference IMGT alleles one by one. For any known allele, the variants are called with the aligned reads. gAIRR-call marks any allele position with more than a quarter of reads supporting bases different from the reference. If there is only one variant in the reference, gAIRR-call simply calls the allele with the variant as a novel allele candidate. If there is more than one variant in the reference, gAIRR-call performs haplotyping and connects the variants with aligned reads. Since most of the AIRR alleles are within 300 bp, there are usually a sufficient number of captured reads (300 bp in length using gAIRR-seq) for haplotyping.

We then collect called haplotypes and check if there are duplicated sequences. If a called haplotype is actually another known allele in IMGT, the haplotype is discarded. Because haplotypes are called from known alleles with different lengths, there can be duplicated haplotypes representing the same allele. In this case, only the shorter haplotype is kept. After de-duplicating and cleaning, the remaining haplotypes are unique. The haplotypes join the known IMGT alleles to be the candidate alleles for the next stage analysis.

Calling AIRR alleles

Compared to the previous step, calling the AIRR allele applies stricter criteria. gAIRR-call aligns capture-based short reads to the allele candidates pool, which includes both known IMGT alleles and the novel ones. To ensure that every candidate allele can be aligned by all potential matched reads, we use the '-a' option of BWA MEM[36] so that one read can be aligned to multiple positions.

After alignment, gAIRR-call removes alignment that contains mismatches or indels. Further, for alleles longer than 100 bp, the reads' alignment length should be more than 100 bp. For alleles shorter than 100 bp, the reads should cover the entire allele. Subsequently, only reads perfectly matched the allele with long coverage length pass the filter. gAIRR-call counts the remaining reads' read-depth of all allele positions. The minimum read-depth in an allele is its final score relative to other candidate alleles. The candidate alleles are sorted according to their supporting minimum read-depth (Figure 3).

We use an adaptive threshold in gAIRR-call to decide the final calls. The goal of this adaptive threshold is to identify the large drop in minimum read-depth, which is observed to be a good indicator of true-positive and false-positive alleles. When two consecutive alleles have a read depth ratio of less than 0.75, we set the adaptive threshold to be the lower read depth. All the candidate alleles with lower minimum read-depth below the adaptive threshold are discarded. We also notice there can be some alleles with extraordinary allele lengths. For example, *TRAV8-5* is a 1355 bp pseudogene that is the longest known TR alleles, where the second-longest allele is 352 bp. When they truly appear in the data, their minimum read-depths are usually slightly lower than other true alleles'. Thus, when calculating the adaptive threshold, we exclude outlier alleles like *TRAV8-5* to provide a more appropriate threshold. After that, we still compare the minimum read-depth of these outlier alleles with the adaptive threshold to decide whether they are positive.

Calling flanking sequences

After allele calling, we perform assembly and haplotyping to call the flanking sequences of the called allele. For each called allele, we group the reads perfectly matched with long coverage length with their mate pairs. These reads are then assembled using SPAdes (v3.11.1)[20] with '–only-assembler' option. We double-check if the allele is actually in the contig by aligning the specific read back to its assembled contigs. The contigs not containing the specific allele are discarded. Mostly, one called allele generates one contig. If multiple contigs are generated, we keep the longest one.

The flanking sequences of an allele can be different in H1 and H2 of a sample. To get the correct phased flanking sequences, we perform haplotyping on the assembled contig. We first align all the capture-based short reads to all the contigs. This time, a read can only be aligned to one position. Here we assume that the flanking sequences of an allele in the two haplotypes are similar and thus the reads of an allele's two flanking sequences can be aligned to the same contig. We mark the contig position with more than a quarter of reads supporting bases different from the contig. Then we haplotype the variants using the pair-end reads. The fragment lengths of the pair-end reads are up to 800 bp; hence the core allele region extended 200 bp from both sides can be fully covered by the pair-end reads. After haplotyping, the extended alleles with 200 bp flanking sequences are reported.

Although the contigs assembled by SPAdes are typically larger than 1000 bp, we do not report the flanking sequences as long as possible for two reasons. First, the contigs' boundary regions are not very reliable because the read-depth is decreasing toward the boundaries. Second, the fragment length of our capture-based reads is up to 800 bp. If the shortest distance between variants on two flanking sides of the allele is longer than 800 bp, it's impossible to haplotype the variants. Considering most AIRR alleles are shorter than 300 bp, it is robust to call the extended alleles with 200 bp flanking sequences from both sides.

The accuracy of flanking sequences is related to novel alleles calling. If the reads can only be aligned to contigs generated by known alleles, the reads of novel alleles will be aligned to the closest known alleles, resulting in false-positive read assignment and lower accuracy.

The gAIRR-annotate pipeline

The gAIRR-annotate pipeline annotates AIRR alleles using publicly available whole-genome assemblies (Table S2). gAIRR-annotate begins with aligning IMGT AIRR alleles to the assemblies. To take novel alleles into consideration, we set the alignment options to allow mismatches and indels, and report all potential sites. Whenever multiple alleles align to the same spot, we keep the alignment with the minimum edit-distance. The gAIRR-annotate pipeline reports both perfectly matched alleles and alignment with edit-distance. Those alleles with edit-distance are considered as novel alleles.

To find RSS of each allele, we align all IMGT RSS to the annotated flanking sequences. For an RSS aligned with mismatch(es), we define the aligned region as a novel RSS. The flanking sequences cannot be aligned with IMGT RSS were re-aligned where IMGT heptamers and nonamers are separated. The second-pass processing thus allows heptamer-nonamer pairs not recorded in IMGT.

For TRV and TRJ alleles, gAIRR-annotate uses the allelic sequences as the query. For TRD alleles, because the lengths of them range from 8 to 16 bp, which would result in a huge number of valid alignments, we extend them by adding the heptamers on both sides of the D region. After the extension with the heptamers, the extended D alleles' length ranges from 21 to 31. They usually appear only once in the whole-genome, suggesting the uniqueness of TRD alleles with appropriate RSS.

Comparing alleles of gAIRR-call and gAIRR-annotate

Because both the capture-based sequence data and personal assembly are available in HG001 and HG002, we can compare the results of gAIRR-call and gAIRR-annotate in these two samples. The known gAIRR-call alleles can be directly compared with gAIRR-annotate results; however, comparing the novel alleles are not as straightforward as comparing known ones. Sometimes the same novel alleles are inferred from gAIRR-call and gAIRR-annotate in different ways. For example, gAIRR-call infer a novel allele from a known allele with SNPs while gAIRR-annotate infer the same novel allele through insertions. For another example, a novel allele can be inferred from two highly similar known alleles with different lengths, such as *TRAV1-1*01* and *TRAV1-1*02*, which are highly similar, but *TRAV1-1*01* is longer by 6 bp. In these two examples, the same novel allele can differ in length in gAIRR-call and gAIRR-annotate. As a result, we verify the gAIRR-call novel alleles by aligning them to the

personal assemblies. If a novel allele can perfectly align to the assembly, then it is a true positive. We also check if gAIRR-call alleles cover all the annotated novel allele positions in gAIRR-annotate. Similarly, since there are novel alleles in extended alleles, we evaluate the flanking sequences with alignment rather than direct comparisons.

Database collecting

We collect the novel alleles and the extended alleles with 200 bp flanking sequences called by gAIRR-annotate into two databases. Usually, the alleles called by gAIRR-annotate using different assemblies from a single sample are concordant. When there are conflicts between different assemblies, the pipeline uses a conservative strategy for high precision– it takes the alleles called from more than a half of the assemblies at a locus when the assemblies of a single sample generated inconsistent results. For example, we included five HG002 assemblies in the analysis, so gAIRR-annotate took an allele when there were calls from at least three assemblies. We also collect novel alleles and extended alleles called by gAIRR-call. Whenever there are multiple forms of the same allele called, which are inferred from different ways, we keep only the shortest one.

References

1. Arstila, T. P. *et al.* A direct estimate of the human $\alpha\beta$ t cell receptor diversity. *Science* **286**, 958–961 (1999).
2. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
3. Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes & Immun.* **13**, 363–373 (2012).
4. Robins, H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr. opinion immunology* **25**, 646–652 (2013).
5. Li, S. *et al.* Imgt/highv quest paradigm for t cell receptor imgt clonotype diversity and next generation repertoire immunoprofiling. *Nat. communications* **4**, 1–13 (2013).
6. Avnir, Y. *et al.* Ighv1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with ighv utilization shifts and varies by ethnicity. *Sci. reports* **6**, 1–13 (2016).
7. Watson, C. T., Glanville, J. & Marasco, W. A. The individual and population genetics of antibody immunity. *Trends immunology* **38**, 459–470 (2017).
8. Chung, W.-H. *et al.* A marker for stevens–johnson syndrome. *Nature* **428**, 486–486 (2004).
9. Pan, R.-Y. *et al.* Identification of drug-specific public tcr driving severe cutaneous adverse reactions. *Nat. communications* **10**, 1–13 (2019).
10. Ishii, H. *et al.* Determination of a t cell receptor of potent cd8+ t cells against simian immunodeficiency virus infection in burmese rhesus macaques. *Biochem. Biophys. Res. Commun.* **521**, 894–899 (2020).
11. Roschmann, E., Wienker, T. F., Gerok, W. & Volk, B. A. T-cell receptor variable genes and genetic susceptibility to celiac disease: an association and linkage study. *Gastroenterology* **105**, 1790–1796 (1993).
12. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. genetics* **45**, 1452–1458 (2013).
13. Parks, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in oceania. *Nat. communications* **8**, 1–10 (2017).
14. Tsai, F.-J. *et al.* Identification of novel susceptibility loci for kawasaki disease in a han chinese population by a genome-wide association study. *PloS one* **6**, e16853 (2011).
15. Johnson, T. A. *et al.* Association of an ighv3-66 gene variant with kawasaki disease. *J. human genetics* 1–15 (2020).
16. Aouinti, S., Malouche, D., Giudicelli, V., Kossida, S. & Lefranc, M.-P. Imgt/highv-quest statistical significance of imgt clonotype (aa) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and t cell receptors. *PLoS One* **10**, e0142353 (2015).
17. Liu, X. & Wu, J. History, applications, and challenges of immune repertoire research. *Cell biology toxicology* **34**, 441–457 (2018).
18. Rosati, E. *et al.* Overview of methodologies for t-cell receptor repertoire analysis. *BMC biotechnology* **17**, 61 (2017).
19. Giudicelli, V., Chaume, D. & Lefranc, M.-P. Imgt/gene-db: a comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic acids research* **33**, D256–D261 (2005).

20. Bankevich, A. *et al.* Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J. computational biology* **19**, 455–477 (2012).

21. Garg, S. *et al.* Efficient chromosome-scale haplotype-resolved assembly of human genomes. *bioRxiv* 810341 (2019).

22. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. biotechnology* **29**, 24–26 (2011).

23. Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).

24. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. biotechnology* **37**, 1155–1162 (2019).

25. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly with phased assembly graphs. *arXiv preprint arXiv:2008.01237* (2020).

26. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human x chromosome. *Nature* **585**, 79–84 (2020).

27. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *BioRxiv* 664623 (2019).

28. Lefranc, M.-P. & Lefranc, G. *The T cell receptor FactsBook* (Elsevier, 2001).

29. Kent, W. J. *et al.* The human genome browser at ucsc. *Genome research* **12**, 996–1006 (2002).

30. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase dna sequence of the human β t cell receptor locus. *Science* **272**, 1755–1762 (1996).

31. Kawasaki, K. *et al.* One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome research* **7**, 250–261 (1997).

32. Tan, K.-T. *et al.* Profiling the b/t cell receptor repertoire of lymphocyte derived cell lines 11 medical and health sciences 1107 immunology. (2018).

33. Klein, J. The hla system. second of two parts. *N Engl J Med* **343**, 782–786 (2000).

34. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. Hla variation and disease. *Nat. Rev. Immunol.* **18**, 325 (2018).

35. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. data* **3**, 1–26 (2016).

36. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* (2013).

37. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome. *bioRxiv* 855049 (2019).

Acknowledgements

We thank to National Core Facility for Biopharmaceuticals (NCFB, MOST 108-2319-B-492 -001) for support and National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources. This study was supported by Taiwan Ministry of Science and Technology grants (MOST 108-2314-B-002-069-MY3) and National Taiwan University Hospital (109-S4521, 108-T05). We also thank Justin Zook for his helpful suggestions in sequencing experiments and analysis.

Author contributions statement

Y.L. and P.C. designed the sequencing experiments and A.C.L. performed the experiments. Y.L. and N.C. designed preliminary analysis pipeline. M.L. designed the final analysis pipelines and wrote the software. M.L., N.C., and Y.L. interpreted all the analysis results with P.C.'s guidance. S.K., J.H., C.C., C.H., W.Y., and P.C. initiated this project and assisted with discussions. M.L., N.C., and P.C. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Competing interesting

The authors declare no competing interest.