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 profiling of germline Adaptive Immune Receptor Repertoire (AIRR), including T cell receptors (TR) and immunoglobulin (IG), is medically important but challenging due to high genetic diversity and complex recombination. While databases such as the international ImMunoGeneTics information system (IMGT) provide a valuable collection of AIRR allele sequences, they are far from complete, and many of them lack flanking sequences outside alleles. It will be beneficial to have technologies that can efficiently capture relevant sequences from DNA samples or genome assemblies, perform accurate allele calling, and discover novel alleles. In this study, we developed gAIRR-seq, a probe capture-based targeted sequencing pipeline, to profile genomic sequences of TR and IG from individual DNA samples. We then proposed computational pipelines — gAIRR-call and gAIRR-annotate — to call alleles from gAIRR-seq reads and validate the results with whole-genome assemblies. Genomic DNA samples from Genome in a Bottle (GIAB) reference materials and public whole-genome assemblies were used for benchmarking. We applied gAIRR-call to genotype TRV and TRJ alleles with 100% accuracy and achieved 98% accuracy when considering flanking regions of V alleles. We used gAIRR-annotate to annotate allele positions and collect alleles and flanking regions into databases. gAIRR-annotate validated the alleles using 13 high-quality whole-genome assemblies on 6 samples and discovered 79 novel TRV alleles and 11 novel TRJ alleles. We provided a summary of the number of TR alleles from 13 high-quality draft assemblies with the proposed pipeline. 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.

Introduction

Adaptive Immune Receptor Repertoire (AIRR)

The Adaptive Immune Receptor Repertoire (AIRR) is a collection of T-cell receptors and B-cell receptors, 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. Thanks to the extensive combination possibility from the V, (D), and J pool, the immune cell for recognizing foreign antigens is enormously diverse.

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 his/her genome is limited[1]. Therefore, it is possible that any given individual's immunological response can be restricted according to the germline IG/TR composition[1]. For example, the response to influenza vaccine was clearly demonstrated to be modulated by *IGHV1-69* polymorphism[2],[3]. Another example is that the *IGHV3-23*03* allele binds Haemophilus influenzae type b polysaccharide more effectively compared with the most frequent allele, *IGHV3-23*01*[1]. An additional example can be found that anti-HIV antibodies have biased usage of *IGHV1-69*[1]. Recently the famous carbamazepine-induced Stevens-Johnson syndrome[4] has an additional twist[5]; in addition to the well-known *HLA-B*15:02* being a disease-causing gene, drug-specific public TR was also critical for this severe cutaneous adverse reactions[5]. However, to determine germline IG/TR for any given individual has been extremely difficult.

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[6]. 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. gDNA includes non-rearranged and rearranged (productive and non-productive) V(D)J segments, which can be either an advantage of giving more information or a disadvantage of resulting background noise to interfere with studying functional IG/TR repertoires[7]. Besides, gDNA is more stable and easier for storage compared to RNA, which is quite vulnerable. Thus, using gDNA as a material is cheaper than mRNA. On the other hand, RNA as the material can provide information on gene expression level, which can interpret immune repertoire expression in different cell stages and statuses. 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[7],[8].

Probe-Based target enrichment and sequencing

Multiplex polymerase chain reaction (PCR) and probe capture-based target enrichment are two commonly used methods for gDNA and mRNA samples. Multiplex PCR incorporates the mixture of primers specific to the known V and J genes (or C genes for mRNA) to amplify the targeted regions. Such approach has at least two drawbacks: the incompetence of detecting novel alleles due to the specific primer sets and the amplification bias due to different efficiencies between primers or preferences of some alleles[7],[8]. Furthermore, difficulty in primer designs and possible allele dropout are also possible issues. Probe capture-based target enrichment begins with the fragmentation of starting materials, followed by using complementary DNA or RNA baits to hybridize with fragmented targets. The hybridized products will undergo amplification and sequencing processes to retrieve targeting regions. This approach's benefit is that fewer PCR cycles are required, which would have less PCR bias[8]. Furthermore, probe hybridization can tolerate sequence mismatch to certain degree, therefore empowering researchers to discover novel alleles. It is also possible to get the sequences from the regions flanking the probe binding regions. In this study, we developed the gAIRR-seq pipeline, which is the combination of probe capture-based target enrichment of genomic AIRR genes and next-generation sequencing (NGS).

Computational tools: gAIRR-call and gAIRR-annotate

To analyze the probe-based target enrichment sequencing data, we develop a computational pipeline, gAIRR-call (Figure 1), to accurately call AIRR alleles. We design 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 IMGT database[9]. gAIRR-call uses error-tolerant alignment in this phase, so both known and novel allele candidates can be identified. In the fine phase, gAIRR-call re-aligns the reads to the candidate alleles with no errors allowed and assigns reads to each allele. It then summarizes the coverage (read depth) of each allele, and uses an adaptive read-depth filter to make final positive calls. We further extend the called alleles using the SPAdes assembler[10] to obtain higher-quality haplotypes. gAIRR-call reports positive alleles and their 200 bp flanking sequences extended from both ends of the core allele.

One major challenge for germline AIRR studies is lack of curated annotations, and thus no ground truth for tool development. Thus, we design a computational pipeline, gAIRR-annotate (Figure 1), to annotate AIRR alleles as well as flanking sequences using high-quality whole-genome assemblies. gAIRR-annotate aligns all known AIRR alleles to an assembly. Alleles that are matched perfectly are annotated. For alleles that are matched with mismatches, we identify the nearest allele and its associated gene in the IMGT database and assign a novel allele. When there are multiple assemblies available for a sample, we only annotate an allele if there are more than half of the assemblies report the same output. Further, we include the sequences outside core allelic regions when annotating. By default, 200 bp flanking sequences extended from both ends of an allele is annotated. Moreover, gAIRR-annotate reports the relative position of each allele, including the novel ones. When phased assemblies are available, we can apply gAIRR-annotate to analyze structural variants in AIRR regions.

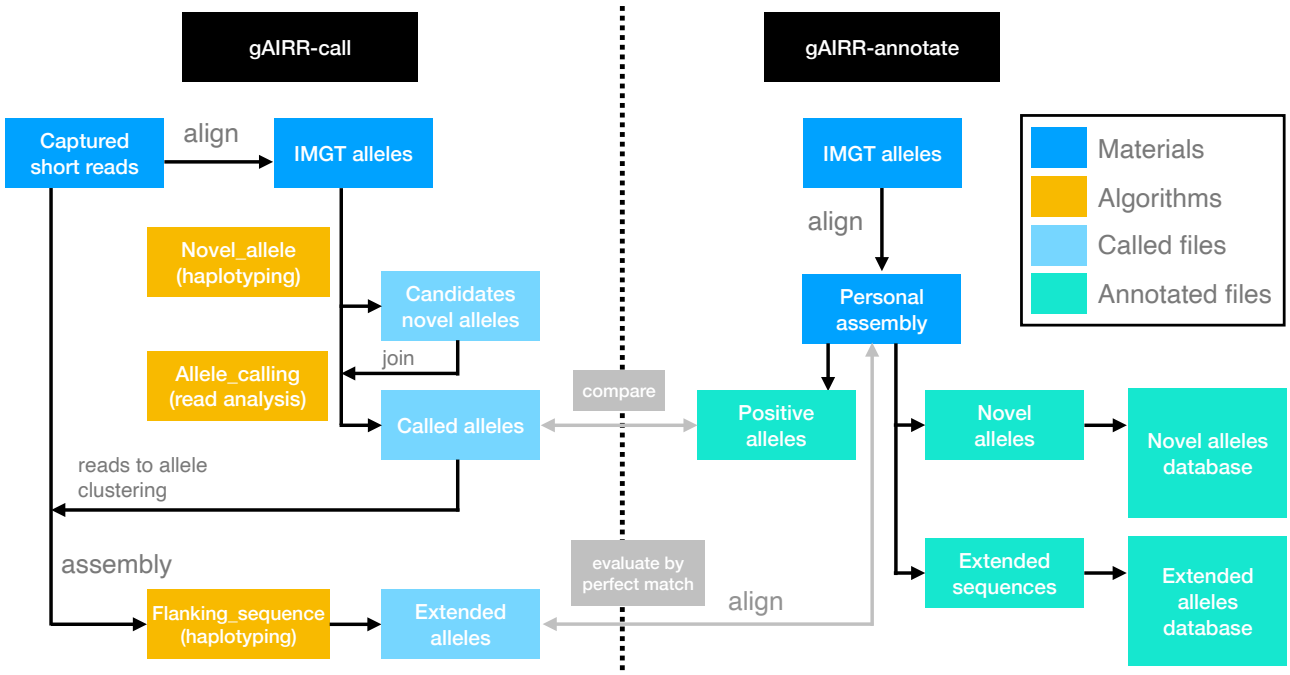


Figure 1. The gAIRR-call and gAIRR-annotate pipelines. The gray blocks and lines show the verification pipeline when both gAIRR-seq reads and personal assembly are available.

Table 1. The number of reads sequenced, fraction of aligned reads and on-target rate of gAIRR-seq using sampels HG001/NA12878 and HG002/NA24385. The target region is defined as the union of annotated allelic regions, where each region includes the core allele and 800 bps extended from both sides.

Sample	# of reads sequenced	% of reads aligned (H1/H2)	% in target region (H1/H2)
HG001 / NA12878	280,579	97.6 / 98.0	83.3 / 83.7
HG002 / NA24385	281,494	98.5 / 96.1	84.5 / 83.3

Results

Probe-based target enrichment sequencing is effective and efficient in AIRR regions

Our gAIRR-seq runs captured more than 280 thousand paired-end reads for each sample. To evaluate the sequencing quality, we aligned the captured reads to HG001 and HG002's whole-genome assemblies[11]. 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 1). 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 makes gAIRR-seq efficient and effective to enrich AIRR sequences and provide high-quality data for computational analyses.

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

We used gAIRR-call to accurately call AIRR alleles using the high-coverage targeted read set sequenced with 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 13 high-quality assemblies). According to the annotations, gAIRR-call called TRV and TRJ alleles with 100% accuracy, including both known and novel alleles, for both

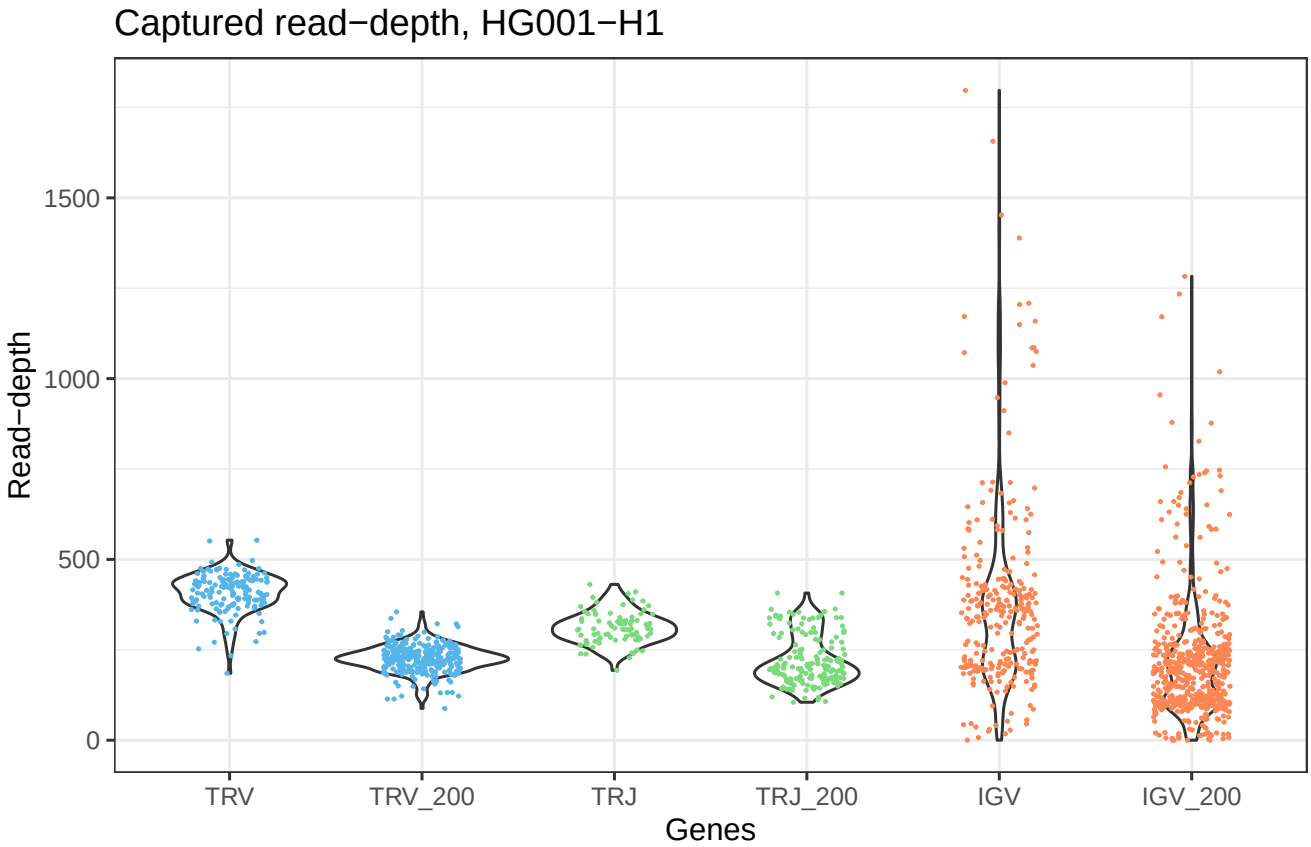


Figure 2. Read depths sequenced with gAIRR-seq in TRV, TRJ and BRV regions using data from HG001/NA12878. Columns without the "_200" suffix shows the average read-depth of a region. Columns with the "_200" suffix shows the read depth 200 bp away from the region boundaries.

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). 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 flanking sequences of TRV alleles. Flanking sequences outside AIRR alleles are challenging and currently not well studied. 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 corresponding assembly of the sample. 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 regarded as 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.

It is worth mentioning that there were three extended alleles not aligned perfectly to the personal assemblies, but we still considered them as true positives after manual inspection. The genes *TRBV20-1* in HG001-H1, *TRBV17* 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[17] and concluded the called alleles were likely to be correct.

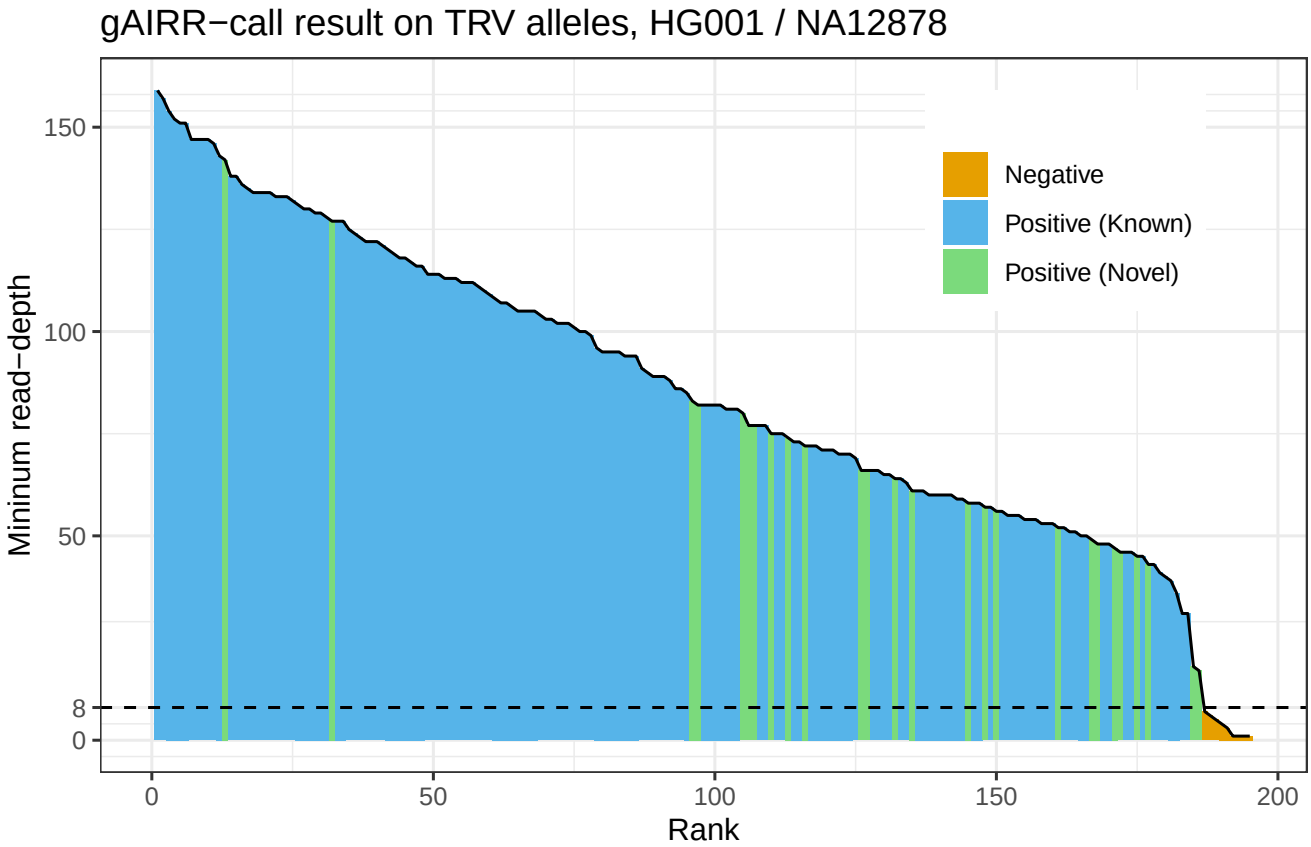
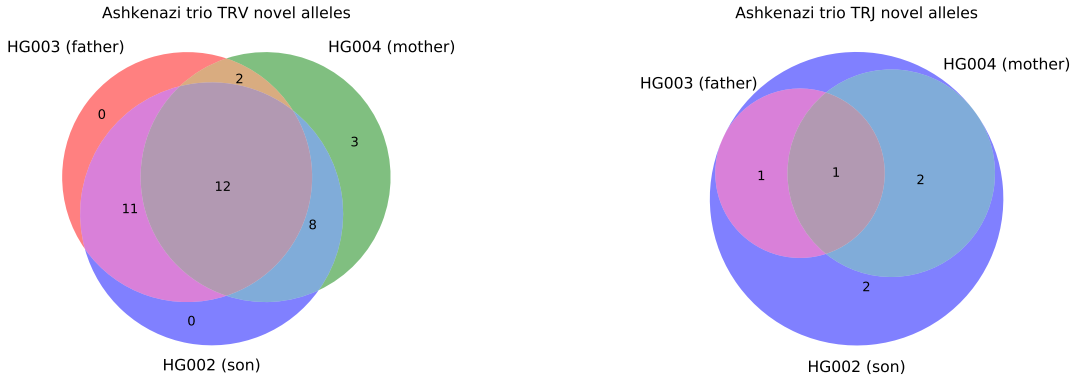


Figure 3. gAIRR-call results using HG001 data. The results are sorted by read-depth and the dash line represents the adaptive threshold in gAIRR-call. All the alleles not annotated by gAIRR-annotate, colored in orange, are below the adaptive threshold and are regarded as nocalls by gAIRR-call. True known 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.



(a) (b)

Figure 4. The (a) TRV and (b) TRJ novel allele relationship in the Ashkenazi family. HG002: son; HG003: father; HG004: mother

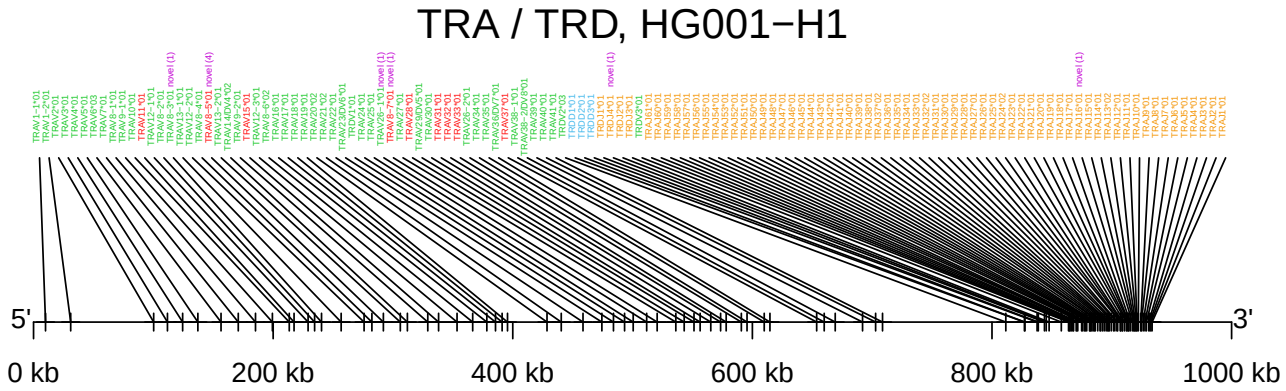


Figure 5. Positions of TRA / TRD alleles called by gAIRR-annotate. The purple text indicates that the allele is novel with its edit-distance compared to the most similar base allele inside parentheses.

Annotating TR alleles using 13 high-quality assemblies

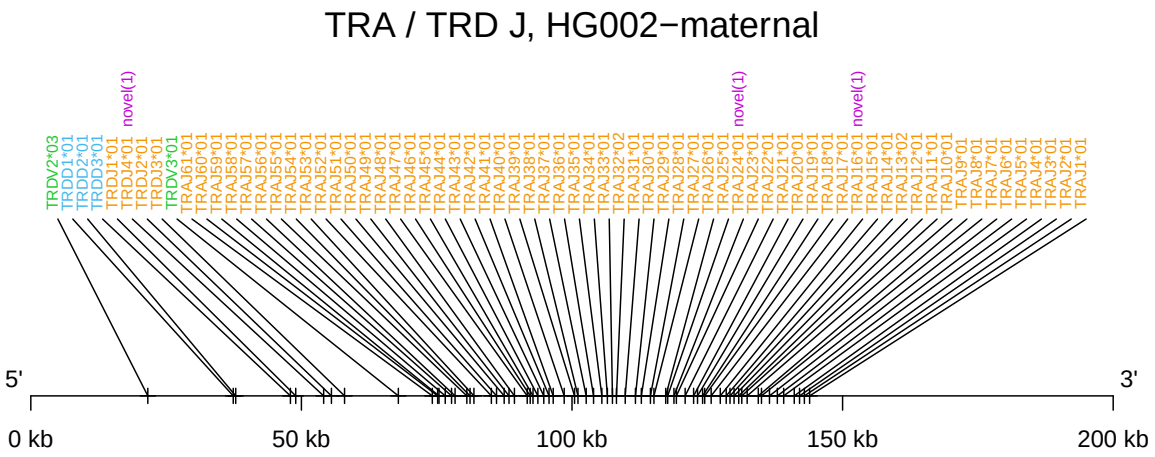
We collected 13 whole-genome assemblies from 6 works of literature and annotated TR alleles using gAIRR-annotate (Table 2). The dataset had 21 haploids from 6 different samples, including diploid samples HG001 (also known as NA12878, Northwest European American from Utah), HG002 (also known as NA24385, Ashkenazi Jewish), HG00733 (Puerto Rican), NA19240 (Yoruban), and two haploid hydatidiform moles CHM1 and CHM13. HG002, HG00733, and CHM13 had multiple assemblies generated from different research groups. We showed that the number of TR alleles were conserved — all samples had 143-150 TRV alleles, and all but HG002 had 5 TRD alleles and 78-90 TRJ alleles. HG002 is known to carry two long deletions in the paternal haplotype in the TRD and TRJ region, resulting in the difference in number of TR alleles([18]; see Section Result: gAIRR-annotate identifies two structural variants for details). The number of annotated alleles was concordant with our knowledge of human TR germline.

The novel alleles called by gAIRR-annotate showed concordance with 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 17 novel V alleles and 3 J alleles in each haplotype (Table 2). 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[19]. It shows that the novel alleles found by gAIRR-annotate, marked by purple text, are known genes with few single nucleotide polymorphisms (SNPs) to the known allele cataloged in IMGT.

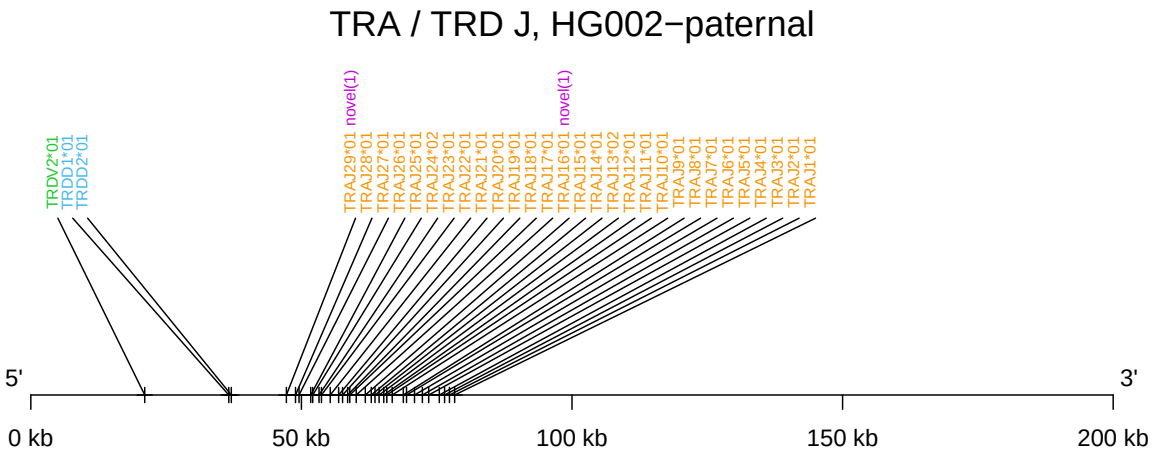
gAIRR-annotate identifies two structural variants in HG002. We further confirmed the structural variants with capture-based sequencing

From Table 2, 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 6 (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 8 (b)).

To further verify the 65 kbp deletion in the TRA/TRD loci of HG002, we aligned capture-based short reads from HG002 to GRCh37 as well as captured reads from HG003 (HG002's father) and HG004 (HG002's mother) to perform trio analysis (Figure 7). We first examined if there were reads aligned across two deletion breakpoints, either in a chimeric form or having two mated 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 a 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



(a)



(b)

Figure 6. The 65 kbp structural variation in TRA / TRD J region of HG002 / NA24385. Upper: maternal haplotype, lower: paternal haplotype

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

We provided two independent investigations using gAIRR-annotate with a diploid HG002 assembly[11] 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 Genome in a bottle[18] 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[18] was similar with the suggestion by gAIRR-call. The deletion of HG002's chromosome contributes to a loss of 32 TRAJ alleles and all 4 TRDJ alleles at the paternal haplotype.

Figure 7. 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.

TR structural variations in GRCh37 and GRCh38

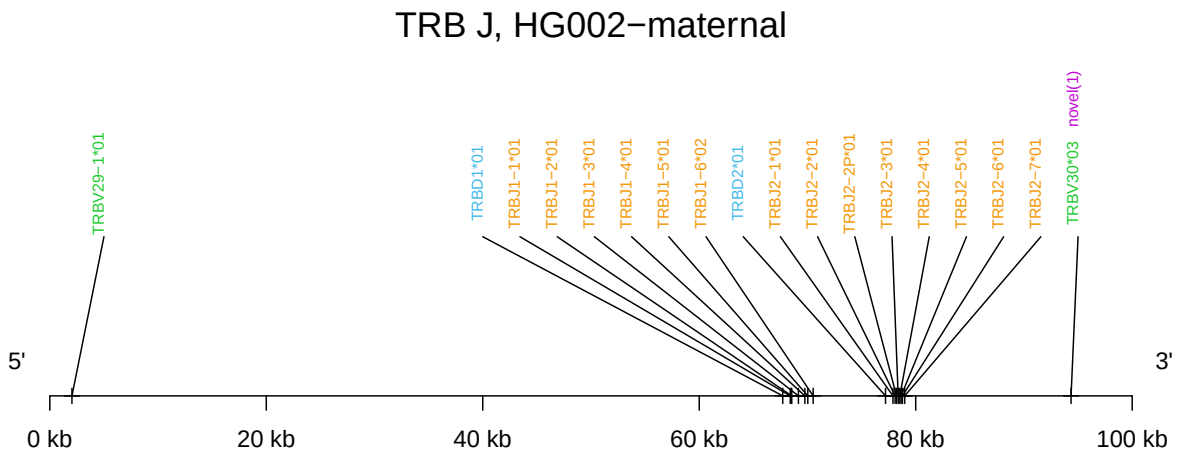
We performed gAIRR-annotate to analyze human reference genomes GRCh37 and GRCh38 and identified structural variants in TR regions (Figure 9). First, we identified an inversion in GRCh37 with respect to the GRCh38 primary assembly that covered TRBV alleles. In this inversion there were 39 TRBV alleles (*TRBV6-2* to *TRBV18* following the allele order provided by IMGT) successfully annotated in GRCh38, whereas 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 10kbp) 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[20]. According to this chain file, the TRBV inversion located around the breakpoint at chr7:142,048,195 (GRCh37). In this region there were 278k bases in GRCh37 unmatched with 270k bases 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.

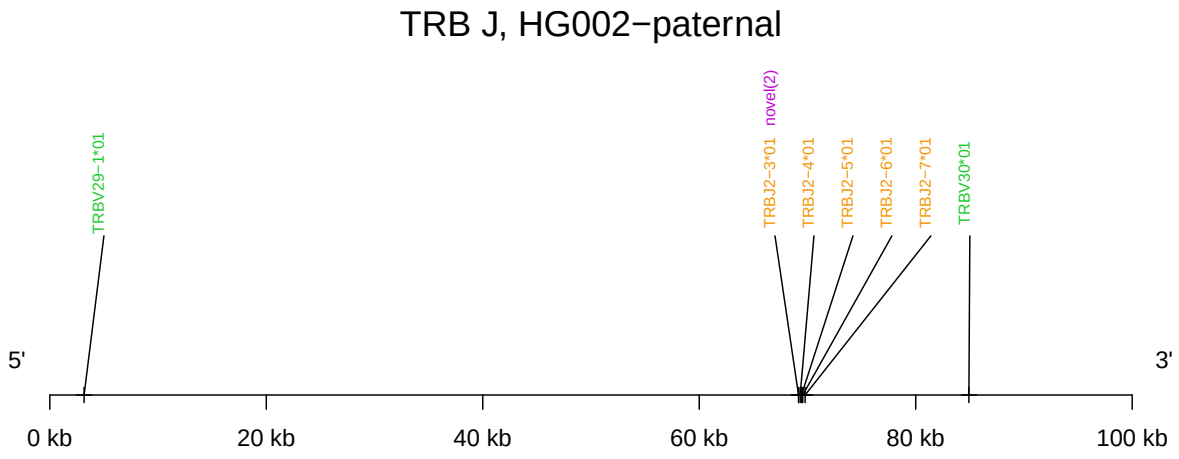
We further analyzed an GRCh38 alternative contig, *chr7_KI270803v1_alt*, which contained TRBV alleles (Figure 9 bottom). Eleven novel TRBV alleles in *chr7_KI270803v1_alt* were identified using gAIRR-annotate. Compared to the GRCh38 primary assembly, *chr7_KI270803v1_alt* had six additional TRBV alleles. Both TRB alleles in chr7 and *chr7_KI270803v1_alt* showed concordance with the Locus representation of IMGT[19]. 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 probe capture-based targeted sequencing pipeline, gAIRR-seq, for germline AIRR genotyping. The probes successfully enriched the AIRR regions with high on-target rate and sufficient read depth. We developed gAIRR-call and gAIRR-annotate pipelines for AIRR genotyping. gAIRR-call can call the alleles both reported by IMGT and novel ones. The



(a)



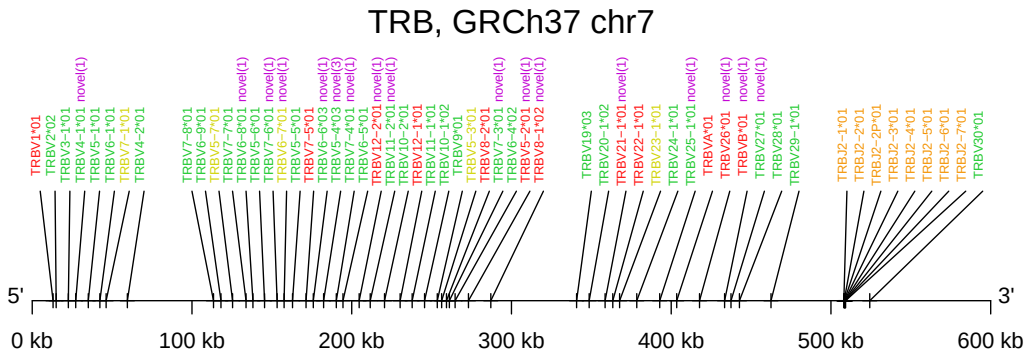
(b)

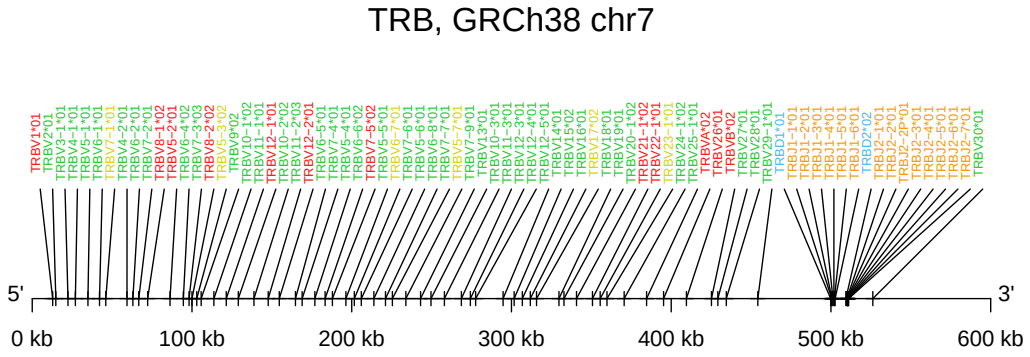
Figure 8. The 10 kbp structural variation in the TRB J region of HG002 / NA24385. Upper: maternal haplotype, lower: paternal haplotype.

gAIRR-call results were then verified by gAIRR-annotate and trio data analysis. gAIRR-annotate can find the relative positions of AIRR genes in a sample. With gAIRR-annotate, we built a novel allele database based on IMGT v3.1.22. The novel allele database was further verified by IMGT v3.1.29. gAIRR-annotate also discovered two structural variants in HG002, which were also verified by capture-based reads. Similarly, gAIRR-annotate showed two large disagreements between reference genome GRCh37 and GRCh38.

Pros and Cons of gDNA analysis

gDNA, as the starting material, is convenient to preserve and obtain[7]. The un-VDJ recombined AIRR genes are relatively simpler to analyze than recombined ones. Moreover, there are already numerous gDNA samples publicly available or preserved by individual organizations or investigators. Hence, the gAIRR-seq and gAIRR-call pipelines can be applied to widely available gDNA samples, and the gAIRR-annotate pipeline can be used for existing genome assemblies with minimal modification. On the other hand, gDNA cannot provide the expression level of the final VDJ recombination product. These pipelines cannot figure out the dynamic changes of lymphocyte clones at specific time points with gDNA, which is the advantage of analysis





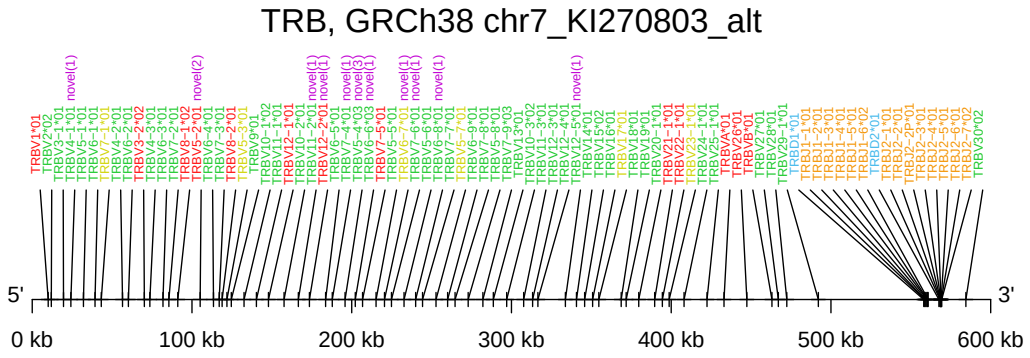


Figure 9. Subfigure from top to bottom is the gAIRR-annotate locus representation of TR beta germline alleles of GRCh37 chr7, GRCh38 chr7, GRCh38, and GRCh38 alternative contig chr7_KI270803v_alt.

Table 2. Number of TR alleles for 13 assemblies (from 6 samples with 21 haplotypes). The novel and total indicates how many novel alleles and total alleles are annotated in a haplotype. From left to right are TRV allele, extended TRD with heptamers on both sides, and TRJ alleles.

	TRV		TRD (+hep)		TRJ	
assembly	#novel	#total	#novel	#total	#novel	#total
HG001-H1[11]	14	145	0	5	2	84
HG001-H2[11]	19	146	0	5	3	84
HG002-M[11]	17	147	0	5	3	84
HG002-P[11]	23	144	0	2	4	39
HG002-CCS-Canu-P[12]	24	144	0	2	4	39
HG002-CCS-Canu-M[12]	19	148	0	5	3	84
HG002-CCS-Falcon-P[12]	20	144	0	2	4	39
HG002-CCS-Falcon-M[12]	21	146	0	5	4	84
HG002-CCS-wtdbg2-P[12]	23	142	0	2	4	39
HG002-CCS-wtdbg2-M[12]	18	145	0	5	3	82
HG002-hifiasm-P[13]	23	144	0	2	4	77
HG002-hifiasm-M[13]	17	147	0	5	3	84
HG00733-hifiasm-H1[13]	13	143	0	5	3	84
HG00733-hifiasm-H2[13]	11	145	0	5	0	84
HG00733-HiFi-v0-H1[14]	13	143	0	5	3	84
HG00733-HiFi-v0-H2[14]	11	145	0	5	0	84
CHM13-hifiasm[13]	11	147	0	5	1	86
CHM13-T2T[15]	11	147	0	5	1	86
CHM13-GCA[16]	11	147	0	5	1	86
CHM1-GCA[16]	15	141	0	5	3	84
NA19240-GCA[16]	31	147	0	5	3	84

using mRNA.

Pros and Cons of capture-based enrichment

The most critical advantages of gAIRR-seq are its comprehensiveness and high resolution. As compared to the multiplex PCR-based methods, gAIRR-seq can cover all known AIRR genes and alleles at the same time. Furthermore, novel alleles can also be identified because the probes can tolerate sequence mismatch to 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 to process 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.

The capture-based enrichment method (the combined gAIRR-seq and gAIRR-call pipeline) also empowers researchers to gather sequences of the flanking regions. The addition of flanking sequences to the core gene/allele sequence may help future alignment-based genotype calling using whole genome sequencing data, and will be beneficial for population genetic study.

One additional advantage of 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.

Library preparation using probe-capture enrichment may to certain degree be biased under different probe preferences. Nevertheless, according to our results, gAIRR-call can still successfully genotype the alleles. The only bias we observe 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 (details in method). The lower read coverage of *TRAV8-5* is likely due to the high allele length-probe ratio. By adding more probes according to the allele length, the bias can probably be solved.

Probe designs for TRD, IGD, and IGJ genes

In this study, we design capture probes for V and J genes based on the alleles cataloged in IMGT databases. We do not include TRD, IGD or IGJ genes because these genes' lengths are shorter than 60 bp. The typical lengths of D genes range from 8 to 16 bp, way shorter than a capture probe. To overcome the issue, we need to take advantage of the flanking sequences to design all

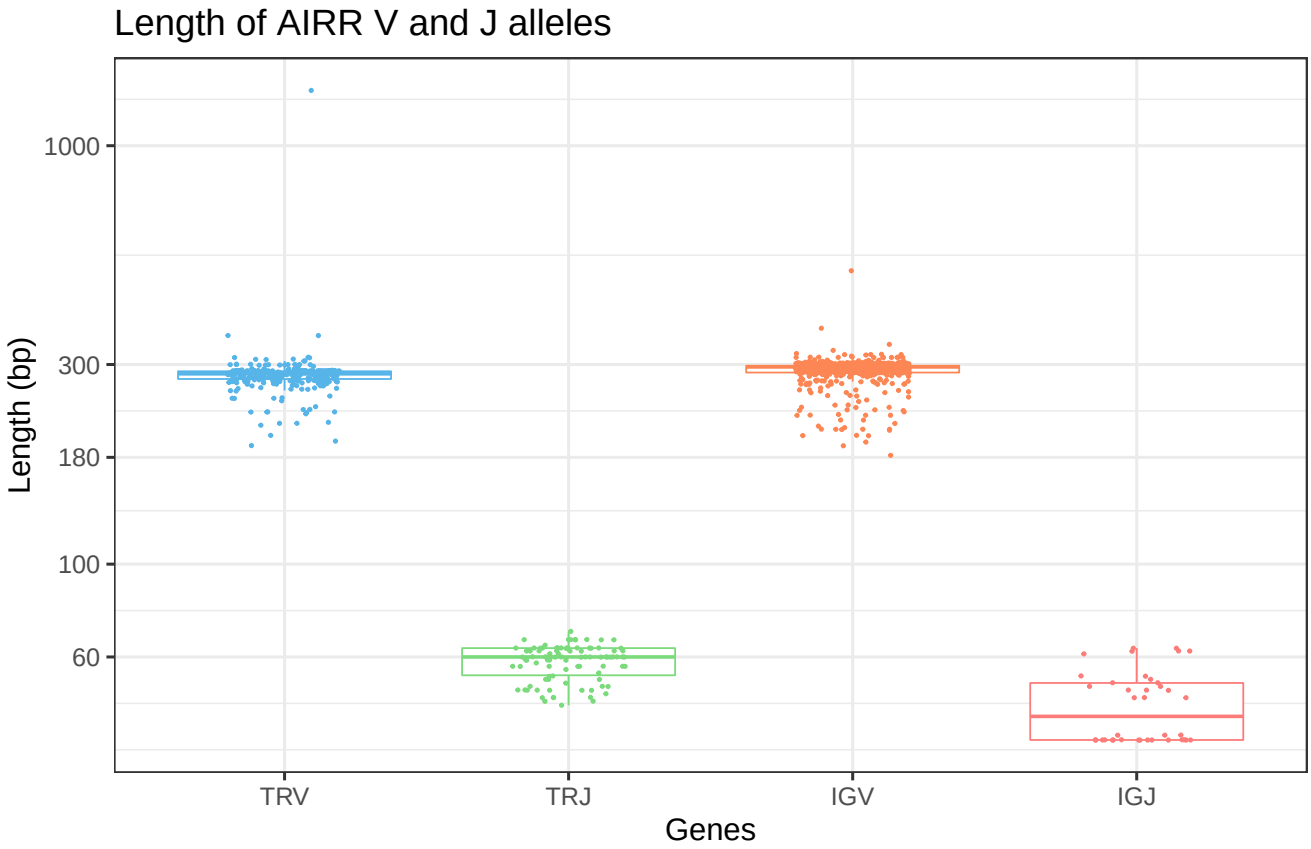


Figure 10. The allele length distribution of AIRR V alleles and J alleles according to IMGT v3.1.22.

the D and J genes' probes. Fortunately, gAIRR-annotate shows that the flanking sequences of D alleles are conservative. There are only 7 types of extended TRD alleles among the 6 samples, while there are only 6 TRD alleles in the IMGT database.

Applying gAIRR-call to IG cell lines

Although the pipelines are technically applicable to IG genes, we find the accuracy lower than TR genes. We hypothesize the reason to be that the 7 RM samples are from EBV-transformed B lymphocytes. The genomic DNA of the EBV-transformed B lymphocyte cell line has lost its diversity during the establishment, growth, and subculture, so it may not be a suitable material for testing the IG genes[21]. In contrast, it is feasible to use this source of gDNA to determine the TR gene. To circumvent the problem, we plan to use EBV-transformed T lymphocytes or primary cells for IG cell line calling in the future.

Methods

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

We obtained 7 genomic DNA reference materials (RMs) from the Genome in a Bottle (GIAB) Consortium 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 7 genomes were well sequenced by more than ten next-generation sequencing platforms by GIAB[22], and therefore were appropriate RMs for testing novel methods.

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)[9]. Each probe was a continuous 60-bp oligo synthesized by Roche/NimbleGen (Roche NimbleGen, Madison, WI, U.S.). Each V allele was covered by three probes while each J allele was covered by one probe based on the lengths of V and J alleles (Figure 10). The probes for J alleles shorter than 60 bps were padded at both ends with uncertain nucleotide "N" to 60 bp in length.

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 bps 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.) according to the standard protocol from the manufacturer. We then used Illumina MiSeq 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 pb flanking sequences on both ends.

Find 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 IMGT alleles. Then gAIRR-call checks the reference IMGT alleles one by one. For any known allele, the variants are called with the aligned reads. gAIRR 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.

Call AIRR alleles

Compared to the previous step, calling AIRR allele applies a 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[23] so that one read can be aligned to multiple positions.

After alignment, gAIRR-call removes alignment that contain mismatches or indels. Further, for alleles longer than 100 bp, the aligned reads 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 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 slighlty 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 theshold to decide whether they are positive.

The gAIRR-annotate pipeline

The gAIRR-annotate pipeline annotates AIRR alleles using publicly available whole-genome assemblies (Table 2). 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.

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 range from 21 to 31 and they usually appear only once in the whole-genome, suggesting the uniqueness of TRD alleles with appropriate RSS sequences.

References

1. 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).
2. 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).
3. Watson, C. T., Glanville, J. & Marasco, W. A. The individual and population genetics of antibody immunity. *Trends immunology* **38**, 459–470 (2017).
4. Chung, W.-H. *et al.* A marker for stevens–johnson syndrome. *Nature* **428**, 486–486 (2004).
5. Pan, R.-Y. *et al.* Identification of drug-specific public tcr driving severe cutaneous adverse reactions. *Nat. communications* **10**, 1–13 (2019).
6. 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).
7. Liu, X. & Wu, J. History, applications, and challenges of immune repertoire research. *Cell biology toxicology* **34**, 441–457 (2018).
8. Rosati, E. *et al.* Overview of methodologies for t-cell receptor repertoire analysis. *BMC biotechnology* **17**, 61 (2017).
9. 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).
10. Bankevich, A. *et al.* Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J. computational biology* **19**, 455–477 (2012).
11. Garg, S. *et al.* Efficient chromosome-scale haplotype-resolved assembly of human genomes. *bioRxiv* 810341 (2019).
12. 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).
13. 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).
14. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome. *bioRxiv* 855049 (2019).
15. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human x chromosome. *Nature* **585**, 79–84 (2020).
16. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. methods* **16**, 88–94 (2019).
17. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. biotechnology* **29**, 24–26 (2011).
18. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *BioRxiv* 664623 (2019).
19. Lefranc, M.-P. & Lefranc, G. *The T cell receptor FactsBook* (Elsevier, 2001).
20. Kent, W. J. *et al.* The human genome browser at ucsc. *Genome research* **12**, 996–1006 (2002).
21. 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).
22. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. data* **3**, 1–26 (2016).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* (2013).

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.