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], and Pei-Lung Chen[1,2,4,8,9,*]

[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
[+]these authors contributed equally to this work
[*]e-mail: paylong@ntu.edu.tw

ABSTRACT

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 complexity. Our gAIRR Suite comprises three modules. gAIRR-seq, a probe capture-based targeted sequencing pipeline, profiles AIRR from individual DNA samples. gAIRR-call and gAIRR-annotate call alleles from gAIRR-seq reads and annotate whole-genome assemblies respectively. We gAIRR-seqed TRV and TRJ of seven Genome in a Bottle (GIAB) DNA samples with 100% accuracy, and discovered novel alleles. We gAIRR-annotated 36 genomes to cumulatively unearth 325 novel TRV alleles and 29 novel TRJ alleles. We could further profile the flanking sequences, including the recombination signal sequence (RSS). We validated two structural variants for HG002. We uncovered substantial conflicts of AIRR genes in references GRCh37 and GRCh38. The gAIRR Suite can potentially benefit genetic study and future clinical applications of various immune-related phenotypes.

Main

Germline AIRR is the collection of TR and IG, which are composed of variable (V), diversity (D) and joining (J) regions. The receptors' final sequences require somatic V(D)J recombination, where one gene from each V, (D) and J genetic region are selected and rearranged into a continuous fragment[1,2], with or without additional somatic hypermutation. Although humans' immune repertoire has theoretically extreme huge variability capable of binding all kinds of antigens[3–5], for each individual the number of V(D)J genes/alleles in one's genome is limited[3]. Therefore, any given individual's immunological response may be restricted according to the germline AIRR composition[3,6,7]. Early data suggested SARS-CoV-2 antibodies prefer to use certain IG genes/alleles[8]. Carbamazepine-induced Stevens-Johnson syndrome[9], besides the well-known HLA-B*15:02 susceptibility in Asians, also depends on specific public TRs for the severe cutaneous adverse reaction[10]. In a simian immunodeficiency virus infection model, the TR alleles responsible for the potent CD8+ T cell response were recently determined[11]. Germline AIRR alleles have started to be found associated with human diseases, including rheumatic heart disease[12], and Kawasaki disease[13,14]. These examples are likely only the tip of the iceberg because determining germline AIRR for any given individual is currently insurmountable.

Both genomic DNA (gDNA) and messenger RNA (mRNA)/complementary DNA (cDNA) can be used to profile AIRR of an individual, albeit yielding quite different information[15,16]. The popular mRNA/cDNA approach provides useful dynamic information of AIRR clonotypes and expression levels, which can be valuable if, and only if, the most relevant lymphocytes can be retrieved from the right site at the right time[15]. mRNA/cDNA-based AIRR sequencing at any given time point can only

cover part of the full spectrum of germline AIRR. Depending on the library preparation strategies, the mRNA/cDNA approach might suffer from allele dropout, lack of full-length coverage and inconsistent results[17]. Furthermore, the non-coding flanking sequences, including RSS which may significantly affect recombination efficiency, cannot be profiled in the mRNA/cDNA approach. On the other hand, bulk gDNA includes non-rearranged and rearranged V(D)J segments and preserves the full spectrum of germline AIRR[15]. Besides, gDNA is more stable and easier for storage compared to mRNA, with millions of gDNA samples already stored in the clinical setting or in previous genetic studies. Although mRNA/cDNA-based AIRR can be profiled by many methods, gDNA-based AIRR has just started to attract attention[18–20].

In this study, we developed the gAIRR Suite comprising three modules. gAIRR-seq, a probe capture-based targeted gDNA sequencing pipeline, profiles germline AIRR from individual DNA samples. The probe hybridization method can tolerate sequence mismatch to a certain degree and therefore can discover novel alleles. gAIRR-call (Figure 1), using a coarse-to-fine strategy, starts from aligning gAIRR-seq reads to known AIRR alleles collected from the international ImMunoGeneTics (IMGT) database[21], to eventually call both known and novel AIRR alleles, as well as their flanking sequences. 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, including novel ones, using whole-genome assemblies (Figure 1). In this study we discovered numerous novel alleles, structural variants, substantial conflicts in different versions of reference genome, and RSS polymorphism across subjects. The gAIRR Suite can potentially benefit genetic study and future clinical applications for various immune-related phenotypes.

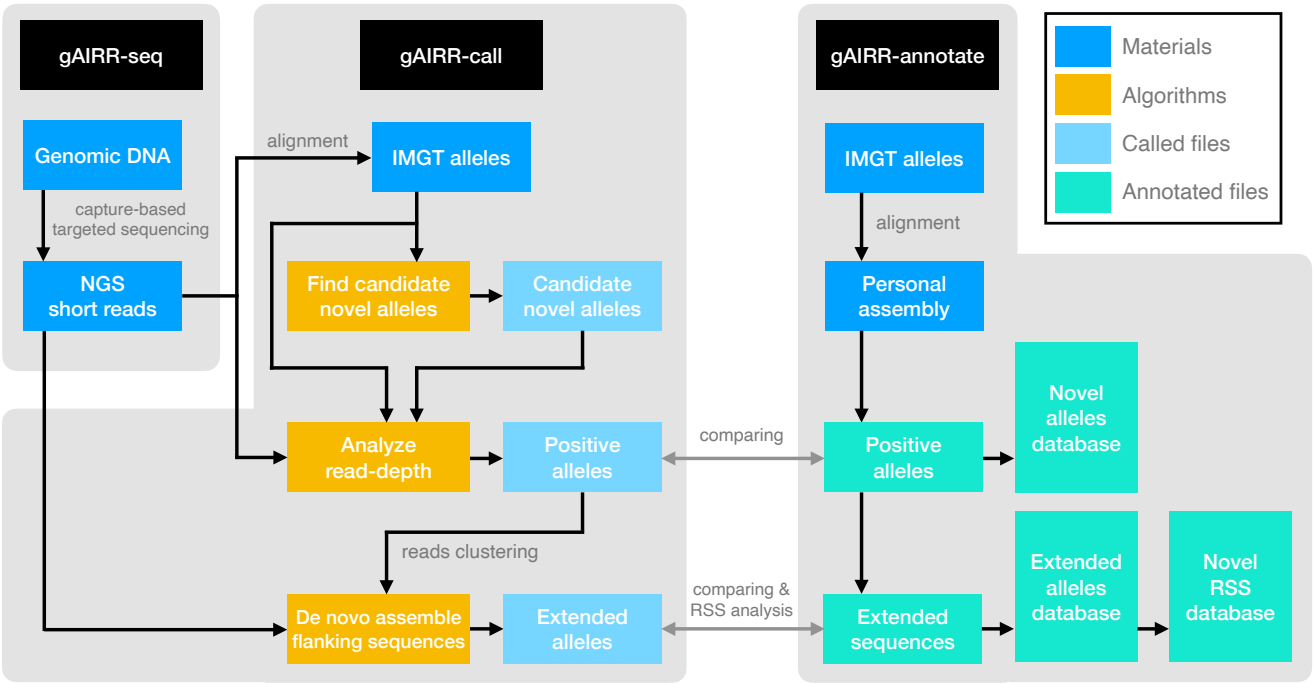

Figure 1. The gAIRR-Suite pipelines. The gray arrows show the verification methods between the two pipelines when both gAIRR-seq reads and personal assembly are available. (Methods: Comparing alleles of gAIRR-call and gAIRR-annotate)

Results

gAIRR-seq achieved efficient targeted sequencing in the AIRR regions

We gAIRR-seqed GIAB reference materials (RMs) HG001-7 (EBV-transformed B lymphocytes) and a primary cell sample (peripheral blood mononuclear cell, PBMC) to generate paired-end reads (2 x 300 bp), with a mean of 310 thousand reads for each DNA sample. To evaluate the sequencing quality and efficiency, we aligned the gAIRR-seq reads of the GIAB RMs, HG001 and HG002, to their individual whole-genome assemblies[22], with over 96% of the reads successfully aligned. Over 83% of the sequenced reads were on-target in the AIRR regions in HG001 and HG002 (Table S1). In core TRV and TRJ allelic regions where probes were designed, the average read depth was above 350; the sequencing depth at positions 200 bp away

from the allele boundaries were typically above 200x (Figure 2a). Much higher read depth variation was found in IGV alleles, which might be biased because the DNA was from EBV-transformed B lymphocytes (see Discussion). High high on-target rate and high coverage make gAIRR-seq efficient to enrich AIRR sequences and provide high-quality data for computational analyses.

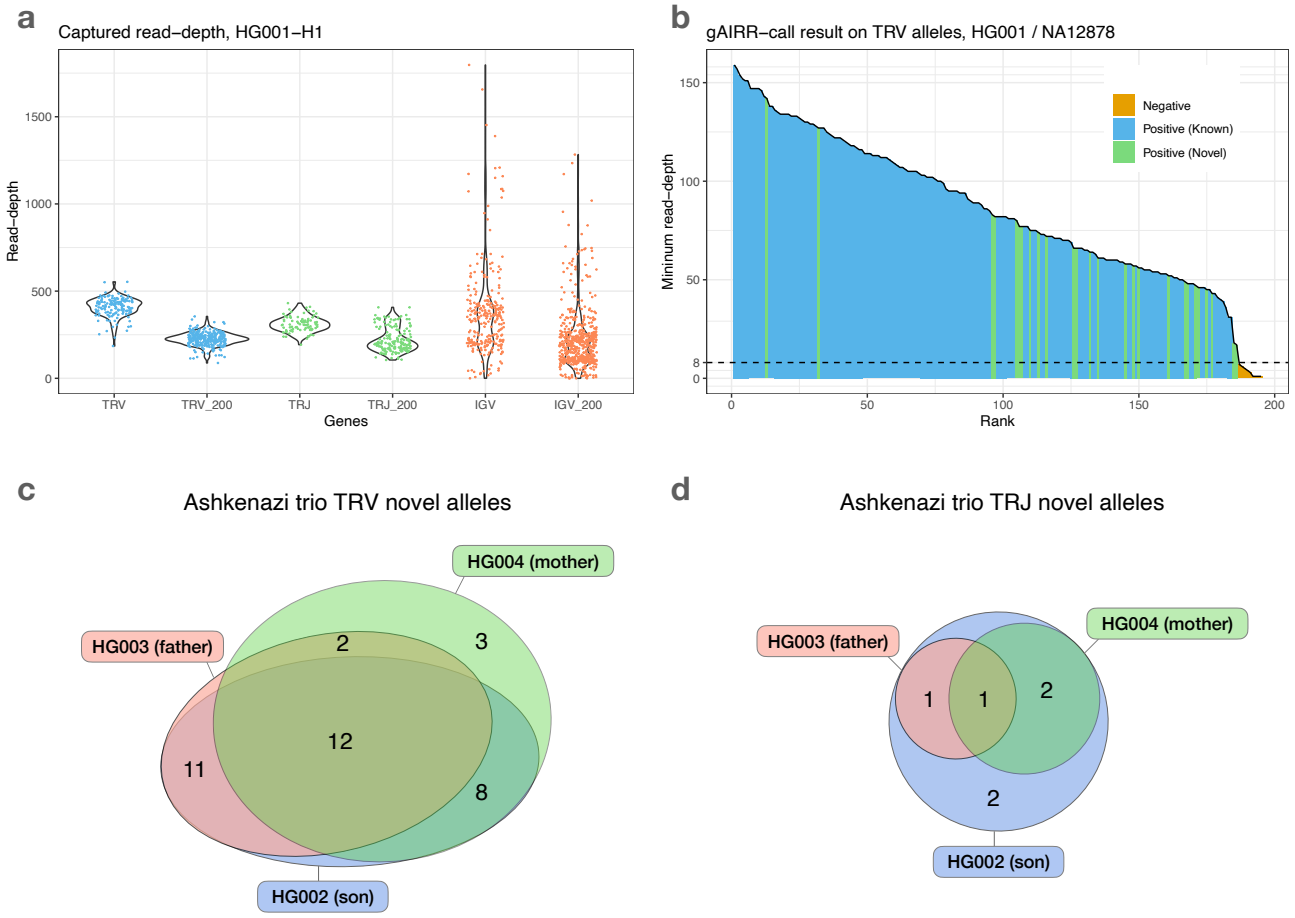

Figure 2. gAIRR-seq and gAIRR-call results. a, Read depths sequenced with gAIRR-seq in TRV, TRJ and IGV regions using data from HG001/NA12878. Columns without the "_200" suffix shows the average read-depth of a region. Columns with the "_200" suffix shows the read depth 200 bp away from the region boundaries. **b,** gAIRR-call results using HG001/NA12878 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 not included in the Figure. **c,** The TRV and **d,** TRJ novel allele relationship in the Ashkenazi family. HG002: son; HG003: father; HG004: mother. In **c** TRV, there are no unique TRV alleles owned by HG002. In other words, all the HG002's novel TRV alleles are also owned by either HG003 or HG004. In **d** TRJ, HG002 owns two unique TRJ alleles, which violates Mendelian inheritance. These two alleles are related to de novo structural variants (See Section gAIRR-annotate identified in HG002 two structural variants which were further confirmed with gAIRR-seq)

gAIRR-call could profile known and novel germline TRV and TRJ alleles

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).

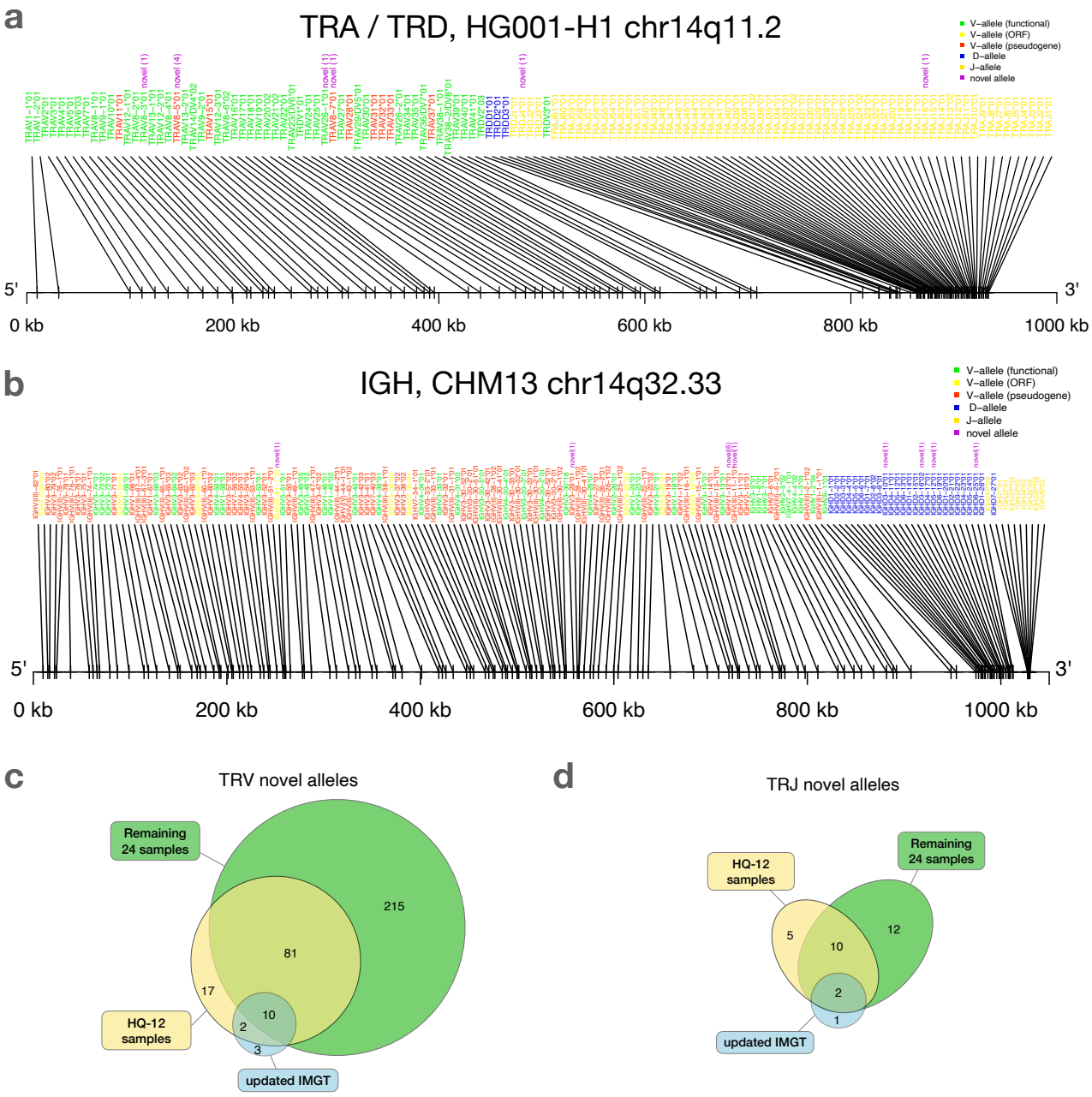

Figure 3. gAIRR-annotate results. a, Positions of HG001's TRA and TRD alleles determined by gAIRR-annotate. The purple text indicates that the allele is novel with its edit-distance compared to the most similar base allele inside parentheses. The text color of the gene names are basically according to IMGT's color menu for genes[21]. **b,** Positions of CHM13[23]'s IGH alleles determined by gAIRR-annotated. The figure settings are the same as in **a**. **c,** TRV and **d,** TRJ 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.

To validate the accuracy of the gAIRR-call results, we compared them with the high-confidence annotations from gAIRR-annotate on HG001 and HG002 using their high-quality whole-genome assemblies (Section Results: Annotating TR alleles using high-quality assemblies). According to the comparison, gAIRR-call called TRV and TRJ alleles with 100% accuracy, including both known and novel alleles, for both HG001 and HG002 (Figure 2b). Furthermore, we provided additional validation using trio analysis, including an Ashkenazi family (son: HG002, father: HG003 and mother: HG004) and a Chinese

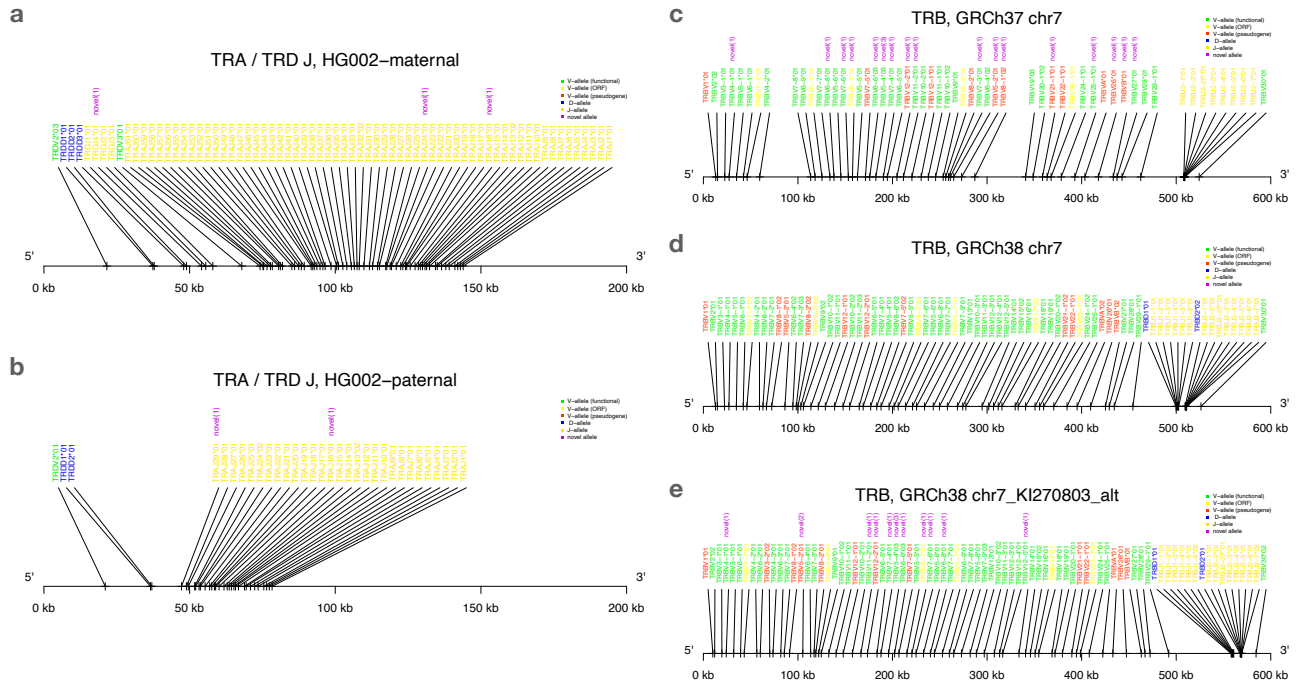

Figure 4. Structural variants found by gAIRR-annotate. a,b, The 65 kbp structural variation in TRA and TRD J region of HG002/NA24385. **a**: maternal haplotype, **b**: paternal haplotype. **c,d,e,** The gAIRR-annotate locus representation of TR beta chain germline alleles of **c** GRCh37 chr7, **d** GRCh38 chr7, GRCh38, and **e** GRCh38 alternative contig chr7_KI270803v_alt.

family (son: HG005, father: HG006 and mother: HG007). We performed gAIRR-call, including assigning novel alleles, independently on each of the samples of the two trios. We showed that there was no Mendelian violation in either family except two calls from HG002 (at *TRAJ29* and *TRBJ2-3*) (Figure 2c,d 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 conclude that the gAIRR-seq pipeline with gAIRR-call is accurate in both known and novel alleles.

Further, we used gAIRR-call to analyze TRV alleles' flanking sequences outside the core region, which are important for the V(D)J recombination mechanism but have not been comprehensively collected in the database. Similar to verifying core alleles, we utilized gAIRR-annotate to provide truth sets from whole-genome assemblies. 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 further checked if the generated flanking sequences carried appropriate RSS, which were critical in V(D)J recombination. All 7 GIAB RMs and the primary cell sample carried identical or nearly identical RSS with known patterns in IMGT.

It is worth mentioning that we considered three gAIRR-seqd/gAIRR-called extended alleles correct though they did not align perfectly to the personal assemblies HG001 and HG002. We found that, after manual inspection, in all three instances the personal assemblies contained indel errors because of PacBio sequencing weakness in homopolymers (Supplementary Section The flanking sequences where gAIRR-seq and gAIRR-call disagree personal assembly).

gAIRR-annotate could profile AIRR alleles using genome assemblies

We collected 106 haplotypes from 6 works of literature[22–27] and annotated TR alleles using gAIRR-annotate. The dataset includes 36 individuals with a broad population background diversity. We first selected 12 extremely high-quality samples to form a "HQ-12" set, where each sample could either be validated with family information or was curated with additional effort (See Section HQ-12 set). We showed that 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). gAIRR-annotate demonstrated HG002 to have fewer TR alleles, which is consistent with the known two long deletions in the paternal haplotype in the TRD and TRJ region[28]; see Section Result: gAIRR-annotate identifies two structural variants for details). We then gAIRR-annotated the remaining 24 samples from

Figure 5. 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.

HGSVC[24] and showed that the allele numbers were similar (Figure 3c,d).

The novel alleles discovered 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*, which resides in the middle of *TRAJ15* and *TRAJ17* according to IMGT. Thus, we were confident that the allele was a novel allele of a known gene. On average, there were 18 novel V alleles and 3 novel J alleles in each haplotype (Table S2). We examined the alleles' relative positions for HG001 (Figure 3a) and HG002, and the TRA and TRD alleles' relative positions were in the same pattern as the *Locus representation of IMGT*[29] (Figure S3). 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 gene locus map (Figure S4).

For IG alleles, after avoiding genome assemblies from long-term cultured EBV transformed B lymphocytes, we only applied gAIRR-annotate to the CHM13 genome. We identified 32 novel alleles from the IGH, IGK, and IGL locus. In addition, there are also 32 novel alleles from the orphons of CHM13 (Figure 3b; Table S5).

Identified novel alleles are consistent with, and far outnumbering, the latest IMGT update

We identified a large number of novel alleles from the 36 samples, as compared with the IMGT database. We used IMGT v3.1.22 (April, 2019) 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, July, 2020) had 15 TRV and 3 TRJ alleles added for human TR since v3.1.22. The novel alleles identified by gAIRR-annotate far outnumbered the IMGT updates (325 vs. 15 for TRV and 29 vs. 3 for TRJ; Figures 3c,d). For the alleles updated by IMGT, a large fraction of TRV and TRJ alleles (14 out of 18) was also discovered by gAIRR-annotate (Figure 3c,d), demonstrating high efficiency and accuracy of gAIRR Suite.

The gAIRR-annotate method can also annotate extended TR alleles, which we defined as the core alleles plus 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 RSS in the extended regions given their important participation in the V(D)J recombination mechanism. According to IMGT gene-DB, most TR RSSs 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. We used gAIRR-annotate to identify known RSS patterns in the 36 genomes and discovered 56 novel TRV and 20 novel TRJ RSSs (Table 1). 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 the nonamer (heterozygous).

We build a novel-allele database collecting all identified novel alleles, an extended-allele database collecting alleles with 200-bp flanking sequences, and an RSS database collecting RSS information in flanking sequences for both TRV and TRJ (novel-allele: Supplementary File S1 to S4; extended-allele: Supplementary File S5 to S8; RSS: Supplementary File S9 to S12). We showed that we could broadly expand the size of known TR 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 traditional methods.

Table 1. Number of known and novel TR RSS in the gAIRR-called and gAIRR-annotated flanking sequences. The number of RSS known in IMGT[21] and novel RSS are shown in columns *known* and *novel* respectively. The functionality *Func* indicates that the RSSs come from functional genes while *P/O* indicates that the RSSs come from pseudogenes or open reading frames (ORFs). The RSS from HG001-7 and the primary cell sample are called from gAIRR-seq and gAIRR-call while the RSS from HQ-12 and the additional 24 samples are called from gAIRR-annotate.

	TRV				TRJ			
	#known		#novel		#known		#novel	
sample	Func	P/O	Func	P/O	Func	P/O	Func	P/O
HG001/NA12878	94	43	4	1	66	10	7	1
HG002/NA24385	95	43	6	1	65	10	9	1
HG003/NA24149	94	44	3	1	66	10	6	1
HG004/NA24143	91	43	7	1	65	10	8	1
HG005/NA24631	93	44	8	2	67	10	7	1
HG006/NA24694	96	46	6	1	67	10	8	1
HG007/NA24695	95	44	7	2	67	10	5	1
Primary cell sample	95	43	6	2	67	10	6	1
HQ-12 set samples	96	46	14	3	67	10	11	3
HGSVC-aditional-24 samples	96	46	31	22*	67	10	15	3

*: In addition to 22 pseudogene and ORF alleles with novel RSS, we didn't find any appropriate nonamers at *TRBV24/OR9-2*03* (ORF) for HG01505.

gAIRR-annotate identified in HG002 two structural variants which were further confirmed with gAIRR-seq

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 4a,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 S5b).

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, as well as gAIRR-seq reads from HG003 (father) and HG004 (mother), to GRCh37 for a trio analysis (Figure 5). By examining the split reads across a certain region, and also the read depth reduction within a continuous region, we confirmed the *de novo* deletion in HG002 (see Section Methods: Verify HG002's structural deletion with gAIRR-seq reads). The deletion contributes to a loss of 32 TRAJ genes and all 4 TRDJ genes from the paternal haplotype.

Our results indicated a long deletion in chr14 from position 22,918,114 to position 22,982,924 (GRCh37) for one HG002 haplotype. The structure variant analysis performed by GIAB[28] 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.

Similarly, there is a 10 kb structural deletion in chromosome 7 q34 related to TRB alleles. We verified the deletion with gAIRR-annotate, gAIRR-seq, and the report from GIAB[28]. (Supplementary Section 8 HG002's structural deletion in

chromosome 7)

gAIRR-annotate uncovered substantial conflicts in GRCh37 and GRCh38

We identified an inversion in GRCh37 with respect to the GRCh38 primary assembly that covered TRBV alleles (Figure 4c,d,e). In this inversion, there were 39 TRBV genes (*TRBV6-2* to *TRBV18* following the gene order provided by IMGT) successfully annotated in GRCh38. In contrast, in GRCh37, only 27 genes from *TRBV8-1* to *TRBV7-8* were successfully annotated, but with a reversed order. We observed another long missing sequence (about 10 kbp) in GRCh37 in the TRB region, covering about half of the expected TRBD and TRBJ genes. We noticed GRCh38 didn't carry this deletion, and we annotated reasonable numbers of TRBD and TRBJ genes. When comparing annotated alleles to the IMGT database, we showed that on chromosome 7, all alleles in GRCh38 are known, and there are 18 novel alleles in GRCh37. In IG alleles, there are also some differences between the reference genomes; however, all the genome structures are in consistence with IMGT's locus representation. (Figure S8; Table S5)

We further analyzed an GRCh38 alternative contig, *chr7_KI270803v1_alt*, which contained TRBV alleles (Figure 4e). We identified eleven novel TRBV alleles in *chr7_KI270803v1_alt* using gAIRR-annotate. Compared to the GRCh38 primary assembly, *chr7_KI270803v1_alt* has six additional TRBV genes. TRB genes in both chr7 and *chr7_KI270803v1_alt* showed concordance with the *locus representation* of IMGT[29]. 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

We have developed the gAIRR Suite that includes probe capture-based targeted sequencing and computational analyses to profile germline AIRR. gAIRR-seq can enrich the AIRR regions with high on-target rates and sufficient read depth, and gAIRR-call can then accurately call both known and novel alleles using gAIRR-seq reads. gAIRR-annotate can take advantage of whole genome assemblies to accurately annotate AIRR genes in both core allelic and flanking regions. We applied gAIRR-annotate to identify 325 novel core alleles in TRV and 29 novel core alleles in TRJ from 36 subjects. We identified RSSs with spacers neither 12-bp nor 23-bp in length, which had been mentioned by[30] and[31] but not yet comprehensively profiled. gAIRR-annotate can also help to discover structural variants. We verified two known structural variants in the TR regions in HG002. We also searched for structural variants in the IG regions of CHM13, and found that the genome structures are consistent with IMGT's locus presentation. Similarly, we used gAIRR-annotate to uncover substantial conflicts between references GRCh37 and GRCh38.

gAIRR-seq is advantageous in its high resolution and comprehensiveness. Compared to multiplex PCR-based methods, gAIRR-seq covers all known AIRR genes and alleles in one experiment. Furthermore, novel alleles can also be identified because the probes can tolerate sequence mismatch to a certain degree. 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 PBMC. We envision that gAIRR-seq capture probe design can be further upgraded by gAIRR-annotate results thanks to the collection of accurate 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 in future gAIRR-seq probe design.

Although the gAIRR-seq and gAIRR-call pipeline is technically applicable to IG genes, we found that in this study the accuracy for IG genes was lower. We reason that the seven 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[32]. To solve this problem, we plan to use primary cells, such as PBMCs or buccal cells, for IG experiments in the future. In this study, we gAIRR-seqed the genomic DNA from PBMCs of a Taiwanese subject (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 TR results are convincing and imply that we can apply the same pipeline to IG genes using primary cells.

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[3,12,14]. 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 successfully used to retrieve AIRR information for human genetic study due to high genomic complexity in the AIRR regions and relatively lower sequencing depth compared to enriched sequencing. 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 do[33,34], is an attractive question awaiting more studies. We badly need a HapMap[35] equivalent for AIRR.

We envision that gAIRR Suite can help in at least the following scenarios. First, the discovered novel alleles can facilitate development of better SNP arrays and better WGS analysis references of AIRR in the future. Second, the personal germline AIRR profile can be directly employed for genetic study and clinical applications. 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 on various immune-related phenotypes. Last but not least, the combined analysis of germline AIRR and dynamic mRNA-based AIRR profiles might help decipher molecular mechanisms underlying immunological responses.

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 HG001/NA12878 and two Personal Genome Project trios - an Ashkenazim Jewish ancestry: HG002/huAA53E0, HG003/hu6E4515, HG004/hu8E87A9; a Chinese ancestry: HG005/hu91BD69, HG006/huCA017E, HG007/hu38168C. The genomic DNA samples were retrieved from EBV-transformed B lymphocytes. These RMs were well-sequenced using more than 10 NGS platforms/protocols by GIAB[36]; therefore, they are appropriate RMs for benchmarking gAIRR-seq.

A primary cell sample

We also extracted the genomic DNA from the PBMC sample of a Taiwanese subject. The subject signed informed consent for participating in the project, and all procedures were approved by the Research Ethics Committee of the National Taiwan University Hospital.

Probes design

We designed probes for all known V and J alleles of IG and TR based on the IMGT database Version 3.1.22 (3 April 2019) (including all functional, ORF, and pseudogenes)[21]. 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 S11). 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[37] 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 2b).

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)[38] 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 annotated AIRR core alleles as well as the extended flanking sequences using publicly available whole-genome assemblies (Table S2). gAIRR-annotate begined 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 reported all potential sites. For any allele aligned with mismatches or indels, we identified its nearest allele in the IMGT database and assigned a novel allele to the associated gene. When there were multiple assemblies available for the same sample, we only annotated an allele if more than half of the assemblies reported the same output.

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 (Figure S12).

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.

HQ-12 set

The HQ-12 set was selected 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.

The alleles called using different assemblies from a single sample were concordant in most cases. The pipeline used a conservative strategy for high-precision 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)

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. 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.

Verify HG002's structural deletion with gAIRR-seq reads

We aligned capture-based short reads of gAIRR-seq from HG002. 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.

We also compared the alignment result of HG002 to those of HG003 (HG002's father) and HG004 (HG002's mother). 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 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 S4). This verified the long deletion in the HG002 TRA/TRD region and further provided evidence that the structural variant was likely a novel mutation.

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. Schultheiß, C. *et al.* Next-generation sequencing of t and b cell receptor repertoires from covid-19 patients showed signatures associated with severity of disease. *Immunity* **53**, 442–455 (2020).
9. Chung, W.-H. *et al.* A marker for stevens–johnson syndrome. *Nature* **428**, 486–486 (2004).
10. Pan, R.-Y. *et al.* Identification of drug-specific public tcr driving severe cutaneous adverse reactions. *Nat. communications* **10**, 1–13 (2019).
11. 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).
12. 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).
13. 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).
14. Johnson, T. A. *et al.* Association of an ighv3-66 gene variant with kawasaki disease. *J. human genetics* 1–15 (2020).
15. Liu, X. & Wu, J. History, applications, and challenges of immune repertoire research. *Cell biology toxicology* **34**, 441–457 (2018).
16. Rosati, E. *et al.* Overview of methodologies for t-cell receptor repertoire analysis. *BMC biotechnology* **17**, 61 (2017).
17. Barennes, P. *et al.* Benchmarking of t cell receptor repertoire profiling methods reveals large systematic biases. *Nat. Biotechnol.* **39**, 236–245 (2021).
18. Rodriguez, O. L. *et al.* A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front. immunology* **11** (2020).
19. Ford, M., Haghshenas, E., Watson, C. T. & Sahinalp, S. C. Genotyping and copy number analysis of immunoglobin heavy chain variable genes using long reads. *Iscience* **23**, 100883 (2020).
20. Khatri, I. *et al.* Population matched (pm) germline allelic variants of immunoglobulin (ig) loci: New pmig database to better understand ig repertoire and selection processes in disease and vaccination. *bioRxiv* (2020).
21. 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).
22. Garg, S. *et al.* Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. biotechnology* **39**, 309–312 (2021).
23. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human x chromosome. *Nature* **585**, 79–84 (2020).
24. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* (2021).
25. 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).
26. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
27. 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).
28. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *BioRxiv* 664623 (2019).
29. Lefranc, M.-P. & Lefranc, G. *The T cell receptor FactsBook* (Elsevier, 2001).

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. *BMC cancer* **18**, 1–13 (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. Thorisson, G. A., Smith, A. V., Krishnan, L. & Stein, L. D. The international hapmap project web site. *Genome research* **15**, 1592–1593 (2005).

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

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

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

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

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.