gAIRR-wgs: An Algorithm to Genotype T Cell Receptor Alleles Using Whole Genome Sequencing Data

Kuan-Ta Huang [1†], Yu-Hsuan Yang [1], Mao-Jan Lin [2], Sheng-Kai Lai [3,4], Ting-Hsuan Chou [1], Chieh-Yu Lee [1], Tsung-Kai Hung [1], Chia-Lang Hsu [5], Ya-Chien Yang [6,7], Chien-Yu Chen [8], **Pei-Lung Chen** [1,3,4,9‡], **Jacob Shu-Jui Hsu** [1‡]

[1] Graduate Inst. of Med. Genomics and Proteomics, Coll. of Med., Natl. Taiwan Univ., Taipei, Taiwan;
[2] Dept. of Computer Sci., Johns Hopkins Univ., Baltimore, MD, USA;
[3] Genome and Systems Biology Degree Program, Academia Sinica and Natl. Taiwan Univ., Taipei, Taiwan;
[4] Dept. of Med. Genetics, Natl. Taiwan Univ. Hosp., Taipei, Taiwan;
[5] Dept. of Med. Res., Natl. Taiwan Univ. Hosp., Taipei, Taiwan;
[6] Dept. of Clinical Lab. Sci. and Med. Biotechnology, Coll. of Med., Natl. Taiwan Univ., Taipei, Taiwan;
[7] Dept. of Lab. Med., Natl. Taiwan Univ. Hosp., Taipei, Taiwan;
[8] Dept. of Biomechatronics Engineering, Coll. of Bioresources and Agriculture, Natl. Taiwan Univ., Taipei, Taiwan;
[9] Div. of Endocrinology and Metabolism, Dept. of Internal Med., Natl. Taiwan Univ. Hosp., Taipei, Taiwan.
† antony1229@gmail.com ; ‡ jacobhsu@g.ntu.edu.tw , paylong@gmail.com

ABSTRACT

T cell receptor (*TR*) genes, including variable (TR_V), diversity (TR_D), and joining (TR_J) segments, exhibit allelic diversity that is critical to adaptive immunity. Growing evidence has identified associations between *TR* genes and immune-related diseases. Germline variants may influence *TR* gene function and subsequent usage, highlighting the importance of accurate *TR* allele profiling. However, accurately identifying germline *TR* from standard WGS data remains challenging due to short read lengths, limited depth, and high sequence similarity. To address these challenges, we developed gAIRR-wgs, for WGS-based *TR* allele typing. By incorporating novel alleles from HPRC individuals, gAIRR-wgs exhibited excellent performance in allele calling, with F1 scores of 100.0% for TR_D, 99.8% for TR_J, and 98.3% for TR_V. Applying this pipeline to 1,492 individuals from the Taiwan Biobank (TWB), we identified 449 novel *TR* alleles, 277 of which overlapped with HPRC release 1 data of mixed ethnicity and are absent in the IMGT database. Further population comparison analysis revealed significant *TR* allele distribution differences across global populations, showing population-specific patterns and diversity variations between ethnic groups. We also discovered TWB-specific deletion polymorphisms affecting contiguous *TRGV* and *TRBV* genes, which are not recorded in the gnomAD database and undetected by standard structural variant callers, highlighting the need for tailored approaches to resolve complex immune gene regions. In conclusion, gAIRR-wgs enables accurate *TR* allele calling from standard WGS data using feasible computational resources and reveals substantial immunogenetic diversity in population cohorts.

INTRODUCTION

T cell receptors, together with immunoglobulins which include both membrane-bound B cell receptors and soluble antibodies, form the adaptive immune receptor repertoires (AIRR), the foundation of adaptive immunity. T cell receptors enable T cells to recognize antigens presented by human leukocyte antigens (HLAs), driving cell-mediated immune responses. T cell receptors consist of α:β or γ:δ peptide chains, encoded by *T cell receptor* (*TR*) genes, including *TRA*, *TRB*, *TRG*, and *TRD*, which are located at distinct genomic loci: *TRA* and *TRD* are co-located on chromosome 14, *TRB* and *TRG* reside on chromosome 7, and an additional orphon *TRB* locus is located on chromosome 9. Each *TR* locus is composed of multiple gene segments: variable (V), diversity (D), and joining (J), which undergo somatic recombination to form functional *TR* genes during T cell development in the thymus. All *TR* loci contain V and J gene segments, while *TRB* and *TRD* additionally possess D gene segments. Within each locus, multiple copies of V, D, and J genes are arranged either in clusters or interspersed along the chromosome, and each individual gene can exist as multiple allelic variants. The germline *TR* (g*TR*) repertoire refers to the set of unrearranged V, D, and J gene segments that collectively generate *TR* diversity through recombination. The international ImMunoGeneTics information system (IMGT) database serves as the nomenclature authority and central repository for g*TR* genes, providing curated reference sequences based primarily on core coding regions (Lane et al., 2010). In this study, we classify g*TR* genes by segment type, given that genes within each segment category share similar lengths and intrinsic sequence complexity: TR_V refers to all variable genes across all *TR* loci, while TR_D and TR_J represent the diversity and joining genes, respectively. The functional diversity

48 arising from g*TR* recombination manifests as the expressed adaptive immune receptor repertoire
49 (exprAIRR), which has been widely associated with disease susceptibility, clinical progression, and
50 prognosis. Non-random selected g*TR* gene combinations, have been observed across a wide range of
51 diseases, including autoimmune diseases (Z. Chen et al., 2021; Fichtner et al., 2020; Jacobsen et al.,
52 2017; Jia et al., 2022; Mitchell & Michels, 2020; Turcinov et al., 2023; Wang et al., 2016; Zheng et
53 al., 2021), infectious diseases (Dolton et al., 2022; Frimpong et al., 2022; Gras et al., 2010; Li et al.,
54 2021; Marín-Benesiu et al., 2024; Mazouz et al., 2021; Miles et al., 2011; Minervina et al., 2021; Shao
55 et al., 2022; Shomuradova et al., 2020), severe cutaneous adverse reactions (SCARs) (Pan et al., 2019),
56 and various cancers (Y. T. Chen et al., 2021; Huda et al., 2021; Keane et al., 2017; Wang et al., 2016;
57 Yu et al., 2024; Zhang et al., 2024). Despite growing evidence for the role of exprAIRR in diseases,
58 the impact of g*TR* allele variation remains insufficiently studied. Distinct g*TR* allelic forms of the same
59 gene can lead to altered V(D)J recombination outcomes, including the production of nonfunctional
60 transcripts or pseudogenes, thereby directly affecting gene availability, contributing to non-random
61 gene usage, and ultimately shaping immune responses and disease susceptibility (Charmley et al., 1993;
62 Omer et al., 2022). Current approaches for g*TR* allele identification can be broadly categorized into
63 two methodologies: inference from rearranged transcripts (exprAIRR) and direct calling from germline
64 DNA sequences. While transcript-based methods capture expressed repertoires, they cannot
65 comprehensively assess the complete germline allelic diversity, as not all alleles are necessarily
66 expressed or detectable in a given sample. Direct germline calling from whole-genome sequencing
67 (WGS) data offers a more complete view of g*TR* variation; however, the high sequence similarity

among *TR* gene segments poses substantial challenges for accurate allele identification from short-read technologies. Existing germline *TR*-calling tools, such as gAIRR-Suite, have demonstrated promising performance but are primarily optimized for targeted panel sequencing with ultra-high depth (>500X), making them incompatible with standard WGS data (~30X) (Lin et al., 2022). This limitation is particularly significant given that biobanks offer an unprecedented opportunity to explore population-level g*TR* diversity, with most containing individual short-read WGS data at moderate sequencing depth (~30X) (Bick et al., 2024; Elfatih et al., 2024; Feng et al., 2022; Flanagan et al., 2024; Halldorsson et al., 2022; Hsu et al., 2024; Pajuste & Remm, 2023). While individual-level g*TR* variants impact immune function, population-level variation provides essential biological context—not only establishing the baseline diversity against which individual deviations are interpreted, but also revealing patterns shaped by evolutionary selection and population-specific immune pressures. However, no existing tools can effectively and accurately identify *TR* alleles from the standard-depth short-read WGS data commonly available in biobanks. To address this critical gap, we developed gAIRR-wgs, a WGS-compatible module for gAIRR-Suite (Lin et al., 2022), to bring high-accuracy *TR* allele identification to standard-depth biobank data.

83 **RESULTS**

84 **1. Development of the gAIRR-wgs pipeline for *TR* genotyping using WGS data**

85 The main challenges in calling g*TR* alleles from WGS data included lower read depth compared to
86 targeted sequencing, shorter read lengths (150 bp), and the limited ability to leverage paired-end
87 information due to the g*TR* structure (**Supplementary Figure 1**). Additionally, WGS data
88 encompasses the entire genome, introducing substantial computational challenges and background
89 noise from non-target regions. To address these limitations, we developed gAIRR-wgs, a specialized
90 module of gAIRR-suite optimized for calling *TR* alleles from WGS data (**Figure 1**).

91 To optimize computational efficiency and reduce background noise, we implemented a targeted read
92 extraction strategy. WGS reads were first aligned to the GRCh38 reference genome using BWA-MEM,
93 followed by extraction of reads mapping to extended *TR* genomic regions using SAMtools
94 (**Supplementary Figure 2**). Rather than using exact *TR* gene coordinates on reference genome, we
95 extended the target regions by approximately 2 kb to capture informative reads potentially located
96 outside strict gene boundaries. Unmapped reads were also collected to account for potential
97 discrepancies between the reference genome and individual-specific sequences. This extraction
98 process dramatically improved computational performance. Prior to extraction, analysis of a single
99 sample required 6-7 hours with maximum memory usage exceeding 500 GB. After extraction, paired-
100 end read files were reduced in size from approximately 26 GB each to 380 MB each, enabling the same
101 analyses to be completed in 12-15 minutes with maximum memory usage of approximately 5 GB,
102 representing a >25-fold improvement in both time and memory efficiency.

103 We expanded the reference database by incorporating 335 novel *TR* alleles previously discovered and
104 validated using long-read sequencing data (Yang et al., 2025) (**Figure 1**). To improve read mapping
105 efficiency and allele discrimination, we extended each reference sequence with 50 bp flanking
106 sequences on both ends (**Figure 1**, **Supplementary Figure 1**). These flanking sequences were derived
107 from HPRC assembly data using a consensus-based strategy that included sequences supported by at
108 least 25% of the HPRC samples, providing a balance between sequence diversity and reliability
109 (**Figure 1**, **Supplementary Figure 3**). Since flanking sequences represented merged consensus
110 versions that might not perfectly match individual-specific sequences, we implemented some
111 flexibilities in subsequent algorithms that allows reads to contain variants in flanking regions while
112 maintaining perfect matches in core gene regions during depth calculation.
113 Following the original gAIRR-panel workflow for targeted-sequencing, we performed initial
114 alignment using BWA-MEM in paired-end mode to map reads to the expanded reference database
115 (**Figure 1**). Candidate novel alleles were identified when variants were supported by at least 25% of
116 reads at specific positions. However, due to the incorporation of flanking regions in the database, novel
117 allele calling was restricted to variants located within core gene regions, as defined by IMGT
118 nomenclature guidelines (Lefranc, 2011), to avoid calling flanking sequence polymorphisms as novel
119 alleles. Candidate novel alleles were then incorporated into the database for following allele calling.
120 To address two common sources of alignment ambiguity, including hybridized-structured alleles and
121 nested-structured alleles, we implemented a comprehensive strategy combining alignment parameter
122 modifications, sophisticated read weighting systems, and validation mechanisms. Hybridized-

structured alleles present a particular challenge during alignment, as their 5' region may share similarity with one allele while their 3' region resembles another allele within the same gene (**Supplementary Table 1**). During alignment, reads derived from other alleles may be misaligned to these alleles with hybridized-structured, creating artificial support and leading to false positive calls. This issue is particularly pronounced in WGS data with 150 bp reads and lower sequencing depth (30x), where alignment ambiguity is more severe. To mitigate these hybridized artifacts, we disabled secondary alignments by removing the "-a" option in BWA-MEM, forcing each read to align to only its best-matching reference. Given the relatively short length of g*TR* reference sequences, many paired-end reads cannot map as proper pairs within the same reference, potentially leading to the exclusion of informative single-end reads if paired-end criteria are strictly enforced. To more effectively leverage information from both single-end and paired-end reads, we developed an edit distance-based weighting system that assigns differential weights based on read type and mapping characteristics. This system requires perfect matches for reads mapping to core regions while allowing tolerance for variants in flanking regions, reflecting the fact that flanking sequences in our expanded database are derived from merged sequences and may legitimately contain variants. Single-end and paired-end reads receive different tolerance levels and weights within this framework, with paired-end reads generally receiving higher confidence scores due to their increased reliability.

We defined allele depth using customized depth-defining regions (DDRs) that calculate the minimum weighted read depth across specified positions within each allele sequence (**Figure 1**). For most alleles, DDRs are restricted to core regions to ensure both adequate read depth and coverage specificity.

143 However, for nested-structured alleles, where the core sequence of a shorter allele is identical to an
144 internal region of a longer allele (**Supplementary Table 2**), we extended DDRs to include flanking
145 regions to enable proper discrimination these two alleles. The depth calculation at each position
146 incorporates our distance-based weighting system, accounting for read length, reference length, read
147 pairing status, and variant location relative to core and flanking regions.
148 For alleles exhibiting hybridized structures, we implemented an additional validation to prevent false
149 positive calls resulting from misaligned reads that could artificially combine variants from different
150 true alleles. We clustered all reference alleles based on sequence similarity using Clustal Omega
151 (version 2.3.0) and performed multiple sequence alignments within each cluster (Madeira et al., 2024).
152 Within these alignments, we manually identified key variants as positions that distinguish between
153 alleles within the same cluster, encompassing both allele-dependent variants that differentiate alleles
154 of the same gene level and novel variants that distinguish from their closest IMGT-documented
155 counterparts. For alleles containing multiple key variants, we required that each key variant be
156 supported by at least one properly paired read group where both reads align primarily to the same
157 reference allele, ensuring that the observed variant combination reflects genuine allelic structure rather
158 than alignment artifacts.
159

160 ## 2. Benchmarking demonstrates high accuracy of the gAIRR-wgs pipeline

161 The gAIRR-wgs pipeline was benchmarked against 44 HPRC subjects with validated *TR* annotations
162 from personal genome assemblies. When the gAIRR-panel, originally designed for targeted
163 sequencing, was directly applied to WGS data, it achieved suboptimal performance with only 72.5%
164 recall for TR_D and 75.4% recall for TR_V genes, demonstrating the limitations of applying panel-
165 optimized algorithms to WGS data (**Table 1**, **Supplementary Table 3**). In contrast, the optimized
166 gAIRR-wgs pipeline achieved excellent performance with 100% recall and 100% precision for TR_D
167 genes, 99.9% recall and 99.6% precision for TR_J genes, and 99.2% recall and 97.4% precision for
168 TR_V genes. When examining the performance by allele type, the pipeline showed distinct patterns
169 for known versus novel alleles. For documented alleles present in the IMGT database (Version 3.1.41;
170 accessed on 2025-04-19), gAIRR-wgs achieved 100.0% recall and 99.6% precision for TR_J genes,
171 and 99.7% recall and 98.1% precision for TR_V genes. For novel alleles not previously catalogued in
172 reference databases, the pipeline maintained robust performance with 98.5% recall and 97.5%
173 precision for TR_J genes, and 98.3% recall and 93.9% precision for TR_V genes. Notably, these results
174 exceeded the performance of the gAIRR-panel on high-depth targeted sequencing data, demonstrating
175 the effectiveness of our WGS-specific optimizations.

176 To further assess the robustness of our pipeline under conditions that mimic real-world scenarios with
177 incomplete g*TR* databases, we performed additional validation by randomly removing 10%, 30%, 50%,
178 70%, and even all of the TR_V and TR_J novel alleles from the reference database according to each
179 sample's annotation results. For TR_J genes, performance remained relatively stable, with both

180 accuracy and sensitivity preserved under 30% removal and high precision still retained even with 50%
181 loss (**Supplementary Figure 4**, **Supplementary Table 4**). In contrast, the more complex and diverse
182 TR_V genes were more vulnerable to database incompleteness: recall dropped substantially once more
183 than 10% of novel alleles were removed, though precision remained relatively high. These findings
184 confirm that our strategy of incorporating previously discovered novel alleles into the database is
185 essential to achieving robust allele calling, and also highlight that incomplete databases can critically
186 constrain performance, particularly for highly diverse gene families such as TR_V. Importantly, the
187 results also demonstrate that once the reference database reaches a sufficient level of completeness,
188 recall of novel alleles can be substantially improved. In other words, a comprehensive database is not
189 only crucial for maintaining accuracy but also indispensable for maximizing sensitivity in novel allele
190 detection.
191 For comparative evaluation of gAIRR-wgs, we accessed it against ImmunoTyper2, a recently
192 developed state-of-the-art tool for *TR* gene analysis from WGS data (Ford et al., 2025). Both pipelines
193 were applied to the same 44 HPRC subjects used in our validation study. The computational efficiency
194 analysis revealed substantial differences between the two approaches. gAIRR-wgs completed the
195 analysis of all 44 samples in about 6 hours (06:42:50), while ImmunoTyper2 required approximate 7
196 days (7-19:51:01) to process the same dataset. This represents approximately a 28-fold improvement
197 in computational speed. For known alleles documented in the IMGT database, both tools demonstrated
198 excellent recall performance (**Supplementary Figure 5**). However, gAIRR-wgs achieved superior
199 precision compared to ImmunoTyper2, resulting in fewer false-positive calls and more reliable allele

200 annotations. The combination of high computational efficiency and enhanced precision makes gAIRR-

201 wgs particularly well-suited for biobank-scale analyses and population immunogenomics research,

202 where both processing speed and accuracy are essential for large-scale genomic studies.

203

204 ## 3. gAIRR-wgs reveals the comprehensive *TR* allele landscape of the Taiwan Han
205 Chinese population Using Taiwan Biobank WGS data

206 We applied the gAIRR-wgs pipeline to analyze *TR* alleles from 1,492 individuals in the Taiwan
207 Biobank (TWB), revealing a comprehensive picture of *TR* allele diversity in the Han Chinese
208 population. Our analysis identified numerous novel alleles not present in the IMGT database, all
209 supported by high-quality sequencing reads (MAPQ ≥20, AS ≥150) (**Figure 2**, **Supplementary**
210 **Figure 6**). All novel alleles were systematically renamed using the IgLabel tool to ensure standardized
211 nomenclature (Lees et al., 2023). On average, each individual carried 185.9 (±9.4) TR_V allele types,
212 including 34.6 (±5.5) novel TR_V alleles, 88.5 (±1.9) TR_J allele types with 2.6 (±1.2) novel TR_J
213 alleles, and either 5 (64.1%) or 6 (35.9%) TR_D alleles with very few novel variants observed (**Table**
214 **2**, **Supplementary Table 6**). Notably, 109 novel TR_V alleles and 15 novel TR_J alleles exhibited
215 cohort frequencies exceeding 1% in the population, indicating their substantial prevalence rather than
216 rare variants. Considering each allele's carrier rate in the population, novel alleles accounted for
217 18.39% of the total TR_V repertoire and 2.77% of the total TR_J repertoire among alleles with cohort
218 frequencies greater than 1% in the TWB cohort.

219 Given the substantial number of novel alleles discovered, we sought to validate these findings using
220 independent analytical approaches. We primarily employed IGV visualization to manually inspect the
221 sequencing reads, confirming that perfectly matched reads covered the core sequences of these novel
222 alleles and extended into the flanking regions, providing strong evidence for their presence in the
223 samples. Additionally, we cross-validated our findings by examining whether these novel alleles could

224 also be detected using DRAGEN variant calling. We converted all discovered novel alleles to variant
225 format relative to GRCh38 and compared them with a DRAGEN Iterative gVCF Genotyper variant
226 callset from the same TWB samples (as described in the **METHODs** section). Remarkably, many
227 novel alleles were not detected in the DRAGEN callset (**Supplementary Figure 7**, **Supplementary**
228 **Table 7**), suggesting that conventional variant calling approaches may face challenges in capturing
229 variation within complex genomic regions such as the *TR* loci. While DRAGEN incorporates
230 specialized algorithms for handling nine complex genes including *HLA* and *CYP2D6*, *TR* genes are
231 not included in this targeted optimization, which may contribute to the observed limitations in
232 detecting *TR*-specific variants (Behera et al., 2025). This highlights the necessity for specialized
233 approaches like gAIRR-wgs that are specifically designed to address the unique challenges posed by
234 the highly repetitive and structurally complex nature of *TR* loci.
235 To gain deeper insights into the diversification of *TR* genes in Taiwan Han Chinese population, we
236 categorized all *TR* genes by cohort frequency into six groups: singleton, doubleton to <0.5%, 0.5% to
237 <1%, 1% to <10%, 10% to <50%, and 50% to 100% (**Figure 3**, **Supplementary Figure 8**). For each
238 *TR* gene, we identified its dominant allele within each cohort frequency group and categorized it as
239 either a IMGT-documented allele or a novel allele (with suffix "_novel"). This allowed us to determine
240 the proportion of genes in each frequency group that were dominated by novel versus known alleles.
241 In the highest frequency group (50-100%), TR_V genes were predominantly represented by *01 alleles
242 (the earliest documented IMGT variants), followed by novel alleles, and then *02 alleles. However, in
243 high-frequency TR_V genes (10-50%), 42.4% were dominated by novel alleles, with *01 and *02 each

representing approximately 20%. This pattern indicated substantial database incompleteness, as many genes with greater than 1% cohort frequency were dominated by novel rather than IMGT-documented alleles.

Among common alleles, with cohort frequencies greater than 1%, 34.17% of TR_V genes (109 out of 319) and 14.02% of TR_J genes (15 out of 107) were dominated by novel alleles (**Figure 4**A, **Supplementary Table 6**). The majority of these high-frequency novel alleles had been previously identified in the 47 HPRC subjects from the first year release data, which were analyzed using long-read sequencing with manual validation (Yang et al., 2025), representing diverse global ancestries and suggesting they represent broader human *TR* diversity missing from current databases. However, a notable proportion appeared to be TWB-specific based on initial comparison: 1.83% of common novel TR_V alleles (2 out of 109) and 26.67% of high-frequency novel TR_J alleles (4 out of 15) were not observed in the initial 47 HPRC subjects. To further investigate this apparent population specificity, we extended our analysis to include 232 HPRC subjects from the HPRC second year data release, which were annotated using gAIRR-annotate (**Figure 4**B). This expanded comparison revealed that most of these previously "TWB-specific" common novel alleles were indeed present in East Asian (EAS) and South Asian (SAS) populations within the larger HPRC cohort (**Figure 4**C, **Supplementary Table 9**). This finding suggests that the initial appearance of population-specific alleles was largely due to limited sample size in the first HPRC release rather than true population restriction. Nevertheless, the broader pattern still highlights the historical underrepresentation of East

263 Asian populations in existing genomic resources and emphasizes the importance of diverse population
264 sampling for comprehensive characterization of *TR* diversity.
265 High-frequency novel alleles (≥1%) shared across populations likely reflect systematic gaps in
266 reference databases, while the prevalence of TWB-specific alleles across frequency ranges underscores
267 the underrepresentation of ethnic diversity in current immunogenomic databases. Validation against
268 the most recent release of 232 HPRC subjects further corroborated these findings, confirming both the
269 systematic database gaps and insufficient ethnic diversity representation (**Figure 4**).
270

4. Identification of structural deletion polymorphisms in *TR* gene loci

Analysis of *TR* gene presence across the 1,492 TWB subjects using gAIRR-wgs revealed that four genes showed missing rates >2%: *TRGV5* (2.95% missing), *TRGV4* (5.03% missing), *TRBV3-2* (37.13% missing), and *TRBV4-3* (38.50% missing) (**Supplementary Figure 8**). These substantial missing rates suggested the presence of structural deletion polymorphisms rather than technical artifacts, prompting further investigation into the genomic architecture of these regions.

The affected genes exhibited distinct chromosomal arrangements. *TRGV4* and *TRGV5* are physically adjacent on chromosome 7 (GRCh38), separated by approximately 4.1 kb, while *TRBV3-2* and *TRBV4-3* are located on the alternative contig chr7_KI270803v1_alt, separated by approximately 3.1 kb. To validate these deletion events, we examined alignment data from TWB WGS samples processed using the GATK best practices pipeline with BWA-MEM mapping. Visual inspection revealed complete absence of read support at the genomic positions of these genes in affected individuals. Furthermore, we observed coverage dropouts spanning approximately 9 kb encompassing *TRGV4* and *TRGV5*, and approximately 21 kb encompassing *TRBV3-2* and *TRBV4-3* (**Figure 5**, **Supplementary Figure 12**). Identical patterns were observed in alignment data generated using DRAGMAP, ruling out mapper-specific artifacts and supporting the hypothesis of genuine deletion events.

***TRGV4*/*TRGV5* Deletion Polymorphism**. Our findings corroborate previous reports of genomic deletions in this region. One study documented homozygous genomic deletion encompassing *TRGV4* and *TRGV5* in 1 individual of South Asian ancestry within a cohort of 45 subjects (Corcoran et al., 2023). Using gAIRR-wgs, we identified a homozygous carrier rate of 2.28% for the absence of both

291 genes on two haplotypes in the TWB cohort, suggesting that the event may arise from a ~9 kb deletion
292 encompassing both genes, as inferred from IGV inspection (**Figure 5**). We compared our findings with
293 the gnomAD SVs database (v4.1.0) and observed notable discrepancies. Only six smaller deletions
294 (4~5 kb) were identified in gnomAD, involving *TRGV4* and *TRGV5* separately. The larger ~9 kb
295 deletion encompassing both genes that we observed was not reported. This discrepancy likely reflects
296 the fact that the *TR* region may not fall within gnomAD's high-confidence callable regions. Indeed,
297 inspection of the gnomAD database revealed that approximately half of the samples showed poor
298 sequencing quality in this region, and although several deletion affecting *TRGV4* and *TRGV5* were
299 detected, they may have been filtered out during quality control procedures, potentially leading to
300 underestimation of the true deletion frequency in this region.

301 ***TRBV3-2*/*TRBV4-3* Insertion/Deletion Polymorphism**. The absence patterns of *TRBV3-2* and
302 *TRBV4-3* align with previously described structural variation in the *TRBV* locus. Literature reports and
303 IMGT documentation describe a 21.5 kb insertion/deletion polymorphism encompassing *TRBV3-2*,
304 *TRBV4-3*, and *TRBV6-3* (Omer et al., 2022; Zhao et al., 1994). Although *TRBV6-3* was excluded from
305 our gAIRR-wgs analysis due to its identical sequence to upstream *TRBV6-2*, making them
306 indistinguishable, we were able to characterize this polymorphism using the presence/absence patterns
307 of *TRBV3-2* and *TRBV4-3* as proxy markers. Our analysis revealed that 38.20% of TWB individuals
308 carried the homozygous deletion haplotype (absence of both *TRBV3-2* and *TRBV4-3* on both
309 haplotypes), while 57.91% carried the insertion haplotype (presence of both genes on at least one
310 haplotype). Furthermore, analysis of the 232 assemblies from the HPRC release 2 dataset revealed that

311 34.48% of individuals carried the homozygous deletion type, with the highest proportion observed in
312 AFR (12.93% of 70 individuals), SAS (9.48% of 54 individuals) and EAS (3.45% of 33 individuals)
313 (**Supplementary Table 10**). Due to the limitations of gAIRR-wgs, only the overall carrier rate of the
314 insertion haplotype could be estimated (57.91%); when applying the same perspective to the 232
315 HPRC assemblies, we observed a comparable carrier rate of approximately 50%. Furthermore, we
316 searched gnomAD SVs (v4.1.0) for corresponding structural variants (**Figure 5**). Since the GRCh38
317 primary contig represents the deletion haplotype, we queried for duplications/insertions between
318 *TRBV4-2* (chr7:142345421-142345985) and *TRBV7-2* (chr7:142352819-142353358). One duplication
319 showed overlap with our observed insertion haplotype: a 10,548 bp duplication
320 (DUP_CHR7_C54B5731) with ~54.72% carrier frequency in East Asians. However, the expected ~21
321 kb insertion originates from a historical duplication, resulting in high sequence similarity between
322 *TRBV4-2*, *TRBV6-2*, and their flanking regions. Consequently, short reads derived from this duplicated
323 segment tend to align back to the GRCh38 primary contig, which represents the deletion haplotype
324 lacking *TRBV3-2*, *TRBV4-3*, and *TRBV6-3*. As a result, many reads are preferentially mapped to
325 *TRBV4-2* and *TRBV6-2*, with some alignments lost, leading to the duplication being interpreted as
326 ~10.5 kb rather than ~21 kb. When considering the GRCh38 alternative contig (chr7_KI270803v1_alt),
327 which represents the insertion haplotype, reads can be correctly aligned to the intended loci, thereby
328 revealing the presence of the full ~21 kb structural variant among individuals carrying the insertion
329 type.

330 In parallel, we compared our observations with structural variants called by the Manta SV caller
331 (DRAGEN pipeline) on the same TWB dataset. No SV segments corresponding to the *TRGV* deletion
332 or the *TRBV* insertion/deletion polymorphism were identified (**Supplementary Table 12**,
333 **Supplementary Table 11**). This discrepancy underscores the inherent difficulty of resolving complex
334 immune gene regions using conventional SV callers.
335

336 ## 5. Population-Specific *TR* Allele Distribution Patterns and Diversity Analysis

337 To systematically characterize population-specific patterns of *TR* allele diversity, we analyzed allele
338 distributions across the TWB and 232 HPRC subjects representing six global populations: African
339 (AFR, n=70), North American (NAM, n=16), South American (SAM, n=28), East Asian (EAS, n=33),
340 South Asian (SAS, n=54), and European (EUP, n=31) (**Supplementary Figure 9**, **Supplementary**
341 **Table 8, Supplementary Table 9**).

342 We employed Hellinger distance calculations to quantify differences in allele frequency distributions
343 between populations, with a threshold of >0.5 indicating significant distributional differences (refer to
344 **METHODS**). Overall, 57 *TR* genes (56 TR_V and 1 TR_J) showed significant distributional
345 differences (Hellinger distance >0.5) in at least one population comparison (**Supplementary Figure**
346 **9**, **Supplementary Table 13**). The African population (AFR) showed the most distinctiveness, with
347 50 genes exhibiting significant differences when comparing AFR against each of the other six
348 populations, totaling 151 gene-population pair comparisons with Hellinger distance >0.5. On average,
349 each significantly different gene in AFR differed from 3.02 other populations. The East Asian
350 population (EAS) demonstrated the second-highest level of differentiation with 44 genes showing
351 significant differences across 60 gene-population comparisons, averaging 1.36 population differences
352 per gene. The TWB exhibited 34 genes with significant differences, generating 78 gene-population
353 pair comparisons where TWB differed from other populations, averaging 2.29 population differences
354 per gene. South American (SAM) and South Asian (SAS) populations showed the least differentiation,
355 with 28 genes (36 comparisons) and 30 genes (33 comparisons) respectively.

356 Among the TWB cohort, 34 *TR* genes (33 TR_V and 1 TR_J) exhibited significantly different allele
357 distributions compared to other global populations (Hellinger distance >0.5). Seven TR_V genes
358 showed particularly pronounced differences: *TRBV20-1*, *TRAV6*, *TRAV35*, *TRBV7-5*, *TRAV12*-2,
359 *TRAV8-4*, and *TRBV5-5*. When comparing TWB specifically with the EAS population, most *TR* genes
360 showed similar allele distributions, reflecting shared ancestry. However, eight TR_V genes maintained
361 significant distributional differences even within this regional comparison, including the seven genes
362 mentioned above plus *TRAV38-1*, suggesting population-specific evolutionary pressures or founder
363 effects within East Asian subpopulations.
364 Moreover, we calculated Shannon's diversity index for each gene locus across all cohorts to quantify
365 allelic richness within populations (see **METHODS**). Using the first quartile (Q1) value of 0.4793
366 across all results as a threshold, we defined high diversity genes, identifying 104 *TR* genes spanning
367 the seven cohorts that exceeded this threshold (**Supplementary Figure 11**, **Supplementary Table**
368 **14**). AFR populations exhibited the highest overall diversity, with 84 genes meeting the high diversity
369 criterion, consistent with established patterns of greater genetic diversity in African populations. In
370 contrast, TWB showed the lowest diversity with only 47 high-diversity genes.
371 Notably, 25 *TR* genes (24 TR_V and 1 TR_J) maintained consistently high diversity across all seven
372 cohorts, representing core genes with universally preserved allelic richness. Conversely, six TR_V
373 genes (*TRBV30*, *TRAV27*, *TRBV20/OR9-2*, *TRBV6-6*, *TRGV5P*, *TRBV5-5*, *TRBV7*, *TRBV25/OR9-2*,
374 *TRAV26-1*, *TRBV7-7*) showed high diversity in other populations but reduced diversity in TWB

375 (Shannon's index <0.4793), potentially indicating population-specific bottlenecks or selective
376 pressures affecting these loci.
377 Notably, consistent with our earlier findings regarding limited allelic variation in TR_D segments,
378 diversity analysis revealed that most TR_D genes were dominated by single prevalent alleles across
379 populations (**Supplementary Figure 8**). However, *TRBD2* emerged as a notable exception, exhibiting
380 the highest Shannon's diversity index among all TR_D genes across different populations. This pattern
381 suggests that while most TR_D segments are highly conserved, *TRBD2* maintains functional diversity
382 that may be important for T cell receptors functionality.

383 **DISCUSSION**

384 This study addresses a critical gap in immunogenomics by developing the specialized pipeline for
385 accurate *TR* allele calling from standard WGS data. The challenges inherent in applying existing *TR*
386 genotyping tools to WGS data, including reduced sequencing depth, shorter read lengths, and limited
387 paired-end information utilization, have prevented large-scale population studies of g*TR* diversity. Our
388 gAIRR-wgs pipeline overcomes these fundamental limitations through three key innovations: targeted
389 read extraction that reduces computational burden, enhanced reference databases incorporating
390 population-specific diversity, and optimized algorithms that prevent hybridized artifacts while
391 maintaining high sensitivity.

392 The improvement in computational efficiency, from 6-7 hours and 500GB memory to 12-15 minutes
393 and 5GB memory per sample, represents an improvement that makes population-scale *TR* genotyping
394 feasible for biobank-sized cohorts. This efficiency gain is crucial for translating immunogenomic
395 discoveries to clinical applications, where rapid and cost-effective analysis is essential. The superior
396 performance of gAIRR-wgs compared to gAIRR-call demonstrates that WGS-specific optimizations
397 are not merely adaptations but fundamental improvements in algorithmic design.

398 To demonstrate the practical utility of gAIRR-wgs and validate its performance in real-world
399 applications, we applied our pipeline to a large-scale population cohort from the Taiwan Biobank. This
400 analysis of 1,492 individuals not only serves as a proof-of-concept for the scalability and robustness
401 of our approach but also provides an unprecedented opportunity to systematically characterize *TR*
402 allelic diversity in a well-defined Han population in East Asia. The choice of TWB as our

403 demonstration cohort is particularly valuable given the historical underrepresentation of East Asian
404 populations in immunogenomic studies, allowing us to assess both the technical performance of
405 gAIRR-wgs and the completeness of existing reference databases in a previously understudied
406 population context.

407 Our comprehensive analysis of 1,492 TWB individuals reveals profound gaps in current
408 immunogenetic reference databases, particularly for East Asian populations. The finding that 42.4%
409 of high-frequency TR_V genes are dominated by novel alleles rather than IMGT-documented alleles
410 represents a systematic bias in global immunogenetic resources. This disparity is particularly striking
411 given that many of these novel alleles were subsequently found in HPRC subjects of diverse ancestries,
412 and further validation using the expanded HPRC release 2 dataset confirmed that most high-frequency
413 novel alleles initially appearing TWB-specific were indeed present in Asian populations. This
414 progression from apparent population specificity to broader population presence underscores how
415 limited sampling in reference databases can create imperfect impressions of population uniqueness.

416 Inaccurate or incomplete g*TR* profiling directly impacts our understanding of immune repertoire biases
417 in disease contexts, potentially leading to misinterpretation of adaptive immune responses in non-
418 European populations. The identification of population-specific high-frequency alleles further
419 emphasizes the critical need for population-specific immunogenomic studies. This is precisely where
420 gAIRR-wgs demonstrates its unique value by enabling systematic characterization of *TR* diversity
421 from existing biobank WGS short read data, our approach can rapidly expand reference databases and
422 correct population-specific biases.

423 Through our gAIRR-wgs post-analysis, we identified structural polymorphisms affecting contiguous
424 *TR* genes, representing a previously underappreciated source of immune diversity. The *TRBV3-*
425 *2*/*TRBV4-3* insertion/deletion polymorphism, present in Taiwan Han Chinese individuals,
426 demonstrates that structural variants can dramatically alter the available germline *TR* repertoire at the
427 population level. Similarly, the *TRGV4*/*TRGV5* deletion polymorphism represents another example of
428 how structural variants contribute to immune genetic diversity. These findings have profound
429 implications for understanding population-specific immune capabilities and disease susceptibilities.
430 Notably, the failure of structural variant callers with standard/default pipeline to detect these
431 immunologically relevant polymorphisms, despite their high frequency and substantial genomic span,
432 underscores the necessity of specialized approaches for complex immune gene regions. These
433 structural variations are incomplete in the gnomAD database, highlighting systematic gaps in structural
434 variant documentation for both Asian populations and immune-relevant genomic regions. The gene-
435 centric approach employed by gAIRR-wgs proves essential for identifying functionally relevant
436 structural variants that conventional genome-wide methods miss.
437
438 While gAIRR-wgs represents a significant advance in *TR* genotyping from WGS data, several
439 technical and methodological limitations warrant consideration and provide directions for future
440 development. The reliance on short-read sequencing technology (PE 150 bp) limits resolution of the
441 most complex structural variants and haplotype phasing. Hybridized-structured alleles, where different
442 segments share similarity with distinct reference alleles, pose particular challenges that could be better

443 resolved with longer reads (≥300 bp) or long-read approaches. Through our benchmark studies, we
444 have identified specific alleles that frequently generate false positives in 30X WGS PE 150 bp data,
445 and we provide a list of these problematic alleles to assist users in recognizing and mitigating potential
446 sources of error.

447 Additionally, gAIRR-wgs currently does not incorporate relative genomic positioning information,
448 making it unable to distinguish alleles with identical sequences but different chromosomal locations
449 (e.g., TRBV6-2*01/TRBV6-3*01, TRBV24-1*02/TRBV24/OR9-2*01, and TRGJ1*02/TRGJ2*01).
450 While evolutionary considerations may justify distinct nomenclature based on gene duplication history,
451 the assignment of completely different gene names to functionally identical sequences can create
452 confusion in downstream application and clinical interpretation.

453 The current implementation focuses on carrier status (presence/absence of genes) rather than providing
454 quantitative assessment of allelic dosage. For population-level studies, carrier status and allelic dosage
455 represent different analytical perspectives: carrier analysis effectively captures population-level allele
456 frequencies and diversity patterns, while dosage assessment provides individual-level genotypic
457 resolution. Future versions should incorporate copy number estimation to provide more comprehensive
458 genotyping information, particularly important for understanding gene expression levels and dosage
459 effects in clinical contexts.

460 Furthermore, gAIRR-wgs currently restricts novel allele calling to core coding regions, following
461 IMGT nomenclature rules. However, we recognize that future implementations should consider non-
462 coding regions, including flanking sequences, for comprehensive allelic characterization, similar to

463 the nomenclature systems used for other complex immune genes such as *CYP*, *KIR*, and *HLA*. While
464 gAIRR-wgs includes algorithms to handle nested structure alleles, some evidence suggests that certain
465 IMGT reference sequences may contain errors. Future work should incorporate systematic validation
466 of reference sequences and provide frameworks for updating problematic entries.
467
468 **Conclusion**. The gAIRR-wgs pipeline addresses fundamental challenges in population
469 immunogenetics by enabling accurate *TR* allele calling from standard WGS data. Our analysis reveals
470 extensive previously uncharacterized *TR* diversity in Asian populations, identifies structural
471 polymorphisms, and provides frameworks for optimal database expansion strategies. These advances
472 lay the foundation for population-scale immunogenomic studies and precision approaches to immune-
473 related diseases. The systematic characterization of human *TR* diversity across global populations will
474 be essential for understanding immune system evolution, disease susceptibility, and therapeutic
475 response in our increasingly interconnected world. Most importantly, this work demonstrates that
476 advanced computational approaches can unlock the immunogenomic potential of existing biobank
477 resources, accelerating translation of population genetic insights into clinical applications. As precision
478 medicine increasingly recognizes the central role of immune system diversity in health and disease,
479 tools like gAIRR-wgs will be essential for realizing the promise of personalized immunotherapy and
480 population-specific medical interventions.

481 **METHODS**

482 1. gAIRR-wgs pipeline development

483 We developed gAIRR-wgs specifically optimized for *TR* genotyping from WGS data. The original
484 gAIRR-Suite, includes a module, gAIRR-call (referred to as gAIRR-panel in this manuscript), was
485 designed for targeted analysis of *IG* and *TR* gene analysis using high-depth panel sequencing data (Lin
486 et al., 2022).

487 **Read Extraction**. WGS reads were first aligned to the GRCh38 reference genome (hs38DH) using
488 BWA-MEM (v.0.7.17) (Li, 2013). Target reads covering *TR* gene loci were extracted, including both
489 mapped reads from target regions and unmapped reads to account for potential reference genome
490 discrepancies.

491 **Reference Database Construction**. The *TR* allele reference database was constructed based on the
492 IMGT database, with *TR* gene sequences last updated in 2020 (Manso et al., 2021). The database was
493 then expanded by incorporating 335 novel *TR* alleles previously identified and validated from 47
494 HPRC personal genome assemblies (Yang et al., 2025).

495 **Allele Calling Algorithm**. The gAIRR-wgs pipeline follows a two-step alignment process. Initial
496 BWA-MEM alignment identifies candidate novel alleles, which are then incorporated into the
497 reference database for a second alignment round to prevent hybridized artifact formation.

498 **Depth Calculation Algorithm**. Allele depth was calculated using the minimum coverage principle
499 across core gene regions, with weighted incorporation of both single-end and paired-end reads. The

500 adaptive threshold from gAIRR-panel, which was designed for high-depth panel data, was removed in
501 gAIRR-wgs as it proved suboptimal for WGS data.
502 Additional details of the computational implementation and parameter optimization are provided in
503 **Supplementary Note 1**.
504

505 2. Benchmark study

506 Benchmark validation used WGS data from 44 HPRC subjects (paired-end 150 bp reads) with
507 validated *TR* allele annotations derived from personal genome assemblies and long read sequencing
508 data (Yang et al., 2025). Performance was evaluated using recall (sensitivity) and precision (positive
509 predictive value) metrics:
510 Recall = True Positives / (True Positives + False Negatives)
511 Precision = True Positives / (True Positives + False Positives)
512 Performance metrics are presented as the mean ± standard deviation across all subjects.
513 Novel alleles were compared based on core sequence identity, with specific exclusions for allele pairs
514 indistinguishable by short-read sequencing (see **Supplementary Note 2** for details).
515

516 3. Population study

517 **Study Cohort and Data Processing**. The population analysis included 1,492 individuals from the
518 Taiwan Biobank with Illumina-based WGS data (paired-end 150 bp reads, sequenced using HiSeq

519 2500 (n=555), HiSeq 4000 (n=634) and NovaSeq 6000 (n=303)) derived from peripheral blood
520 mononuclear cell (PBMC) DNA samples (Hsu et al., 2024). Raw data were provided in CRAM format
521 and processed using our read extraction pipeline to obtain TR-specific reads prior to gAIRR-wgs
522 analysis.

523 **Variant Validation and Database Comparisons**. For variant-based validation, all novel alleles were
524 aligned to the GRCh38 reference genome (hs38DH) using BWA-MEM with secondary alignments
525 excluded. Novel allele sequences were converted to variant format relative to GRCh38, processing
526 various CIGAR operations including matches (M), insertions (I), deletions (D), substitutions, and
527 clipping events (S, H) were processed alongside MD tags to accurately identify sequence differences
528 and generate REF to ALT variant calls for comparison with the DRAGEN Iterative gVCF Genotyper
529 variant callset. Variants were validated based on concordance in genomic positions, allele sequences,
530 and genotype calls with adequate coverage depth (detailed validation criteria in **Supplementary Note**
531 **3**). For reference blocks (regions marked with <NON_REF> in ALT field representing sequences
532 identical to the reference genome), validation was performed by confirming the absence of variant
533 calls within the corresponding genomic intervals.

534 Structural variant analysis was conducted using results generated by the DRAGEN pipeline (version
535 4.0.3) on the TWB WGS dataset. These results are derived on unpublished data, and the related
536 manuscript is currently in preparation. For regions with potential structural variants identified through
537 gene absence patterns, corresponding genomic intervals were systematically queried in the DRAGEN
538 SV callset to identify overlapping or similar-length structural variants. SV calls were considered

concordant when: (1) genomic coordinates showed substantial overlap with predicted deletion/insertion regions, and (2) variant sizes were within reasonable similarity ranges of the expected structural changes based on gene absence frequencies and genomic context.

Shannon's Diversity Index Calculation. We calculated Shannon's diversity index for each *TR* gene locus across all populations to quantify allelic richness and evenness within populations using the formula:

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

where p_i represents the proportion of individuals carrying allele i, and S is the total number of alleles for a given gene. The Shannon's diversity index ranges from $0 \leq H' \leq \ln(S)$, where $H' = 0$ indicates a population dominated by a single allele (no diversity), and higher values indicate greater allelic diversity. We used the first quartile (Q1) value of 0.4793 across all gene-population combinations as a threshold to define high-diversity genes.

Hellinger Distance Calculation. To quantify differences in allele frequency distributions between populations for each *TR* gene, we calculated Hellinger distance using the formula:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

where P and Q represent allele frequency distributions for a given gene in two different populations, p_i and q_i are the frequencies of allele i in populations P and Q respectively, and k is the total number

556 of unique alleles observed across both populations. The Hellinger distance ranges from 0 to 1, where
557 0 indicates identical allele distributions between populations and 1 indicates completely different
558 distributions. We used a threshold of $H > 0.5$ to identify genes with significantly different allele
559 distributions between population pairs.

560

561 ## 4. Statistical Analysis and Visualization

562 The core tool, gAIRR-wgs was primarily developed in Python (version 3.7.13). All the downstream
563 analyses, including benchmark study and population study, were performed using Python (version
564 3.10.14) and R (version 4.4.0).
565 Read alignment visualization and validation were conducted using the Integrative Genomics Viewer
566 (version 2.10.3; IGV) (Robinson et al., 2011). To ensure reproducibility and maintain environment
567 consistency, all computational tools were managed using conda (version 24.11.3), with separate virtual
568 environments created for each major component of the analysis pipeline.

569

570 **DATA ACCESS**

571 Short-read sequencing data for the HPRC individuals (HG002, HG00438, HG005, HG00621,
572 HG00673, HG00733, HG00735, HG00741, HG01071, HG01106, HG01109, HG01175, HG01243,
573 HG01258, HG01358, HG01361, HG01891, HG01928, HG01952, HG01978, HG02055, HG02080,
574 HG02109, HG02145, HG02148, HG02257, HG02572, HG02622, HG02630, HG02717, HG02723,
575 HG02818, HG02886, HG03098, HG03453, HG03486, HG03492, HG03516, HG03540, HG03579,
576 NA18906, NA19240, NA20129, and NA21309) were obtained from the Human Pangenome Reference
577 Consortium (HPRC) project. The corresponding SRA or ENA accession numbers are as follows:
578 HG002 (SRR14724532), HG00438 (ERR3988768), HG005 (SRR14724528), HG00621
579 (ERR3988803), HG00673 (ERR3988814), HG00733 (ERR3988823), HG00735 (ERR3988824),
580 HG00741 (ERR3988826), HG01071 (ERR3988832), HG01106 (ERR3988841), HG01109
581 (ERR3988842), HG01175 (ERR3988851), HG01243 (ERR3988858), HG01258 (ERR3988862),
582 HG01358 (ERR3988875), HG01361 (ERR3988876), HG01891 (ERR3988943), HG01928
583 (ERR3988951), HG01952 (ERR3988958), HG01978 (ERR3988965), HG02055 (ERR3988979),
584 HG02080 (ERR3988986), HG02109 (SRR11124438), HG02145 (ERR3988995), HG02148
585 (ERR3988996), HG02257 (ERR3989003), HG02572 (ERR3989028), HG02622 (ERR3989038),
586 HG02630 (ERR3989040), HG02717 (ERR3989059), HG02723 (ERR3989060), HG02818
587 (ERR3989080), HG02886 (ERR3989091), HG03098 (ERR3989119), HG03453 (ERR3989166),
588 HG03486 (ERR3989170), HG03492 (ERR3989173), HG03516 (ERR3989174), HG03540

589 (ERR3989177), HG03579 (ERR3989180), NA18906 (ERR3989364), NA19240 (ERR3989410),

590 NA20129 (ERR3989456), and NA21309 (SRR10392718).

591 Annotation results and the discovery of 335 novel alleles are provided in the supplementary materials

592 available at https://www.biorxiv.org/content/10.1101/2025.05.24.655452v1.

593 Short-read and alignment data for the Taiwan Biobank (TWB) cohort were accessed under ethical

594 approval TWBR11106-05. The GRCh38 BED reference used for targeted *TR* allele analysis is publicly

595 available at https://github.com/maojanlin/gAIRRsuite.

596

597

598

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank all the participants from the Taiwan Biobank and all investigators who contributed to this study. We are also grateful to the National Center for High-performance Computing (NCHC) of the National Institutes of Applied Research (NIAR) of Taiwan for providing computational and storage resources. This study was supported by the research grants (NSTC 111-2320-B-002-091-MY3, 112-2622-B-002-012) and industry-academia cooperative research program grants (NSTC 112-2622-B-002-012) from the National Science and Technology Council in Taiwan. This project is also supported by the Excellence in Key Advantages Program (NTU Core Consortiums and Competitiveness Programs) from the National Taiwan University (113L892901).

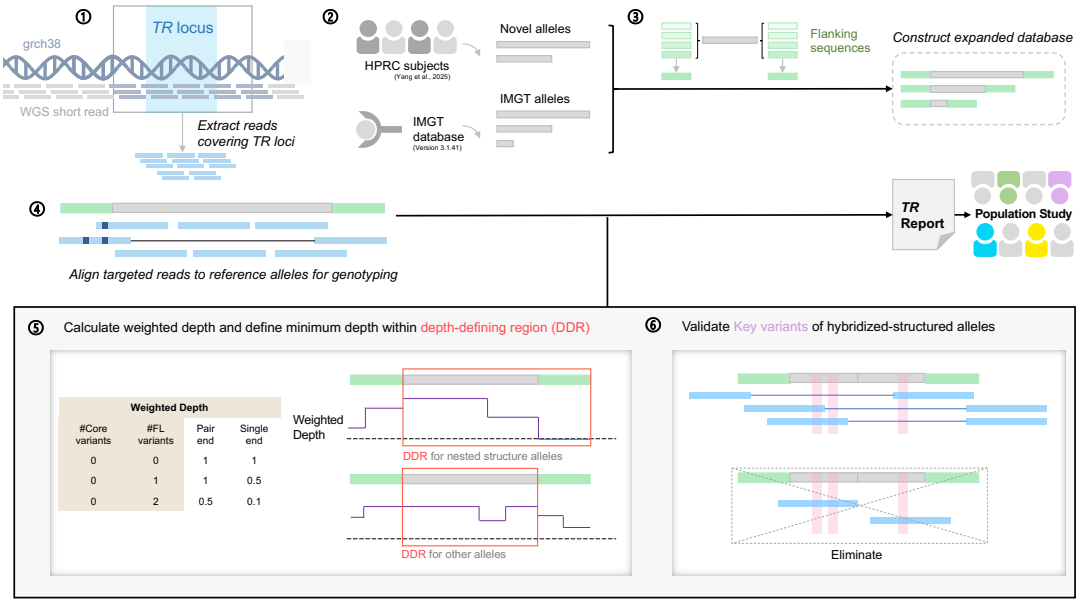


Figure 1 | Overview of the gAIRR-wgs pipeline for *TR* allele calling from WGS data.
TR-related reads were extracted from WGS data (top left; ①), and the allele reference database was expanded by incorporating novel alleles from 47 HPRC subjects (top; ②) and appending consensus-based flanking sequences (top right; ③). Reads were then aligned to the expanded database to identify both candidate novel alleles and IMGT-documented alleles (top; ④). Allele support was quantified using a weighted depth model that integrates both single-end and paired-end reads, with depths calculated either within customized depth-defining regions (DDR) for each allele (bottom left; ⑤). Hybridized-structured alleles were further validated by requiring proper paired reads to support all key distinguishing variants (bottom right; ⑥).

611

612

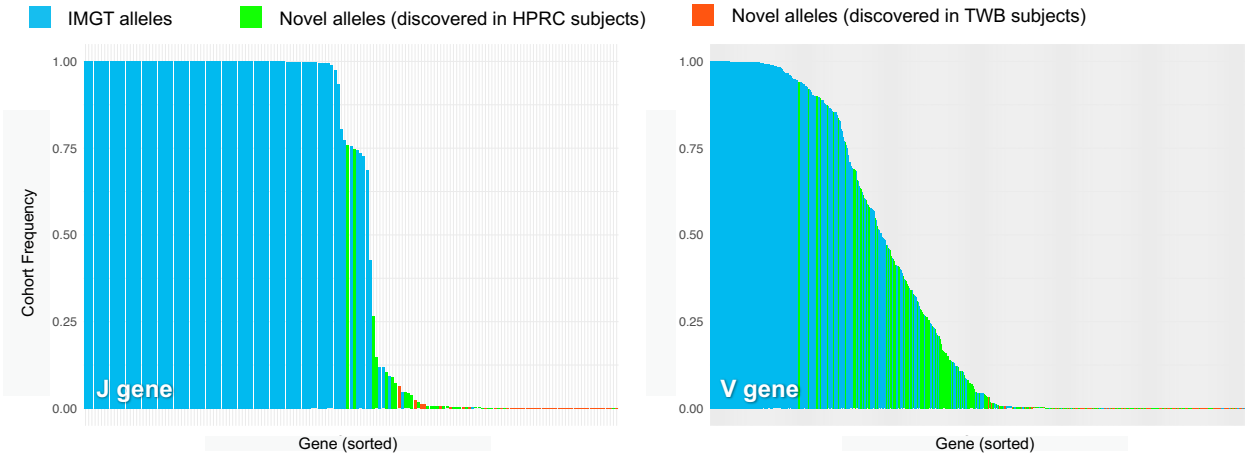
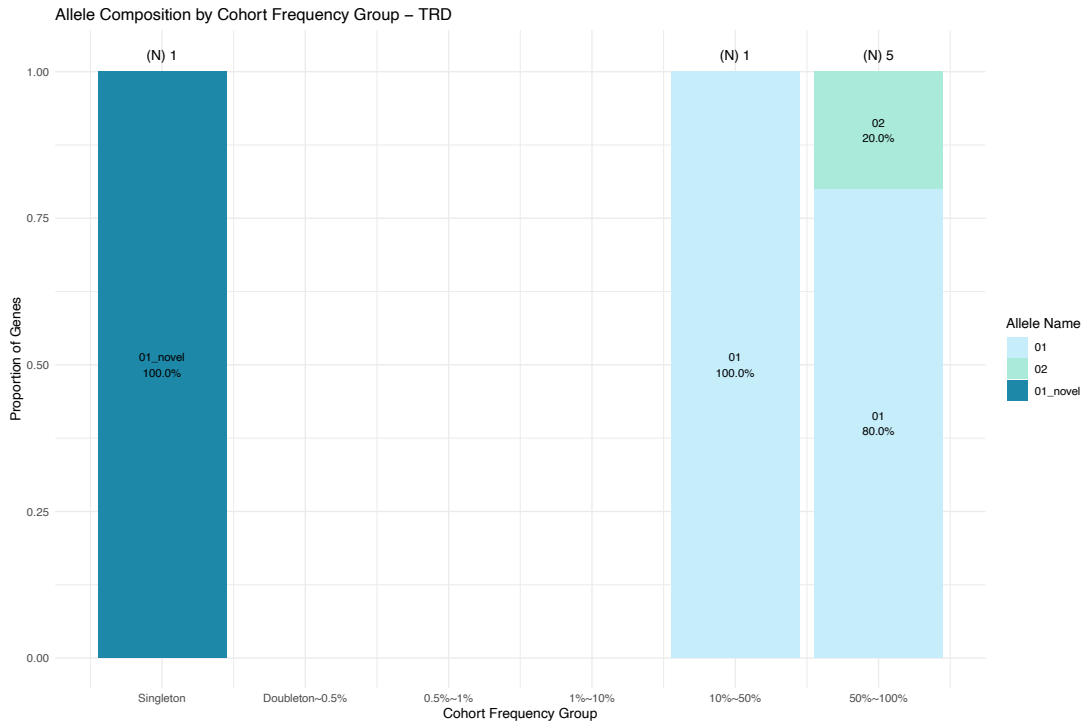


Figure 2 | Cohort frequency distribution of TR_J and TR_V alleles in the Taiwan Biobank. Cohort frequency distributions of TR_J (left) and TR_V (right) alleles identified in Taiwan Biobank (TWB) subjects. The Y-axis represents the cohort frequency, defined as the proportion of individuals carrying each specific allele. The X-axis lists individual alleles, sorted in descending order by cohort frequency. Bars are color-coded based on allele type and discovery origin: Blue bars denote alleles already documented in the IMGT database. Green bars represent novel alleles previously discovered in HPRC subjects. Red bars indicate novel alleles discovered only in the TWB and not observed in HPRC.

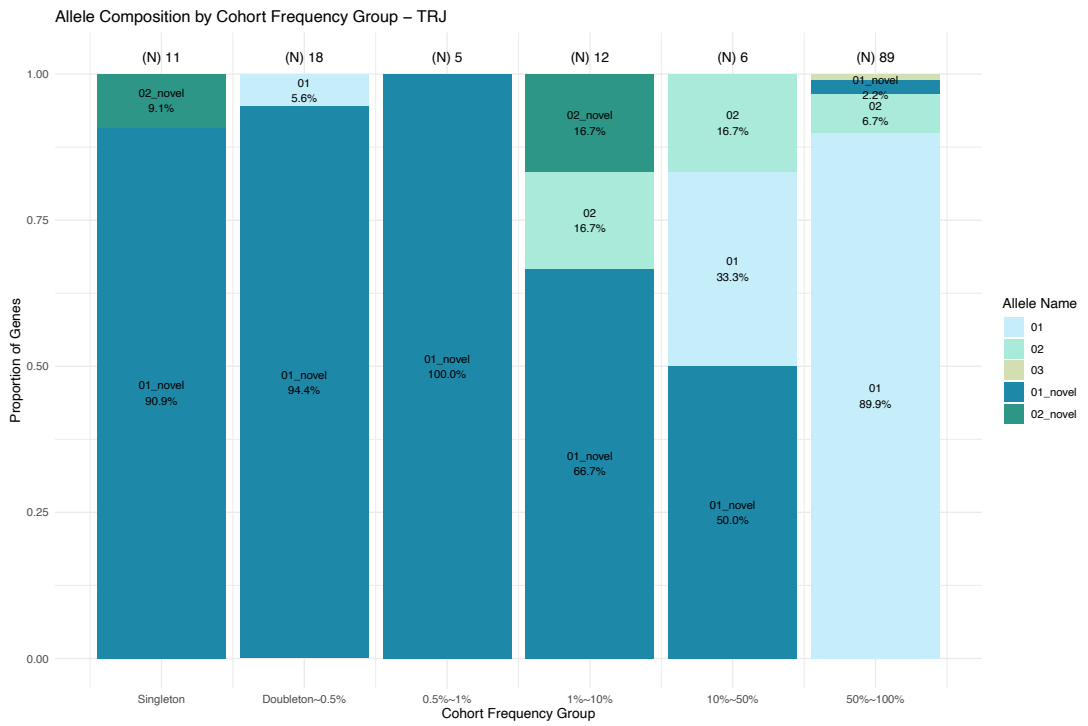
613

614

A.


Allele Composition by Cohort Frequency Group – TRD
(N) 1
(N) 1
(N) 5
1.00
0.75
0.50
0.25
0.00
Proportion of Genes
01_novel
100.0%
01
100.0%
02
20.0%
01
80.0%
Singleton
Doubleton~0.5%
0.5%~1%
1%~10%
10%~50%
50%~100%
Cohort Frequency Group
Allele Name
01
02
01_novel

B.


Allele Composition by Cohort Frequency Group – TRJ
(N) 11
(N) 18
(N) 5
(N) 12
(N) 6
(N) 89
1.00
0.75
0.50
0.25
0.00
Proportion of Genes
02_novel
9.1%
01_novel
90.9%
01
5.6%
01_novel
94.4%
01_novel
100.0%
02_novel
16.7%
02
16.7%
01_novel
66.7%
02
16.7%
01
33.3%
01_novel
50.0%
01_novel
2.2%
02
6.7%
01
89.9%
Singleton
Doubleton~0.5%
0.5%~1%
1%~10%
10%~50%
50%~100%
Cohort Frequency Group
Allele Name
01
02
03
01_novel
02_novel

C.

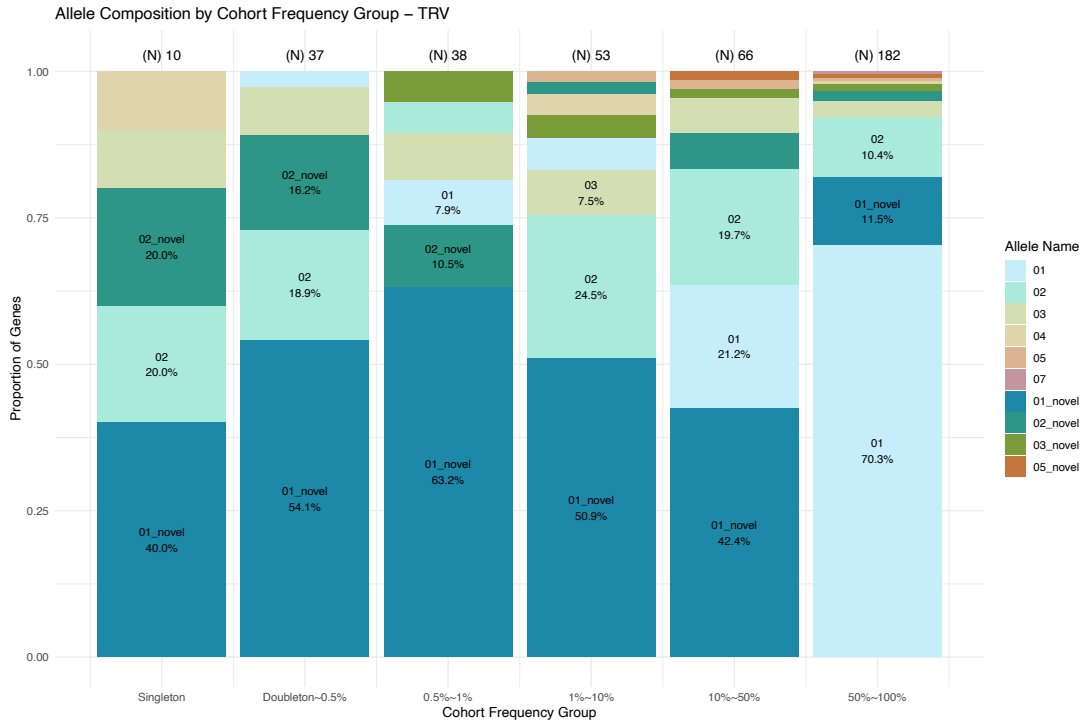


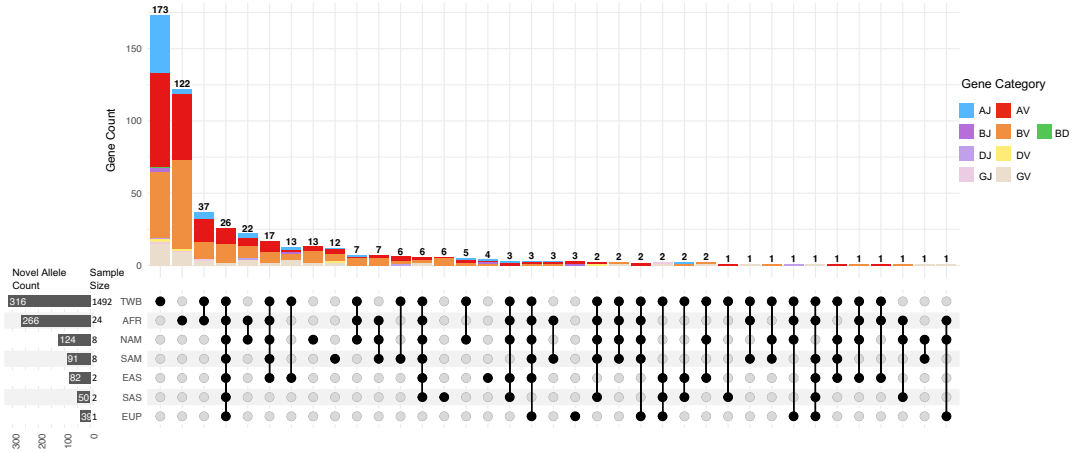
Figure 3 | Distribution of Star Alleles Across *TR* Gene Segments Grouped by Cohort Frequency.

(A), (B), and (C) show TR_D, TR_J, and TR_V genes, respectively. Proportional distributions of dominant star alleles across *TR* gene segments in the TWB cohort, grouped by cohort frequency. Each gene was assigned to one of six frequency intervals as shown along the X-axis. The Y-axis represented the proportion of genes within each frequency category that were dominated by a given star allele type. Each bar was colored by star allele type, with allele assignments based on either direct IMGT annotation or novel allele grouping (e.g., *01_novel was used to represent novel alleles most similar to the canonical *01). These plots illustrated how allelic composition shifted across frequency categories, revealing evolutionary patterns of conservation and diversification in different *TR* segments. While high-frequency groups were dominated by conserved alleles such as *01, lower-frequency groups increasingly featured novel alleles, particularly *01_novel, reflecting ongoing diversification and potential adaptive emergence within the population.

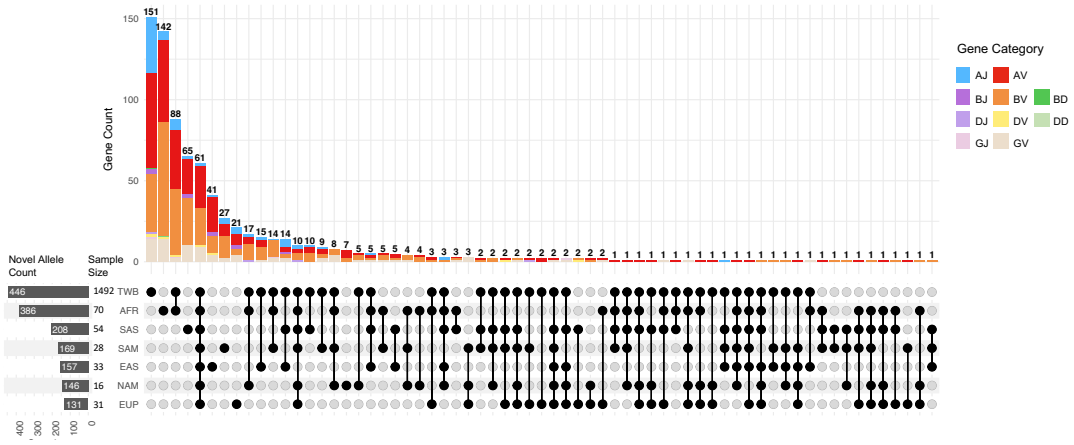
615

616

A.



B.



C.

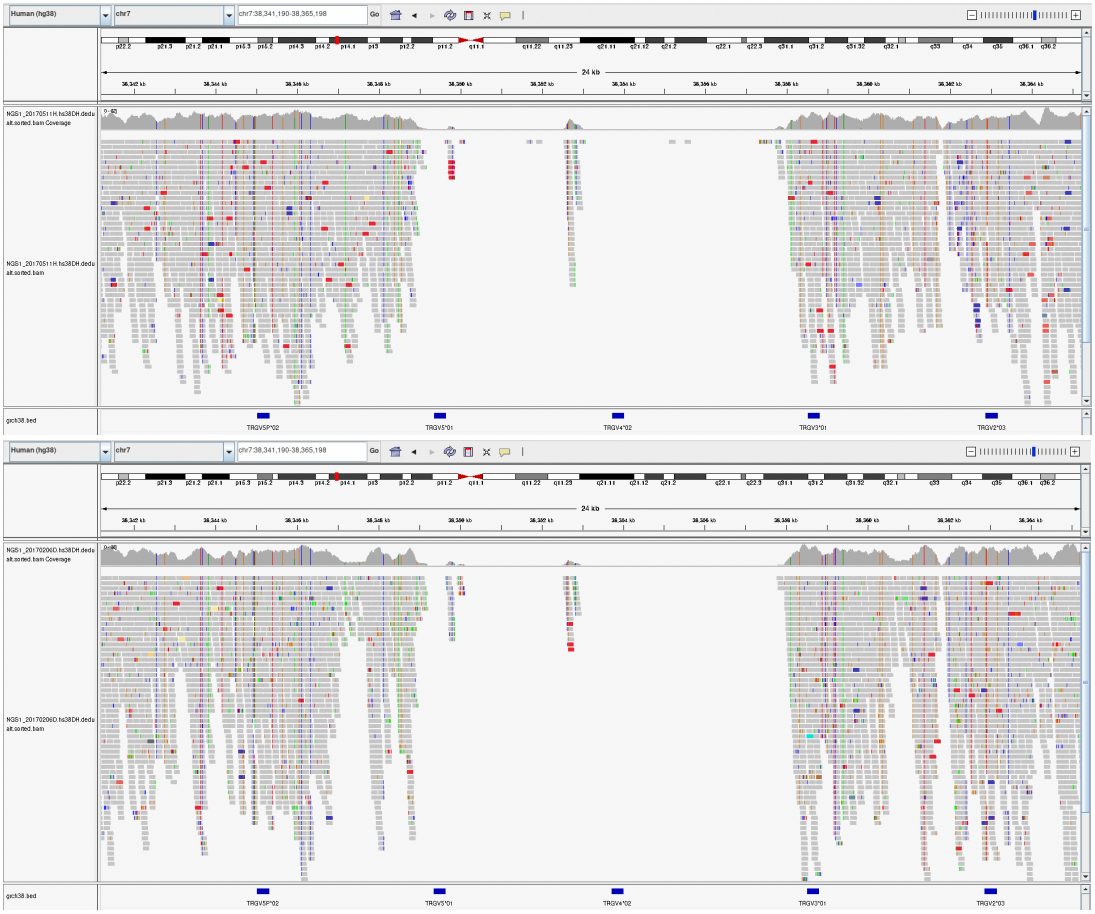
Gene	TWB	AFR	EAS	EUP	NAM	SAM	SAS
TRAJ26*01_6JUE	6.30%		6.06%				14.81%
TRBV10-2*01_5Y3B	3.28%		24.24%				5.56%
TRAJ24*02_ACXJ	2.48%		6.06%				14.81%
TRAV4*01_EYTX	1.61%						1.85%
TRAJ31*01_ESNZ	1.21%		9.09%				1.85%
TRAJ32*02_H2ZT	1.21%		9.09%				
TRAV22*01_4R4J	0.80%						

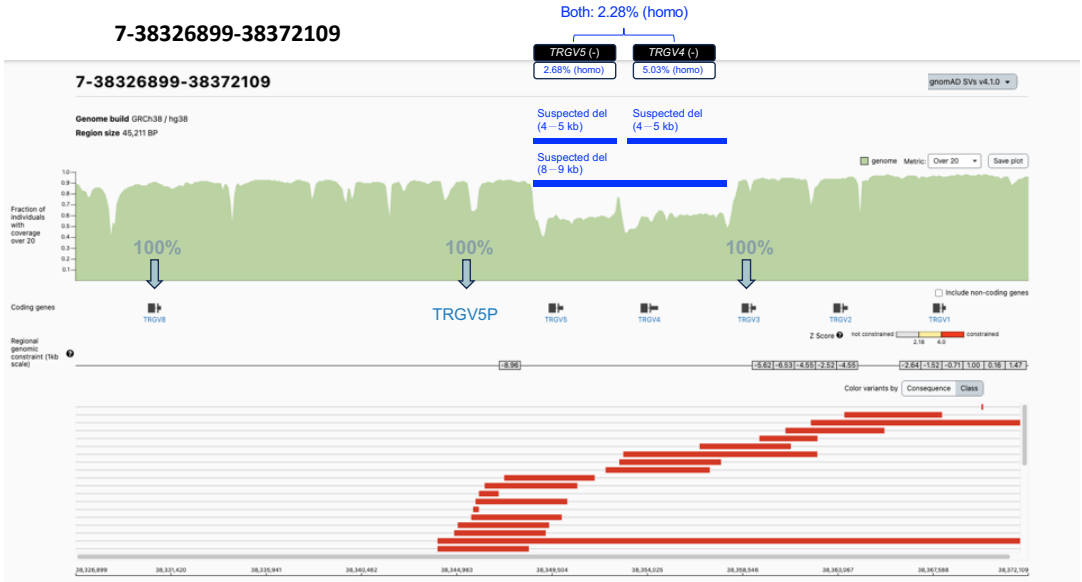
Figure 4 | Distribution of novel *TR* alleles across populations.

Novel alleles were identified from the Taiwan Biobank (TWB) cohort using the gAIRR-wgs genotyping pipeline. Each bar in the upper panel represents the number of genes carrying novel alleles observed across specific population combinations, with colors indicating gene categories (e.g., AV, AJ, BV, etc.). The lower panel is an UpSet plot summarizing population sharing patterns for these novel alleles across cohorts including TWB, AFR (African), NAM (North American), SAM (South American), EAS (East Asian), SAS (South Asian), and EUP (European). The left margin indicates the total number of novel alleles and sample size per cohort.

(A) Novel alleles in TWB were compared against 47 HPRC subjects based on high-confidence annotations from the first HPRC assembly release, using the gAIRR-annotate pipeline with manual validation by long-read alignments.
(B) Novel alleles in TWB were compared against 232 HPRC subjects from the second HPRC assembly release, annotated by gAIRR-annotate without long-read-based manual validation.
(C) Common novel alleles (>1% frequency) initially appearing to be TWB-specific in the first HPRC release were identified in East Asian (EAS) and South Asian (SAS) populations within the larger set of 232 HPRC subjects from the second release.

A.





(sorted by POS)

B.

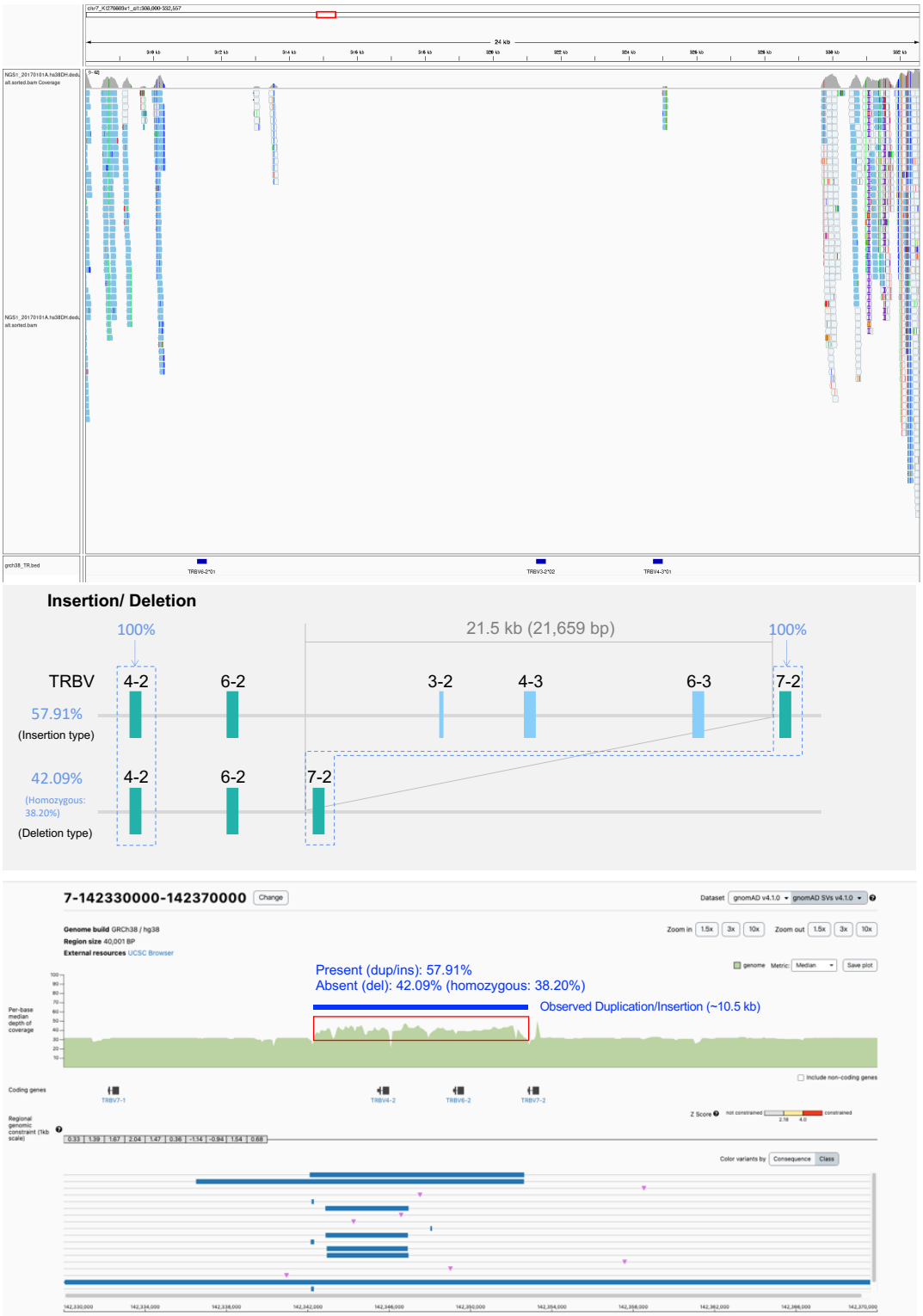


Figure 5 | IGV visualization of structural variation at *TRG* and *TRB* loci in representative TWB WGS samples.

(A) T*RGV4*/*TRGV5* deletion polymorphism.

Read alignments from GATK-processed WGS data mapped to GRCh38 are shown. Affected individuals exhibit complete absence of read coverage across the *TRGV4* and *TRGV5* region, with sharp dropouts in depth, indicating a homozygous deletion. Genes *TRGV4* and *TRGV5* are physically adjacent on chr7, separated by ~4.1 kb. Below, comparison with SV databases highlights

the expected deletion interval (blue box), with TWB carrier frequencies indicated (Present: 97.7%, Absent: 2.28%).

(B) *TRBV3-2*/*TRBV4-3* insertion/deletion polymorphism.

In a representative deletion haplotype, no read coverage is observed at the positions of *TRBV3-2* and *TRBV4-3* (~21 kb span), while neighboring genes *TRBV4-2* and *TRBV7-2* retain full coverage, consistent with their stable presence.

Read alignment visualizations were generated using IGV (Integrative Genomics Viewer). SV database comparison is shown below, with the predicted deletion span marked in blue and TWB carrier frequencies annotated.

621 **Table 1 | Performance comparison of gAIRR-wgs and gAIRR-panel pipelines across different**
622 **sequencing platforms in 44 HPRC subjects**.
623

Gene	Data type	Tool	Recall		Precision		F1 score	
			AV	SD	AV	SD	AV	SD
TR_D	WGS	panel	72.5	17.6	100.0	0.0	82.9	12.0
	WGS	**wgs**	100.0	0.0	100.0	0.0	100.0	0.0
	Panel 150	panel	81.1	1.6	100.0	0.0	89.5	1.0
	Panel 300	panel	81.1	1.6	100.0	0.0	89.5	1.0
TR_J	WGS	panel	96.2	2.2	99.3	0.8	97.7	1.3
	WGS	**wgs**	99.9	0.3	99.6	0.6	99.8	0.4
	Panel 150	panel	99.8	0.7	100.0	0.2	99.9	0.4
	Panel 300	panel	99.8	0.6	99.9	0.2	99.9	0.4
TR_V	WGS	panel	75.4	4.2	91.4	1.9	82.6	2.9
	WGS	**wgs**	99.2	0.9	97.4	0.8	98.3	0.6
	Panel 150	panel	97.7	0.9	96.1	1.3	96.9	0.9
	Panel 300	panel	98.0	0.8	97.8	1.1	97.9	0.8

Abbreviation: AV = average value; SD = standard deviation, both calculated across 44 HPRC subjects and reported as percentages (%). Panel 150 and Panel 300 refer to targeted sequencing data with paired-end 150 bp and 300 bp reads, respectively; both were analyzed using gAIRR-panel (a pipeline for targeted sequencing). WGS (panel) indicates WGS data (paired-end 150 bp) analyzed using gAIRR-panel, while WGS (wgs) indicates the same WGS data analyzed using the gAIRR-wgs pipeline.

624

625

626 **Table 2 | Summary of TR_V and TR_J alleles identified in the TWB cohort**.

627

		Total (IMGT + novel)	Novel		
			Proportion	Novel ^{TWB}	Novel ^{HPRC}
TR_J	**Allele type**	167	44.31% (74)	26.95% (45)	17.37% (29)
	Weighted	131995	2.90% (3823)	0.18% (239)	2.72% (3584)
TR_V	**Allele type**	608	61.68% (375)	20.89% (127)	40.79% (248)
	Weighted	277306	18.62% (51642)	0.11% (307)	18.51% (51335)

"Allele type" represents one-dimensional analysis based solely on allele type. "Weighted" analysis incorporates both allele type and cumulative carrier count across TWB subjects. Values in parentheses indicate the absolute number of alleles or total carriers contributing to each category. "Total" includes both IMGT-annotated alleles and novel alleles. Novel alleles refer to those not present in the current IMGT database, encompassing alleles observed exclusively in TWB as well as those previously identified in the 47 HPRC subjects (Yang et al., 2025).

628

629

630

631 **REFERENCES**

632 Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, Visvanath A, Parnaby G, Roddey C,
633 Onuchic V, Finocchio A, Cameron DL, English A, Mehtalia S, Han J, Mehio R, Sedlazeck
634 FJ. 2025. Comprehensive genome analysis and variant detection at scale using DRAGEN.
635 Nat Biotechnol 43: 1177-1191.
636 Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, Rura S, Carroll RJ, Musick A, Linder JE, Jordan
637 IK, Nagar SD, et al. 2024. Genomic data in the All of Us Research Program. Nature 627:
638 340-346.
639 Charmley P, Wang K, Hood L, Nickerson DA. 1993. Identification and physical mapping of a
640 polymorphic human T cell receptor V beta gene with a frequent null allele. J Exp Med 177:
641 135-143.
642 Chen YT, Hsu HC, Lee YS, Liu H, Tan BC, Chin CY, Chang IY, Yang CY. 2021. Longitudinal
643 High-Throughput Sequencing of the T-Cell Receptor Repertoire Reveals Dynamic Change
644 and Prognostic Significance of Peripheral Blood TCR Diversity in Metastatic Colorectal
645 Cancer During Chemotherapy. Front Immunol 12: 743448.
646 Chen Z, Liu Y, Hu S, Zhang M, Shi B, Wang Y. 2021. Decreased Treg Cell and TCR Expansion Are
647 Involved in Long-Lasting Graves' Disease. Front Endocrinol (Lausanne) 12: 632492.
648 Corcoran M, Chernyshev M, Mandolesi M, Narang S, Kaduk M, Ye K, Sundling C, Färnert A,
649 Kreslavsky T, Bernhardsson C, et al. 2023. Archaic humans have contributed to large-scale
650 variation in modern human T cell receptor genes. Immunity 56: 635-652.
651 Dolton G, Rius C, Hasan MS, Wall A, Szomolay B, Behiry E, Whalley T, Southgate J, Fuller A,
652 Morin T, et al. 2022. Emergence of immune escape at dominant SARS-CoV-2 killer T cell
653 epitope. Cell 185: 2936-2951.
654 Elfatih A, Saad C, Ismail S, Al-Muftah W, Badji R, Darwish D, Fadl T, Yasin H, Ennaifar M, Abdel-
655 latif R, et al. 2024. Analysis of 14,392 whole genomes reveals 3.5% of Qataris carry
656 medically actionable variants. Eur J Hum Genet 32: 1465-1473.
657 Feng Y-CA, Chen C-Y, Chen T-T, Kuo P-H, Hsu Y-H, Yang H-I, Chen WJ, Su M-W, Chu H-W,
658 Shen C-Y, et al. 2022. Taiwan Biobank: A rich biomedical research database of the
659 Taiwanese population. Cell Genomics 2: 100197.
660 Fichtner AS, Ravens S, Prinz I. 2020. Human γδ TCR Repertoires in Health and Disease. Cells 9:
661 800.
662 Flanagan J, Liu X, Ortega-Reyes D, Tomizuka K, Matoba N, Akiyama M, Koido M, Ishigaki K,
663 Ashikawa K, Takata S, et al. 2024. Population-specific reference panel improves imputation
664 quality for genome-wide association studies conducted on the Japanese population. Commun
665 Biol 7: 1665.
666 Ford MKB, Hari A, Yeager M, Mirabello L, Chanock S, Numanagić I, Sahinalp SC. 2025.
667 Genotyping of selected germline adaptive immune system loci using short-read sequencing
668 data. Genome Res doi: 10.1101/gr.280314.124.

669 Frimpong A, Ofori MF, Degoot AM, Kusi KA, Gershom B, Quartey J, Kyei-Baafour E, Nguyen N,
670 Ndifon W. 2022. Perturbations in the T cell receptor β repertoire during malaria infection in
671 children: A preliminary study. Front Immunol 13: 971392.
672 Gras S, Chen Z, Miles JJ, Liu YC, Bell MJ, Sullivan LC, Kjer-Nielsen L, Brennan RM, Burrows JM,
673 Neller MA, et al. 2010. Allelic polymorphism in the T cell receptor and its impact on immune
674 responses. J Exp Med 207: 1555-1567.
675 Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, Palsson
676 G, Hardarson MT, Oddsson A, Jensson BO, et al. 2022. The sequences of 150,119 genomes
677 in the UK Biobank. Nature 607: 732-740.
678 Hsu JS, Wu DC, Shih SH, Liu JF, Tsai YC, Lee TL, Chen WA, Tseng YH, Lo YC, Lin HY, et al.
679 2024. Complete genomic profiles of 1496 Taiwanese reveal curated medical insights. J Adv
680 Res 66: 197-207.
681 Huda TI, Mihyu M, Gozlan EC, Arndt MF, Diaz MJ, Zaman S, Chobrutskiy BI, Blanck G. 2021.
682 Specific HLA alleles, paired with TCR V- and J-gene segment usage, link to distinct multiple
683 myeloma survival rates. Leuk Lymphoma 62: 1711-1720.
684 Jacobsen LM, Posgai A, Seay HR, Haller MJ, Brusko TM. 2017. T Cell Receptor Profiling in Type 1
685 Diabetes. Curr Diab Rep 17: 118.
686 Jia X, Zhai TY, Wang B, Zhang JA, Song RH. 2022. High-throughput T cell receptor sequencing
687 reveals differential immune repertoires in autoimmune thyroid diseases. Mol Cell Endocrinol
688 550: 111644.
689 Keane C, Gould C, Jones K, Hamm D, Talaulikar D, Ellis J, Vari F, Birch S, Han E, Wood P, et al.
690 2017. The T-cell Receptor Repertoire Influences the Tumor Microenvironment and Is
691 Associated with Survival in Aggressive B-cell Lymphoma. Clin Cancer Res 23: 1820-1828.
692 Lane J, Duroux P, Lefranc MP. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT
693 standardized approach for immunoglobulin and T cell receptor gene identification and
694 description in large genomic sequences. BMC Bioinformatics 11: 223.
695 Lees WD, Christley S, Peres A, Kos JT, Corrie B, Ralph D, Breden F, Cowell LG, Yaari G,
696 Corcoran M, et al. 2023. AIRR community curation and standardised representation for
697 immunoglobulin and T cell receptor germline sets. ImmunoInformatics 10: 100025.
698 Lefranc MP. 2011. From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized
699 gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold
700 Spring Harb Protoc 2011: 627-632.
701 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
702 ArXiv doi: 10.48550/arXiv.1303.3997.
703 Li Y, Hu J, Wang Y, Liu D, Shi Y, Zhang J, Liu Y, Lin D, Lin J, Hu W, et al. 2021. T-Cell
704 Repertoire Characteristics of Asymptomatic and Re-Detectable Positive COVID-19 Patients.
705 Front Immunol 12: 769442.
706 Lin MJ, Lin YC, Chen NC, Luo AC, Lai SK, Hsu CL, Hsu JS, Chen CY, Yang WS, Chen PL. 2022.
707 Profiling genes encoding the adaptive immune receptor repertoire with gAIRR Suite. Front
708 Immunol 13: 922513.

709 Madeira F, Madhusoodanan N, Lee J, Eusebi A, Niewielska A, Tivey ARN, Lopez R, Butcher S.
710 2024. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. Nucleic
711 Acids Res 52: W521-W525.
712 Manso T, Folch G, Giudicelli V, Jabado-Michaloud J, Kushwaha A, Nguefack Ngoune V, Georga
713 M, Papadaki A, Debbagh C, Pégorier P, et al. 2021. IMGT® databases, related tools and web
714 resources through three main axes of research and development. Nucleic Acids Res 50:
715 D1262-D1272.
716 Marín-Benesiu F, Chica-Redecillas L, Arenas-Rodríguez V, de Santiago E, Martínez-Diz S, López-
717 Torres G, Cortés-Valverde AI, Romero-Cachinero C, Entrala-Bernal C, Fernandez-Rosado
718 FJ, et al. 2024. The T-cell repertoire of Spanish patients with COVID-19 as a strategy to link
719 T-cell characteristics to the severity of the disease. Hum Genomics 18: 94.
720 Mazouz S, Boisvert M, Abdel-Hakeem MS, Khedr O, Bruneau J, Shoukry NH. 2021. Expansion of
721 Unique Hepatitis C Virus-Specific Public CD8(+) T Cell Clonotypes during Acute Infection
722 and Reinfection. J Immunol 207: 1180-1193.
723 Miles JJ, Thammanichanond D, Moneer S, Nivarthi UK, Kjer-Nielsen L, Tracy SL, Aitken CK,
724 Brennan RM, Zeng W, Marquart L, et al. 2011. Antigen-driven patterns of TCR bias are
725 shared across diverse outcomes of human hepatitis C virus infection. J Immunol 186: 901-
726 912.
727 Minervina AA, Komech EA, Titov A, Bensouda Koraichi M, Rosati E, Mamedov IZ, Franke A,
728 Efimov GA, Chudakov DM, Mora T, et al. 2021. Longitudinal high-throughput TCR
729 repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19
730 infection. Elife 10: e63502.
731 Mitchell AM, Michels AW. 2020. T cell receptor sequencing in autoimmunity. J Life Sci (Westlake
732 Village) 2: 38-58.
733 Omer A, Peres A, Rodriguez OL, Watson CT, Lees W, Polak P, Collins AM, Yaari G. 2022. T cell
734 receptor beta germline variability is revealed by inference from repertoire data. Genome Med
735 14: 2.
736 Pajuste F-D, Remm M. 2023. GeneToCN: an alignment-free method for gene copy number
737 estimation directly from next-generation sequencing reads. Sci Rep 13: 17765.
738 Pan RY, Chu MT, Wang CW, Lee YS, Lemonnier F, Michels AW, Schutte R, Ostrov DA, Chen CB,
739 Phillips EJ, et al. 2019. Identification of drug-specific public TCR driving severe cutaneous
740 adverse reactions. Nat Commun 10: 3569.
741 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011.
742 Integrative genomics viewer. Nat Biotechnol 29: 24-26.
743 Shao MM, Yi FS, Huang ZY, Peng P, Wu FY, Shi HZ, Zhai K. 2022. T Cell Receptor Repertoire
744 Analysis Reveals Signatures of T Cell Responses to Human Mycobacterium tuberculosis.
745 Front Microbiol 13: 829694.
746 Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova KV, Kiryukhin D, Titov A, Peshkova IO,
747 Khmelevskaya A, Dianov DV, Malasheva M, et al. 2020. SARS-CoV-2 Epitopes Are

748 Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity 53:
749 1245-1257.
750 Turcinov S, Af Klint E, Van Schoubroeck B, Kouwenhoven A, Mia S, Chemin K, Wils H, Van Hove
751 C, De Bondt A, Keustermans K, et al. 2023. Diversity and Clonality of T Cell Receptor
752 Repertoire and Antigen Specificities in Small Joints of Early Rheumatoid Arthritis. Arthritis
753 Rheumatol 75: 673-684.
754 Wang CY, Yu PF, He XB, Fang YX, Cheng WY, Jing ZZ. 2016. αβ T-cell receptor bias in disease
755 and therapy (Review). Int J Oncol 48: 2247-2256.
756 Yang Y-H, Yao C-Y, Lin M-J, Huang K-T, Lin Y-H, Chiu I-H, Lai S-K, Chen C-Y, Yang Y-C, Hsu
757 C-L, et al. 2025. Unmasking Human T Cell Receptor Germline Diversity: 335 Novel Alleles
758 Identified in 47 Pangenome Reference Individuals Using the gAIRR Suite. bioRxiv doi:
759 10.1101/2025.05.24.655452.
760 Yu X, Song L, Cen L, Cao B, Tao R, Shen Y, Abate-Daga D, Rodriguez PC, Conejo-Garcia JR,
761 Wang X. 2024. Pan-cancer γδ TCR analysis uncovers clonotype diversity and prognostic
762 potential. Cell Rep Med 5: 101764.
763 Zhang J, Wang Y, Huang Y, Tan X, Xu J, Yan Q, Tan J, Zhang Y, Zhang J, Ma Q, et al. 2024.
764 Characterization of T cell receptor repertoire in penile cancer. Cancer Immunol Immunother
765 73: 24.
766 Zhao TM, Whitaker SE, Robinson MA. 1994. A genetically determined insertion/deletion related
767 polymorphism in human T cell receptor beta chain (TCRB) includes functional variable gene
768 segments. J Exp Med 180: 1405-1414.
769 Zheng F, Xu H, Zhang C, Hong X, Liu D, Tang D, Xiong Z, Dai Y. 2021. Immune cell and
770 TCR/BCR repertoire profiling in systemic lupus erythematosus patients by single-cell
771 sequencing. Aging (Albany NY) 13: 24432-24448.
772

773