Disease associated human TCR characterization by deep-learning framework TCR-DeepInsight

Ziwei Xue[1,2,10], Lize Wu[2,3,4,10], Ruonan Tian[1,2], Zuozhu Liu[5], Yadan Bai[1,2,3], Di Sun[6], Yixin Guo[1], Pengwei Chen[1], Yu Zhao[7], Bing He[7], Lie Wang[2,4,8], Jianhua Yao[7], Linrong Lu[1,2,3,6*], Wanlu Liu[1,2,9,11*]

1. Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University, Haining, Zhejiang 314400, China
2. Future Health Laboratory, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing, 314100, China
3. Institute of Immunology and Department of Rheumatology at Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310058, Zhejiang, China
4. Liangzhu Laboratory, Zhejiang University Medical Center, 1369 West Wenyi Road, Hangzhou 311121, China
5. Zhejiang University-Angel Align Inc. R&D Center for Intelligent Healthcare，Zhejiang University-University of Illinois at Urbana-Champaign Institute (ZJU-UIUC Institute), International Campus, Zhejiang University, Haining, Zhejiang 314400, China
6. Shanghai Immune Therapy Institute, Renji Hospital, Jiao Tong University School of Medicine, Shanghai 200025, China
7. AI Lab, Tencent, Shenzhen, Guangdong Province 518000, China
8. Bone Marrow Transplantation Center and Institute of Immunology, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China
9. Department of Orthopedic of the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China
10. These authors contributed equally to this work.
11. Lead contact

*Correspondence: lu_linrong@zju.edu.cn (L.L.), and wanluliu@intl.zju.edu.cn (W.L.)

SUMMARY

T cell function is defined by both T cell receptors (TCR) and T cell gene expression (GEX). Although single-cell technology enables the simultaneous capture of TCR and GEX information, the lack of a reference atlas and computational tools hinders our ability to uncover the fundamental TCR usage rules and to efficiently characterize disease-associated TCRs (dTCR). Here, through the collection of million-scale single-cell GEX-TCR reference atlas comprising 20 diverse disease conditions, we revealed the intrinsic features of TCR-MHC (Major Histocompatibility Complex) restriction in CD4/CD8 lineages. We observed the higher coherence for TCRα/β chains in memory T cells, and detected widely-existing public TCRα/β pairs across individuals. Building upon the reference atlas, we introduced TCR-DeepInsight, a deep-learning framework featuring a disease specificity scoring system that enables the characterization of dTCR clusters with similar GEX-TCR. Our study provides a valuable tool for researchers to analyze single-cell GEX-TCR data and identify dTCRs comprehensively and robustly.

KEYWORDS

T cell receptors, single-cell immune profiling, disease-associated TCR, deep-learning, computational immunology

INTRODUCTION

T cells play a crucial role in the adaptive immune response by recognizing and eliminating pathogen-infected cells and providing cancer immunosurveillance. On the other hand, dysregulated T cell activation can also promote the pathogenesis of autoimmune diseases. In both protective immune and autoimmune responses, T cells exert their function in an antigen-specific manner, through the recognition of pMHC (peptide Major Histocompatibility Complex) via the T cell receptor (TCR) on the cell surface. The TCR consists of α and β chains with highly variable regions on both chains, contributing to the binding specificity between TCRs and antigen peptides. During their development in the thymus, T cells generate an enormous diversity of TCRs through V(D)J recombination and random nucleotide insertions

and deletions in both α and β chains, enabling the recognition of an almost infinite number of pathogen-derived or self-antigens. Recent estimates of the total diversity of TCRs range from 10^{20} to 10^{61}.[1,2]

In response to infections, cancer progression, or autoimmune diseases, T cells undergo clonal expansion, resulting in the production of a large number of T cells with identical TCRs that inherit the same antigen specificity. The characterization of TCR repertoire is critical for understanding the mechanisms of immune responses and developing effective immunotherapies. There are three complementarity-determining regions (CDRs) on both TCR α and β chain, with CDR3 regions directly binding to antigen peptides and thus exhibiting greater diversity compared to CDR1 or CDR2.[3] Early spectratyping of TCRβ reveals preferential V and J segment usage and perturbations in CDR3β pattern in clonally expanded clonotypes in autoimmune diseases including Rheumatoid Arthritis (RA)[4], Multiple Sclerosis (MS)[5], and infectious diseases [6,7] However, these low-throughput and low-resolution analysis often biases the interpretation of Vβ and Jβ usage and CDR3β features. Next-generation sequencing (NGS) has revolutionized TCR repertoire analysis and led to a more concise understanding of TCRs in different disease conditions. [8–11] However, major drawbacks of NGS lie in the lack of TCRα/β chain pairing information and the incomplete coverage of TCR sequences. Single-cell immune profiling overcomes these limitations and has emerged as a powerful tool for analyzing paired TCRα/β repertoires at a single-cell resolution. Moreover, this technology captures the corresponding gene expression (GEX) profiles, allowing the linking of individual T cell clonotypes to their transcriptome dynamics during immune responses. Example studies using single-cell immune profiling include the identification of T cell clones associated with response to checkpoint inhibitors in patients with basal or squamous cell carcinoma before and after anti-PD-1 therapy[12], the detection of T cell clonal expansion in COVID-19 infection[12–17], and the analysis of TCR repertoire diversity in patients with autoimmune diseases.[18,19] Sequencing-based technologies enabled the rapid accumulation of large TCR repertoire data. However, there are few efforts in integrating the growing single-cell immune profiling datasets across studies, and analyzing linked transcriptome and TCRα/β sequences spanning different individuals and diseases due to the technical batch effects across studies.

TCRs with high sequence similarity in CDR3 regions that were used to directly bind

antigen peptides may indicate their relatively high chance in recognizing the same antigen peptide. Accordingly, various computational tools have been developed to perform unsupervised TCR clustering mostly based on CDR3 amino acids. Based on sequence similarity algorithms, GLIPH/GLIPH2, TCRdist/TCRdist3, iSMART, GIANA, and ClusTCR have been proposed.[20–26] TCRdist3 and ClusTCR could take CDR3α and CDR3β amino acids as input, users could also provide CDR1 and CDR2 sequences as optional.[23,26] While iSMART focuses on CDR3β amino acids, TRBV gene usage information could also be supplied as input.[24] GIANA not only considered the sequencing similarity but also amino acid physicochemical properties.[25] However, it can only take fixed length CDR3β amino acid sequences as input which greatly limited its general application.[25] GLIPH2 could be applied to variable length CDR3α/β as well as TRBV gene usage simultaneously, and thus have been the most widely used tool in this field.[21] The above-mentioned algorithms mostly based on clustering CDR3β amino acid information. Moreover, the computation time for most of these similarity-based algorithms increases non-linearly, which hinders their applications on large million-scale TCR data. Recently, deep-learning based methods have also been proposed especially for TCR-pMHC prediction tasks including DeepTCR, NetTCR-2.0, and TCR-BERT.[27–29] Even though those methods were mainly developed for the prediction of TCR-pMHC binding, embeddings learnt from those methods could also be adapted to unsupervised TCR clustering with paired CDR3α/β sequences. [27–29] Aside from TCR sequence similarities, T cell states also contribute significantly to the identification of potential functional TCRs. Simultaneously capturing transcriptome and full-length paired TCR information from single-cell immune profiling dataset would thus substantially enhance our capability to identify functional TCR with right antigen specificity and the proper cell state. CONGA and Tessa both incorporate transcriptome information in their model to jointly represent TCR and cell state similarity.[30,31] However, Tessa only combined CDR3β with transcriptome information, while CONGA used kernel PCA (Principal Component Analysis) for the transcriptome representation, limiting its scalability for large million-level dataset analysis and its extension for incorporating with other datasets due to potential batch effects.

In order to address major challenges mentioned above, we first collected million-level single-cell immune profiling datasets for T cells with various disease conditions

132 including solid tumor, leukemia, inflammation/autoimmune, infections and healthy
133 with full-length TCRα/β chains, single-cell transcriptome, HLA (Human Leukocyte
134 Antigen) genotype and disease meta-information. With this massive dataset, we
135 systematically analyzed the VJ gene preference and amino acid preferences in CDR
136 regions for different CD4, CD8 cell subtypes. In addition, we discovered in general
137 higher coherence for TCRα and TCRβ chains in antigen-experienced CD8 T cells.
138 We also detected the existence of cross-individual public TCRs for both TCRα and/or
139 TCRβ chains and characterized the general features for public TCRs. With this
140 extensive reference atlas, we also developed a deep-learning based framework
141 called TCR-DeepInsight to jointly represent similarity for TCR and transcriptome
142 information. By incorporating the disease specificity scoring system proposed and
143 the reference atlas established in this study, TCR-DeepInsight would be capable of
144 identifying biologically meaningful disease-associated TCRs (dTCR) clusters
145 effectively.

146 **RESULTS**

147 **Integration of million-level paired TCRα/β repertoire with single-cell**
148 **transcriptome information**

149 In recent years, the advent of high-throughput single-cell sequencing technologies
150 has enabled the profiling of immune repertoires at unprecedented depth and
151 resolution. We previously developed a database called human Antigen Receptor
152 database (huARdb).[32] With our continuous efforts, these data have been expanded
153 to 28 studies and 464 biological samples from 291 individuals, 9 tissue types, and 20
154 disease conditions including solid tumor, leukemia, inflammation/autoimmune,
155 infections and healthy (**Figure1A and 1B; TableS1**). Altogether, we have collected
156 1,017,877 high-confidence T cells with paired transcriptome and full-length α/β TCRs
157 information, which enable us to perform a large and unbiased analysis of the paired
158 TCRα/β repertoire across diseases and individuals. These datasets contain 459,013
159 unique TCRα chains defined by the same TRAV(T cell receptor alpha variable
160 gene)-CDR3α-TRAJ(T cell receptor alpha joining gene), and 604,728 unique TCRβ
161 chains defined by the same TRBV(T cell receptor beta variable gene)-CDR3β-
162 TRBJ(T cell receptor beta joining gene) (**Figure1A**). We defined clonally expanded T
163 cells as T cells with the same V/J genes and CDR3 amino acid sequence (consist of
164 TRAV/TRBV motif, CDR3α/β middle region (mr), TRAJ/TRBJ motif) of both α and β

165 chains in the same individuals (**FigureS1A**). After merging clonal expanded TCRs,
166 we obtained 635,393 unique TCR α/β clonotypes from 291 individuals. In addition,
167 we also extracted the HLA genotypes for each individual from the single cell
168 transcriptome dataset using arcasHLA.[33] For all the 291 individuals collected in this
169 study, A*02, B*15, C*07 were the most frequent MHC class I molecules, while
170 DRB*103, DRB*115, DQB*103, DQA*101, DPB*102, DPB*104 were the most
171 enriched MHC class II molecules (**FigureS1B-S1H**).

172 To gain an accurate cellular phenotype associated with full-length TCRs from gene
173 expression profile, we first applied a rigorous approach to eliminate batch effects
174 across samples and studies involving millions of single cells. We integrated the gene
175 expression datasets with an in-house developed variational autoencoder (VAE)
176 based deep-learning framework, and obtained a latent embedding of transcriptome
177 features representing T cell subtype and functional states (**Figure1C**). We manually
178 categorized T cells into six major states including, naïve (CD8 T_n) and antigen-
179 experienced $CD8^+$ T cells (CD8 T), naïve (CD4 T_n), conventional memory (CD4
180 T_{conv}) and regulatory $CD4^+$ T cells (CD4 T_{reg}), and Mucosal invariant T cells (MAIT),
181 aligned with the expression pattern of key marker genes (**Figure1C and 1D**). T cells
182 including CD8 T, CD4 T_{conv} and CD4 T_{reg} in cycling status also formed a unique
183 cluster in the integrated UMAP, suggesting the distinctive transcriptome features in
184 cycling T cells (**Figure1C and 1D**). Using our VAE model, we successfully integrated
185 data from 28 studies with minimal batch effects, and different cell subtypes were
186 evenly distributed across various studies (**FigureS2A**). We observed a diverse cell
187 type composition across solid tumor, inflammation, CPI-irAE (checkpoint inhibitor
188 associated immune-related adverse events), AML (Acute Myeloid Leukemia),
189 COVID-19, T-LGLL (T-cell large granular lymphocyte leukemia), and healthy
190 conditions (**Figure1B and S2B**). To validate our cell type annotation, we applied the
191 singleR automated cell type annotation tool, using PBMC T cell subtypes from
192 healthy donors as cell type annotation reference, and found consistent cell subtype
193 annotation (**FigureS2C**). [34,35]

194 In this dataset, the percentage of unique TCRα/β pairs derived from CD4 T cells is
195 higher than the percentage of CD4 T cells, indicating overall higher clonal expansion
196 level of CD8 T cells than CD4 T cells (**Figure1E**). Overall, all TCRα/β pairs collected
197 in this study consisted of 43.0% CD8 T cells, 48.0% CD4 T_{conv}, 7.0% CD4 T_{reg}, 2.0%

198 MAIT cells, and 0.2% ambiguous (transcriptome can not be classified into either CD4
199 or CD8 T cells, see **Methods**), while our unique TCRα/β pairs across all individuals
200 (n=634,252) consisted of 22.0% CD8 T, 65.1% CD4 T_{conv} cells, 7.6% CD4 T_{reg}, 1.8%
201 MAIT cells, and 3.3% ambiguous (identical TCRα/β pairs linked with cells with no
202 dominant state, see **Methods**) (**Figure1E**). So far, to our knowledge, this is the most
203 comprehensive atlas containing millions of TCRα/β pairs with transcriptome, full-
204 length paired-TCR, HLA, and disease information.

205 **Association of intrinsic TCR features and major T cell subtypes**

206 Previous studies of TCR repertoire using high-throughput sequencing have revealed
207 preferences in VJ usage selection, and significant bias in VJ segment joining has
208 been identified in single-chain repertoires.[36] By using single-cell TCR sequencing
209 followed by CD4 and CD8 FACS sorting, TCRα/β pairing have been proposed as a
210 more precise way to infer repertoire functionality.[37] With our reference atlas
211 containing paired full-length TCR and transcriptome profiles, we can perform
212 comprehensive analysis of the intrinsic features of TCR in different cell subtypes.

213 We calculated the odds ratio and presented significantly enriched TRAV-TRAJ and
214 TRBV-TRBJ joining and TRAV-TRBV pairing in either CD8 T, CD4 T_{conv}, CD4 T_{reg}, or
215 MAIT cells (**Figure2A; TableS2**). Our results confirmed the well-recognized
216 phenomena that MAIT cells prefer to select TRAV1-2 paired with TRAJ12, TRAJ20
217 and TRAJ33 to form its invariant α chains.[38] For β chain, MAIT cells preferentially
218 select TRBV6-4, TRBV4-2, TRBV6-1, TRBV20-1, and TRBV6-2 as reported
219 before.[39] With the pairing information for TCRα/β chains, we were able to identify the
220 significant preference for TRBJ2-1 and TRBJ2-3 in TRAV1-2^{+}TRAJ33^{+}TRBV6-4^{+}
221 MAIT cells.

222 Our analysis of TCR repertoire in CD8 T cells revealed no significant preference for
223 TRAV-TRAJ genes. However, we identified TRBV27/TRBJ2-7, TRBV7-9/TRBJ2-1,
224 and TRBV27/TRBJ2-1 as the top three significantly enriched TRBV/TRBJ joining
225 pairs. In terms of TRAV/TRBV pairing, we found that TRAV19/TRBV7-9,
226 TRAV14DV4/TRBV7-9, and TRAV14DV4/TRBV9 were the most dominant pairings
227 (**Figure2A**).

228 The most enriched TRAV/TRBV pairings in CD4 T_{conv} cells were
229 TRAV23DV6/TRBV20-1, TRAV23DV6/TRBV5-1, and TRAV8-4/TRBV7-2

230 (**Figure2A**). We observed significant enrichment in various TRAV/TRAJ joining pairs,
231 with TRAV9-2/TRAJ22 being the most dominant pairing. In contrast, we detected
232 only two significantly enriched TRBV/TRBJ joining pairs, namely TRBV3-1/TRBJ1-6
233 and TRBV5-4/TRBJ1-6 (**Figure2A**). Visualizing enriched TRAV/TRAJ, TRBV/TRBJ
234 joining and TRAV/TRBV pairing enriched in CD8 T, CD4 T_{conv}, or MAIT cells on the
235 VAE integrated UMAP (Uniform Manifold Approximation and Projection) showed a
236 diverse distribution with no specificity towards naïve or memory repertoire
237 (**Figure2B**). Marker gene expression in cells with selected VJ gene usage further
238 confirmed the TRAV/TRAJ, TRBV/TRBJ, and TRAV/TRBV preference in CD8 T,
239 CD4 T_{conv}, and MAIT cells (**Figure2C**). For example, MAIT-specific joining coincides
240 with relatively high expression of MAIT-specific gene markers including *KLRB1*
241 (Killer Cell Lectin-Like Receptor Subfamily B Member 1), *ZBTB16* (Zinc Finger and
242 BTB Domain Containing 16), and *SLC4A10* (Solute Carrier Family 4 Member 10)
243 (**Figure2C**). Interestingly, we also observed an enrichment of cells using MAIT
244 specific VJ segments, which were not co-localized with conventional MAIT clusters
245 on the integrated UMAP (**Figure2B**). These groups of cells were derived from a
246 study from T-LGLL patients[40], and displayed transcriptome features of terminal
247 effector T cells including high expression of *CX3CR1* (X3-C Motif Chemokine
248 Receptor 1) and *GZMB* (Granzyme B)(**Figure1D**). Although we do not know the
249 origin of these T cells, nor whether they recognize MR1 (Major Histocompatibility
250 Complex class I-related molecule 1)-restricted antigens as traditional MAIT cells or
251 regular peptide antigens recognized by CD8 T cells, these findings are interesting
252 and the functional role of these T cells in T-LGLL may require further investigation.

253 The CDR1 and CDR2 regions of the TCR typically contact the conserved α-helices
254 of MHCs, and the usage of amino acids in these regions may influence the
255 recognition between TCRs and class I and class II MHCs.[41] Comparing to both CD4
256 T_{conv} cells and CD4 T_{reg} cells, our analysis indicated that CD8 T cells utilize
257 significantly less hydrophobic and special amino acids, and more negatively,
258 positively charged or polar uncharged amino acids in the CDR1α region
259 (**FigureS3A**). For CDR1β region, no significant preferences for hydrophobic amino
260 acids in CD4 T_{reg}, CD4 T_{conv}, or CD8 T cells, while significantly more negatively
261 charged amino acids usage and less polar uncharged or special amino acids usage
262 for CD8 T cells compared with CD4 T_{reg} and CD4 T_{conv} cells (**FigureS3B**). In CDR2α

263 region, significantly more usage of polar uncharged and less usage of positively
264 charged or special amino acids were detected for CD8 T cells comparing with CD4
265 T_{reg} or CD4 T_{conv} cells, while less hydrophobic amino acids and more negatively
266 charged amino acids usage in CD4 T_{reg} comparing with CD8 T and CD4 T_{conv} cells
267 (**FigureS3C**). Compared with both CD4 T_{conv} and CD4 T_{reg} cells, more hydrophobic
268 and negatively charged and less polar uncharged, positively charged or special
269 amino acids usage were observed in CD8 T cells for CDR2β regions (**FigureS3D**).

270 As the CDR3 region has been demonstrated as the most important part of the TCR
271 in recognizing MHC class I or MHC class II displayed peptide, we asked whether the
272 CDR3 repertoire may display different characteristics in various T cell subtypes,
273 including CD4 T_{reg}, CD4 T_{conv}, and CD8 T cells. In our analysis, we find limited amino
274 acid length difference for either CDR3α/β or CDR3α/β-mr in CD4 T_{reg}, CD4 T_{conv}, and
275 CD8 T cells, while MAIT cells used shorter CDR3α or CDR3α-mr due to its relatively
276 invariant α chain usage (**FigureS3E; TableS3**). Interestingly, we found that CD8 T
277 cells use a significantly higher percentage of hydrophobic and negatively charged
278 amino acids, and a lower percentage of polar uncharged and positively charged
279 amino acids in the CDR3-mr for both α and β chain compared with CD4 T_{reg} and
280 CD4 T_{conv} cells (**Figure2D and 2E; TableS4**). CD4 T_{reg} cells showed the lowest
281 percentage of negatively charged amino acids and the highest percentage of
282 positively charged amino acids in CDR3-mr for both chains compared with CD8 T
283 and CD4 T_{conv} cells, while CD4 T_{conv} favored the usage of polar uncharged amino
284 acids in CDR3-mr for both chains compared with CD8 T and CD4 T_{reg} cells
285 (**Figure2D and 2E**).

286 **TCRα/β pairing is more coherent in memory T cells than in naïve T cells**

287 As a comprehensive atlas of paired and full-length TCRα/β, our dataset is useful for
288 understanding the V(D)J recombination biology and TCRα/β pairing patterns. As it
289 has been previously reported that B cell receptors (BCRs) with the same heavy-
290 chain-variable (VH) gene and CDRH3 (heavy-chain complementarity-determining
291 region 3) amino acid sequence in their heavy chains from memory B cells or plasma
292 cells that produce functional antibodies adopt the same light-chain-variable (VL)
293 gene in their light chains.[42] The phenomenon in memory B cells but not in naïve B
294 cells is referred to as light chain coherence in BCRs.[42] We thus hypothesized that a
295 similar phenomenon may also exist in TCRs in memory state (CD4 T_{conv} cells or CD8

296 T cells). In our paired TCRα/β repertoire collections, we can detect TCRα/β chains
297 using the same TRBV-CDR3β paired with the same TRAV gene but different
298 CDR3α, or different TRAV gene and different CDR3α (**Figure3A**). In addition, we
299 also observed TCRα/β chains with the same TRBV gene and similar but mismatched
300 CDR3β pairing with either the same or different TRAV and different CDR3α
301 (**Figure3A**). Similar cases were also observed when analyzing TRBV usage and
302 CDR3β for relatively fixed TRAV-CDR3α (**Figure3B**).

303 To analyze the TCRα/β chain coherence, we grouped TCRs with same TRBV by
304 their deriving T cell subtypes (CD8 T_n, CD8 T, CD4 T_n, CD4 T_{conv}, and CD4 T_{reg}) and
305 CDR3β similarity (CDR3β within 10 Levenshtein distance), and then calculated the
306 percentage of TCRα/β pairs using the same TRAV gene. Our results suggested that
307 TRAV coherence was in general higher in CD8 T cells in memory state than other
308 cell types (**Figure3C**). Consistent phenomena were observed for grouping TCRs
309 with the same TRAV and allowing mismatches in CDR3α, showing strongest TRBV
310 coherence in CD8 T cells (**Figure3D**). Interestingly, when grouping TCRs by their
311 CDR3β allowing more than two mismatches on average, we can no longer observe
312 the TRAV coherence for CD8 T cells (**Figure3C**). However, TRBV coherence in CD8
313 T cells only allows one mismatch in CDR3α (**Figure3D**). The occurrence of the
314 identical V gene usage in a randomly chosen set of TCRs would be a rare event.
315 The increased V gene coherence observed in CD8 memory T cells might suggest
316 these cells have experienced specific antigen. Identifying additional TCR clusters
317 exhibiting V gene coherence could potentially lead to the discovery of TCRs that
318 possess shared antigenic specificity.

319 **Characterization of public TCRs**

320 Discovering TCRs with identical VJ segments and CDR3 amino acid sequences for
321 both chains in different individuals is often considered an extremely rare event due to
322 the total combinatorial diversity of TCR repertoire.[43] These TCRs, hereafter called
323 public TCRs, are thought to be generated by the selection of the same antigen
324 epitope displayed by the same MHC molecule. In our massive paired TCRα/β
325 collection, we identified 858 public TCRα/βs (with identical TRAV-CDR3α-TRAJ, and
326 TRBV-CDR3β-TRBJ), 66,471 public TCRαs (identical TRAV-CDR3α-TRAJ), and
327 10,930 public TCRβs (identical TRBV-CDR3β-TRBJ) (**Figure4A; TableS5-S7**).
328 Given the comparable number of total unique TCRα and TCRβ in the dataset, the

329 significantly higher number of public TCRα indicates a generally more conserved α
330 chain repertoire across populations. Cell states linking to public TCRα/β pairs were
331 mostly enriched in CD8 T, CD4 T_{conv}, and MAIT cells representing 35.8%, 36.0%,
332 and 13.3% respectively, while in both public TCRα and TCRβ, CD4 T_{conv} were most
333 dominant compared to other cell states (**Figure4B**). Public TCRs (either TCRα/β
334 pairs, TCRα, or TCRβ) shared in only two individuals displaying variable cell states
335 were classified into ambiguous (see **Methods**).

336 To understand potential characteristics contributing to TCRα/β recurrence, we
337 annotated each TCRα or TCRβ sequence with generation probability calculated by
338 OLGA[44] using CDR3 amino acid sequence and VJ segment genes. We found public
339 TCRα and TCRβ, or public TCRα/β pairs were all associated with significantly higher
340 generation probability (**Figure4C**). Compared to non-public TCRs, the negative log-
341 transformed generation probability ($-Log_{10}P_{gen}$) is smaller in public TCRα
342 (6.7067±0.0019 compared to 8.4534±0.0027) and TCRβ (7.9750±0.0050 compared
343 to 10.3934±0.0025) (**Figure4C**). Interestingly, when visualizing generation probability
344 for TCRα and TCRβ in public TCRα/β pairs, TCRβ instead of TCRα exhibited greater
345 difference compared to non-public TCRα/β pairs (**Figure4C**).

346 We also analyzed CDR3-mr amino acid length in public TCRs, and observed a
347 generally shortened CDR3mr amino acid length in either public TCRα
348 (2.1834±0.0034 compared to 3.2468±0.0028) or public TCRβ
349 (3.6881±0.0080compared to 5.5380±0.0021) (**Figure4D**). Consistent with the
350 observation for generation probability, when examining the CDR3-mr amino acid
351 length for public TCRα/β pairs, only TCRβ but not TCRα CDR3-mr amino acid length
352 demonstrated a shortening pattern (**Figure4D**). This phenomenon implies that TCRβ
353 may have a more significant role in TCR recurrence resulting from same antigen
354 challenge, which aligns with previous understanding that TCRβ contributes more to
355 antigen specificity determination.

356 **Identification of epitope- and disease-related public TCRs**

357 We hypothesized that public TCRs that occur in different individuals in the same
358 disease setting may bind to disease-specific epitopes. We asked whether public
359 TCRα, TCRβ or TCRα/β pairs would be restricted in common diseases such as solid
360 tumours, autoimmune diseases, or COVID-19. Most of the public TCRs are shared in

361 individuals with different disease settings, and these TCRs may not target disease-
362 specific epitopes (**Figure4E-4G**). However, we indeed identified an enrichment of
363 SARS-CoV-2-specific public TCRα/β pairs (but not in public TCRα or TCRβ), and
364 this phenomenon may be explained by the relatively restrained epitope peptide
365 sequence derived from SARS-CoV-2 (**Figure4E**).

366 We next asked whether the public TCRs would bind epitopes derived from common
367 viruses such as Epstein-Barr virus (EBV), Cytomegalovirus (CMV), and Influenza A
368 virus that remain latently present in target cells after the primary infection. In order to
369 match TCRs with known epitope specificity, we build a reference dataset containing
370 paired CDR3α/β sequences with experimentally validated epitopes from various
371 sources (**FigureS4A; TableS8**).[22,27,45–47] Since CDR3α/β-epitope pairs datasets
372 mostly collected from EBV, CMV, Influenza A, or SARS-CoV-2 viruses, matched
373 TCRα/β pairs with identical CDR3α and/or CDR3β sequences were mostly enriched
374 for antigen listed above and do not harbor significant disease specificity
375 (**FigureS4B**).

376 We performed in-depth analysis on public TCRs that were clonally expanded in at
377 least two individuals across studies. Linking transcriptome features to public TCRα/β
378 defines their potential functional roles in different disease conditions. We identified
379 public and clonally expanded TCRα/β pairs (TRAV5-CAESTGKLIF-
380 TRAJ37/TRBV29-1-CSVGTGGTNEKLFF-TRBJ1-4 and TRAV21-
381 CAVLMDSNYQLIW-TRAJ33/TRBV10-2-CASSEDGMNTEAFF-TRBJ1-1) in 12 or 7
382 HLA-A02:01 positive individuals, with matched epitope specificity to HLA-A02:01
383 restricted BMLF1280-EBV and LMP2-EBV by both CDR3α/β sequences,
384 respectively (**FigureS5A**). We also find another public and clonally expanded
385 TCRα/β pair (TRAV23DV6-CAASIGNFGNEKLTF-TRAJ48/TRBV4-2-
386 CASSPSRNTEAFF-TRBJ1-1) in 5 HLA-B07:01 positive individuals matched to HLA-
387 B07:02 restricted pp65-CMV epitope (**FigureS5A**). Public TCRα/β with previously
388 unknown antigen specificity were also detected in our analysis. For example,
389 TRAV26-2-CILRGAGGTSYGKLTF-TRAJ52/TRBV27-CASSLQGANYEQYF-TRBJ2-
390 7 were identified in five HLA-B07* positive individuals, and TRAV17-
391 CATEGDSGYSTLTF-TRAJ11/TRBV6-5-CASSGQGGYGYTF-TRBJ1-2 were
392 detected in four HLA-A02* individuals (**FigureS5A**). T cells harboring these public
393 TCRα/β pairs were likely to be clonally expanded and enriched in antigen-

394 experienced CD8 T cells (**FigureS5A**).

395 Although observing public TCRs from individuals that are both clonally expanded
396 and associated with the same disease are extremely rare events, we observed
397 several such TCR pairs derived from CD4 or CD8 T cells. For example, the public
398 and clonally expanded TCRα/β pairs enriched on antigen-experienced CD8 T cells
399 (TRAV17-CATDYNQGGKLIF-TRAJ23/TRBV4-1-CASSQDPGAWETQYF-TRBJ2-5)
400 were observed in two individuals with inflammatory diseases (ulcerative colitis and
401 chronic pancreatitis) sharing HLA-A02, HLA-A03, HLA-B07, and HLA-C07
402 (**FigureS5B**). Public and clonally expanded TCR specific to psoriatic arthritis
403 (TRAV17-CATDGYTGNQFYF-TRAJ49/TRBV12-3-CASSLGQNNEQFFTRBJ2-1)
404 was also enriched with antigen-experienced CD8 T cells in two individuals sharing
405 HLA-B15, and HLA-C03 (**FigureS5B**). We have also found other clonally expanded
406 and public TCRs in either CD4 or CD8 T cells in solid tumors, acute myeloid
407 leukemia, and COVID-19 associated with common HLAs (**FigureS5B**). These TCRs
408 are highly likely to play crucial roles in various disease settings, although
409 experimental validation remains necessary to confirm their importance.

410 **TCR-GEX joint-embedding learned by deep learning models facilitates TCRα/β**
411 **clustering with disease-association and shared transcriptome state**

412 The coherence in TCRα or TCRβ chains in antigen-experienced CD8 T cells is
413 dependent on the CDR3α or CDR3β sequence with one or two mismatched amino
414 acids. These results suggest that both CDR3α and CDR3β contributed to the forming
415 of epitope specificity. In most cases, researchers were unable to link a TCR
416 repertoire to a previously known set of antigen peptides. The MHC displayed epitope
417 peptide for patients with cancer and autoimmune diseases, or even specific viral
418 infections could be diverse across individuals, and so does the TCR repertoire. One
419 of the tasks in TCR repertoire analysis is to cluster TCRs with higher probability to
420 recognize the same antigen peptide. It has previously been discussed that using
421 both CDR3α or CDR3β amino acid sequences could improve prediction accuracy in
422 epitope specificity of TCRs.[48] Moreover, it has been recently emphasized that
423 incorporating TCR sequence and cell-specific covariates from single-cell data can
424 enhance the inference of T cell antigen specificity.[49,50]

425 To robustly identify potential disease associated TCRα/β pairs considering both TCR

426 sequence similarity and transcriptome features from million-level paired TCRα/β
427 repertoire, we developed a deep-learning based framework named TCR-
428 DeepInsight. We first extracted the latent embedding learned by the VAE model to
429 represent the GEX features (**Figure5A**). In parallel, we adopted the Bidirectional
430 Encoder Representations from Transformers (BERT) model to learn the underlying
431 relationship between V genes and CDR3 sequence and the pairing pattern of the two
432 chains (**Figure5B**). We used 635,322 unique TCRα/β CDR3 sequences along with
433 TRAV, TRAJ, TRBV, TRBJ gene usage for pretraining BERT by randomly hiding
434 certain VJ gene or amino acids used in either TCRα or β chains (**Figure5B**). The
435 model does not require prior knowledge of the epitope binding specificity and thus
436 clustering on the embedding of the output layer was unsupervised. Given the higher
437 importance of TCRβ chain in its function, we apply the pre-trained model on the
438 same dataset containing 635,322 unique TCRα/β pairs using the TCR embedding
439 pooled from token features from TRAV, TRBV, CDR3β and TRBJ. After principal
440 component analysis transformation to a 64 dimensional TCR embedding space, we
441 found that the TCRs can be primarily clustered by TRBV usage in the 2-dimensional
442 UMAP projection (**Figure5C**). This TCR embedding could be applied to downstream
443 tasks including unsupervised TCRα/β clustering and query for TCRα/β with known
444 epitope specificity (**Figure5C**). To match GEX embedding with TCR embedding and
445 avoid single TCR embedding shared by multiple GEX embedding, we averaged the
446 VAE latent embedding for unique TCRα/β pair by their deriving cells and obtained an
447 aggregated GEX embedding (**Figure5C**). TCRα/β linked with multiple cell states
448 were classified as the most dominant cell state, while TCRα/β linked cells with no
449 dominant state were classified as ambiguous (**Figure5C**). Consistent with the
450 original GEX embedding, similar cell states were clustered together on the UMAP of
451 aggregated GEX embedding (**Figure5C**). We then concatenated the TCR
452 embedding learned from BERT and the aggregated GEX embedding from VAE to
453 obtain a TCR-GEX joint representation of TCRα/β pairs from million-level paired
454 TCRα/β repertoire with single-cell transcriptome information (**Figure5C**). We found
455 that our TCR-GEX joint representation was capable of capturing TCR similarity of
456 amino acids for both CDR3α and CDR3β and a positive correlation between
457 Levenshtein distance of amino acid sequence and Euclidean distance of the joint
458 representation within eight Levenshtein distance (**Figure5D and 5E**).

459 With the massive paired TCRα/β repertoire along with disease information collected
460 in this study, we were able to identify disease associated TCRα/β clusters with
461 shared transcriptome state with TCR-DeepInsight (**Figure5C**). Using the Euclidean
462 distance for unique TCRα/β pairs in TCR-GEX joint representation (referred as TCR
463 distance), we clustered each TCRα/β pairs with 39 most similar TCR neighbors
464 using GPU(Graphics Processing Unit)-accelerated indexed kNN (k-Nearest
465 Neighbors, k=40) search (**FigureS6A**). Then we iteratively selected a unique
466 TCRα/β pair as anchor and ranked its neighboring TCRs by its TCR-GEX joint
467 representation similarity (**FigureS6A**). If an anchor TCR and its top similar
468 neighboring TCRs (minimal 2) were derived from the same disease, these TCRs
469 were then defined as disease-TCR-cluster, while the same count of next ranked
470 TCRs were defined as non-disease-TCR-cluster (**FigureS6A**). TCR distance within
471 disease clusters were then calculated and defined as TCR similarity score, while
472 TCR distance between disease-TCR-cluster and non-disease-TCR-cluster were
473 defined as disease specificity score (**FigureS6A**). Interestingly, we found a positive
474 correlation between the TCR similarity score and the averaged Levenshtein distance
475 of CDR3β but not CDR3α, while a positive correlation was found in the averaged
476 between-cluster difference Levenshtein distance of CDR3β and the disease
477 specificity score (**Figure5F-5I**). For each disease condition, we linked each TCR
478 cluster with cell state information and calculated the TCR similarity score and
479 disease specificity. Consistent with previous definition, disease-associated TCR
480 clusters with dominant cell state were classified into either CD8 T cells, CD4 T_{conv},
481 CD4 T_{reg}, MAIT, or ambiguous. TCR clusters in each disease condition were then
482 classified as dTCR clusters under certain cut-off (**FigureS6A**, see **Methods**). With
483 the GEX information available, we were able to project dTCR clusters onto
484 aggregated GEX and original GEX embedding to visualize their transcriptome
485 features and additional information such as HLA information and clonal expansion
486 level could also be retrieved (**FigureS6A**). Overall, we identified dTCR clusters in
487 various disease conditions, including solid tumors, COVID-19, and Colitis containing
488 relatively large sample numbers, while dTCR clusters could also be detected in
489 Kawasaki disease and Ankylosing Spondylitis which have relatively fewer individuals
490 (**Figure5J; TableS9**).

491 **Identification of dTCR clusters from various disease conditions**

492 In solid tumors, one of the top dTCR clusters ranked by the TCR similarity score
493 contains was derived from three A*33:85 and B*58 positive individuals with
494 nasopharyngeal carcinoma (NPC), breast cancer (BC), and esophagus squamous
495 cell carcinoma (ESCC). This dTCR cluster showed utilization of TCRα/β using
496 TRAV17, TRAJ45, TRBV18 and TRBJ2-1, with a remarkably similar pattern of
497 CDR3α and CDR3β amino acid sequences and no known antigen specificity have
498 been reported (**Figure6A**). Another dTCR cluster associated with basal cell
499 carcinoma (BCC) and breast cancer (BC) utilize TRAV29DV5, TRBV25-1 and
500 TRBJ2-5 with similar CDR3 patterns. Both solid tumor dTCR clusters displayed
501 coherence in the transcriptome states in antigen-experienced CD8 T cells as
502 visualized by the projection to the aggregated and original GEX embedding
503 (**Figure6A**). Interestingly, we also detected a CD4 T_{conv} dTCR cluster with extremely
504 similar CDR3α/β sequences associated with NPC, BCC and SCC (squamous cell
505 carcinoma), shared in 4 individuals with DQA*102 and DRB*107 (**FigureS6B**).

506 We also identified dTCRs derived from COVID-19 patients with high disease
507 specificity and transcriptome similarities. We found a dTCR cluster from 4 A*01
508 individuals utilizing TRAV9-2, and TRBV27 with similar CDR3β but not CDR3α
509 sequence, whose deriving cells were clustered in the antigen-experienced CD8 T
510 state (**Figure6B**). We found that this TCR cluster may recognize the A*01:01
511 restricted QYIKWPWYI derived from the nonstructural protein of SARs-CoV-2
512 reported by Francis et al.[47] using tetramer staining followed by single-cell
513 sequencing. We also found other COVID-19 dTCR clusters with high disease
514 specificity, overlapping HLA alleles, and shared transcriptome states, but without
515 previously known binding SAR-CoV-2 derived epitopes (**Figure6B**).

516 Our method is also applicable to datasets with few individuals. In a study containing
517 two B*27 positive individuals with Ankylosing Spondylitis, we found that the top-
518 ranked TCR cluster utilized TRAV21, TRBV9, and TRBJ2-3 with conserved CDR3β
519 motif but diverse TRAJ usage and CDR3α sequence (**Figure6B**). It has been
520 previously reported as an over-representation of TRBV9-CDR3-TRBJ2-3 sequence
521 motifs in individuals with ankylosing spondylitis (AS) compared to healthy individuals
522 who carry the B*27 allele and further research demonstrates that these TCRs
523 recognize both self-antigen and microbial antigen displayed by the B*27.[51] Cell types
524 associated with these TCRs were located on the border of CD8 T cells, and CD4

525 T_{conv} cells in both aggregated and original GEX embedding, suggesting its potential
526 functional role (**Figure6C**).

527 **Extension of TCR-DeepInsight for TCR query with known antigen specificity**
528 **and model transfer on new dataset**

529 TCR-DeepInsight provided a pre-trained BERT model on massive TCRα/β pairs and
530 the dataset collected in this study provided a comprehensive reference TCR
531 repertoire along with single cell transcriptome information. In order to infer potential
532 antigen specificity for TCRα/β pairs collected, we next extended the TCR BERT pre-
533 trained model to the task of querying TCRα/β with known epitope information. By
534 processing TCRα/β pairs containing VJ gene usage and CDR3α/β sequences with
535 known epitope information through our BERT pre-trained model, we obtained the
536 query TCR embedding with known epitope information. Query TCR embedding and
537 reference TCR embedding were then aligned and kNN search for epitope specific
538 TCRα/β were performed. Under certain cut-off for TCR distance, TCRα/β in the
539 reference dataset similar to query TCRα/β with known epitope information could be
540 identified, and additional information such as HLA information, clonal expansion level
541 and visualization on GEX embedding could be retrieved (**Figure7A**, see **Methods**).
542 By querying two TCRα/β pairs recognizing A*24:01 restricted epitope NYNYLYRLF
543 from SARS-CoV-2 Spike protein and an B*15:01 restricted epitope RVAGDSGFAAY
544 from SARS-CoV-2 membrane glycoprotein, we identified TCR using similar VJ
545 genes and CDR3β. For both queried TCRs, the top three most similar TCRs were
546 both derived from COVID-19 disease condition, and cells harboring these TCRs
547 predominantly were antigen-experienced CD8 T cells (**Figure7B, C**).

548 In general, TCR-DeepInsight provided a comprehensive reference containing
549 millions of TCRα/β pairs with side-by-side transcriptome and disease meta-
550 information, allowing more precise identification of disease-associated TCRs for
551 novel datasets. To demonstrate the model transfer performance for TCR-
552 DeepInsight on new dataset, we leveraged an additional scRNA-seq and scTCR-seq
553 dataset to identify unique clusters of TCRα/β pairs associated with gastric cancer.[52]
554 Our method involved transferring the GEX embedding from the VAE model and the
555 TCR embedding from the BERT model to new datasets. We accomplished this by
556 implementing architecture surgery on the VAE model, a technique that allows the
557 transfer of previously trained weights on a larger reference atlas to new smaller

558 datasets (**Figure7D**). Simultaneously, the query TCR embedding can be obtained
559 from the same pretrained BERT model as used in the reference dataset followed by
560 PCA transformation using the same weight (**Figure7D**).

561 We observed that our query gastric cancer dataset could be projected into a
562 common low dimensional space with our reference atlas using both unaggregated
563 and aggregated GEX embeddings (**Figure7E**). By applying the same strategy as for
564 other disease conditions, we identified two dTCR clusters associated with gastric
565 cancer that exhibited high sequence similarity and disease specificity (**Figure7F;**
566 **TableS10**). Notably, one TCR cluster derived from CD8 T cells and was clonally
567 expanded, characterized by the usage of TRAV3, TRAJ30, TRBV27, and TRBJ2-3
568 genes, and CDR3α and CDR3β sequences that were similar in two individuals who
569 were positive for B*13 and C*03 (**Figure7F**). The other TCR cluster derived from a
570 more similar transcriptome state of CD8 T cells and was also clonally expanded,
571 characterized by the usage of TRAV36DV7, TRAJ28, TRBV28, and TRBJ1-4 genes,
572 and CDR3α and CDR3β sequences that were similar in two individuals who were
573 positive for A*02 and C*03 (**Figure7F**).

574 **DISCUSSION**

575 With the advent of single-cell immune profiling technology, vast amounts of single-
576 cell gene expression (GEX) and T cell receptor (TCR) data have been generated.
577 These data offer tremendous potential for studying TCR biology and identifying
578 potential functional TCRs. However, a major challenge is the lack of a reference
579 atlas that provides easy access to these datasets. On the other side, with the rapid
580 development of TCR engineering, the use of TCR-T in disease immunotherapy is
581 becoming increasingly promising. Single-cell immune profiling datasets provide a
582 valuable platform for the identification of functional TCRs. However, the lack of
583 efficient computational tools to integrate and identify functional TCRs remains a
584 significant bottleneck in this field.

585 In this study, we established a comprehensive dataset comprising millions of T cells
586 obtained from single-cell immune profiling, along with their single-cell transcriptomes,
587 full-length paired TCRs, HLA types, and disease information. This dataset serves as
588 a valuable reference atlas for a wide range of computational tasks in immunology.
589 Using this massive dataset, we comprehensively analyzed the intrinsic features of

590 TCRs in CD4 and CD8 lineages, identified coherence of public TCRα/β chains in
591 memory CD8 T cells, and illustrated the widely-existing public TCRs across
592 individuals. Moreover, we developed TCR-DeepInsight, a powerful and scalable
593 deep-learning framework that can effectively cluster TCRα/β pairs with high
594 transcriptome similarity in single-cell immune profiling datasets. With the disease
595 scoring system embedded in TCR-DeepInsight, researchers can readily identify
596 disease-associated TCR clusters using the dataset collected in this study as a
597 reference. This tool offers a valuable resource for the analysis of single-cell gene
598 expression and TCR data, facilitating the identification of disease-specific TCR
599 repertoires in various disease contexts.

600 Our results highlight the potential of integrative analysis with large cross-samples
601 datasets to gain deeper insights into T cell biology. Additionally, our reference atlas
602 and TCR-DeepInsight tool offer versatile resources for immunologists to identify
603 potential functional TCRs from single-cell immune profiling datasets. Most
604 importantly, the dTCR clusters identified in this study, along with future potential
605 dTCRs characterized with single-cell immune profiling and TCR-DeepInsight, may
606 guide the design of personalized TCR-T (T cell receptor-engineered T cells) for
607 precise treatment of various diseases, including cancers and autoimmune disorders,
608 in the future.

609 We identified significant biases in the selection of VJ segment and amino acid usage
610 in the CDRs of the TCRs associated with different T cell types. A restricted VJ
611 segment usage was found for both TCR α and β chains in MAIT cells by virtue of
612 their evolutionarily conserved recognition of MHC-like molecule MR1.[53] We observed
613 significant enrichment for TRAV/TRAJ, TRBV/TRBJ joining, and TRAV/TRBV pairing
614 in CD8 T, CD4 T_{conv}, and MAIT cells, which is in line with previous findings.[36,37]
615 Additionally, a significant bias was observed in the usage of amino acids in the
616 CDR1 and CDR2 loops. One possible explanation of these observations is that the
617 TRAV/TRBV usage and the composition of amino acid usage after V(D)J
618 recombination predispose TCRs to interact with MHC I or MHC II molecules.[41] While
619 no obvious bias for VJ segments were detected in CD4 T_{reg}, we detected significant
620 differences for the amino acids usage in CD4 T_{reg} including the preference for
621 hydrophobic amino acids in CDR3β compared with CD4 T_{conv} as reported before.[54,55]
622 These findings revealed the intrinsic molecular features difference between MHC

class I and II recognition by TCR, and provide evidence supporting the plausibility of classifying CD4 or CD8 TCRs based on VJ gene usage and TCR amino acid sequence solely.

Given the immense diversity of TCRs, observing public TCRs with identical α and β chains is considered an extremely rare event. These TCRs are believed to share antigen specificity. Previous analysis on public TCRs have shown that TCR sequences are shared across multiple individuals, indicating a potential shared antigenic exposure.[56,57] We comprehensively analyzed public TCRs for both α and β chains, or single chain, and found that public TCR sequences are generally shorter in the middle region of CDR3 for both α and β chains and have a higher generation probability, which is consistent with TCRs in mice.[58] We also observed an enrichment of SARS-CoV-2-specific public TCRα/β pairs and identified public and clonally expanded TCRs with known antigen specificity for viruses such as EBV, CMV, and Influenza A. Meanwhile, we also identified numerous public TCRs shared in solid tumors and inflammation without known epitope specificity, which may imply their roles in recognizing tumor neoantigens or self-antigens. Coherent TCRα/β pairs are also considered to recognize similar antigens. By analyzing a comprehensive atlas of paired and full-length TCRα/β from different T cell subtypes, we found that TRAV coherence, or the tendency for TCRα chains to use the same TRAV gene when paired with the same TRBV-CDR3β, was generally higher in memory CD8 T cells than in other cell types. However, both the public TCR and coherence analysis focused on VJ gene usage and TCR sequence similarity without considering the potential cell transcriptome status, which may not be robust. Hence, developing new methods that can integrate both TCR and GEX data may help facilitate the identification of functional TCRs.

VAE has previously demonstrated its general applicability in scRNA-seq data feature extractions and batch effect removal across samples.[59] Meanwhile, the natural language processing large-language model BERT has also been adapted in representation learning for DNA, protein, and TCR sequences.[29,60,61] Deep-learning based framework allows representation learning for both transcriptome and TCR information. VAE based deep-learning framework in TCR-DeepInsight enables more precise representation learning for T cell transcriptome information, better identification of general features across samples and removal of batch effects, and

model transfer to new dataset. Meanwhile, BERT based deep-learning framework in TCR-DeepInsight obtained accurate TCR representation which allows clustering for CDR3α/β with variable length and amino acid mismatches. The BERT model alone in TCR-DeepInsight could also be applied to unsupervised clustering of TCRα/β or query similar TCRα/β with known epitope specificity when single-cell transcriptome information was not available. Compared with previous TCR clustering methods, our TCR-DeepInsight could integrate large-scale, multimodal, and heterogeneous single-cell transcriptome and full-length paired TCRα/β chains information all at once with relatively high scalability. Most importantly, with the reference dataset collected in this study, the scoring systems embedded in TCR-DeepInsight could overlay additional biological intuitions for TCR clusters identified. Along with additional information from HLA genotype and clonal expansion levels, our TCR-DeepInsight would greatly enhance the biological meaningful information that could be extracted from single-cell immune profiling dataset, and thus shed light on the understanding of T cell biology as well as future precise immunotherapy.

Limitations of the study

Our TCRα/β reference atlas, which includes data from 10^6 T cells derived from 291 individuals, provides a valuable resource for understanding TCR diversity and recurrence, although this number is just a fraction of the theoretical estimation of TCR diversity that could be as high as 10^{20}. Increasing the size of the dataset in the future may significantly enhance our understanding, particularly in terms of TCR recurrence. Furthermore, our current reference atlas focuses primarily on human T cells, and incorporating data from mouse models may further advance our understanding of TCR biology. Meanwhile, our TCR-DeepInsight models currently concatenate GEX and TCR embedding with a default weight hyperparameter, but do not consider their potential underlying interaction. Developing an end-to-end model in the future that better integrates GEX and TCR representation could allow for adaptive learning of model hyperparameters. Additionally, incorporating TCR structural information as model input may further enhance model performance for downstream tasks.

METHODS

688 **Integrating large-scale immune profiling datasets.**

689 We have previously described a preprocessing pipeline of single-cell immune
690 profiling datasets with linked single-cell RNA seq library (GEX library) and single-cell
691 TCR-seq library (TCR library).[32] In brief, cellranger (version 6.1.2) was used to obtain
692 the gene expression count matrix (by cellranger count command) and TCR contig
693 annotations (by cellranger vdj command) using raw sequencing reads as input. We
694 include high-confidence T cells (hcT cells) defined as T cells passing filtering criteria
695 on the number of captured genes and the percentage of mitochondrial genes-derived
696 counts, and with full-length TCR sequences with VJ segment annotation and CDR3
697 sequence in both α and β chains from our previous collection. We further expanded
698 our datasets to the level of millions of high-confidence T cells by appending 249
699 biological samples. We annotated each sample with the individual ID, disease
700 condition, and tissue origin. We used arcasHLA (version 0.5.0, IMGT reference
701 version 3.46.0) to extract the HLA genotype of each individual using the aligned BAM
702 files of the GEX library. After merging biological samples from the same individual,
703 we annotated each individual with genotypes of HLA-A, HLA-B, HLA-C, HLA-DPB1,
704 HLA-DRB1, and HLA-DQB1.

705 We integrated the transcriptome features of the hcT cells including both CD4 and
706 CD8 T cells by learning the batch-corrected latent embedding using VAE (See VAE
707 part of the methods). We annotated the hcT cells with CD4 or CD8 by gene count of
708 *CD4*, *CD8A*, and *CD8B*, followed by the nearest neighbour classifier
709 (n_neighbors=13) to categorise cells with double-positive or double-negative *CD4* or
710 *CD8* expression into CD4 or CD8 T cells. The CD4 T cells were further categorized
711 into naïve CD4 T cells, conventional memory CD4 T cells, and T regulatory cells by
712 key marker genes including *SELL* (Selenoprotein L), *TCF7* (Transcription Factor 7),
713 and *FOXP3* (Forkhead Box P3). The CD8 T cells were further divided into naïve CD8
714 T cells and memory or effector CD8 T cells based on the expression of *SELL* and
715 *TCF7.* We merged the T cells with exactly the same α and β chains defined by
716 TRAV-CDR3α-TRAJ and TRBV-CDR3β-TRBJ in each individual and obtained
717 635,393 unique TCRα/β sequences. Each unique TCRα/β was annotated with T cell
718 subtypes by their deriving cells. The transcriptome feature of each T cell was
719 visualized by projecting the learned latent embedding into 2-dimensional space using
720 the UMAP algorithm (**McInnes et al., 2018**).[62]

721 **Identifying enriched VJ segment in T cell subtype**

722 We calculated the odds ratio and *p*-value using Fisher's exact test to discover T cell
723 subtype-specific VJ joining in α and β chains and TRAV-TRBV combination. The
724 odds ratio of a given VJ combination (C_{VJ}) in a set of T cells of same subtype (T_{type})
725 against all other T cells (T_{others}) is given by:

726 $$\mathrm{OR} = \frac{|C^{+}_{VJ} \in T_{type}| * |C^{-}_{VJ}| \in T_{others}}{|C^{-}_{VJ} \in T_{type}| * |C^{+}_{VJ}| \in T_{others}}$$

727 The calculation of *p*-value followed by multiple testing correction by Bonferroni
728 method was achieved by the SciPy library (www.scipy.org). We reported VJ
729 combinations preferentially selected by certain T cell type by thresholding the
730 adjusted *p*-value < 0.05, odds ratio > 2, and the number supporting cells greater than
731 400. The selected cell type enriched VJ combination are visualized by Sankey plot
732 (https://d3-graph-gallery.com/sankey.html) using number of cells of edges of each VJ
733 combination and density plot in the 2-dimensional embedding of the transcriptome
734 feature of T cells using the *kdeplot* function from the seaborn
735 (https://seaborn.pydata.org/) library.

736 **Analysis of amino acid usage in the middle region of CDR3 sequence**

737 The middle region of the CDR3 sequence in α or β chain (CDR3α-mr and CDR3β-
738 mr) was defined as amino acid encoded by random nucleotide insertions between
739 the V and J segments. After removing the amino acids derived from V and J
740 segment of the CDR3 sequence, the remaining amino acids were annotated with the
741 CDR3-mr sequence. The percentage of amino acid usage grouped by their chemical
742 properties was calculated for each TCR, and then averaged across different T cell
743 subtypes (CD8 T, CD4 T_{conv} and CD4 T_{reg}). The significance of the difference in the
744 percentage between different T cell subtypes was calculated by paired two-tailed
745 Student's t-test.

746 **Building reference of paired TCRα/β with CDR3 amino acid sequence and**
747 **epitope specificity**

748 We collected datasets with experimentally validated TCR-pMHC binding pairs. We
749 included dataset using pMHC-tetramer cell sorting together with single-cell paired
750 TCRα/β amplification.[22] Paired TCRα/β with CDR3 amino acid sequences were
751 obtained from curated databases including McPAS-TCR[45], VDJdb[63,64] and TCRdb[65].

752 We also include recently released datasets using combined DNA-barcoded pMHC
753 tetramer/dextramer and single-cell RNA-sequencing
754 (https://www.10xgenomics.com/cn/resources/datasets/cd-8-plus-t-cells-of-healthy-
755 donor-1-1-standard-3-0-2, https://www.10xgenomics.com/cn/resources/datasets/cd-
756 8-plus-t-cells-of-healthy-donor-2-1-standard-3-0-2,
757 https://www.10xgenomics.com/cn/resources/datasets/cd-8-plus-t-cells-of-healthy-
758 donor-3-1-standard-3-0-2, https://www.10xgenomics.com/cn/resources/datasets/cd-
759 8-plus-t-cells-of-healthy-donor-4-1-standard-3-0-2).[46,47] We merged the curated
760 datasets described in deepTCR.[27] After removing repetitive data records and data
761 curation, we obtained 28,641 TCR-pMHC pairs with paired CDR3α and CDR3β
762 amino acid sequences and 1037 peptide epitopes.

763 **Definition and analysis of public TCRs**

764 Public TCRs were defined independently by either the single α or β chain or paired
765 α/β chains. Public TCRα/βs were TCRs that occur in at least two individuals who
766 share the exact same TRBV, CDR3β amino acid sequence, TRBJ, TRAV, CDR3α
767 amino acid sequence, and TRAJ, while public TCRα and TCRβ only requires exact
768 same VJ gene and CDR3 amino acid sequence of single α or β chain. The V and J
769 segment call adopted the annotation from the output of cellranger. Disease-specific
770 public TCRs were defined as TCRs occurring in at least two individuals with the
771 same disease type.

772 We calculated the generation probability of α and β chains for each public TCRs and
773 non-public TCRs by OLGA (version 1.2.4[44]). The generation probability models for
774 the α and β chains initialized by *GenerationProbabilityVJ* and
775 *GenerationProbabilityVDJ* function accepting both CDR3 amino acid sequence and
776 VJ gene as input. The generation probability was followed by negative log-
777 transformation to get the $\log P_{gen}$ of each TCR. The significance of the difference in
778 $\log P_{gen}$ between public and non-public TCRs was calculated by paired two-tailed
779 Student's t-test.

780 **Analysis of α and β chain coherence of TCRs**

781 The probability of independent TCRs (non-clonally expanded) of the same TRBV
782 and CDR3β amino acid sequence that used the same TRAV gene was defined as
783 the TRAV coherence of TCRs. In contrast, for independent TCRs of the same TRAV

784 and CDR3α amino acid sequence that use the same TRBV gene were defined as
785 the TRBV coherence. The TRAV and TRBV coherence were calculated separately
786 for the CD4 T_{reg}, CD4 T_n, CD8 T_n, CD4 T_{conv}, and CD8 T memory cells separately.
787 The definition of TRAV and TRBV coherence were also expanded to the case where
788 the mismatch of CDR3 amino acid sequence was measured by the Levenshtein
789 distance. The extended TRAV coherence was measured by grouping TCRβ chains
790 with the same TRBV and mismatched CDR3β and calculating the percentage of the
791 same TRAV in the corresponding TCRα chains. The α and β chain coherence were
792 also measured by CDR3 sequence similarity, defined as the mean Levenshtein
793 distance in the CDR3 sequence of one chain given that the V and CDR3 sequence
794 was the same in the other chain.

795 **Training VAE model for universal representation of transcriptome features**

796 We implemented a VAE model to capture the variation in gene expression of the
797 1,017,877 T cells. The probabilistic variational autoencoder accepted raw count
798 matrices as input. The encoder received the gene expression raw count matrix and
799 outputs a scalar mean (μ) and variance (σ^2), which reparameterized and
800 approximated the latent variational distribution (z). The latent distribution was
801 regularized using the Kullback-Leibler divergence with a standard normal distribution
802 prior (N(0,1)). To reflect biological variance and eliminate batch effects, the decoder
803 combined the categorical encoding of the sample name, total library size by the sum
804 of gene expression counts, and the latent embedding of the encoder output. The
805 decoder outputed the mean, variance, and the gene-dropout probability to
806 parameterize the zero-inflated negative-binomial distribution (ZINB) and models the
807 raw count data.

808 We removed the mitochondrial genes (gene name starts with 'MT-'), ribosomal
809 genes (gene name starts with 'RPS' or 'RPL') and heat-shock proteins (gene name
810 starts with 'HSP-'). The top 6000 highly variable genes were selected using scanpy's
811 `pp.highly_variable_genes(flavor='seurat_v3',`
812 `batch_key='sample_name', n_top_genes=6000)` before VAE
813 integration. The VAE model was trained using the AdamW optimizer with a learning
814 rate of 0.0001 and a batch size of 128.

815 **Training BERT model for universal representation of TCRα/β sequences**

816 We adopted the BERT model originated in the field of natural language processing to
817 obtain a scalable representation of the TCR sequence. The transformer-based
818 models were shown to outperform conventional encoding model including MLP,
819 convolutional networks, and recurrent models, using attention mechanism to capture
820 inter-relationship between tokens. The BERT model was implemented in Python
821 using the PyTorch framework and Hugginface's Transformer libraries, with
822 hyperparameters adapted from the default settings (hidden dimensionality = 768,
823 intermediate size = 1536, number of attention heads = 12, number of hidden layers =
824 12). We implemented a tokenizer combining both CDR3α and CDR3β amino acid
825 sequence, adding a classification token (CLS) ahead of CDR3α, gap tokens (GAP)
826 between the two CDR3 sequence, and padding tokens (PAD) after CDR3β to align
827 CDR3α and CDR3β sequence to the same encoding length. During pre-training, we
828 retrieved CDR3α and CDR3β amino acid sequence from both our dataset and the
829 curated datasets with pMHC binding information. We filtered TCRs with abnormally
830 long CDR3α or CDR3β amino acid sequence (>36 aa), or with residues not included
831 in the set of 20 standard amino acids, and finally retained 663,379 TCR sequence for
832 pretraining BERT. We trained the model to predict the masked amino acids after
833 pretraining BERT using a masked language model (MLM) objective in which we
834 randomly hide 15% of amino acids in either CDR3α or CDR3β. We also masked
835 either TRAV or TRAJ gene and require the model to impute the V gene usage given
836 another chain. The MLM objective was to help the model to learn the underlying
837 pairing pattern of TRAV/CDR3α and TRBV/CDR3β. The BERT model was pre-
838 trained using the AdamW optimizer (a first-order stochastic optimizer) with a learning
839 rate of 0.00001 and a batch size of 32.

840 The embeddings of the token from the TRAV, TRBV, CDR3β and TRBJ was used to
841 represent the whole TCRα/β after mean pooling operation. The 768-dimensional
842 embedding BERT output could be projected into 64-dimensional space via PCA by
843 using the implementation of scikit-learn. The 64-dimensional representation of the
844 TCR was then called the TCR embedding. We projected the TCR embeddings of the
845 634,252 TCRs to the 2-dimensional UMAP space via the UMAP algorithm using the
846 implementation of umap-learn[62] and visualized the embedding performance of the
847 pretrained model.

848 **Unsupervised TCRα/β clustering with disease-association and shared**

transcriptome state

We aggregated the GEX embedding from the VAE model for each unique TCRα/β. We concatenated the TCR embeddings and the aggregated GEX embeddings and obtained a TCR-GEX joint representation of TCRα/β pairs. We adopted faiss-gpu, a computational framework for rapid similarity search optimized and accelerated by GPU[66], for indexed *k*-nearest neighbor search of TCR-GEX joint representation based on Euclidean distance.

We use the following strategy to cluster disease-specific TCRs with transcriptome similarity. First, for each TCRα/β in our datasets, we use the TCRα/β as an anchor and find the *k*-nearest neighbor (k=40 in default). We select the top n TCRα/β pairs ranked by the Euclidean distance with the same disease, and group the anchor TCRα/β and the neighbor TCRα/βs as a potential disease-associated cluster. In the following neighbor search, we removed the possibility of the neighbor TCRα/β to prevent repeated neighbor TCRα/βs clusters.

We defined the distance between TCRs in our TCR-GEX joint representation space as the Euclidean distance between the representation of TCRs.

$$d(s_i, s_j) = \mathrm{EuclideanDistance}(s_i, s_j)$$

We defined a similarity score to measure the sequence similarity of TCRα/β within a disease-associated cluster. The similarity score was defined as the mean Euclidean distance in the TCR-GEX joint representation space between all neighboring TCRα/β (s_i) and the anchor TCR (s_{anchor}):

$$D_{\mathrm{Similarity}} = \frac{\sum_{i=1}^{n-1} d(s_i, s_{\mathrm{anchor}})}{n-1}$$

, where *n* is the number of TCRs in the cluster and s_{anchor} is the anchor TCR in the cluster.

We also calculated the mean pairwise Levenshtein distance of either CDR3α or CDR3β within a disease-associated cluster by:

$$D_{\mathrm{Similarity-Levenshtein}} = \frac{\sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} \mathrm{lev}(s_i, s_j)}{n(n-1)}$$

, where *lev* is the Levenshtein distance of two sequences.

We defined a disease-specificity score to measure the distinctiveness of TCRα/β in a disease-associated cluster. The score was computed based on the disparity between the mean Euclidean distance in the TCR-GEX joint representation space. Specifically, the score was calculated by comparing (1) the mean Euclidean distance between the anchor TCRα/β and all neighboring TCRα/β in the cluster, and (2) the mean Euclidean distance between the top-ranked TCRs that are most similar to the anchor TCRα/β, but derived from different diseases. The latter was determined with an equal number of neighbors of the anchor TCRα/β.

$$D_{\text{Specificity}} = \frac{\sum_{k=1}^{m} d(s_k, s_{\text{anchor}}) - \sum_{i=1}^{n-1} d(s_i, s_{\text{anchor}}))}{n-1}$$

, where *m* equals to *n-1*, which is the number of neighbors of the anchor TCRα/β. By using this score, we could better evaluate the uniqueness of TCRα/β in a disease-associated cluster. The proposed score enables a more accurate evaluation of the uniqueness of TCRα/β in a disease-associated cluster. Furthermore, the study also calculated the difference between the mean Levenshtein distance to the anchor TCR of either the CDR3α or the CDR3β from the disease-associated cluster and the top-ranked TCRs.

$$D_{\text{Specificity-Levenstein}} = \frac{\sum \text{lev}(s_j, s_{\text{anchor}}) - \sum \text{lev}(s_i, s_{\text{anchor}}))}{n-1}$$

The correlation between the TCR similarity score and disease specificity and the Levenshtein distance is visualized by nonparametric loess regression.

Selecting top-ranked disease-associated TCRα/β clusters with shared transcriptome features

The selection of disease-associated TCRα/β clusters involved the use of previously established similarity and disease-specificity scores. Specifically, TCRα/β clusters with a combined score (square of similarity score plus square of disease-specificity score) less than 300^2 are selected for further analysis. Additional criteria are applied based on the number of unique individuals and unique TCRα/β in a TCRα/β cluster, which varied depending on the disease type. For disease types with a relatively larger number of individuals, such as solid tumors, COVID-19, and colitis, TCRα/β clusters with more than 2 unique individuals and more than 3 unique TCRα/βs were selected. For diseases with smaller individual numbers and derived from distinct

907 studies, such as ankylosing spondylitis, TCRα/β clusters with more than 1 unique
908 individual and more than 2 unique TCRα/βs were selected. To augment the selected
909 TCRα/β, additional information was included, such as the common HLA genotypes
910 of the individuals from which they were derived and the level of clonal expansion.
911 These information served to provide an additional layer of potentially disease-related
912 features to the TCRα/β clusters.

913 **Querying TCRs with known epitope specificity in the TCR atlas**

914 We adopted the paired TCRα/β with VJ gene usage information and CDR3 amino
915 acid sequences with known epitope specificity to search for similar TCRs in the
916 reference atlas. We used the pre-trained BERT model to transfer TCR embeddings
917 from the reference atlas to the query TCRs, which were then subjected to PCA
918 transformation using the same weights as the reference atlas. We concatenated the
919 resulting TCR embeddings from the query and reference atlas and used the query
920 TCRs as anchors in a kNN search. We ranked the TCRs in the reference atlas by
921 their distance to the anchor TCR for each query TCR, with distance cutoffs of 100
922 and 186 used for TCRs in **Figures7B** and **7C**, respectively.

923 **Query of new paired scRNA-seq and scTCR-seq samples against existing**
924 **reference**

925 We implemented an architecture surgery strategy to apply transfer learning in the
926 VAE model. Specifically, during data transfer, the weights of encoder and decoder
927 were kept the same as the VAE model before transfer learning, and the original set
928 of batch indexes would be extended to allow extended datasets. We adopted a
929 recently published scRNA-seq and scTCR-seq dataset including 59,417 T cells
930 derived from 10 individuals with gastric cancer as example.[52] As the data processing
931 pipeline was the same as our reference atlas, we kept the same 6000 HVGs as in
932 our previous analysis to perform transfer learning. We used a k-Nearest classifier
933 (`n_neighbors=13`) to transfer the original cell type labels including CD8 T_n, CD8 T,
934 CD4 T_n, CD4 T_{conv} and CD4 T_{reg}, and MAIT to the query dataset.

935 We again aggregated the transferred GEX embedding by unique TCRs and
936 concatenated it to the TCR embedding from the BERT model followed by PCA
937 transformation with the same weight as the reference atlas. Categorizing Gastric
938 Cancer associated TCRα/β clusters follows the same strategy described above, and

939 TCRα/β clusters with more than 1 unique individual and more than 2 unique
940 TCRα/βs were selected.

941 **Code and dataset availability**

942 The reference atlas datasets utilized for analysis can be accessed on Figshare
943 (excluding the two controlled access data) at
944 https://figshare.com/collections/_/6471289. The code for TCR-DeepInsight and data
945 analysis is available on GitHub at https://github.com/WanluLiuLab/TCR-DeepInsight,
946 along with a Jupyter Notebook at https://huarc.net/notebook.

947 **ACKNOWLEDGEMENTS**

948 We thank all the researchers who generated the single-cell immune profiling
949 datasets used in this study. We thank Dr. Hussein A. Abbas for kindly sharing the
950 raw data from AML patients generated in their previous publication
951 (EGAD00001007672/EGAD0001007674).[67] We thank Dr. Zheng Wang for kindly
952 sharing the raw data from Kawasaki disease generated in their previous
953 publication.[18] We thank Dr. Hongbo Hu from Sichuan University, Dr. Feng Wang
954 from Shanghai Jiaotong University, Dr. Chaochen Wang from Zhejiang University
955 and all lab members from the Liu lab at ZJU-UoE Institute for their helpful discussion.
956 We would also like to thank the technical support provided by the Core Facilities,
957 especially the ZJE server of ZJU-UoE Institute. This work is supported by National
958 Natural Science Foundation of China 31930038 (to L.L.), 32100718 (to L.L.),
959 U21A20199 (to L.L.), 32170551 (to W.L.), the Fundamental Research Funds for the
960 Central Universities 226-2022-00134 (to W.L.), the Tencent AI Lab Rhino Bird
961 Research Funding RBFR2022015 (to W.L.) and Pre-research Projects of Innovation
962 Center of Yangtze River Delta, Zhejiang University (to L.W., W.L., L.L.).
963

964 **AUTHOR CONTRIBUTIONS**

965 W.L., L.L., J.Y., L.W., Z.X., and L.W., conceived the study and designed
966 experiments. Z.W., W.L., and L.L., wrote the manuscript. Z.X., L.W., Y.B., D.S., Y.G.,
967 and P.C. processed the data. Z.X., L.W., R.T., performed the bioinformatics analysis.
968 Z.X., Z.L., Y.Z., B.H., J.Y. and W.L., conceived the deep-learning framework. Z.W.
969 implemented the deep-learning framework. All authors contributed to the review and
970 corrections of the manuscripts.

971

972 **DECLARATION OF INTERESTS**

973 The authors declare no competing financial interests.

974

975 **REFERENCES**

976 1. Zarnitsyna, V.I., Evavold, B.D., Schoettle, L.N., Blattman, J.N., and Antia, R. (2013).
977 Estimating the Diversity, Completeness, and Cross-Reactivity of the T Cell Repertoire.
978 Front. Immunol. *4*. 10.3389/fimmu.2013.00485.

979 2. Mora, T., and Walczak, A.M. (2019). How many different clonotypes do immune
980 repertoires contain? Curr. Opin. Syst. Biol. *18*, 104–110. 10.1016/j.coisb.2019.10.001.

981 3. Rossjohn, J., Gras, S., Miles, J.J., Turner, S.J., Godfrey, D.I., and McCluskey, J. (2015).
982 T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. Annu. Rev.
983 Immunol. *33*, 169–200. 10.1146/annurev-immunol-032414-112334.

984 4. Wagner, U., Pierer, M., Kaltenhäuser, S., Wilke, B., Seidel, W., Arnold, S., and
985 Häntzschel, H. (2003). Clonally expanded CD4+CD28null T cells in rheumatoid arthritis
986 use distinct combinations of T cell receptor BV and BJ elements. Eur. J. Immunol. *33*,
987 79–84. 10.1002/immu.200390010.

988 5. Matsumoto, Y., Yoon, W.K., Jee, Y., Fujihara, K., Misu, T., Sato, S., Nakashima, I., and
989 Itoyama, Y. (2003). Complementarity-Determining Region 3 Spectratyping Analysis of
990 the TCR Repertoire in Multiple Sclerosis. J. Immun. J. *170*, 4846–4853.
991 10.4049/jimmunol.170.9.4846.

992 6. Kou, Z.C., Puhr, J.S., Rojas, M., Mccormack, W.T., Goodenow, M.M., and Sleasman,
993 J.W. (2000). T-Cell Receptor V^{N}_{L} Repertoire CDR3 Length Diversity Differs within
994 CD45RA and CD45RO T-Cell Subsets in Healthy and Human Immunodeficiency Virus-
995 Infected Children. Clin. Diagn. Lab. Immunol. *7*. 10.1128/cdli.7.6.953-959.2000

996 7. Derfuss, T., Arbusow, V., Strupp, M., Brandt, T., and Theil, D. (2009). The Presence of
997 Lytic HSV-1 Transcripts and Clonally Expanded T Cells with a Memory Effector
998 Phenotype in Human Sensory Ganglia. Ann. N. Y. Acad. Sci. *1164*, 300–304.
999 10.1111/j.1749-6632.2009.03871.x.

1000 8. Reuben, A., Zhang, J., Chiou, S.-H., Gittelman, R.M., Li, J., Lee, W.-C., Fujimoto, J.,
1001 Behrens, C., Liu, X., Wang, F., et al. (2020). Comprehensive T cell repertoire
1002 characterization of non-small cell lung cancer. Nat. Commun. *11*, 603. 10.1038/s41467-
1003 019-14273-0.

1004 9. Musters, A., Klarenbeek, P.L., Doorenspleet, M.E., Balzaretti, G., Esveldt, R.E.E., van
1005 Schaik, B.D.C., Jongejan, A., Tas, S.W., van Kampen, A.H.C., Baas, F., et al. (2018). In
1006 Rheumatoid Arthritis, Synovitis at Different Inflammatory Sites Is Dominated by Shared
1007 but Patient-Specific T Cell Clones. J. Immun. J. *201*, 417–422.
1008 10.4049/jimmunol.1800421.

1009 10. Ye, X., Wang, Z., Ye, Q., Zhang, J., Huang, P., Song, J., Li, Y., Zhang, H., Song, F.,
1010 Xuan, Z., et al. (2020). High-Throughput Sequencing-Based Analysis of T Cell
1011 Repertoire in Lupus Nephritis. Front. Immunol. *11*, 1618. 10.3389/fimmu.2020.01618.

1012 11. Wang, G., Wang, Y., Jiang, S., Fan, W., Mo, C., Gong, W., Chen, H., He, D., Huang, J.,
1013 Ou, M., et al. (2022). Comprehensive analysis of TCR repertoire of COVID-19 patients in
1014 different infected stage. Genes. Genom. *44*, 813–822. 10.1007/s13258-022-01261-w.

1015 12. Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang,
1016 F., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with
1017 COVID-19. Nat. Med. *26*, 842–844. 10.1038/s41591-020-0901-9.

1018 13. Bacher, P., Rosati, E., Esser, D., Martini, G.R., Saggau, C., Schiminsky, E.,
1019 Dargvainiene, J., Schröder, I., Wieters, I., Khodamoradi, Y., et al. (2020). Low-Avidity
1020 CD4+ T Cell Responses to SARS-CoV-2 in Unexposed Individuals and Humans with
1021 Severe COVID-19. Immunity *53*, 1258-1271.e5. 10.1016/j.immuni.2020.11.016.

1022 14. Cheon, I.S., Li, C., Son, Y.M., Goplen, N.P., Wu, Y., Cassmann, T., Wang, Z., Wei, X.,
1023 Tang, J., Li, Y., et al. (2021). Immune signatures underlying post-acute COVID-19 lung
1024 sequelae. Sci. Immunol. *6*, eabk1741. 10.1126/sciimmunol.abk1741.

1025 15. Notarbartolo, S., Ranzani, V., Bandera, A., Gruarin, P., Bevilacqua, V., Putignano, A.R.,
1026 Gobbini, A., Galeota, E., Manara, C., Bombaci, M., et al. (2021). Integrated longitudinal
1027 immunophenotypic, transcriptional, and repertoire analysis delineate immune responses
1028 in patients with COVID-19. Sci. Immunol. *6*, eabg5021. 10.1126/sciimmunol.abg5021.

1029 16. Wang, Z., Zhang, Y., Yang, R., Wang, Y., Guo, J., Sun, R., Zhou, Y., Su, L., Ge, Q., and
1030 Feng, Y. (2022). Landscape of Peripheral Blood Mononuclear Cells and Soluble Factors
1031 in Severe COVID-19 Patients With Pulmonary Fibrosis Development. Front. Immunol.
1032 *13*, 831194. 10.3389/fimmu.2022.831194.

1033 17. Wen, W., Su, W., Tang, H., Le, W., Zhang, X., Zheng, Y., Liu, X., Xie, L., Li, J., Ye, J., et
1034 al. (2020). Immune cell profiling of COVID-19 patients in the recovery stageby single-cell
1035 sequencing. Cell. Discov. *6*, 31. 10.1038/s41421-020-0168-9.

1036 18. Wang, Z., Xie, L., Ding, G., Song, S., Chen, L., Li, G., Xia, M., Han, D., Zheng, Y., Liu,
1037 J., et al. (2021). Single-cell RNA sequencing of peripheral blood mononuclear cells from
1038 acute Kawasaki disease patients. Nat. Commun. *12*, 5444. 10.1038/s41467-021-25771-
1039 5.

1040 19. Simone, D., Penkava, F., Ridley, A., Sansom, S., Al-Mossawi, M.H., and Bowness, P.
1041 (2021). Single cell analysis of spondyloarthritis regulatory T cells identifies distinct
1042 synovial gene expression patterns and clonal fates. Commun. Biol. *4*, 1395.
1043 10.1038/s42003-021-02931-3.

1044 20. Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A.,
1045 Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor
1046 repertoire. Nature *547*, 94–98. 10.1038/nature22976.

1047 21. Huang, H., Wang, C., Rubelt, F., Scriba, T.J., and Davis, M.M. (2020). Analyzing the
1048 Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2
1049 and genome-wide antigen screening. Nat. Biotechnol. *38*, 1194–1202. 10.1038/s41587-
1050 020-0505-4.

1051 22. Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A.,
1052 Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., et al. (2017).
1053 Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature
1054 *547*, 89–93. 10.1038/nature22383.

1055 23. Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J.C., Souquette, A.,
1056 Gaevert, J.A., Hertz, T., Thomas, P.G., Bradley, P., and Fiore-Gartland, A. (2021). TCR
1057 meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public,
1058 HLA-restricted clusters of SARS-CoV-2 TCRs. eLife *10*, e68605. 10.7554/eLife.68605.

1059 24. Zhang, H., Liu, L., Zhang, J., Chen, J., Ye, J., Shukla, S., Qiao, J., Zhan, X., Chen, H.,
1060 Wu, C.J., et al. (2020). Investigation of Antigen-Specific T-Cell Receptor Clusters in
1061 Human Cancers. Clin. Cancer Res. *26*, 1359–1371. 10.1158/1078-0432.CCR-19-3249.

1062 25. Zhang, H., Zhan, X., and Li, B. (2021). GIANA allows computationally-efficient TCR
1063 clustering and multi-disease repertoire classification by isometric transformation. Nat.
1064 Commun. *12*, 4699. 10.1038/s41467-021-25006-7.

1065 26. Valkiers, S., Van Houcke, M., Laukens, K., and Meysman, P. (2021). ClusTCR: a python
1066 interface for rapid clustering of large sets of CDR3 sequences with unknown antigen
1067 specificity. Bioinformatics *37*, 4865–4867. 10.1093/bioinformatics/btab446.

1068 27. Sidhom, J.-W., Larman, H.B., Pardoll, D.M., and Baras, A.S. (2021). DeepTCR is a deep
1069 learning framework for revealing sequence concepts within T-cell repertoires. Nat.
1070 Commun. *12*, 1605. 10.1038/s41467-021-21879-w.

1071 28. Montemurro, A., Schuster, V., Povlsen, H.R., Bentzen, A.K., Jurtz, V., Chronister, W.D.,
1072 Crinklaw, A., Hadrup, S.R., Winther, O., Peters, B., et al. (2021). NetTCR-2.0 enables
1073 accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data.
1074 Commun. Biol. *4*, 1060. 10.1038/s42003-021-02610-3.

1075 29. Wu, K., Yost, K.E., Daniel, B., Belk, J.A., Xia, Y., Egawa, T., Satpathy, A., Chang, H.Y.,
1076 and Zou, J. (2021). TCR-BERT: learning the grammar of T-cell receptors for flexible
1077 antigen-xbinding analyses. bioRxiv 2021.11.18.469186. 10.1101/2021.11.18.469186.

1078 30. Schattgen, S.A., Guion, K., Crawford, J.C., Souquette, A., Barrio, A.M., Stubbington,
1079 M.J.T., Thomas, P.G., and Bradley, P. (2022). Integrating T cell receptor sequences and
1080 transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat. Biotechnol.
1081 *40*, 54–63. 10.1038/s41587-021-00989-2.

1082 31. Zhang, Z., Xiong, D., Wang, X., Liu, H., and Wang, T. (2021). Mapping the functional
1083 landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat. Methods
1084 *18*, 92–99. 10.1038/s41592-020-01020-3.

1085 32. Wu, L., Xue, Z., Jin, S., Zhang, J., Guo, Y., Bai, Y., Jin, X., Wang, C., Wang, L., Liu, Z.,
1086 et al. (2022). huARdb: human Antigen Receptor database for interactive clonotype-
1087 transcriptome analysis at the single-cell level. Nucleic Acids Res. *50*, D1244–D1254.
1088 10.1093/nar/gkab857.

1089 33. Orenbuch, R., Filip, I., Comito, D., Shaman, J., Pe'er, I., and Rabadan, R. (2020).
1090 arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics *36*, 33–40.
1091 10.1093/bioinformatics/btz474.

1092 34. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P.,
1093 Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell

1094 sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.
1095 10.1038/s41590-018-0276-y.

1096 35. Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., Burdin, N., Visan, L.,
1097 Ceccarelli, M., Poidinger, M., et al. (2019). RNA-Seq Signatures Normalized by mRNA
1098 Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. *26*,
1099 1627-1640.e7. 10.1016/j.celrep.2019.01.041.

1100 36. Klarenbeek, P.L., Doorenspleet, M.E., Esveldt, R.E.E., van Schaik, B.D.C., Lardy, N.,
1101 van Kampen, A.H.C., Tak, P.P., Plenge, R.M., Baas, F., de Bakker, P.I.W., et al. (2015).
1102 Somatic Variation of T-Cell Receptor Genes Strongly Associate with HLA Class
1103 Restriction. PLoS ONE *10*, e0140815. 10.1371/journal.pone.0140815.

1104 37. Carter, J.A., Preall, J.B., Grigaityte, K., Goldfless, S.J., Jeffery, E., Briggs, A.W.,
1105 Vigneault, F., and Atwal, G.S. (2019). Single T Cell Sequencing Demonstrates the
1106 Functional Role of αβ TCR Pairing in Cell Lineage and Antigen Specificity. Front.
1107 Immunol. *10*, 1516. 10.3389/fimmu.2019.01516.

1108 38. Reantragoon, R., Corbett, A.J., Sakala, I.G., Gherardin, N.A., Furness, J.B., Chen, Z.,
1109 Eckle, S.B.G., Uldrich, A.P., Birkinshaw, R.W., Patel, O., et al. (2013). Antigen-loaded
1110 MR1 tetramers define T cell receptor heterogeneity in mucosal-associated invariant T
1111 cells. J. Exp. Med. *210*, 2305–2320. 10.1084/jem.20130958.

1112 39. Tilloy, F., Treiner, E., Park, S.-H., Garcia, C., Lemonnier, F., Bendelac, A., Bonneville,
1113 M., and Lantz, O. (1999) An invariant T cell receptor alpha chain defines a novel TAP-
1114 independent major histocompatibility complex class Ib-restricted alpha/beta T cell
1115 subpopulation in mammals. J. Exp. Med. *21*, 1907-21. 10.1084/jem.189.12.1907.

1116 40. Gao, S., Wu, Z., Arnold, B., Diamond, C., Batchu, S., Giudice, V., Alemu, L., Raffo, D.Q.,
1117 Feng, X., Kajigaya, S., et al. (2022). Single-cell RNA sequencing coupled to TCR
1118 profiling of large granular lymphocyte leukemia T cells. Nat. Commun. *13*, 1982.
1119 10.1038/s41467-022-29175-x.

1120 41. Marrack, P., Scott-Browne, J.P., Dai, S., Gapin, L., and Kappler, J.W. (2008).
1121 Evolutionarily Conserved Amino Acids That Control TCR-MHC Interaction. Annu. Rev.
1122 Immunol. *26*, 171–203. 10.1146/annurev.immunol.26.021607.090421.

1123 42. Jaffe, D.B., Shahi, P., Adams, B.A., Chrisman, A.M., Finnegan, P.M., Raman, N., Royall,
1124 A.E., Tsai, F., Vollbrecht, T., Reyes, D.S., et al. (2022). Functional antibodies exhibit
1125 light chain coherence. Nature *611*, 352–357. 10.1038/s41586-022-05371-z.

1126 43. Venturi, V., Price, D.A., Douek, D.C., and Davenport, M.P. (2008). The molecular basis
1127 for public T-cell responses? Nat. Rev. Immunol. *8*, 231–238. 10.1038/nri2260.

1128 44. Sethna, Z., Elhanati, Y., Callan, C.G., Walczak, A.M., and Mora, T. (2019). OLGA: fast
1129 computation of generation probabilities of B- and T-cell receptor amino acid sequences
1130 and motifs. Bioinformatics *35*, 2974–2981. 10.1093/bioinformatics/btz035.

1131 45. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). McPAS-TCR:
1132 a manually curated catalogue of pathology-associated T cell receptor sequences.
1133 Bioinformatics *33*, 2924–2929. 10.1093/bioinformatics/btx286.

1134 46. Minervina, A.A., Pogorelyy, M.V., Kirk, A.M., Crawford, J.C., Allen, E.K., Chou, C.-H.,
1135 Mettelman, R.C., Allison, K.J., Lin, C.-Y., Brice, D.C., et al. (2022). SARS-CoV-2 antigen

1136 exposure history shapes phenotypes and specificity of memory CD8+ T cells. Nat.
1137 Immunol. *23*, 781–790. 10.1038/s41590-022-01184-4.

1138 47. Francis, J., Leistritz-Edwards, D., Dunn, A., Tarr, C., Lehman, J., Dempsey, C., Hamel,
1139 A., Rayon, V., Liu, G., Yuntong, W., et al. (2022). Allelic variation in class I HLA
1140 determines CD8+ T cell repertoire shape and cross-reactive memory responses to
1141 SARS-CoV-2. Sci. Immunol. 7, eabk3070, 10.1126/sciimmunol.abk3070.

1142 48. Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of T Cell Receptor Alpha
1143 and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front.
1144 Immunol. *12*, 664514. 10.3389/fimmu.2021.664514.

1145 49. Fischer, D.S., Wu, Y., Schubert, B., and Theis, F.J. (2020). Predicting antigen specificity
1146 of single T cells based on TCR CDR 3 regions. Mol. Syst. Biol. *16*.
1147 10.15252/msb.20199416.

1148 50. Hudson, D., Fernandes, R.A., Basham, M., Ogg, G., and Koohy, H. (2023). Can we
1149 predict T cell specificity with digital biology and machine learning? Nat. Rev. Immunol. *8*,
1150 1-11. 10.1038/s41577-023-00835-3.

1151 51. Yang, X., Garner, L.I., Zvyagin, I.V., Paley, M.A., Komech, E.A., Jude, K.M., Zhao, X.,
1152 Fernandes, R.A., Hassman, L.M., Paley, G.L., et al. (2022). Autoimmunity-associated T
1153 cell receptors recognize HLA-B*27-bound peptides. Nature *612*, 771–777.
1154 10.1038/s41586-022-05501-7.

1155 52. Sun, K., Xu, R., Ma, F., Yang, N., Li, Y., Sun, X., Jin, P., Kang, W., Jia, L., Xiong, J., et
1156 al. (2022). scRNA-seq of gastric tumor shows complex intercellular interaction with an
1157 alternative T cell exhaustion trajectory. Nat. Commun. *13*, 4943. 10.1038/s41467-022-
1158 32627-z.

1159 53. Treiner, E., Duban, L., Bahram, S., Radosavljevic, M., Wanner, V., Tilloy, F., Affaticati,
1160 P., Gilfillan, S., and Lantz, O. (2003). Selection of evolutionarily conserved mucosal-
1161 associated invariant T cells by MR1. Nature *422*, 164–169. 10.1038/nature01433.

1162 54. Stadinski, B.D., Shekhar, K., Gómez-Touriño, I., Jung, J., Sasaki, K., Sewell, A.K.,
1163 Peakman, M., Chakraborty, A.K., and Huseby, E.S. (2016). Hydrophobic CDR3 residues
1164 promote the development of self-reactive T cells. Nat. Immunol. *17*, 946–955.
1165 10.1038/ni.3491.

1166 55. Lagattuta, K.A., Kang, J.B., Nathan, A., Pauken, K.E., Jonsson, A.H., Rao, D.A., Sharpe,
1167 A.H., Ishigaki, K., and Raychaudhuri, S. (2022). Repertoire analyses reveal T cell
1168 antigen receptor sequence features that influence T cell fate. Nat. Immunol. *23*, 446–
1169 457. 10.1038/s41590-022-01129-x.

1170 56. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-
1171 Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of
1172 Cancer. Immunity *48*, 812-830.e14. 10.1016/j.immuni.2018.03.023.

1173 57. Huisman, W., Hageman, L., Leboux, D.A.T., Khmelevskaya, A., Efimov, G.A., Roex,
1174 M.C.J., Amsen, D., Falkenburg, J.H.F., and Jedema, I. (2021). Public T-Cell Receptors
1175 (TCRs) Revisited by Analysis of the Magnitude of Identical and Highly-Similar TCRs in
1176 Virus-Specific T-Cell Repertoires of Healthy Individuals. Front. Immunol. *24*, 851868,
1177 10.1101/2021.11.29.470325.

58. Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I.R., and Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. *24*, 1603–1612. 10.1101/gr.170753.113.

59. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. 10.1038/s41592-018-0229-2.

60. Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics *37*, 2112–2120. 10.1093/bioinformatics/btab083.

61. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110. 10.1093/bioinformatics/btac020.

62. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 10.48550/ARXIV.1802.03426

63. Shugay, M., Bagaev, D.V., Zvyagin, I.V., Vroomans, R.M., Crawford, J.C., Dolton, G., Komech, E.A., Sycheva, A.L., Koneva, A.E., Egorov, E.S., et al. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. *46*, D419–D427. 10.1093/nar/gkx760.

64. Bagaev, D.V., Vroomans, R.M.A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E.S., Zvyagin, I.V., et al. (2020). VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res. *48*, D1057–D1062. 10.1093/nar/gkz874.

65. Chen, S.-Y., Yue, T., Lei, Q., and Guo, A.-Y. (2021). TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. Nucleic Acids Res. *49*, D468–D474. 10.1093/nar/gkaa796.

66. Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with GPUs. arXiv. 10.48550/ARXIV.1702.08734.

67. Abbas, H.A., Hao, D., Tomczak, K., Barrodia, P., Im, J.S., Reville, P.K., Alaniz, Z., Wang, W., Wang, R., Wang, F., et al. (2021). Single cell T cell landscape and T cell receptor repertoire profiling of AML in context of PD-1 blockade therapy. Nat. Commun.*12*, 6071. 10.1038/s41467-021-26282-z.

Figure 1

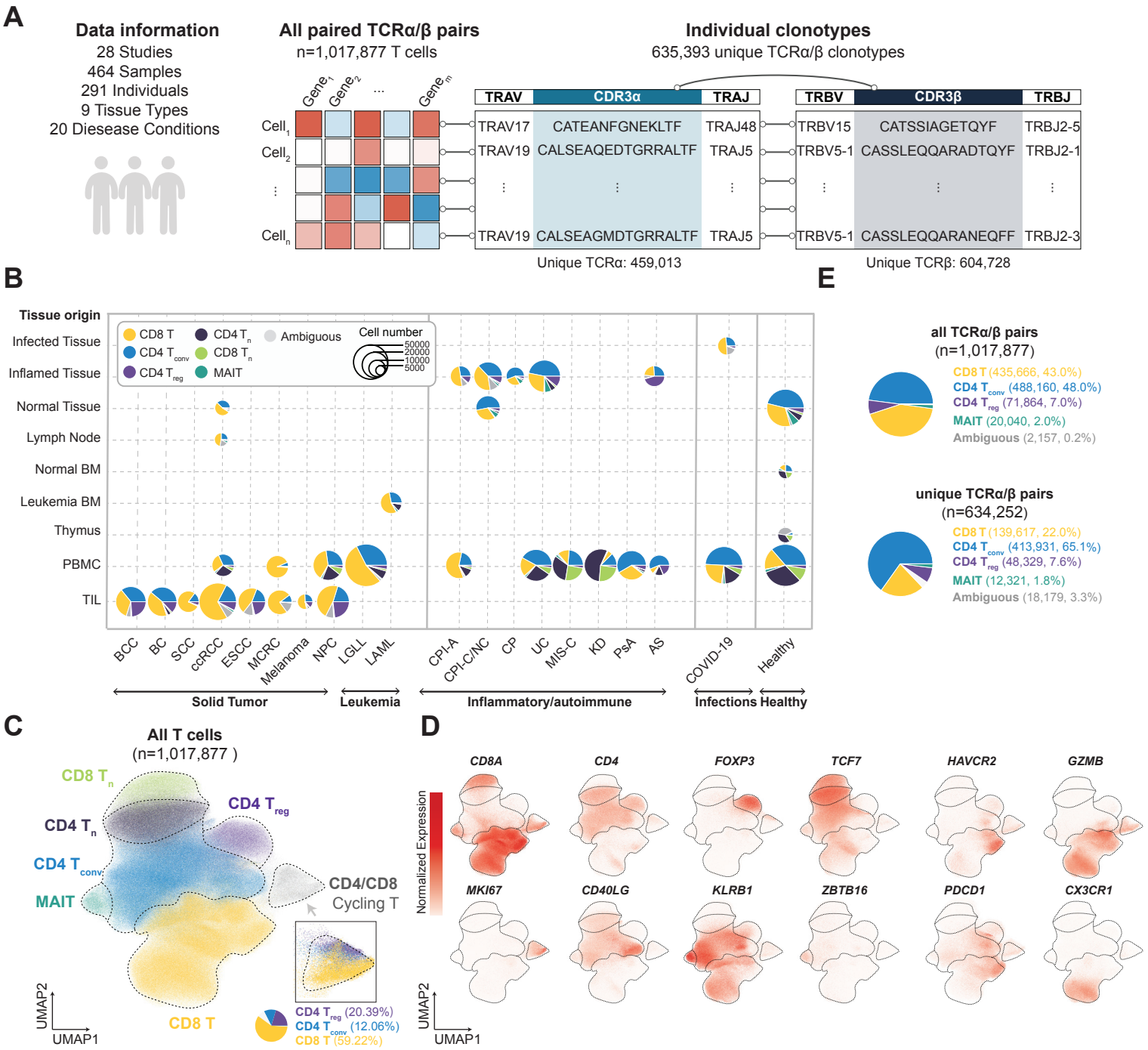

Figure 1. Overview of a million-level single-cell paired transcriptome and full-length TCR sequences dataset.
(A) Schematic representation of the joint single-cell RNA sequencing (scRNA-seq) and full-length single-cell T cell receptor sequencing (scTCR-seq) data collected in this study.
(B) Pie chart illustrating the distribution of T cell lineages from different tissue origins and disease types in the collected data.
(C) Uniform manifold approximation and projection (UMAP) representation and cell lineage annotation for variational autoencoder (VAE)-based data integration for million-level scRNA-seq data. The panel in the bottom right corner shows a zoomed-in UMAP for cycling CD4 T_{conv}, CD8 T, and CD4 T_{reg} cells.
(D) Feature plot for key marker genes, including *CD8A, CD4, FOXP3, TCF7, HAVCR2, GZMB, MKI67, CD40LG, KLRB1, ZBTB16, PDCD1*, and *CX3CR1*, used for cell lineage annotation.
(E) Pie chart depicting the composition of cell lineages for all TCRα/β pairs (upper panel) and unique TCRα/β pairs (lower panel).
CD4 T_{conv}: conventional CD4 T cells; CD4 T_{reg}: regulatory CD4 T cells; CD4 T_n: naïve CD4 T cells; CD8 T_n: naïve CD8 T cells; MAIT: mucosal-associated invariant T cells. BCC: basal cell carcinoma; BC: breast cancer; SCC: squamous cell carcinoma; ccRCC: clear cell renal carcinoma; ESCC: esophagus squamous cell carcinoma; MCRC: metastatic colorectal cancer; BALL: B-cell acute lymphoblastic leukemia; NPC: nasopharyngeal carcinoma; LGLL: large granular lymphocytic leukemia; LAML: acute myeloid leukemia; UC: ulcerative colitis; CPI-C: checkpoint inhibitor associated colitis; CPI-NC: checkpoint inhibitor treatment with no colitis; CP: chronic pancreatitis; MIS-C: multisystem inflammatory syndrome; KD: Kawasaki disease; CPI-A: checkpoint inhibitor associated arthritis; AS: ankylosing spondylitis; PsA: psoriatic arthritis. BM: bone marrow; PBMC: peripheral blood mononuclear cell. TIL: tumor-infiltrating lymphocyte.

Figure 2

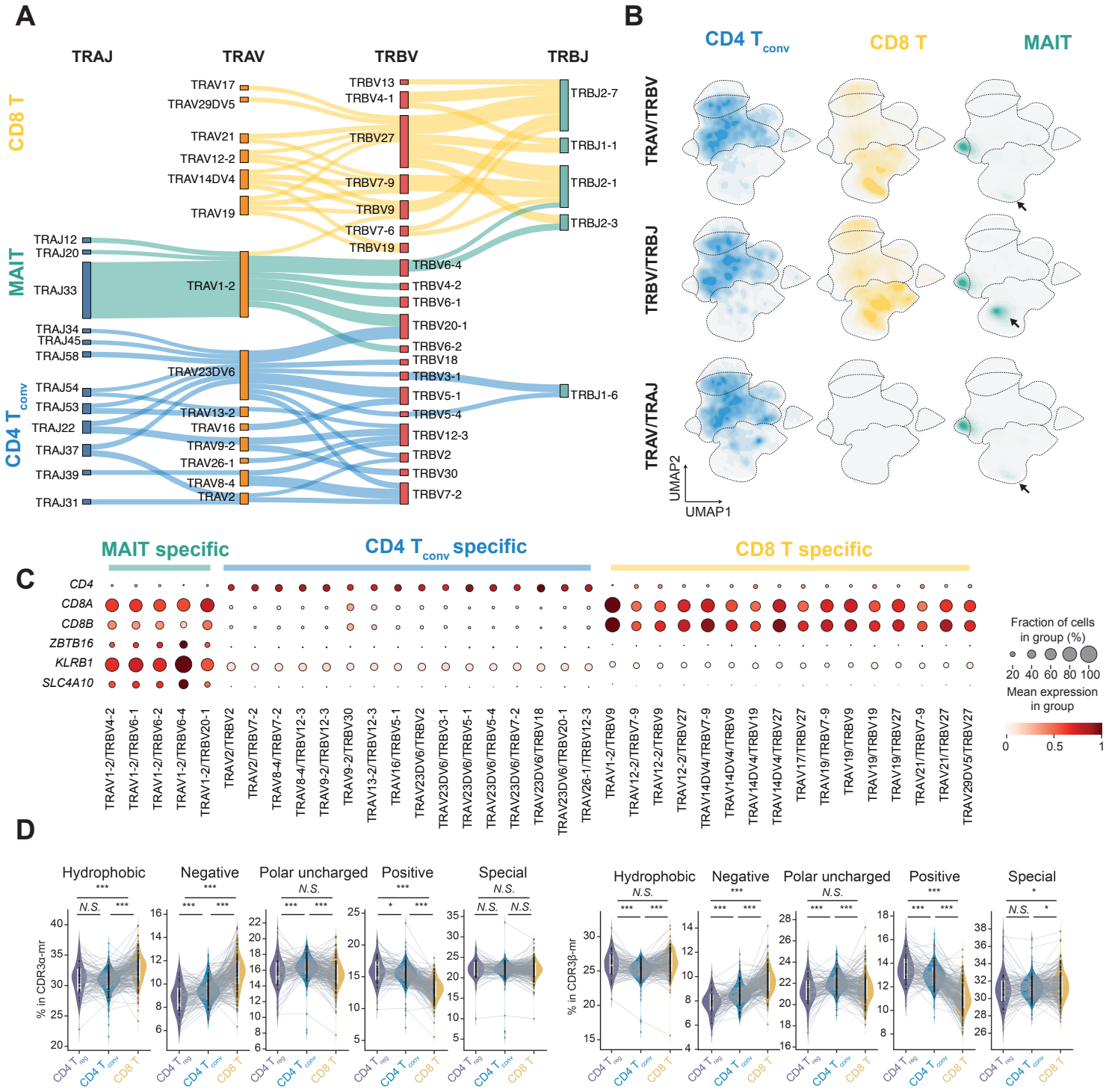

Figure 2. Analysis of T cell receptor (TCR) VJ segment pairing and amino acid usage in different T cell types.
(A) Sankey plot illustrating the enriched TRAV/TRAJ, TRAV/TRBV, and TRBV/TRBJ pairing and joining preferences in CD8 T, MAIT, and CD4 T_{conv} cells.
(B) Density plot showing the distribution of cells with enriched TRAV/TRAJ, TRAV/TRBV, and TRBV/TRBJ pairing in integrated UMAP for CD8, MAIT, and CD4 T_{conv} cells.
(C) Dot plot displaying the expression level of lineage-specific genes in CD8 T, MAIT, and CD4 T_{conv} cells with enriched TRAV/TRAJ, TRAV/TRBV, and TRBV/TRBJ.
(D) Violin plot showing the preference for amino acids with different physicochemical properties in CDR3 middle region (mr) of CD8, CD4 T_{conv}, and CD4 T_{reg} cells (*** *p-value* < 0.001, * *p-value* < 0.01, paired *t*-test; N.S.= not significant). CD8: conventional CD8 T cells; CD4 T_{conv}: conventional CD4 T cells; CD4 T_{reg}: regulatory CD4 T cells; MAIT: mucosal-associated invariant T cells; TRAV: TCR alpha variable; TRAJ: TCR alpha joining; TRBV: TCR beta variable; TRBJ: TCR beta joining. Each grey dot represent one individual, lines connecting grey dots indicate amino acid usage in different cell types within one individual. White boxes within violin plot represent the distribution from 25 percentile to 75 percentile.

Figure 3

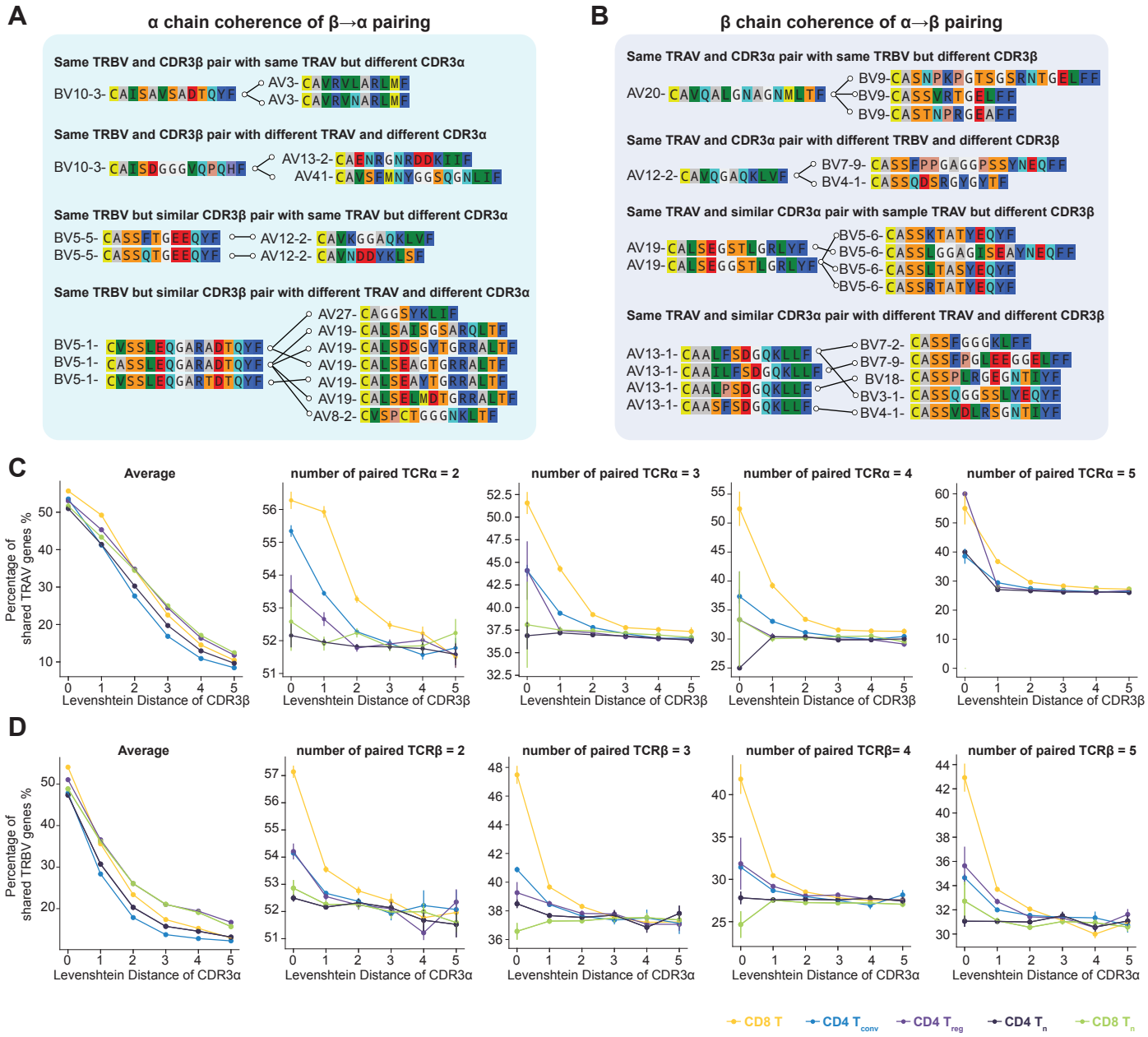

Figure 3. Coherence analysis of TCRα-TCRβ pairing.

(A, B) Example of α coherence of β-α pairing (**A**) and β coherence of α-β pairing (**B**). T cells are grouped if (1) they had the same TRBV (α coherence) or same TRAV (β coherence); (2) they had the CDR3β (α coherence) or CDR3α (β coherence) amino acid sequence with Levenshtein distance differences less than 6; (3) they belong to the same T cell types including CD8 T, CD4 T_{conv}, CD4 T_{reg}, CD4 T_n, and CD8 T_n.

(C, D) TRAV gene (**C**) and TRBV gene (**D**) usage for α coherence of β-α pairing (**C**) and β coherence of of α-β pairing (**D**) at different Levenshtein distance for CDR3α (**C**) or CDR3β (**D**) in different T cell types.

Figure 4

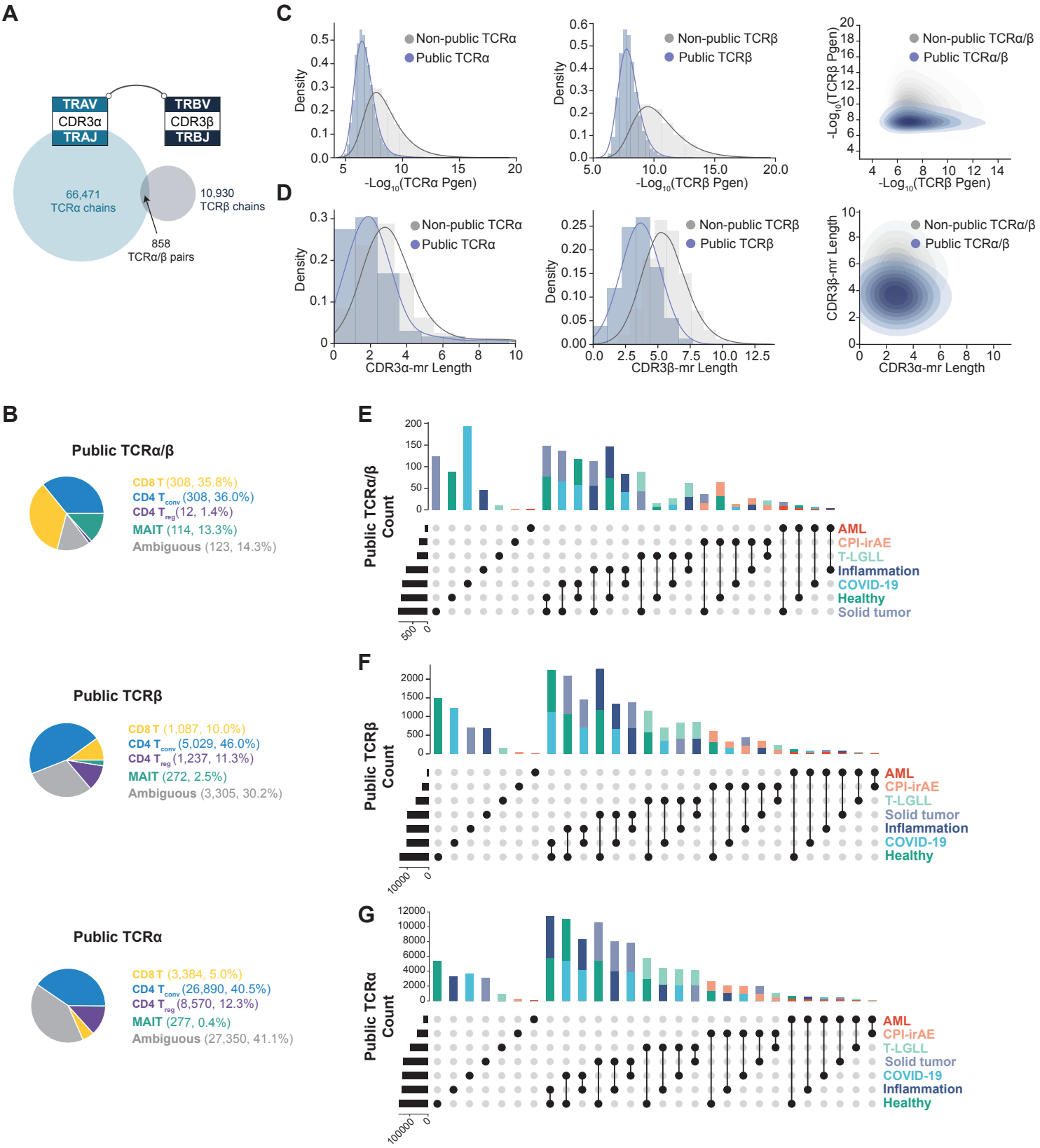

Figure 4. Public TCR analysis across samples and studies.
(A) Venn diagram of public TCR defined by same α chain (same TRAV-CDR3α-TRAJ), β chain (same TRBV-CDR3β-TRBJ) or both chains.
(B) Pie chart for cell type composition of public TCRα/β (upper panel), TCRα (middle panel), or TCRβ (lower panel).
(C) Distribution of TCR generation probability for TCRα (left panel), TCRβ (middle panel), or TCRα/β (right panel) of public and non-public TCRs.
(D) Distribution of CDR3 middle region length for TCRα (left panel), TCRβ (middle panel), or TCRα/β (right panel) of public and non-public TCRs.
(E-G) Upset plot of disease composition of public TCRα/β (**E**), TCRα (**F**), or TCRβ (**G**).

Figure 5

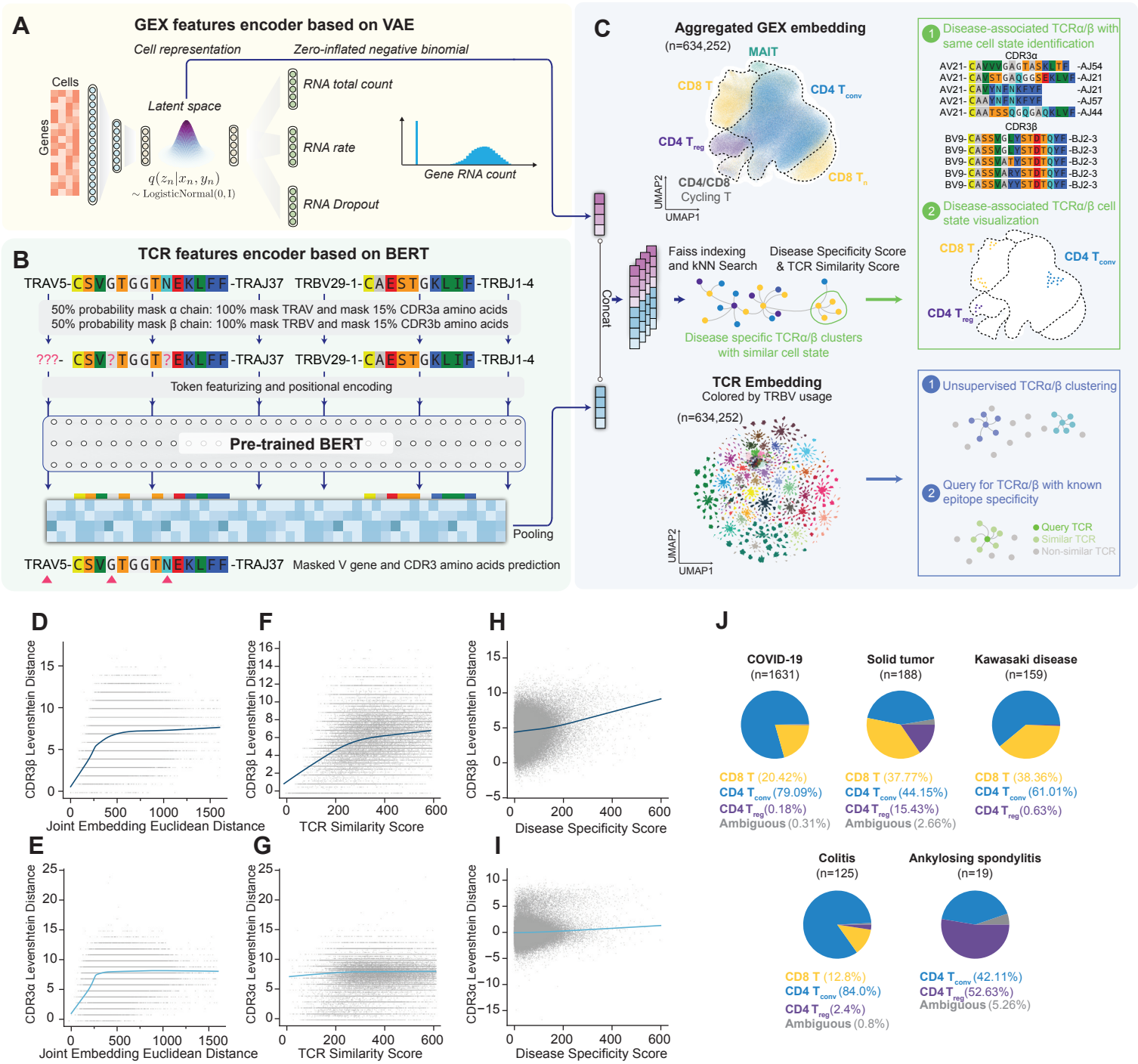

Figure 5. Integrating TCR sequences and transcriptome features to identify dTCR clusters with transcriptome similarity.

(A-C) Schematic illustration for TCR-DeepInsight model including VAE model used in transcriptome features extraction (**A**), TCR pre-trained with BERT model used in TCR features extraction (**B**), TCR embedding used for down-stream tasks including unsupervised TCR clustering, query TCR with known epitope specificity (**C**), and concatenated joint embedding with both TCR and transcriptome features (TCR-GEX joint embedding) used for downstream dTCR clusters identification, and GEX visualization (**C**).

(D, E) Relationship of Levenshtein distance of amino acids in CDR3α (**D**) or CDR3β (**E**) and Euclidean distance of TCR-GEX joint embedding between random selected pairs of TCRs.

(F, G) Relationship of average pair-wise Levenshtein distance of amino acids in CDR3α (**F**) or CDR3β (**G**) and the TCR similarity score of TCR-GEX joint embedding for all potential disease TCR clusters.

(H, I) Relationship of difference in Levenshtein distance of amino acids in CDR3α (**H**) or CDR3β (**I**) and disease-specificity score of TCR-GEX joint embedding for all potential disease TCR clusters.

(J) Pie-chart represent the cell type composition of dTCR clusters in different disease conditions.

Figure 6

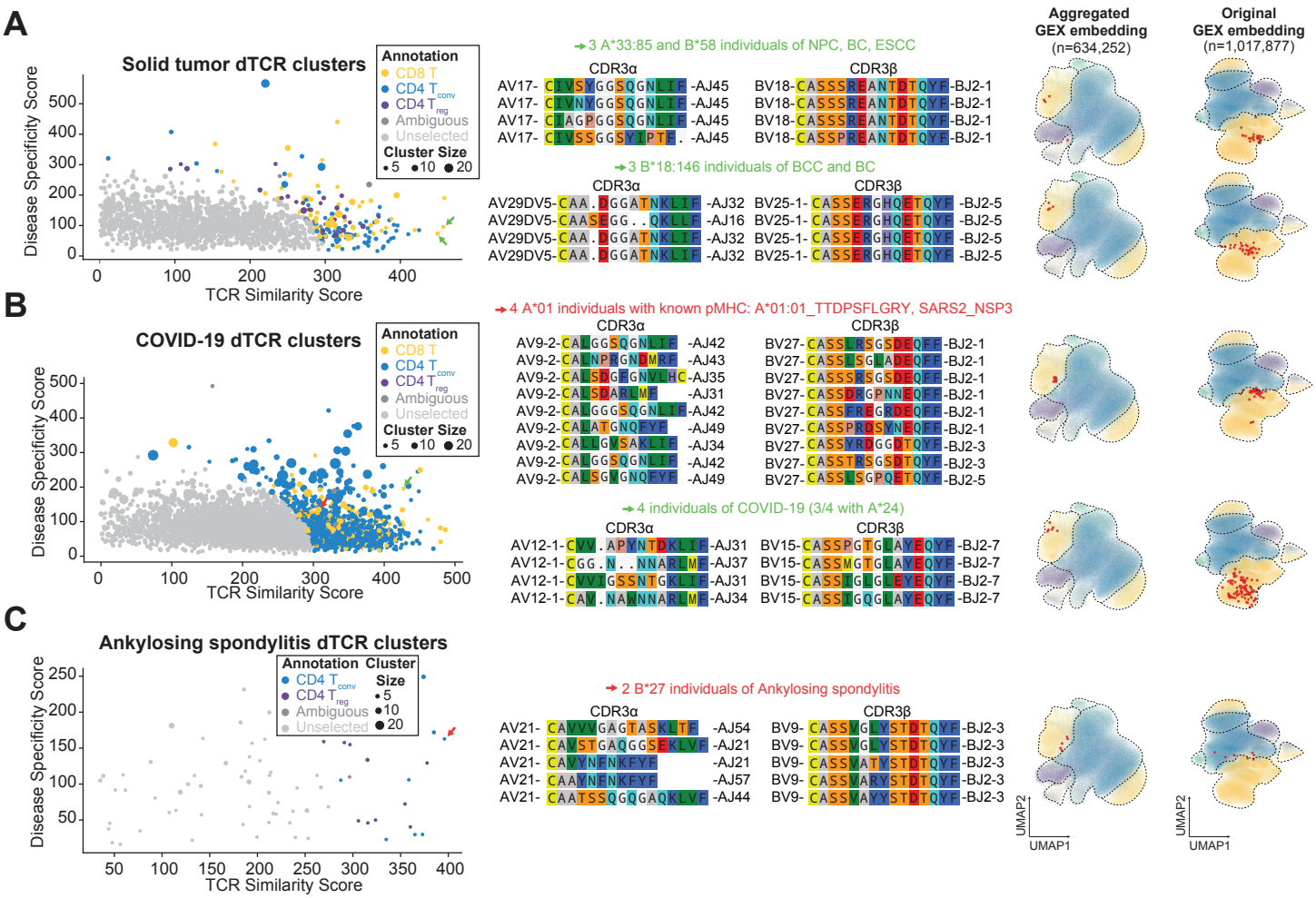

Figure 6. Representative dTCR clusters with transcriptome similarity in solid tumor, COVID-19 and ankylosing spondylitis.

(A-C) Representative dTCR clusters of solid tumor **(A)**, COVID-19 **(B)**, and Ankylosing spondylitis **(C)**. Scatter plot of TCR similarity score, disease specificity score, and cell type annotation for dTCR clusters were shown (left panel), VJ gene usage and CDRα/β amino acid sequences were for selected dTCR clusters were shown (middle panel), and their transcriptome featurs on original and aggregated GEX embedding were visualized (right panel). Green arrows on scatter plot indicate dTCR clusters selected for visualization with no known antigen specificity, red arrows on scatter plot indicate dTCR clusters selected for visualization with previously reported antigen specificity. Red stars on aggregated and original GEX UMAP indiate the cells harboring these dTCRs.

Figure 7

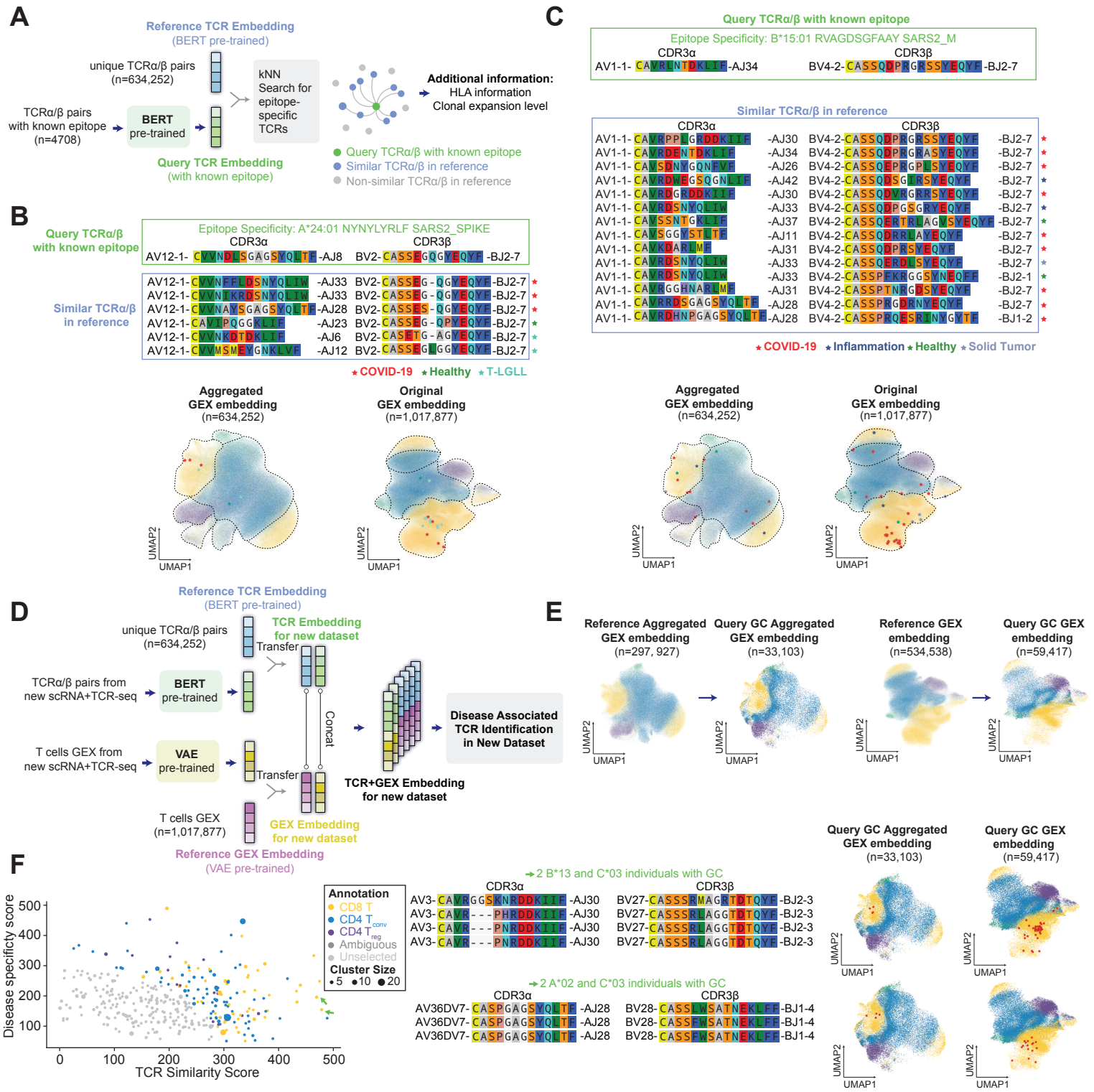

Figure 7. TCR-DeepInsight extended for query TCR with known epitope specificity and transfer on new dataset.
(A) Schematic illustration for TCR-DeepInsight module used for querying TCRs with known epitope specificity in the reference TCR atlas.
(B, C) Representative TCR cluster from the reference TCR atlas queried by an A*24:01 restricted epitope NYNYLYRLF from SARS-CoV-2 Spike protein (**B**) and an B*15:01 restricted epitope RVAGDSGFAAY from SARS-CoV-2 membrane glycoprotein (**C**). Colored stars on aggregated and original GEX UMAP indiate the cells harboring these TCRs similar to the queried TCRs. Color correspond to disease conditions associated with TCRs.
(D) Schematic illustration for transfering BERT pre-trained TCR model and VAE pre-trained GEX model to new scRNA-seq and scTCR-seq dataset to identify dTCRs.
(E) UMAP representation of original GEX and aggregated GEX embedding for queried scRNA-seq and scTCR-seq dataset from gastric cancer and randomly down-sampled reference atlas.
(F) Representative dTCR clusters identified in gastric cancer. Scatter plot of TCR similarity score, disease specificity score, and cell type annotation for dTCR clusters were shown (left panel), VJ gene usage and CDRα/β amino acid sequences were for selected dTCR clusters were shown (middle panel), and their transcriptome featurs on original and aggregated GEX embedding were visualized (right panel). Green arrows on scatter plot indicate dTCR clusters selected for visualization with no known antigen specificity. Red stars on aggregated and original UMAP indiates the cells harboring these dTCRs.

Figure S1

Click here to access/download;Supplemental Text and Figures;20230403TCR-Supplementary-Figure-RS.pdf

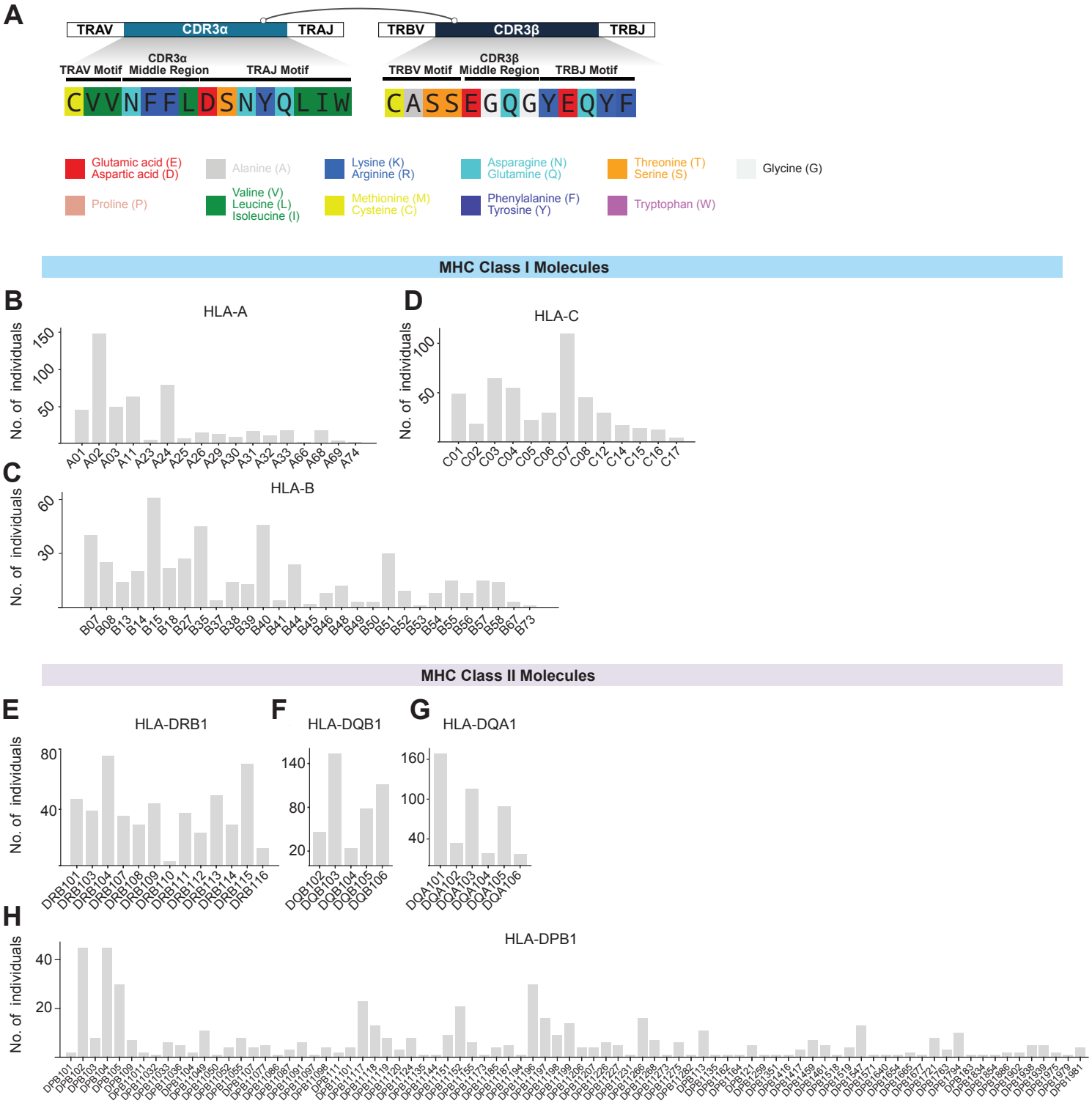

Figure S1. HLA Information in the integrated Paired Transcriptome and Full-Length TCR Sequences Dataset, related to Figure 1.

(A) Schematic illustration for V gene motif, middle region, and J gene motif definition of the CDR3 region. Amino acids in this study were colored by RasMol amino acid color scheme based on Robert Fletterick's "Shapely models".

(B-H) Distribution of number of individuals of called HLA alleles of MHC I HLA-A (**B**), HLA-B (**C**), HLA-C (**D**) and MHC II HLA-DRB1 (**A**), HLA-DQB1 (**F**), HLA-DQA1 (**G**), HLA-DPB1 (**H**) in the integrated datasets.

Figure S2

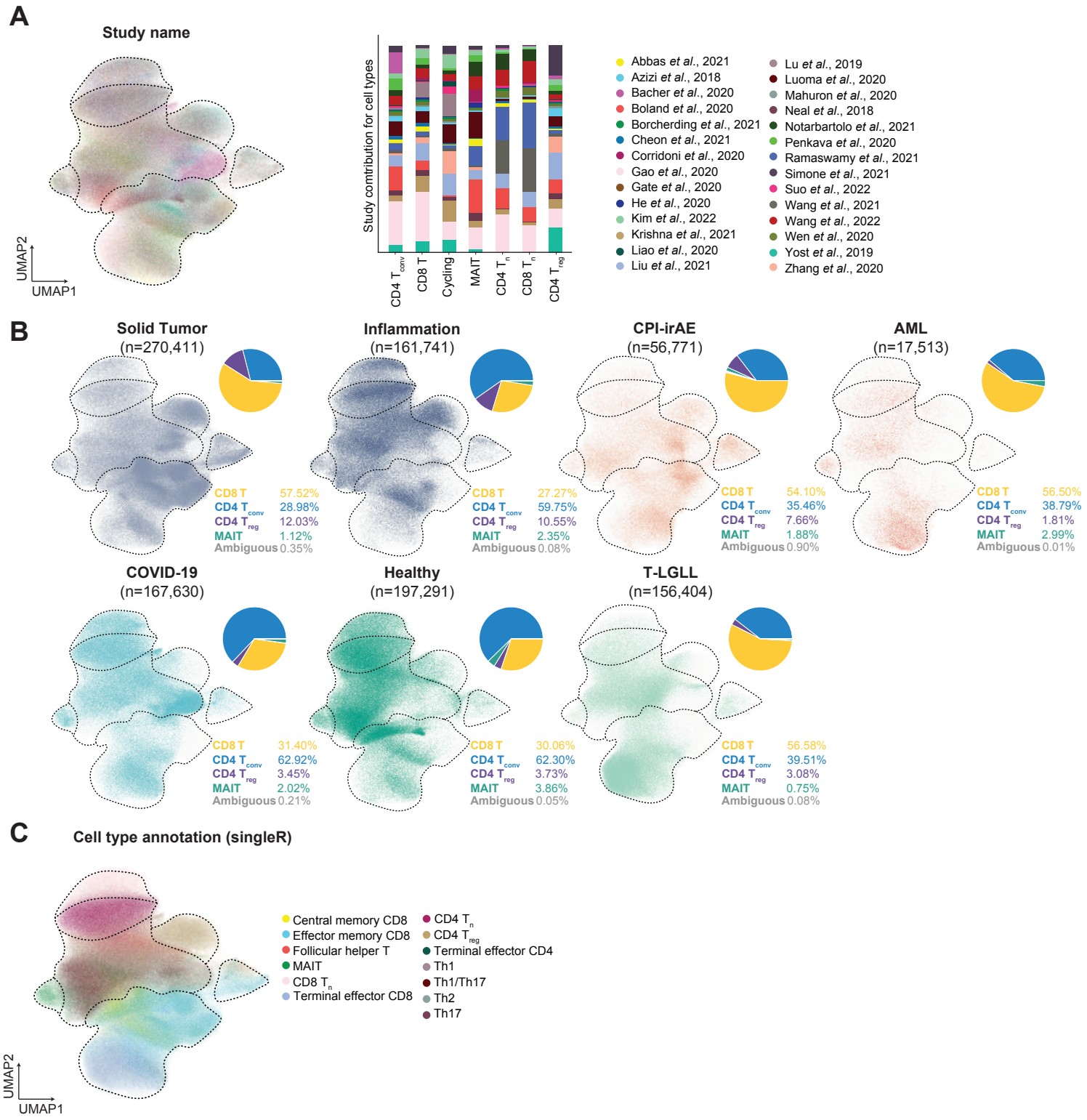

Figure S2. Additional information on the integrated transcriptome features of over one-million T cells, related to Figure 1.

(A) The UMAP projection of the integrated million-level T cells colored by study names (left panel), and bar plot for the distribution of different studies corresponding to major T cell types including CD4 T_{conv}, CD4 T_{reg}, CD4 T_n, CD8 T_n, and MAIT cells (right panel)**.**

(B) Distribution of cells from different disease states including solid tumors, inflammation, CPI-irAE (checkpoint inhibitor associated immune-related adverse events), AML (Acute Myeloid Leukemia), COVID-19, Healthy, and T-LGLL (T cell large granular lymphocytic leukemia) in the integrated UMAP and pie chart for the composition of T cell states in each disease conditions.

(C) Cell type annotation by singleR as previously used in *Wu et al.*, 2021. Th1: T helper 1 cell; Th17: T helper 17 cell; Th2: T helper 2 cell.

Figure S3

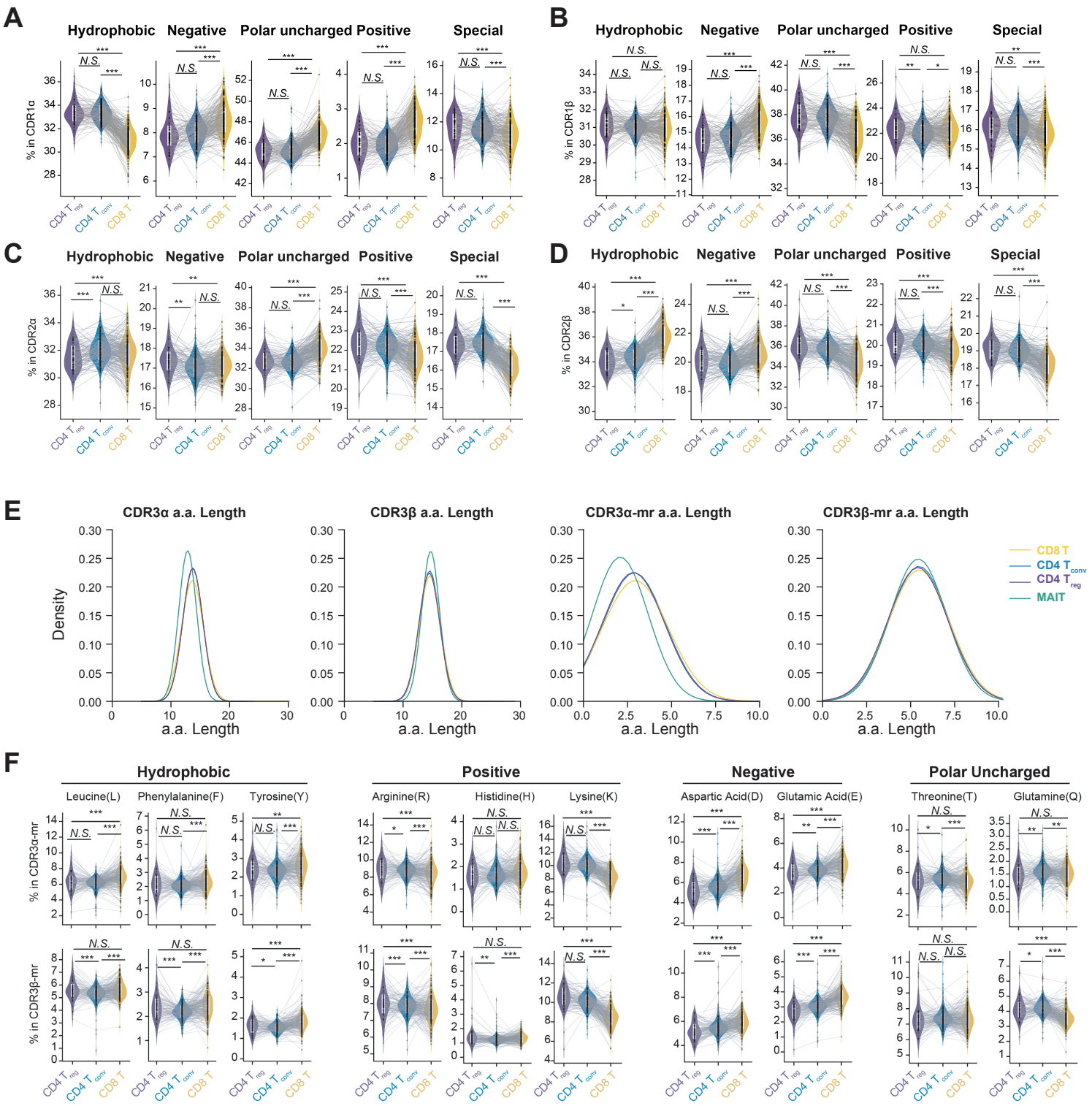

Figure S3. Amino acid usage and length distribution of CDR3 region of different T cell types, related to Figure 2.
(A-D) Violin plot indicating the preference for amino acids with different physicochemical properties in CDR1α (**A**), CDR2α (**B**), CDR1β (**C**), CDR2β (**D**) regions in CD8, CD4 T_{conv}, and CD4 T_{reg} cells (*** *p-value* < 0.001, ** *p-value* < 0.01, * *p-value* < 0.05, paired *t*-test; N.S.= not significant).
(E) Distribution of the length of CDR3 and CDR3 amino acids (a.a.) mr in CD8, CD4 T_{conv}, and CD4 T_{reg} cells.
(F) Violin plot indicating the preference for amino acids with selected amino acids grouped by physicochemical properties in CDR3α-mr (upper panel) and CDR3β-mr (lower panel) in CD8, CD4 T_{conv}, and CD4 T_{reg} cells (*** *p-value* < 0.001, ** *p-value* < 0.01, * *p-value* < 0.05, paired *t*-test; N.S.= not significant). Each grey dot represent one individual, lines connecting grey dots indicate amino acid usage in different cell types within one individual. White boxes within violin plot represent the distribution from 25 percentile to 75 percentile.

Figure S5

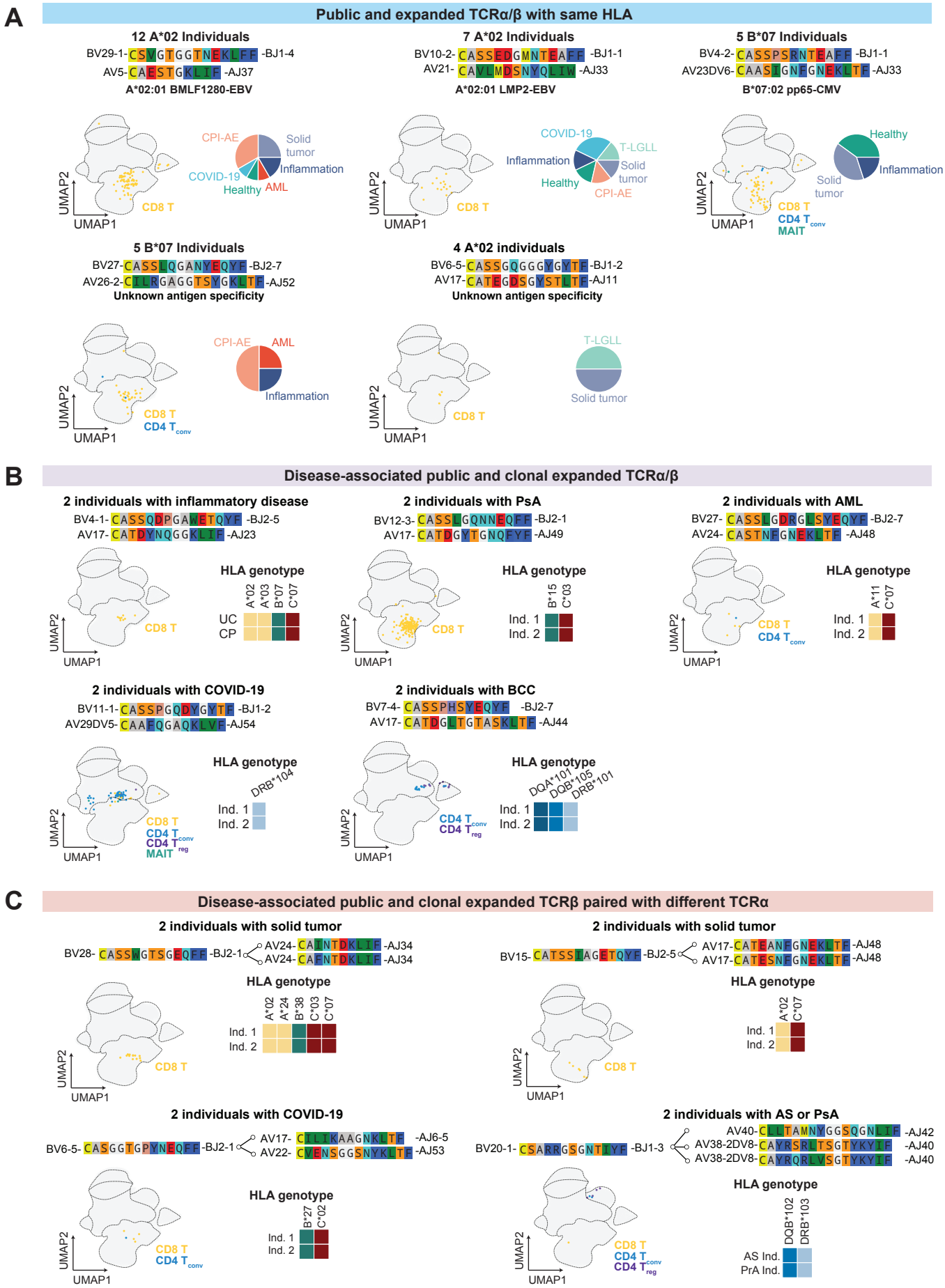

Figure S5. Visualization for public TCRs with their corresponding transcriptome states with at least one common HLA allele, related to Figure 4.

(A-C) Representative public and expanded TCRα/β pairs (**A**), disease-associated public and clonal expanded TCRα/β pairs **(B)** and disease-associated public and clonal expanded TCRβ paired with different TCRα (**C**).

Figure S6

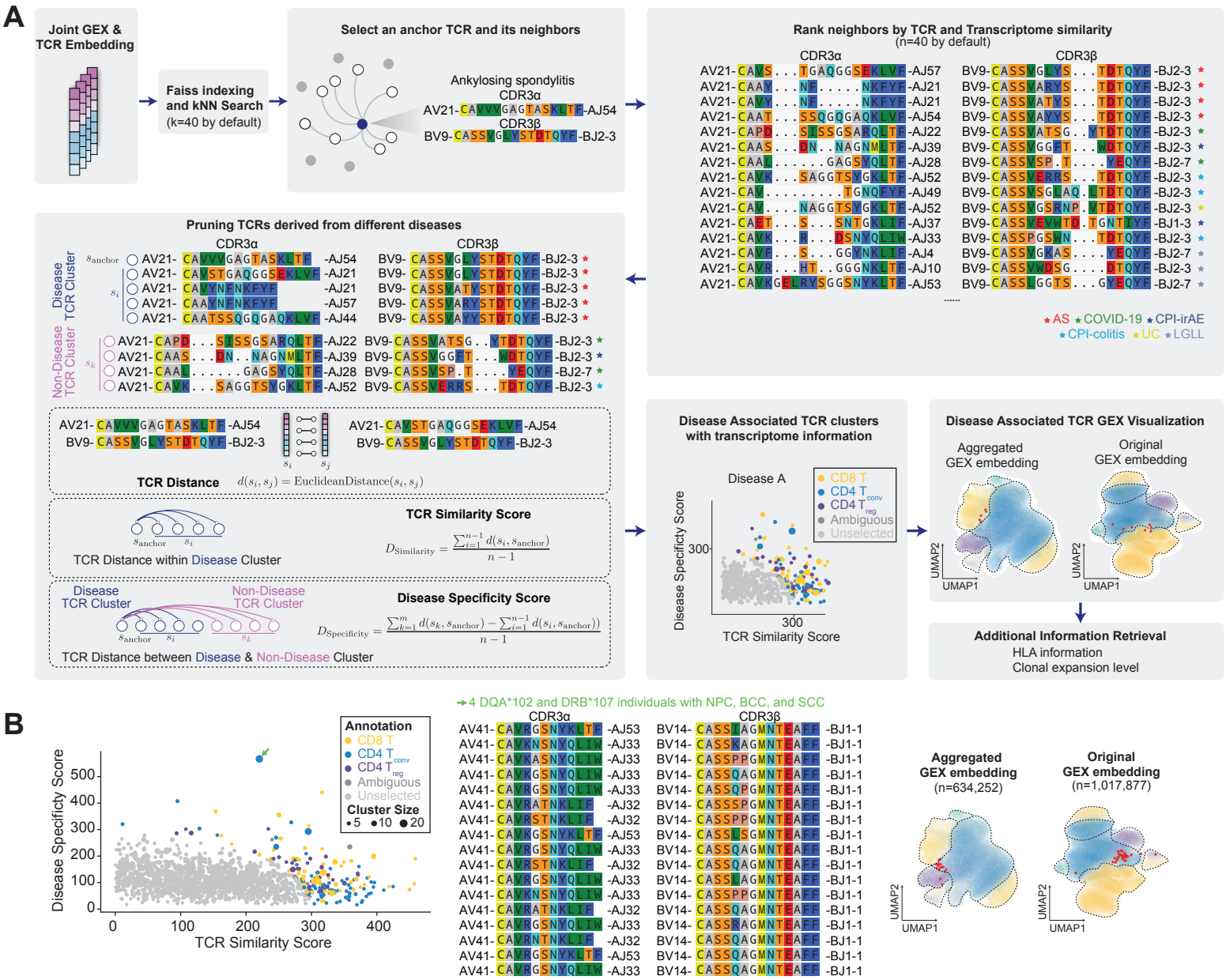

Figure S6. Model details for TCR-DeepInsight and representative CD4 T_{conv} dTCR cluster for solid tumor, relate to Figure 5 and 6.

(A) Schematic illustration for the identification of dTCR clusters with TCR-DeepInsight. Briefly, joint GEX and TCR embedding were faiss indexed for kNN search. Then anchor TCR and its neighbors were identified and ranked by their TCR and transcriptome similarity with anchor TCR. TCRs derived from different disease were pruned to group TCRs into disease TCR cluster and non-disease TCR cluster.TCR-similarity score and disease-specificity score were calculated based on TCR distance, dTCR clusters were then indentified and visualized on GEX embedding, and additional HLA information and clonal expansion level could be retrieved.

(B) Representative CD4 T_{conv} disease-associated TCR cluster identified in solid tumor. Scatter plot of TCR similarity score, disease specificity score, and cell type annotation for dTCR clusters were shown (left panel), VJ gene usage and CDRα/β amino acid sequences were for selected dTCR clusters were shown (middle panel), and their transcriptome featurs on original and aggregated GEX embedding were visualized (right panel). Green arrows on scatter plot indicate dTCR clusters selected for visualization with no known antigen specificity. Red stars on aggregated and original GEX UMAP indiate the cells harboring these dTCRs.