

# 1 Human SARS-CoV-2 challenge resolves local and systemic response dynamics

2

3 Rik G.H. Lindeboom\*<sup>1,3</sup>, Kaylee B. Worlock\*<sup>2</sup>, Lisa M. Dratva<sup>1</sup>, Masahiro Yoshida<sup>2</sup>, David Scobie<sup>6</sup>,  
4 Helen R. Wagstaffe<sup>4</sup>, Laura Richardson<sup>1</sup>, Anna Wilbrey-Clark<sup>1</sup>, Josephine L. Barnes<sup>2</sup>, Krzysztof  
5 Polanski<sup>1</sup>, Jessica Allen-Hyttinen<sup>2</sup>, Puja Mehta<sup>2</sup>, Dinithi Sumanaweera<sup>1</sup>, Jacqueline Boccacino<sup>1</sup>,  
6 Waradon Sungnak<sup>1</sup>, Ni Huang<sup>1</sup>, Lira Mamanova<sup>1</sup>, Rakesh Kapuge<sup>1</sup>, Liam Bolt<sup>1</sup>, Elena Prigmore<sup>1</sup>, Ben  
7 Killingley<sup>6</sup>, Mariya Kalinova<sup>5</sup>, Maria Mayer<sup>5</sup>, Alison Boyers<sup>5</sup>, Alex Mann<sup>5</sup>, Vitor Teixeira<sup>2</sup>, Sam M.  
8 Janes<sup>2</sup>, Rachel C. Chambers<sup>2</sup>, Muzlifah Haniffa<sup>1</sup>, Andrew Catchpole<sup>5</sup>, Robert Heyderman<sup>6</sup>, Mahdad  
9 Noursadeghi<sup>6</sup>, Benny Chain<sup>6</sup>, Andreas Mayer<sup>6</sup>, Kerstin B. Meyer<sup>1</sup>, Christopher Chiu<sup>4</sup>, Marko Z.  
10 Nikolić<sup>†2</sup>, Sarah A. Teichmann<sup>†1</sup>

11

12 1. Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK

13 2. UCL Respiratory, Division of Medicine, University College London, London, WC1E 6JF UK

14 3. The Netherlands Cancer Institute, Amsterdam, the Netherlands

15 4. Department of Infectious Disease, Imperial College London, London

16 5. hVIVO, London, UK

17 6. Research Department of Infection, Division of Infection and Immunity, University College London,  
18 London, UK.

19

20 \*co-first authors

21 †co-senior authors

22

23 Correspondence to: Rik G.H. Lindeboom ([r.lindeboom@nki.nl](mailto:r.lindeboom@nki.nl)), Marko Z. Nikolić

24 ([m.nikolic@ucl.ac.uk](mailto:m.nikolic@ucl.ac.uk)), Sarah A. Teichmann ([st9@sanger.ac.uk](mailto:st9@sanger.ac.uk))

25

26

## 27 Abstract

28 The COVID-19 pandemic is an ongoing global health threat, yet our understanding of the cellular  
29 disease dynamics remains limited. In our unique COVID-19 human challenge study we used single cell  
30 genomics of nasopharyngeal swabs and blood to temporally resolve abortive, transient and sustained  
31 infections in 16 seronegative individuals challenged with pre-alpha SARS-CoV-2. Our analyses

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

32 revealed rapid changes in cell type proportions and dozens of highly dynamic cellular response states  
33 in epithelial and immune cells associated with specific timepoints or infection status. We observed that  
34 the interferon response in blood precedes the nasopharynx, and that nasopharyngeal immune infiltration  
35 occurred early in transient but later in sustained infection, and thus correlated with preventing sustained  
36 infection. Ciliated cells showed an acute response phase, upregulated MHC class II while infected, and  
37 were most permissive for viral replication, whilst nasal T cells and macrophages were infected non-  
38 productively. We resolve 54 T cell states, including acutely activated T cells that clonally expanded  
39 while carrying convergent SARS-CoV-2 motifs. Our novel computational pipeline (Cell2TCR)  
40 identifies activated antigen-responding clonotype groups and motifs in any dataset. Together, we show  
41 that our detailed time series data ([covid19cellatlas.org](https://covid19cellatlas.org)) can serve as a “Rosetta stone” for the epithelial  
42 and immune cell responses, and reveals early dynamic responses associated with protection from  
43 infection.

44

45

## 46 **Main**

47 Coronavirus Disease 2019 (COVID-19) is a potentially fatal disease caused by the severe acute  
48 respiratory syndrome coronavirus 2 (SARS-CoV-2), which gave rise to one of the most severe global  
49 public health emergencies in recent history. Studies by us and others have uncovered that perturbed  
50 antiviral and immune responses to SARS-CoV-2 underlie severe and fatal outcomes, where for example  
51 impaired type I interferon responses<sup>1,2</sup>, decreases of circulating T cell and monocyte subsets<sup>3-5</sup>, and  
52 increased clonal expansion of T and B cells<sup>4</sup> are associated with a more severe outcome. However,  
53 accurate detection and interpretation of the immune response during COVID-19 has been hampered by  
54 heterogeneous responses caused by numerous non-host factors that affect immune and clinical  
55 outcomes that are frequently unmeasurable and uncontrolled. These include infection characteristics  
56 such as viral dose, strain and time since exposure, together with clinical features including  
57 comorbidities, standard of care and pre-existing immunity. In particular, the observed immune response  
58 may represent different phases, from early viral detection to later adaptive responses, depending on the  
59 time between infection and sampling.

60

61 Since the exact time at which patients were exposed to SARS-CoV-2 is nearly always unknown, it can  
62 be challenging to accurately delineate severity-associated and temporal effects such as early interferon  
63 signaling and late adaptive immune responses<sup>1-6</sup>. Determining the dynamics of SARS-CoV-2 infection  
64 and the body’s response is therefore crucial to understand how the immune response is orchestrated and  
65 how risk factors can impact this. In addition, while many studies have investigated responses to COVID-  
66 19 during the course of the disease<sup>7,8</sup>, it has thus far not been possible to study the early phases of

67 exposure and the infection event itself in humans. In particular, studies of natural infection are unable  
68 to capture events in those who are exposed to the virus but do not develop sustained viral infection,  
69 which might be critical in preventing dissemination and disease. Furthermore, the activation and  
70 expansion of antigen-responding T cells (versus bystanders) has been difficult to pinpoint in previous  
71 “snapshot” datasets<sup>4,5</sup>. Here, we trace their development and integrate the paired TCR chains for the  
72 first time.

73

#### 74 **Human SARS-CoV-2 challenge model**

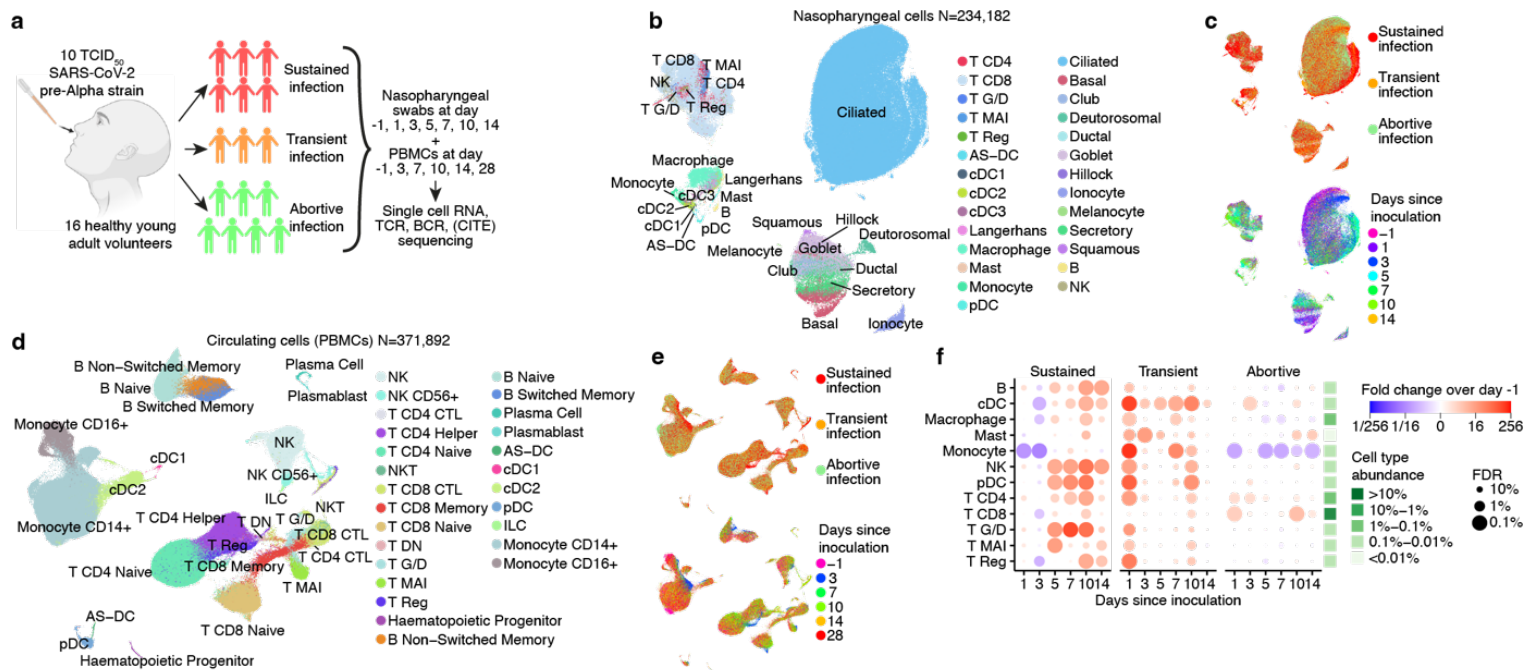
75 To resolve epithelial and immune cell responses over time from SARS-CoV-2 exposure, we conducted  
76 a first of its kind, human COVID-19 challenge study<sup>6</sup>. In this model, young adults seronegative for  
77 SARS-CoV-2 spike were inoculated intranasally with a wild-type pre-Alpha SARS-CoV-2 virus strain  
78 (SARS-CoV-2/human/GBR/484861/2020) in a controlled environment<sup>6</sup>. Prior to challenge, volunteers  
79 underwent extensive screening to exclude risk factors for severe disease and eliminate confounding  
80 effects of comorbidities. As risk mitigation and to maximize physiological relevance, participants were  
81 inoculated with the lowest culture-quantifiable inoculum dose of 10 Tissue Culture Infectious Dose 50  
82 (TCID<sub>50</sub>). There were no serious adverse events and symptoms resolved spontaneously without  
83 treatment.

84

85 We studied local and systemic immune responses at single cell resolution in 16 participants. The highly  
86 controlled nature of this experimental model allowed baseline measurements on the day before  
87 inoculation, followed by detailed time series analyses of cellular responses after inoculation and  
88 subsequent infection, both systemically and in the nasopharynx, to decipher antiviral responses against  
89 SARS-CoV-2 in a precise time-resolved manner.

90

91 Following inoculation, 6 participants from the cohort developed a sustained infection as defined by at  
92 least 2 consecutive quantifiable viral load detections by PCR, along with symptoms (**Fig. 1a and**  
93 **Extended Data Fig. 1**). In contrast, three individuals produced multiple sporadic and borderline-  
94 positive PCR tests between day 1.5 and 7 post-inoculation. While these participants remained symptom  
95 free and did not meet the earlier established criteria to be classified as “sustained infection”, we assigned  
96 them to a separate group of “transient infection” to investigate factors associated with this unique  
97 phenotype.



**Figure 1: Extensive temporal cell state dynamics after SARS-CoV-2 inoculation.**

(a) Illustration of study design and cohort composition. (b-c) UMAPs of all nasopharyngeal cells, color-coded by their broad cell type annotation in (b), by the infection group in the top panel of (c), and by days since inoculation in the bottom panel of (c). Only cells from sustained infection cases are shown in the bottom panel of (c). (d-e) UMAPs as in (b-c), but showing all PBMCs. (f) Fold changes in abundance of nasopharynx resident broad immune cell type categories. Immune cell abundances were scaled to the total amount of detected epithelial cells in every sample prior to calculating the fold changes over days since inoculation compared to pre-infection (day -1) by fitting a GLMM on scaled abundances. The mean cell type proportions over all cells and samples is shown in the green heatmap right of the dotplot to aid the interpretation of changes in cell type abundances.

98

99

100 Seven participants remained PCR negative throughout the quarantine period, indicating that these  
 101 individuals successfully prevented the onset of a sustained or transient infection. Due to the fact that  
 102 these participants all remained seronegative, but were observed to display early innate immune  
 103 responses (see below), we termed these abortive infections (as opposed to uninfected due to for example  
 104 antibody-mediated sterilizing immunity).

105

## 106 Broad cellular transitions observed over time and infection groups

107 To comprehensively identify and time responses to SARS-CoV-2 exposure in these phenotypically-  
 108 divergent groups, we performed single cell RNA sequencing (scRNAseq) and single cell TCR- and  
 109 BCR-seq at up to seven time points (**Fig. 1a**). In addition, we complemented the RNA measurements  
 110 in PBMCs with CITE-seq measurements to quantify 123 surface proteins to aid cell type annotation. At  
 111 each time point, we collected PBMCs and nasopharyngeal swabs to study both the systemic immune



112 response and the epithelial and local immune response at the site of inoculation, respectively. Of note,  
113 while most PBMC and nasopharyngeal time points were matched, we included more early  
114 nasopharyngeal and later PBMC time points as we anticipated more immediate local responses. In total,  
115 we generated over 600K single cell transcriptomes across 181 samples, which include 371,892 PBMCs  
116 and 234,182 nasopharyngeal cells. We used predictive models and marker gene expression to annotate  
117 202 cell states in total (see *Methods*; **Extended Data Fig. 2-5**), including multiple newly identified cell  
118 states that will be discussed throughout this manuscript. Importantly, both datasets contained all  
119 expected cell types (**Fig. 1b,d**; **Extended Data Fig. 2a-b**), including clearly resolved epithelial and  
120 immune compartments in the nasopharyngeal samples, which enabled us to study both the local and  
121 systemic immune response. Strikingly, even when visualizing all cells at once in **Fig. 1b,d**, the  
122 “infection group” and “days since inoculation” mark specific groups of cells (**Fig. 1c,e**), indicating that  
123 there are large changes in cell fate over time and infection groups across the different cell type  
124 compartments.

125

## 126 **Innate and adaptive immune infiltration to site of inoculation**

127 We first investigated how the immune landscape is affected by viral inoculation and subsequent  
128 infection. We used generalized linear mixed models (GLMM) to quantify the changes in cell type  
129 abundances over time since inoculation compared to the day prior to inoculation (-1). This allowed us  
130 to perform paired longitudinal modeling of donor-specific effects while accounting for technical and  
131 biological variation using random effect terms. Analysis of the nasopharyngeal resident immune  
132 compartment revealed that all immune cell types significantly infiltrate the site of inoculation after  
133 exposure to SARS-CoV-2 (**Fig. 1f**). Strikingly, the timing of infiltration strongly differed between  
134 participants with a sustained, transient or abortive infection. During sustained infections, immune  
135 infiltration started at day 5 after inoculation and continued to increase until day 10. In stark contrast,  
136 transient infections led to immediate and robust immune infiltration, followed by a decrease and smaller  
137 secondary infiltration event at day 10. Last, abortive infections do not lead to any obvious patterns of  
138 immune infiltration.

139

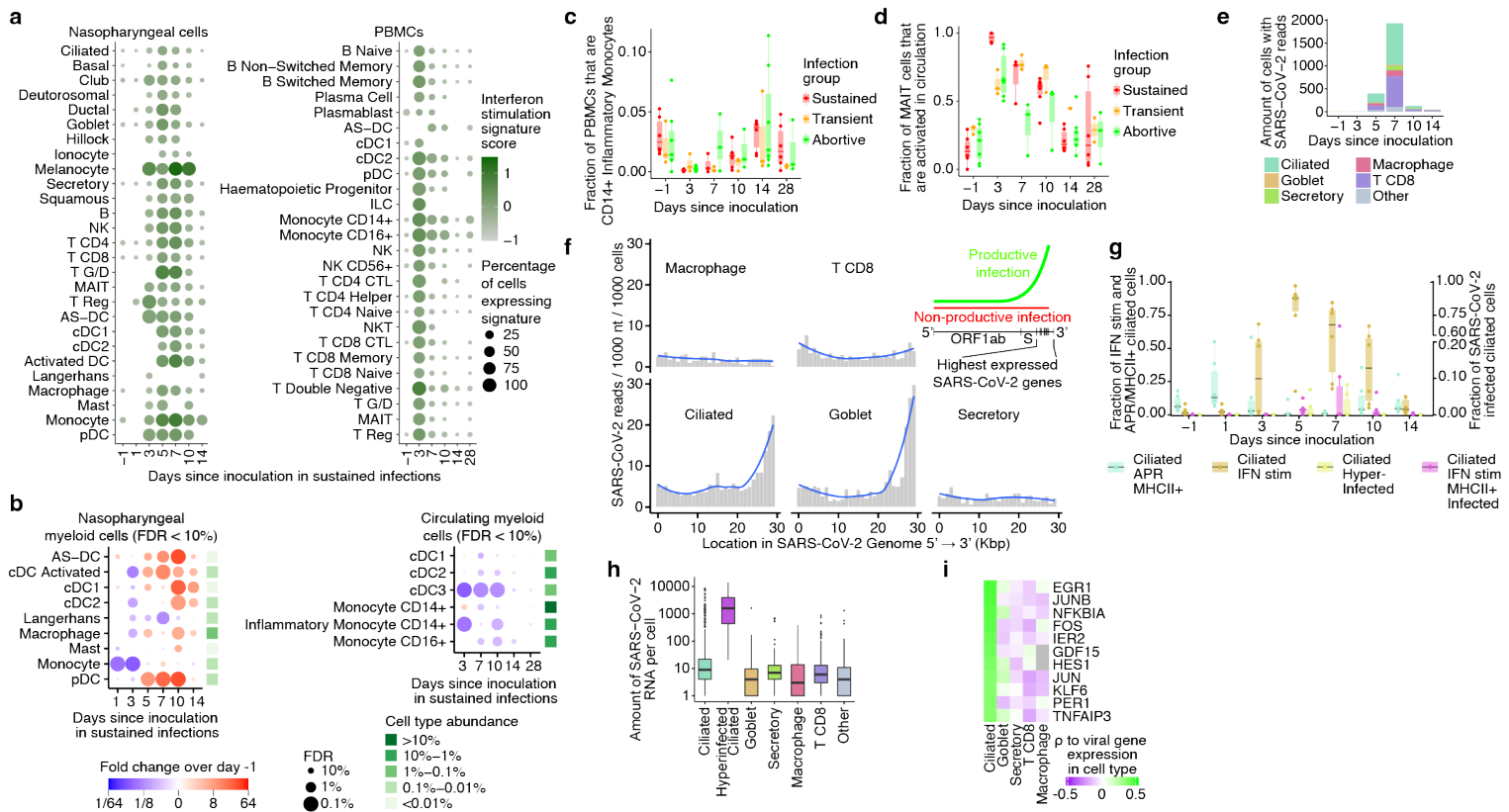
140 Interestingly, both sustained and transient infections lead to infiltration of innate and adaptive immune  
141 cells. However, the increase of innate immune cells such as plasmacytoid dendritic (pDC), natural killer  
142 (NK), gamma/delta T (g/d T), and mucosal-associated invariant T (MAIT) cells was quicker and of  
143 greater magnitude than infiltration by adaptive immune cells in sustained infections. In line with this,  
144 in transient infections, the increase of immune cells at day 1 was also greatest in the innate immune  
145 compartment. The observed difference in timing of immune infiltration between transient and sustained

146 infections suggests that immediate immune recruitment and responses are associated with containing  
 147 SARS-CoV-2 infection and preventing the onset of a sustained infection and COVID-19.

148

149 **Widespread systemic interferon response precedes response at site of inoculation**

150 We next attempted to detect antiviral gene expression programs in any of the tissue resident and  
 151 circulating cells during infection. Gene expression analysis revealed that the interferon response genes  
 152 made up the dominant infection-induced gene expression module in participants with sustained  
 153 infection (**Fig. 2a**). Strikingly, interferon signaling was strongly activated in every cell type of both the  
 154 blood and in the nasopharynx, where up to 100% of some cell types at a given time took on a distinct  
 155 interferon stimulated cell state (**Extended Data Fig. 6a**, annotated in **Extended Data Fig. 2-4** as IFN  
 156 stim), underscoring its widespread and dominant effect. Activation of interferon signaling was absent  
 157 in abortive infection and only short-lived in transient infections (**Extended Data Fig. 6a**). Interestingly,  
 158 at the site of inoculation, we only detected widespread interferon activation from five days post-  
 159 inoculation, whereas the interferon response in the blood peaks at day 3 post-inoculation and appears  
 160 to be stronger. This is unexpected, as we assumed that the cells that reside in the inoculated tissue should  
 161 be the first to respond through direct exposure to the virus and infected cells. Instead, it appears that the  
 162 immediate response to SARS-CoV-2 infection includes informing circulating immune cells before  
 163 tissue-resident cells through interferon signaling.



164

## Figure 2: Cell-state-specific antiviral responses and infection

(a) Dotplot visualizing the mean expression of interferon stimulated genes across cell types and time since inoculation in participants with sustained infections, for nasopharyngeal cells (left plot) and PBMCs (right plot). (b) Dotplot as in (Fig 1f), showing myeloid cell types in sustained infection cases that significantly change at least one time point compared to pre-infection. Nasopharyngeal cells and PBMCs are shown in the left and right plot, respectively. (c) Boxplot showing the relative amounts of circulating inflammatory monocytes over time since inoculation in each infection group. (d) Boxplot showing the fraction of circulating MAIT cells that are activated over time since inoculation in each infection group. (e) Stacked barplot showing the amount of nasopharyngeal cells with at least one SARS-CoV-2 RNA read detected (after background subtraction), split by days since inoculation and color-coded by cell type. (f) Barplots showing the distribution of detected viral reads over the SARS-CoV-2 genome in the five most highly infected cell types. The blue line represents a loess fit over the data. The top-right inset illustration is shown to aid the interpretation of a uniform read distribution versus a 3' biased read distribution. (g) Boxplot showing the fraction of ciliated cells that are annotated into detailed response or infection cell states. Only cells from sustained infection cases are shown and split by days since inoculation. The Y axis for interferon (IFN) and acute-phase response (APR) positive ciliated cells is shown on the left, while the Y axis for infected ciliated cells is shown on the right. (h) Boxplot of the amount of viral sequencing reads per cell type. (i) Heatmap of spearman correlations between host gene expression and the amount of viral reads found in each cell, split by cell type. Shown genes correlate the highest with gene expression in ciliated cells. In all box plots, the central line and the notch are the median and its approximate 95% confidence interval, the box shows the interquartile range and the whiskers are extreme values upon removing outliers.

165

166

### 167 Temporal reduction of myeloid subsets immediately after viral exposure

168 To investigate the potential role of professional antigen presenting cells in the early immune response  
169 to SARS-CoV-2, we next focused on changes in the nasopharyngeal resident and circulating myeloid  
170 compartments during sustained infection. In contrast to most tissue-resident immune cells, myeloid  
171 subsets largely decreased in frequency at the site of inoculation by day 3 after inoculation (**Fig. 2b**). In  
172 particular, monocytes and activated DCs (also often referred to as migratory DCs, mature DCs or DC-  
173 LAMPs) were significantly reduced at the site of inoculation at day 3, consistent with the known role  
174 of activated DCs in trafficking viral antigens to lymph nodes for presentation. After day 3, the number  
175 of DCs at the site of inoculation increased above baseline, along with the global immune infiltration  
176 associated with peak viral load in sustained infections. We also observed a decrease of circulating  
177 myeloid cells during infection across all subsets at day 7, consistent with continued migration of these  
178 cells into inflamed or lymphoid tissues at that time point. However, already at day 3 post-inoculation  
179 there are strong decreases in some circulating myeloid subsets, where inflammatory monocytes (*IL1B*,  
180 *IL6* & *CXCL3* high) and cDC3s (monocyte-like dendritic cells) appear to migrate out of the circulation.

181

182 Strikingly, while all of the above mentioned responses were only observed in sustained infections, the  
183 significant decrease in inflammatory monocytes was also observed in transient and abortive infections  
184 (**Fig. 2c** and **Extended Data Fig. 7a**). This suggests that circulating inflammatory monocytes are able

185 to immediately respond to SARS-CoV-2 exposure, even if the viral infection is rapidly terminated,  
186 implying that exposure alone in the absence of virologically-confirmed infection with SARS-CoV-2  
187 can result in a detectable (but restricted) immune response.

188

### 189 **Novel MAIT subset is activated in both sustained and abortive infections**

190 We next asked if such a detectable immune response across all infection groups could also be observed  
191 in other cell types. When annotating unconventional T cells, we noted that MAIT cells could be further  
192 divided into two subgroups, i.e. classical MAIT cells and activated MAIT cells with higher expression  
193 of cytotoxicity and activation markers such as *PRF1* and *CD27* (**Extended Data Fig. 6b**). These  
194 markers have previously been shown to be indicative of TCR-independent activation<sup>9</sup>. At day 3 post-  
195 inoculation, we observed near complete activation of the entire MAIT cell population in the blood in  
196 sustained infections (**Fig. 2d**). Strikingly, the activation of MAIT cells was also present in abortive and  
197 transient infections, which suggests that MAIT cells may rapidly sense exposure to a virus. Thus both  
198 MAIT cells and inflammatory monocytes might play a key role in the immediate response to SARS-  
199 CoV-2. This further supports the notion that viral exposure, that does not lead to a sustained infection  
200 and subsequent COVID-19, can still induce a detectable, yet restricted immune response.

201

### 202 **Infection in epithelial and immune cells peaks a week after exposure and is most active in** 203 **ciliated cells**

204 To study how the observed immune responses relate to viral infection dynamics, we included the SARS-  
205 CoV-2 ssRNA genome and its transcripts in our analyses. This allowed us to quantify virions and viral  
206 gene expression alongside transcriptome dynamics of infected host cells. As expected, infected cells  
207 were almost exclusively found in the nasopharynx of participants with sustained infections (2505 out  
208 of 2512 cells with viral RNA). We detected infection of multiple cell types at day 5 post-inoculation,  
209 which peaked at day 7 (**Fig. 2e**), followed by a rapid decrease at day 10 - 14 post-inoculation, showing  
210 the narrow time window over which SARS-CoV-2 virion production occurred. These changes over time  
211 were in line with qPCR results (**Extended Data Fig 1b,c and Extended Data Table 1a,b**), albeit with  
212 the latter being more sensitive. Interestingly, we observed viral reads in both immune and epithelial  
213 cells in the nasopharynx (**Fig. 2e**). In contrast to previous studies, we detected large numbers of SARS-  
214 CoV-2-containing CD8<sup>+</sup> T cells, possibly due to the model being able to capture the narrow time  
215 window in which these infected cells are highly abundant. Our results therefore show that, in addition  
216 to epithelial cells, viral transcripts are also detectable at high levels in tissue-resident CD8<sup>+</sup> T cells.

217

## 218 **Non-productive SARS-CoV-2 infection of immune cells**

219 Having identified infected cells, we next asked if these represented productive infections. Because  
220 SARS-CoV-2 has a polyadenylated ssRNA genome, we were able to detect both viral transcripts and  
221 genomes, which allowed us to separate non-productive and productive infections. In non-productive  
222 infections only viral genomes would be present, leading to a fairly uniform distribution of detected viral  
223 RNAs over the length of the viral genome (slightly 5' biased due to 5' tag sequencing). In contrast, a  
224 productive infection requires viral transcription which is known to be highly biased towards the 3' end  
225 of the viral genome<sup>10</sup>. We observed that the viral RNA found in infected ciliated and goblet cells mainly  
226 originated from the 3' end where most genes are encoded, while viral RNA was uniformly distributed  
227 across the genome in infected CD8+ T cells and macrophages (**Fig. 2f**). This suggests that immune cells  
228 are not permissive for or are capable of preventing viral transcription and subsequent replication after  
229 entry into the immune cell, while goblet and ciliated cells are susceptible to proliferative viral infection.  
230 While the detection of inactive SARS-CoV-2 in macrophages could be a consequence of the engulfment  
231 of virions and infected cells, it is unclear how infection of tissue-resident CD8+ T cells is achieved and  
232 how this affects their function.

233

## 234 **Hyper-infected ciliated cells are the main source of SARS-CoV-2 and produce anti-** 235 **inflammatory molecules**

236 Based on the detection of productive viral infections in ciliated and goblet cells, we sought to identify  
237 the cells that contributed the most to viral spread. We noticed a small but distinct cluster of ciliated cells  
238 with an extremely high viral load (**Fig. 2h, Extended Data Fig. 2b**), in which we detected >1000 viral  
239 RNAs per cell on average. Other infected cells typically contained <10 detectable viral RNAs.  
240 Strikingly, while this hyper-infected subcluster of ciliated cells represents only 4% of all infected cells,  
241 they contained 67% of all detectable viral RNA, uncovering an important role for this subset of ciliated  
242 cells in fueling the viral spread.

243

244 To investigate how the varying amounts of virus per cell affects host cell gene expression, and *vice*  
245 *versa*, we correlated the amount of viral RNA with the expression of host genes. This revealed that  
246 ciliated cells exhibited a unique response to high viral amounts, upregulating AP1 and NFkB signaling,  
247 and multiple genes with known anti-inflammatory functions such as ERG1<sup>11</sup>, NFKBIA<sup>12</sup>, GDF15<sup>13</sup>,  
248 HES1<sup>14</sup>, PER1<sup>15</sup>, TNFAIP3<sup>16</sup>, and NR4A1<sup>17</sup> (**Fig. 2i**). This suggests that SARS-CoV-2 is capable of  
249 inducing an unique response state in hyper-infected ciliated cells that is in part anti-inflammatory,  
250 possibly to enhance viral spread and survival. This is further supported by the attenuation of the  
251 interferon response in hyper-infected ciliated cells (**Extended Data Fig. 6c**).

252

## 253 **Ciliated cells exhibit multiple temporally restricted response states resulting in MHC class II** 254 **presentation on infected cells**

255 To further investigate the role of ciliated cells in the local response to SARS-CoV-2 infection, we  
256 delineated the ciliated cell compartment into five distinct cell states. In addition to the above-mentioned  
257 interferon-stimulated, infected, and hyper-infected clusters, we detect a relatively abundant subset of  
258 ciliated cells with high expression of acute-phase response (APR) genes such as SAA1. Interestingly,  
259 these APR+ ciliated cells are present before inoculation, and increase the day after inoculation to up to  
260 50% of all ciliated cells in participants with sustained infections (**Fig. 2g**). At day 3 post-inoculation,  
261 interferon-stimulated ciliated cells emerge and peak at day 5, at which time point APR+ ciliated cells  
262 have disappeared completely. At day 5, infected and hyper-infected ciliated cells start appearing, which  
263 peak at day 7 post-inoculation. At day 10-14, interferon-stimulated cells decrease but remain higher  
264 than baseline, while APR+ ciliated cells reemerge. Of note, APR+ ciliated cells are also immediately  
265 upregulated in abortive but not transient infections, while all other ciliated cell states are uniquely  
266 present in sustained infections only (**Extended Data Fig. 7c**).

267

268 Together, this underscores the highly dynamic nature of the ciliated cell compartment, and uncovers a  
269 potential early response role for APR+ ciliated cells. Interestingly, infected but not hyper-infected,  
270 ciliated cells also activate APR genes, and both APR+ and APR+/SARS-CoV-2 infected ciliated cells  
271 express MHC class II (**Extended Data Fig. 6c**). While epithelial cells normally only express MHC  
272 class I to present antigens to CD8+ T cells, there is evidence that viral infection can also induce MHC  
273 class II expression in epithelial cells<sup>18</sup>, despite classically being thought to be an exclusive feature of  
274 professional antigen presenting cells. The colocalization of MHC class II+ ciliated cells with CD4+ T  
275 helper cells has also previously been reported<sup>19</sup>. This therefore raises the possibility that MHC class II  
276 expression in infected ciliated cells could allow these cells to present SARS-CoV-2 antigens to antigen-  
277 specific CD4+ T cells.

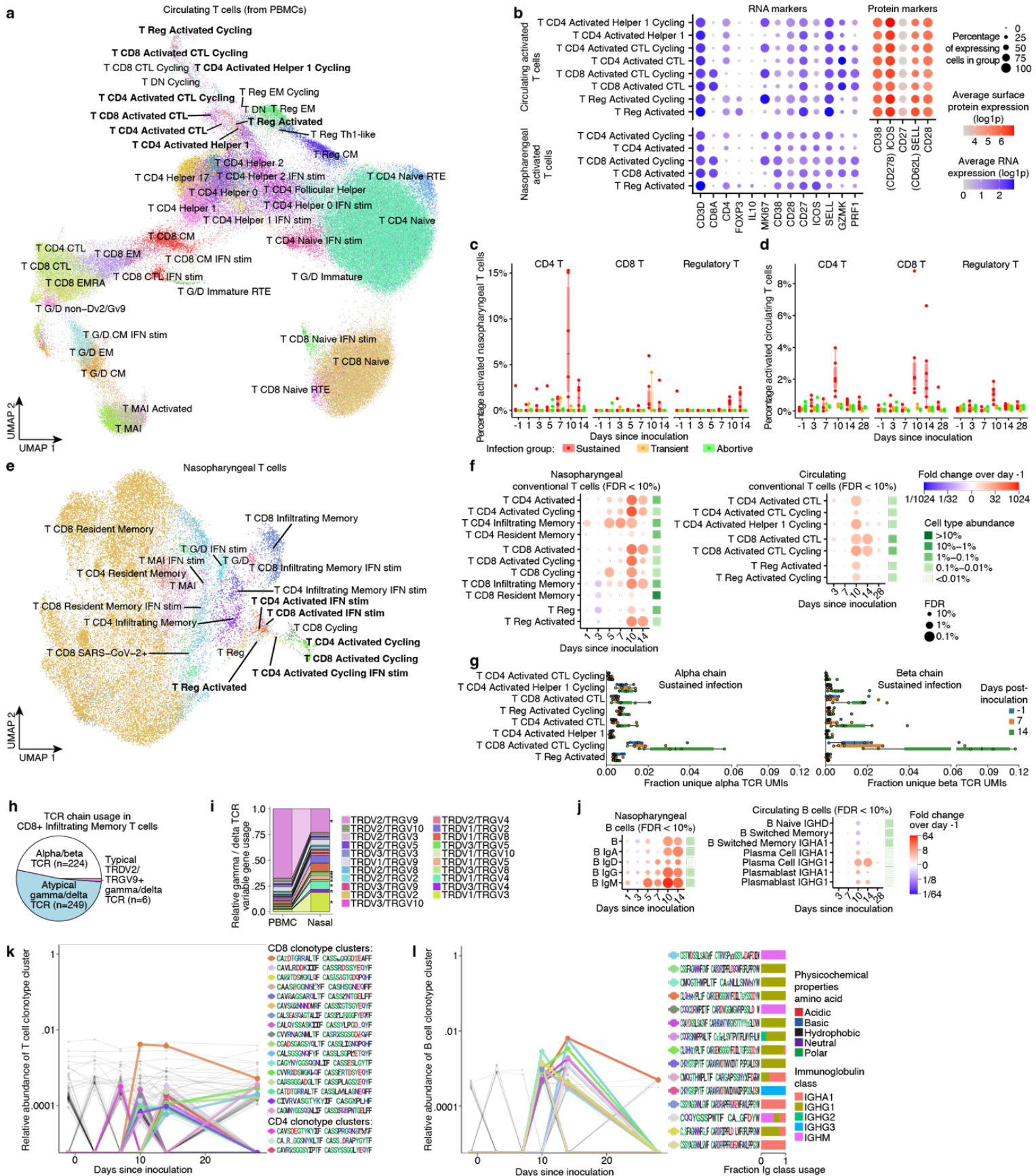
278

## 279 **Identification of activated T cells**

280 To investigate the anatomic and temporal distribution of CD4+ and CD8+ T cells following infection,  
281 we annotated the T cell compartment in the blood and nasopharynx at high resolution into 54 distinct T  
282 cell states (**Fig. 3a,e**). Strikingly, these included subtypes of CD4+, CD8+ and regulatory T cell states  
283 that highly expressed T cell activation markers such as *CD38*, *CD28*, *CD27* and *ICOS* (**Fig. 3b**). While  
284 we and others have previously been unable to separate T cells that become activated during SARS-  
285 CoV-2 infection without enrichment experiments, we detected these activated T cells as distinct clusters



286 in both the circulating and nasopharyngeal T cell compartments. Reassuringly, many nasopharyngeal  
287 and circulating activated T cells expressed the same TCR sequences (**Extended Data Fig. 6d**), showing  
288 that they originated from the same clones found both in circulation and nasopharynx as a response to  
289 infection. In addition, the immune repertoires of activated T cells were significantly more restricted and  
290 clonal than other mature T cell types (**Extended Data Fig. 6k**), suggesting that they were activated and  
291 expanded in a TCR- and antigen-specific manner. As expected from activation through TCR signaling,  
292 we also detected high frequencies of cycling T cells within the activated T cell compartment. Of note,  
293 we also detected many activated T cells that were not cycling, as well as cycling T cells that did not  
294 appear to be activated, implying that our activation signature was at least partially independent of the  
295 cell cycle gene signature. To test if these newly identified activated T cells are antigen-specific and can  
296 recognize SARS-CoV-2 peptides, we performed peptide-MHC-I stainings on PBMCs using DNA-  
297 barcoded Dextramers loaded with SARS-CoV-2 antigens to detect peptide-MHC-I binding in parallel  
298 with scRNAseq and scTCRseq. These experiments reveal that activated T cells are significantly  
299 enriched and indeed specifically bind SARS-CoV-2 peptides compared to unmatched peptide-MHC-I  
300 molecules (**Extended Data Fig. 7d-f, 8b**). Together, the identification of activated T cells and their  
301 transcriptome signature in unsorted PBMC and tissue samples presents a unique opportunity to study  
302 the T cell response to SARS-CoV-2 in unprecedented detail.



**Figure 3: Adaptive immune responses emerge at day 10 post-inoculation.** (a) UMAP of all circulating T cells, highlighting the distinct cluster of activated T cells. Cells are color coded and labeled by their detailed cell state annotation. (b) Marker gene and protein expression of activated T cell subsets are shown in blue and red, respectively. (c) Percentages of nasopharyngeal T cells that were annotated as activated T cells, split over days since inoculation and color coded by infection group. (d) Boxplot as in (c), but showing circulating activated T cells. (e) UMAP as in (a), but showing nasopharyngeal T cells. (f) Fold changes in cell state abundance compared to pre-inoculation of nasopharyngeal and circulating conventional T cells are shown in the left and right plots, respectively. Only cell states that significantly change at a FDR < 10% at least one time point are shown. Nasopharyngeal T cell abundances were scaled to the total amount of detected epithelial cells. Fold changes and significance were calculated by fitting a GLMM as shown in *Figure 1*. The mean cell type proportions over all cells and samples is shown in the green heatmap right of the dotplot to aid the interpretation of changes in cell type abundances. (g) TCR clonality and expansion at day 14 of activated TCRs was validated using bulk TCR sequencing. For TCRs that matched the single cell gene expression, normalized clonality TCR alpha (left) and beta (right) data is separated by type and expressed as the average fraction of total clones in sample contributed by a cell of that type, with changes over time implying clonal expansion or contraction. For activated T cell types of interest, scatterplots for each sustained infection and at each time point sampled (days -1, 7, 14) are drawn. (h) Proportion of CD8+ infiltrating T cells that use alpha/beta TCRs, typical Dv2/Gv9 g/d TCRs, or atypical g/d TCRs is shown. (i) The relative immune repertoire composition of g/d T cells in circulation and nasopharynx after challenge are shown in the left and right bars, respectively. G/d chain pairs that are significantly more or less abundant between circulation and nasopharynx ( $p < 0.05$ ) are highlighted with an asterisks. (j) Dotplot as in (f), showing the fold changes in B cells. Legend for significance and mean cell type proportions as in (f). (k) Abundance of TCR clusters relative to all TCRs are shown over time since inoculation. Activated TCR clusters are color coded and their TCR motifs are shown. Legend for the physicochemical properties of amino acids in shown TCR motifs is shown in panel (l). (l) Plot as in (k), but showing BCR clusters. Immunoglobulin class usage within each activated BCR cluster is shown in the rightmost bars.

304

305

### 306 **Activated T cells expand and peak ten days after inoculation**

307 To better understand the characteristics of the activated T cells described above we quantified their  
308 abundance over time and across infection groups (**Fig. 3c-d,f**). This revealed highly significant  
309 expansions of activated CD4+ and CD8+ T cells peaking in both blood and nasopharynx at day 10 after  
310 inoculation. This expansion was highly time-restricted, only appearing in the circulation after day 7 and  
311 contracting rapidly thereafter. While this decrease meant that activated T cells were barely detectable  
312 at day 28 post-inoculation, the associated TCR clonotypes in circulation could still be identified, having  
313 transitioned into memory and effector T cells (**Extended Data Fig. 6e**). We integrated our single cell  
314 resolved T cell data with highly sensitive bulk TCR sequencing from the blood to validate that activated  
315 T cell-associated TCR sequences indeed clonally expand after day 7 post-inoculation in sustained (**Fig.**  
316 **3g**) but not in abortive infections (**Extended Data Fig. 6f**). The emergence of these cells at day 10 after  
317 inoculation closely resemble the temporal dynamics of a typical antigen-specific adaptive immune  
318 response to vaccination and infection. At this time point we also observed clearance of detectable virus  
319 and a reduction of IFN stimulation in the nasopharynx, suggesting that the onset of an adaptive T cell  
320 response is associated with clearance of the infection. Importantly, activated T cells emerged in all

321 sustained infected participants, but in none of the abortive infections, underscoring their specificity to  
322 infection. We did however detect a small increase of activated T cells in the nasopharynx of two of the  
323 three transiently infected individuals (**Fig. 3c**), which might suggest that a smaller T cell response can  
324 be established without going through a sustained infection.

325

326 In contrast to activated CD4<sup>+</sup> and CD8<sup>+</sup> T cells whose infiltration peaked at day 10, the amount of  
327 activated regulatory T cells was highest at day 14 at the site of infection (**Fig. 3c**), where they strongly  
328 upregulated expression of the anti-inflammatory cytokine IL-10 (**Fig. 3b**). This peak of activated  
329 regulatory T cells coincided with resolution of the observed global immune infiltrate (**Fig. 1f**) and  
330 downregulation of the IFN-stimulated response (**Fig. 2a**), suggesting a role for these regulatory T cells  
331 in suppressing further local inflammation after the infection has been cleared.

332

333 Interestingly, the time window during which activated CD8<sup>+</sup> T cells were increased was broader in  
334 blood (**Fig. 3d**), while activated CD4<sup>+</sup> T cells were detected for longer in the nasopharynx (**Fig. 3c**). In  
335 addition, activated CD4<sup>+</sup> T cells were also significantly more abundant at the site of infection where  
336 they represent up to 15% of all nasopharyngeal-resident T cells at day 10 after inoculation. The  
337 predominance of activated CD4<sup>+</sup> T cells in the respiratory mucosa was surprising, as CD8<sup>+</sup> T cells are  
338 classically understood to be the major effectors in the local cytotoxic response. These results suggest  
339 that CD4<sup>+</sup> T cells may play an unexpected and important role as local effectors.

340

#### 341 **Activated CD4<sup>+</sup> T cells express cytolytic proteins**

342 Activated CD4<sup>+</sup> T cells express high amounts of cytotoxicity genes (e.g. *PRFI*, see **Fig. 3b** and  
343 **Extended Data Fig. 3a & 4a**) that are normally expressed in NK and CD8<sup>+</sup> T cells. As CD4<sup>+</sup> T cells  
344 can only recognise antigens in MHC class II context that is normally exclusive to professional antigen-  
345 presenting immune cells, the function and relevance of cytotoxic CD4<sup>+</sup> T cells remains poorly  
346 understood. However, several studies have reported their emergence during the adaptive immune  
347 response against SARS-CoV-2<sup>20,21</sup>, and they have been reported to have a specific and antiviral effector  
348 function in influenza challenge models<sup>22</sup>. It is therefore conceivable that the activation of cytotoxic  
349 CD4<sup>+</sup> T cells upon SARS-CoV-2 infection could potentially be a response to MHC class II presentation  
350 by infected ciliated cells (**Extended Data Fig. 6c**). This would potentially enable antigen-specific  
351 destruction of infected ciliated cells by CD4<sup>+</sup> T cells (**Fig. 2g**).

352



### 353 **Atypical g/d T cells infiltrate site of infection and dominate the g/d T cell response**

354 In the nasopharynx, we also detected a subset of CD4<sup>+</sup> and CD8<sup>+</sup> T cells lacking both activation and  
355 tissue-residency markers (such as *ITGAE* or CD103, **Extended Data Fig. 3a**) which appeared in the  
356 nasopharynx during sustained infections and which were annotated as “infiltrating memory T cells”  
357 (**Fig. 3e**). We noticed that infiltrating CD8<sup>+</sup> T cells expressed relatively few detectable alpha/beta TCRs  
358 and had heterogenous CD8 expression, similarly to the gamma/delta (g/d) T cells that we had already  
359 detected (**Extended Data Fig. 2a**). To investigate if this infiltrating subset harbors a distinct g/d T cell  
360 population, we performed targeted single cell sequencing of the g/d TCR genes in the nasopharynx and  
361 blood. This revealed that infiltrating CD8<sup>+</sup> T cells indeed predominantly express g/d TCRs (**Fig. 3h**).  
362 As expected, the g/d TCR repertoire found in circulating blood cells consists mostly of TCR chains  
363 containing variable segments TRDV2 and TRGV9 (**Fig. 3i**). Strikingly, the nasopharynx is significantly  
364 depleted for TRDV2/TRGV9<sup>+</sup> T cells, with other variable segments dominating the g/d TCR repertoire.  
365 More than 97% of the g/dTCR expressing infiltrating CD8<sup>+</sup> T cells express these rare non-  
366 TRDV2/TRGV9 TCRs (which we termed atypical g/d T cells; **Fig. 3h**), which means that the atypical  
367 g/d T cell response is four times more abundant than the typical g/d T cell response. While the exact  
368 function of atypical g/d T cells is still poorly understood, their timing alongside other adaptive immune  
369 responses and its restriction to sustained infections, suggests that they might play an underappreciated  
370 role in the immune response against SARS-CoV-2 infection.

371

### 372 **Antibody secreting B cells clonally expand ten days after exposure**

373 Given that the strong T cell response that appears highly time restricted to day 10 post-inoculation, we  
374 hypothesized that there should be a B cell response at a similar time point. To test this, we investigated  
375 the temporal and cell state dynamics of the B cell response to SARS-CoV-2 inoculation. We detected  
376 distinct subtypes of naive, memory and antibody-secreting B cells (plasmablasts and plasma cells), and  
377 used the BCR data to distinguish immunoglobulin class and isotype switching (**Extended Data Fig.**  
378 **4b**). In line with the observed T cell response, we observe a strong and highly time restricted B cell  
379 response from day 10-14 after SARS-CoV-2 exposure (**Fig. 3j**). In blood, this response includes a clear  
380 switch from naive and IgG/IgA memory B cells to mostly IgG1 and some IgA1 secreting plasmablasts  
381 and plasma cells. IgA1 and IgG1 are expected to be the dominant antibody immunoglobulin classes in  
382 blood<sup>23</sup>, and the timing of production of antibodies is in line with B cell responses observed in  
383 vaccination studies<sup>24</sup>, suggesting that these antibody secreting B cells at day 10 after inoculation are  
384 SARS-CoV-2 specific. While numbers of detected B cells in the nasopharynx are limited, we also  
385 observe significant infiltration of both IgA<sup>+</sup> and IgG<sup>+</sup> B cells into the nasopharynx from day 10 post-  
386 inoculation (**Fig. 3j**), indicating that the B cell response leads to antibody production at the site of  
387 infection. Together, these findings suggest that it takes ten days from SARS-CoV-2 exposure for the

388 adaptive immune response to mature and expand to detectable abundances. Importantly, we show a  
389 concerted adaptive immune response of B and T cells at both local and systemic level, which is  
390 facilitated by antibody secreting and activated lymphocytes.

391

### 392 **Cell2TCR: Identification of clonotype groups that are likely SARS-CoV-2 specific**

393 We next set out to leverage the transcriptomically distinct B and T cell states that are associated with  
394 the adaptive immune response, to identify BCR and TCR clonotypes that specifically recognise SARS-  
395 CoV-2 (see *Methods* for details). We designed a cell state-driven approach that enabled us to detect the  
396 permitted divergence between TCR or BCR sequences in an antigen-specific response. To this end, we  
397 quantified the inclusion of naive B and T cells, and the mixing of CD8+ and CD4+ T cells, to quantify  
398 the impurity of TCR and BCR clonotype groups (groups of B and T cells that express highly similar  
399 BCR or TCR sequences). We next selected activated clonotype groups that appear to expand in an  
400 antigen-specific manner (i.e. express multiple independent but highly similar TCR/BCR sequences in  
401 activated T cells or antibody-secreting B cells). Reassuringly, this clonotype selection method  
402 exclusively yields activated clonotypes in participants with sustained infections (**Extended Data Fig.**  
403 **6g-h**). In total, we detect 20 activated TCR clonotype groups and 15 activated BCR clonotype groups  
404 in the six participants with sustained infections (**Fig. 3I**). These clonotype groups first emerge after 1  
405 week, most appear at day 10, and some remain detectable at day 28 post-inoculation. When we applied  
406 Cell2TCR on all activated CD8+ T cells as well as all HLA-matched CD8+ T cells from the Dextramer  
407 assay, we found 14 clonotype groups that contained cells from both datasets, validating the specificity  
408 and annotating the recognized antigen of these clonotype groups (**Extended Data Table 1c**).

409

410 Interestingly, even at the peak of expansion at day 10 post-inoculation, all but one of the activated  
411 clonotype groups have only very low abundance (<0.001% of all T cells), at the detection limit of single  
412 cell genomics approaches. Such low prevalence makes activated clonotypes difficult to detect and  
413 distinguish from bystander cells when simply performing enrichment analysis of the entire immune  
414 repertoire between healthy and infection samples. This highlights the importance of considering single  
415 cell phenotypes in VDJ analyses and the usefulness of our newly identified activated T cell state.

416

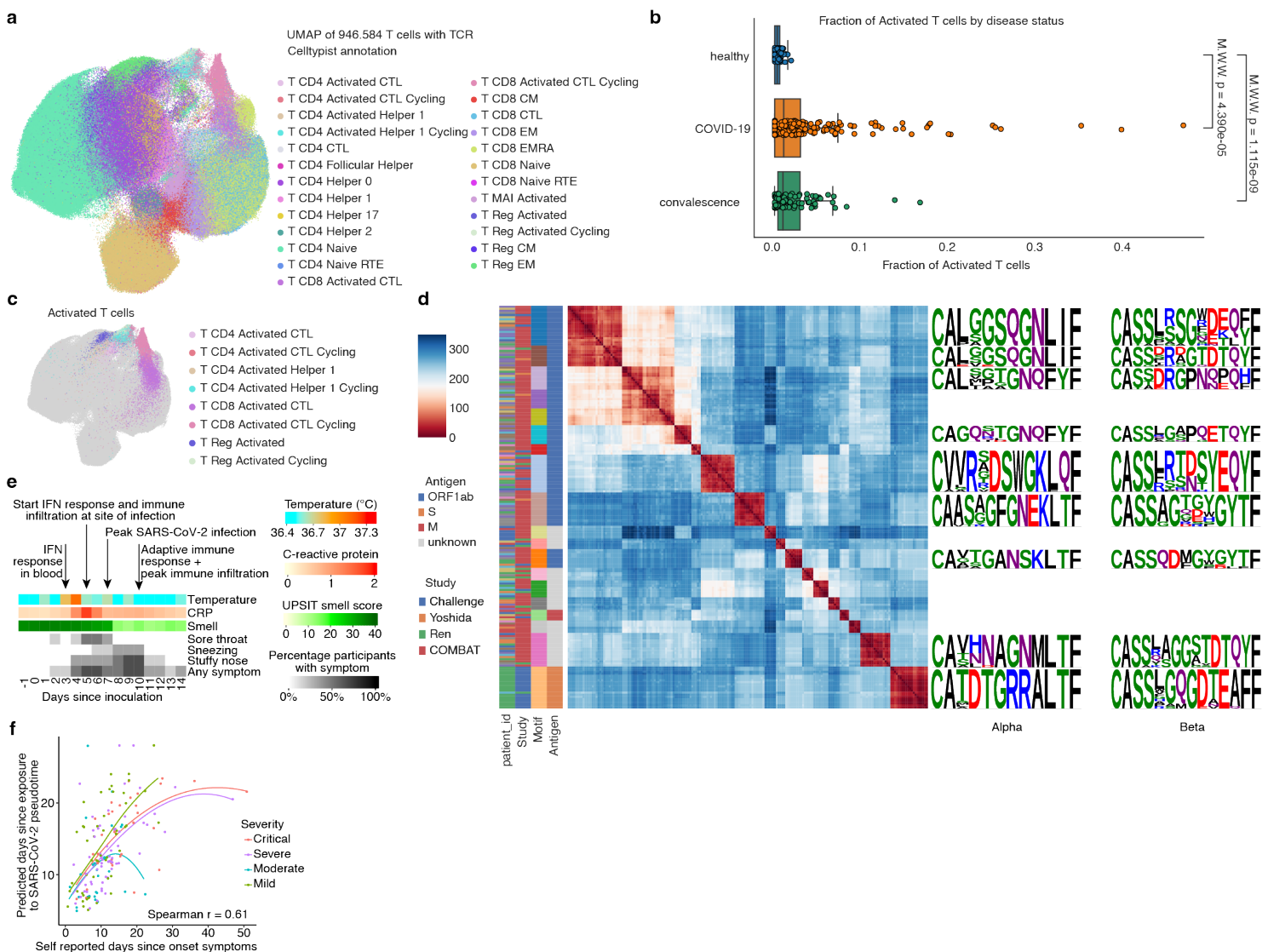
417 Importantly, in contrast to activation or enrichment assays that require *in vitro* incubation with  
418 antigens<sup>25,26</sup>, our Cell2TCR approach for detecting clonotypes that are activated in a disease of interest  
419 is not restricted and biased towards known antigens. Hence, it can be applied to any infection,  
420 inflammatory disease or cancer single cell RNA- and VDJ-seq dataset to extract paired chains  
421 recognising antigens.



422

## 423 Integrating COVID-19 patient data uncovers public SARS-CoV-2 TCR motifs

424 We hypothesized that our deep characterization of the adaptive immune response in PBMCs could be  
 425 leveraged when analyzing patient COVID-19 data, in particular to study activated T cell states and  
 426 associated SARS-CoV-2 specific TCR repertoires. To this end, we integrated our data together with  
 427 single cell RNA-seq data from five large-scale studies that profiled PBMC samples using a deep  
 428 generative model (scVI variational autoencoder, see *Methods*), and obtained just short of one million T  
 429 cells from several hundred individuals, including over 240 acute COVID-19 patients (**Extended Data**  
 430 **Table 1d,e**). We next projected our highly detailed cell type annotation, including the activated T cell  
 431 states, onto the patient data (**Fig. 4a and 4c**). This revealed that activated T cells are also present in  
 432 COVID-19 patients that were sampled outside a viral challenge setting, and that these activated subsets  
 433 also form distinct clusters of cells within the T cell compartment. Importantly, the fraction of activated  
 434 T cells was significantly higher in COVID-19 patients and convalescence samples compared to healthy  
 435 controls, underscoring their involvement in the immune response to COVID-19 (**Fig. 4b**).



436

**Figure 4: Integrating COVID-19 patient data reveals public SARS-CoV-2 TCR motifs**

(a) UMAP representation after integration of five COVID-19 patient datasets with paired RNA and VDJ sequencing data. Cell type labels inferred using a logistic regression classifier (Celltypist) trained on manual annotations of PBMCs from the current work. (b) Fraction of activated T cells across all T cells in sample for COVID-19 (n = 240), convalescent (n = 82) and healthy (n = 88) samples of five COVID-19 patient datasets. Significance levels after Mann-Whitney testing are shown and indicate that COVID-19 and convalescent samples have significantly more activated T cells than healthy samples. (c) Activated T cell types highlighted on UMAP representation from panel (a). Activated CD8<sup>+</sup> T cells were most abundant, followed by CD4<sup>+</sup> and regulatory types, and clustered together in a distinct area of the latent space. (d) Clustermap of pairwise TCR distances with color-coded information for each TCR on patient\_id, study, motif, antigen on the left-hand side, as well as the sequence logos for the nine most common motifs on the right-hand side. Each column/row corresponds to a unique TCR, and the distance to each TCR in the set is indicated by color. Only activated T cells with public motifs (identified in more than one individual) were considered. Low distances indicate similar TCRs, with distances of 40 and less potentially yielding TCRs recognising the same epitopes. For sequence logos, letter height indicates frequency of AA at that position across T cells pertaining to the motif. AAs are colored by side chain chemistry: Acidic (red), basic (blue), hydrophobic (black), neutral (purple), polar (green). AA: amino acid. (e) Recorded symptoms averaged over SARS-CoV-2 challenge participants with sustained infection for days -1 to 14 post-inoculation. Major molecular events in the immune response are highlighted with arrows. (f) Predicted time since viral exposure is plotted against reported time since onset of symptoms. Lines represent loess fits of the data split and color coded by reported severity.

437

438

439 We then employed our cell state aware clonotype group selection approach (Cell2TCR) to identify  
440 activated clonotypes, which resulted in 254 COVID-19 associated clonotype groups (**Extended Data**  
441 **Table 1f**). Strikingly, 211 of these activated clonotype groups were shared between patients (largest  
442 groups shown in **Fig. 4d**), highlighting the antigen-specificity of this approach, with the two most  
443 common motifs being shared by 13 individuals each. This also implies that a relatively small set of  
444 highly immunogenic SARS-CoV-2 peptides results in most of the T cell responses in COVID-19.  
445 Finally, we wanted to validate the antigen-specificity of the COVID-19 associated clonotype groups  
446 that we found in the public and challenge study data. Thus we intersected the CDR3 amino acid  
447 sequences with databases containing experimentally validated SARS-CoV-2 specific TCRs (see  
448 *Methods*). Importantly, this revealed that activated clonotype groups, including groups that contain  
449 TCRs from this study, are 3.75 fold enriched ( $p = 1.68 \times 10^{-21}$ ) for SARS-CoV-2 specific TCRs. This  
450 provides strong validation for activated T cells indeed representing the antigen-specific T cell response  
451 against SARS-CoV-2 (**Extended Data Fig. 6i**). Most of the activated T cell clonotype groups recognise  
452 viral proteins encoded by ORF1ab, but we also find Membrane and Spike specific TCR clonotype  
453 groups. Because our cell state aware clonotype selection method identifies SARS-CoV-2 specific TCRs  
454 without any prior antigen information, our results may also include TCRs that recognise SARS-CoV-2  
455 antigens that have not yet been tested. Together, these results validate the specificity of the adaptive

456 immune response that we observed at day 10, and highlight the power of defining activated T cells for  
457 detecting disease-specific antigens in an unbiased manner.

458

### 459 **Molecular responses precede and are dynamic during clinical manifestations**

460 Last, we wanted to investigate how our single-cell resolved timeline of immune responses relates to  
461 clinical manifestations that are typically observed and measured in COVID-19 patients. The unique  
462 experimental setting of our human challenge model enabled us to collect highly detailed and time-  
463 resolved clinical data for all participants that were profiled using single cell genomics approaches. The  
464 timing of the most relevant and dynamic COVID-19 symptoms show that even the earliest symptoms  
465 appear mostly at day 4 post-inoculation (**Fig. 4e**), which is later than some of the molecular responses  
466 that we described. Here, the upregulation of APR in ciliated cells, the activation of MAIT cells,  
467 depletion of some myeloid cells, and the global activation of IFN signaling in blood are observed before  
468 or at day 3 post challenge (**Fig. 2**). In contrast, a slight rise in temperature is only significantly detectable  
469 at day 4 post-inoculation ( $p = 5 \times 10^{-6}$ ), at which early upper airway-related symptoms such as a stuffy  
470 nose and a sore throat also appear. This is then followed by global immune infiltration and activation  
471 of IFN signaling at the site of infection at day 5, which is also the first time that we detected infected  
472 cells. This coincides with a slight increase in C-reactive protein (CRP) in blood ( $p = 0.04$ ). At day 7  
473 post-inoculation we observed that the amount of detectable infected cells peaked. Interestingly, from  
474 day 8 on we also observed that all but one of the participants with a sustained infection significantly  
475 lost their sense of smell ( $p = 0.004$ ), together with aggravation of sneezing and a stuffy nose. This was  
476 followed by a strong reduction in the amount of infected cells at day 10 and a peak in the amount of  
477 nasopharyngeal immune infiltration, which coincides with the onset and expansion of an adaptive  
478 immune response and clearance of most symptoms. In summary, we observe that clinical manifestations  
479 and different waves of immune responses dynamically change over time, which can aid the molecular  
480 interpretation of COVID-19 based on clinical observations and improves our understanding of the  
481 therapeutic time windows in this disease.

482

### 483 **Human COVID-19 challenge data as a reference atlas for cell dynamics**

484 To maximize the impact of our time-resolved COVID-19 dataset, we build predictive models to infer  
485 time since SARS-CoV-2 exposure. We used Gaussian process regression and latent variable models to  
486 fit the changes in cell state composition during sustained infection. We next applied these predictions  
487 to publicly available PBMC single cell RNA-seq datasets from 361 COVID-19 samples, to infer at  
488 which stage of the immune response each patient was and to predict when this patient was exposed to  
489 SARS-CoV-2. Reassuringly, our Gaussian processes based time inference model predicts that the time  
490 since exposure and the time since onset of symptoms are highly correlated (**Fig. 4f**), and that exposure

491 is predicted to precede onset of symptoms, as expected. Interestingly, the predicted difference between  
492 exposure and symptoms decreases with increased severity (**Extended Data Fig. 6j**), where patients  
493 with more severe COVID-19 are predicted to be in the adaptive immune reaction phase for longer.  
494 While this suggests that patients with severe COVID-19 take longer to clear the virus, it could also  
495 indicate that the cellular composition and immune response timeline in severe cases is perturbed  
496 compared to the relatively mild cases observed in the challenge study. In addition to a temporal model  
497 that could improve the assessment of the disease stage of COVID-19 patients, we also provide  
498 annotation models for a total of 202 cell states including new temporal and rare cell states. These models  
499 are now included in the default models at Celltypist.org, and enable highly detailed cell type annotation  
500 without the need for bioinformatics expertise. In addition, our single cell expression data is freely  
501 available at our COVID19CellAtlas.org web portal for online exploration and analysis.

502

## 503 **Discussion**

504 Our findings have implications for the COVID-19 and broader infection diseases community. We detect  
505 multiple response states that precede the onset of clinical manifestations, including the activation of  
506 MAIT cells and decreases in inflammatory monocytes. Importantly, this represents a newly discovered  
507 immune response that emerges when an individual is exposed to SARS-CoV-2, but manages to prevent  
508 the onset of viral spread. These features of very early and abortive infections can be used as biomarkers  
509 and help to understand the immediate immune response upon viral exposure. In addition, we discovered  
510 a distinct cell state for activated conventional T cells that harbor SARS-CoV-2 specific TCRs, and we  
511 show that this signature can be projected onto patient cohort data to unravel the disease specific T cell  
512 response.

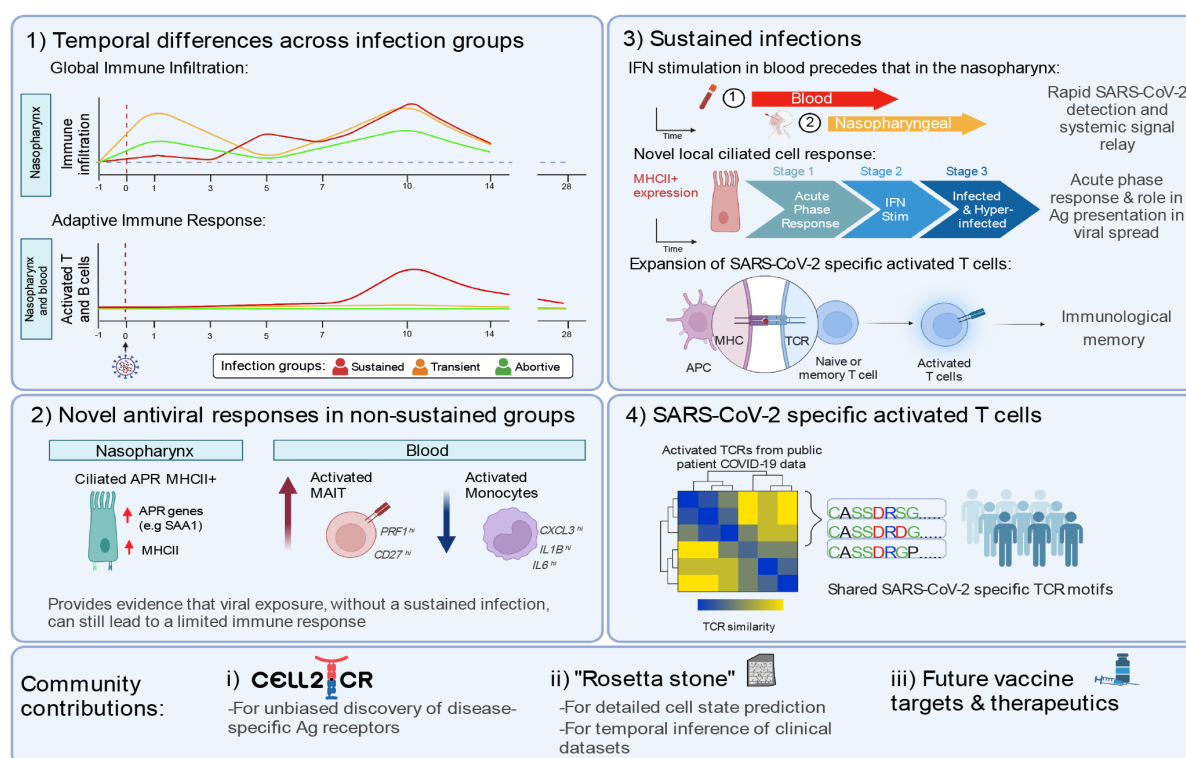
513

514 The timing of our challenge experiments in the early stages of a pandemic with a new virus - before  
515 most of the population acquired immune memory through natural infections and vaccine rollout -  
516 enabled us to recruit and study immune responses in adult participants that were completely naive to  
517 this pathogen. The resulting unique data will be essentially impossible to replicate in future efforts as  
518 the population builds memory to many SARS-CoV-2 strains. In addition to the responses during  
519 sustained infections and COVID-19, we were also able to study abortive and transient infections that  
520 would be extremely challenging to detect outside a controlled challenge setting, revealing novel  
521 immune response signatures associated with successfully preventing sustained infections.

522

523 Our results uncovered a temporally resolved timeline of antiviral responses in epithelial and immune  
524 cells both locally at the site of infection and systemically in circulation (**Fig. 5**). During sustained

525 infections that lead to COVID-19, we observe an immediate and novel acute phase response in ciliated  
 526 cells at the site of infection. In blood, we observe two novel innate immune responses in which MAIT  
 527 cells become activated, and inflammatory monocytes migrate out of circulation. These two responses  
 528 are the only immune responses that are also observed in participants with abortive infections that did  
 529 not develop COVID-19, underscoring their importance and sensitivity, and uncovering a new and  
 530 distinct immune response to SARS-CoV-2 exposure that does not lead to COVID-19.



531

**Figure 5: Temporally resolved epithelial and immune response in SARS-CoV-2 infections**

Summary figure highlighting 1) temporal differences in the distinct infection groups, 2) novel antiviral responses, 3) novel characteristics of sustained infection, and 4) the identification of public motifs in SARS-CoV-2 specific activated T cells. In addition, our work provides community tools for inference of specific TCR motifs (Cell2TCR) in activated T cells, and for temporal assignments of clinical COVID-19 samples underpinning future therapeutic applications.

532

533

534 In sustained infections, we observe global activation of interferon signaling that affects all circulating  
 535 immune cells. Strikingly, the activation of interferon signaling in blood precedes widespread activation  
 536 at the site of inoculation, which suggests that a highly efficient relay to the systemic immune system  
 537 exists, likely through the lymphatic system. The activation of interferon signaling at day 5 to seven  
 538 post-inoculation coincides with global immune infiltration and a peak of detectable virally infected  
 539 cells. This relatively slow immune response at the site of inoculation is in contrast to the immediate  
 540 immune infiltration that we observed in infections that were only transiently detectable. In sustained  
 541 infections, we also detect large amounts of cells containing viral RNA including infection of immune



542 cells, but we provide evidence that only epithelial cells support successful viral replication. Here, we  
543 found that a small subset of hyper-infected ciliated cells becomes anti-inflammatory and the main  
544 source of viral production.

545

546 While our experimental approach included matched preinfection samples and we profiled all expected  
547 cell types from a total of 181 samples from 16 participants, we cannot exclude the possibility that our  
548 infection group sizes remained underpowered to detect very subtle or time-restricted responses. In  
549 addition, we note that the participants enrolled in this study cleared the infection with mild symptoms.  
550 It is possible that COVID-19 patients that require hospitalization exhibit perturbed or exacerbated  
551 immune responses that were not captured in our work, which could mean that caution should be taken  
552 when extrapolating our findings to critically ill COVID-19 patients.

553

554 At day 10 post-inoculation, we detect the onset and expansion of the adaptive immune response. In  
555 addition to antibody-secreting B cells, this response also includes activated conventional T cells. This  
556 is the first time to our knowledge that these cells have been described in single cell transcriptomics  
557 assays, likely because of the limited time window in which these activated T cells are detectable. Two  
558 weeks after inoculation, the amount of activated regulatory T cells at the site of inoculation peaks, while  
559 the abundance of other immune cells normalizes again, which coincides with a near absence of any  
560 remaining infected cells. These activation states have key marker genes, and we can identify these  
561 activated CD4+, CD8+, and regulatory T cell states using machine learning models. We integrate their  
562 prediction into a computational pipeline (Cell2TCR) which includes paired chain TCR motif inference.  
563 This is a tool applicable to any single cell RNA/VDJ dataset of infection, inflammation or tumor  
564 immune response.

565

566 Together, this represents the most comprehensive and detailed time-resolved description of the course  
567 of SARS-CoV-2 infection, or any other infectious disease, and gives unique insights into responses that  
568 are associated with resisting a sustained infection and disease.



## 569 Acknowledgements

570 We are grateful to Sheila Casserly and Sujana Regmi for assistance with collecting samples, and  
571 Tarryn Porter, Agnes Oszlanczi, Yvette Wood and Sabine Eckert for library preparation. We thank Rasa  
572 Elmentaite for helpful discussions on activated T cells. We acknowledge assistance from Rea Dabelić  
573 (10X Genomics), Illumina and 10X Genomics. We are grateful for the support and guidance with  
574 MACS for the Dextramer work provided by Yanping Guo, the Flow Cytometry Translational  
575 Technology Platform Manager at UCL Cancer Institute. We acknowledge assistance provided by the  
576 University College London CL3 facility at the Paul O’Gorman building and the staff at the Sanger  
577 Institute Core Sequencing facility.

578

579 This research was funded in whole, or in part, by the Wellcome Trust Grant 206194, 220540/Z/20/A  
580 and 211276/Z/18/Z, and by Action Medical Research (GN2911, to M.Z.N. and K.B.M.). M.Z.N.  
581 acknowledges funding from a MRC Clinician Scientist Fellowship (MR/W00111X/1). M.Z.N. and  
582 J.L.B. acknowledge funding from the Rutherford Fund Fellowship allocated by the MRC UK  
583 Regenerative Medicine Platform 2 (MR/5005579/1). M.Y. was funded by The Jikei University School  
584 of Medicine and Action Medical Research (GN2911). K.B.W. acknowledges funding from University  
585 College London, Birkbeck MRC Doctoral Training Programme. L.M.D is supported by the European  
586 Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant  
587 agreement No 955321. M.N. acknowledges funding from the Wellcome Trust (207511/Z/17/Z) and by  
588 NIHR Biomedical Research Funding to UCL and UCLH. R.H. is a NIHR Senior Investigator.

589

590 For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author  
591 Accepted Manuscript version arising from this submission.

592

## 593 Author Contributions

594 M.Z.N. and S.A.T. conceived, set up, directed this study and provided funding. C.C. set up the clinical  
595 study and co-ordinated sampling. K.B.W. optimized digestion protocols, processed samples for 10X  
596 and CITEseq, isolated DNA for genotyping, performed Dextramer experiments and assisted with data  
597 analysis and interpretation. R.G.H.L. performed and led the data analysis. L.M.D. assisted with data  
598 analysis. R.G.H.L., K.B.W., L.M.D., K.B.M., M.Z.N. and S.A.T. interpreted the data and wrote the  
599 manuscript. M.Y., J.L.B., J.A.H. assisted with 10X sample processing. V.T., S.M.J. and R.C.C.  
600 provided student supervision to K.B.W. and P.M.. H.R.W. processed blood samples. P.M. collected  
601 nasopharyngeal samples for optimisation of digestion protocols. M.H., M.N. and R.H. assisted in the

602 set up of the study. B.K., M.K., A.C., A.B. and M.M. oversaw sample collection and provided clinical  
603 data.

604

605 These authors contributed equally: Rik G.H. Lindeboom, Kaylee B. Worlock

606 These authors jointly supervised this work: Marko Z. Nikolić, Sarah A. Teichmann

607

#### 608 **Data Availability**

609 The data presented in this study can be explored and analyzed interactively through our COVID-19 Cell  
610 Atlas web portal (<https://covid19cellatlas.org>), currently accessible via early-access at [http://covid19-  
611 challenge-study.cellgeni.sanger.ac.uk/](http://covid19-challenge-study.cellgeni.sanger.ac.uk/) until publication in a peer reviewed journal. The cell by feature  
612 count matrices are also available to download at the web portal. The cell state annotation models are  
613 available in the CellTypist model repository (<https://www.celltypist.org/models>). A reference for our  
614 MT-GPR model to infer time since viral exposure on PBMC data is available at our GitHub repository  
615 ([https://github.com/Teichlab/COVID-19\\_Challenge\\_Study](https://github.com/Teichlab/COVID-19_Challenge_Study)). The raw sequencing data will be made  
616 available before publication in a peer-reviewed journal via the European Genome-Phenome Archive,  
617 which will be made available under managed data access.

618

#### 619 **Code availability**

620 Cell2TCR is available at our GitHub repository (<https://github.com/Teichlab/Cell2TCR>). Code that was  
621 used for data analysis is available at our GitHub repository ([https://github.com/Teichlab/COVID-  
622 19\\_Challenge\\_Study](https://github.com/Teichlab/COVID-19_Challenge_Study)).

623

#### 624 **Competing interest statement**

625 R.G.H.L., L.M.D. and S.A.T. are inventors on a filed patent that is related to the detection and  
626 application of activated T cells. In the past three years, S.A.T. has received remuneration for Scientific  
627 Advisory Board Membership from Sanofi, GlaxoSmithKline, Foresite Labs and Qiagen. S.A.T. is a co-  
628 founder and holds equity in Transition Bio. P.M. is a Medical Research Council (MRC)-  
629 GlaxoSmithKline EMINENT clinical training fellow with project funding unrelated to the topic of this  
630 Comment and receives co-funding from the National Institute for Health Research (NIHR) University  
631 College London Hospitals (UCLH) Biomedical Research Centre. P.M. reports consultancy fees from  
632 SOBI, AbbVie, UCB, Lilly, Boehringer Ingelheim, and EUSA Pharma all unrelated to this submission.  
633 A.M., A.C., M.K., M.M. and A.B. are full time employees at hVIVO Services Ltd.

634 **Methods**

635

636 **Study participants and design**

637 16 healthy adults aged 18-30 years, with no evidence of a previous SARS-CoV-2 infections or  
638 vaccinations (seronegative), were included for this study from the wider cohort (34 participant) enrolled  
639 as part of the Human COVID-19 Challenge study, pioneered by the government task force, Imperial  
640 college london, Royal Free London NHS Foundation Trust and hVIVO<sup>6</sup>. These participants were  
641 enrolled as part of the cohort 5 and 6, from June - August 2021. Reported patient identifiers have been  
642 de-identified and are not known by the participants in this study. Volunteers were tested for the presence  
643 of anti-SARS-CoV-2 protein antibodies via the MosaiQ COVID-19 antibody microarray (Quotient)  
644 prior to enrollment and excluded based upon a positive test, as well as, upon risk factors assessed by  
645 clinical history, physical examinations and screening assessments. See Killingley, Mann et al, (2022)<sup>6</sup>  
646 for the full list of inclusion and exclusion criteria and for further details regarding the challenge set up  
647 and ethics. In short, written informed consent was obtained from all volunteers before screening and  
648 study enrollment. The clinical study was registered with ClinicalTrials.gov (identifier NCT04865237).  
649 This study was conducted in accordance with the protocol, the Consensus ethical principles derived  
650 from international guidelines including the Declaration of Helsinki and Council for International  
651 Organizations of Medical Sciences (CIOMS) International Ethical Guidelines, applicable ICH Good  
652 Clinical Practice guidelines, applicable laws and regulations. The screening protocol and main study  
653 were approved by the UK Health Research Authority – Ad Hoc Specialist Ethics Committee (reference:  
654 20/UK/2001 and 20/UK/0002).

655

656 Participant 674007, who fulfilled enrollment criteria, was later found to have low levels of neutralizing  
657 and spike-binding antibodies on admission to the quarantine upon more sensitive post-study  
658 experiments and analysis. This patient was classified as an abortive infection based on the virus kinetics  
659 (see virology method below), with the exclusion of this individual found not to alter any of our  
660 conclusions.

661

662 The participants were followed for 1 year after inoculation, with continued samples and metadata  
663 collected for the use in future/further studies and to benefit the research community. This study however  
664 focused primarily on the first 28-days post-inoculation (with the exception of 46 days for one participant  
665 as noted below, see Sample collection).

666

## 667 **Challenge virus**

668 Participants were inoculated intranasally with an wild-type pre-alpha SARS-CoV-2 challenge virus (full  
669 formal name: SARS-CoV-2/human/GBR/484861/2020) at dose 10 TCID<sub>50</sub> at day 0. 100 µl per naris  
670 was pipetted between both nostrils and the participant was asked to remain supine (face and torso facing  
671 up) for 10 minutes, followed by 20 minutes in a sitting position wearing a nose clip after inoculation to  
672 ensure maximum contact time with the nasal and pharyngeal mucosa. Mid-turbinate nose and throat  
673 samples were collected twice daily using flocked swabs and placed in 3 ml of viral transport medium  
674 (BSV-VTM-001, Bio-Serv) that was aliquoted and stored at -80 °C in order to evaluate viral kinetics  
675 (infection status) as described in Virology method section below. Participants remained in quarantine  
676 for a minimum of 14 days post-inoculation until the following discharge criteria were met: two  
677 consecutive daily nose and or throat swabs with no viral detection or a qPCR Ct value >33.5 and no  
678 viable virus by overnight incubation viral culture with detection by immunofluorescence. For  
679 protocol/full details and ethics used within the Human SARS-CoV-2 challenge study see the Challenge  
680 virus section of methods in Killingley and Mann., et al (2022)<sup>6</sup>.

681

## 682 **Sample collection**

### 683 *Nasopharyngeal swabs*

684 Samples were collected at the Royal Free Hospital by trained healthcare providers at 7 timepoints; day-  
685 1 (pre-inoculation) and day 1, 3, 5, 7, 10 and 14 post-inoculation. The patients were asked to clear any  
686 mucus from their nasal cavities and nasopharyngeal samples were collected using FLOQSwabs (Copan  
687 flocked swabs, Ref 501CS01) inserted along the nasal septum, above the floor of the nasal passage to  
688 the nasopharynx until a slight resistance was felt. The swab was then rotated in this position in both  
689 directions for 10 seconds and slowly removed whilst still rotating and immediately stored in a pre-  
690 cooled cryovial on wet ice containing freeze media (90% heat inactivated fetal bovine serum (FBS) and  
691 10% dimethyl sulfoxide (DMSO). On wet ice the cryovials vials were transferred to the hospital  
692 chutes where they were sent down to the laboratory (<2 mins at RT) and placed in a slow-cooling device  
693 (Mr. Frosty Freezing Container, Thermo Fisher Scientific) and stored at -20 °C until all samples were  
694 collected, at which point they were moved to -80 °C freezers for at least 48 hours for optimum freezing.  
695 Samples were moved, stored and in liquid nitrogen for later processing.

696

### 697 *PBMC isolation from peripheral blood*

698 Peripheral whole blood was collected at the Royal free hospital in EDTA tubes at 5 timepoints; day-1  
699 (pre-inoculation) and day 3, 5, 10, 14 and 28 post-inoculation. Each day the blood was transferred at  
700 room temperature to Imperial College London for the fresh isolation and collection of peripheral blood

701 mononuclear cells (PBMC) by means of Histopaque Ficoll separation (Merck, H8889-500ML). The  
702 peripheral whole blood was first diluted 1:1 with 1X PBS (Merck, D8662-500ML) before being gently  
703 overlaid onto a maximum of 15 mL of Histopaque, at a ratio of 2:1 (blood:Histopaque). The samples  
704 were then centrifuged at 400xg (with no breaks) for 30 min at room temperature (RT) and the PBMC  
705 white buffer layer collected, washed (with PBS ~50 mL) and spun down (400xg for 10 min at RT),  
706 before the supernatant was carefully discarded and the cell pellet was resuspended in 10 mL PBS. The  
707 cells were filtered using a 40 or 70mm cell strainer and then both the cell number and viability were  
708 assessed using Trypan Blue. The cells were further centrifuged (400xg for 10 mins) and resuspend in  
709 the required volume of Cell Freezing Media; 90% FBS (Sigma, F9665-500ML), 10% DMSO (Sigma,  
710 D2650-100ML), before being cryopreserved at -80 °C using a slow-cooling device. The blood and  
711 nasopharyngeal samples were collected within 2 hours of each other.

712 Of note after their were discharged from quarantine and prior to their 28-day follow up (where additional  
713 blood samples were collected), two participants reported either to have had their first SARS-CoV-2  
714 vaccine (636163 ) or a community infection; 636163 had their first vaccine on day 14 post-inoculation  
715 (2 weeks before the day-28 sample was taken). 677306 tested positive before their day 28 visit was due.  
716 The follow up was therefore delayed by 2 weeks, resulting in the “day 28” sample for this participant  
717 instead being taken day 46 post-inoculation. ELISpot performed on this participant revealed a response  
718 in the day 28 and 90 samples (data not shown). Moreover, participant 677696 tested positive on day 29  
719 post-inoculation, a day after their day 28 sample was taken. However for this participant the ELISpot  
720 showed no response at day 28 and a small response at day 90 suggesting the day 28. See **Extended**  
721 **Data Fig. 1a** for overview of the samples and timepoints included from each participant. These  
722 individuals/timepoints were found not to alter any of our conclusions.

723

## 724 **Clinical assessments**

725 Participants were carefully monitored and assessed daily using an array of blood tests, spirometry,  
726 electrocardiograms and clinical assessments (vital signs, symptom diaries and clinical examination).  
727 Full details of all the safety and clinical data collected with the human SARS-CoV-2 challenge study  
728 can be seen in the methods in Killingley and Mann., et al (2022)<sup>6</sup>, with an overview of select metadata  
729 for the 16 participants enrolled in this study in **Extended Data Table 1g**.

730

## 731 **Virology**

732 Longitudinal measures from the nose and throat (pharyngeal) were carried out daily in order to assess  
733 and quantify the viral kinetics of each participant pre- and post-inoculation. These were measured using  
734 two independent assays: (1) qRT-PCR with N gene primers/probes adapted from the Centers for

735 Disease Control and Prevention (CDC) protocol<sup>27</sup> (updated 29 May 2020) and (2) quantitative culture  
736 by Focus Forming Assay (FFA). For full details of each assay and statistical analysis refer to the  
737 methods in Killingley and Mann., et al (2022)<sup>6</sup>.

738

739 The lower limit of quantification (LLOQ) for RT-qPCR was 3 log<sub>10</sub> copies per milliliter, with positive  
740 detections less than the LLOQ assigned a value of 1.5 log<sub>10</sub> copies per milliliter and undetectable  
741 samples assigned a value of 0 log<sub>10</sub> copies per milliliter. Only samples where participants  
742 presented with a positive RT-qPCR were further tested using the FFA assay. In the FFA the LLOQ was  
743 1.27 FFU ml<sup>-1</sup>; viral detection less than the LLOQ was assigned 1 log<sub>10</sub> FFU ml<sup>-1</sup>; and undetectable  
744 samples were assigned 0 log<sub>10</sub> FFU ml<sup>-1</sup>.

745

746 A sustained laboratory-confirmed infection was defined as quantifiable RT-qPCR detection greater  
747 than the LLOQ from mid-turbinate and/or throat (pharyngeal) swabs on two or more consecutive 12-  
748 hourly time points, starting from 24 hours after inoculation and up to discharge from quarantine.  
749 Participants where only stand alone RT-qPCR tests returned quantifiable results (> LLOQ) were  
750 classified as transient infections. Participants where no RT-qPCR tests returned quantifiable results (>  
751 LLOQ) were classified as abortive infections (See **Extended Data Fig. 1b** and **Extended Data Table**  
752 **1a,b,h,i**).

753

754 Infection intervals for each participant were calculated based on the time of the first and last RT-  
755 qPCR test with detectable virus (across the nose and/or throat), where timepoints in which tests below  
756 the LLOQ (1.5) were also counted if they occurred < 2 days of a quantifiable (> LLOQ) test result.

757

### 758 **Nasopharyngeal swab dissociation and processing**

759 Following freezing, nasopharyngeal swabs were transferred to a category level 3 facility at University  
760 College London were stored and processed in batches of 7-to -8 samples at a time to a single cell  
761 suspension. All work was carried out in a MSC class I hood in compliance with standard category level  
762 3 safety practices. The dissociation and collection of cells from nasopharyngeal swab was carried out  
763 in accordance with the previously described protocol<sup>28,29</sup>, with minor modifications. This approach  
764 involves multiple parallel washes and digestion steps using both the nasopharyngeal swab and collected  
765 freezing/ wash media to help ensure the maximum cells and cellular material is collected. First samples  
766 are exposed to DTT for 15 mins, followed by an accutase digestion step for 30 mins before cells from  
767 the same sample (collected directly from the swab or the freezing media/washes from that swab) are  
768 quenched, pooled and filtered prior to checking the cell number and viability.



769

770 Briefly, samples were rapidly thawed and the liquid collected in an empty 15mL falcon tube (**Tube B**).  
771 The cryovial, lid and swab was then carefully rinsed with 3x 1mL warm RPMI 1640 medium which  
772 was added dropwise to the 15 mL tube whilst gently swirling the tube, in order to slowly dilute the  
773 DMSO from the freezing media to help prevent the cells bursting. After waiting 1 min, the tube (**Tube**  
774 **B**) was then topped up with an extra 2 mL of warm RPMI 1640 media and centrifuged at 400 g for 5  
775 min at 4°C. The cell pellet was then resuspended in RPMI 1640/10 mM DTT (Thermofisher, R0861),  
776 and incubated for 15 min on a thermomixer (37°C, 700 rpm), centrifuged as above and the supernatant  
777 was aspirated and the cell pellet was resuspended in 1 mL Accutase (Merck, A6964-500ML). This was  
778 then incubated for a further 30 min on the thermomixer (37°C, 700 rpm).

779

780 In parallel to the processing of the cell freezing media/washes above, the swab was moved to an new  
781 1.5 mL eppendorf tube (**Tube C**) containing 1 mL RPMI 1640/10 mM DTT and placed on the  
782 thermomixer (37°C, 700 rpm) for 15 minutes. In accordance with the steps above, the swab was next  
783 transferred to a new 1.5 mL eppendorf (**Tube D**) containing 1 mL Accutase and incubated with agitation  
784 (700 rpm) at 37°C. The 1 mL RPMI 1640/10 mM DTT from the nasopharyngeal swab incubation (in  
785 **Tube C**) was centrifuged at 400 g for 5 min at 4°C to pellet cells, the supernatant was discarded, and  
786 the cell pellet was resuspended in 1 mL Accutase and incubated for 30 min at 37°C with agitation (700  
787 rpm).

788

789 Following the Accutase digestion step, all cells were combined (Tube B, C and D) and filtered using a  
790 70 µm nylon strainer (pre-wetted with 3 mL quenching media: RPMI 1640/10% FBS/1 mM EDTA  
791 (Invitrogen, 1555785-038) in a 50 mL conical tube (**Tube E**).

792 The filter, tubes and swab were then further thoroughly rinsed with quenching media in order to collect  
793 all cells and the washes combined. The dissociated, filtered cells (Tube E) were then centrifuged at 400  
794 g for 5 min at 4°C, and supernatant discarded. The cell pellet was resuspended in residual volume (~500  
795 µL) and transferred to a new 1.5 mL eppendorf tube (**Tube F**). **Tube E** was then washed with a further  
796 500 µL of RPMI 1640/10% FBS and combined with Tube F, centrifuged as above, supernatant removed  
797 and cells resuspended in 20 µL of RPMI 1640/10% FBS. Using Trypan Blue, total cell counts and  
798 viability were assessed. The cell concentration was adjusted for 7,000 targeted cell recovery according  
799 to the 10x Chromium manual before loading onto the 10x chip (between 700–1,000 cells per µl) and  
800 processing immediately for 10x 5' single-cell capture using the Chromium Next GEM Single Cell V(D)J  
801 Reagent Kit v1.1 (Rev E Guide). For samples where fewer than 13,200 total cells were recovered, all  
802 cells were loaded.

803

804 Note: Due to the sample type, necessary freezing process and no access to a class 3 flow facility to sort  
805 out viable cells, the majority of the samples processed were seen to have low viability (ranging from  
806 5.4% viability to 57.85, with the average viability of samples processed of 26.89%).

807

### 808 **PBMC CITE-seq staining for single-cell proteogenomics**

809 Frozen PBMC samples were thawed and processed in batches of 16 to allow for a carefully designed  
810 pooling strategy. Here each sample was pooled twice into two unique pools containing up to four PBMC  
811 samples per pool from mixed timepoints. Note: Only one sample from each donor was even pooled  
812 together at a time to assist with the demultiplexing later. This pooling strategy was used to help remove  
813 and correct for any protocol-based batch effects.

814

815 In brief, PBMC samples were thawed quickly at 37 °C in a water bath. Warm RPMI 1640 medium (20–  
816 30 mL) containing 10% FBS (RPMI 1640/FBS) was added slowly to the cells before centrifuging at  
817 300g for 5 min. This was followed by a wash in 5 mL RPMI 1640/FBS. The PBMC pellet was collected,  
818 and the cell number and viability were determined using Trypan Blue.

819

820 PBMCs from 4 different donors were then pooled together ( $1.25 \times 10^5$  PBMCs from each donor) to  
821 make up  $5.0 \times 10^5$  cells in total. The remaining cells were used for DNA extraction (Qiagen, 69504).  
822 The pooled PBMCs were resuspended in 22.5  $\mu$ l of cell staining buffer (BioLegend, 420201) and  
823 blocked by incubation for 10 min on ice with 2.5  $\mu$ l Human TruStain FcX block (BioLegend, 422301).  
824 The PBMC pool was then stained with TotalSeq-C Human Cocktail, V1.0 antibodies (BioLegend,  
825 399905) according to the manufacturer's instructions (1 vial per pool). For a full list of TotalSeq-C  
826 antibodies (130 abs + 7 isotype controls) refer to **Extended Data Table 1j**. Following a 30 min  
827 incubation period with the TotalSeq-C Human Cocktail V1.0 antibodies (at 4 °C in the dark) the PBMCs  
828 were topped up using cell staining buffer and centrifuged down to a pellet (500g for 5 min at 4 °C) and  
829 discarding the supernatant. The pellet was then resuspended and washed in the same manner two more  
830 times using the resuspension buffer (0.05% BSA in HBSS), before finally being resuspended in a 20-  
831 30  $\mu$ L resuspension buffer and counted again. The PBMC pools were then processed immediately for  
832 10x 5' single cell capture (Chromium Next GEM Single Cell V(D)J Reagent Kit v1.1 with Feature  
833 Barcoding technology for cell Surface Protein-Rev D protocol). 25,000 cells were loaded from each  
834 pool onto a 10x chip.

835

## 836 **PBMC Dextramer staining for SARS-CoV-2 antigen specific T cell enrichment and single-cell** 837 **sequencing**

838 In order to further validate and investigate the SARS-CoV-2 antigen-specific T-cell populations in our  
839 single cell dataset Day 10, 14 and 28 post-inoculation PBMCs samples, from all 16 participants, were  
840 further enriched and processed for single cell sequencing using a multi-allele panel of 44 SARS-CoV-  
841 2 antigen specific dCODE™ Dextramer® (10x compatible) (Immudex, see **Extended Data Table 1k**  
842 for full panel). This panel includes; five antigen-specific T-cell populations, spanning four MHCI and  
843 one MHCII alleles (covering a total of 15 participants, see **Extended Data Table 1l**) and several  
844 negative controls. Samples were then stained with several FACS antibodies (for monocyte and T cells)  
845 and sorted using MACSQuant® Tyto® Cell Sorter (Miltenyi Biotec), where PE-dCODE Dextramer®-  
846 positive cells were collected and processed for 10x 5' single cell capture. This allowed for the  
847 quantification of paired clonal TCR sequence and TCR specificity by overlaying single-cell V(D)J  
848 expression onto dCODE Dextramer®-positive cell clusters.

849

850 The dextramer staining protocol was taken from Immudex and optimised/adapted to suit our samples  
851 and pooling/staining strategy. In brief, the PBMC samples were thawed in batches of 7 to 8 samples  
852 and the cell number and viability for each sample calculated using Trypan Blue as previously described  
853 above. All cells from each sample were then pooled together in a fresh 1.5 mL eppendorf tube. Note:  
854 The pooling strategy here was such that only one sample per participant/donor was used per pool in  
855 order to enable de-multiplexing by genotype later on and each pool containing a mixture of timepoints  
856 to help reduce batch effect. In order to ensure the collection of as many cells as possible, each of the  
857 original sample tubes was then washed with 200 µl of staining buffer (1X PBS pH 7.4 containing 5%  
858 Heat inactivated FBS (Thermo Fisher Scientific, 10500064) and 0.1g/l Herring sperm DNA (Thermo  
859 Fisher Scientific, 15634017) and added to the pool. The tube was then topped up to 1.4 mLs with  
860 staining buffer and centrifuged down to a pellet (400g for 5 min at 4 °C). The supernatant was carefully  
861 removed and the cell pellet gently resuspended in a total of 30-40 µl staining buffer depending on  
862 pellet's size, ready for staining.

863

864 In parallel the dCODE™ Dextramer® master mix was prepared (in the dark) as per manufacturer  
865 protocol. To help avoid aggregates, each individual dextramer reagent was first microcentrifuge at full  
866 speed for 5 mins before adding 2 µL from each dCODE™ Dextramer® specificity to a low-bind nucleus  
867 free 1.5 eppendorf tube (Eppendorf, 30108051) containing 8.8 µl 100µM d-Biotin (Avidity science,  
868 BIO200) (0.2 ul d-Biotin per number of dCODE™ Dextramer® specificity i.e. 44).The dCODE™  
869 Dextramer® master mix was mixed by gently pipetting before the total volume (96.8 µl) was added to  
870 the resuspended cells. The sample was then thoroughly mixed and incubated at room temperature for

871 30 mins in the dark. Following the addition of anti-human CD14-FITC (Biolegend, 325603) and CD3-  
872 APC (Biolegend, 300458) (at 1:50) the cells were incubated for a further 20 mins (at room temperature  
873 in the dark) before being topped up to 1.4 mL with wash buffer (1X PBS pH 7.4 containing 5 % heat  
874 inactivated FBS). The cells were centrifuged down to a pellet (400g for 5 min at 4 °C) and the  
875 supernatant discarded. The wash step was then repeated X2, with the latter using the addition of 1.4 mL  
876 wash buffer + 1:5000 DAPI (Sigma) as live/dead stain. The supernatant was removed and the cell pellet  
877 resuspended in 4 mL FACS buffer ; 1X PBS, 1% FBS, 25mM HEPES (Thermo Fisher Scientific,  
878 15630-056) and 1 mM EDTA. The samples were then filtered (35 µm nylon mesh cell strainer) and PE  
879 dCODE Dextramer®-positive cells sorted using a MACSQuant® Tyto® Cell Sorter as per  
880 manufacturer guide (Settings; Mix speed= 800 rpm, Chamber temperature= 4 °C, Pressure= 150hPA,  
881 Noise Threshold =14.40, Trigger Threshold= off). Note: In order to collect as many cells as possible  
882 during sorting the entire sample was run on the MACSQuant Tyto, with the negative run through  
883 collected and re-run a second time to ensure no true positives were lost. See **Extended Data Fig. 8a** for  
884 gating strategy for sorting. The PE dCODE Dextramer®-positive cells were then collected, centrifuged  
885 (400g for 5 min at 4 °C) and resuspended in resuspension media before counting the cells. The entire  
886 sample was then processed for 10x 5' single cell capture (Chromium Next GEM Single Cell V(D)J  
887 Reagent Kit v1.1 with Feature Barcoding technology for cell Surface Protein-Rev D protocol). Where  
888 over 25,000 cells were collected the sample was split equally and loaded over two lanes.

889

890 In order to provide additional controls, participants with non-compatible HLA types, including one  
891 volunteer (674700) matching none of the HLA types for the multi-allele dCODE Dextramer panel, were  
892 also processed and used to determine background noise.

893

#### 894 **Library generation and sequencing**

895 The Chromium Next GEM Single Cell 5' V(D)J Reagent Kit (V1.1 chemistry) was used for single-cell  
896 RNA-seq library construction for all nasopharyngeal swab samples, and the Chromium Next GEM  
897 Single Cell V(D)J Reagent Kit v1.1 with Feature Barcoding technology for cell surface proteins was  
898 used for PBMCs, both to process the PBMCs stained with CITE-sequencing antibodies panel and the  
899 dCODE™ Dextramer® (10x compatible) panel. GEX and V(D)J libraries were prepared according to  
900 the manufacturer's protocol (10x Genomics) using individual Chromium i7 Sample Indices. Additional  
901 TCR  $\gamma/\delta$  enriched libraries were also generated based upon an in-house protocol previously described  
902 in<sup>30</sup>. The cell surface protein libraries were created according to the manufacturer's protocol with slight  
903 modifications used for the creation of libraries generated from the CITE-sequencing antibody panel.  
904 These included doubling the SI primer amount per reaction and reducing the number of amplification  
905 cycles to 7 during the index PCR to avoid the daisy chains effect. GEX, V(D)J and the CITE-sequencing

906 derived cell surface protein indexed libraries were pooled at a ratio of 1:0.1:0.4 and sequenced on a  
907 NovaSeq 6000 S4 Flowcell (paired-end, 150 bp reads) aiming for a minimum of 50,000 paired-end  
908 reads per cell for GEX libraries and 5,000 paired-end reads per cell for V(D)J and cell surface protein  
909 libraries. The dextramer derived cell surface protein indexed libraries were submitted at a ratio of 0.1.

910

### 911 **Single cell genomics data alignment**

912 Single-cell RNA-seq and CITE-seq data from PBMCs was jointly aligned against the GRCh38  
913 reference that 10X Genomics provided with CellRanger 3.0.0, and alignment was performed using  
914 CellRanger 4.0.0. CITE-seq antibody-derived tag (ADT) barcodes were aligned against a barcode  
915 reference provided by the supplier, which we annotated to add informative protein names and made  
916 available in our GitHub repository. Single-cell RNA-seq data from nasopharyngeal swab samples were  
917 aligned against the same reference using STARSolo 2.7.3a, and post-processed with an implementation  
918 of emptydrops extracted from CellRanger 3.0.2. To detect viral RNA in infected cells, we added 21  
919 viral genomes including pre-Alpha SARS-CoV-2 (NC\_045512.2) to the above mentioned reference  
920 genomes for RNA-seq alignment, as described in Yoshida et al <sup>5</sup>. Single cell alpha/betaTCR and BCR  
921 data was aligned using CellRanger 4.0.0 with the accompanying GRCh38 VDJ reference that 10X  
922 Genomics provided. Single cell gamma/delta TCR data was aligned against the GRCh38 reference that  
923 10X Genomics provided with CellRanger 5.0.0, using CellRanger 6.1.2.

924

### 925 **Single cell genomics data processing**

926 Both single cell RNA-seq and ADT-seq data were corrected using SoupX <sup>31</sup> to remove free-floating and  
927 background RNAs and ADTs. To correct ADT counts, SoupX 1.5.2 parameters soupQuantile and  
928 tfidfMin parameters were set to 0.25 and 0.2, respectively, and lowered by decrements of 0.05 until the  
929 contamination fraction was calculated using the autoEstCont function. SoupX on RNA data was  
930 performed using default settings. To confidently annotate SARS-CoV-2 infected cells, we used SoupX  
931 corrected viral RNA counts to remove false positives due to freely floating SARS-CoV-2 virions.  
932 However, when quantifying the amount of reads per cell in **Fig. 2h** and their distribution over the viral  
933 genome in **Fig. 2f**, we used the raw counts and sequencing data. To profile the distribution of viral  
934 reads, we removed PCR duplicates from the aligned BAM files that STARSolo produced with  
935 MarkDuplicates in picard (<https://broadinstitute.github.io/picard/>), and tallied the location within the  
936 SARS-CoV-2 genome using the start of each sequencing read. Aligned single cell RNA-seq data was  
937 imported from the *filtered\_feature\_bc\_matrix* folder into Seurat V4.1.0 for processing, keeping only  
938 cells with at least 200 RNA features detected. Nasopharyngeal and PBMC cells with more than 50%  
939 and 10% of the counts coming from mitochondrial genes were excluded, respectively. SoupX corrected



940 gene expression and ADT counts were normalized by dividing it by the total counts per cell and  
941 multiplying by 10 000, followed by adding one and a natural-log transformation (log<sub>1p</sub>).

942

### 943 **Demultiplexing and patient id assignment**

944 Each PBMC sample was pooled twice into two unique pools containing up to four PBMC samples per  
945 pool, followed by CITE-seq and single cell VDJ sequencing as described above. Souporecell V2.0<sup>32</sup> was  
946 used to demultiplex each pools based on the genotype differences between the mixed samples.  
947 Souporecell analyses were performed with the skip\_remap parameter enabled and using the common  
948 SNP database that was provided by the software. We used two complementary approaches to  
949 confidently assign participant identity to each souporecell cluster. First we compared the cluster  
950 genotypes with SNP array derived genotyping data, generated for all participants and performed using  
951 the Affymetrix UK Biobank Axiom™ Array kit by Cambridge Genomic Services (CGS). Second, the  
952 combinations of samples within each pool was unique, enabling assignment of participant identity based  
953 on the presence of unique participant-specific combinations of identical genotypes in two separate  
954 pools. This multiplexing and replication strategy furthermore enabled us to distinguish library specific  
955 batch effects from participant specific effects in downstream analyses.

956

### 957 **Doublet detection**

958 We used the output from souporecell to identify ground-truth doublets in PBMCs by selecting droplets  
959 that contained two genotypes from different participants. We then included these ground-truth doublets  
960 into the iterative rounds of subclustering and cell state annotation to look for doublet specific clusters  
961 that emerged, which we then subsequently removed. Doublets in the nasopharyngeal data were removed  
962 during iterative rounds of subclustering and cell state annotation by identifying cell clusters that  
963 expressed marker genes from multiple distinct cell types.

964

### 965 **Clustering and cell type annotation**

966 Principal component analysis was run on corrected gene expression counts from selected hypervariable  
967 genes and the first 30 principal components were selected to construct a nearest neighbour graph and  
968 UMAP embedding. We used harmony<sup>33</sup> to perform batch correction on the PBMC data on the  
969 sequencing library identity to remove technical batch effects. Leiden clustering<sup>34</sup> performed at  
970 resolutions of 0.5, 1, 4 and 32, on nearest neighbour graphs and embeddings created with 500, 1000,  
971 2000, 4000, 6000 and 8000 selected hypervariable genes (excluding TCR and BCR genes), was used to  
972 perform iterative rounds of cell type annotation based on marker gene expression and subsetting of  
973 clusters to obtain a highly granular cell state annotation. We used cell type marker genes described in

974 Yoshida et al<sup>5</sup> and Stephenson et al<sup>4</sup> to define cell types. Our cell type annotation was furthermore  
975 guided by predicted cell type labels using Celltypist's<sup>35</sup> provided models and custom trained models  
976 based on annotations in Yoshida et al<sup>5</sup> and Stephenson et al<sup>4</sup>.

977

### 978 **Single cell TCR and BCR data processing**

979 Aligned single cell BCR and alpha/beta TCR sequencing data was imported in scirpy<sup>36</sup> to obtain a cell  
980 by TCR or BCR formatted table, which was then added to Seurat objects containing gene expression  
981 data. Aligned single cell gamma/delta TCR data was reannotated using Dandelion V0.2.4<sup>37</sup>.

982

### 983 **Integration of five COVID-19 studies**

984 All transcriptomic data was processed with the single cell analysis Python workflow Scanpy<sup>38</sup>. Each  
985 data set was individually filtered following best practices outlined in<sup>39</sup> (Between 200 and 3500 genes  
986 per cell, less than 10% mitochondrial genes expressed per cell, genes expressed in at least 3 cells, other  
987 parameters at default). The gene sets were reduced to their intersection before combining data sets. Cells  
988 came from a total of 602 individuals, with 325 acute COVID-19 patients, 110 COVID-19 convalescent  
989 patients, 114 healthy and 53 hospitalized controls (**Extended Data Table 1d**). This resulted in an  
990 integrated embedding containing 946,584 T cells with resolved TCR from 494 samples, made up of 455  
991 donors of which 240 were acute COVID-19 patients, 82 convalescent, 88 healthy and 45 hospitalized  
992 controls (**Extended Data Table 1e**). The total number of donors in the integrated object is smaller as  
993 only samples with matching VDJ sequencing data were kept. A probabilistic scVI model (2 hidden  
994 layers, 128 hidden nodes, 20-dimensional latent space, negative binomial gene likelihood, other  
995 parameters at default<sup>40</sup>) was trained on the data to map cells to a shared latent space, and visualised  
996 using UMAP.

997

### 998 **Identification of activated TCR clonotype groups using Cell2TCR**

999 To identify TCR clonotype groups, we used tcrdist3<sup>41</sup> with the provided human references to compute  
1000 a sparse representation of the distance matrices for all identified TRA and TRB CDR3 sequences, with  
1001 the radius parameter set to 150. We then summed the distances for TRA and TRB to obtain a combined  
1002 distance matrix. Next, we iterated over possible TCR distance thresholds between 5 and 150 with  
1003 increments of 5, to compute TCR clonotype groups at each threshold. We then generated a distance  
1004 adjacency graph of TCRs from different T cells with a distance lower than the threshold, which was  
1005 clustered to identify TCR clonotype groups using leiden<sup>34</sup> clustering through the igraph package<sup>42</sup>, at a  
1006 resolution of 1 and using the RBCConfigurationVertexPartition partition. To find the optimal distance  
1007 threshold at which only TCRs that recognise the same antigen are grouped together, we quantified

1008 clonotype group contamination at each threshold using two approaches. First, we assumed that T cells  
1009 that were annotated as naive should not participate in an expanded clonotype group, and quantified the  
1010 proportion of naive T cells in each clonotype group to determine the largest threshold at which we  
1011 observed minimal participation of naive T cells. Second, we assumed that CD4+ T cells and CD8+ T  
1012 cells should never be part of the same TCR clonotype group, so we set out to quantify the proportion of  
1013 CD4+/CD8+ mixing in each clonotype group to find the largest threshold where mixing is minimal.  
1014 Both approaches revealed the same optimal threshold of 35 at which both naive T cell participation and  
1015 CD4+/CD8+ mixing is minimal, which we then used for downstream analyses. To identify activated  
1016 TCR clonotype groups we assumed that these groups should include activated T cells and that we should  
1017 at least detect multiple independent TCR clonotypes that appear to be raised against the same antigen  
1018 at the same time; we therefore selected clonotype groups that contained at least one participating  
1019 activated T cell and that contained at least two unique CDR3 nucleotide sequences.

1020

### 1021 **Identification of activated BCR clonotype groups**

1022 To identify BCR clonotypes groups that were activated during infection, we used a similar approach as  
1023 described above for T cells. Instead of using tcrdist to compute distances, we used the levenshtein  
1024 distance and iterated over possible thresholds between 1 and 20 to find an optimal threshold by  
1025 quantifying naive B cell participation. This revealed that a levenshtein distance of 2 is optimal to  
1026 identify BCR clonotype groups that only contain B cells that recognise the same antigen. To identify  
1027 activated BCR clonotype groups, we assumed that these groups should include antibody secreting B  
1028 cells (plasmablasts and plasma cells) and that we should at least detect multiple independent BCRs  
1029 clonotypes that appear to be raised against the same antigen at the same time; we therefore selected  
1030 clonotype groups that contained at least one participating antibody secreting B cell and that contained  
1031 at least three unique CDR3 nucleotide sequences.

1032

### 1033 **Generation of VDJ logos**

1034 TCR and BCR logos in Figure X, X and X were generated by providing the CDR3 amino acid sequences  
1035 of each clonotype group to the ggseqlogo R package<sup>43</sup> or the logomaker Python package<sup>44</sup>. When  
1036 clonotype groups contained CDR3 amino acid sequences of variable lengths, we selected the sequences  
1037 with the most frequently occurring length within each group for visualization purposes only.

1038

### 1039 **Generalized linear mixed models of cell state compositional changes over time**

1040 The relative amount of cells per cell type in each sample was modeled using a generalized linear mixed  
1041 model with a poisson outcome. When technical replicates were available (most of the PBMC samples),

1042 these were modeled as separate samples. We modeled participant identifiers, days since inoculation,  
1043 and sequencing library identifiers (of multiplexed libraries), as random effects terms to overcome  
1044 colinearity between these factors. The effect of each clinical/technical factor on cell type composition  
1045 was estimated by the interaction term with the cell type. The glmer function in the lme4 package  
1046 implemented on R was used to fit the model. The standard error of the variance parameter for each  
1047 factor was estimated using the numDeriv package. The conditional distribution of the fold change  
1048 estimate of a level of each factor was obtained using the ranef function in the lme4 package. The log-  
1049 transformed fold change is relative to the pre-inoculation time point (day -1). The statistical significance  
1050 of the fold change estimate was measured by the local true sign rate, which is the probability that the  
1051 estimated direction of the effect is true, that is, the probability that the true log-transformed fold change  
1052 is greater than 0 if the estimated mean is positive (or less than 0 if the estimated mean is negative). We  
1053 calculated P values using a two-sample Z-test using the estimated mean and standard deviation of the  
1054 distribution of the effect (log-transformed fold change). P values were converted into false discovery  
1055 rates using the Benjamini & Hochberg method.

1056

#### 1057 **Gaussian processes regression and latent variable models to infer time since viral exposure**

1058 To infer time from cell state abundances, we first generated a linear regression model using CellTypist  
1059 to predict PBMC or nasopharyngeal cell states based on the highly detailed manual cell state annotation  
1060 presented in this work. Celltypist models were trained and used using default parameters, with  
1061 check\_expression set to false, balance\_cell\_type set to true, feature\_selection set to true, and max\_iter  
1062 set to 150. We next set out to build a predictive model to infer time since viral exposure using the PBMC  
1063 data presented in this work as a training dataset. We used the above mentioned publicly available PBMC  
1064 data from five studies as a test dataset to predict time since viral exposure on. Because we were  
1065 specifically interested in comparing time since viral exposure to reported time since onset of symptoms  
1066 in varying disease severities, we excluded samples for which these features were unknown. To ensure  
1067 that the cell state proportions in the training and test dataset were similar, we used our CellTypist model  
1068 on both datasets to predict relative cell state frequencies, which was used as input for our time prediction  
1069 model. To account for participant-to-participant heterogeneity and continuous variation in the timeline  
1070 of immune responses, we first constructed a gaussian process latent variable model<sup>45</sup> (GPLVM) to  
1071 smooth the time since viral exposure in the training dataset. We applied the Pyro implementation of  
1072 GPLVM<sup>46</sup> across all predicted cell state abundances, and restricted the model to 2000 iterations and a  
1073 single latent variable that was initialized on the square root transformed time since inoculation, which  
1074 resulted in an accurate recapitulation of the mean time since inoculation while smoothing outliers. We  
1075 next used each predicted cell state as input for a task to generate a multi-task gaussian process regression  
1076 model<sup>47</sup> to predict the smoothed time since inoculation using GPyTorch<sup>48</sup>. We used the Adam  
1077 optimiser and allowed for as many iterations for the loss in marginal log likelihood to reach zero. We

1078 next predicted the cell state compositions across the entire tested timeline (day -1 to day 28), and  
1079 compared these cell state compositions to the cell state compositions in our query dataset as predicted  
1080 by our CellTypist model. Last, we selected the time point whose predicted cell state composition had  
1081 the lowest mean squared error compared to the observed cell state composition.

1082

### 1083 **Matching clonotype groups to antigen-TCR database**

1084 Computation of fold change enrichment of SARS-CoV-2 specific TCRs in Activated T cell populations  
1085 compared to other T cell populations: Median p for 10 random draws of n=5000 unique clones of both  
1086 populations = 3.75,  $p = 1.68 \times 10^{-21}$

1087

### 1088 **Bulk TCR sequencing and processing**

1089 Total RNA was extracted from whole blood samples collected in Tempus Blood RNA tubes  
1090 (ThermoFisher #4342792) using the manufacturer's protocol for RNA extraction. TCR  $\alpha$  and  $\beta$  genes  
1091 were sequenced using a pipeline which introduces unique molecular identifiers attached to individual  
1092 cDNA molecules using single-stranded DNA ligation. The unique molecular identifier allows  
1093 correction for sequencing error PCR bias, and provides a quantitative and reproducible method of  
1094 repertoire analysis. Full details for both the experimental TCRseq library preparation (<sup>49,50</sup>) and the  
1095 subsequent TCR annotation (V, J and CDR3 annotation) using Decombinator V4<sup>51</sup> are published. The  
1096 Decombinator software is freely available at <https://github.com/innate2adaptive/Decombinator>.

1097

### 1098 **Memory formation analysis**

1099 T cell phenotypes (naive, activated, effector, memory) were recorded for an antigen-specific TCR clone  
1100 at different time points throughout infection. TCR clones were filtered by having an activated label at  
1101 least once, being observed in at least two samples, one of which had to be at day 28. Unique TCR clones  
1102 are distinguished by color and numbered with their clone\_id identifier. A shaded area is drawn when  
1103 the same clone appeared with several distinct cell type labels, and the size of the shaded area informs  
1104 of their relative ratios.

1105

### 1106 **Quantifying TCR diversity restriction in phenotypic clusters using coincidence analysis**

1107 To quantify the diversity of TCRs found within different phenotypic clusters we determined the  
1108 probability with which two distinct clonotypes within a cluster share an identical CDR3 amino acid  
1109 sequence <sup>52</sup>. For visualization we normalized these probabilities by the same quantity calculated over  
1110 the complete data regardless of phenotype. This ratio of probability of coincidences provides a stringent



1111 measure of convergent functional selection of distinct clonotypes that share the same TCR. The analysis  
1112 is based on clonotypes defined by distinct nucleotide sequences of the hypervariable regions, and does  
1113 not make direct use of clonal abundances as these can also reflect TCR-independent lineage differences.  
1114 We focused our analysis on conventional T cells only, considered only cells with at least one valid  
1115 functional alpha and beta chain, and kept only a single chain for each cell where there were multiple  
1116 chains. We performed the analysis both on the alpha and beta chain separately, as well as on paired  
1117 alpha and beta chain, in each instance requiring exact matching of the CDR3 amino acid sequences.  
1118

1119 **Figure Legends**

1120

1121 **Figure 1: Extensive temporal cell state dynamics after SARS-CoV-2 inoculation.**

1122 (a) Illustration of study design and cohort composition. (b-c) UMAPs of all nasopharyngeal cells, color-  
1123 coded by their broad cell type annotation in (b), by the infection group in the top panel of (c), and by  
1124 days since inoculation in the bottom panel of (c). Only cells from sustained infection cases are shown  
1125 in the bottom panel of (c). (d-e) UMAPs as in (b-c), but showing all PBMCs. (f) Fold changes in  
1126 abundance of nasopharynx resident broad immune cell type categories. Immune cell abundances were  
1127 scaled to the total amount of detected epithelial cells in every sample prior to calculating the fold  
1128 changes over days since inoculation compared to pre-infection (day -1) by fitting a GLMM on scaled  
1129 abundances. The mean cell type proportions over all cells and samples is shown in the green heatmap  
1130 right of the dotplot to aid the interpretation of changes in cell type abundances.

1131

1132 **Figure 2: Cell-state-specific antiviral responses and infection**

1133 (a) Dotplot visualizing the mean expression of interferon stimulated genes across cell types and time  
1134 since inoculation in participants with sustained infections, for nasopharyngeal cells (left plot) and  
1135 PBMCs (right plot). (b) Dotplot as in (Fig 1f), showing myeloid cell types in sustained infection cases  
1136 that significantly change at least one time point compared to pre-infection. Nasopharyngeal cells and  
1137 PBMCs are shown in the left and right plot, respectively. (c) Boxplot showing the relative amounts of  
1138 circulating inflammatory monocytes over time since inoculation in each infection group. (d) Boxplot  
1139 showing the fraction of circulating MAIT cells that are activated over time since inoculation in each  
1140 infection group. (e) Stacked barplot showing the amount of nasopharyngeal cells with at least one  
1141 SARS-CoV-2 RNA read detected (after background subtraction), split by days since inoculation and  
1142 color-coded by cell type. (f) Barplots showing the distribution of detected viral reads over the SARS-  
1143 CoV-2 genome in the five most highly infected cell types. The blue line represents a loess fit over the  
1144 data. The top-right inset illustration is shown to aid the interpretation of a uniform read distribution  
1145 versus a 3' biased read distribution. (g) Boxplot showing the fraction of ciliated cells that are annotated  
1146 into detailed response or infection cell states. Only cells from sustained infection cases are shown and  
1147 split by days since inoculation. The Y axis for interferon (IFN) and acute-phase response (APR) positive  
1148 ciliated cells is shown on the left, while the Y axis for infected ciliated cells is shown on the right. (h)  
1149 Boxplot of the amount of viral sequencing reads per cell type. (i) Heatmap of spearman correlations  
1150 between host gene expression and the amount of viral reads found in each cell, split by cell type. Shown  
1151 genes correlate the highest with gene expression in ciliated cells. In all box plots, the central line and  
1152 the notch are the median and its approximate 95% confidence interval, the box shows the interquartile  
1153 range and the whiskers are extreme values upon removing outliers.

1154

1155 **Figure 3: Adaptive immune responses emerge at day 10 post-inoculation.**

1156 (a) UMAP of all circulating T cells, highlighting the distinct cluster of activated T cells. Cells are color  
1157 coded and labeled by their detailed cell state annotation. (b) Marker gene and protein expression of  
1158 activated T cell subsets are shown in blue and red, respectively. (c) Percentages of nasopharyngeal T  
1159 cells that were annotated as activated T cells, split over days since inoculation and color coded by  
1160 infection group. (d) Boxplot as in (c), but showing circulating activated T cells. (e) UMAP as in (a), but  
1161 showing nasopharyngeal T cells. (f) Fold changes in cell state abundance compared to pre-inoculation  
1162 of nasopharyngeal and circulating conventional T cells are shown in the left and right plots, respectively.  
1163 Only cell states that significantly change at a FDR < 10% at least one time point are shown.  
1164 Nasopharyngeal T cell abundances were scaled to the total amount of detected epithelial cells. Fold  
1165 changes and significance were calculated by fitting a GLMM as shown in *Figure 1*. The mean cell type  
1166 proportions over all cells and samples is shown in the green heatmap right of the dotplot to aid the  
1167 interpretation of changes in cell type abundances. (g) TCR clonality and expansion at day 14 of activated  
1168 TCRs was validated using bulk TCR sequencing. For TCRs that matched the single cell gene  
1169 expression, normalized clonality TCR alpha (left) and beta (right) data is separated by type and  
1170 expressed as the average fraction of total clones in sample contributed by a cell of that type, with  
1171 changes over time implying clonal expansion or contraction. For activated T cell types of interest,  
1172 scatterplots for each sustained infection and at each time point sampled (days -1, 7, 14) are drawn. (h)  
1173 Proportion of CD8+ infiltrating T cells that use alpha/beta TCRs, typical Dv2/Gv9 g/d TCRs, or atypical  
1174 g/d TCRs is shown. (i) The relative immune repertoire composition of g/d T cells in circulation and  
1175 nasopharynx after challenge are shown in the left and right bars, respectively. G/d chain pairs that are  
1176 significantly more or less abundant between circulation and nasopharynx ( $p < 0.05$ ) are highlighted with  
1177 an asterisks. (j) Dotplot as in (f), showing the fold changes in B cells. Legend for significance and mean  
1178 cell type proportions as in (f). (k) Abundance of TCR clusters relative to all TCRs are shown over time  
1179 since inoculation. Activated TCR clusters are color coded and their TCR motifs are shown. Legend for  
1180 the physicochemical properties of amino acids in shown TCR motifs is shown in panel (l). (l) Plot as in  
1181 (k), but showing BCR clusters. Immunoglobulin class usage within each activated BCR cluster is shown  
1182 in the rightmost bars.

1183

1184 **Figure 4: Integrating COVID-19 patient data reveals public SARS-CoV-2 TCR motifs**

1185 (a) UMAP representation after integration of five COVID-19 patient datasets with paired RNA and  
1186 VDJ sequencing data. Cell type labels inferred using a logistic regression classifier (Celltypist) trained  
1187 on manual annotations of PBMCs from the current work. (b) Fraction of activated T cells across all T  
1188 cells in sample for COVID-19 (n = 240), convalescent (n = 82) and healthy (n = 88) samples of five

1189 COVID-19 patient datasets. Significance levels after Mann-Whitney testing are shown and indicate that  
1190 COVID-19 and convalescent samples have significantly more activated T cells than healthy samples.  
1191 (c) Activated T cell types highlighted on UMAP representation from panel (a). Activated CD8<sup>+</sup> T cells  
1192 were most abundant, followed by CD4<sup>+</sup> and regulatory types, and clustered together in a distinct area  
1193 of the latent space. (d) Clustermap of pairwise TCR distances with color-coded information for each  
1194 TCR on patient\_id, study, motif, antigen on the left-hand side, as well as the sequence logos for the nine  
1195 most common motifs on the right-hand side. Each column/row corresponds to a unique TCR, and the  
1196 distance to each TCR in the set is indicated by color. Only activated T cells with public motifs (identified  
1197 in more than one individual) were considered. Low distances indicate similar TCRs, with distances of  
1198 40 and less potentially yielding TCRs recognising the same epitopes. For sequence logos, letter height  
1199 indicates frequency of AA at that position across T cells pertaining to the motif. AAs are colored by  
1200 side chain chemistry: Acidic (red), basic (blue), hydrophobic (black), neutral (purple), polar (green).  
1201 AA: amino acid. (e) Recorded symptoms averaged over SARS-CoV-2 challenge participants with  
1202 sustained infection for days -1 to 14 post-inoculation. Major molecular events in the immune response  
1203 are highlighted with arrows. (f) Predicted time since viral exposure is plotted against reported time since  
1204 onset of symptoms. Lines represent loess fits of the data split and color coded by reported severity.

1205

## 1206 **Figure 5: Temporally resolved epithelial and immune response in SARS-CoV-2 infections**

1207 Summary figure highlighting 1) temporal differences in the distinct infection groups, 2) novel antiviral  
1208 responses, 3) novel characteristics of sustained infection, and 4) the identification of public motifs in  
1209 SARS-CoV-2 specific activated T cells. In addition, our work provides community tools for inference  
1210 of specific TCR motifs (Cell2TCR) in activated T cells, and for temporal assignments of clinical  
1211 COVID-19 samples underpinning future therapeutic applications.

1212

## 1213 **Supplementary Figure 1: Overview of Single Cell Human SARS-CoV-2 Challenge Study cohort.**

1214 (a) Timeline of the samples collected from each of the 16 participants enrolled in our study. Sample  
1215 collections are shown relative to the date of SARS-CoV-2 inoculation (day 0). Samples are shown by  
1216 infection group (sustained, transient and abortive), with their age in years (yrs) and sex (self-identified).  
1217 \*Indicates participants who were either vaccinated (636163) or reported to have developed an  
1218 community infection, before or immediately after blood samples were taken on day 28 (677696 and  
1219 677306). See sample collection section in the methods for more details. Longitudinal measures of nasal  
1220 and pharyngeal (throat) viral kinetics from swabs. Shown for each participant as measured via (b) RT-  
1221 qPCR and (c) quantitative culture by focus forming assay (FFA). Patients were identified as testing  
1222 positive if they had at least one RT-qPCR test where the viral load was able to be quantified ( $\geq$  lower  
1223 limit of quantification (LLOQ)). Six participants were seen to present multiple, sequential, positive RT-

1224 qPCR results and were classified as having a sustained infection. Three participants were seen to have  
1225 standalone positive results and were classified having transient infections. Seven participants never  
1226 presented a single RT-qPCR test result  $\geq$  LLOQ and these were classified as abortive infections. FFA  
1227 tests were only performed for patients identified as having sustained infections, so there is no data for  
1228 participants with transient or abortive infections. Infection intervals for each participant were calculated  
1229 based on the first and last across the nose and throat, where positive tests below the LLOQ were counted  
1230 if they occurred  $< 2$  days of a quantifiable ( $\geq$  LLOQ) test result. \*Indicated where the patient was  
1231 discharged from quarantine prior to testing negative. The black octagon highlights patients that were  
1232 still reporting positive results at day 28 post-inoculation.

1233

1234 **Supplementary Figure 2: All identified and annotated cells.**

1235 (a) UMAP of all PBMCs, color-coded and labeled by detailed cell state annotation. Subsets of B cells  
1236 with differential immunoglobulin chain usage are not shown in full detail for clarity. (b) UMAP of all  
1237 nasopharyngeal cells, color-coded and labeled by detailed cell state annotation.

1238

1239 **Supplementary Figure 3: Marker gene expression used for annotation.**

1240 Marker gene expression of cell states annotated in (a) nasopharyngeal immune cells, (b)  
1241 nasopharyngeal epithelial cells, (c) myeloid and progenitor PBMCs.

1242

1243 **Supplementary Figure 4: Marker gene expression used for annotation of PBMCs.**

1244 Marker gene expression of cell states annotated in (a) T, NK and ILC cells in PBMCs, (b) B cells in  
1245 PBMCs.

1246

1247 **Supplementary Figure 5: Quality control metrics of single cell RNA-seq data.**

1248 Calculate quality control metrics for nasopharyngeal swabs showing (a) cells per sample and (b) gene  
1249 reads per cell per sample. (c) and (d) show the same respectively for PBMCs.

1250

1251 **Supplementary Figure 6: Temporal response states and activated T cells.**

1252 (a) Line plot showing the mean proportions of interferon stimulated cells over time since inoculation  
1253 within cell types with a distinct and annotated cluster of interferon stimulated cells. (b) Marker gene  
1254 expression of activated MAIT cells. (c) Marker gene expression of response states observed in ciliated  
1255 cells. (d) TCR repertoire overlap of nasopharyngeal and circulating conventional T cells. We only



1256 considered the beta TCR chain to identify overlapping T cells to include T cells without a TRA sequence  
1257 detected. (e) Memory analysis in an individual with sustained SARS-CoV-2 infection. Unique TCR  
1258 clones are distinguished by color and numbered with their clone\_id identifier. A shaded area is drawn  
1259 when the same clone appeared with several distinct cell type labels, and the size of the shaded area  
1260 informs their relative ratios. (f) TCR bulk data with matched single cell labels as in **Fig. 3g**, but showing  
1261 data on abortive infections for activated and other T cells. No particular changes are observed across  
1262 the three time points sampled. (g) The fraction of activated T cells that participate in TCR clonotype  
1263 groups versus the fraction of cells in each group that originate from participants with sustained  
1264 infections. (h) Scatterplot as in (g), but showing BCR clonotype groups and the fraction of antibody  
1265 secreting B cells instead of activated T cells. (i) Fraction of unique clones matching SARS-CoV-2  
1266 entries in IEDB across all T cell clones within that broad T cell compartment. Significance level after  
1267 Whitney-Mann testing shown for Activated vs Memory T cells (putative SARS-CoV-2 fraction 3.75  
1268 times higher in Activated T cells,  $p = 1.68 * 10^{-21}$ ) (j) Difference between predicted time since viral  
1269 exposure and reported time since onset of symptoms, split by reported severity. In all box plots, the  
1270 central line and the notch are the median and its approximate 95% confidence interval, the box shows  
1271 the interquartile range and the whiskers are extreme values upon removing outliers (k) Coincidence  
1272 analysis of TCR sequence diversity restriction in phenotypic subsets. Fraction of clonotype pairs within  
1273 each phenotypic cluster that share identical CDR3 amino acid sequences (but distinct nucleotide  
1274 sequences) normalized by the same statistics calculated across all clonotypes, for alpha, beta, and both  
1275 chains together. The ratio of within cluster versus overall sequence coincidence probabilities is a  
1276 measure of the breadth of epitopes targeted by the different clonotypes within a cluster<sup>52</sup>.

1277

### 1278 **Supplementary Figure 7: Detailed temporal dynamics in cell state abundances.**

1279 (a) Fold changes in abundance of cell states in PBMCs. Detailed annotation of interferon stimulated  
1280 subsets and immunoglobulin class specific cell states are not shown for clarity. Immune cell abundances  
1281 were scaled to the total amount of detected PBMCs in every sample prior to calculating the fold changes  
1282 over days since inoculation compared to pre-infection (day -1) by fitting a GLMM on scaled  
1283 abundances. The mean cell type proportions over all cells and samples is shown in the green heatmap  
1284 right of the dotplot to aid the interpretation of changes in cell type abundances. (b) Dotplot as in (a),  
1285 but showing nasopharyngeal immune cells. Immune cell abundances were scaled to the total amount of  
1286 detected epithelial cells in every sample. (c) Dotplot as in (a), but showing nasopharyngeal epithelial  
1287 cells. (d) UMAP of all CD8+ T cells from the Dextramer assay, with cell types predicted by Celltypist  
1288 model trained on previous PBMC data. Activated T cells form a distinct cluster. (e) Cell counts for  
1289 CD8+ T cell types by HLA compatibility of donor with the highest-bound Dextramer. Only Activated  
1290 T cells have positive log<sub>2</sub> fold change for HLA-matched Dextramers. (f) UMAP as in (d), colored by  
1291 HLA compatibility.

1292

1293 **Supplementary Figure 8: Validation of antigen-specific activated T cells.**

1294 (a) Gating strategy used to enrich SARS-CoV-2 antigen specific T cells via MACSQuant Tyto cell  
1295 sorting. Cells were sequentially stained with a multi-allele panel of dCODE dextramer- PE complexes,  
1296 with the addition of anti-human CD3-APC and CD14-FITC FACS antibodies as references to help us  
1297 identify T cell specific binding. Debris and cell aggregates were gated out first using BSB-H  
1298 (backscatter blue laser-height) SSC-H (side scatter-height). From the cells, DAPI+ dead cells were  
1299 excluded. T cells (CD3+) and monocytes (CD14+) were then gated for (CD3+ and/or CD14+  
1300 population) and the sort gate defined from this population as all PE-dCODE Dextramer® positive cells.  
1301 This lenient sorting strategy was decided upon in order to collect enough cells for 10X 5' single cell  
1302 analysis downstream and to ensure we were capturing all SARS-CoV-2 antigen specific cells. Any non-  
1303 specific binding (e.g. to monocytes) and background noise could then be removed computationally. (b)  
1304 Calculated quality control metrics for PE-dCODE Dextramer® positive sequenced cells showing gene  
1305 reads per cell per sample. (c) Proportions of activated T cells bound to Dextramers loaded with selected  
1306 SARS-CoV-2 antigens. The total amount of bound cells to each Dextramer is shown, color-coded by  
1307 predicted cell state. If barcodes from several Dextramers were detected to be bound to the same cell,  
1308 we only selected the Dextramer with the highest signal as bound. As a control to separate background  
1309 and real binding, cells are separated based on the HLA haplotype compatibility with the tested  
1310 Dextramer. Only Dextramers with at least 10 HLA matched bound cells are shown. FDR corrected p  
1311 values were determined by a Fisher-exact test comparing the proportion of HLA matched activated T  
1312 cells in the Dextramer bound cells to the proportion of unbound HLA matched activated T cells. N  
1313 represents the number of cells in each bar. The right-most bar represents the overall distribution of cell  
1314 types across all Dextramer experiments.

1315 **References**

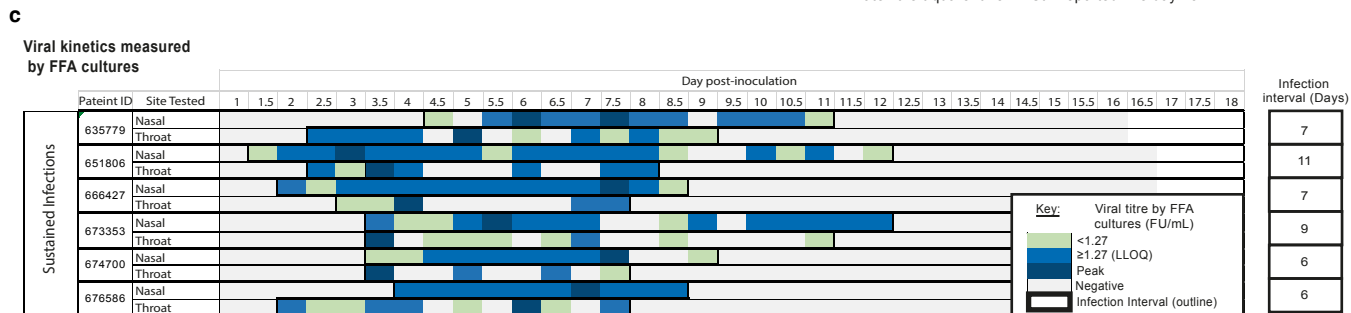
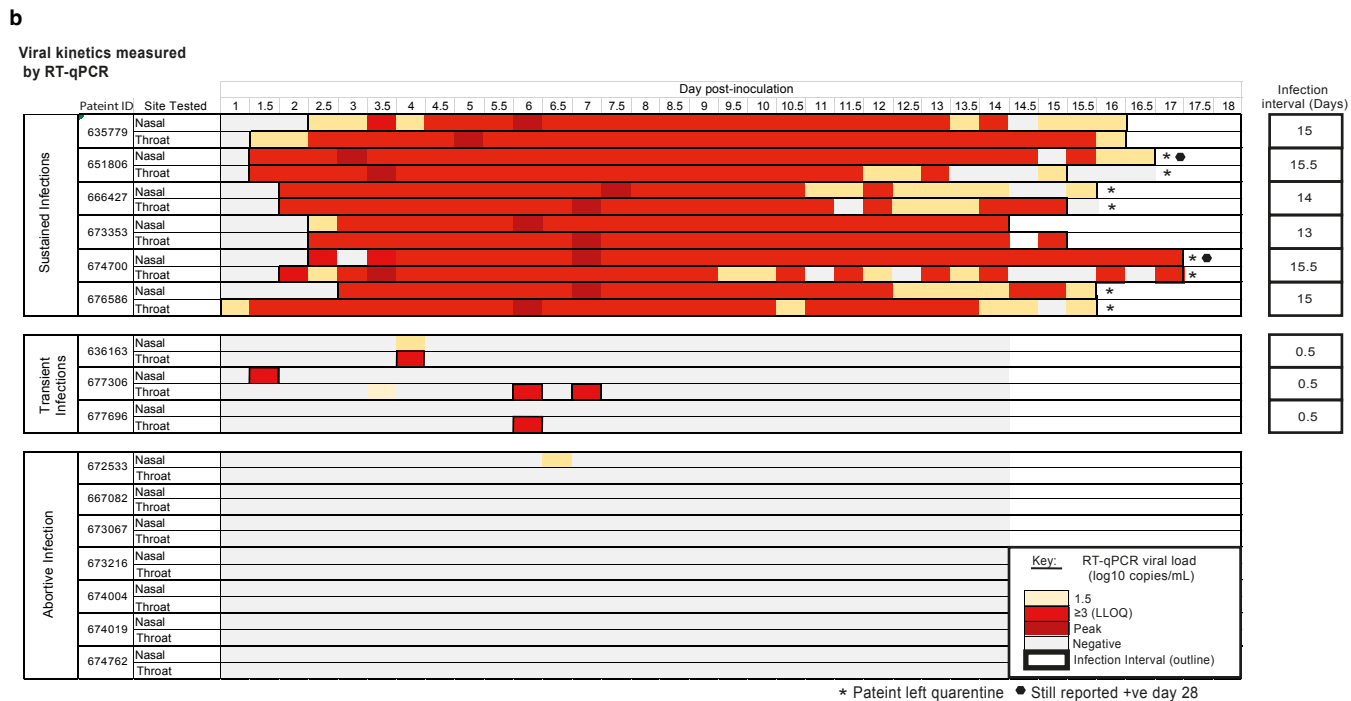
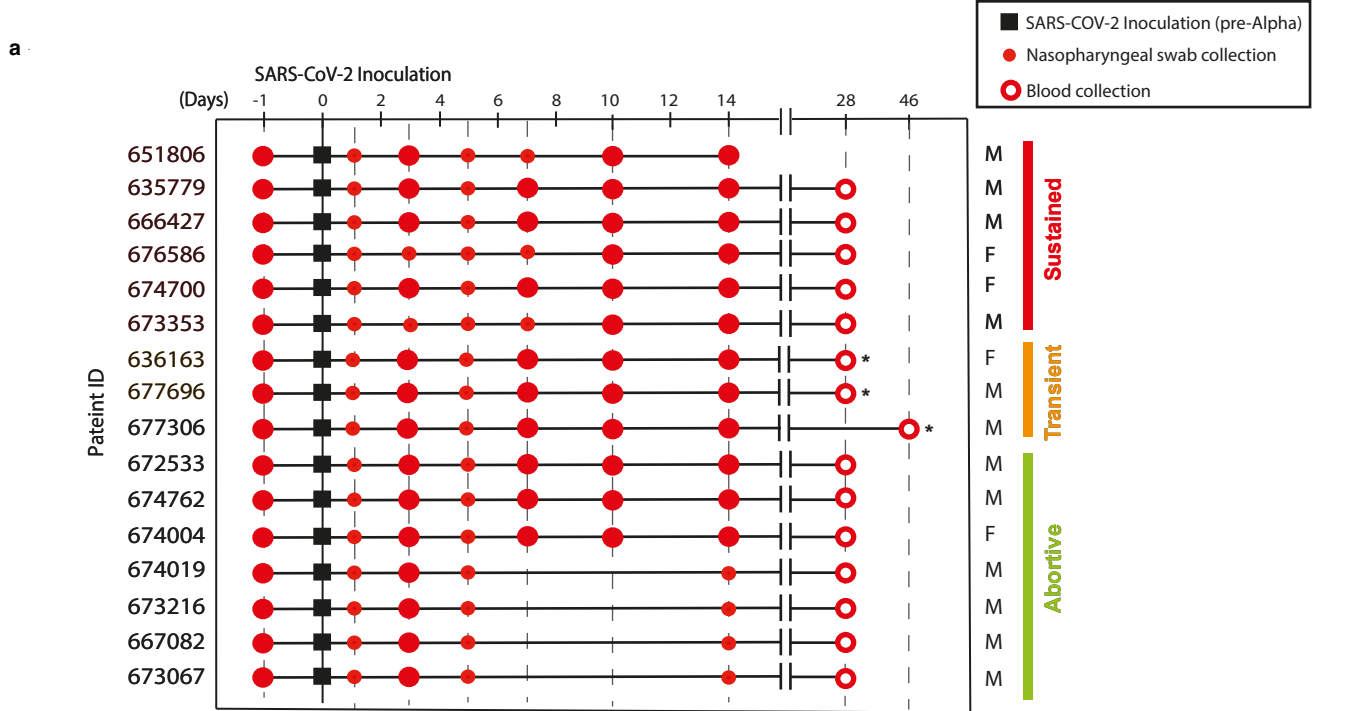
- 1316 1. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of  
1317 COVID-19. *Cell* **181**, 1036–1045.e9 (2020).
- 1318 2. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe  
1319 COVID-19 patients. *Science* **369**, 718–724 (2020).
- 1320 3. Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell  
1321 Compartment. *Cell* **182**, 1419–1440.e23 (2020).
- 1322 4. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19.  
1323 *Nat. Med.* **27**, 904–916 (2021).
- 1324 5. Yoshida, M. *et al.* Local and systemic responses to SARS-CoV-2 infection in children and  
1325 adults. *Nature* **602**, 321–327 (2022).
- 1326 6. Killingley, B. *et al.* Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge  
1327 in young adults. *Nat. Med.* **28**, 1031–1041 (2022).
- 1328 7. Fears, A. C. *et al.* The dynamics of  $\gamma\delta$  T cell responses in nonhuman primates during SARS-  
1329 CoV-2 infection. *Commun Biol* **5**, 1380 (2022).
- 1330 8. Frere, J. J. *et al.* SARS-CoV-2 infection in hamsters and humans results in lasting and unique  
1331 systemic perturbations after recovery. *Sci. Transl. Med.* **14**, eabq3059 (2022).
- 1332 9. Hinks, T. S. C. & Zhang, X.-W. MAIT Cell Activation and Functions. *Frontiers in Immunology*  
1333 vol. 11 Preprint at <https://doi.org/10.3389/fimmu.2020.01014> (2020).
- 1334 10. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
- 1335 11. Trizzino, M. *et al.* EGR1 is a gatekeeper of inflammatory enhancers in human macrophages.  
1336 *Science Advances* vol. 7 Preprint at <https://doi.org/10.1126/sciadv.aaz8836> (2021).
- 1337 12. Li, Q. & Verma, I. M. NF- $\kappa$ B regulation in the immune system. *Nat. Rev. Immunol.* **2**, 725–734  
1338 (2002).
- 1339 13. Luan, H. H. *et al.* GDF15 Is an Inflammation-Induced Central Mediator of Tissue Tolerance.  
1340 *Cell* **178**, 1231–1244.e11 (2019).
- 1341 14. Shang, Y. *et al.* The transcriptional repressor Hes1 attenuates inflammation by regulating  
1342 transcription elongation. *Nat. Immunol.* **17**, 930–937 (2016).
- 1343 15. Wang, T. *et al.* PER1 prevents excessive innate immune response during endotoxin-induced liver  
1344 injury through regulation of macrophage recruitment in mice. *Cell Death Dis.* **7**, e2176–e2176  
1345 (2016).

- 1346 16. Liu, L., Jiang, Y. & Steinle, J. J. TNFAIP3 is anti-inflammatory in the retinal vasculature. *Mol.*  
1347 *Vis.* **28**, 124–129 (2022).
- 1348 17. Scholtyssek, C., Uderhardt, S., Schett, G. & Krönke, G. NR4A1 modulates the inflammatory  
1349 response during murine experimental arthritis. *Ann. Rheum. Dis.* **69**, 37–37 (2010).
- 1350 18. Wosen, J. E., Mukhopadhyay, D., Macaubas, C. & Mellins, E. D. Epithelial MHC Class II  
1351 Expression and Its Role in Antigen Presentation in the Gastrointestinal and Respiratory Tracts.  
1352 *Front. Immunol.* **9**, 2144 (2018).
- 1353 19. Madisson, E. *et al.* A spatially resolved atlas of the human lung characterizes a gland-associated  
1354 immune niche. *Nat. Genet.* **55**, 66–77 (2023).
- 1355 20. Kaneko, N. *et al.* Temporal changes in T cell subsets and expansion of cytotoxic CD4 T cells in  
1356 the lungs in severe COVID-19. *Clinical Immunology* vol. 237 108991 Preprint at  
1357 <https://doi.org/10.1016/j.clim.2022.108991> (2022).
- 1358 21. Meckiff, B. J. *et al.* Imbalance of Regulatory and Cytotoxic SARS-CoV-2-Reactive CD4 T Cells  
1359 in COVID-19. *Cell* **183**, 1340–1353.e16 (2020).
- 1360 22. Wilkinson, T. M. *et al.* Preexisting influenza-specific CD4+ T cells correlate with disease  
1361 protection against influenza challenge in humans. *Nat. Med.* **18**, 274–280 (2012).
- 1362 23. Phad, G. E. *et al.* Clonal structure, stability and dynamics of human memory B cells and  
1363 circulating plasmablasts. *Nat. Immunol.* **23**, 1–10 (2022).
- 1364 24. Fink, K. Origin and Function of Circulating Plasmablasts during Acute Viral Infections. *Front.*  
1365 *Immunol.* **3**, 78 (2012).
- 1366 25. Dan, J. M. *et al.* A Cytokine-Independent Approach To Identify Antigen-Specific Human  
1367 Germinal Center T Follicular Helper Cells and Rare Antigen-Specific CD4+ T Cells in Blood. *J.*  
1368 *Immunol.* **197**, 983–993 (2016).
- 1369 26. Altman, J. D. *et al.* Phenotypic analysis of antigen-specific T lymphocytes. *Science* **274**, 94–96  
1370 (1996).
- 1371 27. Website. Coronavirus Clinical Characterisation Consortium. Site set-up. ISARIC 4C  
1372 <https://isaric4c.net/protocols>.
- 1373 28. Tang, Y. *et al.* Human nasopharyngeal swab processing for viable single-cell suspension v1.  
1374 *protocols.io* (2020) doi:10.17504/protocols.io.bjhkkj4w.
- 1375 29. Ziegler, C. G. K. *et al.* Impaired local intrinsic immunity to SARS-CoV-2 infection in severe  
1376 COVID-19. *Cell* **184**, 4713–4733.e22 (2021).
- 1377 30. Suo, C. *et al.* Mapping the developing human immune system across organs. *Science* **376**,

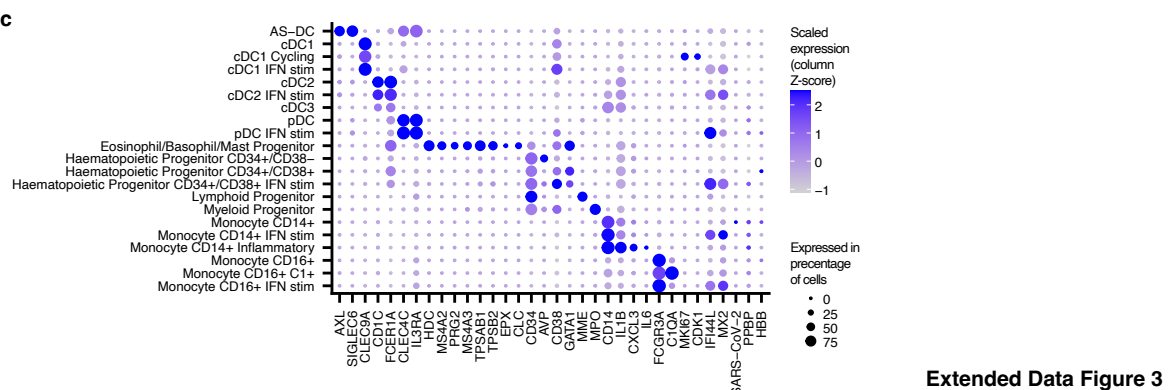
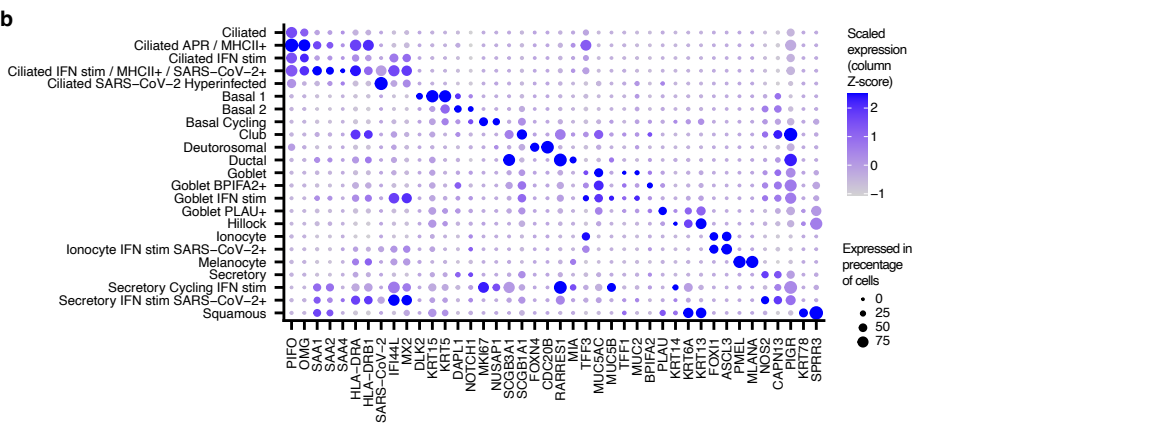
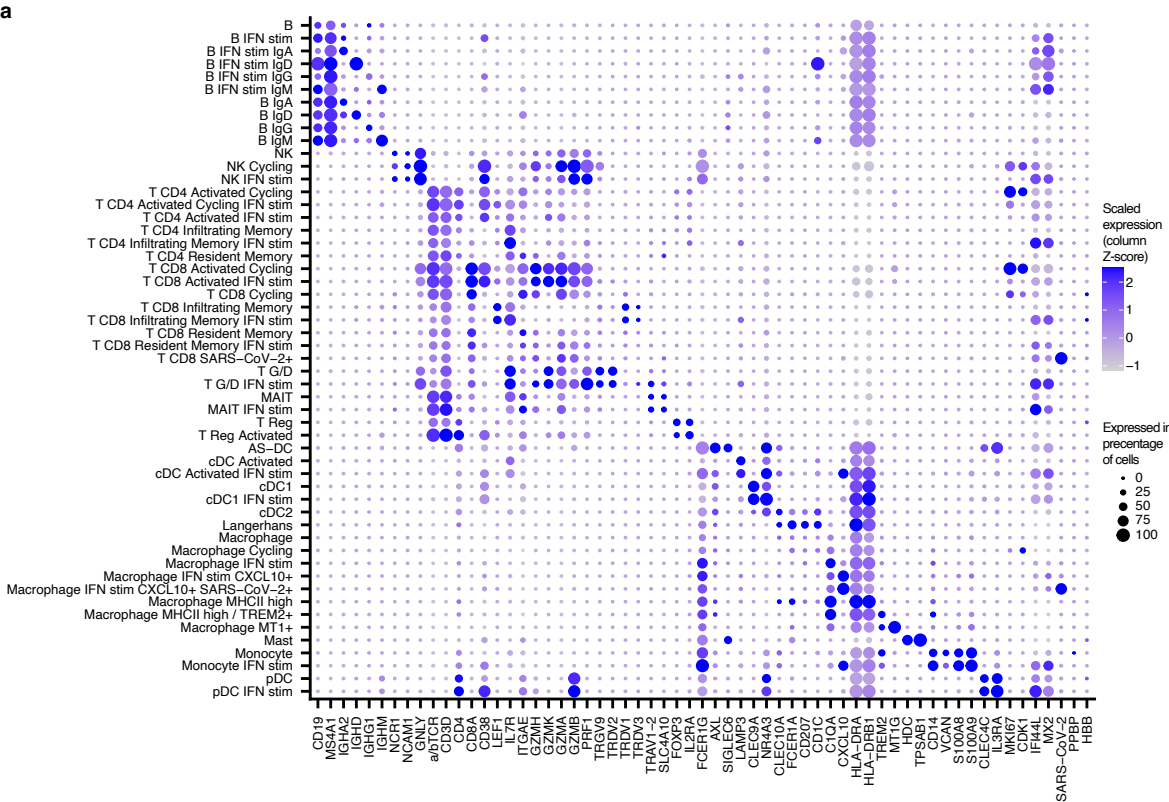
- 1378 eabo0510 (2022).
- 1379 31. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based  
1380 single-cell RNA sequencing data. *Gigascience* **9**, (2020).
- 1381 32. Heaton, H. *et al.* SoupORcell: robust clustering of single-cell RNA-seq data by genotype without  
1382 reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
- 1383 33. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony.  
1384 *Nat. Methods* **16**, 1289–1296 (2019).
- 1385 34. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-  
1386 connected communities. *Sci. Rep.* **9**, 5233 (2019).
- 1387 35. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in  
1388 humans. *Science* **376**, eabl5197 (2022).
- 1389 36. Sturm, G. *et al.* Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing  
1390 data. *Bioinformatics* **36**, 4817–4818 (2020).
- 1391 37. Suo, C. *et al.* Dandelion utilizes single cell adaptive immune receptor repertoire to explore  
1392 lymphocyte developmental origins. *bioRxiv* (2022) doi:10.1101/2022.11.18.517068.
- 1393 38. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data  
1394 analysis. *Genome Biol.* **19**, 15 (2018).
- 1395 39. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial.  
1396 *Mol. Syst. Biol.* **15**, e8746 (2019).
- 1397 40. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for  
1398 single-cell transcriptomics. *Nature Methods* vol. 15 1053–1058 Preprint at  
1399 <https://doi.org/10.1038/s41592-018-0229-2> (2018).
- 1400 41. Mayer-Blackwell, K. *et al.* TCR meta-clonotypes for biomarker discovery with enabled  
1401 identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *Elife* **10**, (2021).
- 1402 42. Csárdi, G. *et al.* *igraph*. (Zenodo, 2023). doi:10.5281/ZENODO.3630268.
- 1403 43. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**,  
1404 3645–3647 (2017).
- 1405 44. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**,  
1406 2272–2274 (2020).
- 1407 45. Bayesian Gaussian Process Latent Variable Model. *Paperpile*  
1408 <https://paperpile.com/app/p/78b378bb-4100-016d-ac7d-6ebbb9b5f9c2>.
- 1409 46. Bingham, E. *et al.* Pyro: Deep universal probabilistic programming. (2018)



- 1410 doi:10.48550/ARXIV.1810.09538.
- 1411 47. Multi-task Gaussian Process Prediction. *Paperpile* [https://paperpile.com/app/p/8ade8ba1-9869-](https://paperpile.com/app/p/8ade8ba1-9869-09eb-9355-5a8e00b75e2c)
- 1412 09eb-9355-5a8e00b75e2c.
- 1413 48. Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q. & Wilson, A. G. GPyTorch: Blackbox
- 1414 Matrix-matrix Gaussian process inference with GPU acceleration. (2018)
- 1415 doi:10.48550/ARXIV.1809.11165.
- 1416 49. Uddin, I. *et al.* An Economical, Quantitative, and Robust Protocol for High-Throughput T Cell
- 1417 Receptor Sequencing from Tumor or Blood. *Methods in Molecular Biology* 15–42 Preprint at
- 1418 [https://doi.org/10.1007/978-1-4939-8885-3\\_2](https://doi.org/10.1007/978-1-4939-8885-3_2) (2019).
- 1419 50. Oakes, T. *et al.* Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and
- 1420 Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is
- 1421 Robust, Economical, and Versatile. *Front. Immunol.* **8**, 1267 (2017).
- 1422 51. Peacock, T., Heather, J. M., Ronel, T. & Chain, B. Decombinator V4: an improved AIRR
- 1423 compliant-software package for T-cell receptor sequence annotation? *Bioinformatics* **37**, 876–
- 1424 878 (2021).
- 1425 52. Mayer, A. & Callan, C. G., Jr. Measures of epitope binding degeneracy from T cell receptor
- 1426 repertoires. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2213264120 (2023).
- 1427



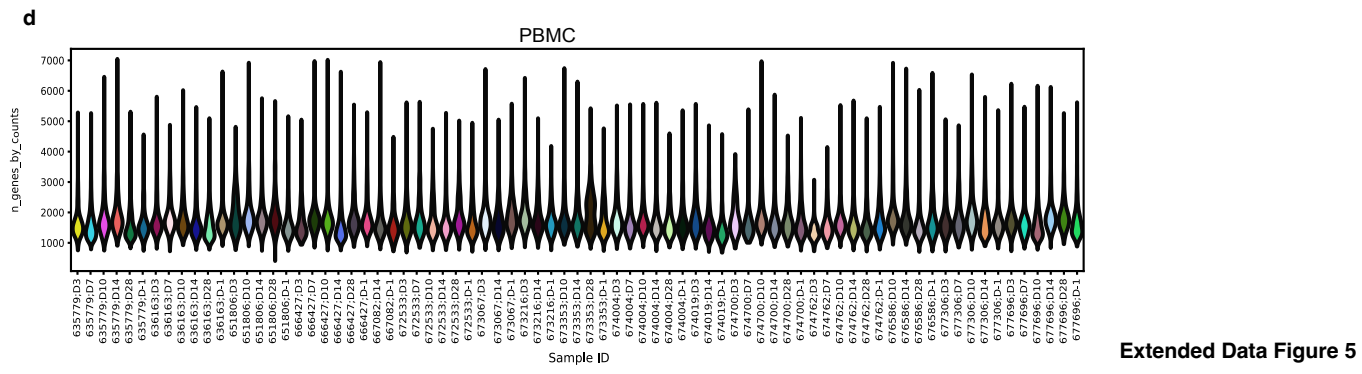
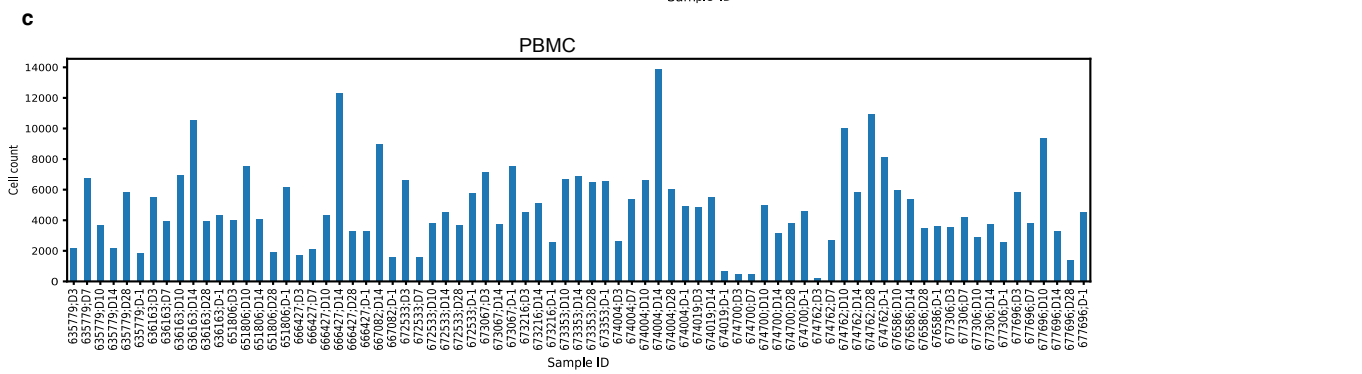
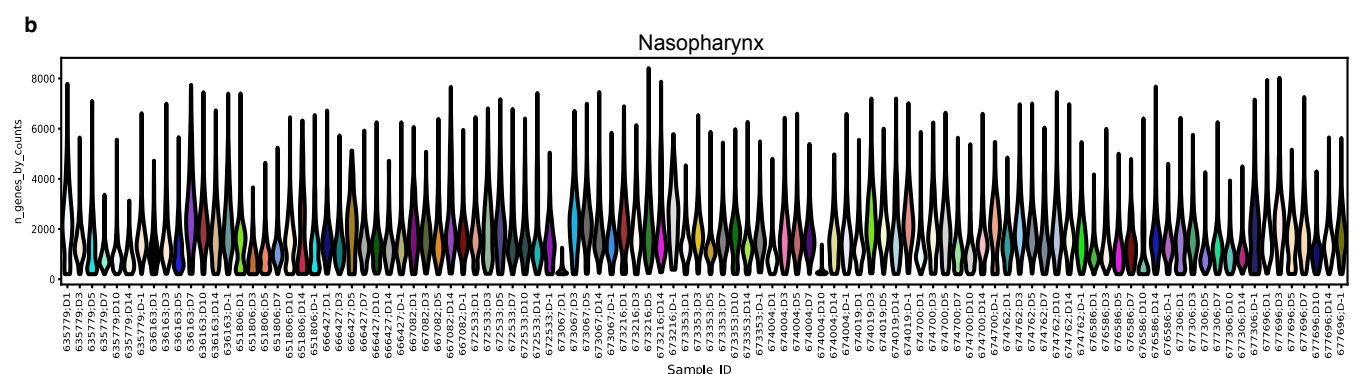
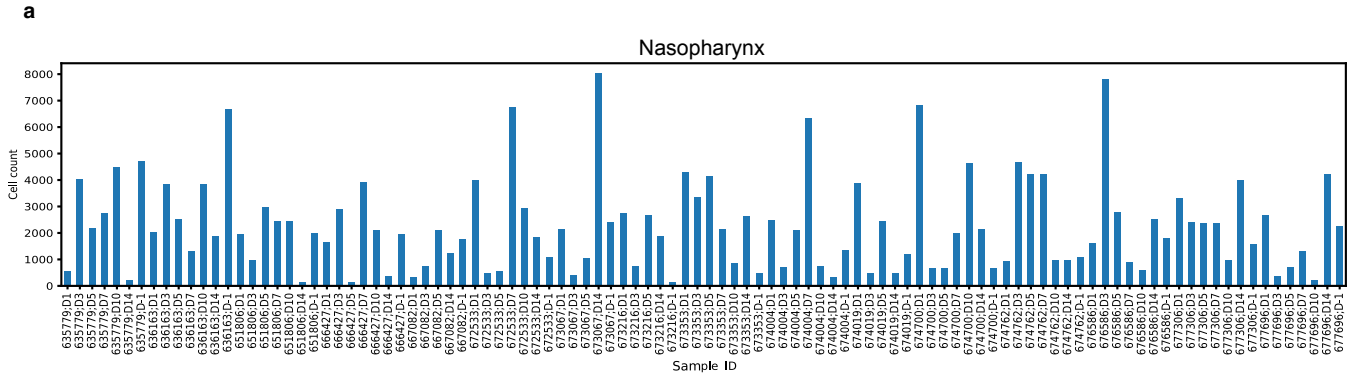


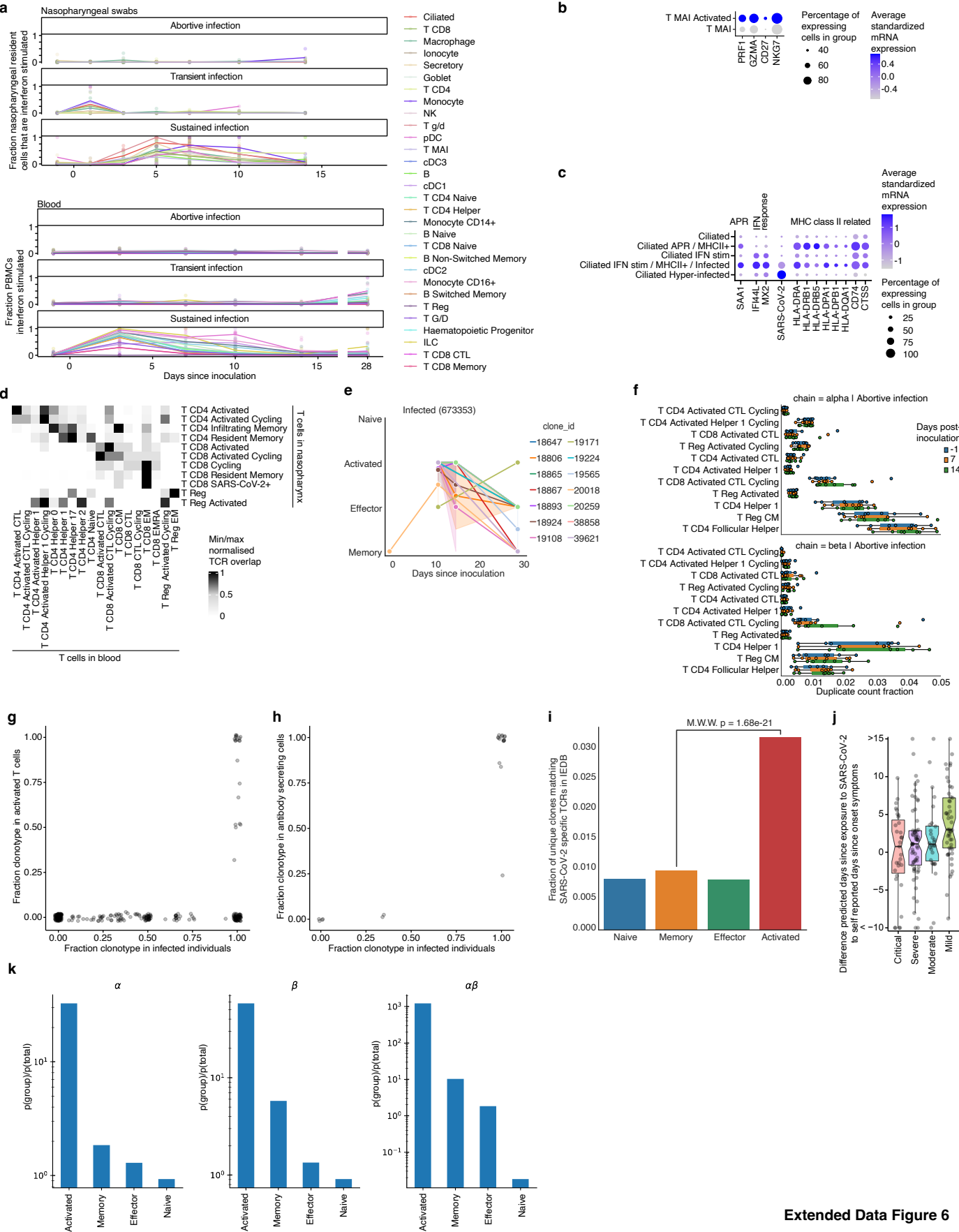


Extended Data Figure 3

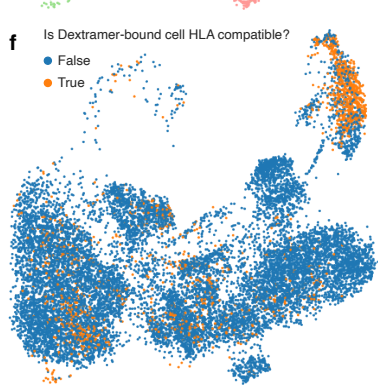
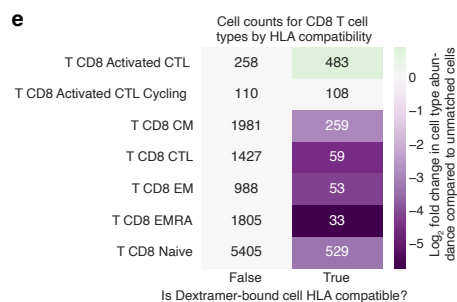
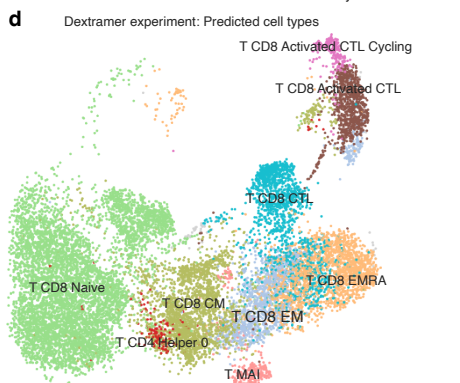
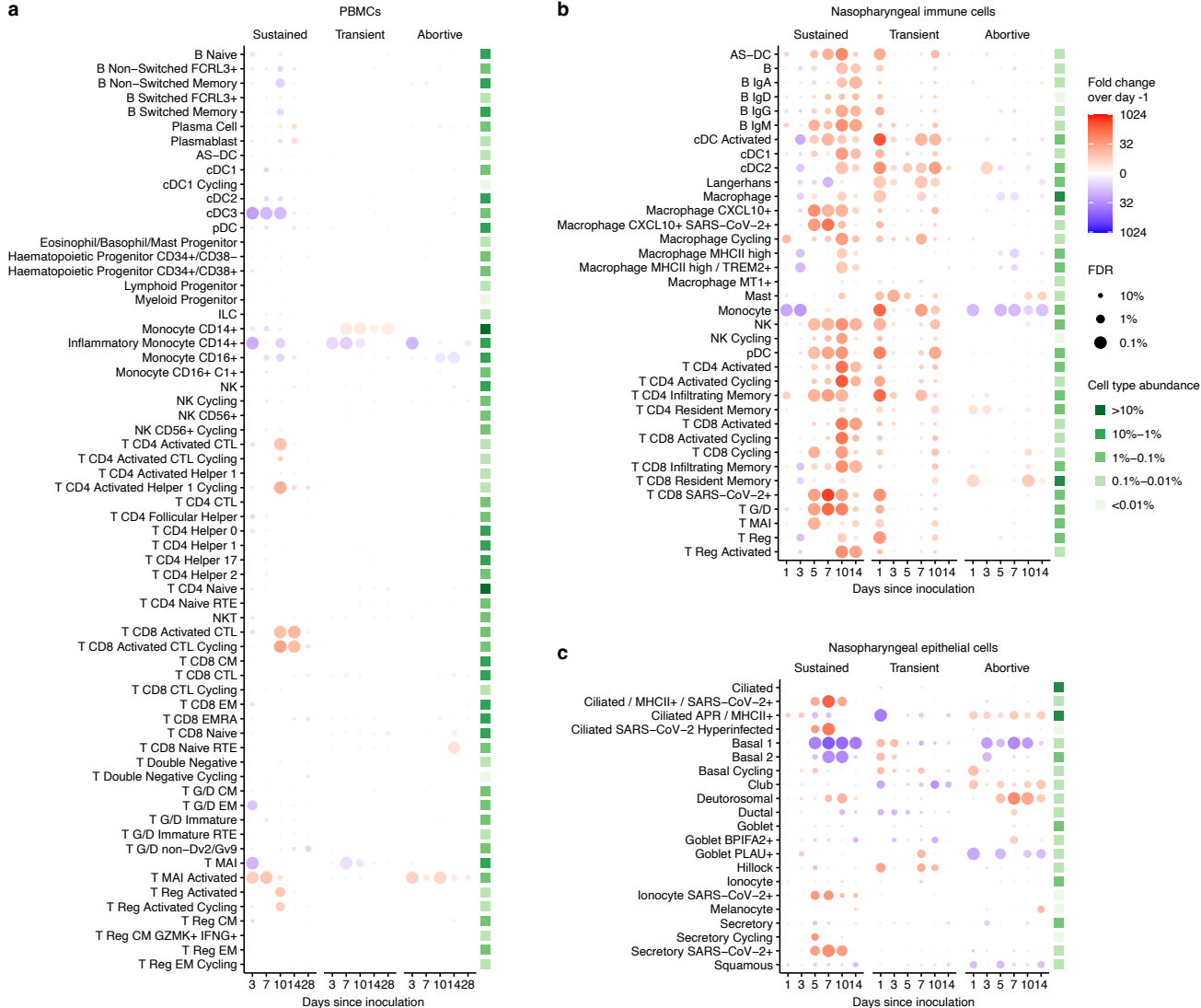


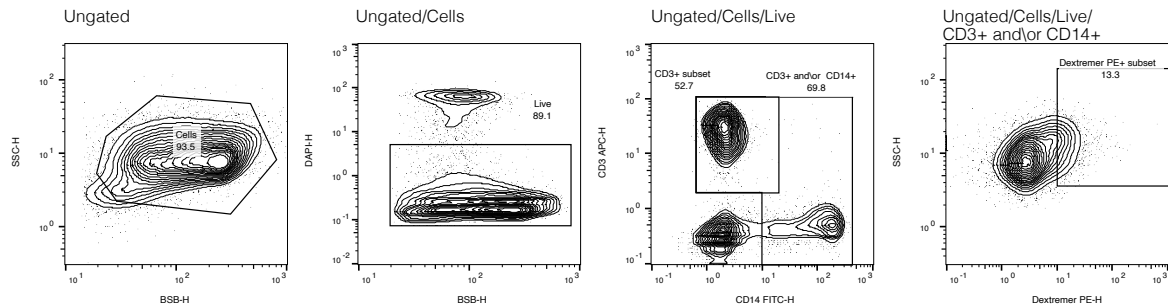






Extended Data Figure 6



**a****b**