Omicron variant of SARS-CoV-2 harbors a unique insertion mutation of putative viral or human genomic origin

A.J. Venkatakrishnan[1], Praveen Anand[2], Patrick J. Lenehan[1], Rohit Suratekar[2], Bharathwaj Raghunathan[3], Michiel J.M. Niesen[1], Venky Soundararajan[1,2,3]*

[1] nference, Cambridge, Massachusetts 02139, USA
[2] nference labs, Bengaluru, Karnataka 560017, India
[3] nference, Toronto, ON M5V 1M1, Canada
*Correspondence to: Venky Soundararajan (venky@nference.net)

Abstract

The emergence of a heavily mutated SARS-CoV-2 variant (B.1.1.529, Omicron) and it's spread to 6 continents within a week of initial discovery has set off a global public health alarm. Characterizing the mutational profile of Omicron is necessary to interpret its shared or distinctive clinical phenotypes with other SARS-CoV-2 variants. We compared the mutations of Omicron with prior variants of concern (Alpha, Beta, Gamma, Delta), variants of interest (Lambda, Mu, Eta, Iota and Kappa), and all 1523 SARS-CoV-2 lineages constituting 5.4 million SARS-CoV-2 genomes. Omicron's Spike protein has 26 amino acid mutations (23 substitutions, two deletions and one insertion) that are distinct compared to other variants of concern. Whereas the substitution and deletion mutations have appeared in previous SARS-CoV-2 lineages, the insertion mutation (ins214EPE) has not been previously observed in any SARS-CoV-2 lineage other than Omicron. The nucleotide sequence encoding for ins214EPE could have been acquired by template switching involving the genomes of other viruses that infect the same host cells as SARS-CoV-2 or the human transcriptome of host cells infected with SARS-CoV-2. For instance, given recent clinical reports of co-infections in COVID-19 patients with seasonal coronaviruses (e.g. HCoV-229E), single cell RNA-sequencing data showing co-expression of the SARS-CoV-2 and HCoV-229E entry receptors (ACE2 and ANPEP) in respiratory and gastrointestinal cells, and HCoV genomes harboring sequences homologous to the nucleotide sequence that encodes ins214EPE, it is plausible that the Omicron insertion could have evolved in a co-infected individual. There is a need to understand the function of the Omicron insertion and whether human host cells are being exploited by SARS-CoV-2 as an 'evolutionary sandbox' for host-virus and inter-viral genomic interplay.

Introduction

A new SARS-CoV-2 variant with an extensively mutated Spike protein was first reported to the World Health Organization (WHO) from South Africa on November 24, 2021, with the first sample collected on November 9, 2021. This strain has since been denoted as the Omicron variant (WHO nomenclature) and B.1.1.529 (PANGO lineage)[1]. The rapid assessment of the variant by *The Technical Advisory Group on SARS-CoV-2 Virus Evolution* and classification of Omicron as a variant of concern by the WHO within 48 hours has facilitated timely epidemiological surveillance. Since the initial discovery of Omicron, the variant has already been detected in over 20 countries across six continents.[2,3]

Thoroughly characterizing the mutational profile of Omicron is the necessary first step to begin interpreting its shared or distinctive clinical phenotypes, sensitivity or resistance to existing vaccines, and whether Omicron-like variants that evolve in the future may have heightened virulence. Indeed, SARS-CoV-2 has evolved into different variants of concern and variants of interest through a combination of missense, deletion, insertion, and other mutations. In the Spike (S) protein that engages the ACE2 receptor on human cells to facilitate viral entry, missense mutations (e.g. E484K) have led to significant changes in the Spike-ACE2 binding affinity, and deletions (e.g. ΔY144) have modulated the effect of neutralizing anti-Spike antibodies.[4–9] Insertion mutations have been less prevalent in SARS-CoV-2 evolution.[10] However, one of the most functionally consequential mutations in the evolutionary history of SARS-CoV-2 till date was the "PRRA" Spike protein insertion in the S1/S2 cleavage site, which introduced the polybasic FURIN cleavage site that mimics the RRARSVAS peptide in human ENAC-alpha.[11–14] The availability of 5.4 million SARS-CoV-2 genomes covering 1523 lineages from over 200 countries/territories in the GISAID database from the beginning of the pandemic gives an opportunity to characterize the mutational profile of the Omicron variant in comparison to other SARS-CoV-2 variants.

In this study, we compare the mutational profile of Omicron with 1523 SARS-CoV-2 lineages including the variants of concern and variants of interest. We highlight that Omicron's Spike protein harbors an insertion mutation ins214EPE that is absent in all other SARS-CoV-2 lineages. Given the salience of viral genetic recombination and the debated plausibility of host genome integration by SARS-CoV-2[15–17], we considered a variety of host-viral and inter-viral genomic matter exchange scenarios that may have contributed to the adoption of this insertion mutation in the precursor variant of Omicron. We discuss potential sources for the origin of the ins214EPE and highlight the need to experimentally characterize the role of ins214EPE for transmission and immune evasion.

Results and Discussion

Comparison of mutations in Omicron to previous SARS-CoV-2 lineages shows the presence of a unique insertion mutation in Omicron's N-terminal domain (NTD)

The Omicron variant harbors 37 mutations in the Spike protein, which include six deletion mutations, one insertion mutation and 30 substitution mutations.[18] 16 of the 37 mutations are surge-associated mutations[19] (**Table 1**), i.e. their mutational prevalence increased during any three-month window when COVID-19 cases surged. Comparing these Spike protein mutations in Omicron with pre-existing variants of concern (Alpha, Beta, Gamma and Delta) shows that 26 mutations are distinct to Omicron and 7 mutations overlap between Omicron and Alpha (**Figure 1**). We analyzed whether any of the 26 Omicron mutations appeared in the prior variants of interest (Lambda, Mu, Eta, Iota and Kappa) or prior SARS-CoV-2 lineages by comparing with mutations from 5,382,852 genomes corresponding to 1523 lineages from the GISAID database (**Table 1; Figure S1**). Interestingly, the insertion mutation ins214EPE (**Figure S2**) has not been previously observed in any SARS-CoV-2 lineage other than Omicron, whereas the substitution and deletions mutations have appeared in previous SARS-CoV-2 lineages (**Table S1**).

The EPE insertion on Omicron maps to the N-terminal domain (NTD) distal from the antibody binding supersite.[9] However, the loop where the insertion is present maps to a known human T-cell epitope on SARS-CoV-2.[20] Further studies will be necessary to understand whether this insertion may help SARS-CoV-2 escape T-cell immunity.[10] Given the importance of the PRRA insertion giving rise to a polybasic FURIN cleavage site in the original SARS-CoV-2 strain, it is important to understand the functional significance and evolutionary origins of the ins214EPE insertion in the Omicron variant.[11–14]

Origin of insEPE in Omicron potentially due to template switching using genome of co-infecting viruses or host

The mutational burden of Omicron is higher in the Spike protein than the rest of the proteome (**Figure S3**). This highly mutated Spike variant harbors a novel insertion mutation ins214EPE. Although the position 214 appears to be an insertion hotspot[10] the EPE insertion in Omicron appears to be novel. Previous analyses of sequences deposited in GISAID suggested that insertions in the SARS-CoV-2 genome likely arise from polymerase slippage or template switching.[10,21] Template switching is a normal event during RNA synthesis for coronaviruses, as this process is used to generate sub-genomic RNAs (sgRNAs).[22,23] In this process, also known as copy-choice recombination, the RNA-directed RNA polymerase (RdRp) and the nascent strand dissociate from the template RNA strand and reassociate with a new template (or the same template at a different position), and then RNA synthesis continues. Typically, such recombination involves templates with high sequence similarity ("homologous recombination"), although non-homologous (or "illegitimate") recombination between dissimilar sequences can also occur.[24]

Recombination between SARS-CoV-2 lineages in the context of simultaneous co-infection has been observed, with particularly high recombination rates seen in the Spike protein sequence.[22,25] The ins214EPE could have been acquired by template switching involving the genomes of SARS-CoV-2, other viruses that infect the same host cells as SARS-CoV-2, or the human transcriptome of host cells infected with SARS-CoV-2. Indeed, there have been clinical reports of COVID-19 patients also being infected with seasonal coronaviruses such as HCoV-229E[26]. Searching the HCoV-229E genome for homology to a nucleotide sequence encoding ins214EPE shows the presence of an identical sequence in HCoV-229E's Spike protein, which could have been exploited for template switching (**Figure 2**). Furthermore, based on analysis of single cell RNA seq data (**Table S2**), we see that the receptors of SARS-CoV-2 (ACE2) and HCoV-229E (ANPEP) are co-expressed in gastrointestinal (e.g. enterocytes) and respiratory tissues (e.g. respiratory ciliated cells). This gives rise to the plausibility of such cells in co-infected individuals being exploited as sites of genomic interplay between different viruses. In addition to co-infection with different coronaviruses, there have been reports of co-infection with SARS-CoV-2 and other respiratory pathogens including non-SARS-CoV-2 Coronaviridae.[27,28]

It has been suggested previously that insertion mutations in the SARS-CoV-2 genomes could have originated from the human host genome[29]. Indeed, numerous fragments of the human genome and transcriptome harbor nucleotide sequences that are identical to the coding sequence of ins214EPE. There are over 750 fragments of the human genome with nucleotide sequences identical to the coding sequence of ins214EPE, which include mRNAs of SLCA7 and TMEM245

as top hits (**Figure S4**). Of these, the transcripts that are expressed specifically in human host cells (e.g. alveolar cells, enterocytes)[30] that are infected by SARS-CoV-2 could be candidates for the origin of the ins214EPE sequence. Thus, the evolution of the unique insertion in Omicron could have been based on template switching during viral co-infections, or from prevalent templates in the human genome.

Even as the production of COVID-19 vaccines is being scaled up, vaccine inequity and vaccine hesitancy have been speculated as contributors to the emergence of Omicron[31,32]. Since achieving global vaccination could take years, it is important to vigilantly monitor the changing mutational landscape that could lead to the emergence of new SARS-CoV-2 variants. Indeed, even among the Omicron variants there are differences in the prevalence of the constituent mutations (**Figure S1**). Finally, there is a need to sequence SARS-CoV-2 genomes from individuals with viral co-infections and in general to develop a "variant warning system" for early detection of variants of concern based on their mutational profile.

Methods

Analysis of surge-associated mutations using GISAID database and OWID

The core SARS-CoV-2 mutations associated with each of the parental strains: Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Lambda (C.37), Mu (B.1.621), Eta (B.1.525), Iota (B.1.526) and Kappa (B.1.617.1) were derived from CoV-RDB database[18]. All the SARS-CoV-2 genome sequences corresponding to the Omicron variants along with the mutations were directly derived from GISAID database[2,3] (121 SARS-CoV-2 genomes annotated as B.1.1.529 as on 29th November 2021). Analyzing the occurrences of ins214EPE in the GISAID database, besides their prevalence in the Omicron variants two occurrences are in genomes assigned to lineages B.1.1 and B.1.1.263, which is likely due to incorrect assignment of lineages. 10 other occurrences outside Omicron are in sequences that were yet to be assigned a lineage.

To understand which SARS-CoV-2 mutations correlates with COVID-19 test positivity, we extracted "Surge-Associated mutations" from GISAID (5,382,852 sequences from 202 countries/territories between December 2019 to November 2021) and Our World in Data (OWID) data[33] as described in previous study[19]. A surge associated mutation satisfies the following criteria: (1) it is present in at least 100 GISAID sequences; (2) it is present in a time window during which there is at least a 5% increase in test positivity, over three consecutive months (in a given country); (3) it is present in time window during which there is at least a 5% increase in sequence deposition with the given mutation over three consecutive months (in a given country).

NCBI BLAST search to identify homologs to nucleotide sequence encoding Omicron's ins214EPE

Aligning the Spike protein nucleotide sequences from Omicron and the reference Wuhan SARS-CoV-2 sequences shows that two different nucleotide sequence insertions can give rise to the sequencing encoding Omicron's ins214EPE. These insertion candidates are GAGCCAGAA and GCCAGAAGA. A NCBI BLAST search was performed for 'CGTGAGCCAGAAGAT' which

includes both the possible insertion candidates (GAGCCAGAA, GCCAGAAGA), using default BLAST parameters (word size: 7; match/mismatch: -1/3; gap costs existence/extension: 5/2). The search against the database of human genome plus transcriptome database (Human G+T) containing 160592 sequences resulted in 763 hits. The search against human coronaviruses — HCoV-OC43 (taxid:31631), HCoV-229E (taxid:11137), HCoV-NL63 (taxid:277944), HCoV-229E (taxid:11137) — resulted in 709 hits.

Declaration of Interests

AJV, PA, PJL, RS, MJMN and VS are employees of nference and have financial interests in the company. nference collaborates with bio-pharmaceutical companies on data science initiatives unrelated to this study. These collaborations had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.
2. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).
3. GISAID - hCov19 Variants. https://www.gisaid.org/hcov19-variants/.
4. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* (2021) doi:10.1038/s41586-021-03398-2.
5. Uriu, K. *et al.* Neutralization of the SARS-CoV-2 Mu Variant by Convalescent and Vaccine Serum. *N. Engl. J. Med.* (2021) doi:10.1056/NEJMc2114706.
6. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).
7. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139–1142 (2021).
8. Motozono, C. *et al.* SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* **29**, (2021).
9. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021).
10. Garushyants, S. K., Rogozin, I. B. & Koonin, E. V. Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring. *Communications Biology* **4**, 1–9 (2021).
11. Anand, P., Puranik, A., Aravamudan, M., Venkatakrishnan, A. J. & Soundararajan, V. SARS-CoV-2 strategically mimics proteolytic activation of human ENaC. *Elife* **9**, (2020).
12. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* **176**, (2020).
13. Jaimes, J. A., Millet, J. K. & Whittaker, G. R. Proteolytic Cleavage of the SARS-CoV-2 Spike Protein and the Role of the Novel S1/S2 Site. *iScience* **23**, (2020).
14. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292.e6 (2020).
15. Zhang, L. *et al.* Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
16. Parry, R., Gifford, R. J., Lytras, S., Ray, S. C. & Coin, L. J. M. No evidence of SARS-CoV-2 reverse transcription and integration as the origin of chimeric transcripts in patient tissues. *Proceedings of the National Academy of Sciences of the United States of America* vol. 118 (2021).
17. Zhang, L. *et al.* Response to Parry et al.: Strong evidence for genomic integration of SARS-CoV-2 sequences

and expression in patient tissues. *Proceedings of the National Academy of Sciences of the United States of America* vol. 118 (2021).

18. Tzou, P. L. *et al.* Coronavirus Antiviral Research Database (CoV-RDB): An Online Database Designed to Facilitate Comparisons between Candidate Anti-Coronavirus Compounds. *Viruses* **12**, (2020).
19. Venkatakrishnan, A. J. *et al.* Antigenic minimalism of SARS-CoV-2 is linked to surges in COVID-19 community transmission and vaccine breakthrough infections. *medRxiv* 2021.05.23.21257668 (2021).
20. Tarke, A. *et al.* Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Rep Med* **2**, 100204 (2021).
21. Chrisman, B. S. *et al.* Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Min.* **14**, (2021).
22. Jackson, B. *et al.* Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* **184**, 5179–5188.e8 (2021).
23. Sawicki, S. G. & Sawicki, D. L. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. *Adv. Exp. Med. Biol.* **380**, (1995).
24. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617–626 (2011).
25. Turkahia, Y. *et al.* Pandemic-Scale Phylogenomics Reveals Elevated Recombination Rates in the SARS-CoV-2 Spike Region. *bioRxiv* 2021.08.04.455157 (2021) doi:10.1101/2021.08.04.455157.
26. Lau, S. K. P. *et al.* Molecular Evolution of Human Coronavirus 229E in Hong Kong and a Fatal COVID-19 Case Involving Coinfection with a Novel Human Coronavirus 229E Genogroup. *mSphere* **6**, (2021).
27. Musuuza, J. S. *et al.* Prevalence and outcomes of co-infection and superinfection with SARS-CoV-2 and other pathogens: A systematic review and meta-analysis. *PLoS One* **16**, e0251170 (2021).
28. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. *JAMA* **323**, 2085–2086 (2020).
29. thomasppeacock. Putative host origins of RNA insertions in SARS-CoV-2 genomes. https://virological.org/t/putative-host-origins-of-rna-insertions-in-sars-cov-2-genomes/761 (2021).
30. Venkatakrishnan, A. J. *et al.* Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *Elife* **9**, (2020).
31. Wroughton, L. Officials: Variants 'haunt' world with vaccine imbalance between rich and poor nations. *The Washington Post* (2021).
32. Head, M. Omicron Is Here: A Lack of COVID Vaccines Is Partly Why. https://www.scientificamerican.com/article/omicron-is-here-a-lack-of-covid-vaccines-is-partly-why1/.
33. Mathieu, E. *et al.* A global database of COVID-19 vaccinations. *Nat Hum Behav* **5**, 947–953 (2021).
34. Doddahonnaiah, D. *et al.* A Literature-Derived Knowledge Graph Augments the Interpretation of Single Cell RNA-seq Datasets. *Genes* **12**, (2021).
35. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, (2019).
36. Chua, R. L. *et al.* COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, (2020).

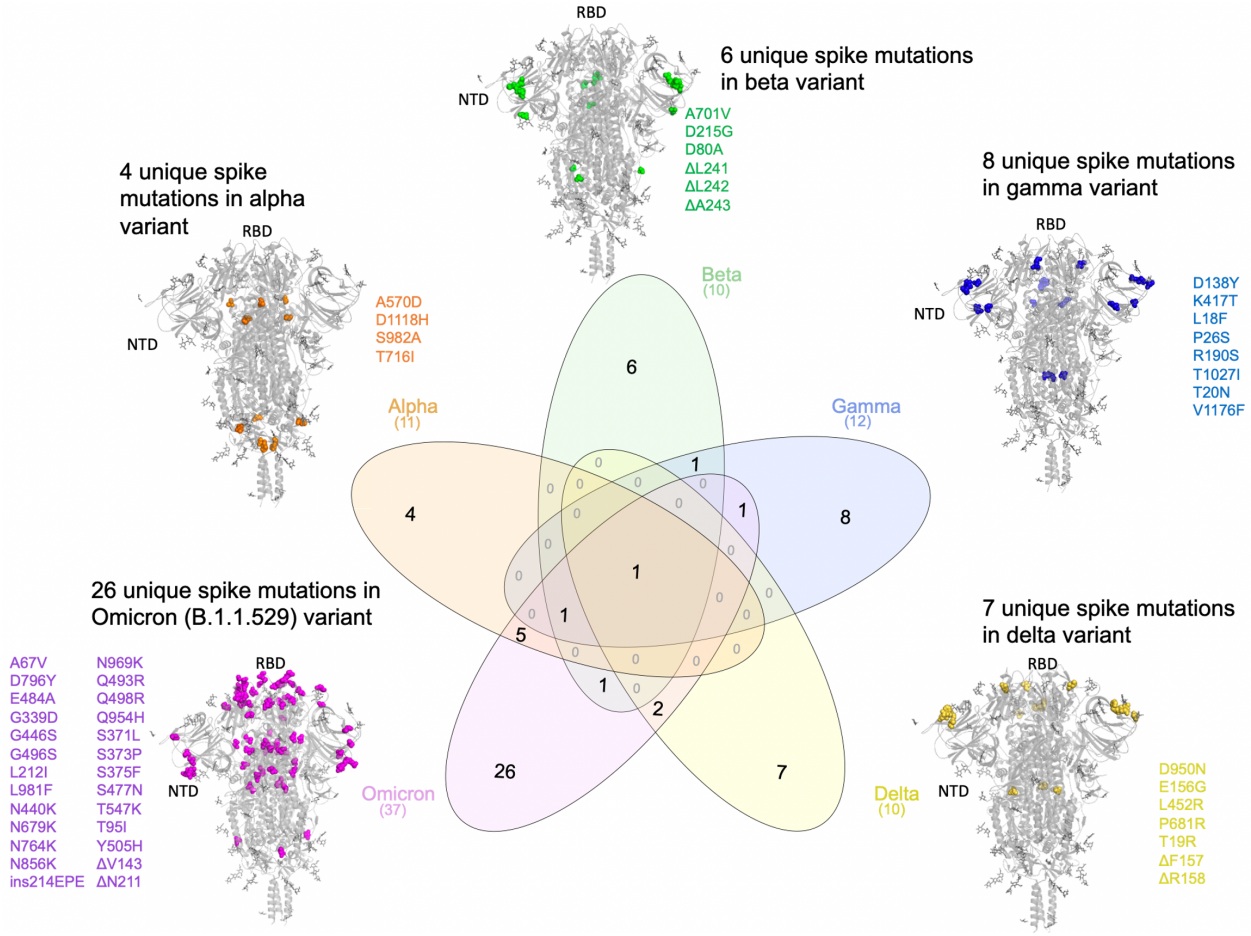
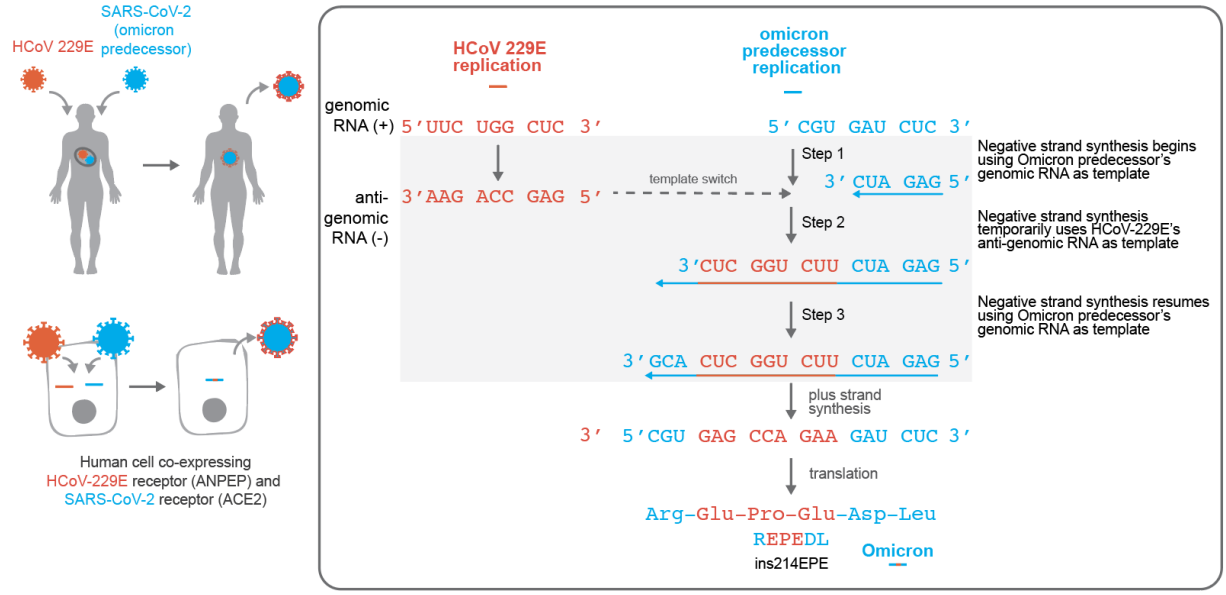


Figure 1. Venn diagram depicting the overlap of lineage specific spike mutations in the SARS-CoV-2 variants of concern. The unique key mutations observed in the spike protein for each of the variants are highlighted (spheres) on the homo-trimeric Spike protein of SARS-CoV-2. The B.1.1.529 (Omicron) variant has the highest number (26) of unique mutations in the spike protein from this perspective, making its emergence a “step function” in evolution of SARS-CoV-2 strains.

a. Potential mechanism of template switching leading to the generation of the ins214EPE in Omicron



b. Comparison of nucleotides corresponding to the Omicron insert with a homologous match from HCoV 229E

```
Omicron Spike (protein)              214        E   P   E        217
Omicron Spike genomic RNA (+)                5′ GAG CCA GAA 3′
                                                ||| ||| |||
HCoV 229E Spike antigenomic RNA(-) 263    3′ GAG CCA GAA 5′    255

HCoV 229E Spike genomic RNA(+)     255    5′ UUC UGG CUC 3′    263
      [AB691766.1]
```

Figure 2. a. Potential mechanism of template switching leading to the generation of the ins214EPE in Omicron. Schematic representations show human body and human cells being infected by Omicron's predecessor variant (blue) and a human coronavirus HCoV-229E (orange). The box shows the potential steps in the template switching involving the genomic RNA (+) of Omicron's predecessor variant and the anti-genomic RNA (-) of HCoV-229E. The steps involving the anti-genomic RNA are shown inside a grey box. **b. Comparison of nucleotides corresponding to the Omicron insert with a homologous match from HCoV 229E.** Sequence alignment of the genomic regions corresponding to Omicron Spike and HCoV-229E Spike are shown.

Table 1. Comparison of the mutations between the Omicron variant and previously identified variants of interest/concern. The first column denotes the protein domain where the mutation is present. NTD represents the N-terminal domain; RBD represents the receptor binding domain (RBM represents the receptor binding motif); PBCS denotes poly-basic cleavage site and CTD denotes C-terminal domain.

Domains	Position(s)	Surge Associated?	Omicron (B.1.1.529)	Alpha (B.1.1.7)	Beta (B.1.351)	Gamma (P.1)	Delta (B.1.617.2)	Lambda (C.37)	Mu (B.1.621)	Eta (B.1.525)	Iota (B.1.526)	Kappa (B.1.617.1)
NTD	5	Yes									L5F	
	18	Yes				L18F						
	19	Yes					T19R					
	20	Yes				T20N						
	26	Yes				P26S						
	52	No								Q52R		
	67	Yes	A67V							A67V		
	69	Yes	H69del	H69del						H69del		
	70	Yes	V70del	V70del						V70del		
	75	No						G75V				
	76	No						T76I				
	80	Yes			D80A							
	95	Yes	T95I						T95I		T95I	
	138	Yes				D138Y						
	142	Yes	G142D				G142D					
	143	Yes	V143del									
	144	Yes	Y144del	Y144del					Y144S	Y144del		
	145	Yes	Y145del	Y145del					Y145N	Y145del		
	156	Yes					E156G					
	157	Yes					F157del					
	158	Yes					R158del					
	190	Yes				R190S						
	211	No	N211del									
	212	No	L212I									
	214	**No**	**R214ins EPE**									
	215	Yes			D215G							
	241	No			L241del							
	242	Yes			L242del							
	243	Yes			A243del							
	246	Yes						R246N				
	247	Yes						S247del				
	248	No						Y248del				
	249	No						L249del				
	250	No						T250del				
	251	No						P251del				
	252	No						G252del				
	253	No						D253del			D253G	
RBD	339	No	G339D									
	346	Yes							R346K			
	371	No	S371L									
	373	No	S373P									
	375	No	S375F									
	417	Yes	K417N		K417N	K417T						
RBD (RBM)	440	No	N440K									
	446	No	G446S									
	452	Yes					L452R	L452Q			L452R	
	477	Yes	S477N									
	478	Yes	T478K				T478K					
	484	Yes	E484A		E484K	E484K			E484K	E484K	E484K	E484Q
	490	No						F490S				
	493	No	Q493K									
	496	No	G496S									
	498	No	Q498R									
	501	Yes	N501Y	N501Y	N501Y	N501Y			N501Y			
	505	No	Y505H									
PBCS	677	No								Q677H		
	679	No	N679K									
	681	Yes	P681H	P681H			P681R		P681H		P681R	

SD1/SD2	547	No	T547K									
	570	Yes		A570D								
	614	Yes	D614G	D614G	D614G	D614G	D614G	D614G	D614G	D614G	D614G	D614G
	655	Yes	H655Y			H655Y						
Others (CTD)	701	Yes			A701V						A701V	
	716	Yes		T716I								
	764	No	N764K									
	796	No	D796Y									
	856	No	N856K									
	859	Yes						T859N				
	888	No								F888L		
	950	Yes					D950N		D950N			
	954	No	Q954H									
	969	No	N969K									
	981	No	L981F									
	982	Yes		S982A								
	1027	Yes				T1027I						
	1071	No										Q1071H
	1118	Yes		D1118H								
	1176	Yes				V1176F						

Supplementary Material

Index:

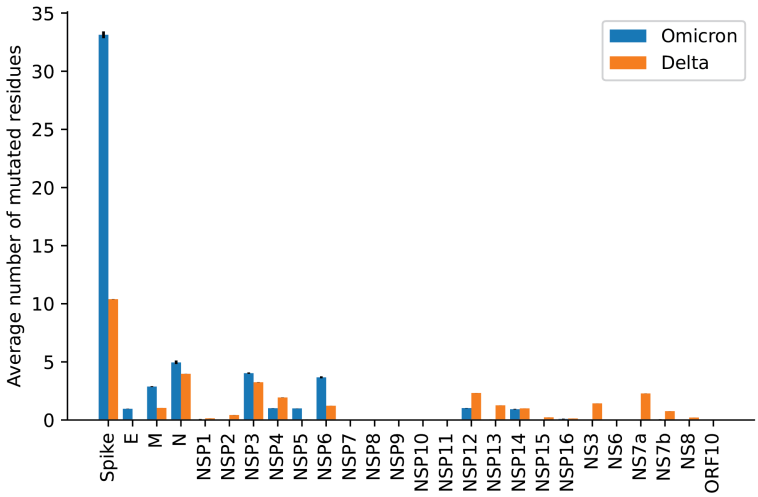


Figure S1: Mutational burden of all SARS-CoV-2 proteins for the Omicron variant (B.1.1.529) compared with the Delta variant (B.1.617.2). A total of 127 Omicron (B.1.1.529) sequences and 169,537 Delta (B.1.617.2) sequences were retrieved from the GISAID on 29 November 2021. Each bar represents the average number of mutations reported for a sequence in each of the SARS-CoV-2 proteins. Error bars represent the standard error of the mean.

```
>P0DTC2          MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS   60
>B.1.1.529       MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS   60

>P0DTC2          NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV   120
>B.1.1.529       NVTWFHVI--SGTNGTKRFDNPVLPFNDGVYFASIEKSNIIRGWIFGTTLDSKTQSLLIV   118

>P0DTC2          NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE   180
>B.1.1.529       NNATNVVIKVCEFQFCNDPFLD---HKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE   175
                                          ins214EPE:
>P0DTC2          GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVR---DLPQGFSALEPLVDLPIGINITRFQT   240
>B.1.1.529       GKQGNFKNLREFVFKNIDGYFKIYSKHTPI-IVREPEDLPQGFSALEPLVDLPIGINITRFQT   237

>P0DTC2          LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK   300
>B.1.1.529       LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK   297

>P0DTC2          CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISN   360
>B.1.1.529       CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFDEVFNATRFASVYAWNRKRISN   357

>P0DTC2          CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD   420
>B.1.1.529       CVADYSVLYNLAPFFTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGNIAD   417

>P0DTC2          YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC   480
>B.1.1.529       YNYKLPDDFTGCVIAWNSNKLDSKVSGNYNYLYRLFRKSNLKPFERDISTEIYQAGNKPC   477

>P0DTC2          NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN   540
>B.1.1.529       NGVAGFNCYFPLKSYSFRPTYGVGHQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN   537

>P0DTC2          FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP   600
>B.1.1.529       FNFNGLKGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP   597

>P0DTC2          GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY   660
>B.1.1.529       GTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEYVNNSY   657

>P0DTC2          ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI   720
>B.1.1.529       ECDIPIGAGICASYQTQTKSHRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI   717

>P0DTC2          SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE   780
>B.1.1.529       SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLKRALTGIAVEQDKNTQE   777

>P0DTC2          VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC   840
>B.1.1.529       VFAQVKQIYKTPPIKYFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC   837

>P0DTC2          LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM   900
>B.1.1.529       LGDIAARDLICAQKFKGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM   897

>P0DTC2          QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN   960
>B.1.1.529       QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNHNAQALN   957

>P0DTC2          TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA   1020
>B.1.1.529       TLVKQLSSKFGAISSVLNDIFSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA   1017

>P0DTC2          SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA   1080
>B.1.1.529       SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA   1077

>P0DTC2          ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP   1140
>B.1.1.529       ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP   1137

>P0DTC2          LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1200
>B.1.1.529       LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL   1197

>P0DTC2          QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD   1260
>B.1.1.529       QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD   1257

>P0DTC2          SEPVLKGVKLHYT                                                  1273
>B.1.1.529       SEPVLKGVKLHYT                                                  1270
```

Figure S2. Protein sequence alignment of the Spike protein of B.1.1.529 (Omicron) and the reference SARS-CoV-2. The uniprot Accession number of the reference SARS-CoV-2 Spike protein is P0DTC2 (https://www.uniprot.org/uniprot/P0DTC2). The insertion mutation ins214EPE is highlighted in blue.

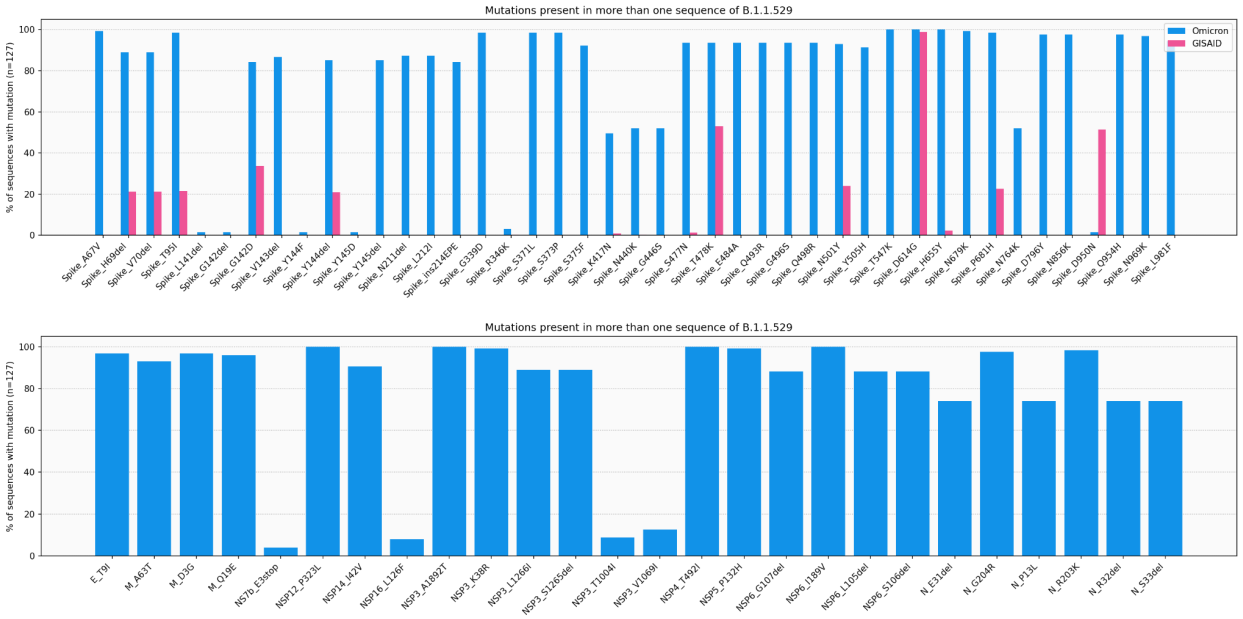


Figure S3: Distribution of prevalence of Omicron proteome mutations. A total of 127 Omicron (B.1.1.529) variant sequences from 8 countries were retrieved from the GISAID on 29 November 2021. Each blue bar represents the percentage of the 127 sequences that contain a given mutation, for the Spike protein (top) or other viral proteins (bottom). As a reference, the red bars represent the percentage of non-Omicron GISAID sequences that contain the exact same Spike-protein mutation. The figure includes only mutations which are present in more than one sequence.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Transcripts								
Homo sapiens solute carrier family 7 member 8 (SLC7A8), transcript variant 4, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	2963	NM_001267037.2
Homo sapiens solute carrier family 7 member 8 (SLC7A8), transcript variant 5, non-coding RNA	Homo sapiens	28.2	28.2	93%	24	100.00%	3077	NR_049767.2
Homo sapiens solute carrier family 7 member 8 (SLC7A8), transcript variant 2, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	3117	NM_182728.3
Homo sapiens transmembrane protein 245 (TMEM245), mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	7999	NM_032012.4
Homo sapiens solute carrier family 7 member 8 (SLC7A8), transcript variant 1, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	4236	NM_012244.4
Homo sapiens apolipoprotein A4 (APOA4), mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	1471	NM_000482.4
PREDICTED: Homo sapiens transmembrane protein 245 (TMEM245), transcript variant X6, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	7747	XM_011518452.2
PREDICTED: Homo sapiens transmembrane protein 245 (TMEM245), transcript variant X4, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	7865	XM_011518449.2
PREDICTED: Homo sapiens uncharacterized LOC101927857 (LOC101927857), transcript variant X2, nc...	Homo sapiens	28.2	28.2	93%	24	100.00%	4327	XR_001753417.2
PREDICTED: Homo sapiens transmembrane protein 245 (TMEM245), transcript variant X7, mRNA	Homo sapiens	28.2	28.2	93%	24	100.00%	7722	XM_017014572.1

Figure S4: Identification of nucleotide sequences in the human genome homologous to Omicron insertion ins214EPE using NCBI BLAST. Subset of 763 hits based on a NCBI BLAST search for 'CGTGAGCCAGAAGAT', which includes the two possible insertion nucleotide sequences (GAGCCAGAA, GCCAGAAGA) that can code for Omicron's ins214EPE. The search was performed against the database of human genome plus transcriptome database (Human G+T) containing 160592 sequences. The top hits include SLC7A8 and TMEM245.

Table S1. Number of PANGO lineages with mutations in the Omicron variant's Spike protein.

AA Substitutions	Number of Pango lineage in which mutation has observed As of 29 Nov 2021
Spike_A67V	250
Spike_H69del	379
Spike_V70del	378
Spike_T95I	509
Spike_G142D	246
Spike_V143del	294
Spike_Y144del	573
Spike_Y145del	149
Spike_N211del	90
Spike_L212I	72
Spike_ins214EPE	**0**
Spike_G339D	62
Spike_S371L	2
Spike_S373P	63
Spike_S375F	48
Spike_K417N	162
Spike_N440K	82
Spike_G446S	87
Spike_S477N	254
Spike_T478K	304
Spike_E484A	87
Spike_Q493K	28
Spike_G496S	18
Spike_Q498R	37
Spike_N501Y	348
Spike_Y505H	15
Spike_T547K	76
Spike_D614G	1464
Spike_H655Y	357
Spike_N679K	122
Spike_P681H	374
Spike_N764K	32
Spike_D796Y	192
Spike_N856K	20
Spike_Q954H	14
Spike_N969K	25
Spike_L981F	15

Table S2. Co-expression analysis of ACE2 and ANPEP in single cell RNA-sequencing datasets. The number and percent of each cell type expressing ACE2, ANPEP, or both ACE2 and ANPEP are shown. The Observed/Expected Ratio is calculated assuming that expression of ACE2 and ANPEP is distributed randomly across the analyzed cells. Specifically, the Observed/Expected Ratio is calculated by dividing the co-expressing percentage by the product of the individual expression percentages and multiplying by 100. In the first row, "All studies" corresponds to the set of studies that are hosted in the previously described Single Cell application at academia.nferx.com.[30,34] The analyzed enterocytes are derived from a study of ileal biopsies from Crohn's Disease patients.[35] The analyzed respiratory ciliated cells and FOXN4+ respiratory epithelial cells are derived from a study of nasopharyngeal and bronchial samples from COVID-19 patients and healthy controls.[36]

Study source (PMID)	Cell Type	Number of cells	Cells expressing ACE2 (%)	Cells expressing ANPEP (%)	Cells co-expressing ACE2 and ANPEP (%)	Observed / Expected Ratio
All studies	*All cells*	*2.8M*	*26.9K (0.96%)*	*221.3K (7.9%)*	*13.5K (0.48%)*	*6.4*
PMID 31474370	*All cells*	*32.5K*	*432 (1.3%)*	*1.4K (4.3%)*	*403 (1.2%)*	*21.7*
	Enterocytes	809	337 (41.7%)	757 (93.6%)	329 (40.7%)	1.04
PMID 32591762	*All cells*	*135.6K*	*3K (2.2%)*	*15.4K (11.4%)*	*1.5K (1.1%)*	*4.4*
	Respiratory ciliated cells	5.8K	827 (14.3%)	1.2K (20.7%)	217 (3.7%)	1.3
	FOXN4+ respiratory epithelial cells	787	89 (11.3%)	302 (38.4%)	56 (7.1%)	1.6