SPEAR: Systematic ProtEin AnnotatoR

Matthew Crown[1], Natalia Teruel[2], Rafael Najmanovich[2], Matthew Bashton[1*]

1. Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK
2. Department of Pharmacology and Physiology, Université de Montréal, H3T 1J4, QC, Canada
   * corresponding author: matthew.bashton@northumbria.ac.uk

Abstract

Summary: We present SPEAR, a lightweight and rapid SARS-CoV-2 variant annotation and scoring tool, for identifying mutations contributing to potential immune escape and transmissibility (ACE2 binding) at point of sequencing. SPEAR can be used in the field to evaluate genomic surveillance results in real-time and features a powerful interactive data visualisation report.

Availability and implementation: SPEAR and documentation are freely available on GitHub: https://github.com/m-crown/SPEAR and is implemented in Python and installable via Conda environment.

Contact: matthew.bashton@northumbria.ac.uk

Supplemental: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The SARS-CoV-2 virus caused a global pandemic with >5.8 million deaths and more than 412 million infections worldwide at time of writing. During this period there have been several variants of concern (VoCs), with enhanced transmissibility and/or immune escape (O'Toole et al., 2021; Kraemer et al., 2021; Twohig et al., 2022; Graham et al., 2021; Elliott et al., 2021; Davis et al., 2021; Chen et al., 2021; Wang et al., 2021). Currently these VoCs are defined by health authorities, World Health Organisation and/or by lineages, such as PANGO Lineages (Rambaut et al., 2020) or Nextstrain (Hadfield et al., 2018). These designations are reactive, based on a set of novel mutations must be observed as a distinct clade before being first

assigned a lineage, and only then can it be labelled as such in sequencing output and the spread of the variant tracked. There is a clear and pressing need to be able to identify evolutionary ingress of potentially problematic variants as they emerge directly from the sequencing data. Especially as we now approach the endemic stage of the pandemic where transmission levels are high, and surveillance and mitigations may no longer be in place.

To this end we present Systematic ProtEin AnnotatoR (SPEAR), a tool to flag potential variants of concern, highlighting samples that show potentially elevated immune escape and enhanced infectivity at point of sequencing. SPEAR is a lightweight functional genomic surveillance discovery tool, that utilises information from protein structure, deep mutational scanning (DMS), and computational molecular biophysics to provide comprehensive full protein product annotation for SARS-CoV-2.

2 Implementation

SPEAR integrates existing tools with its own internal annotation and QC processes, to ensure that annotation is informative, particularly when dealing with low-quality sequences. A more detailed overview of the implementation is available in supplementary information S1 and Fig S1.

2.1 Input

SPEAR is written in Python (version 3.10) and utilises Snakemake (Mölder et al., 2021) for workflow management and parallel job execution. SPEAR is flexible, allowing for single and multi-sample input in the form of either: consensus FASTA sequence, FASTA multiple sequence alignment (MSA), or VCF file(s).

2.2 Variant Detection

SPEAR aligns consensus input files to SARS-CoV-2 reference genome NC_045512.2 using MUSCLE (Edgar, 2004). Single Nucleotide Polymorphisms (SNPs) are obtained from the

alignment. SPEAR detects indels and multi-nucleotide polymorphisms (MNPs) in the alignment and combines linked events (*e.g.* SNP followed by deletion) into a single VCF row for accurate amino acid (AA) consequence description.

2.3 Quality Control (QC)

SPEAR integrates several QC checks for input samples. All consensus and alignment inputs are checked for unknown base (N) content in the genome (default 50%), which is a sign of poor sequence quality and may bias downstream score estimates. SNPs can also be optionally filtered to remove commonly problematic positions.

SPEAR also provides a Spike protein dropout detection system with user configurable parameters to flag gaps in Spike coverage (>150bp) due to amplicon dropout, as well as flagging high levels of global N content (>25%), and Ns in the Spike receptor binding domain (RBD) (>12nt). Dropout detection is critical as missing mutations in Spike can't be scored and need to be drawn to the user's attention.

2.4 SPEAR Annotation

SnpEff (Cingolani et al., 2012) is leveraged to annotate basic variant consequences. Compound HGVS.p format variants are then expanded by SPEAR to individual AA variants, *e.g.*:

S: p.G142_Y145delinsD to G142D,V143del,Y144del,Y145del

SPEAR utilises the amino acid and gene annotations from SnpEff for its downstream functional and structural annotations. The full SnpEff annotation is retained in the final VCF file produced for each sample. SPEAR examines the AA variants in the Spike and evaluates these to show potential increases in immune escape relative to that of the original "wild-type" for the different Barns classes (Barnes et al., 2020) of antibody binding epitopes using DMS data (Greaney, Loes, et al., 2021; Dong et al., 2021; Greaney, Starr, et al., 2021; Starr, Greaney, Dingens, et al., 2021; Starr, Greaney, Addetia, et al., 2021; Starr, Czudnochowski,

et al., 2021; Tortorici et al., 2021). ACE2 binding is assessed using DMS scores available for the Spike Receptor Binding Domain (RBD, Starr et al., 2020) which show the likely impact of each mutation, and Vibrational Difference Scores (VDS) (Teruel et al., 2021) which shows the propensity for the open conformational state (that correlates with infectivity). Scoring operates over the RBD (AA:331-531) for all scores except VDS, which covers a larger region of Spike (AA:14-913). Descriptions of the scoring system can be found in Supplemental material S1.

SPEAR also annotates all structural, non-structural, and accessory proteins for which structural information is present, including protein-protein interaction interfaces (such as replicase complex subunits and oligomerisation), ligand binding and active sites residues as well as domain boundaries, using a manually curated set of annotations derived from protein structures. Example per sample output for BA.1 is found in Supplemental S2.

2.5 Summary Output

SPEAR will provide a per run score summary which sums immune escape and ACE2 interaction scores for each sample. This is the sum of scores for all variants within the sample. A terminal output table of these scores is produced using the Rich python package.

SPEAR produces a HTML report with interactive heatmaps of scores for all samples and metrics as well as sortable tables of mutations and associated scores, built using Plotly and Bootstrap. This report is distributed with all dependencies required for offline viewing. Per sample ORF plots can also be viewed that enable the mutations to be shown against the protein product they reside in along with associated scores.

Supplemental S3 contains example HTML reports and ORF plots for the Alpha, Delta, Omicron, BA.1, BA.1.1 and BA.2 lineages.

2.6 Baseline comparison

SPEAR is distributed with a set of baseline scores for lineages and VoCs which can be selected to compare samples against. The default is to compare to BA.1, (Alpha, Delta, Omicron, BA.1, BA.1.1 and BA.2 can also be selected) or a user provided baseline can be used. Where scores exceed the baseline, these are highlighted in both the HTML and terminal sample summary tables. This enables samples scoring higher for any one metric than the current predominant lineage to be highlighted.

3 Application

SPEAR can be used to identify new sets of mutations within samples of interest from genomic surveillance. The example report baselined to Delta (Supplemental S3) highlights this for historical and current lineages/VoCs. Both the heatmaps and scores summary table show that Omicron and its BA sub-lineages score higher for immune escape than both Alpha and Delta, immediately drawing attention to the emergence of new threats in a surveillance setting (Fig. 1a). Using a report baselined to BA.1 (predominant at time of writing in UK) (Supplemental S3), it is apparent that BA.1.1 (a recently designated sublineage) has increased immune escape specifically for epitope classes 2 and 3 (Fig. 1b,c), driven by S:R346K. BA.2 (a Variant Under Investigation) shows increased escape in classes 1 (driven by S:D405N) and 4 (driven by S:R408S, Fig. 1d), but reduced escape in classes 2 and 3 owing to the lack of S:G446S (Fig. 1b,c).

4 Conclusion

SPEAR provides rapid assessment of mutations in SARS-CoV-2 samples and can be run without reliance on external web servers. Together with its lightweight implementation this allows for deployment both in the field and in pathogen surveillance labs worldwide.

Funding

MB and MC are funded by Research England's Expanding Excellence in England (E3) Fund. MB is funded by the UK Health Security Agency. This work is supported by COG-UK. RJN a

member of the Réseau Québécois de Recherche sur les Médicaments (RQRM) and the Quebec Network for Research on Protein Function, Engineering and Applications (PROTEO).

Supplemental material

S1. Further details of SPEAR implementation (Fig S1). Definitions of SPEAR scores and summary of scores Tables S1 and S2.
S2. Example of SPEAR annotation output for BA.1
S3. Example Summary Reports baselined to Delta and BA.1

References

Barnes,C.O. *et al.* (2020) SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, 588, 682–687.
Chen,R.E. et al. (2021) Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat Med*, 27, 717–726.
Cingolani,P. et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6, 80–92.
Davis,C. et al. (2021) Reduced neutralisation of the Delta (B.1.617.2) SARS-CoV-2 variant of concern following vaccination. *Plos Pathog*, 17, e1010022.
Dong,J. et al. (2021) Genetic and structural basis for SARS-CoV-2 variant neutralization by a two-antibody cocktail. *Nat Microbiol*, 6, 1233–1244.
Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics*, 5, 113.
Elliott,P. et al. (2021) Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science*, 374, eabl9551.
Graham,M.S. et al. (2021) Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Heal*, 6, e335–e345.
Greaney,A.J., Loes,A.N., et al. (2021) Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe*, 29, 463-476.e6.
Greaney,A.J., Starr,T.N., et al. (2021) Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat Commun*, 12, 4196.
Hadfield,J. et al. (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinform Oxf Engl*, 34, 4121–4123.
Kraemer,M.U.G. et al. (2021) Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science, 373, 889–895.
Mölder,F. et al. (2021) Sustainable data analysis with Snakemake. *F1000research*, 10, 33.
O'Toole,Á. et al. (2021) Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. *Wellcome Open Res*, 6, 121.
Rambaut,A. et al. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol, 5, 1403–1407.
Starr,T.N., Greaney,A.J., Dingens,A.S., et al. (2021) Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016.

Cell Reports Medicine, 2, 100255.
Starr,T.N. et al. (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell, 182, 1295-1310.e20.
Starr,T.N., Greaney,A.J., Addetia,A., et al. (2021) Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science, 371, 850–854.
Starr,T.N., Czudnochowski,N., et al. (2021) SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. Nature, 597, 97–102.
Teruel,N. et al. (2021) Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants. *Plos Comput Biol*, 17, e1009286.
Tortorici,M.A. et al. (2021) Broad sarbecovirus neutralization by a human monoclonal antibody. Nature, 597, 103–108.
Twohig,K.A. et al. (2022) Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect Dis*, 22, 35–42.
Wang,P. et al. (2021) Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature, 593, 130–135.