PLOS ONE

RESEARCH ARTICLE

A general approach for predicting protein epitopes targeted by antibody repertoires using whole proteomes

Michael L. Paull*, Tim Johnston, Kelly N. Ibsen, Joel D. Bozekowski, Patrick S. Daugherty*

Department of Chemical Engineering, University of California Santa Barbara, California, United States of America

* michael.paull444@gmail.com (MLP); psdaug@gmail.com (PSD)

OPEN ACCESS

Citation: Paull ML, Johnston T, Ibsen KN, Bozekowski JD, Daugherty PS (2019) A general approach for predicting protein epitopes targeted by antibody repertoires using whole proteomes. PLoS ONE 14(9): e0217668. https://doi.org/10.1371/journal.pone.0217668

Editor: Natalia V Cheshenko, Albert Einstein College of Medicine, UNITED STATES

Received: May 14, 2019

Accepted: August 22, 2019

Published: September 6, 2019

Data Availability Statement: All 278 antibody-binding peptide files are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.v7d0350).

Funding: This work was funded in part by grant M2016219 from the Brightfocus Foundation (https://www.brightfocus.org/) to PSD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Antibodies are essential to functional immunity, yet the epitopes targeted by antibody repertoires remain largely uncharacterized. To aid in characterization, we developed a generalizable strategy to predict antibody-binding epitopes within individual proteins and entire proteomes. Specifically, we selected antibody-binding peptides for 273 distinct sera out of a random library and identified the peptides using next-generation sequencing. To predict antibody-binding epitopes and the antigens from which these epitopes were derived, we tiled the sequences of candidate antigens into short overlapping subsequences of length k (k-mers). We used the enrichment over background of these k-mers in the antibody-binding peptide dataset to predict antibody-binding epitopes. As a positive control, we used this approach, termed K-mer Tiling of Protein Epitopes (K-TOPE), to predict epitopes targeted by monoclonal and polyclonal antibodies of well-characterized specificity, accurately recovering their known epitopes. K-TOPE characterized a commonly targeted antigen from *Rhinovirus A*, predicting four epitopes recognized by antibodies present in 87% of sera (n = 250). An analysis of 2,908 proteins from 400 viral taxa that infect humans predicted seven enterovirus epitopes and five Epstein-Barr virus epitopes recognized by >30% of specimens. Analysis of *Staphylococcus* and *Streptococcus* proteomes similarly predicted 22 epitopes recognized by >30% of specimens. Twelve of these common viral and bacterial epitopes agreed with previously mapped epitopes with p-values < 0.05. Additionally, we predicted 30 HSV2-specific epitopes that were 100% specific against HSV1 in novel and previously reported antigens. Experimentally validating these candidate epitopes could help identify diagnostic biomarkers, vaccine components, and therapeutic targets. The K-TOPE approach thus provides a powerful new tool to elucidate the organisms, antigens, and epitopes targeted by human antibody repertoires.

Introduction

Immunological memory allows for rapid antibody responses towards diverse antigens long after initial exposure. For example, the adaptive immune response to many vaccinations is often sustained throughout an individual's lifetime [1]. This immunological information is

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: PSD is a Director, Officer, and Stockholder, and TJ and JDB are Employees of immune repertoire mapping company SerImmune. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

archived within the genes encoding B-cell and T-cell receptors along with the corresponding receptor structures, but has proven difficult to characterize in a comprehensive manner. The ability to more fully interrogate immunological memory could reveal exposures to pathogens, commensal organisms, and allergens. Such information has proven useful for correlating antibody responses with disease outcomes to design more effective vaccines [2]. A detailed record of immune exposures can also facilitate the identification of biomarkers to diagnose infectious [3], autoimmune [4], and allergic conditions [5]. Furthermore, the capability to broadly characterize antibody repertoires at the epitope level could be used to identify conserved pathogen epitopes [6] and tumor specific antigen epitopes [7] to aid in therapeutic discovery.

A disease with prominent antibody responses is the common viral infection HSV, which causes human infections in the orofacial region ("cold sores") and the genital region ("genital ulcers") [8]. In 2012, the global prevalence of HSV1 was 3.7 billion people ages 0–49 [9] and the global prevalence of HSV2 was 417 million people ages 15–49 [10]. Diagnostic discovery generally focuses on diagnosing HSV2, since HSV2 infections can exacerbate HIV infections [10]. However, HSV1 and HSV2 contain the same genes [11] and the protein-coding regions of the HSV1 and HSV2 genomes share 83% sequence homology [12]. Therefore, researchers have often analyzed HSV glycoprotein G, since it differs substantially between the two HSV species [13]. In general, efforts have been limited to analyses of the surface-exposed envelope glycoproteins [14–17], using approaches such as microarrays [18]. Therefore, it would be novel to probe immunological memory using the entire proteomes of HSV1 and HSV2.

Immunological memory has been investigated extensively through sequencing the variable regions of B- and T-cell receptor encoding genes amplified from circulating cells [19]. These methods have proven useful for identifying receptor-encoding genes that associate with vaccination [20]. Nevertheless, such genetic information has not generally provided insight into the specific environmental antigens and epitopes targeted, unless they are known *a priori*. Furthermore, these methods require large specimen volumes (>10 mL) to obtain a sufficient quantity of cells [20]. Thus, there remains a need for methods that identify the diverse antigen targets of adaptive immunity.

Several methods have been developed to profile the protein epitopes of the secreted antibody repertoire [21]. Approaches have often focused on linear epitopes since 85% of epitopes contain at least one contiguous stretch of five amino acids [22]. By analyzing linear epitopes, researchers have identified sensitive and specific diagnostic epitopes for numerous diseases [21]. One common approach to epitope mapping is to generate short overlapping peptides by tiling candidate antigens. These peptides are then assayed for serum antibody reactivity in peptide microarray [23] or bacteriophage display library [24] formats. However, because these methods are biased towards specific organisms, they do not enable comprehensive or hypothesis-free immune evaluation. One strategy to overcome the limitations of tiling experiments is to use fully random peptide libraries [5,25,26]. Here, experiments are less biased and methods can analyze epitopes corresponding to a variety of organisms and antigens. A disadvantage of microarrays is that they are typically several orders of magnitude less diverse than peptide display libraries (e.g. 10^5 [25] versus 10^{10} [5]), limiting the effectiveness with which current methods can achieve epitope discovery for low titer antibodies. In random library experiments, epitopes are typically discovered using *de novo* motif discovery by unsupervised clustering [27]. The most widely used algorithm for this purpose, MEME, scales approximately quadratically with the number of input sequences, making it less useful for analyzing large datasets resulting from next generation sequencing (NGS). While full-length antibody-binding peptides can be analyzed, the majority of the binding energy is typically derived from just 5–6 amino acids [28], thus other amino acids within the peptide will contribute noise. To rectify

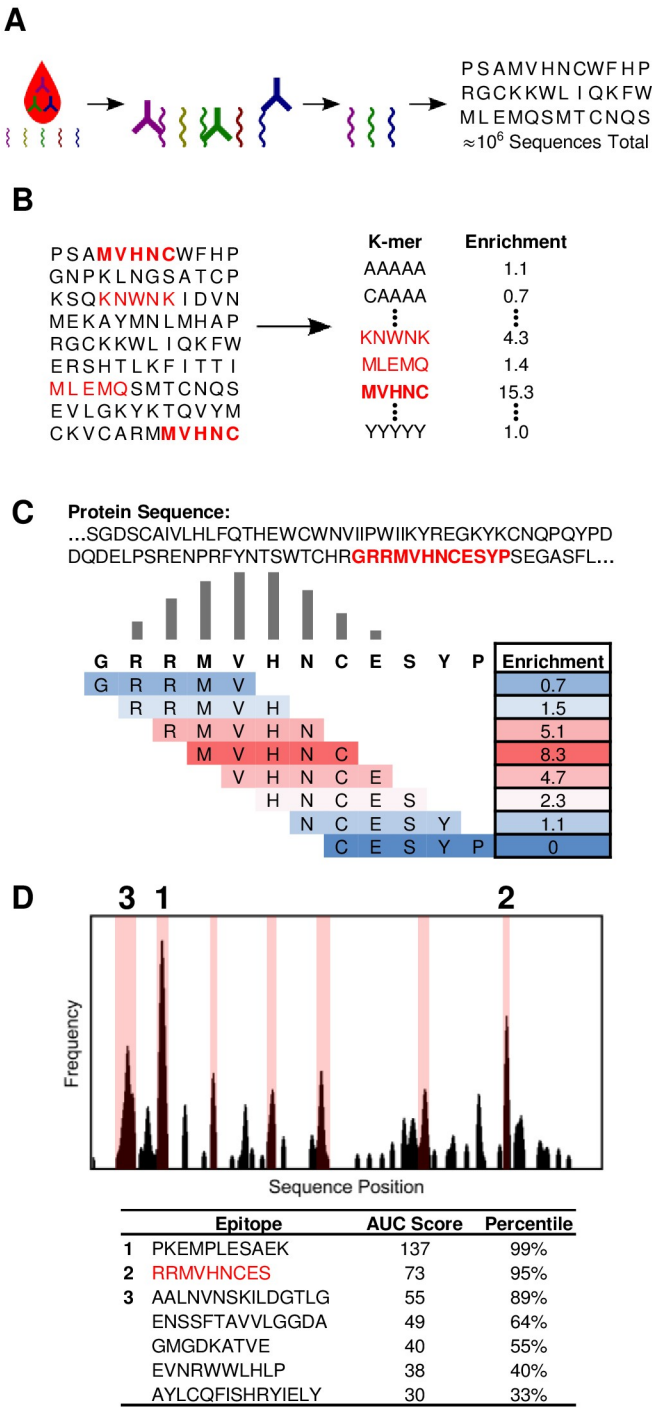
this problem researchers developed the IMUNE algorithm to reduce peptide datasets into statistically enriched patterns and cluster these patterns to build motifs [29].

A significant challenge for epitope mapping approaches is the association of epitopes and motifs with their corresponding antigens. Neither MEME nor IMUNE have the integrated capability to connect motifs to plausible antigens. Also, motifs identified through these methods often fail to reach the seven amino acids requirement for unambiguous identification of antigens within the full database of protein sequences [30]. Fundamentally, linear stretches in epitopes are typically less than seven amino acids in length [22], therefore, protein database searches of individual epitopes (such as through BLAST [31]) often fail to achieve statistical significance. Using multiple epitope matches within a single candidate antigen can increase the confidence of antigen prediction [26,32]. However, this method is insufficient for antigens with a single important epitope. Additionally, protein database searches are conducted using short amino acid sequences, therefore these searches do not fully leverage large quantitative binding datasets. To address these challenges, we present a general approach for associating epitopes with antigens using large peptide datasets. The K-mer Tiling of Protein Epitopes (K-TOPE) algorithm identifies epitopes by computationally tiling candidate antigens into k-mers, which are then evaluated within large datasets of antibody-binding peptides. Here, we demonstrate the utility of this approach by predicting linear epitopes within the proteomes of several prevalent infectious pathogens.

Results

To enable the prediction of protein epitopes bound by serum antibodies, we developed a method that uses a database of antibody-binding peptides to predict epitopes in known protein sequences (Fig 1). First, we selected peptides binding to an individual antibody repertoire within a specimen (serum or plasma) from a bacterial display peptide library with 10^{10} random 12-mer members. Then, we identified antibody-binding peptide sequences using NGS. To allow for the manipulation of 20^5 (3.2 million) k-mers rather than full-length peptides, we processed peptides into subsequences and evaluated the enrichments of all k-mers of length 5 [29]. We chose 5-mers because virtually all 5-mers were found in the peptide library at least once (S1 Text). Next, K-TOPE tiled candidate antigen sequences, such as from a proteome, into overlapping k-mers. K-TOPE used the enrichment values for these k-mers to construct an enrichment histogram across the length of each protein sequence. The frequency value at each sequence position in the histogram was proportional to the enrichment of k-mers that included that position. Specifically, for all k-mers overlapping a position, we summed the log base 2 of the k-mer enrichment. Thus, higher frequency values at a position in a protein sequence corresponded to a greater probability that a position was included in an epitope. All subsequences between two minima in the histogram with non-zero frequency values were considered "potential epitopes". These potential epitopes were scored based on the area under the curve (AUC). Next, potential epitopes were assigned an "epitope percentile" based on the rank of the epitope's AUC score in a list of AUC scores generated by analyzing random proteins. Finally, a threshold was set on the epitope percentile to determine whether an individual epitope was considered bound or simply noise. For this study an epitope percentile threshold of 95% was used, which corresponds to a p-value of 0.05. The prevalence of each epitope was calculated as the proportion of specimens that bound the epitope.

To assess the utility of K-TOPE, we first determined epitopes for monoclonal and polyclonal antibodies that bind specific, well-defined epitopes in cMyc, V5, and amyloid beta. We spiked these antibodies into serum at a final concentration of 25 nM and then selected and identified binding peptides. K-TOPE predicted epitopes that had greater than 60% overlap with the



	Epitope	AUC Score	Percentile
1	PKEMPLESAEK	137	99%
2	RRMVHNCES	73	95%
3	AALNVNSKILDGTLG	55	89%
	ENSSFTAVVLGGDA	49	64%
	GMGDKATVE	40	55%
	EVNRWWLHLP	38	40%
	AYLCQFISHRYIELY	30	33%

Fig 1. K-TOPE determines epitopes by tiling proteins into k-mers. (A) The input to the algorithm is a dataset of approximately 10^6 peptides that were bound by serum antibodies. (B) All 5-mers are evaluated for their enrichment in the list of peptides. (C) A portion of a protein sequence is tiled into 5-mers which are weighted by their enrichment. This determines a "frequency" value for each position in the sequence. (D) The frequency value for each position in a protein sequence is plotted as a histogram. Possible epitopes are highlighted in pink on the graph. Epitope sequences, area under the curve (AUC) scores, and significance percentiles are displayed.

https://doi.org/10.1371/journal.pone.0217668.g001

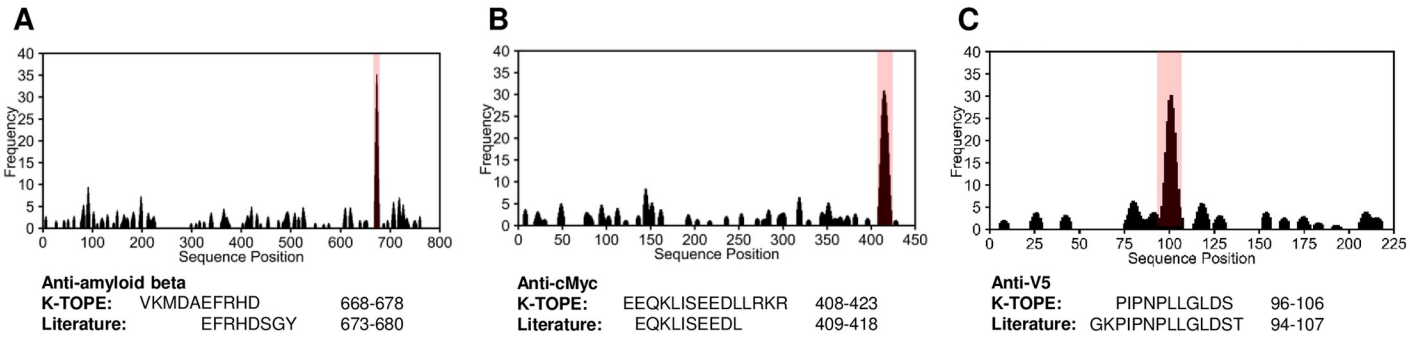


Fig 2. K-TOPE found epitopes for antibodies with known specificity spiked into serum. Histograms for antibodies with known specificity against amyloid beta (P05067), cMyc (P01106), and V5 (P11207) had prominent epitopes with epitope percentiles > 99.9% (in pink). (A) K-TOPE analysis of amyloid beta determined the epitope VKMDAEFRHD (668–678). This antibody was raised to whole protein and is known from literature to have a conformation-specific discontinuous epitope that maps to segments EFRHDSGY (673–680) and ED (692–693). (B) K-TOPE analysis of cMyc determined the epitope EEQKLISEEDLLRKR (408–422). This antibody was raised to AEEQKLISEEDLLRKRRE (407–424). (C) K-TOPE analysis of V5 determined the epitope PIPNPLLGLDS (96–106). The antibody was raised to GKPIPNPLLGLDST (94–107).

https://doi.org/10.1371/journal.pone.0217668.g002

previously reported epitopes of these antibodies (Fig 2). Importantly, the enrichment histograms generated by antibodies spiked into background serum or buffer were nearly identical (S1 Fig), suggesting that the noisy serum environment minimally affected epitope identification.

To predict "public epitopes" conserved across many individuals, epitopes were predicted for each specimen individually and then clustered. Although many private epitopes were predicted for each specimen in this process, we focused on the far smaller set of public epitopes to facilitate comparison with previous literature. Given the ubiquity of exposure to the upper respiratory pathogen *Rhinovirus A*, we validated the approach by predicting epitopes within its genome polyprotein. The true epitope percentile indicative of antibody binding and the true prevalence suggesting clinical relevance vary by antibody and the determination of the optimal values of these parameters would require additional experimental validation. However, for the purposes of this study, the epitope percentile threshold (S2 Fig) and prevalence (S3 Fig) were varied and an epitope percentile threshold of 95% and a prevalence of 30% were chosen. These values were chosen to ensure that the total number of epitopes predicted was of order one with the goal of decreasing the inclusion of false positives. Using a unique set of 250 serum specimens, we predicted four epitopes within *Rhinovirus A* that were targeted by 30% or more of the specimens (Fig 3A). Of the 250 specimens, 87% exhibited binding to at least one of these consensus epitopes (Fig 3B). Three of these epitopes were located within positions 570–620 (Fig 3C), in the antigenic attachment region of VP1. A fourth epitope within the VP2 region of the *Rhinovirus A* genome polyprotein was targeted by 43% of the population.

To assess trends in the population, each specimen was assigned into one of 16 groups based on which of the four *Rhinovirus A* epitopes were bound (Fig 3D). Notably, epitope binding was not independent, since 5 of the 16 groups of specimens were at least 50% larger than expected and the group targeting epitopes '1+3' was 60% smaller than expected (S1 Table). The average age of the subset of specimens of known age (n = 138) was 35 years. However, the specimens targeting 3 or more epitopes had an average age of 17, which was approximately 50% lower than the average age of 35 and the epitope group targeting none of the epitopes had an average age of 52, which was approximately 50% higher than average age of the population (S2 Table). Thus, people who targeted fewer *Rhinovirus A* epitopes tended to be older.

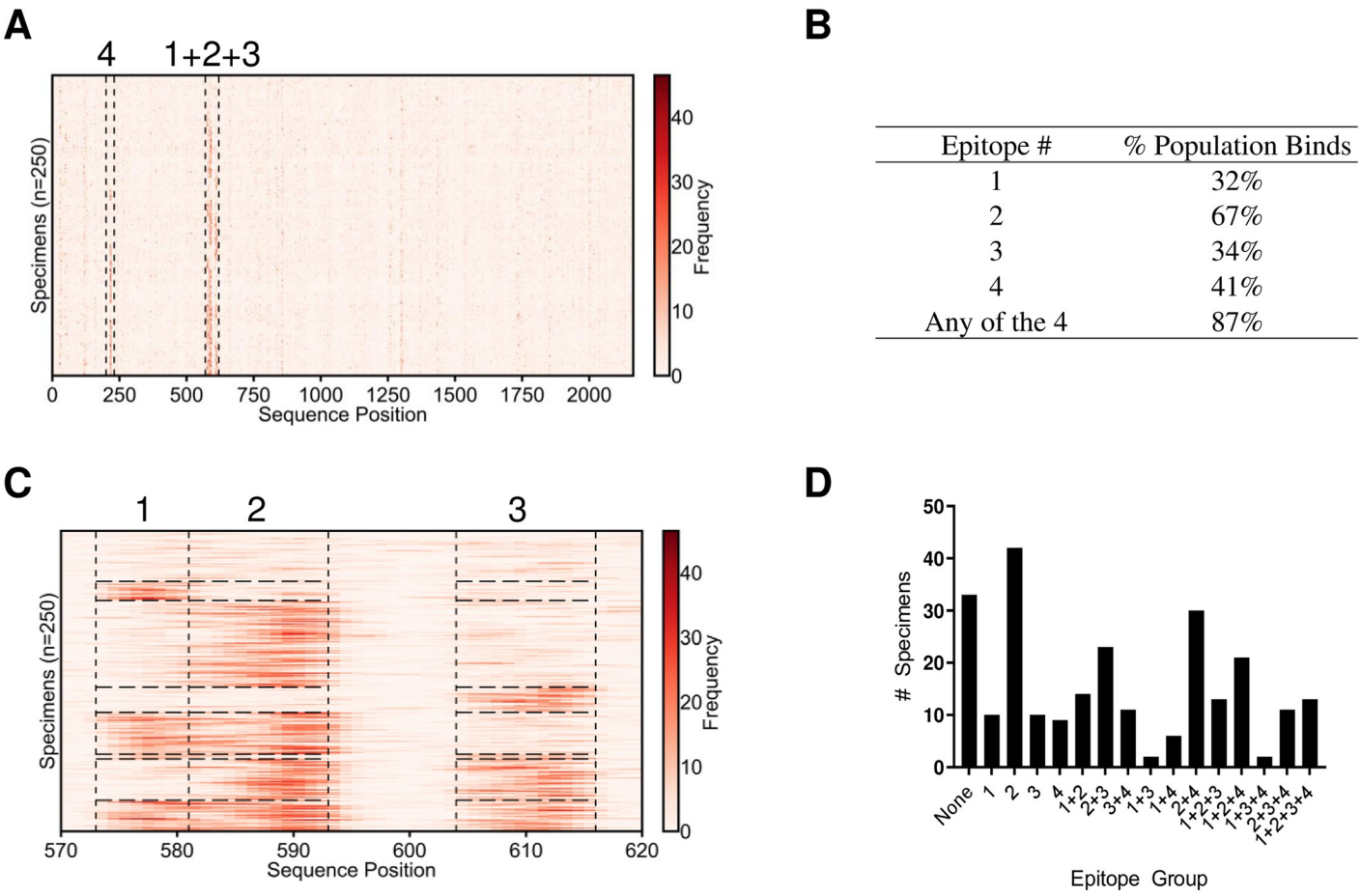


Fig 3. K-TOPE predicted four epitopes in the *Rhinovirus A* genome polyprotein. (A) K-TOPE was applied to the *Rhinovirus A* genome polyprotein (P07210) for 250 specimens. Histograms for all specimens are shown as rows in a heat map. The specimens have been clustered such that specimens that bind the same epitopes are adjacent. Regions that contain epitopes are outlined by dotted lines. (B) A table of the percentage of the population that bound each epitope. (C) The region from positions 570–620 is divided into 3 sections that correspond to distinct epitopes. These epitopes are consensus epitopes which were present in >30% of the 250 specimens. (D) Bar graph showing membership in different epitope groups. For example, a specimen that binds epitopes 2 and 3 will belong to epitope group "2+3". In this population, 87% of the specimens bound at least one of the consensus epitopes. The sequences of the epitopes were 1: QNPVENYI, 2: DSVLEVLVVPN, 3: APALDAAETGHT, and 4: NHTHPGEQG.

https://doi.org/10.1371/journal.pone.0217668.g003

To establish whether these *Rhinovirus A* epitopes were strain-specific, we predicted epitopes using KTOPE for 43 different *Enterovirus* strains (S3 Table). The epitopes predicted for these *Enterovirus* strains were similar to the 4 epitopes predicted in Fig 3, as illustrated by bands in the heat map showing the positions of each epitope (S4 Fig) Epitopes 1, 2, and 4 from Fig 3 were only found in *Rhinovirus*, whereas epitope 3 from Fig 3 was found in many *Enterovirus* strains. These results suggest that the epitopes predicted for *Rhinovirus A* may be relevant to multiple other *Enterovirus* strains.

Next, we investigated the utility of using K-TOPE to predict epitopes within a set of 2,908 proteins from 400 viral taxa with human tropism. This approach yielded 29 epitopes that were bound by at least 30% of all specimens (Table 1). Some of these epitopes have been reported previously [6,33–35]. Thus, a modest number of prominent linear viral epitopes were bound by >30% of the specimens analyzed. A common antigen identified from this analysis was Epstein-Barr nuclear antigen 1 (EBNA1) from Epstein-Barr virus (EBV), which is expressed in EBV-infected cells [36]. Additionally, the epitopes predicted for the enterovirus genus were

Table 1. A collection of 29 viral epitopes to which >30% of 250 specimens bound.

Epitope	Protein	Taxon	Accession	Prevalence
DSVLNEVLVVPN	Genome polyprotein	Enterovirus	P07210	0.668
PALTAAETG	Genome polyprotein	Enterovirus	Q66575	0.588
GRRPFFHPV	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	Q1HVF7	0.524
AGAGGGAGA	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	Q1HVF7	0.516
KYTHPGEA	Genome polyprotein	Enterovirus	Q82122	0.492
VRRPFFSD	Protein UL84	Human cytomegalovirus	P16727	0.452
NPVERYVDE	Genome polyprotein	Enterovirus	Q82122	0.428
MVVPEFK	DNA-binding protein	Human mastadenovirus C	P03265	0.428
EVKLPHWTPT	Glycoprotein 42	Epstein-Barr virus (strain GD1)	P03205	0.42
KPQPEKPK	Structural polyprotein	Mayaro virus	Q8QZ72	0.416
GGAGAGGAGAGGG	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	P03211	0.412
ININRPLE	Large structural protein	Lyssavirus	Q9QSP0	0.412
RPSCIGCKG	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	P03211	0.404
GAGAGAGGG	Packaging protein UL32	Simplexvirus	P89455	0.376
LEEVIVEKTK	Genome polyprotein	Enterovirus	Q82081	0.352
KHTHPGI	Replication origin-binding protein	Human herpesvirus 3	P09299	0.352
AETGHTNKI	Genome polyprotein	Enterovirus	Q82122	0.344
YVFPHWITK	Envelope glycoprotein gp63	Primate T-lymphotropic virus 3	Q0R5Q9	0.34
KTTNTTTNT	Immediate-early protein 2	Roseolovirus	Q9QJ16	0.34
MAADKPTL	Genome polyprotein	Murray Valley encephalitis virus	P05769	0.34
SFIVPEFA	Virion membrane protein A16	Orthopoxvirus	P16710	0.332
LVLPHWYMA	Cytoplasmic envelopment protein 1	Simplexvirus	P89430	0.328
YVDDMLNDI	Large tegument protein deneddylase	Human herpesvirus 6A (strain Uganda-1102)	P52340	0.328
SSGPKHTQKV	Genome polyprotein	Enterovirus	P03303	0.324
PVPEFQA	Non-structural polyprotein	Semliki forest virus	P08411	0.316
VPVTPNIAI	Genome polyprotein	Hepatitis C virus	Q68749	0.304
LHRPALTA	Minor capsid protein L2	Human papillomavirus type 34	P36758	0.304
EHILNRPTG	RNA-directed RNA polymerase L	Crimean-Congo hemorrhagic fever orthonairovirus	Q6TQR6	0.304
GEFIGSE	Shutoff alkaline exonuclease	Human herpesvirus 8	Q2HR95	0.3

K-TOPE was used to analyze 2,908 proteins from viruses with human tropism. This search demonstrated that only a few prominent linear viral epitopes were bound by a large proportion of the population.

https://doi.org/10.1371/journal.pone.0217668.t001

consistent with the epitopes predicted for *Rhinovirus* A, which is a species in that genus (Fig 3). Several of the epitopes were likely due to false discovery (e.g., Mayaro virus and Lyssavirus), since these viruses are uncommon in a general population. There is an intrinsic lower limit on false positives since antibodies only bind 5–6 amino acids, which is not enough information to uniquely specify a protein subsequence. This limitation is especially pronounced among evolutionarily related proteins in closely related species. To decrease the incidence of false positives, K-TOPE should only be used to analyze biologically relevant proteins. Ultimately, the epitopes predicted by K-TOPE require experimental validation to eliminate spurious results.

We performed a similar analysis for the proteomes of the genera *Streptococcus* and *Staphylococcus*, which are common bacterial human pathogens with 2,976 and 3,071 proteins in their respective proteomes. K-TOPE was used with each of these proteomes to determine epitopes bound by >30% of a population of 250 specimens, yielding 9 epitopes for *Streptococcus* and 13 epitopes for *Staphylococcus* (Table 2). The epitope LIPEFIG(R) in ATP-dependent Clp

Table 2. Epitopes in the proteomes of the genera *Staphylococcus* and *Streptococcus* which were bound by >30% of 250 specimens.

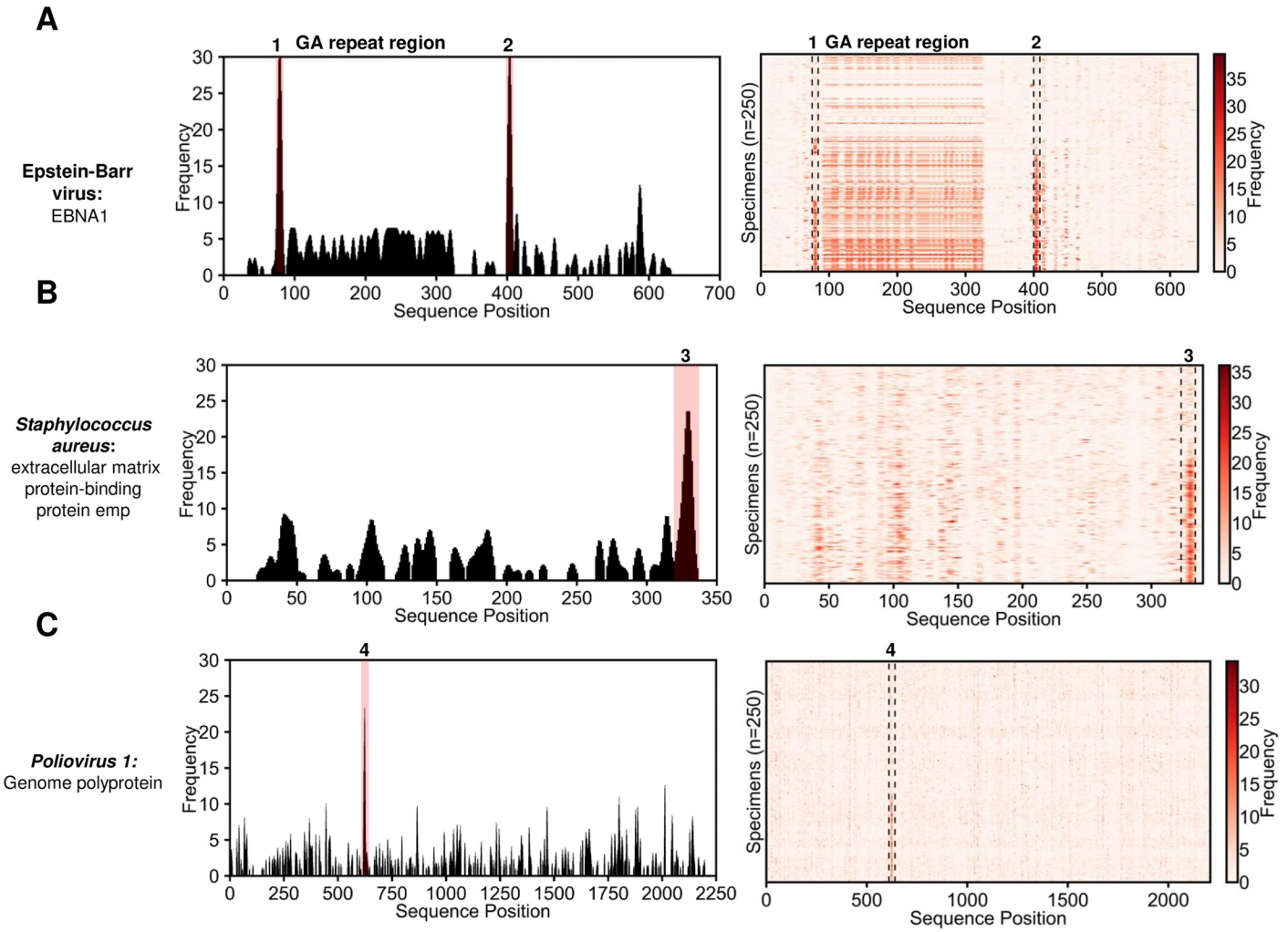
Epitope	Protein	Accession	Prevalence
Streptococcus			
LIPEFIGR	ATP-dependent Clp protease ATP-binding subunit ClpX	P63793	0.512
GQKMDDMLNS	Streptolysin O	Q5XE40	0.436
QIPALDKPL	FMN-dependent NADH-azoreductase	A4W2Z7	0.416
IADKPILD	UPF0154 protein SSU05_1707	A4VX34	0.392
TVADKPVA	Phenylalanine- -tRNA ligase beta subunit	Q5XCX3	0.360
RTPDKPT	Agglutinin receptor	P16952	0.324
VVPNIWR	Putative 2-dehydropantoate 2-reductase	P65666	0.320
LLNRPIHD	CCA-adding enzyme	Q5M153	0.320
TLADKPEF	Autolysin	P06653	0.308
Staphylococcus			
PTHYVPEFKGS	Extracellular matrix protein-binding protein emp	Q2FIK4	0.572
LIPEFIG	ATP-dependent Clp protease ATP-binding subunit ClpX	B9DNC0	0.508
NKPEFSGAT	3-isopropylmalate dehydratase small subunit	Q4L7U3	0.436
NKNNKNNKN	Translation initiation factor IF-2	Q4L5X1	0.372
KLGNIVPEYK	Extracellular matrix protein-binding protein emp	P0C6P1	0.360
KLCRICFRE	30S ribosomal protein S14 type Z	Q5HM12	0.352
DFLNRPVD	Proline- -tRNA ligase	Q4L5W5	0.348
EKNNNNNNNS	Alkaline shock protein 23	Q4L860	0.320
GVVPNISR	UvrABC system protein A	Q5HHQ9	0.312
LIPEFNQV	Homoserine kinase	Q8CSQ2	0.308
SPEFLGSQ	Undecaprenyl-diphosphatase	B9DK59	0.308
VGINRPTY	Putative glycosyltransferase TagX	O05154	0.308
VIPEFNND	Peptide chain release factor 2	Q4L4H9	0.300

K-TOPE was used to analyze 2,976 proteins from *Streptococcus* and 3,071 proteins from *Staphylococcus*.

https://doi.org/10.1371/journal.pone.0217668.t002

protease ATP-binding subunit ClpX was the most prevalent *Streptococcus* epitope and second most prevalent *Staphylococcus* epitope. Therefore, K-TOPE could not determine which genus generated this epitope. The most prevalent *Staphylococcus* epitope was PTHYVPEFKGS from extracellular matrix protein-binding protein emp, which is a known virulence factor [37]. For *Streptococcus*, the second most prevalent epitope was GQKMDDMLNS from the highly antigenic Streptolysin O protein [38]. This epitope falls within a 70 amino acid range in Streptolysin O that is known to bind antibodies [39]. The sequence "DKP" was present in 5/9 *Streptococcus* epitopes and the sequence "PEFXG" was present in 6/13 *Staphylococcus* epitopes (Table 2). Therefore, there are multiple candidate antigens that may correspond to these highly enriched sequences.

We searched IEDB ([40]) to determine which of the 51 viral and bacterial epitopes predicted by KTOPE were previously identified (S4 Table). Twelve of the 51 epitopes were similar to epitopes found in prior studies ([6,23,33,35,39,41–45]). However, 30 of the epitopes were in proteins with no reported epitopes, and 3 epitopes were in organisms with no reported epitopes. Only 6 of the epitopes were in well-characterized proteins but were not found in the literature, suggesting that these epitopes were false positives or novel epitopes. Additionally, only two bacterial epitopes were in previously described proteins, suggesting that the remainder of the bacterial proteins were false positives or novel antigens. Literature validation is shown in Fig 4 for the viral proteins EBNA1 from EBV and the *Poliovirus 1* genome polyprotein, as well as the bacterial protein Extracellular matrix protein-binding protein emp from *Staphylococcus*.



D

Peak #	Literature	K-TOPE	Prevalence	P-Value	Reference
1	KRPSCIGCK	RPSCIGCKG	0.40	0.0056*	[6]
2	RRPFF	PGRRPFFHP	0.52	0.0031	[33]
3	VPEFKGSLP	PTHYVPEFKGS	0.57	0.019	[45]
4	PALTAVETGATNPL	EIPALTAVETG	0.39	0.003	[35]

Fig 4. Epitopes predicted through proteome searches were validated using literature-reported epitopes. In (A), (B), and (C), histograms are shown for typical individual specimens (epitopes with percentiles > 99.7% are highlighted in pink). To the right of the histogram is a heat map for 250 specimens. For (A), there is a region of antigenic GA-repeats from positions 100–350. The table in (D) provides the statistical significance of agreement between literature epitopes and K-TOPE epitopes for the labeled peaks in (A), (B), and (C). The UniProt accessions used for this analysis were P03211 for EBNA1, Q8NXI8 for extracellular matrix protein-binding protein emp, and P03300 for *Poliovirus 1* Genome Polyprotein. Statistical tests where epitopes with >50% GA content were removed are denoted by an asterisk "*". All predicted epitopes had p-values below 0.05.

https://doi.org/10.1371/journal.pone.0217668.g004

In these cases, K-TOPE found prominent peaks in the histograms that corresponded to reported epitopes (Fig 4) [6,33,35,45]. Additionally, K-TOPE identified an immunogenic region of GA-repeats from positions 100–350 in the analysis of EBNA1 [23]. We used a non-parametric statistical test to assign significance to the overlap between K-TOPE epitopes and known epitopes. Using this method, all epitopes evaluated using K-TOPE had P-values below 0.05 (Fig 4C).

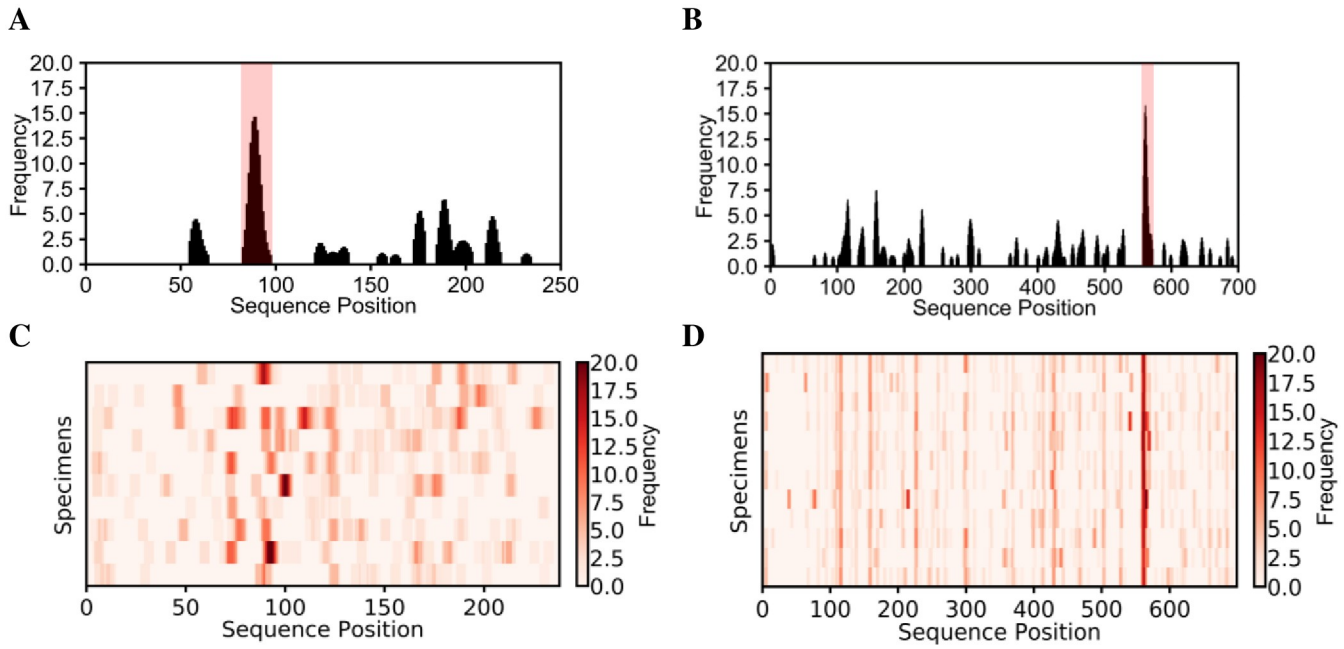


Fig 5. K-TOPE predicted epitopes for glycoprotein G1 using HSV1 specimens and for glycoprotein G2 using HSV2 specimens. For glycoprotein G1, a representative histogram for a single specimen is shown in (A) and a heat map for all HSV1 specimens is shown in (C). For glycoprotein G2, a representative histogram for a single specimen is shown in (B) and a heat map for all HSV2 specimens is shown in (D). There was a single epitope predicted for each protein. Epitopes with percentiles > 97% are highlighted in pink.

https://doi.org/10.1371/journal.pone.0217668.g005

To predict HSV species-specific epitopes, we analyzed 12 HSV2 specimens and 10 HSV1 specimens. Since these viruses share many of the same proteins in their proteomes [11], HSV1 specimens were appropriate controls for HSV2 specimens and vice-versa. To begin, we predicted species-specific epitopes in glycoprotein G, which is a protein that varies significantly between the two species (Fig 5) [46]. There was a single HSV1 epitope, PMPSIGLEE, bound by 40% of HSV1 specimens and a single HSV2 epitope, GGPEEFEGAGD, bound by all HSV2 specimens. This HSV2-specific epitope aligned well with previous epitopes found for glycoprotein G2 [13,47,48] (Table 3). Also, this epitope has been validated as an HSV2-specific diagnostic [49,50]. The HSV1-specific epitope was also similar to the previously reported epitope DHTPPMPSIGLE [18]. Interestingly, the two HSV-specific epitopes terminated in an identical 7-mer sequence EGAGDGE (PMPSIGLEEEEEEE**EGAGDGE** and GGPEEF**EGAGDGE**) [47]. This suggests that the regions containing these epitopes may be evolutionarily or structurally related targets of the immune system.

To predict candidate HSV species-specific epitopes, we analyzed the HSV1 and HSV2 proteomes. We predicted 30 HSV2-specific epitopes that were 100% specific with

Table 3. Alignment of an HSV2-specific glycoprotein G2 epitope with previously reported epitopes.

Peptides																												Reference
								G	G	**P**	**E**	**E**	**F**	**E**	**G**	**A**	**G**	**D**										K-TOPE
										P	**E**	**E**	**F**	**E**	**G**	**A**	**G**	**D**	G	E	P	P	E	D	D	D	S	G	[13]
	P	P	P	P	E	H	R	G	G	**P**	**E**	**E**	**F**	**E**	**G**	**A**	**G**	**D**	G	E	P	P	E						[47]
A	P	P	P	P	E	H	R	G	G	**P**	**E**	**E**	**F**	**E**	**G**	**A**	**G**	**D**	G										[48]

https://doi.org/10.1371/journal.pone.0217668.t003

Table 4. HSV2-specific epitopes were predicted.

Epitope	Protein	Accession	Prevalence
GGPEEFEGAGD	Envelope glycoprotein G	P13290	1
PLYARTTPAKF	Tegument protein UL47	P89467	1
VDSQRLTPGGSVS	Tegument protein UL21	P89444	1
KARKKGTSAL	Envelope glycoprotein B	P08666	1
TPLRYACVL	Tegument protein UL47	P89467	1
ANSPWAPVL	mRNA export factor	P28276	1
RYSPLHN	Envelope glycoprotein B	P08666	1
EAMLNDAR	Large tegument protein deneddylase	P89459	1
QRLTPH	Large tegument protein deneddylase	P89459	1
LRYTPAGEV	Envelope glycoprotein H	P89445	1
RTPSMR	Major viral transcription factor ICP4 homolog	P90493	1
LATNNA	Small capsomere-interacting protein	P89458	0.917
LRTNNL	Ribonucleoside-diphosphate reductase small subunit	P69521	0.917
PRTTPTPPQ	Envelope glycoprotein C	Q89730	0.833
HRLYAVVA	Inner tegument protein	P89460	0.833
PSTPAMLNLG	Ribonucleoside-diphosphate reductase large subunit	P89462	0.667
VTKHTALCAR	Large tegument protein deneddylase	P89459	0.583
TRDYAGL	Envelope glycoprotein I	P13291	0.583
RLTVAQ	Envelope glycoprotein I	P13291	0.583
RSLGIA	Protein UL20	P89443	0.583
IRDLARTFA	Thymidine kinase	P89446	0.5
DITAKHRCL	Major capsid protein	P89442	0.5
ETPAQPPRY	Capsid scaffolding protein	P89449	0.5
VSGITPTQ	Tripartite terminase subunit 1	P89451	0.5
HEELYYGPVS	Tegument protein VP22	P89468	0.417
IQDLAYAIV	Ribonucleoside-diphosphate reductase large subunit	P89462	0.417
GPAQRHTY	DNA polymerase catalytic subunit	P89453	0.417
YFEEYAYS	Envelope glycoprotein B	P08666	0.417
LDDFDL	Tegument protein VP16	P68336	0.417
AARLIDALYAEFLGG	Envelope glycoprotein H	P89445	0.333

A total of 30 epitopes were predicted that were 100% specific against HSV1.

https://doi.org/10.1371/journal.pone.0217668.t004

prevalence > 30% (Table 4). Notably, 11 of these epitopes were bound by all HSV2 specimens. K-TOPE predicted a glycoprotein C epitope PRTTPTPPQ with 83% prevalence which was contained in a previously identified epitope RNASA**PRTTPTPPQ**PRKATK [18]. In contrast to the numerous HSV2-specific epitopes, only 4 HSV1-specific epitopes were predicted, and the highest prevalence achieved was only 40% (Table 5). One of these epitopes, RIRLPHI, overlapped with the previously identified epitope HRRTRKAPK**RIRLPHI**R [51] in the well-described antigen glycoprotein D [17]. One possible explanation for the discovery of fewer HSV1-specific epitopes is that the HSV2 specimens had high IgM levels, whereas the HSV1 specimens had high IgG levels. Since high IgM levels occur with severe recurrent herpes infections [52], we would expect the high IgM HSV2 sera to yield more epitopes.

We sought to determine whether the HSV2-specific epitopes were contained in proteins that differed between the HSV species [46]. We determined 8 HSV2-specific epitopes with sequences that were contained in both HSV proteomes (S5 Table). Our analysis suggested that these epitopes were only targeted by HSV2 specimens, despite their presence in the HSV1

Table 5. HSV1-specific epitopes were predicted.

Epitope	Protein	Accession	Prevalence
RIRLPHI	Envelope glycoprotein D	Q69091	0.4
PMPSIGLEE	Envelope glycoprotein G	P06484	0.4
CAAFVNDYSLV	Major capsid protein	P06491	0.3
EMADTFLDT	ICP47 protein	P03170	0.3

Only 4 epitopes were predicted that were 100% specific against HSV2.

https://doi.org/10.1371/journal.pone.0217668.t005

proteome. Thus, even sequences that are conserved between species could serve as species-specific targets.

Since HSV1 is common in the general population, we were interested in identifying similarities in the epitopes predicted using the HSV1 specimens (Table 5) and the 250 specimens. Through this analysis we predicted 30 epitopes that were found in at least 10% of the 250 specimens (S6 Table). Notably, epitope 1 (FVLPHWYM) contains the 3-mer LPH which is also in the first epitope of the HSV1-specific epitopes, RIRLPHI (Table 5). Additionally, epitope 21 (PMPSLTA) contains the 4-mer PMPS which is also in epitope 2 of the HSV1-specific epitopes, PMPSIGLEE. A nearly exact match was found for epitope 18 (AAFVNDYS), which is highly similar to epitope 3 in the HSV1-specific epitope list, CAAFVNDYSLV. Thus, 3 of the 4 epitopes discovered from an HSV1-infected population had similarities to epitopes found using a general population. Additionally, 3 of the 4 antigens predicted from analyzing the 250 specimens were also predicted using the HSV1-infected specimens (major capsid protein, envelope glycoprotein G, and envelope glycoprotein D). Fewer epitopes were identified for the HSV1-infected specimens than the 250 specimens due to the group size disparity (10 specimens vs 250 specimens) and since the epitopes predicted from the HSV1-infected specimens used HSV2-infected specimens as controls. Although the prevalence of HSV1 is nearly 50% [9], we did not find epitopes with a prevalence this high, likely because there are a variety of HSV1 epitopes that collectively indicate a prevalence of 50%. Hence, individual epitopes may only have prevalence values of 10–30%.

Discussion

Here, we present a generalizable methodology for predicting epitopes within candidate immunogenic proteins. By tiling proteins into k-mers and evaluating those k-mers in a database of antibody-binding peptides, we determined epitopes for individuals and a population. Importantly, we have demonstrated that K-TOPE can predict disease-specific epitopes and antigens. One of the main features of this approach is that it combines k-mers to determine composite epitopes that may not explicitly exist in the peptide dataset. Another important element is using an antigen sequence to predict epitopes, thereby surmounting the 7 amino acid requirement for successful antigen identification [30].

The K-TOPE approach to epitope mapping differs from reported methods in several important ways. While proteome-derived peptide libraries have been used to predict disease-specific epitopes [33,53], these methods lack the flexibility to examine multiple proteomes. For instance, separate libraries would be required to analyze both HSV1 and HSV2. Even a library that contains peptides spanning all viral proteomes cannot easily be extended to much larger bacterial or parasitic proteomes [24]. A disadvantage of microarrays is that they have far lower 5-mer coverage (~27% [32]), than surface display (~100%) which could limit the application of k-mer approaches. Other algorithms have been developed that predict binding motifs in

peptide datasets, but they lack the integrated capability to connect motifs to protein antigens [54,55]. Also, the direct method of aligning peptides to sequences becomes computationally infeasible with a large number of peptides and candidate antigens [56].

The heterogeneity of experimental approaches complicates the validation of putative epitopes and their associated antigens. The Immune Epitope Database (IEDB) has an all-inclusive representation of information [57], which may not reflect important distinctions in experimental platforms, specimens, and data analysis techniques. For instance, there are likely numerous false positive epitopes for highly studied organisms and few identified epitopes for poorly studied organisms. Also, there is a lack of quantitative data reported for epitopes [58], such as the proportion of a given population that binds an epitope. To address this lack of information, we first used K-TOPE to analyze specimens for responses to common pathogens in a general population. This allows newly predicted "public epitopes" to be benchmarked by nearly any set of serum specimens. We required that a proportion of the population bind an epitope to reduce false positives. Although analysis of the variation in private epitopes could be valuable for understanding the variation in immune responses, it would complicate validation. We determined public epitopes in *Rhinovirus A* and showed that people who targeted fewer *Rhinovirus A* epitopes tended to be older, perhaps due to immunosenescence [59], reduced pathogen exposure, or a lower incidence of rhinovirus infections [60]. With a diverse group of specimens, it was possible to confirm that the RRPFF epitope in EBV's protein EBNA1 is a very commonly targeted epitope [33]. Since the specimens used to determine public epitopes were not assayed for responses to pathogens, acute and chronic infections could not be readily distinguished from prior infections. These public epitopes could be further validated using specimens with acute infections or using longitudinal studies to determine if these epitopes appear upon vaccination [61]. We did not find epitopes corresponding to measles or rubella vaccination, which is consistent with a recent study that comprehensively predicted viral epitopes [62]. This implies that for these viruses, high titer antibodies targeting linear epitopes may not be present. For HSV1 and HSV2, we determined whether an epitope was specific by analyzing specimens infected by both virus species. Unexpectedly, we demonstrated that even epitopes present in the conserved regions of both species' proteomes could be species-specific. The difference in binding was likely due to differences in the structure and post-translational modifications of the proteins. For the HSV analysis, we validated epitopes using previous studies, however, it was difficult to know *a priori* whether a non-validated epitope was novel or spurious. In general, since studies use different specimens, experiments, and computational analyses, it is unlikely for the epitopes of two studies to completely coincide.

K-TOPE provides a new tool for identifying diagnostic biomarkers, vaccine components, and candidate therapeutic targets. This approach could be used in the iterative process of designing a vaccine, since it would be useful to know which epitopes are elicited in a population by vaccination. Vaccine formulation could be altered to maximize the percentage of the population that targets epitopes associated with a positive disease outcome [2]. K-TOPE could also enable the development of diagnostics that assign disease based on the presence of epitopes. Since this method only involves a single experimental screen, in principle multiple diseases could be simultaneously diagnosed [63]. By searching for consensus epitopes in a disease group that are absent in a control group, K-TOPE can discover disease-specific epitopes. For an autoimmune disease, the entire human proteome could be analyzed to determine autoantigen epitopes [33]. Similarly, using clinical histories of viral infection, K-TOPE can analyze the proteomes of suspected pathogens to link epitopes to infections [24]. With specimens that have HLA information, it could be possible to detect a correlation between HLA type and bound epitopes [64]. This could have implications for how we determine genetic predisposition to immunological disease.

There are important limitations to the conditions in which this approach could be successful. First, this approach is currently limited to the prediction of linear epitopes. However, since 85% of epitopes have at least one linear stretch of five amino acids [22], conformational epitopes with linear segments may be represented in the datasets. We chose to focus on linear epitopes since methods that predict conformational epitopes often require 3D protein structures, which are scarce relative to the number of protein sequences. This report focuses on epitopes from common pathogens which are high-titer, but it could be difficult to detect rare antibody epitopes. Methods that selectively deplete out high-titer antibodies could prove effective for probing rare antibodies [65]. Another limitation is that protein sequences tend to have a large degree of conservation and redundancy [66], as demonstrated by the false positives found in the viral epitope search. Thus, even for analyses of non-immunogenic proteomes, false positives will occur due to evolutionary or coincidental sequence overlap with immunogenic proteomes. The issue of false positives can be partially allayed by deliberately choosing the set of investigated proteins, such that all proteins are plausible candidate antigens. Sequence conservation was demonstrated with the Enterovirus epitope PALTAVETGATNPL [35], as well as with the *Human herpesvirus 6A* epitope YVDDMLNDI (Table 1) which shares the k-mer "DDMLN" with the *Streptococcus* epitope GQKMDDMLNS (Table 2). Generally, if an epitope sequence is present identically in multiple antigens, all candidate antigens should be considered equally plausible without further biological, epidemiological, or experimental information. It is important to note that one of the purposes of K-TOPE is to reduce thousands of candidate proteins to a small set of proteins that can be experimentally validated.

In summary, the present approach enables the discovery of epitopes within the proteomes of any organism whose sequence is deposited into the protein database. The challenge of associating epitopes with antigens can be surmounted by transforming sets of antibody-binding peptides to k-mers and tiling proteins of interest. Advancements upon this paradigm may enable comprehensive immunological evaluations from serum and other biological tissues.

Materials and methods

Strains and reagents

E. coli strain MC1061 was used with surface display vector pB33eCPX for all library screening experiments. Protein A/G magnetic beads were from Thermo Scientific Pierce. Antibodies with known specificity included C3956 rabbit anti-c-Myc polyclonal antibody (Sigma), anti-beta amyloid 1–42 antibody [mOC31]—conformation-specific (ab201059) (Abcam), and rabbit V8137 Anti-V5 polyclonal antibody (Sigma). Antibodies were spiked into healthy donor serum at a concentration of 25 nM. All sera (n = 273) were obtained as deidentified specimens from biobanks according to institutional guidelines (Table 6), (Biosafety authorization numbers #201417, #201713), and handled according to CDC-recommended BSL2 guidelines.

Bacterial peptide display and sequencing

The bacterial peptide display screening protocol was carried out as previously described [29,67]. Briefly, an *E. coli* library displaying approximately 8 billion different 12-mer peptides was combined with 1:100 diluted serum. We used magnetic selection with Protein A/G beads to isolate bacterial cells with bound antibodies. Then, we confirmed that this isolated fraction of bacteria bound antibodies using flow cytometry. Amplicons were prepared from the isolated fraction for sequencing using the Illumina NextSeq.

Table 6. Serum sources.

Received From	# Specimens	Ages
UCLA	26	67 ± 10.5
UCSF	32	62 ± 7.2
Mayo Clinic	9	N/A
National Institute for Health and Welfare (Helsinki, Finland)	90	10 ± 5
NIH	40	N/A
Santa Barbara Cottage Hospital	12	N/A
Johns Hopkins University	15	30 ± 11.8
Max Delbrück Center for Molecular Medicine	26	N/A
BioreclamationIVT (HSV2)	12	N/A
Discovery Life Sciences (HSV1)	10	N/A
UCSB (mAb study)	1	N/A

Specimens that were provided without age information are noted by "N/A" under "Ages". Of the 90 specimens from the National Institute for Health and Welfare (Helsinki, Finland), 25 specimens were provided without age information. The average age is given with the standard deviation.

https://doi.org/10.1371/journal.pone.0217668.t006

Protein databases

Protein sequences were obtained from UniProt or by using the Biopython module [68]. Accessions for proteins are noted in figures and figure captions. For the epitope validation in Fig 4, accessions were chosen that reference the most highly annotated version of the proteins identified in Tables 1 and 2. The list of random proteins used for statistical analysis was obtained through a UniProt search of "reviewed:yes". The list of *Enterovirus* strains was obtained from a UniProt search of "enterovirus NOT organism:"homo sapiens" AND reviewed:yes". The viral proteome search used a Uniref search of "uniprot:(host:"homo sapiens" reviewed:yes fragment:no) AND identity:0.9" and yielded 2,908 proteins. The *Staphylococcus* proteome search used a Uniref search of "uniprot:(taxonomy:"Staphylococcus [1279]" fragment:no reviewed:yes) AND identity:0.9" and yielded 3,071 proteins. The *Streptococcus* proteome search used a Uniref search of "uniprot:(taxonomy:"Streptococcus [1301]" fragment:no reviewed:yes) AND identity:0.9" and yielded 2,976 proteins. HSV analysis used a UniProt search of "reviewed:yes AND organism:"Human herpesvirus 1 (strain 17) (HHV-1) (Human herpes simplex virus 1) [10299]" AND proteome:up000009294" for HSV1, yielding 73 proteins and a Uniprot search of "reviewed:yes AND organism:"Human herpesvirus 2 (strain HG52) (HHV-2) (Human herpes simplex virus 2) [10315]" AND proteome: up000001874" for HSV2, yielding 72 proteins.

Literature epitopes

For EBNA1, it was noted that RRPFF antibodies were found in the serum of healthy individuals [33]. KRPSCIGCK was noted as an EBNA1 epitope that was preferentially targeted by preeclamptic women, but was also targeted by healthy controls [6]. The motif XPEFXGSXX was discovered and inferred to correspond to VPEFKGSLP in *Staphylococcus aureus* using protein database searches [45]. For *Poliovirus 1*, the epitope PALTAVETGATNPL was found to be a cross-reactive epitope in many enteroviruses [35]. The remainder of the literature epitopes were obtained directly from IEDB.

Sequence processing

All software files are posted on GitHub (https://github.com/mlpaull/KTOPE) and all 278 antibody-binding peptide files are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.v7d0350). The imune-processor.jar file is available for research, non-profit, and non-commercial use, but requires a license for commercial use. All other software is available under the MIT license. The algorithms for generating nonredundant sequence lists from FASTQ files, outputting enrichment values for subsequences, and exhaustively calculating k-mer statistics were adapted from IMUNE (imune-processor.jar and calculate-patterns.jar) [29]. We added the capability to start with lists of peptides rather than NGS data. The enrichment of a k-mer is defined as the ratio of the number of observations of the k-mer to the "expected" number of observations. The "expected" value is calculated as the product of the total number of sequences, the number of frames the k-mer could fit in the sequences, and the probability of the k-mer appearing based on amino acid usage. If a k-mer's enrichment is above the "enrichment minimum", it is used in K-TOPE. The enrichment minimum was chosen as 2.0 for this study to reduce the dataset to only k-mers observed at least twice as often as expected. K-mers need to be calculated only once per specimen. All interaction with IMUNE-derived code is through a Python module which sets up a folder hierarchy and acts as a wrapper for IMUNE-derived code (imuneprocessor.py). These programs are memory and hard-drive intensive and it is recommended to have at least 16 GB of free RAM and 100 GB of hard-drive space. Analysis was carried out on a Dell Optiplex 9020 with an Intel® Core™ i7-4790 CPU @ 3.60 GHz, 64-bit operating system, and 32.0 GB of RAM. Processing FASTQ files into subsequences from 12 specimens, each containing approximately 1.5 million unique sequences, required 2.3 hours and calculating k-mer enrichment required 7.7 minutes. The duration of these calculations scales approximately linearly with the number of specimens and sequences.

K-TOPE algorithm

The K-TOPE algorithm (S1 Code) is written in Python 3.6 (KTOPE.py). A usage guide for KTOPE is available (S2 Text). First, there is a RAM-intensive step of loading k-mer enrichment data into memory as a dictionary. The enrichment dictionary for 250 specimens required approximately 4 GB of RAM. Then, a protein of interest is chosen for analysis and its sequence is loaded. This protein is tiled into k-mers of a set length. For this study, 5-mers were used. Each position in the protein sequence is assigned a frequency counter that starts at 0. The frequency counter of each sequence position contained in an enriched k-mer is incremented by the logarithm base 2 of the k-mer's enrichment. For instance, if 3 k-mers that overlapped at a position had enrichments of 2, 4, and 8, the frequency for that position would be $\log_2 2 + \log_2 4 + \log_2 8 = 6$. The frequency counters are compiled into a histogram which is smoothed using a moving window. For this analysis, to provide adequate smoothing, the window had width 7 and used linear weighting with 1 in the center and 0.1 at the edges. Minima and maxima are identified in the smoothed histogram. All intervals between 2 minima that contain a maximum are used to define epitopes. Epitopes were limited to a minimum length of 6 and a maximum length of 15 to roughly approximate the size of actual linear epitopes [22]. Epitopes are scored using the area under the curve of the un-smoothed histogram. To assign statistical significance to each epitope, the epitope's score is ranked in a list of scores for epitopes of the same length generated through an analysis of 10,000 random proteins. This rank is reported as a percentile in the distribution of random protein epitope scores. For this study, an epitope percentile threshold of 95% was used. For 12 specimens, analysis of 10,000 random proteins required 10.0 minutes.

After determining epitopes for individual specimens, K-TOPE can determine consensus epitopes for a population. Each epitope is characterized by a "centroid" which is the weighted central position of the epitope, indexed as a position in the protein sequence. Centroids for all epitopes that meet the epitope percentile threshold are compiled. They are then clustered using k-means to associate close centroids with the KMeans function from scitkit-learn [69]. A representative epitope is made for each cluster and kept if it meets a minimum prevalence in the population. Closely overlapping epitopes are removed and the final list is sorted by prevalence. Consensus epitopes can be determined for each protein in a proteome, generating a list of epitopes prevalent in a population. Determination of consensus epitopes for the *Rhinovirus A* genome polyprotein (P07210) for 250 specimens required 24.4 seconds. The proteome searches for viruses with human tropism, *Staphylococcus*, and *Streptococcus* for 250 specimens required 3.1, 2.3, and 1.9 hours, respectively.

We calculated expected membership of epitope groups by multiplying the proportions of the population that bound each epitope. For example, if epitope 1 was bound by 32% of the population and epitope 2 was bound by 67%, then the expected membership of epitope group '1+2' would be 21%. We ranked the overlaps between K-TOPE derived epitopes and literature epitopes in a list of 10,000 randomly generated epitope overlaps to determine a p-value. To remove redundant epitopes found in the proteome searches, we used the PAM30 similarity matrix to align two epitopes and compare each position to calculate a similarity score. Epitopes that had similarity scores >10, were in the same protein, and were from different organisms were considered redundant. We removed the less prevalent of the two redundant epitopes.

The HSV analysis used "disease" group specimens to predict epitopes and "control" group specimens to subtract epitopes. Epitopes were predicted in the disease group that met the epitope percentile threshold (95%) and the minimum prevalence (30%). Then, all disease epitopes were evaluated in the control group. For an epitope to be considered disease-specific, its score had to be below the epitope percentile threshold (80%) in all control specimens. To predict HSV2-specific epitopes that were also in the HSV1 proteome, we identified epitopes that exactly matched a subsequence in an HSV1 protein.

Data visualization

Fig 1 was created using Inkscape. Histograms and heat maps were generated using the Matplotlib python module [70]. Bar graphs were generated using GraphPad Prism 7.

Supporting information

S1 Fig. A comparison of histograms generated by K-TOPE when antibodies were added to serum or buffer. Histograms were generated for antibodies against cMyc (P01106), V5 (P11207), and amyloid beta (P05067). The most prominent peaks were present regardless of whether antibodies were added to serum or buffer. This suggests that the binding signature of a single antibody was not obscured by the many other antibody specificities present in serum.
(TIF)

S2 Fig. The number of epitopes generated as a function of varying the epitope percentile threshold. Epitopes were generated for 250 specimens using the *Rhinovirus A* genome polyprotein (P07210) with the prevalence fixed at 30%. The base 10 logarithm of the number of epitopes appeared to decrease linearly with increasing epitope percentile threshold. The value 95% was chosen for analysis because it corresponds to a p-value of 0.05 and ensures that the

total number of epitopes predicted was of order one. By predicting a total number of epitopes of order one, fewer false positives should to be included in this analysis.
(TIF)

S3 Fig. The number of epitopes generated as a function of varying the prevalence. Epitopes were generated for 250 specimens using the *Rhinovirus A* genome polyprotein (P07210) with the epitope percentile threshold fixed at 95%. The base 10 logarithm of the number of epitopes appeared to decrease exponentially with increasing prevalence. There were 123 epitopes bound by at least one member of the group. The value 30% was chosen arbitrarily from the prevalence values that predicted a total number of epitopes of order one. By predicting a total number of epitopes of order one, fewer false positives should to be included in this analysis.
(TIF)

S4 Fig. Heat map of epitopes predicted for 43 *Enterovirus* strains. Analyzing multiple strains of *Enterovirus* revealed that the epitopes found for the *Rhinovirus A* strain analyzed in Fig 3 were found in multiple enteroviruses. The 4 epitopes in Fig 3 were similar to epitopes in other *Enterovirus* strains, as demonstrated by the bands at approximately positions 212–221, 569–578, 577–590, and 602–613 (respectively corresponding to epitopes 1, 2, 3, and 4). Epitopes 1, 2, and 4 were only found in *Rhinovirus*, whereas epitope 3 was found in many *Enterovirus* strains. The heat map was restricted to positions 0–700 to show relevant epitopes. A binary decision was made for each position in each protein to determine whether it was in an epitope.
(TIF)

S1 Table. The expected and actual membership of different epitope groups. The expected membership of epitope groups was calculated by multiplying the proportions of the population that bound each epitope. For example, if epitope 1 was bound by 32% of the population and epitope 2 was bound by 67%, then the expected membership of epitope group '1+2' would be 21%. Note that specimens in groups *only* bound the epitopes in the groups e.g. specimens in group '1' did not bind '2' or '3'. Most of the actual and expected membership values agreed except for the '1+2+3', '3+4', '1+2+4', '1+2+3+4', and the 'None' groups which had higher membership than expected and the '1+3' group which had lower membership than expected. Additionally, the group targeting only epitope 4 was 40% smaller than expected suggesting that it was generally bound along with other epitopes. All groups that had percent differences equal to or greater than 50% are in bold.
(DOCX)

S2 Table. The average age for each epitope group. The average age for the 138 specimens for which there was age data was 35. The 'None' group had an average age of 52 which was approximately 50% higher than the average age of 35 (in bold). Additionally, specimens targeting 3 or more epitopes had an average age of 17 (in bold), which was approximately 50% lower than the average age of 35. This discrepancy suggests that older people targeted fewer *Rhinovirus A* epitopes. The average age is given with the standard deviation.
(DOCX)

S3 Table. Epitopes predicted for 43 *Enterovirus* strains. Analyzing multiple strains of *Enterovirus* revealed that the epitopes found for the *Rhinovirus A* strain analyzed in Fig 3 were found in multiple enteroviruses. Particularly, there were 31 strains with epitopes similar to epitope 3 in Fig 3 (APALDAAETGHT). Additionally, there were 3, 3, and 5 strains with epitopes similar to epitopes 1 (QNPVENYI), 2 (DSVLEVLVVPN), and 4 (NHTHPGEQG) from Fig 3, respectively. Epitopes 1, 2, and 4 were found in multiple rhinovirus strains suggesting that these

epitopes were *Rhinovirus*-specific, but not *Enterovirus*-specific. Similarity comparisons used the PAM30 similarity matrix with similarity defined as a similarity score > 10.
(DOCX)

S4 Table. Validating K-TOPE epitopes with prior studies. Twelve of the epitopes (in bold) were similar to epitopes found in prior studies with p-values of < 0.05. However, 30 of the epitopes were in proteins with no reported epitopes, and 3 epitopes were in organisms with no reported epitopes. Only 6 of the epitopes were in well-characterized proteins but were not found in the literature, suggesting that these epitopes were false positives or novel epitopes. Additionally, only two bacterial epitopes were in previously described proteins, suggesting that the remainder of the bacterial proteins were false positives or novel antigens. Epitopes 2, 3, and 4 differ slightly from those in Fig 4 because Fig 4 shows analysis for the most annotated accessions of these antigens, rather than the accessions used in K-TOPE analysis. Epitopes 4 and 11 were noted as part of the "GAGA" repeat region of EBNA1 and due to their frequency in the sequence, were not tested for significance. For epitopes 7, 17, and 31, the literature protein sequence did not match the protein sequence used for KTOPE searches. Instead, for these epitopes, the following accessions were used, respectively, P07210, A0A455KI32, and P0DF97 generating new epitopes DSVLNEVLVVPN, PALTAVETGHT, and KTDDMLNSND. Epitopes 1 and 7 matched the same literature epitope (NPVENYIDSVL-NEVLVVPNIQ) and epitopes 2 and 17 matched similar literature epitopes (PALTAVET-GATNPL and EAIPALTAVETGHTSQV). This suggests that the epitopes within each pair may be highly similar or identical. IEDB had 13 overlapping epitopes recorded for the protein streptolysin O, although for this analysis we chose the first of these epitopes (epitope 31). Surprisingly, epitope 20 in Murray Valley encephalitis had a corresponding literature epitope, but given the rarity of this virus, this was likely a coincidence. Each epitope was searched in IEDB by specifying the sequence with 70% BLAST similarity, the organism, the antigen name, positive assays only, and B Cell assays.
(DOCX)

S5 Table. Eight HSV2-specific epitopes were also in the HSV1 proteome.
(DOCX)

S6 Table. Analysis of the HSV1 proteome predicted 30 epitopes that were bound by at least 10% of the 250 specimens. Notably, epitope 1 (FVLPHWYM) contains the 3-mer LPH which is also in the first epitope of the HSV1 specific epitopes, RIRLPHI (Table 5). Additionally, epitope 21 (PMPSLTA) contains the 4-mer PMPS which is also in epitope 2 of the HSV1-specific epitopes, PMPSIGLEE. A nearly exact match was found for epitope 18 (AAFVNDYS), which is highly similar to epitope 3 in the HSV1-specific epitope list, CAAFVNDYSLV. Thus, 3 of the 4 epitopes discovered from an HSV1-infected population had similarities to epitopes found using a general population. Additionally, 3 of the 4 antigens predicted from analyzing the 250 specimens were also predicted using the HSV1-infected specimens (major capsid protein, envelope glycoprotein G, and envelope glycoprotein D). Fewer epitopes were identified for the HSV1-infected specimens than the 250 specimens due to the group size disparity (10 specimens vs 250 specimens) and since the epitopes predicted from the HSV1-infected specimens used HSV2-infected specimens as controls. It is likely that the epitopes predicted for the HSV1 specimens and the 250 specimens did not match exactly because epitopes predicted from people with active HSV1 infections may not be identical to those predicted from people with latent HSV1 infections. Although the prevalence of HSV1 is nearly 50%, we did not find epitopes with a prevalence this high, likely because there are a variety of HSV1 epitopes that collectively

indicate a prevalence of 50%. Hence, individual epitopes may only have prevalence values of 10–30%.
(DOCX)

S1 Code. KTOPE software, written in Python 3.6.
(TXT)

S1 Text. Justification of conducting analysis with 5-mers.
(DOCX)

S2 Text. KTOPE usage guide.
(DOCX)

Acknowledgments

The authors acknowledge the use of the Biological Nanostructures Laboratory within the California NanoSystems Institute, supported by the University of California, Santa Barbara and the University of California, Office of the President. We would like to acknowledge the work of Jack Reifert, Robert Pantazes, Chia-In Lin, Serra Elliott, and Kiho Song. We would like to acknowledge Linc Johnson for help with the initial conceptualization of this project.

Author Contributions

Conceptualization: Michael L. Paull, Patrick S. Daugherty.

Data curation: Michael L. Paull, Tim Johnston.

Formal analysis: Michael L. Paull.

Funding acquisition: Patrick S. Daugherty.

Investigation: Michael L. Paull, Kelly N. Ibsen, Joel D. Bozekowski.

Methodology: Michael L. Paull, Kelly N. Ibsen, Joel D. Bozekowski.

Project administration: Michael L. Paull, Patrick S. Daugherty.

Resources: Michael L. Paull, Kelly N. Ibsen, Joel D. Bozekowski.

Software: Michael L. Paull, Tim Johnston.

Supervision: Michael L. Paull, Patrick S. Daugherty.

Validation: Michael L. Paull, Patrick S. Daugherty.

Visualization: Michael L. Paull.

Writing – original draft: Michael L. Paull.

Writing – review & editing: Michael L. Paull, Patrick S. Daugherty.

References

1. Amanna IJ, Carlson NE, Slifka MK. Duration of humoral immunity to common viral and vaccine antigens. N Engl J Med. 2007; 357: 1903–1915. https://doi.org/10.1056/NEJMoa066092 PMID: 17989383
2. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. J Biomed Inform. Elsevier Inc.; 2015; 53: 405–414. https://doi.org/10.1016/j.jbi.2014.11.003 PMID: 25464113
3. Lemon SM, Gates NL, Simms TE, Bancroft WH. IgM antibody to hepatitis B core antigen as a diagnostic parameter of acute infection with hepatitis B virus. J Infect Dis. 1981; 143: 803–809. https://doi.org/10.1093/infdis/143.6.803 PMID: 6788859

4. Schellekens GA, Visser H, De Jong BA, Van Den Hoogen FH, Hazes JM, Breedveld FC, et al. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. Arthritis Rheum. 2000; 43: 155–163. https://doi.org/10.1002/1529-0131(200001)43:1<155::AID-ANR20>3.0.CO;2-3 PMID: 10643712
5. Spatola BN, Murray JA, Kagno M, Kaukinen K, Daugherty PS. Antibody repertoire profiling using bacterial display identifies reactivity signatures of celiac disease. Anal Chem. 2012; 85: 1215–1222. https://doi.org/10.1021/ac303201d PMID: 23234559
6. Elliott SE, Parchim NF, Kellems RE, Xia Y, Soffici AR, Daugherty PS. A pre-eclampsia-associated Epstein-Barr virus antibody cross-reacts with placental GPR50. Clin Immunol. 2016; 168: 64–71. https://doi.org/10.1016/j.clim.2016.05.002 PMID: 27181993
7. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature. 2014; 515: 577–581. https://doi.org/10.1038/nature13988 PMID: 25428507
8. Taylor TJ, Brockman MA, Mcnamee EE, Knipe DM. Herpes simplex virus. Front Biosci. 2002; 7: 752–764. https://doi.org/10.2741/taylor
9. Looker KJ, Magaret AS, May MT, Turner KME, Vickerman P, Gottlieb SL, et al. Global and regional estimates of prevalent and incident herpes simplex virus type 1 infections in 2012. PLoS One. 2015; 10: 1–17. https://doi.org/10.1371/journal.pone.0140765 PMID: 26510007
10. Looker KJ, Magaret AS, Turner KME, Vickerman P, Gottlieb SL, Newman LM. Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012. PLoS One. 2015; 10: 1–23. https://doi.org/10.1371/journal.pone.0114989 PMID: 25608026
11. Bowden RJ, Mcgeoch DJ. Evolution of herpes simplex viruses. Herpes simplex viruses. CRC Press; 2017.
12. Gupta R, Warren T, Wald A. Genital herpes. Lancet. 2007; 370: 2127–2137. https://doi.org/10.1016/S0140-6736(07)61908-4 PMID: 18156035
13. Marsden HS, MacAulay K, Murray J, Smith IW. Identification of an immunodominant sequential epitope in glycoprotein G of herpes simplex virus type 2 that is useful for serotype-specific diagnosis. J Med Virol. 1998; 56: 79–84. https://doi.org/10.1002/(SICI)1096-9071(199809)56:1<79::AID-JMV13>3.0.CO;2-R PMID: 9700637
14. Clo E, Kracun SK, Nudelman AS, Jensen KJ, Liljeqvist J-A, Olofsson S, et al. Characterization of the viral O-glycopeptidome: A novel tool of relevance for vaccine design and serodiagnosis. J Virol. 2012; 86: 6268–6278. https://doi.org/10.1128/JVI.00392-12 PMID: 22491453
15. Pan M, Wang X, Liao J, Yin D, Li S, Pan Y, et al. Prediction and identification of potential immunodominant epitopes in glycoproteins B, C, E, G, and i of herpes simplex virus type 2. Clin Dev Immunol. 2012; 2012. https://doi.org/10.1155/2012/205313 PMID: 22649465
16. Liu K, Jiang D, Zhang L, Yao Z, Chen Z, Yu S, et al. Identification of B- and T-cell epitopes from glycoprotein B of herpes simplex virus 2 and evaluation of their immunogenicity and protection efficacy. Vaccine. Elsevier Ltd; 2012; 30: 3034–3041. https://doi.org/10.1016/j.vaccine.2011.10.010 PMID: 22008818
17. Whitbeck JC, Huang Z-Y, Cairns TM, Gallagher JR, Lou H, Ponce-de-Leon M, et al. Repertoire of epitopes recognized by serum IgG from humans vaccinated with herpes simplex virus 2 glycoprotein D. J Virol. 2014; 88: 7786–7795. https://doi.org/10.1128/JVI.00544-14 PMID: 24789783
18. Risinger C, Sørensen KK, Jensen KJ, Olofsson S, Bergström T, Blixt O. Linear multiepitope (glyco)peptides for type-specific serology of herpes simplex virus (HSV) infections. ACS Infect Dis. 2017; 3: 360–367. https://doi.org/10.1021/acsinfecdis.7b00001 PMID: 28238255
19. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. Nature Publishing Group; 2014; 32: 158–68. https://doi.org/10.1038/nbt.2782 PMID: 24441474
20. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat Med. 2016; 22: 1456–1464. https://doi.org/10.1038/nm.4224 PMID: 27820605
21. Paull ML, Daugherty PS. Mapping serum antibody repertoires using peptide libraries. Curr Opin Chem Eng. Elsevier Ltd; 2018; 19: 21–26. https://doi.org/10.1016/j.coche.2017.12.001
22. Kringelum JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in antibody:protein complexes. Mol Immunol. Elsevier Ltd; 2013; 53: 24–34. https://doi.org/10.1016/j.molimm.2012.06.001 PMID: 22784991
23. Hecker M, Fitzner B, Wendt M, Lorenz P, Flechtner K, Steinbeck F, et al. High-density peptide microarray analysis of IgG autoantibody reactivities in serum and cerebrospinal fluid of multiple sclerosis patients. Mol Cell Proteomics. 2016; 15: 1360–80. https://doi.org/10.1074/mcp.M115.051664 PMID: 26831522

24. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Comprehensive serological profiling of human populations using a synthetic human virome. Science (80-). 2015; 348: aaa0698. https://doi.org/10.1126/science.aaa0698 PMID: 26045439

25. Legutki JB, Zhao Z-G, Greving M, Woodbury N, Johnston SA, Stafford P. Scalable high-density peptide arrays for comprehensive health monitoring. Nat Commun. Nature Publishing Group; 2014; 5: 4785. https://doi.org/10.1038/ncomms5785 PMID: 25183057

26. Liu X, Hu Q, Liu S, Tallo LJ, Sadzewicz L, Schettine CA, et al. Serum antibody repertoire profiling using in silico antigen screen. PLoS One. 2013; 8: e67181. https://doi.org/10.1371/journal.pone.0067181 PMID: 23826227

27. Bailey T, Elkan C. Unsupervised learning of multiple motifs using expected minimization. Mach Learn. 1995; 21: 51–80. https://doi.org/10.1007/BF00993379

28. Sykes KF, Legutki JB, Stafford P. Immunosignaturing: A critical review. Trends Biotechnol. 2013; 31: 45–51. https://doi.org/10.1016/j.tibtech.2012.10.012 PMID: 23219199

29. Pantazes RJ, Reifert J, Bozekowski J, Ibsen KN, Murray JA, Daugherty PS. Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. Sci Rep. 2016; 6: 30312. https://doi.org/10.1038/srep30312 PMID: 27481573

30. Bastas G, Sompuram SR, Pierce B, Vani K, Bogen SA. Bioinformatic requirements for protein database searching using predicted epitopes from disease-associated antibodies. Mol Cell Proteomics. 2007; 7: 247–256. https://doi.org/10.1074/mcp.M700107-MCP200 PMID: 17897933

31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

32. Richer J, Johnston SA, Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. Mol Cell Proteomics. 2015; 14: 136–147. https://doi.org/10.1074/mcp.M114.043513 PMID: 25368412

33. Larman HB, Laserson U, Querol L, Verhaeghen K, Solimini NL, Xu GJ, et al. PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. J Autoimmun. 2013; 43: 1–9. https://doi.org/10.1016/j.jaut.2013.01.013 PMID: 23497938

34. Sundström P, Nyström M, Ruuth K, Lundgren E. Antibodies to specific EBNA-1 domains and HLA DRB1*1501 interact as risk factors for multiple sclerosis. J Neuroimmunol. Elsevier B.V.; 2009; 215: 102–107. https://doi.org/10.1016/j.jneuroim.2009.08.004 PMID: 19733917

35. Samuelson A, Forsgren M, Johansson BO, Wahren B. Molecular basis for serological cross-reactivity between enteroviruses. Clin Diagn Lab Immunol. 1994; 1: 336–341. PMID: 7496972

36. Young LS, Rickinson AB. Epstein-Barr virus: 40 years on. Nat Rev Cancer. 2004; 4: 757–768. https://doi.org/10.1038/nrc1452 PMID: 15510157

37. Romero Pastrana F, Neef J, Koedijk DGAM, de Graaf D, Duipmans J, Jonkman MF, et al. Human antibody responses against non-covalently cell wall-bound Staphylococcus aureus proteins. Sci Rep. 2018; 8: 3234. https://doi.org/10.1038/s41598-018-21724-z PMID: 29459694

38. Kaplan EL, Huwe BB. The sensitivity and specificity of an agglutination test for antibodies to streptococcal extracellular antigens: A quantitative analysis and comparison of the streptozyme test with the anti-streptolysin O and anti-deoxyribonuclease B tests. J Pediatr. 1980; 96: 367–373. https://doi.org/10.1016/s0022-3476(80)80674-3 PMID: 6987354

39. Mortensen R, Nissen TN, Fredslund S, Rosenkrands I, Christensen JP, Andersen P, et al. Identifying protective Streptococcus pyogenes vaccine antigens recognized by both B and T cells in human adults and children. Sci Rep. Nature Publishing Group; 2016; 6: 22030. https://doi.org/10.1038/srep22030 PMID: 26911649

40. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015; 43: D405–D412. https://doi.org/10.1093/nar/gku938 PMID: 25300482

41. Niespodziana K, Napora K, Cabauatan C, Focke-Tejkl M, Keller W, Niederberger V, et al. Misdirected antibody responses against an N-terminal epitope on human rhinovirus VP1 as explanation for recurrent RV infections. FASEB J. 2012; 26: 1001–1008. https://doi.org/10.1096/fj.11-193557 PMID: 22121050

42. Sam Narean J, Glanville N, Nunn CM, Niespodziana K, Valenta R, Johnston SL, et al. Epitope mapping of antibodies induced with a conserved rhinovirus protein generating protective anti-rhinovirus immunity. Vaccine. Elsevier; 2019; 37: 2805–2813. https://doi.org/10.1016/j.vaccine.2019.04.018 PMID: 31003914

43. Härkönen T, Lankinen H, Davydova B, Hovi T, Roivainen M. Enterovirus infection can induce immune responses that cross-react with β-cell autoantigen tyrosine phosphatase IA-2/IAR. J Med Virol. 2002; 66: 340–350. https://doi.org/10.1002/jmv.2151

44. Roehrig JT, Hunt AR, Johnson AJ, Hawkes RA. Synthetic peptides derived from the deduced amino acid sequence of the E-glycoprotein of Murray Valley encephalitis virus elicit antiviral antibody. Virology. 1989; https://doi.org/10.1016/0042-6822(89)90509-6

45. Weber LK, Palermo A, Kügler J, Armant O, Isse A, Rentschler S, et al. Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays. J Immunol Methods. Elsevier B.V; 2017; 443: 45–54. https://doi.org/10.1016/j.jim.2017.01.012 PMID: 28167275

46. McGeoch DJ, Moss HWM, McNab D, Frame MC. DNA sequence and genetic content of the HindIII l region in the short unique component of the herpes simplex virus type 2 genome: Identification of the gene encoding glycoprotein G, and evolutionary comparisons. J Gen Virol. 1987; 68: 19–38. https://doi.org/10.1099/0022-1317-68-1-19 PMID: 3027242

47. Liljeqvist JÅ, Trybala E, Svennerholm B, Jeansson S, Sjögren-Jansson E, Bergström T. Localization of type-specific epitopes of herpes simplex virus type 2 glycoprotein G recognized by human and mouse antibodies. J Gen Virol. 1998; 79: 1215–1224. https://doi.org/10.1099/0022-1317-79-5-1215 PMID: 9603337

48. Grabowska A, Jameson C, Laing P, Jeansson S, Sjögren-Jansson E, Taylor J, et al. Identification of type-specific domains within glycoprotein G of herpes simplex virus type 2 (HSV-2) recognized by the majority of patients infected with HSV-2, but not by those infected with HSV-1. J Gen Virol. 1999; 80: 1789–1798. https://doi.org/10.1099/0022-1317-80-7-1789 PMID: 10423148

49. Oladepo DK, Klapper PE, Marsden HS. Peptide based enzyme-linked immunoassays for detection of anti-HSV-2 IgG in human sera. J Virol Methods. 2000; 87: 63–70. https://doi.org/10.1016/S0166-0934(00)00152-X PMID: 10856753

50. Nilsen A, Ulvestad E, Marsden H, Langeland N, Myrmel H, Matre R, et al. Performance characteristics of a glycoprotein G based oligopeptide (peptide 55) and two different methods using the complete glycoprotein as assays for detection of anti-HSV-2 antibodies in human sera. J Virol Methods. 2003; 107: 21–27. https://doi.org/10.1016/S0166-0934(02)00185-4 PMID: 12445934

51. Eisenberg RJ, Long D, Ponce de Leon M, Matthews JT, Spear PG, Gibson MG, et al. Localization of epitopes of herpes simplex virus type 1 glycoprotein D. J Virol. 1985; 53: 634–644. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2578577 PMID: 2578577

52. Kalimo KO, Marttila RJ, Granfors K, Viljanen MK. Solid-phase radioimmunoassay of human immunoglobulin M and immunoglobulin G antibodies against herpes simplex virus type 1 capsid, envelope, and excreted antigens. Infect Immun. 1977; 15: 883–9. PMID: 192678

53. Carmona SJ, Nielsen M, Schafer-Nielsen C, Mucci J, Altcheh J, Balouz V, et al. Towards high-throughput immunomics for infectious diseases: Use of next-generation peptide microarrays for rapid discovery and mapping of antigenic determinants. Mol Cell Proteomics. 2015; 14: 1871–1884. https://doi.org/10.1074/mcp.M114.045906 PMID: 25922409

54. Kim T, Tyndel MS, Huang H, Sidhu SS, Bader GD, Gfeller D, et al. MUSI: An integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. Nucleic Acids Res. 2012; 40: e47. https://doi.org/10.1093/nar/gkr1294 PMID: 22210894

55. Liu H, Han F, Zhou H, Yan X, Kosik KS. Fast motif discovery in short sequences. 2016 IEEE 32nd Int Conf Data Eng ICDE 2016. 2016; 2016: 1158–1169. https://doi.org/10.1109/ICDE.2016.7498321

56. Halperin RF, Stafford P, Emery JS, Navalkar K, Johnston SA. GuiTope: An application for mapping random-sequence peptides to protein sequences. BMC Bioinformatics. BioMed Central Ltd; 2012; 13: 1. https://doi.org/10.1186/1471-2105-13-1 PMID: 22214541

57. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, et al. The immune epitope database and analysis resource: From vision to blueprint. PLoS Biol. 2005; 3: 0379–0381. https://doi.org/10.1371/journal.pbio.0030091 PMID: 15760272

58. Caoili SEC. Benchmarking B-cell epitope prediction with quantitative dose-response data on antipeptide antibodies: Towards novel pharmaceutical product development. Biomed Res Int. 2014; 2014. https://doi.org/10.1155/2014/867905 PMID: 24949474

59. Alter-Wolf S, Blomberg BB, Riley RL. Deviation of the B cell pathway in senescent mice is associated with reduced surrogate light chain expression and altered immature B cell generation, phenotype, and light chain expression. J Immunol. 2009; 182: 138–47. https://doi.org/10.4049/jimmunol.182.1.138 PMID: 19109144

60. Heikkinen T, Järvinen A. The common cold. Lancet. 2003; 361: 51–59. https://doi.org/10.1016/S0140-6736(03)12162-9 PMID: 12517470

61. Stafford P, Wrapp D, Johnston SA. General assessment of humoral activity in healthy humans. Mol Cell Proteomics. 2016; 15: 1610–1621. https://doi.org/10.1074/mcp.M115.054601 PMID: 26902205

62. Burnham CAD, McAdam AJ. Your viral past: A comprehensive method for serological profiling to explore the human virome. Clin Chem. 2016; 62: 426–427. https://doi.org/10.1373/clinchem.2015.245027 PMID: 26769753

63. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. Proc Natl Acad Sci. 2014; 111: E3072–E3080. https://doi.org/10.1073/pnas.1409432111 PMID: 25024171

64. Palermo A, Weber LK, Rentschler S, Isse A, Sedlmayr M, Herbster K, et al. Identification of a tetanus toxin specific epitope in single amino acid resolution. Biotechnol J. 2017;1700197: 1700197. https://doi.org/10.1002/biot.201700197 PMID: 28922578

65. Bozekowski JD, Graham AJ, Daugherty PS. High-titer antibody depletion enhances discovery of diverse serum antibody specificities. J Immunol Methods. Elsevier; 2018; 455: 1–9. https://doi.org/10.1016/j.jim.2018.01.003 PMID: 29360471

66. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23: 1282–1288. https://doi.org/10.1093/bioinformatics/btm098 PMID: 17379688

67. Ibsen KN, Daugherty PS. Prediction of antibody structural epitopes via random peptide library screening and next generation sequencing. J Immunol Methods. Elsevier; 2017; 451: 28–36. https://doi.org/10.1016/j.jim.2017.08.004 PMID: 28827189

68. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

69. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2012; 12: 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

70. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007; 9: 99–104. https://doi.org/10.1109/MCSE.2007.55