

1 **Protein-based Immunome Wide Association**  
2 **Studies (PIWAS) for the discovery of**  
3 **significant disease-associated antigens**

4 Winston A. Haynes<sup>1</sup>, Kathy Kamath<sup>1</sup>, Patrick S. Daugherty<sup>1</sup>, John C. Shon<sup>1,\*</sup>

5 <sup>1</sup>Serimmune, Inc. Santa Barbara, CA, USA.

6 \*Corresponding author: [john.shon@serimmune.com](mailto:john.shon@serimmune.com)

## 7 **Abstract**

8 Identification of the antigens associated with antibodies is vital to understanding immune  
9 responses in the context of infection, autoimmunity, and cancer. Discovering antigens at a  
10 proteome scale could enable broader identification of antigens that are responsible for  
11 generating an immune response or driving a disease state. Although targeted tests for known  
12 antigens can be straightforward, discovering antigens at a proteome scale using protein and  
13 peptide arrays is time consuming and expensive. We leverage Serum Epitope Repertoire  
14 Analysis (SERA), an assay based on a random bacterial display peptide library coupled with NGS,  
15 to power the development of Protein-based Immunome Wide Association Study (PIWAS).  
16 PIWAS uses proteome-based signals to discover candidate antibody- antigen epitopes that are  
17 significantly elevated in a subset of cases compared to controls. After demonstrating statistical  
18 power relative to the magnitude and prevalence of effect in synthetic data, we apply PIWAS to  
19 systemic lupus erythematosus (SLE, n=31) and observe known autoantigens, Smith and  
20 Ribosomal P, within the 22 highest scoring candidate protein antigens across the entire human  
21 proteome. We validate the magnitude and location of the SLE specific signal against the Smith  
22 family of proteins using a cohort of patients who are positive by predicate anti-Sm tests.  
23 Collectively, these results suggest that PIWAS provides a powerful new tool to discover disease-  
24 associated serological antigens within any known proteome.

## 25 **Author Summary**

26 Infection, autoimmunity, and cancer frequently induce an antibody response in patients with  
27 disease. Identifying the protein antigens that are involved in the antibody response can aid in

28 the development of diagnostics, biomarkers, and therapeutics. To enable high-throughput  
29 antigen discovery, we present PIWAS, which leverages the SERA technology to identify antigens  
30 at a proteome- and cohort- scale. We demonstrate the ability of PIWAS to identify known  
31 autoantigens in SLE. PIWAS represents a major step forward in the ability to discover protein  
32 antigens at a proteome scale.

## 33 **Introduction**

34 Antibodies present in human specimens serve as the primary analyte and disease biomarker for  
35 a broad group of infectious (bacterial, viral, fungal, and parasitic) and autoimmune diseases. As  
36 such, hundreds of distinct antibody detecting immunoassays have been developed to diagnose  
37 human disease using blood derived specimens. The development of high-throughput  
38 sequencing technologies has enabled sequencing of numerous proteomes from diverse  
39 organisms. However, methods for antigen discovery within any given proteome remain  
40 relatively low throughput. The serological analysis of expression cDNA libraries (SEREX) method  
41 has been applied frequently to identify a variety of antigens, but high quality cDNA library  
42 construction remains technically challenging and time consuming [1–3]. Alternatively, entire  
43 human and pathogen derived proteomes can be segmented into overlapping peptides, and  
44 displayed on phage or solid-phase arrays and probed with serum [4–6]. Fully random peptide  
45 arrays of up to 300,000 unique sequences have also been used successfully to detect antibodies  
46 towards a range of organisms [7–9]. Even so, the limited molecular diversity of array based  
47 libraries can reduce antibody detection sensitivity and hinder successful mapping of peptide  
48 motifs to specific proteome antigens [7]. Thus, a general, scalable approach to identify  
49 serological antigens within arbitrary proteomes is needed.

50

51 In autoimmune diseases and cancers, autoantigen discovery is further complicated by the size  
52 of the proteome, heterogeneity of disease, and variability in immune response. Patient  
53 genetics, exposures, and microbiomes contribute to this heterogeneity, which in turn yields  
54 disparate responses to diverse antigens and epitopes [10,11]. In such cases, the mapping of

55 multiple epitopes to one antigen can increase confidence in a candidate antigen [7,12]. Even for  
56 diseases with conserved autoantigens, epitope spreading can lead to a diversified immune  
57 response against additional epitopes from the same protein or other proteins from the same  
58 tissue [13,14]. In cancer patients, neoepitopes can arise in response to somatic mutations that  
59 yield conformational changes or abnormal expression [15,16].

60

61 In complex autoimmune diseases like systemic lupus erythematosus (SLE), autoantibodies play  
62 an important role in diagnosis, patient stratification, and pathogenesis. SLE autoantigens  
63 include double-stranded DNA, ribonuclear proteins (Smith), C1q,  $\alpha$ -actinin,  $\alpha$ -enolase, annexin  
64 II, annexin AI, and ribosomal protein P [17–19]. In particular, anti-Smith antigen antibodies are  
65 present in 25-30% of SLE patients [20,21]. The Smith antigen consists of a complex of U-rich  
66 RNA U1, U2, U4/U6, and U5, along with core polypeptides B', B, D1, D2, D3, E, F, and G. Not all  
67 components of this complex are equally antigenic, and there are multiple epitopes within the  
68 complex [22,23].

69

70 One approach for antigen discovery, serum epitope repertoire analysis (SERA), uses bacterial  
71 display technology to present random 12mer peptides to serum antibodies [24–26]. Peptides  
72 that bind to serum antibodies are separated using magnetic beads and sequenced using next-  
73 generation sequencing. For each of these peptides and their kmer subsequences, enrichment  
74 can be calculated by comparing the actual number of observations to that expected based on  
75 amino acid frequencies [24]. Mapping these peptide epitopes to their corresponding protein  
76 antigens requires protein structure and/or sequence. Structure-based epitope mapping

77 methods (e.g., 3DEX, MIMOX, MIMOP, Pepitope) are not yet feasible at a proteome scale, due  
78 in part to the large number of undetermined structures [27–30]. However, since 85% of  
79 epitope-paratope interactions in crystal structures have a linear stretch of 5 amino acids,  
80 sequence information alone can be sufficient to identify many antigens [31–33]. The K-TOPE  
81 (Kmer-Tiling of Protein Epitopes) method has demonstrated the ability of tiled 5-mers to  
82 identify known epitopes in a variety of infections at proteome scale [34]. Here, we present a  
83 method, Protein-based Immunome Wide Association Study (PIWAS), which leverages the SERA  
84 assay to discover disease relevant antigens within large cohorts and at proteome scale. We  
85 evaluate PIWAS with synthetic data to examine the magnitude and prevalence of the effect  
86 needed for robust detection. We validate PIWAS using specimens from individuals with SLE and  
87 controls, identifying established anti-Smith and anti-Ribosomal P autoantibodies. We further  
88 validate the anti-Smith epitopes identified in our analysis using specimens positive for anti-  
89 Smith autoantibodies by predicate tests.

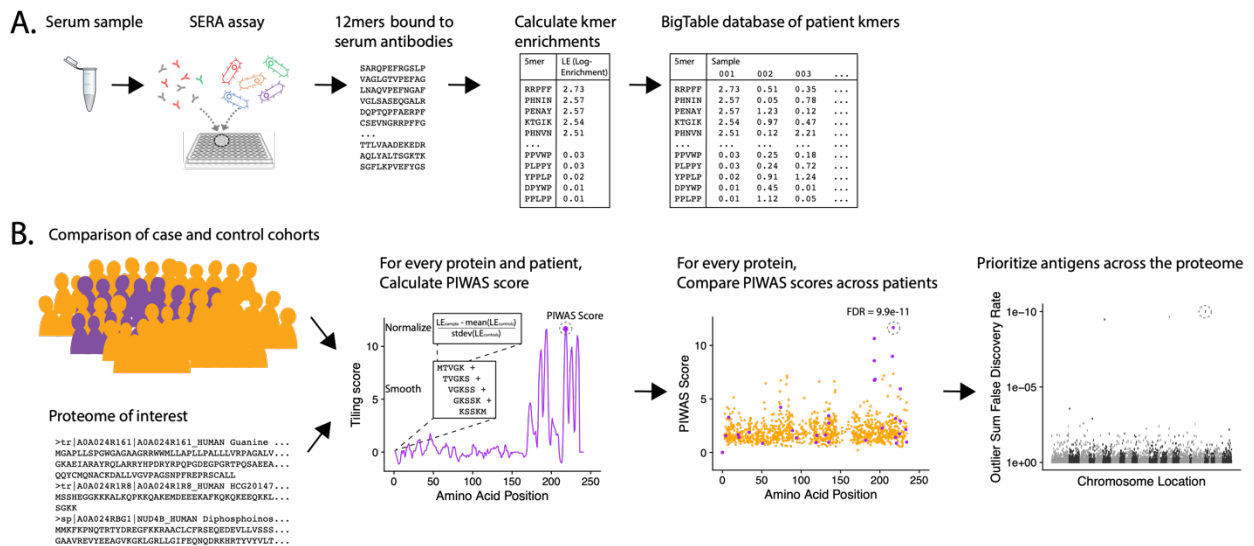
90

## 91 **Results**

### 92 **PIWAS allows identification of proteome-based signals**

93 To identify candidate serological antigens from arbitrary proteomes, we developed a robust,  
94 cohort-based statistical method to analyze peptide sequence data from the SERA assay. SERA  
95 uses a large bacterial display random peptide library of 10 billion member 12mers to identify  
96 binding to the epitopes recognized by antibodies species in a biospecimen (e.g. serum, plasma,  
97 cerebrospinal fluid) [Figure 1A]. From a typical specimen, we acquire 1-5 million unique 12mers.  
98 We break these 12mers into their constituent kmers, calculate log-enrichments (observed

99 divided by expected counts), and store the results in a BigTable database. To identify disease-  
 100 specific antigens from these data, PIWAS compares kmer data from case and control cohorts  
 101 against a proteome of interest (Figure 1B). For each protein and specimen dataset, we calculate  
 102 tiled kmer enrichments (normalized to the controls as a background) and smooth across a  
 103 sliding window. For each protein, we leverage statistics such as the outlier sum and Mann-  
 104 Whitney test to compare the case and control populations. At a proteome scale, we prioritize  
 105 candidate antigens based on these statistics (see Methods).



106  
 107 **Figure 1. PIWAS discovers candidate disease antigens through proteome-wide analysis.** (A)  
 108 Case and control specimens are processed using SERA to generate a dataset of 12mer amino  
 109 acid sequences bound by serum antibodies. Each 12mer is broken into kmer components and  
 110 log-enrichments of these kmers are calculated, where enrichment indicates the number of  
 111 observations compared to expectation based on amino acid frequency. (B) As input for the  
 112 PIWAS algorithm, case and control cohorts are identified (purple, cases; gold, controls) as well  
 113 as the target proteome. For each individual in the case and control cohorts and protein in the  
 114 proteome, PIWAS scores are calculated by tiling kmers onto the protein sequence, smoothing  
 115 over a window of these kmers, normalizing to the background signal in the controls, and  
 116 calculating the maximum value. PIWAS scores are compared across all case and control samples  
 117 to detect proteins whose scores are significantly greater in some subset of the case population  
 118 than in the control population. Antigens are then rank-ordered by one or more statistics across  
 119 the entire proteome.

120

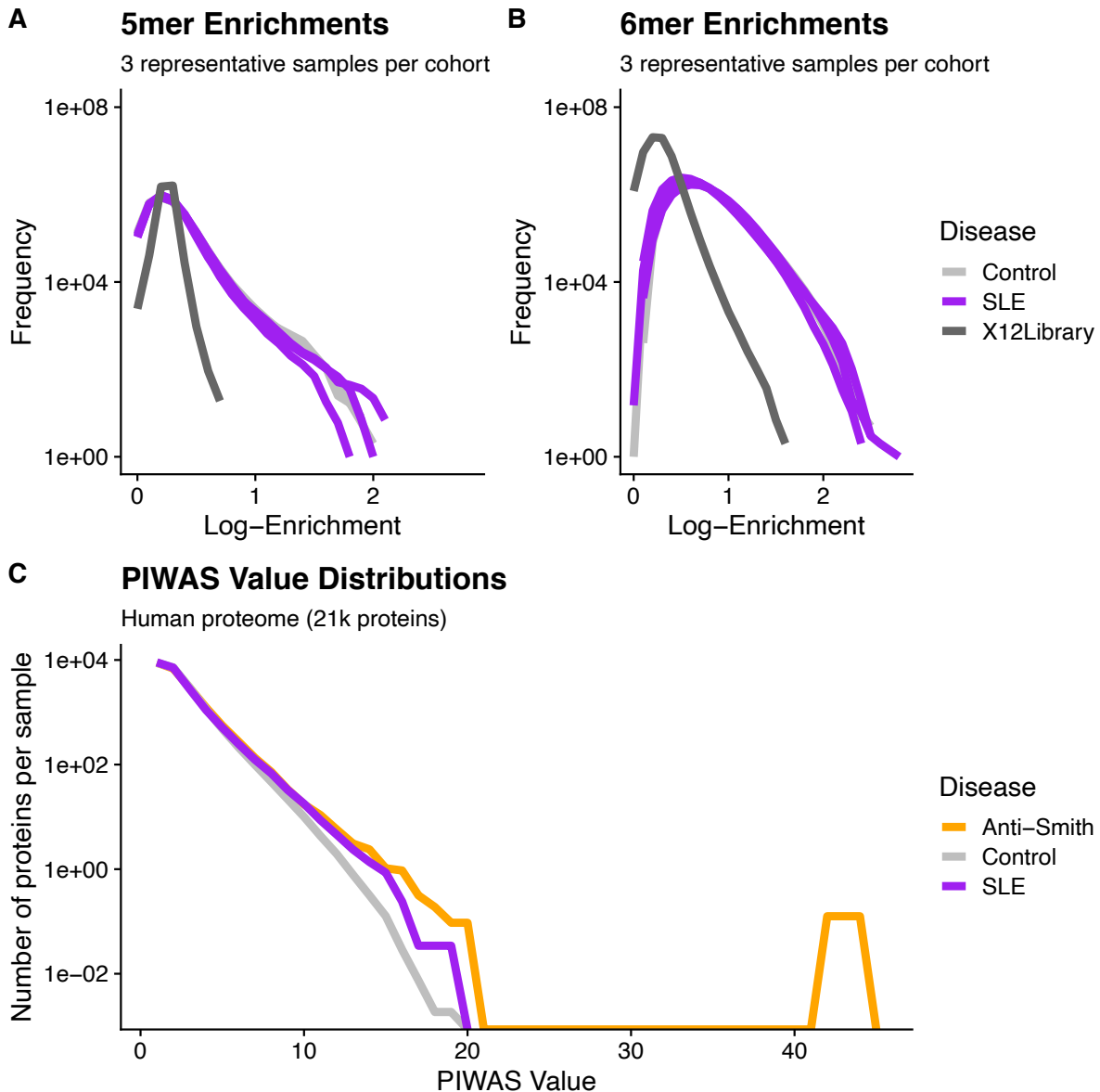
121 **Kmer enrichment in samples with serum compared to enrichment in a random library**

122 We first compared SERA library sequence composition before and after library selection with  
123 serum from healthy controls and SLE patients (Figure 2 A,B). Both the control and SLE serum  
124 yielded larger enrichments for both 5mers and 6mers compared to the unselected library. The  
125 enrichment of 5mers and 6mers in samples incubated with serum demonstrates the effects of  
126 antibody selection on the peptide library composition. We also compared the distribution of  
127 PIWAS values when 5mers were mapped to the human proteome. Interestingly, both SLE and  
128 anti-Smith cohorts yielded PIWAS value distributions with longer tails when analyzed against  
129 the entire human proteome when compared to those of healthy controls (Figures 2C). These  
130 findings confirm the general basis for using 5mers and 6mers for identifying both enriched  
131 signal in serum relative to a random library and enriched autoantigen signal using PIWAS in an  
132 example disease population relative to healthy controls.

133

134





135  
136

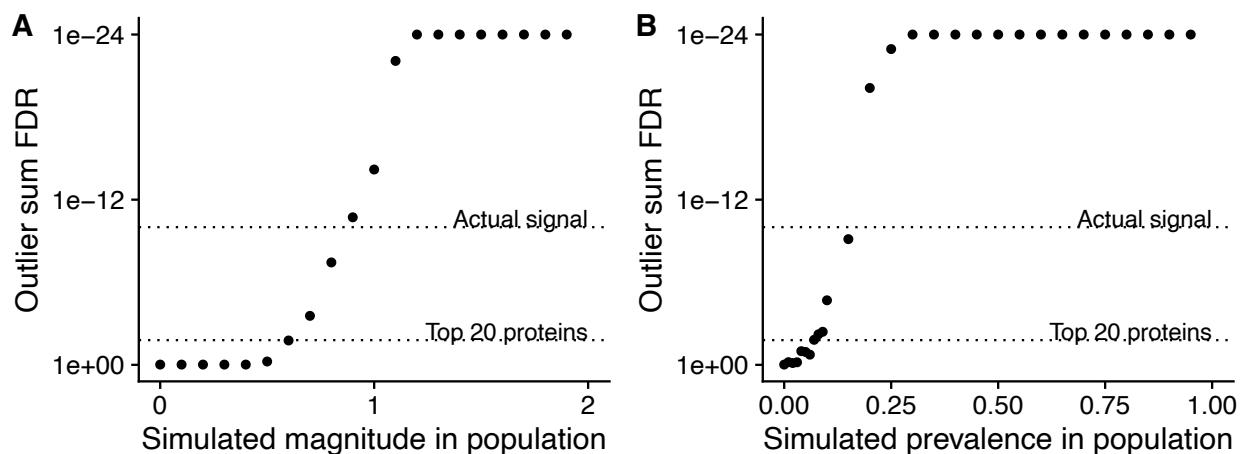
137 **Figure 2. Distributional differences in kmer enrichments and PIWAS values between the**  
138 **unselected library and after selecting with SLE and control specimens.** 5mer (A) and 6mer (B)  
139 Kmer frequency (y-axis) vs. Log-enrichment score (x-axis) for 6 subjects and the naïve library  
140 demonstrates species with large enrichments are found exclusively in those SERA assays  
141 incubated with serum. All 5mers or 6mers from three representative samples per cohort are  
142 evaluated for enrichment. Dark-gray lines = naïve 12-mer peptide library, purple lines = SLE  
143 cohort, gray lines = control cohort. (C) A comparison of PIWAS values (x-axis) vs. the number of  
144 proteins per sample with the corresponding PIWAS value (y-axis) reveals differences in both the  
145 range and distribution of PIWAS values between SLE and control samples. Distributions are  
146 based on 31 SLE cases and 1,157 controls. Purple = SLE cohort, gray = control cohort, orange =  
147 anti-Smith cohort.

148

## 149 PIWAS Power Simulations

150 In order to assess the statistical power of PIWAS to detect enriched antigens in a cohort, we  
151 performed computational experiments where we adjusted the magnitude and prevalence of  
152 known autoantigenic signal against Sm antigens (specifically small nuclear ribonucleoprotein-  
153 associated proteins B and B') in a cohort of SLE patients. Unsurprisingly, as the magnitude of  
154 the effect increases, so does the significance of the antigenic signal (Figure 3A). At an effect of  
155 only 60% of the SERA signal obtained with true SLE biospecimens, Sm antigens are significant at  
156 FDR=0.017 using the outlier sum FDR, still ranking within the top 20 proteins. Similarly, as the  
157 prevalence of the anti-Sm signal increases in the case population, so too does the significance  
158 of the outlier sum p-value (Figure 3B). At a prevalence of 7% (less than half of the actual  
159 biological prevalence in this cohort), anti-Sm is significant at FDR= 0.015 and remains within the  
160 top 20 scoring proteins. These results indicate an ability to detect signals well below the  
161 prevalence of many established autoantigens.

162  
163



164  
165 **Figure 3. Simulations of magnitude and prevalence of autoantigenic signal to assess statistical**  
166 **limits of detection for PIWAS.** SERA datasets from a cohort of SLE patients and kmer  
167 enrichments on small nuclear ribonucleoprotein-associated proteins B and B' were used as the  
168 actual biological signal (magnitude = 1 and prevalence = 19%). The magnitude (A) and

169 prevalence (B) of the kmer signal in this cohort was synthetically modulated to understand the  
170 statistical limits of detection for PIWAS.

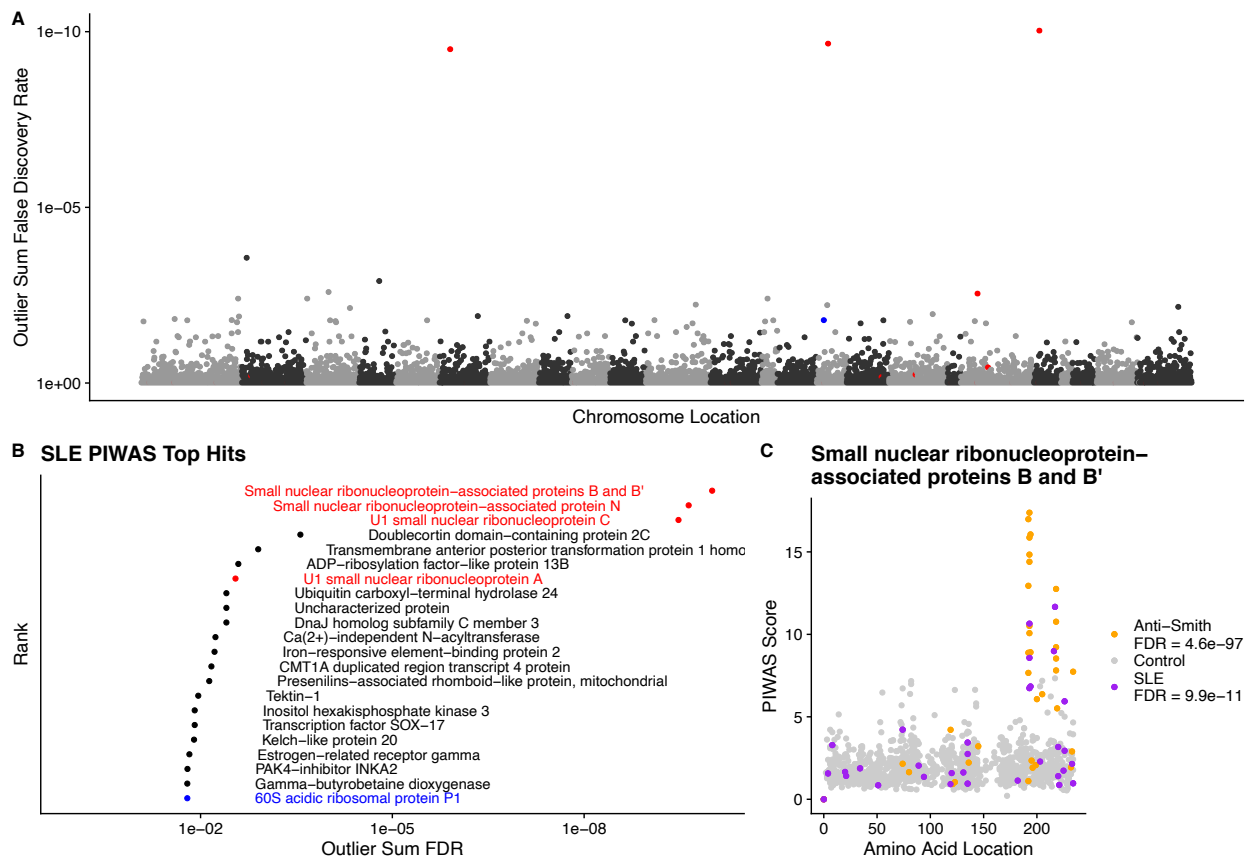
171

## 172 **PIWAS analysis of SERA datasets from SLE specimens**

173 We performed PIWAS to identify candidate autoantigens using specimens obtained from SLE  
174 patients. PIWAS results from individuals with SLE (n=31) were compared to those from controls  
175 (n=1,157) and proteins were ranked based on outlier sum FDR as a measure of significance  
176 across the human proteome (21,057 proteins) (Figure 4A-B). The highest scoring 22 proteins  
177 had outlier sum FDRs ranging from 1.6e-2 to 9.9e-11 and included multiple established  
178 autoantigens. Four Smith complex antigens were among the top seven hits with small nuclear  
179 ribonucleoprotein-associated proteins B and B' exhibiting the highest significance (outlier sum  
180 FDR = 9.9e-11). In addition, 60S acidic ribosomal protein P1, another known SLE autoantigen  
181 [20,35], was highly significant. Multiple highly significant epitopes were evident within nuclear  
182 ribonucleoprotein-associated proteins B and B' (Figure 4C, Table 1). The most significant  
183 enrichments occurred at two different locations near the C-terminus.

184

185



186

187

188

189

190

191

192

193

194

195

196

197

**Figure 4. Literature reported and putative autoantigens are detected in SLE samples by PIWAS.** (A) PIWAS results from a comparison of SLE samples to controls against the human proteome were prioritized using outlier sum FDR as a measure of significance (y-axis, see Methods). For visualization, proteins were laid out according to chromosome location. (B) Among the top set of 22 ranked proteins, 5 are established autoantigens (Smith family in red, others in blue). (C) Strength (y) and location (x) of PIWAS scores for the small nuclear ribonucleoprotein-associated proteins B and B' within SLE (n=31, purple) vs. control (n=1,157, grey). A cohort of anti-Sm predicate positive patients (n=35, orange) were compared to the same controls to validate the signal obtained using SLE specimens with unknown anti-Sm serostatus.

**Table 1. Dominant epitopes for highest scoring antigens from SLE PIWAS.**

Protein Name	Outlier Sum FDR	Dominant epitope(s)
Small nuclear ribonucleoprotein-associated proteins B and B'	9.9E-11	GGPSQQVMTPQ, PGMRPPMGPPM
Small nuclear ribonucleoprotein-associated protein N	2.3E-10	GGPSQQVMTPQ, PPGMRPPPPGI
U1 small nuclear ribonucleoprotein C	3.3E-10	GMRPPMGGHMP
Doublecortin domain-containing protein 2C	0.00027	IKPVVHCDINV, YWKSPPVPSEV
Transmembrane anterior posterior transformation protein 1 homolog	0.0013	LLQPAQVCDIL
ADP-ribosylation factor-like protein 13B	0.0026	IASVIINEGK

U1 small nuclear ribonucleoprotein A	0.0028	PPGMIPPPGLA, PGMIPPPGLAP
Ubiquitin carboxyl-terminal hydrolase 24	0.0039	None
Uncharacterized protein	0.0039	None
DnaJ homolog subfamily C member 3	0.0039	None
Ca(2+)-independent N-acyltransferase	0.0058	LIEGNCEHFVN
Iron-responsive element-binding protein 2	0.006	None
CMT1A duplicated region transcript 4 protein	0.0068	YVTYTSQTVKR, RLIEKSKTREL, SSKSSGKAVFR
Presenilins-associated rhomboid-like protein, mitochondrial	0.0073	GRRFNFFIQK
Tektin-1	0.011	KKLEQRLEEVQ, NSVSLEDWLDLDF
Inositol hexakisphosphate kinase 3	0.012	YDGPDPGYIFG
Transcription factor SOX-17	0.012	QSPSPPEALPC, MGLPYQGHDSG
Kelch-like protein 20	0.013	None
Estrogen-related receptor gamma	0.015	None
PAK4-inhibitor INKA2	0.016	MDCYLRLRKQE, LQDQMNCMMGA, TKFPSHRVCG
Gamma-butyrobetaine dioxygenase	0.016	TTGKLSFHTDY, DYCDFSVQSKH
60S acidic ribosomal protein P1	0.016	MGFGLFD

198

## 199 **PIWAS in an independent cohort of Smith antigen positive subjects**

200 To investigate the ability of PIWAS to identify Smith antigens in an independent cohort positive  
 201 for anti-Smith using validated clinical tests, we applied PIWAS to a cohort of 35 Smith antigen  
 202 positive samples. In this anti-Smith seropositive cohort, PIWAS again clearly identifies Smith  
 203 antigens at the top of the ranked list of antigens (Table 2). The dominant C-terminal, anti-Smith  
 204 epitope was identical between the two independent cohorts. The statistical significance within  
 205 the second cohort is greatly increased relative to the general SLE cohort as might be expected,  
 206 given the 100% seroprevalence of anti-Smith within this second specimen set. The unbiased  
 207 identification of known SLE autoantigens in independent cohorts validates the ability of PIWAS  
 208 to identify shared autoantigens in a data-driven way.

209

**Table 2. Dominant PIWAS epitopes for top antigens from anti-Smith seropositive specimens.**

Protein Name	Outlier Sum FDR	Dominant epitope(s)
Small nuclear ribonucleoprotein-associated protein N	1.1e-98	PGMRPPPPGIR

Small nuclear ribonucleoprotein-associated proteins B and B'	4.6e-97	PGMRPPMGPPM
U1 small nuclear ribonucleoprotein A	1.2e-67	PPGMIPPPGLA
U1 small nuclear ribonucleoprotein C	1.3e-47	PGMMPVGPAPG

210

## 211 **Discussion**

212 We demonstrate the utility of a general and scalable methodology to identify serological  
213 antigens within arbitrary proteomes using Protein-based Immunome Wide Association Studies  
214 (PIWAS). The power of PIWAS derives from cohort-based statistical analyses within large  
215 datasets of antibody-binding epitopes. PIWAS analyzes the enrichments of proteome spanning  
216 overlapping 5mers and 6mers that are observed amongst a peptide library selected for binding  
217 to antibody repertoires from cases and controls. We show that the kmer enrichment space  
218 demonstrates enriched signals compared to the unselected libraries. Further, the PIWAS space  
219 is enriched in SLE patients compared to control samples. Using synthetic data, we found that  
220 PIWAS has power to detect significant antigens at a signal of only 60% of the signal of a known  
221 autoantigen. When applied to experimental datasets from SLE cases and controls, PIWAS ranks  
222 SLE-specific Smith antigens highly in a proteome-wide search of candidate antigens. Finally, the  
223 epitopes from this antigen family were validated using a cohort of anti-Sm autoantibody  
224 positive patients.

225

226 Previous approaches to proteome-scale antigen identification rely on wet lab approaches that  
227 require a priori knowledge of the target proteome when the assay is performed [1–6]. In  
228 contrast, the use of random peptide library data with PIWAS enables analyses against arbitrary

229 proteomes. In addition to the reference human proteome utilized here, the same SERA data  
230 can be reanalyzed against proteomes of infectious agents, patient-specific mutations, and splice  
231 variants, without performing additional wet lab assays. Indeed, we have identified previously  
232 validated epitopes for multiple bacterial, viral, and fungal infectious diseases using this method  
233 [data not shown].

234

235 PIWAS is an immunological analog to widely employed genome-wide association studies  
236 (GWAS) that employ statistical association of gene variants in large disease and control cohorts  
237 to identify disease-associated loci. Like GWAS, PIWAS employs a data-driven statistical  
238 approach to scan entire genomes and proteomes for statistically significant differences  
239 between case and control cohorts. Advancements in GWAS methods such as burden testing has  
240 enabled multiple variants within a single gene to be collapsed, thereby increasing the power to  
241 detect disease-associated genes [36,37]. Similarly, PIWAS scans each protein to find a maximum  
242 signal and allows for the contributions of multiple distinct epitopes to identify candidate  
243 antigens associated with disease. By leveraging the outlier sum statistic [38], we are able to  
244 further highlight antigens with signals that are strong, but present in only a subset of the  
245 patient population, or derive from unique epitopes within the same antigen.

246

247 Just as GWAS must consider a variety of biological and technical limitations, effective PIWAS  
248 must consider and address pre-assay, assay, and post-assay factors that can impact  
249 performance. The most significant pre-assay issues relate to the selection of cohorts for disease  
250 and control populations. Our analyses using synthetic data demonstrated that magnitude and

251 prevalence of autoantigenic signal affects the ability of PIWAS to prioritize antigens. Thus, clean  
252 case and control cohorts are more likely to yield genuine autoantigens. In this study we were  
253 able to detect known antigens using a small cohort of SLE cases. As the cohort size grows, we  
254 anticipate even greater power to identify known and novel autoantigens.

255

256 Application of PIWAS to a cohort of SLE subjects identified known autoantigens, with 5 of 16 of  
257 the highest ranking hits across the entire human proteome being validated and clinically  
258 significant autoantigens. In particular, Smith antigens stood out as top hits in the SLE analysis.  
259 To validate this particular hit, we analyzed specimens from a second independent cohort of  
260 patients that tested positive for anti-Sm using clinical predicate tests. We found that the anti-  
261 Sm positive cohort exhibited reactivity against the same antigens and epitopes as the less  
262 homogeneous SLE discovery cohort. PIWAS identified an anti-Sm epitope occurring within a  
263 proline rich region in agreement with multiple prior studies [20,39].

264

265 Other highly ranked proteins identified using PIWAS could represent novel candidate antigens  
266 associated with SLE. PIWAS ranks antigens based on the maximum signal observed across a  
267 cohort, however it is not always possible to determine which antigens are biologically  
268 significant due to sequence similarity between proteins. Therefore antigens ranked highly in  
269 PIWAS should be considered candidate antigens, and orthogonal experimental validation is  
270 generally necessary to establish a *bona fide* antigen. If these candidate autoantigens are  
271 validated, they could be incorporated into multi-analyte autoantigen panels for diagnostic or  
272 prognostic purposes.



273

274 Although many known antibody epitopes contain a linear or contiguous segment, those with  
275 purely conformational epitopes or mimotopes may not be identified using PIWAS. PIWAS as  
276 presented, is limited to identifying linear epitopes at a proteome scale. Thus, we are developing  
277 PIWAS with degenerate positions that leverage motif patterns identified by IMUNE [24].  
278 Furthermore, the current method uses the maximum signal observed within the protein  
279 sequence for a particular patient, but some antigens have multiple antibody epitopes [40]. The  
280 use of multiple signals within a protein is another avenue of development to improve both  
281 sensitivity and specificity of PIWAS.

282

283 In conclusion, we developed PIWAS to enable robust, proteome-wide, cohort-based antigen  
284 discovery. PIWAS analyzes the datasets resulting from random peptide library selections against  
285 case and control cohorts (e.g., SERA) to discover shared candidate antigens, regardless of  
286 whether the epitopes therein are public or private. Since SERA employs random libraries,  
287 PIWAS can be applied to multiple proteomes utilizing the same physical assay. As the size of  
288 case and control datasets continue to increase, PIWAS may uncover previously undiscovered  
289 antigens with potential utility in diagnostic and therapeutic applications. Finally, PIWAS may be  
290 useful to investigate, in an unbiased manner, the association of autoantigens, human  
291 pathogens, and commensal organisms with human disease.

292

## 293 **Materials and Methods**

294 **Serum epitope repertoire analysis (SERA)**

295 Development and preparation of the *Escherichia coli* random 12-mer peptide display library

296 (diversity  $8 \times 10^9$ ) has been described previously [24]. SERA was performed as described [24].

297 Briefly, serum was diluted 1:25 and incubated for 1 hr with a 10-fold oversampling of the library

298 ( $8 \times 10^{10}$  cells/well) in a 96-well plate format at 4°C with orbital shaking (800 rpm) during which

299 time serum antibodies bind to peptides on the bacterial surface that mimic their cognate

300 antigens. Cells were then collected by centrifugation (3500 rcf x 7 min), the supernatant was

301 removed, and the cell pellets were washed by resuspending in 750  $\mu$ L PBS + 0.05% Tween-20

302 (PBST). The cells were again collected by centrifugation (3500 rcf x 7 min) and the supernatant

303 was removed. Cell pellets were resuspended in 750  $\mu$ L PBS and mixed thoroughly with 50  $\mu$ L

304 Protein A/G Sera-Mag SpeedBeads (GE Life Sciences, 17152104010350) (6.25 % the beads'

305 stock concentration). The plate was incubated for one hour at 4°C with orbital shaking (800

306 rpm). Bead-bound cells were captured in the plate using a Magnum FLX 96-ring magnet

307 (Alpaqua, A000400) until all beads were separated. Unbound cells in the supernatant were

308 removed by gentle pipetting, leaving only those cells bound to A/G beads. Beads were washed

309 5X by removing from the magnet, resuspending in 750  $\mu$ L PBST, and then returning to the

310 magnet. The supernatant was removed by gentle pipetting after the beads were securely

311 captured. Cells were resuspended in 750  $\mu$ L LB with 34  $\mu$ g/mL chloramphenicol and 0.2% wt/vol

312 glucose directly in the 96-deep-well plate and grown overnight with shaking (300 rpm) at 37°C.

313 **Amplicon library preparation for sequencing.** After growth, cells were collected by

314 centrifugation (3500 rcf for 10 min) and the supernatant was discarded. Plasmids encoding the

315 selected peptides were isolated in 96-well format using the Montage Plasmid Miniprep<sub>HTS</sub> Kit

316 (MilliPore, LSKP09604) on a Multiscreen<sub>HTS</sub> Vacuum Manifold (MilliPore, MSVMHTS00)  
317 following the “Plasmid DNA—Full Lysate” protocol in the product literature. For amplicon  
318 preparation, two rounds of PCR were employed; the first round amplifies the variable “X12”  
319 peptide region of the plasmid DNA. The second round barcodes each patient amplicon library  
320 with sample-specific indexing primers for data demultiplexing after sequencing. KAPA HiFi  
321 HotStart ReadyMix (KAPA Biosystems, KK2612) was used as the polymerase master mix for all  
322 PCR steps. Plasmids (2.5 µL/well) were used as template for a first round PCR with 12.5 µL of  
323 KAPA ReadyMix and 5 µL each of 1 uM forward and reverse primers. The primers (Integrated  
324 DNA Technologies) contain annealing regions that flank the X12 sequence (indicated in bold)  
325 and adapter regions specific to the Illumina index primers used in the second round PCR.  
326 Forward primer: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGVBHDV**CCAGTCTGGCCAGGG**  
327 Reverse primer: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**GTGATGCCGTAGTACTGG**  
328 A series of five degenerate bases in the forward primer, VBHDV (following IUPAC codes),  
329 provide base diversity for the first five reads of the sequencing on the NextSeq platform. The  
330 five base pairs were designed to be non-complementary to the template to avoid bias during  
331 primer annealing. To reduce non-specific products, a touchdown PCR protocol was used with an  
332 initial annealing temperature of 72°C with a decrease of 0.5°C per cycle for 14 cycles, followed  
333 by 10 cycles with annealing at 65°C. The 25 uL primary PCR product was purified using 30uL  
334 Mag-Bind TotalPure NGS Beads (Omega Bio-Tek, M1378-02) according to the manufacturer’s  
335 protocol. The second round PCR (8 cycles, 70°C annealing temperature) was performed using  
336 Nextera XT index primers (Illumina, FC-131-2001) which introduce 8 base pair indices on the 5’  
337 and 3’ termini of the amplicon for data demultiplexing of each sample screened. The PCR 1

338 product (5uL) was used as a template for the second PCR with 5uL each of forward and reverse  
339 indexing primers, 5uL PCR grade water and 25uL of KAPA ReadyMix. The PCR product (50uL)  
340 was cleaned up with 56 uL Omega Mag-Bind TotalPure NGS Beads per reaction. A 96-well  
341 quantitation was performed using the Qubit dsDNA High Sensitivity assay (Invitrogen, Q32851)  
342 adapted for a microplate fluorimeter (Tecan SPECTRAFlour Plus) measuring fluorescence  
343 excitation at 485 nm and emission at 535 nm. Positive (100 ng) and negative (0 ng) controls,  
344 included with the Qubit kit, were added to the plate as standards along with 2uL of each PCR  
345 product diluted 1:100 for quantitation. The fluorescence data were used to calculate DNA  
346 concentration in each well based on the kit standards. To normalize the DNA and achieve equal  
347 loading of each patient sample on NGS, the DNA in each well was diluted with Tris HCl (pH 8.5,  
348 10 mM) to 4 nM and an equal volume from each well was pooled in a Lo-Bind DNA tube for  
349 sequencing.

350 The sample pool was prepared for sequencing according to specifications of the Illumina  
351 NextSeq 500. Due to the low diversity in the adapter regions of the amplicon after the first five  
352 bases, PhiX Run Control (Illumina, FC-110-3001) was included at 40% of the final DNA pool. The  
353 pool was sequenced using a High Output v2, 75 cycle kit (Illumina, FC-404-2005).

354 **Naïve Library Sequencing.** An aliquot of the naïve X12 library representing 10-fold  
355 oversampling of the diversity was divided into 10 tubes, and the plasmids were purified and  
356 amplicons prepared as described above. Each prep was barcoded with a unique set of indices  
357 and sequenced on the NextSeq 500 to yield approximately 400 million unique sequences.

## 358 Cohorts

359 **Control cohort.** Specimens from 1,157 apparently healthy individuals were used as a control  
360 cohort.

361 **SLE cohort.** De-identified specimens from 31 individuals diagnosed with SLE, and primarily  
362 female (27), were acquired from Proteogenex (9) and BioIVT (22). The mean age within this  
363 cohort was 43 years, with a range of 22-72.

364 **Anti-Smith cohort.** Samples from 34 subjects that tested positive for Anti-SM RNP (4) or Anti-  
365 Smith (30) antibodies by predicate ANA multiplex testing were obtained from Discovery Life  
366 Sciences. Subjects ranged in age from 18 -74, with the majority (26) being female.

## 367 PIWAS Calculation

368 We define case ( $T$ ), and control ( $U$ ), cohorts of samples and begin with 12mer amino acid  
369 sequences for each sample generated by SERA (minimum of 1e6 total unique sequences per  
370 sample).

371 **Enrichment calculation.** We decompose each 12 mer from SERA into constituent  $k$ mers (where  
372  $k=5$  and  $k=6$  consecutive amino acids). For every  $kmer$  in each sample ( $S$ ), we calculate  
373 enrichment as:

$$374 \quad E_S(kmer) = n_S(kmer)/e_S(kmer)$$

375 where  $n(kmer)$  is the number of unique 12mers containing a particular  $kmer$  and  $e_S(kmer)$  is  
376 the expected number of  $kmer$  reads for the sample, defined as:

$$377 \quad e_S(kmer) = N_S(L_{seq} - k + 1) \prod_{i=1}^k p_i$$

378 where  $N_S$  is the number of 12mer reads generated for  $S$ ,  $L_{seq}$  is the length of the amino acid  
379 reads (12),  $k$  is the kmer length, and  $p_i$  is the amino acid proportion for the  $i$ th amino acid in  
380 *kmer* in all 12mers from  $S$ .

381 **Number of standard deviation normalization.** For every kmer, we normalize enrichment values  
382 to a control population. We define the control enrichment values as:

$$383 \quad C = \{E_v(kmer): w \in W\}$$

384 where  $W$  is the control cohort ( $U$ ).

385 The normalized enrichment is calculated as:

$$386 \quad F_S(kmer) = \frac{E_S(kmer) - \mu(C)}{\sigma(C)}$$

387 where  $\mu(C)$  is the mean of  $C$  and  $\sigma(C)$  is the standard deviation of  $C$ .

388 **PIWAS score calculation.** For each protein  $p$  and sample  $s$ , we calculate a PIWAS score  $P(s,p)$ ,  
389 defined as:

$$390 \quad P(s,p) = \max_{1 \leq i \leq len(p)} \sum_{k=5}^6 \sum_{j=i}^{\min(i+w, len(p)-k)} G_S(kmer(j,k,p))$$

391 where  $w$  is the width of the smoothing window,  $len(p)$  is the length of protein  $p$ ,  $kmer(j,k,p)$  is  
392 the kmer of length  $k$  at location  $j$  in protein  $p$ , and  $G_S$  is either  $E_S$  or  $F_S$ . Similarly, we record the  
393 location of this maximum statistics value,  $P_{loc}(s,p)$ , as:

$$394 \quad P_{loc}(s,p) = \operatorname{argmax}_{1 \leq i \leq len(p)} \sum_{k=5}^6 \sum_{j=i}^{\min(i+w, len(p)-k)} G_S(kmer(j,k,p))$$

395 **Cohort comparison statistics.** For each protein  $p$ , we define our case enrichments as:

$$396 \quad A(p) = \{P(t,p): t \in T\}$$

397 Similarly, we define our control enrichments as:

398 
$$B(p) = \{P(u, p): u \in U\}$$

399 We use several statistical tests to compare  $A(p)$  and  $B(p)$ , including traditional tests like the  
400 Mann-Whitney U and Kolmogorov-Smirnov. We calculate effect size as the Hedges'  $g$  statistic.  
401 We calculate the Outlier Sum, which we define as  $O(p)$ , statistic defined in Tibshirani and Hastie  
402 [38]. We perform 1,000 random permutations of the samples in  $A(p)$  and  $B(p)$  and calculate the  
403 Outlier Sum to calculate  $O^0(p)$ , the null distribution of the Outlier Sum for protein  $p$ . We  
404 calculate the z-score as:

405 
$$z_{O(p)} = \frac{O(p) - \mu_{O^0(p)}}{\sigma_{O^0(p)}}$$

406 Since the Outlier Sum is a sum of i.i.d. variables, we can apply the Central Limit Theorem and  
407 calculate a p-value for  $z_{O(p)}$  using the normal distribution.

408 We define the sets of case and control locations as:

409 
$$A_{loc}(p) = \{P_{loc}(t, p): t \in T\}$$

410 
$$B_{loc}(p) = \{P_{loc}(u, p): u \in U\}$$

411 We perform a Kolmogorov-Smirnov test comparing  $A_{loc}(p)$  and  $B_{loc}(p)$  to identify proteins  
412 with locational conservation of epitopes.

### 413 **Proteome description**

414 The reference *Homo sapiens* proteome was downloaded from Uniprot[41] on February 28,  
415 2019.

## 416 **Kmer Enrichment Analysis**

417 We compared the count of unique kmer species vs. enrichment scores for 5 and 6 mers in  
418 assays with a random library vs. those incubated with serum. We also compared the  
419 distribution of PIWAS values and average PIWAS values across control and SLE samples.

## 420 **Autoantigen Simulation Experiments**

421 To simulate the effects of changing the magnitude and prevalence of autoantigenic signal, the  
422 real PIWAS signal against one of the Smith antigens in the SLE cohort was selected for use in a  
423 series of simulations (P14678: Small nuclear ribonucleoprotein-associated proteins B and B').  
424 For every sample, the PIWAS values were calculated. To simulate different magnitudes of  
425 effect, the SLE PIWAS values were multiplied by scaling factors ranging from [0.1,2] and the  
426 outlier sum statistics were calculated relative to unscaled control values. To simulate different  
427 prevalences of effect, the SLE PIWAS values were divided into “high” (PIWAS > 6) and  
428 “low”(PIWAS < 6) values, 1000 random samplings with replacement of the SLE cohort were  
429 taken to simulate prevalences of “high” ranging from [0.01, 1], and the outlier sum statistics  
430 were calculated relative to unaffected control values.

## 431 **Data Availability**

432 PIWAS scores for the the human proteome in the SLE, anti-Smith, and control samples have  
433 been provided as a supplemental file.

## 434 **Author contributions**

435 Conceptualization: WH, KK, PD, JS; Data curation: WH; Formal analysis: WH; Funding  
436 acquisition: PD, JS; Software: WH; Visualization: WH; Writing- Review & Editing: WH, KK, PD, JS;  
437 Writing- Original draft preparation: WH, JS.



## 438 **Acknowledgements**

439 We thank the Serimmune team for supporting the development of PIWAS and processing of

440 samples: Becky Waitz, Jack Reifert, Joel Bozekowski, Burak Himmetoglu, Brian Martinez,

441 Gregory Jordan, Timothy Johnston, Cameron Gable, Steve Kujawa, Elisabeth Baum-Jones.

442 Special thanks to Elizabeth Stewart for editing this manuscript.

443

## 444 References

- 445 1. Zwick C, Pfreundschuh M. SEREX. *Encyclopedia of Cancer*. 2011. doi:10.1007/978-3-  
446 642-16483-5\_5252
- 447 2. Shi YY, Wang HC, Yin YH, Sun WS, Li Y, Zhang CQ, et al. Identification and analysis  
448 of tumour-associated antigens in hepatocellular carcinoma. *Br J Cancer*. 2005.  
449 doi:10.1038/sj.bjc.6602460
- 450 3. Zhou FL, Zhang WG, Chen G, Zhao WH, Cao XM, Chen YX, et al. Serological  
451 identification and bioinformatics analysis of immunogenic antigens in multiple myeloma.  
452 *Cancer Immunol Immunother*. 2006. doi:10.1007/s00262-005-0074-x
- 453 4. H.B. L, Z. Z, U. L, M.Z. L, A. C, M.A.M. G, et al. Autoantigen discovery with a synthetic  
454 human peptidome. *Nature Biotechnology*. 2011.
- 455 5. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Comprehensive serological  
456 profiling of human populations using a synthetic human virome. *Science* (80- ). 2015;348:  
457 aaa0698. doi:10.1126/science.aaa0698
- 458 6. Zandian A, Forsstro rn, Ha A, Schwenk JM, Uhle M, Nilsson P, et al. Whole-Proteome  
459 Peptide Microarrays for Profiling Autoantibody Repertoires within Multiple Sclerosis and  
460 Narcolepsy. 2017 [cited 27 Feb 2020]. doi:10.1021/acs.jproteome.6b00916
- 461 7. Richer J, Johnston SA, Stafford P. Epitope identification from fixed-complexity random-  
462 sequence peptide microarrays. *Mol Cell Proteomics*. 2015.  
463 doi:10.1074/mcp.M114.043513
- 464 8. Hansen LB, Buus S, Schafer-Nielsen C. Identification and Mapping of Linear Antibody  
465 Epitopes in Human Serum Albumin Using High-Density Peptide Arrays. *PLoS One*.  
466 2013;8: e68902. doi:10.1371/journal.pone.0068902
- 467 9. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-  
468 resolution mapping of linear antibody epitopes using ultrahigh-density peptide  
469 microarrays. *Mol Cell Proteomics*. 2012;11: 1790–1800. doi:10.1074/mcp.M112.020800
- 470 10. Finn OJ. Human tumor antigens yesterday, today, and tomorrow. *Cancer Immunol Res*.  
471 2017;5: 347–354. doi:10.1158/2326-6066.CIR-17-0112
- 472 11. Zhao Z, Ren J, Dai C, Kannapell CC, Wang H, Gaskin F, et al. Nature of T cell epitopes  
473 in lupus antigens and HLA-DR determines autoantibody initiation and diversification.  
474 *Ann Rheum Dis*. 2018. doi:10.1136/annrheumdis-2018-214125
- 475 12. Liu X, Hu Q, Liu S, Tallo LJ, Sadzewicz L, Schettine CA, et al. Serum Antibody  
476 Repertoire Profiling Using In Silico Antigen Screen. *PLoS One*. 2013.  
477 doi:10.1371/journal.pone.0067181
- 478 13. Chan LS, Vanderlugt CJ, Hashimoto T, Nishikawa T, Zone JJ, Black MM, et al. Epitope  
479 spreading: Lessons from autoimmune skin diseases. *Journal of Investigative Dermatology*.  
480 1998. doi:10.1046/j.1523-1747.1998.00107.x
- 481 14. Didona D, Di Zenzo G. Humoral epitope spreading in autoimmune bullous diseases.  
482 *Frontiers in Immunology*. Frontiers Media S.A.; 2018. doi:10.3389/fimmu.2018.00779
- 483 15. Rosen A, Casciola-Rosen L. Autoantigens in systemic autoimmunity: Critical partner in  
484 pathogenesis. *Journal of Internal Medicine*. 2009. pp. 625–631. doi:10.1111/j.1365-  
485 2796.2009.02102.x
- 486 16. Zaenker P, Gray ES, Ziman MR. Autoantibody Production in Cancer-The Humoral  
487 Immune Response toward Autologous Antigens in Cancer Patients. *Autoimmunity*

- 488           Reviews. 2016. doi:10.1016/j.autrev.2016.01.017
- 489   17.   Yaniv G, Twig G, Shor DBA, Furer A, Sherer Y, Mozes O, et al. A volcanic explosion of  
490       autoantibodies in systemic lupus erythematosus: A diversity of 180 different antibodies  
491       found in SLE patients. *Autoimmunity Reviews*. 2015. doi:10.1016/j.autrev.2014.10.003
- 492   18.   Riemekasten G, Hahn BH. Key autoantigens in SLE. *Rheumatology (Oxford)*. 2005;44:  
493       975–82. doi:10.1093/rheumatology/keh688
- 494   19.   Ching KH, Burbelo PD, Tipton C, Wei C, Petri M, Sanz I, et al. Two major autoantibody  
495       clusters in systemic lupus erythematosus. *PLoS One*. 2012;7.  
496       doi:10.1371/journal.pone.0032001
- 497   20.   Dema B, Charles N. Autoantibodies in SLE: Specificities, Isotypes and Receptors. [cited  
498       30 Dec 2019]. doi:10.3390/antib5010002
- 499   21.   Kalinina O, Louzoun Y, Wang Y, Utset T, Weigert M. Origins and specificity of auto-  
500       antibodies in Sm+ SLE patients. *J Autoimmun*. 2018. doi:10.1016/j.jaut.2018.02.008
- 501   22.   Deshmukh US, Bagavant H, Lewis J, Gaskin F, Shu MF. Epitope spreading within lupus-  
502       associated ribonucleoprotein antigens. *Clinical Immunology*. Academic Press; 2005. pp.  
503       112–120. doi:10.1016/j.clim.2005.07.002
- 504   23.   James JA, Harley JB. Linear epitope mapping of an Sm B/B' polypeptide. *J Immunol*.  
505       1992;148: 2074–9. Available: <http://www.ncbi.nlm.nih.gov/pubmed/1372022>
- 506   24.   Pantazes RJ, Reifert J, Bozekowski J, Ibsen KN, Murray JA, Daugherty PS. Identification  
507       of disease-specific motifs in the antibody specificity repertoire via next-generation  
508       sequencing. *Sci Rep*. 2016;6. doi:10.1038/srep30312
- 509   25.   Bozekowski JD, Graham AJ, Daugherty PS. High-titer antibody depletion enhances  
510       discovery of diverse serum antibody specificities. *J Immunol Methods*. 2018;455: 1–9.  
511       doi:10.1016/j.jim.2018.01.003
- 512   26.   Getz JA, Schoep TD, Daugherty PS. Peptide discovery using bacterial display and flow  
513       cytometry. *Methods in Enzymology*. 2012. doi:10.1016/B978-0-12-396962-0.00004-5
- 514   27.   Schreiber A, Humbert M, Benz A, Dietrich U. 3D-Epitope-Explorer (3DEX): Localization  
515       of conformational epitopes within three-dimensional structures of proteins. *J Comput*  
516       *Chem*. 2005. doi:10.1002/jcc.20229
- 517   28.   Moreau V, Granier C, Villard S, Laune D, Molina F. Discontinuous epitope prediction  
518       based on mimotope analysis. *Bioinformatics*. 2006. doi:10.1093/bioinformatics/btl012
- 519   29.   Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T, Freund NT, et al. Pepitope: Epitope  
520       mapping from affinity-selected peptides. *Bioinformatics*. 2007.  
521       doi:10.1093/bioinformatics/btm493
- 522   30.   Huang J, Gutteridge A, Honda W, Kanehisa M. MIMOX: A web tool for phage display  
523       based epitope mapping. *BMC Bioinformatics*. 2006. doi:10.1186/1471-2105-7-451
- 524   31.   Kringelum JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in  
525       antibody: Protein complexes. *Mol Immunol*. 2013;53: 24–34.  
526       doi:10.1016/j.molimm.2012.06.001
- 527   32.   Sun J, Xu T, Wang S, Li G, Wu D, Cao Z. Does difference exist between epitope and non-  
528       epitope residues? Analysis of the physicochemical and structural properties on  
529       conformational epitopes from B-cell protein antigens. *Immunome Research*. 2011.
- 530   33.   Sun P, Ju H, Liu Z, Ning Q, Zhang J, Zhao X, et al. Bioinformatics resources and tools for  
531       conformational B-cell epitope prediction. *Computational and Mathematical Methods in*  
532       *Medicine*. 2013. doi:10.1155/2013/943636
- 533   34.   Paull ML, Johnston T, Ibsen KN, Bozekowski JD, Daugherty PS. A general approach for

- 534 predicting protein epitopes targeted by antibody repertoires using whole proteomes. *PLoS*  
535 *One*. 2019;14: e0217668. doi:10.1371/journal.pone.0217668
- 536 35. Gerli R, Caponi L. Anti-ribosomal P protein antibodies. *Autoimmunity*. 2005.  
537 doi:10.1080/08916930400022699
- 538 36. Madsen BE, Browning SR. A groupwise association test for rare mutations using a  
539 weighted sum statistic. *PLoS Genet*. 2009;5. doi:10.1371/journal.pgen.1000384
- 540 37. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs  
541 and statistical tests. *American Journal of Human Genetics*. Cell Press; 2014. pp. 5–23.  
542 doi:10.1016/j.ajhg.2014.06.009
- 543 38. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis.  
544 *Biostatistics*. 2007;8: 2–8. doi:10.1093/biostatistics/kxl005
- 545 39. Sundar K, Jacques S, Gottlieb P, Villars R, Benito ME, Taylor DK, et al. Expression of  
546 the Epstein-Barr virus nuclear antigen-1 (EBNA-1) in the mouse can elicit the production  
547 of anti-dsDNA and anti-Sm antibodies. *J Autoimmun*. 2004.  
548 doi:10.1016/j.jaut.2004.06.001
- 549 40. Zhang L. Multi-epitope vaccines: A promising strategy against tumors and viral  
550 infections. *Cell Mol Immunol*. 2018;15: 182–184. doi:10.1038/cmi.2017.92
- 551 41. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45: D158–D169.  
552 doi:10.1093/nar/gkw1099  
553