Published in final edited form as:

Nat Med. 2018 November ; 24(11): 1762–1772. doi:10.1038/s41591-018-0203-7.

Antigen Discovery and Specification of Immunodominance Hierarchies for MHCII-Restricted Epitopes

Daniel B. Graham^{1,2,3,4,+,*}, Chengwei Luo^{1,2,5,+}, Daniel J. O'Connell¹, Ariel Lefkovith¹, Eric M. Brown¹, Moran Yassour¹, Mukund Varma¹, Jennifer G. Abelin¹, Kara L. Conway^{2,5}, Guadalupe J. Jasso^{1,6}, Caline G. Matar¹, Steven A. Carr¹, and Ramnik J. Xavier^{1,2,3,4,5,*}

¹Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

²Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

³Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

⁴Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁵Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

⁶Immunology Program, Harvard Medical School, Boston, MA 02115, USA

Abstract

Identifying immunodominant T cell epitopes remains a significant challenge in the context of infectious disease, autoimmunity, and immuno-oncology. To address the challenge of antigen discovery, we developed a quantitative proteomic approach that enabled unbiased identification of MHCII-associated peptide epitopes and biochemical features of antigenicity. Based on these data, we trained a deep neural network algorithm for genome-scale predictions of immunodominant MHCII-restricted epitopes. This algorithm is entitled bacteria-origin T cell antigen (BOTA) predictor. In validation studies, BOTA accurately predicted novel CD4 T cell epitopes derived from the model pathogen *Listeria monocytogenes* and the commensal microorganism *Muribaculum intestinale*. To conclusively define immunodominant T cell epitopes predicted by

*Ramnik Xavier, Center for Computational and Integrative Biology, Massachusetts General Hospital, 185 Cambridge St. Boston, MA 02114, Tel: 617-726-7411, Fax: 617-726-2394, xavier@molbio.mgh.harvard.edu, Daniel Graham, Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142, Tel: 617-714-8572, Fax: 617-714-8957, dgraham@broadinstitute.org.

Author Contributions

Conceptualization, D.B.G., C.L., and R.J.X.; Methodology, D.B.G., C.L., J.G.A., K.L.C., and S.A.C.; Software, C.L. and M.Y.; Formal Analysis, C.L., M.V., and J.G.A.; Investigation, D.B.G., J.G.A., C.G.M., A.L., G.J.J., E.M.B., D.J.O., and K.L.C.; Resources, S.A.C.; Writing – Original Draft, D.B.G.; Supervision, D.B.G., and R.J.X.; Funding Acquisition, R.J.X.

⁺Equal contribution

Data Availability. Source data is available from Figures 1, 2, 5, and 7 and can be found in Supplementary Information. There are no restrictions of source data availability. Data for Figure 7 can be accessed through GEO Series accession number GSE117166 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117166>).

Code Availability. All source code for BOTA is available at <https://bitbucket.org/luo-chengwei/bota>

Competing Financial Interests Statement

The authors declare no competing financial interests.

BOTA, we developed a high-throughput approach to screen DNA-encoded peptide-MHCII libraries for functional recognition by TCRs identified from single cell RNA sequencing. Collectively, these studies provide a framework for defining the immunodominance landscape across a broad range of immune pathologies.

Introduction

Canonical CD4 T cell responses are restricted to peptide antigens presented by MHCII, which in turn is dictated by sequence-specific interactions between the peptide backbone and the MHCII binding groove^{1,2}. Degeneracy in this interaction allows for presentation of a broad spectrum of peptide antigens and promotes diverse responses to potential antigens. However, CD4 T cell responses tend to be constrained to a limited set of immunodominant epitopes, even when the pool of available peptide epitopes is not limiting. Thus, the T cell response must be balanced with respect to the magnitude of the response to any given epitope (to maximize efficacy) and specificity towards multiple epitopes (to combat epitope escape). Despite intensive investigation, the factors controlling immunodominance and antigenicity are incompletely understood.

At the level of antigen-presenting cells, several lysosomal pathways contribute to antigen processing and epitope selection³. In this context, numerous thiol reductases⁴ and proteases^{5,6} perform redundant functions in converting native proteins into MHCII ligands. Sequence- and structure-specific preferences of these enzymes for their substrates can bias which peptides are ultimately processed and loaded onto MHCII. Furthermore, kinetic parameters of the peptide-MHCII interaction impact the stability of the complex such that weak-binding peptides are replaced by means of HLA-DM editing^{7,8}. Taken together, the complexity of antigen processing and limited availability of model immunodominant antigens has posed a barrier to understanding the biochemical features of antigenicity.

Defining immunodominant CD4 T cell antigens has benefited from the development of unbiased approaches for epitope discovery. In this context, autologous self-epitopes have been identified by immunoaffinity purification of MHCII-associated peptides followed by Edman degradation sequencing or mass spectrometry^{9–17}. Later advances in mass spectrometry allowed for identification of tumor antigens¹⁸ and autoantigens^{19,20}. While detection of MHCII-associated peptides derived from extracellular proteins has been achieved²¹, identifying epitopes from complex microorganisms remains challenging. Previous successes identified commensal T cell epitopes from within antigens that are recognized by immunoglobulins^{22–24}. Here, we develop a technology for CD4 antigen discovery and demonstrate its ability to systematically define the immunodominance hierarchy of antigens and discover novel bacterial epitopes based on genomic sequence.

Results

MHCII peptidomics in primary murine dendritic cells defines biochemical features of antigenicity

To define features of antigenicity, we developed a proteomics-based approach to identify MHCII-associated peptides in murine bone marrow-derived dendritic cells (BMDCs) from C57BL/6J mice expressing I-A^b. In these experiments, mature peptide-MHCII complexes were immunoprecipitated from primary BMDCs with monoclonal Y3P antibody²⁵. Associated peptides from biological replicates were acid-eluted, chemically labeled with iTRAQ isobaric mass tags for relative quantification and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Fig. 1A). This sensitive approach identified 3,671 unique peptides derived from 1,088 autologous murine proteins (Fig. 1B and Supplementary Table 1).

Having generated an expansive catalogue of the MHCII immunopeptidome in murine BMDCs, we sought to identify key biochemical features associated with antigenicity. Initially, we derived the optimal I-A^b-binding motif from primary sequence features in autologous murine peptides (Fig. 1C). The resulting 9mer core peptide sequence resembles previous predictions based on *in vitro* binding kinetics between synthetic peptide libraries and purified I-A^b²⁶. In contrast, we detected a strong preference for proline in the P4 position, as was suggested by previous work^{27,28}.

Autophagy shapes the MHCII immunopeptidome

The primary sequence of antigenic peptides confers binding to MHCII but does not represent the only factor governing antigenicity. Given that lysosomal pathways contribute to antigen processing and epitope selection, we hypothesized that autophagy shapes the immunopeptidome by dictating which antigens gain access to the lysosomal compartment. The core autophagy protein ATG16L1 is required for macroautophagy and xenophagy, which divert cytosolic cargo to lysosomes for disposal. To selectively perturb autophagy in antigen-presenting cells, we generated *Atg16l1*^{fl/f} x *CD11c-Cre* mice (*Atg16l1*^{-/-}). MHCII peptidomics experiments demonstrated that *Atg16l1*^{-/-} BMDCs were severely impaired in their ability to present peptides from organelle-derived and cytosolic proteins (Fig. 2A and 2B). In contrast, *Atg16l1*^{-/-} cells displayed elevated levels of lysosome- and endosome-derived peptides compared to WT BMDCs (Fig. 2B). These findings are consistent with the known role for Atg16l1 in macroautophagy and highlight the degree to which lysosomal trafficking impacts the spectrum of peptides presented by MHCII.

Given the abundance of MHCII-associated peptides derived from endolysosomal proteins, we queried these sequences to define additional features that may impart antigenicity. Using the high-resolution I-A^b-binding motif as a reference sequence (Fig. 1C), we scanned endolysosomal proteins identified by proteomics for consensus. As expected, we were able to retrospectively predict peptides that we identified by mass spectrometry; however, we found many instances of consensus sites that were not presented on MHCII. In the vast majority of these cases, consensus sites located within structured protein domains were not detected by proteomics, whereas those located in unstructured inter-domain regions were

(Supplementary Figure 1A). For example, MHCII peptidomics identified one peptide derived from the murine cation-independent mannose 6-phosphate receptor (CIMR). This peptide is located between two CIMR domains. In contrast, we found three additional peptides conforming to the I-A^b-binding consensus motif that were located within CIMR domains and were not detected as processed peptides bound to MHCII (Fig. 2C). Across the dataset, epitopes preferentially originated from interdomain regions of greater than 20 amino acids or protein domains greater than 30 amino acids (Supplementary Figure 1B). Thus, epitope accessibility in the context of native protein structure is an important factor that impacts MHCII binding and protease processing.

Training a deep neural network on MHCII peptidomics data

Having demonstrated that MHCII peptidomics can identify antigenic features conferred by lysosomal processing, we leveraged this dataset to devise a neural network-based algorithm for epitope prediction. The concept of epitope accessibility, together with MHCII affinity, adds to the complexity of formulating predictions for immunodominant epitopes. Therefore, we explored the use of a deep neural network-based algorithm to predict these epitopes using peptidomics data as a training set. We developed a model, bacteria-origin T cell antigen (BOTA) predictor, which generates a list of candidate peptides utilizing information including protein cellular location, transmembrane structure (if applicable), and domain distribution (Fig. 2D). BOTA requires only an annotated genome input to extract amino acid sequences from genes. It then relies on outputs from three algorithms: HMMTOP, PSORT, and HMMER's search against pfam. First, BOTA identifies secreted and cell wall proteins²⁹. Second, it masks intracellular regions and transmembrane domains of cell wall proteins (with an 8-amino acid flanking buffer)³⁰. Third, it excludes regions that fall within small domains or between a series of adjacent domains (inaccessible, compact folding)³¹. The metric used for domain mapping includes the distance to the up/downstream domains, the density of flanking domains, and the domain size. Finally, candidate protein regions are piped into the deep neural network predictor to generate their probabilities to be MHCII binders (Fig. 2D). We note that this pipeline is modular and can be used to generate allele-specific models in both human and model animals.

The core algorithm of BOTA is a deep neural network, which employs a sparse representation of input peptides and multiple pre-trained binder models to generate a feature map. The feature map includes many nonlinear summarizations of peptide features that are used to make predictions. The output from the deep neural network is a score (f) that serves as an indicator of MHCII binding. In the case of mouse I-A^b, we used MHCII peptidomics data to train the BOTA model. The model was trained using randomly sampled parameters 30 times on a 3-fold cross-validation scheme. The optimal parameter calibration was selected based on their receiver operating characteristic curves (ROC AUCs).

The advantages of BOTA over current methods are that BOTA (1) requires only whole genome, not polypeptides, as input, (2) considers epitope accessibility by focusing the search on extracellular regions of proteins and epitope location relative to protein domain organization, (3) trained on peptidomics data from antigen-presenting cells, which are significantly more accurate than *in vitro* peptide binding data, and (4) easily scales to

thousands of genomes for different alleles. Thus, BOTA is capable of predicting immunogenic epitopes from medically relevant bacterial pathogens.

Validation of BOTA epitope predictions for *Listeria* with MHCII peptidomics

To validate BOTA, we tested its performance in predicting MHCII-restricted epitopes from *Listeria monocytogenes*, one of the most common foodborne pathogens associated with a relatively high mortality rate³². Despite the epidemiological burden of listeriosis, the number of experimentally validated CD4 T cell epitopes encoded within the *Listeria* genome is limited. Thus, we sought to identify MHCII-associated peptides in BMDCs exposed to live *Listeria* for 10 minutes or 6 hours prior to immunoprecipitation of mature peptide-MHCII complexes. Associated peptides from biological replicates of the 10-minute and 6-hour timepoints were analyzed using mass spectrometry (LC-MS/MS) (Fig. 3A and B). These experiments discovered 48 unique peptides derived from exogenous *Listeria* proteins. Twenty-nine of these peptides represented nested sets derived from 7 unique proteins (Supplementary Table 2). The second-most enriched peptide was the previously identified immunodominant epitope from listeriolysin O (LLO₁₉₀₋₂₀₅), while the remaining peptides represented novel candidate antigens. Notably, all of the detected peptides were derived from secreted proteins or cell wall proteins. These observations highlight the importance of epitope accessibility for antigen presentation.

We next compared *Listeria* peptides identified by MHCII peptidomics with BOTA predictions. Nine out of 17 BOTA-predicted epitopes were validated by MHCII peptidomics (Supplementary Table 2). Of the 35 unique *Listeria* peptides identified by proteomics with an adjusted p value < 0.05, BOTA-predicted epitopes were present in 28. We compared the prediction accuracy of the BOTA model between training with two data sets: peptides we captured in murine dendritic cells ("Peptidomics Training") and MHCII-associated peptides from the Immune Epitope Database³³ ("IEDB Training") (Fig. 3C and D). Both BOTA pre-training accuracies plateaued after 200 epochs, but BOTA training with our peptidomics data increased prediction accuracy over training with IEDB data by 15%. Notably, BOTA showed strong improvement for mouse MHCII alleles compared to current prediction methods, including state-of-the-art netMHCIIpan²⁶ (Fig. 3E). To test BOTA predictions for human HLA alleles, we focused on the inflammatory bowel disease risk allele HLA-DRB1*01:03³⁴. We first searched IEDB for HLA-DRB1*01:03-restricted epitopes validated by MHCII binding assays or T cell reactivity assays (www.iedb.org). We then used full-length protein sequences associated with these epitopes as input for BOTA upstream modules (HMMTOP and Pfam) and used the output as input for epitope prediction using netMHCIIpan with HLA-DRB1*01:03 specificity. This approach yielded 76% accuracy (38 out of 50) for predicting IEDB validated epitopes, with 12 missed epitopes from BOTA (Fig. 3F). In contrast netMHCIIpan predictions on full-length proteins without using BOTA upstream modules performed with 22% accuracy (11 out of 50), while 39 epitopes were missed by NetMHCIIpan (Fig. 3F).

Improvement in prediction accuracy achieved by BOTA signifies successful application of a deep neural network to a complex biomedical problem. Previous efforts using traditional neural networks or hidden Markov models were limited in their ability to extract highly

abstract features, leading to insufficient insights into epitope prediction³⁵. In contrast, BOTA was designed to predict CD4 T cell epitopes for virtually any MHCII allele and any antigen source, including commensal microbes, pathogens, autoantigens, and tumor antigens.

BOTA and MHCII peptidomics predict immunodominance of *Listeria* epitopes *in vivo*

Given that MHCII peptidomics and BOTA successfully identified *Listeria* peptides and features associated with efficient antigen presentation, we sought to measure T cell responses to these epitopes *in vivo*. To quantify the CD4 T cell response to the top eight candidate *Listeria* epitopes, mice were sacrificed seven days after intraperitoneal (i.p.) infection with *Listeria*. T cells were restimulated *in vitro* with synthetic peptides, and IFN- γ responses were enumerated by ELISPOT. Robust responses were detected against four of the eight candidate epitopes (lmo0202, lmo2558, lmo2185, lmo0135) (Fig. 4A). The remaining four epitopes elicited weaker responses near the limit of detection for ELISPOT, suggesting subdominance (Fig. 4B). Notably, the T cell response to these eight candidate epitopes correlated remarkably well with their abundances detected by MHCII peptidomics (Fig. 4C). To determine if the abundance of *Listeria* antigens drives immunogenicity, we evaluated previously published microarray data measuring RNA expression of *Listeria* genes after infection of macrophages³⁶. Accordingly, *Listeria* antigen expression correlated modestly with our measurement of the corresponding T cell response in ELISPOT assays (Fig. 4D), although low expression did not appear to preclude antigenicity (Fig. 4E).

Integration of T cell phenotype with TCR specificity in the *Listeria* response

Having demonstrated the utility of BOTA for predicting bacterial epitopes, we next sought to rigorously characterize the antigen-specific T cell response to *Listeria*. Previous approaches for defining immunodominance, such as ELISPOT, rely on ranking candidate epitopes based on the magnitudes of the T cell responses they elicit *in vivo*. However, such approaches require prior knowledge of TCR specificity and T cell phenotype (cytokine profile). To address these limitations, we developed a technique for single cell analysis of T cell phenotypes matched to their corresponding TCR sequences. This TCRseq approach allows for simultaneous enumeration of T cell clonal frequency and cytokine profiles (Supplementary Figure 2).

Towards this end, we infected mice with *Listeria* and FACS-sorted single CD4 T cells for transcriptomics and TCR sequencing. Based on unbiased clustering of transcriptional profiles, T cells partitioned into 4 distinct subsets (Fig. 5A and B). To identify the activated T effector (Teff) cluster, we generated a per-cell Teff score that was based on the gene expression signature derived from ImmGen datasets comparing CD4 Teff splenocytes (day 8 after lymphocytic choriomeningitis virus, LCMV, infection) versus naïve CD4 T splenocytes³⁷. These analyses identified cluster 2 as being enriched for Teff cells expressing signature activation markers (Ccl5, Nkg7, Ikzf1), genes that regulate ER homeostasis (Calr, Pdia3), and genes that control cellular metabolism (Acly, Akr1a1) (Fig. 5A). Amongst these T cells, the TCR repertoire was remarkably diverse with a maximal clonal frequency of ~3% for conventional T cells. The most abundant conventional TCR (Trav14–2Traj25|Trbv5Trbj2–7) was detected exclusively in cells residing in Teff cluster 2. By contrast, NKT

cells expressing the invariant alpha chain Trv11Traj18 were found scattered throughout clusters (Fig. 5C and Supplementary Table 3).

Using TCRseq and whole transcriptome data (Supplementary Table 4), we prioritized the most abundant TCR to test for reactivity against BOTA predictions (Fig. 5C). In designing a screening modality, we took into consideration the low affinity interaction between TCRs and peptide antigen presented on MHCII. This interaction occurs between multiple TCRs and pMHCII complexes within lipid bilayers on adjacent cells. Avidity and kinetics of engagement/disengagement elicit a TCR signaling cascade that amplifies the input signal. Therefore, we designed a heterologous expression system in which TCR-negative BW5147-CD4-CD28 cells are transduced to express single chain TCR α and TCR β fused to the cytoplasmic tail of CD3 ζ . Functional TCR $\alpha\beta$ proteins in BW5147-CD4-CD28 cells engage cognate peptide-MHCII on antigen presenting cells to initiate a TCR signaling response that results in expression of IL-2. As a source of surrogate antigen presenting cells, we utilized HEK 293T cells transfected to express I-A^b α -CD3 ζ and candidate peptide antigens fused to I-A^b β -CD3 ζ . For these experiments, we selected the most abundant TCR clone from TCRseq to test for reactivity with 4 *Listeria* pMHCII complexes predicted by BOTA (Fig. 5D). As a positive control, we demonstrated that the OT2 TCR expressed in BW5147-CD4-CD28 cells reacted with ovalbumin peptide in the context of I-A^b by inducing IL-2 secretion (Fig. 5E). Similarly, a previously identified TCR (LLO_118)³⁸ reacted robustly with LLO (Fig. 5E). The top TCR candidate identified by TCRseq also reacted with LLO, according to induction of IL-2 and binding to LLO-I-A^b tetramers (Fig. 5E and F and Supplementary Figure 3). Taken together, these experiments establish the feasibility of integrating population-level TCR repertoires with antigen specificity by screening individual TCRs for reactivity against candidate antigens predicted *in silico*.

Identification of commensal epitopes

Having validated BOTA epitope predictions for a common pathogenic bacterium, we sought to explore the complex relationship between commensal microbes and host adaptive immunity. In this context, the intestinal microbiome fine-tunes inflammatory thresholds, primes innate immune effector function, and shapes the adaptive immune response through selection and tolerization of lymphocytes³⁹. Thus, the dynamic role of the microbiome in immune education impacts organ systems throughout the body, and in turn, many disease states⁴⁰. We reasoned that because the host adaptive immune system continuously interacts with the microbiome, monitoring the magnitude and nature of the T cell response to specific commensal antigens can reveal the health status of the immune system. In the state of health, the adaptive immune system maintains tolerance to local gut antigen exposure and protection from systemic infection. The sheer number of bacterial species and diversity of protein-coding elements in the microbiome represents an enormous search space. Towards identifying immunogenic commensals, we developed “serum immunoglobulin commensal capture and sequencing” (SICC-seq). Mice were administered a course of dextran sodium sulfate (DSS) to induce barrier breach and allowed to recover for 7 days. Serum was then harvested from mice and incubated with stool to opsonize commensals. IgG-positive microbes were enriched by selection with Protein A/G-coupled magnetic beads and analyzed along with total stool by 16S rRNA sequencing (Supplementary Table 5). Using this

approach, we demonstrated that induction of colitis with dextran sulfate sodium (DSS), an epithelial injury model, elicits a systemic T cell-dependent IgG response that preferentially targets bacteria in the genus *Bacteroidales* (Fig. 6A). In contrast, *Akkermansia* evade this response, likely due to their ability to induce T cell tolerance and IgA responses under homeostatic conditions (before colitis) (Fig. 6A). Having established the immunogenicity of *Bacteroidales*, we employed BOTA to predict T cell epitopes from the dominant species inhabiting mice⁴¹ (Fig. 6B) and identified a highly abundant epitope in a SusC-like protein that is conserved across the *Bacteroidales* genus and often duplicated within species (Fig. 6B). To determine if T cells recognize this SusC epitope *in vivo*, we harvested splenocytes from a naïve mouse and stimulated them with SusC peptide *in vitro*. Importantly, the SusC peptide induced IL-10 production by T cells, indicating a homeostatic relationship between host T cells and *Bacteroidales* in a normal healthy mouse (Fig. 6B). We hypothesize that these interactions extend to many other commensals, and that tumultuous relationships between T cells and commensals typify immune pathologies and autoimmunity.

Discussion

Host-microbe interactions cooperatively influence the specificity and diversity of the T cell response. A deeper understanding of this relationship requires developing new approaches for unbiased antigen discovery, defining features of antigenicity, and elucidating host pathways that underlie preferential selection of these features. Toward these objectives, we report a highly quantitative adaptation of MHCII peptidomics. By recapitulating the interaction between primary murine BMDCs and live *L. monocytogenes*, we identified four dominant CD4 T cell epitopes and established their immunodominance hierarchy *in vivo*. In addition, MHCII peptidomics identified over 3,600 autologous mouse peptides, which provided an exceptionally detailed view of lysosomal function.

As a consequence of deep profiling of the MHCII immunopeptidome, we generated a rich dataset for identifying features associated with antigen processing and developed BOTA as a predictive algorithm, incorporating several important attributes of immunodominant epitopes revealed by proteomics. Based on observations from MHCII peptidomics, optimal epitopes tend to (1) derive from secreted proteins or extracellular regions of cell wall proteins in bacteria, (2) have a primary sequence structure that conforms to a defined MHCII binding motif, (3) be located more than 8 amino acids away from transmembrane domains, and (4) have tertiary structure characteristics that promote accessibility to the MHCII binding groove and the enzymes required for proteolytic processing. Importantly, all of the features identified by peptidomics are readily identifiable at the level of DNA sequence; therefore, the only input users need to run a BOTA immunodominance analysis is an annotated genome. Consequently, BOTA enables large-scale mapping of the immunodominant epitope landscape for any bacterial species or collection of species, such as the human microbiome. Taken together, we demonstrate the utility of MHCII peptidomics for training a deep neural network that specifically identifies candidate epitopes with key features associated with immunodominance and antigenicity. Furthermore, MHCII peptidomics serves as a powerful tool for unbiased discovery of complex pathogen antigens and for interrogation of host pathways underlying human disease.

Inherent to antigen discovery is the significant challenge of validating epitope predictions. Conclusive validation of epitope immunogenicity requires demonstration that a measurable T cell response is elicited *in vivo*. To address this challenge, we developed an approach for identifying and integrating TCR repertoire, phenotype, and antigen reactivity. Coupling TCRseq with whole transcriptome profiling at the single cell level enabled assignment of transcriptional phenotypes to individual TCRs. While we employed this approach to identify TCRs associated with CD4 T effector cells derived from *Listeria*-infected mice, it is applicable to any immune cell type that can be defined at the level of the transcriptome, including Treg or Th17. Such determinations are challenging in T cell hybridomas because the primary T cell phenotype is not preserved after immortalization. Moreover, generating T cell hybridomas is an inefficient process that is further biased by chromosome loss, drug selection, and screening for antigen reactivity. In contrast, TCRseq is comparatively efficient, which is an important consideration in the context of limited T cell input from precious clinical specimens. A small tissue biopsy is sufficient to generate a permanent archive of TCR sequences matched with transcriptional profiles that can be used to screen defined TCRs for reactivity against candidate epitopes. Thus, immunodominance can be unambiguously defined in the context of epitopes that elicit the strongest T cell responses, as determined by clonal frequency and/or absolute numbers within a relevant tissue or organ.

Here, we develop a technology with broad utility for antigen discovery. BOTA is capable of predicting CD4 T cell epitopes from multiple sources, including genomes or transcriptomes derived from pathogens, the microbiome, allergens, tumors, and tissue biopsies (Supplementary Figure 4). By coupling BOTA predictions with TCRseq from matching specimens, we enable generation of DNA-encoded epitope and TCR libraries that can be screened to define antigen specificity and quantify immunodominance. The screening platform we developed can be implemented in arrayed format (well-based screening of individual TCRs by peptide-MHCII antigens) or in a pooled format where cells expressing TCR libraries are cocultured with cells expressing peptide-MHCII libraries. For example, we have performed TCR fingerprinting by coculturing stimulator cells expressing the OT2 TCR with responder cells displaying a library of ovalbumin peptide mutants expressed as a fusion protein with MHCII ectodomain linked to the CD3 ζ cytoplasmic domain. After FACS-sorting activated responder cells (41bb⁺) that had engaged in a productive interaction with stimulator cells, we sequenced the Ova-MHCII ζ libraries to recover the known cognate peptide ligand for the OT2 TCR (Supplementary Figure 5). In this context, it is possible to simultaneously screen many epitopes for reactivity with a given TCR.

With new approaches to effectively predict T cell epitopes and validate TCR reactivity, it is possible to identify antigenic determinants in bacterial pathogens and even complex communities such as the intestinal microbiome. We identified an I-A^b-restricted epitope in a SusC-like protein derived from *Bacteroidales* that is associated with IL-10 producing T cells derived from mouse spleen. Notably, these experiments identified systemically circulating T cells with reactivity to a benign commensal species. Our findings highlight the intimate relationship between commensals and the host adaptive immune system. In this context, previous studies in mouse models have demonstrated that a systemic T cell-dependent IgG response to commensals is cross-protective against pathogen infection⁴². Similarly, IgG-reactivity with commensals has been shown to be widespread in healthy humans and

qualitatively perturbed in the context of autoimmunity^{43,44}. While our understanding of host-commensal mutualism is in its infancy, future research stands to reveal how this relationship promotes homeostatic immune-regulation or precipitates immune dysfunction. New approaches to antigen discovery offer new opportunities for biomedical research, such as tracking antigen-specific immune responses in clinical studies, vaccine design, and understanding the host-microbiome relationship in the cancer, autoimmunity, and inflammatory disease.

Online Methods

Immuno-affinity purification of MHCII complexes.

Bone marrow was harvested from C57BL/6J (WT), *Atg16l1*^{f/f} x CD11c-Cre⁴⁵. BMDCs were differentiated from bone marrow for seven days in antibiotic-free DMEM supplemented with 20% FBS and GM-CSF (2% conditioned TOPO medium). In parallel, *L. monocytogenes* strain EGDe (ATCC BA-679) was cultured overnight in BHI medium, washed in PBS, and cocultured with BMDCs at an MOI of approximately 100:1 for 30 minutes. Tissue culture dishes containing cocultures were washed in PBS and cultured for an additional 10 minutes or 6 hours in DC media containing gentamicin (30 µg/ml). BMDCs were then harvested by scraping, washed in PBS, and lysed in 1% NP-40, 4 mM MgCl₂, 6 µg/ml DNaseI from bovine pancreas (Sigma Aldrich), and PBS pH 7.4 at 250 × 10⁶ cells/4ml lysis buffer/sample. Clarified lysates were subjected to immunoprecipitation overnight at 4°C with gentle rotation. Immunoprecipitation was performed with NHS Mag Sepharose (GE Healthcare) covalently coupled to anti-H2-IA (clone Y3P²⁵). Each sample contained 130 µl packed beads and approximately 650 µg antibody. Beads were then washed twice in PBS containing 0.1% NP-40 and three times in PBS.

MHC-peptide elution and desalting.

Peptides were eluted from MHC complexes and desalted on in-house-built Empore C18 StageTips (3M, 2315) as described previously⁴⁶. Sample loading, washes, and elution were performed on a tabletop centrifuge at a maximum speed of 2,000–3,500 x g. Briefly, StageTips were equilibrated with 2 × 100 µL washes of methanol, 2 × 50 µL washes of 50% acetonitrile (ACN)/0.1% formic acid (FA), and 2 × 100 µL washes of 1% FA. In a tube, the dried beads from MHC-associated peptide IPs were thawed at 4°C, reconstituted in 50 µL 3% ACN/5% FA, and loaded onto StageTips. The beads were washed with 50 µL 1% FA, and peptides were further eluted using two rounds of 5-minute incubations in 10% acetic acid. The combined wash and elution volumes were loaded onto StageTips. The tubes containing the IP beads were washed again with 50 µL 1% FA, and this volume was also loaded onto StageTips. Peptides were washed twice on the StageTip with 100 µL 1% FA. Peptides were eluted using a step gradient of 20 µL 20% ACN/0.1% FA, 20 µL 40% ACN/0.1% FA, and 20 µL 60% ACN/0.1% FA. Step elutions were combined and dried to completion.

iTRAQ 4-plex labeling for quantitative proteomics.

Quantitative proteomics was performed as described previously⁴⁷. Briefly, each peptide mixture was reconstituted in 20 µL dissolution buffer labeled with 0.5 units (40 µL of 80 uL

iTRAQ4 reagent in ethanol) of iTRAQ 4 reagent for 1 hour at room temperature (RT) (~22°C). Excess reagent was quenched with 5 µL Tris/HCl for 30 minutes at RT. iTRAQ 4-plexes were combined, dried to completion, and acidified by adding 1% FA (150 µL). Samples were desalted on in-house built Empore SCX/C18 StageTips (3M, 2315, 2251) as described previously⁴⁶. Peptides were eluted from the C18/SCX StageTips with two pH cuts (5.5, 11), and the remaining 20% ACN was diluted with 1% FA. Samples were then loaded on in-house built Empore C18 StageTips, desalted, and dried to completion as described above.

MHC-peptide sequencing by tandem mass spectrometry.

All nanoLC-ESI-MS/MS analyses employed the same LC separation conditions described below. Samples were chromatographically separated using a Proxeon Easy NanoLC 1000 (Thermo Scientific, San Jose, CA) fitted with a PicoFrit (New Objective, Woburn, MA) 75-µm inner diameter capillary with a 10-µm emitter was packed under pressure to ~20 cm with of C18 Reprosil beads (1.9 µm particle size, 200 Å pore size, Dr. Maisch GmbH) and heated at 50°C during separation. Samples were reconstituted in 9 µL 3% ACN/5% FA, and 3 µL (~100 × 10⁶ cell equivalents) was injected for analysis. Peptides were eluted with a linear gradient from 7–30% of Buffer B (0.1% FA/90% ACN) over 82 min, 30–90% Buffer B over 6 min, and then held at 90% Buffer B for 15 min at 200 nL/min (Buffer A: 0.1% FA/3% ACN) to yield ~11 sec peak widths. During data-dependent acquisition, eluted peptides were introduced into a Q-Exactive HF mass spectrometer (Thermo Scientific) equipped with a nanoelectrospray source (James A. Hill Instrument Services, Arlington, MA) at 2.15 kV. A full-scan MS was acquired at a resolution of 60,000 from 300 to 1,800 *m/z* (AGC target 1e6, 20 ms Max IT). Each full scan was followed by top 15 data-dependent MS2 scans at resolution 15,000, using an isolation width of 1.7 *m/z* with a 0.3 *m/z* offset, a fixed first mass at 100 *m/z*, a collision energy of 29, an ACG Target of 5e4, and a max fill time of 100 ms max injection time. An isolation offset of 0.3 *m/z* was used so that doubly charged precursor isotope distributions would be centered in the isolation window. Some MHC-associated peptides tend to be short (<15 amino acids) so the monoisotopic peak is nearly always the tallest peak in the isotope cluster and the mass spectrometer acquisition software places the tallest isotopic peak in the center of the isolation window in the absence of a specified offset. Dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 10 sec. Charge state screening was enabled along with monoisotopic precursor selection using Peptide Match Preferred to prevent triggering of MS/MS on precursor ions with charge state 1, >6, or unassigned.

Interpretation of MS/MS data.

Mass spectra were interpreted using the Spectrum Mill software package v5.1 pre-Release (Agilent Technologies). MS/MS spectra were excluded from searching if they did not have a precursor MH⁺ in the range of 600–4000, had a precursor charge > 5, or had a minimum of < 5 detected peaks. Merging of similar spectra with the same precursor *m/z* acquired in the same chromatographic peak was disabled. High resolution MS/MS spectra were searched against a UniProt database containing reference proteome sequences (including isoforms and excluding fragments) from human mouse (41,157 entries) with a set of common

laboratory contaminant proteins (150 sequences) to yield a total of 41,307 redundant sequences. The sequences were downloaded from the UniProt website in April 2013.

Prior to both search rounds all MS/MS had to pass the spectral quality filter with a sequence tag length > 3, i.e., minimum of 4 masses separated by the in-chain mass of an amino acid. In the first round search, all spectra were searched using a no-enzyme specificity, fixed modification of cysteine as unmodified, fixed modification of partial iTRAQ labeling, the following variable modifications (oxidized methionine, deamidation, acetyl-N-term), a precursor mass tolerance of ± 10 ppm, product mass tolerance of ± 20 ppm, and a minimum matched peak intensity of 50%. Peptide spectrum matches (PSMs) for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to apply target-decoy based FDR estimation at the PSM level to set scoring threshold criteria. Peptide autovalidation was performed with an auto thresholds strategy using a minimum sequence length of 7, automatic variable range precursor mass filtering, and score and delta Rank1 – Rank2 score thresholds optimized across all LC-MS/MS runs. This yielded a PSM level FDR estimate for precursor charges 1 thru 7 of < 1.0% for each precursor charge state. In the second round search, all remaining spectra that that were not confidently identified in the first round were searched using the above parameters against a RefSeq database containing *L. monocytogenes* EGD-e reference protein sequences (2,867) downloaded in December 2014. An additional round of FDR thresholding as described above was applied to PSMs from the second round search to estimate FDR by species (mouse vs. *L. monocytogenes* EGD-e). The combined PSMs from each round had a peptide level FDR < 2.0%. Only *L. monocytogenes* EGD-e peptides that did not overlap with human peptides were reported.

Listeria infection and ELISPOT.

L. monocytogenes strain EGDe (ATCC BA-679) was grown to log phase in BHI, washed in PBS, and used to inoculate mice by i.p. injection with a dose of 1×10^4 CFU/200 μ l/mouse. Seven days after infection, spleens were harvested, red blood cells were lysed, and single cell suspensions were replated at 2.5×10^5 /200 μ l/well on 96-well ELISPOT plates (Millipore MAIPS4510) precoated with anti-IFN γ (BD Pharmingen 551881) at a concentration of 5 μ g/ml in PBS. Peptides were added at a concentration of 100 nM to restimulate splenocytes overnight. After stimulation, plates were washed, blocked, and incubated with biotinylated anti-IFN γ (BD Pharmingen 551881) at a concentration of 2 μ g/ml and streptavidin-AKP (BD Pharmingen 554065) at a dilution of 1:1000. After 60 minutes, plates were washed and developed with 3-amino-9-ethylcarbazole (AEC, Sigma-Aldrich, A6926). IFN γ -secreting cells were scanned and enumerated using an Immunospot Analyzer ELISpot reader (CTL). All samples were run in triplicate. The following peptides were synthesized by New England Peptide: lmo2185_293–312 ADFRYVFDTAKATAASSYPG, lmo0202_189–204 WNEKYAQAYPNVSAKI, lmo0135_150–169 VDDTTVKFTLPTVAPAFENT, lmo2558_533–553 APGQETQHYYGLPVADSAIDR, lmo2360_289–306 GGINQAYTGSTALSDGLN, lmo0186_285–303 GTKEKVVATPVSNSVSTSSA, lmo0582a_26–46 STVVVEAGDTLWGIAQSKGTT, lmo0582b_12–25 IAVTAFAPTIAASA. To calculate the frequency of CD4⁺ cells in each splenocyte sample from infected mice, an aliquot of cells

from each mouse was stained for flow cytometric analysis. 1×10^6 cells were incubated with 2.4G2 Mouse Fc block in PBS/FBS (BD Pharmingen, 553142) for 20 min at 4°C. Cells were then washed and stained with FITC-conjugated anti-mouse CD4 (Biolegend, 100510) for 20 min at 4°C. Fluorescently labeled cells were acquired on the FACSVerse flow cytometer (BD Biosciences) and analyzed using FlowJo Analysis Software (Tree Star).

Cytokine assays.

Splenocytes from 8–12 week old C57BL/6 mice were isolated and cultured in complete media (DMEM with 10% FBS, 1% L-glutamine, 2.5% sodium bicarbonate, and 1% penicillin/streptomycin) at a concentration of 1×10^7 cells/mL. Cells from each well were pulsed with a final concentration of 10 μ M of the SusC 14mer peptide (VLKDASAAIYGSR) or a vehicle control (DMSO) for 24 hrs, at 37°C. Supernatants were collected, and cytokine production was measured by cytometric bead array (BD Biosciences, Flex-set), for IFN- γ , IL-17A and IL-10.

BOTA algorithm architecture.

BOTA starts with an input genome and the associated genome annotation in GFF3 format. It first extracts the amino acid sequences of the protein-coding genes and then performs the following analysis: (i) domain identification using hmmscan of HMMER3.1b2⁴⁸ against Pfam⁴⁹; (ii) cellular localization prediction using PSORTb v3.0.2²⁹ with default settings; and (iii) transmembrane topology prediction by HMMTOP³⁰. This information is then integrated to generate a list of candidate peptides following three criteria: (i) the protein should be located in the outer membrane, cell wall (if applicable), or extracellular space; (ii) if located in the outer membrane or cell wall, only the out-facing part of the protein will be considered; and (iii) the peptide should present sufficient accessibility as decided by three rules: (iii-a) it cannot be fewer than eight amino acids away from anchoring domain in the cell wall or outer membrane; (iii-b) it cannot be located in domains shorter than 30 amino acids due to the fact that small domains are usually tightly folded; and (iii-c) it cannot be flanked by two domains that are fewer than 20 amino acids apart. These criteria were based on observations about the vast majority of *Listeria* epitopes captured by MHCII peptidomics. These candidate peptides are then scored by the deep neural network as described below. For each peptide classified as a candidate epitope, BOTA further validates it with the motif score as defined previously and a randomized score. For the motif score validation, it calculates all the 9-mers within the peptide and requires the maximum to be larger than 5×10^{-11} ; for the randomized score, BOTA shuffles the amino acids in the 9-mer and calculates the motif score for each of the randomized 9-mers. The original 9-mer's motif score requires a rank in the top 30% of all randomized 9-mers.

Deep neural network for MHCII binder prediction.

We employed a deep neural network scheme to develop an MHCII-binder predictor. In brief, we first encoded every amino acid into a p -dimensional binary vector b , with half of its elements being 1 and the rest being 0, chosen at random. Therefore, given a peptide with length l longer than k , it is first converted to an $l \times b$ descriptor matrix S , in which $S_{ij}=1$ if the i -th amino acid's 1-valued indices overlap with j , otherwise $S_{ij}=0$. The matrix S is then normalized by row sum to become S' such that

$$S'_{ij} = S_{ij} / \sum_{j=1}^b S_{ij}$$

S' is then convoluted into an $(l-k+1) \times d$ matrix X , where d is the number of pre-trained motif network models within the overall model and k is the length of such binding cores. X_{ij} represents the score of motif network model j aligned to position i . The cohort of motif network models are arranged in a $d \times k \times b$ array H with H_{ijk} being the d -th motif network model aligned to the k -th position of the b -th set of amino acids (Fig. 2D). This sequence conversion and convolution setup is similar to model developed by Alipanahi et al⁵⁰.

With the convoluted matrix X , we filter it with a max-rectified linear unit (ReLU) layer, and the rectified matrix Y is then fed into a max pooling stage to be transformed into d -dimensional vector Z , in which

$$z_j = \max(Y_{1j}, Y_{2j}, \dots, Y_{nj})$$

This d -dimensional vector Z is then used as input for a neural network with Maxout dropout model. Z is then used as the input for a standard output layer for final prediction calculation.

The goal is to minimize the prediction error as measured by 1-norm distance of all the peptides. Back propagation with a stochastic gradient descent method using mini batch size at 64 was used to reach optimal weights. To train the weights, the mouse epitope peptides were used as training set true positive; we also added *in silico* true negative peptides by surveying an equal number of the peptides that are not part of the peptidomics data readout at random. We first randomly constructed 100 replicates of 3-fold cross validation datasets using the epitopes and the *in silico* true negative sequences. For each replicate, BOTAs' initial state weights were assigned randomly and then trained until performance plateaued (<0.1% improvement in ten iterations). The validation's average AUC was used as a measure of model fitness; and the one with the highest average AUC among the 100 replicates was chosen to generate the final optimal parameters by using all mouse peptides (Fig. 3C). Then the optimal model constructed from the previous step was used to predict the unique *Listeria* peptides, and netMHCIIpan was used to predict the affinity of the same set of peptides with default settings.

TCRseq and 5' digital gene expression (DGE).

Single T Cells from mice were sorted into 384 well plates containing capture buffer, Ultrapure BSA (0.5mg/mL, MCLAB) and 2uM Template Switch Oligo (TSO, 384 unique TSO, 1 unique TSO per well from Integrated DNA Technologies). Oligo sequences can be found in Supplementary Table 6. A reverse transcription reaction was performed with a polydT oligo and Maxima Reverse Transcriptase (Thermo Fisher). Unique TSO were composed of a 5' Biotin, a truncated Illumina adaptor, a 7 base pair unique molecular identifier, and 3 riboguanosines at the 3' end. The 5' end of the polydT oligo was comprised of a 5' Biotin and a truncated Illumina adaptor. Following reverse transcription the reactions were treated with Exonuclease I (NEB) and then in-well whole transcriptome amplification

(WTA) was performed with Herculanase II (Agilent). The primers in the WTA targeted the two truncated Illumina adaptor sequences used in the TSO and in the polydT oligo from the RT reaction. After WTA all 384 reactions were pooled and purified over a single Zymoclean gel DNA recovery column (Zymo Research Corporation). The volume of DNA was reduced by using Agencourt AMPure XP (Beckman Coulter) beads after the column elution.

In order to capture the TCRA and TCRB chain sequences, 0.5–1ng of WTA product was used in two separate PCR reactions (PCR #1) using Herculanase II (Agilent). The TCRA reaction used a primer to target the TCRA constant region (CGGCACATTGATTTGGGAGT) and a primer to target the truncated Illumina adaptor in the TSO. The TCRB reaction used a primer to target the TCRB constant region (CTTGCCATTCACCCACCAGC) and a primer to target the truncated Illumina adaptor in the TSO. The DNA from TCRA PCR#1 and TCRB PCR#1 was isolated by Agencourt AMPure XP beads (Beckman Coulter) and used separately in a second set of PCR reactions (PCR#2). PCR#2 further targeted the TCR regions of interest by nesting the TCR specific primer within the region already captured. Additionally, PCR#2 served to add Illumina adaptors for sequencing to the 5' and 3' ends of the targeted regions. The primer targeting the TCR for TCRA PCR#2 contained the sequence CACAGCAGGTTCTGGGTCTGGATG with an Illumina P5 sequencing adaptor appending the 5' end. The primer targeting the TCR for TCRB PCR#2 contained CAAGGAGACCTTGGGTGGAGTCACA with an Illumina P5 sequencing adaptor appending the 5' end. The second primer used in both PCR#2 reactions targets the TSO-labeled end and includes an Illumina P7 sequencing adaptor and 8 base pair sample barcode. These 2 reactions were then cleaned using Agencourt AMPure XP beads (Beckman Coulter) and the amplicons generated were selected and isolated by zymoclean gel DNA recovery column (Zymo Research Corporation) after the PCR products were run on a 2% E-Gel EX agarose gel (Thermo Fisher). TCRA and TCRB amplicons were pooled at equimolar concentrations for sequencing. A final concentration of 14nM with 10% PhiX Control (Illumina) was sequenced with a 600- cycle MiSeq kit (Illumina) by paired-end reads of 305 and 305 cycles and one adapter read.

To generate a 5'DGE library, WTA product (0.6–1ng) was tagmented for 10 minutes using the Nextera XT Library Prep Kit (Illumina) followed by a 14 cycle PCR to add adapters and amplify the fragmented library. The 14 cycle PCR incorporates the provided Nextera XT i5 adapters (Illumina) and a custom universal primer that targets the TSO-labeled end and includes an Illumina P7 sequencing adapter and 8 base pair sample barcode. The tagmentation reaction was cleaned with Agencourt AMPure XP beads (Beckman Coulter) and the library size was selected by zymoclean gel DNA recovery column (Zymo Research Corporation) after the library was run on a 2% E-Gel EX agarose gel (Thermo Fisher). The library, at a final concentration of 2pM with 10% PhiX Control (Illumina), was sequenced with a 75-cycle NextSeq 500/550 kit (Illumina) by paired-end reads of 46 and 21 cycles and one adapter read.

TCR-seq reads were first deconvoluted according to well identifier and unique molecular identifier (UMI). We required that at least 6 reads were detected for each UMI. For each UMI, we determined the consensus sequence at each nucleotide position for all reads associated with that UMI. Consensus sequences for each UMI were mapped by BLAST to

the TCR database (IMGT reference set; <http://www.imgt.org/vquest/refseqh.html>). The best hit for each UMI-associated TCR consensus sequence (representing TCR V and J gene segments) was quantified based on read counts and tabulated in association with each distinct UMI on a per-well basis.

5'DGE RNAseq data were analyzed using an in-house data analysis pipeline. Reads were aligned to the mm10 genome using bwa⁵¹. The downstream analysis was carried out using the Seurat package (<http://www.satijalab.org/seurat>). The dataset was filtered to remove genes that were expressed in less than 10% of the data or cells that expressed transcripts that mapped to fewer than 500 unique genes. Further, to remove doublet cells, cells that displayed more heterogeneity than the 90th quantile (1,846 unique genes) were filtered out. Highly variable, highly expressed genes ($\log(\sigma/\mu) > 0.75$ and $\log(\mu) > 1.5$) were identified from a mean-variance dispersion analysis. These genes were then used for clustering the single cells using the SNN-Cliq algorithm, which maps the cells onto a k-nearest neighbor graph and then finds “cliques” of cells expressing similar genes⁵². The cells were clustered on the basis of highly expressed, highly varying genes, using a graph-based method, SNN-Cliq^{52,53}.

TCR Screening.

TCR-negative BW5147 cells were obtained from ATCC. The stable sub-line BW_4–28 was produced by introduction of murine CD4 and CD28 by lentiviral transduction. The Orf encoding CD4-P2A-CD28 was synthesized (IDT) and cloned in place of spCas9-P2A-BlastR into pXPR_BRD101 (Genetic Perturbation Platform, Broad Institute) by Gibson assembly. The stable sub-line BW_B7/4 was produced by introduction of chimeric murine CD86 cytoplasmic domain fused in-frame with CD4 transmembrane and cytoplasmic domains by lentiviral transduction. The Orf encoding CD86/CD4 was synthesized (IDT) and cloned in place of spCas9-P2A-BlastR into pXPR_BRD101 (Genetic Perturbation Platform, Broad Institute) by Gibson assembly.

scTCRz_v3 vector was derived from pLX_TRC307 (Genetic Perturbation Platform, Broad Institute) by replacing the stuffer sequence with the following codon-optimized Orf: mTrbc1-hCD3zeta(transmembrane and cytoplasmic domains)-P2A-hIgKleader-mTrac-hCD3zeta(transmembrane and cytoplasmic domains). Single-chain TCRs were synthesized as gBlocks (IDT) comprising Trav-Traj-linker(3xG₄SGGGG)-Trbv-Trbj and cloned by Gibson assembly into NheI and BsrGI sites upstream of Trbc1 in scTCRz_v3 vector. MHCIIz_v2 vector was derived from pLX_TRC307 (Genetic Perturbation Platform, Broad Institute) by replacing the stuffer sequence with the following sequence: H2Ab1leader-XmaI-linker(3xG₄S)-H2Ab1(extracellular and transmembrane domains)-hCD3zeta(cytoplasmic domain)-P2A-H2Aa(extracellular and transmembrane domains)-hCD3zeta(cytoplasmic domain). Peptide epitope-encoding sequences were synthesized as ultramers (IDT) and cloned by Gibson assembly into the XmaI site in the MHCIIz_v2 vector. scTCRs were introduced into BW_B7/4, by lentiviral transduction and maintained under puromycin selection (3ug/ml). To define antigen specificity of TCRs, BW_4–28 cells expressing scTCRs were cocultured with HEK 293T cells transfected with pMHCII and CD86/CD4. Cells were combined at a 1:1 ratio in 96-well flat-bottom plates (total 100,000

cells/well). Culture supernatants were harvested after 18 hours, and IL-2 was measured by cytokine bead array (BD Pharmingen).

Serum Immunoglobulin Commensal Capture and Sequencing.

Mice were administered 2% dextran sulfate sodium (DSS, MP Biomedicals) dissolved in drinking water for 7 days while control mice were administered untreated drinking water. Subsequently, mice receiving DSS were given untreated drinking water for the next 7 days, while the control group remained on the untreated water. Mice were euthanized on day 14. Blood was collected from each mouse via cardiac puncture. Stool was collected from each mouse by flushing the colon with PBS. Serum was collected from each blood sample and pooled according to experimental group (DSS treatment or water). The stool was homogenized in PBS, pooled according to experimental treatment group, and filtered through 70µm filters. The remaining debris was removed from the stool suspension by centrifugation. Stool suspensions were incubated with serum, overnight at 4°C. After centrifugation, pellets isolated from the overnight incubations were incubated with protein A/G magnetic beads (Life Technologies) and PBS 0.5% BSA for 2hrs at 4°C. The beads were isolated from supernatant by use of a magnet and washed with PBS. DNA was isolated from bacteria bound to beads using a QIAamp DNA mini kit (Qiagen) and rLysozyme Solution (Millipore) while following the manufacturer's instructions for isolation of genomic DNA from Gram-positive bacteria. Amplification and 16S rRNA gene library preparation was performed as previously described^{54,55} and sequenced on the Illumina MiSeq platform.

Animals.

Mice were bred in specific-pathogen-free facilities at Massachusetts General Hospital (Boston, MA) prior to infection and transferred to biosafety level 2 (BL2) housing upon *L. monocytogenes* infection. All animal studies were conducted in compliance with ethical regulations and were approved by the Institutional Care and Use Committee (IACUC) at Massachusetts General Hospital. *Atg16l1^{f/f}* mice were generated as previously described⁴⁵ and bred with mice expressing *Cre* recombinase under the control of the CD11c promoter (CD11c-Cre) (Jackson Laboratories). All mice were maintained on food and water *ad libitum*, used between 8–12 weeks of age, and age- and sex-matched for each experiment.

Statistical analysis.

For data analysis, iTRAQ 4 ratios for the two biological replicates were filtered to retain only those deemed reproducible as described previously⁵⁶. Reproducibility was based on replicates being confined within the 95% limits of agreement of a Bland–Altman plot⁵⁷. Reproducible replicates were then subjected to a moderated *t*-test to assess statistical significance⁵⁸. To generate the I-A^b-binding motif, all mouse peptides were aligned by MultAlin⁵⁹ and visualized with WebLogo. The frequencies of each amino acid at each position of the 9-mer core were used to generate a position-weight matrix. I-A^b binding scores for *Listeria* peptides were produced from this matrix and represented as the product of amino acid frequencies for each position of the 9-mer core (Supplementary Table 1). Epitope mapping and protein domain overlay were performed with NCBI's Conserved Domain Database³¹. Subcellular localization of antigenic *Listeria* proteins was predicted by

PSORTb v3.0.2²⁹ and by COMPARTMENTS⁶⁰ for mouse proteins. Transmembrane topology was predicted with HMMTOP³⁰. Additional details can be found in the “Life Sciences Reporting Summary”.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank H. Vlamakis, T. Reimels, and I. Latorre for scientific input, J. Gracias for technical assistance, and P. Rogers for FACS. This work was supported by funding from The Leona M. and Harry B. Helmsley Charitable Trust; National Institutes of Health grants DK043351, AI109725, AT009708, and DK092405; and the Juvenile Diabetes Research Fund (JDRF) to R.J.X.

References

1. Babbitt BP, Allen PM, Matsueda G, Haber E & Unanue ER Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature* 317, 359–361 (1985). [PubMed: 3876513]
2. Stern LJ, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368, 215–221 (1994). [PubMed: 8145819]
3. Kim A & Sadegh-Nasseri S Determinants of immunodominance for CD4 T cells. *Curr. Opin. Immunol* 34, 9–15 (2015). [PubMed: 25576665]
4. Arunachalam B, Phan UT, Geuze HJ & Cresswell P Enzymatic reduction of disulfide bonds in lysosomes: characterization of a gamma-interferon-inducible lysosomal thiol reductase (GILT). *Proc. Natl. Acad. Sci. U. S. A* 97, 745–750 (2000). [PubMed: 10639150]
5. Hsieh CS, deRoos P, Honey K, Beers C & Rudensky AY A role for cathepsin L and cathepsin S in peptide generation for MHC class II presentation. *J. Immunol* 168, 2618–2625 (2002). [PubMed: 11884425]
6. Hsing LC & Rudensky AY The lysosomal cysteine proteases in MHC class II antigen presentation. *Immunol. Rev* 207, 229–241 (2005). [PubMed: 16181340]
7. Miyazaki T, et al. Mice lacking H2-M complexes, enigmatic elements of the MHC class II peptide-loading pathway. *Cell* 84, 531–541 (1996). [PubMed: 8598040]
8. Schulze MS & Wucherpfennig KW The mechanism of HLA-DM induced peptide exchange in the MHC class II antigen presentation pathway. *Curr. Opin. Immunol* 24, 105–111 (2012). [PubMed: 22138314]
9. Rudensky A, Preston-Hurlburt P, Hong SC, Barlow A & Janeway CA, Jr. Sequence analysis of peptides bound to MHC class II molecules. *Nature* 353, 622–627 (1991). [PubMed: 1656276]
10. Hunt DF, et al. Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* 256, 1817–1820 (1992). [PubMed: 1319610]
11. Chicz RM, et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 358, 764–768 (1992). [PubMed: 1380674]
12. Chicz RM, et al. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp. Med* 178, 27–47 (1993). [PubMed: 8315383]
13. Sette A, et al. Invariant chain peptides in most HLA-DR molecules of an antigen-processing mutant. *Science* 258, 1801–1804 (1992). [PubMed: 1465617]
14. Lippolis JD, et al. Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. *J. Immunol* 169, 5089–5097 (2002). [PubMed: 12391225]
15. Sofron A, Ritz D, Neri D & Fugmann T High-resolution analysis of the murine MHC class II immunopeptidome. *Eur. J. Immunol* 46, 319–328 (2016). [PubMed: 26495903]

16. Mommen GP, et al. Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. *Mol. Cell. Proteomics* 15, 1412–1423 (2016). [PubMed: 26764012]
17. Dongre AR, et al. In vivo MHC class II presentation of cytosolic proteins revealed by rapid automated tandem mass spectrometry and functional analyses. *Eur. J. Immunol* 31, 1485–1494 (2001). [PubMed: 11465105]
18. Depontieu FR, et al. Identification of tumor-associated, MHC class II-restricted phosphopeptides as targets for immunotherapy. *Proc. Natl. Acad. Sci. U. S. A* 106, 12073–12078 (2009). [PubMed: 19581576]
19. Suri A, Walters JJ, Rohrs HW, Gross ML & Unanue ER First signature of islet beta-cell-derived naturally processed peptides selected by diabetogenic class II MHC molecules. *J. Immunol* 180, 3849–3856 (2008). [PubMed: 18322192]
20. Seamons A, et al. Competition between two MHC binding registers in a single peptide processed from myelin basic protein influences tolerance and susceptibility to autoimmunity. *J. Exp. Med* 197, 1391–1397 (2003). [PubMed: 12756272]
21. Nelson CA, Roof RW, McCourt DW & Unanue ER Identification of the naturally processed form of hen egg white lysozyme bound to the murine major histocompatibility complex class II molecule I-Ak. *Proc. Natl. Acad. Sci. U. S. A* 89, 7380–7383 (1992). [PubMed: 1323833]
22. Brandwein SL, et al. Spontaneously colitic C3H/HeJBir mice demonstrate selective antibody reactivity to antigens of the enteric bacterial flora. *J. Immunol* 159, 44–52 (1997). [PubMed: 9200437]
23. Lodes MJ, et al. Bacterial flagellin is a dominant antigen in Crohn disease. *J. Clin. Invest* 113, 1296–1306 (2004). [PubMed: 15124021]
24. Cong Y, Feng T, Fujihashi K, Schoeb TR & Elson CO A dominant, coordinated T regulatory cell-IgA response to the intestinal microbiota. *Proc. Natl. Acad. Sci. U. S. A* 106, 19256–19261 (2009). [PubMed: 19889972]

References for Methods

25. Janeway CA, Jr., et al. Monoclonal antibodies specific for Ia glycoproteins raised by immunization with activated T cells: possible role of T cellbound Ia antigens as targets of immunoregulatory T cells. *J. Immunol* 132, 662–667 (1984). [PubMed: 6228596]
26. Andreatta M, Schafer-Nielsen C, Lund O, Buus S & Nielsen M NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS ONE* 6, e26781(2011). [PubMed: 22073191]
27. Zhu Y, Rudensky AY, Corper AL, Teyton L & Wilson IA Crystal structure of MHC class II I-Ab in complex with a human CLIP peptide: prediction of an I-Ab peptide-binding motif. *J. Mol. Biol* 326, 1157–1174 (2003). [PubMed: 12589760]
28. Liu X, et al. Alternate interactions define the binding of peptides to the MHC molecule IA(b). *Proc. Natl. Acad. Sci. U. S. A* 99, 8820–8825 (2002). [PubMed: 12084926]
29. Yu NY, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615 (2010). [PubMed: 20472543]
30. Tusnady GE & Simon I Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol* 283, 489–506 (1998). [PubMed: 9769220]
31. Marchler-Bauer A, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–226 (2015). [PubMed: 25414356]
32. Scallan E, et al. Foodborne illness acquired in the United States--major pathogens. *Emerg. Infect. Dis* 17, 7–15 (2011). [PubMed: 21192848]
33. Vita R, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43, D405–412 (2015). [PubMed: 25300482]

34. Goyette P, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet* 47, 172–179 (2015). [PubMed: 25559196]
35. Wang P, et al. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol* 4, e1000048(2008). [PubMed: 18389056]
36. Chatterjee SS, et al. Intracellular gene expression profile of *Listeria monocytogenes*. *Infect. Immun* 74, 1323–1338 (2006). [PubMed: 16428782]
37. Heng TS, Painter MW & Immunological Genome Project, C. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol* 9, 1091–1094 (2008). [PubMed: 18800157]
38. Weber KS, et al. Distinct CD4+ helper T cells involved in primary and secondary responses to infection. *Proc. Natl. Acad. Sci. U. S. A* 109, 9511–9516 (2012). [PubMed: 22645349]
39. Palm NW, de Zoete MR & Flavell RA Immune-microbiota interactions in health and disease. *Clin. Immunol* 159, 122–127 (2015). [PubMed: 26141651]
40. Hall AB, Tolonen AC & Xavier RJ Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet* 18, 690–699 (2017). [PubMed: 28824167]
41. Ormerod KL, et al. Genomic characterization of the uncultured Bacteroidales family S24–7 inhabiting the guts of homeothermic animals. *Microbiome* 4, 36(2016). [PubMed: 27388460]
42. Zeng MY, et al. Gut Microbiota-Induced Immunoglobulin G Controls Systemic Infection by Symbiotic Bacteria and Pathogens. *Immunity* 44, 647–658 (2016). [PubMed: 26944199]
43. Christmann BS, et al. Human seroreactivity to gut microbiota antigens. *J. Allergy Clin. Immunol* 136, 1378–1386 e1371–1375 (2015). [PubMed: 26014812]
44. Stoll ML, et al. Altered microbiota associated with abnormal humoral immune responses to commensal organisms in enthesitis-related arthritis. *Arthritis Res. Ther* 16, 486(2014). [PubMed: 25434931]
45. Conway KL, et al. Atg1611 is required for autophagy in intestinal epithelial cells and protection of mice from *Salmonella* infection. *Gastroenterology* 145, 1347–1357 (2013). [PubMed: 23973919]
46. Rappsilber J, Mann M & Ishihama Y Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc* 2, 1896–1906 (2007). [PubMed: 17703201]
47. Mertins P, et al. Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics* 13, 1690–1704 (2014). [PubMed: 24719451]
48. Eddy SR Accelerated Profile HMM Searches. *PLoS Comput. Biol* 7, e1002195(2011). [PubMed: 22039361]
49. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–285 (2016). [PubMed: 26673716]
50. Alipanahi B, Delong A, Weirauch MT & Frey BJ Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol* 33, 831–838 (2015). [PubMed: 26213851]
51. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009). [PubMed: 19451168]
52. McDavid A, et al. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Comput. Biol* 10, e1003696(2014). [PubMed: 25032992]
53. Xu C & Su Z Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980 (2015). [PubMed: 25805722]
54. Caporaso JG, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A* 108 **Suppl** 1, 4516–4522 (2011). [PubMed: 20534432]
55. Caporaso JG, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624 (2012). [PubMed: 22402401]
56. Kronke J, et al. Lenalidomide induces ubiquitination and degradation of CK1alpha in del(5q) MDS. *Nature* 523, 183–188 (2015). [PubMed: 26131937]

57. Bland JM & Altman DG Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310 (1986). [PubMed: 2868172]
58. Smyth GK Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol* 3, Article3 (2004).
59. Corpet F Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16, 10881–10890 (1988). [PubMed: 2849754]
60. Binder JX, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014, bau012 (2014).

Figure 1. MHCII peptidomics in primary murine dendritic cells results in more than 3,700 distinct peptide identifications and defines the I-A^b-binding motif.

(A) Experimental workflow for immunopurification and sequencing of MHCII-associated peptides from murine dendritic cells. MHCII-peptide complexes were immunopurified from WT and *Atg16l1*^{-/-} cells. Associated peptides were then acid-eluted, labeled with iTRAQ4 reagents, desalted with SCX and C18, and analyzed using high resolution LC-MS/MS. (B) A database search strategy for MHCII-peptide sequencing. All MS/MS spectra were searched against a database containing mouse proteins using Spectrum Mill software with a “no enzyme” specificity. Mouse peptides were validated using a 1% FDR cutoff, and the total numbers of peptides quantified across all samples were reported. (C) The I-Ab-binding motif was derived from endogenous mouse peptides bound to MHCII. Heatmap color coding represents the frequencies of each amino acid at each respective position.

Figure 2. Antigen processing pathways and epitope features revealed by MHCII peptidomics. (A) Deficiency in the autophagy protein Atg16l1 skews the spectrum of MHCII-associated peptides. MHCII-bound peptides quantified in *Atg16l1*^{-/-} dendritic cells relative to wild type (WT). Replicate (rep) samples were compared based on log2 fold change (FC) between mouse strains. Each dot represents a unique peptide sequence. Peptides that were observed to be significantly upregulated or downregulated are shown in red, while peptide measurements that were not reproducible across both biological replicates are shown in cyan. Dot plot axes: Log2FC. Histogram axes: number of distinct peptides. n=2 biologically independent samples per genotype in a single experiment. Reproducible replicates (95% limits of agreement of a Bland–Altman plot) were subjected to a moderated *t*-test to assess

statistical significance. **(B)** Abundance and subcellular sources of MHCII-associated peptides derived from *Atg611*^{-/-} and WT dendritic cells. **(C)** Epitope mapping relative to domain structure of endogenous antigens indicates preferential presentation of epitopes derived from the luminal/extracellular domains of transmembrane proteins and epitopes positioned between structurally-defined domains. **(D)** Immunodominant epitope prediction with BOTA. Workflow of BOTA with input as genome and output as a binding score. The upper panel shows the extraction of candidate peptides, and the lower panel shows the deep neural network core of the BOTA algorithm to assign a binding score to each candidate peptide. To extract candidate peptides, predicted genes from input genome are processed by HMMTOP, pfam domain search using HMMER3.0, and PSORT to define various features, which are later integrated, and candidate peptides are selected based on criteria previously described. To encode each candidate peptide with length l , amino acids are first encoded using a b -bit vector at random, and thus, the total amino acid space could be presented by a $20 \times b$ binary matrix. During the encoding process, a window of length k slides through the input peptide, thus forming $l-k+1$ windows for an input peptide with length l . This process is repeated for all d detectors, forming a d -HMM model scoring matrix X . Matrix X will then go through the regular rectify-max pool-neural network prediction route to generate a final output score, f , for each input peptide.

Figure 3. Validation of BOTA epitope predictions with MHCII peptidomics.

(A) Experimental workflow for immunopurification and sequencing of MHCII-associated peptides from murine dendritic cells. MHCII-peptide complexes were immunopurified from WT cells after a 10-minute or 6-hour *Listeria* treatment. Associated peptides were then acid-eluted, labeled with iTRAQ4 reagents, desalted with SCX and C18, and analyzed using high resolution LC-MS/MS. (B) MHCII-bound peptides (mouse and *Listeria*) detected before and after *Listeria* exposure. Biological replicates (rep) were compared based on log2 fold change (FC) between time 10 min and 6 hr after exposure to bacteria. Each dot represents a unique peptide sequence. Peptides that were observed to be significantly upregulated or

downregulated are shown in red, while peptide measurements that were not reproducible across both biological replicates are shown in cyan. Dot plot axes: Log2FC. Histogram axes: number of distinct peptides. n=2 biologically independent samples per treatment in a single experiment. Reproducible replicates (95% limits of agreement of a Bland–Altman plot) were subjected to a moderated *t*-test to assess statistical significance. (C) Predictions for *Listeria* epitopes were made using the deep neural network core of BOTA. (D) The BOTA model pre-training accuracy plateaus after 200 epochs in cross-validation. The model was trained using the mouse peptides captured in BMDCs infected by *Listeria* (blue line); in contrast, the same model trained solely on Immune Epitope Database (IEDB) data reached a plateau at approximately 70%, signifying a 15% gap in accuracy. (E) Comparison of predictions for *Listeria* epitopes in proteins identified by proteomics. Peptides are split into categories based on the protein's subcellular localization using PSORTb. (F) BOTA was used to predict epitopes for the human IBD risk allele (HLA-DRB1*01:03) that were annotated in IEDB as validated by MHCII binding assays or T cell reactivity assays. Full-length protein sequences associated with these epitopes were used as input for BOTA upstream modules. The output of BOTA upstream modules (trimmed protein sequences) was used as input for epitope prediction using netMHCIIpan with HLA-DRB1*01:03 specificity. This approach yielded 76% accuracy (Venn diagram BOTA predictions overlapping with validated IEDB epitopes) for predicting IEDB validated epitopes, with 12 epitopes missed by BOTA. In contrast netMHCIIpan predictions on full-length proteins without using BOTA upstream modules performed with 22% accuracy (Venn diagram netMHCIIpan predictions overlapping with validated IEDB epitopes), while 39 epitopes were missed by NetMHCIIpan.

Figure 4. BOTA and MHCII peptidomics accurately predict immunodominance *in vivo*. (A) Epitope mapping and domain structure of *Listeria* antigens indicate preferential presentation of surface-exposed and secreted proteins. (B) Immunodominance of epitopes was determined by infecting mice with *Listeria*. At day 7, splenocytes were harvested and restimulated with the indicated peptides for quantification of the T cell response by IFN γ ELISPOT. Data represent the mean number of spots per 1×10^5 CD4 T cells $\pm$ sd for $n = 6$ mice. (C, D and E) Immunodominance *in vivo* (IFN γ ELISPOT data from Figure 4B) correlates with fold change (FC) of *Listeria* peptides quantified by MHCII peptidomics (data from Figure 3B), and to a lesser extent, with mRNA expression of the corresponding peptide-encoding genes in *Listeria* derived from infected macrophages (previously published microarray³⁶).

Figure 5. Single cell RNAseq integrates T cell phenotype with TCR repertoire in the *Listeria* response.

Mice were inoculated with *Listeria* by i.p. injection on days 0 and 11. On day 18, FSCHICD4+CD8-B220-MHCII- T cells were FACS-sorted from spleens for single cell RNAseq and TCRseq. Sorted T cells were derived from 2 mice, sequenced separately, and combined for analysis. **(A)** Violin plots displaying Teff signature score derived from ImmGen. Each dot represents a single cell classified by clusters defined by tSNE. n=1920 cells from 2 mice. Limits of violin plot capture minima and maxima. **(B)** tSNE plot derived from T cell transcriptomes identifies distinct cell states that cluster according to unique signatures. Each dot represents a single cell that is color coded according to gene signature. n=1920 cells from 2 mice. **(C)** Circos plots of the linkages between the TCR alpha chain CDR3 and TCR beta chain CDR3. Ribbons link two chains with thickness proportional to the number of corresponding TCR pairs observed. Dominant TCR clones are labeled

according to TCR gene segment usage. **(D)** Specifying antigen-reactivity by screening TCRs for reactivity with *Listeria* epitopes predicted by BOTA. HEK 293T cells were transfected to express peptide epitopes fused in-frame with the I-Abbeta chain bearing CD3zeta cytoplasmic domains. BW5147_CD4–28 cells were transduced to express chimeric single chain TCRs bearing the transmembrane and cytoplasmic domains of CD3zeta. In this coculture system, cognate antigen recognition results in T cell activation characterized by production of IL-2. **(E)** The most abundant TCR identified in mice infected with *Listeria* (lmo_R6) was screened for reactivity against *Listeria* epitopes as described above. IL-2 was detected in culture supernatant by cytometric bead array. As controls, OT2 TCR (reactive with Ova) and LLO_118 TCR (reactive with LLO) were included. **(F)** HEK 293T cells were transfected with constructs encoding single chain TCRs. Cells were analyzed by FACS for expression of TCRb and binding to LLO I-Ab tetramers (NEKYAQAYPNVS I-Ab). Data represent a single experiment.

Figure 6. Computational prediction and validation of a dominant commensal antigen.

(A) Mice were administered DSS to induce colitis prior to analysis by SICC-seq. At day 14, serum was collected and incubated with stool to allow binding of IgG with commensals. IgG-positive and -negative fractions were separated with magnetic beads covalently attached to protein A/G. The immunogenicity of *Bacteroidales* was demonstrated by IgG-reactivity score (relative abundance in IgG-positive minus IgG-negative fractions) derived from 16S sequencing. (B) BOTA identified SusC, a highly represented epitope within and across *Bacteroidales* genomes, including the murine commensal *Muribaculum intestinale*. Splenocytes from naive mice were harvested and stimulated *in vitro* with SusC peptide or DMSO control. Cytokines were measured 24 hours later by cytometric bead array. Data represent the mean cytokine concentration $\pm$ sd for $n = 7$ mice. ** $P < 0.0001$ as determined by unpaired two-tailed Student's T test.