Supplemental Experimental Procedures:

Derivation of the sampling density. Based on Hoof et al. (1), we define the presentation density (D) of each protein sequence as $D = K/(L-8)$, where K is a weighted count of HLA-I presented peptides mapped to the protein in question, L is the length of the protein and L-8 is therefore the number of potential 9-mer peptides. If $X = \{x_1, x_2, \ldots, x_H\}$ is the set of H presented HLA-I binding peptides (x_i) in protein sequence S and N(x) is the number of expressed protein sequences sharing peptide x then we define K(S) as:

$$K(S) = \sum_{x \in X} P(x|N(x))$$

The term P(x|N(x)) is the probability of cleaving peptide x from protein sequence S. We then constructed a simple ad-hoc weighting-scheme that adds a contribution to the baseline cleavage probability (q) as a function of N:

$$P(x|N) = q + (1-q)^N$$

The a priory expectation is that about 1/5 peptides that can be generated from a protein sequence will be generated via the proteasome (2). Thus, q=0.20. The posterior probability of cleaving x from a single sequence S from N possible sequences is P(x|N) and the posterior number of expected cleavage events knowing x is observed is therefore P(x|N)*N. The purpose of the ad-hoc estimate of the posterior cleavage probability is to ensure that P(x|N)*N>=1 for any N to account for x being observed. If N=1 then P(x|1)=1 i.e. the weighting-scheme automatically counts uniquely assigned peptides with both a weight and a count of 1 since P(x|1) = q + (1-q)^1 = 1. The table

below shows the effect of P(x|N) as N grows. The asymptotic value of P(x|N) is of course q as N approaches infinity because then the a priory expectation is to observe x at least one time already.

N	(1-q)^N	P(x|N)
1	0.80000	1.00000
2	0.64000	0.84000
3	0.51200	0.71200
4	0.40960	0.60960
5	0.32768	0.52768
6	0.262144	0.462144
7	0.209715	0.409715
8	0.167772	0.367772
9	0.134218	0.334218
10	0.107374	0.307374
20	0.011529	0.211529
30	0.001238	0.201238
40	0.000133	0.200133
50	0.000014	0.200014

The effect of P(x|N) as N grows.

1. Hoof, I., van Baarle, D., Hildebrand, W. H., and Kesmir, C. (2012) Proteome sampling by the HLA class I antigen processing pathway. *PLoS computational biology* 8, e1002517.
2. Yewdell, J. W., and Bennink, J. R. (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annual review of immunology* 17, 51-88.