Molecular Detection of Targeted MHC I-bound Peptides using a Probabilistic Measure and Nanospray MS^3 on a Hybrid Quadrupole-LIT

Supporting Information

Bruce Reinhold[1,2], Derin B. Keskin[2], and Ellis L. Reinherz[1,2,]*

[1]Cancer Vaccine Center and [2]Laboratory of Immunobiology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

Contents:

Random processes, relative entropy, probabilistic distance and chemical noise. The following is an expanded discussion of the material in the theoretical section of the manuscript. The objective here is to connect the probabilistic structure underlying MS measurement and the detection problem introduced by substantial chemical noise. The presentation will be divided into two parts. First the multinomial and Poisson models of random process sampling on a finite outcome space will be stated and their relation to the relative entropy and the theory of large deviations shown. Second, the appearance of chemical noise as events that cannot be identified with an *a priori* underlying distribution will be structured as a restriction of the probability measure on the outcome space. The overall goal will be to present MS detection with chemical noise as a straightforward calculation of probability with well-established roots in information theory**.**

A spectrum is represented as a finite sampling of an independent, identically distributed (in sequence) random process defined on a finite outcome space D. The elements of D correspond to the m/z data points established by the acquisition hardware. Let O(D) be the number of elements in D. The random process on D is asymptotically characterized by the empirical measure or probability distribution $\vec{\mathrm{P}}=\left(p_1,p_2,...,p_{O(D)}\right)$ where $\sum_{j=1}^{O(D)} p_j = 1$. Sampling of this process can be embedded in physical time in two different ways leading to either a multinomial or Poisson model of the underlying random process. For the multinomial model, consider N sequential ion events acquired in a sampled spectrum. If the ion counts in the data channels are $\left\{n_1,n_2,...,n_{O(D)}\right\}:\sum_{j=1}^{O(D)} n_j = N$ and the underlying probability for an event in the jth channel is p_{j}, then the probability of observing the spectrum $\left\{n_1,n_2,...,n_{O(D)}\right\}$ is given by the multinomial distribution function

$$\mathrm{P}\left\{n_1,n_2,...,n_{O(D)}\right\} = \frac{N!}{n_1!n_2!\cdots n_{O(D)}!} p_1^{n_1} p_2^{n_2} \cdots p_{O(D)}^{n_{O(D)}} \qquad \text{(S1)}$$

As a Poisson model, consider each data channel j in D to be associated with an ion arrival rate α_j. This rate describes the number of ions arriving per unit time and is formally defined by $\lim_{\Delta T \to 0} \frac{1}{\Delta T} P(\Delta T) = \alpha_j$. The arrival rates α_j can be related to a normalized distribution $\vec{P}=(p_1, p_2, ..., p_{O(D)})$ on D by introducing a unit arrival period T such that $\sum_{j=1}^{O(D)} T\alpha_j = 1$ or $p_j = T\alpha_j$. To relate the Poisson model to a measured spectrum, suppose the random process is sampled for a period NT and one counts the event distribution $\{n_1, n_2, ..., n_{O(D)}\}$. For a Poisson process on a single channel characterized by an arrival rate α and a sampling period NT, the probability of measuring n events is $P(n) = \frac{1}{n!}(\alpha NT)^n e^{-\alpha NT}$. As event arrivals are independent, and translating arrival rates into underlying probabilities by $p_j = T\alpha_j$, the probability of the event distribution $\{n_1, n_2, ..., n_{O(D)}\}$ after an NT sampling period is given by

$$P\{n_1, n_2, ..., n_{O(D)}\} = \prod_{j=1}^{O(D)} \frac{1}{n_j!}(Np_j)^{n_j} e^{-Np_j} \qquad \text{(S2)}$$

The multinomial and Poisson models of the underlying random process do not describe the same situation. Although both models describe a sampling error, that is, a probability $P(\vec{n} \square \vec{p})$ for measuring an empirical outcome $\vec{n} = \{n_1, n_2, ..., n_{O(D)}\}$ in a random process that has an underlying probability distribution $\vec{p}=(p_1, p_2, ..., p_{O(D)})$, for the Poisson model, the event sum $\sum_{j=1}^{O(D)} n_j = N$ after a time period NT is not fixed while in the multinomial model it is the time period for N events to be sampled that is not specified. The two models can be formally connected by taking the coupled limit of N going to infinity and p going to zero such that Np is finite (the arrival rate). The Poisson or finite period sampling more directly reflects the experimental acquisition of spectra and is the form used in our data analyses but as

far as we understand either model could be used and would give essentially the same results. For the multinomial model (S1) we have

$$\ln\left[\mathrm{P}\left(\vec{n}\,\square\,\vec{p}\right)\right]=\ln N!\;-\sum_{j}^{O(D)}\ln\left(n_j!\right)\;+\sum_{j}^{O(D)}n_j\ln\left(p_j\right)$$

Applying Stirling's approximation for the factorial $\mathrm{n!}=\sqrt{2\pi\mathrm{n}}\;\mathrm{n}^{\mathrm{n}}\;e^{-\mathrm{n}}$ and introducing the normalized measure $\vec{\nu}$ by $N\vec{\nu}=\left(n_1,n_2,\cdots,n_{O(D)}\right)$ we have

$$\ln\left[\mathrm{P}\left(\mathrm{N}\vec{\nu}\,\square\,\vec{p}\right)\right]=\frac{\ln\left(2\pi N\right)\left(1-O(D)\right)}{2}\;-\frac{1}{2}\sum_{j}^{O(D)}\ln\left(\nu_j\right)\;-\;N\sum_{j}^{O(D)}\nu_j\ln\left(\frac{\nu_j}{p_j}\right)\tag{S3}$$

For the Poisson distribution (2) a similar calculation gives

$$\ln\left[\mathrm{P}\left(N\vec{\nu}\,\square\,\vec{p}\right)\right]=-\frac{O(D)}{2}\ln\left(2\pi N\right)-\frac{1}{2}\sum_{j=1}^{O(D)}\ln\left(\nu_j\right)-N\sum_{j=1}^{O(D)}\nu_j\ln\left(\frac{\nu_j}{p_j}\right)\tag{S4}$$

Identifying the relative entropy by

$$\mathrm{I}_2\left(\vec{\nu},\vec{p}\right)=\sum_{j=1}^{O(D)}\nu_j\ln\frac{\nu_j}{p_j}\tag{S5}$$

the results (S3) and (S4) show by elementary algebra that for the Poisson and multinomial models the probability of empirically measuring a distribution $\vec{\nu}$, when the underlying distribution is $\vec{p}$, asymptotically decays with an exponential rate given by the relative entropy: $\mathrm{P}\left(\vec{\nu}\,\square\,\vec{p}\right)=Ae^{-N\mathrm{I}_2\left(\vec{\nu},\vec{p}\right)}$. For MS detection, given a reference spectrum and a measurement of that spectrum (without chemical noise), the probabilistic 'difference' is asymptotically an exponentially decaying function of the number of events times the relative entropy.

In a practical setting calculating the probabilistic distance between a measured and reference spectrum is not an effective detection algorithm. Chemical noise as ion fragments from co-selected molecular components are expected in the MS^n spectra of complex mixtures, but any data channel where finite

chemical background events are measured and where reference events are not expected (specifically where $\nu_j >> p_j$) results in a high entropy cost in (S5) and a low probability in (S3) or (S4). This is not wanted; reflecting the expectation of chemical noise, low detection probabilities are to be associated only with $\nu_j << p_j$. The strategy will be to shift perspective, but only slightly. Instead of representing the measured events as a finite sampling of the reference distribution we will consider detection by calculating the probability that *M* reference events are contained in the measured events and focus on the decrease in probability as *M* increases. Restated in the context of MS data, some peaks in the measured spectrum will limit the amount of the target that could be present; with the limiting amount of target assumed, other peaks in the measured spectrum may have too many events and these additional events will be identified as chemical noise and not due to the target.

As previously defined, the measured spectrum is $N\vec{\nu}$ and the reference distribution of the target is $\vec{p}$. Let *M* be an integer representing a potential number of target events, i.e., events distributed by $\vec{p}$. For fixed $N\vec{\nu}$ and $\vec{p}$, every *M* splits the outcome space D into D_M (on which $N\nu_j < Mp_j$) and its complement ($D\text{-}D_M$). Events in the complement correspond to channels with too many events for *M* total reference events. That is, they are peaks obscured by chemical noise and are ignored. For fixed $\vec{n} = N\vec{\nu}$ and $\vec{p}$, the probability P_M strictly decreases as *M* increases. The multinomial (S1) and Poisson models (S2) are exactly the same except they are evaluated over D_M only, e.g., for the Poisson model

$$\ln\left[P_M\left(\vec{n}\,\square\,\vec{p}\right)\right] = \sum_{j\in D_M}\left(-\ln\left(n_j!\right) + n_j \ln N + n_j \ln p_j - Np_j\right)$$

Following the same algebra as in the derivation of (S3) and (S4) we have an expression similar to (S5): $N\sum_{j\in D_M}\left(\nu_j \ln\left(\frac{\nu_j}{p_j}\right)\right)$. This expression is not immediately an entropy function since $\vec{\nu}$ and $\vec{p}$ are not normalized over the subset D_M. But one can introduce an *N*-dependent factor in the exponential

convergence by renormalizing the measures over the restricted domains. Explicitly, define α such that $\nu_j = \alpha \nu_j^\circ$ and $\sum_{j \in D_M} \nu_j^\circ = 1$ (likewise for β) so that the restricted sum has a proper form for a relative entropy:

$$N \sum_{j \in D_M} \left(\alpha \nu_j^\circ \ln\left(\frac{\alpha \nu_j^\circ}{\beta p_j^\circ} \right) \right) = N\alpha \sum_{j \in D_M} \left(\nu_j^\circ \ln\left(\frac{\nu_j^\circ}{p_j^\circ} \right) \right) + N\alpha \ln\left(\frac{\alpha}{\beta} \right). \quad \text{(S6)}$$

Appropriately, as *M* increases, D_M converges back to D, α and β converge to 1 and the restricted relative entropy (S6) returns to (S5).

Detection Confidence Score. Detection confidence in a single $\mathsf{M}_{\mathsf{P}_0}(\tau)$ plot is related to ratios between the number of events that can be associated with the reference distribution in its original position, i.e., $\mathsf{M}_{\mathsf{P}_0}(0)$, and the number of events that can be associated with the reference distribution when it is translated by τ, $\mathsf{M}_{\mathsf{P}_0}(\tau)$. A reasonable score parameter is the ratio of the '0-offset' peak $\mathsf{M}_{\mathsf{P}_0}(0)$ to some measure of the fluctuation in the translated peaks. Specifically, let k be a set of discrete +/- 1 amu translations of the reference pattern, e.g., discrete translations from -50 to +50 amu, calculate an averaged event number $\bar{\mathsf{M}}_{\mathsf{P}_0} = \frac{1}{101} \sum_{k=-50}^{50} \mathsf{M}_{\mathsf{P}_0}(k)$ and subtract this from the 0-offset peak. This difference is then normalized by the averaged fluctuation (Equation S7).

$$s = \left(\mathsf{M}_{\mathsf{P}_0}(0) - \bar{\mathsf{M}}_{\mathsf{P}_0} \right) \Big/ \sqrt{\frac{1}{101} \sum_{k \in S_M} \left(\mathsf{M}_{\mathsf{P}_0}(k) - \bar{\mathsf{M}}_{\mathsf{P}_0} \right)^2} \quad \text{(S7)}$$

Q-TOF MS spectrum of HLA-A2 peptides from BEAS-2B cells.

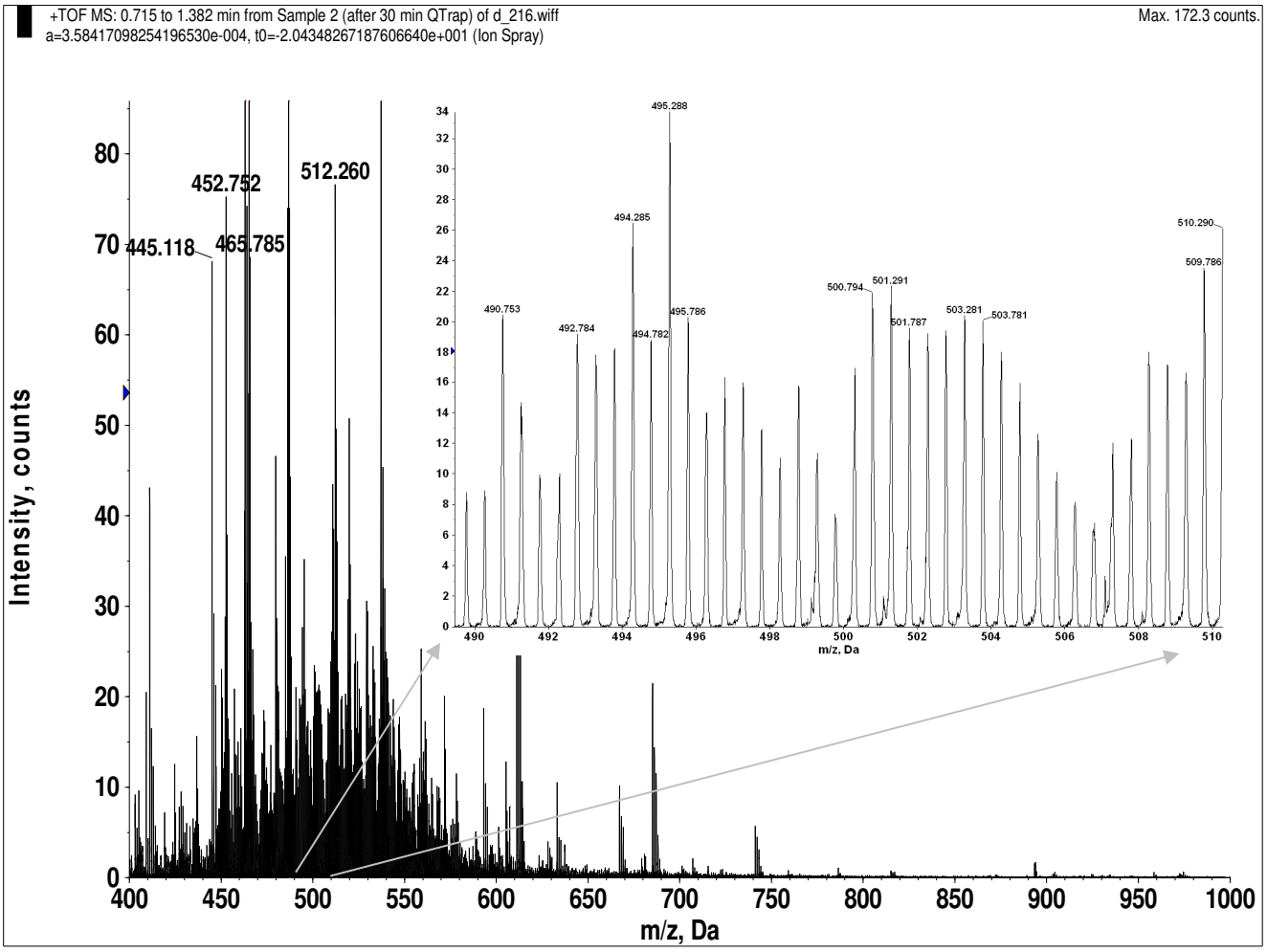

Figure S1. Q-TOF MS spectrum of peptides recovered from immunoaffinity purified HLA-A2 complexes from 10^7 infected BEAS-2B cells. Inset shows a segment of the spectrum with an expanded m/z scale to illustrate the uniform double charge signature reflecting the length restriction (8, 9 and 10mers) associated with MHC I binding.

Numerical generation of sampled spectra from arrival rate arrays. A number of circumstances arise in MS^3 Poisson detection where one would like to scale the intensity of a reference spectrum to represent a changed ion flux, sample concentration or collection period, and add this new arrival rate pattern on top of an existing spectrum. However, one cannot simply add point by point the data array containing the scaled reference spectrum to the data array of the ion background spectrum. Such an exercise ignores sampling fluctuations (shot noise), and generates fine detail in the detection plot which would not be observed in actual data. Scaling the reference spectrum generates an m/z-dependent array of real numbers describing the event arrival rate α per collection period T, i.e., the product αT. In order to generate a properly sampled instance spectrum, for each m/z data point an integer-valued event number n must be randomly drawn from a Poisson distribution $P(n) = \frac{1}{n!}(\alpha T)^n e^{-\alpha T}$ where $\alpha T \doteq \alpha T$ (m/z) is the scaled reference spectrum (Figure S2). Numerically generated spectra containing scaled reference patterns yet realistic shot noise and ion backgrounds can be applied to limit of detection estimates, receiver operating characteristic (ROC) curves and confidence techniques. To add target to background spectra a few steps are required. First, and trivially, the background spectrum must be m/z translated to correspond to the selected m/z of the target. In these studies ion background is due to other co-selected peptides and the neutral losses in this spectrum reflect peptide dissociation chemistry and this should be preserved as typical background features. Another requirement is to determine the MS^3 signal that corresponds to a given target concentration in the sample. For this a known amount of target is added to an MHC I workup (with a carrier system) and one measures the corresponding ion flux in the MS^3 spectrum. The reference spectrum measured at one concentration and collection period is then scaled to match the new concentration and the collection period of the MS^3 spectrum used for the ion background. Once a sampled instance of the scaled reference distribution spectrum has been generated, it can be

added to the translated MS^3 ion background and the summed spectrum analyzed by the probabilistic metric.

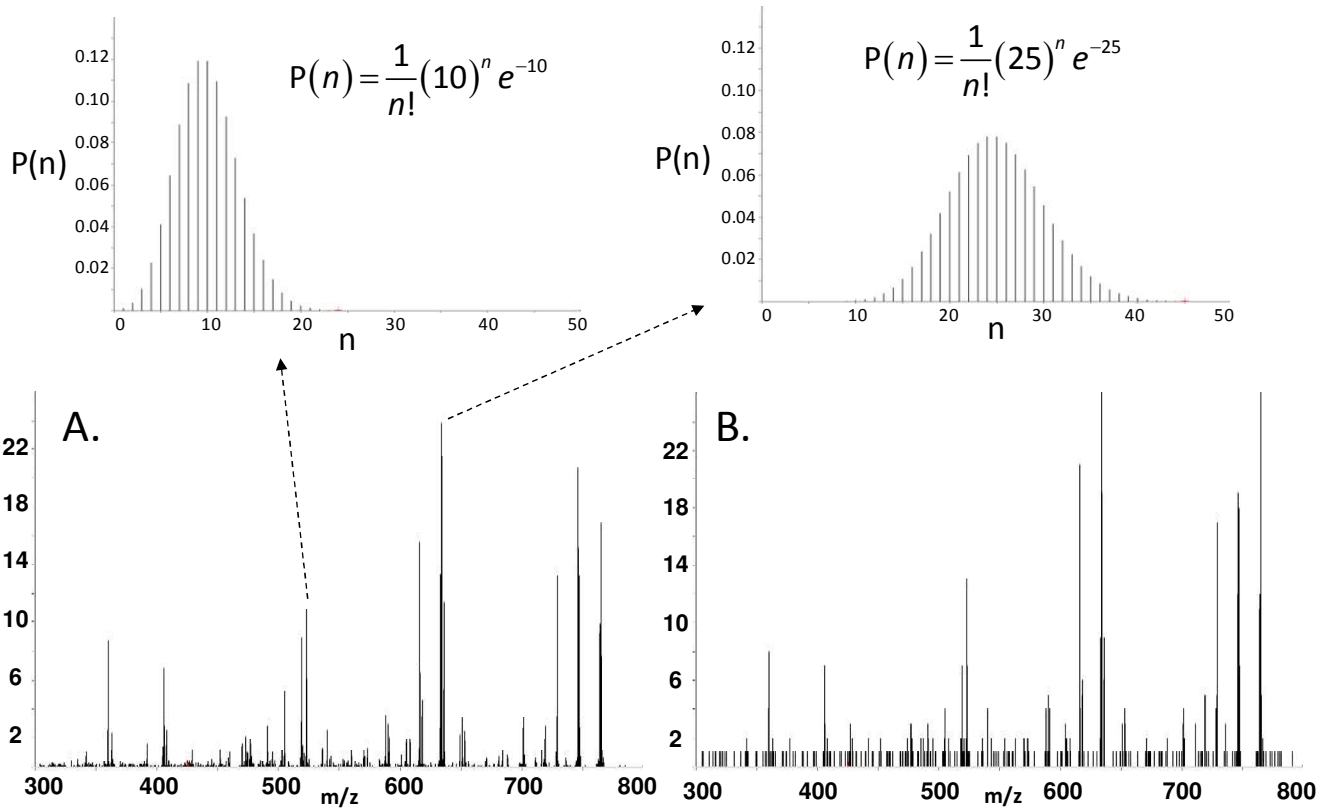

Figure S2. Generating sampled spectra from a reference spectrum. A) The $E7_{11\text{-}19}$ reference spectrum from a 9 fmol/µl sample collected for 2 minutes is scaled to describe a 10 amol/µl concentration collected for 50 minutes. To capture sampling fluctuations or shot noise the Poisson distribution function $P(n) = \frac{1}{n!}(\alpha T)^n e^{-\alpha T}$ for each arrival rate must be sampled for the event count. Distribution functions are shown as insets for αT = 10 and 25 events. B) An instance of the generated spectrum with shot noise from sampling the scaled reference spectrum.

Multiple MS3 detection spectra. In collecting reference patterns for a single target, the different MS^3 reference spectra are acquired using alternating scans so that relative ion fluxes among the reference MS^3 spectra are determined. If one can subsequently determine the level of one MS^3 reference pattern in a sample containing the target then from the reference collection one knows the levels at which other MS^3 reference patterns should be observed. This information can be used to identify false positives. For example, Poisson detection of the dissociation pattern of the y_7 fragment from the peptide GILGFVFTL (MS^3 483.79 : 796.46) against a background of total HLA-A2 peptides from 10 million T1 cells generated a score (Eqn. S7) of 6.1 at 74 events ($M_{P_0}(0)$ or 0-offset amplitude) while the evidence for the b_6 fragment of GILGFVFTL (MS^3 483.79 : 587.36) is weaker with a score of 3.6 and 46 events (Figure S3).

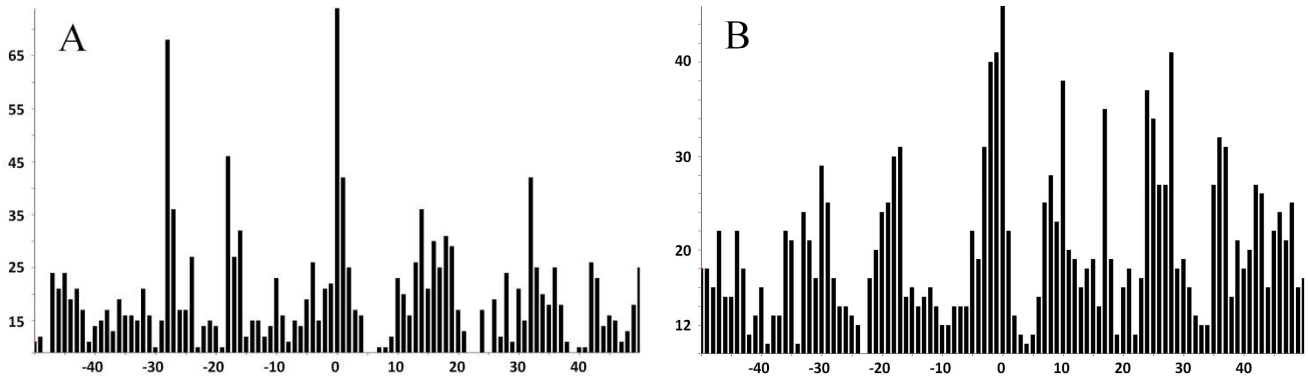

Figure S3. Ambiguous Poisson plots for detecting influenza M1 peptide GILGFVFTL from the y_7 fragment LGFVFTL (MS^3 483.79 : 796.46, A) and the b_6 fragment GILGFV (MS^3 483.79 : 587.3, B) in the HLA-A2 peptides extracted from 10 million T1 cells. The 0-offset amplitude in panel A (74 events) relative to the translated event amplitudes generates a score of 6.1 (Eqn. S7) which could suggest detection while the 0-offset amplitude in B (46 events) generates a lower score of 3.6.

MS^3 analysis using alternating scans with the synthetic peptide produced an event ratio of 0.6 for the MS^3 spectra of the y_7 to b_6 fragments. That is, Poisson fits of the y_7 reference spectrum in MS^3 483.79 :

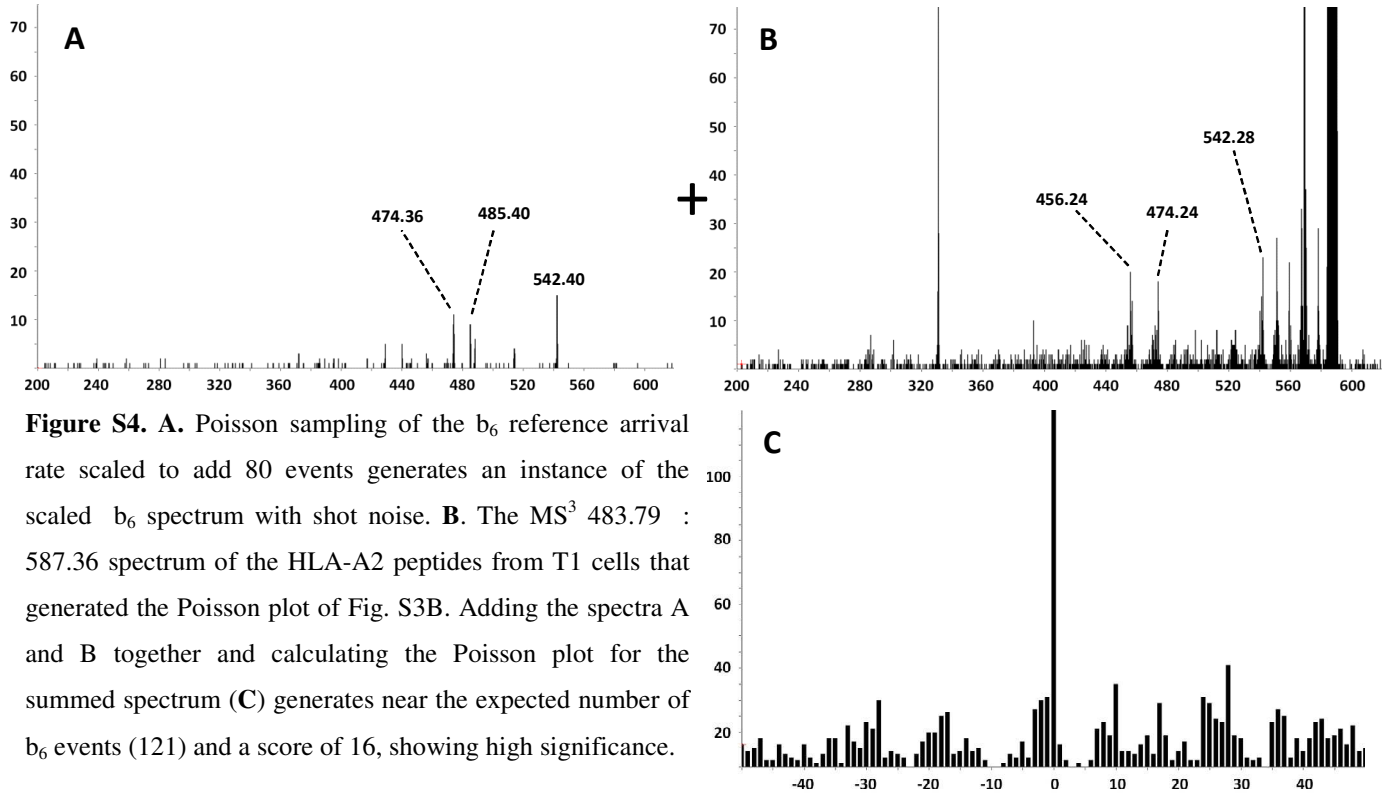

Figure S4. A. Poisson sampling of the b_6 reference arrival rate scaled to add 80 events generates an instance of the scaled b_6 spectrum with shot noise. **B**. The MS^3 483.79 : 587.36 spectrum of the HLA-A2 peptides from T1 cells that generated the Poisson plot of Fig. S3B. Adding the spectra A and B together and calculating the Poisson plot for the summed spectrum (**C**) generates near the expected number of b_6 events (121) and a score of 16, showing high significance.

796.46 spectra of the synthetic peptide generated 0.6N events as a 0-offset amplitude while Poisson fits of the b_6 reference spectrum in MS^3 483.79 : 587.36 spectra of the peptide generated N events as a 0-offset amplitude. Hence if the Poisson fit of the y_7 reference in the T1 sample (Fig. S3A) is a true positive, one would expect 123 events (0-offset) for the Poisson fit of the b_6 reference pattern in the MS^3 483.79 : 587.36 spectrum. However only 46 events are measured (Fig. S3B), indicating the detection of the Fig. 3A is false. If one adds 80 MS^3 483.79 : 587.36 b_6 reference events to the T1 MS^3 spectrum and generates a new Poisson plot (Fig. S4C) with near the expected 123 0-offset amplitude, the score of 16 indicates a high significance. Essentially, if the M1 peptide was present as suggested by the weak detection signature of its y_7 fragment, it would have been easily detected by its b_6 fragment. It was not and so the y_7 signature is identified as a false positive.

Quantitation with negative control: T1 cells loaded with influenza M1 peptide GILGFVFTL. To estimate the number of copies per cell when T1 cells were loaded by incubation with a dilute peptide solution, three samples were prepared: 5 million unloaded T1 cells, 5 million unloaded T1 cells with 829 amol (100 peptide copies/cell) of GILGFVFTL added to the affinity beads containing the HLA-A2 complexes immediately after adding 17.5 μl 10% acetic acid, and 5 million T1 cells loaded with GILGFVFTL peptide at a concentration of 62 pg/ml. For each sample, MS^3 spectra 483.8 : 587.4 (b_6 fragmentation) and 483.8 : 796.5 (y_7) were acquired in alternating sequence. The optimal sensitivity is to extract sample loading from the same spectrum that is used for target quantitation. Again, a fundamental strength of Poisson fitting is the quantitation stability in the presence of overlapping. The unloaded T1 cell sample produced a weak ion background for the MS^3 483.8 : 587.4 spectrum (good for detection, but not quantitation), so the MS^3 483.8 : 796.5 spectrum was used to generate a T1 background reference and the sample loads (ion flux times collection period) for the MS^3 spectra of the other samples were normalized by Poisson fit to this reference pattern. The Poisson fits to both the T1 background spectrum and the GILGFVFTL y_7 reference for the 5 million T1 cells with 829 amols GILGFVFTL peptide added are shown in Figure S5. Table S1 gives the Poisson fit amplitudes for all three T1 samples to all three reference patterns.

Table S1	**T1 background**	**GILGFVFTL 483/796**	**GILGFVFTL 483/587**
T1 Blank	616	50	30
T1 + 829 amols	577	361	556
T1 + 62 pg/ml	1855	250	256

The relation between the scaled reference amplitudes (shown as gray bars in Figure S5) and the numbers in the Table S1 is as follows. Divide the reference peak amplitudes by the largest amplitude (scale so the largest reference peak is 1.0), multiply all reference peaks by the number in the table and then multiply all peaks by the factor 0.32. The last factor accounts for a transition between peak integrals (the event measure for Poisson fitting) and peak heights which are plotted over the spectrum. For example, the largest peak in the T1 background fit (Figure S5A) is at m/z 536.3 and the gray bar height is 1 x 577 x 0.32 = 185 events.

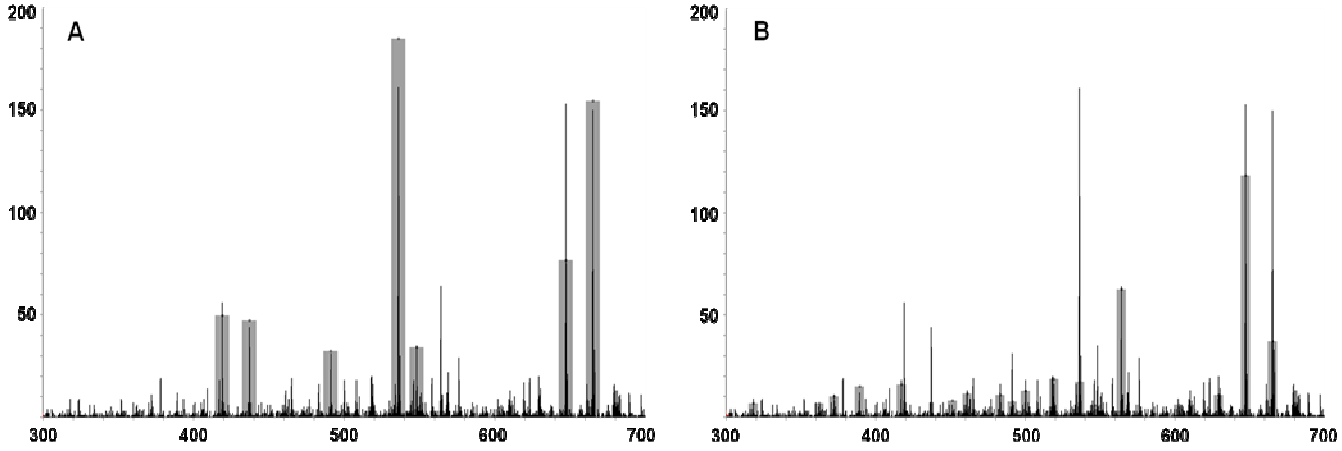

Figure S5. The T1 background in the MS^3 483.8 : 796.5 spectrum is represented by seven major reference peaks whose relative amplitudes were measured in the unloaded T1 sample. The Poisson fit to this background for the MS^3 483.8 : 796.5 spectrum of the T1 sample with 829 amols GILGFVFTL peptide added is shown in **A** as gray bars. The Poisson fit to the GILGFVFTL y_7 reference peaks in the same MS^3 483.8 : 796.5 spectrum is shown in **B**.

The T1 background contribution to the 0-offset amplitudes is corrected by scaling the 0-offset events assigned in the blank T1 sample. For example, the T1 background events for the sample T1 + 829 amols GILGFVFTL is 94% (577/616) of the background events assigned in the T1 blank sample, hence 47

(0.94 x 50) of the 361 MS^3 483 : 796 events are due to background and subtracted (314 events assigned to y_7 fragment of peptide). Correspondingly, 28 (0.94 x 30) of the 556 MS^3 483 : 587 events are background (528 assigned to b_6 fragment). For the 'T1+62 pg/ml' sample the background contribution to the y_7 0-offset amplitude is 1855/616 x 50 = 151, hence 250-151 = 99 events are assigned to peptide. For b_6 one subtracts 1855/616 x 30 = 90 from 256 to get 166 events. This gives Table 1 of the manuscript which is reproduced here.

Table 1	**T1 background**	**y_7 MS^3 483:796**	**b_6 MS^3 483:587**
T1 Blank	616	50	30
T1 + 829 amols	577	314	528
T1 + 62 pg/ml	1855	99	166

Two partially independent estimates can be made, one from the y_7 and one from the b_6 fragment. For the y_7 fragment, 99 events in the 'T1+62 pg/ml' sample, scaled (divided) by the 1855 T1 background events, is 9.8% of the 314 events in the 'T1+829 amols' sample scaled by the 577 T1 background. For the b_6 fragment, 166 divided by 1855 ('T1+62 pg/ml' sample) is also 9.8% of 528 divided by 577 ('T1+829 amols' sample). As the 'T1+829 amols' sample added peptide calibrated to 100 copies per cell, both the y_7 and b_6 MS^3 spectra indicate 9.8 copies of GILGFVFTL per cell is loaded on T1 cells by incubation in a solution of peptide at 62 pg/ml. That the two measurements gave two digit correspondence in this case is an accident, we are certainly not suggesting this is generally expected. Quoting a percentage accuracy for these measurements in general is not realistic since it obviously depends on the signal intensities.

Comparing translated inner product (correlation function) with Poisson plots. Figures 4 and 5 of the manuscript illustrate in part the contrast between detection using the probabilistic Poisson plots and detection using the metric translated inner product or correlation function. As discussed in the manuscript, there are two concepts involved in both of these detection algorithms. The first is the notion of difference between a measured and a reference spectrum. The Poisson difference is probabilistic while there are a number of implementations that use metric differences ($L^p(M)$ spaces in general, although $L^2(M)$ or Euclidean metric spaces are most common). The second concept is to compare the expected optimal fit with a set of other candidates and identify the confidence in detection with some function of the relative amplitudes. For our implementation of Poisson detection and for the correlation function, the set of other candidates is generated by translation in m/z. The optimal candidate is then the untranslated spectrum or 0-offset peak. Again we emphasize that the common implementation of the metric algorithm (e.g., in SEQUEST) is to compare LC-MS/MS spectra against (primarily) low confidence *in silico* MS/MS spectra generated from a database of sequences. If the computer model predicts some fragments that are not generated in the experimental dissociation the consequence for metric scoring is secondary - one is more interested in the peaks that are expected and observed. On the other hand, Poisson detection with an inaccurate spectral model would generally fail - the failure to observe even minor predicted peaks will eliminate candidates even if there is substantial correspondence with other predicted peaks.

At reviewer request we have included another comparison of the probabilistic and metric algorithms of the detection data in the manuscript (Figure S6). This is of the influenza peptide FVANFSMEL in the infected and uninfected epithelial cells and the Poisson detection illustrated in Fig. 3 of the manuscript.

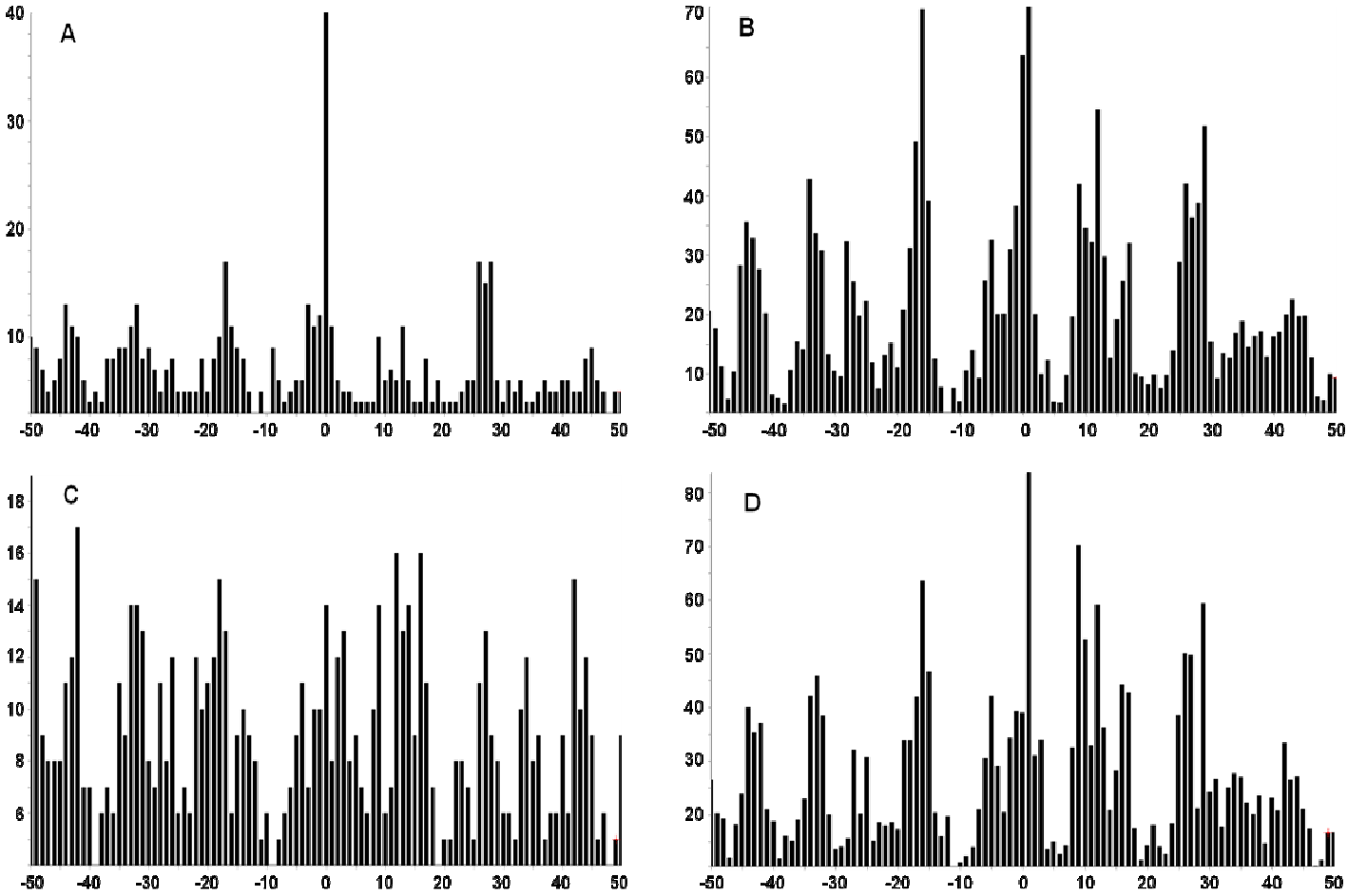

Figure S6. Comparing detection between Poisson and correlation function algorithms. This uses the same data shown in Figure 3 of the manuscript and further details are given there. **A** is the Poisson detection plot corresponding to the influenza infected BEAS sample, **B** is the translated inner product or correlation function of the same MS^3 spectrum using the same reference spectrum. **C** is the Poisson plot corresponding to the uninfected BEAS sample, **D** is its translated inner product or correlation function.