Automated selected reaction monitoring software for accurate label-free protein quantification

Teleman, J., Karlsson, C., Waldemarson, S., Hansson, K., James, P., Malmström, J., Levander, F.

May 29, 2012
Contents

1 Supplementary Methods: Anubis algorithm 2
2 Supplementary Figure 5 11
3 Supplementary Figure 6 12
4 Supplementary Table 1: Comparison of Anubis analysis with previously published manual analysis. 13
5 Supplementary Tables 2a-b and Supplementary Figure 7: Biological and technical replicates 14
6 Supplementary Table 3: Measured S. pyogenes proteins and peptides 16
7 Supplementary Results: Null distribution size effect on Anubis results. 20
8 Supplementary Results: Number of transitions used effect on Anubis results. 21
9 Supplementary Table 4: Anubis analysis times. 25
10 Supplementary Table 5: Protein yield from S.pyogenes cultures. 26
11 Supplementary Table 6: CVs and success rates for biological and technical replicate sets 27
12 Supplementary Notes: Missing values 28
13 Supplementary Results: Dilution series analysis 34
14 Supplementary Figure 24: normality of peptide replicate measurement 37
1 Supplementary Methods: Anubis algorithm

Problem Formulation
To simplify the description of the algorithm we first introduce some basic notations. The signal from a transition, $s$, will be referred to as a fragment and viewed as a function of time: $s(t)$. The set of all $n$ fragments $s_i(t), i = 1..n$ arising from the same precursor is called a precursor signal. For this paper we will also use the convention that a bar, $\bar{a}$, will denote an interval containing $a > 0$ which is in every point larger than 0, and the length of such an interval is denoted with surrounding vertical bars, $|\bar{a}|$

$$\bar{a} := [start(\bar{a}), end(\bar{a})], \quad 0 < start(\bar{a}) \leq a < end(\bar{a}) \quad (1)$$
$$|\bar{a}| := end(\bar{a}) - start(\bar{a}) \quad (2)$$

Algorithm Outline
Before going into the details of the algorithm, the importance of signal processing should be mentioned. We have baseline reduced and smoothed our signals with the method and reasoning specified in Signal Processing before analysis. This is very important if aiming at recovering signals close to the noise level. The proposed algorithm shows similarity with many other suggested algorithms for peak picking, in the sense that it can be divided into a search stage and a sort stage. In the search stage, a search is done for time intervals in the precursor signals where peaks might exist. Any such place will be referred to as a peak candidate (PC). When the search is done, the sort stage commences as Anubis filters the results by a number of aspects to retrieve the most probable PC for each precursor signal. This PC is chosen as the elution time of the target peptide in the current precursor signal, and will during the sort stage receive an estimate of its quality and a quantity.

Zooming in one level, we present an overview of the stages and the substages implemented in Anubis in Supplementary Figure 1. In the rest of the method, each substage is described under its own heading, in the order of execution.

PC Finding
We have mentioned earlier that peptides are identifiable by their fragment ratios on the used mass spectrometer, which we can retrieve from our reference chromatogram. Thus finding a peptide in our precursor signal is actually a case of searching for time intervals where the measured fragment ratios are close enough to our reference fragment ratios.

There are two different ways to consider the ratios between the signal fragments. If one chooses to consider each signals part of the total, $p_i(t)$

$$p_i(t) = \frac{s_i(t)}{\sum_i s_i(t)}$$

the number of ratios is equal to the number of transitions, but since every ratio is defined by all the transitions, interference in one transition will cause distortion in all ratios.
Search stage - PC Finding

1. Find PCs
2. Calculate PC estimates

Sort stage

1. Sanity check
   Remove PCs where any transition sum is beneath
   
   \[ \text{sanityTolerance} \times \sum_t e_i(t) \]

1. Estimate a local p-value for each PC
   Pick the PC with minimal p-value, or all PCs with
   
   \[ p < 0.01 \]

1. Quantify the found PCs
   Choose the largest PC

Supplementary Figure 1: Summary of the Anubis algorithm.

Therefore \( p_i(t) \) are too vulnerable for our analysis, and we instead use the pairwise fragment ratios, \( r_{ij}(t) \)

\[
 r_{ij}(t) = \frac{s_i(t)}{s_j(t)} \quad (3)
\]

This will make a total of \( n(n-1) \) untrivial ratios, and of these each signal is involved in \((n-1)\) ratios. Thus, one erroneous transition distorts only \((n-1)\) of \( n(n-1) \) ratios instead of \( n \) of \( n \), and the vulnerability to interference is reduced. Having picked this pairwise definition of fragment ratios as the base of analysis, the two concepts of the target fragment ratio \( R_{ij} \), and target fragment ratio range \( \bar{R}_{ij} = [\frac{1}{1.5} R_{ij}, 1.5 R_{ij}] \) will be needed. These reflect the value of each fragment ratio that is expected for the targeted peptide. The interval size was experimentally chosen, and the individual values are retrieved from the reference chromatogram.

In the worst conceivable scenario where there still is a theoretical possibility to identify a peptide using ratios, all but two transitions in the peak are hidden behind interference. In order to find these peaks we need to accept PCs where only one fragment ratio stays inside the target fragment ratio range. This requirement may seem very loose at first but when we later grade our PC quality, low quality PCs can be discarded. Therefore we can afford a few over-optimistic candidates initially.

PC Finding Algorithm

The backbone of Anubis' PC finding, is the novel concept of correct ratio areas \((I_{ij})\). \( I_{ij} \) is a length of time during which a the fragment ratio between fragments \( i \) and \( j \) is within its target ratio range, which is demonstrated by the constructed
Supplementary Figure 2: Constructed example of correct ratio areas. Consider the fragments a, b and c. From our reference (a) we know that these should have the pairwise fragment ratios $r_{ab} = 3$, $r_{ac} = 6$ and $r_{bc} = 2$ as shown in (c). In (b) is shown a fictional precursor signal, for which the correct ratio areas are shown in (d), ordered by ratio. Demonstrating some correct ratio areas, CRA1 is formed because $r_{bc}$ is within $\bar{R}_{bc}$ while CRA2 is formed because $r_{ab}$ is within $\bar{R}_{ab}$. CRA3 is shorter than the others because a is hidden in interference to the left, and to the right c is hidden.
example in Supplementary Figure 2. Mathematically, \( \bar{I}_{ij} \) can be defined as

\[
\bar{I}_{ij} := \begin{cases} 
|I_{ij}| \geq L_{\text{min}} \\
 r_{ij}(t) \in \bar{R}_{ij}, & \forall t \in \bar{I}_{ij}
\end{cases}
\]

where \( L_{\text{min}} \) is a parameter defining the minimal required length.

With the concept of correct ratio areas, the following algorithm is used to extract PCs from a precursor signal \( S \). A graph is created where each \( \bar{I} \) in \( S \) is a node and an edge \( E(\bar{I}_{ij}, \bar{I}_{kl}) \) is formed between CRAs \( \bar{I}_{ij} \) and \( \bar{I}_{kl} \) if they share exactly one fragment, and overlap in time in at least one data point.

After having built the graph the \( \bar{I} \)-nodes are effectively divided into \( \bar{I} \)-groups, where each \( \bar{I} \)-group consists of all the \( \bar{I} \)-nodes that are reachable from any node in the group. Each \( \bar{I} \)-group is translated into a PC, and since there can exist single node groups (not connected to any other node) there can be single \( \bar{I} \) PCs.

### PC Complementary Estimation

According to our logic, \( \bar{I}_{ij} \) specify intervals where a pure signal might exist, and groups of sufficiently overlapping \( \bar{I} \)’s form PCs. That, however, raises the question how one should treat the points in the PC that are not part of any \( \bar{I} \). In order to handle this problem we have formed the concept of estimates, \( e_i \). The idea is to estimate untrusted parts\(^1\) of a fragment using the trusted fragments at that point, as graphically presented in Supplementary Figure 3.

More specifically, an estimate is defined using the target ratios \( R_{ij} \) as

\[
\mathcal{I}(t) := \{ \forall \bar{I}_{ij} \mid t \in \bar{I}_{ij}, \bar{I}_{ij} \in \mathcal{P}C \} \\
\mathcal{P}(t) := \{ \forall i \mid \bar{I}_{ij} \in \mathcal{I}(t) \} \cup \{ \forall j \mid \bar{I}_{ij} \in \mathcal{I}(t) \} \\
e_i(t) := \begin{cases} 
s_i(t) & \text{if } i \in \mathcal{P}(t) \\
\sum_{k \in \mathcal{P}(t)} R_{ik} s_k(t) & \text{otherwise}
\end{cases}
\]

In plain text: if any \( \bar{I} \) in the PC includes the \( i \)-th fragment at the time \( t_x \), then \( e_i(t_x) = s_i(t_x) \). Otherwise \( e_i(t_x) \) is calculated from the \( s_j(t_x) \) that are in a \( \bar{I} \) at the time \( t_x \), namely as the mean of all the products \( R_{ij} s_j(t_x) \). Estimates will prove useful further on, when handling sanity checks and peak quantification.

### Sort Stage

#### Sanity Check

The above presented method of constructing PCs allows candidates where as few as one of the fragment ratios \( r_{ij} \) is within its target range \( \bar{R}_{ij} \) for a longer length of time. We allowed these weak PCs in order to tackle the problem of interference. Interference is always additive though, which leaves a possibility open to remove PCs where a fragment is much lower than expected. Introducing a new tolerance value, we thus use our estimates \( e_i(t) \) to make certain that all fragments are

\(^1\) = parts not validated by any \( \bar{I} \)
Supplementary Figure 3: A fictional estimate example. The figure shows a precursor signal with the fragments a, b and c during a PC. Below the precursor signal all the I of the PC are shown, organized by fragment ratio. There is interference to the left in a and to the right in c. This coincides with a and c not being part of any I during the interference, whereas any fragment that is part of a I is close to the pure peptide signal. Therefore a’s estimate (dark green) is calculated from b and c, and c’s estimate (purple) from a and b. These estimates, which are theoretical signals, prove useful later when analyzing the PC.
large enough to possibly be from the target peptide. Described in logic: only keep the candidate PC if $H_0$ is true for $PC$, where $H_0$ is defined as

$$H_0 : \neg \exists d_i | \int_{\text{start}(PC)}^{\text{end}(PC)} s_i(t) \ dt < d_{\text{tol}} \quad \int_{\text{start}(PC)}^{\text{end}(PC)} e_i(t) \ dt, \ d_{\text{tol}} \leq 1$$

$$H_1 : \neg H_0$$

where $d_{\text{tol}} = 10$ is our chosen tolerance level for considering PCs sane. The sanity check is a computationally cheap way of avoiding unnecessary quality calculations, which are expensive. It’s also complementary to the quality estimation of the next section in the sense that it considers the entire PC, whereas the quality estimation only considers the parts of the PC that are in a $\bar{I}$.

**Quality Estimation**

To determine the quality of a peak candidate in a chromatogram, we would like to know how often a peak of similar or better fragment ratios and signal to noise would appear if there was no endogenous peptide in the chromatogram. We therefore use a p-value as quality measure, representing the significance of the candidate peak features in a null distribution, $f_0$, of 1,000 false chromatograms. We define $p$ as

$$p = \frac{|a(f_0)|}{|f_0|}$$

where $a(X)$ denotes the subset of $X$ containing the chromatograms with good enough peaks. The null distribution $f_0$ is generated for each chromatogram with a new algorithm based on wavelet analysis, described in detail in *Generating Uncorrelated Signals*. $a(f_0)$ is taken as the subset of chromatograms in $f_0$ with at least one time interval fulfilling the target peak criteria. We considered the criteria to be fulfilled if there in the false chromatogram is a segment of the target PC length $[a, a + |PC|)$, such that it for every $\bar{I}$ in the PC contains a respective correct ratio area $\bar{I}^*$ of at least the same length

$$\bar{I}^* := \left\{ \bar{I}^* \in [a, a + |PC|) \mid |\bar{I}^*| \geq |\bar{I}| \right\}$$

and also has a signal to noise value larger that that of the target PC, times a tolerance of 0.8. The signal to noise for a interval $\bar{X}$ is calculated as

$$\bar{X}_{s/n} := \max(\bar{X}_{i,s/n})$$

$$\bar{X}_{i,s/n} := \max(x \mid x \in \bar{X}) / \text{median}(x \mid x \notin \bar{X})$$

**Peak Quantification**

To achieve precise quantification while avoiding interference we used our previously calculated estimates $e_i(t)$. Since interference is by nature only additive, any contaminated fragment will be larger than its estimate. Therefore the quantity of a transition peak $q_i$ can be simply calculated as

$$q_i = \int_{t \in PC} \text{min}(q_{\text{tol}} \ast e_i(t), s_i(t)), \ q_{\text{tol}} \geq 1$$
We chose a tolerance of $q_{\text{tot}} = 1.5$, which cuts out substantial contributions from interference while allowing reasonable variation in the ratios. The total quantity $q$ is unsurprisingly taken as the sum of the parts,

$$q = \sum_{i=1}^{n} q_i \quad (11)$$

**Signal Processing**

There are a number of error sources in SRM, biological and chromatographical, as well as the technical limits of the mass spectrometer itself. In order to improve the SRM data analysis it is thus important to suitably refine the precursor signals to, if possible, compensate for or reduce the noise.

On the MS setup used for our experiments there is some background current, as no fragment intensity is ever zero but the minimum lies a little below 1 unit of intensity. This offset is inconvenient when looking at the fragment ratios, so we remove it using the concept of the minimal signal $M_s$ for the precursor signal $s$. This is defined as

$$M_s(t) = \min(s_i(t)), \forall i \quad (12)$$

Since it is very unlikely to have interference in all fragments simultaneously, at all times when no real signal exist $M_s$ will quite accurately equal our background. This means that the typical background size $l_s$ can be taken as $l_s = \text{median}(M_s)$. During signal peaks, $M_s$ will be far above $l_s$ since all fragments form a synchronized peak, and so $l_s$ does not accurately model our background. In fact if we would use $M_s$ unmodified to remove background current we would likely distort any signal beyond recognition. Therefore we choose as our background model the limited minimal signal, $L_s$, defined as

$$L_s(t) = \min(M_s(t), 2l_s) \quad (13)$$

After having removed the background current we also applied a smoothing filter, namely a Savitzky & Golay 9-point filter (SG9) Savitzky and Golay, 1964. Also considered were different wavelet approximations and fourier-transformation based filtering, but SG9 provided the best mix of smoothing and faithfulness to the original signal. This filter is in essence a moving least-squares quadratic regression, but with some convenient computational optimization possible with equidistant samplings.

**Generating Uncorrelated Signals**

The separating strength of Anubis is individual for each PC depending on correct ratio area lengths and input signal, so to measure the quality of a PC a unique null distribution $f_0$ of false precursor signals $s'$ is needed for each precursor signal $s$. For every $s'$ it's fragments $s'_i$ also need similarity to their respective $s_i$, both in frequency and time domains. Generating $f_0$ was accomplished with a new method as described below, based on wavelet analysis.

A wavelet analysis of level $k$ decomposes a signal into the $k$ detail signals $D_i$, $1 < i \leq k$ and the approximation signal $A_k$. $A_k$ and $D_k$ are defined recursively as

$$A_k = l(A_{k-1})$$
$$D_k = h(A_{k-1}) \quad (14)$$
where \( l \) and \( h \) are half band low-pass respective high-pass filters. The original signal is set as \( A_0 \).

We decompose \( s_i \) using wavelet analysis, and then generate \( s'_i \) by a randomized reconstruction from \( s_i \)'s composition. Wavelets do not generally allow lossless reconstruction, so to accommodate this the decomposition was modified to

\[
A_k = \ l(A_{k-1}) \\
D_k = A_{k-1} - A_k
\]

(15)

A lowpass filter wavelet was designed using ScopeFIR (http://www.iowegian.com/scopefir), by using the Parks-McClellan algorithm, with 21 taps, Cutoff = 0.15, \( L = 2 \), Stopband Attn = 70 dB and Grid Density = 16, with Zeroize checked. In real data the intensities for different wavelet levels tend to correlate, which is why the \( s'_i \) were divided into sections where each section would only choose intensities from the corresponding section in \( s_i \).

A false precursor signal \( s' \) is generated from the real precursor signal \( s \) by applying this algorithm to all \( s_i \) in \( s \). An example of a fragment generated with our method is shown in Supplementary Figure 4, compared to two fragments generated with alternative methods.

**Supplementary Figure 4:** Example of transition reconstructed using three methods. a) The original chromatogram. b-d) Green transition reconstructed using a) simple empirical distribution, c) uniformly randomized wavelet decomposition and d) sectioned randomized wavelet decomposition.
Implementation details

Anubis was implemented in Scala 2.8.1 and consequently runs on the Java Virtual Machine. Development was performed on a desktop computer running OpenSuSE 11.3, with Eclipse Hellios with the scala IDE Eclipse plug-in. Program runtime is heavily dependent on the number of transitions and samplings in the data file. A typical scheduled, 45 min gradient, 800 transition SRM file is analyzed in 1:00-1:30 minutes, using all of 1 3Ghz processor and about 1Gb of RAM.
2 Supplementary Figure 5

Examples of fragment ratios for a dilution series of synthetic peptides spiked in a SILAC human cell line background in triplicates. The peptides are spiked with relative strengths of 1, 3, 10, 30 and 100 and are ALEAANGELEVK, ATFYGEQVDYYK, LSFLVNALAK and VDFTEEEINNMK.
3 Supplementary Figure 6

Examples of instability of elution time for 3 peptides from 3 streptococcal proteins in 2 sets of 10 technical replicates. Peptides are VFSVGGDLVEMQEAVAK (NP_269778.1), AGVAAAEAEIIK (RL7_STRP1) and LVDIINPTQK (RL10_STRP1).
Supplementary Table 1: Comparison of Anubis analysis with previously published manual analysis.

| Statistic                             | Anubis label-free | SIS   | Manual SIS |
|---------------------------------------|-------------------|-------|------------|
| CVs at 2.92 fmol / µl                 | 9.8 - 21.5%       | 3.7 - 17.5% | 4.3 - 14.1% |
| Median CVs across conc. range         | 6.6 - 19.3%       | 2.6 - 7.4%  | 2.8 - 5.0%  |
| Number of intra-lab CVs < 15%         | 60 / 80           | 74 / 80 | 78 / 80    |
| $R^2$ median                          | 0.942             | 0.993  |            |
| $R^2$ average                         | 0.906             | 0.929  |            |

Supplementary Table 1: Comparison of Anubis analysis with and without SIS-reference information, and Anubis analysis vs. manual analysis, with both using SIS-reference information.
5 Supplementary Tables 2a-b and Supplementary Figure 7: Biological and technical replicates

| Growth medium | Preparation batch | Time of analysis | n grown | n injected | n final | note |
|---------------|-------------------|-----------------|---------|------------|---------|------|
| 0% human plasma | 1 | 19-21 nov 2010 | 10 | 9 | 9 | 1 excluded because of failed protein extraction |
| 10% human plasma | 1 | 19-21 nov 2010 | 10 | 10 | 9 | 1 excluded because of synthetic peptide carry over |

Supplementary Table 2a: Summary of biological replicates grown with or without 10% human plasma.

| Replicate set name | Biological replicate | Preparation batch | Preparation sample | Time of analysis | n injected | n final | note |
|-------------------|----------------------|-------------------|-------------------|-----------------|------------|---------|------|
| A1 nov | 0% plasma 5 | 1 | A | 19-21 nov 2010 | 10 | 10 | |
| A2 apr 1 | 0% plasma 5 | 2 | B | 27-29 apr 2011 | 10 | 7 | 3 excluded because of failing column |
| A2 apr 2 | 0% plasma 5 | 2 | C | 27-29 apr 2011 | 10 | 10 | |
| A2 jun 1 | 0% plasma 5 | 2 | B | 17-20 jun 2011 | 10 | 9 | 1 excluded because of low TIC signal |
| B mar | 0% plasma 10 | 1 | D | 15-23 mar 2011 | 10 | 10 | |

Supplementary Table 2b: Summary of technical replicates.
Supplementary Figure 7: Summary of *S. pyogenes* replicates. We cultured 20 biological replicates, 10 with and 10 without human plasma supplement. Out of these one without plasma had to be discarded. All of these were first prepared for mass spectrometry and injected once, giving 1 successful measurement of all biological replicates except one. Two of the sample preparations without plasma were injected an additional 10 times each giving the two sets of technical replicates A1nov and B1mar. From one of the two biological replicates chosen for creating technical replicates, two additional sample preparations were made in batch 2. The batch 2 samples were injected 10 times each at two separate times, giving A2.1apr, A2.2apr, A2.1jun and A2.2jun.
6 Supplementary Table 3: Measured *S. pyogenes* proteins and peptides

| Uniprot protein name or refseq id | Peptide sequence                  |
|-----------------------------------|-----------------------------------|
| RL29_STRA3                        | GLSQEELAK                         |
| RL29_STRA3                        | TVQSEMK                           |
| RL29_STRA3                        | FQAAGQLEK                         |
| RL7_STRP1                         | EGVAAAEAEIK                       |
| RL7_STRP1                         | GLVDGAPANVK                       |
| RL7_STRP1                         | EASILELNDLVK                      |
| RL18_STRP1                        | GISEVVFDR                         |
| RL18_STRP1                        | GGYLYHGKR                         |
| RS10_STRA1                        | LDDLPSGVNVEIK                    |
| RS10_STRA1                        | LVDIINPTQK                       |
| RS10_STRA1                        | TGATVAGPVPPLPTER                  |
| RL22_STRP1                        | TLNSAIAANAENNFLEK                 |
| RL22_STRP1                        | ANLVSETFANEGPTMK                  |
| RL22_STRP1                        | VADAIAILK                        |
| RL15_STRP1                        | LGFEGGQTPLFR                      |
| RL15_STRP1                        | VLGNGEKT                         |
| RL1_STRP1                         | AGNVQALIGK                        |
| RL1_STRP1                         | AYSVEAAVALVK                      |
| RL1_STRP1                         | AAGADFYGEDDLVAK                   |
| RS17_STRA1                        | TITTVVETK                        |
| RS17_STRA1                        | LVEVVEK                           |
| RS17_STRA1                        | EEIQALATLPNR                     |
| NP_269238.1                       | GLTDQDVTLR                       |
| ACPS_STRP1                        | LFQDIEILNDVR                     |
| ACPS_STRP1                        | VQEIIVEELK                       |
| ACP_STRP1                         | YGNTSAASIPILLSEAVQK               |
| FABH_STRP1                        | SILTLMAQSDITK                    |
| FABZ_STRP1                        | LFYAGMDK                         |
| FABZ_STRP1                        | YPMLLVDISR                      |
| FABZ_STRP1                        | VLEVSDDHIAIKeK                   |
| NP_269766.1                       | NGGQPNEGYR                       |
| NP_269766.1                       | NLMEMSDLK                        |
| NP_269767.1                       | MQEGIMSLMQMAK                    |
| NP_269767.1                       | GSVSHVPEVPDELFAK                  |
| NP_269768.1                       | IQVEHPVTEFVTGVDIVK                |
| NP_269768.1                       | INAENTAFNFAPSPGK                 |
| NP_269768.1                       | FIGPSAVMDK                       |
| NP_269770.1                       | DLMAQFDTSSLR                     |
| NP_269770.1                       | MNIQEIK                          |
| NP_269770.1                       | TNEGELIFSK                       |
| NP_269771.1                       | ALPNMAAGNVAMSLK                   |
| NP_269771.1                       | FGTDVMIVGGSEAAITK                |
| NP_269771.1                       | FAIAGFQSALTSTTEDPSR               |
NP_269772.1  MTEEDFER
NP_269772.1  NIFITGSTR
NP_269773.1  TAFLFAGQGAQK
NP_269773.1  LIPLNVSQFHTALLEPASR
NP_269773.1  IPIVGNTEANIMEK
NP_269774.1  TATDIEEMGAGSLR
NP_269774.1  LGVDAVIAEGMEAGGHIYGK
NP_269774.1  DIDTVISAQVVGHPVR
NP_269777.1  ELMVTLGTVTTLNSLNK
NP_269777.1  YDQVTIPSQDR
NP_269777.1  TSQFSDVSLK
NP_269777.1  VFSVGGDLVEMQEAVAK
NP_269777.1  ATHLVMTGEGITADK
NP_269777.1  SLWQSFTGWEDYAK
NP_269777.1  DQGPQPK
NP_269777.1  GAQGPVGPAGK
NP_269777.1  TPEPQNPDTAPHTPK
STRP_STRP1  SEQLILASER
STRP_STRP1  TILPMDQEFTYHVK
STRP_STRP1  DGSVTPLTQPQVEFLLK
C5AP_STRP1  VVANGTYTYR
C5AP_STRP1  NGYFLEGFVR
C5AP_STRP1  AYDYAYANR
NP_269972.1  GIDFEGFQYGWSDEYK
NP_269972.1  SGVGLSQYGWSK
NP_269972.1  NPQESDFDR
NP_269972.1  VPEKIPQHTPQK
NP_269972.1  YGWSSDK
NP_269972.1  EEWPDWPDWDSDDK
NP_269972.1  TRPPYGEALGTGYEK
DLTC_STRP1  VPISEFGR
DLTC_STRP1  MSIEETVIELFDR
NP_268944.1  ATFHSIQLQDR
NP_268944.1  VTPHYSQEVVPR
NP_268944.1  GHLVGQFQCLNDDEPR
NP_269065.1  AIYVTDSDNSASIGMK
NP_269065.1  INFNGEQMFDVK
NP_269065.1  YFVGVNAGK
SPEC_STRP1  SDLLYAYITIPYDYK
SPEC_STRP1  NFSHFDIYLEK
SPEC_STRP1  IYDATSPYVSGR
NP_268627.1  VVITAPGGNDVK
NP_268627.1  ALHDAFGIQK
NP_268627.1  AIGLVPELNGK
SPEC_STRP1  GHHYNQGNPYNLTPVIEK
SPEC_STRP1  SPEILGYSTSGSDNGK
NP_269973.1  EQLTIEK
NP_269973.1  EAELEAIDQASR
NP_269973.1  DLANLTAELDK
NP_269973.1  ASDSQTPDKPNGK
NP_269973.1  ESTSWDR
NP_269973.1  GQAPQAGTKPNQNKR
NP_269973.1  LENAMEVAGR
NP_269973.1  LELDQLSSEK
NP_269973.1  AANNPAIQNIR
NP_269973.1  ELETTTR
NP_269973.1  ELQQDYDLAK
NP_269973.1  ALELAIDQASQDYNR
NP_269973.1  ALEEEANSK
NP_269973.1  QALEDQRR
NP_269973.1  EVIEDLAANPAPIQR
DHLA_STRP1  TLINEVNNR
DHLA_STRP1  EMVDNYTR
NP_269464.1  HIPNGGTLTNLLGNAPEK
NP_269464.1  AATPIAALDVK
17981973  HFNINYGFVDVK
17981973  NIPVINIFIDGVQQK
17981973  NSYGGITLSDGNR
NP_268729.1  NGAMTHRPVEVR
NP_268729.1  HSFMSTTALK
NP_268729.1  GCQLEVGVAYVSDQDK
NP_268723.1  ALDSSGIVR
NP_268723.1  NPLWVSPESFK
NP_268723.1  TAQQDLIPLIQSLVVPK
SPEH_STRP1  GTVTDLDNSGK
SPEH_STRP1  YEAFGGITLTNSEK
SPEH_STRP1  VTAQQEVDIK
NP_269959.1  LVFHTNDNSDK
NP_269959.1  QIPVPYNQIESTNK
NP_268735.1  HESFNLYDATSSSTR
NP_268735.1  GQLEIAINGNK
NP_268735.1  NFSVDSEVESYTTK
SPEG_STRP1  SLQFVGILNQDGK
SPEG_STRP1  YNIYDSESRR
SPEG_STRP1  QFTLQEFDFK
TACY_STRP1  TQYTESMVYSK
TACY_STRP1  TSPFSTVPIPLGANSR
TACY_STRP1  ILDGTLGIDFK
NP_269203.1  ENWWVVYEIGTPR
NP_269203.1  DYSASANITK
NP_269203.1  FNAQSHVAGIEALR
NP_269818.1  VDFFSVPADK
NP_269818.1  GLPSIDSLHYOLEN
NP_269818.1  LDLAPGTEKR
NP_269947.1  AEVSSTTMSSQR
NP_269947.1  GETGAQGPVGPGQGEK
NP_269947.1  SPEGEAQPGEK
NP_269402.1  VLGKIPFDVYNNR
NP_269402.1  AIVFSTTQLETHK
NP_269402.1  VEALSDAIK
| Accession     | Sequence                  |
|--------------|---------------------------|
| NP_268544.1  | IELGDTYTK                 |
| NP_268544.1  | YSDGVYQYEYDSELTK          |
| NP_268544.1  | AVLAEIDAK                 |
| NP_269520.1  | FNPPGWHNYK                |
| NP_269520.1  | NFPDTTEILLGTK             |
| NP_269520.1  | GHLVGYQFSGLNDEPK          |
| HASA_STRP1   | HAQAWAFAER                |
| HASA_STRP1   | LGLSFLYEPFK               |
| HASA_STRP1   | SFNDETVYAAATGHLNAR        |
| NP_269989.1  | TLGTSQITPALFPK            |
| NP_269989.1  | AVVSMQSSDNTINEK           |
Supplementary Results: Null distribution size effect on Anubis results.

To characterize the Anubis p-value calculation further we have analyzed the same 2 samples using null distribution sizes of 100, 300, 1000, 3000 and 10000, in triplicates (because of the stochastic null distribution generation). This analysis indicated that there is very little effect of the null distribution size on the overall results. Looking at the standard deviation of the p-values across all the null distribution sizes, a large majority of peptides had standard deviations of less than 0.001 (Supplementaty Figure 8).

The advantage of a larger null distribution size becomes apparent when calculating the standard deviation of each peptide in each triplicate measurement of the same null distribution size. Plotting the average standard deviation of each null distribution size shows that increasing null distribution size gives increasing p-value precision (Supplementary Figure 9). Our choice of 1000 as the default null distribution size is a tradeoff between the need for p-value precision and inconveniences of long analysis times. As shown in Supplementary Table 4, analysis times are rapidly increasing for large null distribution sizes.

Supplementary Figure 8: The distribution of average standard deviations of the p-values of 257 successfully detected peptides, analysed in triplicate with the null distribution sizes 100, 300, 1000, 3000 and 10000. The low standard deviations suggest that the reported p-value is not dependent on the null distribution size.
Supplementary Figure 9: The average standard deviations of the p-values of 257 successfully detected peptides, analyzed with different null distribution sizes. An inverse correlation between null distribution size and p-value variation is seen.

8 Supplementary Results: Number of transitions used effect on Anubis results.

The effect of the number of transitions used on Anubis results is slightly more complicated than the null distribution size effect. To investigate this we repeatedly analyzed 2 samples each measuring 161 peptides. The number of transitions used was altered programmatically by only using the top-most abundant transitions in the reference chromatogram. To see how increasing numbers of transitions affected peptide quantification, analyses of the two samples were paired and fold-changes calculated, which gave three fold-changes for each transition limit. We then calculated the coefficients of variation for the combined fold-changes of 2 and 3 transitions, 3 and 4 transitions and 4 and 5 transitions. Plotting these indicates discrepancies in the quantifications achieved using 2, 3 and 4 transitions, and only the 4 and 5 transition quantifications seem to agree (Supplementary Figure 10). Therefore we recommend using a minimum of 4 transitions for acquiring accurate quantifications.

The Anubis p-values respond to the transition number in 3 different ways. To visualize this behaviour we performed k-means clustering (k=4) on the means of the p-values at each transition level, which resulted in Supplementary Figure 11. In this data the majority of peptides are detected with a low p-value even at 2 transitions, and this p-value stays unchanged as the number of transitions is increased (cluster 4). However, 8 peptides (cluster 1) are classified as unreliable at the 2 transition level, but at 3 transitions the confidence in them is improved.
and they get low p-values from 3-5 transitions. These peptides would have been missed if only 2 transitions were measured. On the other hand cluster 2 and 3 are reasonably reliable (especially cluster 2) at the 2 transition level, but are discarded with increasing confidence as the number of transitions is increased. These peptides could have been false positives if measured with only 2 transitions. It appears from this data that 3 transitions would be the critical number of transitions to gain reliable p-values.

Finally we investigated whether the same peaks were picked when different numbers of transitions were used. To control for this we looked at the absolute difference in the averaged retention times between the transition levels (Supplementary Figure 12), in the same samples as above. This demonstrates that the same peak is chosen (difference ≤ 30s) between all the transitions levels for the vast majority of peptides measured.

In summary, we find that Anubis is reasonable stable to the number of transitions used for peptide measurement. Our analysis however suggests that a minimum of 3 transitions should be used to get the most reliable p-values, and for the most accurate quantification at least 4 transitions should be used. Lastly, the number of transitions used does not greatly affect Anubis analysis times, see Supplementary Table 4.

Supplementary Figure 10: To investigate the effect of the number of transitions measured on the anubis results, we analyzed the same two samples (PL3 and TH5) multiple times using from 2 transitions to all the 5 measured, in triplicates at each transition level. P-values at each level were averaged a k-means clustering performed (k=4) giving the 4 clusters displayed. While the majority of peptides have reliable p-values independent of the transition level (cluster 4), some peptides are falsely rejected (cluster 1) and some falsely accepted (cluster 2) at the 2 transition level, before the p-values stabilize with more transitions.
Supplementary Figure 11: To investigate the effect of the number of transitions measured on the anubis results, we analyzed the same two samples (PL3 and TH5) multiple times using from 2 transitions to all the 5 measured, in triplicates at each transition level. Pairing adjacent transition level results we calculated the CVs of the foldchanges between the PL and TH sample for each peptide. This data is shown for the 3 pairings. As seen there was not much difference between using 4 and 5 transitions on the fold change, whereas between 2 and 3 or 3 and 4 there is more disagreement between the correct fold change. This suggests that 4 transitions is the optimum number for anubis operation.
Supplementary Figure 12: To investigate the effect of the number of transitions measured on the anubis results, we analyzed the same two samples (PL3 and TH5) multiple times using from 2 transitions to all the 5 measured, in triplicates at each transition limit. Pairing adjacent transition level results we calculated the absolute difference in average retention time between the halves of the pairs. This data is shown for the 3 pairings.
9 Supplementary Table 4: Anubis analysis times.

| null distribution size | max transitions used | analysis time |
|------------------------|----------------------|---------------|
| 100                    | 6                    | 00:00:20      |
| 300                    | 6                    | 00:01:00      |
| 1000                   | 6                    | 00:03:50      |
| 3000                   | 6                    | 00:29:00      |
| 10000                  | 6                    | 01:30:00      |
| 100                    | 2                    | 00:01:00      |
| 1000                   | 3                    | 00:01:50      |
| 1000                   | 4                    | 00:02:50      |
| 1000                   | 5                    | 00:03:50      |

Supplementary Table 4: Anubis analysis time for different combinations of transitions used and null distribution size. Each combination was analyzed in triplicate and shown is the average time. Each analysis was carried out on 2 different samples, each measuring the same 161 peptides.
10 Supplementary Table 5: Protein yield from *S.pyogenes* cultures.

| Sample | Average A595 | Conc. mg/ml | Error (±mg/ml) |
|--------|--------------|-------------|----------------|
| TH1    | 0.3178875    | 0.25        | 0.08           |
| TH2    | 0.4804022    | 0.47        | 0.08           |
| TH3    | 0.4470633    | 0.42        | 0.08           |
| TH4    | 0.4300515    | 0.4         | 0.08           |
| TH5    | 0.4929202    | 0.49        | 0.09           |
| TH6    | 0.1187693    | 0.02        | 0.04           |
| TH7    | 0.474905     | 0.46        | 0.08           |
| TH8    | 0.4264271    | 0.39        | 0.08           |
| TH9    | 0.474461     | 0.46        | 0.08           |
| TH10   | 0.4650089    | 0.45        | 0.08           |
| Pl.1   | 0.5085806    | 0.51        | 0.09           |
| Pl.2   | 0.5013723    | 0.5         | 0.09           |
| Pl.3   | 0.5103161    | 0.51        | 0.09           |
| Pl.4   | 0.4913757    | 0.49        | 0.09           |
| Pl.5   | 0.4987511    | 0.5         | 0.09           |
| Pl.6   | 0.535176     | 0.55        | 0.09           |
| Pl.7   | 0.4730394    | 0.46        | 0.08           |
| Pl.8   | 0.4659583    | 0.45        | 0.08           |
| Pl.9   | 0.4311152    | 0.4         | 0.08           |
| Pl.10  | 0.5389933    | 0.56        | 0.09           |

**Supplementary Table 5:** Protein yield from the 20 biological *S.pyogenes* cultures. Protein concentrations were estimated with Pierce Coomassie Protein Assay kit, using a standard curve of Bovine Serum Albumin in concentrations from 0.25 to 2.0 mg/ml. The table shows the raw optical density value A595 as well as the translated concentration.
### Supplementary Table 6: CVs and success rates for biological and technical replicate sets

| Replicate set | Peptide median CV | Peptide measurement average success rate |
|---------------|-------------------|-----------------------------------------|
|               | RIB | FAS | Virulome | RIB | FAS | Virulome |
| **Individual replicate sets** |     |     |           |     |     |           |
| A1nov         | 4.1% | 7.2% | 14.5% | 98.8% | 100.0% | 80.0% |
| A2.1apr       | 10.7% | 15.1% | 19.7% | 100.0% | 98.6% | 80.7% |
| A2.2apr       | 10.3% | 14.4% | 19.9% | 100.0% | 95.2% | 74.7% |
| A2.1jun       | 5.7% | 6.7% | 17.0% | 99.5% | 97.1% | 78.6% |
| A2.2jun       | 9.0% | 10.3% | 26.4% | 99.6% | 97.4% | 76.5% |
| B1mar         | 4.5% | 6.6% | 22.2% | 97.4% | 96.8% | 77.7% |
| Biological replicates | 17.6% | 18.6% | 37.8% | 99.1% | 95.1% | 83.8% |
| **Two sample preparations** |     |     |           |     |     |           |
| A2.1apr + A2.2apr | 15.6% | 18.2% | 28.8% |
| A2.1jun + A2.2jun | 7.5% | 9.6% | 28.7% |
| **Two times of injection** |     |     |           |     |     |           |
| A2.1apr + A2.1jun | 35.0% | 32.3% | 42.9% |
| A2.2apr + A2.2jun | 44.2% | 41.4% | 52.3% |
| **Two samples preparations and two times of injection, unnormalized** |     |     |           |     |     |           |
| 46.2% | 43.8% | 55.8% |
| **Two samples preparations and two times of injection, normalized** | 11.7% | 16.4% | 21.6% |

**Supplementary Table 6:** Summary of CVs and success rates for biological and technical replicate sets, and different combinations of these. The detection success rate for a peptide is measured as the number of replicates were the peptide was detected below with $q < 0.01$, divided by the total number of replicates in that replicate set. The reported success rate for each protein group is the average of the peptides representing proteins from that group.
Supplementary Notes: Missing values

A few examples of peptides were the target peptide was detected in all, some of none of the replicates. The elution time of the detected peak is marked with two vertical black lines, as well as a horizontal black line indicating the quantity of the peak. Detection was defined at $p \leq 0.05$.

Detection in all replicates

**Supplementary Figure 13a:** DLMAQFDTSSLR reference

**Supplementary Figure 13b-k:** DLMAQFDTSSLR replicates
Supplementary Figure 14a: GGYLYHGR reference

Supplementary Figure 14b: GGYLYHGR replicates

Supplementary Figure 15a: TGATVAGPVPLTER reference

Supplementary Figure 15b-k: TGATVAGPVPLTER replicates
Detection in some replicates

Supplementary Figure 16a: IELGDTYTK reference

Supplementary Figure 16b-k: IELGDTYTK replicates

Supplementary Figure 17a: NIPVNIFGDVQQK reference

Supplementary Figure 17b-k: NIPVNIFGDVQQK replicates
Detection in no replicates

Supplementary Figure 20a: GIDFEGFQYGGWSDEYK reference

Supplementary Figure 20b-k: GIDFEGFQYGGWSDEYK replicates

Supplementary Figure 21a: LIPLNVSGPFHALLPEASR reference

Supplementary Figure 21b-k: LIPLNVSGPFHALLPEASR replicates
Supplementary Figure 22a: AIYVTDSNASIGMK reference

Supplementary Figure 22b-k: AIYVTDSNASIGMK replicates
13 Supplementary Results: Dilution series analysis

To validate the quantification properties of Anubis versus a current SRM software, a dilution series experiment was devised. An intra-laboratory standard mixture of 42 synthetic peptides representing tryptic peptides from human proteins (listed below) was diluted in a tryptic digest of a SILAC labelled human cell line, at relative strengths of 1, 3, 10, 30 and 100, with three replicates of each dilution. Dilution strength 100 contained 500 fmol of each synthetic peptide per 300 ng SILAC digest. Five transitions were chosen for each peptide using Skyline based on Orbitrap analysis of the peptide mixture, and SRM assays developed using synthetic peptides. The peptides were treated and the SRM assay was created as described for the S. pyogenes synthetic peptides and each sample was analyzed by scheduled SRM with 5 minute windows.

Supplementary Figure 23: a) Comparison between Anubis and Skyline of reported quantities for the MSGNGAAATAEENSPK peptide. b) Log/log linear regression calculated for reported quantities from Anubis and Skyline for the MSGNGAAATAEENSPK, $m/z = 824.865$ peptide. c,d) Anubis versus Skyline quantification performance, measured in c) relative standard deviation from regression ($\sigma$) and d) coefficient of determination $R^2$.

The data was analyzed separately with both Anubis and Skyline. On a precursor basis we plotted the resultant quantities from Skyline and Anubis, accepting p-
values $\leq 0.01$ from Anubis. Supplementary Figure 23 shows a typical example of such a plot, where we can see a clear correlation between dilution strength and measured quantity as well as a relatively low technical variance. One can also distinguish a slightly better quantification by Anubis, primarily in sample 1dil_01, but also in 100dil_01 and 3dil_01.

To assess their relative performance, a least-squares linear regression was made on the quantities from each program respectively. The quantity versus theoretical quantity dataset was first transformed to log-log space to transform any multiplicative variation into additive and thus try to minimize the relative error of the regression. A least-squares linear regression was made on the transformed values, and two measures of quality were calculated (in log space): the estimated standard deviation from the linear regression $\sigma$ and the coefficient of determination $R^2$ (Fig. 23). Comparing the statistics of all the 40 successfully measured peptides, Anubis had better $\sigma$ in 30 of 40 peptides, and equal or better $R^2$ in all peptides.

The 42 tryptic peptides from human proteins used in the dilution series.

| Peptide sequence |
|------------------|
| YLGPQYVAGITNLK  |
| VVSEDFLQDVSASTK |
| EELGFRPEYSASQLK |
| GDVFTEMPEDEYTVYDDGEK |
| VDFTEEINNMK |
| ATFYGEQVDYYK |
| TEVSLTTLTNK |
| IQLMNSGIGWFQPDVLK |
| ILTQDTPEFFIDQGHAK |
| LEFDLLYPAIK |
| LSFLVNALAK |
| SFPLHFDENSSFAGDK |
| LGAVDSELTSETQK |
| MLLLEILHEIK |
| YLLQETWLEK |
| LLESSLSSSEEGEPEYK |
| VTGLNCTTNHPNPK |
| GLELDPEGSLHHQQK |
| GQVGGQVSVEVDASAPGDLAK |
| ALEAANGELEVK |
| FQTVDDSNIDGFVNCSTK |
| LTQLEILTSGGYIEK |
| GQYCYELDEK |
| SSGCAYQDVGVTCEQPDK |
| EHQPSLQCQCEAVYK |
| QLCYHHSLNNWTK |
| YPIEHGIVTNWDDMEK |
| DSYVGDEAQSK |
| HQGVMVGMGQK |
| AGAHLQGGAK |
| M$\ddot{S}$NGNAAATAEENSPK |
| ENPHDAVVFPK |
VIGGDDLSTLTGK
NVLLIEDIDIATGK
FFADLLDYIK
VPPIKPNAGEESVMNLK
IDEYDYSKPIQGQQK
GFHFTVDNSK
AEIYEAFENIYPILK
TITLEVEPSDTIENVK
TLDYNIQK
GIVDQSQQAYQEAFEISK
Supplementary Figure 24: normality of peptide replicate measurement

Supplementary Figure 24: a) Q-Q plot of 4 lower CV peptides, displaying rough normality. b) Q-Q plot of 4 higher CV peptides, displaying somewhat worse normality.