Quantifying the ~75-95% of Peptides in DIA-MS Files that have been Historically Ignored, Despite their Value

Gautam Saxena[1]; Qin Fu[2]; Aleksandra Binek[2]; Jennifer E Van Eyk[2]
[1]GoldenHaystack Lab, Bethesda, MD; [2]Cedars-Sinai Medical Center, Los Angeles, CA

Abstract

We demonstrate an algorithm termed GoldenHaystack (GH) that, compared to the leading DIA-MS algorithm (DIA-NN), (a) quantifies and identifies with better FDR accuracy the peptides found in FASTA search spaces (~5-25% of analytes in DIA-MS datasets), (b) quantifies the remaining ~75-95% of analytes that were previously unquantified, and (c) runs ~40-200x faster (or ~1-10x faster than the LC-MS). Specifically, without a FASTA or spectral library, GH can deconvolute and accurately quantify chimeric LC-MS spectra. The central idea that enables this claim is: for sufficiently sized projects (e.g., ≥ ~50 LC-MS files), pairs of peptides that co-elute in one subset of LC-MS files do not exactly co-elute in a different subset of files. GH thus analyzes a project holistically: it uses *multi*-partite matching to match fragment ions across all samples, separates and regroups the fragment ions into unique analyte signatures, reduces stochastic noise, and then quantifies those unique analyte signatures (UASs).

Methods (Computational)

We take advantage of the natural variation in LC: pairs of peptides that co-elute in one subset of LC-MS files do not exactly co-elute in another subset of LC-MS files (Fig. 1). This natural "jitter" in the LC domain allows us to computationally separate (i.e., in silico deconvolve) the MS2 (not just MS1) fragments into UASs, which can then be accurately quantified, all without using (and therefore being constrained to) spectral libraries (e.g., FASTA files) search spaces, which are known (a) to be substantially incomplete and (b) to not contain, by definition, the "unexpected" sequences (e.g,. SNPs, disease-specific proteolytic cleavages etc.) and/or dozens of possible PTMs (glycosolations, methylations etc.)

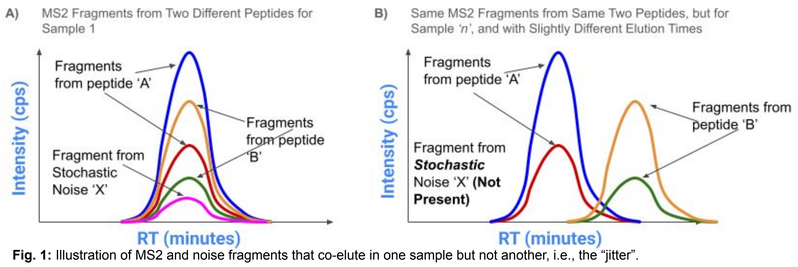


Fig. 1: Illustration of MS2 and noise fragments that co-elute in one sample but not another, i.e., the "jitter".

Results

We performed a dilution series experiment (11 concentrations with 5 technical replications/concentration using pooled human plasma on a Thermo Astral MS) to verify that GH can quantify with linear accuracy the peptides present in human plasma samples. First, we compared the R^2 linearity response for 6 peptides that DIA-NN had shown high R^2 linearity response (Table 1 & Fig. 2). The results were comparable, even though a) these peptides were pre-selected from DIA-NN's perspective as "excellent" peptides and b) DIA-NN performed a custom normalization within each batch of 5 replicates which GH did not do.

Sequence	Charge	GH R^2	DIA-NN R^2
CEACPPGYSGPTHQG VGLAFAK	3	0.98	0.99
ESDTSYVSLK	2	0.97	0.99
IADVTSGLIGGEDGR	2	0.97	0.99
ILEGFQPSGR	2	0.98	1.00
SDVVYTDWK	2	0.99	0.99
YVGGQEHFAHLLILR	2	1.00	1.00
	Mean:	~0.982	~0.993

Table 1: 6 peptides pre-selected from DIA-NN's perspective as having high quantitative R^2 linearity response

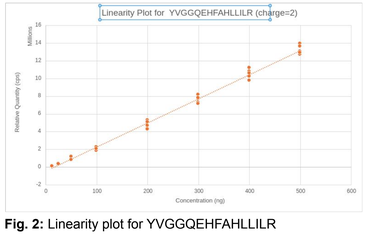


Fig. 2: Linearity plot for YVGGQEHFAHLLILR

Results (Continued)

We also inspected the XIC plots for each of those six pre-selected peptides, one of which is shown in Fig 3.

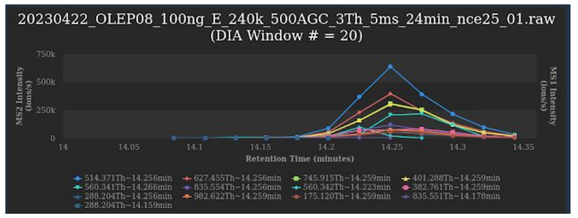


Fig. 3: XIC plot for YVGGQEHFAHLLILR

Next, we analyzed the linearity R^2 response for all the ~6000 PSMs at 1% FDR (Table 2). Interestingly, GH had a noticeably higher percentage of peptides quantified with R^2 above 0.90 than DIA-NN (83% vs 75%), despite the fact that DIA-NN had tremendous advantage in quantitation since it a) used knowledge from the FASTA file for quantitation, b) performed normalization within each batch of 5 technical replicates (which required custom running of DIA-NN) and c) focused exclusively on peptides identified in the FASTA search space.

Identified Analytes				
	GH		DIA-NN	
R^2 Correlation Range	# Quantified	% Quantified	# Quantified	% Quantified
0.90 to 1.00	5292	**83%**	6,611	75%
-1 to 0.90	1089	17%	2,245	25%

Table 2: Number of Identified Analytes with R^2 ≥ 0.90

Finally, we count the number of *unidentified* peptides whose R^2 linearity quantitation response was above 0.90. We note that (a) there is >10x ((33890+27708+4014) / 5292) more unidentified peptides at quantitative R^2 linearity response ≥ 0.90 than identified peptides and (b) the high-quality unidentified analytes had approximately the same distribution of peptides (83%) with R^2 ≥ 0.90 as the identified peptides (80%).

Unidentified Analytes						
	High-Quality		Medium-Quality		Low-Quality	
R^2 Correlation	# Quantified	% Quantified	# Quantified	% Quantified	# Quantified	% Quantified
0.90 to 1.00	33890	**80%**	27708	48%	4014	23%
-1 to 0.90	8285	20%	29771	52%	13471	77%

Table 3: Number of Unidentified Analytes with R^2 ≥ 0.90

Bonus Results

Though the focus of GH is quantitation – and specifically, quantitation of the unidentified peptides – we can also use the "cleaned-up spectra" to identify spectra from a FASTA search space. We do so for a 61 sample Covid-19 plasma study analyzed on a Thermo Exploris. Interestingly, while GH finds all 11 of 11 iRT peptides (and correctly identifies one peptide that elutes twice), DIA-NN only finds 10. Also, when performing the search, we request both GH and DIA-NN to search for variable modifications: oxidation, phosphorylation, as well as stably isotopically label (SIL) modifications. However, we never spiked in any SIL peptides into the samples. Therefore, the true number of SIL peptides is 0, though acceptable answers would be between 0 and ~1% FDR. GH found close to 0 SIL peptides (i.e., 2) wheas DIA-NN claimed to find 237, a 11750% increase.

A) "Positive Control" -- 11 iRT Peptides Spiked In

	GH		DIA-NN	
Peptides (all charge 2)	Avg RT (min)	# Samples	Avg RT (min)	# Samples
ADVTPADFSEWSK	12.97	61	13.22	61
DGLDAASYYAPVR	11.52	61	11.65	61
DGLDAASYYAPVR	12.09	60	n/a	0
GAGSSEPVTGLDAK	5.79	61	5.90	61
GTFIIDPAAVIR	16.24	61	16.92	61
GTFIIDPGGVIR	14.68	61	n/a	0
LFLQFGAQGSPFLK	18.35	34	18.79	32
LGGNEQVTR	3.23	60	3.23	61
TPVISGGPYEYR	9.60	61	9.74	61
TPVITGAPYEYR	10.06	59	10.26	61
VEATFGVDESNAK	7.31	61	7.47	61
YILAGVENSK	8.42	61	8.61	61

B) "Negative Control" -- 0 SIL Peptides Spiked In

	GH	DIA-NN	% Increase
Claimed FDR Rate	1%	1%	n/a
Claimed # True PSMs	1629	2822	73%
True # SIL PSMs	0	0	n/a
Claimed # SIL PSMs	2	237	**11750%**

Table 4: iRT (i.e., "Positive Control") and SIL analysis ("Negative Control") for 61 sample human plasma Covid 19 study

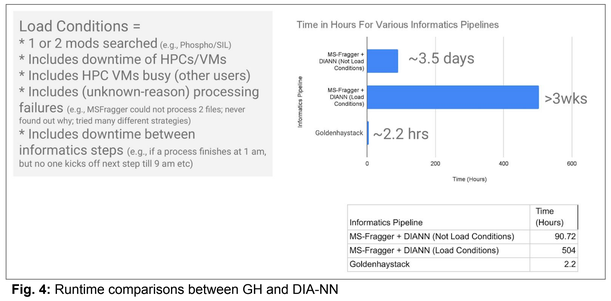
Bonus Results (Continued)

We also ran GH against one of three fractions of a 57 sample human heart cell tissue study that was run on a Thermo Exploris. Interestingly, although DIA-NN claimed to find ~100% more PSMs than GH, when we looked for the *top* PSMs in DIA-NN but not in GH, the first five of those top DIA-NNs were self-evidently false matches: they all had a SIL modification, even though no SIL peptides were spiked in (and SIL modifications are not known to occur naturally).

Peptide	Charge	RT (min)	Q-value (sorted asc.)	Quant (Norm.) (sorted desc.)
VADALTNAVAHVDDM(UniMod:35)PNALSALS DLHAHK(UniMod:259)LRVDPVNFK	5	20.698	5.28E-5	1.48E8
DAK(UniMod:259)MK(UniMod:259)APSSLAVSPD GTLYVADLGNVR	3	73.979	5.28E-5	5.57E7
SYELPDGQVITIGNER(UniMod:267)FRC(UniMod: 4)PEALFQPC(UniMod:4)FLGMESC(UniMod:4)GI HK(UniMod:259)	5	48.754	5.28E-5	5.53E7
SEVAHR(UniMod:267)FKDLGEENFK	4	18.053	5.28E-5	4.44E7
GATPAR(UniMod:267)ELFR	2	35.259	3.24E-4	5.41E7

Table 5: Top 5 peptides in DIA-NN but not in GH (all self-evidently false matches)

Finally, although not the initial focus of this first iteration of GH, we discuss speed of execution: GH was up to ~10x faster than the LC-MSs, and up to ~200x faster than DIA-NN when considering real-life load conditions (Fig. 3).



Informatics Pipeline	Time (Hours)
MS-Fragger + DIANN (Not Load Conditions)	90.72
MS-Fragger + DIANN (Load Conditions)	504
Goldenhaystack	2.2

Fig. 4: Runtime comparisons between GH and DIA-NN

Summarizing Table

Data Set #	GH Quantification Metrics:			Quality Metrics:				Runtimes (In Hours):		
				GH		DIA-NN				
	# PSMs In FASTA	# Analytes Not In FASTA	% Analytes Not In FASTA	# iRT	# SIL	# iRT	# SIL	LC-MS	GH (on a single 180 core computer)	DIA-NN (on cluster) no load conditions
1	5,938	117,560	**95**	11	n/a	11	n/a	~24	**22.8**	~68
2	2,388	33,138	**93**	12	**2**	10	237	~24	**2.5**	>90
3	13,271	40,208	**75**	11	204	11	653	~86	**4.1**	>90
4	6,495	121,969	**95**	11	95	(could not run)		~24	**23.4**	(could not run)

Table 6: Summary Table of Key GH Attributes