Experiment-Specific Spectral Libraries Unlock Additional Peptide Identifications from Public Orbitrap Astral DIA Data

P. Askovich[1], B. Wen[2], K. Teplitz[1], M. MacCoss[2]

[1]EuropaDX, Inc., Redmond, WA [2]Department of Genome Sciences, University of Washington, Seattle, WA

TAKE-HOME Carafe spectral libraries recovered 5–6% more precursors and 3% more protein groups from the same Orbitrap Astral runs. The additional IDs were lower-abundance proteins, and quantification reproducibility was unchanged.

Methods

Public Orbitrap Astral dataset, reprocessed with DIA-NN.

BACKGROUND

DIA-NN searches DIA data using either an in-silico library predicted from FASTA (library-free) or a supplied spectral library. Carafe trains deep-learning models directly on DIA spectra to produce experiment-specific libraries [Wen 2025]; published benchmarks report 5–38% gains over DDA-trained predictions. We asked whether this benefit extends to public Orbitrap Astral data already analyzed with DIA-NN library-free.

DATA

PXD054015 [Hendricks 2024] was retrieved from ProteomeXchange via the EuropaXp OmicsCloud built-in importer (server-to-server transfer). Two tissues: depleted human plasma (n = 9, three days × three replicates, 2 Th windows) and mouse lung (n = 9, three 2 Th + six 3 Th). All runs: 24 min gradient on an Orbitrap Astral. Raw files were converted to mzML on upload.

PROCESSING

Each raw file was processed twice with DIA-NN 1.8.1: once library-free, and once with a Carafe-trained library. All search parameters were held constant: 1% FDR, trypsin (no P rule), 1 missed cleavage, peptide length 7–35, charge 2–4, fixed Carbamidomethyl (C), variable Met-Ox (max 1), MBR (--reanalyse), RT-dependent cross-run normalization (--rt-profiling). Reference proteomes from UniProt: UP000005640 (Homo sapiens) for plasma, UP000000589 (Mus musculus) for lung, each appended with the cRAP common contaminants list. Only the spectral library differed between the two runs.

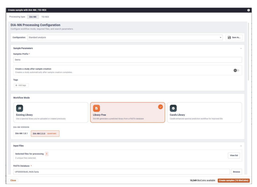
ANALYSIS

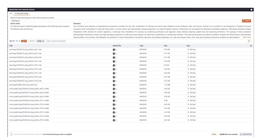
Data were processed both in EuropaXp OmicsCloud and manually using the Carafe 2.0 UI. Per-run identification counts were taken from DIA-NN stats.tsv; set membership, abundance distributions, and CV across 9 replicates were computed from pg_matrix.tsv and pr_matrix.tsv. Figures were prepared in EuropaXp, Python (pandas, numpy, matplotlib), and Free Online Bioinformatics Tools (europadx.com/tools).

MATCHED DIA-NN SETTINGS

Parameter	Library-free	Carafe library
DIA-NN version	1.8.1	1.8.1
FDR	1%	1%
Precursor charge	2–4	2–4
Peptide length	7–35	7–35
Missed cleavages	1	1
Variable mods	Met-Ox (max 1)	Met-Ox (max 1)
Spectral library	**DIA-NN predictor**	**Carafe-trained**

EUROPAXP WORKFLOW



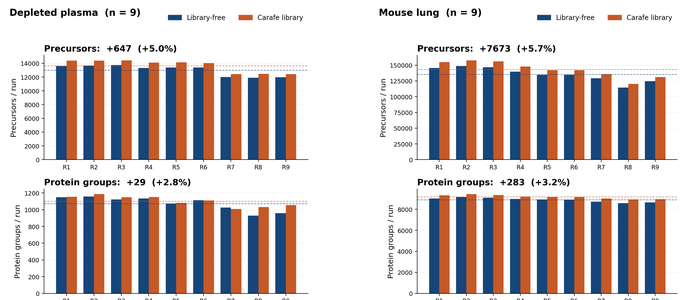


Identification gains

Carafe identified more precursors and protein groups in every run.

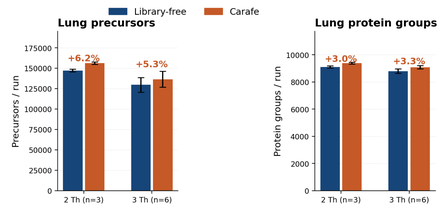
Across 18 runs in two tissues, Carafe libraries identified more precursors and more protein groups than DIA-NN library-free. The gain was present in every replicate, not driven by outliers.

PER-RUN IDENTIFICATIONS (1% FDR)

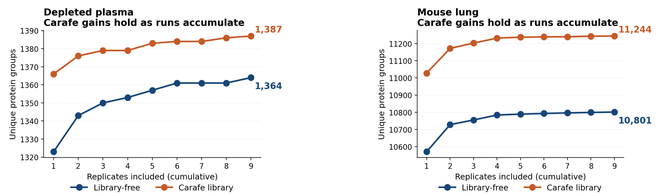


Per-run means from DIA-NN stats.tsv. The +29 / +283 gains above are mean PG differences per single run, averaged across the 9 replicates of each condition (see Set Composition for union-of-runs view).

ROBUST ACROSS DIA WINDOWS (lung)



CUMULATIVE DISCOVERY (gains persist as runs accumulate)

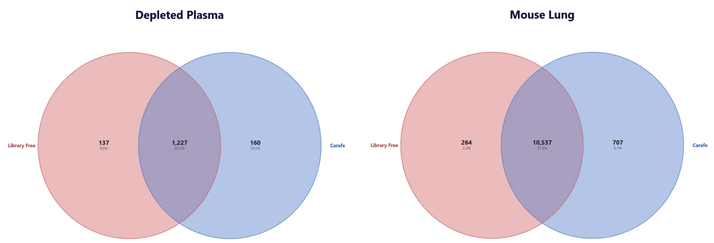


Gains were observed in all 18 runs and at both DIA window widths, indicating the additional IDs come from the library rather than the acquisition method.

Composition and quality of the additional IDs

The additional IDs are lower-abundance proteins. CV is unchanged.

SET COMPOSITION (protein groups across 9 replicates)

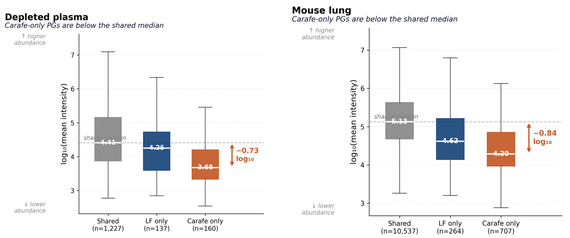


	Library-free	Carafe	Shared	LF only	Carafe only	Net gain
Depleted plasma	1,364	1,387	1,227	137	160	+23
Mouse lung	10,801	11,244	10,537	264	707	+443

Each Venn shows unique Protein Group IDs pooled across the 9 replicates per condition. This pooled view counts a protein group once if it was seen in any run, so it measures total proteome coverage rather than per-run yield. Carafe added 23 protein groups in plasma and 443 in lung at this pooled level.

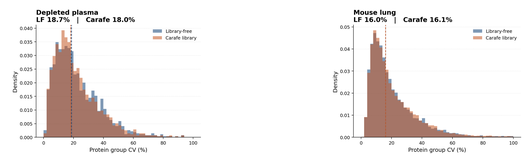
ABUNDANCE OF METHOD-SPECIFIC PROTEIN GROUPS

For each protein group we computed the $\log_{10}$ mean intensity across its 9 replicates, then split PGs into three sets from the Venn above: Shared, LF only, Carafe only. The dashed line marks the Shared median, the abundance level common to both methods.



Both tissues show the same pattern. Carafe identifies proteins library-free does not, rather than re-quantifying existing PGs at the noise floor.

REPRODUCIBILITY — CV ACROSS 9 REPLICATES



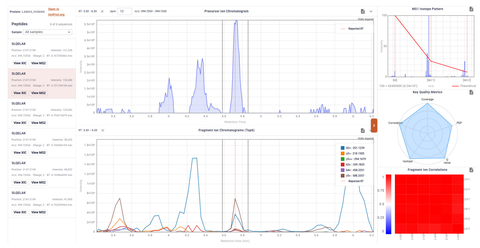
Median CV is essentially identical between methods in both tissues. The extra IDs do not add quantification noise.

Implications

The Carafe-only hits show clean chromatographic peaks.

EuropaXp OmicsCloud single-protein view for SLQELAK (LAMA3_HUMAN), a Carafe-only identification. The precursor ion chromatogram shows a clear peak at the reported retention time, with the top-6 fragment ions co-eluting at the same RT. The MS1 isotope pattern matches theoretical, and quality metrics (coverage, correlation, isotope, PEP, Q-value) confirm a high-confidence identification.

EXAMPLE: CARAFE-ONLY PRECURSOR



CONCLUSIONS

Increased identification yield.

Carafe libraries identified +5.0%/+5.7% more precursors and +2.8%/+3.2% more protein groups (plasma/lung), with all DIA-NN parameters held constant.

Gains are enriched at low abundance.

Carafe-only PGs sit 0.75–0.85 $\log_{10}$ below shared PGs: closer to the detection limit.

No loss of quantification quality.

Median CVs across 9 replicates were 18.7 to 18.0% in plasma and 16.0 to 16.1% in lung.

REFERENCES

[1] Hendricks NG et al. J Proteome Res 23:4163–4169 (2024).
[2] Wen B et al. Nat Commun 16:9815 (2025).
[3] Demichev V et al. Nat Methods 17:41–48 (2020).

KEY FINDINGS

+5.0% / +5.7%
Precursors per run
plasma / lung

+2.8% / +3.2%
Protein groups per run
plasma / lung

−0.75 / −0.85
$\log_{10}$ abundance shift
Carafe-only PGs

ANALYSIS PLATFORM

DIA-NN searches were executed on EuropaXp OmicsCloud (EuropaDX, Inc., Redmond, WA). The platform supports DIA-NN library-free, existing-library, and Carafe library modes through a single configuration. Manual local DIA-NN runs were performed to confirm identical outputs.

CONFLICT OF INTEREST

P. Askovich and K. Teplitz are affiliated with EuropaDX. M. MacCoss is a scientific advisor for EuropaDX. B. Wen is affiliated with the Department of Genome Sciences, University of Washington. Carafe is open-source software developed at the University of Washington. Dataset PXD054015 was generated by Cedars-Sinai Medical Center.

CONTACT & ACKNOWLEDGMENTS

Acknowledgments: We thank Cedars-Sinai for PXD054015 and the Carafe team at the University of Washington. The MacCoss lab receives support from Agilent, Sciex, Shimadzu, ThermoFisher, and Waters.
Contact: peter.askovich@europadx.com | europadx.com