quantmsdiann: a scalable SDRF-driven DIA-NN workflow for reanalysis of single-cell, spatial, and bulk proteomics datasets

Qi-Xuan Yue
Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China. https://orcid.org/0009-0004-8120-9339

Yufei Shen
Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China.

Chengxin Dai
State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing, China https://orcid.org/0000-0001-6943-5211

Asier Larrea-Sebal
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK https://orcid.org/0000-0001-9107-4299

Henry Webel
Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark https://orcid.org/0000-0001-8833-7617

Jose Nimo
Spatial Proteomics Group, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany https://orcid.org/0000-0002-1565-7799

Fabian Coscia
Spatial Proteomics Group, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany https://orcid.org/0000-0002-2244-5081

Orhun Kok
Department of Bioengineering and Barnett Institute, Northeastern University, Boston, Massachusetts, USA

Nikolai Slavov
Department of Bioengineering and Barnett Institute, Northeastern University, Boston, Massachusetts, USA https://orcid.org/0000-0003-2035-1820

Juan Antonio Vizcaíno
EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK https://orcid.org/0000-0002-3905-4335

Timo Sachsenberg

Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen, Germany https://orcid.org/0000-0002-2833-6070

Mingze Bai

Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China https://orcid.org/0000-0002-9782-2056

Vadim Demichev

Quantitative Proteomics Laboratory, Charité – Universitätsmedizin Berlin, Berlin, Germany https://orcid.org/0000-0002-2424-9412

Yasset Perez-Riverol

yperez@ebi.ac.uk

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK https://orcid.org/0000-0001-6579-6941

Research Article

Keywords: DIA-NN, single-cell proteomics, spatial proteomics, phosphoproteomics, plexDIA, data-independent acquisition, Nextflow, nf-core, SDRF, reproducibility, QPX, pmultiqc

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10319687/v1

Additional Declarations: The authors declare potential competing interests as follows: V.D. holds shares of Aptila Biotech. The other authors declare no competing interests.

Abstract

Public proteomics archives now hold thousands of data-independent acquisition (DIA) datasets, but reusing them is difficult: each was processed with a different software configuration, and most lack standardized metadata. Here, we present quantmsdiann, an open-source Nextflow/nf-core workflow that runs DIA-NN in parallel across cloud and high-performance computing (HPC) infrastructure, guided by the experimental design declared in SDRF format. The workflow provides pinned container profiles and builds recipes for each supported DIA-NN version under BioContainers, resolving dependencies automatically, and reads all major vendor formats and the HUPO-PSI mzML standard. It exports harmonized quantification tables as MSstats input, in the Quantitative Proteomics eXchange (QPX) format, and a pmultiqc quality-control report. The parallel design reanalyzes a 2,300-run single-cell dataset in 2.2 hours on 300 HPC nodes. We performed multiple experiments and benchmarks of quantmsdiann on single-cell datasets; ProteoBench DIA-NN single-machine submissions or public datasets in ProteomeXchange. The benchmark against ProteoBench single-machine DIA-NN modules demonstrated no differences between DIA-NN single-machine runs and parallelization in quantmsdiann; while upgrading DIA-NN from 1.8.1 to a current release increased protein-group identifications by up to 17% in single-cell datasets. Remarkably, reanalysis of public DIA datasets with quantmsdiann and the latest version of DIA-NN always exceeds the originally deposited counts, recovering up to 59% more protein groups, with the largest gains on deposits processed with older or non-DIA-NN engines and smaller gains where a recent DIA-NN release was already used. quantmsdiann is a step toward scalable, reproducible reanalysis of the growing DIA data archive and a foundation for large-scale DIA meta-analysis and atlas building. It is available at https://github.com/bigbio/quantmsdiann.

Introduction

Data-independent acquisition (DIA) is a rapidly growing mass-spectrometry strategy for proteomics at scale, now approaching parity with data-dependent acquisition among public proteomics deposits, and DIA-NN (1) is among the most widely adopted analysis engines across timsTOF (2, 3), Astral (4), and Orbitrap platforms. Recent benchmarks confirm DIA-NN's competitiveness for single-cell DIA workflows (5), and recent releases have expanded support for plexDIA (6, 7), timsTOF parallel accumulation–serial fragmentation (PASEF), and vendor-native raw-file reading. At the same time, DIA dataset sizes and run counts are growing rapidly (8, 9), and the proteomics community is increasingly interested in reanalyzing public deposits to recover additional identifications, integrate across cohorts, and build large-scale DIA atlases (10, 11). DIA-NN's command-line interface, simplicity, and analytical power make it a natural engine for mining public archives, but realizing it at scale requires distributing the computation: extracting more information from public data means reprocessing cohorts of hundreds to thousands of runs fast enough that reanalysis can be iterative rather than a one-off effort. The same need extends beyond reanalysis: laboratories generating their own large cohorts increasingly have access to cloud and high-performance computing (HPC) infrastructure and benefit from distributing the analysis rather than running it on a single workstation. The steps of a DIA-NN run are heterogeneous and map naturally onto heterogeneous hardware: many lightweight per-file passes that fit small nodes, and a few memory- and

compute-heavy cohort-level steps that need large nodes. Efficient large-scale analysis therefore depends on software containers and workflow engines that can place each task on appropriate cloud or HPC resources (12). Despite this analytical maturity, most DIA-NN users run the tool as a desktop application, which limits its scalability for en masse processing of very large datasets.

In 2024, we published bigbio/quantms (11), a Nextflow pipeline that standardized multi-engine DIA and DDA analyses behind a Sample and Data Relationship Format (SDRF) (13) input on the nf-core framework (14). quantms was, however, built around DIA-NN 1.8.1 with a fixed parameter set, and its DDA/DIA combination within the same workflow made extending DIA-specific functionality difficult. New DIA-NN parameters conflicted with DDA options, versions were hard-coded, sample-level acquisition parameters were not properly propagated to per-run DIA-NN invocations, and raw-file conversion was mandatory. Iterative reanalysis at archive scale requires a workflow that combines the reproducibility of nf-core, the sensitivity of current DIA-NN releases, the standardized sample metadata required by the SDRF format, and a parallelization strategy capable of processing thousands of raw files.

Here, we present quantmsdiann, an independent SDRF-driven Nextflow pipeline dedicated to DIA-NN that extends DIA-specific features beyond what the original bigbio/quantms pipeline could accommodate. While quantmsdiann shares its SDRF-first design philosophy and parts of the configuration-generation tooling with quantms, its workflow, container set, and module library are developed and released independently, with container-pinned profiles for every supported DIA-NN version (1.8.1 through 2.5.1, including the enterprise build) selectable at runtime under Docker or Singularity. Downstream, it exports a harmonized QPX (Quantitative Proteomics eXchange) Parquet archive alongside the standard DIA-NN reports and MSstats (15) input. We demonstrate the pipeline along three axes: ProteoBench (16) benchmarks across four DIA modalities and successive DIA-NN releases; independent reanalyses of public DIA deposits spanning bulk cell lines (NCI-60, ProCan), single-cell datasets, a spatial Deep Visual Proteomics cohort, and a phosphoproteomics cohort; and a pan-cohort of five public DIA cell-line reanalyses integrated under a single SDRF-driven invocation. Across the single-cell and bulk cell-line cohorts, reanalysis with the current DIA-NN build recovers up to 59% more protein groups, with the largest gains on the oldest deposits processed with older or non-DIA-NN engines and smaller gains where a recent DIA-NN release was already used, so reprocessing archived data with an up-to-date engine yields measurably more from each deposit.

Experimental Procedures

Pipeline implementation

quantmsdiann is implemented in Nextflow DSL2 (> = 25.10.4) and follows nf-core community standards (14) for pipeline structure, testing, documentation, and continuous integration. It supports both DIA-NN library-free analysis (predicting the spectral library in silico from a FASTA reference) and library-based analysis against a user-supplied spectral library; all analyses in this study used the library-free mode; the reanalysis cohorts used the UniProt human reference proteome (17) (UP000005640, downloaded March

2026), while the ProteoBench HYE benchmarks used the corresponding three-species (human, yeast, *E. coli*) FASTA. Each computational step is packaged as a versioned container image (Docker, Singularity, or Apptainer). Open-source community tools ThermoRawFileParser (18), pridepy (19), QPX, and the wiffconverter SCIEX reader are pulled from public registries BioContainers (20) or ghcr.io/bigbio. Because the DIA-NN license restricts redistribution, only version 1.8.1 is shipped as an image (the default, from BioContainers); releases from 1.9 onward (1.9.2–2.5.1, including the 2.5.1-enterprise build) are built locally from the per-release recipes in the companion quantms-containers repository (https://github.com/bigbio/quantms-containers) and tagged so the workflow selects them automatically (**Supplementary Note 1**). DIA-NN is free for academic use, and version 1.8.1 is freely redistributable; for releases after 1.8.1, academic users should contact the DIA-NN team before cloud deployment, and the 2.5.1-enterprise build requires a separate commercial license (see the DIA-NN license for the applicable terms). The repository at https://github.com/bigbio/quantmsdiann is organized as a standard nf-core pipeline with main workflow (main.nf), subworkflows, modules, test data and CI configurations, and documentation. Pipeline version v2.2.0 (release codename “Chongqing”, 2026-06-16) was used for the analyses reported in this paper.

SDRF extensions for quantmsdiann

quantmsdiann accepts SDRF (13) as primary input. SDRF parsing is performed using the sdrf-pipelines Python package developed within the quantms project, which validates the SDRF against the proteomics-specific subset of EFO, PSI-MS, and PRIDE ontology terms and emits a normalized per-sample channel for downstream Nextflow processes. To express DIA- and DIA-NN-specific acquisition settings at the sample level, we extended the SDRF-Proteomics format (version 1.1.0; specification and templates at https://github.com/bigbio/proteomics-sample-metadata) with additional comment columns: the precursor (MS1) and fragment (MS2) m/z scan ranges and, for multiplexed experiments, the plexDIA channel definitions and their isotopologue labels. Because these settings are declared per sample, independent datasets acquired with different scan-range schemes are each parametrized from their own metadata and can be processed together in one invocation, with no change to the workflow. We added a convert-diann command to sdrf-pipelines that reads these extended columns together with the standard fields (enzyme, fixed and variable modifications, instrument, precursor mass tolerance, fragment mass tolerance, and missed-cleavage settings) and writes a per-run DIA-NN command-line configuration; where SDRF columns are absent, DIA-NN defaults appropriate for the inferred instrument are applied.

Quality control and QPX export

Per-run and per-cohort quality control reports are generated by pmultiqc (21), a proteomics extension of MultiQC (22) that summarizes DIA-NN-specific metrics including peptide and protein counts, missed cleavages, charge-state distributions, precursor and fragment mass accuracy, and per-run identification yield. The QPX archive is produced by the qpxc CLI from the qpx package

(https://github.com/bigbio/qpx), which writes Parquet files for identifications (peptide-spectrum matches (PSMs), peptides, protein groups, features), metadata (sample, run, ontology), and provenance, and optionally exports MuData (.h5mu) containers for scverse-based (23) downstream analysis. The aim of exporting to QPX is to bundle the SDRF, identifications, quantifications, and pipeline provenance into a single queryable artifact for reanalysis and long-term archival, and to provide a standardized input for downstream tools that support the format (e.g. MSstats, Perseus, or custom scripts).

Datasets benchmarked and compute environment

Supplementary Note 2 lists all benchmark and reanalysis datasets with their instrument, acquisition mode, and deposition year, and an SDRF prepared from the original sample metadata. They fall into four experiment classes: *ProteoBench* DIA benchmarks (modules 5, 7, 9, and 10); *single-cell and spatial proteomics*, comprising the HeLa single-cell cohort PXD046357 (Orbitrap Astral), the oocyte plexDIA cohort MSV000093870, the PXD064049 spatial Deep Visual Proteomics cohort, and the PXD071075 single-cell scalability sweep; *bulk cell-line panels* (PXD003539, PXD030304, PXD004701, PXD041421, and PXD017199); and *phosphoproteomics* (PXD034128, PXD034623, and PXD049692). For each cell-line reanalysis cohort, the original published quantification was downloaded from a public repository as the comparator. quantmsdiann counts follow a single global rule (protein groups at Lib.PG.Q.Value < = 0.01 precursors at Lib.Q.Valu e <= 0.01, decoys dropped; no contaminant or positive-quantity filter); the like-for-like-threshold rule applied to each original headline count is described in **Supplementary Note 3**. Runtime benchmarks were run on the EMBL-EBI HPC cluster (LSF scheduler), with per-task CPU and memory dispatched by Nextflow and memory profiled from the Nextflow trace report (-with-trace).

Benchmarking and cross-study comparison

A distributed workflow is only useful if it reproduces a single-machine DIA-NN run, so we benchmarked quantmsdiann against the ProteoBench community submissions (16) on the four public DIA modules (Orbitrap Astral, single-cell Astral, timsTOF diaPASEF, and SCIEX ZenoTOF 7600). All quantmsdiann runs were library-free (the library predicted in silico from the FASTA), so for a like-for-like comparison we kept only predicted-from-FASTA DIA-NN community submissions and report the oldest supported release (1.8.1) and the current 2.5.1-enterprise build at DIA-NN's recommended per-version q-values under a uniform 1% global precursor cut-off. For the public reanalyses, each deposit's original published quantification served as the comparator, counted under that study's own threshold.

Results

quantmsdiann pipeline architecture

Built on the nf-core framework (14), quantmsdiann accepts a single SDRF file as primary input, which encodes per-sample metadata (cell line or tissue origin, factor values, instrument, post-translational modifications) and per-run file references. We extended the SDRF with DIA-relevant columns (MS1 and

MS2 scan ranges, and plexDIA channel definitions and labels) and added an adapter to sdrf-pipelines (13) (convert-diann) that translates each run's declared metadata into a per-run DIA-NN command-line configuration. Every run is therefore parametrized from its own metadata with no manual configuration or parameter-harmonization step, removing a common source of analysis-script divergence between groups: independent datasets with different scan-range schemes, or a multiplexed plexDIA design, are each analyzed directly from their declared SDRF settings without any workflow change. The workflow comprises four mandatory and three optional modules organized around the DIA-NN analysis core (Fig. 1), alternating between per-file parallel stages (red; raw-file preparation, preliminary calibration, and individual search) and collective stages (green; in-silico library generation, empirical library assembly, final quantification, MSstats conversion (15), and pmultiqc (21) quality-control reporting) operating on the full cohort.

quantmsdiann follows DIA-NN's distributed multi-pass flow, alternating parallel per-file work with serial cohort-wide steps (Fig. 1). In summary, each raw file is first prepared *in parallel* (vendor-format conversion, decompression, or indexing) where needed; conversion to mzML is mandatory for early DIA-NN versions and for some vendor formats such as SCIEX .wiff. A spectral library is then predicted in silico from the FASTA once (*serial*; skipped when a library is supplied), and every file is calibrated against it in a *parallel* first pass. The first-pass results are merged into a single empirical library (*serial*); each file is searched again against that library *in parallel*; and a final consolidation pass merges all per-file results into the cohort quantification, applying match-between-runs, normalization, and protein inference (*serial*). MSstats conversion, pmultiqc reporting, and QPX export then run once at the end (*serial*). The two per-file search passes carry essentially all of the compute and scale across cluster nodes, driving the throughput reported below; the serial steps set a largely fixed wall-clock floor. Outputs include the standard DIA-NN report tables: report.tsv (or report.parquet), report.pr_matrix.tsv, and report.pg_matrix.tsv, alongside MSstats-formatted input and a pmultiqc quality-control report. A harmonized QPX archive bundles identifications, metadata, and provenance into a Parquet-based artifact, queried directly with DuckDB and, for the quantification layers, exported to MuData (.h5mu) for scverse-compatible downstream analysis.

quantmsdiann throughput, scaling, and analytical equivalence

To characterize how quantmsdiann scales, we reanalyzed an entire single-cell experiment of ~ 2,300 raw files (PXD071075, Orbitrap Eclipse) across cluster sizes from 10 to 300 nodes and measured end-to-end wall-clock time (Fig. 2a). Because the per-file work is parallelized across nodes, a single SDRF-driven invocation completed the whole cohort within hours, falling from 37.4 h at 10 nodes to 2.2 h at 300. Scaling stays close to the ideal $1/N$ reference up to ~ 100 nodes for this dataset and then bends away, with a practical plateau near 200 nodes beyond which added nodes return progressively less. This is unlikely to be an intrinsic limit of the workflow; it probably reflects the fixed serial steps that run once

regardless of node count, together with contention for shared-cluster I/O and scheduler slots on this allocation.

We then compared time-to-finish across every dataset in the study (Fig. 2b), each reported net of the one-time in-silico library-prediction step, which cannot be parallelized and whose cost depends on the library-generation parameters. Across cohorts spanning three orders of magnitude in file count, wall-clock time stayed within hours because the per-file passes absorb additional files in parallel: the 5,798-run ProCan cohort (PXD030304) finished in 6.2 h, in the same band as cohorts orders of magnitude smaller. Time-to-finish is therefore governed by per-run search cost and cohort-level consolidation rather than by file count; the longest reanalysis was a 206-run cohort on an older Q Exactive instrument (PXD017199, 11.4 h), exceeding the far larger ProCan cohort. Peak memory during consolidation stayed within the 64–128 GB available on commodity HPC nodes, so no specialist high-memory hardware is required. The plexDIA and phosphoproteomics cohorts, processed to demonstrate workflow support rather than for cross-engine benchmarking, likewise complete within hours (Fig. 2b), with per-step runtime and resource envelopes for the benchmark and sweep cohorts in **Supplementary Note 4 Figs. S1–S2**. To confirm the workflow is not tied to one infrastructure, an independent group ran the workflow on a separate (non-EBI) SLURM cluster with the same container-pinned profiles and observed the same per-step runtime structure (**Supplementary Fig. S1b**).

One concern when parallelizing a tool such as DIA-NN is whether the distributed results remain comparable to those from a single-machine, serial run. On the two ProteoBench modules with predicted-from-FASTA DIA-NN community submissions (Orbitrap Astral, n = 15; timsTOF diaPASEF, n = 8), quantmsdiann reproduces the expected HYE per-species $\log_2$ fold-changes, and across both releases its measured ratios sit within the spread of the standalone desktop submissions (Fig. 2c). Distributed quantmsdiann thus matches desktop DIA-NN quantification accuracy on these two modules, within the dispersion of the community submissions.

Identification depth, by contrast, increases with each DIA-NN version on every module. Per-run protein groups rise consistently from 1.8.1 through 2.5.1 to the enterprise build across all four DIA modalities (Fig. 2d), a gain of 10–12%: for example, from 7,305 to 8,190 protein groups per run on the single-cell Astral module (Module 9) and from 10,192 to 11,432 on Orbitrap Astral (Module 7). Precursors and the number of proteins quantified in every run increase in parallel (per-version precursor and complete-profile counts in **Supplementary Note 5**). This tracks the broader ProteoBench community trend (16) and motivates reanalysis of public deposits with the current build.

Single-cell DIA reanalysis improves with current DIA-NN releases

Single-cell proteomics has rapidly advanced, with significant contributions from computational methods (24) and community recommendations for performing, benchmarking, and reporting experiments (25). To assess quantmsdiann on low-input data, where sensitivity and match-between-runs matter most, we

reanalyzed two label-free Orbitrap Astral single-cell cohorts, HeLa (PXD046357) and a larger A549/H460 cohort of 146 single cells (PXD049412), through the workflow under DIA-NN 1.8.1 and the current 2.5.1-enterprise build, both parametrized from the same SDRF (Fig. 3). Upgrading from 1.8.1 to 2.5.1-enterprise increased identifications on both cohorts (Fig. 3a). On HeLa Astral, precursors and protein groups each rose by ~ 17%, and the gain held per cell. On the deeper A549/H460 cohort, the reanalysis reproduced the study's reported Astral single-cell depth (4), a median of ~ 3,400 protein groups per cell; here the precursor gain exceeded the protein-group gain, consistent with the near-saturated protein depth of this larger cohort. The improvement is largest in the data-completeness profile (the number of protein groups quantified across an increasing fraction of cells; Fig. 3b), consistent with stronger cross-run propagation, while quantitative precision is preserved (Fig. 3c). As in the cross-deposit comparisons, these version gains are identifications recovered under a current, FDR-controlled engine rather than strictly FDR-matched, and DIA-NN calibration in library-free, low-input mode remains an open question (26, 27).

The value of reanalyzing public DIA deposits at scale

Public DIA deposits are typically processed once, with the search engine available at the time of deposition, and are seldom revisited, leaving recoverable identifications in raw data that has already been collected and published (28). Reanalyzing these deposits with a current DIA-NN build recovers them: across every deposit reanalyzed here (single-cell, bulk cell-line, and spatial), reprocessing yielded more protein groups than the originally deposited or published analysis, and more precursors wherever the original was itself DIA-NN (Fig. 4). Three caveats bound how these cross-deposit comparisons should be read. First, *FDR calibration*: the comparisons assume that recent DIA-NN (2.5/2.5.1) exercises largely correct false-discovery-rate (FDR) control, whereas older releases (the 1.7-series and 1.8.1) and other search engines may carry substantially higher true FDR at the same nominal 1% q-value (5, 26, 27), and verifying each deposit's FDR calibration is beyond the scope of this work; the reported gains should therefore be read as identifications recovered under a current, FDR-controlled engine, not as strictly FDR-matched differences. Second, *counting criterion*: for the same reason, we compare counts only under each original study's own threshold. Third, *database and search settings*: the reanalysis applies one harmonized FASTA and parameter set across all cohorts, whereas deposited analyses used their own databases and settings, including contaminant databases that are often not deposited (as for the spatial Deep Visual Proteomics cohort, PXD064049); residual count differences attributable to such undocumented choices cannot be fully excluded.

To show the value of running one SDRF-driven workflow across many independent deposits, we reanalyzed five public cell-line DIA panels in a single quantmsdiann invocation (29–33): NCI-60 (originally OpenSWATH), ProCan-DepMapSanger (originally DIA-NN 1.7.x), the Sun PCT-SWATH breast-cancer panel, the MultiPro batch-effect testbed, and Tognetti (67 breast-cancer lines); accessions are listed in the figure legend and Data Availability (**Supplementary Fig. S5**). For the two cohorts with a published DIA protein-group headline, the reanalysis recovered more protein groups under the original studies' >=2-unique-peptide criterion (11), applied to the deposited count and to the quantmsdiann count

under its own global protein-group rule: NCI-60 6,812 versus 4,284 (+ 59%) and ProCan 8,993 versus 6,698 (+ 34%, Fig. 4a). This > = 2-peptide gain exceeds the change in the total protein-group count: current DIA-NN releases identify more peptides per protein, so proteins that the original analyses supported with a single peptide now clear the stricter two-peptide threshold, expanding the high-confidence protein set rather than only the marginal one. The Sun PCT-SWATH panel, whose deposited OpenSWATH analysis reports proteotypic proteins rather than a > = 2-peptide count, shows a comparable improvement at the standard 1% protein-FDR rule: 8,304 quantmsdiann protein groups (Lib.PG.Q.Value < = 0.01) versus 6,183 in the deposited analysis (+ 34%). The size of the gain reflects the age and engine of the deposited analysis, largest for the oldest, OpenSWATH-era NCI-60 (+ 59%); the return is therefore greatest exactly where deposits are most numerous: the large back-catalog of DIA data processed with older or non-DIA-NN engines. Each original count is recomputed from public data (NCI-60 and Sun from the deposited PRIDE OpenSWATH analyses, ProCan from the authors' figshare peptide-count matrix), so every bar is reproducible rather than a transcribed paper headline (Experimental Procedures). (PXD041421 and PXD017199 carry no published DIA headline and contribute reanalysis-only counts.) Across the five cohorts the reanalysis spans 12,914 unique UniProt accessions, of which ~ 44% form a pan-cohort core proteome detected in all five, resolved across 29 tissue categories (**Supplementary Fig. S5**). Because every count comes from a single SDRF-parameterized invocation, the resulting protein matrix is queryable directly from the published QPX archive and can be filtered or re-grouped by Cellosaurus, NCIt, or Disease Ontology terms without manual harmonization, turning independently deposited cohorts into a single reusable cross-study resource. The same pattern extends beyond bulk cell lines to another acquisition mode: a spatial Deep Visual Proteomics cohort of MYCN-amplified neuroblastoma yielded more identifications than its original DIA-NN 1.8.1 analysis (PXD064049; precursors 21,535 versus 17,958, + 20%, protein groups 3,301 versus 2,947, + 12%, both counted from the per-precursor reports under the same 1% Lib q-value rather than the deposited count matrices, which are written at different run-specific q-values per DIA-NN version). Both sides count the identical 12 quantified runs. Notably, the deposited DIA-NN 1.8.1 analysis additionally loaded runs beyond these 12, which would tend to raise its globally identifiable protein count; even with that advantage, the 2.5.1 reanalysis identified more. Residual protein-group inference inflation is also higher in the deposited 1.8.1 analysis (3.2% multi-accession groups) than in the reanalysis (1.3%), so the protein-group gain is a conservative estimate. Per-cohort peptide-per-protein distributions and accession overlap for the reanalyzed cohorts are provided in **Supplementary Note 6 Figs. S3–S5**.

Discussion

Public proteomics archives such as PRIDE and iProX are increasingly functioning as discovery resources, not merely as static collections of datasets deposited to support individual publications. Reuse of repository data can increase statistical power, connect independent cohorts, and test biological signals across laboratories (28, 34); workflow-based reanalysis has already shown how public deposits can be converted into reusable quantitative resources (11), and resources such as ProteomicsDB show the value of aggregating protein-expression evidence across studies (10). In particular, DIA datasets are

becoming an increasingly valuable resource for biological discovery as their volume in public archives grows (8, 9), and DIA-NN has become one of the central engines for their analysis (1). quantmsdiann addresses this need with a scalable, SDRF-driven Nextflow workflow that parametrizes every run from its own metadata and runs any supported DIA-NN release from a container-pinned profile.

Because the per-file work is parallelized across cloud or HPC nodes, this design makes archive-scale reanalysis practical: a 2,300-run single-cell cohort was reanalyzed in 2.2 hours on 300 nodes, and the 5,798-run ProCan cohort finished in 6.2 h (Fig. 2a,b). Across ProteoBench (16), distributed quantmsdiann reproduces the quantification of the corresponding single-machine DIA-NN submission (Fig. 2c), so distributing the analysis across a cluster is an infrastructure choice rather than a change in the result. Reproducibility also holds across releases: because every DIA-NN version runs as a container-pinned profile driven by the same SDRF, a reanalysis can be reproduced exactly or repeated on a later release without reinstallation or reconfiguration. To our knowledge, this is the first pipeline-level treatment of reproducible, version-pinned DIA-NN reanalysis, which grows in importance as the engine evolves and archived data are revisited.

Beyond reproducing the original analysis, reanalysis with the current DIA-NN builds used here adds depth, an effect already reported when public data were reprocessed with quantms. The gain is clearest in low-input single-cell data, where match-between-runs matters most: upgrading the engine raised single-cell precursor and protein-group identifications by up to ~ 17%. Bulk cell-line deposits processed with older or non-DIA-NN engines show the same effect at larger magnitude, with reanalysis recovering 34% to 59% more protein groups per cohort (for example, NCI-60 + 59% and ProCan + 34%) and the largest gains on the oldest deposits. The largest recoverable gains therefore lie in the older DIA back-catalog, where many datasets were processed with earlier software generations.

Depth and speed are only useful if the results can be interpreted and trusted, and biological interpretation requires knowing what each sample is, how it was acquired, and how those attributes map onto the quantification matrix; the scarcity of structured sample metadata remains a major barrier to proteomics reuse (35). quantmsdiann treats the SDRF as both the analytical control layer and the annotation source: every per-run DIA-NN setting is derived from it, and the same file annotates the resulting QPX archive, so identifications, quantifications, and provenance travel together in one queryable artifact. This is what allowed the five-cohort cell-line reanalysis to be regrouped by Cellosaurus, NCIt, or Disease Ontology terms directly from the published output, with no manual harmonization. Because each run is parametrized from its own metadata, the same design spans acquisition modes without modification, from single-cell to plexDIA, spatial Deep Visual Proteomics, and phosphoproteomics (3, 7, 36, 37), and other DIA-NN applications such as immunopeptidomics (38) follow the same path. This combination of current DIA-NN releases, SDRF-first configuration, scalable execution, and standardized QPX output extends earlier DIA workflows (39).

quantmsdiann therefore provides actively maintained infrastructure for DIA meta-analysis and archive-scale proteomics reuse. Its value is amplified by the broader quantms ecosystem (https://quantms.org)

rather than ending at a single workflow: quantmsdiann produces harmonized QPX and MSstats-ready outputs, pmultiqc provides metadata-aware quality control (21), downstream single-cell ecosystems support batch correction, cell-type assignment, and differential analysis (23, 40), and the sibling quantms workflow extends the same SDRF/nf-core model to large-scale quantitative proteomics beyond DIA, including DDA and TMT-style analyses (11). Applied across the growing public archive, this ecosystem can integrate independently deposited cohorts into reusable resources and move DIA data toward protein-expression knowledge bases (10) built from the community's accumulated measurements. quantmsdiann is open-source under the MIT license and developed within the nf-core and bigbio proteomics ecosystem, allowing it to track new DIA-NN releases, acquisition modes, and metadata requirements as they appear.

Data Availability

All datasets reanalyzed in this study are publicly available through ProteomeXchange. Cell-line bulk DIA reanalyses: PRIDE PXD003539 (Guo 2019, NCI-60), PXD004701 (Sun 2023), PXD030304 (ProCan-DepMapSanger), PXD017199 (Tognetti 2021), PXD041421 (Wang 2023). Spatial Deep Visual Proteomics reanalysis: PRIDE PXD064049 (MYCN-amplified neuroblastoma, diaPASEF). Single-cell and plexDIA: PRIDE PXD046357 (HeLa, Orbitrap Astral) and MassIVE MSV000093870 (oocyte plexDIA). Phosphoproteomics: PRIDE PXD034128, PXD034623, and PXD049692. Scalability experiment: PRIDE PXD071075 (Wu 2025). ProteoBench benchmarks: PXD049412 (single-cell, Module 9; also, the A549/H460 cohort of Fig. 3), PXD062685 (diaPASEF, Module 5), PXD070049 (ZenoTOF, Module 10), and Module 7 (Astral) datasets. The full list of PRIDE accessions used in the pan-cohort of DIA cell-line reanalyses is provided in **Supplementary Note 2 Table S1**. quantmsdiann is available at https://github.com/bigbio/quantmsdiann under the MIT license; documentation is hosted at https://quantmsdiann.quantms.org/. The pipeline is implemented in Nextflow DSL2 ($\geq$ 25.10.4) and packages each step as a versioned Docker, Singularity, or Apptainer container image. DIA-NN 1.8.1 is distributed publicly via BioContainers; because the DIA-NN license restricts redistribution of later releases, containers for DIA-NN 1.9 and above are built locally from the recipes in https://github.com/bigbio/quantms-containers (see Experimental Procedures). Pipeline version v2.2.0 (release codename "Chongqing", 2026-06-16) was used for the analyses reported in this paper. Analysis scripts that generated the figures in this manuscript are available at https://github.com/ypriverol/quantmsdiann-paper (this repository).

Supplemental Data

This article contains supplemental data. The Supplementary Notes provide the DIA-NN container-build recipes (**Supplementary Note 1**), the full PRIDE accession list (**Supplementary Note 2, Table S1**), the filtering and protein-counting rules (**Supplementary Note 3**), per-process runtime and resource envelopes (**Supplementary Note 4, Figs. S1–S2**), the ProteoBench per-version identification breakdown (**Supplementary Note 5**), per-cohort peptide-per-protein and accession-overlap figures (**Supplementary**

Note 6, Figs. S3–S5), and a table of software, formats, and public resources with their repositories (**Supplementary Note 7, Table S2**).

Abbreviations

Abbreviation
Definition
CLI
command-line interface
DDA
data-dependent acquisition
DIA
data-independent acquisition
DIA-NN
data-independent acquisition neural networks (DIA search and quantification software)
diaPASEF
data-independent acquisition with parallel accumulation–serial fragmentation
DVP
Deep Visual Proteomics
EFO
Experimental Factor Ontology
FDR
false discovery rate
HPC
high-performance computing
HUPO-PSI
Human Proteome Organization Proteomics Standards Initiative
HYE
human, yeast, and E. coli (ProteoBench three-proteome mixture)
IQR
interquartile range
LSF
Load Sharing Facility (HPC scheduler)
MS
mass spectrometry
MS1/MS2
first-/second-stage mass spectrometry (precursor/fragment scans)
PASEF
parallel accumulation–serial fragmentation
PSI-MS

HUPO-PSI mass spectrometry-controlled vocabulary
PSM
peptide-spectrum match
QPX
Quantitative Proteomics eXchange
SDRF
Sample and Data Relationship Format

Declarations

Conflict of Interest

V.D. holds shares of Aptila Biotech. The other authors declare no competing interests.

Use of Generative AI

During the preparation of this manuscript, the authors used generative AI tools (Anthropic Claude, OpenAI Codex, and ChatGPT) to assist with software code generation and manuscript editing. All content was reviewed and edited by the authors, who take full responsibility for it.

Author Contributions

QXY, YS, CD, and Y.P-R. developed the pipeline. HW helped build the containers and contributed to the workflow implementation and manuscript writing. AL annotated the SDRF metadata and helped run the workflow. TS, FC, NS, and JN assisted with the DIA-NN reanalyzes and manuscript writing; TS additionally contributed to the workflow design and to the design of QPX and the SDRF format, and JN tested the tool and identified multiple issues. OK provided feedback on the plexDIA datasets. VD provided DIA-NN expertise and guidance on cross-version reproducibility. JAV provided proteomics data-resource infrastructure and funding, supervised the EMBL-EBI contributors, and contributed to manuscript review and editing. MB contributed to manuscript writing and supervised the students; MB and Y.P-R. conceived and supervised the project. QXY and Y.P-R. wrote the original draft, with input from all authors. All authors read and approved the final manuscript.

Acknowledgments

M.B. is supported by the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2023TIAD-STX0002). A.L. acknowledges support from MCIN (Recovery, Transformation and Resilience Plan); the Basque Government “Biotechnology Complementary Plan Applied to Health” funded by the European Union NextGeneration EU (PRTR-C17.I1; PRTR-C17.I01.P01.S13)

(AAAA_ACG_AY_2539/22_05); and an EMBO Scientific Exchange Grant (number 10015/2023). Y.P-R. acknowledges support from the Wellcome Trust [223745/Z/21/Z], BBSRC [BB/V018779/1, BB/X001911/1], and EPSRC [EP/Y035984/1]. Y.P-R., N.S., O.K. and J.A.V. acknowledge support from a single-cell proteomics (SCP) grant [UKRI701]. Y.P-R. and J.A.V. acknowledge EMBL core funding. V.D. was funded by the German Ministry of Education and Research (BMBF), as part of the National Research Node “Mass spectrometry in Systems Medicine” (MSCoreSys), under grant agreement 161L0221. We thank the EMBL-EBI HPC team for cluster support, and the ProteoBench community for providing a reproducible benchmark of DIA search engines.

References

1. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods 17:41–44
2. Meier F, Brunner A-D, Frank M, Ha A, Bludau I, Voytik E (2020) diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. Nat Methods 17:1229–1236
3. Brunner A-D, Thielert M, Vasilopoulou C, Ammar C, Coscia F, Mund A (2022) Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol Syst Biol 18:e10798
4. Bubis JA, Arrey TN, Damoc E, Delanghe B, Slovakova J, Sommer TM, Kagawa H, Pichler P, Rivron N, Mechtler K, Matzinger M (2025) Challenging the Astral mass analyzer to quantify up to 5,300 proteins per single cell at unseen accuracy to uncover cellular heterogeneity. Nat Methods 22:510–519
5. Wang J, Huang Y, Lu F, Xu Q, Yang Z, Jiang Y (2025) Benchmarking informatics workflows for data-independent acquisition single-cell proteomics. Nat Commun 16:10276
6. Derks J, Slavov N (2023) Strategies for increasing the depth and throughput of protein analysis by plexDIA. J Proteome Res 22:697–705
7. Derks J, Leduc A, Wallmann G, Huffman RG, Willetts M, Khan S (2023) Increasing the throughput of sensitive proteomics by plexDIA. Nat Biotechnol 41:50–59
8. Perez-Riverol Y, Bandla C, Kundu DJ, Kamatchinathan S, Bai J, Hewapathirana S, John NS, Prakash A, Walzer M, Wang S, Vizca\'ino JA (2025) The PRIDE database at 20 years: 2025 update. Nucleic Acids Res 53:D543–D553
9. Deutsch EW, Bandeira N, Perez-Riverol Y, Sharma V, Carver JJ, Mendoza L, Kundu DJ, Bandla C, Kamatchinathan S, Hewapathirana S (2026) The ProteomeXchange consortium in 2026: making proteomics data FAIR. Nucleic Acids Res 54:D459–D469
10. Lautenbacher L, Samaras P, Schmidt TK, Kuster B, Wilhelm M (2022) ProteomicsDB: toward a FAIR open-source resource for life-science research. Nucleic Acids Res 50:D1541–D1552
11. Dai C, Pfeuffer J, Wang H, Zheng P, K\all L, Sachsenberg T (2024) quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. Nat Methods 21:1603–1607

12. Perez-Riverol Y, Moreno P (2020) Scalable Data Analysis in Proteomics and Metabolomics Using BioContainers and Workflows Engines. Proteomics 20, e1900147
13. Dai C, F\ullgrabe A, Pfeuffer J, Solovyeva EM, Deng J, Moreno P (2021) A proteomics sample metadata representation for multiomics integration and big data analysis. Nat Commun 12:5854
14. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A (2020) The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 38:276–278
15. Kohler D, Staniak M, Tsai TH, Huang T, Shulman N, Bernhardt OM, MacLean BX, Nesvizhskii AI, Reiter L, Sabido E, Choi M, Vitek O (2023) MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. J Proteome Res 22:1466–1482
16. Devreese R, Jachmann C, Van Puyvelde B, Webel H, Bittremieux W, Chiva C (2025) ProteoBench: the community-curated platform for comparing proteomics data analysis workflows. *bioRxiv preprint*
17. UniProt C (2025) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 53:D609–D617
18. Hulstaert N, Shofstahl J, Sachsenberg T, Walzer M, Barsnes H, Martens L (2020) ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. J Proteome Res 19:537–542
19. Kamatchinathan S, Hewapathirana S, Bandla C, Insua S, Vizca\'ino JA, Perez-Riverol Y (2025) pridepy: a Python package to download and search data from PRIDE database. J Open Source Softw 10:7563
20. da Veiga Leprevost F, Gr\uning BoA, Aflitos A, R\ost S, Uszkoreit HL, Barsnes J, H., and, Vaudel M (2017) BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics 33:2580–2582
21. Yue Q-X, Dai C, Kamatchinathan S, Bandla C, Webel H, Larrea A, Bittremieux W, Uszkoreit J, M\uller TD, Xiao J, Cox J, Yu F, Ewels P, Demichev V, Kohlbacher O, Sachsenberg T, Bielow C, Bai M, Perez-Riverol Y (2026) pmultiqc: an open-source, lightweight, and metadata-oriented QC reporting library for MS proteomics. Mol Cell Proteom 25:101530
22. Ewels P, Magnusson Ma, Lundin S, K\aller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048
23. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A (2023) The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol 41:604–606
24. Slavov N (2026) Single-Cell Proteomic Technologies: Tools in the Quest for Principles. Annual Rev Biophys 55:253–275
25. Gatto L, Aebersold R, Cox J, Demichev V, Derks J, Emmott E, Franks AM, Ivanov AR, Kelly RT, Khoury L, Leduc A, MacCoss MJ, Nemes P, Perlman DH, Petelski AA, Rose CM, Schoof EM, Van Eyk J, Vanderaa C, Yates JR, Slavov N (2023) Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments. Nat Methods 20:375–386

26. Wen B, Freestone J, Riffle M, MacCoss MJ, Noble WS, Keich U (2025) Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. Nat Methods 22:1454–1463
27. Krull KK, Ali SA, Krijgsveld J (2024) Enhanced feature matching in single-cell proteomics characterizes IFN-γ response and co-existence of cell states. Nat Commun 15:8262
28. Perez-Riverol Y (2022) Proteomic repository data submission, dissemination, and reuse: key messages. Expert Rev Proteom 19:297–310
29. Guo T, Luna A, Rajapakse VN, Koh CC, Wu Z, Liu W, Sun Y, Gao H, Menden MP, Xu C (2019) Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. *iScience* 21, 664–680
30. Gon\ccalves E, Poulos RC, Cai Z, Barthorpe S, Manda SS, Lucas N, Beck A (2022) Pan-cancer proteomic map of 949 human cell lines. Cancer Cell 40:835–849e838
31. Sun R, Ge W, Zhu Y, Sayad A, Luna A, Lyu M, Liang S (2023) Proteomic Dynamics of Breast Cancer Cell Lines Identifies Potential Therapeutic Protein Targets. Mol Cell Proteom 22:100602
32. Wang H, Lim KP, Kong W, Gao H, Wong BJH, Phua SX, Guo T, Goh WWB (2023) MultiPro: DDA-PASEF and diaPASEF acquired cell line proteomic datasets with deliberate batch effects. Sci Data 10:858
33. Tognetti M, Gabor A, Yang M, Cappelletti V, Windhager J, Rueda OM (2021) Deciphering the signaling network of breast cancer improves drug sensitivity prediction. Cell Syst 12:401–418e412
34. Hewapathirana S, Bai J, Bandla C, Kamatchinathan S, Kundu DJ, John NS, Brown-Harry B, Madhusoodanan N, Duocastella R, Vizca\'ino JM, J. A., and, Perez-Riverol Y (2026) Quantifying data reuse in proteomics using PRIDE download statistics and a semi-supervised LLM-based framework. *bioRxiv preprint*
35. Perez-Riverol Y, Mass EBC (2020) S. Toward a Sample Metadata Standard in Public Proteomics Repositories. *J Proteome Res* 19, 3906–3909
36. Mund A, Coscia F, Kriston A, Hollandi Re, Kov\'acs F, Brunner A-D (2022) Deep Visual Proteomics defines single-cell identity and heterogeneity. Nat Biotechnol 40:1231–1240
37. Makhmut A, Qin D, Fritzsche S, Nimo J, K\onig J, Coscia F (2023) A framework for ultra-low-input spatial tissue proteomics. Cell Syst 14:1002–1014e1005
38. Scheid J, Lemke S, Hoenisch-Gravel N, Dengler A, Sachsenberg T, Declercq A (2025) MHCquant2 refines immunopeptidomics tumor antigen discovery. Genome Biol 26:290
39. Bichmann L, Gupta S, Rosenberger G, Kuchenbecker L, Sachsenberg T, Ewels P (2021) DIAproteomics: a multifunctional data analysis pipeline for data-independent acquisition proteomics and peptidomics. J Proteome Res 20:3758–3766
40. Vanderaa C, Gatto L (2023) The current state of single-cell proteomics data analysis. Curr Protoc 3:e658

Figures

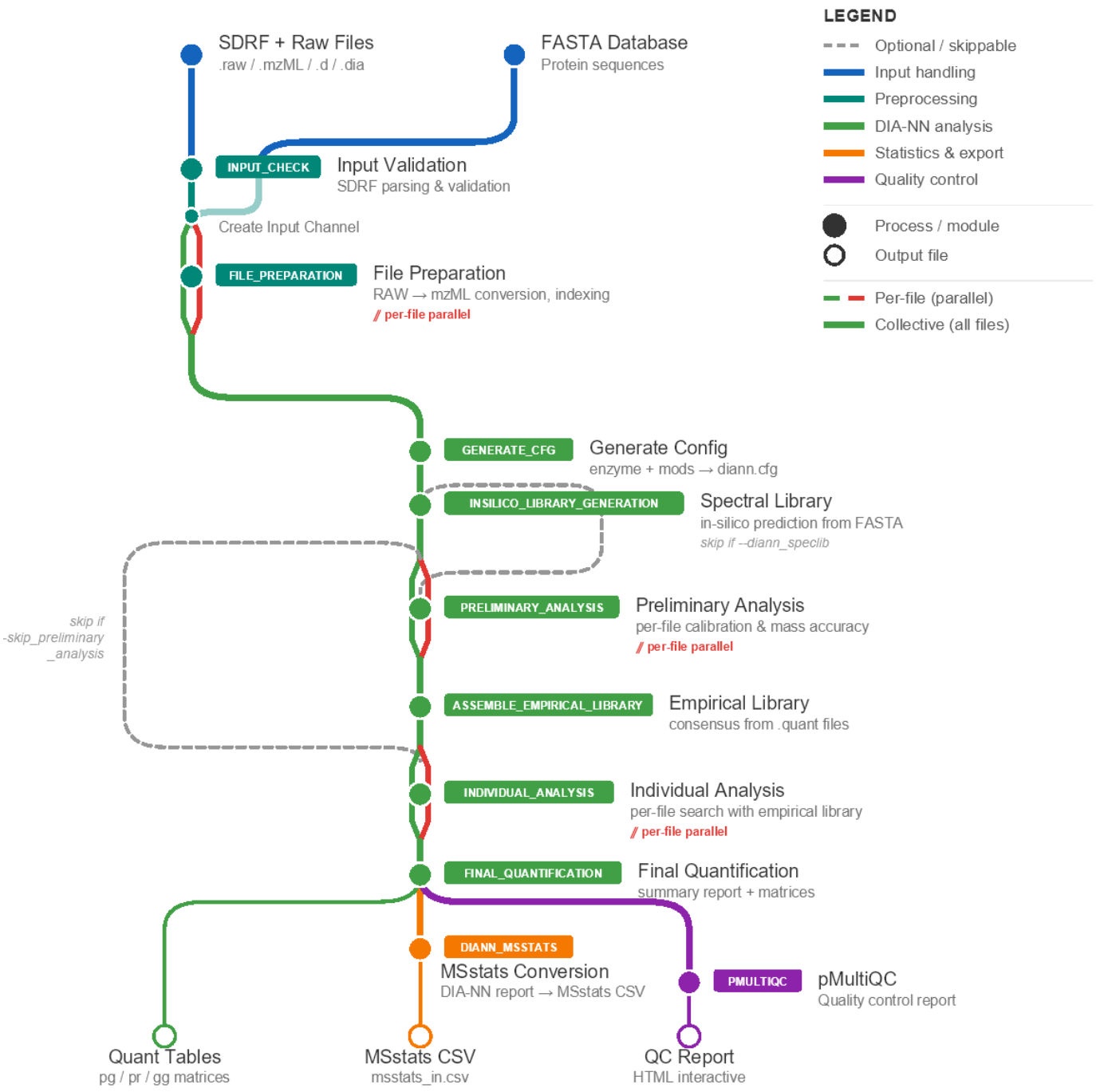

Figure 1

The quantmsdiann workflow. Subway diagram: mandatory modules as filled circles, optional/skippable modules as open circles. Per-file parallel stages (red; raw-file preparation, preliminary calibration, individual search) run as one task per raw file, whereas collective stages (green; in-silico library generation, empirical library assembly, final quantification, MSstats conversion, pmultiqc reporting, and QPX export) run once over the full cohort.

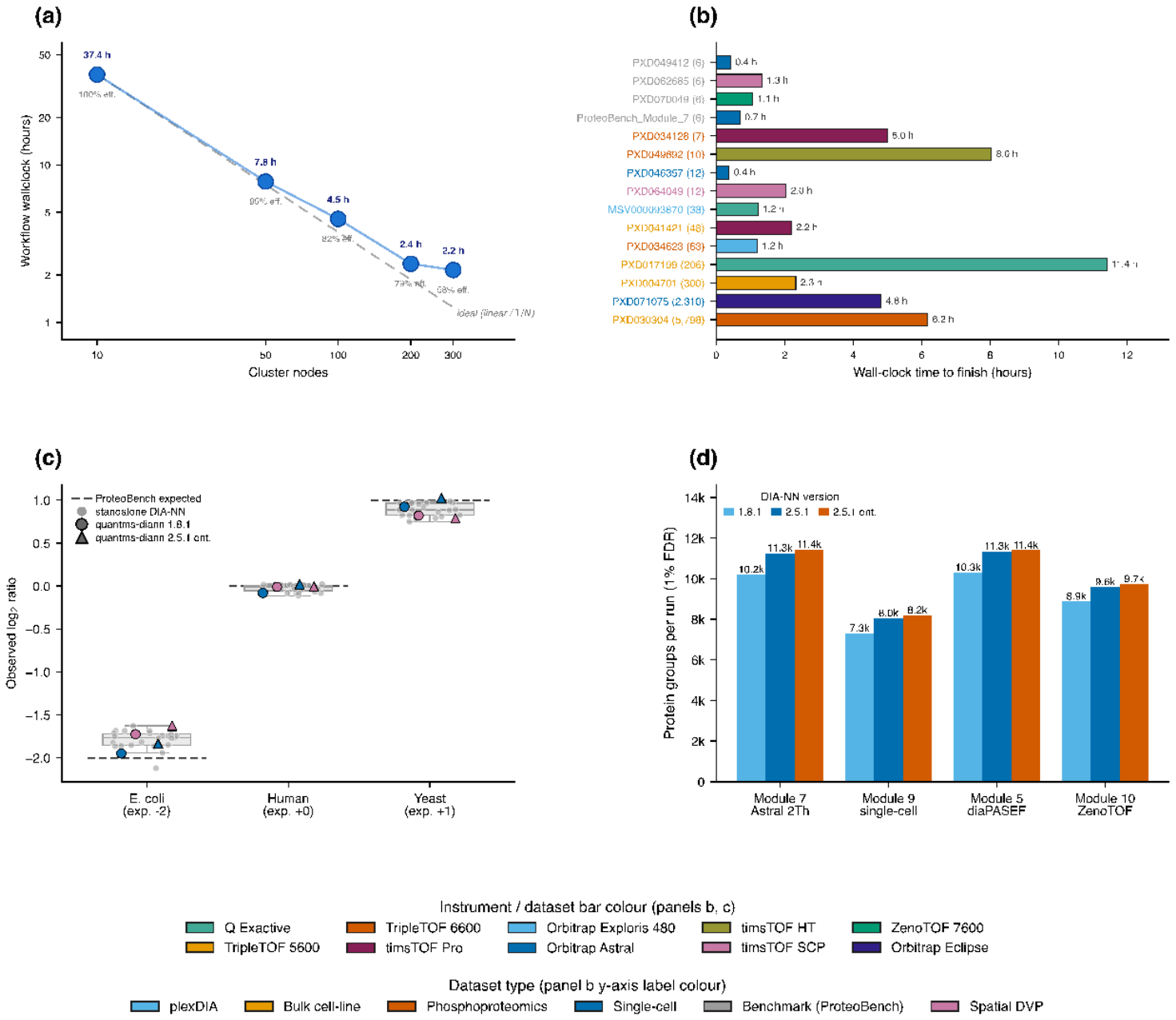

Figure 2

Runtime scaling and ProteoBench quantification accuracy. **(a)** Workflow wall-clock versus number of cluster nodes (Nextflow executor.queueSize) on the PXD071075 single-cell cohort (log–log; one run per point). The dashed gray line is ideal scaling anchored at 10 nodes; parallel efficiency is printed under each dot. The measured curve rises above the ideal line beyond ~100 nodes (parallel efficiency falling from 100% to 58%), reflecting fixed serial cohort-level steps and shared-cluster contention on this EMBL-EBI allocation. **(b)** Wall-clock per reanalysis, one bar per dataset, ordered by cohort size and colored by instrument family; the dataset-type metadata. **(c)** ProteoBench accuracy per HYE species: observed $\log_2$ fold-change against the expected ratio (dashed) for the two DIA modules with predicted-from-FASTA DIA-NN submissions (Orbitrap Astral, Module 7; timsTOF diaPASEF, Module 5, PXD062685). Standalone community runs (all versions) form the gray strip and box; quantmsdiann points are colored by

instrument and shaped by DIA-NNversion (circle, 1.8.1; triangle, 2.5.1-enterprise), tracking the community distribution, i.e. distributed orchestration does not change quantification accuracy. **(d)** Protein-group identification depth on ProteoBench across three DIA-NN configurations (1.8.1, 2.5.1, and the 2.5.1-enterprise build) and four DIA modalities, as protein groups per run (averaged over the six replicates).

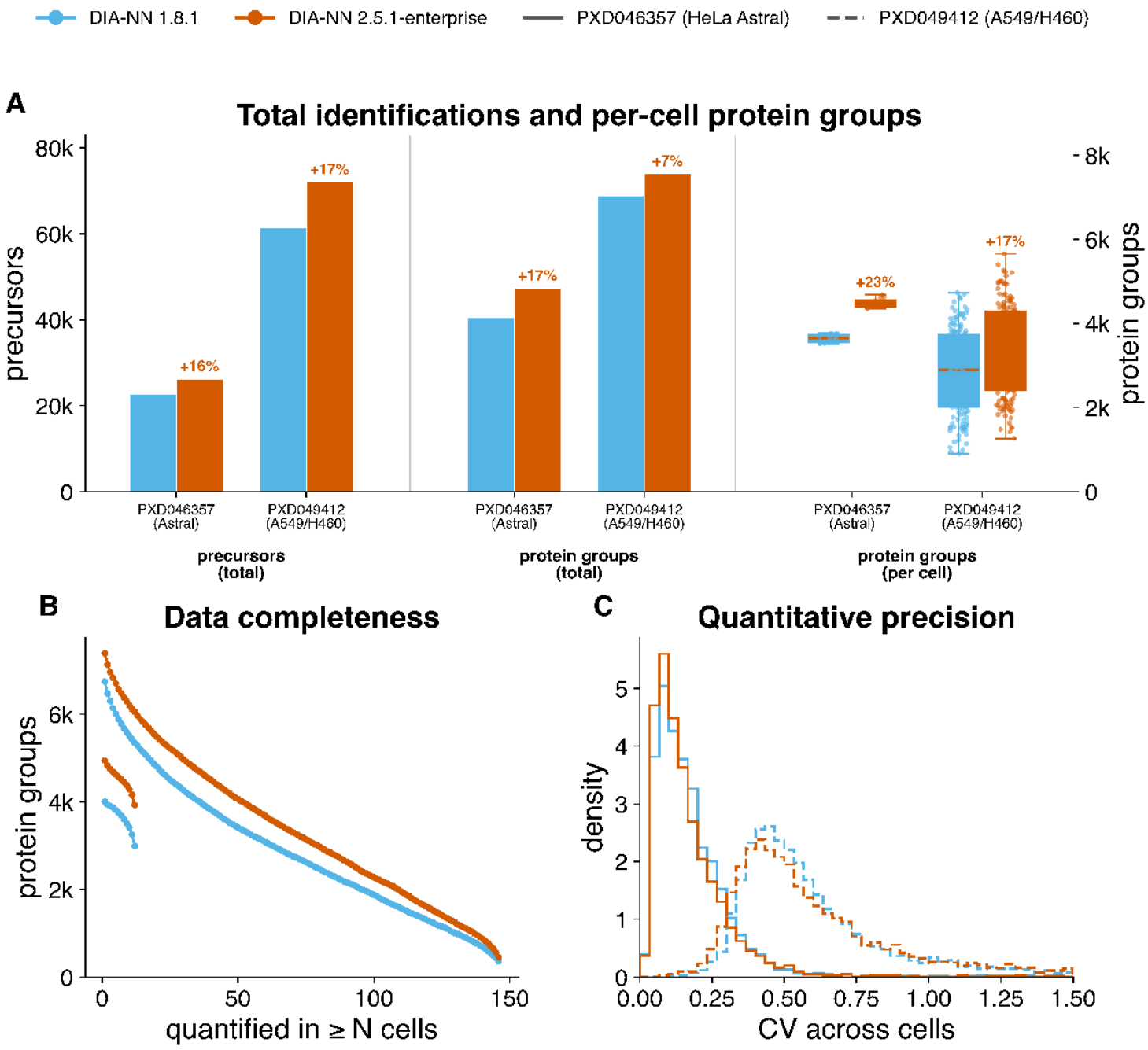

Figure 3

Single-cell DIA reanalysis by quantmsdiann, comparing DIA-NN 1.8.1 with the current 2.5.1-enterprise build on identical raw data. Two label-free Orbitrap Astral cohorts: HeLa (PXD046357) and a larger A549/H460 cohort of 146 single cells (PXD049412). (a) Total identifications and per-cell depth per build:

total precursors (left axis), total protein groups, and per-cell protein-group distribution (box and jittered points, right axis), with per-build percentage change annotated. A549/H460 reaches comparable per-cell depth on many more cells but is already saturated in total protein groups, so the version effect is concentrated in precursor depth. (b) Data-completeness curves (HeLa solid, A549/H460 dashed): protein groups quantified in at least cells (match-between-runs signature; right endpoint is the complete-profile count). (c) Per-protein coefficient of variation across cells (HeLa solid, A549/H460 dashed; axis truncated at CV), summarizing quantitative precision; the larger A549/H460 cohort shows the expected higher cell-to-cell variability.

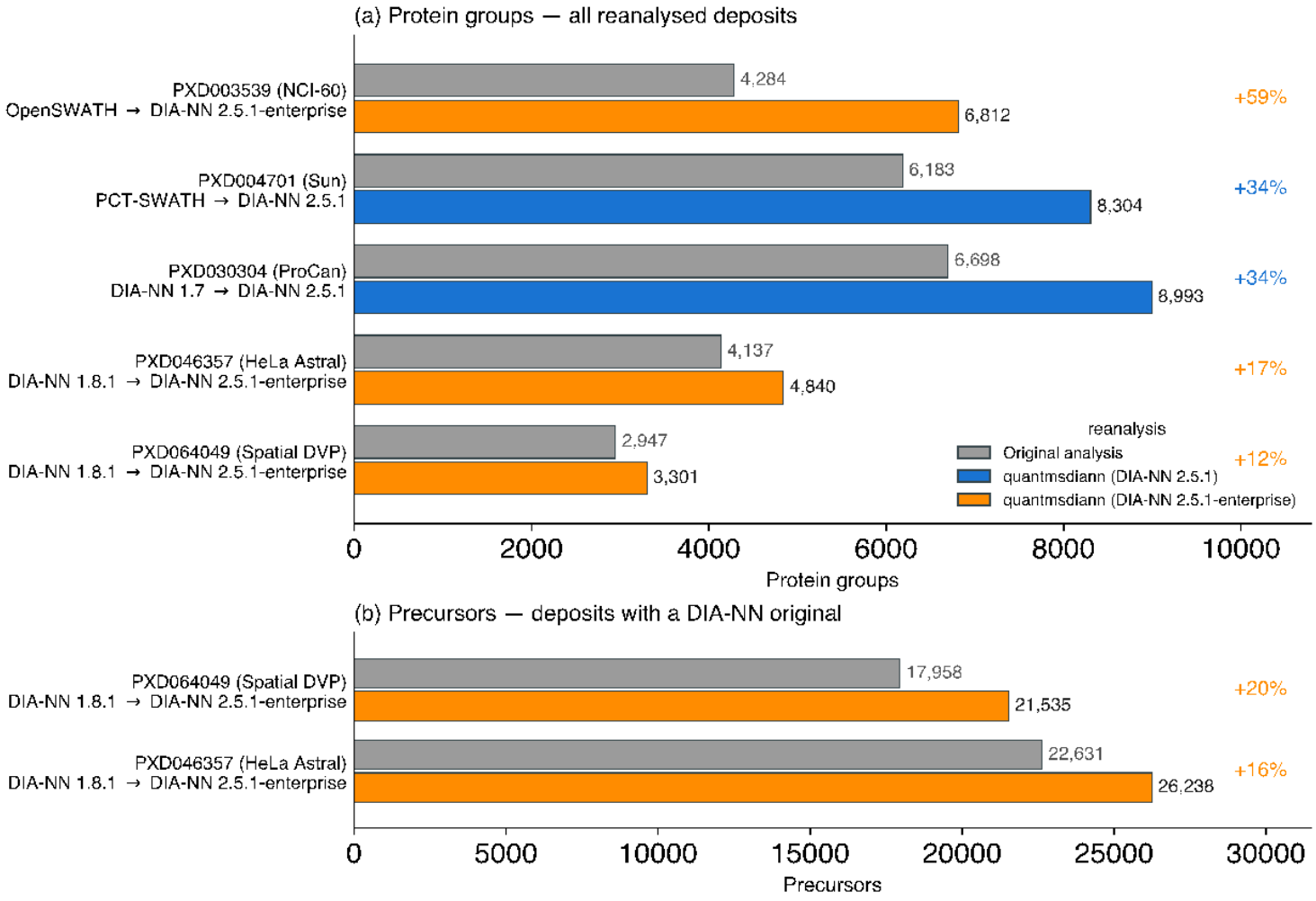

Figure 4

Reanalysis with current DIA-NNidentifies more than the originally deposited analyses. Each public DIA deposit is shown as the originally deposited or published count (gray) versus the quantmsdiann reanalysis (colored by the DIA-NN version used: 2.5.1 or 2.5.1-enterprise), sorted by gain. **(a)** Protein groups for the five reanalyzed deposits, spanning single-cell (HeLa Astral, PXD046357), bulk cell-line (NCI-60, PXD003539; ProCan, PXD030304; Sun, PXD004701) and spatial Deep Visual Proteomics (PXD064049). **(b)** Precursors, restricted to the two deposits with a DIA-NN-based original precursor count (HeLa Astral and spatial DVP); the OpenSWATH/PCT-SWATH originals provide no comparable

precursor count. The quantmsdiann side is counted from the per-precursor report (parquet) for all cohorts; the original side is the deposit's count matrix or deposited analysis, except for the spatial cohort, where both the deposited DIA-NN 1.8.1 and quantmsdiann counts are taken from their per-precursor reports under the same global rule (**Supplementary Note 3**). The five-cohort pan-cohort integration (per-tissue coverage) is in **Supplementary Fig. S5**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryquantmsdiannpreprint.docx