Summary Reports – CPTAC Common Data Analysis Pipeline (CDAP)

v. 09/14/2018

Summary

The purpose of this document is to describe the summary reports generated as part of the CPTAC Common Data Analysis Pipeline (CDAP) of the Clinical Proteomics Tumor Analysis Consortium (CPTAC). The summary reports are generated by the Edwards lab for the CPTAC Data Coordinating Center from the peptide-spectrum-matches generated by the CDAP PSM Analysis. The reports provide protein identification and quantitation summaries for 'label-free' and isobaric-labeling quantitation workflows, phosphopeptide and phosphosite quantitation for phosphopeptide enriched isobaric-labeling quantitation workflows, and N-linked glycopeptide and glycosite quantitation for N-linked glycopeptide enriched isobaric-labeling workflows. QC metrics and a QC report are also provided. Human and human-in-mouse xenograft samples are supported, consistent with the studies carried out by the consortium.

Authors

Nathan J. Edwards*
CPTAC Data Coordinating Center, and
Department of Biochemistry and Molecular & Cellular Biology
Georgetown University

Sanford P. Markey and Stephen E. Stein
NIST
Mass Spectrometry Data Center
Biomolecular Measurement Division
Material Measurement Laboratory

* Contact: nje5@georgetown.edu

Table of Contents

Conventions used in this document

`This font is used to highlight a software program, command-line options, or to display the contents of a file.`

Bold is used to highlight a program name, file type or data file format.

Major output files

Report	Filename	Description
Experimental Design	*.sample.txt	Mapping from analytical samples and isotopic labeling reporter-ion to (biological) sample identifiers, and reporter-ion log-ratios to report.
Labeling Reagent Corrections	<reagent-batch>.txt	Correction values for isobaric-labeling reagent batch.
Protein Identification Summary	*.summary.tsv	Protein identification summary report.
Precursor Area Quantitation	*.precursor_area.tsv	Label-free workflow protein quantitation report for relative quantitation by precursor peak area integration.
Spectral Count Quantitation	*.spectral_count.tsv	Label-free workflow protein quantitation report for relative quantitation by spectral counts.
Isobaric-Labeling Protein Quantitation	*.itraq.tsv *.tmt10.tsv	iTRAQ 4-plex / TMT 10-plex workflow protein relative quantitation report.
Isobaric-Labeling Phosphopeptide Quantitation	*.phosphopeptide.tsv *. phosphopeptide.tmt10.tsv	iTRAQ 4-plex / TMT 10-plex workflow phosphopeptide relative quantitation report.
Isobaric-Labeling Phosphosite Quantitation	*.phosphosite.tsv *.phosphosite.tmt10.tsv	iTRAQ 4-plex / TMT 10-plex workflow phosphopeptide relative quantitation report.
Isobaric-Labeling Glycopeptide Quantitation	*.glycopeptide.tsv *.glycopeptide.tmt10.tsv	iTRAQ 4-plex / TMT 10-plex workflow N-linked glycopeptide relative quantitation report.
Isobaric-Labeling Glycosite Quantitation	*.glycosite.tsv *.glycosite.tmt10.tsv	iTRAQ 4-plex / TMT 10-plex workflow N-linked glycosite relative quantitation report.

Peptide Identification Summary	*.peptides.tsv	Identified peptide summary report.
QC Metrics	*.qcmetrics.tsv	QC metrics computed by the CDAP PSM Analysis.
QC Report	*.qcmetrics.html	Study QC Report.

Terminology

The use of multiplexed quantitation strategies in many CPTAC studies has necessitated the use of terminology to distinguish two types of samples: biological samples, the original samples of interest; and analytical samples, for which a number of biological samples are tagged with isobaric labels and mixed before analysis by mass-spectrometry. Label-free quantitation strategies analyze one biological sample in each analytical sample. Isobaric-labeling quantitation workflows using a common reference sample with plex-k will analyze (k-1) biological samples (plus the common reference) in each analytical sample, so for iTRAQ workflows (4-plex) there are 3 biological samples (plus the common reference) in each analytical sample, and for TMT-10 workflows (10-plex) there are 9 biological samples (plus the common reference) in each analytical sample.

Quantitation reports, whether computed for label-free or isobaric-labelling workflows always refer to biological samples. The protein identification summary, peptide identification summary, and QC reports always refer to analytical samples.

Gene-Based Generalized Parsimony Analysis

The protein reports are based on a gene-based generalized parsimony analysis developed by the Edwards lab. PSMs, including decoy identifications, are conservatively filtered; peptides are associated with genes, rather than protein identifiers; and genes with at least two unshared peptide identifications are inferred. The inferred genes have at most 1% gene FDR, as estimated using the MAYU method[1]. A summary of the gene-based generalized parsimony analysis is provided in the *.summary.tsv report.

Mapping Proteins to Genes

The association between identified peptides, filtered at 1% spectral FDR, and their protein accessions is taken from the mzIdentML format peptide-spectral-matches computed by the CDAP PSM Analysis. Peptides may be associated with RefSeq Human, RefSeq Mouse (xenograft samples only), UniProt Human Reference Proteome, and UniProt Mouse Reference Proteome (xenograft samples only) as per the mzIdentML re-formatting methods documentation.

The association between protein accessions and NCBI gene names are taken from i) the RefSeq Gene project, and ii) the UniProt gene-name annotations. Peptides associated only with protein accessions with no gene association are not retained for parsimony analysis, but are listed in the peptide summary report. Typically, the number of peptides lost by mapping of proteins to genes is very small, about 0.1% of all peptides.

All gene names are associated with their NCBI Gene description (derived from HGNC for human genes), cytoband, and organism (human or mouse).

[1] Reiter, L., M. Claassen, S. P. Schrimpf, M. Jovanovic, A. Schmidt, J. M. Buhmann, M. O. Hengartner, and R. Aebersold (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics 8*, 2405-2417.

For xenograft samples, we sometimes observe the same gene-name in both human and mouse e.g. C3. For clarity, we append the taxonomy id to distinguish these gene-names only e.g. C3(9606), C3(10070) – ensuring that the human and mouse genes can be distinguished. This affects a small number of gene names.

For xenograft samples, we also create faux genes to represent identified peptides common to orthologous human and mouse genes. These faux genes are selected by the parsimony analysis when the human and mouse genes have the same set of identified peptides, and receive compound names, e.g. FBL|Fbl, descriptions, cytoband, and organism designations. Gene ortholog relationships are extracted from NCBI Homologene.

Generalized Parsimony Analysis

Generalized parsimony analysis is applied globally to peptides and their genes. The 1% spectrum FDR peptide-spectrum matches are filtered further to ensure at most 1% gene FDR, estimated using decoy peptide-spectrum matches and the MAYU technique[1], following the parsimony analysis. Peptides matched by a single PSM are disregarded and the spectral FDR of PSMs tightened until the desired gene FDR is reached. All 1% spectrum FDR PSMs of inferred genes are used in the protein identification and quantitation reports. All inferred genes must be supported by at least two unshared peptides. Peptides that are not found in inferred genes and their peptide-spectrum-matches do not contribute to the resulting summary reports, other than the peptide summary report (*.peptides.tsv). This approach ensures that all reported genes have their own peptide evidence in addition to any shared peptide evidence.

The parsimony algorithm prefers losing the lowest quality peptide-spectrum-matches, minimizing the sum of the –log(min FDR) of lost peptides. Peptides with min FDR of 0 are given a pseudo-FDR of half of the minimum non-zero FDR for the entire dataset. Spectra with tied rank 1 peptide-spectrum-matches to more than one peptide sequence, typically due to isobaric amino-acid residues, define peptide groups. The peptide groups are associated with all of the peptides' genes and are treated like shared peptides. Spectra with tied rank 1 peptide-spectrum-matches to more than one peptide sequence in which either peptide is sometimes the sole peptide-spectrum-match to some other spectrum are dropped from the analysis. Most tied rank 1 PSMs to more than one peptide sequence are represented as peptide groups, and very few peptides and spectra (about 0.2%) are removed from the analysis in this cleanup.

Genes with identical peptide evidence cannot be distinguished by the filtered identified peptides. For this case, we select a single gene, preferring genes without certain "stop" words in the description or uninformative gene-names. The current list of stop words is: "predicted", "hypothetical", "isoform", "uncharacterized", "putative", "cDNA", "homolog", "pseudogene", "readthrough", "like", "*-like", and "LOC*." For the xenograft samples, we prefer faux genes representing ortholog pairs over species specific genes when they cannot be distinguished by the filtered identified peptides.

Protein Quantitation

Protein quantitation analysis is provided for label-free, and isobaric-labeling workflows using iTRAQ or TMT labels, for global (no sub-proteome enrichment) proteome workflows. For label-free workflows, proteins are quantitated by spectral count and integrated precursor elution peak area. For iTRAQ and TMT workflows, proteins are quantitated using the reporter ions.

Spectral Count Protein Quantitation

For each inferred gene, the peptide-spectrum-matches are tallied by peptide and sample. Peptides are marked as shared, if associated with more than one inferred gene, or unshared. Peptides which participate in peptide groups (see Generalized Parsimony Analysis above) are also marked shared. Gene spectral counts from acquisitions associated with a sample are then summed over associated peptides. Spectral counts with ("Spectral Count") and without ("Unshared Spectral Count") shared peptides are provided.

Precursor Area Protein Quantitation

The integrated area of the elution peak of each identified peptide ion is computed by the CDAP for label-free workflows and provided with the filtered peptide-spectrum-match results. The annotation may be repeated for multiple identifications of the same peptide ion from the same acquisition, due to repeated sampling of the peptide ion in data-dependent acquisition mode. We choose only the largest precursor area associated with a peptide ion in each LC-MS/MS acquisition. Precursor areas from acquisitions associated with a sample are then summed over peptides associated with a gene. Precursor areas aggregated with ("Area") and without ("Unshared Area") shared peptides are provided.

Isotopic-Labeling Protein Quantitation

Most CPTAC studies using isobaric-labeling workflows use a common reference sample, called POOL, in every acquisition. The POOL sample is made up of a mixture of a small amount of many of the clinical samples, ensuring that peptides can be identified more consistently, and that individual peptide-ion abundance can be determined relative to the common reference. For most CPTAC studies, we report the ratio of each (biological) sample to this reference sample, but in some cases, other experimental designs are used. The *.sample.txt file provides the mapping of sample identifiers (including POOL) to reporter ions in each analytical sample, and specifies the (log-)ratios to be computed for the summary reports.

As with precursor area quantitation, the same peptide ion may be identified multiple times in the same acquisition. We choose the spectrum with the maximum total reporter ion intensity, presumed to provide the most reliable quantitation, for each peptide ion in each LC-MS/MS acquisition. Isobaric labelling reagent batch corrections are applied to the reporter ion intensities for each selected spectrum, and we compute the $\log_2$ of each requested ratio. The log-ratios associated with specific sample pairs and genes are collected, subject to outlier removal and averaged. Outlier removal follows the strategy of the "Libra" software from the Trans-Proteomic-Pipeline, in which values more than two standard-deviations from the mean are removed. Log-ratio aggregation with ("Log Ratio") and without ("Unshared Log Ratio") considering shared peptides are provided.

The aggregated log-ratios for each sample are normalized by subtracting the median from all genes' values. The before-normalization sample median, mean, and standard deviation are provided in the first three rows of the report.

Phosphopeptide and Phosphosite Quantitation

Phosphopeptide and phosphosite quantitation analysis is provided for phosphopeptide enriched samples. Available only for isobaric-labeling workflows, phosphopeptides and phosphosites are quantitated using the reporter ions in each phosphopeptide enriched analytical sample.

Isotopic-Labeling Phosphopeptide Quantitation

For phosphopeptide quantitation analysis, the peptide identifications, after gene-based generalized parsimony, are further filtered by the use of a RefSeq-focused generalized parsimony analysis requiring two distinct peptides per protein. Next, only fully-tryptic peptides with respect to at least one protein alignment are retained. Semi-tryptic peptides are excluded, as their presence is expected to correlate more with sample degradation and may consequently produce less reliable reporter ion signals. Only peptide ions with one or more fully-localized phosphorylation modifications, as reported by PhosphoRS, are considered. Redundant peptide ions, per spectra file, are reduced to a single observation using the total reporter ion intensity and corrected for labeling reagent batch, as for protein quantitation analysis, and requested log-ratios computed. The $\log_2$ ratios are then associated with their canonical phosphopeptide sequence, without charge state, or deamidation or oxidation modifications if present on original peptide ion. The log-ratios associated with each ratio's samples and the canonical phosphopeptide sequence are then subject to outlier removal and averaged. Outlier removal follows the strategy of the "Libra" software from the Trans-Proteomic-Pipeline, in which log-ratios more than two standard-deviations from the mean are removed.

NOTE: Unlike the protein relative quantitation, and protein and peptide summary reports, the phosphopeptide report considers and reports only fully-tryptic phosphopeptide identifications.

Isotopic-Labeling Phosphosite Quantitation

For phosphosite quantitation analysis, the peptide identifications, after gene-based generalized parsimony, are further filtered by the use of a RefSeq-focused generalized parsimony analysis requiring two distinct peptides per protein. Next, only fully-tryptic peptides with respect to at least one protein alignment are retained. Semi-tryptic peptides are excluded, as their presence is expected to correlate more with sample degradation and may consequently produce less reliable reporter ion signals. Only peptide ions with one or more fully-localized phosphorylation modifications, as reported by PhosphoRS, are considered. Redundant peptide ions, per spectra file, are reduced to a single observation using the total reporter ion intensity and corrected for labeling reagent batch, as for the protein quantitation analysis, and requested log-ratios computed. The phosphorylated amino-acids of the peptide are mapped to combinations of the corresponding proteins' sites, for each protein with a fully-tryptic alignment of the peptide sequence. The $\log_2$ ratios are then associated with the protein site combinations. The log-ratios associated with each ratio's samples and protein site combination are then subject to outlier removal and averaged. Outlier removal follows the strategy of the "Libra" software

from the Trans-Proteomic-Pipeline, in which log-ratios more than two standard-deviations from the mean are removed.

NOTE: Unlike the protein relative quantitation and protein summary reports, which report gene-name protein identifiers, the phosphosite report indicates amino-acid residues with respect to RefSeq protein accessions. Following gene-based generalized parsimony, RefSeq-focused generalized parsimony is used to select RefSeq proteins for this report.

NOTE: Unlike the protein relative quantitation, and protein and peptide summary reports, the phosphosite report considers only fully-tryptic phosphopeptide identifications.

NOTE: The phosphosite report may summarize the ratios of multiple fully-tryptic phosphopeptides spanning a given set of phosphosites on the RefSeq protein due to missed cleavages. Note that some phosphopeptides considered fully-tryptic in the phosphopeptide report (with respect to at least one protein) may not be fully-tryptic with respect to the specific RefSeq protein chosen by the RefSeq-focused generalized parsimony analysis.

N-linked Glycopeptide and Glycosite Quantitation

N-linked glycopeptide and glycosite quantitation analysis is provided for N-linked glycopeptide enriched samples. Available only for isobaric-labeling workflows, N-glycopeptides and N-glycosites are quantitated using the reporter ions in each N-glycopeptide enriched analytical sample.

Isobaric-Labeling N-linked Glycopeptide Quantitation

For N-glycopeptide quantitation analysis, the peptide identifications, after gene-based generalized parsimony, are further filtered by the use of a RefSeq-focused generalized parsimony analysis requiring two distinct peptides per protein. Next, only fully-tryptic peptides with respect to at least one protein alignment are retained. Semi-tryptic peptides are excluded, as their presence is expected to correlate more with sample degradation and may consequently produce less reliable reporter ion signals. Tied rank 1 peptide identifications to the same spectrum are also excluded, as these usually indicate ambiguously placed deamidation modifications. Redundant peptide ions, per spectra file, are reduced to a single observation using the total reporter ion intensity and corrected for labeling reagent batch, as for protein quantitation analysis, and requested log-ratios reported. The $\log_2$ ratios are then associated with their canonical glycopeptide sequence, without charge state, oxidation modifications, or deamidation modifications not at an N-linked glycosylation motif site, if they present on original peptide ion. The log-ratios associated with each ratio's samples and the canonical glycopeptide sequence are then subject to outlier removal and averaged. Outlier removal follows the strategy of the "Libra" software from the Trans-Proteomic-Pipeline, in which log-ratios more than two standard-deviations from the mean are removed.

NOTE: Unlike the protein relative quantitation, and protein and peptide summary reports, the N-linked glycopeptide report considers and reports only fully-tryptic N-glycopeptide identifications.

Isotopic-Labeling N-linked Glycosite Quantitation

For N-glycosite quantitation analysis, the peptide identifications, after gene-based generalized parsimony, are further filtered by the use of a RefSeq-focused generalized parsimony analysis requiring two distinct peptides per protein. Next, only fully-tryptic peptides with respect to at least one protein alignment are retained. Semi-tryptic peptides are excluded, as their presence is expected to correlate more with sample degradation and may consequently produce less reliable reporter ion signals. Tied rank 1 peptide identifications to the same spectrum are also excluded, as these usually indicate ambiguously placed deamidation modifications. Redundant peptide ions, per spectra file, are reduced to a single observation using the total reporter ion intensity and corrected for labeling reagent batch, as for the protein quantitation analysis, and requested log-ratios computed. Deamidated Asn residues in the correct N-linked glycosylation motif are mapped to combinations of the corresponding proteins' sites, for each protein with a fully-tryptic alignment of the peptide sequence. The $\log_2$ ratios are then associated with the protein site combinations. The log-ratios associated with each sample and protein site combination are then subject to outlier removal and averaged. Outlier removal follows the strategy of the "Libra" software from the Trans-Proteomic-Pipeline, in which log-ratios more than two standard-deviations from the mean are removed.

NOTE: Unlike the protein relative quantitation and protein summary reports, which report gene-name protein identifiers, the N-linked glycosite report indicates amino-acid residues with respect to RefSeq protein accessions. Following gene-based generalized parsimony, RefSeq-focused generalized parsimony is used to select RefSeq proteins for this report.

NOTE: Unlike the protein relative quantitation, and protein and peptide summary reports, the N-linked glycosite report considers only fully-tryptic N-glycopeptide identifications.

NOTE: The N-linked glycosite report may summarize the ratios of multiple fully-tryptic N-glycopeptides spanning a given set of N-glycosites on the RefSeq protein due to missed cleavages. Note that some N-glycopeptides considered fully-tryptic in the N-linked glycopeptide report (with respect to at least one protein) may not be fully-tryptic with respect to the specific RefSeq protein chosen by the RefSeq-focused generalized parsimony analysis.

Summary Report Files

This section describes the contents of the protein reports. Each report is in tab-separated-values (TSV) format with a header row with field names.

Experimental design report

The experimental design report is provided in the file with a sample.txt extension. These reports provide the following fields.

`FileNameRegEx`
: Regular expression for matching spectral datafiles to their Analytical Sample.

`AnalyticalSample`
: The Analytical Sample identifier.

`[Reporter Ions]`
: Reporter ions. For iTRAQ: 114, 115, 116, 117. For TMT-10: 126 , 127N, 127C, 128N, 128C, 129N, 129C, 130N, 130C, 131.

`LabelReagent`
: The (TMT) label reagent batch identifier.

`Ratios`
: Reporter ion (log-)ratios to compute, specified using forward slash ("/") and separated by commas.

Notes:

1. Repeated biological sample identifiers are automatically augmented by the addition of a trailing .1, .2, etc. when summary reports are generated.
2. The sample identifier POOL is used to designate the common reference sample.
3. Summed reporter ion intensities, rather than averaged (log-)ratios, can be indicated by the use of "1.0" as the denominator of the ratio.
4. The order of the (log-)ratios specifies the column order in the generated sample reports.

Labeling reagent batch correction report

The labeling reagent batch correction report is provided in the file <batch-identifier>.txt. The format mirrors PDF documents available from Thermo for TMT reagents. These reports provide the following fields.

`[First column]`
: Reporter ion. For TMT-10: 126 , 127N, 127C, 128N, 128C, 129N, 129C, 130N, 130C, 131.

`[Isotopic Offset]`
: One of -2, -1, +1, +2. Values represent the extent of reporter's convolution with neighboring reporter ions.

Protein identification summary report

The protein identification summary report is provided in the file with a summary.tsv extension. These reports provide the following fields.

`Gene`
: NCBI Gene name.[1]

`[Analytical-Sample] Spectral Counts`
: Number of spectra matched to peptides associated with the gene in acquisitions from a specific analytical sample.[2]

`[Analytical-Sample] Distinct Peptides`
: Number of distinct peptide sequences associated with the gene in acquisitions from a specific analytical sample.[2]

`[Analytical-Sample] Unshared Peptides`
: Number of unshared peptide sequences associated with the gene in acquisitions from a specific analytical sample.[2]

`Spectral Counts`
: Number of spectra matched to peptides associated with the gene over all analytical samples.

`Distinct Peptides`
: Number of distinct peptide sequences associated with the gene over all analytical samples.

`Unshared Peptides`
: Number of unshared peptide sequences associated with the gene over all analytical samples.

`Description`
: NCBI Gene description.[3]

`Organism`
: NCBI Gene organism.[4]

`Chromosome`
: NCBI Gene chromosome.[5]

`Locus`
: NCBI Gene cytoband.[6]

`Proteins`
: Semi-colon separated list of protein accessions associated with the gene.[7]

Notes:

1. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
2. For label-free workflows, the Analytical-Sample is the biological sample identifier. For iTRAQ workflows, the Analytical-Sample indicates the (three, for 4-plex iTRAQ) biological sample identifiers and the common reference sample identifier, separated by colons, in reporter ion order (114, 115, 116, 117, for 4-plex iTRAQ). For TMT workflows, the analytical sample identifier comes from the *.sample.txt file. The analytical sample headers are listed in lexicographic order, left to right. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
3. For xenograft analyses, indistinguishable ortholog gene-pairs descriptions' will show the human description and the mouse description separated by a semi-colon, unless they are identical.
4. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.

5. For xenograft analyses, indistinguishable ortholog gene-pairs' chromosomes will show the human chromosomes and the mouse chromosomes separated by a semi-colon, unless they are identical.
6. For xenograft analyses, indistinguishable ortholog gene-pairs' locus will show the human cytoband and the mouse cytoband separated by a semi-colon, unless they are identical.
7. Protein accessions do not necessarily associate with all of a gene's peptides. Provided gene-based counts should not be assumed to apply to each protein accession.

Spectral count protein quantitation report

The spectral count quantitation report is provided in the file with a spectral_counts.tsv extension. These reports provide the following fields.

`Gene`
: NCBI Gene name.[1]

`[Sample] Spectral Counts`
: Number of spectra matched to peptides associated with the gene in acquisitions from a specific biological sample.[2]

`[Sample] Unshared Spectral Counts`
: Number of spectra matched to unshared peptides associated with the gene in acquisitions from a specific biological sample.[2]

`Total Spectral Counts`
: Number of spectra matched to peptides associated with the gene over all analytical samples.

`Total Unshared Spectral Counts`
: Number of spectra matched to unshared peptides only associated with the gene over all analytical samples.

`Description`
: NCBI Gene description.[3]

`Organism`
: NCBI Gene organism.[4]

`Chromosome`
: NCBI Gene chromosome.[5]

`Locus`
: NCBI Gene cytoband.[6]

Notes:

1. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
2. The biological sample identifier. The biological sample headers are listed in lexicographic order, left to right. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
3. For xenograft analyses, indistinguishable ortholog gene-pairs descriptions' will show the human description and the mouse description separated by a semi-colon, unless they are identical.
4. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.
5. For xenograft analyses, indistinguishable ortholog gene-pairs' chromosomes will show the human chromosomes and the mouse chromosomes separated by a semi-colon, unless they are identical.
6. For xenograft analyses, indistinguishable ortholog gene-pairs' locus will show the human cytoband and the mouse cytoband separated by a semi-colon, unless they are identical.

Precursor area protein quantitation report

The precursor area quantitation report is provided in the file with a precursor_area.tsv extension. These reports provide the following fields.

`Gene`
: NCBI Gene name.[1]

`[Sample] Area`
: Total precursor area of peptide ions associated with the gene in acquisitions from a specific biological sample.[2]

`[Sample] Unshared Area`
: Total precursor area of peptide ions of unshared peptides only associated with the gene in acquisitions from a specific biological sample.[2]

`Description`
: NCBI Gene description.[3]

`Organism`
: NCBI Gene organism.[4]

`Chromosome`
: NCBI Gene chromosome.[5]

`Locus`
: NCBI Gene cytoband.[6]

Notes:

1. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
2. The biological sample identifier. The biological sample headers are listed in lexicographic order, left to right. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
3. For xenograft analyses, indistinguishable ortholog gene-pairs descriptions' will show the human description and the mouse description separated by a semi-colon, unless they are identical.
4. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.
5. For xenograft analyses, indistinguishable ortholog gene-pairs' chromosomes will show the human chromosomes and the mouse chromosomes separated by a semi-colon, unless they are identical.
6. For xenograft analyses, indistinguishable ortholog gene-pairs' locus will show the human cytoband and the mouse cytoband separated by a semi-colon, unless they are identical.

Isobaric-labeling protein quantitation report

The iTRAQ/TMT workflow protein quantitation report is provided in the file with a itraq.tsv or tmt10.tsv extension. These reports provide the following fields.

`Gene`
: NCBI Gene name.[1,2]

`[Sample] Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions associated with the gene in acquisitions from a specific biological sample.[3]

`[Sample] Unshared Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions of unshared peptides only associated with the gene in acquisitions from a specific biological sample.[2]

`Description`
: NCBI Gene description.[4]

`Organism`
: NCBI Gene organism.[5]

`Chromosome`
: NCBI Gene chromosome.[6]

`Locus`
: NCBI Gene cytoband.[7]

Notes:

1. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
2. Special gene names, “median”, “mean”, and “stddev”, are used to provide before normalization sample median, mean, and standard deviation. See Isobaric-Labeling Protein Quantitation above.
3. The biological sample identifier. The biological sample headers are listed in analytical sample order (see Protein Identification Summary Report, note 2 above), and then by reporter ion within an analytical sample. iTRAQ reporter ions are ordered numerically. TMT reporter ions are ordered by their nominal mass: 126, 127N, 127C, …, 130N, 130C, 131, for example. Experimental Design files, if available, specify the order of reported ratios from an analytical sample. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
4. For xenograft analyses, indistinguishable ortholog gene-pairs descriptions’ will show the human description and the mouse description separated by a semi-colon, unless they are identical.
5. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs’ organisms show human and mouse scientific names, separated by a semi-colon.
6. For xenograft analyses, indistinguishable ortholog gene-pairs’ chromosomes will show the human chromosomes and the mouse chromosomes separated by a semi-colon, unless they are identical.
7. For xenograft analyses, indistinguishable ortholog gene-pairs’ locus will show the human cytoband and the mouse cytoband separated by a semi-colon, unless they are identical.

Isobaric-labeling phosphopeptide quantitation report

The iTRAQ/TMT phosphopeptide quantitation report is provided in the file with a phosphopeptide.tsv or phosphopeptide.tmt10.tsv extension. These reports provide the following fields.

`Phosphopeptide`
: The canonical phosphopeptide sequence. Phosphorylated amino-acids indicated in lower-case.

`[Sample] Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions associated with the canonical phosphopeptide in acquisitions from a specific biological sample.[1]

`Protein`
: The RefSeq accessions of proteins containing the peptide sequence, after protein-based generalized parsimony.[2]

`Gene`
: The NCBI Gene genes associated with the peptide sequence.[2,3,4]

`Organism`
: NCBI Gene organism.[2,5]

Notes:

1. The biological sample identifier. The biological sample headers are listed in analytical sample order (see Protein Identification Summary Report, note 2 above), and then by reporter ion within an analytical sample. iTRAQ reporter ions are ordered numerically. TMT reporter ions are ordered by their nominal mass: 126, 127N, 127C, …, 130N, 130C, 131, for example. Experimental Design files, if available, specify the order of reported ratios from an analytical sample. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
2. Multiple values separated by semi-colons.
3. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
4. The provided proteins and genes are associated with each peptide only. These fields should not be presumed to indicate that the protein accessions are necessarily associated with the genes.
5. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.

Isobaric-labeling phosphosite quantitation report

The iTRAQ/TMT phosphosite quantitation report is provided in the file with a phosphosite.tsv or phosphosite.tmt10.tsv extension. These reports provide the following fields.

`Phosphosite`
: The RefSeq protein accession, after protein-based generalized parsimony, with phosphorylated site combinations.

`[Sample] Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions associated with phosphorylated site combinations in acquisitions from a specific biological sample.[1]

`Peptide`
: Canonical phosphopeptide sequences with fully-tryptic alignment to the protein and corresponding phosphorylated amino-acids.[2]

`Gene`
: The NCBI Gene genes associated with the RefSeq protein.[2,3]

`Organism`
: NCBI Gene organism.[2,4]

Notes:

1. The biological sample identifier. The biological sample headers are listed in analytical sample order (see Protein Identification Summary Report, note 2 above), and then by reporter ion within an analytical sample. iTRAQ reporter ions are ordered numerically. TMT reporter ions are ordered by their nominal mass: 126, 127N, 127C, …, 130N, 130C, 131, for example. Experimental Design files, if available, specify the order of reported ratios from an analytical sample. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
2. Multiple values separated by semi-colons.
3. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
4. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.

Isobaric-labeling N-linked glycopeptide quantitation report

The iTRAQ/TMT N-linked glycopeptide quantitation report is provided in the file with a glycopeptide.tsv or glycopeptide.tmt10.tsv extension. These reports provide the following fields.

`Glycopeptide`
: The canonical N-glycopeptide sequence. Deglycosylated Asn residues indicated in lower-case.

`[Sample] Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions associated with the canonical N-glycopeptide in acquisitions from a specific biological sample.[1]

`Protein`
: The RefSeq accessions of proteins containing the peptide sequence, after protein-based generalized parsimony.[2]

`Gene`
: The NCBI Gene genes associated with the peptide sequence.[2,3,4]

`Organism`
: NCBI Gene organism.[2,5]

Notes:

6. The biological sample identifier. The biological sample headers are listed in analytical sample order (see protein identification summary, note 2 above), and then by reporter ion within an analytical sample. iTRAQ reporter ions are ordered numerically. TMT reporter ions are ordered by their nominal mass: 126, 127N, 127C, …, 130N, 130C, 131, for example. Experimental Design files, if available, specify the order of reported ratios from an analytical sample. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
7. Multiple values separated by semi-colons.
8. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
9. The provided proteins and genes are associated with each peptide only. These fields should not be presumed to indicate that the protein accessions are necessarily associated with the genes.
10. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.

Isobaric-labeling N-linked glycosite quantitation report

The iTRAQ/TMT N-linked glycosite quantitation report is provided in the file with a glycosite.tsv or glycosite.tmt10.tsv extension. These reports provide the following fields.

`Glycosite`
: The RefSeq protein accession, after protein-based generalized parsimony, with deglycosylated Asn residue glycosite combinations.

`[Sample] Log Ratio`
: Average log-ratio of sample reporter-ion to common reference of peptide ions associated with deglycosylated N-glycosylation site combinations in acquisitions from a specific biological sample.[1]

`Peptide`
: Canonical N-glycopeptide sequences with fully-tryptic alignment to the protein and corresponding deglycosylated Asn residues.[2]

`Gene`
: The NCBI Gene genes associated with the RefSeq protein.[2,3]

`Organism`
: NCBI Gene organism.[2,4]

Notes:

5. The biological sample identifier. The biological sample headers are listed in analytical sample order (see protein identification summary, note 2 above), and then by reporter ion within an analytical sample. iTRAQ reporter ions are ordered numerically. TMT reporter ions are ordered by their nominal mass: 126, 127N, 127C, …, 130N, 130C, 131, for example. Experimental Design files, if available, specify the order of reported ratios from an analytical sample. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
6. Multiple values separated by semi-colons.
7. For xenograft analyses, the gene field may indicate the NCBI taxonomy id of the gene or indistinguishable human/mouse ortholog gene-pairs. See Mapping Proteins to Genes above.
8. For xenograft analyses, indistinguishable human/mouse ortholog gene-pairs' organisms show human and mouse scientific names, separated by a semi-colon.

Peptide identification summary report

The peptide identification summary report is provided in the file with a peptides.tsv extension. These reports provide the following fields.

`Peptide`
: The peptide sequence.

`Charge`
: The charge state(s) of the peptide's identified ions.[1]

`Mods`
: Post-translational modifications of the peptide's identified ions.[1,2]

`MinFDR`
: The minimum of the FDR values of the peptide's peptide-spectrum-matches.

`SpectralCount`
: The total number of the peptide's peptide-spectrum-matches.

`AmbigSpectralCount`
: The number of the peptide's peptide-spectrum-matches that were rank 1 ties with a peptide-spectrum-match with a different peptide sequence.

`Sample`
: The analytical samples of the peptide's peptide-spectrum-matches.[1,3]

`Protein`
: The accessions of protein sequences containing the peptide.[1,4]

`Gene`
: The gene names of genes associated with the peptide.[1,4,5]

Notes:

1. Multiple values separated by semi-colons.
2. A specific peptide ion may have multiple modifications, these are separated by commas. Each modification indicates the amino-acid, the position of the amino-acid in the peptide sequence, and the delta-mass of the modification. N-terminal modifications are indicated with an amino-acid of "[" and a position of 0.
3. For label-free workflows, the analytical-sample is the biological sample identifier. For iTRAQ workflows, the Analytical-Sample indicates the (three, for 4-plex iTRAQ) biological sample identifiers and the common reference sample identifier, separated by colons, in reporter ion order (114, 115, 116, 117, for 4-plex iTRAQ). For TMT workflows, the analytical sample identifier comes from the *.sample.txt file. The analytical sample headers are listed in lexicographic order, left to right. Biological sample replicates are indicated by a trailing .1, .2, etc. on the sample identifier.
4. The provided proteins and genes are associated with each peptide only. These fields should not be presumed to indicate that the protein accessions are necessarily associated with the genes.
5. Peptides found only in protein sequences with no associated genes will be listed with an empty gene field.

QC metrics report

The QC metrics report is provided in the file with a qcmetrics.tsv extension. Each row of the report consists of statistics from one spectral datafile. Each column represents some summary statistic derived from all MS/MS and identified MS/MS spectra from the datafile. This file is intended to be consumed by the QC report generator, and as such, we do not list the header fields here. This file will ultimately be described in detail in a separate document.

QC Report

The QC report is provided in the file with a qcmetrics.html extension – indicating that it is a HTML file, suitable for viewing using a web-browser. This report is completely self-contained. This file will ultimately be described in detail in a separate document.

Protein Report Version History

Version 1.0 | R1 (July 3, 2014)

Initial CDAP protein reports.

Version 2.0 | R2 (February 17, 2015)

More stringent PSM filtering applied, prior to generalize parsimony, to control gene FDR estimated using decoy PSMs in the parsimony analysis. Gene FDR now computed using the MAYU method.

Version 3.0 | R3 (April 6, 2016)

Phospopeptide and phosphosite reports.

Version 3.1 | R3 (May 3, 2016)

Glycopeptide and glycosite reports.

Version 4.0 (September 14, 2018)

CPTAC-3 updates, including TMT10, reporter ion correction, experimental design files, and QC reports.

Document Version History

Version 1.0.0 (July 3, 2014)

Initial documentation of version 1.0 of CDAP protein reports in preparation for release.

Version 1.0.1 (July 8, 2014)

Small edits as suggested by Sanford Markey (NIST).

Version 2.0.0 (February 17, 2015)

Changes to generalized parsimony discussion to reflect changes in release R2 version protein reports.

Version 3.0.0 (March 1, 2016)

Documentation of phosphopeptide and phosphosite reports.

Version 3.0.1 (April 6, 2016)

Small edits to the phosphopeptide and phosphosite report descriptions as suggested by Paul Rudnick.

Version 3.1.0 (May 3, 2016)

Documentation of glycopeptide and glycosite reports.

Version 4.0.0 (September 14, 2018)

Updated to reflect CPTAC3 and TMT workflows. Terminology “protein reports” replaced with “summary reports”. Additional summary reports files, *.sample.txt, *.qcmetrics.tsv, *.qcmetrics.html described.