

Protein Reports – CPTAC Common Data Analysis Pipeline (CDAP)

v. 05/03/2016

Summary

The purpose of this document is to describe the protein reports generated as part of the CPTAC Common Data Analysis Pipeline (CDAP) of the Clinical Proteomics Tumor Analysis Consortium (CPTAC). The protein reports are produced at the Edwards lab based on the peptide-spectrum-matches from the CDAP run at NIST. The reports provide protein identification and quantitation for both 'label-free' and 4plex iTRAQ™ workflows with a common reference sample, phosphopeptide and phosphosite quantitation for phosphopeptide enriched 4plex iTRAQ™ workflows with a common reference sample, and N-linked glycopeptide and N-linked glycosite quantitation for N-linked glycopeptide enriched 4plex iTRAQ™ workflows with a common reference sample. Both human and human/mouse xenograft samples are supported, representing the TCGA sample analyses and the system suitability (CompRef) analyses 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, steve.stein@nist.gov

Table of Contents

Summary	1
Authors.....	1
Conventions used in this document	4
Major output files	4
Gene-Based Generalized Parsimony Analysis.....	5
Mapping Proteins to Genes	5
Generalized Parsimony Analysis	5
Protein Quantitation	7
Spectral Count Protein Quantitation	7
Precursor Area Protein Quantitation.....	7
iTRAQ™ Protein Quantitation.....	7
Phosphopeptide and Phosphosite Quantitation	8
iTRAQ™ Phosphopeptide Quantitation.....	8
iTRAQ™ Phosphosite Quantitation.....	8
N-linked Glycopeptide and Glycosite Quantitation	9
iTRAQ™ N-linked Glycopeptide Quantitation	9
iTRAQ™ N-linked Glycosite Quantitation	9
Summary Report Files	11
Protein identification summary report.....	11
Spectral count protein quantitation report	13
Precursor area protein quantitation report.....	14
iTRAQ™ protein quantitation report	15
Peptide identification summary report	16
iTRAQ™ phosphopeptide quantitation report	17
iTRAQ™ phosphosite quantitation report.....	18
iTRAQ™ N-linked glycopeptide quantitation report	19

iTRAQ™ N-linked glycosite quantitation report	20
Protein Report Version History	21
Version 1.0 R1 (July 3, 2014)	21
Version 2.0 R2 (February 17, 2015).....	21
Version 3.0 R3 (April 6, 2016).....	21
Version 3.1 R3 (May 3, 2016)	21
Document Version History.....	22
Version 1.0.0 (July 3, 2014).....	22
Version 1.0.1 (July 8, 2014).....	22
Version 2.0.0 (February 17, 2015)	22
Version 3.0.0 (March 1, 2016)	22
Version 3.0.1 (April 6, 2016)	22
Version 3.1.0 (May 3, 2016).....	22

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 file extension	Description
.summary.tsv	Protein identification summary report
.precursor_area.tsv	Label-free workflow protein quantitation report for relative quantitation by precursor peak area integration
.spectral_count.tsv	Label-free workflow protein quantitation report for relative quantitation by spectral counts
.itraq.tsv	iTRAQ™ workflow protein relative quantitation report
.peptides.tsv	Identified peptide summary report
.phosphopeptide.tsv	iTRAQ™ workflow phosphopeptide relative quantitation report
.phosphosite.tsv	iTRAQ™ workflow phosphopeptide relative quantitation report
.glycopeptide.tsv	iTRAQ™ workflow N-linked glycopeptide relative quantitation report
.glycosite.tsv	iTRAQ™ workflow N-linked glycosite relative quantitation report

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 method¹. A summary of the gene-based generalized parsimony analysis is provided in the .summary.tsv protein report.

Mapping Proteins to Genes

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

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).

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 technique¹, following the parsimony analysis. Peptides matched by a single PSM are disregarded and the spectral FDR of PSMs tightened until the desired gene

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.

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. This constraint typically removes a small number of peptides and peptide-spectrum-matches from the protein reports. This approach ensures that each gene has its 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 $-\log(\text{min FDR})$ 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 both label-free and iTRAQ™ 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™ workflows, proteins are quantitated using the iTRAQ™ reporter ions relative to a common reference in each analytical sample.

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.

iTRAQ™ Protein Quantitation

All CPTAC iTRAQ™ 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. In some datasets, the 117 iTRAQ™ reporter ion is used for the POOL sample, and in others, the 114 iTRAQ™ reporter ion is used. 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. We report the \log_2 of the ratio of each (non-POOL) reporter ion with respect to the POOL reporter ion and associate it with the appropriate peptide and sample. The log-ratios from the reporter ion associated with a sample and the peptides associated with a gene are collected, subject to outlier removal and averaged. Log-ratio 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. Log-ratio aggregation with (“Log Ratio”) and without (“Unshared Log Ratio”) shared peptides are provided.

The aggregated iTRAQ™ log-ratios for each sample are normalized by subtracting the median from all values. The before-normalization sample median, mean, and standard deviation are provided as special “genes” 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 iTRAQ™ workflows, phosphopeptides and phosphosites are quantitated using the iTRAQ™ reporter ions relative to the common reference POOL in each phosphopeptide enriched analytical sample.

iTRAQ™ 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 iTRAQ™ 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, as for protein quantitation analysis, and all ratios reported relative to a common reference sample, called POOL. The log₂ 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 sample and canonical phosphopeptide sequence are then subject to outlier removal and averaged. Log-ratio 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 lists only fully-tryptic phosphopeptide identifications.

iTRAQ™ 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 iTRAQ™ 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, as for the protein quantitation analysis, and all ratios reported relative to a common reference sample, called POOL. 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 log₂ 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. Log-ratio 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 aggregates 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 iTRAQ™ workflows, N-glycopeptides and N-glycosites are quantitated using the iTRAQ™ reporter ions relative to the common reference POOL in each N-glycopeptide enriched analytical sample.

iTRAQ™ 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 iTRAQ™ 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, as for protein quantitation analysis, and all ratios reported relative to a common reference sample, called POOL. 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_2 ratios associated with each sample and canonical glycopeptide sequence are then subject to outlier removal and averaged. Log-ratio 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 lists only fully-tryptic N-glycopeptide identifications.

iTRAQ™ 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 iTRAQ™ 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, as for the protein quantitation analysis, and all ratios reported relative to a common reference sample, called POOL. 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 log₂ 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. Log-ratio 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 aggregates 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.

Protein identification summary report

The protein identification summaries are provided in protein reports with .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). 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.

iTRAQ™ protein quantitation report

The iTRAQ™ protein quantitation report is provided in the file with a .itraq.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 iTRAQ™ Quantitation Analysis above.
3. The biological sample identifier. The biological sample headers are listed in analytical sample order (see protein identification summary, note 2 above), and then by iTRAQ™ reporter ion within 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.

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). 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.

iTRAQ™ phosphopeptide quantitation report

The iTRAQ™ phosphopeptide quantitation report is provided in the file with a .phosphopeptide.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, note 2 above), and then by iTRAQ™ reporter ion within 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.

iTRAQ™ phosphosite quantitation report

The iTRAQ™ phosphosite quantitation report is provided in the file with a .phosphosite.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, note 2 above), and then by iTRAQ™ reporter ion within 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.

iTRAQ™ N-linked glycopeptide quantitation report

The iTRAQ™ N-linked glycopeptide quantitation report is provided in the file with a .glycopeptide.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 iTRAQ™ reporter ion within 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.

iTRAQ™ N-linked glycosite quantitation report

The iTRAQ™ N-linked glycosite quantitation report is provided in the file with a .glycosite.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 iTRAQ™ reporter ion within 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.

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)

Phosphopeptide and phosphosite reports.

Version 3.1 | R3 (May 3, 2016)

Glycopeptide and glycosite 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.