CPTAC, TCGA Cancer Proteome Study of Breast Tissue Naming Conventions

I. Sample Names

TCGA breast cancer tumor samples have a human readable bar code and a Universally Unique Identifier (UUID). Both identifiers are tracked in the TCGA Data Portal Biospecimen Metadata Browser (https://tcga-data.nci.nih.gov/uuid/uuidBrowser.htm) and in the CPTAC DCC.

The human readable bar code is composed of 7 parts (codes) separated by a hyphen (-)
TCGA-A2-A0CM-01A-11-A21V-30

1. TCGA project
2. A2 tissue source site, location where the samples and clinical metadata were collected
3. A0CM Study participant
4. 01A Sample type and vial, 01 = a solid tumor, A = The first vial in a sequence of samples
5. 11 Portion, order of portion in a sequence of 100 – 120 mg samples, here it is portion 11
6. A21V Plate, order of plate in a sequence of 96-well plates
7. 30 Center that will receive the sample, Center 30 = Washington University School of Medicine Proteomics

The codes (parts) in the human readable identifier can be looked up in the Code Table Report
https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=center

The UUID is an identifier standard used in software construction, standardized by the Open Software Foundation (OSF) as part of the Distributed Computing Environment (DCE). A UUID is a 16-byte (128-bit) hexadecimal value. In its canonical form, a UUID consists of 32 hexadecimal digits, displayed in 5 groups separated by hyphens, in the form 8-4-4-4-12 for a total of 36 characters (32 digits and 4 hyphens). For example:

330f7598-824c-4cd6-9303-a27fe74a6695

Information on TCGA UUIDs can be found on this page,
https://wiki.nci.nih.gov/display/TCGA/UUID+Migration+Plan

The TCGA bar codes and UUIDs are listed in the file, “Clinical Data for CPTAC, TCGA Cancer Proteome Study of Breast Tissue.”

The TCGA bar code is used in the file, “CPTAC, TCGA Breast Cancer iTRAQ Sample Mapping.”

II. File Names

Each iTRAQ raw data file records the 4 samples that are assayed in the file name. There are 10 elements that are separated by an underscore (_). The format is shown here in this example:

TCGA_AO-A12D_C8-A131_AO-A12B_117C_W_BI_20130208_H-PM_f01.raw

1.	TCGA	Project
2.	AO-A12D	Sample used with the isobaric reagent mass = 114, only tissue source site (AO) and study participant (A12D) are included
3.	C8-A131	Sample used with the isobaric reagent mass = 115
4.	AO-A12B	Sample used with the isobaric reagent mass = 116
5.	117C	Internal reference control used with the isobaric reagent mass = 117
6.	W, P	(W) whole, global proteome, (P) phosphoproteome
7.	BI	Site generating the mass spectrometry file
8.	20130208	Date of the experiment, yyyymmdd format
9.	H-PM	Site specific codes
10.	f01.raw	(f) fraction, (01) number, raw data instrument file

III. Folder Names

Each set of iTRAQ raw data files are grouped in a folder and the 3 TCGA samples assayed in the experiment are listed in the folder name. Eight elements, separated by an underscore (_), are present. The format is shown here in this example:

TCGA_AO-A12D-01A_C8-A131-01A_AO-A12B-01A_Proteome_BI_20130208_raw

1.	TCGA	Project
2.	AO-A12D-01A	Sample used in the iTRAQ experiment only tissue source site (AO) and study participant (A12D), solid tumor and vial (01A)
3.	C8-A131-01A	Sample used in the iTRAQ experiment
4.	AO-A12B-01A	Sample used in the iTRAQ experiment
5.	Proteome, Phosphoproteome	Type of analysis
6.	BI	Site generating the mass spectrometry file
7.	20130208	Date of the first mass spectrometry raw file, yyyymmdd format
8.	raw	Data type, (**raw**) the original mass spectrometry(MS) instrument files; (**mzML**) HUPO-PSI standard raw data files generated from the original MS instrument files; (**PSM**): Peptide-Spectrum Match data