LINCS Standardized Unique Small Molecule IDs (LSM)

Release Notes

Release Number: 27

Release Date: September 18, 2016

##
Description:
##
Small molecule information submitted by the LINCS Data and Signature Generation Centers (DSGCs) are standardized and registered by the Data Coordination and Integration Center (DCIC). The DCIC developed a standardization and registration pipeline that generates a unique chemical structure representation for each submitted record. The LINCS standardized small molecule IDs (SM_LINCS_ID) correspond to the standardized small molecule structure. Via the standardized chemical structure representation, LINCS compounds are mapped to PubChem CIDs and other resources.

# LINCS Small Molecule Specifications: http://www.lincsproject.org/data/data-standards/

##
Release Notes for Files:
##
# CompoundTable_LINCS_StandardizedCmpds_LSMIDs.txt

Compound table with chemical structure of the standardized small molecules including LINCS ID, PubChem CID, parent SMILES, parent InChI, parent InChI Key, and molecular mass

# LincsID2FacilityID_LINCS_StandardizedCmpds_LSMIDs.txt

Mapping table between SM_LINCS_ID and SM_Center_Canonical_ID

# SampleTable_LincsID2FacilityID2CenterBatchID_LINCS_StandardizedCmpds_LSMIDs.txt

Sample table with the chemical information provided by the source (DSGC).

##
Summary:
##
# Total unique standardized compound count 41,847

# Total sample count submitted by DSGCs: 45,069

# Table: Sample count per data source (DSGC)

Sample	Number of Samples	Number of Compounds
Broad	44,178	41,934
HMS	611	373

PCCSE	95	95
DToxS	52	52
CGC	128	128
MEP	3	3
NeuroLINCS	2	2

##
Changes since the last release (v26 - May 2016):
##

All small molecule structures submitted by the LINCS DSGCs were reprocessed by a substantially improved in-house standardization pipeline to handle various special cases such as metal complexes, multi-fragment compounds, salts. The pipeline in most cases generates a single fragment molecular representation by removing common salt and addend forms and neutralizing salts were possible (protonating acids, deprotonates basic groups). A canonical tautomer is generated using an improved protocol. The previous standardization pipeline relied on a standardization service provided by PubChem. We discovered and reported several errors that occur in certain situations. The current process uses Biovia Pipeline Pilot and ChemAxon components to generate and QC valid unique tautomeric forms and canonical standardized chemical structure representations.
The process includes several QC and validation steps and substantial expert curation to resolve conflicts and errors. The SM_LINCS_ID is assigned to each unique chemical compound based on its canonical chemical structure representation. LINCS standardized compounds are also cross-referenced to PubChem and other small molecule resources. In the process several previously unrecognized errors were corrected.

The Sep 2016 release has the following changes compared to May 2016 release:
# New sample IDs submitted by DSGCs: 133

# New SM_LINCS_ID: 212

# Expunged SM_LINCS_ID: 197

Previously obtained and standardized compounds that do not associate to any SM_Center_Canonical_ID received the status expunged and are no longer included in the file CompoundTable_LINCS_StandardizedCmpds_LSMIDs.txt.

# Deprecated SM_LINCS_ID: 255

If a SM_Center_Canonical_ID was previously associated with one SM_LINCS_ID and now corresponds to a different SM_LINCS_ID, the previous SM_LINCS_ID received a deprecated status. Out of the 255 deprecated SM_LINCS_IDs, 197 SM_LINCS_ID have the expunged status. Deprecated, but not expunged SM_LINCS_ID remains associated with other SM_Center_Canonical_ID. The purpose of reporting deprecated records is to quickly identify changes in associations of submitted records and standardized SM_LINCS_ID. Reasons for such changes include corrections of chemical structures of a sample at the source (e.g. an error was discovered and corrected) or a change in the processing pipeline that results in a different standardized chemical structure (note, not a different representation of the same chemical structure, but an actual different chemical entity).

The file DeprecatedSampleTableWxpg_Sep2016.txt includes all submitted sample records associated with deprecated SM_LINCS_IDs; records include these fields: SM_Center_Canonical_ID (submitted compound ID), SM_Center_Batch_ID (submitted sample ID), SM_LINCS_ID (current valid standardized SM_LINCS_ID), SM_LINCS_ID_Previous (previous SM_LINCS_ID that is now deprecated from the sample), Expunge (a boolean tag if the previous SM_LINCS_ID has been expunged).

# SM_LINCS_ID for which SM_PubChem_CID association changed: 79

Because of the change on the standardization pipeline, in some cases the SM_LINCS_ID is now associated with a different PubChem CID, compared to previous release.

All fields are defined in the file Property_Description.csv.

For further information about LINCS Standardized Small Molecules, for general queries/feedback or report any problems with data content please contact: metadata {at} lincs {dash} dcic {dot} org