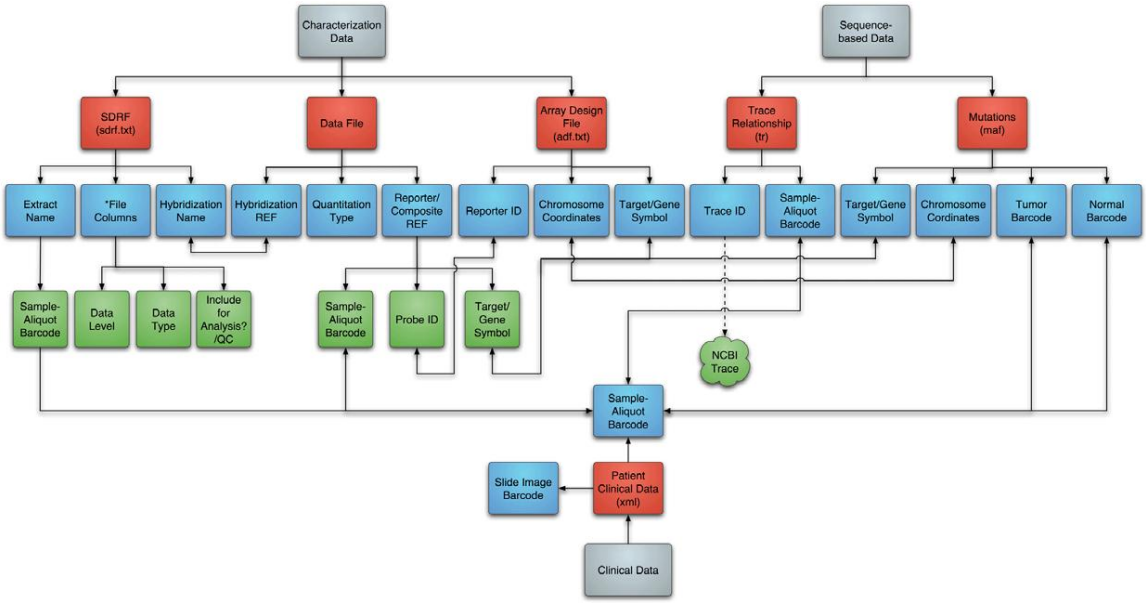
DCC Metadata Normalization

08/09/2011

Introduction

Firehose provides its algorithms with data that are up-to-date and follow a regularized format. To do this, it mirrors the DCC site nightly, scans for new SDRF files, and transforms each file referenced by the SDRF file into a highly regular format, containing one sample per file. The transformation process eliminates two types of variation: that which is explicitly allowed by the spec (single vs multisample files, filenames, hybridization ids) and that on which the spec is silent (line termination styles, spaces in IDs, files with no data, uneven number of fields per line, duplicated samples, and other novel variants). Next, a collection of samples is identified, using criteria such as tumor type and exclusion lists from Disease Working Groups (DWGs) or the Biospecimen Core Resource (BCR); we would like to also use clustering group membership. Finally, the chosen collection of per-sample files is merged together into a single file, providing Firehose-hosted algorithms with the latest submitted data in a consistent format.



Basic Process

Mirror new data files from DCC (nightly)

- 6TB, 90 minutes if no new data – 6 hours if lots of new data.
- Mirror already exploded files but not tarfiles.
- File deletions at DCC must be manually deleted in mirror.
- Caching – mirror only new or newly modified files

Clinical Data Normalization

- convert xml into 2 column parameter-value file
- aggregate columns, using superset of parameter names
- map selected parameter names to shorter name; Volatile. Lookup table used, which can contain functions of multiple parameters.

Processing MAF and WIG files

- Add data files that do not yet have home - maf/wig files from GSC
- Generate SDRF files for BCRs and GSCs

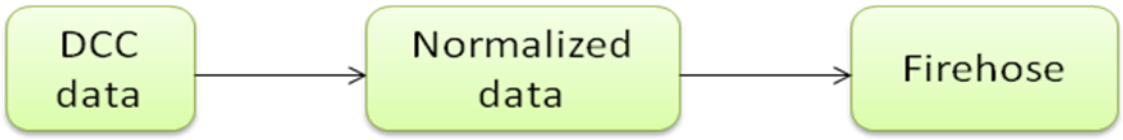
Processing SDRF files

Delete pathological files (early submission attempts, batch 27 sdrf file, malformatted data) Current items eclipsed: LAML bcr early attempts, hudson-alpha early attempts, autogenerated batch 27 bcr SDRF, mskcc gbm ‘archive’ directory

Map normalized data to Firehose ID

- GBM-01-2345-Tumor, GBM-01-2345-Normal
- based on latest good SDRF
- drop dupes, cell lines, mets

Normalization



Workflow diagram

1 Processing maf and wig files

- Pull wig and maf files from cga Firehose into dcc_mirror_overlay
- Generate sdrf files for wig and maf. Dicer will get the sdrf files and process wig and maf files.
- One sdrf file for each tumor type and center

2 Turn off BCR Clinical

- Before each GDAC Firehose Run, turn off BCR clinical data downloading.

3 Dicer

a. Data located: /xchip/gdac_data/dcc_mirror3/

/xchip/gdac_data/dcc_mirror_overlay/

/xchip/gdac_data/normalized/

b. Process clinical data from xml to tsv

Because of the complexity of XML files, and because there is one XML file for each patient rather than for each of the patient's data elements, the DCC parses each XML file into separate comma-separated value (CSV) files, each of which represents a major BCR XML element, with metadata for multiple patients.

Treating biospecimen and clinical as separate data types

clin__bio__intgen_org__Level_1__biospecimen__clin
clin__bio__intgen_org__Level_1__clinical__clin

c. Process transcriptome, Copy number, SNP, miRNA and methylation data for each sdrf files.

d. Find all of the sdrf files and Process sdrf file one by one.

For each sdrf file

- Process filetype and samples
- Process one file (extract the file if they have more than one samples in this file)

e. Generate samplestamps for the files

f. Generate firehose load files which will load data into DEV workspace

4 Connector to FireHose

- mRNA, copy number, miRNA, SNP, methylation and mutation files will be loaded into dev workspace of Firehose.

5 Data will be pushed through the pipelines for each RUN (fiss)

Glossary of Terms

Data Coordinating Center (DCC)

Sample and Data Relationship Files (SDRF)

Mutation annotation Format (MAF)