UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Tag Storm

Metadata made simple

CIRM CALIFORNIA'S STEM CELL AGENCY

Introducing the Tag Storm format

In bioinformatics we deal with the lowest common denominator formats. We want something that is easily readable by all computer programs: usually this means the data are stored in a spreasdsheet or as plain text organized into rows and columns, often with tabs delimiting the columns. While this is very easy for a computer to parse, sometimes it's a bit confusing for us to read and interpret.

The Tag Storm format is a way of overcoming this challenge: they are easy for computers to parse, reduce the redundancy of a tab-separated file, and they are human readable.

The problem being addressed

Traditional top-down approach

A top-down approach to metadata starts with the big picture and organizes into smaller and smaller segments.

This makes it difficult to read about individual experiments.

```
lab  ucsc_db  access  species  assay  data_set_id
life_stage  meta  enriched_in  treatment
differentation  donor
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_1  exon  X  10 hour
A
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_2  exon  Y  10 hour
B
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_3  exon  Y  10 hour
C
kent  hg38  group  sapiens  loHomo ng-RNA-seq
kentTest1  postpartum  kent1_4  exon  X  1 day
A
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_5  exon  Y  1 day
B
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_6  exon  Y  1 day
C
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_7  exon  X  5 days
A
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_8  exon  Y  5 days
B
kent  hg38  group  Homo sapiens  long-RNA-seq
kentTest1  postpartum  kent1_9  exon  Y  5 days
C
```

A bottom-up approach to metadata

The people performing experiments spend a lot of time thinking about samples, and can generally write the most useful metadata to describe them. A bottom-up design to metadata makes it easier for scientists to capture data about their individual experiments, then organize that into a big picture.

Applying a bottom-up approach: a Tag Storm

```
lab kent
data_set_id kentTest1
access group
assay long-RNA-seq
enriched_in exon
life_stage postpartum
species Homo sapiens
ucsc_db hg38

    treatment X
    donor A

        differentiation 10 hour
        meta kent1_1

        differentiation 1 day
        meta kent1_4

        differentiation 5 days
        meta kent1_7

    treatment Y

        donor B

            differentiation 10 hour
            meta kent1_2

            differentiation 1 day
            meta kent1_5

            differentiation 5 days
            meta kent1_8

        donor C

            differentiation 10 hour
            meta kent1_3

            differentiation 1 day
            meta kent1_6

            differentiation 5 days
            meta kent1_9
```

The top level stanza contains data inherited by all child stanzas

Tag Storm allows you to choose which fields to divide on, helping scientists to represent data from their perspective

The lowest level stanzas represent individual samples, in this case

Tags can appear on different stanza levels, such as *donor*, and can even override a parent stanza's value

Experimental design visualized

Compare the experimental tree above (two separate treatments and three donors) to the tab-separated metadata to the left, and see which is easier to read.

Schemas to validate

Schemas allow you to define constraints within the Tag Storm to validate. Each line has a tagSpec (tag, or field name) followed by a type (integer %, floating point #, or string $) and then optional constraints, or controlled vocabulary.

```
age %
age_unit $ day week
*date* $ ????-??-??
file $ ex*.*
lab $
lab_*_* $
organ $ “gall bladder”
part # 1 10
```

A package of open source utilities is available

tagStormCheck – validates a Tag Storm against a schema

tagStormDeleteTags – deletes tags from a Tag Storm

tagStormFromTab – generates Tag Storm representation of a tab-separated values file

tagStormHoist – raises tags from child stanzas to parent stanzas

tagStormInfo – prints out statistics about a Tag Storm, and can generate a schema based on that Tag Storm

tagStormJoinTab – joins tab-separated data into a Tag Storm

tagStormQuery – prints out stanzas which match a SQL-like query

tagStormReformat – reformats a Tag Storm, for example sorting tags within stanzas

tagStormRenameTags – renames a tab-separated list of tags from one thing to another

tagStormToHtml – generates an html Tag Storm with controls to expand and contract

tagStormToJson – converts a Tag Storm to a .json file

tagStormToTab – converts a Tag Storm to a tab-separated values file

Source code is available through the UCSC Genome Browser source tree:
https://github.com/ucscGenomeBrowser/kent/tree/master/src/tagStorm

Funding and Acknowledgements

We would like to thank the California Institute for Regenerative Medicine for providing grant GC1R-06673-C. DCM staff involved: Clay Fischer, Nick Fong, Parisa Nejad, Maximilian Haeussler, Josh Stuart and Jim Kent.

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

CIRM CALIFORNIA'S STEM CELL AGENCY