





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.

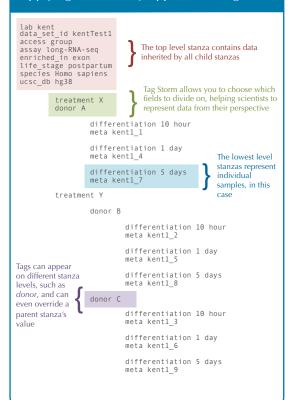
This makes it difficult to read about individual experiments.

```
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 kentl_1 exon Y 10 hour A kent hg38 group Homo sapiens long-RNA-seq kentTest1 postpartum kentl_2 exon Y 10 hour B long-RNA-seq kentTest1 postpartum kentl_3 exon Y 10 hour C long-RNA-seq kentTest1 postpartum kentl_3 exon Y 10 hour C long-RNA-seq kentTest1 postpartum kentl_4 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_5 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 1 day C long-RNA-seq kentTest1 postpartum kentl_6 exon Y 5 days A long-RNA-seq kentTest1 postpartum kentl_8 exon Y 5 days B long-RNA-seq kentTest1 postpartum kentl_8 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_8 exon Y 5 days B long-RNA-seq kentTest1 postpartum kentl_8 exon Y 5 days B long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpartum kentl_9 exon Y 5 days C long-RNA-seq kentTest1 postpart
```

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



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 tabseparated 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.



