# ISA-TAB for CSSI DCC

What is ISA-TAB and how it fits for CSSI DCC

*Jianjun Wang*

*Feb 26, 2016*

# What is ISA-TAB

The challenge:

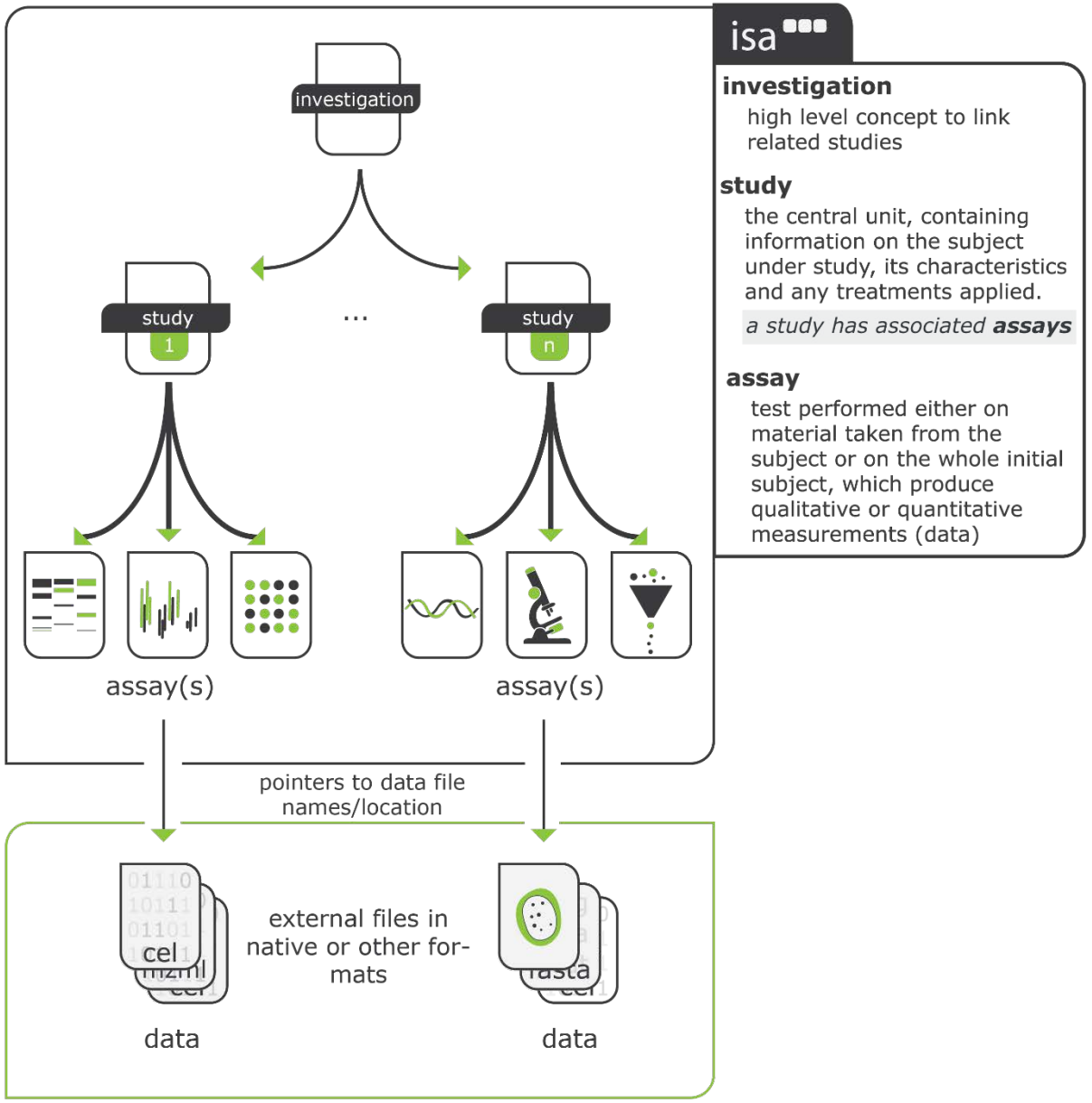
multi-omic experimental datasets to be submitted to repositories.

The solution:

- A general purpose, flexible, and domain agnostic format.
- A metadata framework.
- XSD>XML>XLS
- Open source tools hosted on github and licensed under MPL

# ISA-TAB Structure

[ISA-TAB specification]



# ISA-TAB Investigation Configuration

- Investigation file:
  - Investigation description
    - Fields and sections
    - Hierarchical components: Studies and Assays
  - Components of Multiple instances are listed in columns, such as ontology resources, assays and authors, etc.
  - Studies are in Study blocks
- Section
- Field

# ISA-TAB Investigation-level Sections – to be cont'd

| Section Name | Fields | Notes |
|---|---|---|
| ONTOLOGY SOURCE REFERENCE - | Term Source Name, Term Source File, Term Source Version, Term Source Description | |
| INVESTIGATION - | Investigation Identifier/Title/Description/Submission Date/Disease/Disease Term Accession Number/Disease Term Source REF/Outcome | |
| INVESTIGATION PUBLICATION - | Investigation PubMed ID/Publication DOI/Author List/Title/Status/Term Accession Number/Term Source REF | |
| INVESTIGATION CONTACTS - | Investigation Person Last Name/First Name/Mid Initials/Email/Phone/Fax/Address/Affiliation/Roles/Term Accession Number/Term Source REF | |

# ISA-TAB Investigation-level Sections – to be cont'd

| Section Name | Fields |
|---|---|
| STUDY | Study Identifier/Title/Description/Submission Date/Public Release Date/Disease/Disease Term Accession Number/Disease Term Source REF/Outcome/File Name |
| STUDY DESIGN DESCRIPTORS | Study Design Type/Design Type Term Accession Number/Term Source REF |
| STUDY PUBLICATION | Study PubMed ID/Publication DOI/Author List/Title/Status/Term Accession Number/Term Source REF |
| STUDY FACTORS | Study Factor Name/Type/Term Accession Number/Term Source REF |
| STUDY ASSAYS | Study Assay Measurement Type/Term Accession Number/Source REF/Technology Type/Term Accession Number/Source REF/Plaftorm/Measurement Name/Term Accession Number/Term Source REF |

# ISA-TAB Investigation-level Sections

| Name | Fields |
| --- | --- |
| STUDY PROTOCOLS | Study Protocol Name/Type/Term Accession Number/Term Source REF/Description/URI/Version/Parameters Name/Term Accession Number/Term Source REF/Component Name/Type/Type Term Accession Number/Term Source REF |
| STUDY CONTACTS | Study Person Last Name/First Name/Mid Initials/Email/Phone/Fax/Address/Affiliation/Roles/Term Accession Number/Term Source REF |
| Comments | |

# ISA-TAB Study-level Sections – to be cont'd

| Name | Definition | Example | Notes |
|---|---|---|---|
| Study | Identifier/Title/Description/Submission Date/Public Release Date/Disease/Outcome/File Name | | Field Group |
| Sample Name | | | |
| Source Name | | rat | |
| Characteristics | Such as Characteristics[genotype] | KanMx4 MATa/MATalpha ura3-52/ura3-52 leu2-1/+trp1-63/+his3-D200/+ hoD KanMx4/hoD | Optionally decorated with "Term Accession Number" and "Term Source REF", and optionally "Unit". Recurrent. |

# ISA-TAB Study-level Fields

| Name | Definition | Example | Notes |
|---|---|---|---|
| Parameter Value | Such as Parameter Value[blood sample volume] | | Optionally and usually decorated with “Term Accession Number”, “Term Source REF”, and optionally “Unit”. Recurrent. |
| Provider | | | |
| Protocol REF | | | |
| Factor Value | Such as Factor Value[rate] | 0.07 | Optionally decorated with “Term Accession Number” and “Term Source REF”, and/or “Unit” and the associated “Term accession Number” and “Term Source REF” . Recurrent. |
| Performer | | | |
| Date | | | |
| Comment | | This is a comment | |

# ISA-TAB Study Configuration

- Study Level
- Assay Level
  - Measurement
  - Technology
  - Protocols

# ISA-TAB Assay-level Fields – to be cont'd

| Name | Definition | Example | Notes |
|---|---|---|---|
| Sample Name | | NZ_0hrs_Grow1_Sample_1 | |
| Protocol REF | | mRNA extraction | |
| Assay Name | | | |
| Parameter Value[parameter name] | | | Optionally decorated with "Term Accession Number" and "Term Source REF", and optionally "Unit". Recurrent. |
| Unit | | | Often decorated with "Term Accession Number" and "Term Source REF". |
| Measurement Value[statistic(measurement name)] | Measurement Value[mean(injected dose fraction)] | | Often followed by Unit and the associated "Term Accession Number" and "Term Source REF". |
| Material Type | | Biological specimen<br>or<br>Internal/external in a_metabolome | Optionally decorated with "Term Accession Number" and "Term Source REF":<br>EHDAA<br>http://purl.obolibrary.org/obo/EHDAA_7477 |

# ISA-TAB Assay-level Fields

| Name | Definition | Example | Notes |
|---|---|---|---|
| | Such as Parameter Value[blood sample volume] | | Optionally and usually decorated with “Term Accession Number”, “Term Source REF”, and optionally “Unit”. Recurrent. |
| Provider | | | |
| | | | |
| Factor Value[factor name] | Such as Factor Value[rate] | 0.07 | Optionally decorated with “Term Accession Number” and “Term Source REF”, and/or “Unit” and the associated “Term accession Number” and “Term Source REF” . Recurrent. |
| Image File | | | |
| Raw Data File | | | |
| Derived Data File | | | |
| Performer | | | |
| Date | | | |
| Comment | | This is a comment | Can be recurrent: Comment[ArrayExpress Accession], Comment[ArrayExpress Raw Data URL] |

| Name | Definition | Example | Notes |
|---|---|---|---|
| Protocol REF | | mRNA extraction | Microarray/proteome |
| Extract Name | | NZ_0hrs_Sample_1_Extract | Microarray/proteome |
| Protocol REF | | biotin labeling | Microarray/proteome |
| Labeled Extract name | | NZ_0hrs_Sample_1_Labelled | Microarray/proteome |
| Label | | biotin | Microarray/proteome |
| Term Source REF | | CHEBI | Microarray/proteome |
| Term Source Number | | http://purl.obolibrary.org/obo/CHEBI_15956 | Microarray/proteome |
| Protocol REF | | EukGE-WS4 | |
| Hybridization Assay Name | | NZ_0hrs_Sample_1_Labelled_Hyb1 | Microarray/transcriptome |
| Comment[ArrayExpress Accession] | | E-MEXP-115 | Microarray/transcriptome |
| Comment[ArrayExpress Raw Data URL] | | E-MEXP-115 | Microarray/transcriptome |
| Comment[ArrayExpress Processed Data URL] | | E-MEXP-115 | Microarray/transcriptome |
| Array Design REF | | A-AFFY-27 | Microarray/transcriptome |
| Scan Name | | NZ_0hrs_Sample_1_Labelled_Hyb1_Scan1 | Microarray/transcriptome |
| Array Data File | | E-MAXD-4-raw-data-426648549.txt | Microarray/transcriptome |
| Data Transformation Name | | data processing | Microarray/transcriptome |
| Derived Array Data File | | E-MAXD-4-processed-data-1342566476.txt | Microarray/transcriptome |
| Factor Value[dose] | | 0 | |
| Unit | | ng /ml | |

| Name | Definition | Example | Notes |
|---|---|---|---|
| Acquisition Parameter Data File | | | Metabolite profile NMR |
| Protocol REF | | | |
| NMR Assay Name | | | |
| Free Induction Decay Data File | | | |
| Protocol REF | | | |
| Normalization Name | | | NMR/proteome |
| Protocol REF | | | |
| Data Transformation Name | | | NMR/proteome |
| Derived Spectral Data File | | | NMR/proteome |
| Metabolite Assignment File | | | |
| Data_collection method/data processing operation/ sample preparation for assay | | FT-ICR mass spectrometry/Signal processing of mass spectra (or Apodisation, zero-filling&Fourier Transformation) /Sample preparation | |
| Metabolome | | | METABOLOME |
| MS Assay Name | | | Metabolome/proteome |
| Raw Spectral Data File | | | proteome |
| Comment[PRIDE Processed Data Accession] | | | proteome |
| Protein Assignment File | | | proteome |
| Peptide Assignment File | | | proteome |
| Post Translational Modification Assignment File | | | proteome |

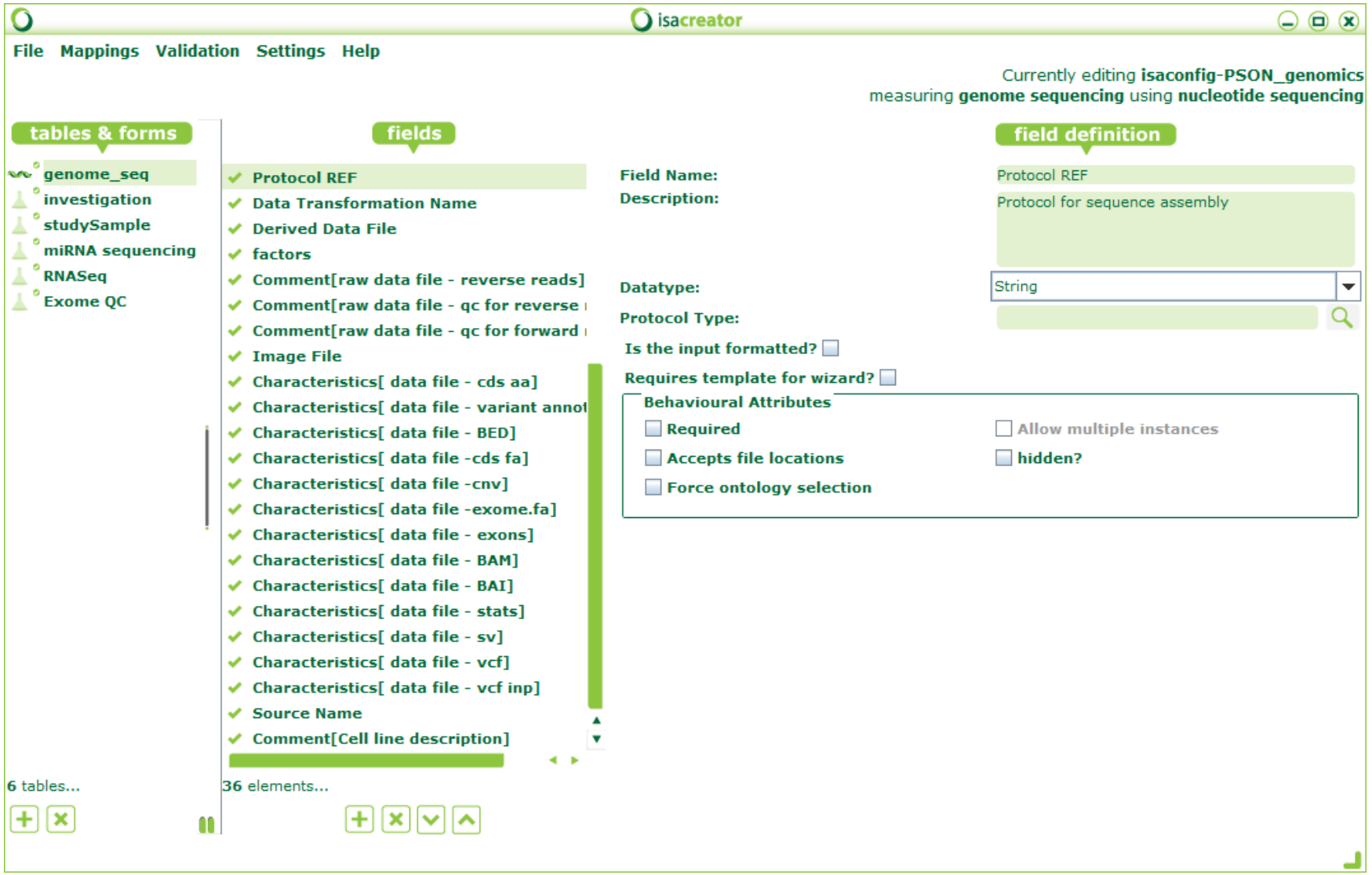
# ISA-TAB Software Suit

- Viewer:
  - Investigation: Samples, Assays, Publications, Contacts
  - Study: Factors, Protocols
  - demo
- Creator
- Validator
- Configurator
- Others: Converter, BII, etc.

# A Look at CSSI DCC Dataset

| Cell Line Name | PBCF Cell Line Catalog Number | Exome | | | | RNASeq | | | | miRNA | | | | Summary Data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Raw | BAM | Annotation | Putative Mutations | Raw | BAM | Expression | Fusion | Raw | BAM | Mature Unique Known list | Predicted miRNAs | Exome Stats | WES Coverage Plot | RNASeq Stats | Fusion Summary page | Gene Expression Heatmap | miRNA Summary page |
| MCF-10A-JSB | NCI-PBCF-1000 | yes | NCI-PBCF-1000.mark.bam | NCI-PBCF-1000.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| MCF-B7-TS | NCI-PBCF-1001 | yes | NCI-PBCF-1001.mark.bam | NCI-PBCF-1001.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| SW620 | NCI-PBCF-CCL227 | yes | NCI-PBCF-CCL227.mark.bam | NCI-PBCF-CCL227.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| SW480 | NCI-PBCF-CCL228 | yes | NCI-PBCF-CCL228.mark.bam | NCI-PBCF-CCL228.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| LoVo | NCI-PBCF-CCL229 | yes | NCI-PBCF-CCL229.mark.bam | NCI-PBCF-CCL229.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| HCT116 | NCI-PBCF-CCL247 | yes | NCI-PBCF-CCL247.mark.bam | NCI-PBCF-CCL247.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| NCI-H2126 | NCI-PBCF-CCL256 | yes | NCI-PBCF-CCL256-1.mark.bam | NCI-PBCF-CCL256.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| NCI-BL2126 | NCI-PBCF-CCL256.1 | yes | NCI-PBCF-CCL256.mark.bam | NCI-PBCF-CCL256-1.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| RWPE-1 | NCI-PBCF-CRL11609 | yes | NCI-PBCF-CRL11609.mark.bam | NCI-PBCF-CRL11609.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| PC-3 | NCI-PBCF-CRL1435 | yes | NCI-PBCF-CRL1435.mark.bam | NCI-PBCF-CRL1435.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| Panc-1 | NCI-PBCF-CRL1469 | yes | NCI-PBCF-CRL1469.mark.bam | NCI-PBCF-CRL1469.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| ZR-75-1 | NCI-PBCF-CRL1500 | yes | NCI-PBCF-CRL1500.mark.bam | NCI-PBCF-CRL1500.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| A375 | NCI-PBCF-CRL1619 | yes | NCI-PBCF-CRL1619.mark.bam | NCI-PBCF-CRL1619.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| WM266-4 | NCI-PBCF-CRL1676 | yes | NCI-PBCF-CRL1676.mark.bam | NCI-PBCF-CRL1676.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| T98G | NCI-PBCF-CRL1690 | yes | NCI-PBCF-CRL1690.mark.bam | NCI-PBCF-CRL1690.vcf.out | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |

# PS-ON Genomics Data and ISA-TAB



# Conclusion & Brief Discussion

1. ISA-TAB framework is sound and moving forward.
2. Software not commercial products, not very robust, steep learning curve for biologists.
3. Not supporting omics analysis – ISA-TAB 2.0
4. Moving away from desktop tools and toward web services.

What do all of these mean for CSSI DCC project?

# Challenges for ISA-TAB with Archive Submission

1. Current ISA-TAB model doesn't accommodate "above-assay-level" analysis
   - Cross sample normalization
   - Differential gene expression studies
   - Population genetics studies
2. Burden on users for omics submission – requiring configuration and automation for batch processing
3. User experience suffers and maintenance is heavy because of the existence of 3rd party tools of various sources/versions, and the quality.
4. ISA-TAB development is moving away from desktop tools and toward web services.