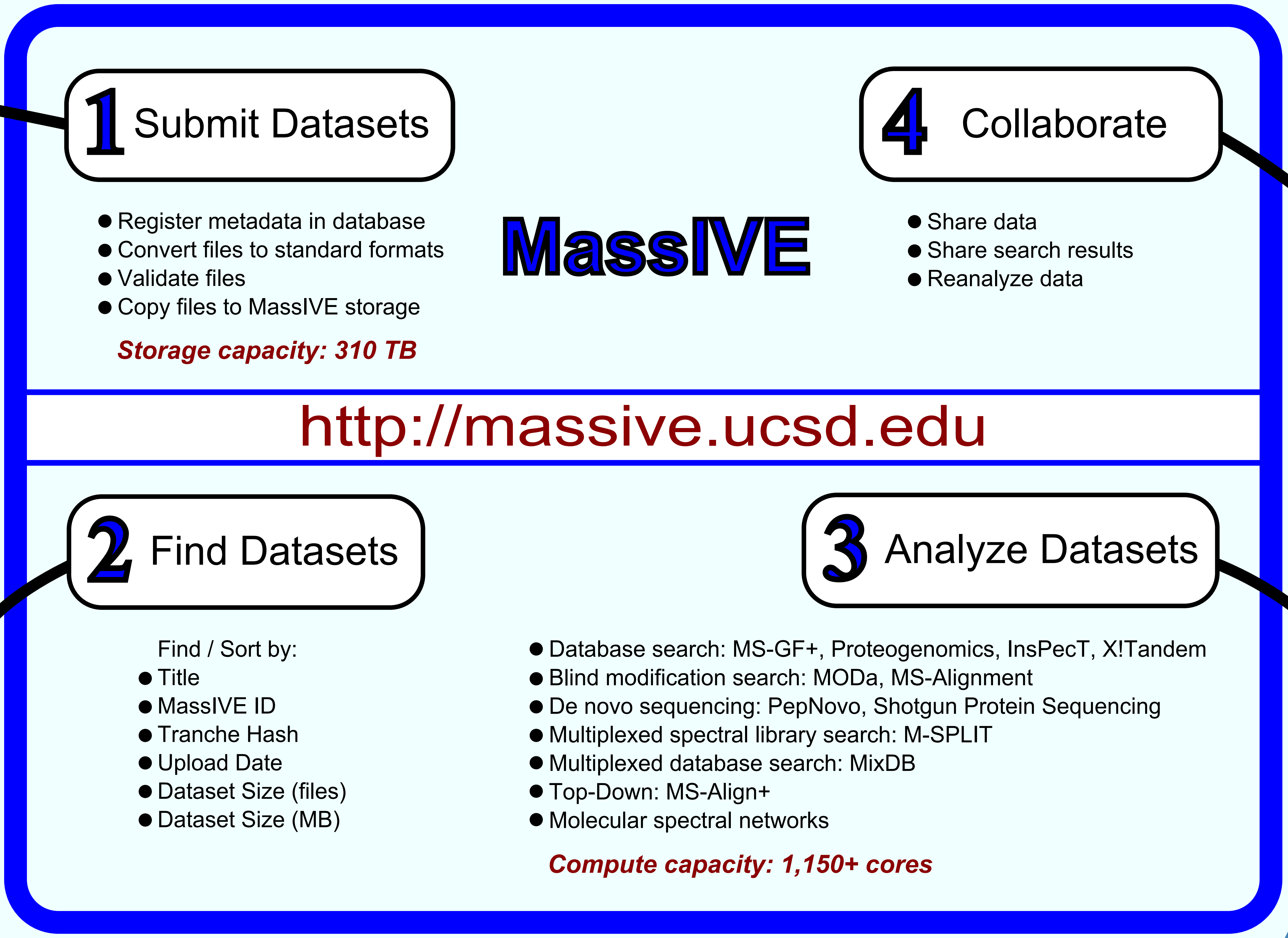
Center for
Computational
Mass
Spectrometry

MassIVE: *Mass* Spectrometry *I*nteractive *V*irtual *E*nvironment for Data Sharing in Proteomics

Jeremy Carver, Ian Kaufman, Claudiu Farcas, Vineet Bafna, Nuno Bandeira
Center for Computational Mass Spectrometry
University of California, San Diego, USA

The public availability of mass spectrometry data is increasingly recognized as indispensable for the future of computational mass spectrometry and, consequently, for the collective ability of the proteomics community to progress towards efficiently reutilizing the billions of spectra already acquired and the many millions more being generated on a daily basis. But despite the pioneering efforts of the Tranche/ProteomeCommons system and the essential integration services under development by the ProteomeXchange consortium, there is still a pressing need for data repositories designed not only for reliable sharing of primary data but also offering user-friendly tools for advanced data analysis and visualization. MassIVE is a new mass spectrometry data repository designed to store, browse, redistribute, re-analyze and integrate all publicly available mass spectrometry data.



MassIVE

1 Submit Datasets

- Register metadata in database
- Convert files to standard formats
- Validate files
- Copy files to MassIVE storage

Storage capacity: 310 TB

4 Collaborate

- Share data
- Share search results
- Reanalyze data

http://massive.ucsd.edu

2 Find Datasets

Find / Sort by:
- Title
- MassIVE ID
- Tranche Hash
- Upload Date
- Dataset Size (files)
- Dataset Size (MB)

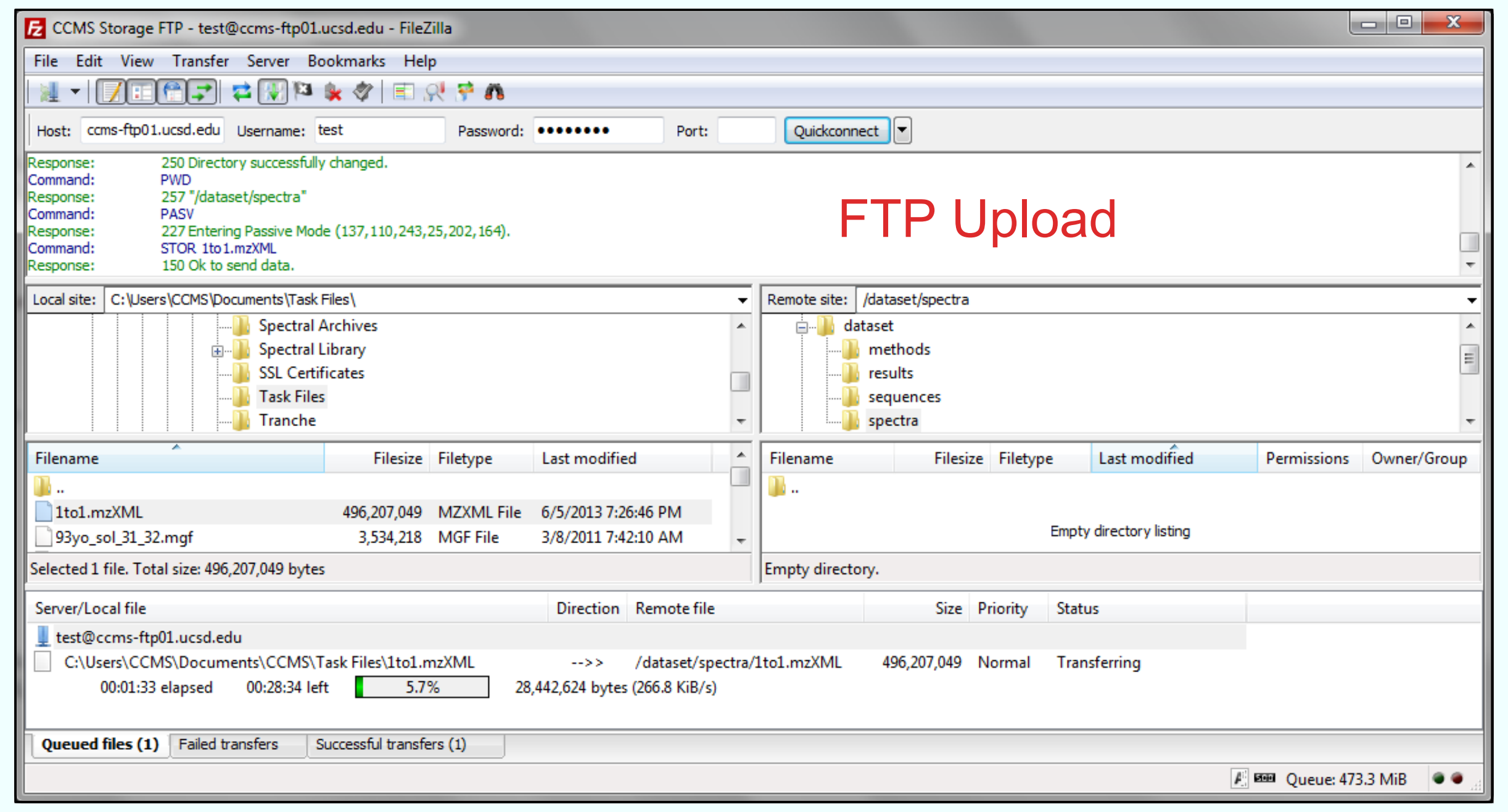
3 Analyze Datasets

- Database search: MS-GF+, Proteogenomics, InsPecT, X!Tandem
- Blind modification search: MODa, MS-Alignment
- De novo sequencing: PepNovo, Shotgun Protein Sequencing
- Multiplexed spectral library search: M-SPLIT
- Multiplexed database search: MixDB
- Top-Down: MS-Align+
- Molecular spectral networks

Compute capacity: 1,150+ cores

MassIVE Submission Interface

MassIVE dataset submission is implemented as a workflow within the ProteoSAFe platform, consisting of two important steps: first, dataset files are uploaded via FTP, and then these files are validated and securely copied to the MassIVE dataset repository.



FTP Upload

Host: ccms-ftp01.ucsd.edu Username: test Password: •••••••• Port:

```
Response:   250 Directory successfully changed.
Command:    PWD
Response:   257 "/dataset/spectra"
Command:    PASV
Response:   227 Entering Passive Mode (137,110,243,25,202,164).
Command:    STOR 1to1.mzXML
Response:   150 Ok to send data.
```

Local site: C:\Users\CCMS\Documents\Task Files\

Remote site: /dataset/spectra

Filename	Filesize	Filetype	Last modified
..			
1to1.mzXML	496,207,049	MZXML File	6/5/2013 7:26:46 PM
93yo_sol_31_32.mgf	3,534,218	MGF File	3/8/2011 7:42:10 AM

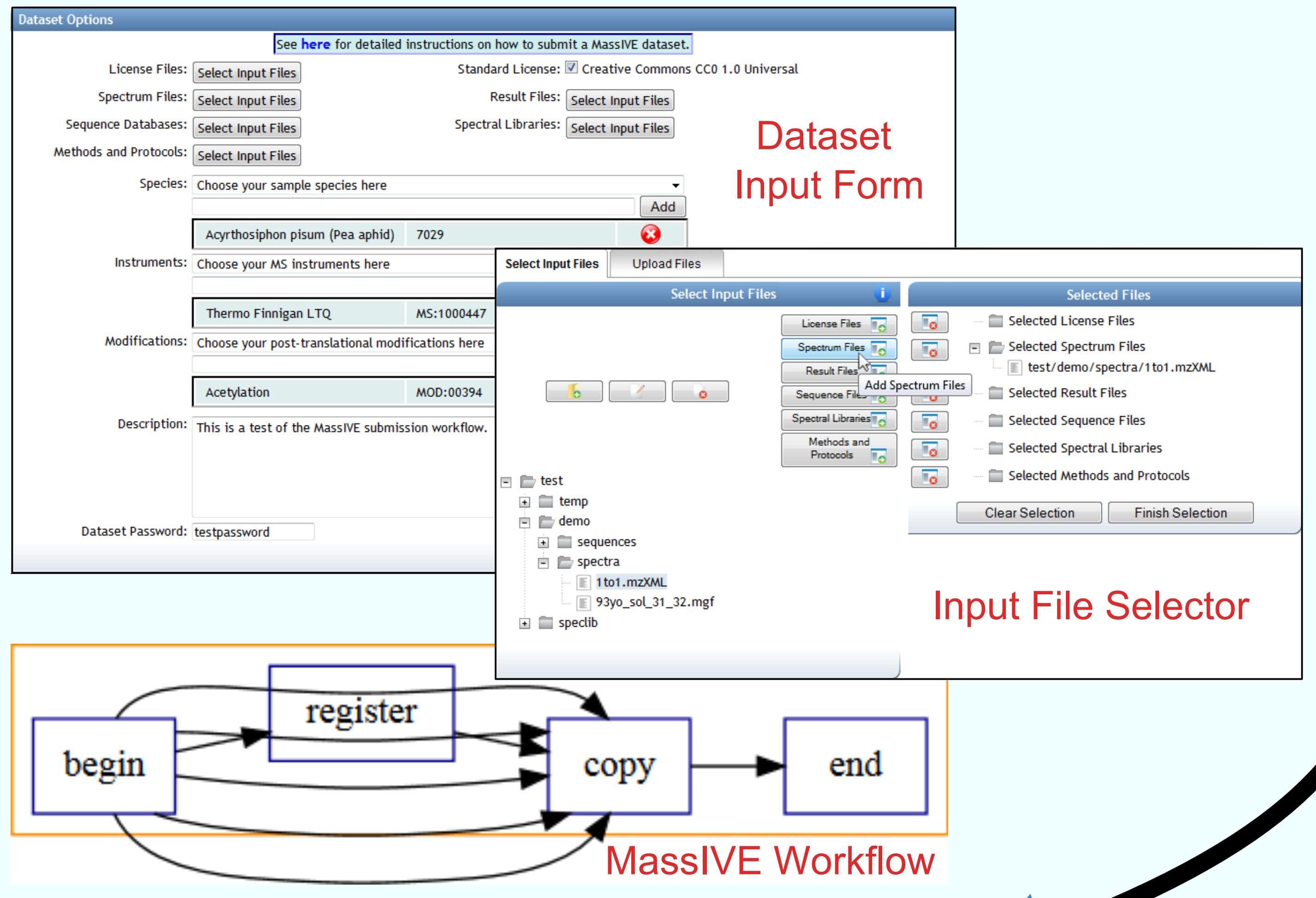
Selected 1 file. Total size: 496,207,049 bytes

Empty directory listing

Empty directory.

Server/Local file	Direction	Remote file	Size	Priority	Status
test@ccms-ftp01.ucsd.edu					
C:\Users\CCMS\Documents\CCMS\Task Files\1to1.mzXML	-->>	/dataset/spectra/1to1.mzXML	496,207,049	Normal	Transferring
00:01:33 elapsed 00:28:34 left 5.7%		28,442,624 bytes (266.8 KiB/s)			

Queued files (1) | Failed transfers | Successful transfers (1) | Queue: 473.3 MiB



Dataset Input Form

Dataset Options

See here for detailed instructions on how to submit a MassIVE dataset.

License Files: Select Input Files
Spectrum Files: Select Input Files
Sequence Databases: Select Input Files
Methods and Protocols: Select Input Files
Standard License: ☑ Creative Commons CC0 1.0 Universal
Result Files: Select Input Files
Spectral Libraries: Select Input Files
Species: Choose your sample species here — Acyrthosiphon pisum (Pea aphid) 7029
Instruments: Choose your MS instruments here — Thermo Finnigan LTQ MS:1000447
Modifications: Choose your post-translational modifications here — Acetylation MOD:00394
Description: This is a test of the MassIVE submission workflow.
Dataset Password: testpassword

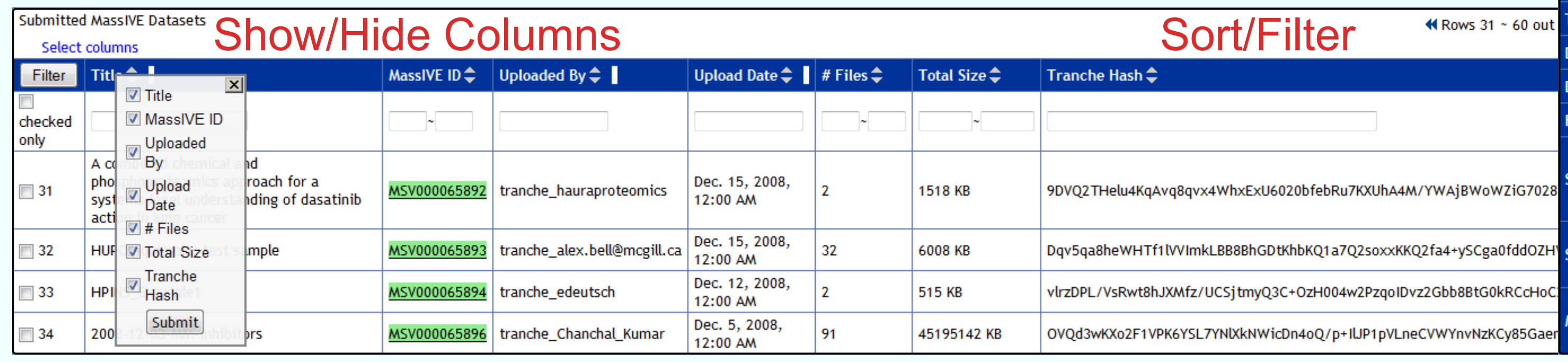
Input File Selector

MassIVE Workflow

MassIVE Dataset Browser

After a dataset is submitted, its contents are private until they are released by the submitting user. Password protection is used to allow reviewer access. All datasets that have been made public can be browsed and searched via ProteoSAFe.

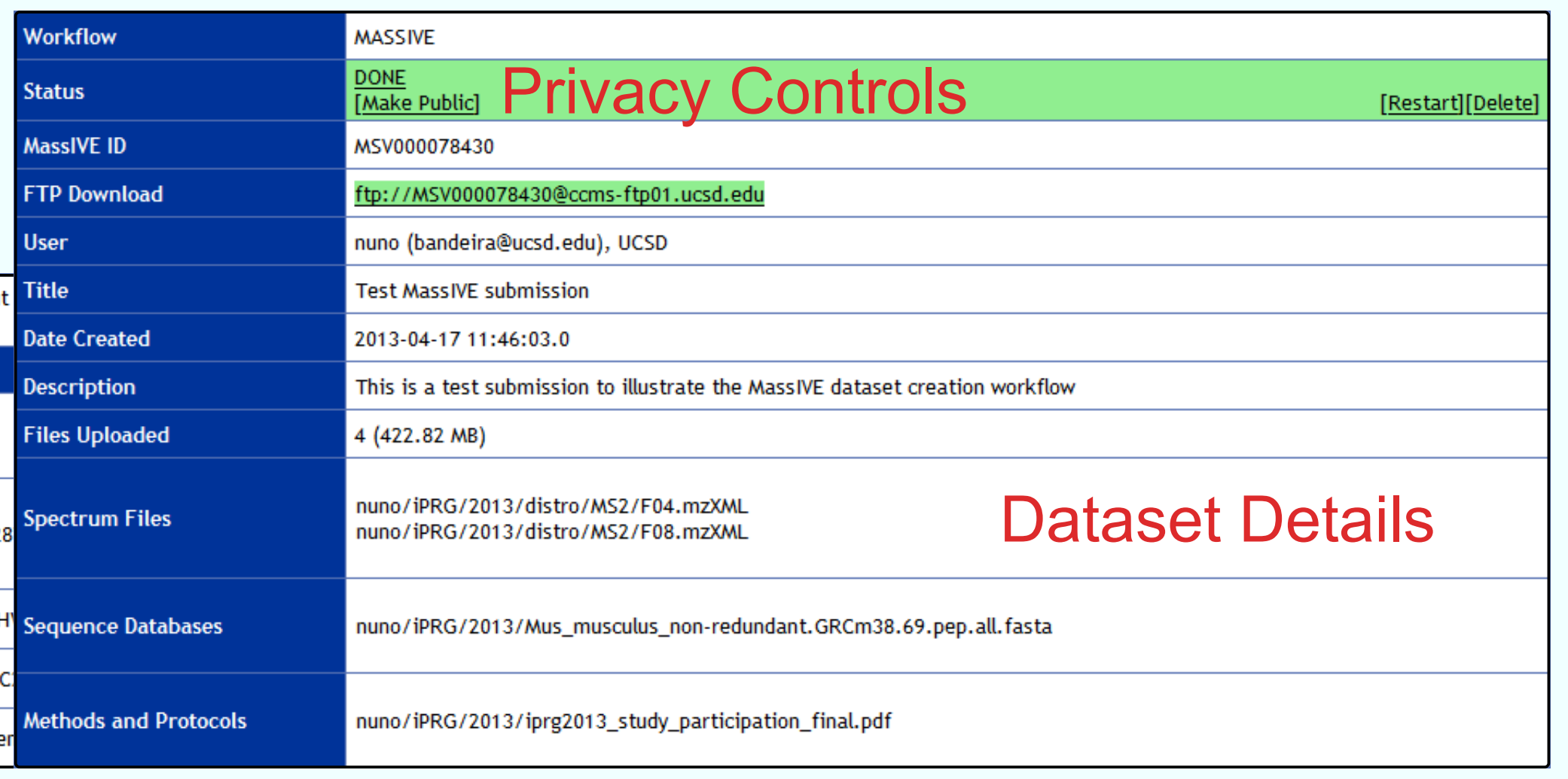
Dataset List



Show/Hide Columns

Sort/Filter

	Title	MassIVE ID	Uploaded By	Upload Date	# Files	Total Size	Tranche Hash
☐ checked only							
☐ 31	A c… pho… syst… acti… …roach for a …ding of dasatinib	MSV000065892	tranche_hauraproteomics	Dec. 15, 2008, 12:00 AM	2	1518 KB	9DVQ2THelu4KqAvq8qvx4WhxExU6020bfebRu7KXUhA4M/YWAjBWoWZiG7028
☐ 32	HUF… …mple	MSV000065893	tranche_alex.bell@mcgill.ca	Dec. 15, 2008, 12:00 AM	32	6008 KB	Dqv5qa8heWHTf1lVVImkLBB8BhGDtKhbKQ1a7Q2soxxKKQ2fa4+ySCga0fddOZH
☐ 33	HPI…	MSV000065894	tranche_edeutsch	Dec. 12, 2008, 12:00 AM	2	515 KB	vlrzDPL/VsRwt8hJXMfz/UCSjtmyQ3C+OzH004w2PzqoIDvz2Gbb8BtG0kRCcHoC
☐ 34	200… …rs	MSV000065896	tranche_Chanchal_Kumar	Dec. 5, 2008, 12:00 AM	91	45195142 KB	OVQd3wKXo2F1VPK6YSL7YNlXkNWicDn4oQ/p+lUP1pVLneCVWYnvNzKCy85Gaer



Privacy Controls

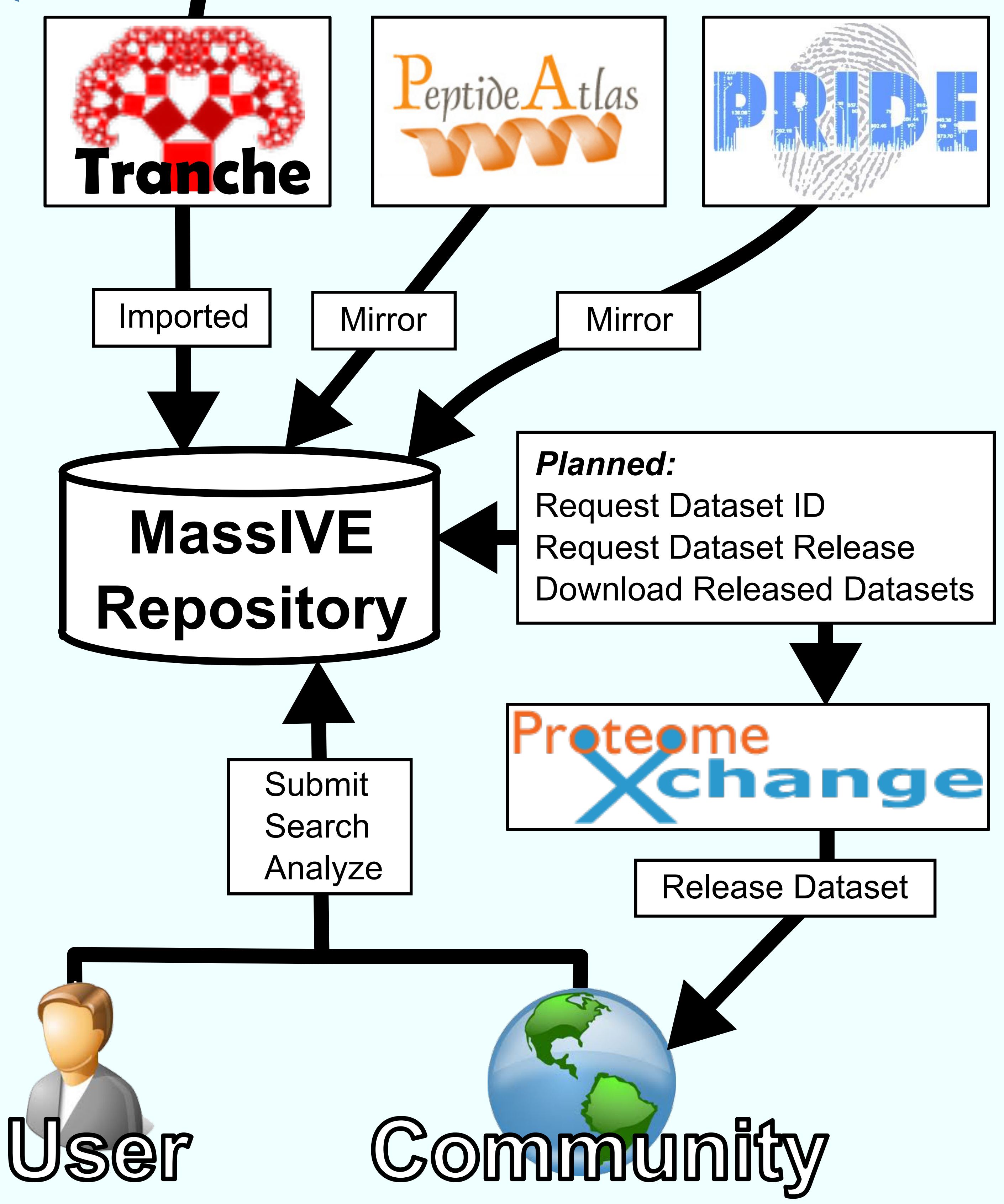
Dataset Details

Workflow	MASSIVE
Status	DONE [Make Public] [Restart][Delete]
MassIVE ID	MSV000078430
FTP Download	ftp://MSV000078430@ccms-ftp01.ucsd.edu
User	nuno (bandeira@ucsd.edu), UCSD
Title	Test MassIVE submission
Date Created	2013-04-17 11:46:03.0
Description	This is a test submission to illustrate the MassIVE dataset creation workflow
Files Uploaded	4 (422.82 MB)
Spectrum Files	nuno/iPRG/2013/distro/MS2/F04.mzXML nuno/iPRG/2013/distro/MS2/F08.mzXML
Sequence Databases	nuno/iPRG/2013/Mus_musculus_non-redundant.GRCm38.69.pep.all.fasta
Methods and Protocols	nuno/iPRG/2013/iprg2013_study_participation_final.pdf

Repository Collaboration

***Tranche → MassIVE* Import**

12,634 13,299,264 25.36	Datasets Files Terabytes	→	*Originally in Tranche*
6,307 11,903,244 18.98	Datasets Files Terabytes	→	*Not recovered from Tranche*
3,049 827,891 5.20	Datasets Files Terabytes	→	*Recovered incompletely from Tranche*
3,278 568,129 1.18	**Datasets Files Terabytes**	→	***Recovered intact, copied to MassIVE***



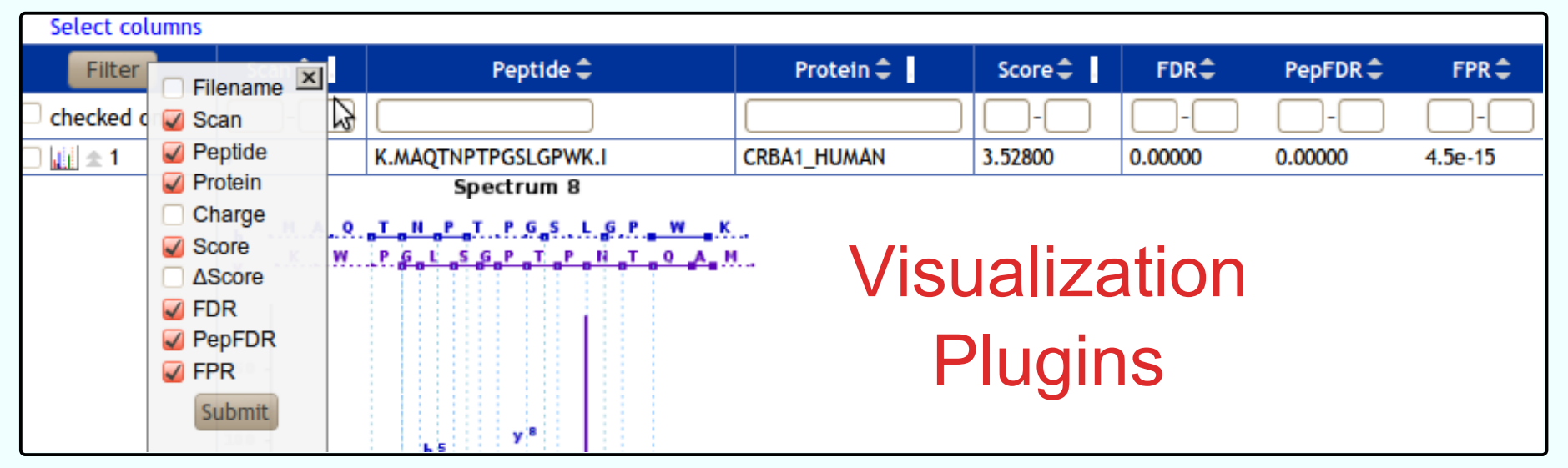
MassIVE Repository

Planned:
Request Dataset ID
Request Dataset Release
Download Released Datasets

ProteoSAFe Analysis Workflows

Once data is submitted to the repository and made public, it can then be searched and re-analyzed using the existing tools offered by ProteoSAFe.

Tabular Result Views



Visualization Plugins

	Peptide	Protein	Score	FDR	PepFDR	FPR
☐ checked						
☐ 1	K.MAQTNPTPGSLGPWK.I	CRBA1_HUMAN	3.52800	0.00000	0.00000	4.5e-15