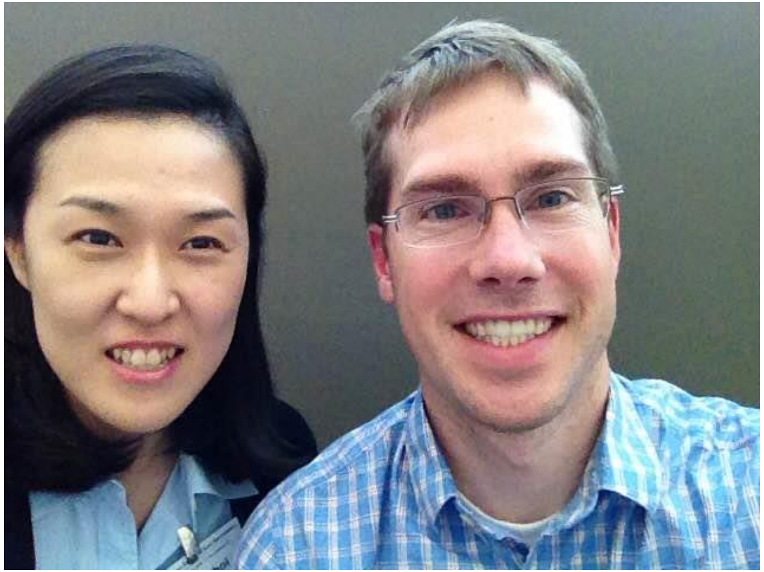
Open Source Software Packages: Using and Making your conbributions

June 5, 2017

Bioinformatics MS Interest Group

Your hosts



Meena Choi

Post doc.

Northeastern University

Statistical methods for quantitative proteomics

Samuel Payne

Scientist

Pacific Northwest National Lab

Integrative Omics

Outline

- General Intro – Meena Choi
- mzRefinery/proteowizard - Sam Payne
- openMS - Oliver Kohlbacher
- Skyline - Brendan MacLean
- General discussion on open source
  - Ask questions for the General Discussion http://bit.ly/2qNZVBU
  - Shout-out for Open Source tool http://bit.ly/2qVHVo7

Oliver Kohlbacher

- The chair of Applied Bioinformatics at University of Tübingen & fellow at the Max Plank Institute
- OpenMS (openms.de)

Brendan MacLean

- Principal developer for Skyline (skyline.ms)
- University of Washington

Skyline

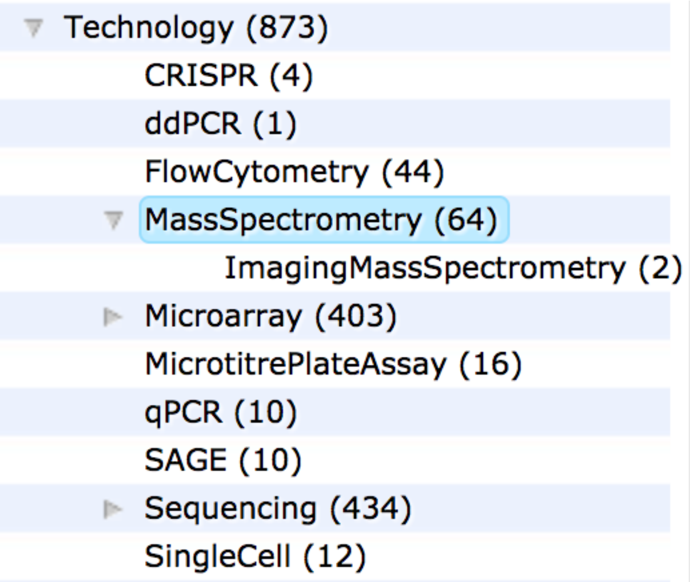
Ask questions or comments : http://bit.ly/2qNZVBU

- Why have open source?
- What are the advantages and disadvantages between open source and private closed-source software?
- How should a developer consider the question of making a project open source or not?
- What is appropriate level of guide/documentation to help new developers?
- How to incentivize people to contribute to open source software?

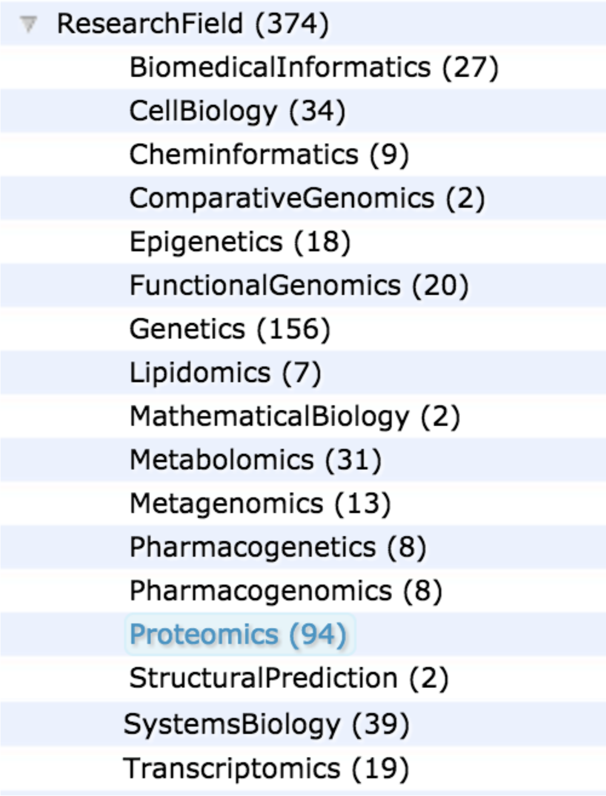
Bioconductor.org

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

biocViews search



- Technology (873)
  - CRISPR (4)
  - ddPCR (1)
  - FlowCytometry (44)
  - MassSpectrometry (64)
    - ImagingMassSpectrometry (2)
  - Microarray (403)
  - MicrotitrePlateAssay (16)
  - qPCR (10)
  - SAGE (10)
  - Sequencing (434)
  - SingleCell (12)



- ResearchField (374)
  - BiomedicalInformatics (27)
  - CellBiology (34)
  - Cheminformatics (9)
  - ComparativeGenomics (2)
  - Epigenetics (18)
  - FunctionalGenomics (20)
  - Genetics (156)
  - Lipidomics (7)
  - MathematicalBiology (2)
  - Metabolomics (31)
  - Metagenomics (13)
  - Pharmacogenetics (8)
  - Pharmacogenomics (8)
  - Proteomics (94)
  - StructuralPrediction (2)
  - SystemsBiology (39)
  - Transcriptomics (19)

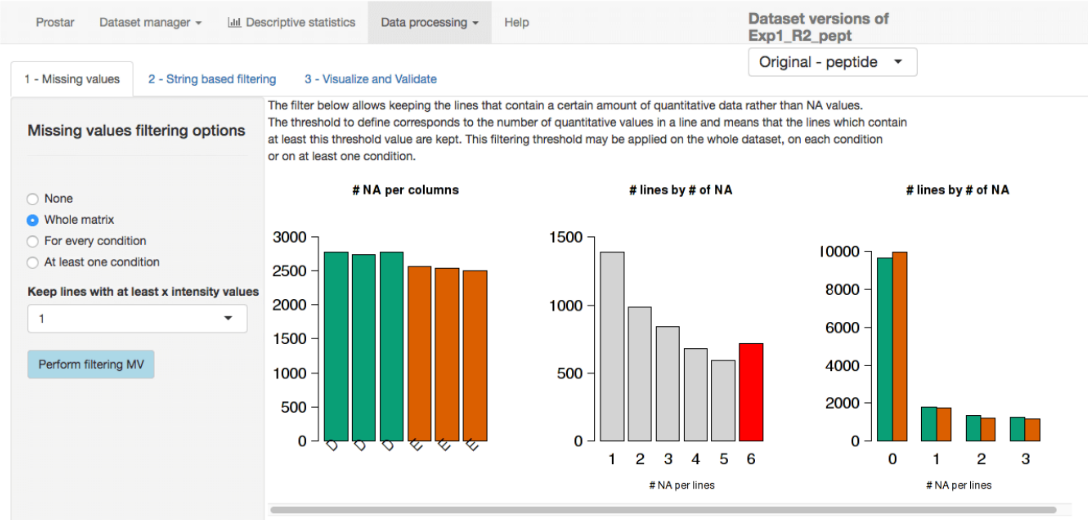
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

R package development

- Provide the framework for developing package : basic structure, requirements…
  - Requirements :
  1. pass check or BiocCheck on all supported platforms (their own checking system)
  2. Documents
     - DESCRIPTION, NAMESPACE, vignette, help file, NEWS
  3. Review process (2-5 weeks)
     - submit a GitHub repository
     - a reviewer will be assigned and a detailed package review is returned.
     - the process is repeated until the package is accepted to Bioconductor.
- Maintaining the packages across release cycles (twice a year) + deprecate packages
- Import or depend on other packages in Bioconductor or CRAN

R package as software

- Easy to make open source software for new method development.
- Reproducible : R script, R markdown
- Can improve GUI with Shiny
- Not easy to work with other language

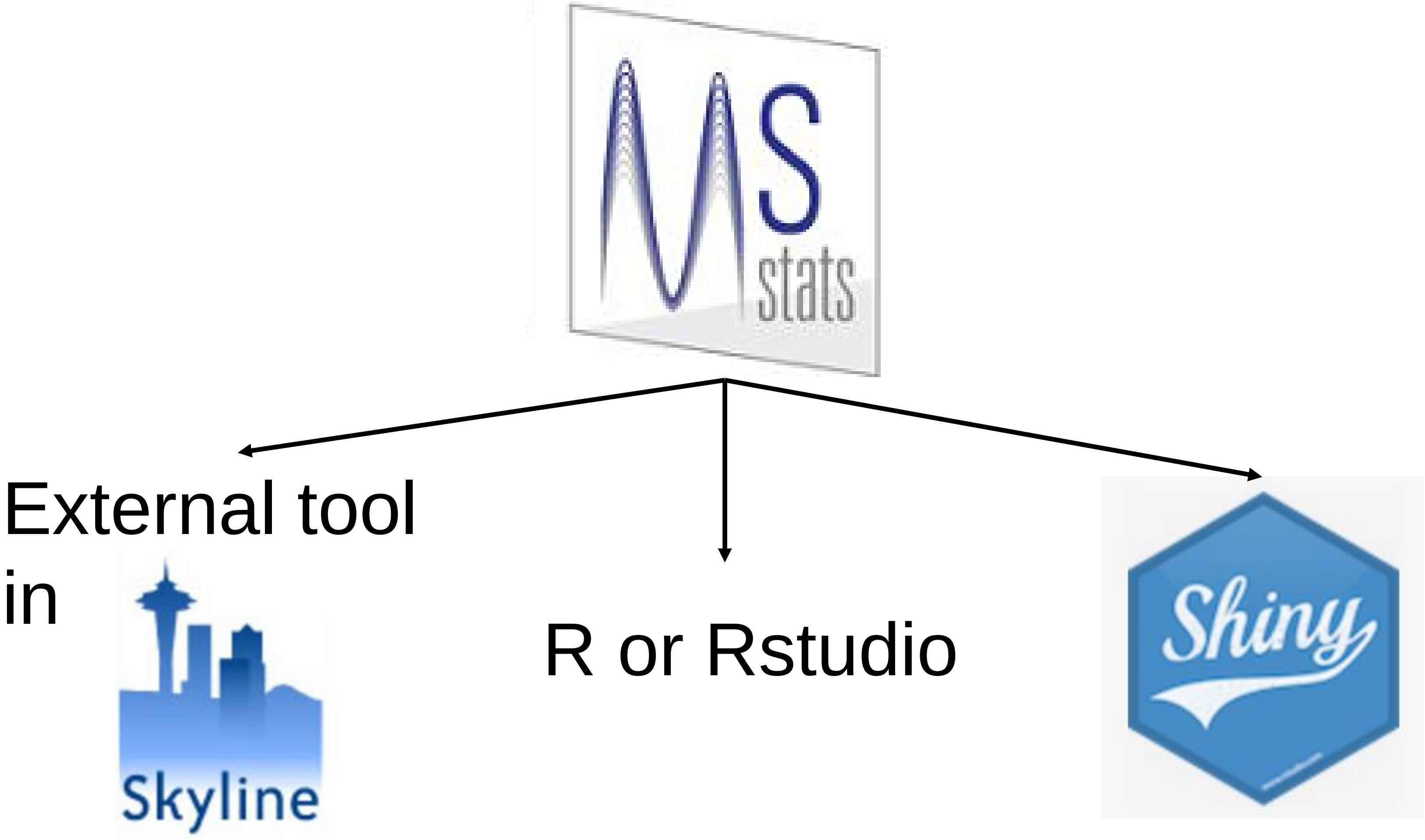


Missing values filtering options

○ None
● Whole matrix
○ For every condition
○ At least one condition

Keep lines with at least x intensity values: 1

The filter below allows keeping the lines that contain a certain amount of quantitative data rather than NA values. The threshold to define corresponds to the number of quantitative values in a line and means that the lines which contain at least this threshold value are kept. This filtering threshold may be applied on the whole dataset, on each condition or on at least one condition.



External tool in

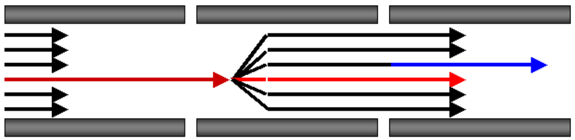
R or Rstudio

Skyline

Targeted Mass Spec Environment

Reflections on open source projects

Brendan MacLean
MacCoss Lab
University of Washington

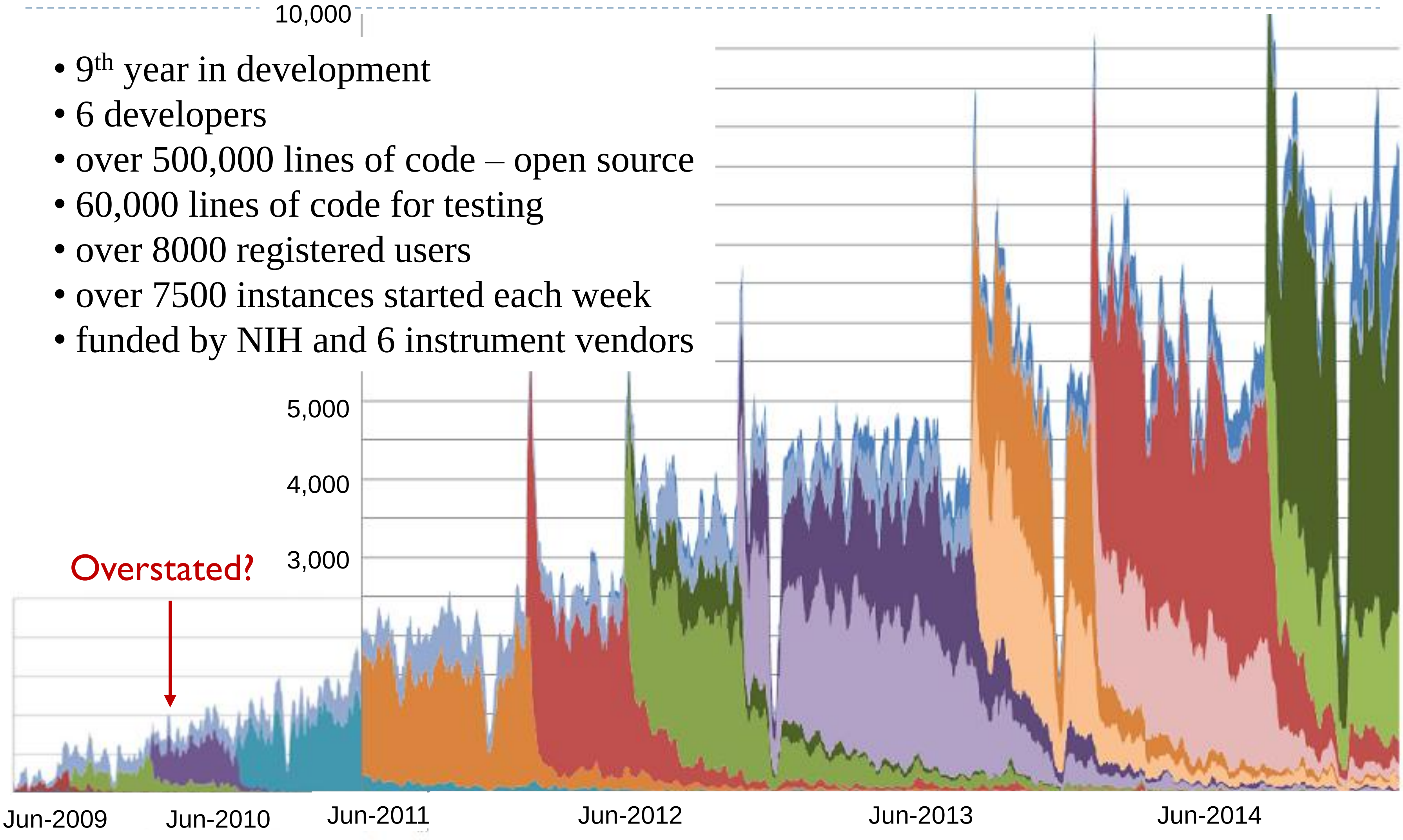


Personal Open Source History

- Microsoft 1991 – Microsoft Foundation Classes
- BEA Systems 2001 – Apache XML Beans
  - Last release 2012 – Retired 2014
- LabKey Server 2003 (originally CPAS)
  - X! Tandem contributions – pluggable scoring & k-score
  - TPP X! Tandem pipeline
- Skyline as a ProteoWizard subproject
  - Drove vendor acceptance of open source licensing
- Panorama as a module in LabKey Server

Skyline Project Overview

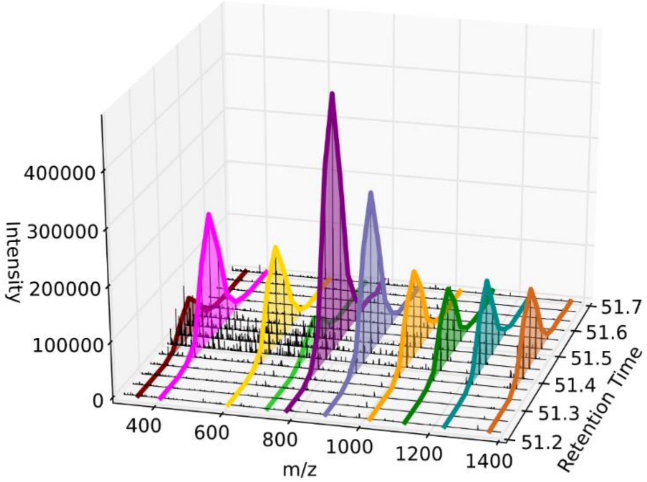
- 9th year in development
- 6 developers
- over 500,000 lines of code – open source
- 60,000 lines of code for testing
- over 8000 registered users
- over 7500 instances started each week
- funded by NIH and 6 instrument vendors

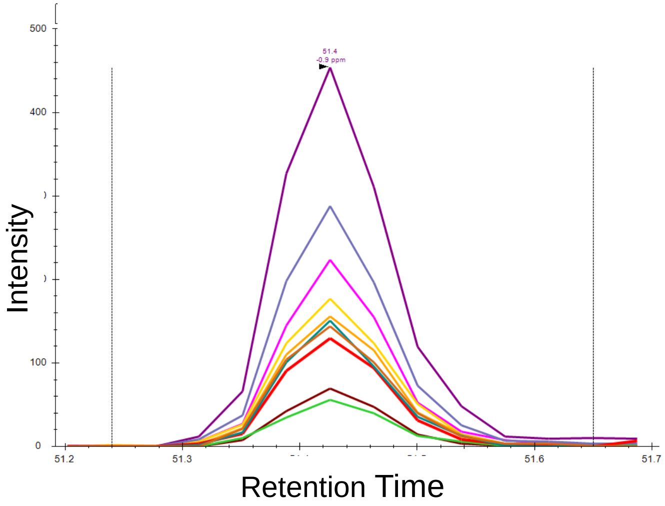


Chromatography-based Quantification

- SRM – Selected ion chromatograms
- PRM – Extracted ion chromatograms
- DIA – Extracted ion chromatograms
- DDA – Extracted ion chromatograms from MS1-only

Acquisition	Targeted	Survey
More Selective	PRM	DIA
Less Selective	SRM	DDA





Aggregating and Publishing

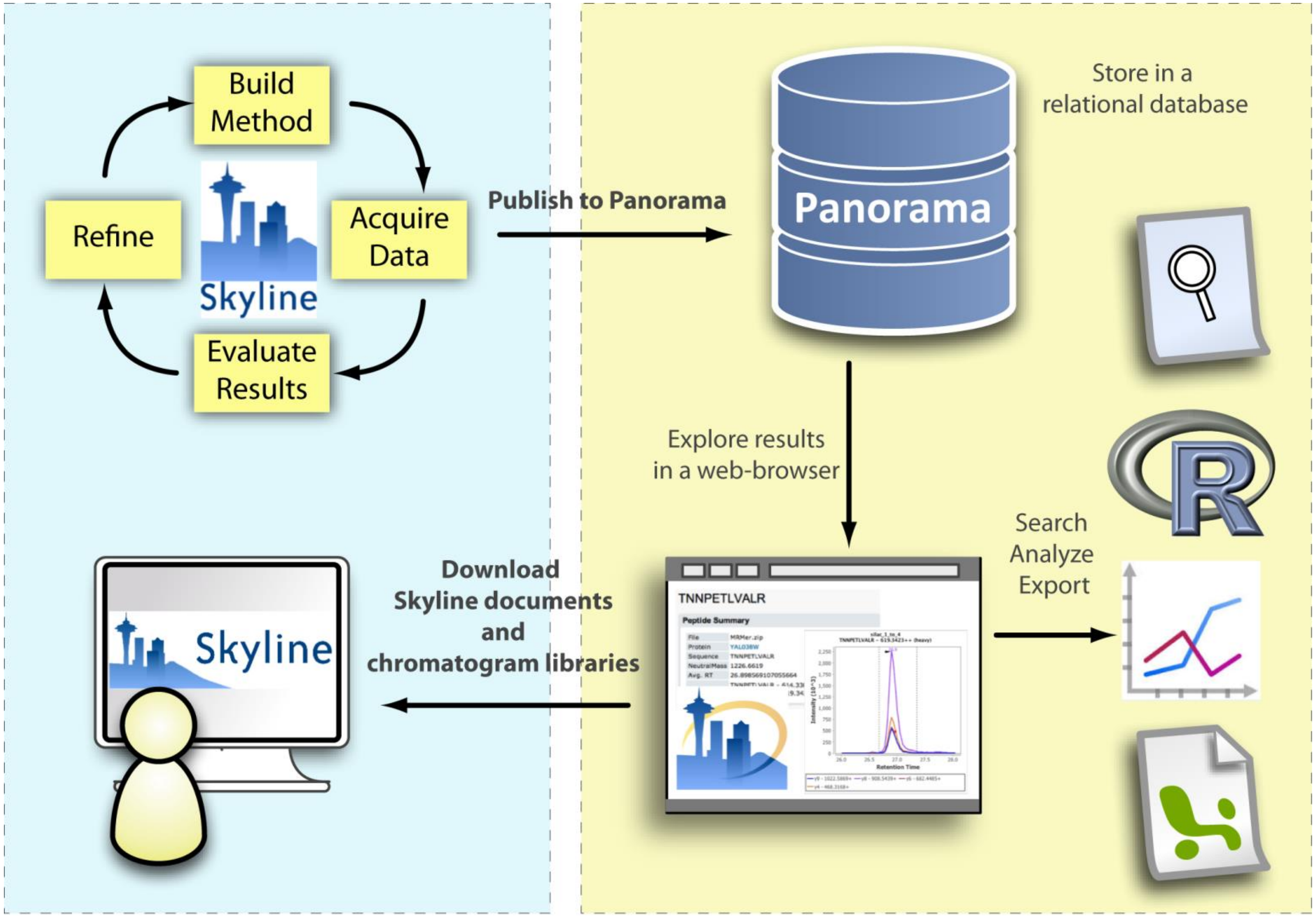
- Publish fully annotated Skyline documents
- Build chromatogram libraries
- Aggregate lab QC data

Panorama

- Free hosted version (http://panoramaweb.org)
  - >220 separate projects so far (CPTAC, LINCS & ABRF sPRG)
  - >2500 data sets uploaded (+7000 QC documents)
  - User controlled security
- Locally installable server application
  - Roche, Genentech, [unnamed], Merck, CPTAC, Amgen
- Free and open source (Apache 2.0)

Sharma, et. al, *J. Proteome Res. 2014*

Skyline/Panorama Workflow



On Open Source Licensing

Most permissive

↓

Most restrictive

- Public Domain
- Apache 2.0
  - Explicit rights to patents
- Berkeley Software Distribution (BSD)
  - Use as you please
- Artistic, Mozilla …
  - No branching allowed, adaptations must be public
- LGPL
  - Backwards architecting of using software
- GPL
  - “Viral” – users must open their own source

More Permissive Has Benefits

- May inspire broader adoption
  - Adoption is critical to all software
- May inspire trust from funders
  - Public grants
  - For profit companies
    - Instrument vendors
    - Pharma companies
- May inspire more outside contribution

Why people worry about going open

- Loss of "control"
  - Others will jump in and push the code places I don't want
- Loss of revenue opportunity
  - Having a free offering limits revenue potential
- Loss of advantage
  - Others can see and steal my best ideas
- Exposure
  - My code is not ready to share with others

The biggest open source fallacy

- Overestimating interest in your project
  - If I open source the project, it will go faster or last longer…
  - More likely benefits:
    - Extra scrutiny
    - Occasionally, inspired contributions
    - Broader adoption and interest
  - If I open source, others will read the code and find it lacking…
  - If it is important enough to read, they may offer improvements.
  - Let's open source that project that grad student left behind

What motivates involvement?

Skyline Team

- Nick Shulman

- Don Marsh

- Brian Pratt

- Max Horowitz-Gelb

- Vagisha Sharma

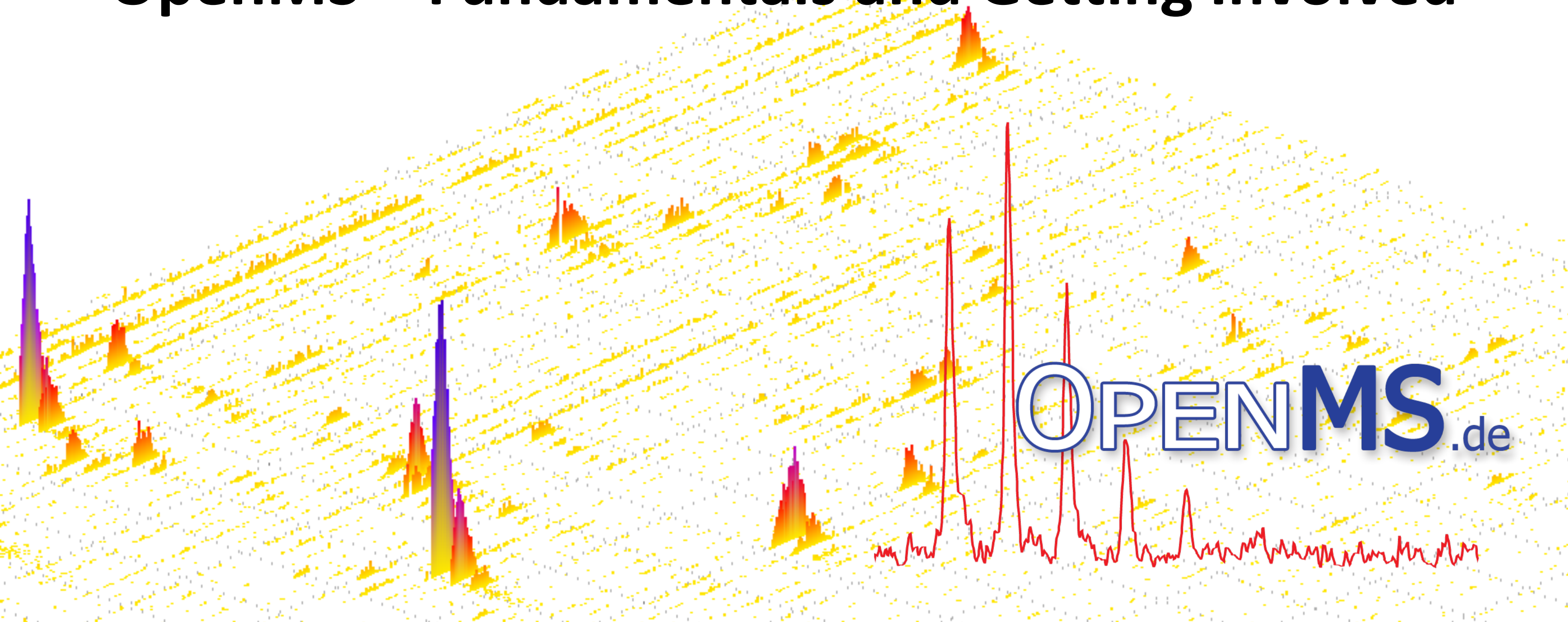
- Nat Brace

- Kaipo Tamura

- Yuval Boss

ASMS 2017 – Open Source Software Packages

OpenMS – Fundamentals and Getting Involved



Oliver Kohlbacher

University of Tübingen and
MPI for Developmental Biology
KohlbacherLab.org | @okohlbacher

MAX-PLANCK-GESELLSCHAFT

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

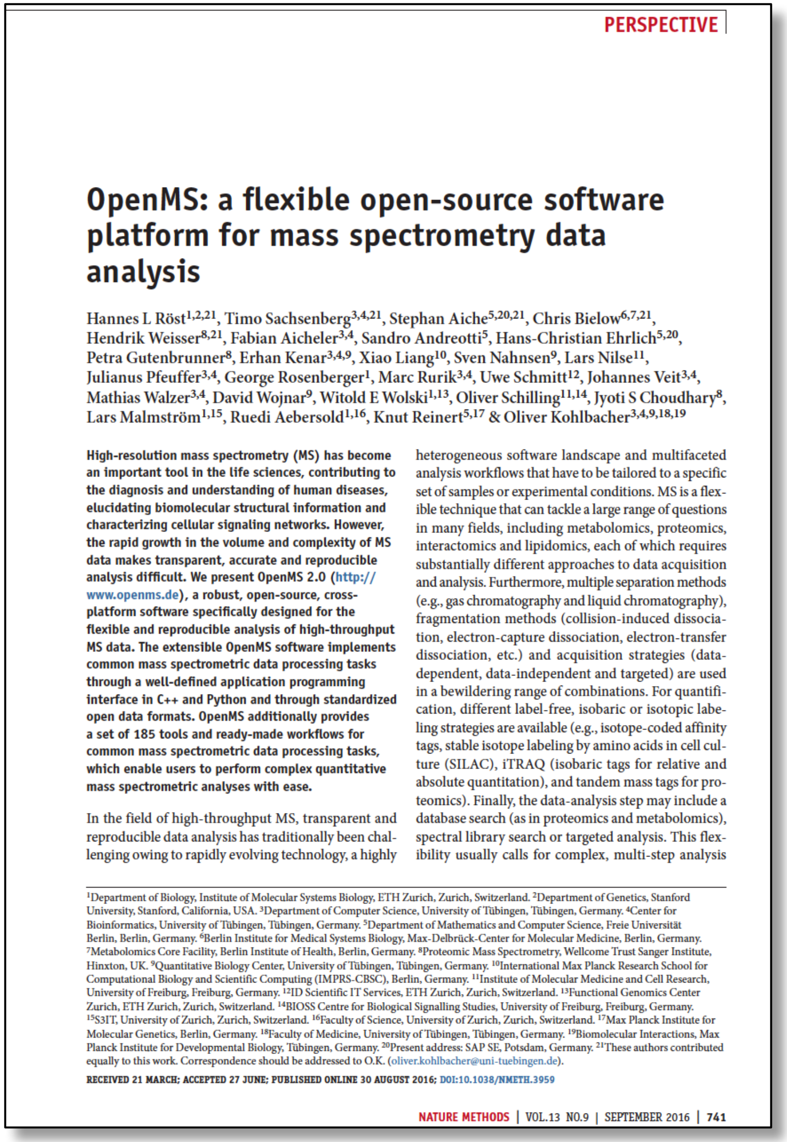
OpenMS

- **OpenMS** – an open-source C++ framework for computational mass spectrometry
- Jointly developed at ETH Zürich, FU Berlin, University of Tübingen
- **Open source**: BSD 3-clause license
- **Portable**: available on Windows, OSX, Linux
- **Vendor-independent**: supports all standard formats and vendor-formats through proteowizard
- **TOPP – The OpenMS Proteomics Pipeline**
  - Building blocks: One application for each analysis step
  - All applications share **identical user interfaces**
  - Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats
- **Tools** can be integrated in various **workflow systems**
  - TOPPAS – TOPP Pipeline Assistant
  - Galaxy
  - WS-PGRADE/gUSE
  - Proteome Discoverer/Compound Discoverer
  - **KNIME**

Kohlbacher et al., Bioinformatics (2007), 23:e191

OpenMS 2.x - Features

- **Currently 185 distinct tools**
- **Utilities** – extract information from files, file conversion, visualization
- **PTX identification** – interface to DB search engines, de novo search, RNA-protein XL MS, protein inference, RT prediction, proteotypicity prediction
- **PTX quantification – label-free,** TMT, iTRAQ, SILAC, MRM, OpenSWATH (DIA), ProteinSIP (metaproteomics), RT alignment
- **MTX quantification – nontargeted metabolomics**, MRM
- **MTX identification** – accurate mass DB search, spectral matching, composition
- **Miscellaneous** – MRM scheduling, LC-MS simulator, …



PERSPECTIVE

OpenMS: a flexible open-source software platform for mass spectrometry data analysis

Hannes L Röst[1,2,21], Timo Sachsenberg[3,4,21], Stephan Aiche[5,20,21], Chris Bielow[6,7,21], Hendrik Weisser[8,21], Fabian Aicheler[3,4], Sandro Andreotti[5], Hans-Christian Ehrlich[5,20], Petra Gutenbrunner[8], Erhan Kenar[3,4,9], Xiao Liang[10], Sven Nahnsen[9], Lars Nilse[11], Julianus Pfeuffer[3,4], George Rosenberger[1], Marc Rurik[3,4], Uwe Schmitt[12], Johannes Veit[3,4], Mathias Walzer[3,4], David Wojnar[9], Witold E Wolski[1,13], Oliver Schilling[11,14], Jyoti S Choudhary[8], Lars Malmström[1,15], Ruedi Aebersold[1,16], Knut Reinert[5,17] & Oliver Kohlbacher[3,4,9,18,19]

High-resolution mass spectrometry (MS) has become an important tool in the life sciences, contributing to the diagnosis and understanding of human diseases, elucidating biomolecular structural information and characterizing cellular signaling networks. However, the rapid growth in the volume and complexity of MS data makes transparent, accurate and reproducible analysis difficult. We present OpenMS 2.0 (http://www.openms.de), a robust, open-source, cross-platform software specifically designed for the flexible and reproducible analysis of high-throughput MS data. The extensible OpenMS software implements common mass spectrometric data processing tasks through a well-defined application programming interface in C++ and Python and through standardized open data formats. OpenMS additionally provides a set of 185 tools and ready-made workflows for common mass spectrometric data processing tasks, which enable users to perform complex quantitative mass spectrometric analyses with ease.

In the field of high-throughput MS, transparent and reproducible data analysis has traditionally been challenging owing to rapidly evolving technology, a highly heterogeneous software landscape and multifaceted analysis workflows that have to be tailored to a specific set of samples or experimental conditions. MS is a flexible technique that can tackle a large range of questions in many fields, including metabolomics, proteomics, interactomics and lipidomics, each of which requires substantially different approaches to data acquisition and analysis. Furthermore, multiple separation methods (e.g., gas chromatography and liquid chromatography), fragmentation methods (collision-induced dissociation, electron-capture dissociation, electron-transfer dissociation, etc.) and acquisition strategies (data-dependent, data-independent and targeted) are used in a bewildering range of combinations. For quantification, different label-free, isobaric or isotopic labeling strategies are available (e.g., isotope-coded affinity tags, stable isotope labeling by amino acids in cell culture (SILAC), iTRAQ (isobaric tags for relative and absolute quantitation), and tandem mass tags for proteomics). Finally, the data-analysis step may include a database search (as in proteomics and metabolomics), spectral library search or targeted analysis. This flexibility usually calls for complex, multi-step analysis

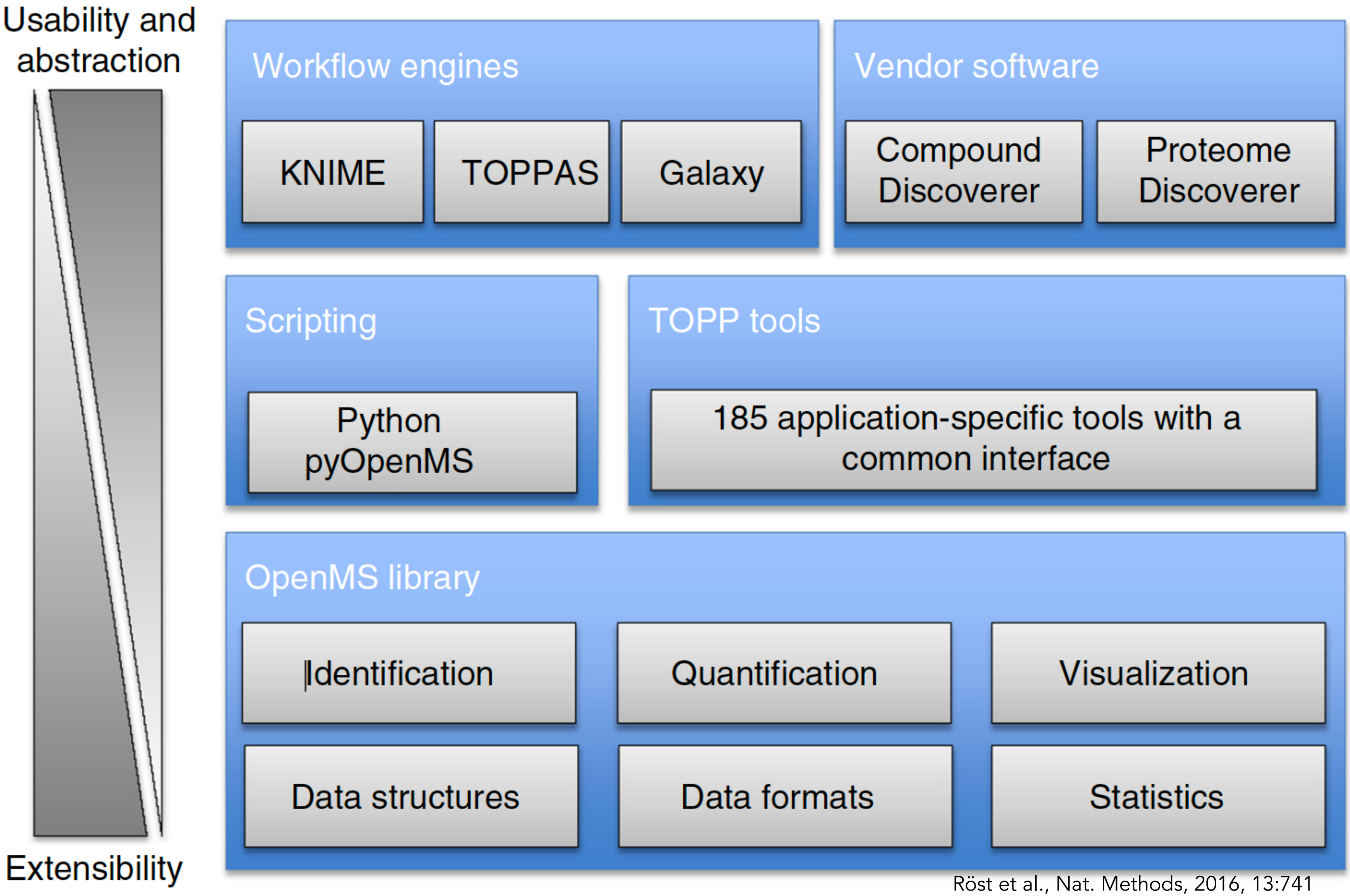
[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [2]Department of Genetics, Stanford University, Stanford, California, USA. [3]Department of Computer Science, University of Tübingen, Tübingen, Germany. [4]Center for Bioinformatics, University of Tübingen, Tübingen, Germany. [5]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [6]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany. [7]Metabolomics Core Facility, Berlin Institute of Health, Berlin, Germany. [8]Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Hinxton, UK. [9]Quantitative Biology Center, University of Tübingen, Tübingen, Germany. [10]International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC), Berlin, Germany. [11]Institute of Molecular Medicine and Cell Research, University of Freiburg, Freiburg, Germany. [12]ID Scientific IT Services, ETH Zurich, Zurich, Switzerland. [13]Functional Genomics Center Zurich, ETH Zurich, Zurich, Switzerland. [14]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [15]S3IT, University of Zurich, Zurich, Switzerland. [16]Faculty of Science, University of Zurich, Zurich, Switzerland. [17]Max Planck Institute for Molecular Genetics, Berlin, Germany. [18]Faculty of Medicine, University of Tübingen, Tübingen, Germany. [19]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen, Germany. [20]Present address: SAP SE, Potsdam, Germany. [21]These authors contributed equally to this work. Correspondence should be addressed to O.K. (oliver.kohlbacher@uni-tuebingen.de).

RECEIVED 21 MARCH; ACCEPTED 27 JUNE; PUBLISHED ONLINE 30 AUGUST 2016; DOI:10.1038/NMETH.3959

NATURE METHODS | VOL.13 NO.9 | SEPTEMBER 2016 | 741

Röst et al., Nat. Methods, 2016, 13:741

OpenMS - Architecture



Usability and abstraction

Extensibility

Workflow engines

KNIME

TOPPAS

Galaxy

Vendor software

Compound Discoverer

Proteome Discoverer

Scripting

Python pyOpenMS

TOPP tools

185 application-specific tools with a common interface

OpenMS library

Identification

Quantification

Visualization

Data structures

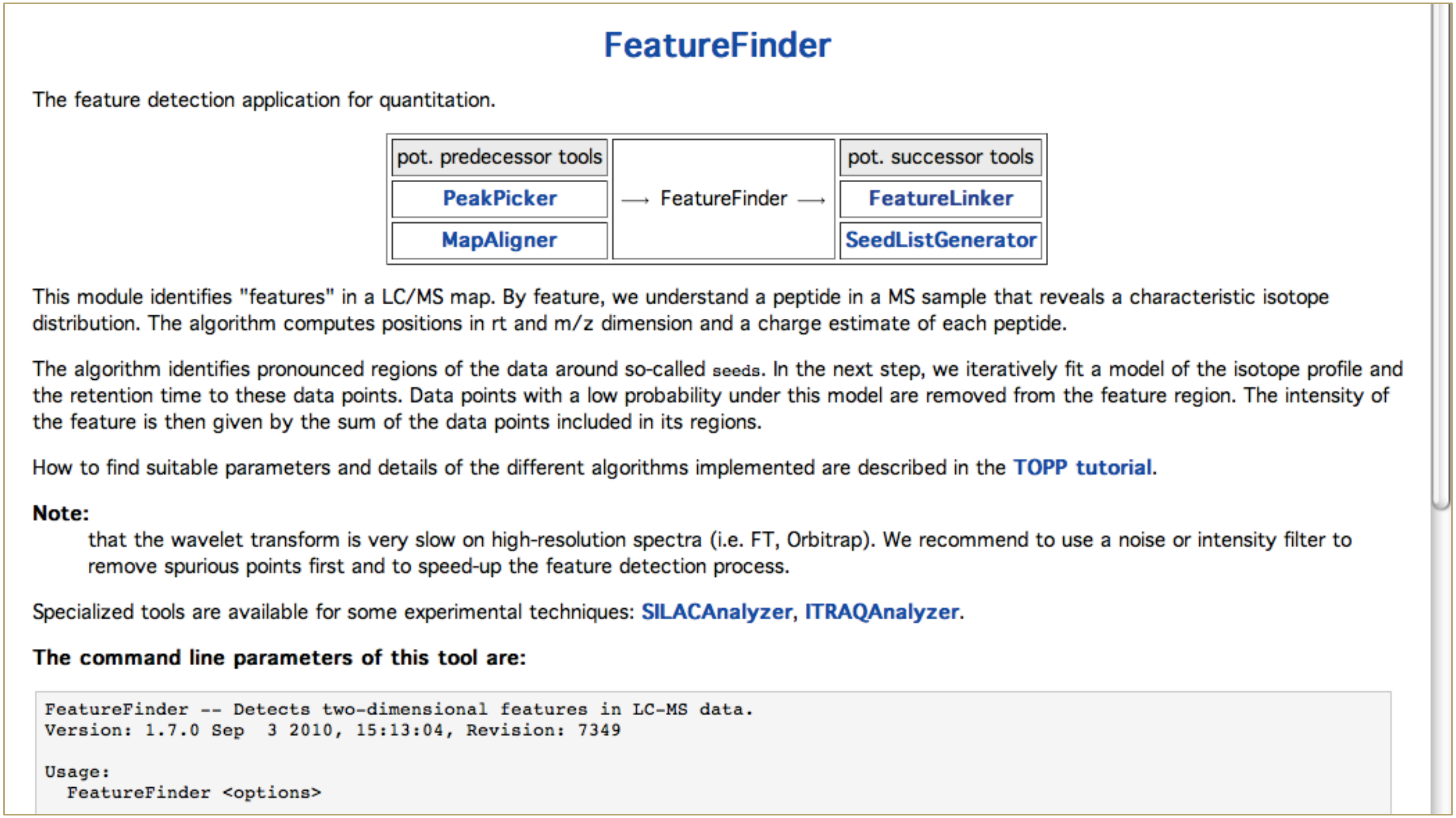
Data formats

Statistics

Röst et al., Nat. Methods, 2016, 13:741

Tool Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.OpenMS.org)



FeatureFinder

The feature detection application for quantitation.

pot. predecessor tools		pot. successor tools
PeakPicker	→ FeatureFinder →	FeatureLinker
MapAligner		SeedListGenerator

This module identifies "features" in a LC/MS map. By feature, we understand a peptide in a MS sample that reveals a characteristic isotope distribution. The algorithm computes positions in rt and m/z dimension and a charge estimate of each peptide.

The algorithm identifies pronounced regions of the data around so-called `seeds`. In the next step, we iteratively fit a model of the isotope profile and the retention time to these data points. Data points with a low probability under this model are removed from the feature region. The intensity of the feature is then given by the sum of the data points included in its regions.

How to find suitable parameters and details of the different algorithms implemented are described in the **TOPP tutorial**.

Note:

> that the wavelet transform is very slow on high-resolution spectra (i.e. FT, Orbitrap). We recommend to use a noise or intensity filter to remove spurious points first and to speed-up the feature detection process.

Specialized tools are available for some experimental techniques: **SILACAnalyzer**, **ITRAQAnalyzer**.

The command line parameters of this tool are:

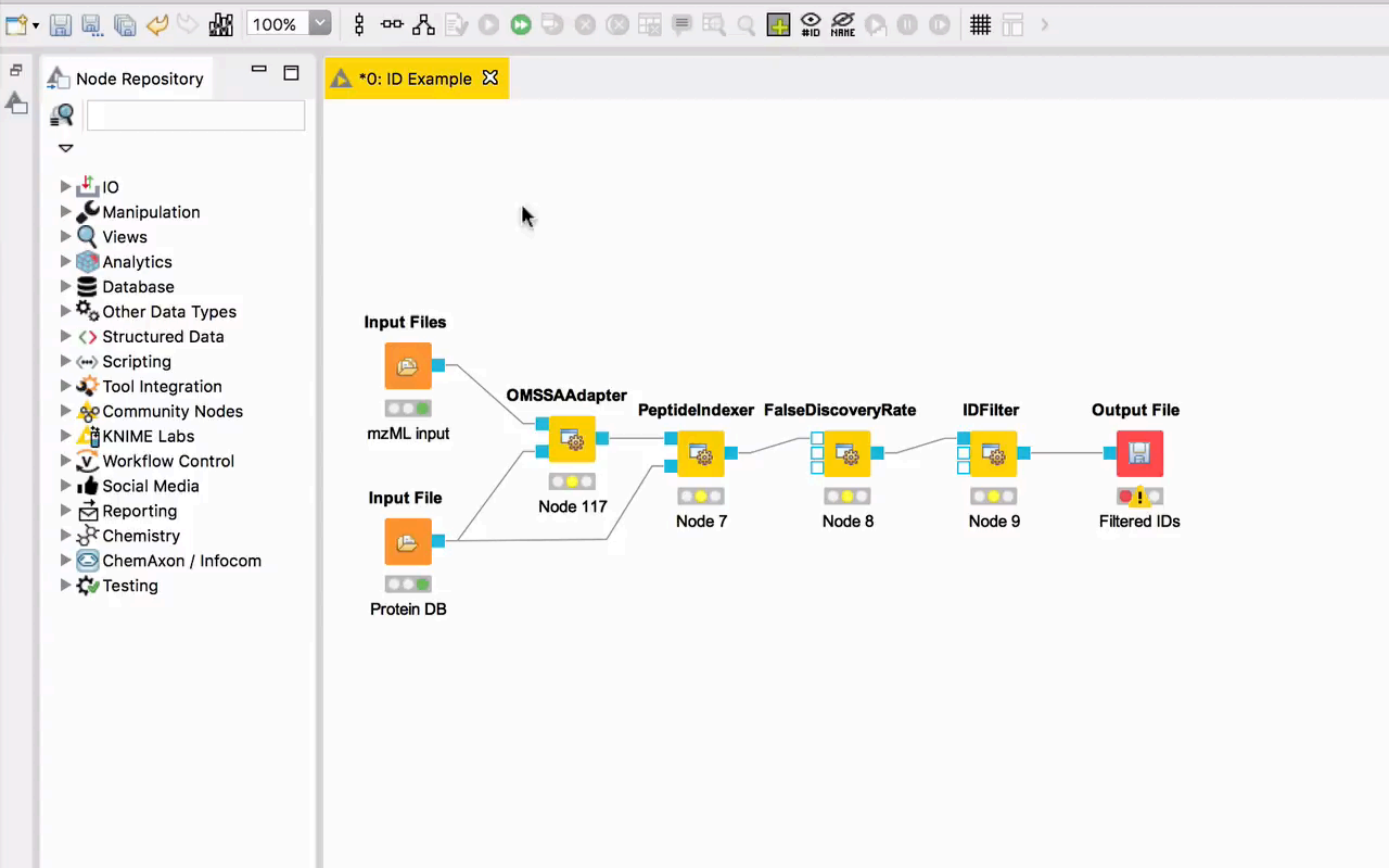
```
FeatureFinder -- Detects two-dimensional features in LC-MS data.
Version: 1.7.0 Sep  3 2010, 15:13:04, Revision: 7349

Usage:
  FeatureFinder <options>
```

Tool Implementation

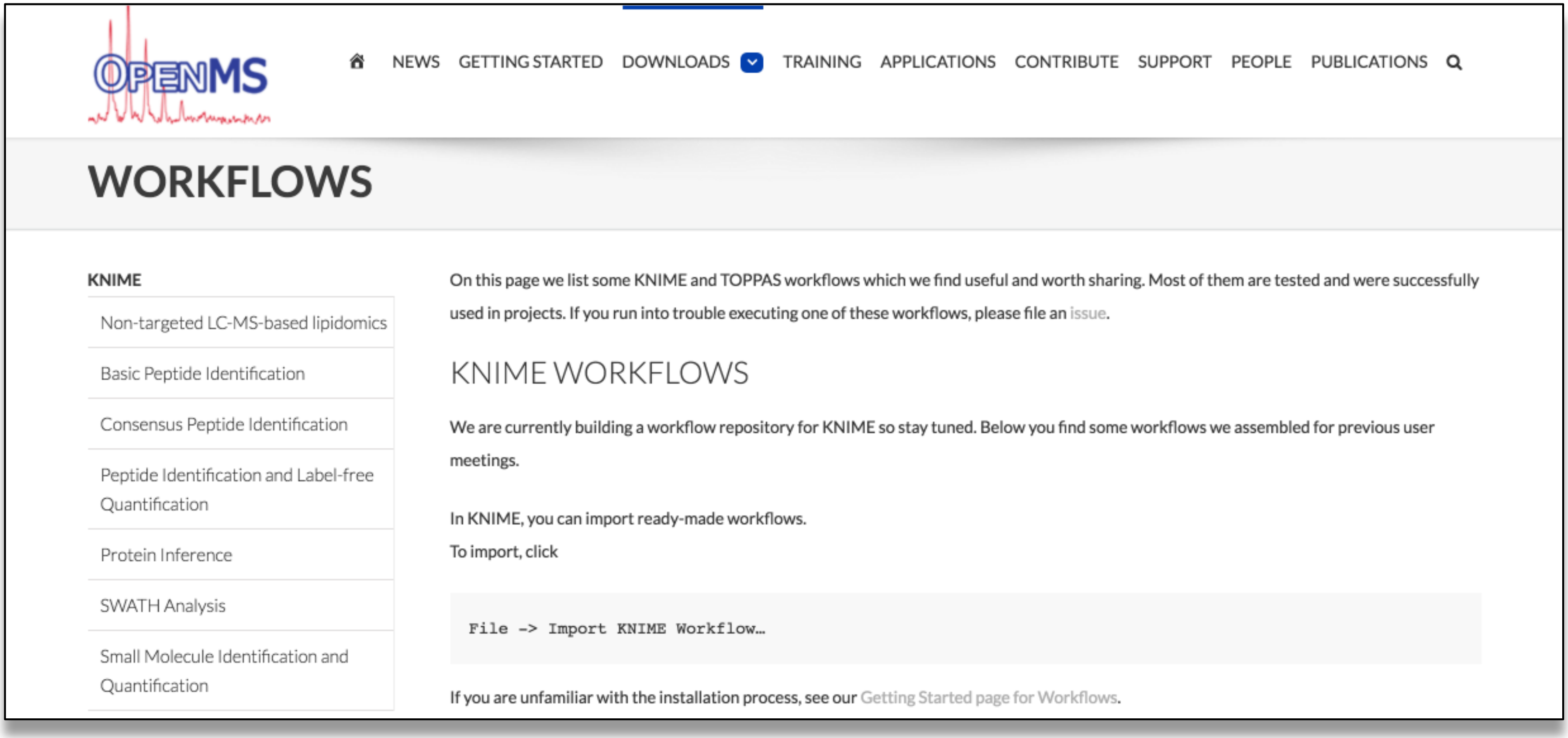
- Very easy to implement thanks to the OpenMS framework
- Usually short (200 lines of code on average, mostly concerned with parameter handling)
- Use of the OpenMS core library

```
IDMapper.C:
  [...]
  vector<ProteinIdentification> protein_ids;
  vector<PeptideIdentification> peptide_ids;
  String document_id;
  IdXMLFile().load(getStringOption_
     ("id"), protein_ids,peptide_ids, document_id);
  IDMapper mapper;
  [...]
  ConsensusXMLFile file;
  ConsensusMap map;
  file.load(in, map);
  mapper.annotate(map, peptide_ids, protein_ids, false);
  file.store(out, map);
```

Workflows as an Abstraction



Workflow Repositories


OPENMS
NEWS
GETTING STARTED
DOWNLOADS
TRAINING
APPLICATIONS
CONTRIBUTE
SUPPORT
PEOPLE
PUBLICATIONS
KNIME
Non-targeted LC-MS-based lipidomics
Basic Peptide Identification
Consensus Peptide Identification
Peptide Identification and Label-free Quantification
Protein Inference
SWATH Analysis
Small Molecule Identification and Quantification

WORKFLOWS

On this page we list some KNIME and TOPPAS workflows which we find useful and worth sharing. Most of them are tested and were successfully used in projects. If you run into trouble executing one of these workflows, please file an issue.

KNIME WORKFLOWS

We are currently building a workflow repository for KNIME so stay tuned. Below you find some workflows we assembled for previous user meetings.

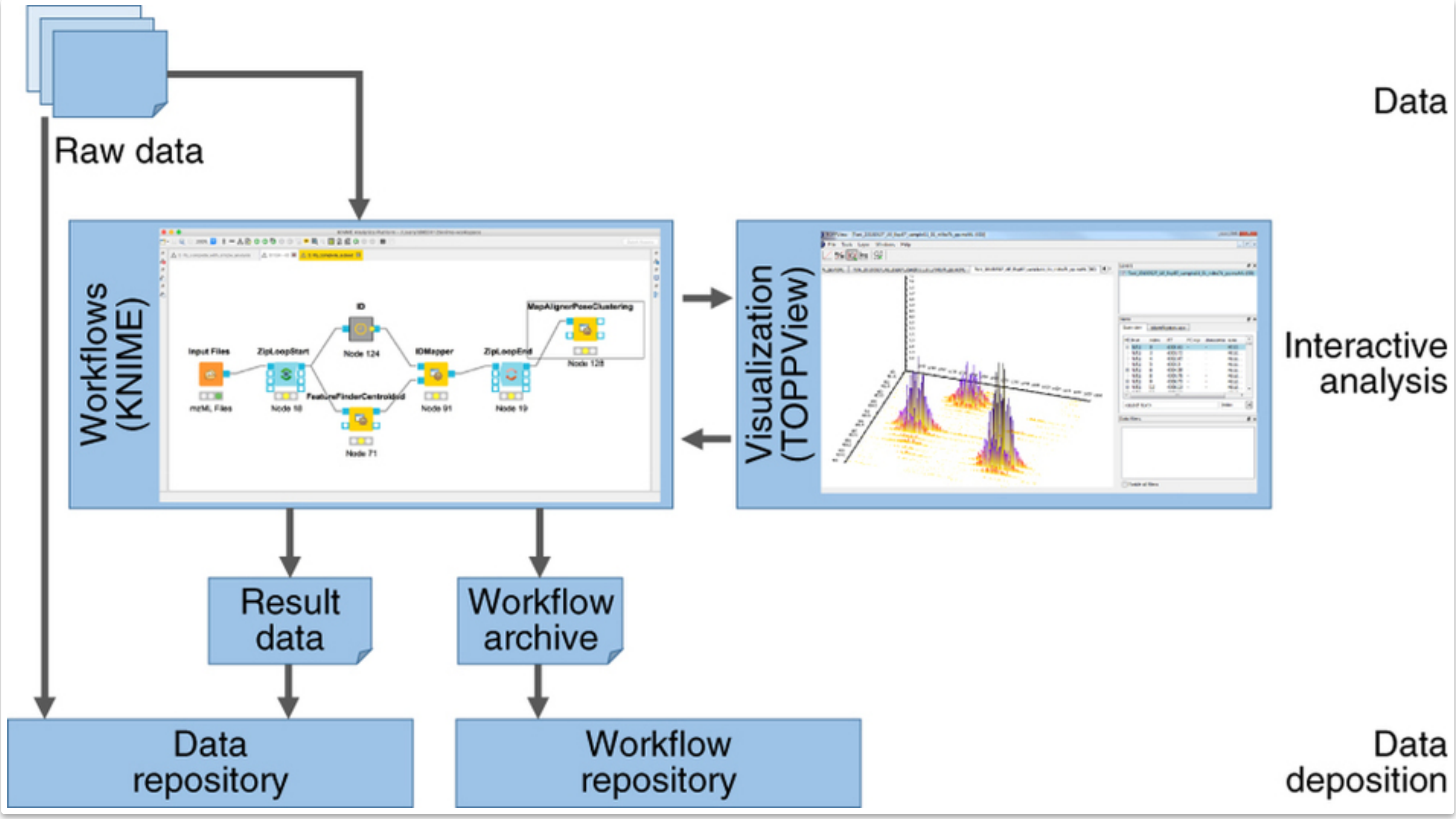
In KNIME, you can import ready-made workflows.
To import, click

```
File -> Import KNIME Workflow…
```

If you are unfamiliar with the installation process, see our Getting Started page for Workflows.

- OpenMS website contains a workflow repository with selected example workflows (**www.OpenMS.org**)
- General-purpose workflow repository: **www.myexperiment.org**
  - Collects workflows from arbitrary workflow engines
  - Numerous applications, can be used to document data analysis

Open (Data | Source | Science)



Raw data

Workflows (KNIME)

Visualization (TOPPView)

Result data

Workflow archive

Data repository

Workflow repository

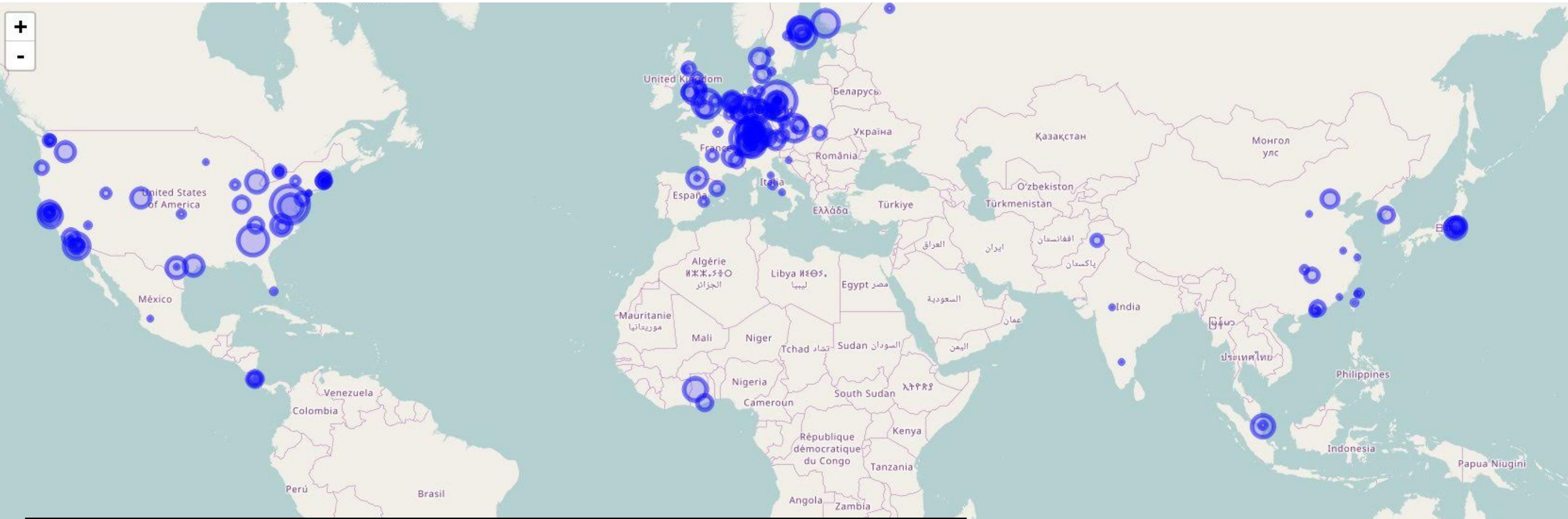
Data

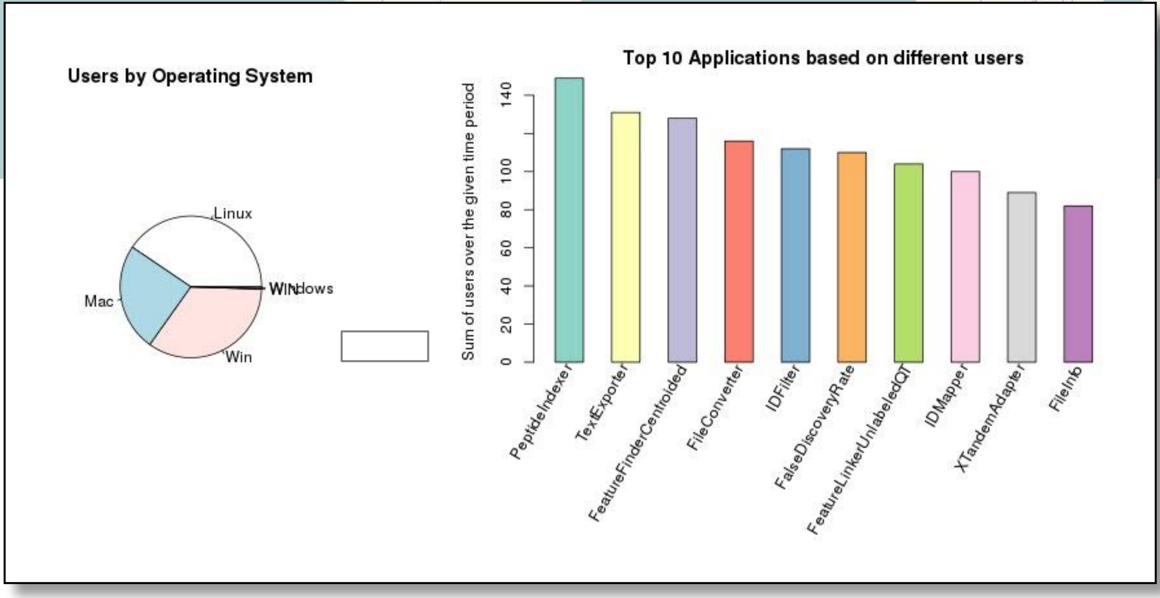
Interactive analysis

Data deposition

Röst et al., Nat. Methods, 2016, 13:741

Who uses OpenMS? And what parts of it?





Getting Involved

- **Contributing algorithms**
  - Got interesting algorithms you developed?
  - Anything missing in the library/tools?
- **Contributing interfaces**
  - Got a tool we have no other solution for?
  - Let’s discuss interfaces and levels of integration
- **Contributing scripts**
  - Solved an interesting problem once?
  - Need help hacking a quick prototype for something?
- **Contributing workflows**
  - Put something together that solves a particular problem?

Get in touch with us – ideally via the mailing list or the talk to me during the conference!

Materials

- **OpenMS Website**: http://www.OpenMS.org
  - Documentation
  - Tutorials
  - Online lecture ‘Computational Proteomics and Metabolomics’ (Kohlbacher, Reinert, Nahnsen) http://bit.ly/2d2kBSq
  - Downloads
    - Binaries
    - Source code
    - Plugins for Proteome Discoverer
  - Access to mailing lists – this is where you can get help 24/7

OpenMS.de