Pivoting Towards Users: BioPAX, Paxtools, and Pathway Commons with PaxtoolsR

COMBINE

8/22/2014

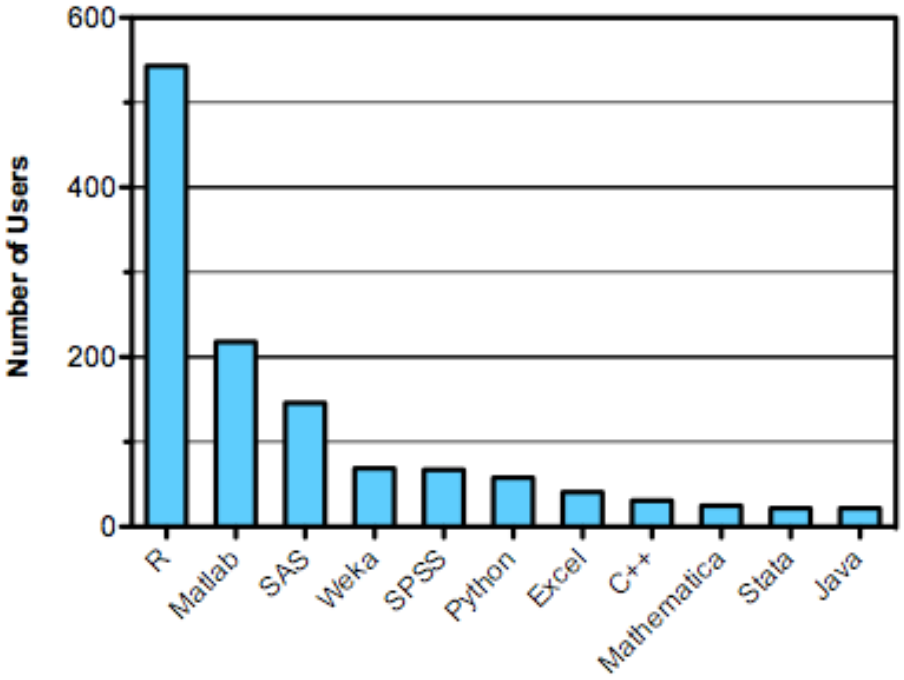
Augustin Luna

lunaa@cbio.mskcc.org

cBio@MSKCC

Where are users?

- Are COMBINE standards talking the same language as users?
- COMBINE software developments tend to be in Java and C++
- There are large and growing R and Python communities



Languages used in Kaggle data analysis competitions

R Language

- Free, open source
- Started in 1993
- Geared towards scientific computing
  - Created by Ross Ihaka and Robert Gentleman (statisticians)
- Functional; influenced by Scheme
- Interpreted; similar to MATLAB

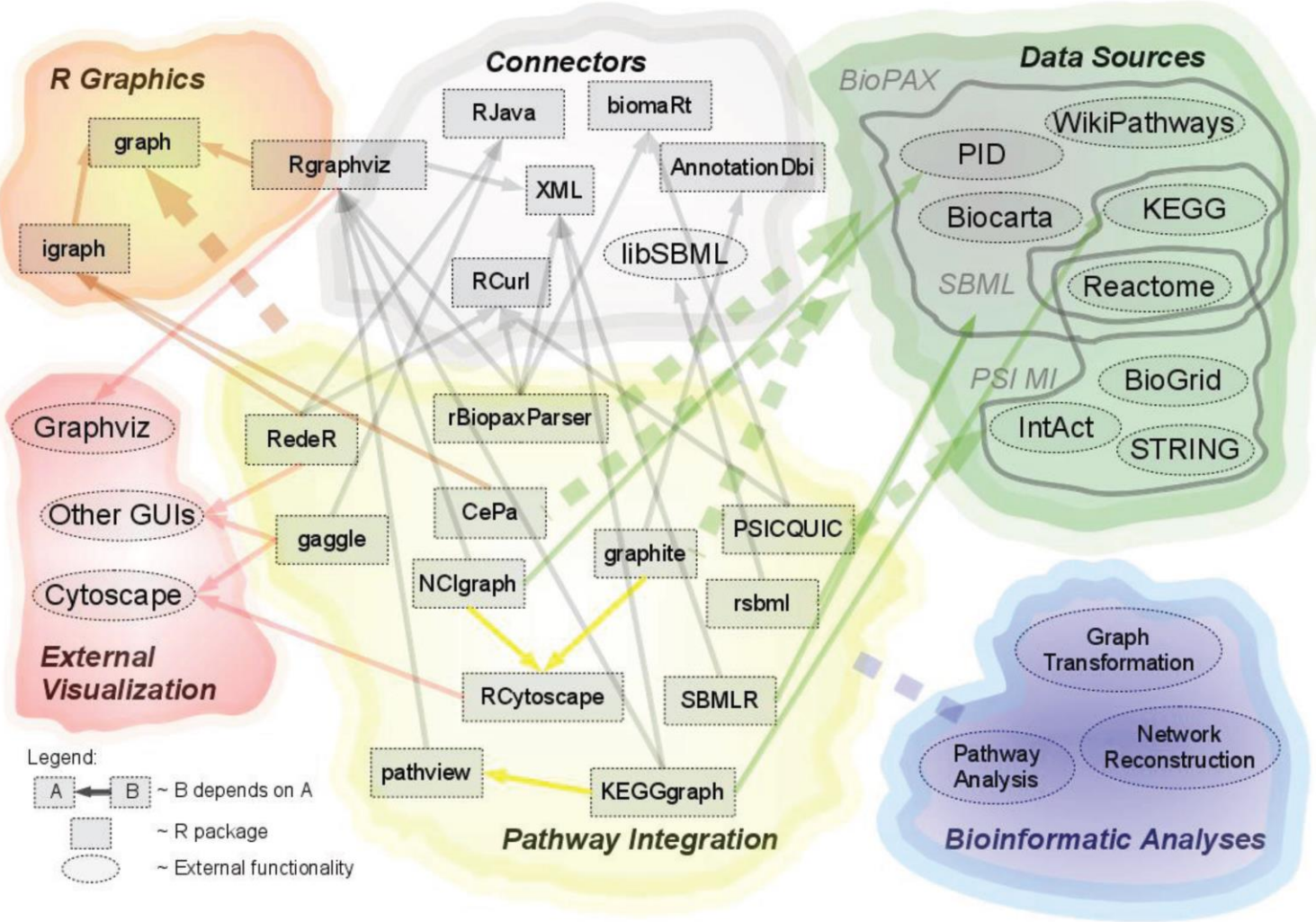
Why is R Popular?

- Free, open source
- Easy for users to try out ideas
- Interactive data analysis
  - Script-driven rather than menu-driven helps reproducibility
- Code is less verbose than Java, C++, etc.
- Flexible and powerful plotting support
- Excellent package management system
  - Large and growing collection of statistical analysis methods
  - Simple package installation; dependency management
  - R scripts usually portable to other platforms
  - Package repositories ensure functionality, documentation, and interoperability
  - Vignettes (tutorials) provided as runnable analyses

Extending R and Package Repositories

- Comprehensive R Archive Network (CRAN)
  - 5,800 R packages (as of June 2014)
  - Many packages call C, C++, Fortran, or Java code for speedups
- Bioconductor
  - R packages focused on bioinformatics
  - Currently, 824 R packages
  - 56 packages dedicated to pathway analysis
- Devtools
  - R package that allows package installation from code repositories

Pathway Data and Analysis in R



From review by Kramer et al (2014)

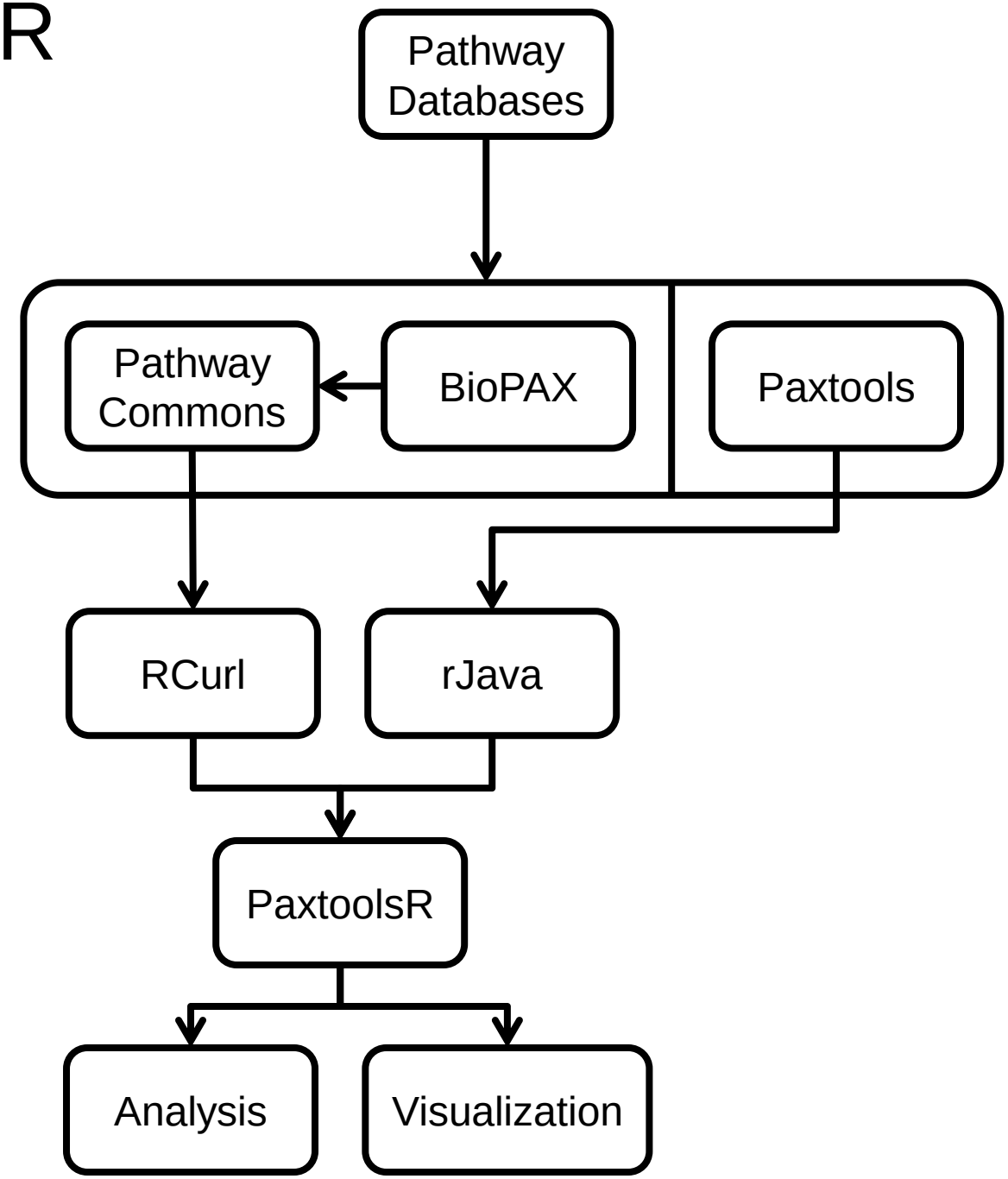
Data Sources of Pathway Data in R

- Pathguide lists 500+ pathway and molecular interaction resources
  - A large amount of pathway data is not accessible easily in R
- Many well-known databases missing
  - HPRD, PantherDB, etc.
- Many biomolecules underrepresented
  - Drugs, transcription factors, metabolites, miRNA, etc.

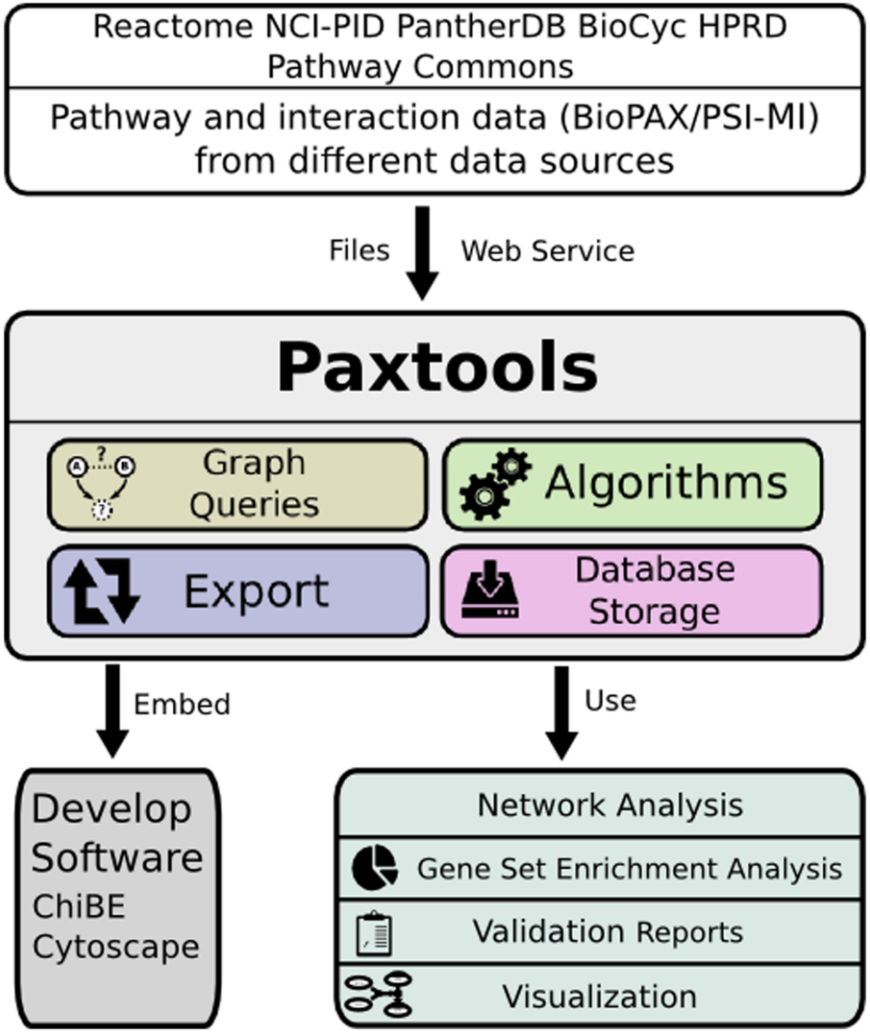
Package	Datasource(s)
rBiopaxParser	BioPAX API (mainly L2)
Graphite	KEGG, BioCarta, PID, Reactome, SPIKE
NCIGraph	PID
Pathview	KEGG
KEGGgraph	KEGG
rsbml	SBML API
PSICQUIC	PSIMI

From review by Kramer et al (2014)

PaxtoolsR

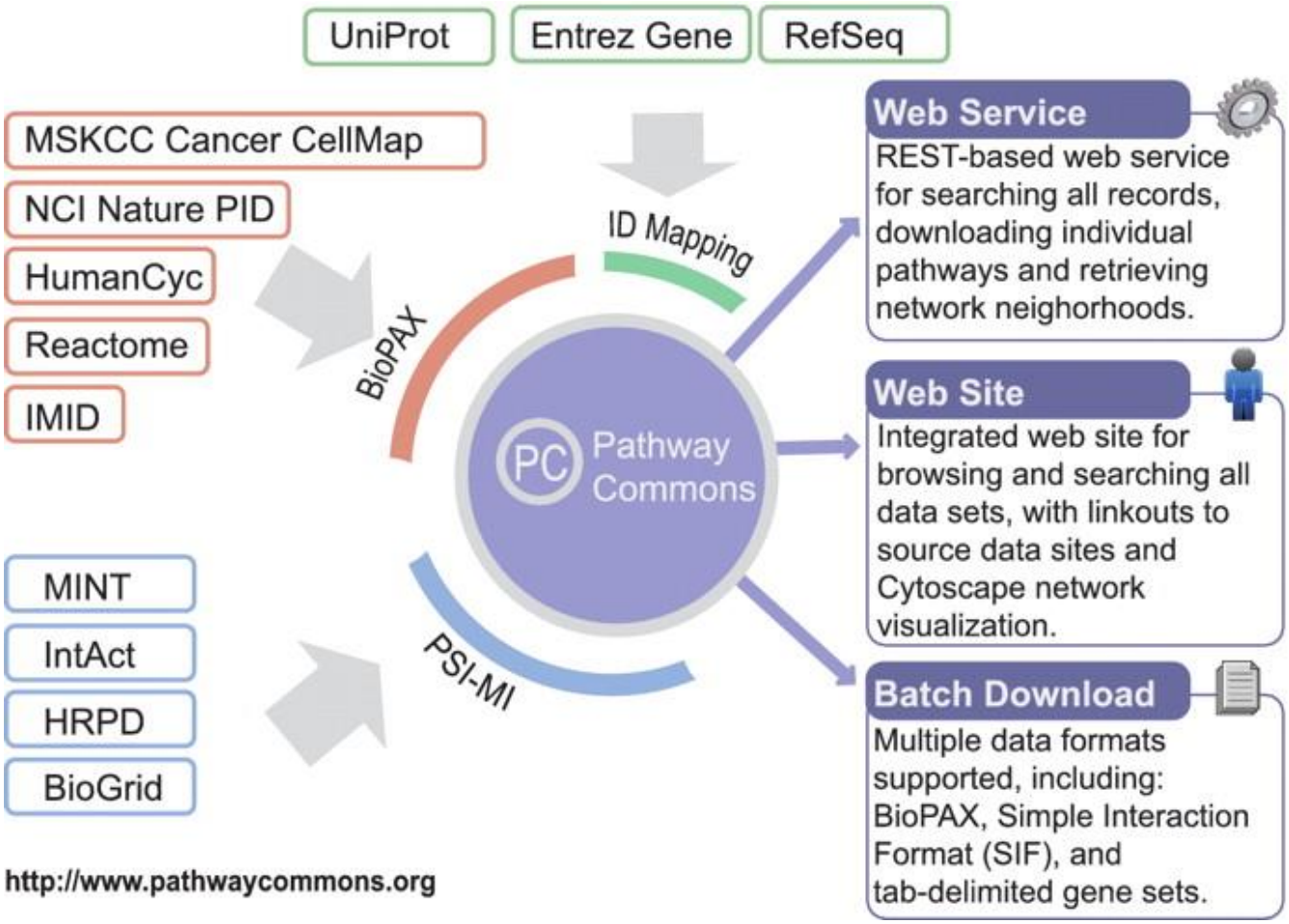

Pathway Databases
Pathway Commons
BioPAX
Paxtools
RCurl
rJava
PaxtoolsR
Analysis
Visualization

Paxtools



Demir et al (2013)

Pathway Commons



Cerami et al (2011)

PaxtoolsR

- Features
  - Validate, merge multiple subnetworks, extract subnetworks from local BioPAX files and PC
  - Query and extract subnetworks from aggregated PC data
  - Works primarily with expanded binary interactions; simplified from BioPAX representation
- Providing a tutorial with example workflows
  - Visualization; data overlay
  - GSEA

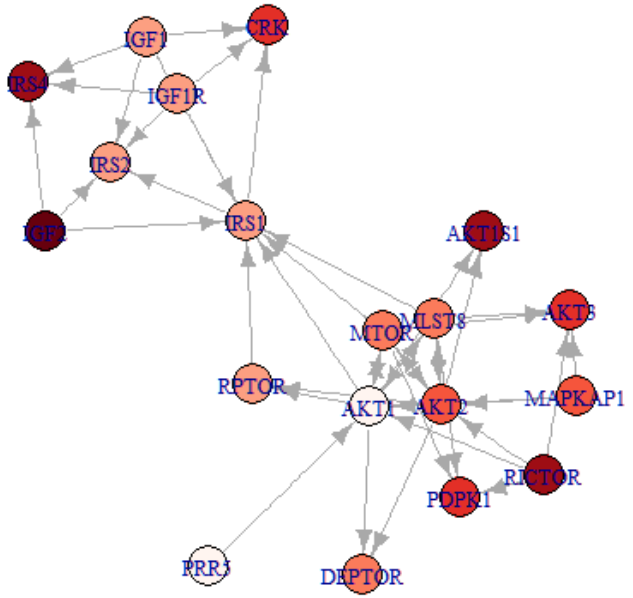


Image from paxtoolsR tutorial; a network queried from PC, converted to binary network using Paxtools, and overlayed with example data

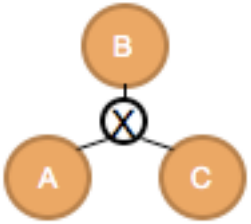
Luna A, et al. in preparation

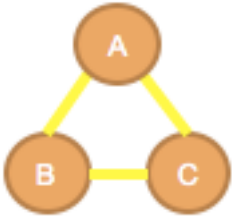
Dealing with the Complexity of BioPAX

- BioPAX contains 60+ classes and 90+ properties
- Simple Interaction Format (SIF)
  - An edgelist with interaction type
- Extracted using graph queries that detect biologically interesting interaction patterns in Pathway Commons data
  - Complexes, metabolic, modification, control interactions
  - Generates binary interactions and integrates them across databases

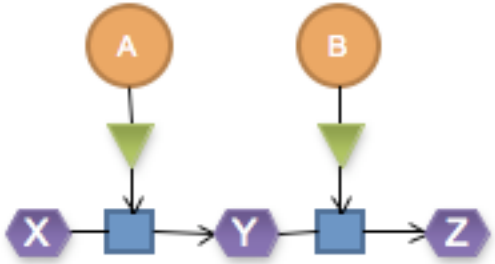
Examples BioPAX to SIF Conversions

in-complex-with

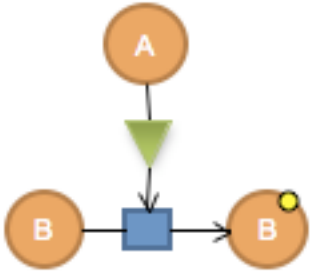


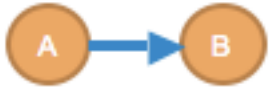


catalysis precedes



controls-state-change-of





14 Interaction Types Total

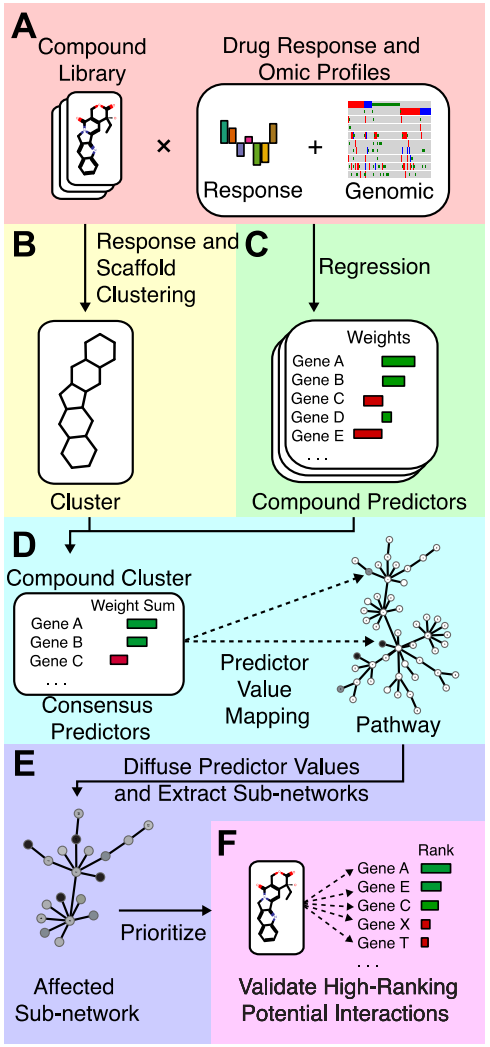
Babur et al (in preparation)

Bringing BioPAX and PC to Analysis Packages in R

- Levels of functionality
  1. Basic data requests
  2. Data traversals and graph queries
  3. Interaction reduction methodologies for specific datatypes
  4. More advanced analysis algorithms that make use of experimental data
     - NetBox: Find modules in networks of altered (mutations/copy number alterations) genes
     - SPIA (Signaling Pathway Impact Analysis): Find relevant pathways given differentially expressed genes

NCI Drug Set Analysis

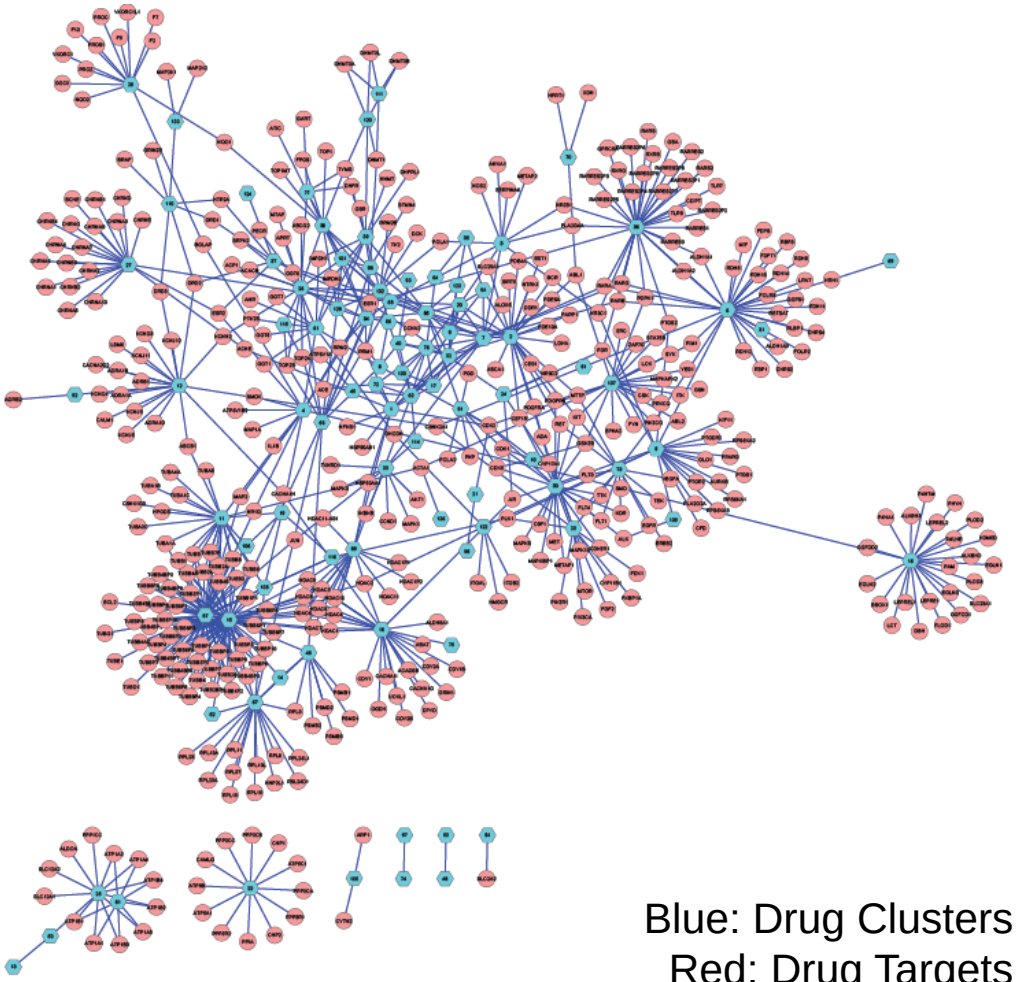
- Objectives:
  - Identify novel drug-target interactions using compound activity and genomic profiling data
  - Prioritize compounds for further development using understanding of pharmacology, target pathway context, relevance to cancer, and novelty
- Data:
  - ~42K drug compounds
  - 60 cell line panel developed as an anti-cancer drug efficacy screen by the NCI
  - Cell lines have been characterized to include copy number variations, mutations, gene expression, microRNA expression, and protein levels
- Key R packages:
  - glmnet (regression), rcdk (structure), and paxtoolsr (pathways)



Work with Vinodh Rajapakse

Addressing Knowledge Deficits in Pathway Commons

- Deficits in the PC data
  - Drug target, metabolic, transcription regulation are areas were we can bolster available data
  - Examples: Pharmgkb and CTD
- Example Usage:
  - Compounds clustered by activity
  - Many drug clusters possess drugs with known target interactions
  - Drugs in cluster may share target interactions

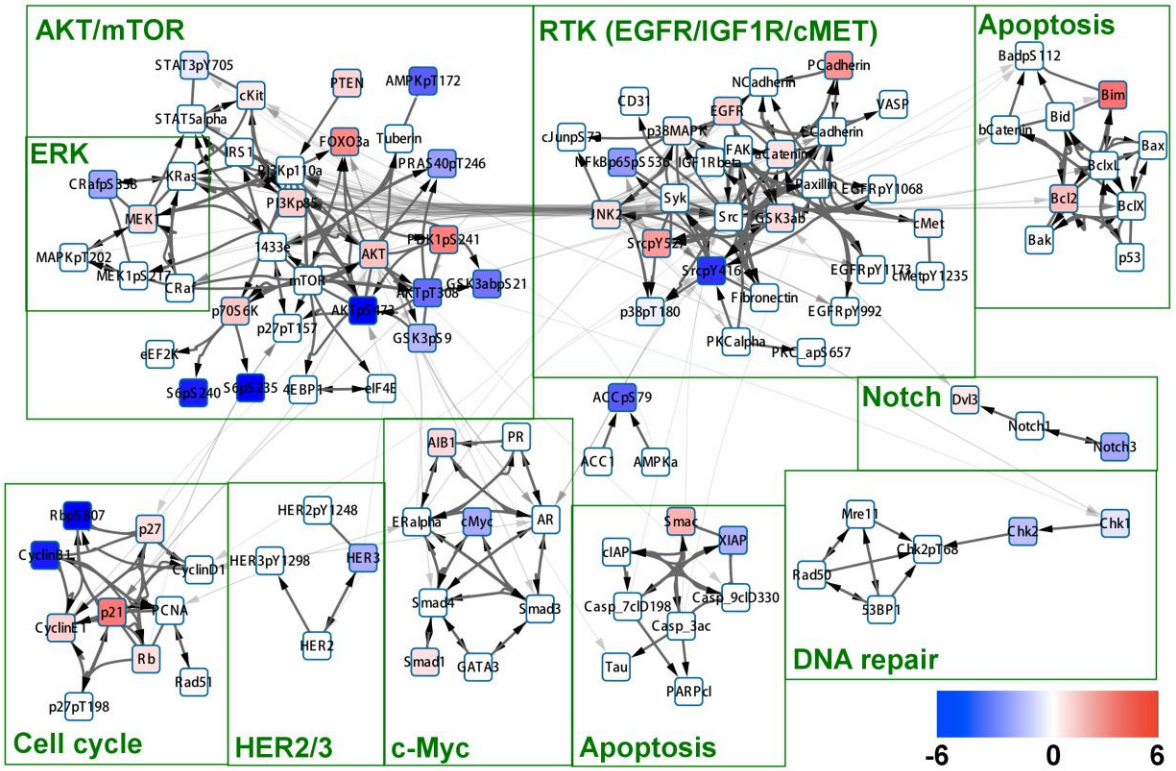


Blue: Drug Clusters
Red: Drug Targets

Drug-target interactions from PiHelper; Aksoy A et al. (2013)

Pathway Analysis in Post-Treatment Ovarian Cancer Cell Lines

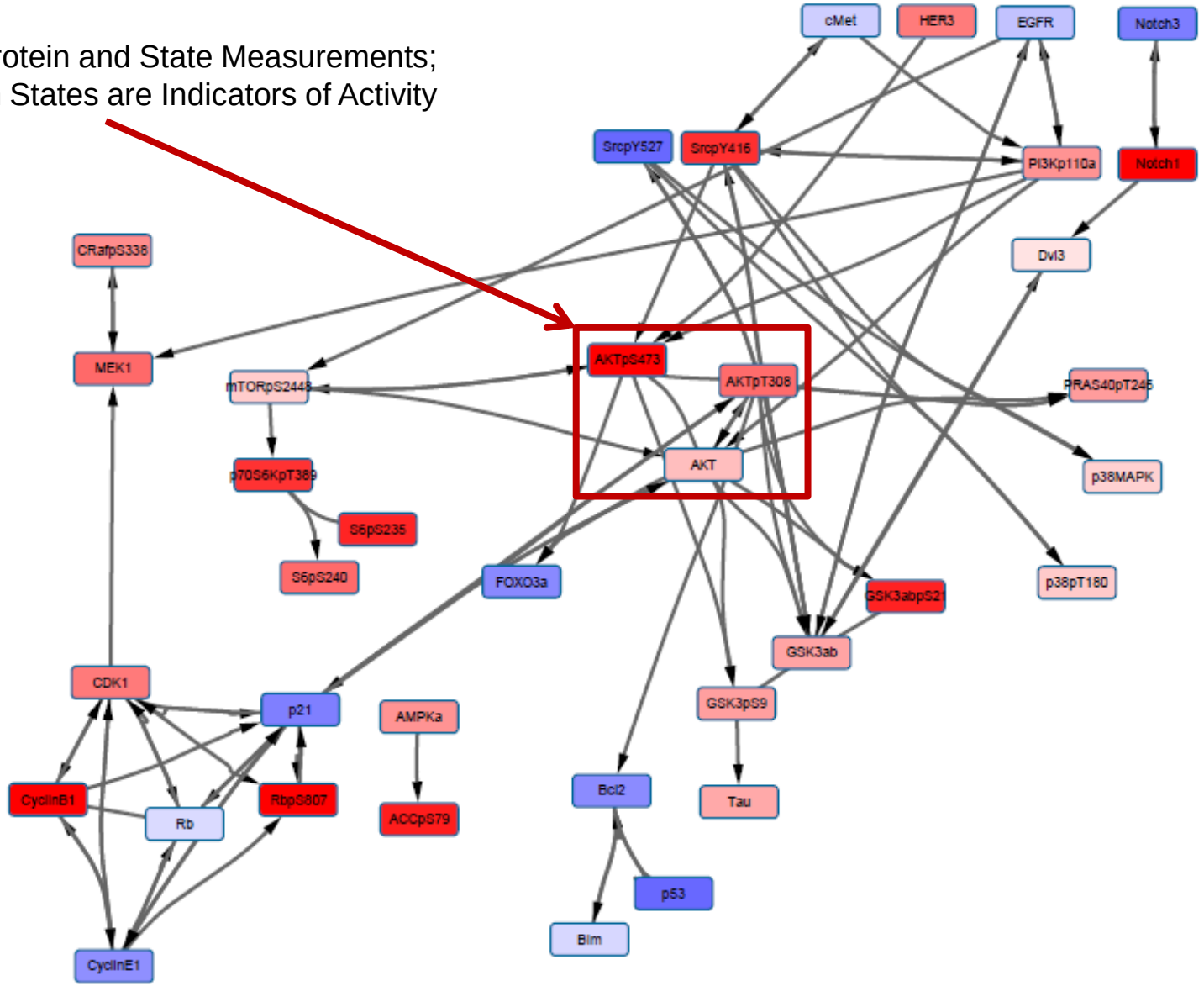
- Cell lines treated with MYC inhibiting drug
- Challenge: Antibodies target specific phosphorylation states
- One goal: Propose possible drug combinations



Protein production in sensitive cell line; Work with Anil Korkut

Modified Pathway Reduction

Whole Protein and State Measurements;
Phosphorylation States are Indicators of Activity



Tool BP-Prior; Arman Aksoy

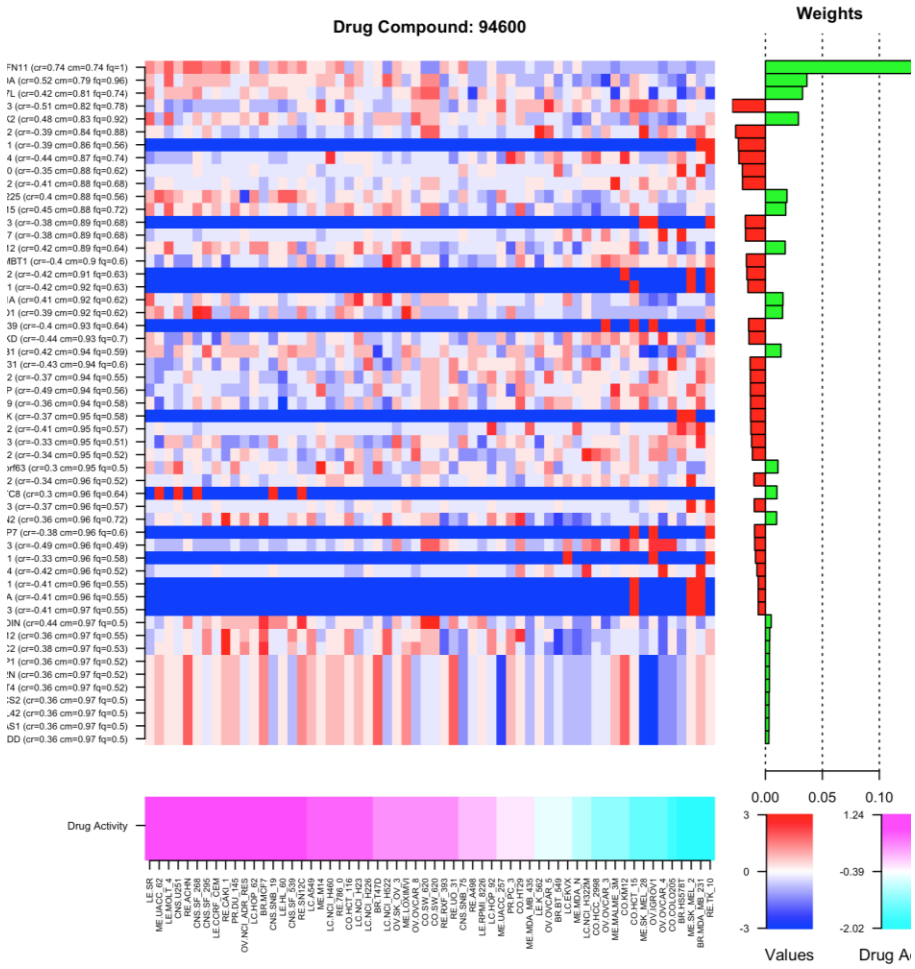
Summary/Future Work

- As COMBINE standards stabilize, they need to pivot to simplify usage/analysis by users
  - R includes a large community of bioinformaticians
  - Programming interface availability does not necessarily translate into utility; understanding of projects developed in other languages is important
- PaxtoolsR is a tool that makes local BioPAX datasets/Paxtools algorithms and those available through PC accessible in R
- Further work in PC/Paxtools
  - Extend data sources used by PC
  - Work on BioPAX reduction methods for various experimental data types
    - Metabolic, proteomic, epigenetic, etc.
- Further work is need in several areas for paxtoolsR
  - Integrate graph analyses methods
  - Extension of webservice commands to local data

Acknowledgements

- MSKCC
  - Arman Aksoy
  - Ozgun Babur
  - Emek Demir
  - Anil Korkut
  - Onur Sumer
  - Igor Rodchenkov (U of Toronto)
  - Gary Bader (U of Toronto)
  - Chris Sander
- NCI
  - Vinodh Rajapakse
  - Bill Reinhold
  - Yves Pommier
- COMBINE Organizers
- NHGRI Funding

Example Regression Analysis Results



Outline

- BioPAX infrastructure
- R introduction
- Pathway data in R
- PaxtoolsR overview
- Related research projects
- Summary