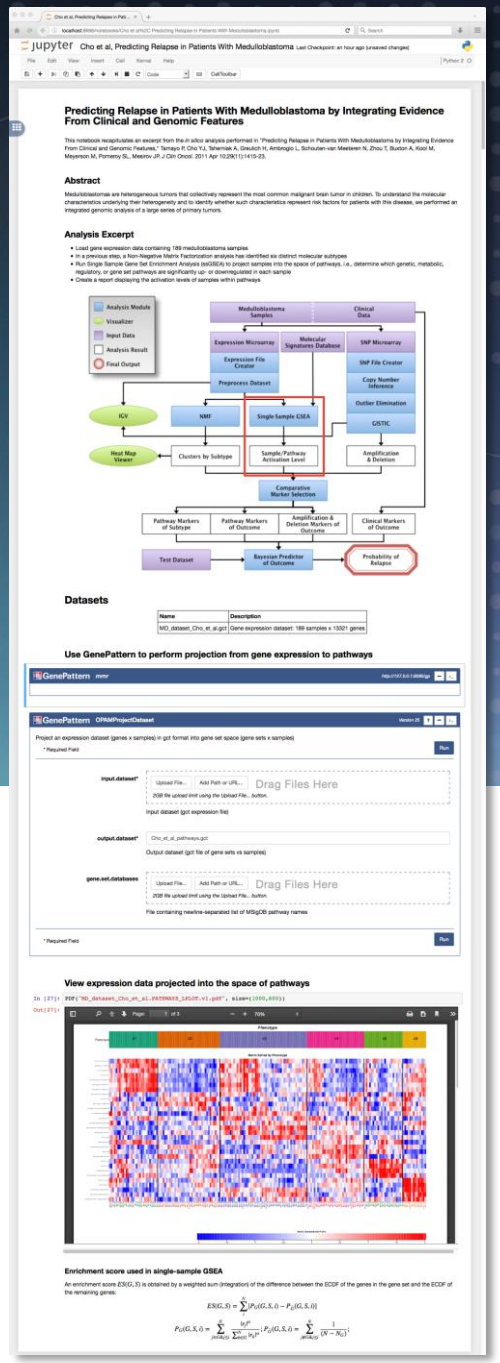
GenePattern Notebooks for Cancer Research



GenePattern

Comprehensive
Extensible
Reproducible
Accessible
Available

Jill P. Mesirov

ITCR PI Meeting
June 14, 2016

Outline

Informatics Technology for Cancer Research

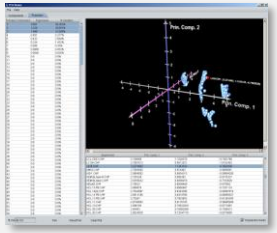
- GenePattern: overview and activity
- The GenePattern electronic notebook: overview and status
- GenePattern notebook video demo
- GenePattern notebook repository
- New GenePattern modules to address cancer data complexity.

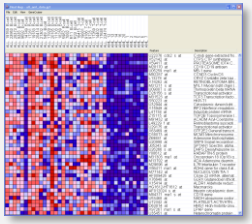
The GenePattern Ecosystem: Architecture

Module Repository

NMF | Cuffdiff | MuTect | ...

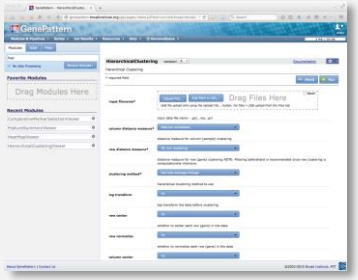
GISTIC | GSEA | Tophat | ...





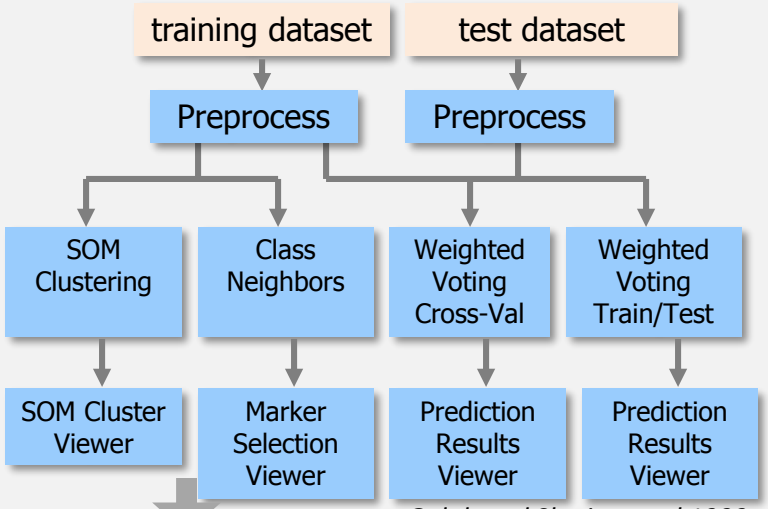
Hundreds of analysis and visualization tools

Clients

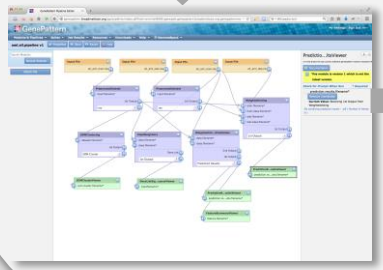


Web

Pipeline Environment



Golub and Slonim et. al 1999



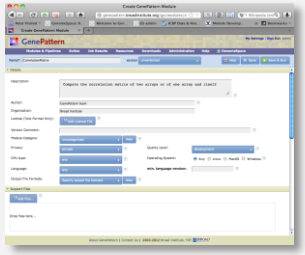
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

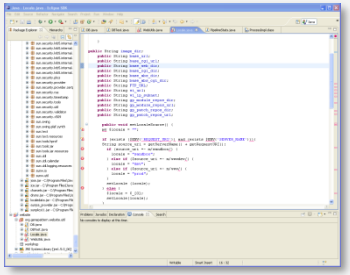
Analysis Engine

Record/replay analyses
Versioning of methods
Web service access

Module Integrator



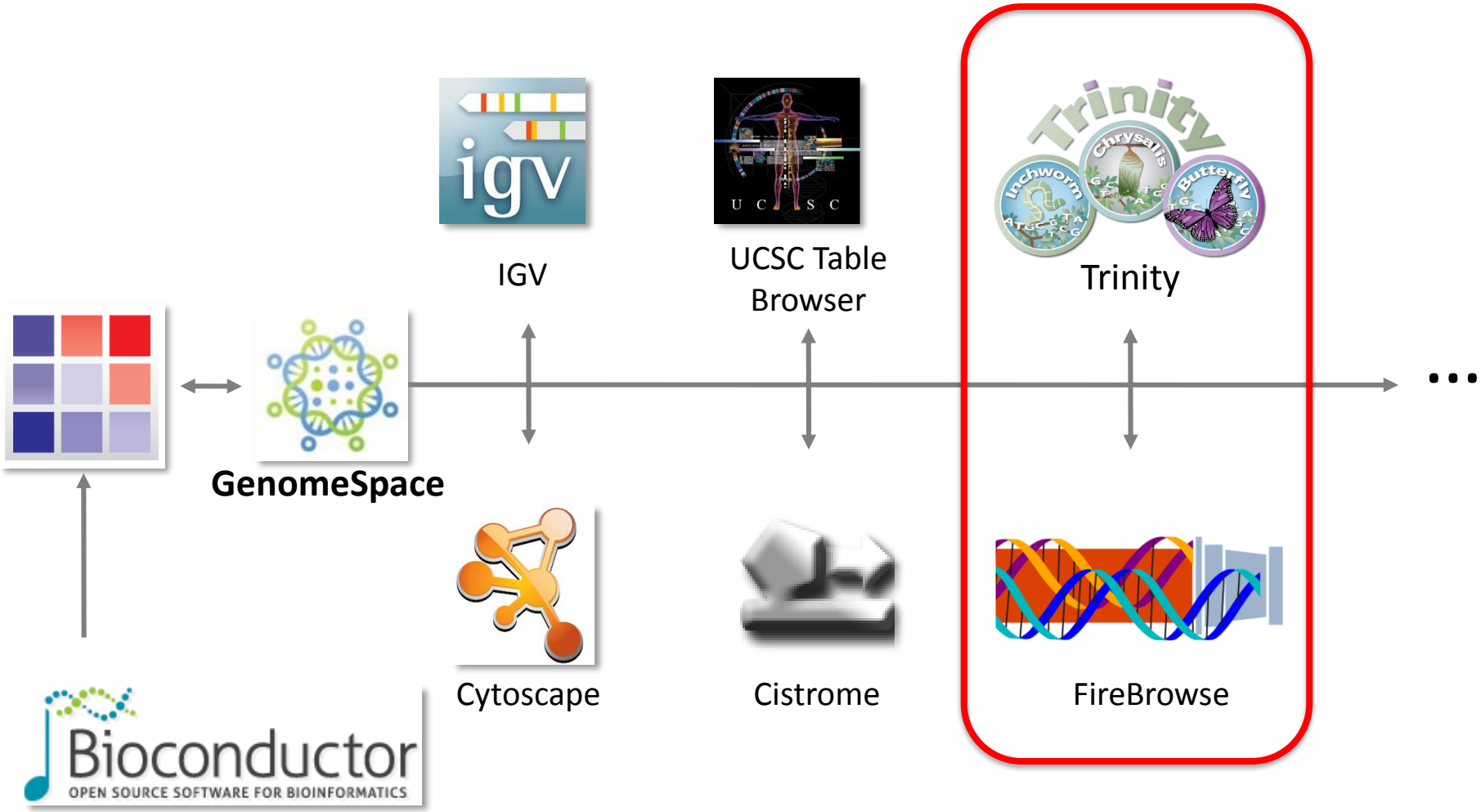
Easy addition of new tools



Programming

Access for all levels of user

Connectivity to Software and ITCR

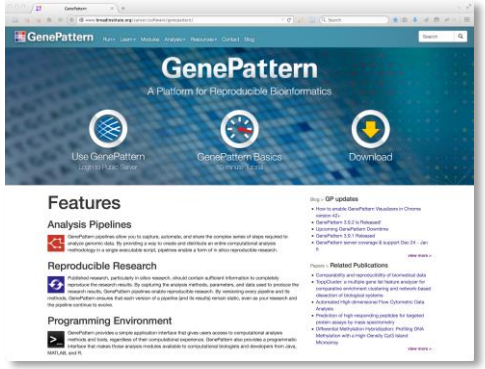


Availability

- Public server at Broad Institute (genepattern.broadinstitute.org)
- Public server at Indiana U, backed by Mason HPC cluster (gp.indiana.iu)
- Downloadable server (laptop to compute farm) (www.genepattern.org)
- Amazon Machine Image (AMI)
- *In development*
  - Gateway connecting XSEDE resources including Pittsburgh Supercomputing Center (PSC) and Texas Advanced Computing Center (TACC)

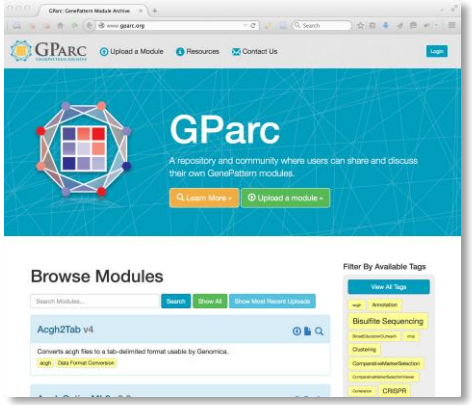
Activity and Outreach

www.genepattern.org



- Current version: 3.9.7 (4/2016)
- Open source, BSD-style license
- ~47,000 registered users
- Public server runs ~4,000 analyses/week
- GParc: GenePattern community repository
  - ~100 community-contributed methods
  - CRISPR analysis
  - Bisulfite sequencing
  - Flow cytometry
  - RNAi screens
- GenePattern workshops: 10/1/2015, 1/26/2016

www.gparc.org

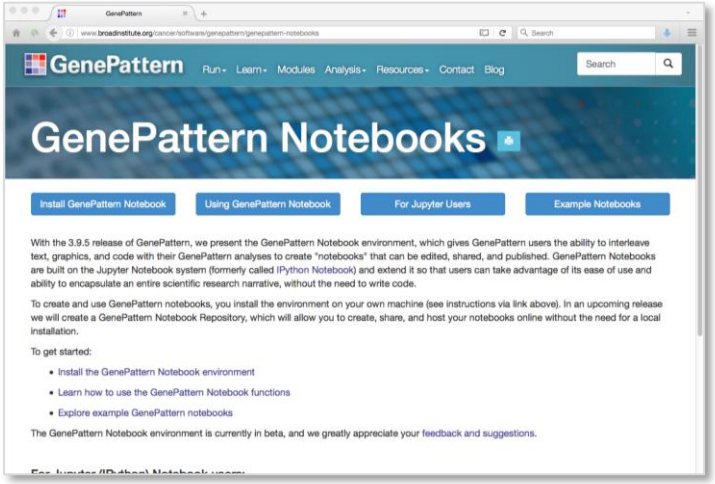


GenePattern Notebooks for Cancer Research

for collaborative and reproducible *in silico* research

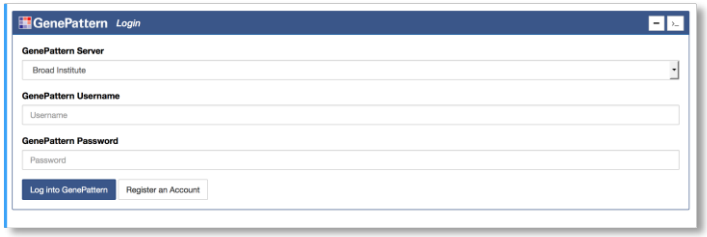
- Support *in silico* research from inception through publication
- Shareable, executable notebook documents
- No programming required
- Seamless integration of GenePattern with Jupyter notebook

GenePattern Notebook Environment



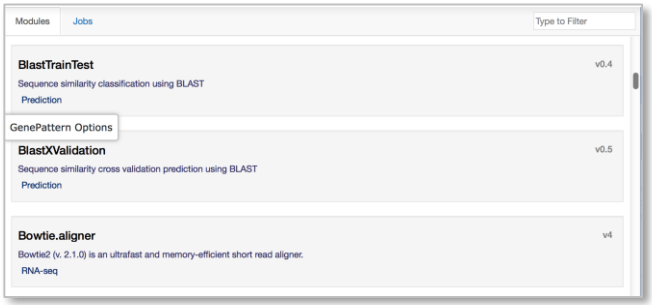
- Announced 2/17/2016
- Available at www.genepattern.org/genepattern-notebooks
- Extension to Jupyter Noteboook system
- Install via Docker (non-programmers) or pip/conda package managers (programmers)
- Outreach:
  - AACR annual meeting, 4/20/2016: poster, ITCR session demo
  - JupyterDays conference, 3/17/2016: talk
  - Open Data Science Conference, 5/31/2015: Talk

Software Development



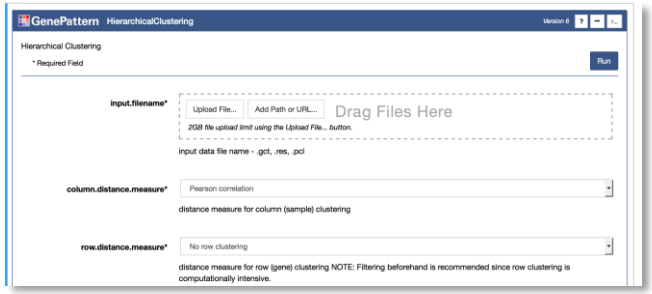
Added a GenePattern login cell

- Input user credentials
- Access any GenePattern server
- Connect through GenePattern API



Added Module browser window

- View, search, and select available modules and pipelines



Added GenePattern analysis cell

- Upload datasets and enter parameters
- Submit jobs to GenePattern server

Added GenePattern job cell

- View job status
- Link to result files and share results

Packaging and distribution

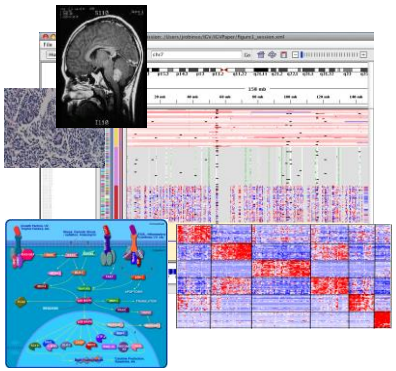
- Packaging of genepattern-notebook extension for use by pip and conda package managers
- Docker image containing complete GP Notebook environment with all dependencies

Predict medulloblastoma outcome

Background: Most common pediatric brain tumor
65% of patients cured via chemotherapy and radiation
Treatment causes **life-altering neurologic side effects**

Challenge: Improved stratification of high/low risk patients

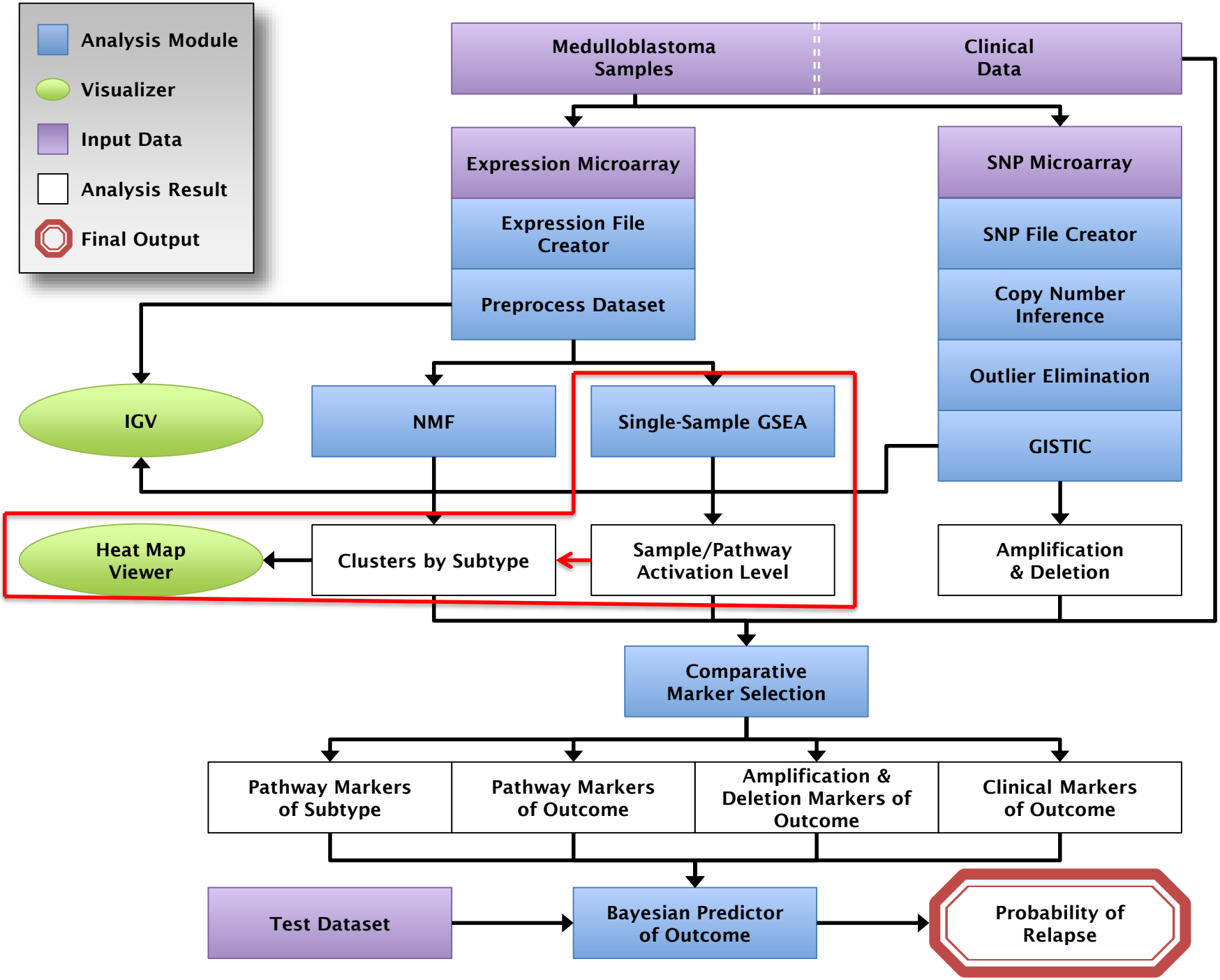
Approach: Integrate 5 clinical and molecular data types
Identify and leverage differences in molecular subtype
Use pathways/signatures for biological/treatment insights
Use sophisticated Bayesian/machine learning algorithm



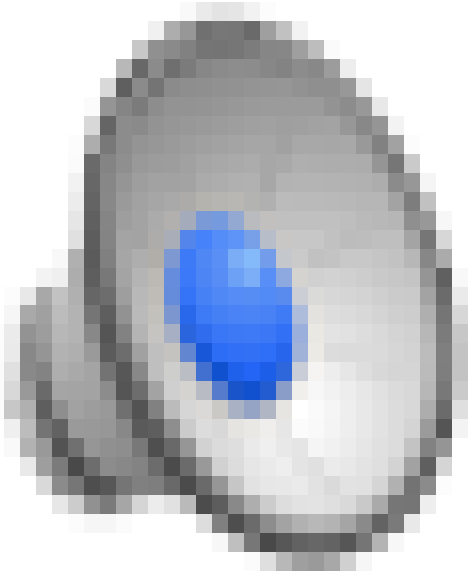
Result: Better accuracy and first to generalize to independent patient cohort

6/15 clinically low risk that relapsed moved to high risk
Characterization subtypes can lead to treatment
Method can apply to other tumor types

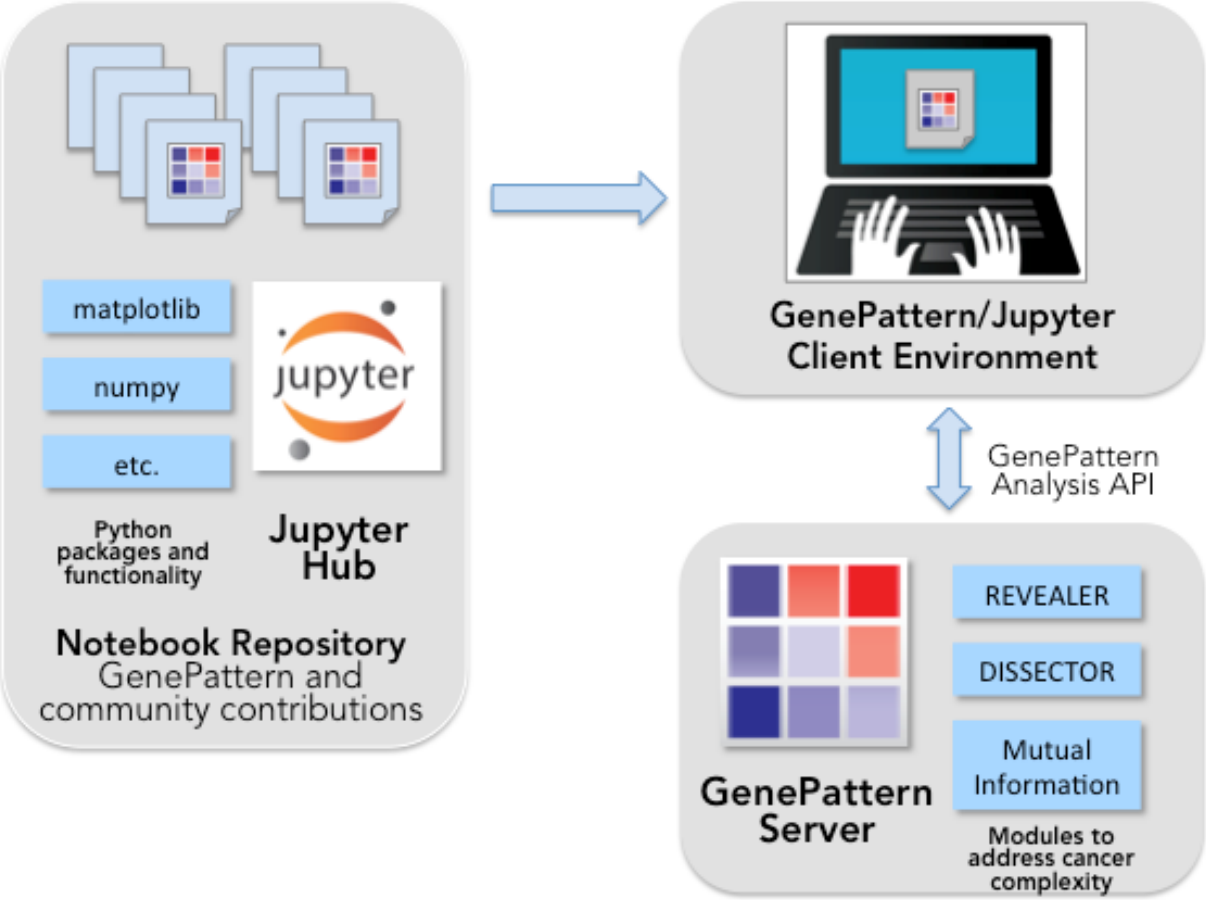
Tamayo & Cho et al. JCO 2011
Cho & Tamayo et al. JCO 2011


Analysis Module
Visualizer
Input Data
Analysis Result
Final Output
Medulloblastoma Samples
Clinical Data
Expression Microarray
Expression File Creator
Preprocess Dataset
SNP Microarray
SNP File Creator
Copy Number Inference
Outlier Elimination
GISTIC
IGV
NMF
Single-Sample GSEA
Heat Map Viewer
Clusters by Subtype
Sample/Pathway Activation Level
Amplification & Deletion
Comparative Marker Selection
Pathway Markers of Subtype
Pathway Markers of Outcome
Amplification & Deletion Markers of Outcome
Clinical Markers of Outcome
Test Dataset
Bayesian Predictor of Outcome
Probability of Relapse

GenePattern Notebook Demo



GenePattern Notebook Repository

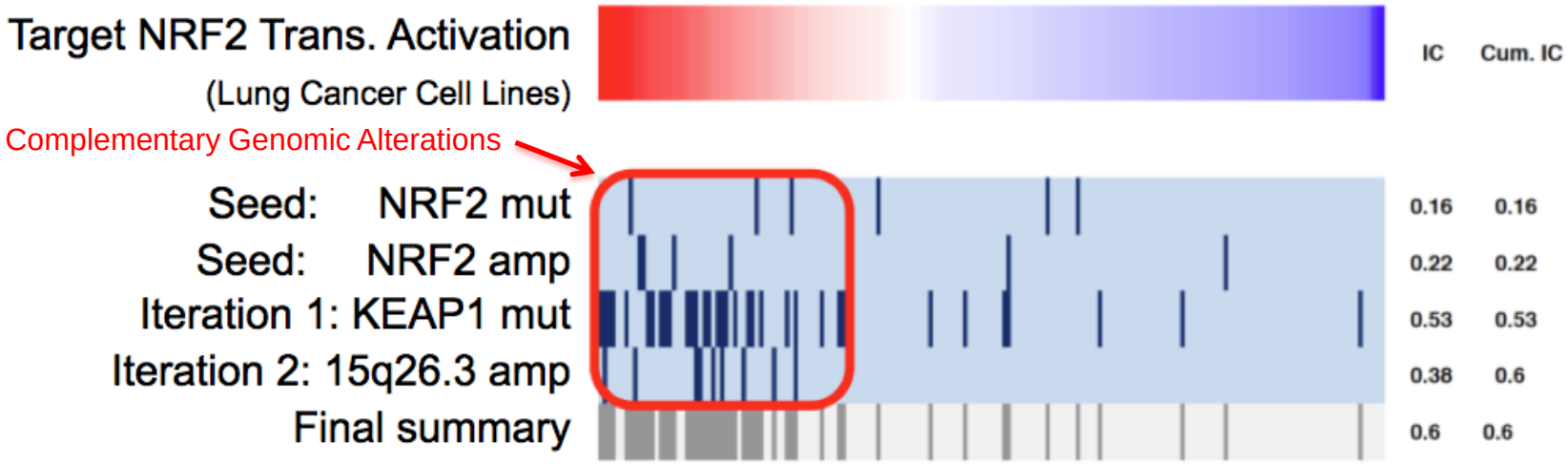


- Create, execute, and share their GenePattern notebooks
- Enable true zero-install usage of GenePattern Notebook
- Python mathematical statistical machine learning and graphics packages
- Comment, save, and adapt shared notebooks
- Alpha prototypes are running at the Broad and UCSD

- Beta release in Q3-Q4 of 2016
- Early adopters contact us

Revealer Functional Characterization of Cancer Genomes

- Shedding light on the relationships between genomic features and biological function
- Maps alterations to profiles of pathway activation, gene dependency and drug sensitivity/resistance
- GenePattern module released, notebook to come



- Keap1 is a master regulator of antioxidant response, shown to interact with NRF2
- 15q26.3 contains NOX5 implicated in production of reactive oxygen species and activates NRF2 (validated in lab)

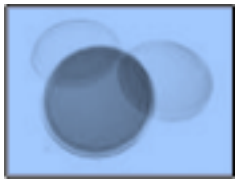
Kim et al., Nature Biotech 2016

DiSCoVER: Combine disease models and *in silico* analysis to identify novel therapeutics

UC San Diego
MOORES CANCER CENTER

Example: **A**ggressive, MYC-driven medulloblastoma subtype

Method:



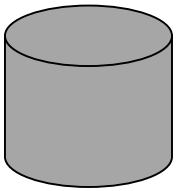
Cell-line model

Mouse xenograft

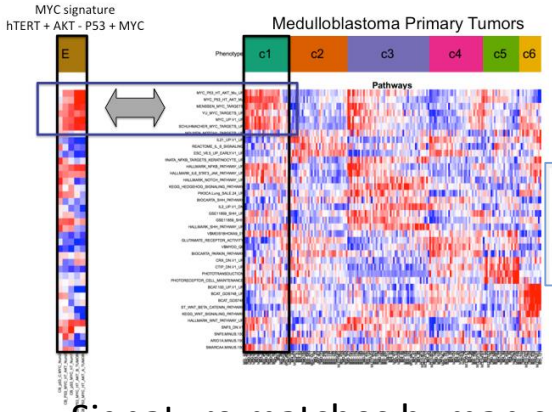
Profiling & Analysis

MYC-driven pathway signature

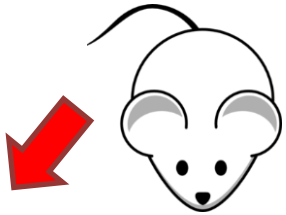
Public data bases
Drug Sensitivity



Associates high MYC signature with CDK inhibitor sensitivity



Signature matches human subtype



Palbociclib extends mouse survival candidate treatment for carefully selected patients

Module and notebook in development

Hannaford, Archer, et al., Clinical Cancer Research 2016

Acknowledgements

GenePattern Team

Peter Carr

Ted Liefeld

Thorin Tabor

Helga Thorvaldsdottïr

Bobbie Hill

Michael Reich

Pablo Tamayo

Collaborators

Hahn Lab
(Dana-Farber Cancer Institute)

Regev Lab
(Broad Institute)

Van Allen Lab
(Dana-Farber Cancer Institute)

Fernando Perez
(Lawrence Berkeley National Laboratory)

Brian Granger
(Cal Poly San Luis Obispo)

Pittsburgh Supercomputing Center

Indiana University National Center for Genome Analysis Support

Texas Advanced Computing Center