



Accessible Reproducible Research

Michael Reich

Director of Cancer Informatics Development
Broad Institute of MIT and Harvard

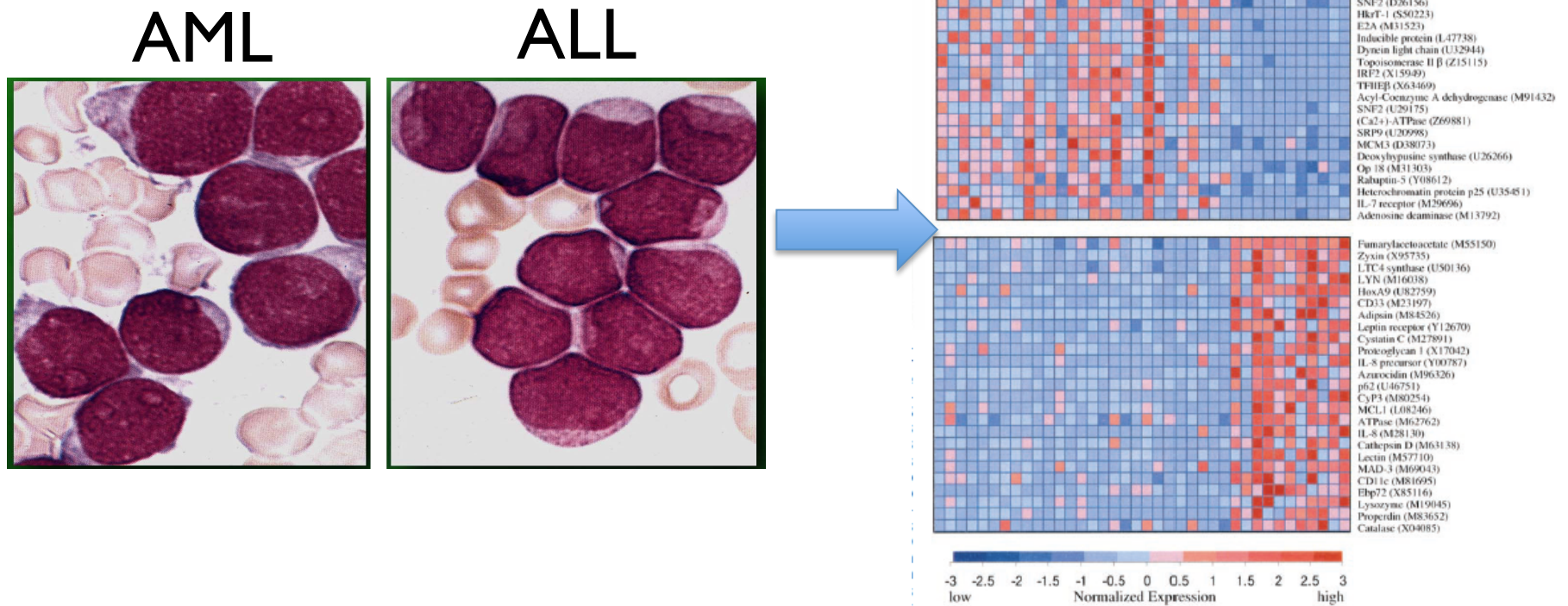
February 19, 2011



Overview

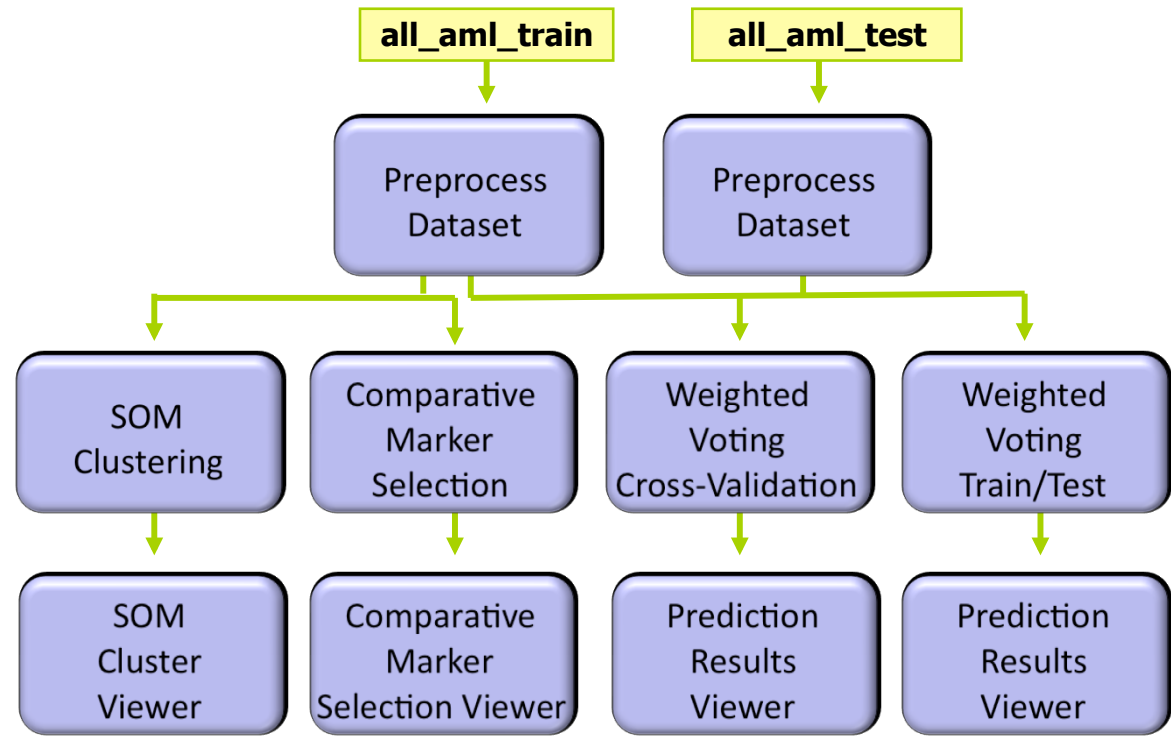
- Background: Molecular Classification of Cancer
- Challenges to Reproducible Research
- GenePattern: A platform for reproducible genomic research
- GenePattern Reproducible Research Document

Background: Molecular Classification of Cancer



Golub et al. *Science* 1999

Cancer Classification Workflow



8 analysis tools/scripts
10 analysis steps

Dataset

Analysis
Step

A reproducible research environment is
required to support multi-tool, multi-dataset
research

Challenges of Reproducible Research

- Claerbout (2000) – “An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

- How did you get those results?
 - New versions of methods, algorithms, implementations
 - Varying parameter settings

A Reproducible Research Environment

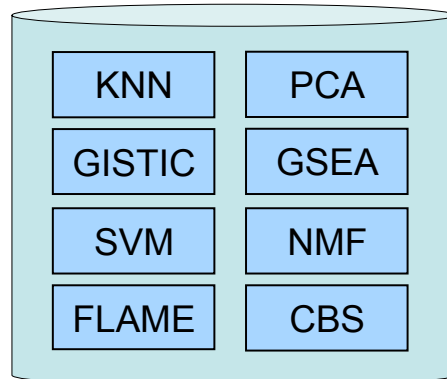
- The community is using large, unshared data
 - Same data called different things by different “owners”
 - Same name for different objects in different repositories or organisms
- Every new paper should provide sufficient information to completely reproduce the results:
 - Data
 - Tools or programs

A Reproducible Research Document

- The paper should include the code and data without specialized knowledge

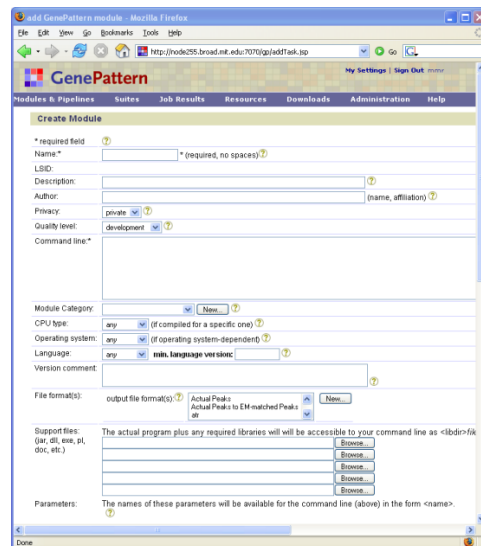
Reproducible Research Environment: GenePattern

Module Repository

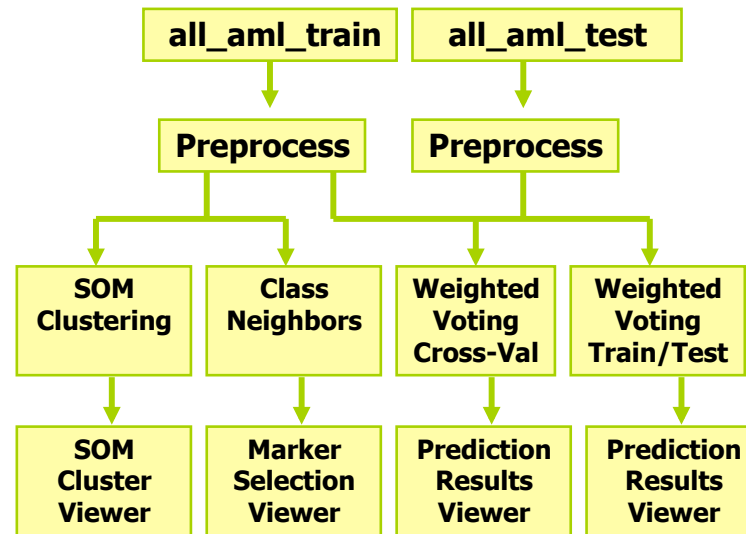


~150 modules

Module Integrator



Pipeline Environment



Golub and Slonim et. al 1999

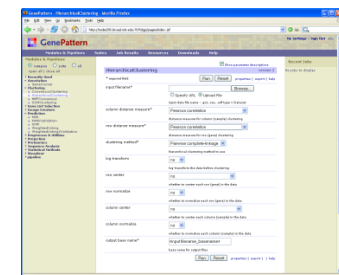


Golub, Slonim, 1999, Nature,
all_aml.pipeline.zip

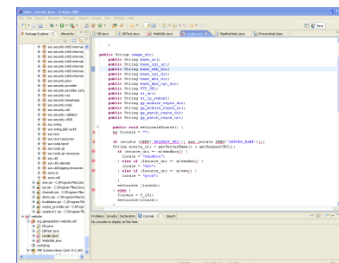
Client User Interfaces



Visualizer



Web



Programming

Gene Pattern Interface

GenePattern - HierarchicalClustering

http://genepattern.broadinstitute.org/gp/pages/index.jsf

Most Visited Broad - INTRANET Broad WebMail Cancer Inf Drupal Download Utility GP Public Home - GenomeSpa... ICBP Portal MIT MMGP MTM Wikipedia

GenePattern - HierarchicalCl... +

GenePattern My Settings | Sign out admin

Modules & Pipelines Suites Job Results Resources Downloads Administration Help

Modules & Pipelines

☒ category ☐ suite ☐ all
open all | close all

- Recently Used
 - ComparativeMarkerSelectionViewer
 - ExpressionFileCreator
 - FLAMEViewer
 - GeneListSignificanceViewer
 - HeatMapView
 - TopHat
- Annotation
 - GeneCruiser
- Clustering
 - ConsensusClustering
 - HierarchicalClustering
 - KMeansClustering
 - NMFConsensus
 - SOMClustering
 - SparseHierarchicalClustering
 - SubMap
 - SubMapBrowser
- Data Format Conversion
 - CSV2FCS
 - GctToPcl
 - PclToGct
- Flow Cytometry
 - FLAMEChooseOptimalClusterNumber
 - FLAMEContourDataGenerator
 - FLAMEMetacluster
 - FLAEMMixtureModel
 - FLAMEPreprocess
 - FLAMEPreviewTransformation
- Gene List Selection
 - ClassNeighbors
 - ComparativeMarkerSelection
 - COPA
 - copyOfComparativeMarkerSelection
 - ExtractComparativeMarkerResults
 - GeneNeighbors
 - GSEA
 - SelectFeaturesColumns
 - SelectFeaturesRows
- Image Creators
 - HeatMapImage
 - HierarchicalClusteringImage
- Missing Value Imputation
 - ImputeMissingValuesKNN
- Pathway Analysis

HierarchicalClustering version 5 [Show parameter descriptions](#)

☒ Show parameter descriptions [Run](#) [Reset](#) [properties](#) | [export](#) | [help](#)

* required field

input filename* [Browse...](#)

☐ Specify URL ☒ Upload File

input data file name - .gct, .res, .odf type = Dataset

column distance measure* [Pearson correlation](#)

distance measure for column (sample) clustering

row distance measure* [No row clustering](#)

distance measure for row (gene) clustering NOTE: Filtering beforehand is recommended since row clustering is computationally intensive.

clustering method* [Pairwise complete-linkage](#)

hierarchical clustering method to use

log transform [no](#)

log-transform the data before clustering

row center [no](#)

whether to center each row (gene) in the data

row normalize [no](#)

whether to normalize each row (gene) in the data

column center [no](#)

whether to center each column (sample) in the data

column normalize [no](#)

whether to normalize each column (sample) in the data

output base name*

Recent Jobs

- TopHat (321141) Feb 18 01:59:52 PM
[stderr.txt](#)
- TopHat (321140) Feb 18 02:05:52 PM
[mm9.p1](#)
[stdout.txt](#)
[stderr.txt](#)



GenePattern Community

GenePattern Users

-  First released in 2004
-  Current release: version 3.2.3 6/2010
-  Over 15,000 registered users
-  Users mailing list: over 2000 members
-  www.genepattern.org

Project collaborations

-  TGCA: The Cancer Genome Atlas Project
-  ICBP: Integrative Cancer Biology Program
-  NCBC: National Centers for Biomedical Computing
-  caBIG: cancer Biomedical Informatics Grid
-  Stanford University: SMD
-  European Nutrigenomics Organization (NuGO)

Provenance in GenePattern

- All analyses are recorded
 - Input files
 - Parameters
 - Code version (via LSID)
- Any previous analysis can be reloaded and rerun
- Pipelines can be cloned, edited, imported/exported
- Each pipeline edit creates a new version
- Pipelines can be inferred from result files, eliminating the need to know all the steps beforehand

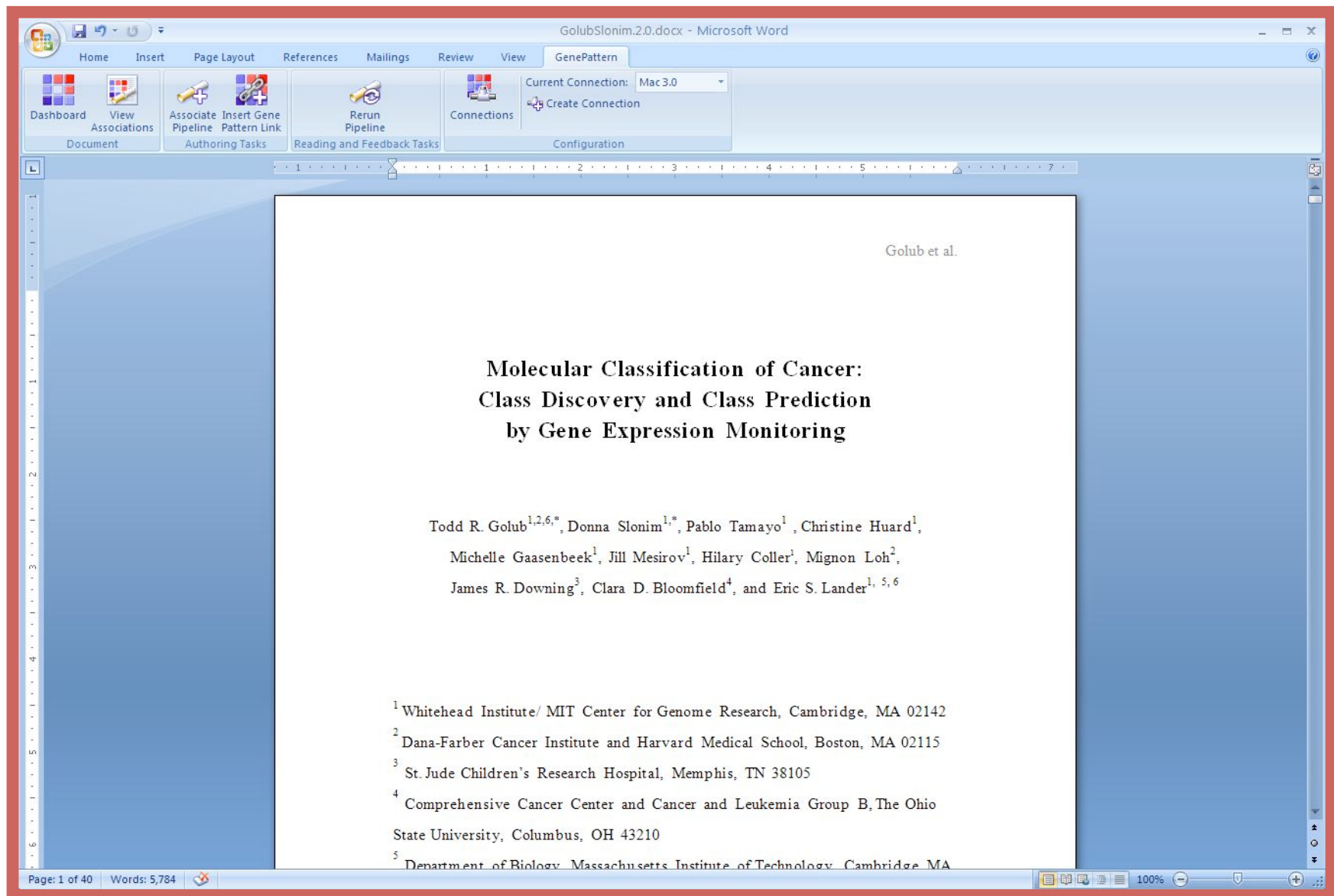
Previous Approaches to Reproducible Research Documents

- Don Knuth: Literate Programming
- Claerbout: use of Makefiles
- Buckheit and Donoho: plots should be reproducible
- Duncan Temple Lang: extensible dynamic documents
- Tony Rossini: Literate Data Analysis
- Fritz Leisch: Sweave
- Robert Gentleman: Compendium combining R & LaTeX

Adapted from Robert Gentleman

Reproducible Research Document Collaboration

Broad, Microsoft, Infusion Development, Persistent Systems



GolubSlonim.2.0.docx - Microsoft Word

Home Insert Page Layout References Mailings Review View GenePattern

Dashboard View Associations Associate Insert Gene Pipeline Pattern Link Rerun Pipeline Connections Current Connection: Mac 3.0 Create Connection

Document Authoring Tasks Reading and Feedback Tasks Configuration

1 2 3 4 5 6 7

Golub et al.

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Todd R. Golub^{1,2,6,*}, Donna Slonim^{1,*}, Pablo Tamayo¹, Christine Huard¹,
Michelle Gaasenbeek¹, Jill Mesirov¹, Hilary Coller¹, Mignon Loh²,
James R. Downing³, Clara D. Bloomfield⁴, and Eric S. Lander^{1, 5, 6}

¹ Whitehead Institute/ MIT Center for Genome Research, Cambridge, MA 02142
² Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115
³ St. Jude Children's Research Hospital, Memphis, TN 38105
⁴ Comprehensive Cancer Center and Cancer and Leukemia Group B, The Ohio State University, Columbus, OH 43210

Page: 1 of 40 Words: 5,784

GenePattern Dashboard

Fig. 2.

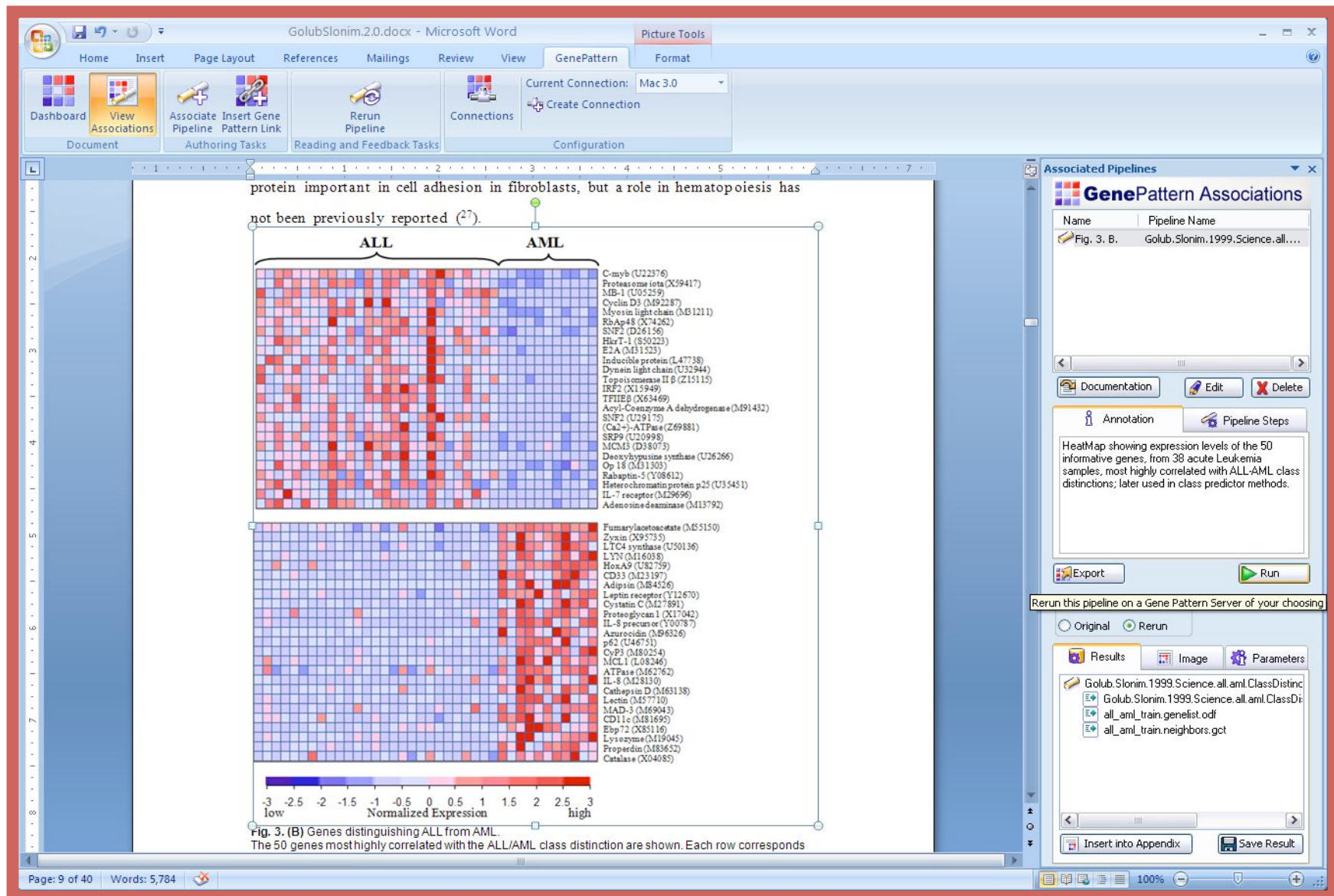
Fig. 3. B.

Scroll to Picture

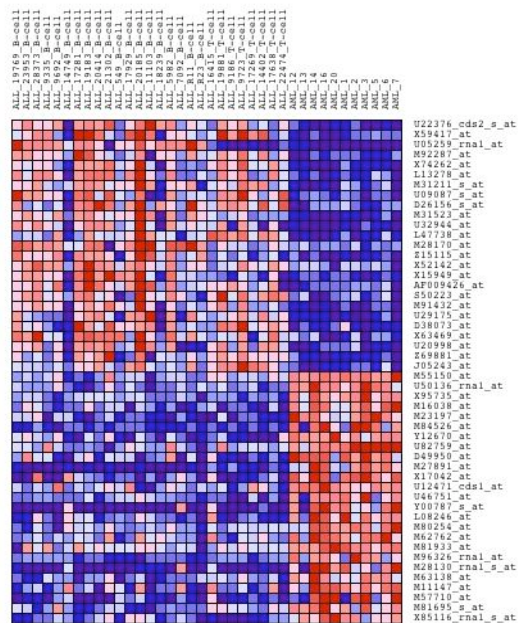
Document Contents

Name	Belongs To	Size (kB)
Golub.Slonim.1...	Fig. 2.	2,159
ClassNeighbors	Fig. 2., Fig. 3. B.	642
GeneListSignific...	Fig. 2.	6,376
Golub.Slonim.1...	Fig. 3. B.	2,159
HeatMapImage	Fig. 3. B.	2,013

Total size: 13.35 MB About...



Rerun of Fig. 3. B. added at 3/24/2008 4:51:48 PM



User: John Smith. Rerun executed with default parameters and datasets.

GenePattern Reproducible Research Document

Golub et al.

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Todd R. Golub^{1,2,6,*}, Donna Slonim^{1,*}, Pablo Tamayo¹, Christine Huard¹,
Michelle Gaasenbeek¹, Jill Mesirov¹, Hilary Coller¹, Mignon Loh²,
James R. Downing³, Clara D. Bloomfield⁴, and Eric S. Lander^{1, 5, 6}

¹ Whitehead Institute/ MIT Center for Genome Research, Cambridge, MA 02142
² Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115
³ St. Jude Children's Research Hospital, Memphis, TN 38105
⁴ Comprehensive Cancer Center and Cancer and Leukemia Group B, The Ohio State University, Columbus, OH 43210

Page: 1 of 41 Words: 5,814

Potential Future Directions

- PDF implementation
- Cross-platform MS Word implementation
- Interoperability with Web-based publication resources, e.g. myExperiment
- Inference engine-based assertion of dataset constraints

Acknowledgements

- Helga Thorvaldsdottir
- Barbara Hill
- Jon Bistline
- Ted Liefeld
- Peter Carr
- Marc-Danie Nazaire
- Microsoft
- Infusion
- Persistent Systems

PI Jill P. Mesirov

<http://genepatternwordaddin.codeplex.com>

<http://www.genepattern.org>

<http://www.genomespace.org>