Cancer Genomics Data Portal

Author: Ethan Cerami, cBio @ MSKCC.

Introduction 2

About this Document 2

Disclaimers 2

To Do List 2

Part I: Functional Specification 2

Prototype 3

Overview of Functionality 3

Non-Goals 5

Competitive Analysis 6

Lowering the Barriers to Use 7

Functionality Details 8

Selection of Pathways and Gene Sets 8

Filtering by Clinical Attributes 8

Threshold Configuration 9

Network Pruning 9

User Registration and Account Management 9

Distribution, Documentation, Outreach and User Feedback 10

Connection to Back-End Services 10

Open Functionality Issues: 10

Part II: Technical Architecture 10

Loosely Coupled Sub-Systems 10

Available Tool-Kits and AJAX Libraries 11

Technical Models 12

Browser Stats 12

Open Technical Issues: 14

Part III: Development Plan 14

Prioritized Feature List 14

Development Schedule 15

Revision History 16

Feedback 16

# Introduction

## About this Document

This document summarizes the proposed functionality of the cBio Cancer Genomics Data Portal. The document is broken into three parts:

* Part I, Functional Specification focuses on end user-functionality.
* Part II, Technical Architecture focuses on technical implementation issues.
* Part III, Development Plan identifies the proposed order of development, and will soon include a proposed schedule.

## Disclaimers

1) Nothing in this document is set in stone. Feedback is welcome.

2) The name, *Cancer Genomics Data Portal,* is intended to be a placeholder. The final name will be determined after further brainstorming with the group.

## To Do List

* Email Carl Schaefer re: CMA plans.
* Update network neighborhood query, based on ideas in UniHi.

## Feedback

* Rule Builder: there should be some way to exclude specific types of alterations for specific genes. For example, amplification of a tumor suppressor gene does not mean that the pathway is altered. (Niki)
* Can we specify gene families, e.g. RAS = NRAS, KRAS, HRAS? How would this be reflected in the heatmap view? (Niki)
* When picking a gene set or pathway, we should have some kind of auto-completion / search feature. (Emek)
* Change patients to samples. (Chris)
* Perhaps represent the matrix as some kind of 3D/4D matrix? (Chris)
* Support very nice printable, paper ready heatmaps (Chris)
* Can we support comparison of gene sets across different cancer types? (Chris)
* Should add links to Mondrian (Chris)
* Things to check out (Gary):
  + Yahoo Pipes
  + Prefuse Flair

# Part I: Functional Specification

## Prototype

A prototype of the cBio Cancer Genomics Data Portal is available at:

<http://cbio.mskcc.org/~cerami/portal/>

## Overview of Functionality

The goal of the cBio Cancer Genomics Data Portal is to enable researchers to selectively visualize and extract multi-dimensional genomic data for all genes within a predefined biological pathway or gene set. As such, the tool enables selective and exploratory data analysis of very large sets of multi-dimensional cancer genomics data.

The following are example questions that a researcher could pose and answer via the data portal:

* Which genes in Glioblastoma are mutated in the Rb signaling pathway, and how frequently are they mutated across all samples?
* How frequently is the Rb signaling pathway altered across all samples, taking into account mutation and copy number data?
* What is the copy number status, mutation status, and expression status for TP53 for all Glioblastoma patients with clinical attribute X?

Users will begin access by interactively building a query via a simple web interface (See Figure 1). Each query is defined by the following four-step procedure:

1. **Select Cancer Type:** For example, select TCGA Glioblastoma or MSKCC Prostate.
2. **Select Genomic Data Type(s):** For example, select Copy Number Alteration and/or Mutation.
3. **Define Gene Set or Pathway:** Enter a user-defined gene set of interest or select from a list of pre-loaded pathways or gene sets.
4. **(Optional) Filter by Clinical Criteria:** For example, only include treated Glioblastoma patients.

Once a query is built, users will have the option to:

1. **Download** tab-delimited text files containing their requested slice of data; this data can then be imported to a third-party application, such as Cytoscape Mondrian, the Broad Integrative Genomics Viewer (IGV) or the statistical R package for further analysis.
2. **Visualize** the data directly within the web browser.

Visual summaries will be presented in four forms:

* **Summary statistics** indicating percentage of samples where the gene set or pathways is altered, along with a frequency plot, showing percentage of samples where each gene is altered.
* **Heat-map view** of all genomic alterations across all patients; and
* **Network view** of the selected pathway with node size or color proportional to percentage of samples where altered.
* **Gene correlation view** of all genes pairs, indicating all correlated and anti-correlated gene pairs.

With a few clicks, and no software installation, researchers will be able to interactively download and visualize individual slices of genomic data. A primary goal of the data portal is to lower the current barriers that prevent researchers from accessing large-scale genomics data and pathway information.

The following additional requirements will also be implemented:

* User access and group control, enabling all users to view public data, and approved users to view MSKCC-specific private data. This will require a registration, login and user-approval process.

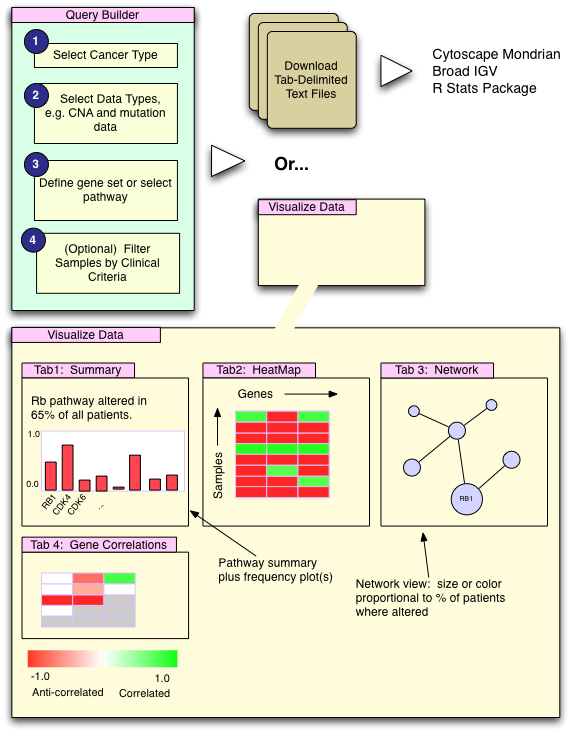


Figure : Overview of Cancer Genomics Data Portal.

## Non-Goals

At this point, it is useful to explicitly state non-goals, e.g. features that we do not intend to build. This helps focus development efforts, and limits the scope of the project.

While all of the following are worthy goals, the data portal will not support the following functionality:

* **Automatic discovery** of commonly perturbed pathways or sub-networks, or automatic discovery of pathways that best distinguish clinical sub-groups. This is a larger research project, which Ethan will be pursuing separately.
* **Session or project management.** For example, the tool will not remember a list of the user’s previously selected data sets or pathways.
* **Statistical tests of significance.** Permutation testing, Fisher’s exact tests and other statistical tests will only be possible by exporting the selected data sets to an external tool, such as R.
* **Loading of user-defined data files.**  For simplicity, users will only be able to load data from the Cancer Genomics Data Server, and will not be able to load their own tab-delimited data files.

As described above, none of these non-goals are set in stone, and subject to additional discussion with the group.

## Competitive Analysis

We have currently identified two major platforms, which overlap in functionality with that proposed here. These are summarized below.

|  |  |  |
| --- | --- | --- |
| **Platform / Tool** | **Pros** | **Cons** |
| **NCI caBIG Cancer Molecular Analysis Portal: Pathway Visualization**  **URL:** [https://cma.nci.nih.gov/cma/](https://cma.nci.nih.gov/cma/graph.do?method=setup)  **Summary:** Enables a user to select and visualize any pathway from a pre-defined list. Pathways are presented visually along with a summary of pathway gene anomalies. | * Does not require installation of third-party software, e.g. a Java WebStart application. * Simple user interface. | * Cannot currently select by cancer type, genetic alteration type, or patient ID. * Genomic data is not overlayed onto pathways. * Summary statistics for patient population not available. * Source of pathways is not clearly defined. * Users cannot select or define their own gene sets. * No user documentation available. |
| **Mondrian, the Genomic Data Mapper**  **URL:** <http://cbio.mskcc.org/mondrian>  **Summary:** Enables a user to visualize genomic alterations in the context of a hand-curated pathway or any pathway or network selected from Pathway Commons. | * Auto-loads with a pathway of interest and genomic data. * Dynamic visualization of networks and heatmaps. * Ability to load genomic data from a remote data server or local files. * Integration with Cytoscape and other third-party analysis plugins. * Direct connection to Pathway Commons via the Pathway Commons Plugin. * User documentation available. | * High-barrier to entry, due to separate WebStart application. * High-barrier to entry, due to complexity of Cytoscape user interface. * High-barrier to entry, due to buried web service functionality. * Summary statistics for patient population not available. |
| **Broad Integrative Genomics Viewer (IGV)**  **URL:** <http://www.broad.mit.edu/igv/>  **Summary:** fast, flexible viewer for genomic data. IGV visually integrates datasets from various platforms and sources, including: genetic variation data (copy number, loss of heterozygosity, somatic mutations), gene / microRNA expression data  epigenetic data; RNAi screens;  genomic features; clinical and phenotype annotation of samples. | * Integrate multi-dimensional genomic data. * Fast performance. * Powerful sorting and filtering options. * Good documentation. | * Does not include support for gene sets or biological pathways. |

## Lowering the Barriers to Use

A primary aim of the proposed tool is to lower the barrier of entry for researchers. Based on prior experience in developing bioinformatics tools and data portals, we have identified the following set of barriers:

* **Separate Software Installation:** Requiring a user to install a standalone application or even a Java webstart application can impose a barrier for many end-users. A web-based interface, which requires no additional software is likely to attract the greatest number of new users.
* **Complex User Interfaces:** Cytoscape includes a complex user interface, which overwhelms many new users. With the complex interface comes the ability to do more complex analysis, but may types of analysis and research do not require such complexity.
* **User-generated data files:** New users do no want to reformat their genomics data into the proper format and have a separate step simply to load that data into the system. Rather, the system should provide seamless access to a back-end data server, thereby enabling the user to simply select from a menu of existing data sources.
* **Visual Complexity:** The visual complexity of the original Cytoscape BioPAX plugin has now given way to the simplified binary view. Until we have a robust way of handling compound graphs, this simplified binary view will probably remain as the most understandable and approachable view.
* **Lack of training and outreach:** We have failed to reach out to collaborators and provide brief tutorials on how to get started. We can also improve the method by which we solicit feedback from collaborators.

Addressing these issues, the Cancer Genomics Data Portal will be entirely web-based (requiring no additional software installation), and will include a radically simplified user interface. It will also use the simplified binary view of networks, and provide direct access to multiple cancer data sets.

## Functionality Details

### Selection of Pathways and Gene Sets

The user can define gene sets or pathways by any of the following methods:

* User defined gene sets entered via a text box.
* Any pathway chosen from Pathway Commons.
* Network neighborhood of any gene or gene set, as defined in Pathway Commons.
  + Neighborhood can be filtered by interaction type and/or data source. For example, use the network neighborhood of all protein-protein interactions of BRCA1, as defined in HPRD.
* Any gene set from the Broad Molecular Signature Database (MSigDB).

### Filtering by Clinical Attributes

Users will be able to filter data sets by any combination of clinical attributes. For example, download data for primary v. secondary Glioblastoma patients only.

The user interface for interactively building filter rules remains to be defined. We need to carefully balance the tradeoffs between a fully configurable interface capable of representing any boolean query with the complexity that such an interface would require. One model interface that may server to inspire us is the filter interface provided by the Mac OS X Automator tool (See Figure 2 for a sample screenshot).

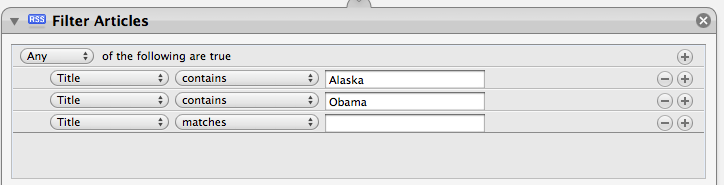


Figure : Screenshot of the Filter Articles Action in the Mac OS X Automator software. The interface enables user to build and/or logical filter queries, and may serve as a model user interface for building complex filter and threshold rules.

### Threshold Configuration

Threshold configuration can be used to define criteria for significant genomic alterations. This threshold information is then used to build the heatmap and network views. For example, a gene can be marked as “altered” if any of the following conditions occur:

* Copy number alteration value > |1|
* Mutation value equals MUTATED

As is the case with the filter interface, we need to carefully balance the tradeoffs between a fully configurable interface capable of representing any boolean query with the complexity that such an interface would require. Again, the Mac OS X Automator filter interface may serve as a useful model.

### Network Pruning

Having selected a pathway or network neighborhood, the user should have the option to easily “prune” the network of genes. For example, remove Genes C and F, and redo the analysis.

### User Registration and Account Management

The site will need to support the following functionality:

* New user registration, including the option to request access to restricted data sources.
* New user management, whereby an administrator can grant/deny requests to restricted data sources.
* Password recovery.
* Account management, including ability to update user name and/or password.

Only approved users will be able to view MSKCC-private data. User credentials and login information for the data portal should also apply to all subpages within cbio/cancergenomics. For example, a user approved to query MSKCC prostate data should also automatically have access to all restricted pages cbio/cancergenomics/prostate.

## Distribution, Documentation, Outreach and User Feedback

Once the portal is built, we propose to:

* Create HTML and PDF documentation for using the portal.
* Create HTML and PDF documentation for using the CGDS Web Service API.
* Create a 5 minute intro tutorial, and place it right on the home page for the tool.
* Create a priority list of collaborators at MSKCC, and provide 15-20 minute tutorials for these collaborators and their lab members. Possible collaborating labs may include: Cameron Brennan, William Gerald, etc.
* Create a one-page PDF brochure explaining what the tool does, and email it to all TCGA collaborators.
* Create a mailing alias for submission of feature requests and bug reports.

## Connection to Back-End Services

In order to function, the Data Portal assumes the following back-end services:

* A cancer genomics data repository, as currently provided by the Cancer Genomics Data Server (CGDS).
* Additional web services, to be implemented and made available on the Pathway Commons web site.

## Open Functionality Issues:

1. How exactly will the user prune a network? See “Networking Pruning” section above.

# Part II: Technical Architecture

## Loosely Coupled Sub-Systems

The data portal will be built with three separate, loosely coupled sub-systems (see Figure 3):

* **Data Portal:** User interface and front-end system.
* **Cancer Genomics Data Server (CGDS):**  Back-end database with all cancer genomics data and web services for accessing such data.
* **Pathway Commons:** Back-end database with all pathway data and web services for accessing such data.

Architecturally, the data portal will never access genomic or pathway data directly from the CGDS or Pathway Commons databases. Rather, all access will be made via available web service calls. There are multiple advantages to building a loosely-coupled architecture, such as the one proposed above, including:

* Clearly defined modules, enabling parallel work by multiple developers on any of the three sub-systems.
* The further extension of the existing web service APIs for CGDS and Pathway Commons. This will help drive Pathway Commons work and enable other groups to build on CGDS and Pathway Commons and build their own applications.

That said, having a loosely-coupled system does present some challenges. Specifically:

* The Data Portal will require new web API calls within Pathway Commons and CGDS. Since these calls do not yet exist, the completion of the data portal will be dependent on Pathway Commons and CGDS work and production deployment.
* It is not yet clear how we will restrict user Web API access to protected data sources within CGDS, nor is it clear where authentication of users will rest.

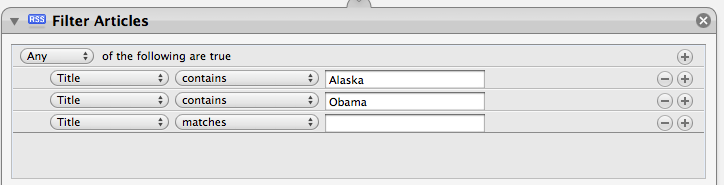


Figure : Proposed architecture of loosely-coupled sub-systems.

## Available Tool-Kits and AJAX Libraries

The portal outlined in this document will probably need to leverage several recent advances in building Rich Internet Applications (RIA). These include advances in AJAX (Asynchronous JavaScript and XML), SVG (Scalable Vector Graphics), modern web browser support for Javascript and CSS, and Adobe Flex. The following tool-kits and platforms may prove useful:

|  |  |
| --- | --- |
| **Tool-Kit** | **URL** |
| **The Yahoo! User Interface Library (YUI):** a set of utilities and controls, written in JavaScript, for building richly interactive web applications using techniques such as DOM scripting, DHTML and AJAX. | <http://developer.yahoo.com/yui/> |
| **Raphaël:** small JavaScript library that simplifies work with vector graphics on the web. For example, you can draw circles, rectangles, etc. directly within a browser. | <http://raphaeljs.com/> |
| **Google Charts:** API for dynamically and easily creating all sorts of web-based charts. | <http://code.google.com/apis/chart/> |
| **Prototype:** one of several Javascript / AJAX libraries. | <http://www.prototypejs.org/>` |
| **Adobe Flex:** a collection of technologies released by Adobe Systems for the development and deployment of cross platform rich Internet applications based on the proprietary Adobe Flash platform (Source: Wikipedia) | <http://www.adobe.com/products/flex/> |

## Technical Models

The following web sites may contain useful technical implementations that we can borrow and/or extend:

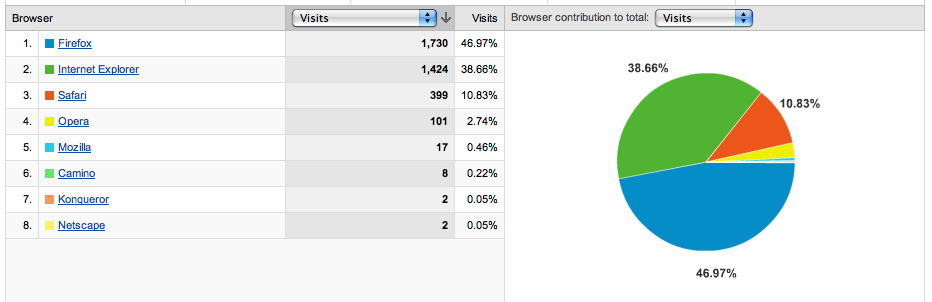
* Reactome beta site: built with multiple AJAX toolkits. <http://www.reactome.org/>.
* JSViz: javascript-based network drawing library. <http://www.jsviz.org/blog/>.

## Browser Stats

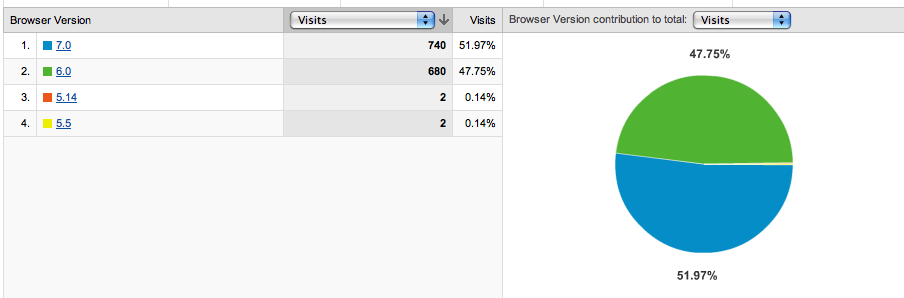
We will need to balance the need for rich internet applications with potential cross-browser incompatibility problems. Such cross-browser incompatibility problems could add significantly to the development effort.

For baseline information, I have compiled browser information, as recorded on PathwayCommons.org for June 1- August 31, 2008.

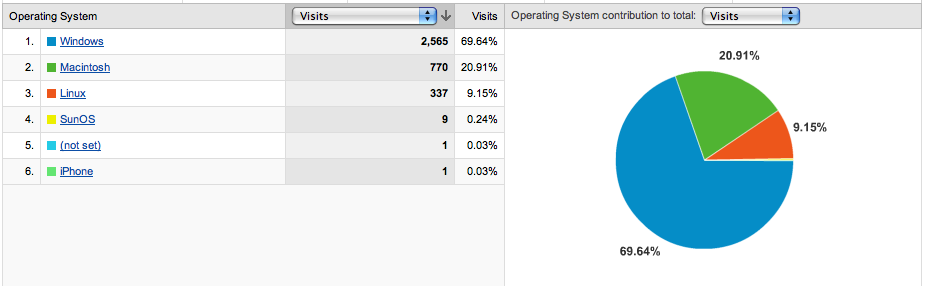
Overall Web Browser Stats:



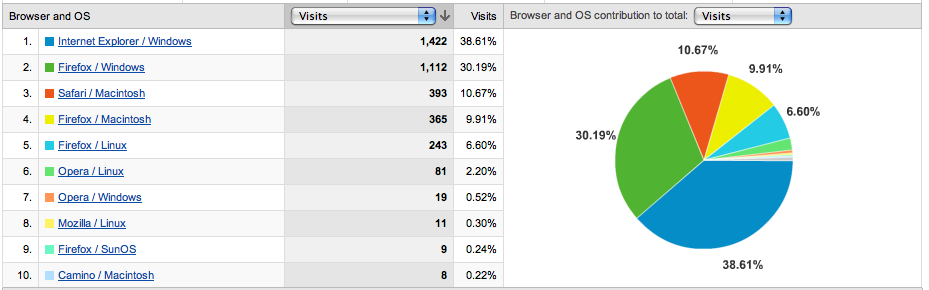
Internet Explorer Stats:



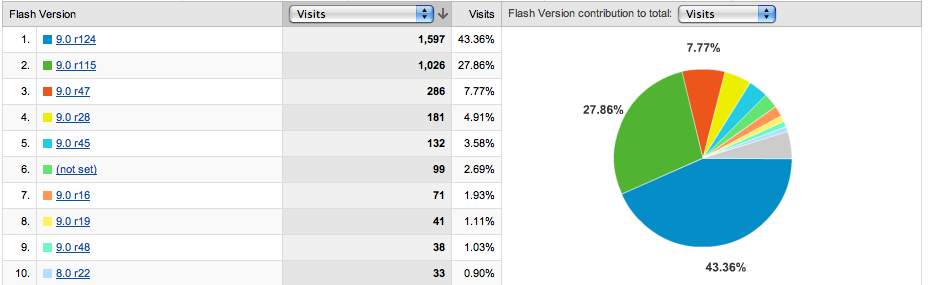
Operating System Stats:



Browser plus Operating System Stats:



Flash Versions Stats:



## Open Technical Issues:

1. Where exactly will the network visualization code live? Do we want to reuse the same network visualization code in the data portal and in Pathway Commons?

# Part III: Development Plan

## Prioritized Feature List

We cannot build the entire data portal in one go-around, and therefore need to prioritize functionality, based on usefulness to the research community. Ideally, we deploy a minimal working portal in 2-3 weeks, and update the portal with new functionality each month.

1. Basic Query Builder + User Defined Gene Sets🡪 tab-delimited files.
2. Selection of Pathways 🡪 tab-delimited files.
3. Network Neighborhood 🡪 tab-delimited files.
4. Gene Sets from the Broad Molecular Signature Database (MSigDB).

🡪 tab-delimited files.

1. Filter by clinical attribute 🡪 tab-delimited files.
2. Visual interface 1: Summary stats
3. Visual interface 2: Heatmap view
4. Visual interface 3: Gene Correlation view
5. Visual interface 4: Network view
   1. May require extensive UI development
6. Threshold configuration
7. Access control and user account management
8. Network pruning

## Development Schedule

Once the functionality specification is finalized, and we determine who exactly will be working on this, we will finalize a development schedule. In the meantime, Figure 4 summarizes the main development tasks, broken down by feature.

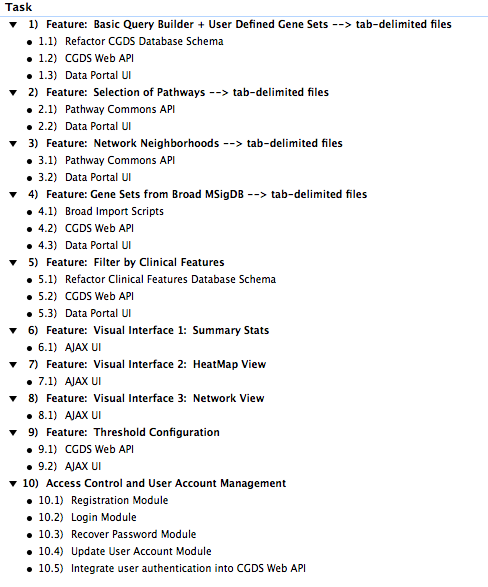


Figure : Main Development Tasks, broken down by feature.

# Revision History

|  |  |  |
| --- | --- | --- |
| 1.0 | August 21, 2008 | Ethan Cerami: initial draft created. |
| 1.1 | August 26, 2008 | Updated, based on much feedback from Niki Schultz:   * Changed tool name for Pathway Explorer to Data Portal. * Integrated visualization tools with access to flat tab-delimited text files. * Added new section: Threshold configuration. * Added new section: User Registration and Account Maintenance. * Added new section: Prioritized Feature List. |
| 1.2 | August 27, 2008 | Added Parts II and III:   * Technical Architecture * Development Plan |
| 1.3 | September 2, 2008 | Multiple changes, based on feedback from Gary, Emek, Ben and Nicki:   * Added Broad IGV to competitive analysis section. * Added Flex as possible tool, instead of AJAX. * Added new technical models section with reference to Reactome beta and jsviz.org. * Added new section on browser stats. * Added details re: gene correlation view. * Added Emek’s idea re: network pruning. * Added note re: where will mini-map code live? Should this be a pc web service? * Added URL for prototype. |

# Feedback

The following people have been consulted and have provided feedback:

* Nicholas Schultz
* Gary Bader
* Emek Demir
* Benjamin Gross