GenePattern 3.1.1 Integration Guide

Software Copyright

The Broad Institute
SOFTWARE COPYRIGHT NOTICE AGREEMENT

This software is supplied without any warranty or guaranteed support whatsoever. Neither the Broad Institute nor MIT can be responsible for its use, misuse, or functionality.

Integration Guide

Download pdf

This document provides guidance to system administrators interested in integrating GenePattern into the analysis tools at their site.

Typographical conventions:

	Tables like this describe implementation on the GenePattern public server.

Audience for this guide

This document is intended for system administrators. It assumes that you are comfortable installing and configuring client-server applications. It assumes that you have a basic understanding of GenePattern, how it's installed, and how it works. This guide highlights issues that might arise and how to address them. It provides links to relevant portions of the GenePattern documentation, supplementing that documentation as needed.

This document assumes that you are familiar with:

- GenePattern fundamentals as described in the Concepts Guide
- Basic operations as described in the Tutorial
- Installing GenePattern and its analysis modules as described on the download page: http://www.genepattern.org/download/

Installing GenePattern

The standard installation procedure uses Install Anywhere to install the server on Windows, Mac, or Unix using a Tomcat web server. To install on a different web server or on another platform, use the WAR file installer. Instructions for both the standard installation and the WAR file installation are on the download page: http://www.genepattern.org/download/.

Hardware and software requirements for GenePattern are described in the Release Notes.

GenePattern Database

The GenePattern server runs against a database. The GenePattern installation creates an HSQL database. For instructions on how to build and use an Oracle database instead, see Changing the GenePattern Database (HSQL to Oracle) in the *Web Client Guide*.

	We use an Oracle database for the GenePattern public server.

Securing the GenePattern Server

Access | User Accounts | Authentication | Permissions | SSL

The Web Client Guide discusses how to secure your GenePattern server, including access to the server from client machines, GenePattern user accounts, authentication (e.g. username & password) and authorization (e.g. permissions). The following sections briefly summarize this information. For more detail, see Securing the Server in the *Web Client Guide*.

Access

By default, any client machine can access a GenePattern server. Optionally, you can configure your GenePattern server to restrict access to selected domains. See Securing the Server.Access Filtering. in the *Web Client Guide*.

	Access to the GenePattern public server is not restricted.

User Accounts

A user must have a GenePattern account to log into the GenePattern server. By default, when a user first logs into the server, GenePattern automatically create an account for that username.

To enable registration, in the genepattern.properties file, set require.password=true. This setting adds a registration link (and password prompt) to the GenePattern login page. The first time users log into GenePattern, they must click the registration link to create an account. User account information is stored in the GenePattern Database.

Alternatively, configure the GenePattern server to not allow users to create GenePattern accounts (create.account.allowed=false). In this case, new user accounts must be explicitly created by editing the GenePattern database.

See Securing the Server.Password Protection in the *Web Client Guide*.

▦	Registration (and passwords) are enabled on the GenePattern public server

Authentication

Each GenePattern user must register to access the GenePattern server. By default, GenePattern requires only a username for authentication. Optionally, you can configure the GenePattern server to require both a username and a password for authentication. See Securing the Server.Password Protection in the *Web Client Guide*.

GenePattern user authentication is performed by a servlet filter installed in front of the GenePattern web application in its web.xml file. To provide site-specific authentication for the GenePattern server, write your own servlet filter. See Securing the Server.User Authentication in the *Web Client Guide*.

▦	The username and password authentication provided by the GenePattern installation is the authentication used by the GenePattern public server hosted at the Broad Institute.
Collaborator	A large university uses Kerberos to provide username and password authentication for their network. They wrote their own servlet filter to have the GenePattern server also authenticate using Kerberos.

Permissions

GenePattern permissions are based on two configuration files:

- userGroups.xml defines user groups
- permissionMap.xml defines which user groups have which permissions; the permissions themselves (e.g. CreateModule, adminModules, and so on) are predefined and cannot be added or removed

GenePattern user authorization is performed by a servlet filter installed in front of the GenePattern web application in its web.xml file. By default, users are assigned permissions based on the username they provide at login. To provide site-specific authorization for the GenePattern server, write your own servlet filter. See Securing the Server.User Permissions in the *Web Client Guide*.

▦	The user authorization provided by the GenePattern installation is the authorization used on the GenePattern public server. The following permissions are restricted to a small number of users in the Administrator group: ● createModule – we restrict this to prevent malicious code on the server ● createPublicPipeline – we restrict this to prevent proliferation of untested pipelines ● adminJobs, AdminModules, adminPipelines, adminSuites – we restrict these to preserve privacy ● adminServer – we restrict this to secure the server

SSL

This section of the Web Client Guide describes how to modify the GenePattern web application to run on a web server that is configured to use the HTTPS protocol. See Securing the Server.Secure Sockets Layer (SSL) Support in the *Web Client Guide*.

▦	The GenePattern public server is not running under SSL.

Other Security Considerations

We take the following additional steps to secure the machine running the GenePattern public server (these steps may not be necessary on less public servers):

- create an operating system user account with limited permissions from which to run the GenePattern server
- disable JSP compilation and remove the compiler
- prevent users from entering an input file path (file:// urls) as an input file for a module; to do so, edit genepattern.properties and set allow.input.file.paths=false

  By default allow.input.file.paths=true, which allows users to input an arbitrary network file path (such as file:///server/directory/file.gct) as the value for an input file parameter. When allow.input.file.paths=true, you can use the server.browse.file.system.root property to set a root directory where the GenePattern server begins browsing for the specified network file path.
- prevent LSF log files from being displayed with other job results in the Web Client; to do so, edit genepattern.properties and set jobs.FilenameFilter=.lsf* (further discussion in Running Modules in a Cluster)

Modules

Installing Modules | Creating Modules | Running Modules in a Cluster | Managing Memory for Modules | Module Notes

This section discusses how to install, create, and manage modules.

Installing Modules

By default, you install modules, pipelines, and suites from the Broad repository. The module repository contains more than 100 modules and pipelines. Suites are stored in a separate suite repository. For instructions on how to install modules from the repository, see Managing Modules, Pipelines, and Suites in the *Web Client Guide*.

The repository is updated on a regular basis. We recommend checking for new modules on a weekly basis.

Create your own repository: Optionally, you can select an alternate repository from which to install modules, pipelines, and/or suites. See Repositories in the *Web Client Guide.*

> At the Broad, we maintain a development repository for modules in development and a production repository for released modules. Only the production repository is available from the GenePattern public server.

Creating Modules

For instructions on how to create modules, as well as a step-by-step tutorial for creating a module, see Creating Modules in the *Web Client Guide*.

Running Modules in a Cluster

Queuing systems such as the Load Sharing Facility (LSF) and the Sun Grid Engine (SGE) allow computational resources to be used effectively. If you have such a queuing system, you typically want the GenePattern server to use it. The *Web Client Guide* includes instructions for configuring the GenePattern server to use a queuing system: see Using a Queuing System.

As described in the instructions, you click *Administration>Server Settings* and use the Command LIne Prefix page to define the command prefix that runs the module on the cluster. The instructions use the *Default Command Prefix* field of the Command Line Prefix page to define one command prefix for all modules, which sends all modules to one queue. You can use that same page to define unique command line prefixes for specific modules. This allows you to send different modules to different queues, which helps to address hardware and memory issues. For example, certain modules (such as SNPFileCreator or HierarchicalClustering) require significant amounts of RAM.

The script described in the instructions writes the LSF log file into the job results directory. To prevent the Web Client from displaying the lsf log files with the rest of the job results, edit the genepattern.properties file and set jobs.FilenameFilter=.lsf*.

> The GenePattern public server uses two queues: one for most modules and one for modules that require large amounts of memory. Modules sent to the ‘bigmem’ queue are run on a cluster of large memory machines. LSF log files are hidden.

Managing Memory for Modules

In GenePattern, you manage memory for modules in one of two ways:

- Add appropriate parameters to the command line that executes the module. Click *Administration>Server Settings* and use the Programming Languages page to add memory parameters (e.g. –Xmx512M) for
  - modules written in Java (Java VMOptions field)
  - modules written in R (R Options field)
  - visualizer modules (Java Visualizer VMOptions field) – Visualizer modules are run on the client machine; therefore, individual users can override this setting by clicking My Settings and specifying the Java Visualizer VMOptions appropriate for their machine.
- Override memory parameters on a per-module basis. Create a file, java_flags.properties, in the GenePattern /resources directory. Each line of the file lists the LSID of a module and the memory setting for that module. To find the LSID of a module, in the Web Client click the module and then click the *Properties* link. Following is an example file:

```
# Here is an example which allocates extra RAM for some of the modules:
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00087=-Xmx2500m
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00086=-Xmx10G
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00085=-Xmx2500m
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00106=-Xmx2500m
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00096=-Xmx2500m
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00094=-Xmx2500m
urn\:lsid\:broad.mit.edu\:cancer.software.genepattern.module.analysis \:00093=-Xmx2500m
```

Modules that Require Extra Memory

The following modules frequently require additional memory:

- HierarchicalClustering
- PreprocessDataset
- SNP Analysis modules
  - CopyNumberDivideByNormals
  - GISTIC
  - GLAD
  - SNPFileCreator
  - SNPFileSorter
  - SNPMultipleSampleAnalysis
  - XChromosomeCorrect

On the GenePattern public server, these modules are sent to a cluster of large memory machines.

genepattern.properties

Most server configuration options are in the genepattern.properties file /resources directory. Most of the options in this file can be set through the Web Client interface by clicking *Administration>Server Settings*. For descriptions of the options, see Modifying Server Settings in the *Web Client Guide.*

The options listed in the following table can only be set by editing the genepattern.properties settings. We recommend editing the properties through the Web Client when possible.

require.password create.account.allowed	See User Accounts
GenePatternURL fqHostName	See http://www.genepattern.org/doc/faq/index.html#35

fullyQualifiedHostName gpServerHostAddress	

allow.input.file.paths=true
server.browse.file.system.root=/
	See Other Security Considerations

input.file.mode=path	Determines how GenePattern handles network file paths: ● **path** (default) leaves the network file in place ● **move** copies the network file to the job directory on the GenePattern server before beginning the job and copies the file back to its original location after the job completes
soap.attachment.dir=../temp/attachments	Used for the GenePattern SOAP interface. Specify a temporary directory to be used for SOAP messages with attachments.

Web Service Interface

All GenePattern server functionality is available programmatically. There are two basic access methods:

- Option 1, use one of our programming libraries (Java, MATLAB, or R). The libraries are designed to allow you to use the programming environment as a client, running modules and retrieving results from the GenePattern server. See the Programmer's Guide for more information and examples of how to use the libraries.
- Option 2, use SOAP calls directly to http://genepattern.broad.mit.edu/gp/services. This is a standard SOAP interface. The SOAP client interacts with the GenePattern interface as it would with any other SOAP interface.

Documentation Update History

Version	Release date	Comments
3.1.1	July 2008	Initial draft, GenePattern 3.1.1 Beta.