GenePattern Concepts Guide

GenePattern provides access to a broad array of computational methods used to analyze genomic data. Its extendable architecture makes it easy for computational biologists to add analysis and visualization modules, which ensures that GenePattern users have access to new computational methods on a regular basis.

The GenePattern **Concepts Guide** provides a brief introduction to GenePattern: its primary objects (modules, pipelines, suites) and its client-server architecture. All other GenePattern documentation assumes that you are familiar with these concepts.

Analysis and Visualization Modules	Describes the GenePattern modules, which you use to perform computational analysis against your data.
Pipelines	Describes GenePattern analysis pipelines, which you can use to create workflows guiding the execution of multiple analysis modules.
Suites	Describes GenePattern suites, which you use to organize modules and pipelines.
GenePattern Platform	Describes the GenePattern client/server architecture, how GenePattern runs analysis jobs, and the Web and Desktop Clients, which you use to interact with GenePattern.
GenePattern Security and Permissions	Describes how a system administrator can limit access to the GenePattern server and how that affects the operations available to you.
Version Numbers	Describes how GenePattern ensures reproducible analysis results by using Life Science Identifiers (LSIDs) to uniquely identify every version of every module, pipeline, and suite.

Once you are familiar with the GenePattern fundamentals described in the **Concepts Guide**, refer to the following GenePattern documentation for more information:

- The Tutorial provides a hands-on tour of the Web Client and Desktop Client.
- The Desktop Client Guide describes the Desktop Client and how to use it.
- The Web Client Guide describes the Web Client and how to use it.
- The Programmers Guide provides guidelines for writing modules and instructions for accessing GenePattern from the Java, MATLAB, and R programming environments.
- The Modules page lists the modules and pipelines in the Broad repository, with links to their documentation.
- The File Formats Guide describes all file formats and provides instructions for creating input files.
- The Release Notes describe new features and known issues in this release.
- Frequently Asked Questions answers commonly asked questions about GenePattern.

Analysis and Visualization Modules

Analysis and visualization modules are at the heart of GenePattern:

- **Analysis modules** provide computational methods and tools for gene expression analysis, proteomics data analysis, SNP analysis, and data preprocessing and conversion.
- **Visualization modules** display your data and analysis results graphically. GenePattern includes two types of visualization modules:
  - Visualizers display your data graphically and allow you to manipulate that view interactively. By convention, these modules have "Viewer" in the name. These modules are run on your desktop computer; all other modules are run on the GenePattern server.
  - Image creators create static graphics (for example, gif and pdf files) for display in other applications. By convention, these modules have "Image" in the name.

Each module includes its own documentation, which is supplied by the module developer. For a list of the modules in the repository maintained by the Broad Institute, with links to their documentation, see the Modules page of the GenePattern web site.

Pipelines

Pipelines combine analysis and visualization modules into a single workflow. Pipelines can also include other pipelines. Running a pipeline runs the sequential series of modules and pipelines defined by that pipeline.

Pipelines can be defined to analyze a particular dataset; for example, you might create a pipeline to reproduce published analysis results. They can also be parameterized, which allows the person running the pipeline to provide datasets and other configuration values for the analysis. Pipelines can be used to run multiple analyses against a single data file or to run a progressive series of analyses, where the output from one analysis is used as input for the next.

When you create a pipeline, you select the modules (and pipelines) to be executed by the pipeline. When you add a module to your pipeline, you also select the parameter values for that module. You can enter a parameter value as a static value, variable value, or chained value:

- Static: you specify the parameter value.
- Variable: you have GenePattern prompt the user for the parameter value each time the pipeline is run.
- Chained: you use an output file from one module as the input parameter value for a subsequent module. This allows you to create a pipeline that runs a progressive series of analyses.

Pipelines are easily exported and imported from GenePattern, which allows you to share pipelines with colleagues who would like to reproduce your analysis results. They can even be used to document your research. By providing a way to create and distribute an entire computational analysis methodology in a single executable script, pipelines enable a form of *in silico* reproducible research.

The repository maintained by the Broad Institute includes a number of pipelines that document analysis methodologies published by Broad researchers. For a list of the pipelines in the repository maintained by the Broad Institute, with links to their documentation, see the Modules page of the GenePattern web site.

Suites

Suites group modules and pipelines into convenient packages. For example, if you tend to work with a particular set of modules, you might find it helpful to create a suite that contains those modules. The suite provides easy access to those frequently accessed modules. Or, if you have an analysis methodology that you want to share with colleagues, you might create a suite that contains the pipelines and modules that define that methodology. You can export the suite and its modules to a zip file, which you send to your colleagues. Your colleagues can then use the zip file to install the suite and its modules on their own GenePattern servers.

The repository maintained by the Broad Institute includes a number of useful suites. For a list of the suites in the repository maintained by the Broad Institute, see the Suites page of the GenePattern web site.

GenePattern Platform

GenePattern has a client/server architecture designed for flexibility and ease of use:

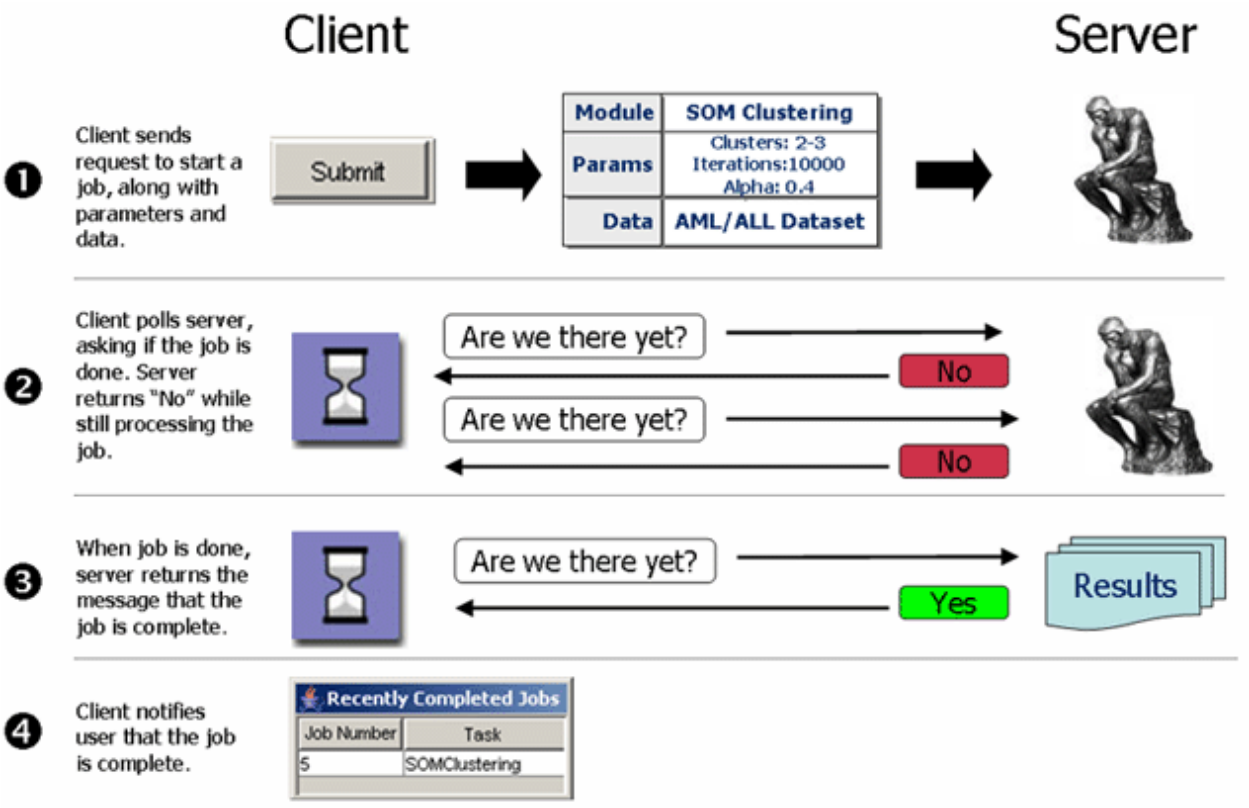
- The **server** is the GenePattern engine. Modules, pipelines, and suites are installed on the GenePattern server.
- The **client** is the interface that you use to interact with the GenePattern server. The GenePattern client has access to the modules, pipelines, and suites installed on the GenePattern server. Most analyses are run on the GenePattern server, rather than on the client.

This section describes the GenePattern platform:

GenePattern Workflow

Modules and pipelines are stored on the GenePattern server. As shown in the following figure, you use the GenePattern client to run analyses on the GenePattern server:

1. From the client, you select the module or pipeline that you want to use, specify the necessary parameters, and run the analysis. The client sends your request to server, where the analysis will be run with the specified parameters.
2. After sending the request, the client periodically polls the server asking whether the analysis has finished running. While the server works on the analysis, it returns “No” to the client.
3. When the server has finished running the analysis, it returns “Yes” when polled by the client.
4. The client notifies you that the analysis has finished. You can then view the analysis results, which are in temporary storage on the server.



GenePattern Clients

GenePattern provides three different user interfaces (three different clients): the Web Client, the Desktop Client, and the programmatic interfaces. The Web Client is a graphical user interface available from any web browser. The Desktop Client is a graphical user interface written in Java, which you can install on your desktop computer. The programmatic interfaces provide support for software programmers interested in calling GenePattern modules from their applications, calling GenePattern modules from the Java, MATLAB, or R programming environments, or writing modules for GenePattern.

Typically, you use the Web Client or the Desktop Client. Most GenePattern operations are available from both clients; however, each client offers unique functionality, as described below. If you are interested in the programmatic interfaces, see the GenePattern Programmer's Guide.

- The Web Client provides administrative operations, which are not available in the Desktop Client. You must use the Web Client to do the following:
  - Modify GenePattern server settings
  - Install modules, pipelines, and suites from the Broad repository
  - Create modules
  - Delete modules, pipelines, and suites
  - View analysis jobs other than your own
- The Desktop Client provides a few ease-of-use features, which are not available in the Web Client. In the Desktop Client, you can do the following:
  - Add project directories for easy access to your dataset files
  - Run an analysis on every file in a directory by specifying that directory as an input parameter
  - Apply filters, based on suites, to hide modules and pipelines that you do not need

The following table summarizes the primary similarities and differences between the Web and Desktop Clients:

	Modules	**Pipelines**	**Suites**	**Server**
Run	Web Desktop	Web Desktop	--	--
Create	Web	Web Desktop	Web Desktop	--
Install from repository	Web	Web	Web	--
Install from zip	Web Desktop	Web Desktop	Web Desktop	--
Export to zip	Web	Web Desktop	Web Desktop	--
Delete	Web	Web	Web	--
Modify server settings	--	--	--	Web
View jobs	--	--	--	Web (all jobs) Desktop (your jobs)

GenePattern Jobs

When you run a module or pipeline, the GenePattern client sends your request to the GenePattern server. The server starts a job to run the analysis. If you are running a pipeline, all of the modules/pipelines in that pipeline are run as part of one job. Job results (analysis results files and execution logs) are stored on the GenePattern server for a period of time (by default, one week) and then deleted. From the client, you view the job results stored on the server.

Note: When you run a visualizer module, the client runs the module. The client does not send a request the server, the server does not start a job, and no job results are stored on the server.

Every job run on the GenePattern server has an owner and is persistent:

- The owner of the job is the person logged into the GenePattern client that submitted the job. An owner is identified by the GenePattern username used to log into the client.
- Every job is persistent, which means:
  - Jobs run independently of the client. Once you start a job on the server, the job continues to run, even if you exit from the client that started the job. The next time you start the client, the client shows you the status of any jobs that were running when you exited from the client.
  - Jobs are restarted if necessary. If the GenePattern server is shut down or interrupted while executing a job, the next time you start the server, the server restarts the interrupted job.

GenePattern Security and Permissions

GenePattern provides a flexible architecture that allows a server administrator to control access to the GenePattern server in three ways:

- Access filtering defines which computers have access to the GenePattern server. If you are running a GenePattern client on a computer that does not have access to the GenePattern server, you cannot use GenePattern.
- User authentication defines who can log into the GenePattern server. Depending on how your GenePattern server is configured, you might need a username (the default) or a username and password. In rare cases, a GenePattern administrator might have implemented additional user authentication requirements. If you cannot log into the GenePattern client, you cannot use GenePattern.
- User permissions define which operations you can perform. If your username is assigned all permissions, you can perform all operations. If your username is assigned limited permissions, you will not be able to perform certain operations. Following are a few examples of operations that require permissions:
  - You can always view and delete jobs that you have run. However, to view and/or delete jobs run by other users requires adminJobs permission.
  - You can always run public modules. However, to create or install modules requires createModule permission. To view and/or delete private modules owned by other users requires adminModules permission.
  - You can always view public suites. However, to create or install suites requires createSuites permission. To install a suite that contains new modules (modules not yet installed on your server) requires both createSuites and createModules permissions because you are installing both suites and modules.

By default, when you install the GenePattern server, all computers have access to the server, only a username is required to log into a client, and all users have all permissions. For more information about modifying server security and for a complete list of permissions, see Securing the Server in the *Web Client Guide*.

The GenePattern documentation describes all of the GenePattern operations. If you see user interface elements (such as menu commands and buttons) in the documentation that are not available the Web Client or Desktop Client, or are unable to perform an operation that appears to be available to you, you most likely do not have the permission required for that operation. Ask your GenePattern server administrator which permissions you have been assigned.

Version Numbers

GenePattern uses Life Science Identifiers (LSIDs) to uniquely identify objects, such as modules, pipelines, and suites. When you create an object, GenePattern automatically assigns the object an LSID with a version number of one (1). When you update an object, GenePattern automatically updates the version number of the object's LSID. As an example, consider pipeline version numbers.

When you create a pipeline, GenePattern assigns it a new LSID. If you modify that pipeline, GenePattern increments the version number of its LSID. Similarly, each module within the pipeline has its own LSID. If you modify the module, GenePattern increments the version number of its LSID. The module in the pipeline, identified by the original version number, is not changed. By using LSIDs, GenePattern preserves the originally created pipeline and every modified version of that pipeline. This careful versioning process allows you to accurately reproduce any previous analysis, even if you have modified the pipeline used for that analysis.

When you view and edit pipelines, you can see the pipeline's version number (the version number of the pipeline's LSID). Typically, you update the latest version of a pipeline, which increments its version number. For example, editing version 1 of a pipeline creates version 2 of that pipeline. At times, you may need to edit an older version of a pipeline, which creates a "point version" of that pipeline. For example, if you have versions 1 and 2 of a pipeline, editing version 1 of the pipeline creates version 1.1 of that pipeline.