Gigwa v2 – Documentation

A/ USER DOCUMENTATION

A1/ IMPORTING DATA

Choosing "Manage data" then "Import data" from the main horizontal menu leads to a page dedicated to data imports.

Anonymous users and users with no particular permissions are limited to importing genotyping data into temporary databases that remain accessible for 24h. These databases are hidden (only visible to people knowing their precise URL and to administrators).

In the case where the importing person is logged in as administrator or as a user with permissions to write into a permanent database, a tabset allows choosing between importing genotyping data or metadata. Providing individual metadata is only supported for existing databases and aims at enabling users to select them by filtering on that metadata. This is convenient for cases where the individual list is long and / or individual names are not meaningful.

Genotyping data may be provided in various formats (VCF, HapMap, PLINK, BrAPI) and in various ways:

- By specifying an absolute path on the webserver filesystem (convenient for administrators managing a production instance used as data portal);
- By uploading files from the client computer (with an adjustable size limit: see section B4.2);
- By providing an http URL, linking either to data files or to a BrAPI v1.1 base-url.

Genotyping data import progress can be watched in real time from the upload page, or via a dedicated asynchronous progress page (convenient for large datasets), as imports run as background processes and may not be interrupted by users.

A2/ WORKING WITH GENOTYPING DATA

From the home page, select a database and a project. Note the presence of an "Enable browse and export" checkbox that toggles between a mode where only variant counts are displayed, and one where users may browse, visualize and export selected data.

A2.1/ GENERAL FILTERING FEATURES

By default, a single grey filter panel appears, providing means to select variants based on their inherent attributes:

- variant type (i.e. SNP, INDEL…);
- sequence;
- position;

- if applicable, number of known alleles (when several are represented);
- if applicable, functional annotations (if the data was provided as a SnpEff or VEP-annotated VCF file).

As a general rule, a filter widget where no selection has been made will behave as if all its items were selected (no filtering applied on that field).

A2.2/ GROUP COMPARISON

At the bottom of this panel, a dropdown button may be used to display one or two additional panels that allow for more advanced filtering features based on a subset of individuals either in one or two distinct groups.

Both of these panels have the same contents and each one lets users select a list of individuals (thus defining groups 1 and 2). Additionally, some handy tools are available on the right side of the “Individuals” drop down menu as tooltipped icons:

- Activating the **disk** icon allows the current selection of individuals to be memorized within the web browser;
- The **magnifier** icon (only present when metadata are available) helps selecting individuals according to the metadata attached to them;
- The copy **icon** adds currently selected individuals to clipboard
- Clicking the **paste** icon opens a textbox for pasting / editing a list of individuals to select.

All other widgets in the same panel will then let users apply genotype-level filters on the selected list of individuals:

- In the case where data was provided in VCF format containing numeric genotype-level fields (e.g. depth, genotype quality), a minimum acceptable value may be provided for each of them. Any genotype not respecting thus defined constraints is treated as missing for the rest of the query;
- “Max missing data” ratio defines how many individuals (among the selected ones) may have a missing genotype;
- “Minor allele frequency” can be provided as a range (only supported for bi-allelic data);
- The “Genotype pattern” dropdown provides a list of genotype patterns that may be applied within the group.

Useful tip: To identify variants for which genotypes are steady within each group but different between them, set both groups’ genotype pattern to “All or mostly the same”. In this case, in each group panel a similarity ratio lets users specify how many of the current group’s selected individuals must have the same genotype. Additionally, an extra pink panel appears between both group panels, containing a checkbox labelled “Discriminate groups”. Checking this box will ensure that the most frequent genotype in group 1 is different from that in group 2.

Note that genotype-level filters are applied in the order they appear: the maximum missing data filter applies first, taking into account truly missing data and genotypes treated as missing because of low quality. MAF and genotype pattern queries are then applied at the same time, on the remaining (non-missing) genotypes only.

If the “Enable browse and export” box is ticked when clicking the Search button, then one may browse online the selection (i.e. list of variants that match the query). Clicking a variant line opens a dialog providing variant details along with individuals’ genotypes and optional complementary information like quality data or annotations.

Above the variant list:

- A density chart icon leads to a dialog in which the variant distribution may be observed for each sequence represented in the selection. In the case where data was provided in VCF format containing numeric genotype-level fields (e.g. depth, genotype quality) an additional series can be displayed for each of these fields, on top of the default density series;
- A download button opens a panel where users may select an output format, refine the list of individuals to export, and choose between directly downloading the output, or creating a file on the server (in which case its URL may be used later, shared or passed to external tools);
-

 An “External tools” box provides means to setup the application for pushing data into external tools: an online genome browser can be configured for viewing each variant in its genomic context (via an extra icon at the end of each table row); a running instance of IGV can be fed with a VCF export file (refer to tooltip for details); online tools (e.g. Galaxy) can also be fed using exported files (click online-output-tools icon for details).

A3/ WORKING WITH REST APIs

Any data imported into Gigwa is automatically interfaced via two standard REST APIs, documented in a Swagger page available from the main menu:

The GA4GH v0.6.0a5 implementation has by design a single base-url. Listing available databases can be achieved by posting an empty body to /rest/ga4gh/referencesets/search. Thus obtained values can be then passed to other calls as referencesetId or datasetId. GA4GH’s VariantSet, Reference and CallSet concepts respectively correspond to Gigwa’s project, sequence and individual entities.

The BrAPI v1.1 implementation has by design a separate base-url for each database, constructed as follows: /{database}/brapi/v1/token. Each database’s BrAPI base-url can be deducted from the mentioned /rest/ga4gh/referencesets/search call’s response, and by convenience, the main Gigwa interface provides a link to the corresponding BrAPI base-url when a new database is selected.

Please refer to http://ga4gh-schemas.readthedocs.io/en/latest/ and https://brapi.org/ for more details about each API.

B/ ADMINISTRATOR DOCUMENTATION

By default, a fresh instance of Gigwa comes with a single pre-defined administrator account (login: gigwadmin, password: nimda). It is of course strongly advised to **change this password upon first connection** (see section B2 below).

B1/ TOMCAT CONFIGURATION

Ready-to-use bundled packages should not require any changes in Tomcat configuration since the settings below have already been applied to them. However, if you install Gigwa in a production environment from fresh Tomcat binaries, it is necessary to apply the following modifications:

- The bin/setenv.bat or bin/setenv.sh (depending on the platform) script must contain a line as follows in order to reserve enough RAM for Tomcat:
  set "JAVA_OPTS=%JAVA_OPTS% -Xms512m -Xmx2048m"
  (This setting may of course be adapted to get the best out of the hardware configuration)

- In the conf/server.xml file, the main Connector element must be configured as below:
  maxHttpHeaderSize="65536" maxParameterCount="-1" maxPostSize="-1"

B2/ MANAGING DATA

The visibility of a database is defined using two flags (default values in **bold**):

- public / **private:** if public, anyone (even anonymous users) can search this database; if private, only administrators and users who were explicitly granted permissions may do so;

- hidden / **exposed**: if hidden, only administrators will see the database in the main menu list; if exposed, any entitled user (that is, anyone if database is public, otherwise any user with permissions on at least one of the database’s projects) will see it in the list.

In other words, the first flag defines visibility on the server side while the second defines exposure on the client side. Typically, a temporary database created by an anonymous user or a user without any management permissions will be public and hidden (searchable by anyone, listed to the administrators only), and will be made accessible to its creator via a specific URL referring to the database name (thus accessible to anyone if shared).

Administrators can see all databases (even if private and/or hidden) and have all privileges on them. Only administrators may create permanent databases. This can be done either at import time, or via the main menu’s “Manage data” link, by subsequently clicking on “Manage databases”. A simple interface allows then to create an empty database on a selected MongoDB host, set the public and hidden flags on existing databases, and delete existing projects and databases.

B3/ MANAGING USER ACCOUNTS AND PERMISSIONS

By choosing the “Administer existing data and user permissions” link from the “Manage data“ menu item and subsequently clicking on “Manage users and permissions”, administrators may access an interface for creating / deleting users, setting their password (even their own), and setting their permissions:

- at the database level by allowing to import new projects into it (any user importing a new project will be automatically granted the MANAGER role on it);

- at the project level by granting either the READER role (only makes sense for projects in private databases) which allows to search this project's data, or the MANAGER role which allows to search project data, import metadata, and grant roles to existing users on that project.

Indeed, a user with the MANAGER role on a project can administer that project in the same way as an administrator, via the main menu's "Manage data" menu item also available to him after authentication.

B4/ CONFIGURING ADVANCED SETTINGS (FOR SYSTEM ADMINISTRATORS: REQUIRES WRITE PERMISSIONS ON INSTALLED FILES)

Although a Gigwa instance installed via a distribution package is functional out of the box, some configuration settings can only be adjusted by editing text files. Most of them only need to be set once.

B4.1/ Managing data hosts

Declaring MongoDB hosts is done via the WEB-INF/classes/applicationContext-data.xml file following provided examples. Only hosts running with authentication enabled (refer to MongoDB documentation if needed) must be declared along with a UserCredentials bean. Note that Gigwa associates them internally using their IDs: for example, a host named myMongoHost will expect a UserCredentials bean named myMongoHostCredentials. Those credentials must be provided for a user declared in MongoDB's admin collection, who has readWriteAnyDatabase and dbAdminAnyDatabase roles. The web-application needs to be reloaded for such changes to be taken into account (please refer to Tomcat documentation if needed).

B4.2/ Setting configuration properties

The WEB-INF/classes/config.properties file may be used to set values for the following parameters:

- **dbServerCleanup** - You may specify under this property, a csv list of hosts for which this instance will drop temporary variant collections on startup (e.g. 127.0.0.1:27017, another.server.com:27018). Temporary variant collections are often used once a search has been completed, for browsing/exporting results. They are normally dropped upon user interface unload, but some may remain if the web-browser is exited ungracefully or the application goes down while someone is using the search interface. If this property does not exist then the instance will drop all found temp collections, if it exists but is empty, none will be dropped.

- **adminEmail** - If Gigwa is being used as a multi-user data-portal you may specify via this property an email address for users to be able to contact your administrator, including for applying for account creation.

- **igvDataLoadPort** - Defines the port at which IGV listens for data submission. No IGV connection if missing / invalid.

- **igvGenomeListUrl** - Defines the URL from which to get the list of genomes that are available for IGV. No IGV connection if missing / invalid.

- **sessionTimeout** - Web session timeout in seconds. Default: 3600 (1h)

- **forbidMongoDiskUse** - MongoDB's allowDiskUse option will be set to the opposite of this parameter's value when launching aggregation queries. Default: false

- **tempDbHost** - Tells the system which MongoDB host to use when importing temporary databases (for anonymous users). Only used when several hosts have been configured in applicationContext-data.xml. If unspecified all connected hosts will be available for use. If invalid, no import will be possible for users without specific permissions.

- **maxUploadSize_anonymousUser** - Defines the maximum allowed size (in megabytes) granted to anonymous users for data file upload. Default: 500Mb

- **maxUploadSize_USERNAME** - Defines the maximum allowed size (in megabytes) granted to the USERNAME user for data file upload. Default: 500Mb

- **serversAllowedToImport** - CSV list of external servers that are allowed to import genotyping data.

- **genomeBrowser-MODULE_NAME** - Any property named genomeBrowser-MODULE_NAME is a way for defining a default genome browser URL for a module called MODULE_NAME. This is optional as users may define their own genome browser URL, thus overriding the default one if it exists.

- **onlineOutputTool_N** - Any property named onlineOutputTool_N with N being an integer >= 1 is a way for defining an online output tool for datasets exported to server. N accepts consecutive values (if only onlineOutputTool_1 and onlineOutputTool_3 exist then only onlineOutputTool_1 will be taken into account). The property value must consist in semi-colon-separated values. The first one is the label to display for this tool, the second one is the tool URL (in which any * character will be replaced at run time with the exported file’s URL). The third value is optional and may contain a comma-separated list of file-formats (must match some of those that the Gigwa instance is able to export: BED, DARWIN, EIGENSTRAT, FLAPJACK, GFF3, HAPMAP, PLINK, VCF), thus defining those accepted by the tool (if unspecified, files in any format will be made available for this tool).

- **maxSearchableBillionGenotypes** - Defines the maximum estimated size (in billions) of the genotype matrix (#individuals * #markers) within which genotype-level filters may be applied. This property may be tuned according to server performance. #markers is estimated by calculating an average marker count per sequence. Whatever value is set here, Gigwa will at least allow searching on one sequence for all individuals. Default: 1 billion

- **maxExportableBillionGenotypes** - Defines the maximum size (in billions) of the genotype matrix (#individuals * #markers) that may be exported. This property may be tuned according to server performance. It aims at limiting system overhead in situations where numerous users may be working on very large databases. Default: 1 billion

- **googleAnalyticsId** - If set, a Google Analytics tag is automatically added into the main page.