

Bioinformatics Integration Support Contract (BISC), Phase II

SYSTEM ARCHITECTURE AND SOFTWARE DESIGN SPECIFICATION



Version 4.0

Period Of Performance: September 30, 2004 to December 31, 2011

Developed Under Contract Number: HHSN266200400076C

ADB Contract Number: N01-A1-40076

Delivered: November 15, 2011

Project Sponsor:

National Institutes of Health (NIH)

National Institute of Allergy and Infectious Diseases (NIAID)
Division of Allergy, Immunology, and Transplantation (DAIT)

Prepared by:

NORTHROP GRUMMAN

Northrop Grumman Corporation
Information Systems Sector
Health IT
2101 Gaither Road, Suite 600
Rockville, MD 20850
(301)527-6400 Main
(301)527-6401 Fax
Jeff.Wiser@ngc.com

Contents

1.0 Introduction	7
1.1 Scope	7
1.2 Purpose	7
1.3 Key Objectives	7
1.4 Identification.....	8
1.5 References	8
2.0 System Architecture	9
2.1 Overview	9
2.2 Hardware Architecture	9
2.3 Software Architecture.....	13
2.4 Hardware and Software Relationship	15
2.5 ImmPort System Security	16
3.0 ImmPort Database Architecture and Design.....	18
3.1 Overview of the ImmPort Database Architecture	18
3.2 Overview of the ImmPort Database Design and Content.....	20
4.0 Research Data Models	26
4.1 ImmPort Administration.....	26
4.2 Research Projects and Collaborative Projects, and the Semi-Public Workspace	27
4.3 Experimental Research Data	29
4.4 Lookup Tables	33
4.5 Analysis Tools Tables - Update for Added SFVT.....	38
4.6 Clinical Data Model.....	46
5.0 Reference Data Models.....	50
5.1 NCBI Entrez Gene and Taxonomy Information.....	51
5.2 Genomic Sequence and Gene Structure Information	52
5.3 Pathway Information	53
5.4 Basic Protein Information.....	55
5.5 Protein Interactions.....	56
5.6 Protein-Protein Interaction Network	58
5.7 SNP Polymorphism Data.....	59
5.8 MHC and Microsatellite Polymorphism Data	60
5.9 Ontology Data Models.....	62
6.0 Software Design Overview	63
6.1 Summary of Capabilities	63
6.2 Product Perspective	64
6.3 Software Design Presentation.....	64
6.4 User Communities	64

6.5 Project Roles.....	66
7.0 Manage User Design Packages.....	67
7.1 Create/View/Update/Search User.....	67
7.2 Users Registration	70
7.3 Approve/Reject User	74
7.4 Deactivate User	76
7.5 Activate User	78
7.6 Search Users	80
7.7 Forgot Password	82
7.8 Log in/Off.....	84
7.9 User Administration Class Diagram.....	85
8.0 Manage Program Design Packages	86
8.1 Create Program.....	86
8.2 Update Program.....	88
8.3 Search Program	90
8.4 Program Administration Class Diagram.....	92
9.0 Manage Contract Design Packages	93
9.1 Add/Update Contract.....	94
9.2 Search Contract	97
9.3 Contract Administration Class Diagram.....	99
10.0 Manage Project Design Packages – Update for Admin process	100
10.1 Add/Update Projects.....	100
10.2 Adding/Updating A Project Class Diagram	102
10.3 Display Projects.....	104
10.4 Project Administration Class Diagram	106
11.0 Data Submission Design Packages.....	107
11.1 Data Submission Workflow Overview	107
11.2 Data Submission Templates	107
11.3 Data Submission Validator.....	109
11.4 Data Submission History	110
11.5 Data Submission Processing Workflow	111
11.6 Data Submission Class Diagrams.....	113
11.7 Study Design and Clinical Data Submission	118
12.0 Advanced and Basic Search Design Packages.....	122
12.1 General Overview.....	123
12.2 Presentation Layer.....	126
12.3 Form Processing, Business Object Layer	126
12.4 Query Construction	127

12.5 Query Processing – Results Display.....	127
12.6 Search Class Diagrams	127
13.0 Research Data Detail Page Design Packages	132
13.1 Class Diagram – Subject Detail.....	134
13.2 Class Diagram – Subject Panels	135
13.3 Sequence Diagram.....	136
13.4 Table of Entities and Panels	137
14.0 Research Data Management Design Packages	139
14.1 Displaying the Edit Screen	139
14.2 Update Subject.....	142
14.3 Delete Subject.....	143
14.4 Add Treatment.....	144
14.5 Delete Treatment	145
14.6 Update Treatment	146
14.7 Display Non-Associated Entities.....	147
14.8 Associate Entities	150
14.9 Disassociate Entities	150
14.10 Edit Subject Class Diagram	152
15.0 List Design Packages	153
15.1 Saving Lists	153
15.2 List Management.....	156
15.3 List Detail	158
15.4 List Delete	160
15.5 List Management Class Diagram	161
16.0 Research and Clinical Data Sharing Design Packages	162
16.1 Associate Entities	165
16.2 Process.....	167
16.3 Share Project Class Diagram	168
17.0 Reference Data Design Packages	169
17.1 Data Sources for Reference Data.....	169
17.2 Gene.....	171
17.3 Protein.....	177
17.4 SNP	182
17.5 Pathway	187
17.6 Protein Networks	192
17.7 MHC Allele	197
17.8 MHC Sequence Feature Structure Viewer (MSFSV).....	213
18.0 Flow Analysis Design Packages	220

18.1 Data Management: File Upload.....	221
18.2 Data Management: Create Analysis Set	223
18.3 FLOCK Analysis	225
18.4 Cross Sample Analysis	227
18.5 FLOCK Analysis Results	230
18.6 Cross Sample Analysis Task Results.....	233
18.7 FLOCK Results Visualization	236
18.8 Flow Analysis Class Diagrams.....	239
19.0 HLA Analysis Design Packages	243
19.1 Upload an HLA File From Your Computer	244
19.2 Create a an HLA Data Set From ImmPort Data.....	247
19.3 Re-Analyze HLA Data Set Previously Created in ImmPort.....	252
19.4 Run SFVT Analysis.....	254
19.5 Viewing Pipeline Results.....	258
19.6 HLAVT Pipeline	262
19.7 HLA Analysis Class Diagrams.....	264
19.8 HLA VT Pipeline Code Base Description.....	266
20.0 GenePattern	268
20.1 GenePattern Architecture	269
20.2 GenePattern Analysis Submission.....	271
21.0 PED File Generation Design Packages	277
21.1 Choose Subject Lists	277
21.2 Subject Validation Results.....	281
21.3 SNP Selection.....	285
21.4 PED Generation Results	289
21.5 PED File Generation Class Diagrams.....	291
22.0 Haplovew Design Packages	295
22.1 Submit Haplovew Analysis	295
23.0 TagSNP Design Packages	302
23.1 TagSNP Analysis Gene	303
23.2 TagSNP Analysis Chromosome	305
23.3 Analysis History	307
23.4 Analysis Result	308
23.5 TagSNP Class Diagram	310
24.0 GBrowse	311
25.0 Ontology Browser	312
25.1 UI Overview	312
25.2 Ontology Class Diagram	314

25.3 Sequence Diagram.....	315
----------------------------	-----

BISC System Architecture and Software Design Specification Version History

Version	Date	Description
1.0	4/29/2005	Initial release of the BISC Phase II System Architecture and Software Design Specification
1.1	6/1/2005	Completed Sections 3, 11, 14, 15, and 16, including addition of data models to Section 3. Updated sections 12 and 13. Added Section 17. Added acronym list to Appendix A. Added preliminary design and storyboards for future ImmPort versions to Appendix B.
1.2	12/28/2005	<p>Revised the Purpose section to note that most requirements for the ImmPort system version 1.0 were derived primarily from working with the Population Genetics Program (Section 1.2).</p> <p>Updated assumptions (Section 1.3).</p> <p>Updated references (Section 1.6).</p> <p>Updated descriptions and diagrams of hardware and software architecture (Section 2).</p> <p>Updated the database design to reflect the current database instances and servers across the development environment, test environment, and production environment (Section 3).</p> <p>Updated the previous Experimental and Reference Data Models Description: Supplement to System Architecture and Software Design Specification Version 1.1 and incorporated it into the sections on experimental data and reference data models (Sections 4 and 5, respectively).</p> <p>Added an ImmPort Ontology Browser schema (Section 5).</p> <p>Made minor changes to correct design diagrams in the following sections: 7, 8, 9, 10, 14, 16, and 17.</p> <p>Updated prototypical mock-ups to ImmPort system screenshots in the following sections: 7, 8, 9, 10, 11, 13, 14, 16, and 19.</p> <p>Expanded reference data query use cases and corresponding screenshots (Section 16).</p> <p>Removed preliminary design artifacts related to an upload wizard approach to experimental data (SASDS v.1.1, Appendix B).</p>
2.0	6/1/2007	<p>Updated section 1.2 Purpose - includes updated information regarding new sections and update details.</p> <p>Updated section 1.4 Key Objectives - updated relevant version information and clarification of objectives.</p> <p>Updated section 2.1 System Architecture Overview.</p>
2.2	2/6/2008	<p>Updated section 5.0 Reference Data Models.</p> <p>Updated section 3.2.1 Overall Database Design - provided more detail.</p> <p>Updated Table 3-1 ImmPort Database Schemas.</p>
3.0	10/29/2010	Re-factored the majority of the document to reflect updated data model and definitions, as well as updating the software design section to reduce redundancy and add new functionality.
4.0	11/01/2011	Updates to reflect changes in admin, flow cytometry, SFVT (Section 19) areas

1.0 INTRODUCTION

1.1 SCOPE

The scope of the Bioinformatics Integration Support Contract (BISC) Phase II is to provide advanced information technology support in the production, analysis, archiving, and exchange of scientific data for a diverse community of life science researchers.

1.2 PURPOSE

The Immunology Database and Analysis Portal (ImmPort) system has been developed by a team led by Northrop Grumman Information Systems (NGIS) in partnership with the science team from the University of Texas-Southwestern (UTSW) through the BISC Phase II contract. The ImmPort system is intended to serve as a long-term, sustainable archive of data generated by investigators funded through the Division of Allergy, Immunology and Transplantation (DAIT) of the National Institute of Allergy and Infectious Disease (NIAID), National Institutes of Health (NIH). The ImmPort system consists of an extensive data warehouse containing an integration of experimental and clinical data supplied by NIAID/DAIT-funded investigators and reference data extracted from a variety of public databases. Additionally, the resource hosts both internally and externally developed best of breed open source data analysis and data set generation tools identified as useful by Immunology researchers. The ImmPort system is freely accessible as a resource to all investigators funded through NIAID/DAIT as well as any other DAIT-approved scientist in the research community.

This System Architecture and Software Design Specification (SASDS) defines the overall ImmPort architecture and software design specification identified by the NGIS BISC Team (hereinafter referred to as the BISC Team) for the ImmPort system that has been developed for NIAID/DAIT. The architecture and design described in this document focuses on the capabilities that are implemented in the ImmPort system as of version 2.12 in November 2011.

1.3 KEY OBJECTIVES

The key objective of the SASDS version 4.0 is to provide an update to the hardware specifications of the system, provide the database table models used in functional areas systems, and to provide a design specification for the full extent of functionality now completed in ImmPort.

Database Entity Relationship Diagrams (ERD) are provided in the context of the primary functional area in which the tables are utilized. Tables may appear in multiple diagrams where appropriate. A full ERD diagram may be provided on request in a different form, since the breadth of the database schema cannot be clearly illustrated in document form.

In the software design packages, prospective use case diagrams from earlier versions of the document have been replaced with screen shots from implemented interfaces and descriptions and packaged with the applicable sequence diagrams and class diagrams. A new tool has been utilized for sequence diagram generation that has been found to be more cost effective and easier to maintain due to its scripting capabilities. Additionally, overarching class diagrams representing code for the presentation (Struts Actions), business model(EJB) and persistence (DAO + Hibernate) layers have been included to provide a context for the reduced class diagrams used to illustrate method calls utilized in that particular software feature.

This document is to be considered a “work in progress” and will evolve during the life of the BISC effort as additional requirements are implemented, new requirements are identified, and others are modified or deleted. The largest impending change to the ImmPort architecture involves the migration of the ImmPort application to hardware at a facility run by the Office of Cyber Infrastructure and Computational Biology

(OCICB) within NIAID. This transfer is in progress and will be completed by the end of 2011. As such, some aspects of the hardware and software architecture will need to be modified following the migration to ensure accuracy. Notes will be made throughout the document highlighting where there may be upcoming changes.

1.4 IDENTIFICATION

Northrop Grumman IT provides development of the BISC system (ImmPort) under Contract No. HHSN266200400076C, ADB Contract No. N01-A1-40076.

1.5 REFERENCES

- Northrop Grumman IT Contract No. HHSN266200400076C, ADB Contract No. N01-A1-40076

2.0 SYSTEM ARCHITECTURE

2.1 OVERVIEW

The ImmPort system is implemented as a n-tier distributed architecture, providing a scalable, pluggable, and reusable componentized solution where additional layers or components can be added to accommodate current and future needs. The architecture is comprised of BigIP F5 hardware load balancers, a pair of application server nodes running the Redhat Enterprise Linux 5 OS, a 2-node Real Application Cluster (RAC) Oracle database, a Network Attached Storage (NAS) and a 3-node grid computing environment using the open source Sun Grid Engine (SGE).

The software organization is implemented using a conventional 3-tier approach, which decouples software components into three major areas: the presentation, the business, and the persistence layers. Each layer specializes in specific orchestrated functions. The presentation layer separates the Graphical User Interface (GUI) from the complexity of the business rules. The persistence layer abstracts data retrieval and storage from yet another complex layer such as that of a Relational Database Management Systems (DBMS).

The presentation layer utilizes key web technologies, included but not limited to, the Hypertext Markup Language (HTML), Extensible Markup Language (XML), Cascading Style Sheets (CSS), Java Server Pages (JSP), Asynchronous JavaScript and XML (Ajax), Extjs, and the Struts Framework. The business tier leverages the Stateless Enterprise Java Beans (EJBs) and the Message Driven Beans (MDBs) stack of the Java2 Enterprise Edition (J2EE) framework. Using the Facade Object Oriented Design pattern, the EJBs abstracts the complexity of the underling complex business rules. Lastly, the persistence layer manages all database access and storage and it relies on the successful object-relational mapping (ORM) Hibernate library for the Java platform.

The 3-node Sun Grid computing environment is a key layer of the ImmPort system architecture. It is a number crunching environment, where heavy duty, computationally intense jobs are sent to a queue for processing. These jobs or tasks are resource intensive, requiring significant amount of Random Access Memory (RAM) and Central Processing Unit (CPU) time as well as Input/Output (IO) resources, including network traffic. Aspects of this will be disabled in initial installations of the deployment of ImmPort to OCICB, but details will be specified after the migration is complete.

The ImmPort System consists of hardware and software configuration items. Hardware configuration items (HWCIs) and the ImmPort system hardware architecture are described in Section 2.2. Software configuration items (SWCIs) and the ImmPort system software architecture are described in Section 2.3.

2.2 HARDWARE ARCHITECTURE

Table 2-1 lists the ImmPort system hardware configuration items on which the ImmPort system software configuration items will reside. This hardware architecture will change slightly in the migration to OCICB's facility.

Table 2-1. ImmPort System Hardware Configuration Items

Hardware Configuration Item	Description/Purpose
DMZ	In a DMZ configuration, most computers on the LAN run behind a firewall connected to the Internet. One or more computers also run in a DMZ; these are Web and FTP servers that require fewer restrictions to the Internet.

Hardware Configuration Item	Description/Purpose
Firewall	A firewall protects a computer network from unauthorized access. Network firewalls may be hardware devices, software programs, or a combination of the two. A network firewall typically guards an internal network (intranet) against malicious access from outside the firewall as well as unauthorized access to the outside from within the network.
Fibre Channel RAID Array Storage Device	High-speed storage in RAID 5 configuration, including Oracle database storage and raw reference data storage (source data).
Database Servers (node1 and node2)	The bottom-tier servers (node1 and node2) supporting all aspects of the database layer software, as part of a 2-node RAC server configuration. Configuration of both Database Servers are identical.
Application Server	The mid-tier server supporting application layer software.
BigIP F5 Hardware load balancers	Top-tier hardware load balancers responsible for managing and distributing HTTP requests.
Router	A hardware device that forwards data packets along networks . A router is connected to at least two networks, commonly two LANs or WANs or a LAN and its ISP's network. Routers are located at gateways , the places where two or more networks connect. A router is located outside the firewall connecting the Internet with the firewall.
Switch	A network device that controls the flow of traffic between multiple network nodes.
Fibre Switch	A network device that controls the flow of traffic between the two database servers and the backup tape device via fiber optic cable.
Tape Backup	Backup and restore and archive capabilities.
T1 line	Broadband cable for data communication.
Sun Grid Compute Engine (SGE)	A 3-node grid computing environment using the open source Sun Grid Engine. The grid supports the queuing and processing of resource intensive jobs such as identifying significant co-clustering of genes with similar functional properties using the Gene Ontology. The SGE will be disabled on initial deployment at OCICB to be re-instated at a later time.
Network Attached Storage	A file-level hardware storage attached to the network providing data access to heterogeneous client computers on the network.

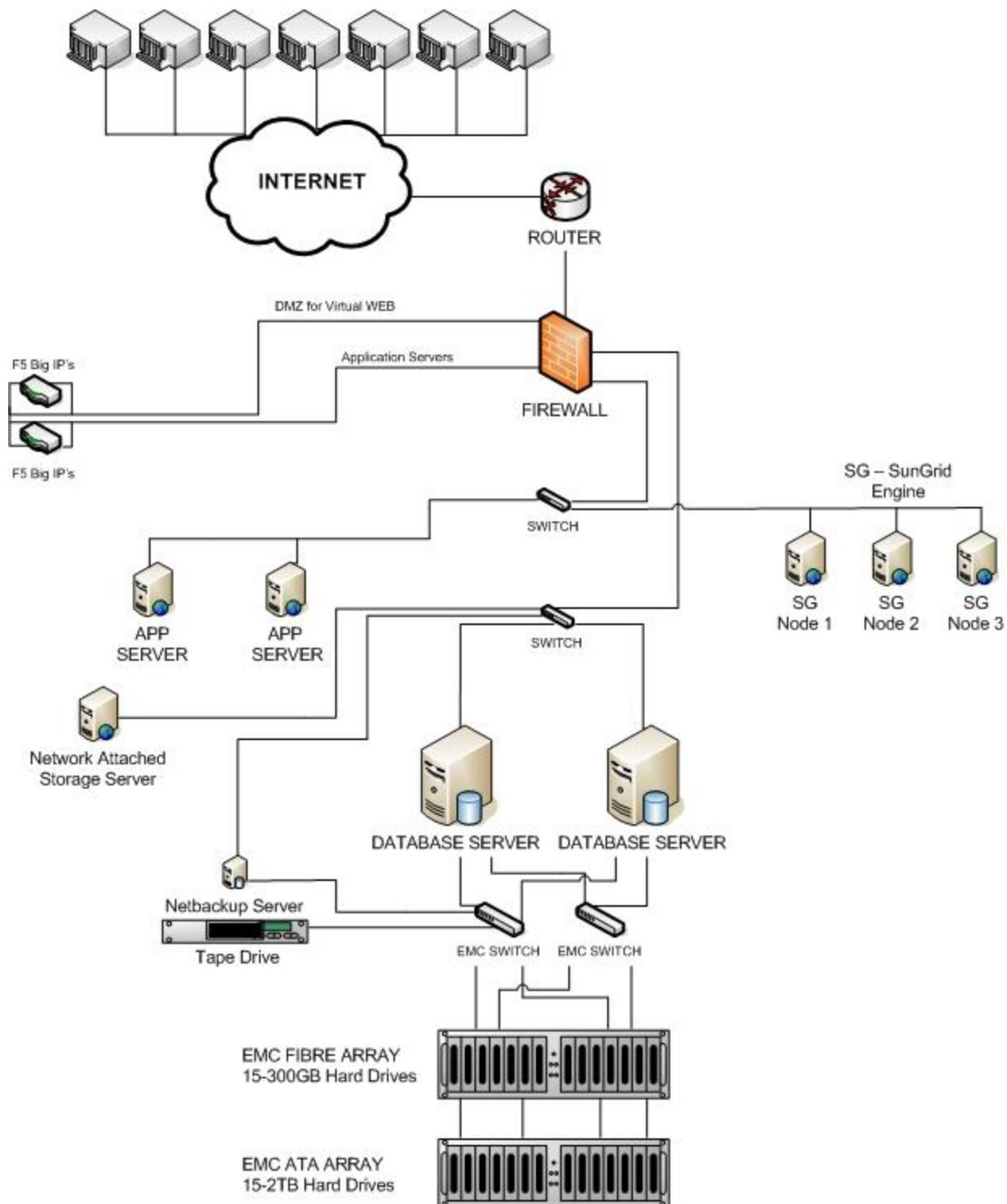
Figure 2-1 shows the ImmPort system hardware and network diagram, illustrating the hardware components at each layer and the relationships of all components. Note that this configuration is based on the current implementation at the Northrop Grumman facility. The intention is to have a similar architecture at OCICB, while making use of the load balancing, data backup, networking, environmental, and storage capabilities already available at NIAID. Vendor differences will need to be fully specified in a later iteration of this document, however there are known differences in vendors for storage (NetApp instead of EMC), firewalls, application servers and database servers (Hewlett Packard instead of Dell) and data backup.

As depicted in Figure 2-1, the ImmPort system is shielded from the Internet cloud by a hardware firewall. The firewall is the first line of defense, providing a filtering mechanism based on the Deny All/Grant Some paradigm. Suspicious or unrecognized traffic is discarded and logged where as valid web site requests are forwarded to the BigIP F5 hardware load balancers. The pair of BigIP load balancers provide a high availability (HA) configuration and transparent traffic distribution. Traffic distribution is based on a Round Robin Least Connections fashion, where requests are forwarded and scheduled based on the load of the underlying application servers. The pair of application servers depicted in Figure 2-1 in turn process the user requests either synchronously or asynchronously. In other words, the response sent back to the user is either immediate or queued for later processing, depending on the anticipated amount of resources needed to carry out a unit of computational work. Generally, analysis jobs such as those involving the identification of significant co-clustering of genes with similar functional properties are

queued for later processing. In this case, the application servers delegate the work to the Sun Grid Engine (SGE) environment depicted in the diagram. SGE is responsible for managing the execution of standalone, parallel or distributed jobs. The Network Access Storage (NAS) server provides a reliable and secure network file server.

The database layer is the bottom layer in the architecture, adding additional levels of security assurance and controls. The database layer is an Oracle Real Cluster Application enabled solution, and consists of dual database servers managing the database and other data on the Fiber Channel RAID array. The Fiber array devices support rapid access, high throughput, and fail-over capabilities in case of disk drive failure. A tape drive is also attached, providing backup and recovery capabilities, including off-site storage.

Figure 2-1. ImmPort Production Network and Hardware Architecture



2.3 SOFTWARE ARCHITECTURE

The ImmPort system software architecture diagram shown in Figure 2-2 depicts all the components comprising the BISC software architecture. Each software component is described in Table 2-2, which follows the figure.

Figure 2-2. ImmPort System Software Architecture

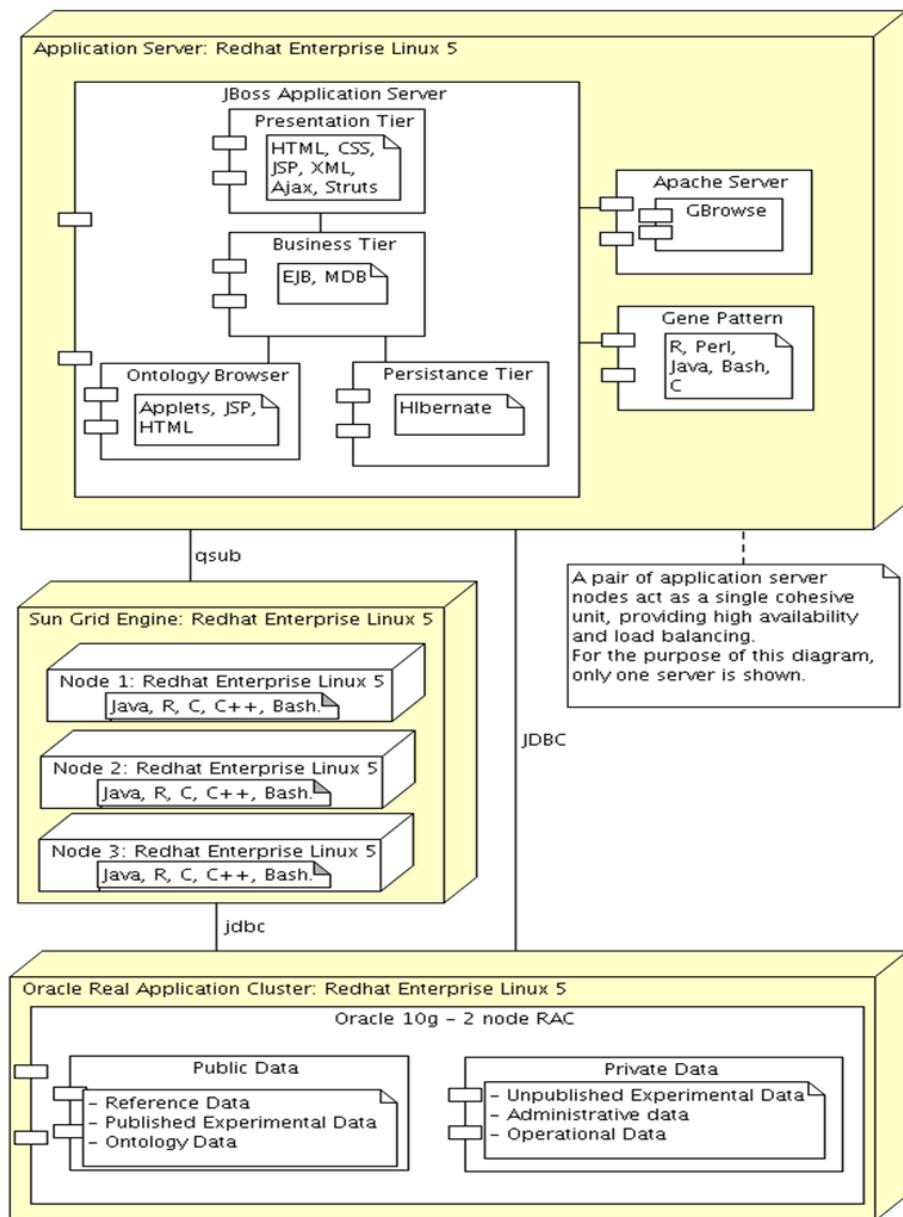


Table 2-2. Software Architecture Components

Component Name	Description
Application Server	The Application Server contains JBoss Application Server instance(s) and an Apache Web Server. The JBoss App Server is responsible for hosting the presentation tier, business logic tier, and the data access tier. For purposes of performance and scalability, multiple application server instances may be utilized to support running a cluster of JBoss application servers. The Apache Web Server hosts the GBrowse application. The GenePattern instance runs outside the JBoss container.
Business Tier	The business tier abstracts the complexity of the business rules and manages the communication between the presentation and database tiers.
Database Servers (2 nodes)	This software component contains two major database areas: Public Data Warehouse and Private Project Workspace. There is also a Raw File Repository for staging and backup/archive purposes. To ensure proper performance and scalability, there will be at least two of these servers running Oracle 10g in Real Application Cluster mode.
EJB Container	The application tier consists of the Enterprise Java Bean (EJB) container and the Java Messaging Service. The EJB container houses the majority of the application logic in terms of data loading/publishing, browsing/querying, journaling, and any administrative/management activities. The JMS component, on the other hand, is responsible for the execution of any application logic that is asynchronous in nature. This helps reduce the load and distribute the work across the network.
GenePattern	The GenePattern application is a powerful genomic analysis platform, providing a number of modules for performing computational tasks involving proteomics, micro array analysis, and SNP analysis.
GBrowse	The Generic Genome Browser (GBrowse) is a popular genome browser that is implemented in Perl. GBrowse displays graphical representations of gene sequences and allows the user to navigate throughout an organism's genome. It uses the Apache mod-perl library to access the database.
Hibernate	The data access tier utilizes the open source Hibernate library for data access. Hibernate provides fine-grained access/update, caching, and transactional persistence to the Private Project Workspace database where transactional support is required.
Java Messaging Service	Java Messaging Service (JMS) is the standard API for sending and receiving messages.
Java Server Pages	Java Server Pages (JSP) is server-side technology for building presentation tier components.
JBoss Application Server	JBoss Application Server is an open-source J2EE application server.
Oracle 10g Database (may be upgraded to 11g on deployment at OCICB)	Oracle 10g is a relational database management system (RDBMS). This software component contains all subcomponents that make-up the server-side RDBMS software. This includes, but is not limited to: Cluster Ready Services (CRS) software, Real Application Clusters (RAC) software, and RAC database server software. The server configuration is a 2-node RAC environment with shared Automatic Storage Management (ASM) utilizing ASMLib software.
Oracle 10g Client (may be upgraded to 11g on deployment at OCICB)	This software component contains all subcomponents and utilities that are installed as part of the Oracle 10g Client software, currently residing on the ImmPort Application Server. This SQL*Plus Client subcomponent is used by the GBrowse application software for establishing a remote client database connection for retrieval of data for the GBrowse Visualization Tool.
Persistence Tier	The persistence tier manages all database access and storage using the open source object relational mapping Hibernate library.
Presentation Tier	The presentation tier manages and renders Graphical User Interface (GUI) components using web technologies such as HTML, CSS, JSP, XML, Ajax, and the Struts framework.

Component Name	Description
Research Project (or Private Project Workspace)	The Research Project (RP) or Private Project Workspace (PPW) contains all types of experimental research data uploaded by ImmPort users for analysis and collaboration. The PPW will contain operational data (such as audit, archive, and session activity logging data), and administrative support data. The ImmPort System Research Project data is comprised of many Research Projects. The ImmPort System allows certain users such as Principal Investigators the ability to create Research Projects, assign users to them, and then allows batch upload of experimental research data into that Research Project. All Research data uploaded and created in the ImmPort System as part of a Research Project is given a unique identifier (PROJECT_ID). Additional Analysis Tool data created on Research Project data is also given the same unique identifier (PROJECT_ID).
Reference Data (or Public Data Warehouse)	The Reference Data or Public Data Warehouse (DW) contains any immunology-related reference data (extracted from other public data warehouses) and any experimental data published by the ImmPort user community. This data is maintained in separate schemas grouped and divided by data source.
Raw File Repository	The Raw File Repository provides a single location to stage any data (experimental or reference) that are about to be published into the DW (directly or via the PPW) or uploaded into a user's PPW.
Struts	Struts is an open-source Model View Controller (MVC) framework for developing J2EE web applications.
Sun Grid Engine (SGE)	A 3-node grid computing environment is implemented using the open source Sun Grid Engine (SGE). The grid supports the queuing and processing of resource intensive jobs such as identifying significant co-clustering of genes with similar functional properties using the Gene Ontology. The Sun Grid will be initially disabled after migration to OCICB with an evaluation made of later reactivating.
Analysis Tool - TagSNP Analysis software code	Software code written that: <ul style="list-style-type: none"> - creates Prettybase file for input into TagSNP. - performs TagSNP analysis using LDSelect algorithm. - re-formats and appends data to original analysis output.
Analysis Tool - Haplovview Analysis software code	Software code that performs Haplovview Analysis. <ul style="list-style-type: none"> - allows uploading of Haplovview dataset files (.ped and .info). - allows creation of Haplovview dataset files (.ped and .info) from ImmPort data. - allows update of Affection Status for ImmPort created of Haplovview dataset files (.ped and .info). - performs Haplovview analysis using of Haplovview dataset files either uploaded or created as input.

2.4 HARDWARE AND SOFTWARE RELATIONSHIP

Each SWCI will be used in conjunction with one or more HWCI. While some HWCI may not be used directly by any SWCI, every SWCI requires at least one HWCI. Table 2-3 identifies the HWCI on which each of the ImmPort system SWCIs will be utilized.

Table 2-3. Software Configuration Item to Hardware Configuration Item Relationship

Software Component Name	Hardware Component Name
Apache Web Server	Web Server
Application Server	Application Server
Business Tier	Application Server
Database Tier (Hibernate)	Application Server
Database Servers	Database Servers (node 1 and node2)
EJB Container	Application Server
Gene Pattern	Application Server

Software Component Name	Hardware Component Name
GBrowse	Application Server
Hibernate	Application Server
Java Messaging Service	Application Server
JavaServer Pages	Application Server
JBoss Application Server	Application Server
Oracle 10g Database	Database Servers (node 1 and node2)
Oracle 10g Client	Application Server
Research Project (or Private Project Workspace)	Fibre Channel RAID Array Storage Device
Reference Data (or Public Data Warehouse)	Fibre Channel RAID Array Storage Device
Presentation Tier (HTML, CSS, JSP, XML, Ajax, Struts)	Application Server
Raw File Repository	Database Servers (node 1 and node2)
Sun Grid Engine (SGE)	Analysis Server
Analysis Tool - TagSNP Analysis software code	Application Server
Analysis Tool - Haplovie Analysis software code	Application Server

2.5 IMMPORT SYSTEM SECURITY

The ImmPort system security is based on industry standards and best practices. It consists of multiple levels of security access controls, including logical and physical security models that mitigates risks and minimizes threats. Logical access controls protects the ImmPort software and related subsystems, while physical security controls protects the tangible assets. Collectively, these controls aligned with business process, policies and procedures strive to persistently meet the three commons goals of Information Systems Security: confidentiality, integrity, and availability (CIA).

In ImmPort, confidentiality is achieved by guaranteeing the disclosure of information or data to only those users of the system who have been properly identified and authorized to have it. By enforcing and validating user names and strong passwords, the ImmPort system validates the identity of a user who claims to be who she or he declares to be. The system then proceeds onto granting the appropriate level of access control, therefore correctly authorizing the user to perform or preclude the user from executing specific system functionality. User credentials (i.e., username/passwords) travel across the wires using strong encryption and proven web technologies such as HTTP over the Secure Sockets Layer (SSL) protocol.

Integrity, on the other hand, is attained in ImmPort by ensuring the data submitted by the users of the system is not tampered with or changed either by accident or malicious activities. Because the data is stored securely on the backend databases, which are behind the firewall and other network devices, the risks are minimized and the integrity of the data is kept as it travels from one layer onto another.

Lastly, the availability of the system is key and part of the ImmPort system security and assurance. ImmPort is implemented using redundancy, where possible, especially in those areas that are most critical. Examples of this include a pair of application server nodes, a pair of database nodes, and a pair of hardware load balancers. At any point, if one of the nodes in each pair goes down, the other node takes

over transparently. In addition, RAID-5 for redundancy is commonly deployed across advanced data and network storages.

To focus on confidentiality, integrity and availability, the ImmPort systems implements logical access controls using hardware and software solutions. On the front line of defense, a combination of hardware firewalls, switches, routers and other network devices protects the internal network from the Internet cloud. Advanced Intrusion Prevention Systems and Intrusion Detection Systems network appliances are in place to identify and block malicious activities, and log and report such activities. A ***Deny All, Allow Some*** set of rules is enforced in multiple layers of the network. This in turn grants processes, services and users access to network resources, while discarding and logging untrusted requests. Valid HTTP requests are furthered filtered using a Role Based Access Control (RBAC) mechanism, which allocates specific roles to users. Each role is assigned one or more privileges. For example, a user who has been correctly authenticated and authorized as a Principal Investigator can perform tasks such as creating a project workspace and assigning users to such project. Furthermore, there are specific security filters that cannot be enforced using traditional roles. In this case, a programmatic security approach is in place to verify business rules such as validating a contract/grant is still active or a user is member of a project.

Physical security controls are also in place, which include monitoring devices such as cameras and temperature and humidity sensors. These sensors trigger alerts by email or other means to network staff when configurable thresholds exceeds desirable limits. The server rooms are equipped with electronic locks and only authorized personnel is allowed access to the server room facilities. After move to the OCICB facility, all of these capabilities will be utilized for ImmPort as is already available to all NIAID computational resources.

3.0 IMMPORT DATABASE ARCHITECTURE AND DESIGN

3.1 OVERVIEW OF THE IMMPORT DATABASE ARCHITECTURE

The ImmPort system database architecture is stored and maintained in an Oracle 10gR2 Enterprise Edition database utilizing Real Application Clusters (RAC), installed on a Linux EL5 operating system. Installed database options include Oracle Partitioning, On-Line Analytical Processing (OLAP), and Oracle Data Mining (ODM) options. The RAC environment provides necessary system load distribution and load balancing, while at the same time providing system redundancy and failover capabilities.

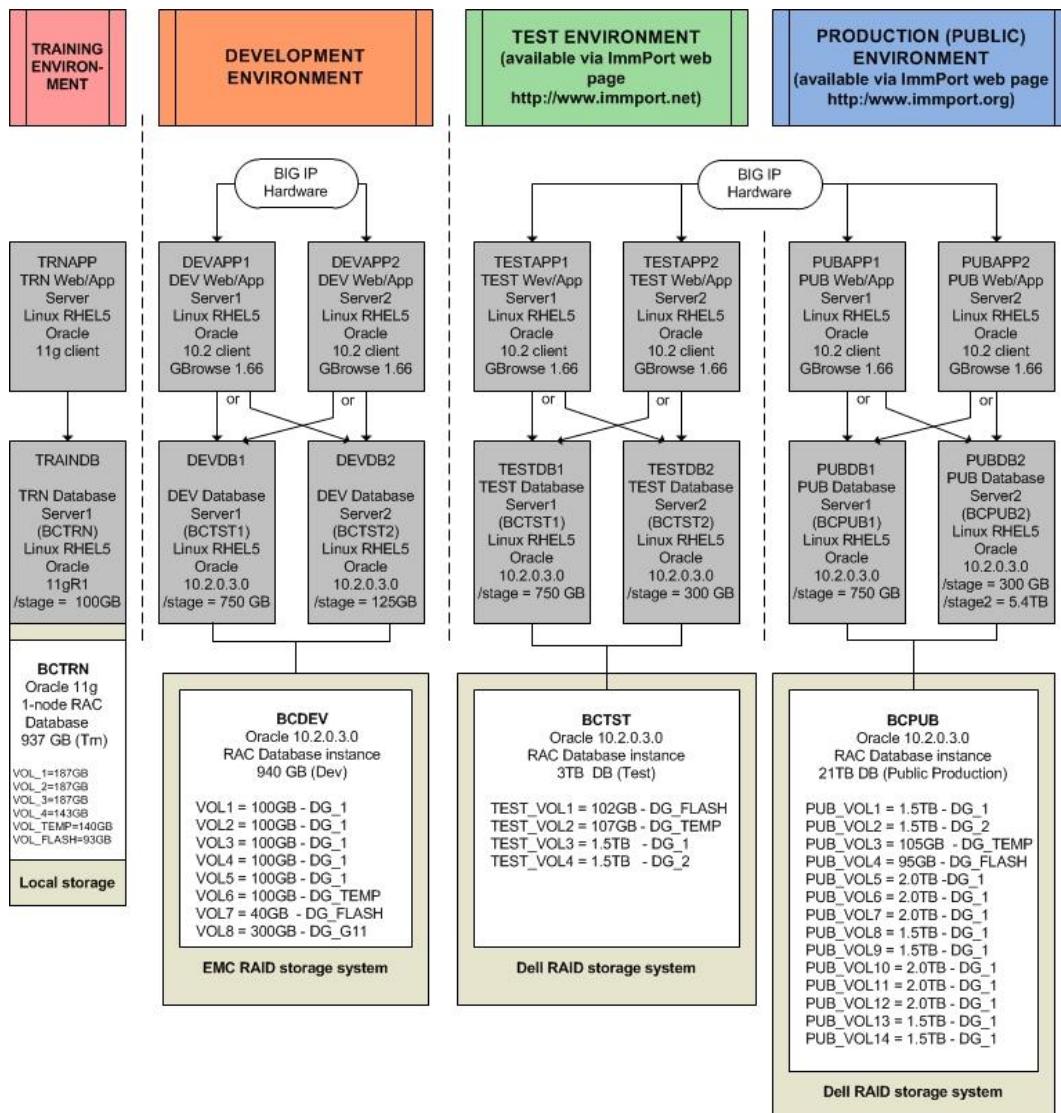
The ImmPort database architecture consists of four RAC database system environments as described below:

- **BCDEV** - ImmPort Development RAC database instance - Oracle 10gR2 (default for BISC project team), consisting of two database server nodes (BCDEV1 and BCDEV2)
- **BCTST** - ImmPort Test RAC database instance – Oracle 10gR2 (default for pre-production, data staging, and testing activities), consisting of two database server nodes (BCTST1 and BCTST2)
- **BCPUB** - ImmPort Production RAC database instance – Oracle 10gR2 (default for all ImmPort system users), consisting of two database server nodes (BCPUB1 and BCPUB2).
- **BCTRН** - ImmPort Training One-Node RAC database instance – Oracle 11gR1 (default for training BISC project team, and for system software installation and upgrade testing), consisting of a one-node RAC database server (BCTRН1).
- Upon deployment to OCICB, in addition to the BCPUB database instances being deployed at the NIAID facility, an additional QA database instance will be housed at that facility to perform installation testing. Specifications for this system can be provided in a later iteration of this document.

These four RAC databases contain multiple schemas and data marts that support all aspects of the development, training, deployment, and maintenance lifecycle of the ImmPort system. These RAC environments have the capability to support additional database servers if needed in the future for expanded scalability to meet performance needs required as ImmPort system user growth and ImmPort system load increases.

Below is a diagram depicting the four RAC database environments along with the high-level hardware server components. The production environment going forward will exist at the OCICB facility, and this diagram will necessarily change.

Figure 3-1. Server Architecture Diagram for Production, Development, Test, and Training Environments



3.1.1 ImmPort Development Database (BCDEV)

The ImmPort Development database supports the database development-related activities that are performed on a daily basis by the BISC project team. This is the default database for the BISC project team, allowing separate connection route to the ImmPort database, and supports daily deployments of new development and maintenance software releases. This separate connection and database ensures that the BISC project team activities do not impact normal ImmPort production users and their activities, which are performed on the ImmPort Production database (BCPUB), or that they also do not impact the ImmPort testing activities, or the monthly data refreshing data staging activities performed on the ImmPort Test database (BCTST) by the BISC project team.

Common activities on BCDEV database include:

- Connections to the database for all aspects of ImmPort system design and development
- Design, development, and loading of new/revised ImmPort data structures.
- Provide an amount of ImmPort database content sufficient for front-end, middle-tier, analytic tool development, and for systems integration development purposes.
- Design, development, and implementation of database objects for administration data, experimental data, and reference data.
- Design and development related to database programming, system-level programming, DBA activities, system administration activities, data processing, and data publishing activities.

3.1.2 ImmPort Test Database (BCTST)

The ImmPort Test database is used for system and user acceptability testing, and also serves as a staging environment for staging, pre-processing, and loading (ETL activities) of ImmPort Reference data that is refreshed on a monthly basis. A dual-schema design is used for all types of Reference data provides 100% up-time capability while ETL activities are being performed. This database is the default instance for end-users participating in testing new functionality and/or enhanced functionality.

Common activities on BCTST database include:

- System testing, Regression testing, and user acceptability testing for new development and maintenance software releases.
- Testing related to System defects, both in current production releases, and in development software releases.
- Establishment of standardized data collection of reference data via FTP from public data sources to the FTP server and subsequent transfer of data to the ImmPort Test instance (BCTST).
- Initial creation, update, and maintenance of complete database staging environment and all database objects, data streams, and server dependencies that are associated with the extraction, transformation, and loading (ETL) of public reference data.

3.1.3 ImmPort Production Instance (BCPUB)

The ImmPort Production database supports the user community for their normal ImmPort connections, serving as the default connection to the ImmPort database via the application server.

3.2 OVERVIEW OF THE IMMPORT DATABASE DESIGN AND CONTENT

Database design is defined as, "Design of the logical and physical structure of one or more databases to accommodate the information needs of the users in an organization for a defined set of applications." The

ImmPort design process consists of planning and analysis, conceptual design, logical design, physical design, and implementation. Due to space constraints, however, the data models in this document have been limited in scope to show necessary and active areas of database environment. The areas include experimental research data and clinical data, reference data, and analytic tools. The document includes Entity-Relationship (ER) diagrams for the relevant tables across these general functional areas.

3.2.1 Overall Database Design

The ImmPort database consists of multiple schemas and, where needed, specific data marts within those schemas. These schemas and data marts contain many types of Oracle database objects and sub-objects including, but not limited to: tables, views, materialized views, triggers, sequences, types, functions, procedures, packages, package bodies, database links, indexes, table partitions, index partitions, lobs, lob partitions, synonyms, etc. The ImmPort system database currently contains over 103,000 database objects, with system, schema, and object traceability available inherently built into the Oracle 10g database via the suite of Oracle DBA and USER views. (Some database view examples include: DBA_TABLES, USER_TABLES, DBA_VIEWS, USER_VIEWS, DBA_MVIEWS, USER_MVIEWS, etc.).

The current ImmPort system schemas, and the total number of main database objects per schema are listed in Table 3-1 below. The table shows both primary and matching alternate database schemas (where applicable). The alternate database schemas are utilized for data refresh of Reference data on a monthly basis, and on a weekly basis for some loading activities, allowing for instantaneous switching (pointing) to new monthly database objects content, and 100% up-time for content provided in the Reference data module. It also allows for easy QC of the current month's data against previous month's data, and the unique ability to switch back to the previous month's data, in the event of a problem with the new data, or the data refresh process or publishing. The main database objects include: tables, views, and materialized views. All database objects are tracked and maintained using various DBA tools such as the suite of tools available with Oracle Enterprise Manager 10g and with TOAD for Oracle Suite with DBA module.

Table 3-1 ImmPort Database Schemas

ImmPort Area	Primary Database Schemas	Alternate Database Schemas (if applicable - used for Reference Data Refresh)	Total Number of Main Database Objects (Tables-Views-Materialized views)	Schema Description
DBA Administration	Administration		5	Contains Oracle DBA administration information.
Reference Data (or Public Data Warehouse)	BISC Data 1	BISC Data 2	21 Tables 16 Materialized Views	Contains the main data tables and views of the DW. This includes both reference data extracted from public sources, and in the future, published experimental data released to the ImmPort user community.
	AFCS **	AFCS1 **	6	Contains Alliance for Cellular Signaling (AfCS) yeast two-hybrid data.
	AFFYMETRIX **	AFFYMETRIX1 **	6	Contains tables to support Affymetrix platform information.
	AGI	AGI1 (data refresh weekly)	15	Contains Animal data loaded from external partner
	BIND	BIND1	4	Contains Biomolecular Interaction Network Database (BIND) protein interactions data and complex-to-subunits data.

ImmPort Area	Primary Database Schemas	Alternate Database Schemas (if applicable - used for Reference Data Refresh)	Total Number of Main Database Objects (Tables-Views-Materialized views)	Schema Description
ImmPort Area	BIOCYC **	BIOCYC1 **	2	Contains Pathway data from BioCyc data source.
	CLASSIFI	CLASSIFI1	10	Contains the main CLASSIFI database source information as part of Gene Ontology database (GO Database). This schema is used for CLASSIFI algorithm within gene expression analysis
	DBMHC **	DBMHC1 **	8	Contains data obtained (downloaded) from dbMHC.
	DBSNP		11	Contains dbSNP flat data, dbSNP sequence data, dbSNP HapMap data, and dbSNP Hapmap genotype data for 5 populations. Current SNP data are for the human species only.
	DIP **	DIP1 **	0 Tables	FUTURE USE - NOT USED. May contain DIP schema data in the future.
	ENTREZ	ENTREZ1	19	Contains Entrez Gene data publicly available at ftp://ftp.ncbi.nih.gov/gene/DATA . Also contains MIM title information for Online Mendelian Inheritance in Man.
	GB_HG17	GB_HG171	7	Contains tables in the GBrowse schema for Human Hg17 (NCBI Build 35). Data currently includes all human Hg17 data, Ref gene track, Known gene track, 5 LDSelect TagSNP HapMap Population tracks, 6 specific SNP tracks, and attributes including ImmPort Genes list, and gene symbols.
	GB_HG18	GB_HG181	7	Contains tables in the GBrowse schema for Human Hg18 (NCBI Build 36). Data currently includes all human Hg18 data, Ref gene track and Known gene tracks only.
	HOMOLOGENE	HOMOLOGENE1	1	Contains homolog data from NCBI.
	KEGG	KEGG1	9	Contains Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information.
	MHC_SEQ_VAR	MHC_SEQ_VAR1	39	Contains HLA Sequencing data.
	MIPS **	MIPS1 **	4	Molecular Interaction Protein System (MIPS) information.
	ONTOLOGY	ONTOLOGY1	4	Contains 3 tables for the ImmPort Ontology browser, and 1 table for the Protégé v3.0 Ontology table.
	REACTOME	REACTOME1	2	Contains protein interaction data from Reactome.
	TAXONOMY	TAXONOMY1	3	Contains taxonomy (names.dmp) data from NCBI.
	UCSC	UCSC1	5	Contains known gene data and Refgene data for several species from UCSC.
	UNIPROT	UNIPROT1	14	Contains all Swiss-Prot protein data from UniProt Knowledgebase data source.
Research Data (includes Research Project / Private Project Workspace)	Research Data Schema		260 Tables 4 Materialized views 44 Views	Contains all objects defined for Research Data. This includes all experimental data, metadata, and results batch uploaded, or created on-line via Analysis Tools. All Research Data is located in a defined Research Project or Collaborative Project area. The Research Data schema also includes data tables associated with user registration, user logon, contract/grants, programs, research projects, experiments, analysis tool tables, lookup tables (containing coded values), and other miscellaneous administration tables required for the ImmPort System application.

ImmPort Area	Primary Database Schemas	Alternate Database Schemas (if applicable - used for Reference Data Refresh)	Total Number of Main Database Objects (Tables-Views-Materialized views)	Schema Description
Clinical Research Data (includes Research Project / Private Project Workspace)	Clinical Data Schema		65 Tables 17 Views	Contains all objects defined for Clinical Research Data. This includes all data tables related to Studies, both clinical and non-clinical. The Clinical data schema includes data areas such as: Study Design, Biomaterial Transforms, Adverse Events, Assessments, Lab Tests, Lab Test Analysis & Derived Measures, and Outcome Measures.
Clinical Research Data (includes Research Project / Private Project Workspace)	Clinical Staging Schema		400+ External Tables 2 Tables	Contains all objects related to Staging and ETL of Clinical Research Data. This includes all external tables related to Studies, both clinical and non-clinical. The Clinical Staging schema includes data included in both Study Design and Clinical Data batch upload data submission packages prepared for each Study. Each Study may contain up to approx. 50 files/tables of data. They include areas such as: Study Design, Biomaterial Transforms, Adverse Events, Assessments, Lab Tests, Lab Test Analysis & Derived Measures, and Outcome Measures.
Research Data (includes Database Audit History and Archive tables for DEVELOPER schema)	Audit schema		150 Tables	Contains all objects defined for ImmPort's Database Audit History and Archive features and capabilities. This includes a large subset of tables in the Audit schema with audit and archive columns added to each table. ImmPort System application inserts data into this schema as Research data is updated or deleted. With this schema, a chronological database audit and archive history and logging activity is achieved.
DB connectivity for Batch Upload	Data Submission schema		0 Tables	This schema is used exclusively to support Sign On (SSO) connectivity to the database for Batch Upload processing.
DB connectivity for Login etc.	General user schema		0 Tables	This schema is used exclusively to support Sign On (SSO) connectivity to the database for the ImmPort system. The front-end application utilizes this schema and associated roles and privileges to perform database connection and to support the various database actions requested by the user. It provides independent database-connectivity support for the ImmPort managed connection pool.
DB connectivity for Gene Expression Analysis Module (GenePattern)	GenePattern schema		0 Tables	This schema is used exclusively to support Sign On (SSO) connectivity to the database for the Gene Expression Analysis Module (GenePattern), and provides independent database-connectivity support for the GenePattern application managed connection pool.
DB connectivity for HLA Analysis Tool	HLA Analysis schema		0 Tables	This schema is used exclusively to support Sign On (SSO) connectivity to the database for the HLA Analysis Module, and provides independent database-connectivity support for the HLA application managed connection pool.

**** NOTE:** Schema is currently NOT USED.

3.2.2 Overview of the ImmPort Database Design, Implementation, and Deployment Process

All database design for the ImmPort system has progressed from conceptual models, through logical models, to physical models. This database modeling takes place at the DBA level utilizing several data modeling and development tools. Once data models have been developed and their designs meet project goals and requirements, these models are then used as a reference to create the actual physical database. The physical database schemas and objects are then created using the models as reference in the design and development process. Initially all physical data models are created and implemented in the Development database environment. Additionally, mock-ups, Powerpoint slides, use case diagrams, and other materials are used as design tools and aid in the overall design and development. At this point, the main development phase of the process begins, where database software, database applications, database schemas, and all types of database objects are created, modified, linked or associated, removed and/or replaced, refined, and ultimately constantly maintained. Once development is complete, this process is implemented in the Test database environment. As changes and necessary additions are identified, they are once again designed and implemented in the Development database environment, and then when they are ready for testing, they are re-deployed to the Test database environment, where they are once again tested, many times incrementally. Regression testing, where applicable and necessary, is also performed as an integral part of the testing and acceptance process. Once final acceptance is given, then the database design is deployed to the Production database environment, at a scheduled time, where once again it is tested before ImmPort users are granted access. This completes the database design, implementation and deployment process into the Production database environment.

3.2.3 Overall ImmPort System Architecture

For database performance, optimization, or visualization purposes, customized data marts are created that may span more than one subject area. These data marts can be logical and/or physical views consisting of summary-level, horizontal and/or vertical views of the atomic-level data. The OLAP option provides the capability to create different dimensions and views of the data without the need to replicate data at different summarization levels or organization levels. Below is a summary of the subjects areas and their object content. For visualized content, these subject areas as described as data models in subsequent Sections 4 and 5.

Research Data subject areas (covered in Section 4) include:

- ImmPort Administration
- Research Projects and Collaborative Projects
- Research Data (experimental and mechanistic)
- Research Data (clinical)
- Lookup tables
- Analysis Tools
- Other

Reference Data subject areas (covered in Section 5) include:

- NCBI Entrez Gene and Taxonomy
- Genomic gene and gene structure

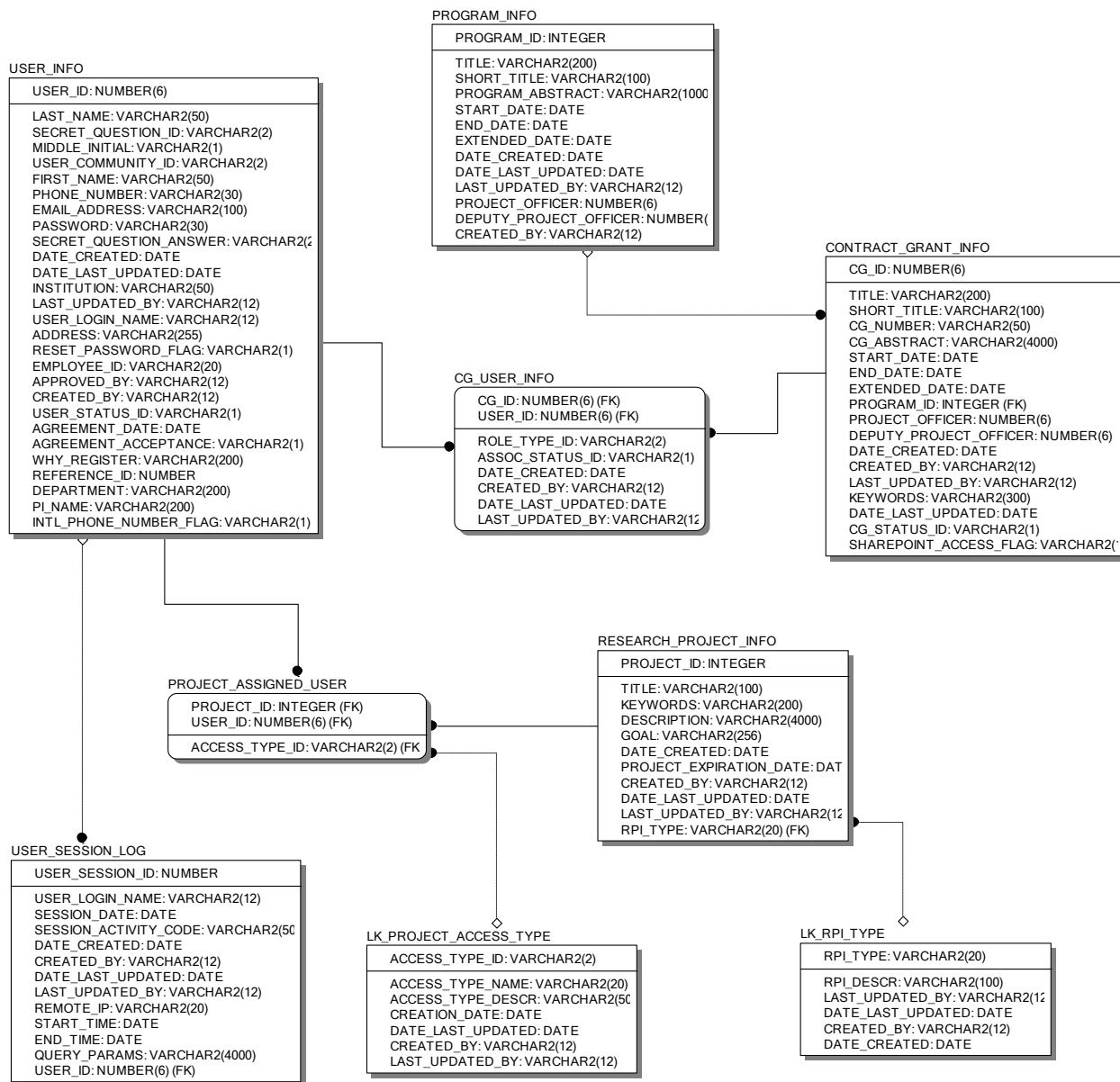
- Gene expression platform annotation (Affymetrix, Illumina, Agilent)
- Pathways
- Basic Protein
- Basic Protein Interactions
- Protein-protein Interaction Networks
- SNP polymorphism
- MHC and microsatellite polymorphism

4.0 RESEARCH DATA MODELS

4.1 IMMPORT ADMINISTRATION

The ImmPort system contains a core set of database objects pertaining to the administration of the system. The table objects in Figure 4-1 stores the information needed to administer the use of ImmPort and enable the basic user login and system functions required.

Figure 4-1. ImmPort Administration Data Model



4.2 RESEARCH PROJECTS AND COLLABORATIVE PROJECTS, AND THE SEMI-PUBLIC WORKSPACE

Project areas provide DAIT-funded scientists a place where they can store pre-publication experimental research data, experimental results, clinical data and other related data (e.g., biological sample data, subject data, study design, etc.). This Project data storage provides ImmPort users the ability to conduct data sharing, perform data queries, and utilize analysis tools integrated in ImmPort. PI users and PM users have the ability to create and update Projects, and also to assign/de-assign users to Projects. Upon Project creation, each Project is given a unique identifier (PROJECT_ID) which is then used exclusively on all data submitted and created within ImmPort.

Projects can have four different types: Research Projects (also known as Private Project Workspaces), Collaborative Projects (also known as Collaborative Workspaces), Semi-Public Workspace, and the Public Workspace.

Authorization and access control to a physical data record within a Research Data table, is done through the exclusive use of the PROJECT_ID identifier and its linkage to the accompanying RPI table, and at a lower level, the ImmPort User's USER_LOGIN_NAME for access within a Project. For example, to determine if a given Subject row from the SUBJECT_ORG_INFO table belongs to a specific Research Project, Collaborative Projects or the Semi-Public Workspace one would check the RPI_2 SUBJECT table for the appropriate records.

The Research Project area provides a private and secure workspace for DAIT-funded users to submit, upload, and create research and analysis data online in ImmPort. A row of research data must be associated with one and only one Research Project and the associated Research Project is chosen at the time of data submission.

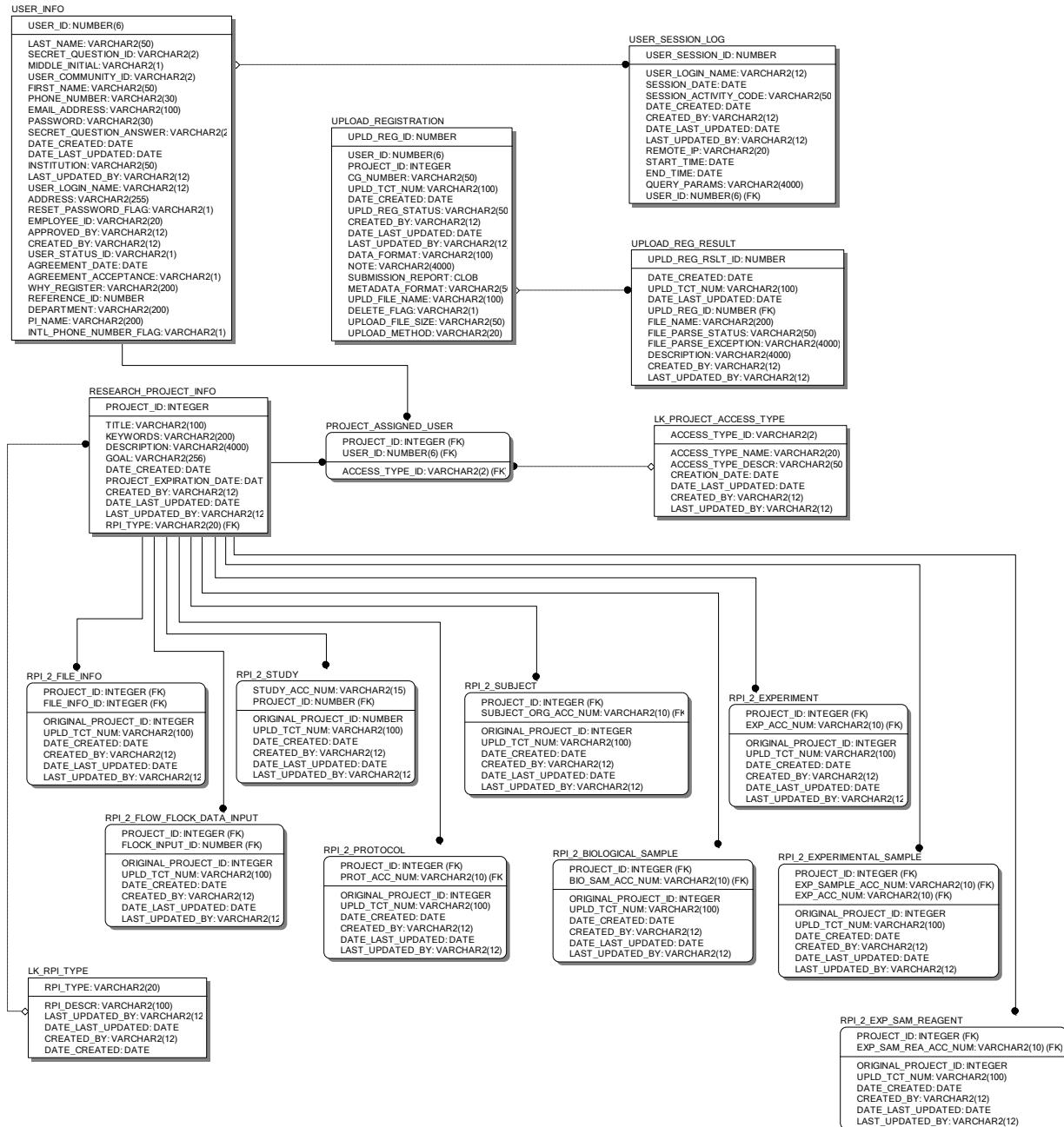
Collaborative Project areas provide DAIT-funded scientists a place where they can share any type of pre-publication experimental research data that is initially stored in a Research Project. PI users and PM users have the ability to create and update Collaborative Projects, and also to assign/de-assign users to Collaborative Projects.

After research data is shared to the Semi-Public Workspace, it is available to all registered users. At the time of sharing, an additional record is created in the accompanying RPI table for each research data entity shared that is linked to the record in the RESEARCH_PROJ_INFO table with type = "SPW".

Data that can be made available to users in the general public (i.e., those without an ImmPort login) can be associated with the Public Workspace.

Figure 4-2 shows tables designed to store the Research Project, Collaborative Project, and Semi-Public workspace information and linkages to the Research Data and RPI tables.

Figure 4-2. Research Projects, Collaborative Projects and Experiment Groups Data Model



4.3 EXPERIMENTAL RESEARCH DATA

The Experimental Research data area consists of DAIT-funded pre-publication experimental research data, experimental results, and other related data areas. This includes the following major data areas:

- batch upload registration and raw data file storage information
- experiment information
- experimental samples
- experimental results
- protocol information
- subject information
- biological samples
- experiment sample reagents
- annotation information on experiment sample reagents
- analytes
- treatments

The Experiment Research data area is defined as a central data repository for all data that is received through ImmPort's Experiment Data Submission module. All data submission is principally done via batch upload processing submitted online by ImmPort users. ImmPort users can upload all types of research data into their respective Research Project(s). The Experiment Research data model is broken in three parts as shown in Figure 4-3, Figure 4-4, and Figure 4-5. They capture all of the major data areas as bulleted above.

Figure 4-3. Batch Upload Tables - Experimental Research Detail - Part 1 of 3

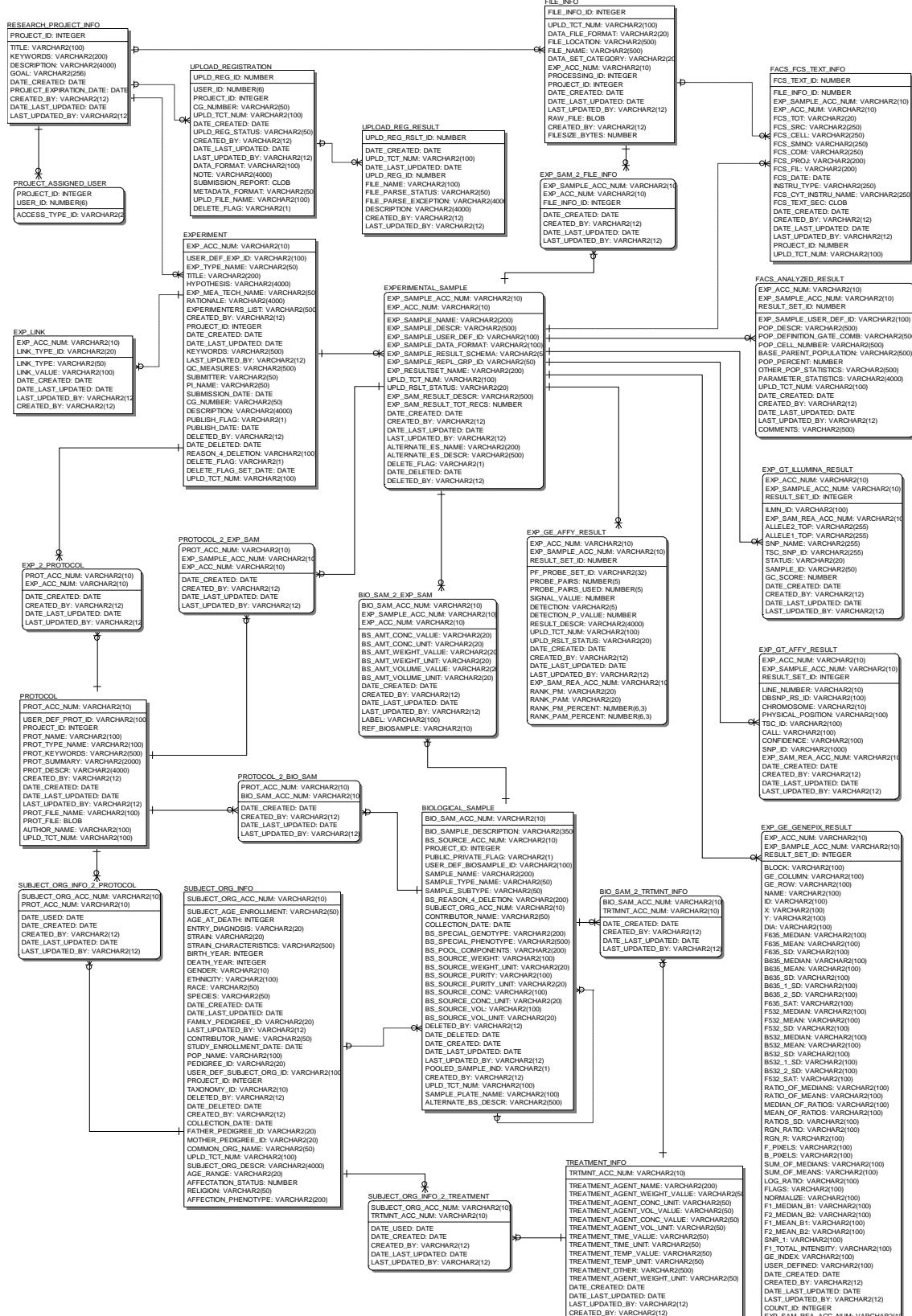


Figure 4-4. Batch Upload Tables – Protocol, Study, Subject, Biological Sample - Part 2 of 3

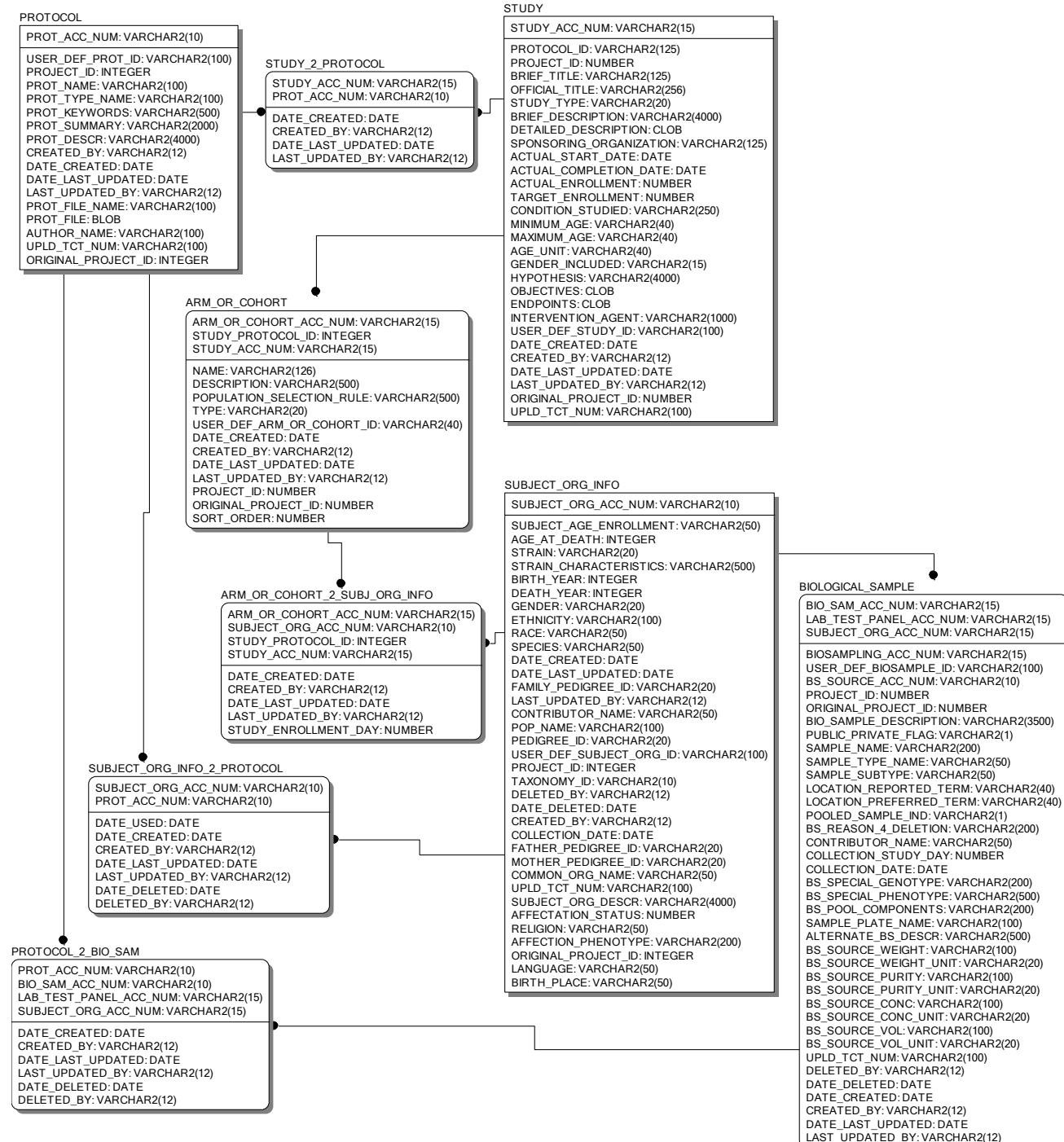
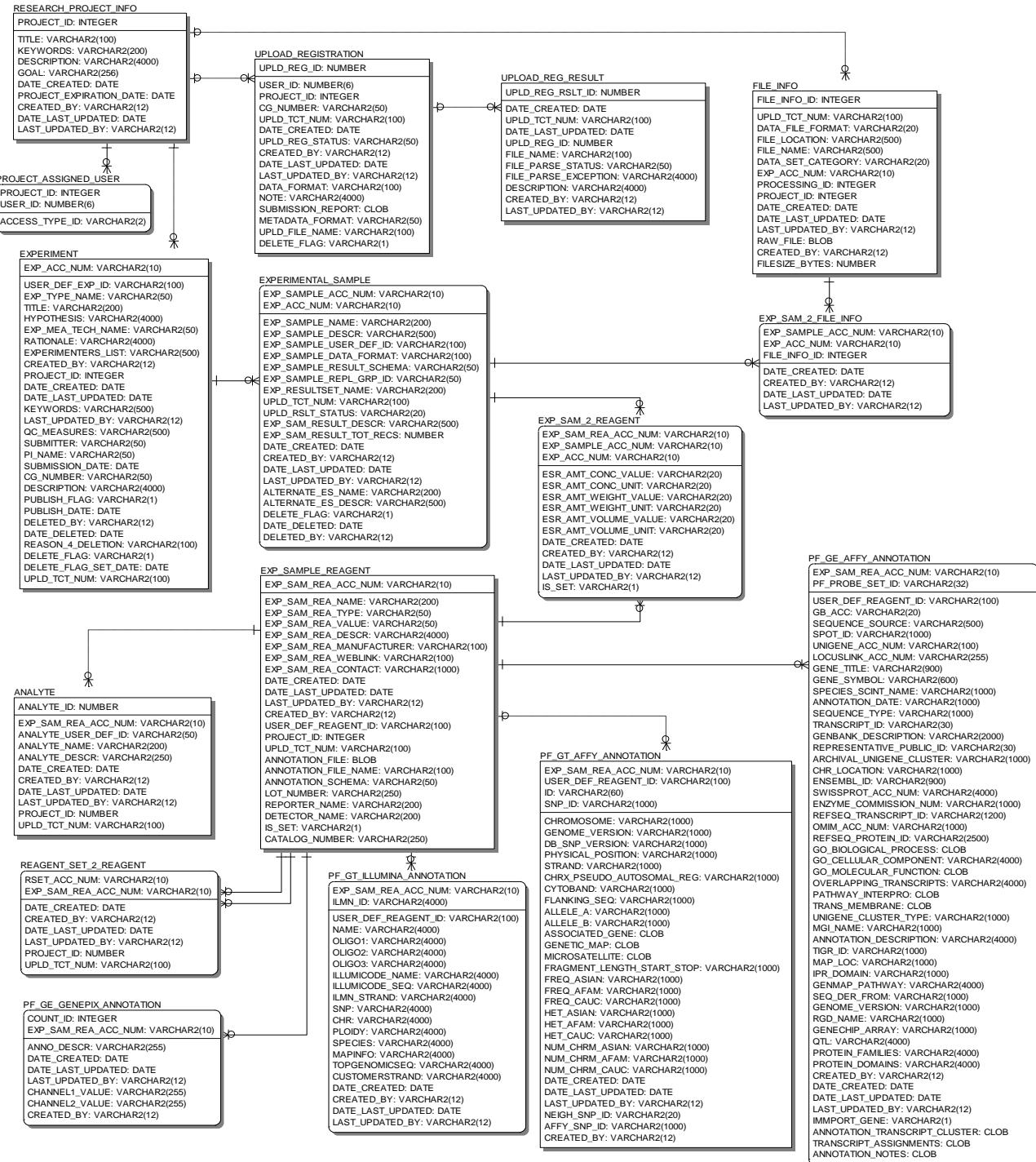


Figure 4-5. Batch Upload Tables - Experiment Sample Reagents and Annotation - Part 3 of 3



4.4 LOOKUP TABLES

The ImmPort system architecture contains an area of tables used exclusively for Lookup tables containing coded values and definitions for column data contained in ImmPort. The coded values and definitions developed and stored in ImmPort system provide the DAIT and DAIT-funded user community with a systematic way to collect, and establish a controlled vocabulary (CV) for the immunology-based Research studies and analyses, and other system-related information. The Lookup tables listed in this section are organized by data-related area.

Figure 4-6. Administration Lookup Tables

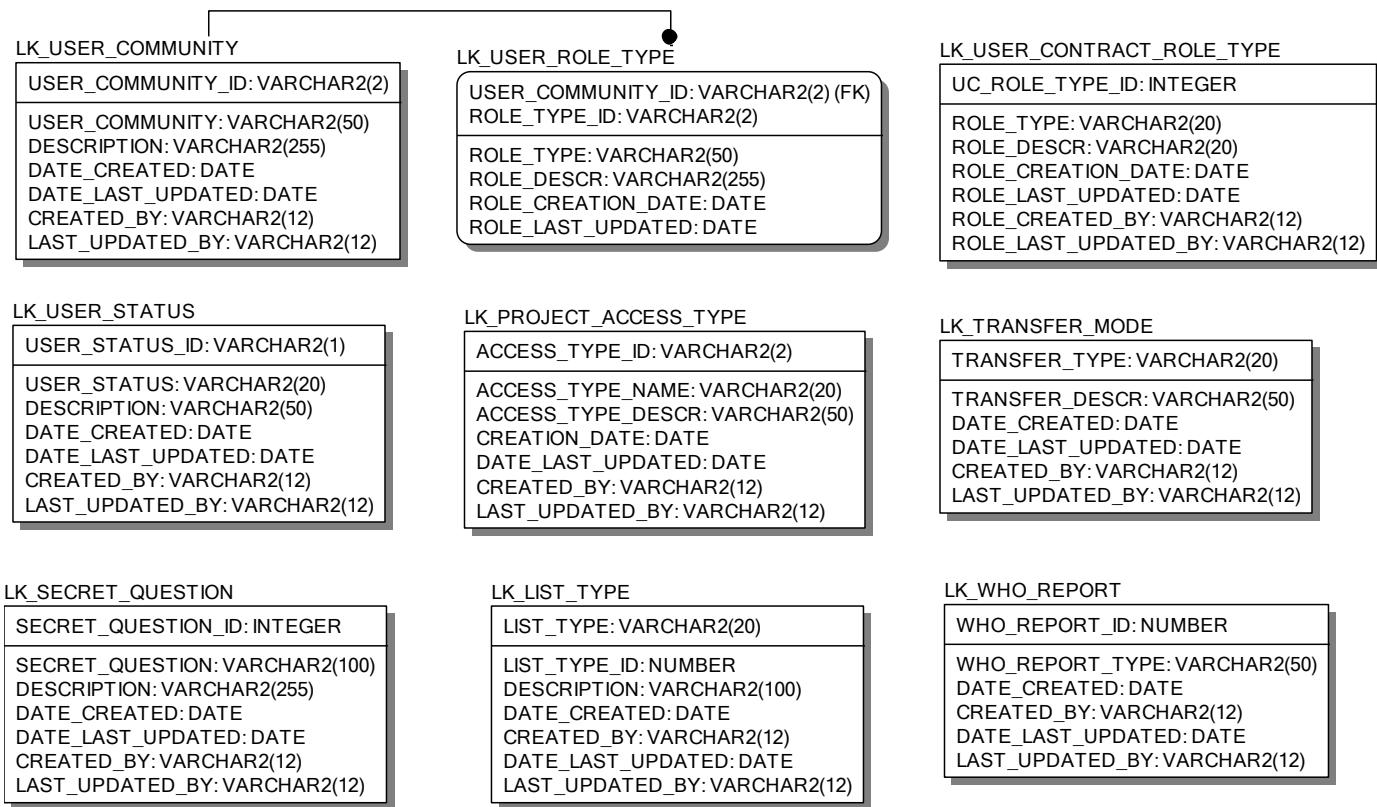


Figure 4-7. Research Experimental related Lookup Tables

LK_RPI_TYPE	LK_EXPERIMENT_TYPE	LK_EXP_SAM_TYPE
RPI_TYPE: VARCHAR2(20) RPI_DESCR: VARCHAR2(100) LAST_UPDATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE CREATED_BY: VARCHAR2(12) DATE_CREATED: DATE	EXP_TYPE_ID: INTEGER EXP_TYPE_NAME: VARCHAR2(50) EXP_TYPE_DESCR: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)	ES_TYPE: VARCHAR2(50) ES_NAME: VARCHAR2(100) ES_TYPE_COMMENT: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)
LK_DATA_FORMAT	LK_SAMPLE_TYPE	LK_PROTOCOL_TYPE
DATA_FORMAT_ID: INTEGER DATA_FORMAT_NAME: VARCHAR2(100) DATA_FORMAT_DESCR: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)	SAMPLE_TYPE_ID: VARCHAR2(10) SAMPLE_TYPE_NAME: VARCHAR2(50) SAMPLE_DESCR: VARCHAR2(100) CREATED_BY: VARCHAR2(12) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	PROTOCOL_TYPE_ID: VARCHAR2(10) PROTOCOL_TYPE_NAME: VARCHAR2(100) PROTOCOL_DESCR: VARCHAR2(500) CREATED_BY: VARCHAR2(12) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)
LK_EXP_MEASUREMENT_TECH	LK_ATTR_CATEGORY	LK_REAGENT_ROLE
EXP_MEASUREMENT_ID: INTEGER EXP_MEA_TECH_NAME: VARCHAR2(50) EXP_MEA_TECH_DESCR: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12) MQ_DISPLAY: VARCHAR2(1) MQ_ORDER: NUMBER(6)	ATTR_CATEGORY_ID: VARCHAR2(2) CATEGORY_TYPE: VARCHAR2(30) DESCRIPTION: VARCHAR2(100) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE CREATED_BY: VARCHAR2(12) LAST_UPDATED_BY: VARCHAR2(12)	ESR_ROLE_TYPE: VARCHAR2(30) ESR_ROLE_NAME: VARCHAR2(100) ESR_ROLE_COMMENT: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)
LK_REAGENT_2_SNP	LK_EXP_SAM_REA_TYPE	
EXP_SAM_REA_ACC_NUM: VARCHAR2(12) SNP_ID: VARCHAR2(20) CHR_NUMBER: VARCHAR2(10) CHR_POSITION: NUMBER GLOBAL_COORDINATE: VARCHAR2(30) TWO_ALLELES: NUMBER ORI_TO_DBSPN: NUMBER STRAND_ORI: NUMBER ALLELE_A: VARCHAR2(10) ALLELE_B: VARCHAR2(10) ANNOTATION_DATE: DATE BUILD_CODE: VARCHAR2(25)	EXP_SAM_REA_TYPE_ID: NUMBER EXP_SAM_REA_TYPE: VARCHAR2(250) EXP_SAM_REA_TYPE_NAME: VARCHAR2(250) EXP_SAM_REA_TYPE_DESCR: VARCHAR2(250) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	
LK_ANNOTATION_SCHEMA	LK_EXP_SAM_RESULT_SCHEMA	
ANNOTATION_SCHEMA_ID: NUMBER ANNOTATION_SCHEMA_DESCR: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	EXP_SAM_RESULT_SCHEMA_ID: NUMBER(2) EXP_SAMPLE_RESULT_SCHEMA: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	

Figure 4-8. Research Subject related Lookup Tables

LK_GENDER	LK_RACE	LK_ETHNICITY
GENDER_TYPE: VARCHAR2(20)	RACE_ID: INTEGER	ETHNICITY_ID: VARCHAR2(2)
GENDER_ID: VARCHAR2(2)	RACE_TYPE: VARCHAR2(50)	ETHNICITY_TYPE: VARCHAR2(50)
GENDER_DESCR: VARCHAR2(200)	DATE_CREATED: DATE	DATE_CREATED: DATE
DATE_CREATED: DATE	CREATED_BY: VARCHAR2(12)	CREATED_BY: VARCHAR2(12)
CREATED_BY: VARCHAR2(12)	DATE_LAST_UPDATED: DATE	DATE_LAST_UPDATED: DATE
DATE_LAST_UPDATED: DATE	LAST_UPDATED_BY: VARCHAR2(12)	LAST_UPDATED_BY: VARCHAR2(12)
LAST_UPDATED_BY: VARCHAR2(12)		
LK_SPECIES_TYPE	LK_MARITAL_STATUS	LK_RELATIONSHIP
SPECIES_ID: INTEGER	MARITAL_STATUS_ID: VARCHAR2(2)	RELATIONSHIP_ID: INTEGER
SPECIES_TYPE: VARCHAR2(30)	MARITAL_STATUS_TYPE: VARCHAR2(20)	RELATIONSHIP: VARCHAR2(50)
TAXONOMY_ID: VARCHAR2(10)	DATE_CREATED: DATE	DATE_CREATED: DATE
SPECIES_DESCRIPTION: VARCHAR2(100)	CREATED_BY: VARCHAR2(12)	CREATED_BY: VARCHAR2(12)
DATE_CREATED: DATE	DATE_LAST_UPDATED: DATE	DATE_LAST_UPDATED: DATE
CREATED_BY: VARCHAR2(12)	LAST_UPDATED_BY: VARCHAR2(12)	LAST_UPDATED_BY: VARCHAR2(12)
DATE_LAST_UPDATED: DATE		
LAST_UPDATED_BY: VARCHAR2(12)		
LK SUBJECT_POPULATION	LK_AFFECTATION_STATUS	
POP_NAME: VARCHAR2(20)	AFFECTATION_STATUS: NUMBER	
POP_SOURCE: VARCHAR2(20)	AFFECTATION_STATUS_DESCR: VARCHAR2(50)	
DATE_CREATED: DATE	DATE_CREATED: DATE	
DATE_LAST_UPDATED: DATE	CREATED_BY: VARCHAR2(12)	
LAST_UPDATED_BY: VARCHAR2(12)	DATE_LAST_UPDATED: DATE	
CREATED_BY: VARCHAR2(12)	LAST_UPDATED_BY: VARCHAR2(12)	

Figure 4-9. Research Study and Clinical related Lookup Tables

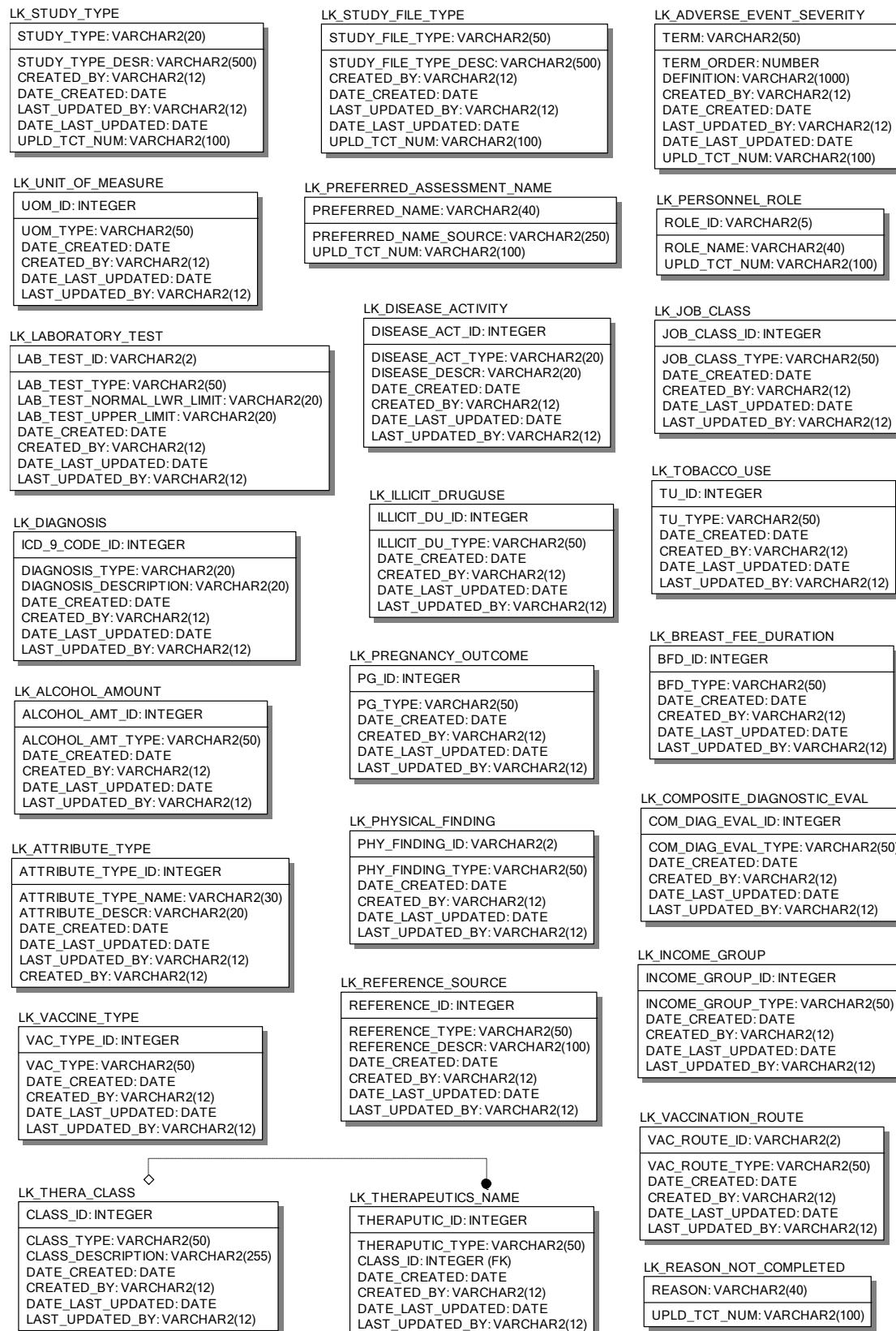


Figure 4-10. Analysis Tools and HLA related Lookup Tables

LK_TASK_TYPE	LK_GP_MODULE	LK_HLAVT_ANALYSIS_RSLT_STATUS
TASK_TYPE: VARCHAR2(50) TASK_TYPE_NAME: VARCHAR2(100) TASK_TYPE_DESCR: VARCHAR2(250) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)	GPM_ID: VARCHAR2(15) MODULE_NAME: VARCHAR2(250) MODULE_DISPLAY_NAME: VARCHAR2(250) MODULE_DESCRIPTION: VARCHAR2(4000) MODULE_CATEGORY: VARCHAR2(250) MODULE_DISPLAY_CATEGORY: VARCHAR2(250) MODULE_ENABLE: VARCHAR2(1) MODULE_VERSION: VARCHAR2(20) MODULE_VERSION_DATE: DATE SORT_ORDER: NUMBER DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	HLAVT_STATUS_NAME: VARCHAR2(50) HLAVT_STATUS_ID: NUMBER HLAVT_OLD_STATUS_NAME: VARCHAR2(50) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE CREATED_BY: VARCHAR2(12) LAST_UPDATED_BY: VARCHAR2(12)
LK_HAPLO_DATASET_TYPE		
HD_TYPE_ID: INTEGER HD_TYPE_NAME: VARCHAR2(50) HD_TYPE_DESCR: VARCHAR2(1000) DATE_CREATED: DATE DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12) CREATED_BY: VARCHAR2(12)		
LK_LOCUS_NAME	LK_ALLELE_STATUS	LK_FEATURE_SEQUENCE_TYPE
LOCUS_NAME_ID: NUMBER LOCUS_NAME: VARCHAR2(100) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	ALLEL_Status_ID: NUMBER ALLEL_Status_Type: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	FEATURE_SEQUENCE_TYPE_ID: NUMBER FEATURE_SEQUENCE_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)
LK_LOCUS_TYPING_METHOD	LK_FEATURE_STRAND	LK_FEATURE_TYPE
LOCUS_TYPING_METHOD_ID: NUMBER LOCUS_TYPING_METHOD_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	FEATURE_STRAND_ID: NUMBER FEATURE_STRAND_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	FEATURE_TYPE_ID: NUMBER FEATURE_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)
LK_EXON_INTRON_INTERROGATED	LK_FEATURE_LOCATION	
EXON_INTRON_INTERROGATED_ID: NUMBER EXON_INTRON_INTERROGATED_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	FEATURE_LOCATION_ID: NUMBER FEATURE_LOCATION_TYPE: VARCHAR2(50) DATE_CREATED: DATE CREATED_BY: VARCHAR2(12) DATE_LAST_UPDATED: DATE LAST_UPDATED_BY: VARCHAR2(12)	

4.5 ANALYSIS TOOLS TABLES - UPDATE FOR ADDED SFVT

The ImmPort system architecture contains an area of tables dedicated for use by the suite of Analytic Tools and visualization tools available in ImmPort.

These include:

- Flow Analysis Tools
- TagSNP Analysis Tool
- HLAVT Analysis Toolset
- ImmPort Genetic Analysis Tool (IGAT)
- Gene Expression Analysis (GenePattern) Tool
- ImmPort Genome Browser (GBrowse) Visualization Tool

The Analysis Tools database architecture is based on a design methodology that a unique set of 3 types of tables for each toolset including:

1. common system processing tables (ANALYSIS_TASK and LK_TASK_TYPE)
2. data input and processing tables
3. analysis output and result tables.

All of the Tools utilize two common tables ANALYSIS_TASK and LK_TASK_TYPE that are used to process all analysis jobs, and their statuses. As a high-level, both Project access control and User access control is used and maintained in the ANALYSIS_TASK table and is customized for Tool-specific application managed designs.

The **Flow Analysis Tool** data model (Figure 4-11) contains a set of tables that are used for FLOW analyses and provide for specific user actions such as:

1. uploading and storing flow cytometry datasets for flow analysis
2. creating and storing flow datasets for flow analysis
3. creating and storing flow datasets from existing uploaded flow datasets
4. creating and storing flow cross sample datasets from existing uploaded flow datasets
5. performing flow analysis
6. performing flow cross sample analysis

The **TagSNP Analysis Tool** (Figure 4-12) is available to all registered ImmPort users, and does not require that the user be assigned to any particular Research Project. Outside of Project access control, it uses its own TAGSNP_ANALYSIS_TASK table, in addition to creating and managing statuses in the ANALYSIS_TASK table, so that job management and content is standardized with all other Analysis Tools. The TagSNP Tool allows users the unique ability to submit a TagSNP analysis job that may contain up to 10 separate TagSNP tasks at a time. Each task (within the job) can be performed on a single Gene, or on a genomic range by specifying Chromosome start and end positions. For gene-specific tasks, the user is provided a gene discovery/query page to search and identify the genes they wish to select for analysis. A prettybase file is created and stored for each tagSNP task, and is used as input to the tagSNP analysis tool.

The **HLAVT Analysis Toolset** (Figure 4-13) is available to DAIT-funded users only. Specific tasks normally require user selection of a Research Project(s) or Collaborative Projects(s) that they wish to select data from (sources), and also selection of a Research Project for the data they create to go into (target).

The user-specific analysis tasks available include:

Task Type	Task Type Name	Task Type Description
hlqc	HLA QC	Run the HLA Quality Control pipeline from HLA typing results
hlava	Analyze VF	Use generated Sequence Feature Variant Type vector files in HLA Analysis
hlavd	HLA Ambiguity Reduction	Validate and disambiguate alleles and genotypes using IMGT data from HLA typing results
hlavf	Validate Alleles	Validate alleles using IMGT data from HLA typing results
hlavt	Generate VF	Generate a Sequence Feature Variant Type vector file from HLA typing results

The **IGAT Analysis Tool** (Figure 4-14) allows users to perform Genetic Association Analyses. The tool requires the user to select a Research Project(s) or Collaborative Projects(s) that they wish to select data from (sources), and also to select a Research Project for the data they create to go into (target). Users may create PED data files from data stored within ImmPort across multiple assay runs and assay platforms; due to the need to validate the merged SNP content across multiple vendor platforms, the data model for this function requires references to research data results (SNP), reference data content (dbSNP), and annotation data (SNP platforms and reagent information). Users can create and store Haplovew datasets based on data selected from the ImmPort System (contained in projects they have access to). Users can then perform IGAT Haloview analysis using the Haplovew (ped) datasets they have created.

The **Gene Expression Analysis (GenePattern) Tool** allows users to perform Gene Expression Analyses, using GenePattern developed by the Broad Institute. User can perform GenePattern Single Stage Analysis, or GenePattern Pipeline Multiple Stage Analysis. The ImmPort support team provides and maintains a list of GenePattern Modules that can be used for analysis. Task load balancing is performed for both single and pipeline jobs. The GenePattern database structure consists of the GenePattern v3.2.1 suite of tables, sequences, and triggers, in addition to the common system processing tables (ANALYSIS_TASK and LK_TASK_TYPE), and the LP_GP_MODULE table containing the available GenePattern module available for user selection and analysis usage. Currently there are 57 modules available to the user.

The **ImmPort Genome Browser (GBrowse)** is a visualization Tool available within the Resources section of ImmPort, and is available publicly to all users. The GBrowse tool is setup in an enterprise data architecture with modification to the Oracle adapter, and minor database enhancements for its usage in an Oracle environment. The database architecture and configuration of GBrowse Data Sources supports easy growth of Data Sources content, and added growth of available Tracks within each Data Source.

Figure 4-11. Flow Analysis Toolset Data Model

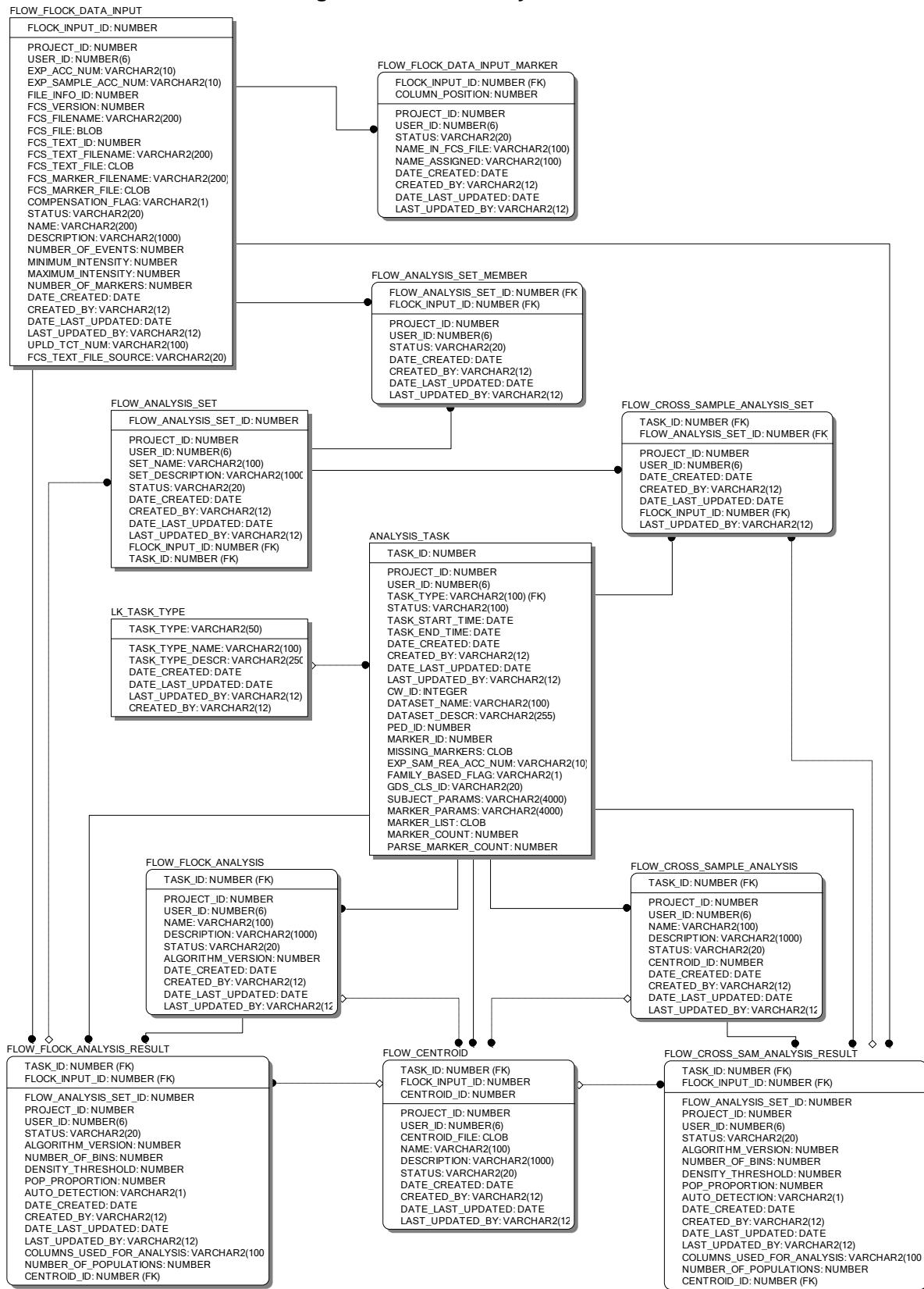


Figure 4-12. TagSNP Analysis Tool Data Model

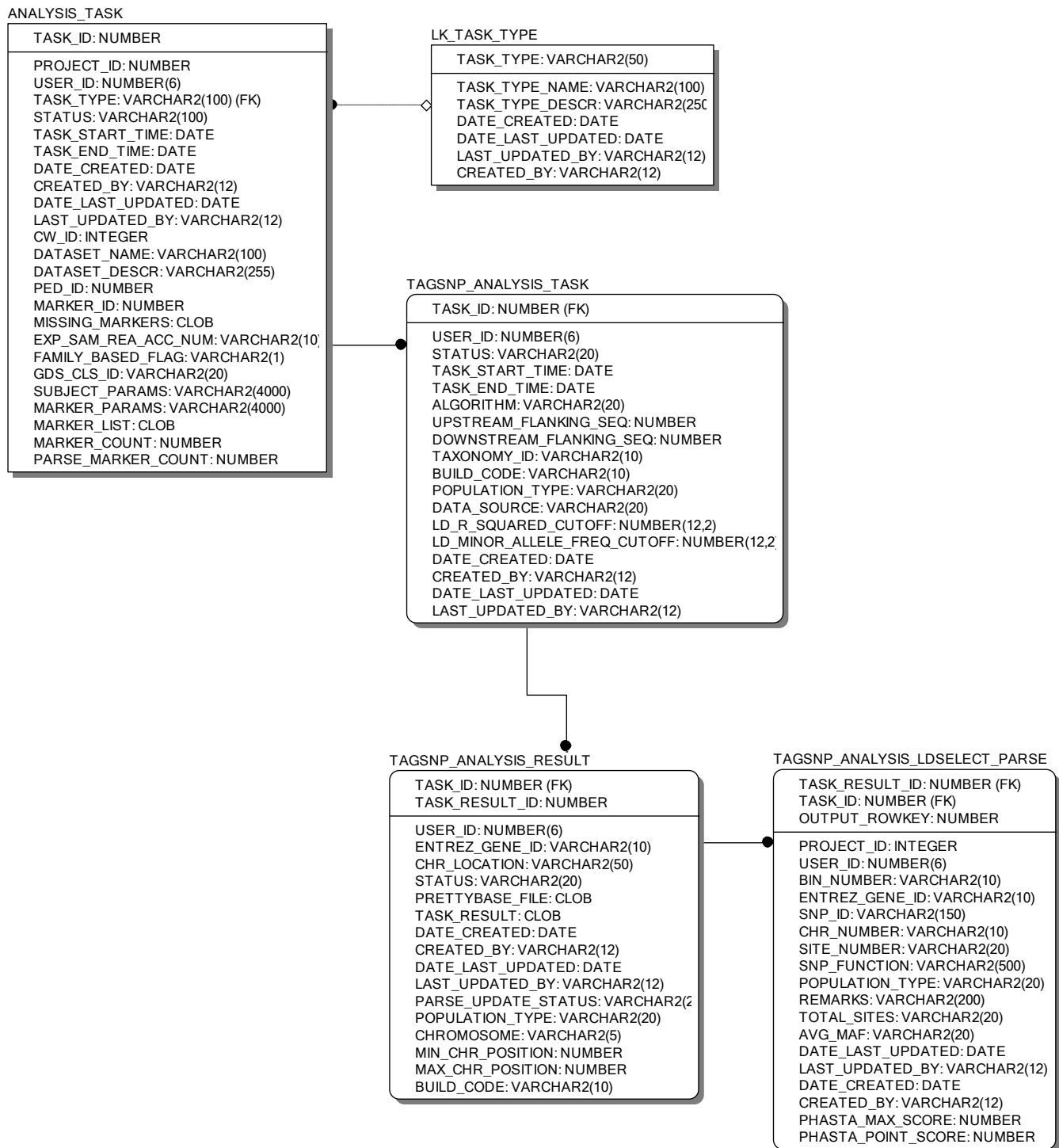


Figure 4-13. HLAVT Analysis Data Model

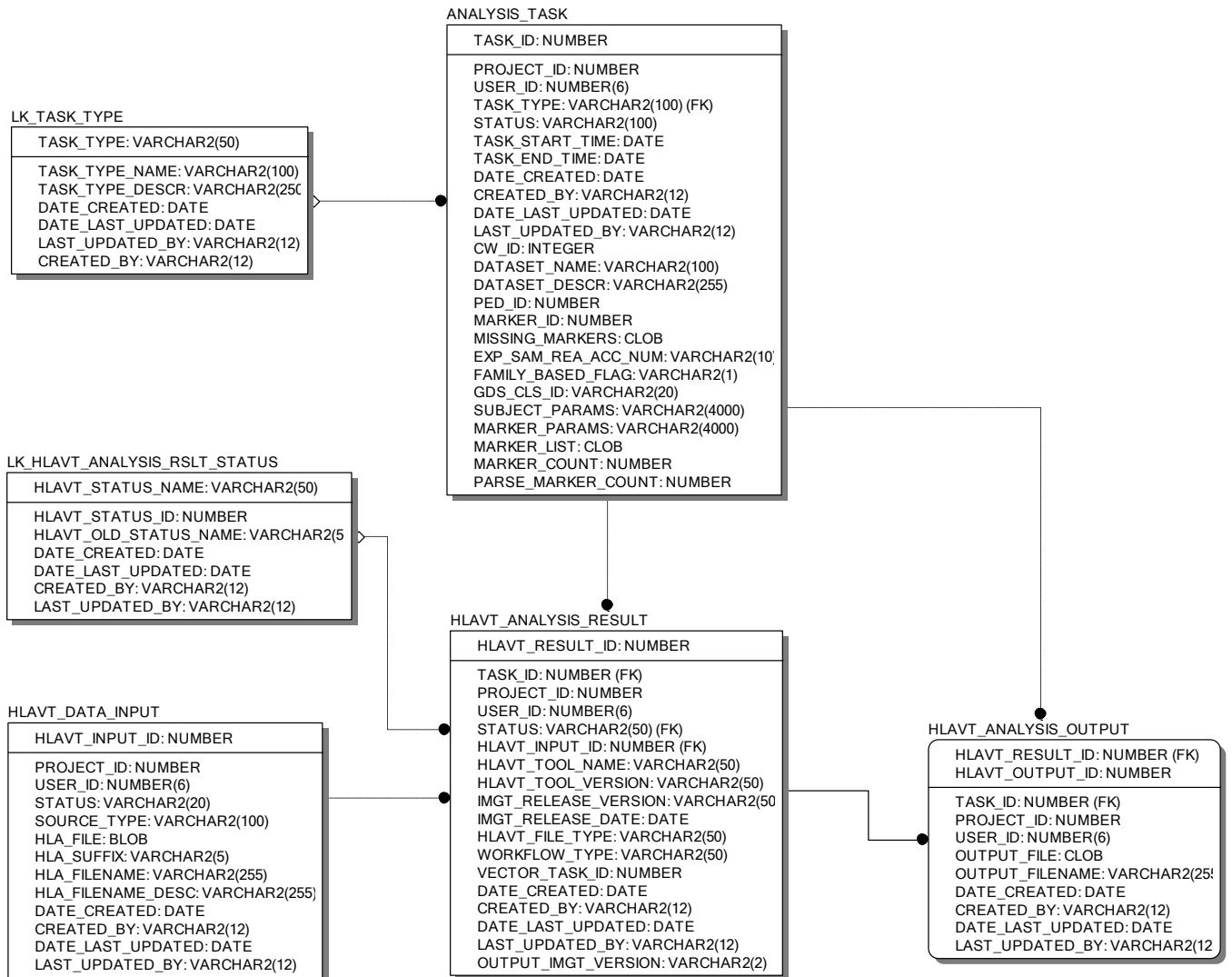


Figure 4-14. IGAT Analysis Tool Data Model (Haplovie dataset creation and Haplovie analysis)

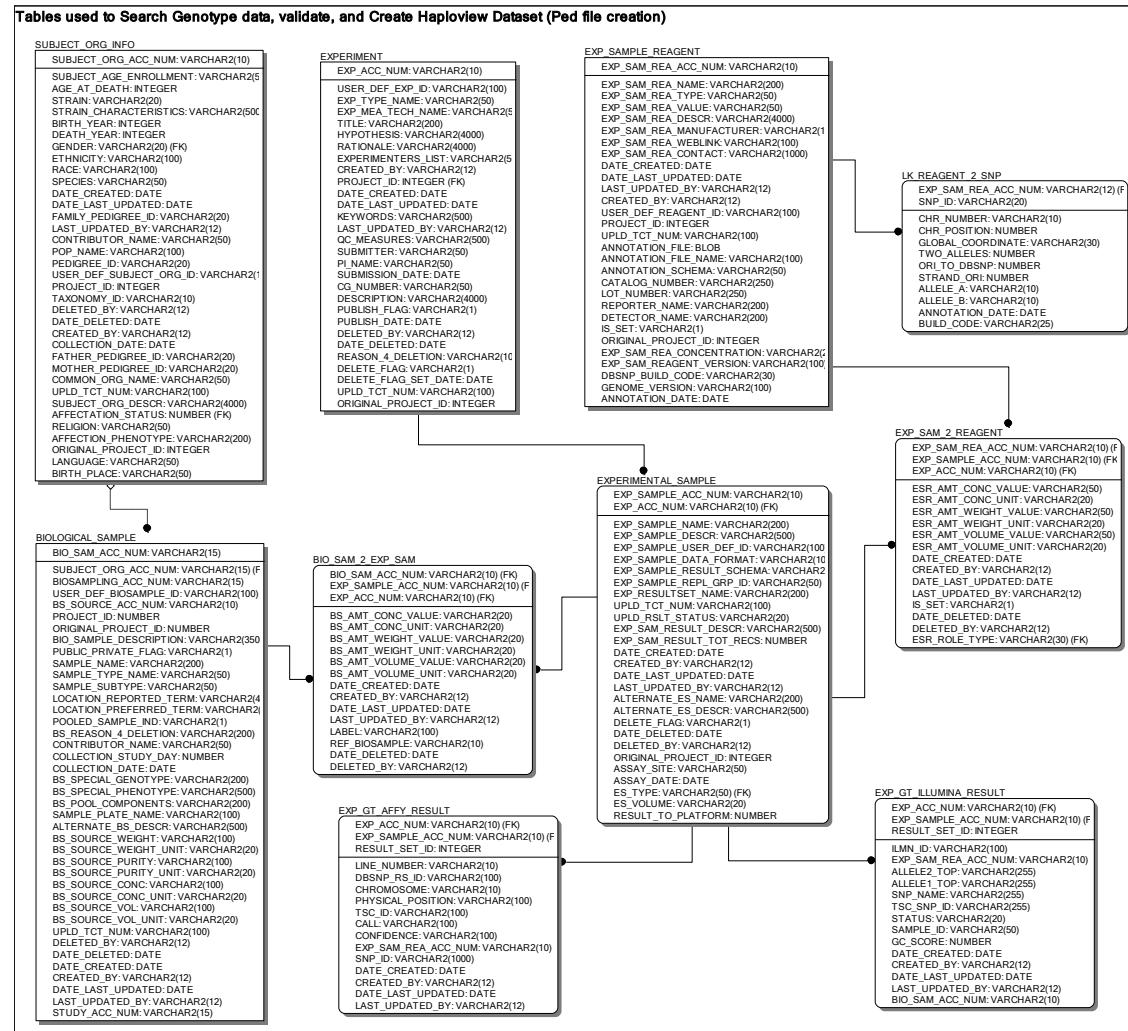
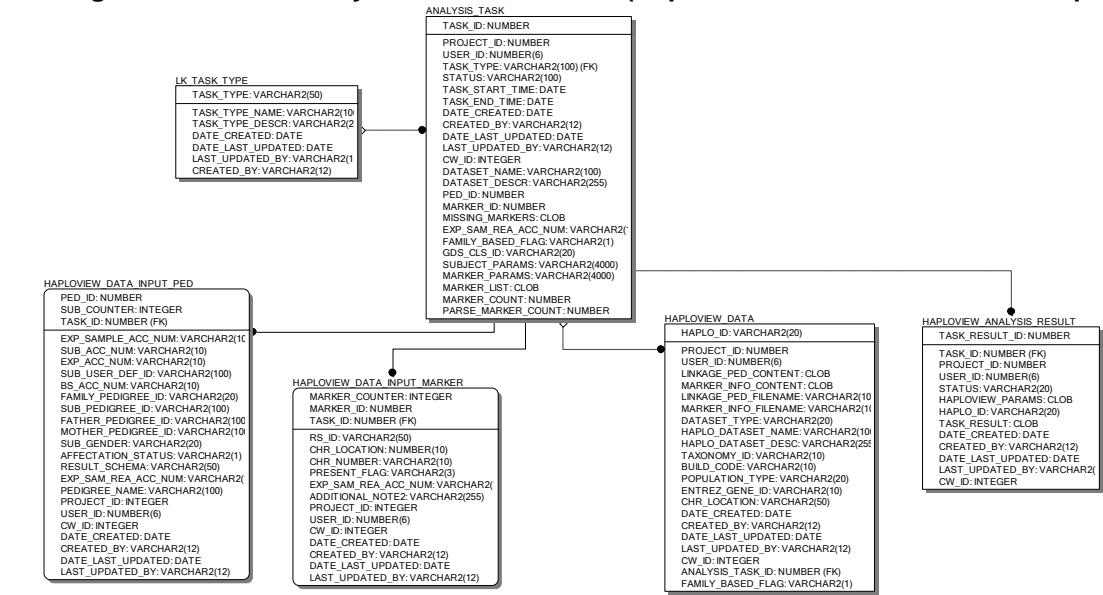


Figure 4-15. Gene Expression Analysis (GenePattern) Tool Data Model

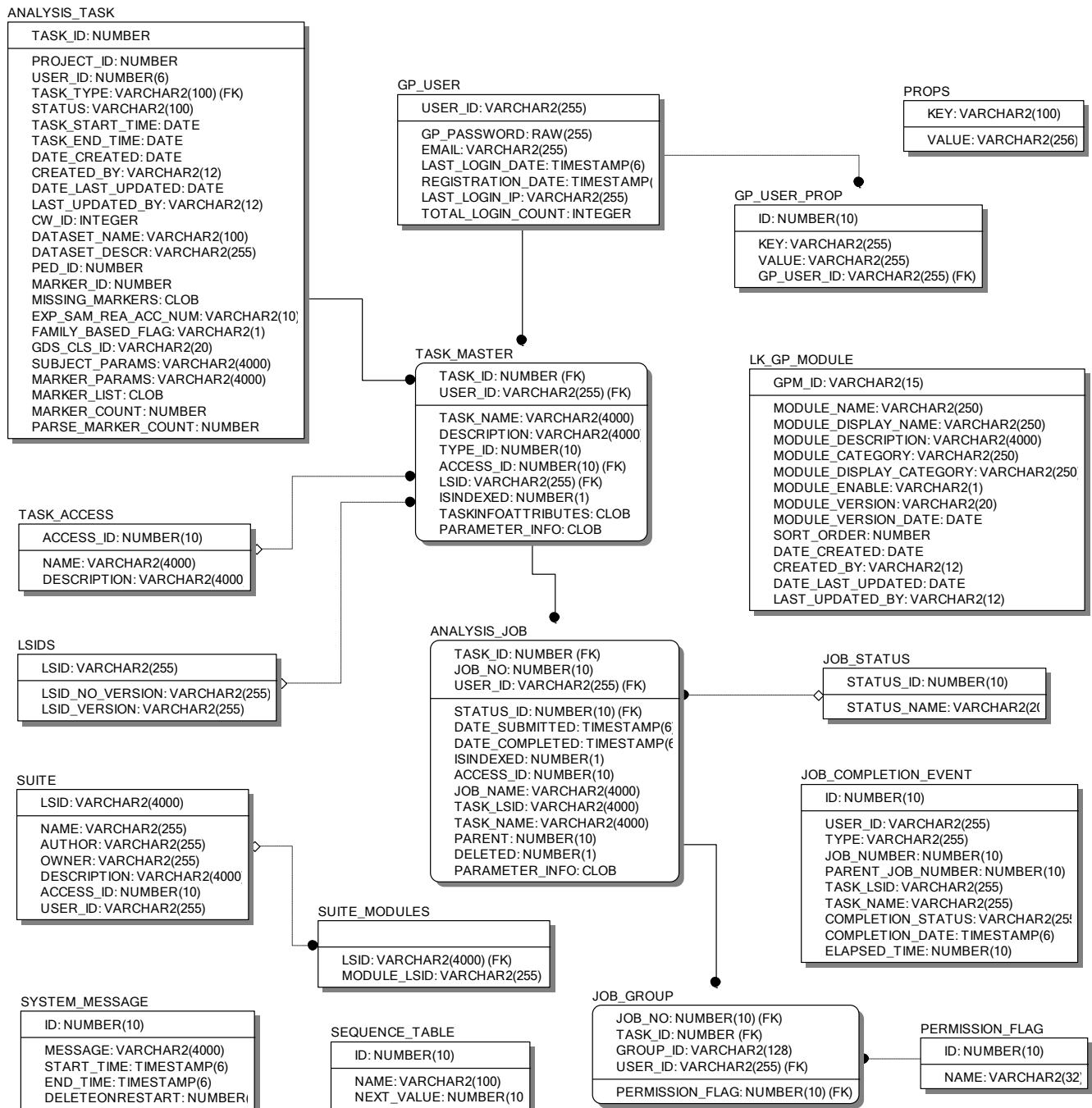
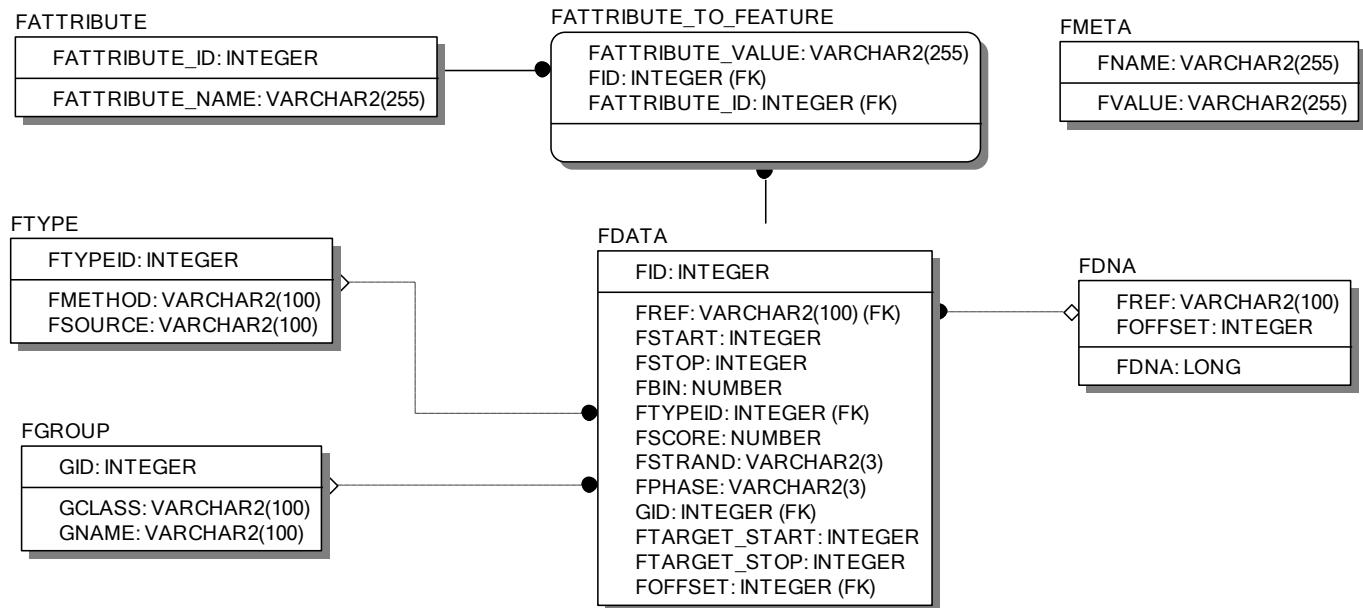


Figure 4-16. ImmPort Genome Browser (Gbrowse) Data Model



4.6 CLINICAL DATA MODEL

The BISC repository for clinical studies was developed to provide an interoperable representation of clinical study data. Interoperability constitutes storing and querying clinical study designs and descriptions, clinical datasets and mechanistic assays from clinical studies and trials. The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) and content from ClinicalTrials.gov was used extensively as a source for standards, entities and entity attributes contributing to interoperability. The process for modeling was published in the Kong, Dahlke et al in the Journal of Biomedical Informatics at http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed&cmd=Retrieve&list_uids=20460173.

The BISC Clinical Repository schema creates syntactic commonality among studies by categorizing data into object and event classes and employing CDISC code lists for controlled vocabulary study annotation. Mapping study data to the schema involves extracting data implicit in the study documentation (e.g., units of measure printed on CRF forms) and making it explicit for loading purposes. Capturing clinical study protocols is done by parsing the study description, purpose, plans and methods for loading into the study design area of the repository schema. The study design capture process is highly interactive and involves coordinated efforts by both researchers and BISC staff—this process is challenging due to the inherent variability between study types and study designs.

Data provided by either researchers or contract research organizations (CRO's) typically includes study protocols, data sets derived from Clinical Report Forms (CRFs) and mechanistic assay result files including microarray, flow cytometry, ELISA, etc. This data is presented in a uniform content view via the web interface--browsing studies, downloading study data, creation and download of data sets with access to analysis tools enhances study data accessibility and utility.

Although BISC may receive data from unregulated institutions, the repository will be organized as though all data submitted were subject to HIPAA regulation. The primary constraint imposed by HIPAA regulations on the BISC repository is that the data contained therein must be de-identified in accordance with HIPAA Safe Harbor standards. The Safe Harbor standards require that data which would allow identification of the subject of a study be removed from study records before being made publicly available.

To control the proliferation of entities in the model, the BISC team used a combination of row models for selected entities with a well defined set of attributes while modeling the remaining entities as EAV classes.

Figure 4-17. Clinical - Study Design Data Model

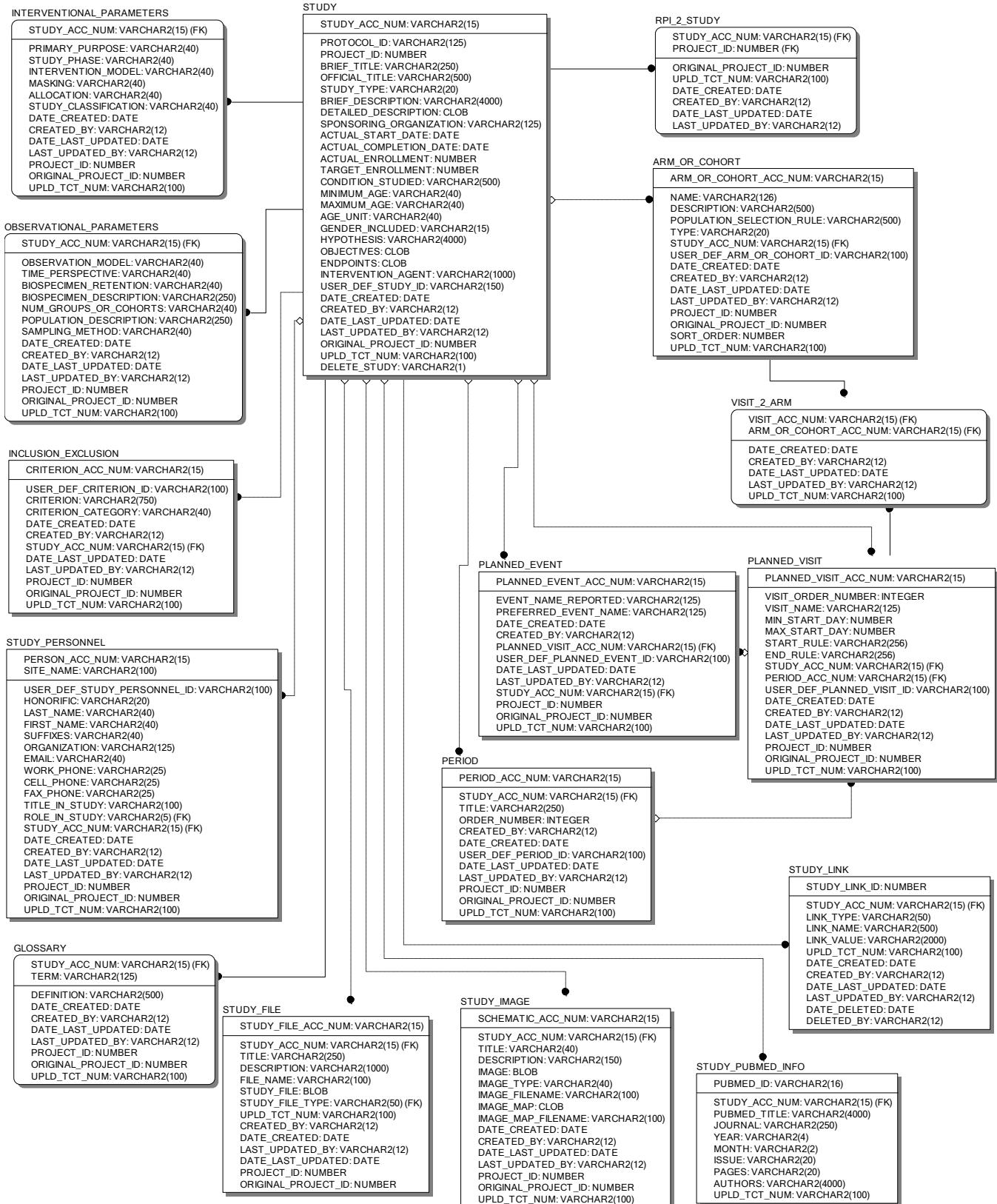


Figure 4-18. Clinical - Assessments, Adverse Events, Substance Merge Data Model

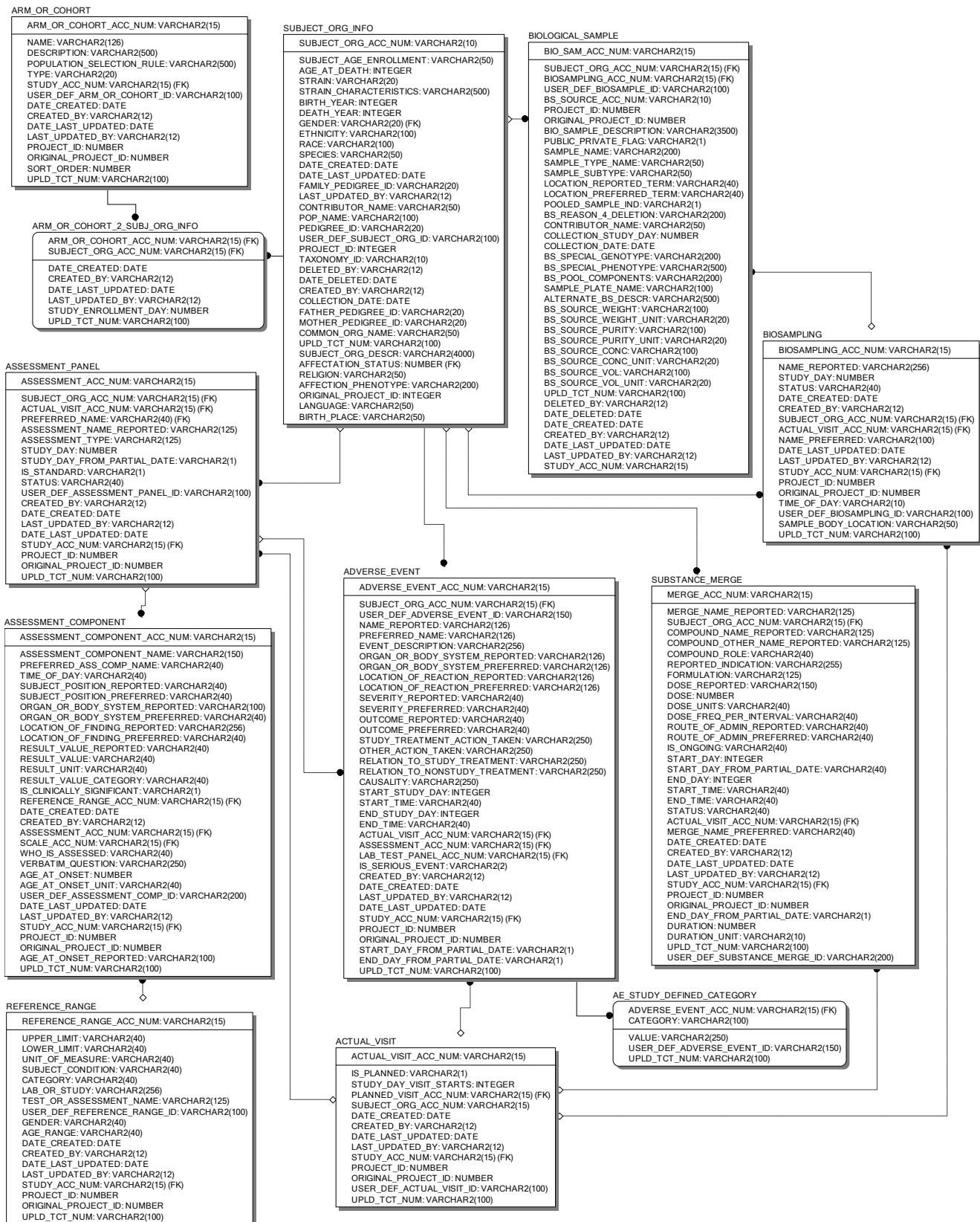


Figure 4-19. Clinical – Lab Tests Data Model

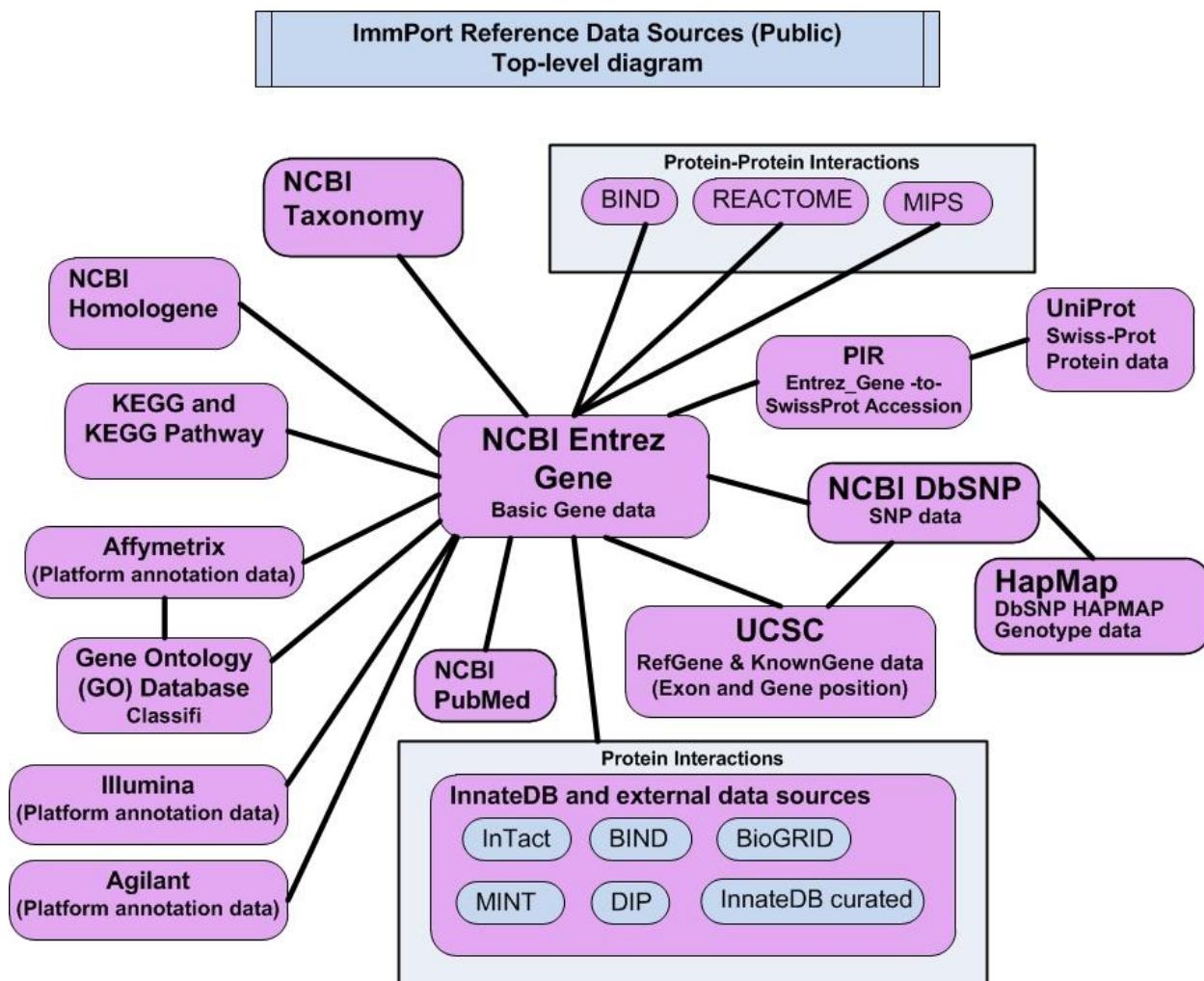


5.0 REFERENCE DATA MODELS

The reference data models in this section represent a careful selection of immunology-related data from publicly available, reputable data sources whose accessibility and integration within ImmPort serves as a value-added resource to the ImmPort user community. The following data models provide an overview of ImmPort's reference data, which are accessible to end-users via ImmPort's query interface.

Note that when possible, the designs of ImmPort's reference data tables closely resemble those of the original data sources from which they were modeled. However, data standardization (done in the ETL process) with respect to data structures, field naming, field attributes, etc. is done as needed to support the incorporation of the Reference database data within ImmPort.

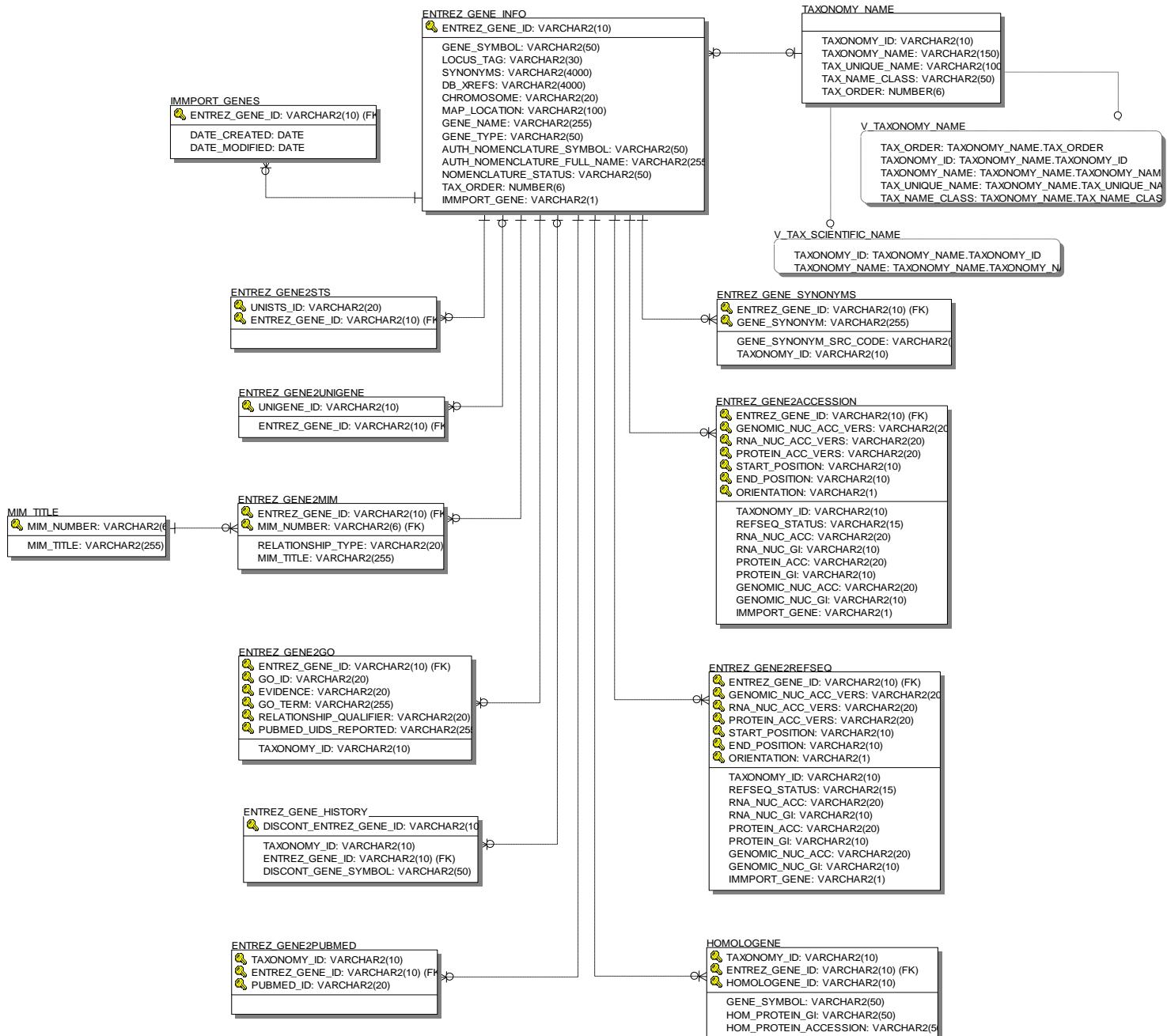
Below is a top-level diagram of Reference Data Sources in ImmPort:



5.1 NCBI ENTREZ GENE AND TAXONOMY INFORMATION

Figure 5-1 depicts the ImmPort reference data tables containing basic gene and taxonomy information. All data are extracted from NCBI Entrez Gene and NCBI Taxonomy information resources.

Figure 5-1. NCBI Entrez Gene and Taxonomy Information

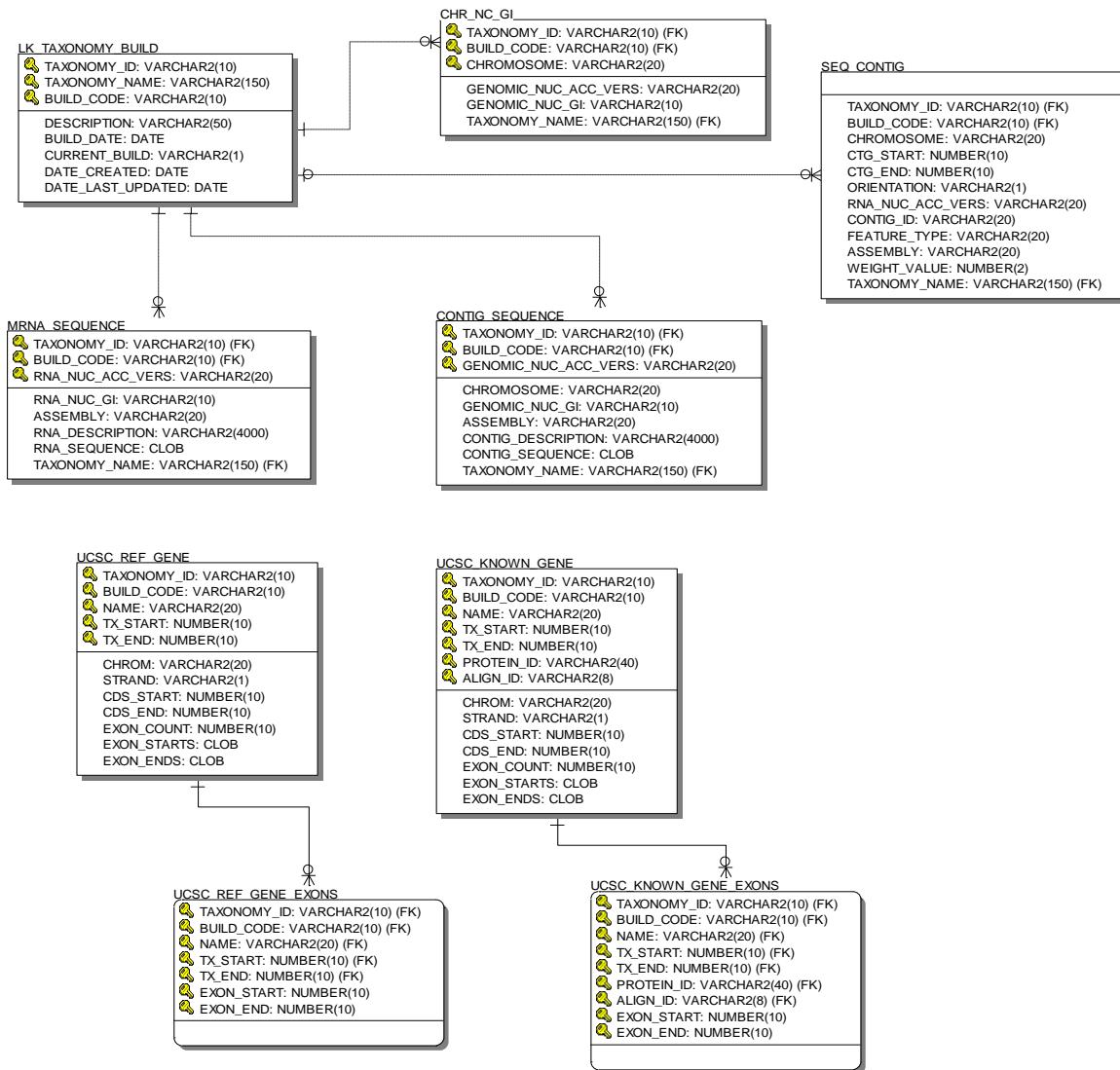


5.2 GENOMIC SEQUENCE AND GENE STRUCTURE INFORMATION

Figure 5-2 depicts the ImmPort tables containing genomic sequence and gene structure information. The data stored in these tables are extracted from both UCSC Genome Bioinformatics and NCBI Entrez Gene information resources. The LK_TAXONOMY_BUILD table and the UCSC-related tables include the taxonomy IDs and multiple builds for the selected species/taxonomies contained in ImmPort. Specifically, all UCSC RefGene and KnownGene data is provided for all taxonomies and genome builds listed below (where applicable), and is refreshed/updated monthly. These include:

TaxID	Taxonomy_Name	Genome Builds (UCSC)
9606	Homo sapiens	Hg15, Hg16, Hg17, Hg18
10090	Mus musculus	mm3, mm4, mm5, mm6, mm7, mm8, mm9
10117	Rattus rattus	
10116	Rattus norvegicus	rn2, rn3, rn4
9031	Gallus gallus	galGal2, galGal3
7227	Drosophila melanogaster	dm1, dm2, dm3
9544	Macaca mulatta	rheMac1, rheMac2
9541	Macaca fascicularis	

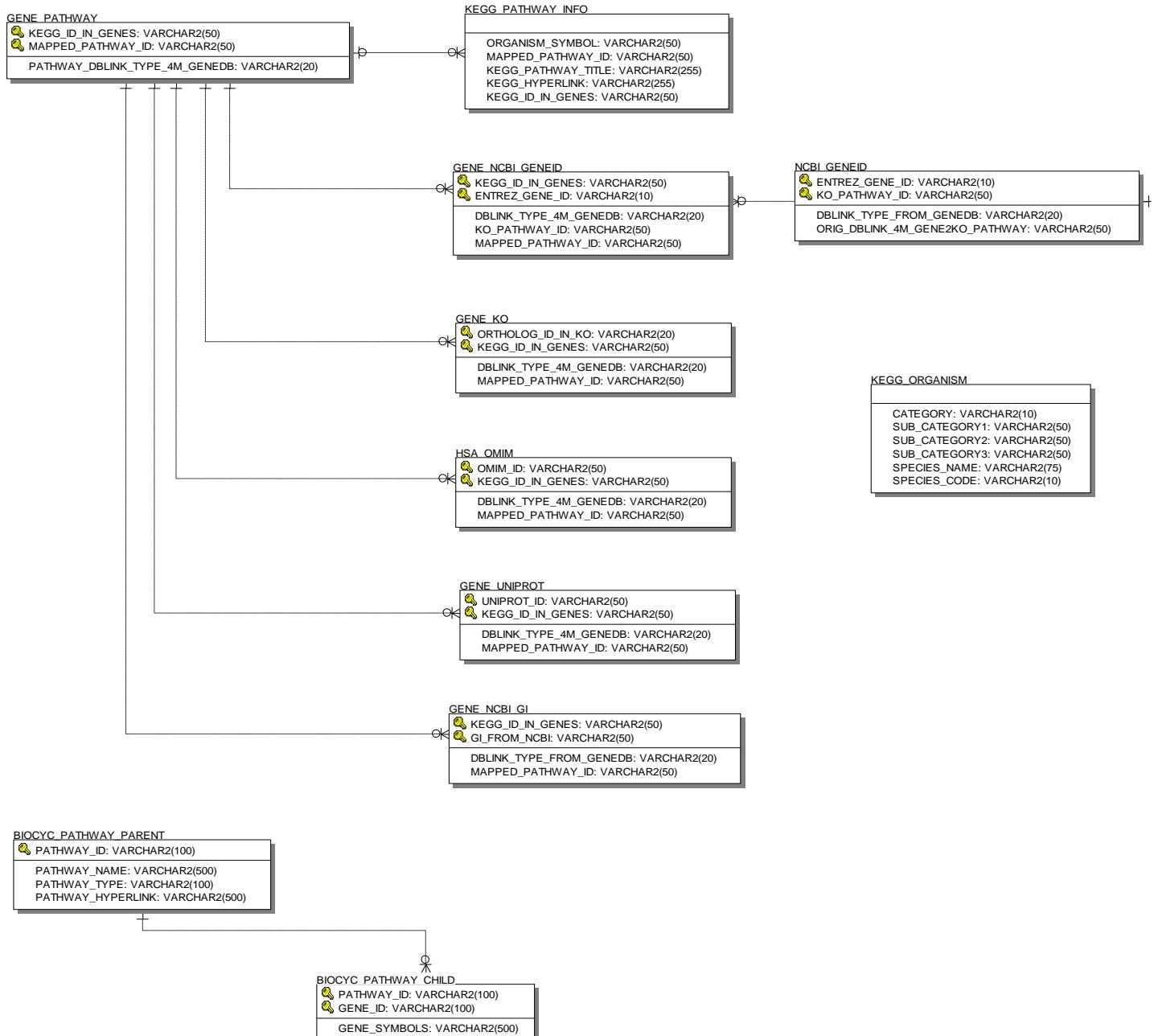
Figure 5-2. Genomic Sequence and Gene Structure Information



5.3 PATHWAY INFORMATION

The tables in Figure 5-3 store pathway data and provide linkages to NCBI gene data and genetic disorder data (via the Online Mendelian Inheritance in Man). Pathway data in ImmPort are extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Additional pathway data from BioCyc is also stored in ImmPort.

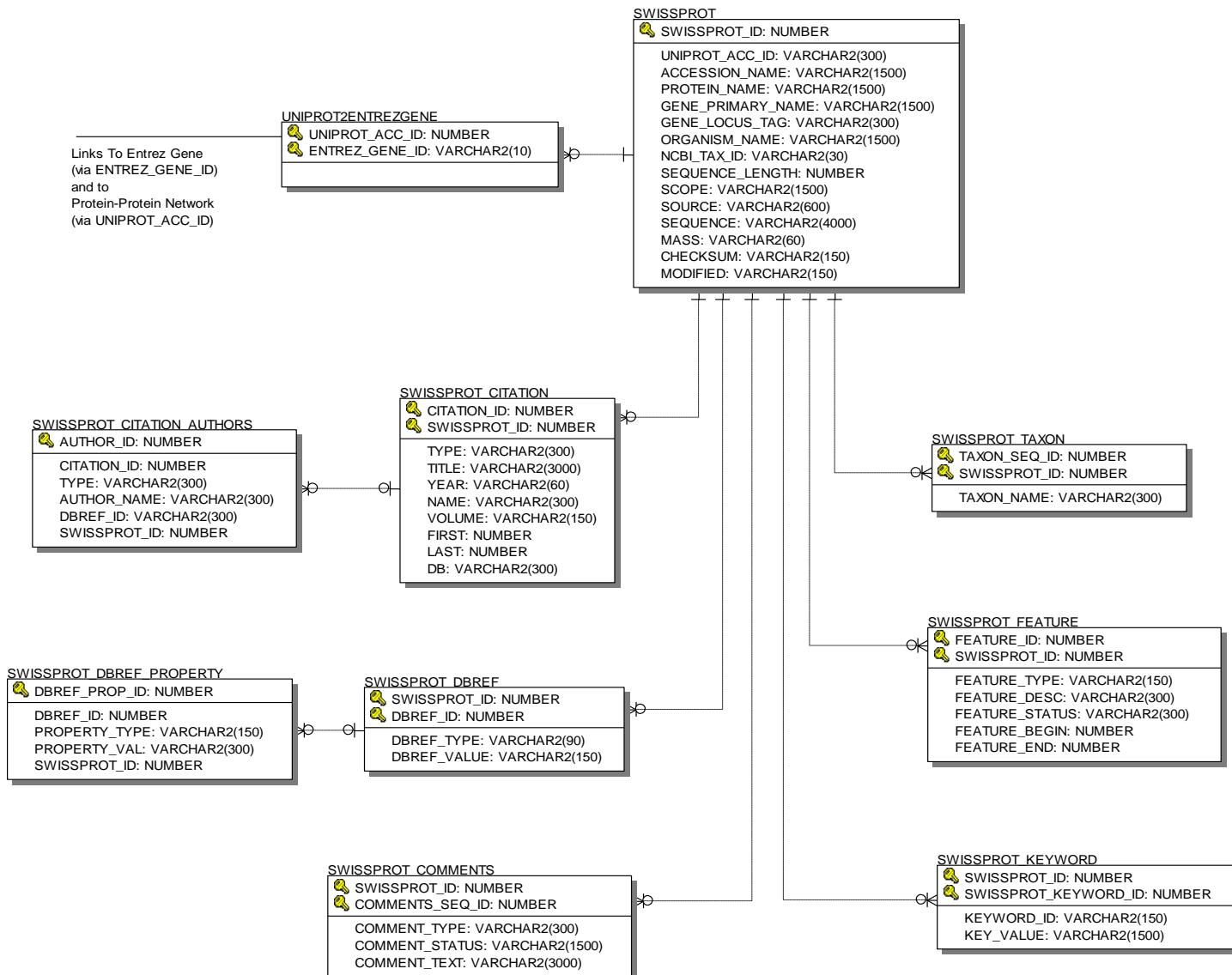
Figure 5-3. Pathway Information



5.4 BASIC PROTEIN INFORMATION

Basic protein information are extracted from the Universal Protein Resource (UniProt) and the Protein Information Resource (PIR). The Swiss-Prot data model are the result of cooperative efforts between the Bioinformatics Resource Center (BRC) and the BISC Phase II projects at Northrop Grumman Health Solutions.

Figure 5-4. Basic Protein Information



5.5 PROTEIN INTERACTIONS

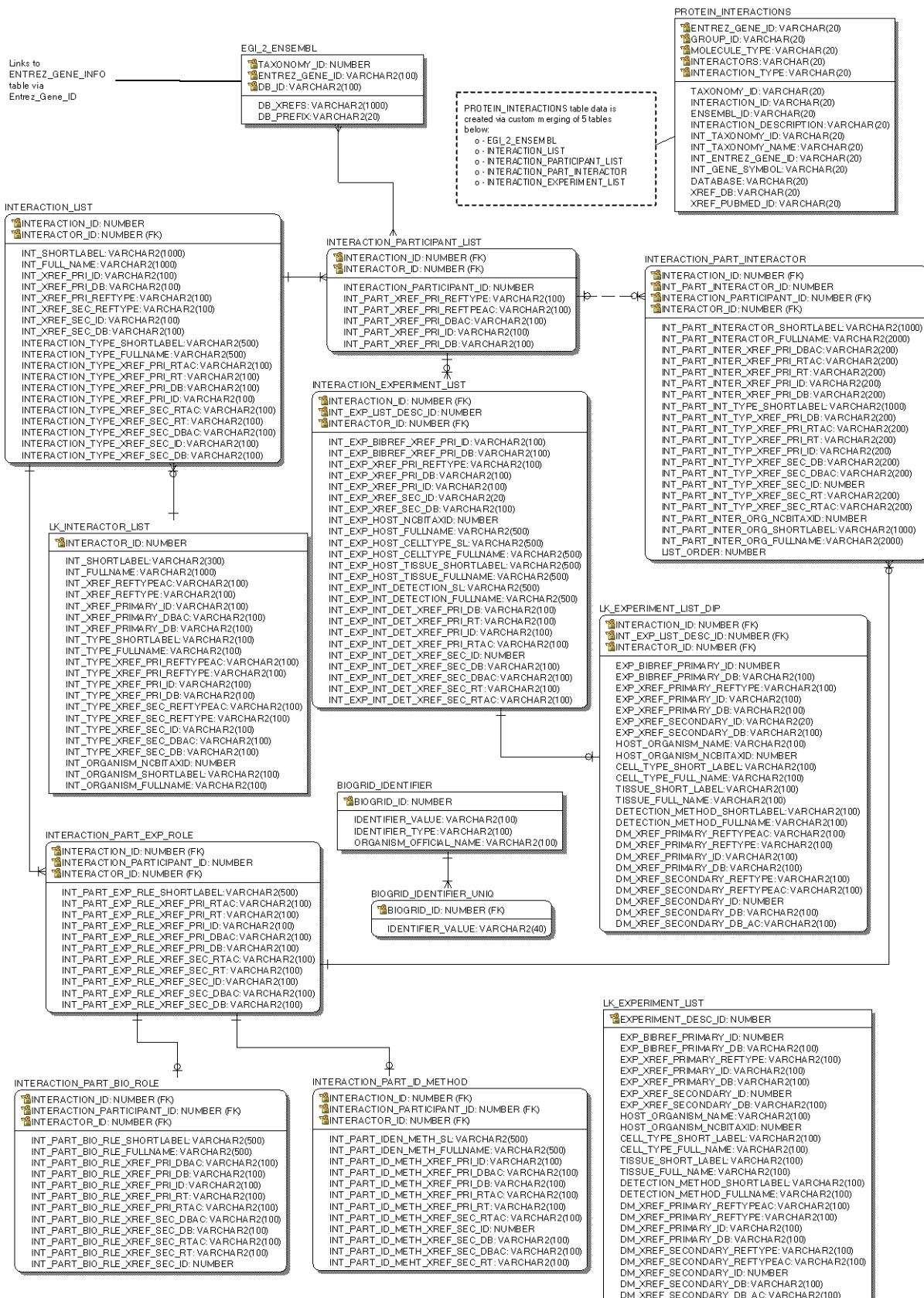
Protein interactions data, as reflected in Figure 5-5, are extracted from the InnateDB database, and ETL processing is performed to consolidate all data provided InnateDB that are made available for download in Proteomics Standards Initiative (PSI) Molecular Interaction 2.5 XML format ([PSI MI 2.5](#)). Interaction data is refreshed when data source updates become available, usually bi-annually.

The 6 data sources are:

- InTact - The IntAct database (Kerrien et al, 2007)
- MINT - Molecular Interaction database (MINT) (Chatr-aryamontri et al, 2007)
- DIP - The Database of Interacting Proteins (DIP) (Salwinski et al, 2004)
- BioGRID - The General Repository for Interaction Datasets (BioGRID) (Breitkreutz et al, 2007)
- BIND - The Biomolecular Interaction Network Database (BIND) (Alfarano et al, 2005)
- InnateDB - Curated InnateDB Data

Within ImmPort, the source content is provided as protein interactions apply to genes, and is shown as a separate block within the Reference data Gene Detail page. The PROTEIN_INTERACTIONS table is created via custom merging/consolidation of data from several tables in the Protein Interactions schema of tables. This final table is used in Reference queries, and is displayed in ImmPort as a separate block on the Gene Detail page.

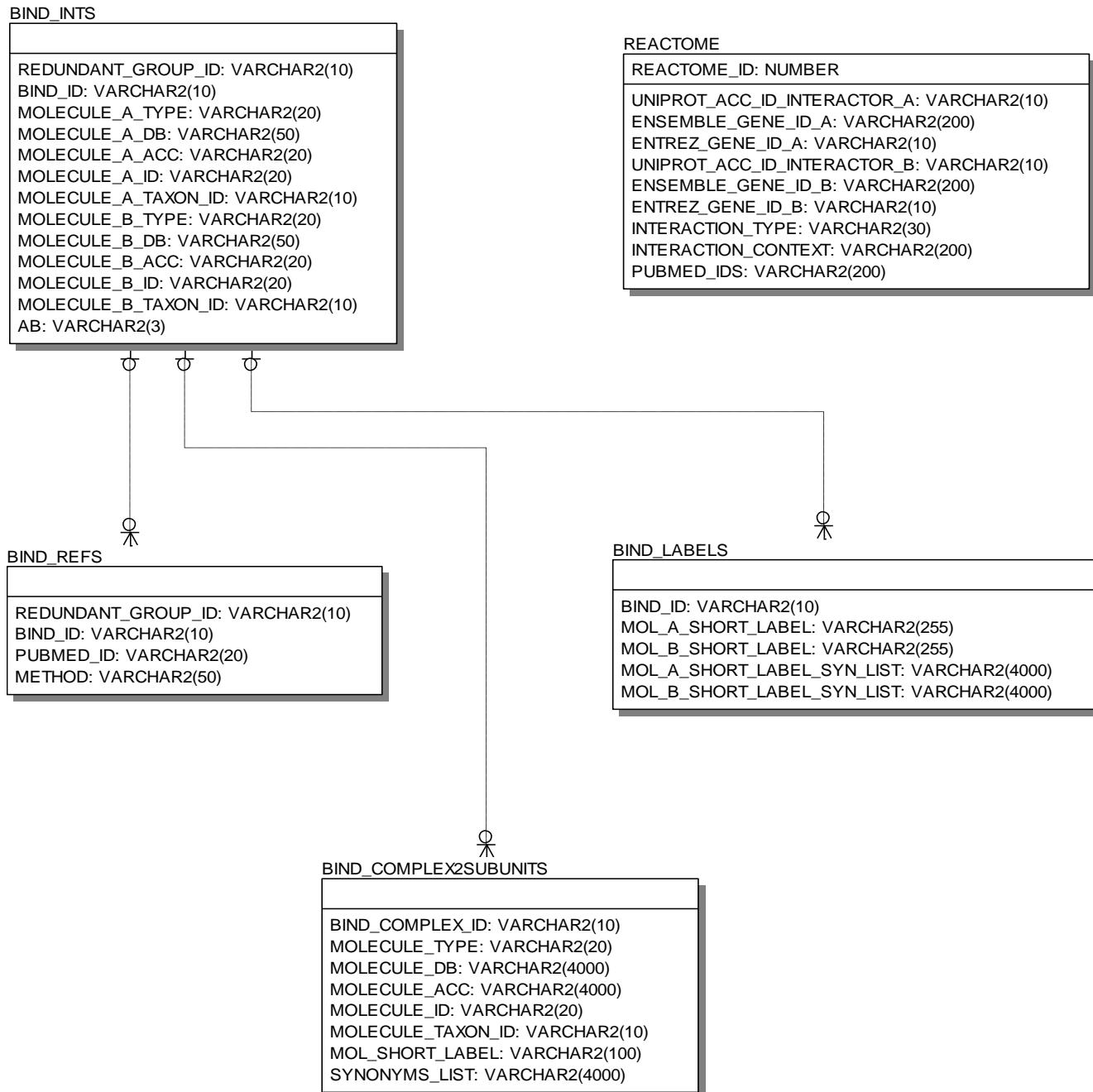
Figure 5-5. Protein Interactions



5.6 PROTEIN-PROTEIN INTERACTION NETWORK

Protein-protein interaction network data, as reflected in Figure 5-6, are extracted from the Biomolecular Interaction Network Database (BIND) and the Reactome Database. Additional protein-protein interaction data from the Munich Information Center for Protein Sequences (MIPS).

Figure 5-6. Protein-Protein Interaction Network

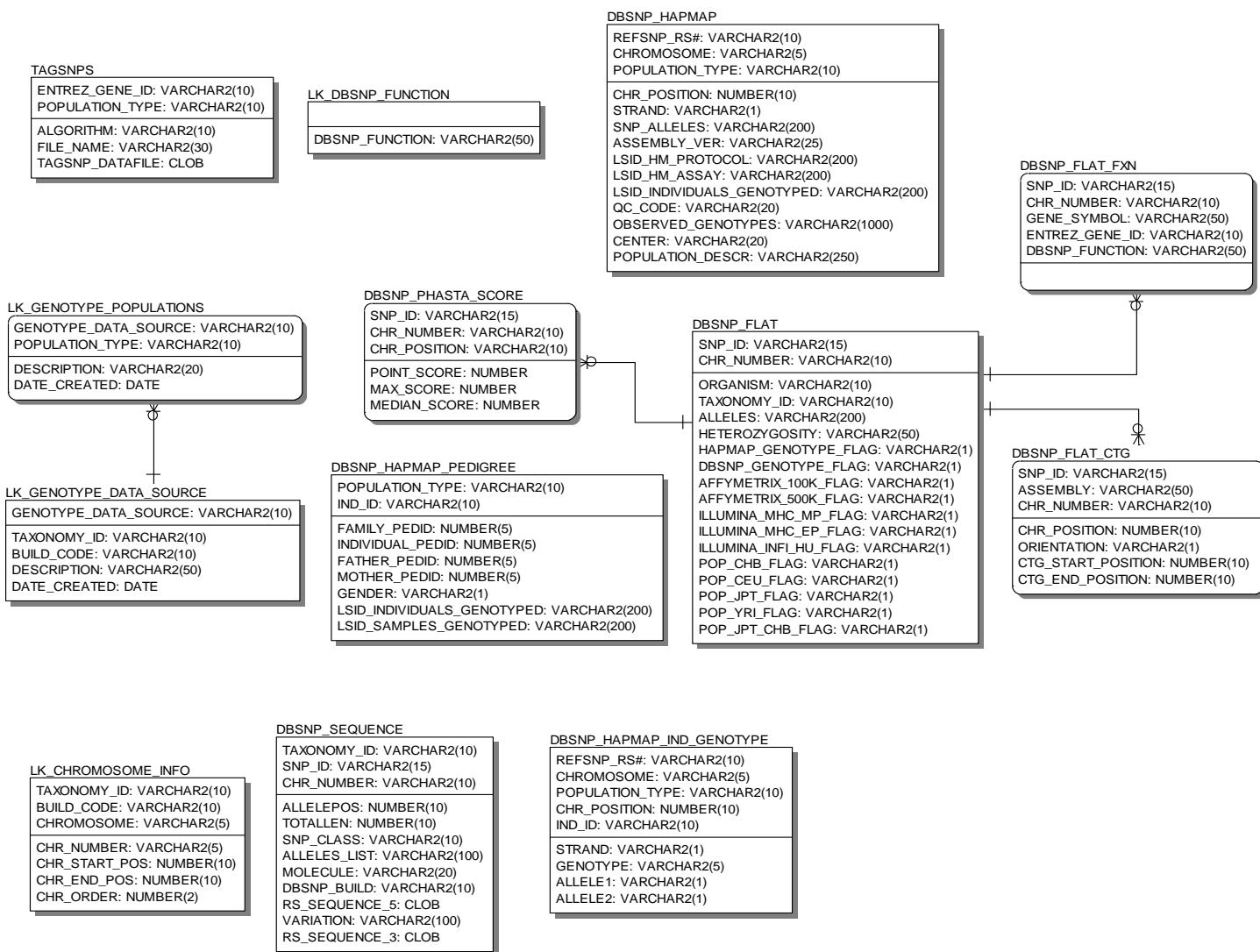


5.7 SNP POLYMORPHISM DATA

SNP polymorphism data are stored according to the data model in Figure 5-7.

Data extracted from NCBI dbSNP includes: basic SNP chromosome and position data, contigs, dbSNP functions, and dbSNP sequence data. Builds included are: NCBI genome build 35 (dbSNP build v125), and NCBI genome build 36.1 (dbSNP build 126). Data extracted from the NCBI International HapMap Project include both Phase I and Phase II data for all HapMap populations - Release 22 Genotype data (NCBI Build 36 - Hg18), and HapMap SNP data, non-redundant (dbSNP v126) for all populations as well. In addition, dbSNP phastacon score and dbSNP Hapmap pedigree information is also available.

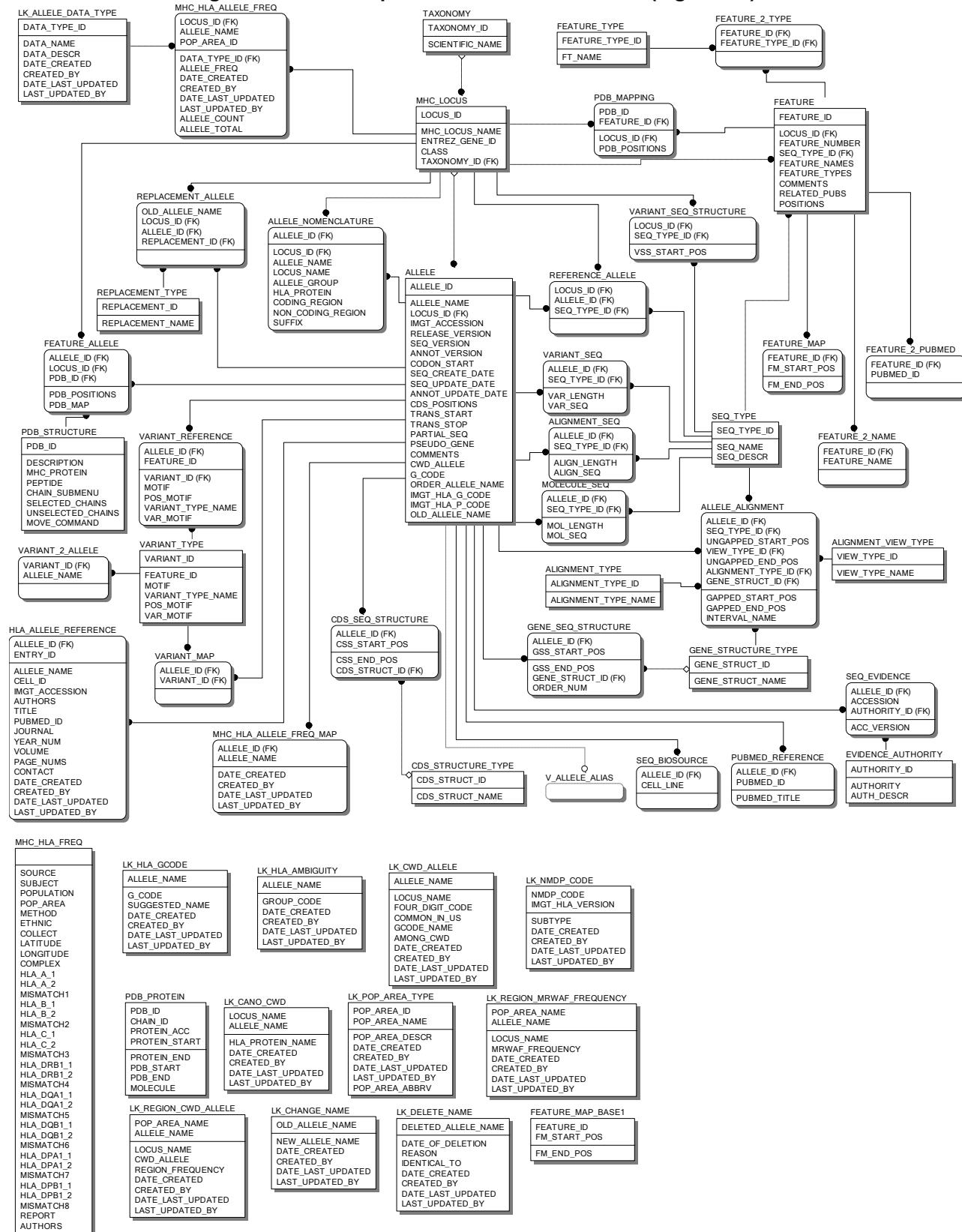
Figure 5-7. SNP Polymorphism Data



5.8 MHC AND MICROSATELLITE POLYMORPHISM DATA

The MHC data are stored according to the data model in Figure 5-8. These data are provided on a frequent basis by the IMGT HLA data source (from Anthony Nolan).

Figure 5-8. MHC Sequence Variance Data Model (high-level)



5.9 ONTOLOGY DATA MODELS

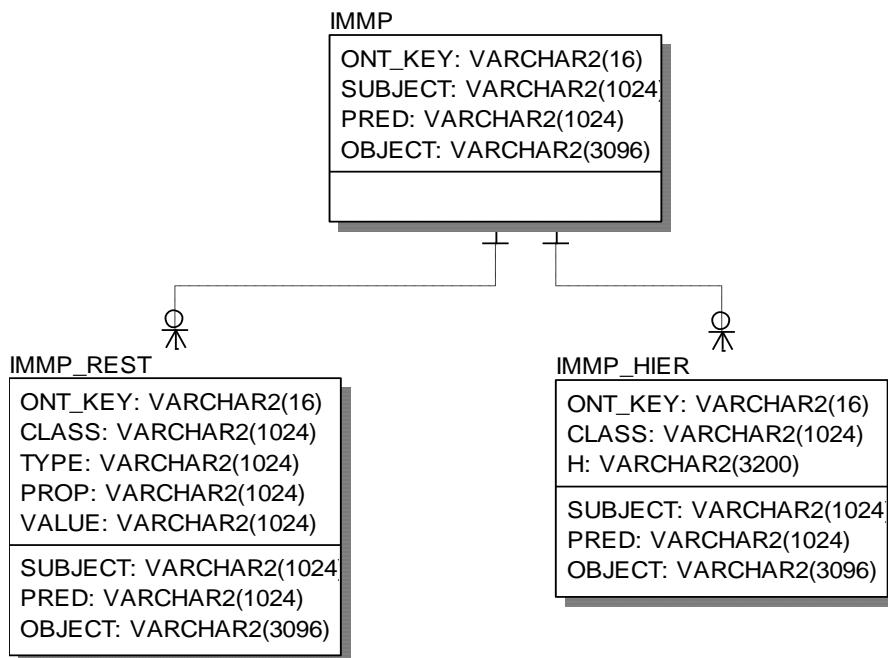
The ImmPort system maintains one database schema in Oracle 10g for the ImmPort Ontology, which includes the Ontology Browser schema. The Ontology Browser schema is actively used as part of the ImmPort system.

NOTE: All Oracle database schemas pertaining to the **Unicorn Server and Central Repository version 3.8** database are now deprecated, and have since been removed from the ImmPort system. For historical documentation and informational purposes only, all database schemas, and all data content have been archived.

5.9.1 Ontology Browser

The Ontology Browser schema, shown in Figure 5-9, consists of three partitioned tables to store the ontology data that was developed and edited in Protégé. The Ontology Browser and its query functionality has been integrated into the ImmPort system.

Figure 5-9. Ontology Browser Schema



6.0 SOFTWARE DESIGN OVERVIEW

6.1 SUMMARY OF CAPABILITIES

Extensive functionality has been built into the ImmPort web application during the duration of the contract period. Table 6-1 summarizes the capabilities and features that have been developed to date.

Table 6.1: Summary of ImmPort Capabilities and Features

#	Capabilities/Features	Capability/Feature Description
1	Manage ImmPort User	System administration capabilities, which include allowing users to request system access (register), approve registration requests, create user accounts, update user information, query user information, and deactivate users.
2	Log In/Off	Authentication and authorization capabilities, which include allowing users to login, logoff, and retrieve login information when the account or password is forgotten.
3	Manage NIAID/DAIT Programs	Program management capabilities, which include creating, deleting, modifying, searching, and viewing programs and associating NIAID/DAIT-funded contracts and grants with DAIT programs.
4	Manage NIAID/DAIT Contracts and Grants	Contract/grant management capabilities, which include creating, searching, deleting, modifying, and viewing contracts and grants and assigning a PI for a contract or grant
5	Manage Research Project (RP) /Private Project Workspace (PPW)	Manage Research Projects (RP), which include allowing a user designated as a PI or PM on a contract or grant to create a project and its associated RP, and update the information associated with the project, manage user access to the RP.
6	Manage Collaborative Project (CP)	Manage the CP capabilities, which include allowing a user designated as a PI or PM on a contract or grant to create a CP, update the information associated with the CP, control user access to the CP, and share datasets in a CP.
7	Ontology Browser	A web-based browser to view the data in the ImmPort Ontology.
8	Load Experimental Data	Batch loading of experimental data into a RP for multiple types of relevant metadata to provide the minimum information for multiple experimental assay types.
9	Browse, Simple Search, Advanced Search, Download Experimental and Clinical Research Data	Experimental and Clinical research data querying, which includes allowing users to set up highly advanced multiple attribute search criteria and to retrieve the details of the experimental data from either a RP or from the semi-public part of the ImmPort database. Query results can be saved in the users RP as lists of entities or results downloaded in multiple files formats. Result files themselves can also be downloaded real-time online, through a queued mechanism, and as a request to receive a physical data delivery.
10	Browse, Simple Search, Advanced Search, Export Reference Data	Reference data querying, which includes allowing users to set up highly advanced multiple attribute search criteria. Query results can be saved in the users RP as lists of entities or results downloaded in multiple files formats.
11	Analysis Dataset Generation and Analysis Tools	Analysis or visualization of reference data and experimental research data from the RP, CP and SPW which includes visualization tools such as GBrowse (Genome Browser), JMOL and Analysis tools including the Flow Analysis Tools, GenePattern for gene expression analysis, TagSNP Analysis Tool, HLA Typing Ambiguity Reduction tools and IGAT. Also allows for the creation of analysis data sets by combining disparate research data including interactive interfaces for creating PED files for genotyping analysis and HLA typing results.
12	Database Audit History and Archiving	Maintain and retain a complete audit history of research data (including analysis toolset data created) that is both updated and deleted. The audit history is defined as the ability to capture "who", "what", and "when" of the data involved in a change or deletion to Research data contained in the ImmPort System. Additionally, audit and collect limited summary information with respect to auditing/tracking of user session activity on a limited number of database areas. The focus is to obtain summary information on system activity such as logins, information created and updated in the several areas of the Administration Module, and usage of baseline Analysis Tools. Additionally, provide the capability to audit and track user session log information.
13	Share Research Data	Provide the capability to subset clinical and experimental metadata from RP areas to be shared into CP through an interactive user interface.

6.2 PRODUCT PERSPECTIVE

The BISC team during the contract period has developed a comprehensive suite of tools and functionality into the ImmPort system that provides useful capabilities for Immunology researchers. The system provides a means to submit, QC, store, query, combine, export and download detailed clinical and experimental research metadata across a wide spectrum of study types from human clinical trials to observational research studies with animal subjects. Novel analytical tools such as the Flow analysis tools and FLOCK have been developed and best of breed open source analytical frameworks such as GenePattern have been integrated to provide researchers the ability to aggregate results from multiple studies into datasets for immediate analysis. Tools within the HLA focused area of the system provide novel capabilities to researchers performing HLA typing to perform quality control, allele and genotype ambiguity reduction, and to convert their results into respective sequence features for further analysis. Finally, the reference data content from public resources are collected, aggregated, curated (as in the ImmPort Gene lists) and made available through query and visualization tools to provide users already visiting ImmPort capabilities in one location instead of requiring navigation away from ImmPort to other public resources.

6.3 SOFTWARE DESIGN PRESENTATION

The intent of the Software Design section of this document is to provide a roadmap for the design and code organization for the functionality developed within the ImmPort system. It is not intended as a comprehensive low level design of the system but to provide a technical professional the ability to understand the nature of ImmPort and how the interfaces presented map to the underlying code tiers.

The remainder of the document is divided into functional groups (Sections 7-23) with each functional group consisting of sets of design packages. Each package represents a particular system feature (or set of features) and is documented with a screen shot of the web interface along with a functional description, a reduced class diagram illustrating the relevant methods, and one or more sequence diagrams illustrating the code flow from the front end tier to the persistence layer and back. The intention is to highlight unique design areas in ImmPort, so in some cases one representative design package is used to illustrate the approach for multiple areas. In the closing section for each functional group a class diagram for the presentation (Struts Actions), business model (EJB) and persistence (DAO + Hibernate) layers are provided to put the earlier design packages into the large class perspective.

6.4 USER COMMUNITIES

The ImmPort system accepts four types of communities which include DAIT Community, DAIT-Funded Research Community, Life Science Researcher, and Public Users. The DAIT Community contains two role types, DAIT BISC Project Officer and DAIT Users. The DAIT BISC Project Officer has permissions to manage users, programs, contract/grants, and view system reports. The DAIT user has the permissions to view the users, programs, and contracts/grants. The DAIT-Funded Research Community contains three role types, Principal Investigator, Project Manager and Other Staff. The Life Science Research Community has access to their own private project that allows data to be uploaded and analyzed using the ImmPort visualization tools. Currently, the Life Science Research community does not have the ability to associate users into their private project workspace, but the plan is to allow this in the future. The public user community has access to the public areas of ImmPort which include the reference queries.

Table 6-1 lists the ImmPort functionalities and capabilities for each user role type that resides in the system.

Table 6.1 User Functionality and Capability Matrix

Functionality	Capability/Permission	DAIT Community		DAIT Funded Research Community	Life Science Research	Public User
		DAIT BISC Project Officer	DAIT User	All DAIT Funded Research Users	Life Science Research	Public User
Manage Users	Register User (self)	✓	✓	✓	✓	✓
	Approve/Reject Registration	✓				
	Create User	✓				
	Query User	✓	✓			
	View/Update User (self)	✓	✓	✓	✓	
	Deactivate User	✓				
	Manage User Role	✓				
	Delete User	✓				
Log In/Off	Log In	✓	✓	✓	✓	
	Retrieve Login Information	✓	✓	✓	✓	
	Log Off	✓	✓	✓	✓	
	Change Password (self)	✓	✓	✓	✓	
	Reset Password (others)	✓				
Manage Contracts/Grants	Create Contract/Grant	✓				
	Deactivate Contract/Grant	✓				
	Modify Contract/Grant	✓				
	Search for Contract/Grant	✓	✓			
	View Contract/Grant	✓	✓	✓	✓	
Manage Programs	Create Program	✓				
	Deactivate Program	✓				
	Modify Program	✓				
	Search for Program	✓	✓			
System Reports	System Reports	✓				
Use Analysis Tools	tagSNP	✓	✓	✓	✓	
	Genetic Expression			✓	✓	
	FLOCK			✓	✓	
	Genetic Analysis			✓	✓	
	MHC Analysis			✓	✓	
	GBrowse	✓	✓	✓	✓	
	Resources	✓	✓	✓	✓	✓
	News & Events	✓	✓	✓	✓	✓
Manage Saved Lists	Export Experiment Data			✓	✓	
	Save Private Data	✓		✓	✓	

Functionality	Capability/Permission	DAIT Community		DAIT Funded Research Community	Life Science Research	Public User
		DAIT BISC Project Officer	DAIT User	All DAIT Funded Research Users	Life Science Research	Public User
	Delete Private Data	✓		✓	✓	

6.5 PROJECT ROLES

Table 6-2 lists the ImmPort functionalities and capabilities for the DAIT Funded Research community that resides in the system. The DAIT-Funded Research Community contains three role types, Principal Investigator, Project Manager and Other Staff, which are assigned to a contract/grant. The Principal Investigator and Project Manager obtain the permissions to create, update, and view projects while associating all types of users to their projects. Once a project is created, data is uploaded into the project by users associated to the project so that analysis may be conducted utilizing the ImmPort visualization tools. The Other Staff obtains access to the assigned projects which allows them to upload and view data while utilizing the ImmPort visualization tools.

Table 6.2 ImmPort Project Roles and Functionalities

Functionality	Capability/Permission	DAIT Funded Research		
		PI	PM	OS
Manage Private Project Workspace	Create Private Project Workspace	✓	✓	Available to be added
Manage Collaborative Project	Create Collaborative Project Workspace	✓	✓	Available to be added
Manage Private Project Data	Create Private Project Workspace	✓	✓	
	Add/Remove User Access	✓	✓	
	Download Project Summary	✓	✓	✓
	Upload Data into Private Project Workspace	✓	✓	✓
	Update Data in Private Project Workspace	✓	*	*
	Delete/Update Data from User's View	✓	*	*
	Publish Data to SPW	✓		
	Share Data to Collaborative Data Area	✓	✓	✓
Manage Collaborative Project Data	Create/Edit Collaborative Project Workspace	✓	✓	
	Add/Remove User Access	✓	✓	
	Download Collaborative Project Summary	✓	✓	✓
	Delete Data in Collaborative Project from User's View	✓	*	*

* - if the user submitted the data originally

7.0 MANAGE USER DESIGN PACKAGES

The Manage User design artifacts model system administration capabilities, which include allowing users to request system access (register), approving registration requests, creating user accounts, updating user information, querying users, and deactivating users from the system.

7.1 CREATE/VIEW/UPDATE/SEARCH USER

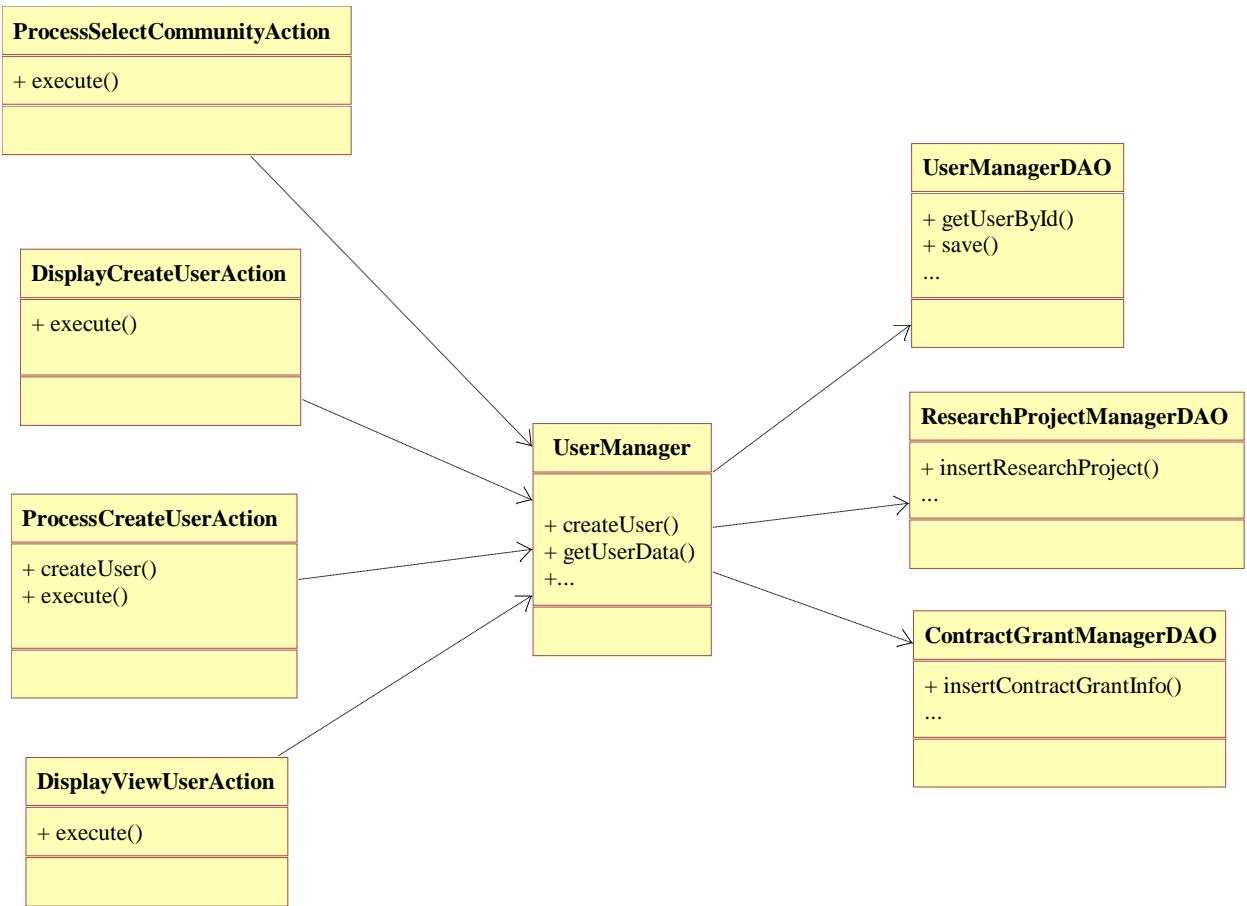
The Create/View/Update/Search User depicts the events for creating/viewing/updating/searching a user account. The system collects the user information and creates a user account, provided the submitted data are valid. If the system encounters errors, then it will display the appropriate message. Once the user information is collected the system will display the data while giving the user an option to update the information.

The screenshot shows a web-based application interface for creating a new user account. The title bar reads "Admin / Create User". Below the title bar is a navigation menu with links: My Profile, Programs, Grants - Contracts, Users (selected), and Reports. The main content area is titled "Create User: Life Science Research Community" and includes a note that fields marked with an asterisk (*) are required. The form contains the following fields:

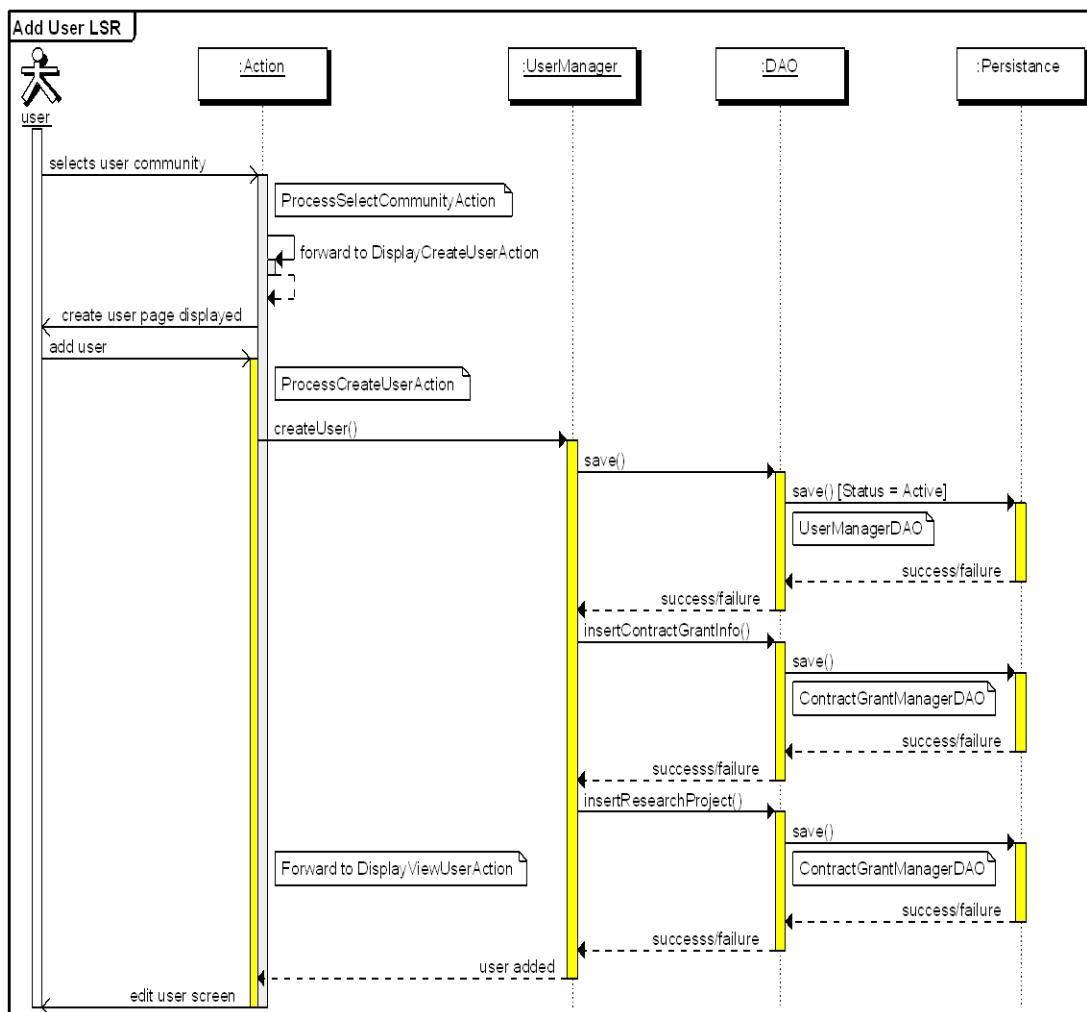
- ImmPort Login Name*: (up to 12 characters allowed) [Text input field]
- Password*: [Text input field with placeholder text: "Please enter a minimum of 8 characters for your password."]
- Re-Enter Password*: [Text input field]
- Last Name*: [Text input field]
- First Name*: [Text input field]
- Middle Initial: [Text input field]
- Organization*: [Text input field]
- Department: [Text input field]
- Principal Investigator Name: [Text input field]
- Phone Number: (US/Canada xxx-xxx-xxxx)* [Text input field] [Radio buttons for US and Canada or International]
- Email Address*: [Text input field]
- How can ImmPort assist your research efforts?: [Text input field]
- How did you learn of ImmPort*: [Text input field] [Dropdown menu showing "Colleague"]

At the bottom of the form are two buttons: "Create" and "Cancel".

7.1.1 Create New User Class Diagram



7.1.2 Create New User Sequence Diagram



7.2 USERS REGISTRATION

The User Registration depicts the events for requesting access to ImmPort. First, the individual initiates a registration request.

This screenshot shows the ImmPort registration process. At the top, the ImmPort logo is displayed with the tagline 'BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY'. A navigation bar includes links for 'About ImmPort', 'Access Data', 'Tools', 'Resources', and 'News & Events'. Below the navigation bar, a section titled 'Register User: Notice' contains a detailed text about access restrictions and a link to a 'data sharing and access agreement'. At the bottom of this section are 'Continue' and 'Cancel' buttons. The footer of the page includes links for 'Privacy Policy', 'Disclaimer', 'Accessibility', 'Contact Us', and 'Site Map', along with acknowledgments for 'National Institute of Allergy and Infectious Diseases (NIAID)', 'National Institutes of Health (NIH)', and 'Health and Human Services (HHS)'.

Depending upon the type of access requested, the system prompts the individual for additional information. The user may select to be a general member of the life sciences community, a member of the DAIT-funded research community, or an extramural NIH employee.

This screenshot shows the 'Select Community' step in the ImmPort registration process. The top navigation bar and footer are identical to the previous screenshot. The main content area is titled 'Register User - Select Community'. It lists three options with radio buttons: 'Life Science Research Community' (selected), 'Division of Allergy, Immunology, and Transplantation (DAIT) Funded Research Community', and 'Extramural NIH Employee'. Each option has a corresponding explanatory text block. At the bottom are 'Continue' and 'Cancel' buttons. The footer includes links for 'Privacy Policy', 'Disclaimer', 'Accessibility', 'Contact Us', and 'Site Map', along with acknowledgments for 'National Institute of Allergy and Infectious Diseases (NIAID)', 'National Institutes of Health (NIH)', and 'Health and Human Services (HHS)'.

The system validates the user information and saves the registration request, provided the data submitted are valid.

The screenshot shows a registration form for 'Life Science Research Community'. The form includes fields for ImmPort Login Name, Password, Re-Enter Password, Last Name, First Name, Middle Initial, Organization, Department, Principal Investigator Name, Phone Number, Email Address, Password Retrieval Question, Password Retrieval Answer, and How can ImmPort assist your research efforts. Error messages are displayed for the password and phone number fields. Buttons for 'Register' and 'Cancel' are at the bottom.

Register User: Life Science Research Community ⓘ

Fields marked with an asterisk * are required.

ImmPort Login Name *
(up to 12 characters allowed)
johnd

Please enter a minimum of 8 characters for your password.

Password *
Re-Enter Password *

Last Name *
Doe

First Name *
John

Middle Initial

Organization *
BISC

Department

Principal Investigator Name

Phone Number
(US/Canada xxx-xxx-xxxx) *
301-527-0000

ImmPort recommends using an email address from an education or research institution to expedite registration approval.
Email Address *
johndoer@research.org

What is your favorite color?
blue

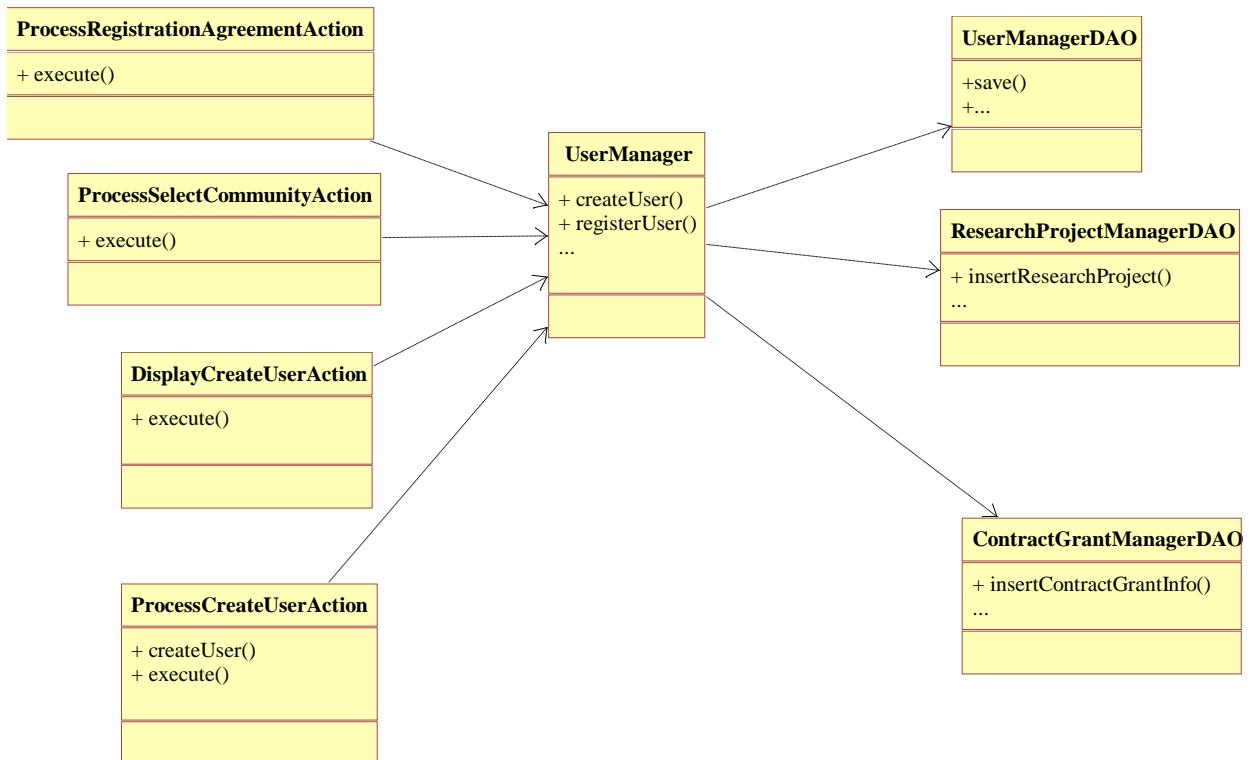
How did you learn of ImmPort?
Colleague

How can ImmPort assist your research efforts?
Research

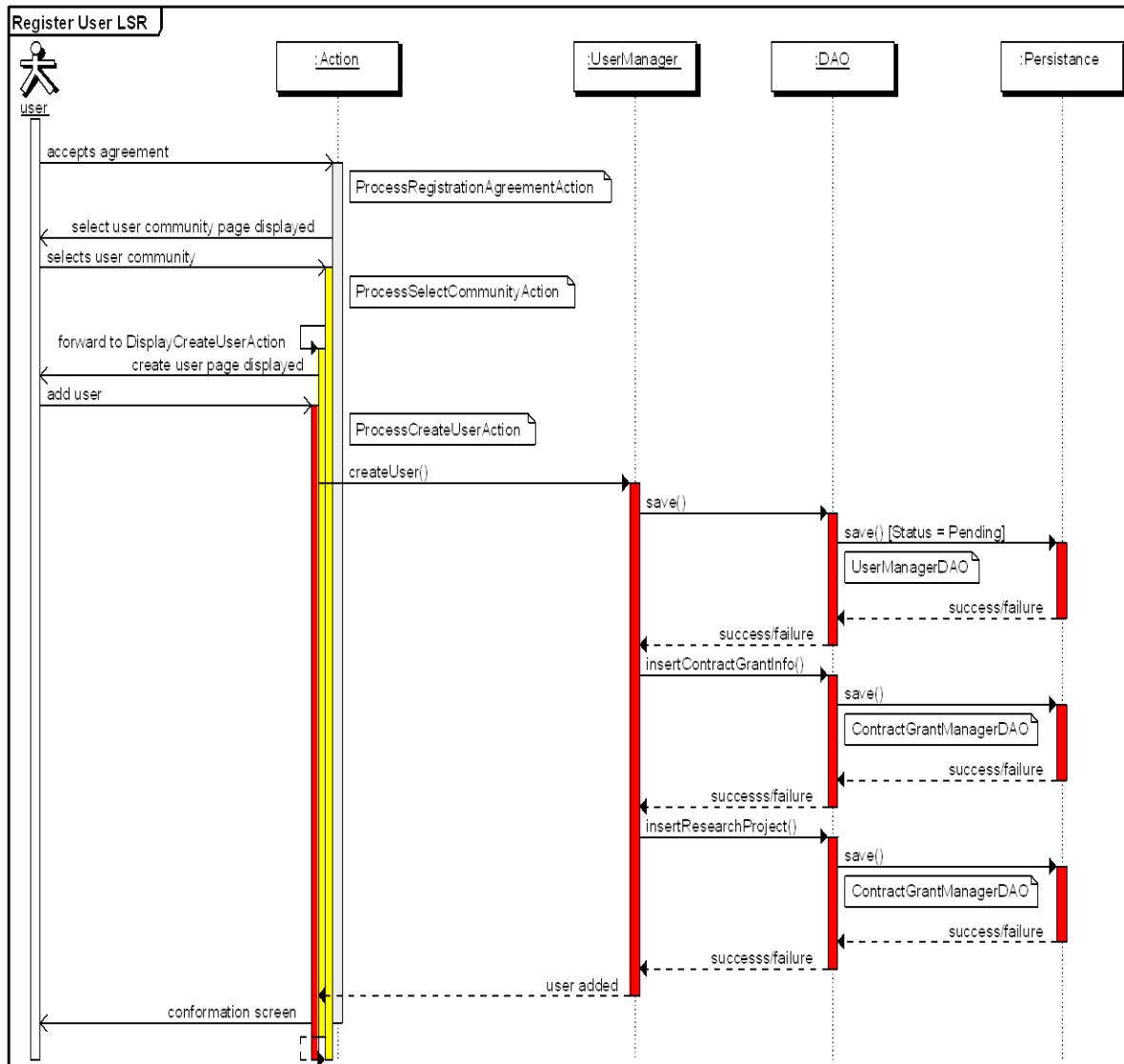
US and Canada International

Appropriate error messages are displayed if errors are encountered. At the end of the process, the system notifies those users responsible for reviewing registration requests.

7.2.1 User Registration Class Diagram



7.2.2 User Registration Sequence Diagram



7.3 APPROVE/REJECT USER

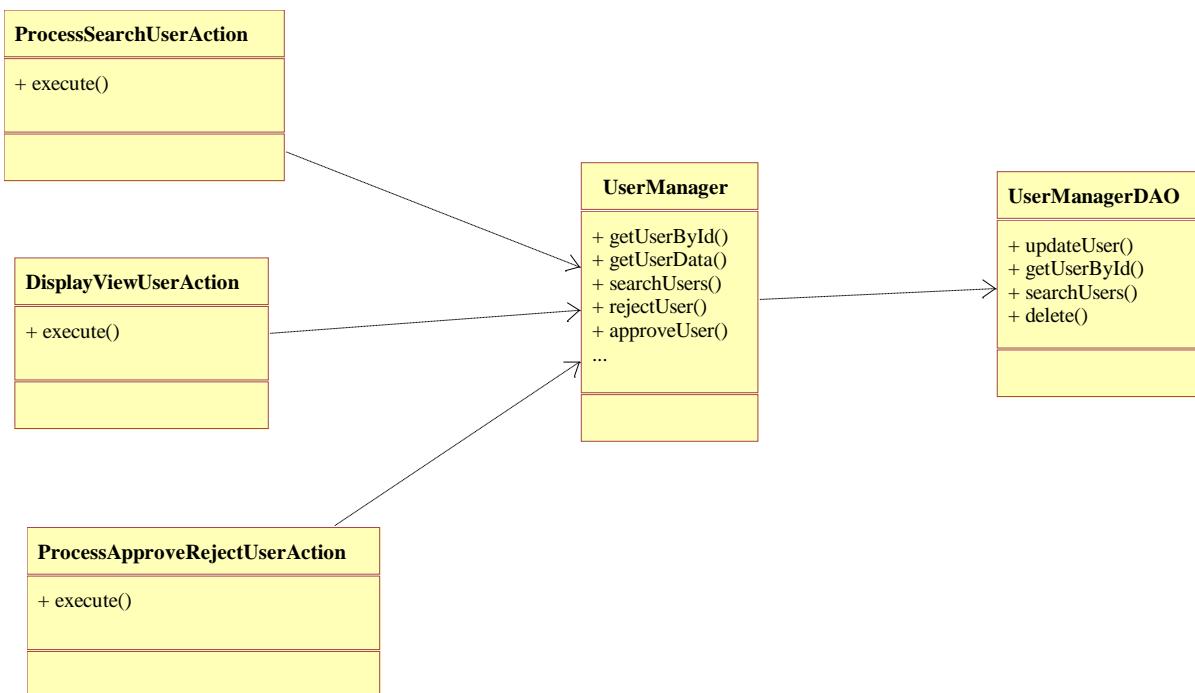
The Approve/Reject User depicts the events for approving/rejecting a user account. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web-based administrative interface titled "Admin / Review Registration Request". The page includes a navigation bar with links for "My Profile", "Programs", "Grants - Contracts", "Users", and "Reports". A note at the top states: "Fields marked with an asterisk * are required." The form contains the following fields:

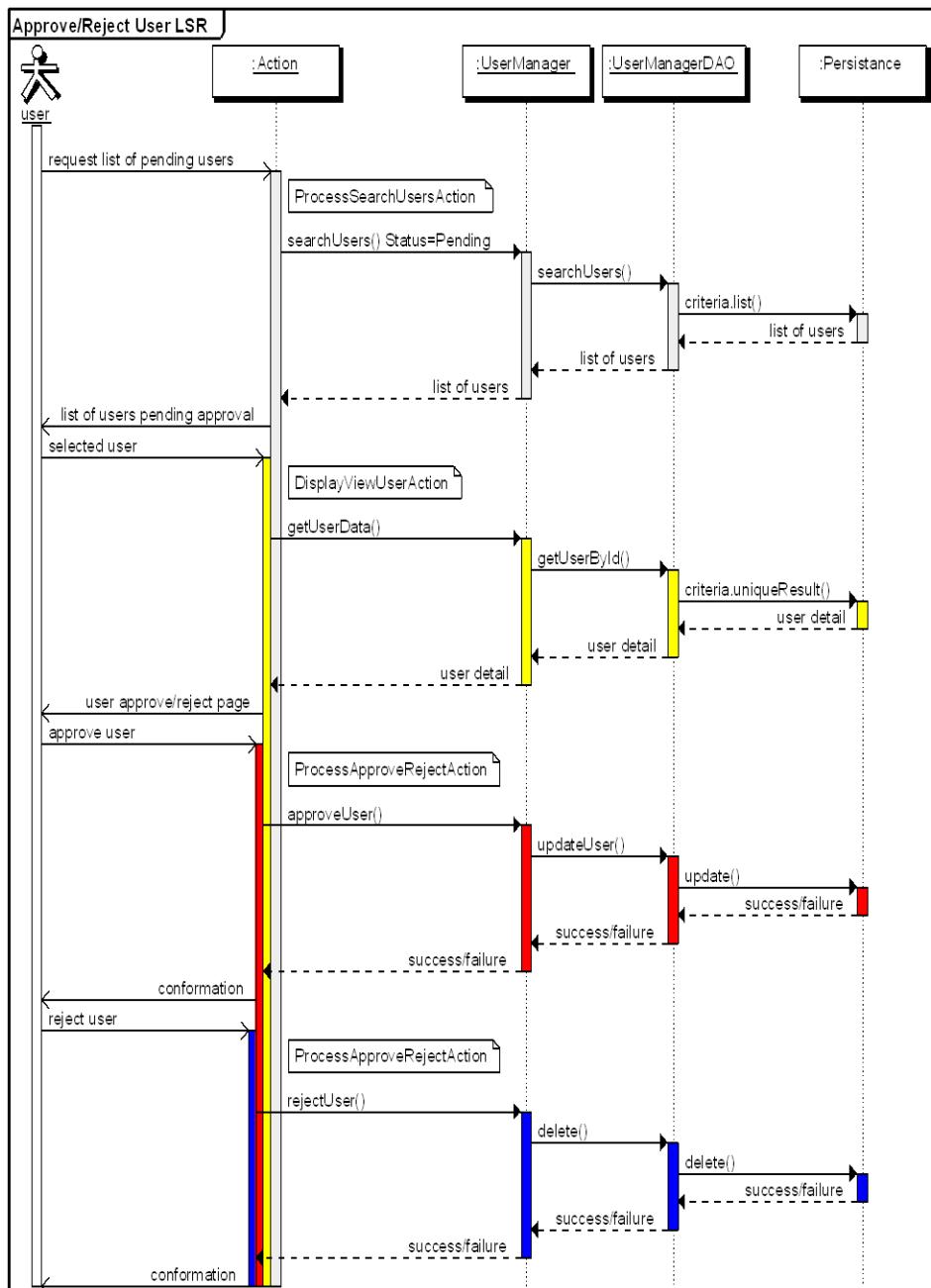
User ID	johnd
Last Name *	Doe
First Name *	John
Middle Initial	
Organization *	BISC
Department	
Principal Investigator Name	
Phone Number (US/Canada xxx-xxx-xxxx) *	<input checked="" type="radio"/> US and Canada <input type="radio"/> International 301-527-0000
ImmPort recommends using an email address from an education or research institution to expedite registration approval.	
Email Address *	john.doe@research.org
How did you learn of ImmPort*	Colleague
How can ImmPort assist your research efforts	Research
User Community*	Life Science Research Community

At the bottom of the form are three buttons: "Approve", "Reject", and "Cancel".

7.3.1 Approve/Reject User Class Diagram

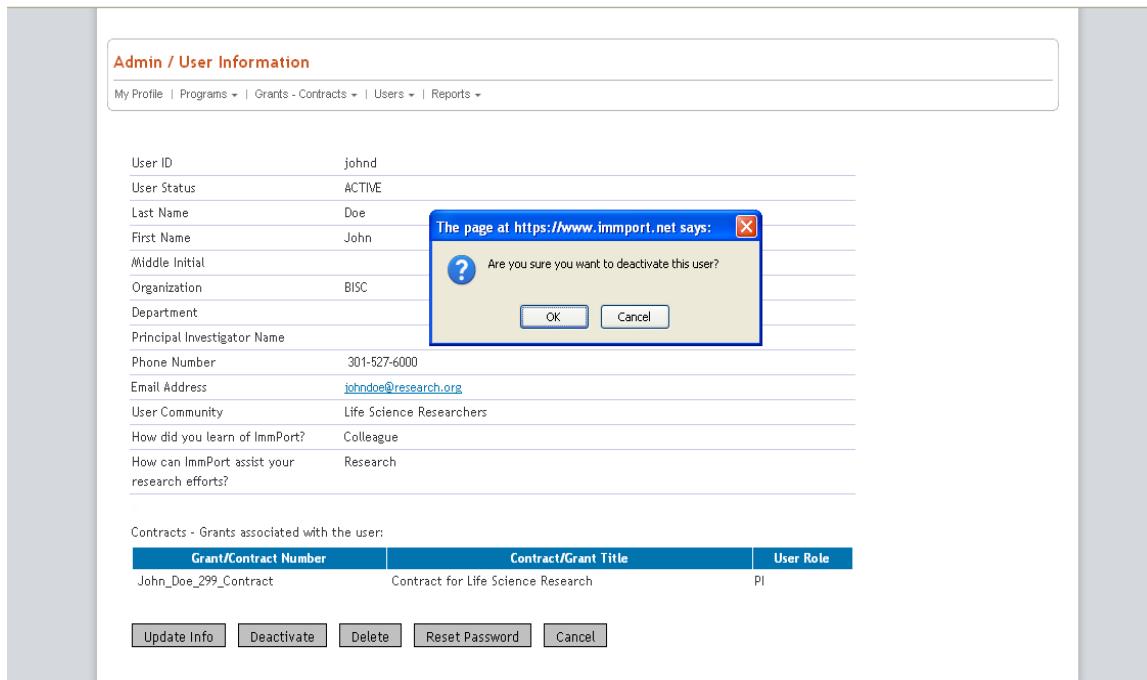


7.3.2 Approve/Reject User Sequence Diagram

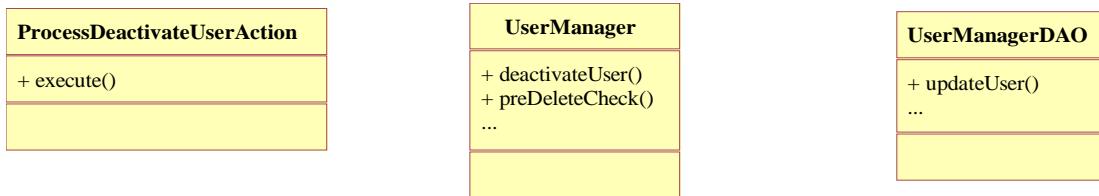


7.4 DEACTIVATE USER

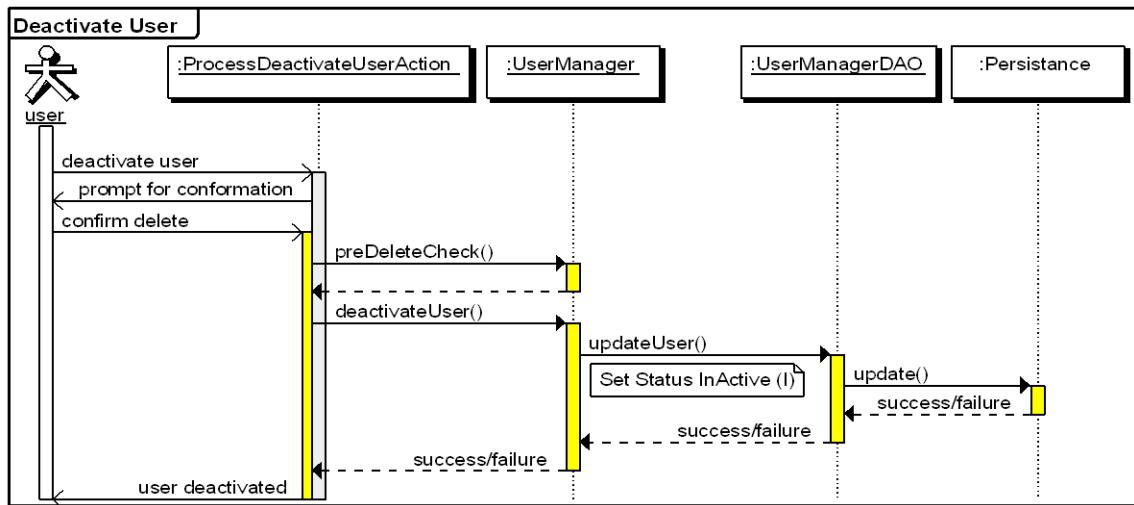
The Deactivate depicts the events for deactivating a user account. Appropriate error messages are displayed if errors are encountered.



7.4.1 Deactivate User Class Diagram



7.4.2 Deactivate User Sequence Diagram



7.5 ACTIVATE USER

The activate user depicts the events for activating a user account that has been deactivated in the past. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web-based administrative interface titled "Admin / User Information". At the top, there is a navigation bar with links: "My Profile", "Programs", "Grants - Contracts", "Users", and "Reports". Below the navigation, a message states: "User was deactivated by patty-po on 2010-10-26 16:04:28.0". The user profile information is listed in a table:

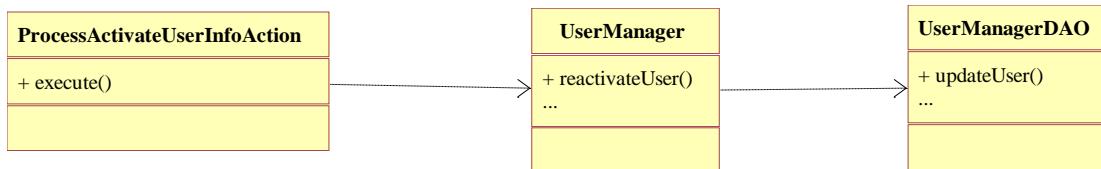
User ID	johnd
User Status	INACTIVE
Last Name	Doe
First Name	John
Middle Initial	
Organization	BISC
Department	
Principal Investigator Name	
Phone Number	301-527-6000
Email Address	johndoe@research.org
User Community	Life Science Researchers
How did you learn of ImmPort?	Colleague
How can ImmPort assist your research efforts?	Research

Below the profile table, there is a section titled "Contracts - Grants associated with the user:" with a table:

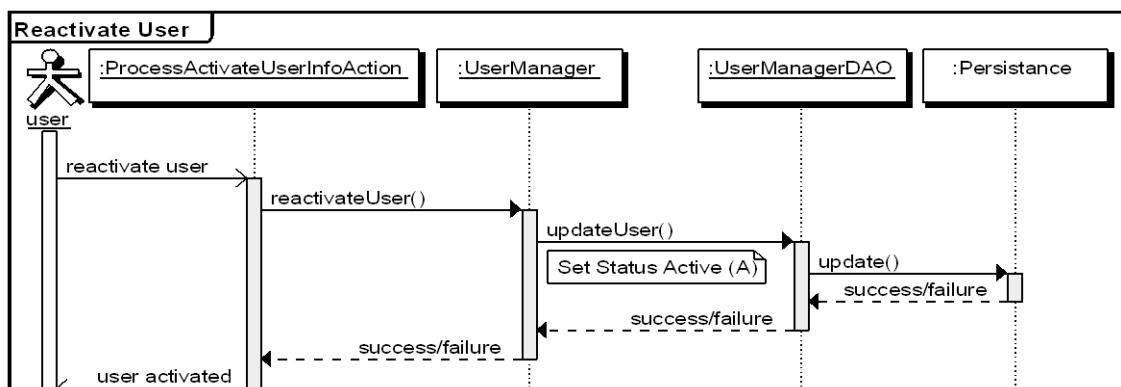
Grant/Contract Number	Contract/Grant Title	User Role
John_Doe_299_Contract	Contract for Life Science Research	PI

At the bottom of the page is a single button labeled "Activate User".

7.5.1 Activate User Class Diagram



7.5.2 Activate User Sequence Diagram



7.6 SEARCH USERS

The search user depicts the events for searching a user account. The user is first presented with a search screen.

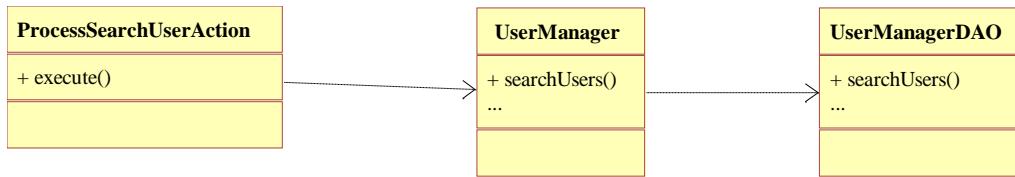
Last Name: Doe
First Name:
Grant/Contract Number:
Status: Active Inactive Active & Inactive Pending

[Submit](#) [Reset](#)

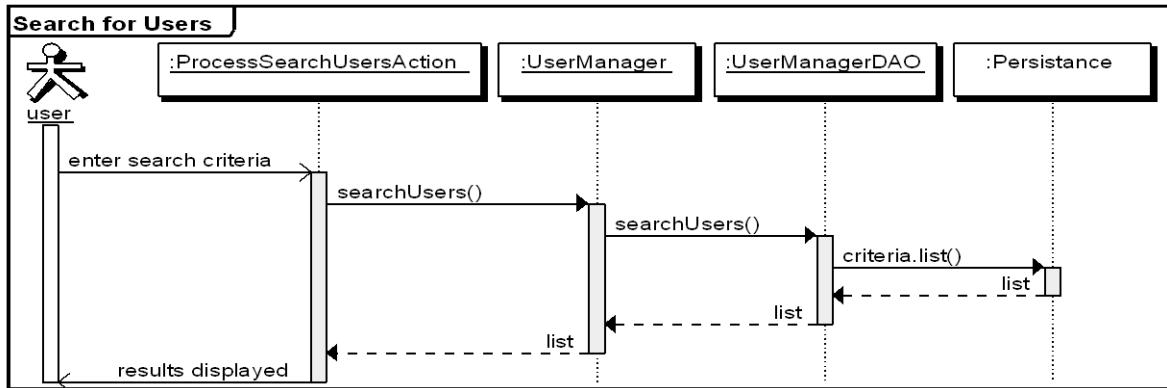
After the search, results are presented and the user may click to view more details about the selected user.

Last Name	First Name	MI	Status	Organization	Agreement Acceptance	Data Agreement Date	Community
Doe	Jane		ACTIVE	University			Life Science Researchers
Doe	John		ACTIVE	BISC			Life Science Researchers

7.6.1 Search Users Class Diagram



7.6.2 Search Users Sequence Diagram



7.7 FORGOT PASSWORD

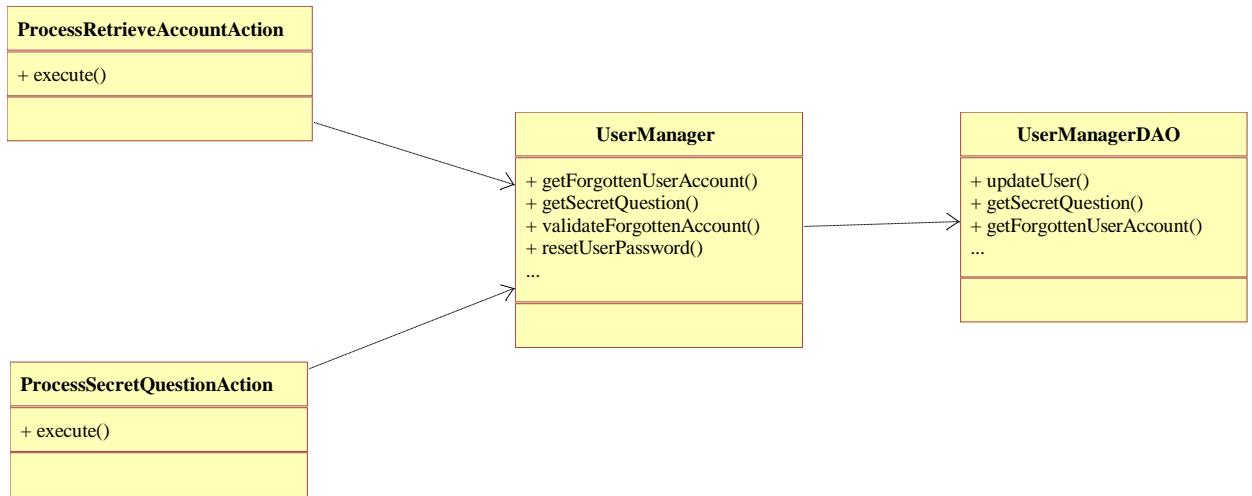
The Forgot Password Sequence Diagram depicts the sequence of events for a user that has forgotten their username/password.

The screenshot shows the ImmPort website's login page. At the top, there is a logo for ImmPort with the tagline "BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY". Below the logo is a navigation bar with links: Home, About ImmPort, Access Data, Tools, Resources, and News & Events. The main content area is titled "Retrieve User Name and Password" in red. It contains a message: "Please enter the following information to confirm your identity. The security question you provided during registration will be displayed." Below this, it says "Fields marked with an asterisk * are required." There are four input fields: "First Name *" with value "John", "Last Name *" with value "Doe", "Middle Initial" with empty value, and "Registered Email Address *" with value "johndoe@research.org". At the bottom of the form are two buttons: "Submit" and "Cancel". At the very bottom of the page, there is footer text: "Sponsored by: National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Health and Human Services (HHS)" and links to "Privacy Policy", "Disclaimer", "Accessibility", "Contact Us", and "Site Map".

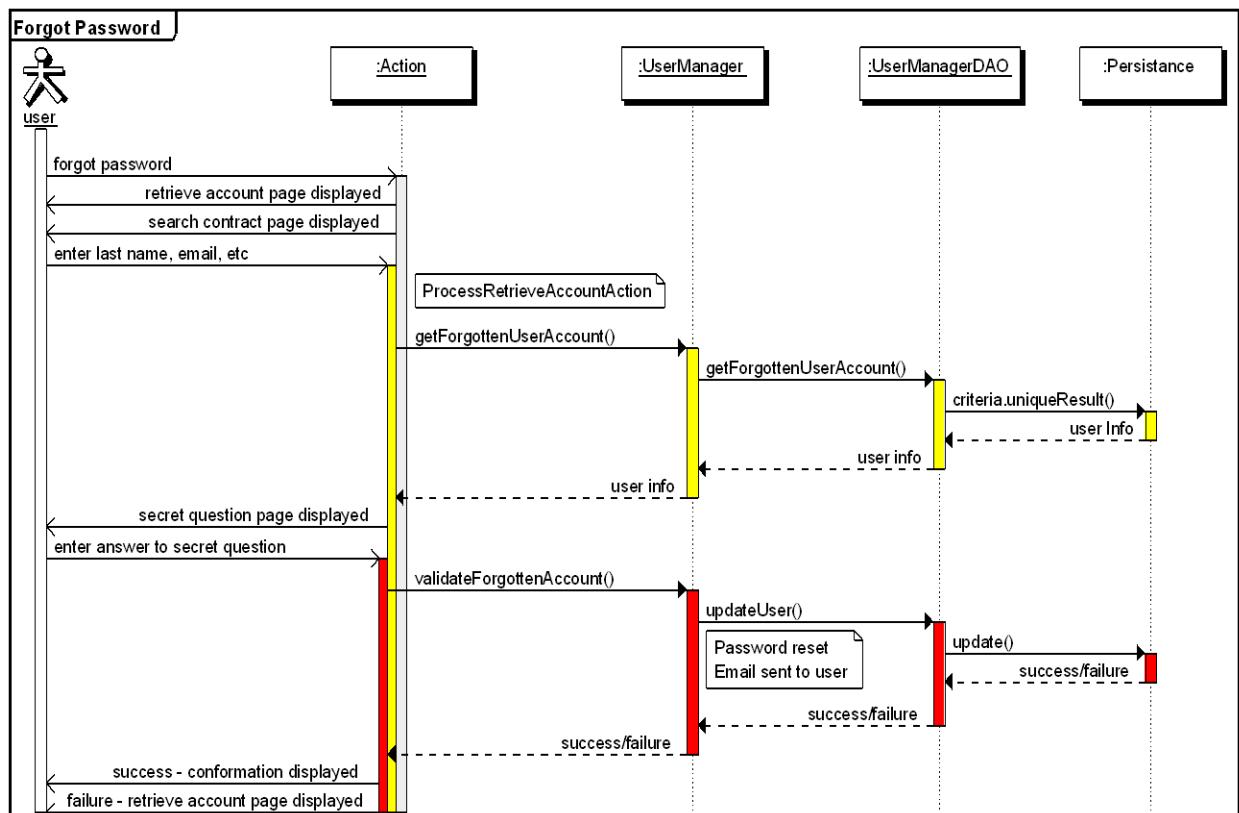
A security question is utilized to verify the user's identity, and the password is emailed upon success.

This screenshot shows the same ImmPort login page as above, but with a different step. It displays a security question: "What is your favorite color?" followed by an answer field containing "blue". The "Submit" and "CANCEL" buttons are at the bottom. The footer information is identical to the previous screenshot.

7.7.1 Forgot Password Class Diagram



7.7.2 Forgot Password Sequence Diagram



7.8 LOG IN/OFF

In order to log into the ImmPort system, a unique user-id is required with a password. These are created during the registration process or creation of an account. The password is restricted to 8-12 characters with one being a numeric value. Appropriate error messages are displayed, if applicable.

Sign In

User Name:

Password:

SIGN IN [Forgot Password?](#)

Register for ImmPort

Benefits of Registration:

- » Access to data visualization tools
- » Compare data sets to other researchers
- » Access to data analysis tools

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

[Edit Profile](#) | [Sign Out](#)

[About ImmPort](#) [Admin](#) [Access Data](#) [Tools](#) [Resources](#) [News & Events](#)

[Browse Data](#) [Search Data](#) [Visualize Data](#) [Analyze Data](#)

ImmPort Research Data

[NIAID Programs in ImmPort](#)

[Shared Research Data](#)

My Workbench

[Private Research Data](#)

[Saved Lists](#)

[Analysis Results](#)

[Data Submission Status](#) (Coming Soon)

News

Recent Publications in Immunology

[Dorner et al.](#) studies a mouse model for blocking hepatitis C virus infection.

[Hansen et al.](#) reported the use of an effector memory T cell

Events

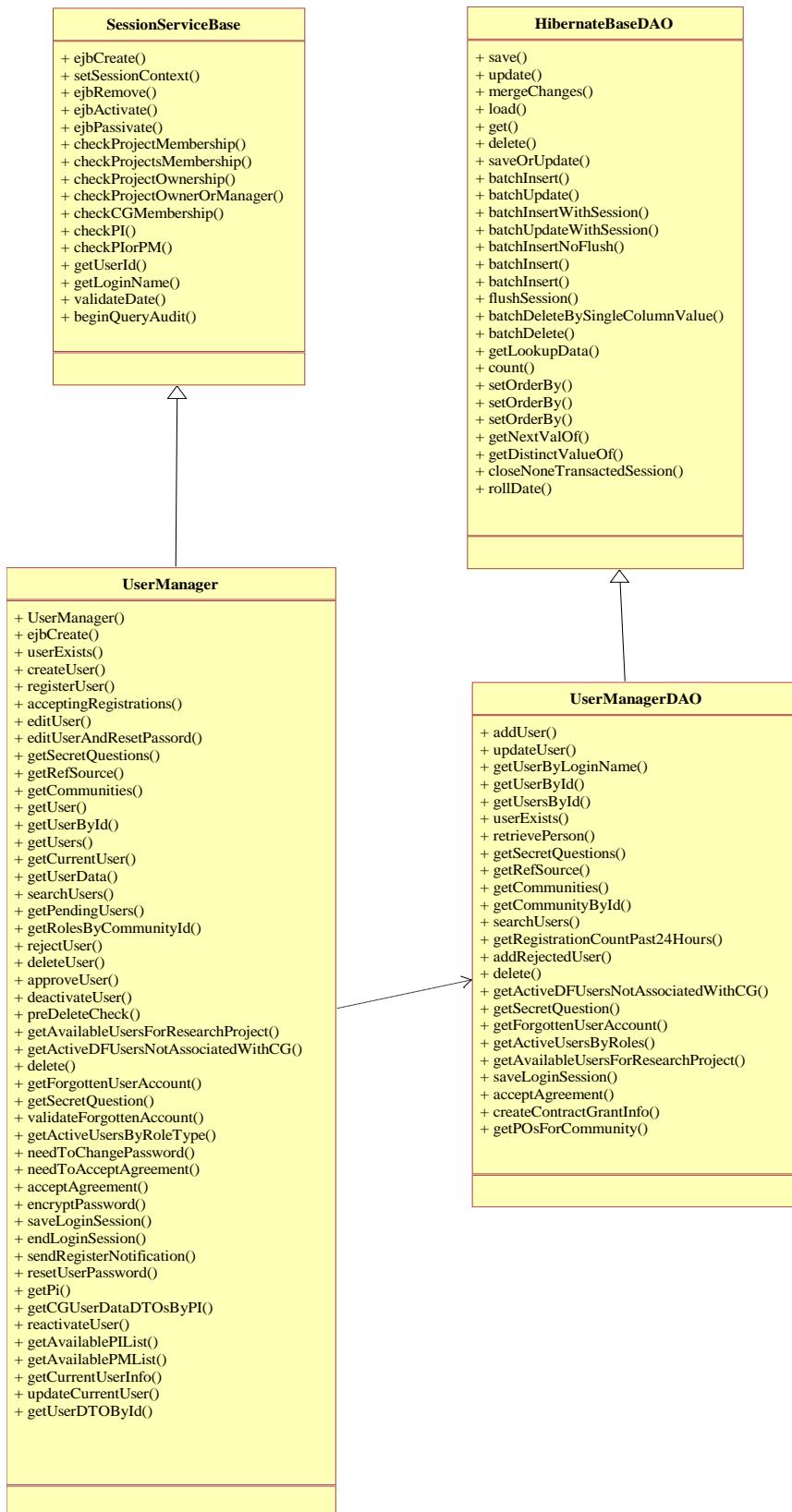
ImmPort Events

July 21, 2011 [ImmPort 2.12 release](#)

Notifications

- » Release of research data from the study by Dr.Thomas Casale: 'Allergen immunotherapy Co-administered with Omalizumab' is available [now](#).
- » Doing flow cytometry? [FLOCK](#) has been upgraded for improved automated population identification.
- » Analyze HLA typing data with ImmPort's [new ambiguity reduction tool](#) for typing data.
- » Download summary tables, pathways, protein-protein interaction networks and GO terms for immunologically relevant genes from the [ImmPort Gene Lists](#).
- » Submitting data to ImmPort? We want to help! We're available to assist you in all aspects of data submission, including formatting, validating and uploading your data. Send us an [email](#), and we'll work with you to ensure everything runs smoothly.

7.9 USER ADMINISTRATION CLASS DIAGRAM



8.0 MANAGE PROGRAM DESIGN PACKAGES

The Manage Program Module high-level design artifacts model program management capabilities, which include searching, creating, modifying, and viewing programs.

8.1 CREATE PROGRAM

The create program diagrams depict the events for creating a program. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web-based application window titled 'Create Program'. At the top, there is a navigation bar with links: 'My Profile', 'Programs', 'Contracts-Grants', 'Users', and 'Reports'. Below the title, there is a note: 'Fields marked with an asterisk * are required.' The form contains the following fields:

- Title***: The Program
- Short Title (or Acronym)***: Short Title for Program
- Description***: Program Description
- Category***: NIAD Program that will share data: ***will make Program metadata viewable on Home page
- Start Date (mm/dd/yyyy)***: 09/30/2004
- End Date (mm/dd/yyyy)***: 09/30/2011
- Import Access End Date (mm/dd/yyyy)**: (empty field)

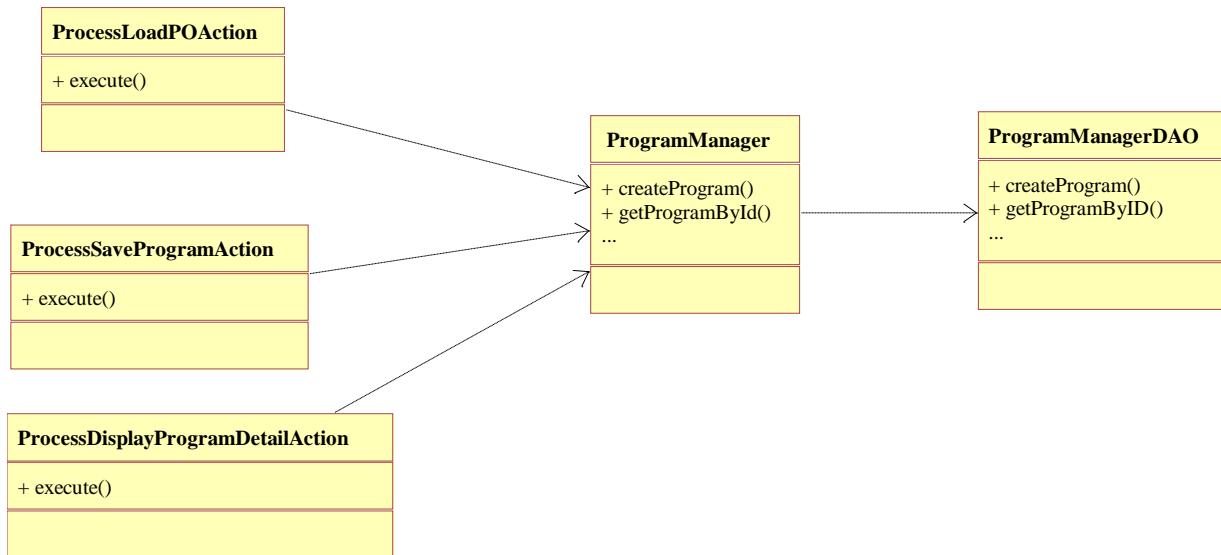
Below these fields is a section titled 'DAIT Information' containing:

- Program/Project Officer ***: Cheryl Kraft
- Deputy Program/Project Officer**: (empty field)

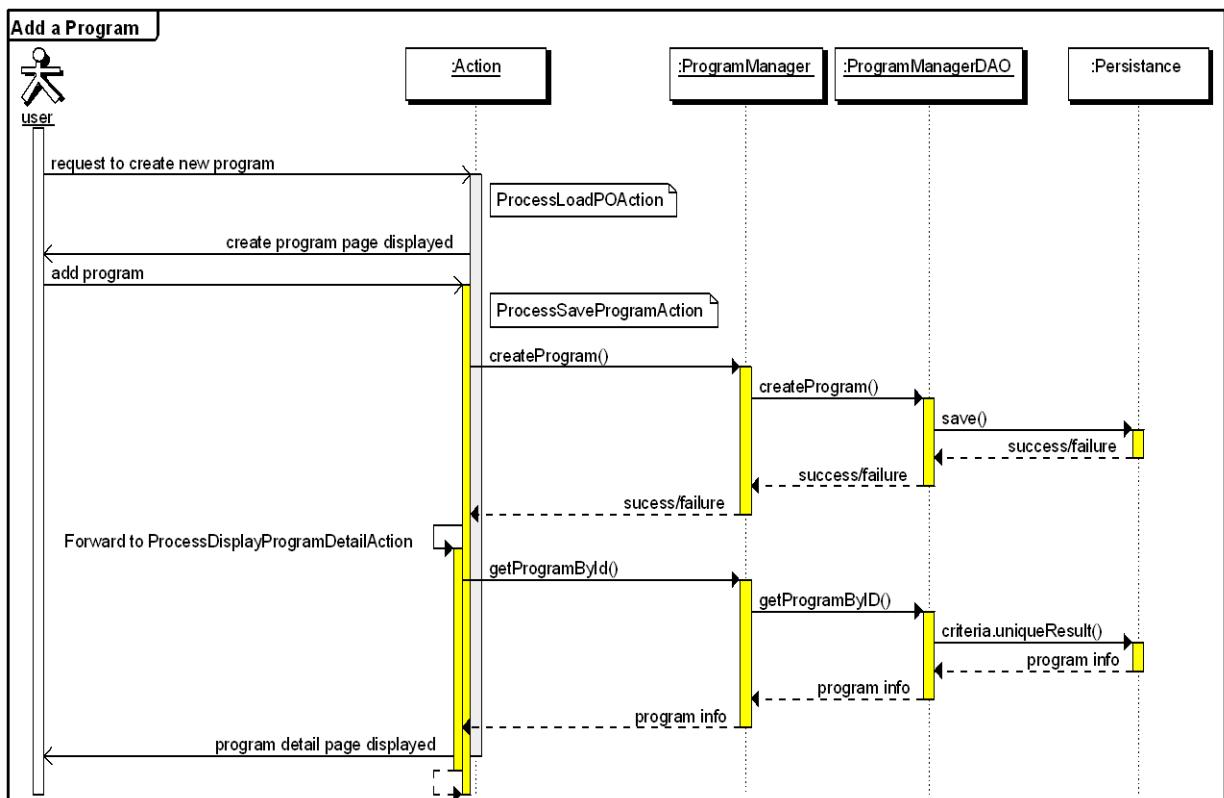
At the bottom of the form are two buttons: 'Create' and 'Reset'.

Once the program is created, contracts are associated.

8.1.1 Create Program Class Diagram



8.1.2 Create Program Sequence Diagram



8.2 UPDATE PROGRAM

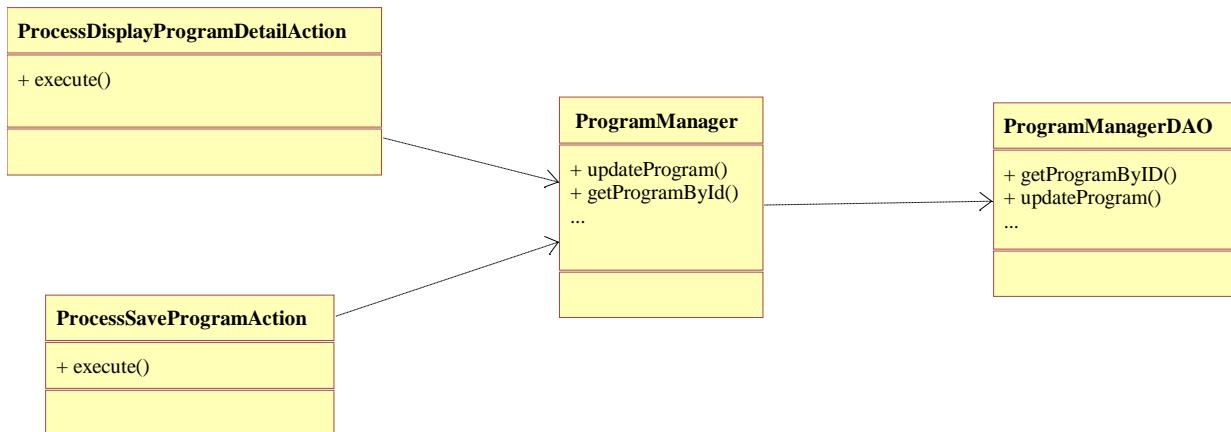
The Update Program diagrams depict the events for updating a program. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web-based application interface titled "Update Program". At the top, there is a navigation bar with links: My Profile, Programs, Contracts-Grants, Users, and Reports. Below the navigation, the main form has the following fields:

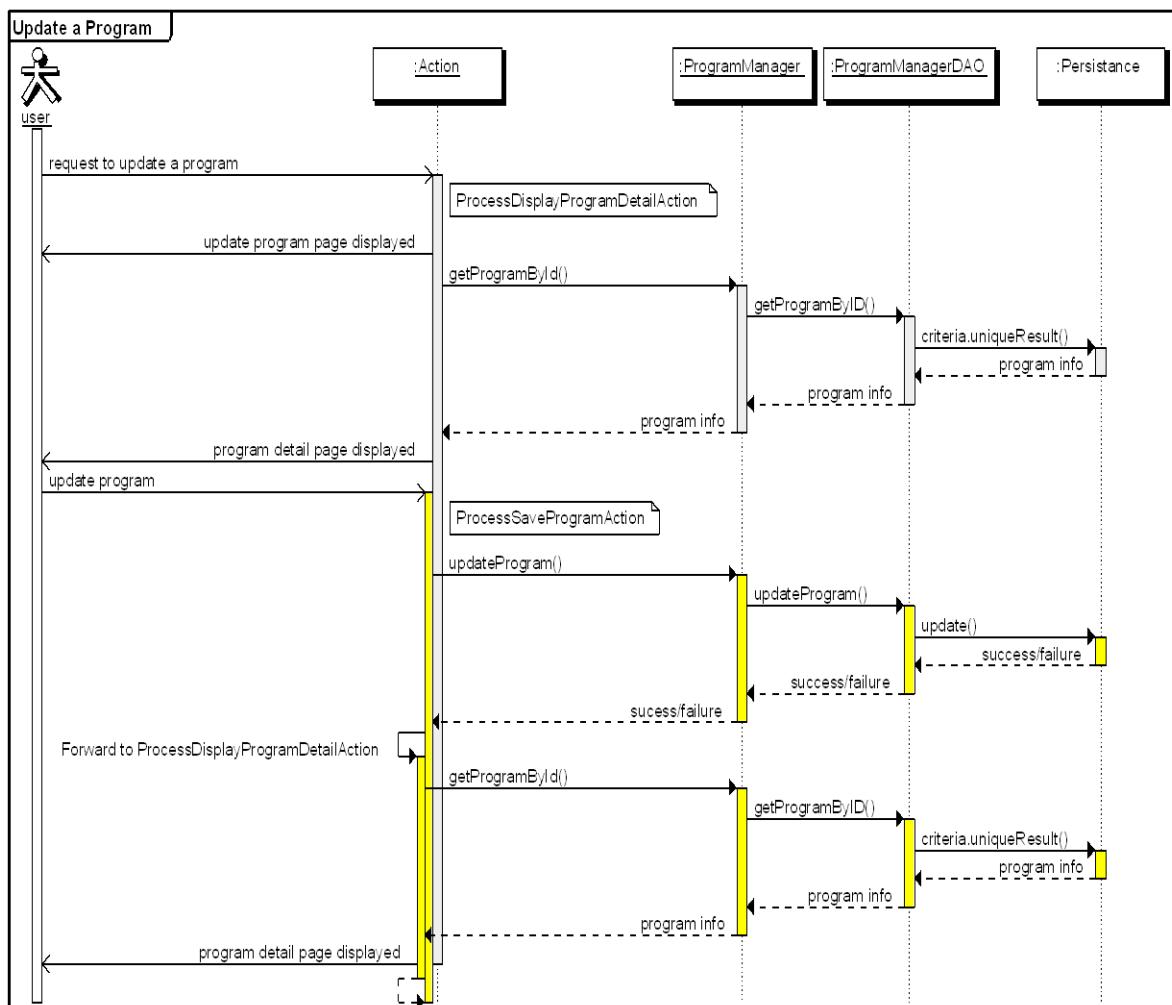
- Title***: The Program
- Short Title (or Acronym)***: Short Title for Program
- Description***: Program Description
- Category***: NIAID Program that will share data: ***will make Program metadata viewable on Home page
- Start Date (mm/dd/yyyy)***: 09/30/2004
- End Date (mm/dd/yyyy)***: 09/30/2011
- Import Access End Date (mm/dd/yyyy)**: 09/30/2011

Below these fields is a section titled "DAIT Information" containing two dropdown menus for "Program/Project Officer*" and "Deputy Program/Project Officer", both currently set to "Cheryl Kraft". At the bottom of the form are "Save" and "Reset" buttons.

8.2.1 Update a Program Class Diagram



8.2.2 Update a Program Sequence Diagram



8.3 SEARCH PROGRAM

The search program diagrams depict the events for searching a program. Appropriate error messages are displayed if errors are encountered.

Admin / Programs / Search Programs

My Profile | Programs | Contracts-Grants | Users | Reports

Enter criteria in the form below to search for a specific program. If you submit the form without entering search criteria, you will retrieve records for all programs.

Title

Short Title (or Acronym)

Category

Start Date

End Date

ImmPort Access End Date

Program/Project Officer (Last Name)

Status Active Inactive All

The user is able to search on the Title, Short Title, Start Date, End Date, ImmPort Access End Date, or Program/Project Officer.

Admin / Programs / Program Search Results

My Profile | Programs | Contracts-Grants | Users | Reports

[Modify Search](#)

Programs matching your search criteria are listed below. Click on a program title to view program details.

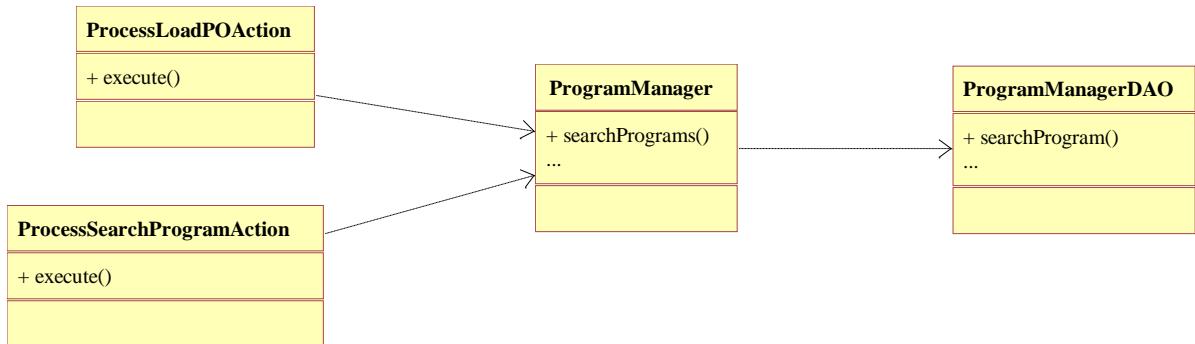
One item found.

Title	Short Title	Category	Status	Program/Project Officer
The Program	Short Title for Program	DAIT	Active	Cheryl Kraft

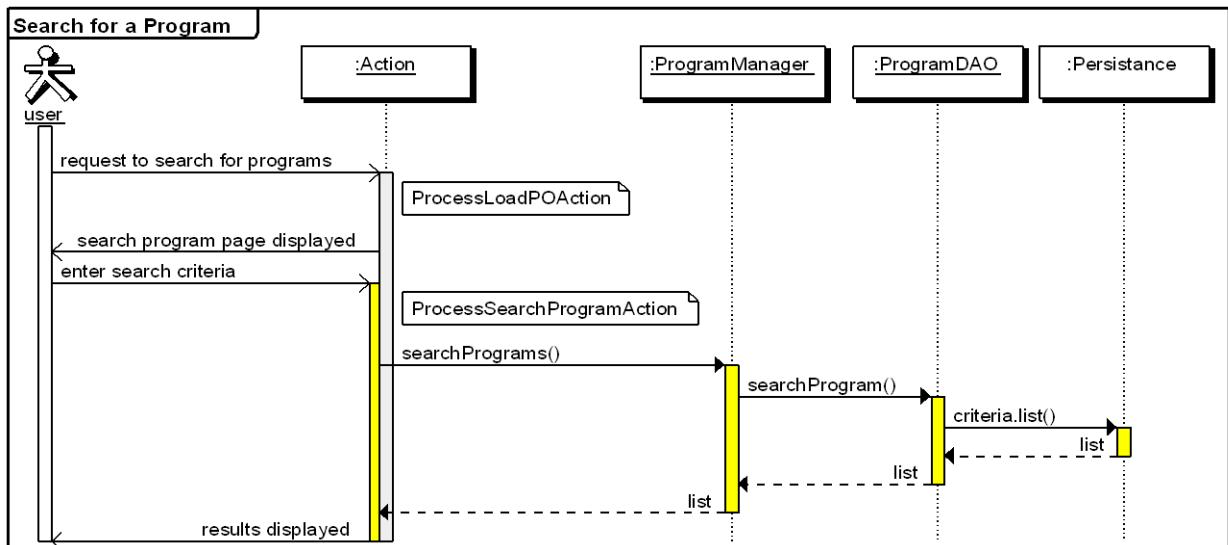
One item found.

A list of program search results is displayed. This allows the user to select a program to view or modify. In order to modify the search, click the "Modify Search" link.

8.3.1 Search Program Class Diagram



8.3.2 Search Program Sequence Diagram



8.4 PROGRAM ADMINISTRATION CLASS DIAGRAM



9.0 MANAGE CONTRACT DESIGN PACKAGES

The Manage Contract Module high-level design artifacts model contract management capabilities which include adding, updating, searching and viewing contracts.

9.1 ADD/UPDATE CONTRACT

The Add/Update Contract diagrams depict the events for adding and updating a contract. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web-based form for adding or updating a contract. The form includes fields for:

- Title*, Short Title*, Contract-Grant Number*, Abstract*
- Keywords*
- Category*: NIAID Contract-Grant that will share data: ***makes Contract metadata viewable on Home page
- Start Date (mm/dd/yyyy)*: 09/30/2004
- End Date (mm/dd/yyyy)*: 09/30/2011
- ImmPort Access End Date (mm/dd/yyyy)
- DAIT Information:
 - Program *: Other Programs
 - Program/Project Officer*: Ken KTest
 - Deputy Program/Project Officer
- Submit button

Once the contract is added, a Principal Investigator, Program Manager and Other Staff is associated with the contract.

The screenshot shows the 'Associated Users' section of the contract management interface. It includes a table of users and a user selection interface:

Name	Phone	Email	Role
Dunn, Patrick	301-527-6603	pdunn@imimport.org	OS
Wiser, Jeff	301-527-6673	jwiser@imimport.org	PM
Xia, Ashley	301-496-7551	axia@niam.nih.gov	PI

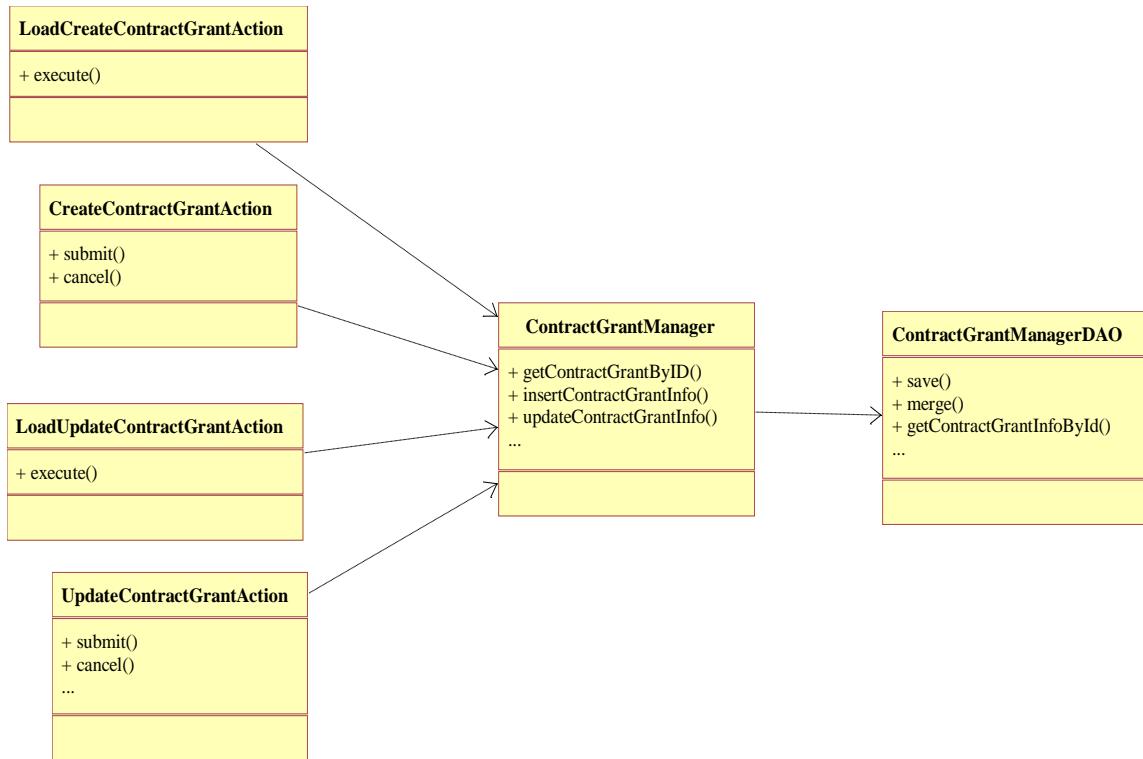
Buttons for 'Add...' and 'Remove' are shown below the table.

The 'Associated Projects' section includes a 'Create Default Project' button.

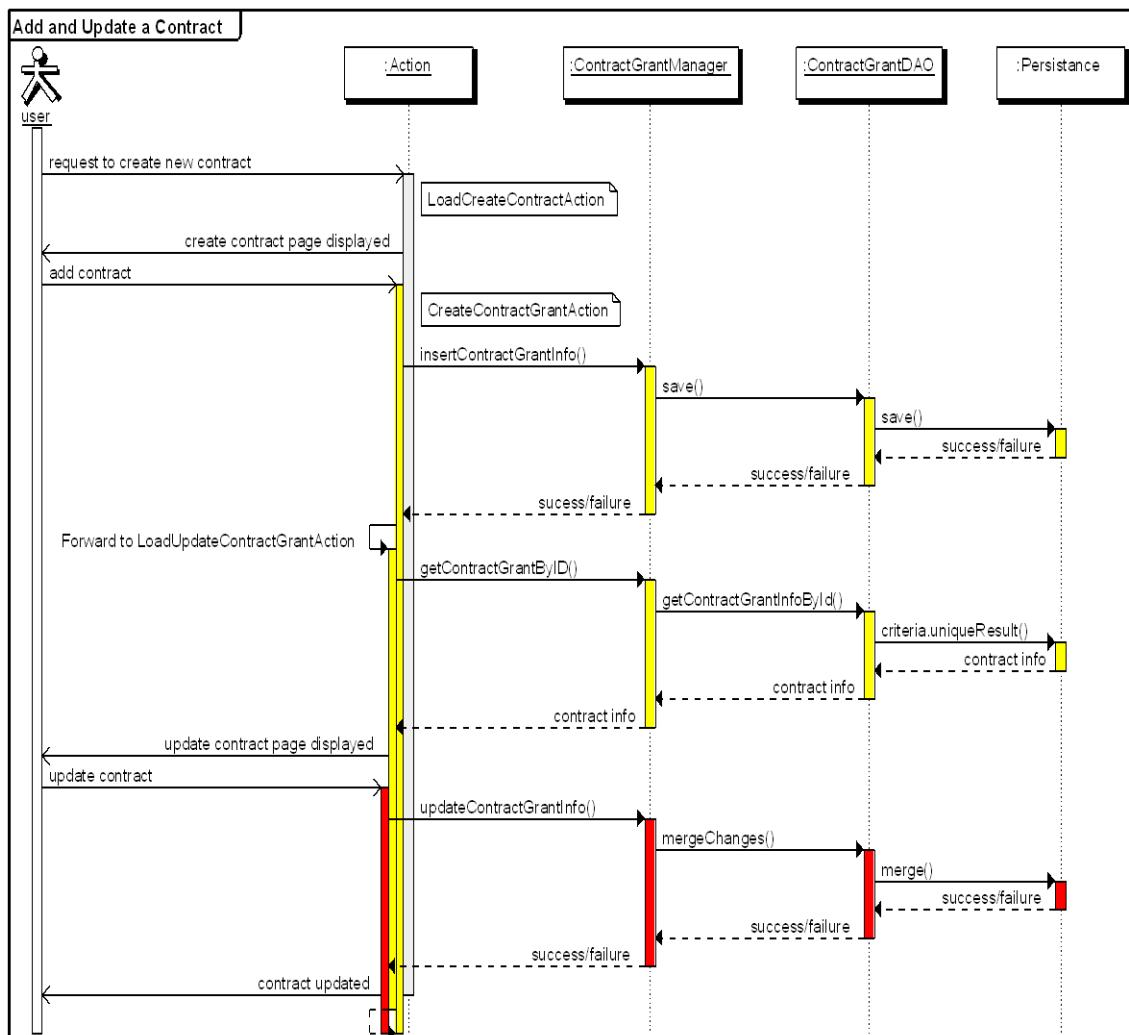
Once the associated users are added, a default project may be created. The associated users to the contract will obtain access to the default project.

Program	<input type="text" value="OTHER PROGRAMS"/> <input type="text" value="Ken KTest"/> <input type="button" value="..."/> <input type="button" value="..."/>																				
Program/Project Officer*	<input type="text" value="Deputy Program/Project Officer"/> <input type="button" value="Submit"/> <input type="button" value="Cancel"/>																				
Associated Users <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>Name</th> <th>Phone</th> <th>Email</th> <th>Role</th> </tr> </thead> <tbody> <tr> <td><input type="checkbox"/></td> <td>Dunn, Patrick</td> <td>301-527-6603</td> <td>pdunn@immp.org</td> <td>OS</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Wiser, Jeff</td> <td>301-527-6673</td> <td>jwiser@immp.org</td> <td>PM</td> </tr> <tr> <td><input type="checkbox"/></td> <td>Xia, Ashley</td> <td>301-496-7551</td> <td>axia@nifld.nih.gov</td> <td>PI</td> </tr> </tbody> </table> <input type="button" value="Add..."/> <input type="button" value="Remove"/>			Name	Phone	Email	Role	<input type="checkbox"/>	Dunn, Patrick	301-527-6603	pdunn@immp.org	OS	<input type="checkbox"/>	Wiser, Jeff	301-527-6673	jwiser@immp.org	PM	<input type="checkbox"/>	Xia, Ashley	301-496-7551	axia@nifld.nih.gov	PI
	Name	Phone	Email	Role																	
<input type="checkbox"/>	Dunn, Patrick	301-527-6603	pdunn@immp.org	OS																	
<input type="checkbox"/>	Wiser, Jeff	301-527-6673	jwiser@immp.org	PM																	
<input type="checkbox"/>	Xia, Ashley	301-496-7551	axia@nifld.nih.gov	PI																	
Associated Projects <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Project Title</th> <th>Project Category</th> <th>Project Description</th> </tr> </thead> <tbody> <tr> <td>Contract Title</td> <td>DAIT</td> <td>Contract Abstract</td> </tr> </tbody> </table>		Project Title	Project Category	Project Description	Contract Title	DAIT	Contract Abstract														
Project Title	Project Category	Project Description																			
Contract Title	DAIT	Contract Abstract																			

9.1.1 Add/Update Contract Class Diagram



9.1.2 Add/Update Contract Sequence Diagram



9.2 SEARCH CONTRACT

The Search Contract diagrams depict the events for searching a contract. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a search form titled "Admin / Search Contracts-Grants". The form includes fields for Title, Contract-Grant Number, Contract-Grant Category (a dropdown menu), Start Date, End Date, ImmPort Access End Date, Principal Investigator (Last Name) (a dropdown menu), Associated Program Name (a dropdown menu showing options like "Immune Function and Biodefense in Children, Elderly, and Imm... Population Genetics II", "Human Immunology Project Consortium", "Modeling Immunity for Biodefense II", and "Atopic Dermatitis Research Network (ADRN)"), and Status (radio buttons for Active, Inactive, or All). Below the form are "Submit" and "Reset" buttons.

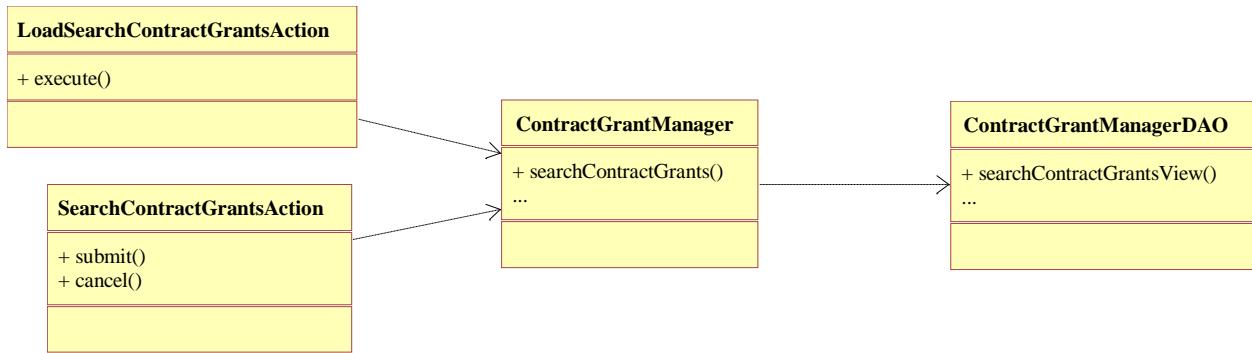
A user is able to search on the fields Title, Grant – Contract Number, Category, Start Date, End Date, ImmPort Access End Date, Principal Investigator, Associated Program Name or Status.

The screenshot shows search results for "Admin / Contracts-Grants Search Results". It includes a "Modify Search" link and a table of search results. The table has columns: Contract-Grant Number, Title, Status, Category, Program, Principal Investigator, and End Date. One row is shown with the following data:

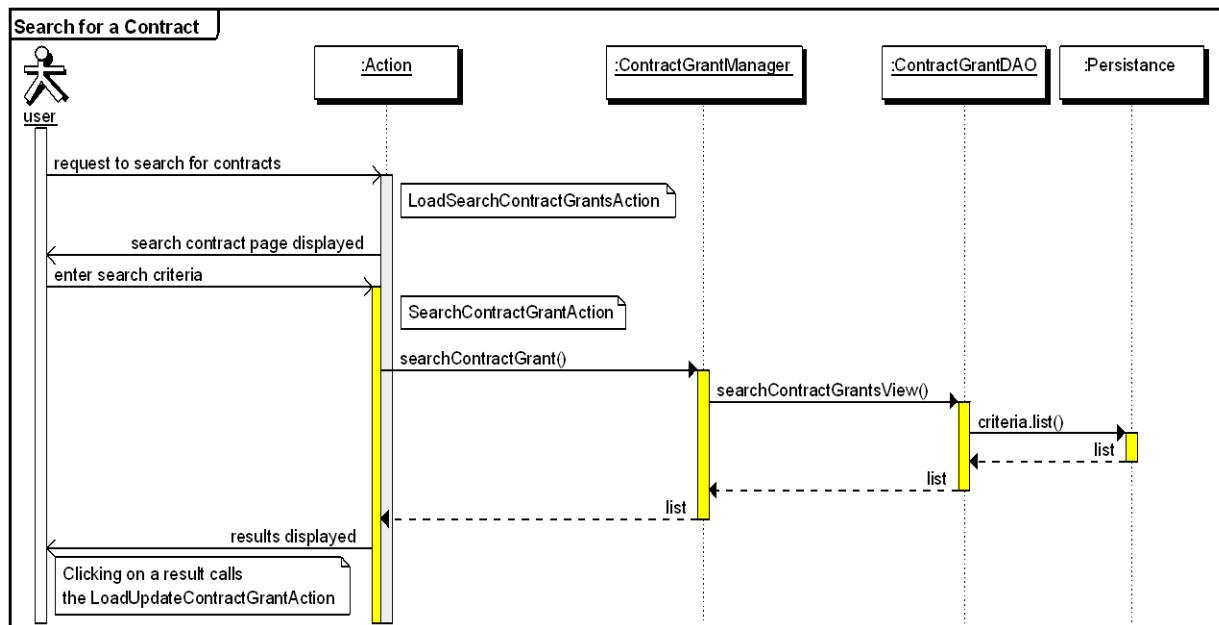
Contract-Grant Number	Title	Status	Category	Program	Principal Investigator	End Date
Contract Number	Contract Title	Active	DAIT	Other Programs	Xia Ashley	2011-09-30 00:00:00.0

A list of Grants - Contracts search results is displayed. This allows the user to select a Grants - Contracts to view or modify. In order to modify the search, click the “Modify Search” link.

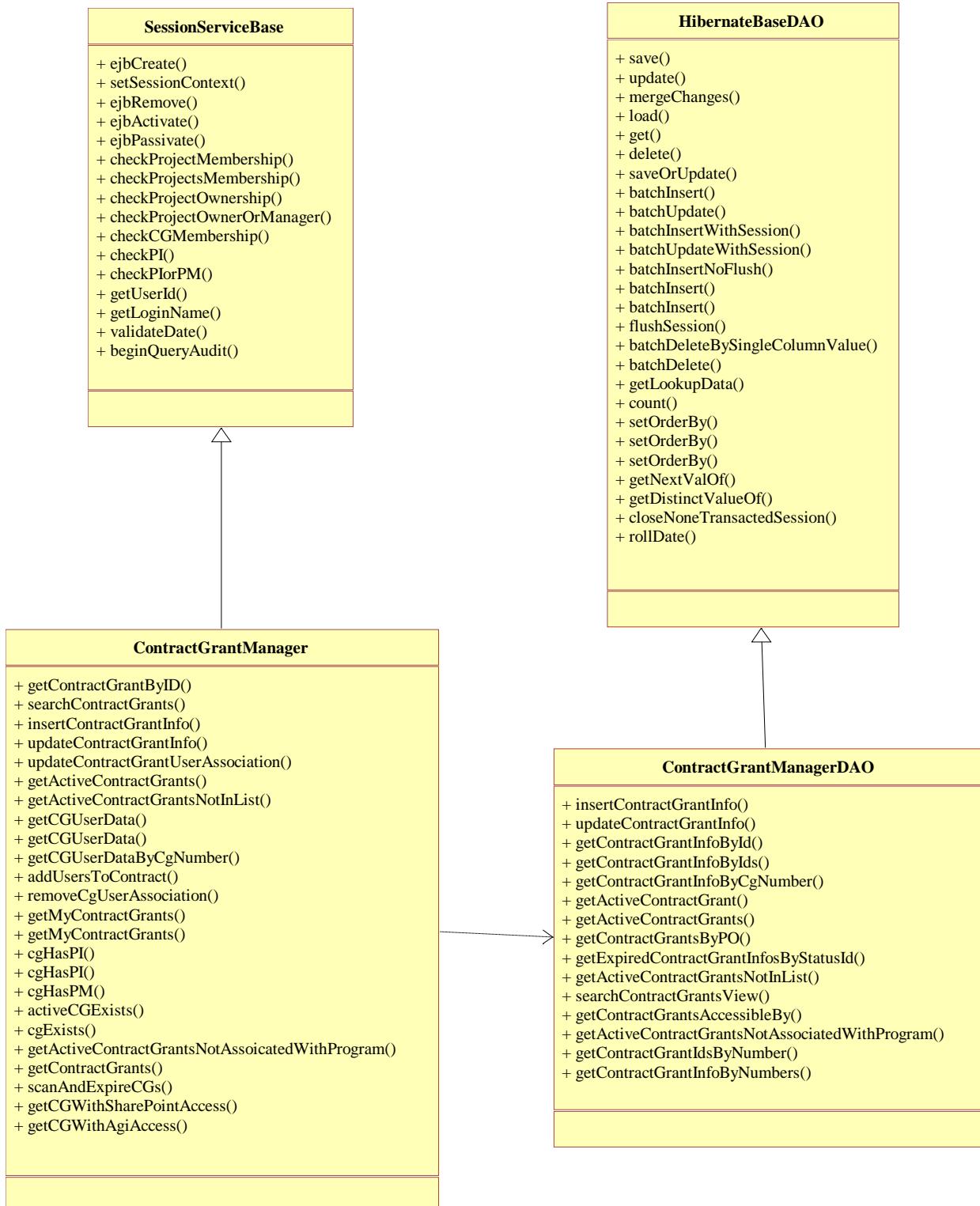
9.2.1 Search Contract Class Diagram



9.2.2 Search Contract Sequence Diagram



9.3 CONTRACT ADMINISTRATION CLASS DIAGRAM



10.0 MANAGE PROJECT DESIGN PACKAGES – UPDATE FOR ADMIN PROCESS

The Manage Project Module high-level design artifacts model project management capabilities which include adding, updating, searching and viewing contracts.

10.1 ADD/UPDATE PROJECTS

The Add/Update project diagrams depict the events for adding and updating a project. Appropriate error messages are displayed if errors are encountered.

The screenshot shows a web application interface for 'IMMPORT BIINFORMATICS FOR THE FUTURE OF IMMUNOLOGY'. At the top right are links for 'Edit Profile' and 'Sign Out'. A navigation bar includes 'About ImmPort', 'Admin', 'Access Data', 'Tools', 'Resources', and 'News & Events'. Below the header, a breadcrumb trail shows 'Admin / Create Research Project: Notice'. Underneath, there's a link to 'My Profile' and dropdown menus for 'Grants - Contracts' and 'Projects'. A section titled 'Terms and Conditions for Creating and Using Private Project Workspaces' contains text about the system's purpose and a bulleted list of guidelines. At the bottom, a note states 'By continuing with the creation of a private project workspace, you agree to these terms and conditions governing the use of private project workspaces.' Two buttons, 'Continue' and 'Cancel', are at the bottom left.

In order to create a project, a user must read the terms and conditions for creating a Private Project Workplace.

The screenshot shows a web-based application interface for creating a project. At the top, there's a header bar with the text "Admin / Create Project". Below the header, a navigation bar includes links for "My Profile", "Contracts-Grants", and "Projects". The main area is titled "Research Project" and contains several input fields:

- Title ***: A text input field labeled "Project Title".
- Description ***: A text area labeled "Project Description" with the placeholder "(max 4000 chars)".
- Keywords ***: A text input field labeled "Project Keywords".
- Category ***: A dropdown menu currently set to "Test Project: can be used for any testing purpose".
- Contract-Grant***: A dropdown menu currently set to "Sally-Contract (TEST)".
- Project Owner ***: A text input field labeled "Patty-PI Tester".
- Project Manager**: A text input field containing "Patty-PI Tester". To its right is a "Remove Project Manager" link and a "...".

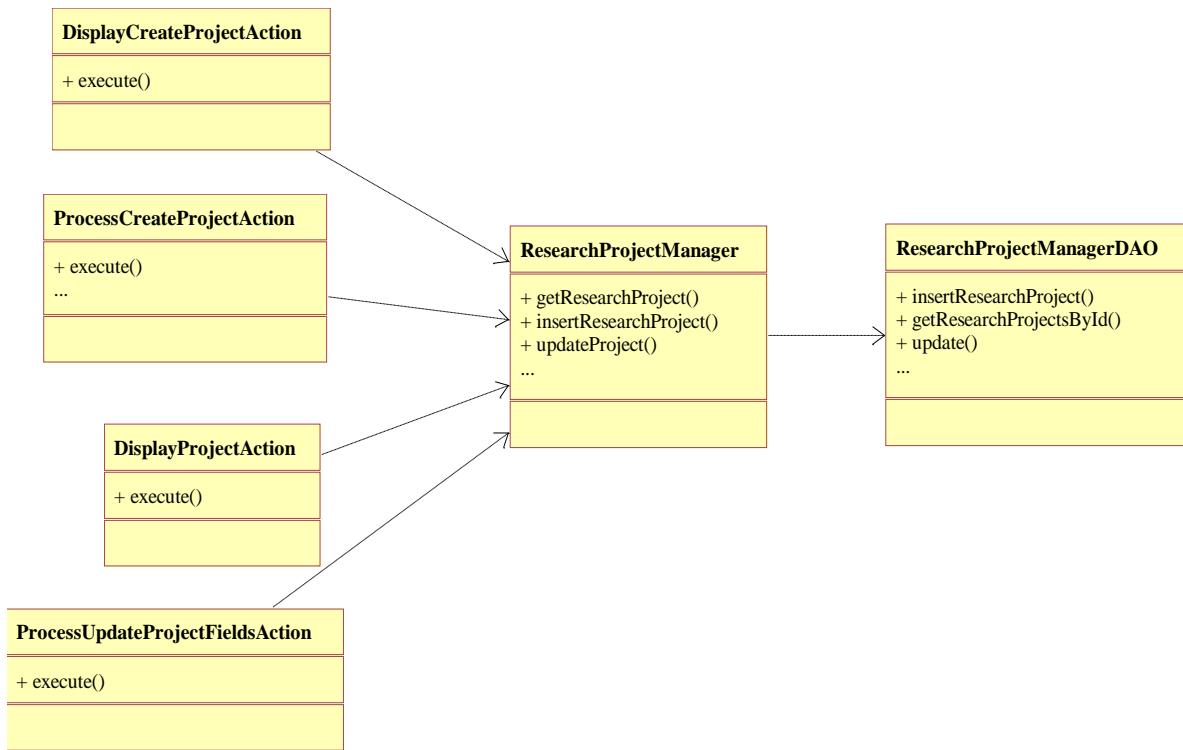
At the bottom of the form is a "Save" button.

The user enters the required fields, Title, Description, Keywords, Category, and Contract-Grant. The Project Owner is auto generated. An appropriate error message will display if an error has occurred.

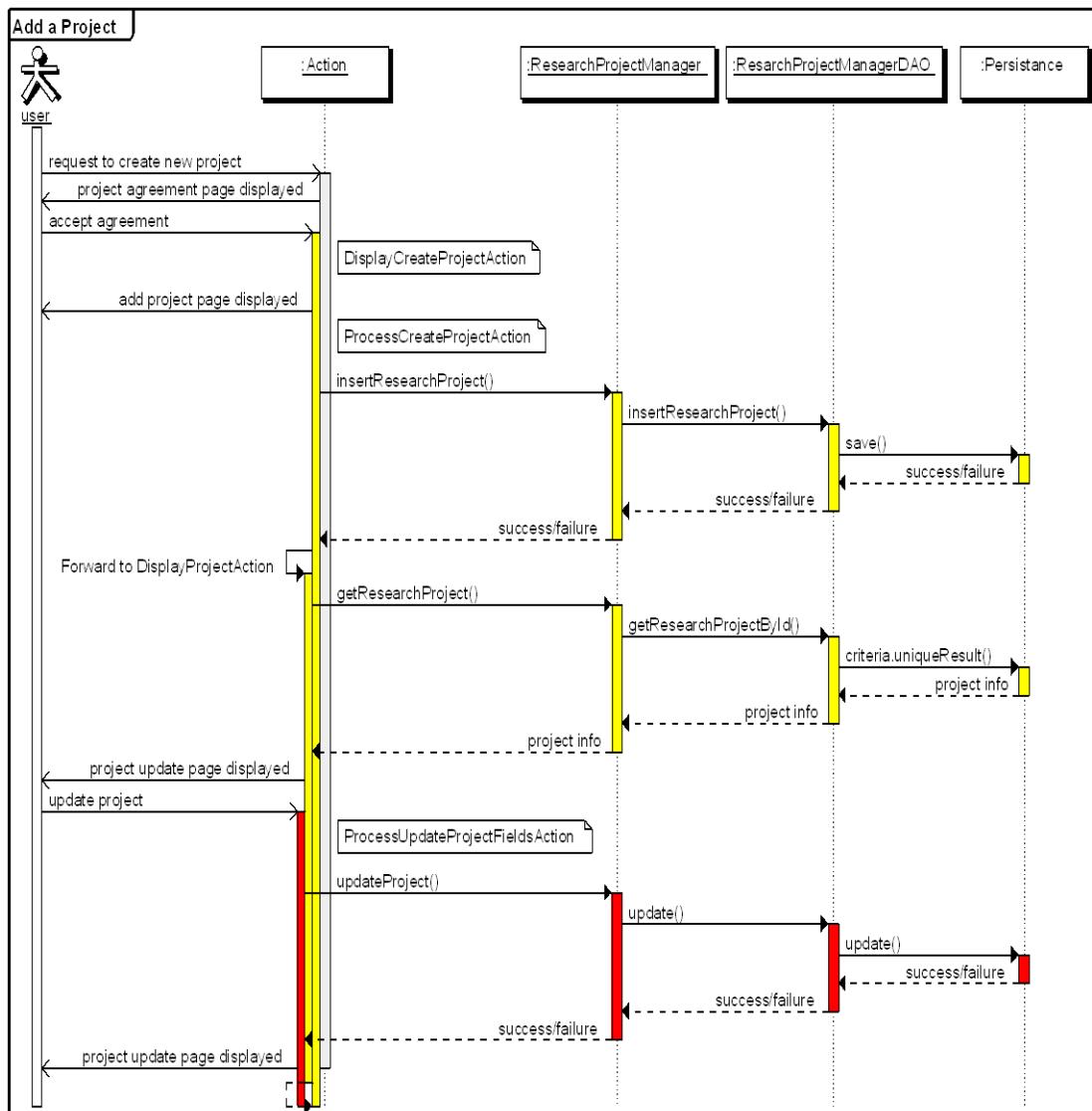
The screenshot shows the same project creation interface after a successful submission. The "Project Manager" section now displays "Your workspace has been created." Below this, the "Associated Users" section is visible, stating: "The following users are associated with this research project. Add users by clicking "Add..." Remove users by clicking on the corresponding checkboxes and clicking "Remove."". It also says "No results found." At the bottom are "Add...", "Remove", and "Finish" buttons.

Once the project is created, the user is able to update the project and/or associate users to their project.

10.2 ADDING/UPDATING A PROJECT CLASS DIAGRAM



10.2.1 Adding/Updating a Project Sequence Diagram



10.3 DISPLAY PROJECTS

The display projects diagrams depict the events for viewing a project.

The screenshot shows a user interface titled "My Research Projects". It displays two tables: "My Research Projects" and "My Collaborative Projects".

My Research Projects:

Project Title	Project Category	Project Description
Atopic Dermatitis & Vaccinia Network (ADVN) Cli...	DAIT	Atopic Dermatitis & Vaccinia Network (ADVN) Clinical Studies Consortium
ITN019AD - Clinical Trial	DAIT	Phase II, Double-Blinded, Placebo-Controlled, Efficacy and Safety Evaluation...
McMaster Test Project	DAIT	McMaster
Mt. Sinai VIP004	DAIT	for testing Mt Sinai load files
Project Title	TEST	Project Description
Test Example Packages	TEST	testing
Test Project	DAIT	Test Project
UAB PopGen	DAIT	study design subs, bioSams, etc
UWash PoPGen test GT data	TEST	UWash PoPGen test GT data
p2	BISC	a
rene test	DAIT	rene test
t	DAIT	t

My Collaborative Projects:

Project Title	Project Category	Project Description
Collaborating	DAIT	Collaborating with users
Collaborative Project 123	DAIT	Sandbox for me to play
Rene's cw	BISC	a cw for test
hey	DAIT	what.
rene cp 1	LSR	rene cp2 k

Buttons:

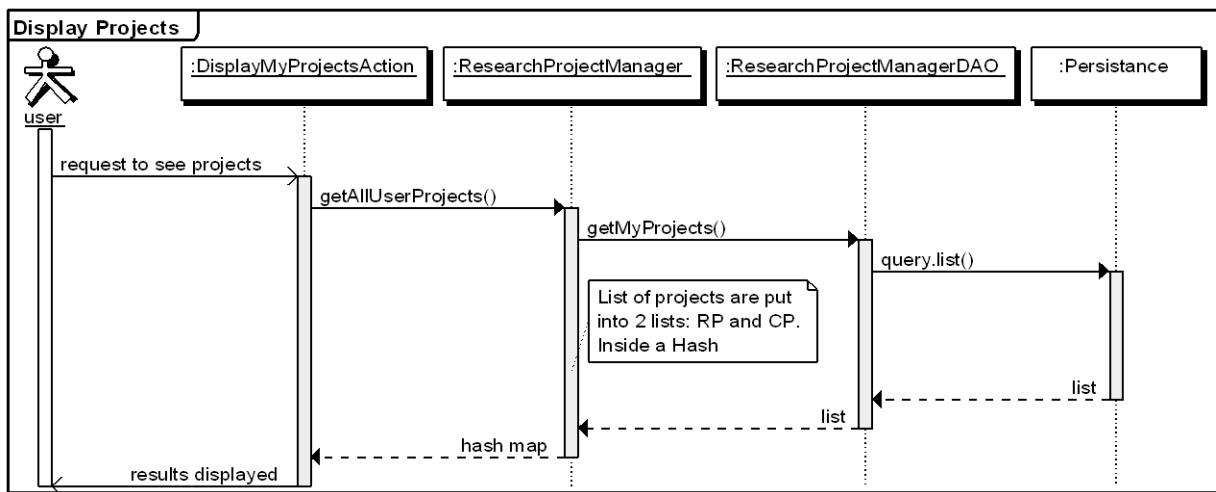
- View Project Details**
- Manage Project Data**

A user is able to select the project while viewing the project details or manage project data.

10.3.1 Display Projects Class Diagram



10.3.2 Display Projects Sequence Diagram



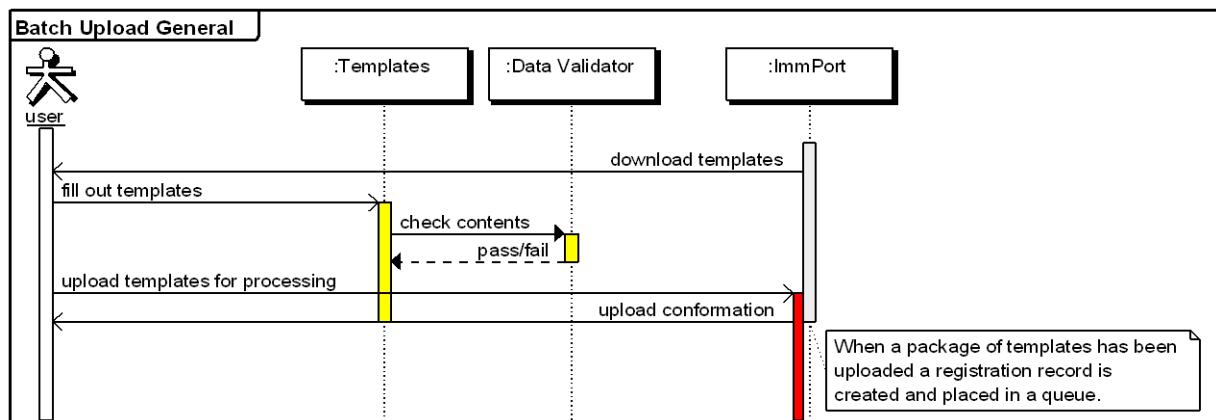
10.4 PROJECT ADMINISTRATION CLASS DIAGRAM



11.0 DATA SUBMISSION DESIGN PACKAGES

One of the primary goals of the BISC project is the sharing of research data at a level of detail that is beyond the current published literature dissemination mechanism. The ImmPort data upload module is designed to capture information regarding experimental results and the accompanying metadata, including subjects, biological samples, experiments details, etc. Uploaded information is access controlled by linking to a private research project and the user must be authorized to submit information for this project. Experimental results and metadata submission is currently handled as an online process and can be performed independently by users on ImmPort. That code is detailed in sections 11.1 to 11.6. To capture detail study design information as well as clinical data such as assessments, lab tests, adverse events, etc, a more interactive and manual process is utilized that involves more direct interaction with the data providers. The complexity and diversity of clinical data content necessitates more back and forth communication, data extraction, data transformation and quality control in order to perform final data loading. This process is detailed in section 11.7.

11.1 DATA SUBMISSION WORKFLOW OVERVIEW



11.2 DATA SUBMISSION TEMPLATES

The ImmPort system has developed a series of templates using Excel and XML to define how information is to be submitted for each entity (<https://immport.org/immportWeb/display.do?content=DataTemplates>).

11.2.1 Data Submission Templates Links

The screenshot shows the ImmPort interface for data submission. At the top, there's a navigation bar with 'Data Submission / Resource / Data Submission Templates'. Below it is a breadcrumb trail: 'Submit Data' > 'Submission History' > 'Resources'. A large central area displays a process flow: 'Step 1: Download and Fill Template' (with a link to 'Template'), 'Step 2: Check Data in Database' (with a link to 'Check Data in Database'), 'Step 3: Send Data in ZIP File' (with a link to 'Send Data in ZIP File'), and 'Step 4: Enter Submission Data & Results' (with a link to 'Enter Submission Data & Results'). Below this flow, there's a list of instructions:

1. Which Data Submission Templates do you need?
Please see the [User Guide](#) to determine which templates need to be completed.
2. Complete the templates that are needed.
Note: Please save spreadsheet (.xls) template as tab delimited (.txt) files.
3. Create a .zip file that contains the files you want to submit (e.g. results, protocols, bioclinical template, experimentSamples template, etc.).
4. Please check that you are using the [latest version](#) of the ImmPort data transfer templates.

Below the instructions is a table titled 'ImmPort Research Data Class' with columns for 'ImmPort Research Data Class', 'Spreadsheet Template', 'Required Data to Complete Metadata Form', 'XML Template', and 'Latest Version Date Available'. The table contains three rows:

ImmPort Research Data Class	Spreadsheet Template	Required Data to Complete Metadata Form	XML Template	Latest Version Date Available
Protocol	Protocol.xls	Protocol User Defined ID Protocol File Name	Protocol.xml	January 2010
Reagent	Reagent.xls	Reagent User Defined ID Assay Specific Details	Reagent.xml	January 2010
Subject	Subject.xls SubjectDetails.xls	Human User Defined ID Species Name Protocol Phenotype	Subject.xml	January 2010

Excel templates are used to annotate columns and provide features such as lists of controlled vocabulary terms. The Excel templates are the primary method used by most data providers, so support for XML templates may be deprecated. The research data content is organized into generally understandable biological domains including study, subjects, samples, protocols, reagents, and experiments. There is a template for each domain. The records in one domain often reference user defined records in other domains (e.g. subject records reference protocols for subject assessment and treatment). The data upload process supports a modular or incremental approach so that subsets of a study's content may be uploaded to ImmPort at the discretion of the data provider. Data uploaded in an earlier upload session may be referenced in a subsequent session without having to redefine the records. The data uploaded by users may be referenced by their own identifiers or by ImmPort accessions. User defined identifiers must be unique within a research project. ImmPort accessions are unique across the data repository. Each upload template has required fields that represent the minimum information that must be submitted (as defined by the NIAID DAIT minimum information guidelines-
<https://immpor.org/tutorials/MinimumInformationGuidelines.doc>), plus additional fields that are optional. In some cases, references from a record in one domain to a record in another domain are required and in other cases they are optional. This establishes links between uploaded data and ensures that the data model is effectively populated.

A commonly adopted practice is for the data provider to upload their data when their study is completed. The user fills out one or more templates. If the templates are in the Excel format, they saved from Excel into tab-delimited text files. This eliminates the need to support multiple versions of the Excel spreadsheets. These templates are bundled together into one ZIP archive with additional documents such as protocols and assay results. The ZIP archive is submitted to ImmPort via web based file upload technology. This generates an upload ticket which is used to track status and audit uploaded data. After

the package is submitted, the system displays a confirmation screen and the package is queued for later processing.

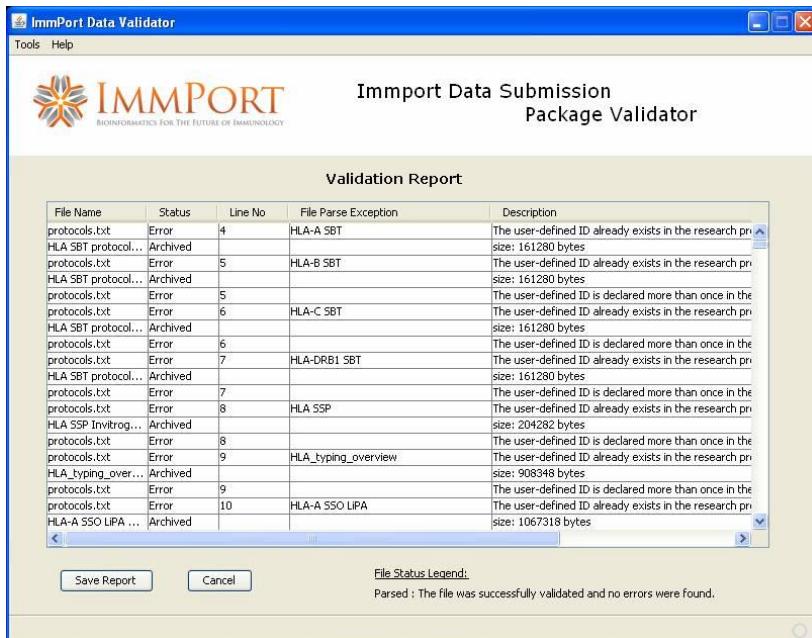
There are alternate procedures to deal with large data upload requirements that do not scale well with web-based technology. An upload ticket may be generated without uploading a ZIP archive. This ticket is used to link data transferred to the data upload staging area via non-web based means. For large data sets, an external drive may be sent to a data provider who uploads data to the drive and then sends it back to ImmPort. The data is transferred to the ImmPort repository and the standard data processing that parses and links the data to the content in the repository is completed. In some cases, a hybrid approach is taken whereby descriptive data is uploaded via the web interface and additional data is sent to ImmPort by offline methods.

11.3 DATA SUBMISSION VALIDATOR

To assist the user with checking the validity of the templates they have filled out, ImmPort has developed a stand-alone Java application (DataValidator) that can be downloaded and run on the user's desktop before they upload the ZIP archive for processing. The DataValidator checks content for referential integrity between the templates, controlled vocabulary fields, previously assigned Project based identifiers, etc. Using the DataValidator is optional, but if used helps find most of the common data submission errors, before the package is submitted for final processing.



The Validator uses a HSQL database that is populated from content downloaded by the user in a Project content ZIP archive from their ImmPort project. The HSQL database stores a subset of the content in the project necessary to validate referential integrity. The Validator synchronizes with the ImmPort repository by loading the Project content ZIP archive. After synchronization, the user selects the data upload archive to evaluate and the Validator provides a report on the success of the validation or issues it encountered.



11.4 DATA SUBMISSION HISTORY

The user may review the data submission history to their project(s) to get a status update and a log of their data submission sessions.

Data Submission / Data Submission History

Submit Data | Submission History | Resources +

Submit Data Main Page → Step 1: Download and Fill Templates → Step 2: Check Data in .zip file → Step 3: Send Data in .zip file → Step 4: Review Submission Status & Results

This page is a summary overview of the data submissions to a project. If you have access to more than one project, use the project filter to view the submission queue from other projects.

The data submitted to ImmPort can be queried and reviewed in [Research Data Search](#).

Filter By Project: Public Gene Expression Data Set Filter

Click on a Ticket Number below to view a submission's details.

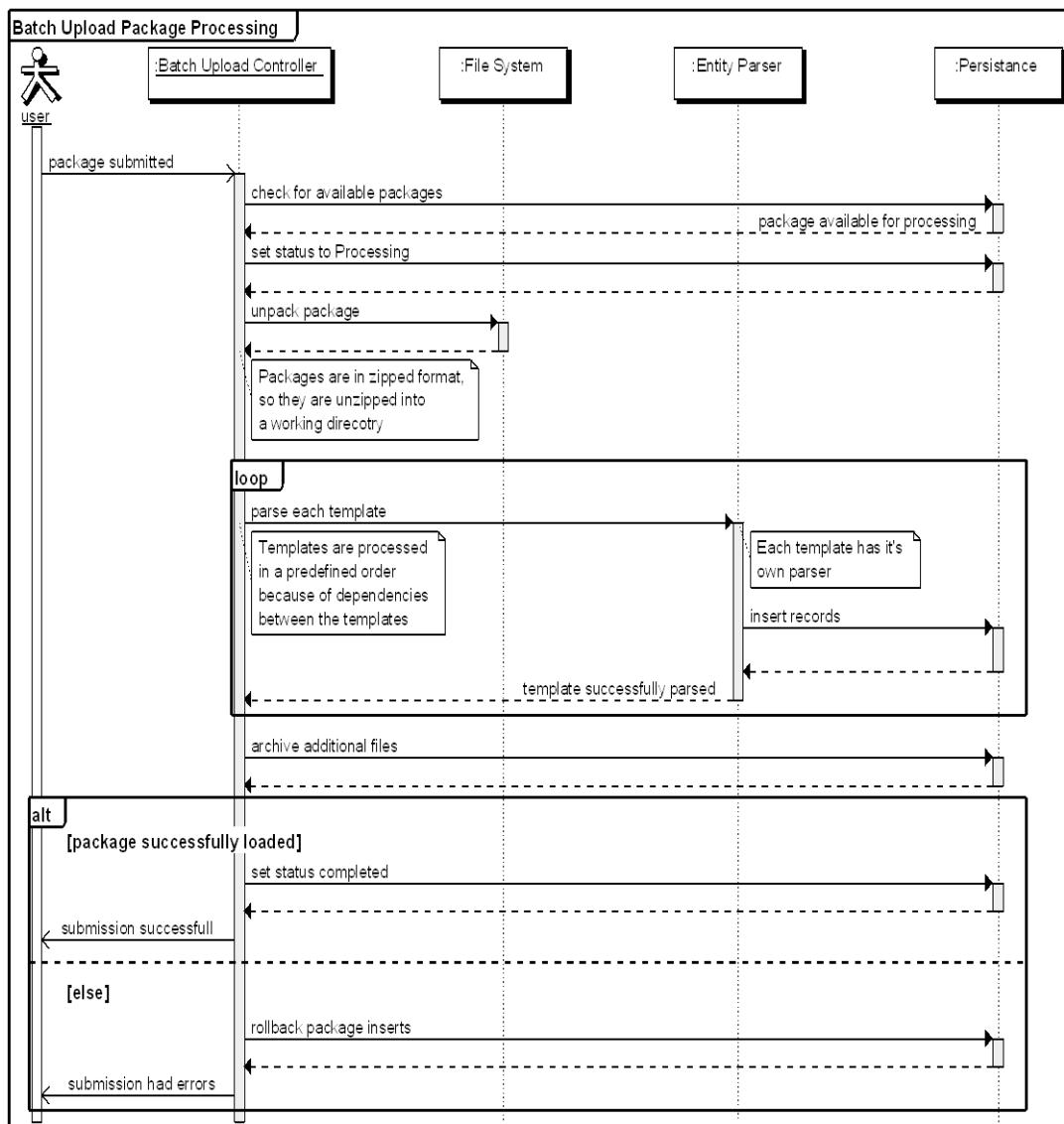
105 items found, displaying 1 to 25. [First/Prev] 1, 2, 3, 4, 5 [Next/Last]

Ticket Number	ZIP File Name	Format	Status	Submitter	Submit Date
Bea_20090504_1166	HLA_1reagentperlocus_Example_Package.tab-delimited.zip	HLA_typing_summary	Completed	Bea	05/04/2009 10:51:28
Bea_20090504_1165	Illumina_BeadStudio_Final_Report_SubjectIDs.tab-delimited.zip	Affy_GCOS-GDAS-GTYPE_Genotyping_output	Completed	Bea	05/04/2009 10:50:46
Bea_20090504_1164	ELISPOT_example_Package.tab-delimited.zip	ELISPOT_report_file	Completed	Bea	05/04/2009 10:50:40
Bea_20081219_1084	Flow_Cytometry_example_Package.tab-delimited.zip	FCM_report_file	Completed	Bea	12/19/2008 17:04:26
Bea_20081125_1059	Genotyping_Small_Example_Package_111kb_XML.zip	Affy_GCOS-GDAS-GTYPE_Genotyping_output	Failed	Bea	11/25/2008 15:58:35

11.5 DATA SUBMISSION PROCESSING WORKFLOW

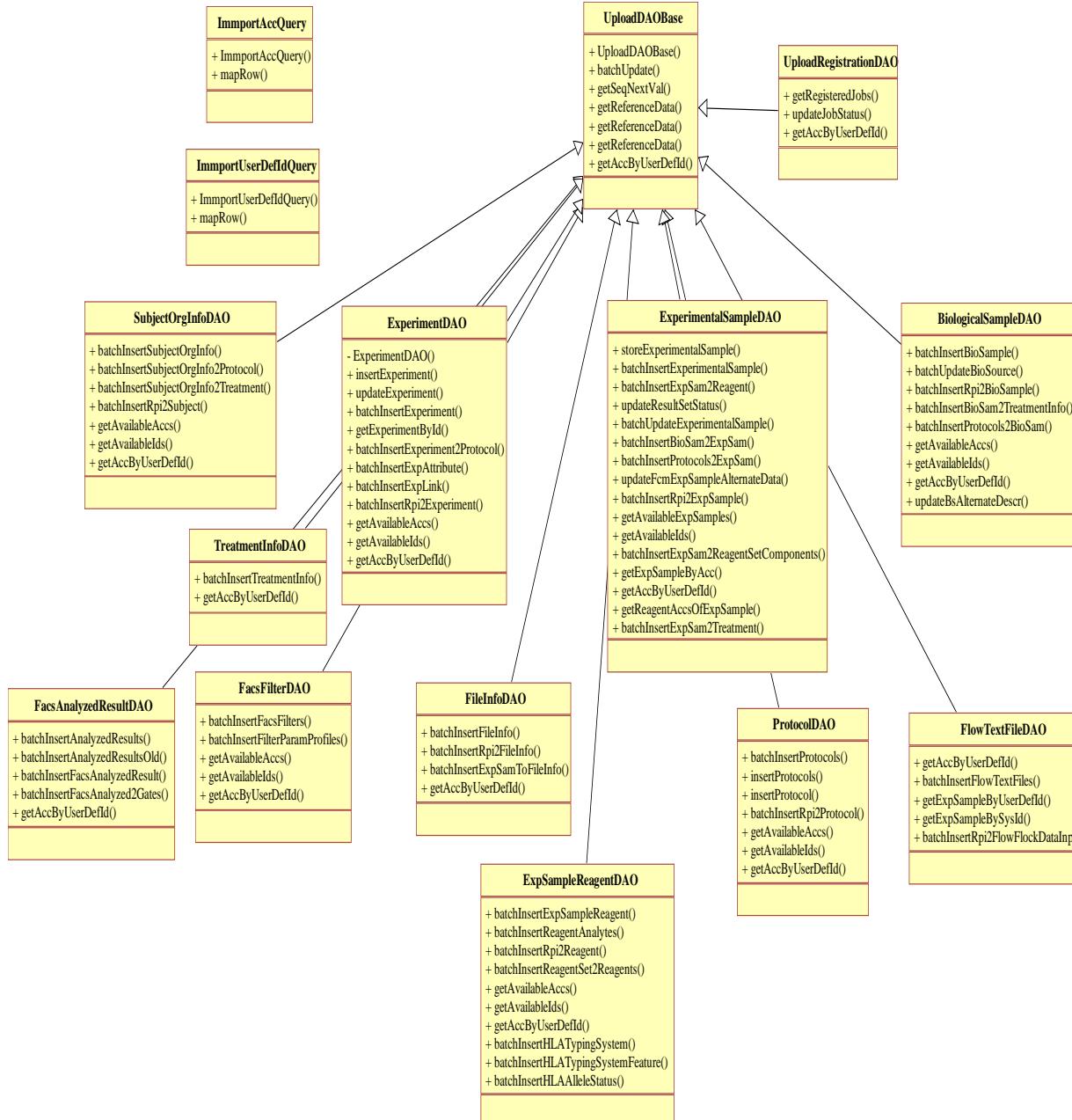
When a data package is uploaded, the system assigns a ticket number and records an entry in the upload registration table for this package. The batch upload daemon runs every 5 minutes and checks the database registration table looking for packages to process. When the upload daemon finds a package to process, it does the following:

1. Marks the package as processing
2. Unpacks the package in a working directory
3. Parses each template in predefined order, because there can be dependencies between the templates. For example, every biological sample record must link to a subject record, so the subject template, must be loaded before processing the biological sample template
4. As each template is parsed, the records are inserted into the ImmPort database.
5. Any file in the package that do not match the template names, is treated as a file to be archived and linked to the project
6. If the package is successfully loaded, the user is notified that the submission was completed successfully. If any error occurs during the load, the entire package is rolled back and the user is notified of the errors.

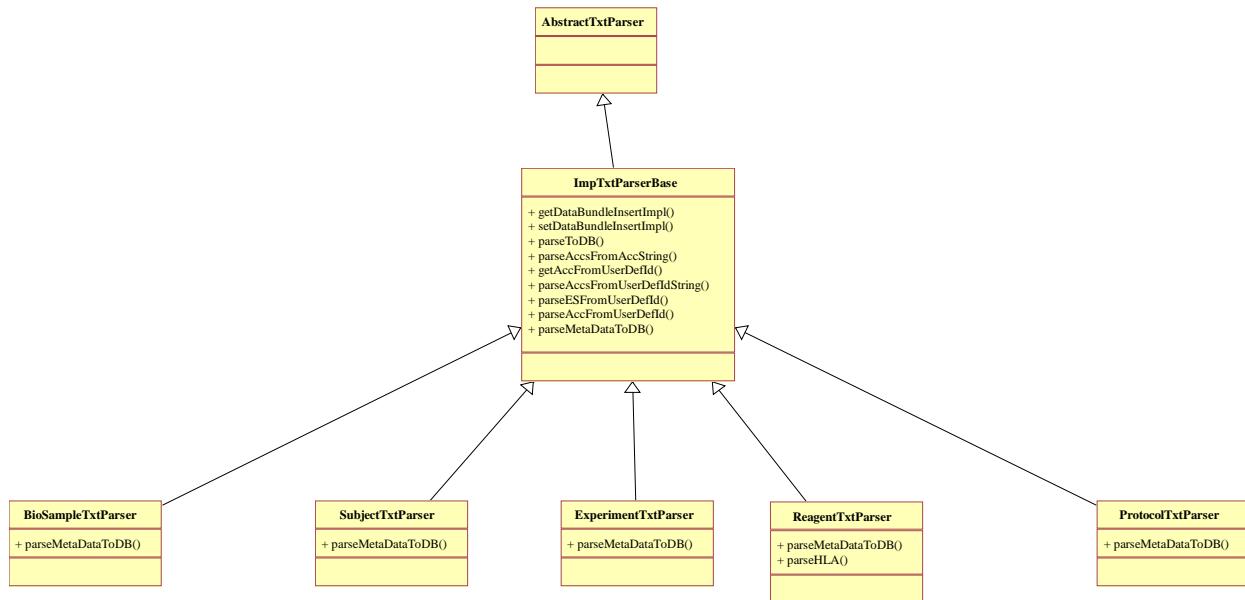


11.6 DATA SUBMISSION CLASS DIAGRAMS

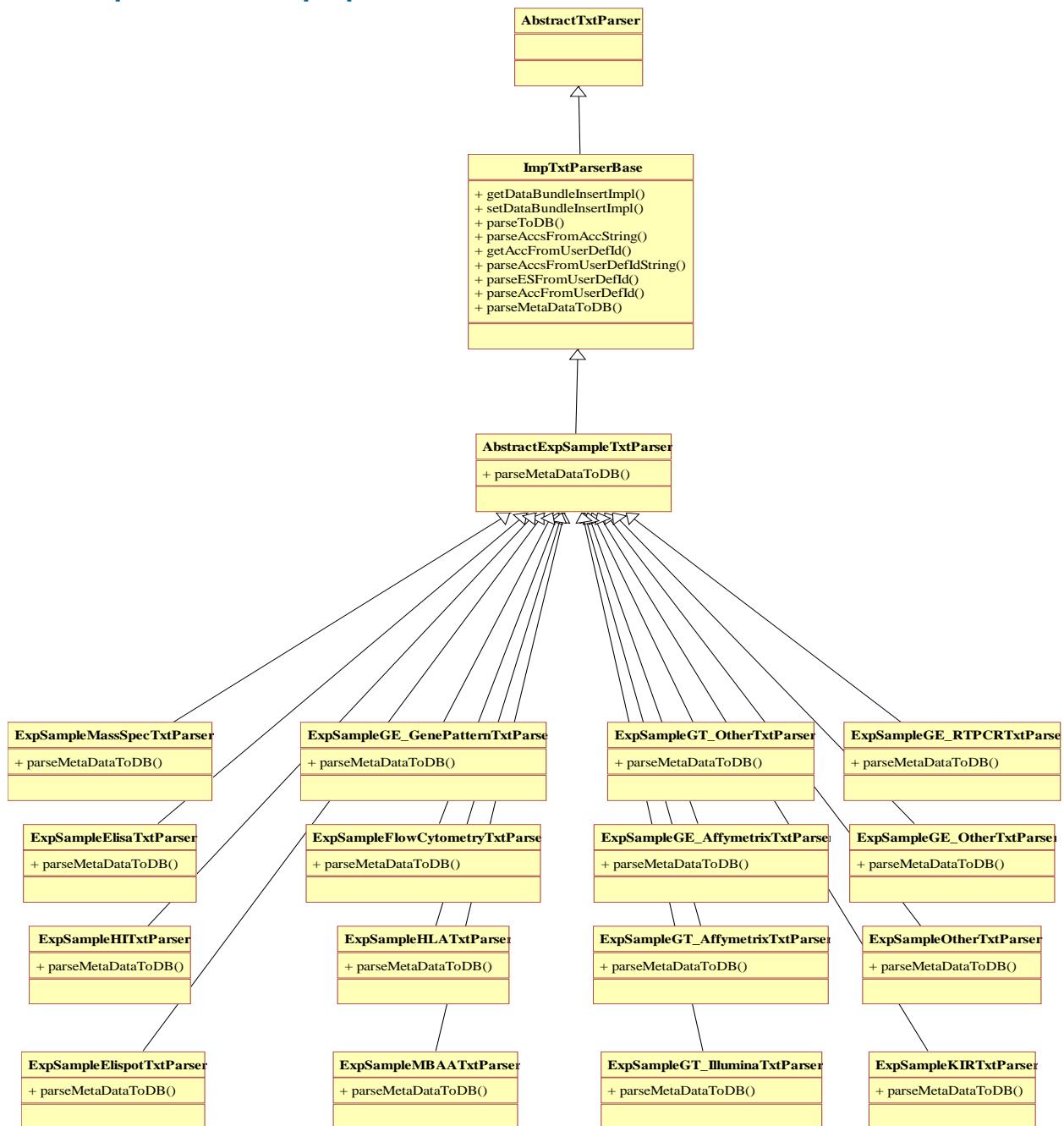
11.6.1 DAO



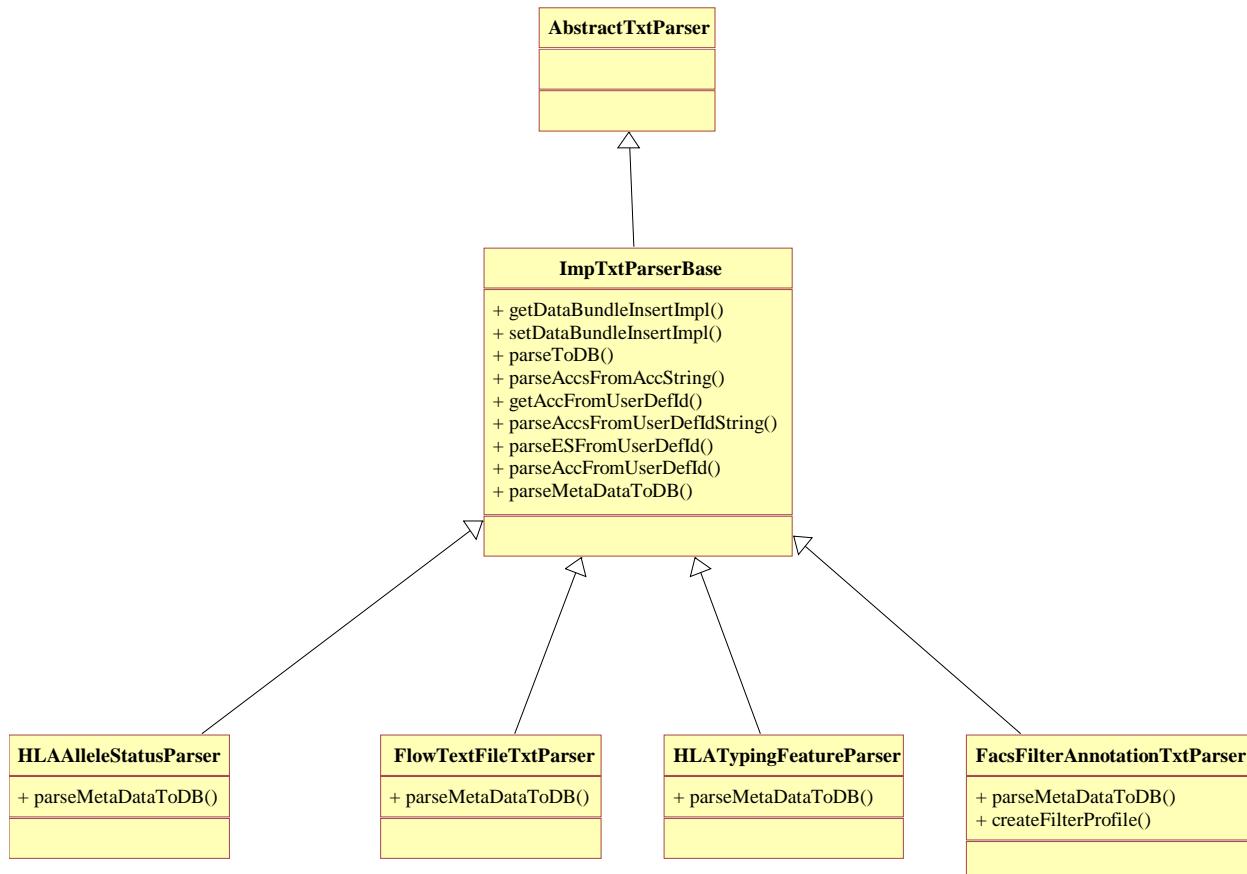
11.6.2 Entity Parsers



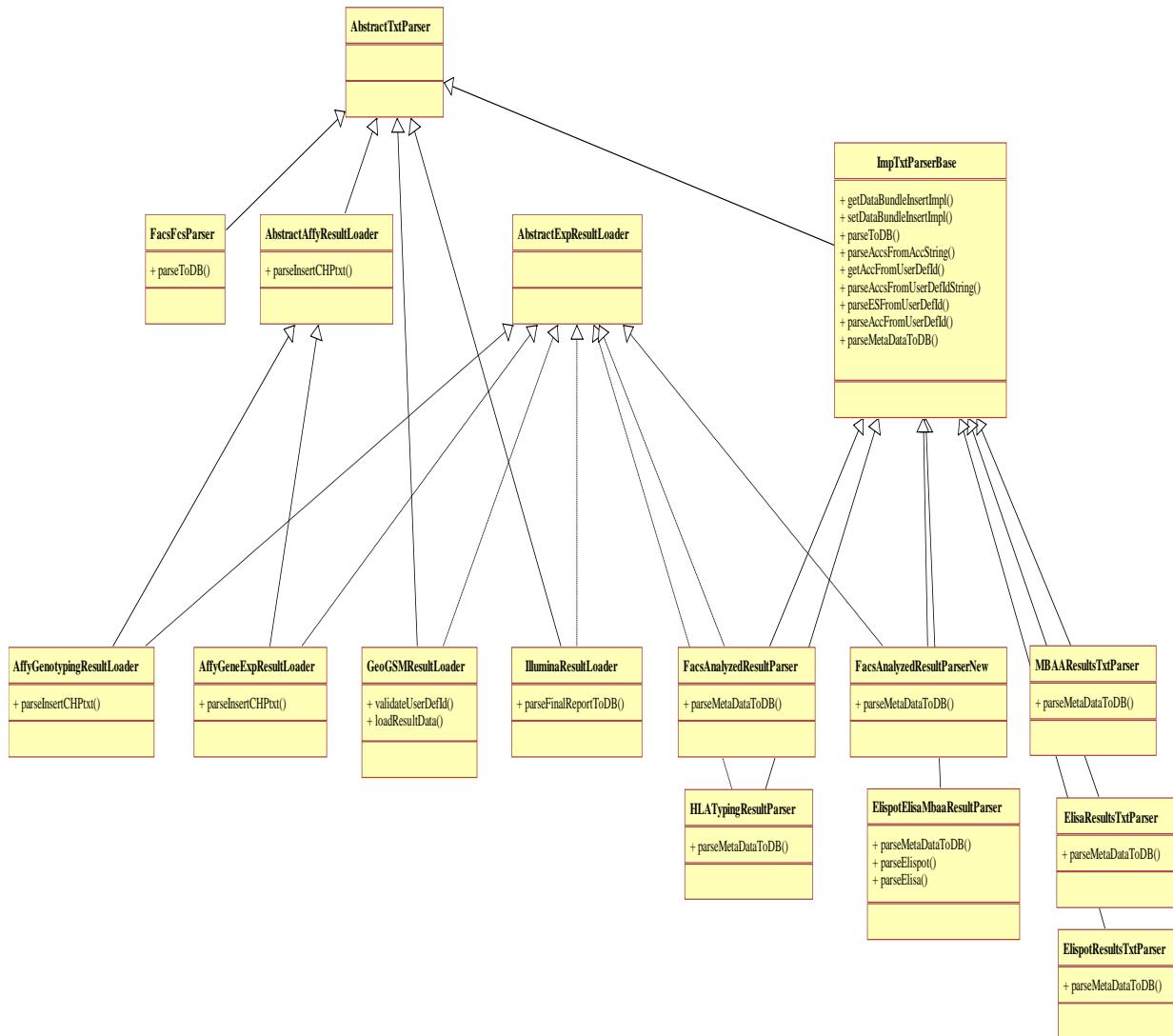
11.6.3 Experimental Sample parsers



11.6.4 Other Parsers



11.6.5 Result parsers



11.7 STUDY DESIGN AND CLINICAL DATA SUBMISSION

The process for loading the study design and clinical data, such as the content captured in Case Report Forms (CRF's) in clinical trials, is more interactive than the process for submitting experimental metadata and results. As noted earlier, the process for extracting, transforming and loading this data content is more time-consuming for the data provider as well as the BISC staff due to the breadth of data, the complexity of the data and the lack of standardization. It is assumed that there will be back and forth discussions and quality control checks and verifications throughout the process.

In general, the process for loading study design and clinical data results follows these steps:

1. The data provider sends the Study Protocol along with annotated CRF's to the BISC team along with either example or actual clinical data files in SAS or text or an alternative readable format.
2. The BISC team evaluates the Study Protocol to obtain an understanding of the Study Design and expected clinical data and experimental (mechanistic) assay results that might also be collected.
3. The BISC team populates a preliminary set of Study Design templates that are shared with the data provider for validation. This information includes information such as the study purpose and aims, study personnel, study arms, planned visits and the linkage of arms and visits, a glossary, inclusion and exclusion criteria, etc. These templates are detailed below in Table 11-1.
4. The data provider validates the Study Design templates with the BISC team.
5. The BISC team then maps the clinical data files provided to the Clinical Data templates used to load the CRF question and answers into ImmPort. A spreadsheet is utilized to illustrate column by column the matchup between the two file formats as well as to detail any rules that need to be applied to modify the content. Examples could include splitting one column into two to reflect value and unit and standardizing the values to a more standard vocabulary.
6. It is then decided which team (BISC or the data providers) would create the translation code to create the load files based on the data provider's clinical data file and the destination template format. This depends on the availability of staff on both teams as well as the technical sophistication of the data provider team.
7. After the load files are created and run through local QC scripts by the BISC team, the study design and clinical load files are submitted into ImmPort via the typical upload mechanism where a destination project is selected. These load files are flagged as clinical load packages automatically and diverted to a set of separate loading scripts.
8. The Load files are stored as Oracle external tables and further QC'ed within the database.
9. Following this QC, the data is loaded from the external tables into the clinical database schema noted earlier, and is available for review within ImmPort.

The clinical data templates are not currently listed on the website due to their evolving nature and since this content is only relevant for a subset of the programs at DAIT. Over time these will also be moved to the website. Each of the files listed in the tables below is a text file taken and loaded via standard Oracle loading mechanisms into the database. The load files generally map to the tables defined earlier in the document.

When communicating with the user community, these files are packaged into an Excel workbook with multiple worksheets for each template below. In these worksheets the columns are annotated comments and descriptions and simple validation is accomplished with Excel drop down values for fields with controlled vocabulary restrictions. Via color coding, the worksheets detail which columns are required to be populated and which are not required. Additionally, columns with more complex rules are itemized; in some cases a set of columns must exist and at least one of those columns must have data. For example, in

a load file defining reference ranges there must be an upper limit and a lower limit and one of them must have data. Table 11-1 summarizes the Study Design Data Loading Templates and the type of content contained in the files, and Table 11-2 summarizes the same information for the Clinical Data Loading Templates.

Table 11-1. Study Design Data Loading Templates

Template Name	Purpose
study.txt	Provides overall description of a study in conjunction with the Observational_Parameters and Intervention_Parameters. Many attributes of a study are taken from the clinicaltrials.gov Protocol Data Element Definitions including the entire contents of the ancillary Observational_Parameters and Observational_Attributes. Note that the slightly inconsistent nomenclature of Brief_Summary and Detailed_Description are taken directly from the Protocol Data Element document.
study_personnel.txt	This is essentially a contact information table for persons who are involved in conducting studies. The contact information is intended to be the latest contact information for a person. Titles and roles of a person involved in conducting the study are stored in the table that associates people with studies since these attributes are properties of the person's association with the study and not of the person. The person's organizational affiliation during the study is also recorded in the study association table since a person's current organizational affiliation may not be the same as their affiliation during the study.
visit_2_arm.txt	If all ARMs or Cohorts have the same set of visits, then every ARM or Cohort will be associated with every visit. In interventional studies it is possible for the different ARMs to be differentiated only by the drug regimens (e.g., placebo vs. experimental drug) occurring during the visits, so the ARMs may be distinguishable at the event level only by differences in the attributes of actual substance merge events for each subject.
subject_measure_definition.txt	Defines computed measures on subject. For example, a daily average allergy score may be computed from a set of specific allergy scores.
planned_visit.txt	Represents both the classical CDISC visit which is a clinical encounter with a patient, and a set of planned activities that may occur between clinical encounters with a subject. The beginning of the planned event set is defined in terms of a range (Min_Start_Day and Max_Start_Day) within which the event can occur. If the event must occur on a specific study day then Min_Start_Day must equal Max_Start_Day.
planned_event.txt	An EAV representation of all the possible events. Table_Name is the name of the table that will record the actual event. Category is one of "Assay", "Biomaterial Transform", "Complex Event", or "Data Transform". Name is the identifier for a specific kind of event (for example the name of a lab test). If the Category of event is "Complex Event" then it is a composite of multiple events. The set of events grouped as a complex event is defined by the "Event Panel" and "Event Panel Component" tables. Hence if a planned event has the category "Complex Event" then the event's identifier must appear as a foreign key in the Event_Panel table. An example is an "Allergen Skin Reaction Test" which consists of a series of substance merge events (inoculating the skin with an allergen). Each of the inoculations is followed by assessments at 15 minute and 24 hour intervals. Hence the "Allergen Skin Reaction Test" is a composite of many substance merge and assay events.
period.txt	Period captures the ClinicalTrials.gov definition of study stages used for reporting participant flow through a study. Period is defined in the "Basic Results" Data Element Definitions. "Period" at ClinicalTrials.gov is equivalent to sed for a CDISC "Epoch". If there is only one period in a study its value should be "Overall Period". This table is designed to capture participant flow at the level of aggregation used by ClinicalTrials.gov.
outcome_measure_definition.txt	Description of an outcome measure without actual measure values. This description will be used across all ARMs of the study. The outcome measures defined in this table are AGGREGATE measures for one of the populations defined in the study, and therefore do not apply to individuals.
observational_parameters.txt	Attributes descriptive of observational studies

Template Name	Purpose
interventional_parameters.txt	Attributes descriptive of interventional studies.
inclusion_exclusion.txt	Defines the criteria by which subjects are excluded from or included in a study.
glossary.txt	Captures term definitions provided by a study.
event_panel_components.txt	If the Event_Panel is visualized as a table, this table defines each event that appear in one row of the Event_Panel.
event_panel.txt	Defines a complex event. A complex event can be viewed as a table in which each row is a sequence of events. The event panel describes the entire collection of event sequences in its Event_Panel_Description attribute and describes a single row in the Component_Set_Description attribute. The actual events comprising a row are described in the Event_Panel_Components.
arm_or_cohort.txt	An ARM is a specialized population selection rule

Table 11-2. Clinical Data Loading Templates

Template Name	Purpose
Actual_Visit.txt	A clinical encounter during which events such as assessments, biosampling, and substance merging occur for a particular subject. An actual visit is always associated with a single subject
Adverse_Event.txt	Records all information typically collected for adverse events. If the adverse event is detected as part of a planned subject assessment, then the Assessment_Acc_Num foreign key will be non-null. If that key is null then the adverse event is detected through unplanned means such as subject self-reporting or clinician observation.
AE_Study_Defined_Category.txt	Provides a means to categorize and further annotate Adverse Events
Arm_or_Cohort_2_Subj_Org_Info.txt	Provides the linkage between subjects and the Arms to which they may belong in many studies.
Assessment_Component.txt	It is expected that an assessment component may be either a study specific component or a standard component (such as a standard blood test or an erythema component of a EASI assessment)
Assessment_Panel.txt	Note that some assessments, such as subject self-assessments performed outside any clinical encounter, do not have an actual visit. In order to place the assessments not associated with an actual visit on the study timeline, the assessment panel must have a study day. An example of a self-assessment with a study day but no visit is the allergy symptom diary cards which subject filled out for allergy studies.
Biosampling.txt	Contains BioSampling event records prepared that pertain to BioSampling events such as Clinical Lab Tests and Assessments.
Biosampling_2_bs.txt	This prepared file is used to directly link biosampling records to a biological_sample record. For most Clinical data such as Lab Tests and Assessments, there is a 1:1 correlation between biosampling and biological_sample records.
Lab_Test.txt	A laboratory process operating on a biological sample that produces a single value. Reference ranges specify the bounds of the determined value that are considered normal.
Lab_Test_Panel.txt	A group of lab test conducted on a given sample (e.g., a Complete Blood Count panel). A panel may contain only one test (e.g., and IgE level). In that case the panel and the lab test may have the same name.
Protocol_Deviation.txt	Reports any deviations from the study protocol for a given subject along with the reason, resolution, timing and adverse event linkage.
Reference_Range.txt	Defines value ranges for a lab test value that are considered normal. If a single value is considered normal rather than a range then upper and lower limits will be the same. If only an upper or lower bound is defined then the other bound shall be null. If a Category is defined for a reference range then all values between lower and upper bound will be represented by the category.

Template Name	Purpose
Reported_Early_Termination.txt	Data regarding any early terminations for a given study subject with the day reported, reason, and linkage to adverse event if appropriate
Study_2_Protocol.txt	Links any protocol information and documentation to the study.
Study_PubMed_Info.txt	This table contains PubMed information content (PubMed_ID, Title, Date, Authors, etc.) that a researcher or research group wishes to link to a Study.
Subject_Measure_Results.txt	Set of results for a defined Subject_Measure_Definition record.
Substance_Merge.txt	A compound is merged with a subject via some route of administration. The compound can be either for test purposes (e.g., an allergen for an allergy test) or for intervention (a drug treating a condition).

12.0 ADVANCED AND BASIC SEARCH DESIGN PACKAGES

The search module was designed to allow the user to query for summary information about objects by entering ad-hoc queries, based on the properties of each class of object, plus associations between objects. The module was used to provide support for both advanced search and basic search UI's. Examples of the types of objects represented in ImmPort are: subjects, experiments, experimental results, studies, genes, proteins, etc.

During the design process we had several goals and constraints that are outlined below:

- Develop the system to be easily maintained.
- Develop the system to be extensible to other objects within the ImmPort schema.
- Reuse the tabular result page design already developed.
- Reuse the Hibernate object model.
- Make much of the system configurable or meta-data driven.
- Support the use of previously saved query results (lists) in the complex queries.
- Keep the interfaces between the layers clean and simple, allowing reuse in other parts of the application.

Below is a screen shot of the Advance Search module for Research Queries. The screen has 4 major components that will be discussed on more detail later in the document.

- Toolbar.
- Query Tree Panel.
- Criteria Panel.
- Summary Results Panel

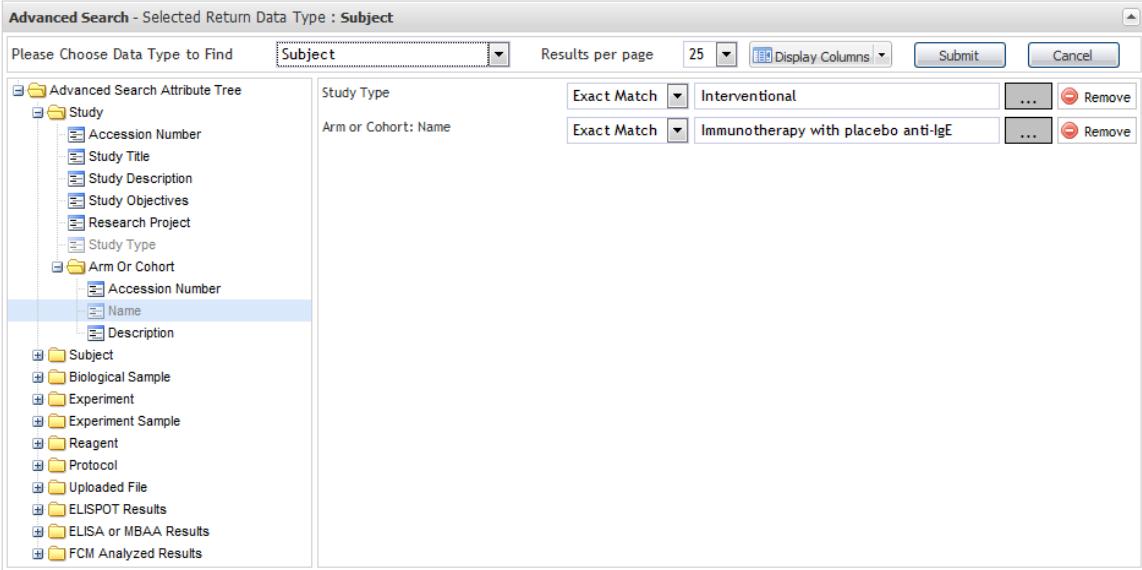
This query and UI framework has been implemented for basic and advanced research search interfaces and for advanced reference query in the same manner. It is also used in selected queries across the system when it is useful to return a subset of objects based on a criteria specific to a given page. This framework re-use allows for greater maintainability and expedites the addition of new search interfaces to the system.

12.1 GENERAL OVERVIEW

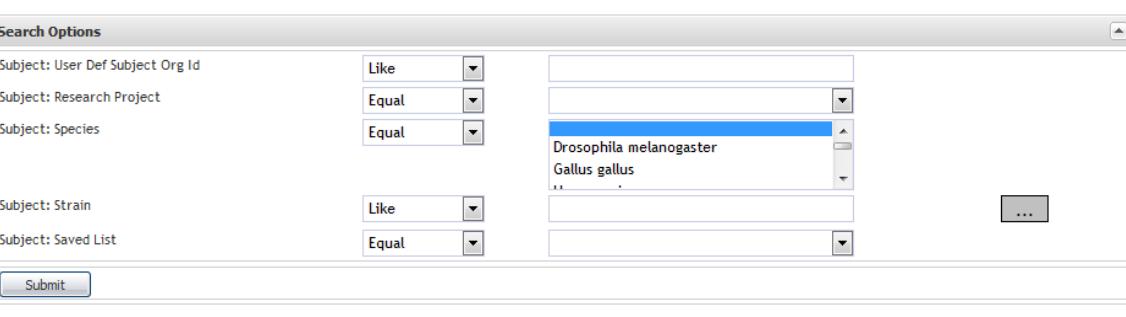
Multiple search screens provide the entry point into the search framework described, but all utilize the same presentation layer technology. The Advanced Search research page appears with a tree view with available search attributes on the left side, the selected search attributes on the right. The user is prompted to return the type of object that they wish to return at the top.

Using Advanced Search:

- ▶ Choose the type of data you want to find (e.g. subjects, samples) from the drop down list.
- ▶ Choose a folder to refine your search. It can be the same or different data type from the data you want to find.
- ▶ Click on the data feature(s) you want to search.
- ▶ Enter search terms for one or more features.



Basic Search screens have a fixed and minimal set of attributes with which to query for a fixed object to return, as shown for the Subject Search page.



After a search is submitted, the attribute panels disappear and the query results summary grid appears. This result summary page provides extensive functionality, including:

1. Sorting by each column
2. Re-ordering of columns by drag and drop
3. Hiding and Un-hiding columns
4. Page through result records
5. Selection of individual object records to view further details
6. Save Items or Save All Item rows to a list for later re-use
7. Export of the summary query results to Excel or PDF

Viewing of detail records and list management and use will be discussed in later sections.

Using Advanced Search:

- ▶ Choose the type of data you want to find (e.g. subjects, samples) from the drop down list.
- ▶ Choose a folder to refine your search. It can be the same or different data type from the data you want to find.
- ▶ Click on the data feature(s) you want to search.
- ▶ Enter search terms for one or more features.

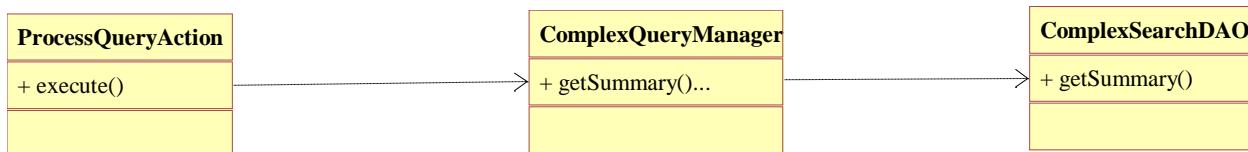
To modify your search, please expand the search box ~

Advanced Search - Selected Return Data Type : Subject

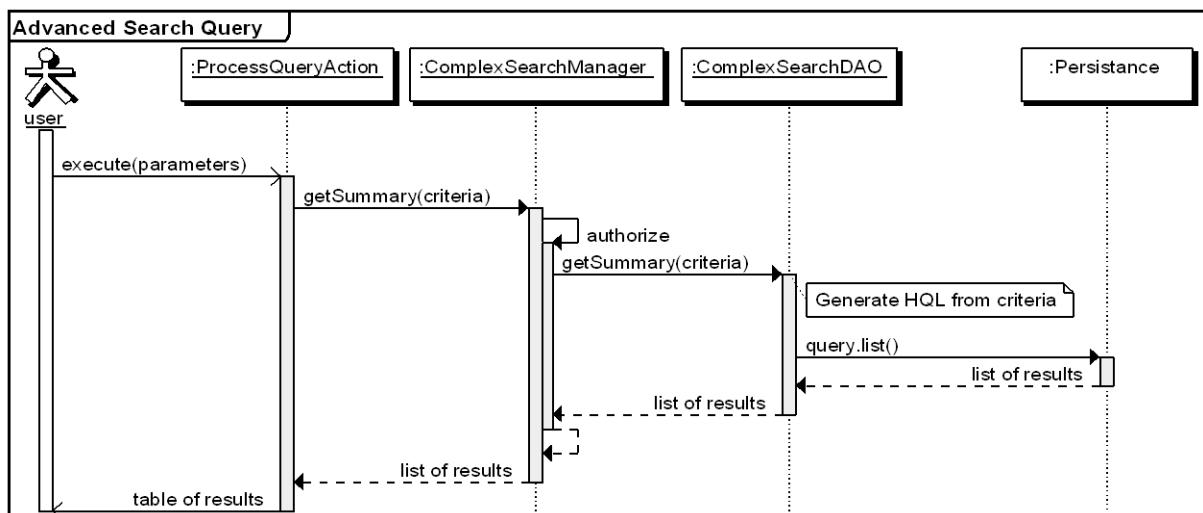
Tip: You can select a range of rows by holding down the 'Shift' key and then clicking the first and last rows you wish to select. This works best if you do not click on the check boxes. [More tips.](#)

Sub Org Accession	Species	Race/Stra	Gender	Description	Subject Phenot	Project Title	Study Title	Arm
SUB73370	Homo sapiens	White	Male	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with placebo anti-IgE
SUB73373	Homo sapiens	Asian	Male	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with placebo anti-IgE
SUB73377	Homo sapiens	White	Female	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with placebo anti-IgE
SUB73378	Homo sapiens	White	Female	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with placebo anti-IgE
SUB73384	Homo sapiens	White	Male	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with placebo anti-IgE
SUB73389	Homo sapiens	White	Male	Placebo + Ragweed IT	Ragweed-indu	ITN019AD - Clinical Trial	Allergen immunotherapy	Immunotherapy with placebo anti-IgE

12.1.1 Class Diagram



12.1.2 Sequence Diagram



12.1.3 UI generation

The query tree and form attributes are defined in a metadata file using JavaScript Object Notation (JSON). JSON is an emerging standard used to configure JavaScript web components, and used as the data interchange format between components. The metadata file is input into a JavaScript query tree engine, which generates the tree view in the UI. When a property is chosen from the query tree, the node is passed to a JavaScript component to generate the form input field.

12.1.4 Form processing

When the form is submitted, an AJAX call is made to a Struts Action sending the form parameters entered by the user. The Struts Action class generates a complex query criteria object that is passed to the DAO layer to construct the query. The complex query constructor generates a Hibernate HQL query that submits a SQL query to the Oracle database.

12.1.5 Results processing

The results of the query are packaged into a JSON object and passed back the UI. The results are then displayed using a grid control.

12.2 PRESENTATION LAYER

12.2.1 Technologies

For the presentation layer we use a mix of JSP, JavaScript, Ajax and metadata files to present the query interface to users. By using metadata files and generic code, we are able to present alternative versions of the query interface, without the need for additional coding. We have chosen the EXT-JS JavaScript libraries, to build the query tree shown in the Advance Search screens, and to build the summary table grids.

12.2.2 Query UI

The property query tree on the left panel of the page is built using a combination of JavaScript and metadata file, which represents the entities and the properties that can be queried by the user. As the users clicks on query properties from the tree, an input field is displayed in the right side panel. The input field can be a simple input text box, drop down list for controlled vocabulary terms, or the user can click on a link to display a popup that lists all terms previously entered for this property.

12.2.3 Query Results

The results of the query are displayed in a tabular grid, using EXT-JS components. The user has several options:

1. Sort results based on column selection.
2. Selecting rows, for detail view.
3. Selecting rows to save in a result set.
4. Export the results to a file, that can be downloaded

12.3 FORM PROCESSING, BUSINESS OBJECT LAYER

Because the number of different properties that can be chosen by the user to qualify the query, is subject to change, we have developed a generic methodology for processing the parameters and constructing the SQL to be submitted. In the Action class, based on the parameters supplied, a ComplexSearchCriteria will be constructed. The ComplexSearchCriteria object has properties to contain the project Id's and the Name of the primary object of interest (Subject, Experiment, etc). In addition, it contains a collection of items that represents the properties and values passed in by the form submission. Each item will contain the value entered for this property, the name of the property and the search operator used to qualify the join to this property. The field name contains information representing the object and the property of the object that is being queried. A query operator can be “Exact Match”, “Like”, “Greater Than”, etc. For example, if the user is looking for Subjects that are “Male” and have Biological_Samples where the sample_type_name is “Tissue”, the action would be passed the following parameters:

StartingObject: Soi
SoiGenderValue: Male
SoiGenderOperator: “Exact Match”
BsSampleTypeNameValue: “Tissue”
BsSampleTypeNameOperator: “Exact Match”

12.4 QUERY CONSTRUCTION

The query constructor engine uses metadata that represents the paths between objects and information from the Hibernate mapping files to construct a query similar to the one in the section below. The Hibernate criteria classes are utilized to dynamically build the Hibernate Query language where clauses based on the property, value, and operator passed from the UI through form submission. The names of the properties in the UI tier are named to match the object properties in the Hibernate mapping files to simplify this construction.

12.4.1 Hibernate Query

An example of the resulting Hibernate query is show below.

```
select distinct soi.id
from SubjectOrgInfo soi
inner join soi.biologicalSampleSet bs
where soi.researchProjectInfo.id in (:projectId)
    and soi.gender = 'Male'
    and bs.bsType = 'Tissue'
```

12.4.2 SQL Query

An example of the resulting SQL query generated by Hibernate from the Hibernate query is shown below:

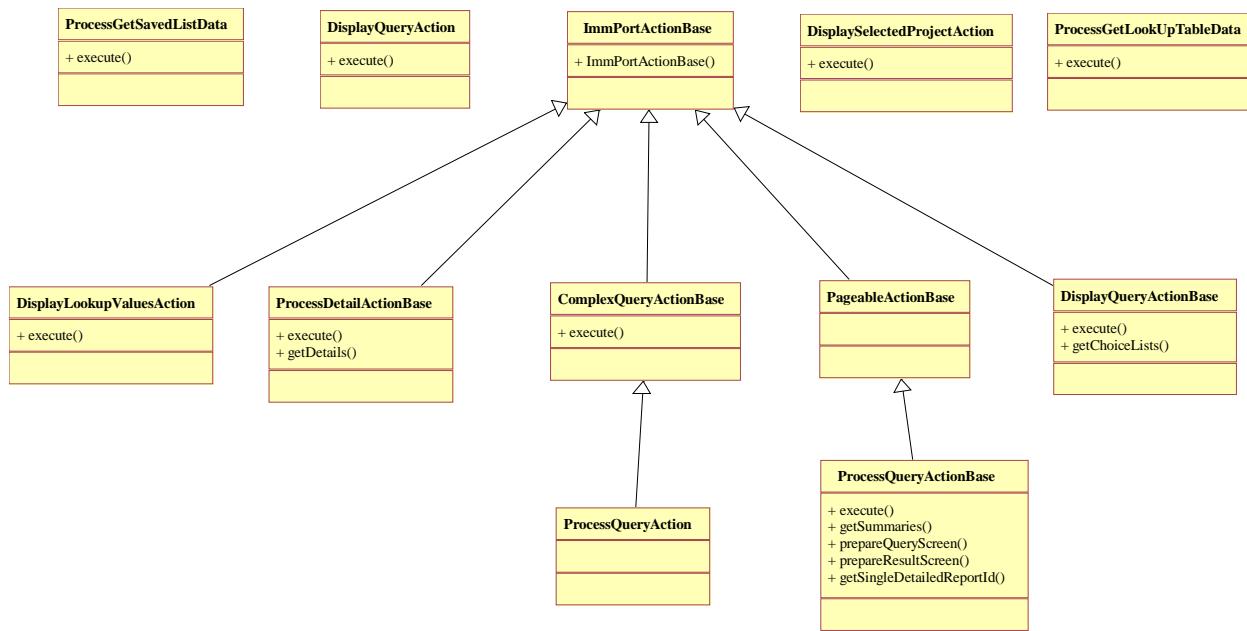
```
select * from
( select distinct subjectorg0_.SUBJECT_ORG_ACC_NUM as col_0_0_
  from SUBJECT_ORG_INFO subjectorg0_
  inner join BIOLOGICAL_SAMPLE biological1_
    on subjectorg0_.SUBJECT_ORG_ACC_NUM=biological1_.SUBJECT_ORG_ACC_NUM
   where (subjectorg0_.PROJECT_ID in (?))
     and (subjectorg0_.GENDER) = 'Male'
     and (biological1_.SAMPLE_TYPE_NAME = 'Tissue') )
where rownum <= ?
```

12.5 QUERY PROCESSING – RESULTS DISPLAY

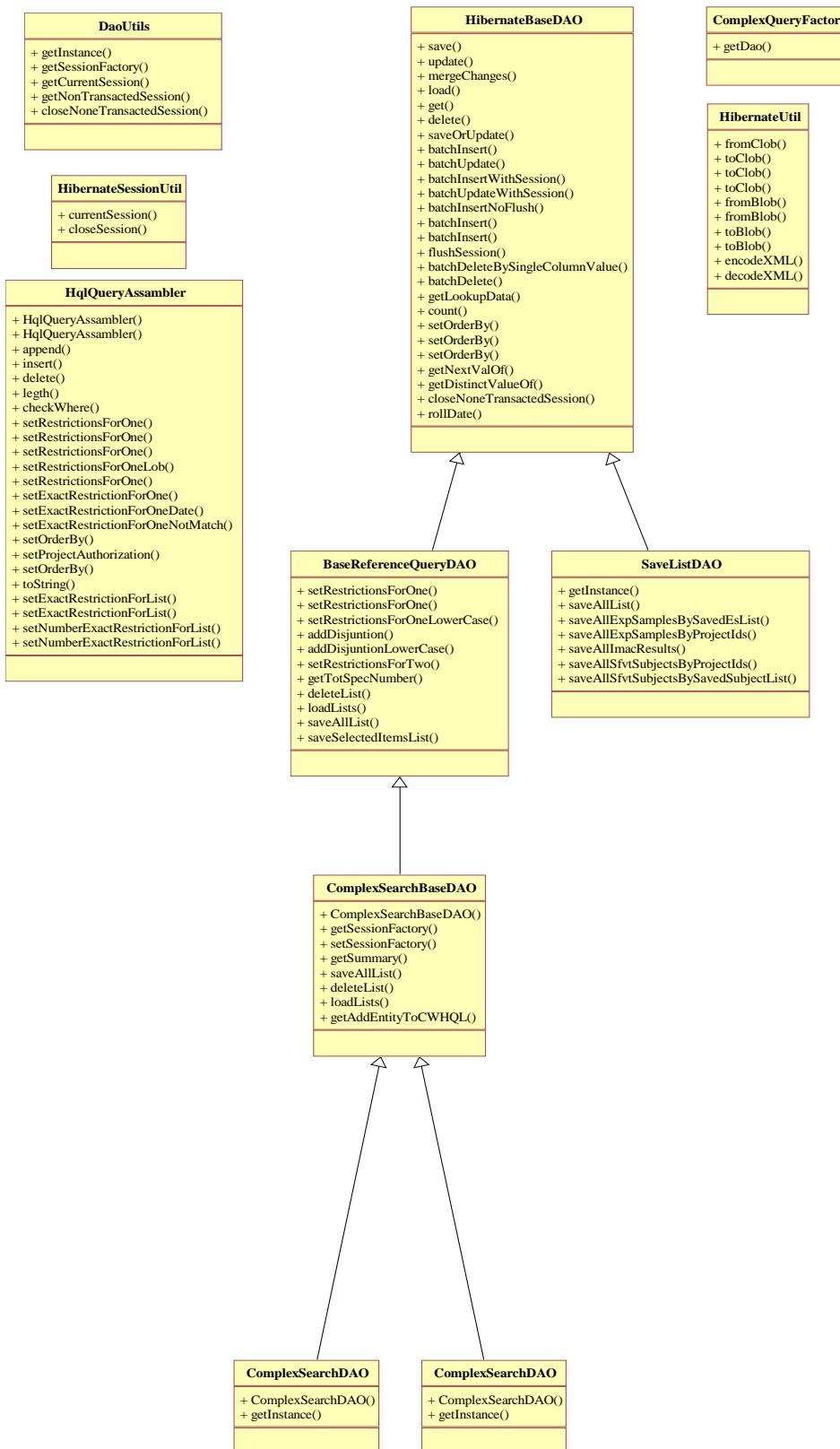
The SQL query generated by query constructor is then submitted to the Oracle database. The results are processed and packaged into a JSON object. The JSON object is then passed back to the AJAX call that initiated the form submission, and displayed using the Table Grid control.

12.6 SEARCH CLASS DIAGRAMS

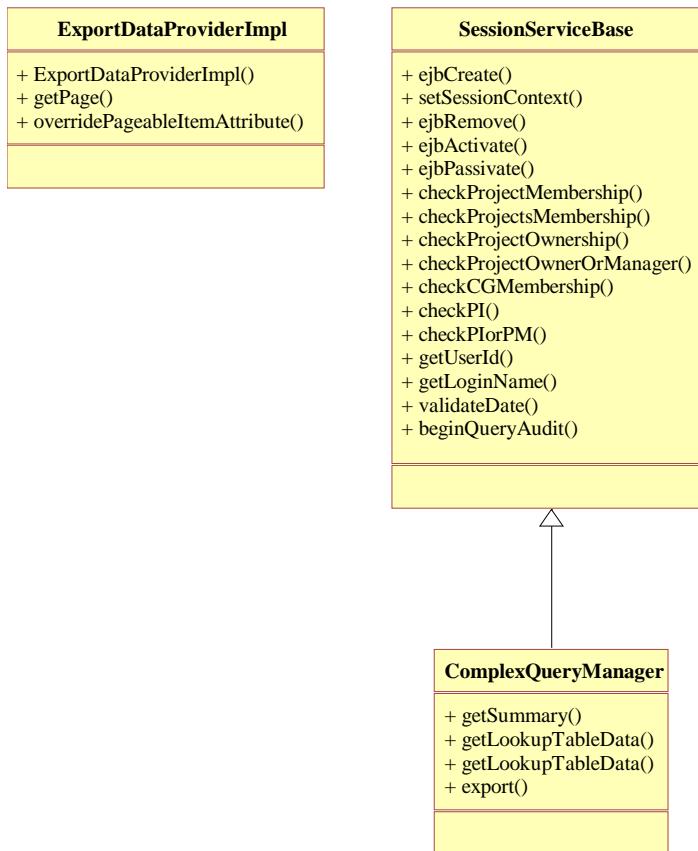
12.6.1 Action – Presentation Layer



12.6.2 DAO



12.6.3 EJB



12.6.4 DTO



13.0 RESEARCH DATA DETAIL PAGE DESIGN PACKAGES

The detail screens for the research entities, with the exception of the study detail screen, all follow a similar pattern and use the same technology stack. The top panel on the detail screen shows the detailed information for the entity. Below this panel there are a series of panels that can be expanded or collapsed and are loaded on demand. These panels re-use the same technology used to display the summary grids in the advanced search module described previously.

The screenshot displays a detailed view of a research subject's information. At the top, two tabs are visible: 'SUB73366' and 'SUB73367'. The main content area is titled 'Subject Summary' and contains the following data:

User-Defined ID:	ITN019AD_001_001						
Study / ARM / Study Pl:	<table border="1"><tr><td>Title</td><td>ARM Name</td><td>Study PI</td></tr><tr><td>Allergen immunotherapy Co-administered with Omalizumab</td><td>Immunotherapy with anti-IgE</td><td>Thomas Casale</td></tr></table>	Title	ARM Name	Study PI	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with anti-IgE	Thomas Casale
Title	ARM Name	Study PI					
Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with anti-IgE	Thomas Casale					
Subject-Organism Accession No:	SUB73366						
Description:	Omalizumab + Ragweed IT						
Organism:	Homo sapiens						
Taxonomy ID:	9606						
Enrollment Age:	47						
Gender:	Male						
Religion:							
Race:	Asian						
Ethnicity:	Not Hispanic or Latino						
Population:							
Affection Status:	0						
Subject Phenotype:	Ragweed-induced seasonal allergic rhinitis						
Subject Pedigree ID:							
Family Pedigree ID:							
Mother Pedigree ID:							
Father Pedigree ID:							
Project Title:	ITN019AD - Clinical Trial shared to Semi-Public Workspace (SPW) Project						

Below the summary are expandable sections: 'Treatments', 'Protocols', 'Biological Samples', and 'Assessments'. Under 'Assessments', the 'Clinical Assessments' tab is selected, showing a grid of vital signs data:

Page	1 of 116	Export	Displaying 1 - 25 of 2889						
Subject Acc Nu	Name	Clinically Signif	Study Day	Time Of Day	Result	Value	Unit	Organ	
SUB73366	Vital Signs	Baseline Weight	-21		69.4008	KG	69.4008	KG	
SUB73366	Vital Signs	Body Mass Index	-21		23.9633	KG/M ²	23.9633	KG/M ²	

Many of the panels in the study detail screen represent aggregated information and cannot be easily satisfied with a simple SQL query. Therefore many of these panels require the aggregation of results to construct in the business layer.

The user may request to view one or more detail pages for a specific entity. For this discussion we will assume the user requested to see the details for 2 subjects: SUB1 and SUB2. This request starts by returning a screen containing “tabs” across the top, where tab 1 one represents SUB1 and tab 2 represents SUB2. As this page is displayed an AJAX call is submitted to an Action

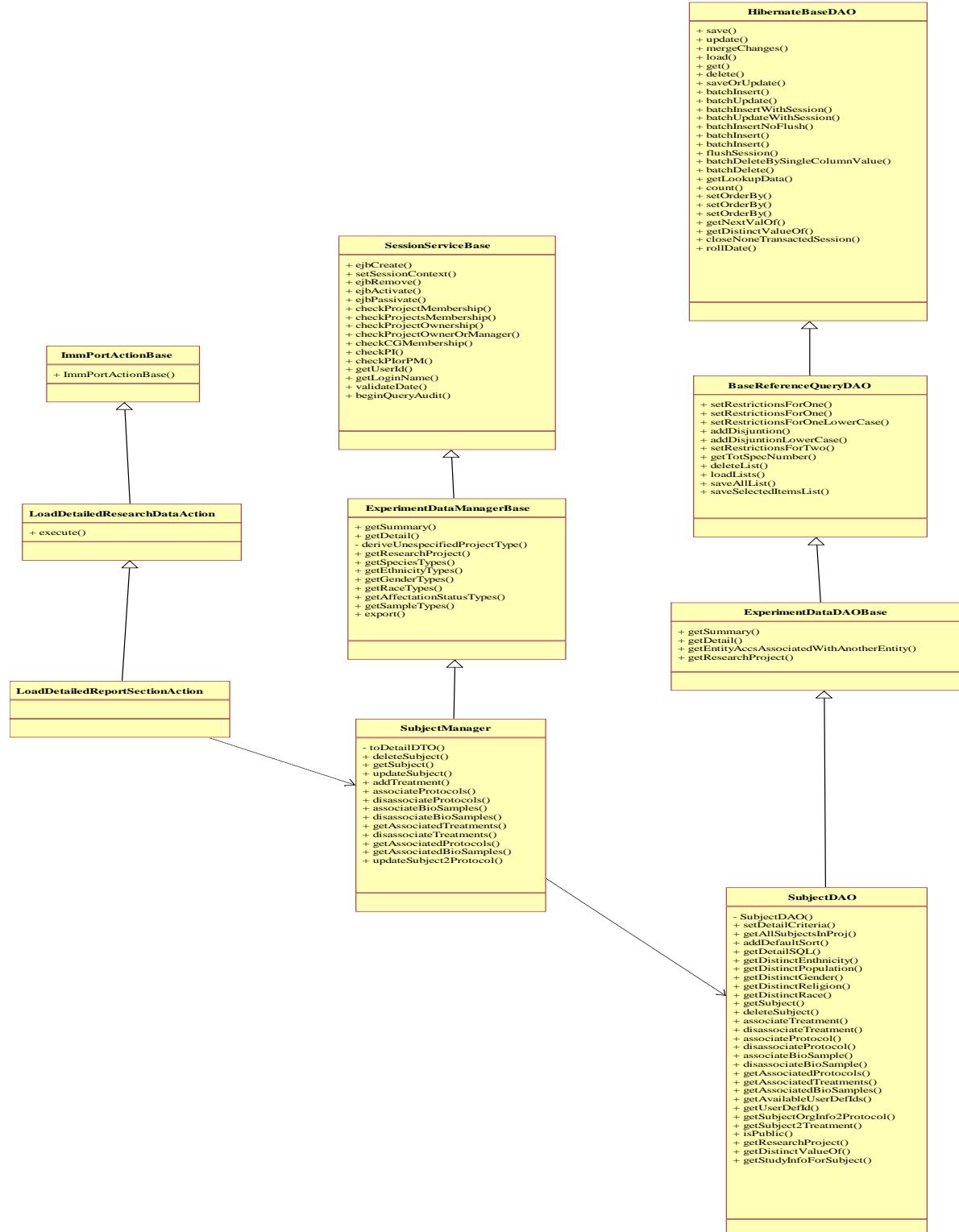
class requesting the detailed information for the first tab. The action class calls an ejb class, which calls a dao class. The results are then merged with the JSP page, and returned to the AJAX call as HTML. This HTML is then placed in the DOM and displayed. The remainder of the page displays collapsed panels that represent entities that can be associated to SUB1. The contents of these panels are not available at this time. To retrieve the content for a panel the user clicks on the expand button, which initiates an AJAX call to retrieve a summary grid representing the associated entities. This AJAX call uses the same technology as outlined in the advanced search portion of this document, to return the summary grid.

If the user clicks on tab 2, then the whole process described above repeats. If the user then clicks on tab 1 again, results are shown immediately since this content was cached in the DOM.

The design package that follows utilizes the subject object detail page as an illustrative example of how all of the research data detail pages are implemented given that the same search and UI framework is used in all pages. The final section summarizes the detail page entities and the panels that are implemented in each one.

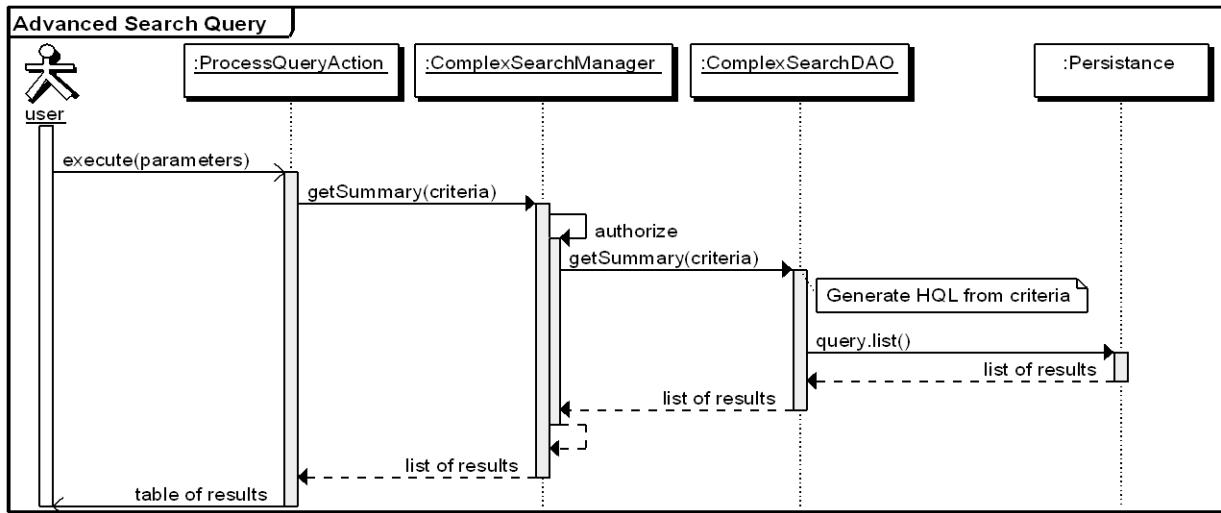
13.1 CLASS DIAGRAM – SUBJECT DETAIL

This diagram for subject detail is representative for how all the other research detail screens are constructed.

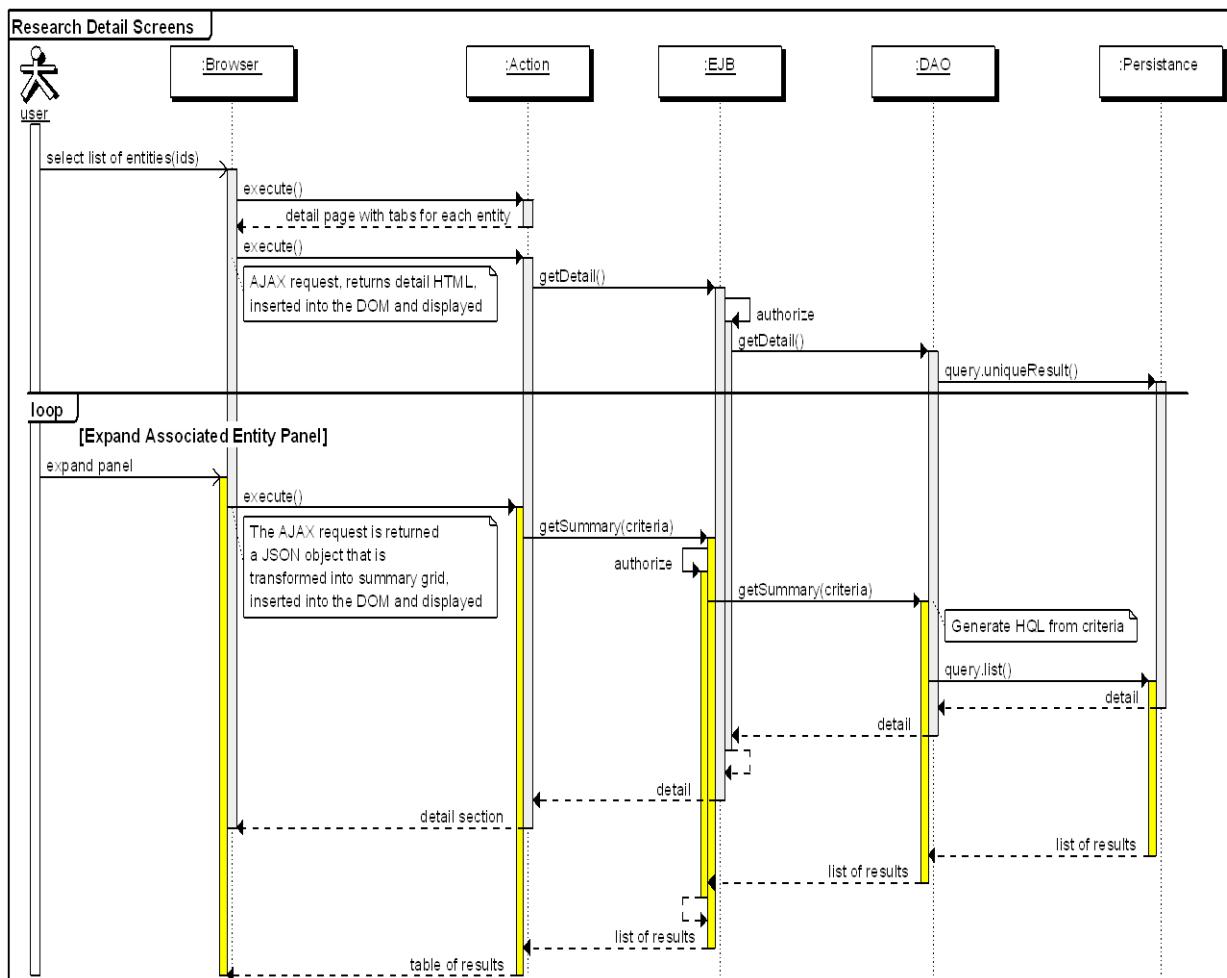


13.2 CLASS DIAGRAM – SUBJECT PANELS

The classes used to construct the panel grids are the same as the classes used to construct the advanced search grids.



13.3 SEQUENCE DIAGRAM



13.4 TABLE OF ENTITIES AND PANELS

Entity	Panel
Subject	Subject Detail
	Treatments
	Protocols
	Biological Samples
	Assessments
	Lab Tests
	Adverse Events
	Concomitant Medications
Biological Sample	Biological Sample Detail
	Treatments
	Protocols
	Associated Experiments
	Biological samples derived from this Biological Sample
	Clinical Lab Tests
Experiment Sample	Experiment Sample Detail
	Result Files
	Protocols
	Biological Samples
	Reagents
Experiment	Experiment Detail
	Experiment Samples associated with this Experiment
	Protocols
Protocol	Protocol Detail
	Experiments
	BioSamples
	Experiment Samples
	Subjects
	Studies
Reagent	Reagent Detail
	HLA Typing System
	Analytes
	Reagent Sets
Study	Study Detail
	Inclusion Exclusion Criteria
	Demographics
	Assessments
	Concomitant Medications
	Treatments
	Adverse Events
	Clinical Lab Tests

	Mechanistic Assays
	Documentation
	Glossary

14.0 RESEARCH DATA MANAGEMENT DESIGN PACKAGES

Based on the users permissions they may have the ability to edit the research entities by either updating the primary information of an entity or by creating or removing a relationship to another entity. Because editing of all research entities works in a similar fashion, we will use the editing of the subject entity to illustrate the process.

14.1 DISPLAYING THE EDIT SCREEN

From the subject detail page, the “Edit Subject Information” link is displayed.

The screenshot shows a software interface for managing research subjects. At the top, there are three tabs labeled SUB1461, SUB1463, and SUB1462. The SUB1461 tab is selected. Below the tabs is a toolbar with 'Collapse All:' and 'Expand All:' buttons. A 'Subject Summary' section is displayed, containing a link 'Edit Subject Information'. Underneath is a table with the following data:

Subject-Organism Accession No:	SUB1461
Description:	Lupus erythematosus with nephritis, ICD-9 710.5
Organism:	Homo sapiens
Taxonomy ID:	9606
Enrollment Age:	40
Gender:	Female

After choosing to update a subject the edit subject detail screen will be displayed. In the top section of the screen the previously entered values will appear in input text boxes or in drop down lists. The user makes the necessary changes and then clicks the Save Changes button.

Fields marked with an asterisk * are required.

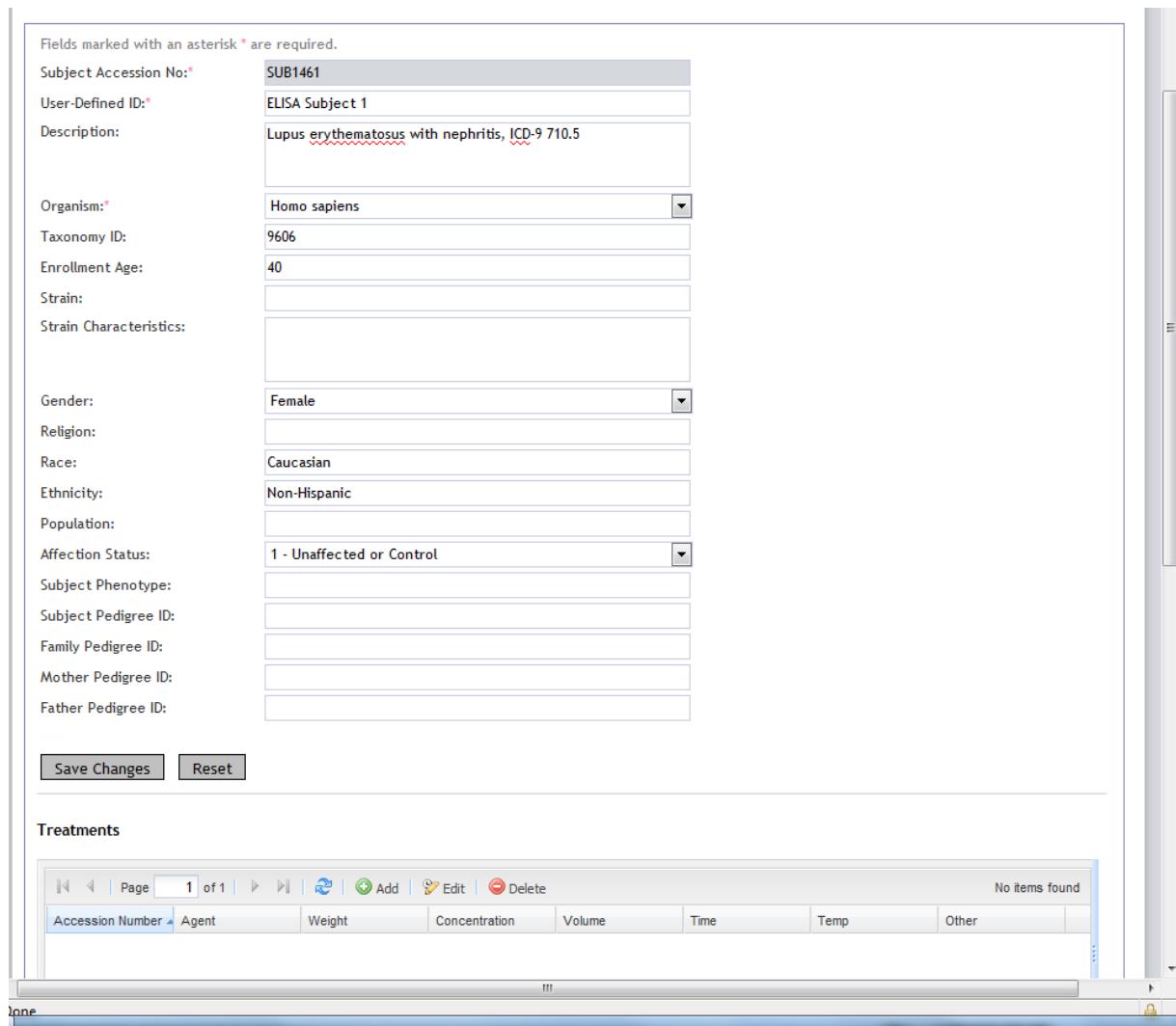
Subject Accession No:*	SUB1461
User-Defined ID:*	ELISA Subject 1
Description:	Lupus erythematosus with nephritis, ICD-9 710.5
Organism:*	Homo sapiens
Taxonomy ID:	9606
Enrollment Age:	40
Strain:	
Strain Characteristics:	
Gender:	Female
Religion:	
Race:	Caucasian
Ethnicity:	Non-Hispanic
Population:	
Affection Status:	1 - Unaffected or Control
Subject Phenotype:	
Subject Pedigree ID:	
Family Pedigree ID:	
Mother Pedigree ID:	
Father Pedigree ID:	

Treatments

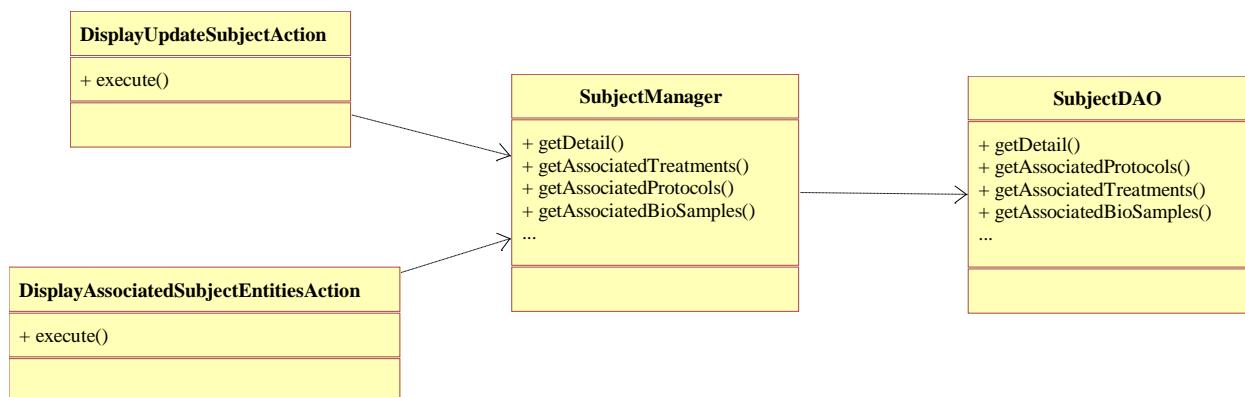
Accession Number	Agent	Weight	Concentration	Volume	Time	Temp	Other
No items found							

None

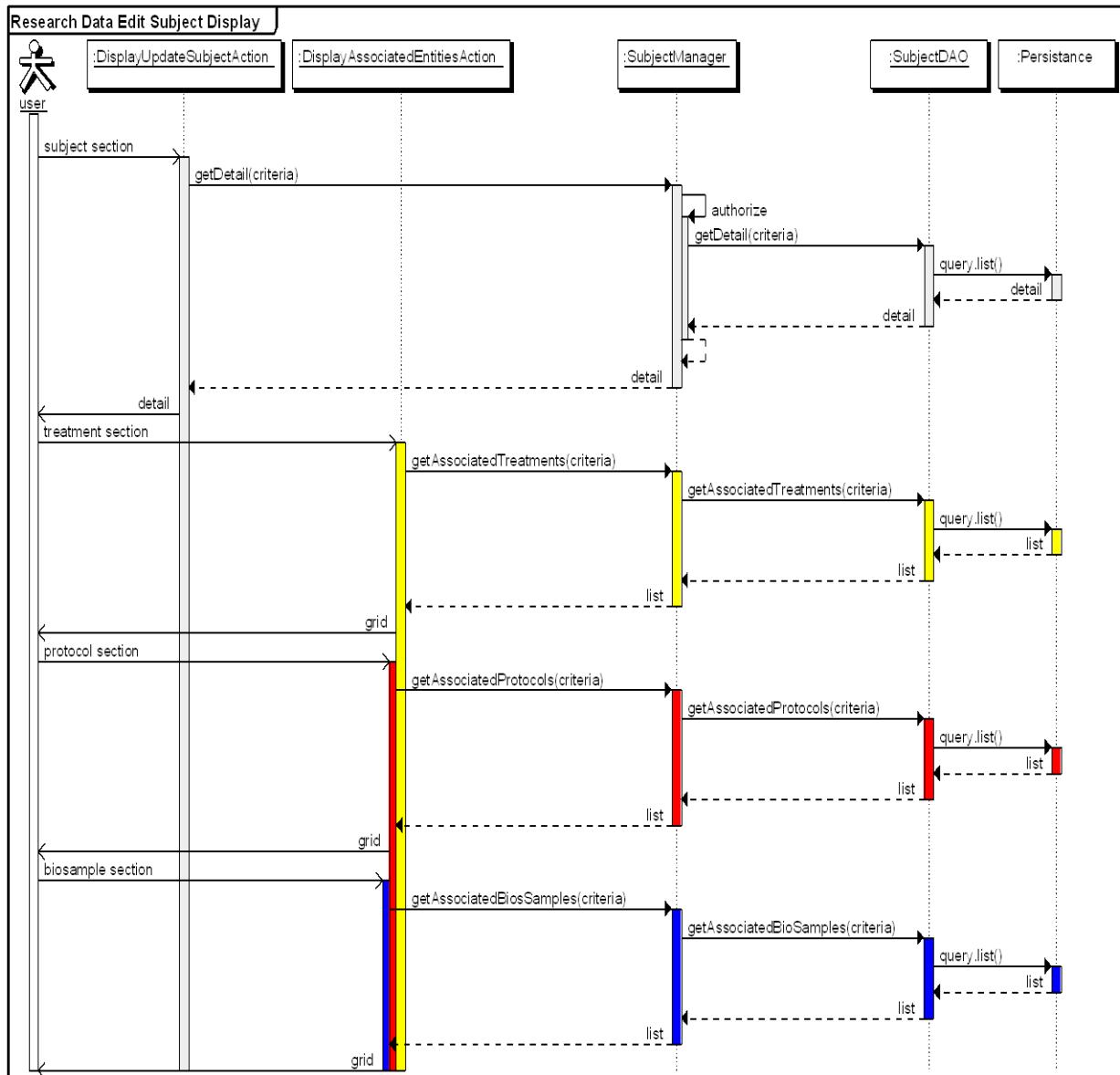
Save Changes **Reset**



14.1.1 Display Edit Screen Class Diagram



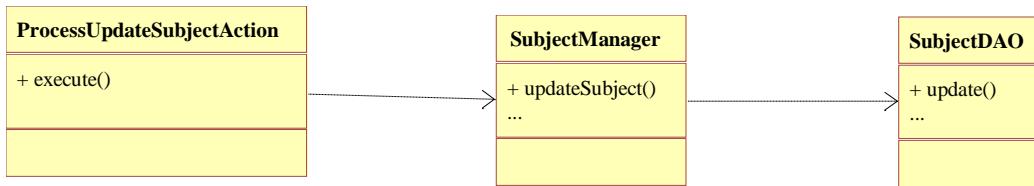
14.1.2 Display Edit Screen Sequence Diagram



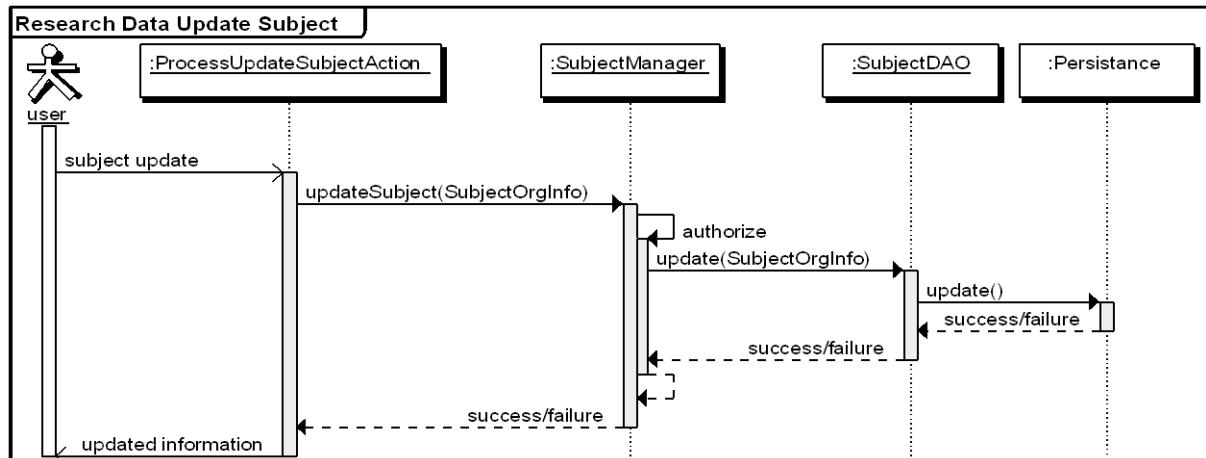
14.2 UPDATE SUBJECT

After the save changes button is clicked, the transaction is sent to the database to update the appropriate fields.

14.2.1 Update Subject Class Diagram



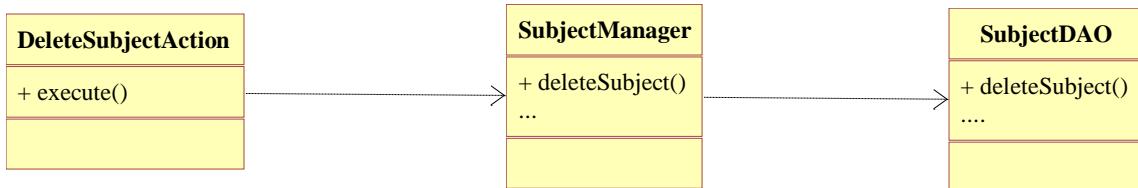
14.2.2 Update Subject Sequence Diagram



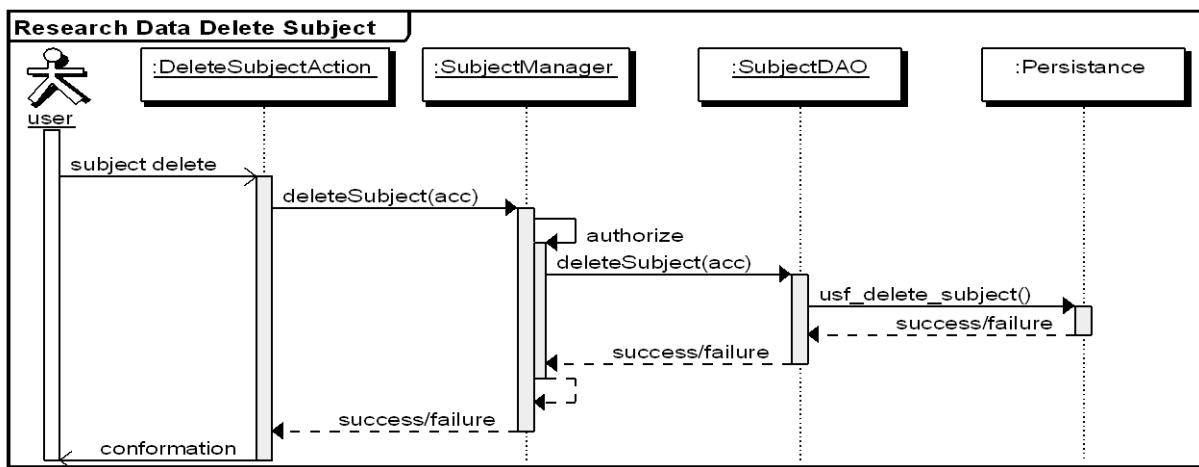
14.3 DELETE SUBJECT

Additionally, if the user has appropriate permissions he may delete the subject from the detail page.

14.3.1 Delete Subject Class Diagram



14.3.2 Delete Subject Sequence Diagram



14.4 ADD TREATMENT

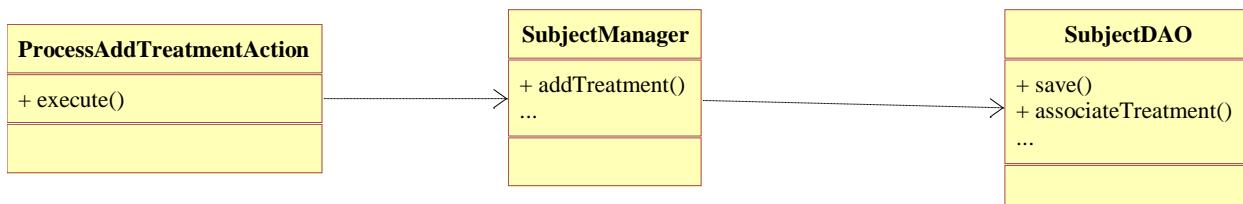
The next section shows any treatments currently linked to this subject. In ImmPort treatments are not considered primary research entities, but are considered as children of subjects. In this section you can add a new treatment, delete a treatment or edit an existing treatment.

A screenshot of a Mozilla Firefox browser window showing the ImmPort: Immunology Database and Analysis Portal. The title bar says 'ImmPort: Immunology Database and Analysis Portal - Mozilla Firefox'. The address bar shows the URL 'immpport.net https://www.immpport.net/immpportWeb/workspace/subject/displayAddTreatment.do?entityPath=subject&...'. The main content is a form titled 'Add Treatment' with the following fields:

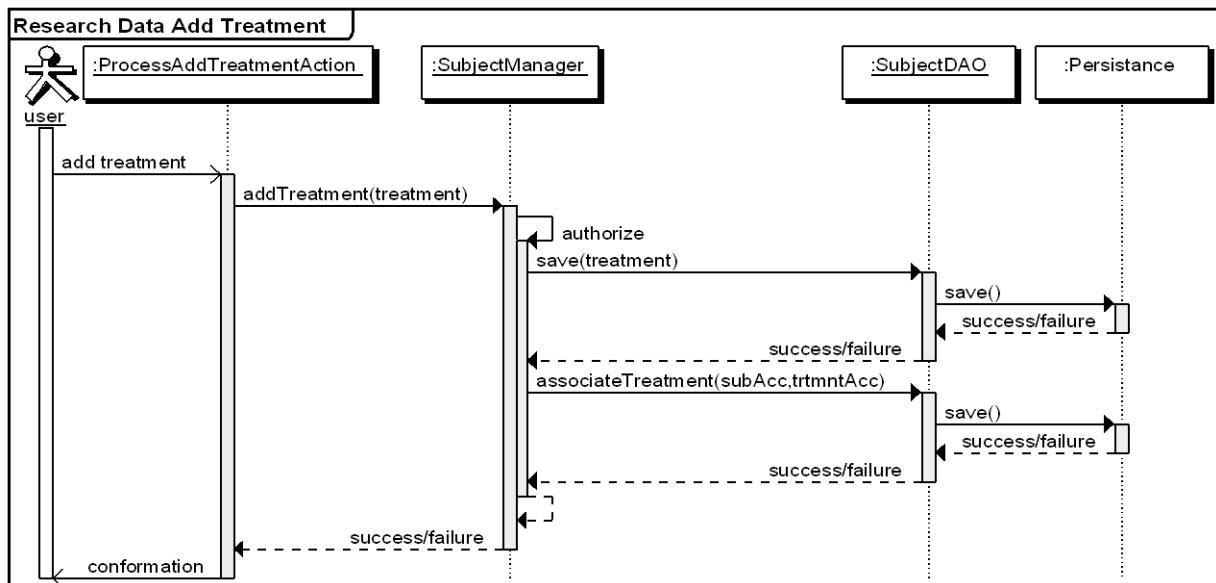
Agent:	[Text Input]
Weight Value:	[Text Input]
Weight Unit:	[Text Input]
Concentration Value:	[Text Input]
Concentration Unit:	[Text Input]
Volume Value:	[Text Input]
Volume Unit:	[Text Input]
Time Value:	[Text Input]
Time Unit:	[Text Input]
Temperature Value:	[Text Input]
Temperature Unit:	[Text Input]
Other:	[Text Input]

At the bottom are two buttons: 'Create' and 'Cancel'.

14.4.1 Add Treatment Class Diagram



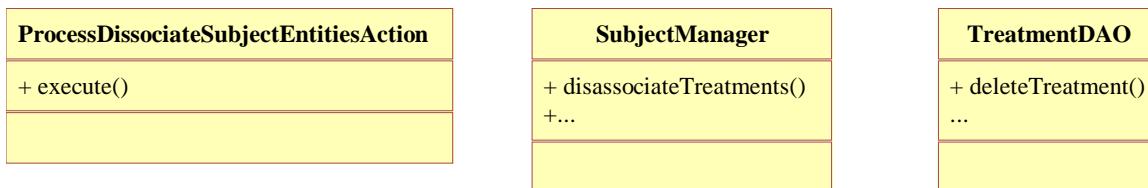
14.4.2 Add Treatment Sequence Diagram



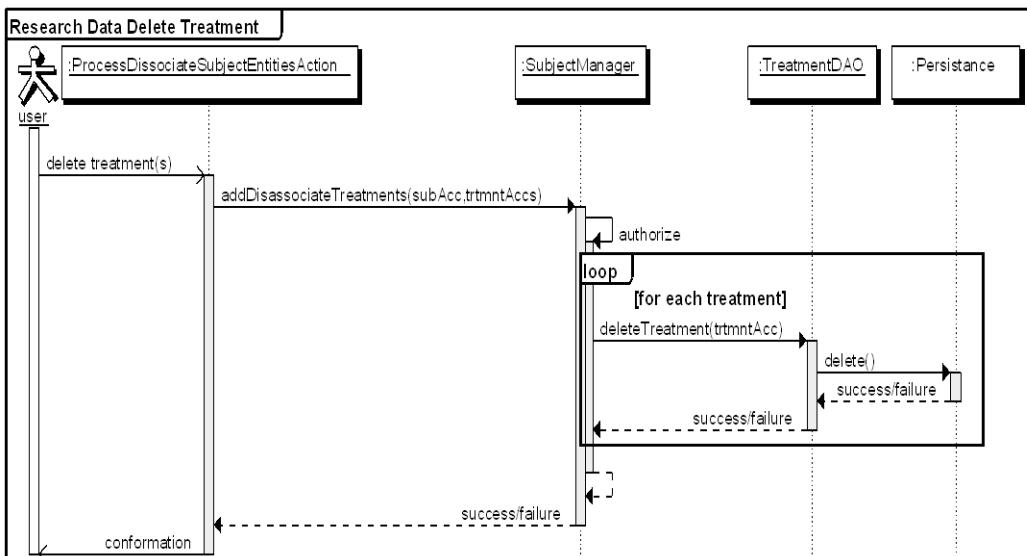
14.5 DELETE TREATMENT

From the summary grid for treatments in the main subject update page, the delete treatment button allows the user to remove a treatment from the database.

14.5.1 Delete Treatment Class Diagram



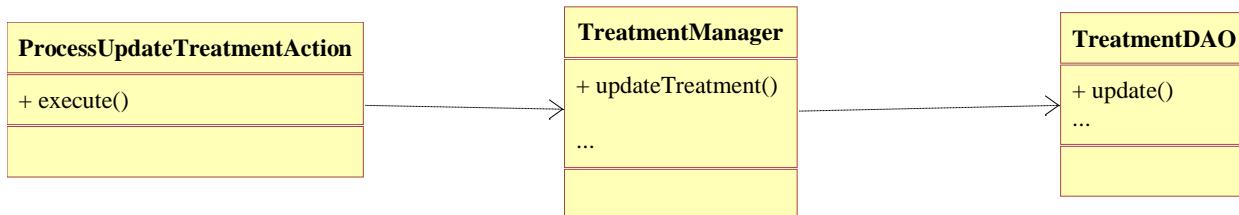
14.5.2 Delete Treatment Sequence Diagram



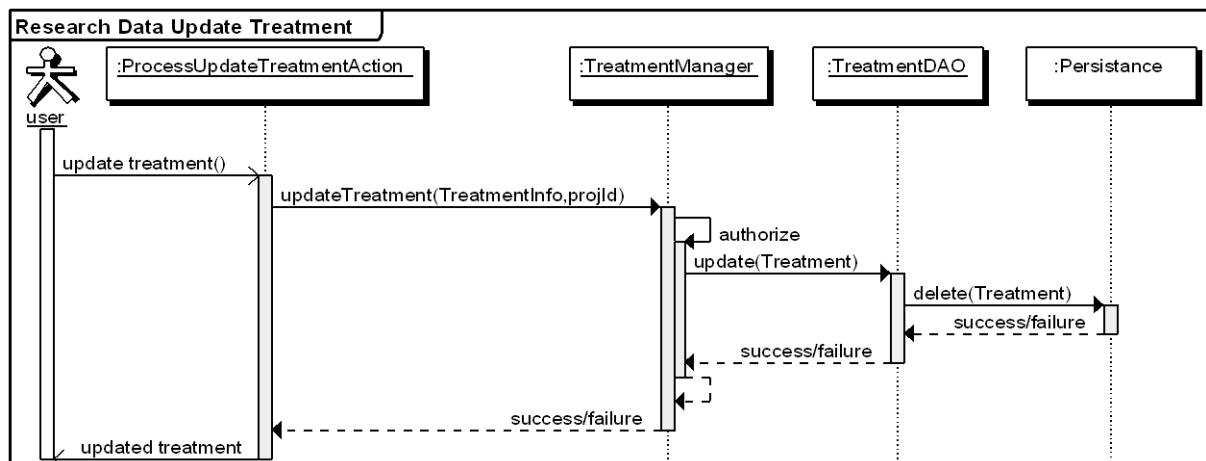
14.6 UPDATE TREATMENT

Treatments may also be updated in the same interface as was shown earlier for entering a new treatment with the prior values now populated.

14.6.1 Update Treatment Class Diagram



14.6.2 Update Treatment Sequence Diagram



14.7 DISPLAY NON-ASSOCIATED ENTITIES

A subject record may be associated with many protocols and many biological samples.

The screenshot shows a software interface with two main sections: "Protocols used" and "Biological Samples Obtained from this Subject/Organism".

Protocols used:

Accession Number	User-Defined ID	Name
PTL29	ELISA Protocol	ELISA Protocol Information

Biological Samples Obtained from this Subject/Organism:

Selected items: BS1602

Accession Number	Name
BS1602	Blood from Subject 1
BS1603	Lymphocytes from Subject 2

Finished

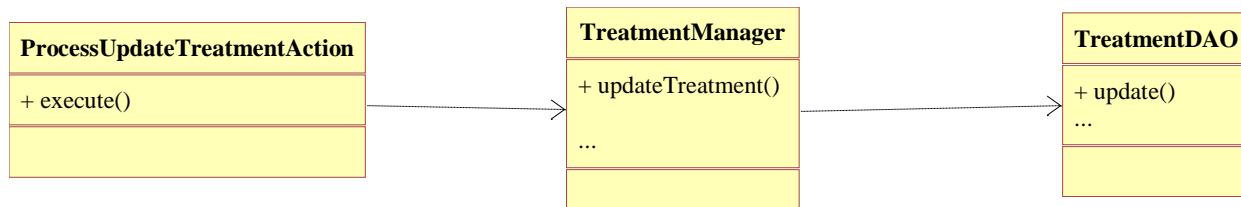
To add additional associations, the user clicks the Add button in the appropriate grid and a window is displayed with a grid containing the available protocols or biological samples that may be associated to the subject. The sequence diagram below will only show how the non-associated biological samples are retrieved and displayed.

Biosamples

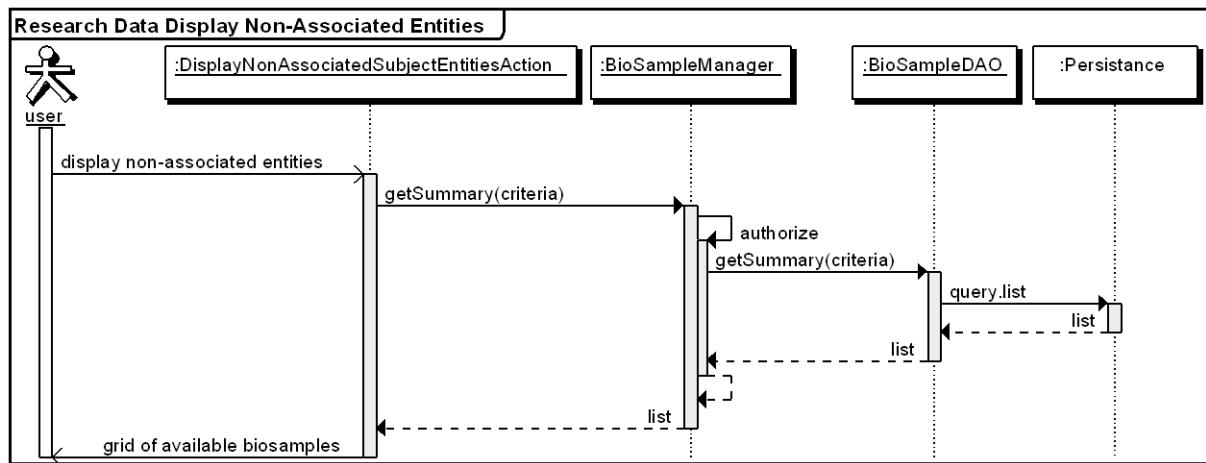
Select one or more Biosamples. Then, click on the Associate Button. Close the dialog box when done.

Accession Number	Study Date	Subject ID	Species	Type	Sub Type	Study Title	Arm
BS1604			Homo sapien	Cell	B cells		null
BS1605			Homo sapien	Tissue	Blood		null
BS1606			Homo sapien	Cell	Lymphocytes		null
BS1607			Homo sapien	Cell	B cells		null
BS1608			Homo sapien	Tissue	Blood		null
BS1609			Homo sapien	Cell	Lymphocytes		null
BS1610			Homo sapien	Cell	B cells		null
BS1611			Homo sapien	Tissue	Blood		null
BS1612			Homo sapien	Cell	Lymphocytes		null
BS1613			Homo sapien	Cell	B cells		null
BS1614			Mus musculus	Cell	B cells		null
BS1615			Mus musculus	Cell	B cells		null
BS1616			Mus musculus	Cell	B cells		null

14.7.1 Display Non-Associated Entities Class Diagram



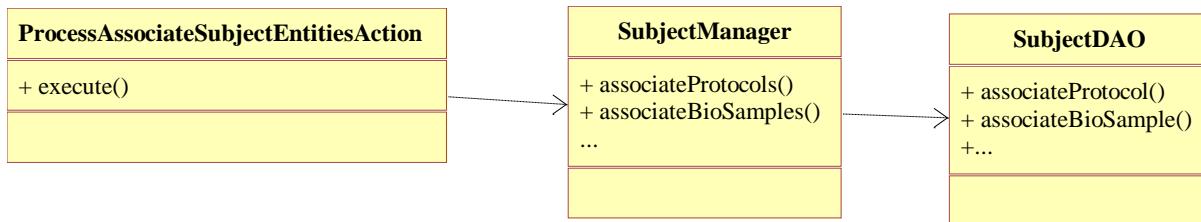
14.7.2 Display Non-Associated Entities Biological Sample Sequence Diagram



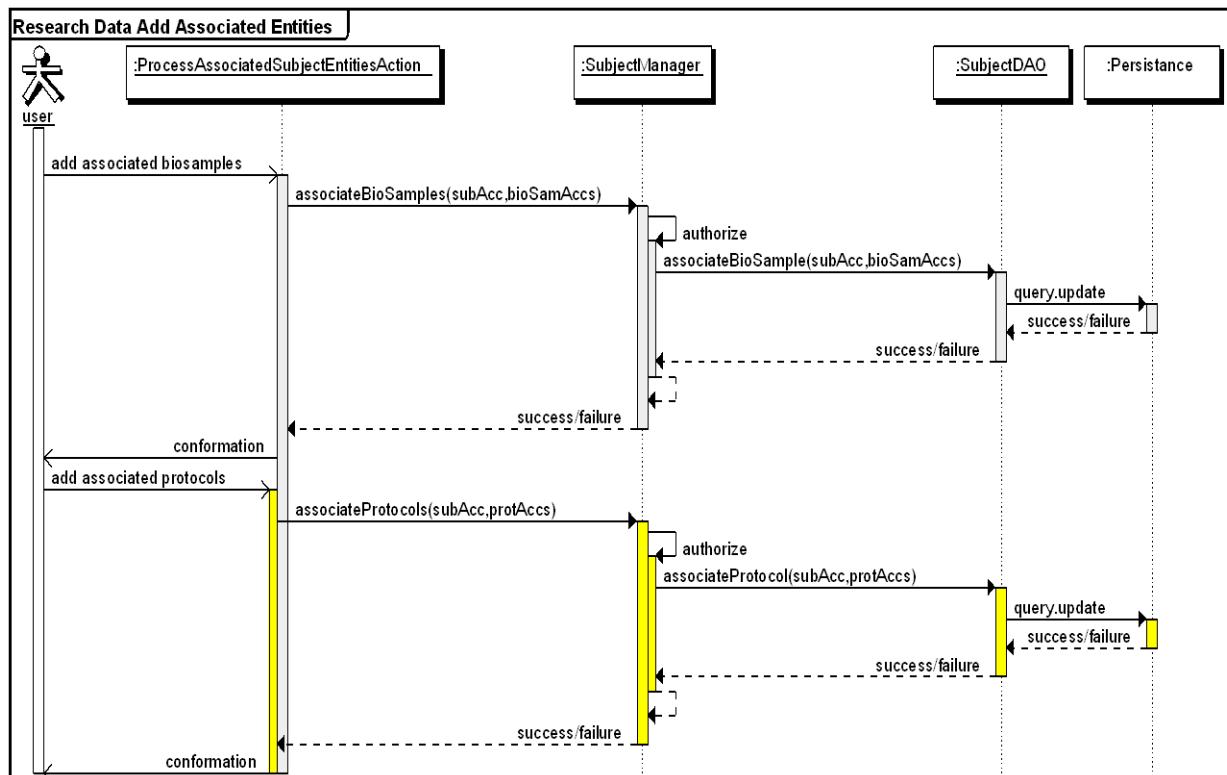
14.8 ASSOCIATE ENTITIES

After the user has chosen biological sample or protocols to link to the subject they submit the request using the following process.

14.8.1 Associate Entities Class Diagram



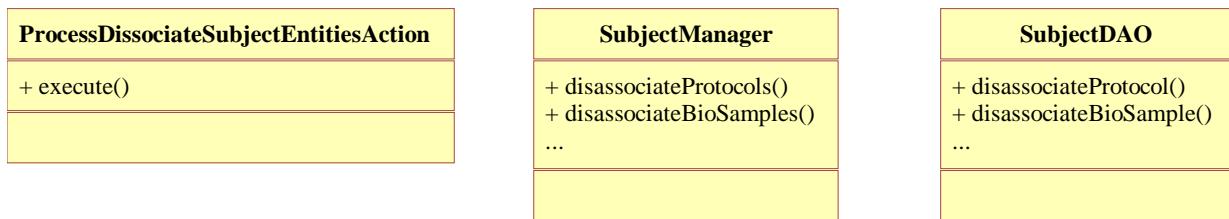
14.8.2 Associate Entities Sequence Diagram



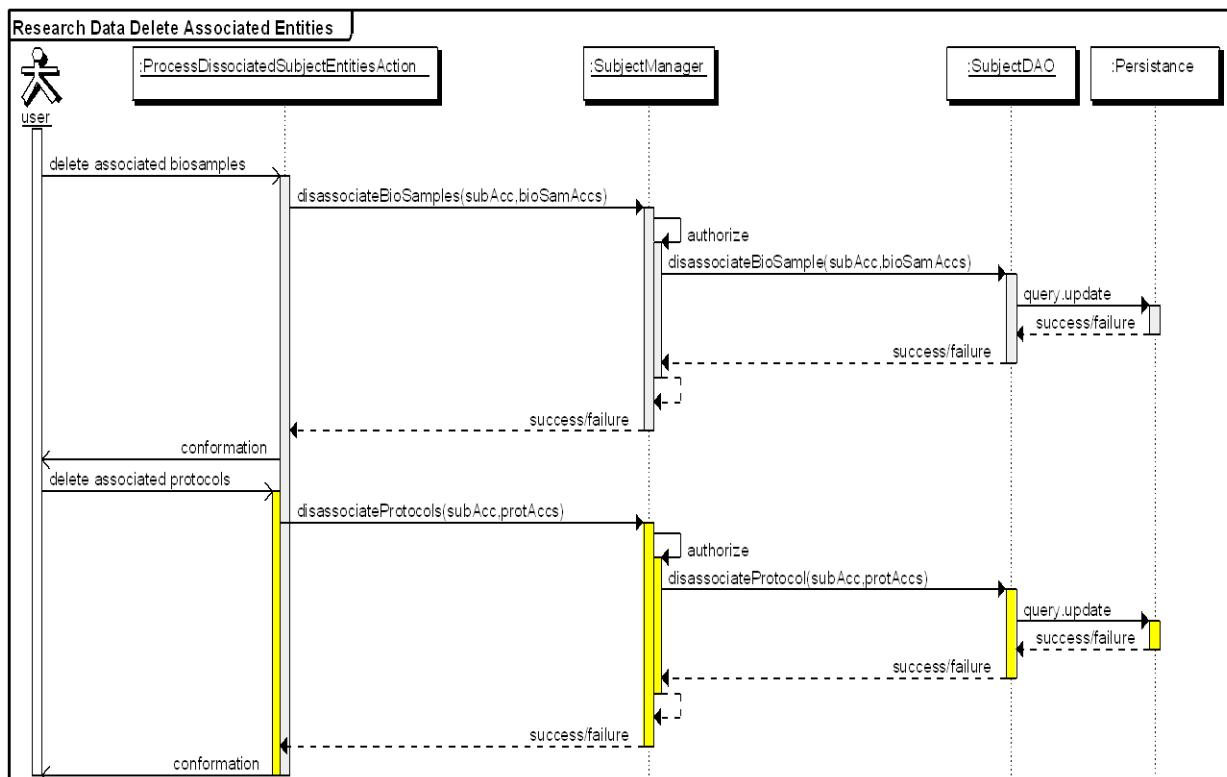
14.9 DISASSOCIATE ENTITIES

On the edit summary page the user may choose biological sample or protocols displayed in the associated entities grid to mark them for unlinking from the subject. The process is outlined below.

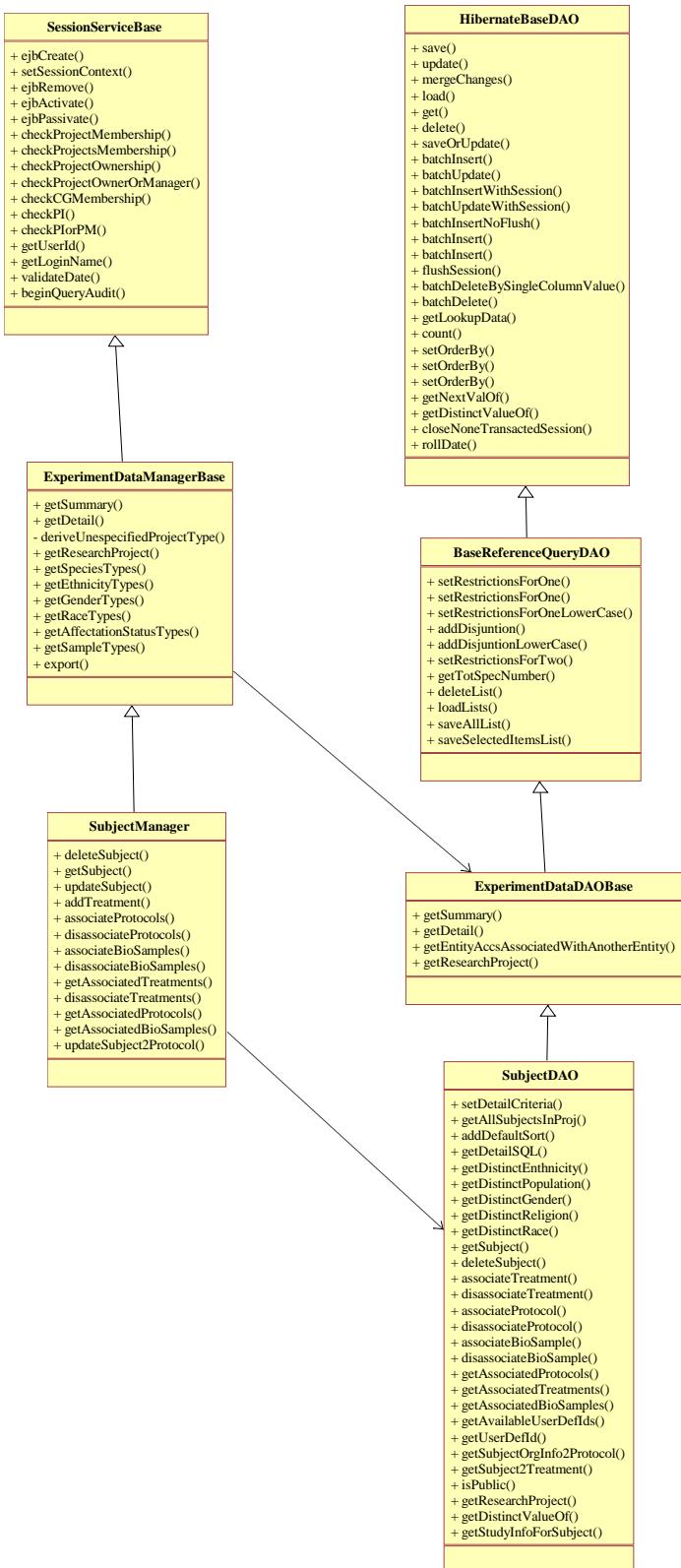
14.9.1 Disassociate Entities Class Diagram



14.9.2 Disassociate Entities Sequence Diagram



14.10 EDIT SUBJECT CLASS DIAGRAM



15.0 LIST DESIGN PACKAGES

The ability to create and manage lists of entities, and then use them in other modules is a key feature in the ImmPort system. Before going into detail on the technology, here is a typical scenario on how lists may be used. The user starts by using the advanced search module to find all subjects that were in study 1 and arm1 and are linked to flow cytometry experimental samples. The user now has the option of saving selected rows from the result set or saving all the results into a list. The user is prompted for a list name and description and the results are stored for later use. Now the user enters the flow cytometry analysis module and uses the previously saved list to query for flow cytometry files to create a set of files to be analyzed. The user may also return to the advanced query screen and use the list to filter the results to only those subjects in the list and then add additional filters, like gender equal male. These results may then be saved as another list.

The user can do the following with lists:

1. Save individual rows from a result set to a list.
2. Save the entire result set to a list
3. Use the list to restrict new advanced search queries
4. Serve as input to analysis modules or data set generation steps
5. Union lists to create new lists
6. Intersect lists to create new lists
7. Delete a list

15.1 SAVING LISTS

When a user has completed an advanced search they can save selected rows from the result set or the entire result set to a list. A list can contain only one entity type like subject, gene, experimental sample, etc. When they choose to save the list they are prompted to enter a name and a description, to help them identify the list for later user.

A maximum of 10,000 records will be saved to the list.

Select a project to save the list *

RP : ITN019AD - Clinical Trial

List Name*

My Favorite Subjects

Description

Subjects from this arm

Note: The project determines who else can see your list.

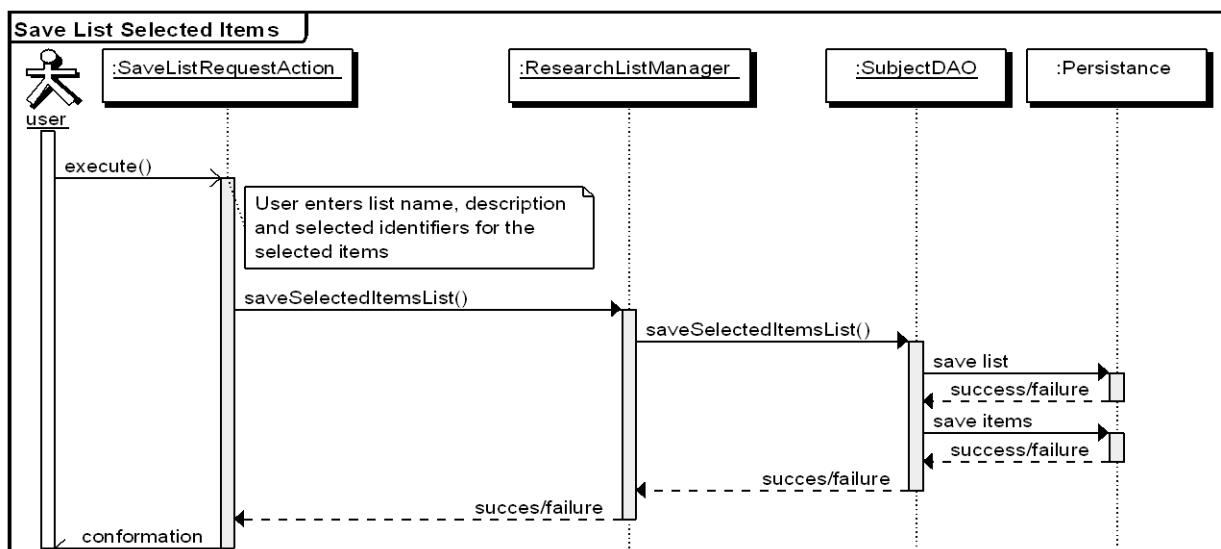
Save Cancel

This screenshot shows a 'Save List' dialog box. At the top, it says 'A maximum of 10,000 records will be saved to the list.' Below that, there's a label 'Select a project to save the list *' followed by a dropdown menu showing 'RP : ITN019AD - Clinical Trial'. Next is a 'List Name*' field containing 'My Favorite Subjects' and a 'Description' field containing 'Subjects from this arm'. At the bottom, a note says 'Note: The project determines who else can see your list.' and there are 'Save' and 'Cancel' buttons.

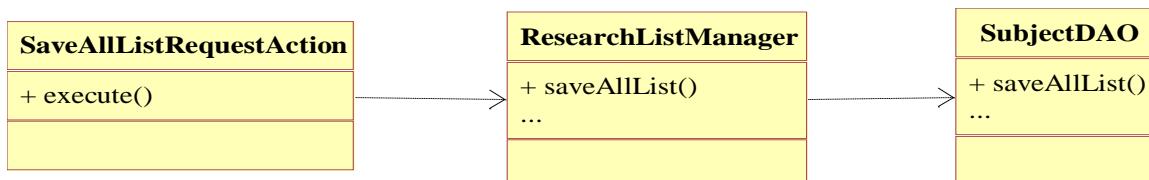
15.1.1 Class Diagram – Save List Selected Items



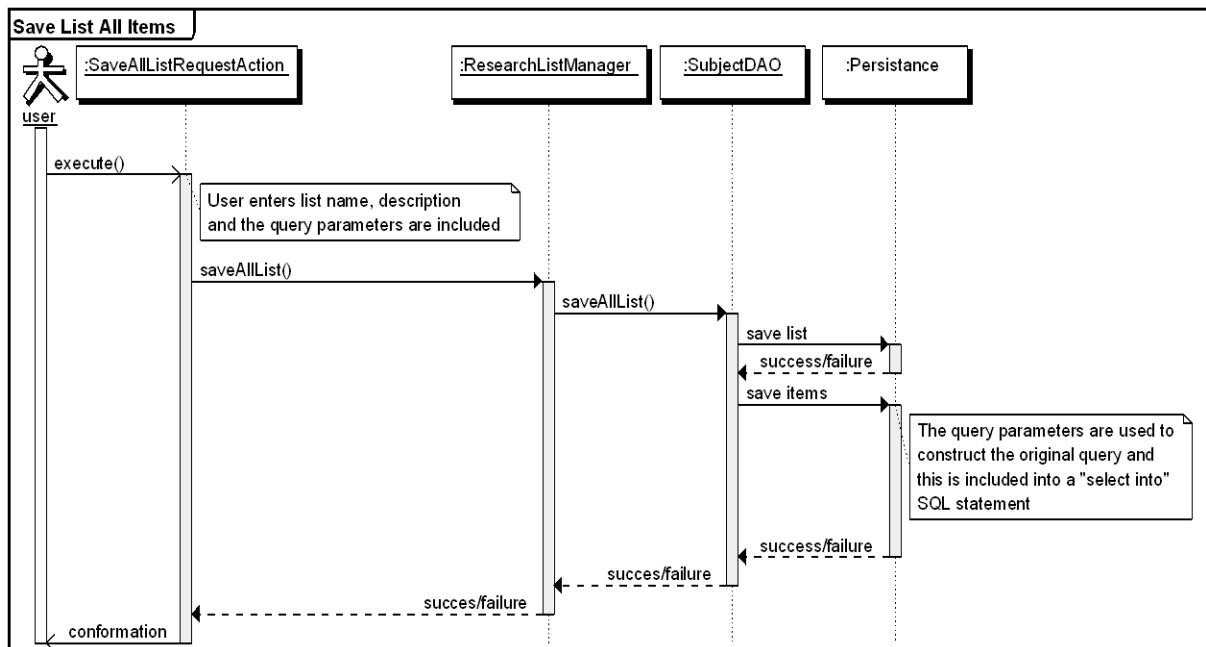
15.1.2 Sequence Diagram List Save Selected Items



15.1.3 Class Diagram List Save All Items



15.1.4 Sequence Diagram List Save All Items



15.2 LIST MANAGEMENT

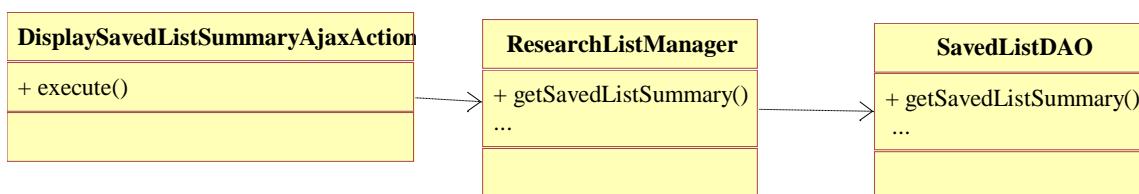
After a list has been saved the user can query for all lists, view the detailed content of a list, delete a list, export a list, union lists together and intersect lists.

A set of an individual's saved lists are available from the "My Saved Lists" menu option.

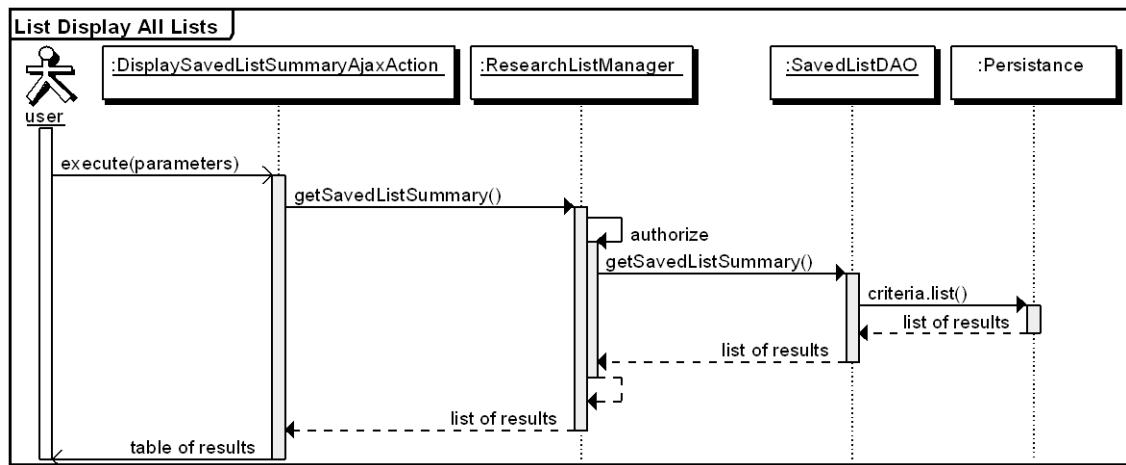
The screenshot shows a web-based application interface titled "My Saved List / Summary". At the top, there are two buttons: "Selected Projects" and "Change Projects". Below the buttons, a message states: "This page provides a summary overview of various types of saved lists within your selected research and collaborative projects. You may view detailed information about specific saved lists by selecting one or more saved lists and clicking the 'View Details' button. Please note that the items displayed in a list are dependent on the projects you selected." A table below the message displays 25 of 303 saved lists. The columns in the table are: List Name, List Description, List Type, Total Items, and Accessible Items. The table data is as follows:

List Name	List Description	List Type	Total Items	Accessible Items
casale fcs results	save all	EXPERIMENT_SAMPLE	10000	10000
2 CEU case subjects	2 case subjects modified from	SUBJECT	2	2
2 list names		SUBJECT	2	2
20090710 AS protein all	total - 2848	PROTEIN	2803	0
20090710 AS protein all - 2	total - 2848	PROTEIN	2803	0
20090710 AS protein items - 3	5th and 6th items	PROTEIN	2	0
20090710 protein union SPW	Test	PROTEIN	10000	0
20090710 save all - 2 items	copy of list 55	PROTEIN	10000	0
20090717 AS Exp RPs save to CP	total - 8	EXPERIMENT	8	6
20090722 SFVT sub PUBLIC	total - 4	SUBJECT	4	4
20090722 SFVT sub SPW	total - 106	SUBJECT	106	106
20090813 AS sub SPW	all - total - 3 thomsls	SUBJECT	3	0
20090813 PO AS subject	all - 2	SUBJECT	2	0

15.2.1 List Management Display Lists Class Diagram



15.2.2 List Management Display Lists Sequence Diagram



15.3 LIST DETAIL

The process to list the detail about the list, plus the entities in the list is a three step process. The initial action class displays tabs across the top of the screen, with one tab for each list selected. Then an AJAX call is submitted to retrieve the details for this list. Next an AJAX call is submitted to retrieve summary information for the individual entities in the list. The AJAX call to retrieve the individual entities in the list uses the same classes as outlined in the search module for returning summary grids.

The screenshot shows a web-based application interface for managing saved lists. At the top, there is a header bar with the title "My Saved List / Detail" and a "Selected Projects" button. Below the header, there are three tabs: "2 list nam", "2 CEU case", and "casale fc", with "casale fc" being the active tab. A summary table provides details about the selected list:

List ID:	691
List Name:	casale fcs results
List Description:	save all
Project Name:	ITN019AD - Clinical Trial
List Type:	EXPERIMENT_SAMPLE
Total Unique Items Saved:	10000
Number of accessible Items:	10000
Date Created:	2010-04-27 12:24:49.0

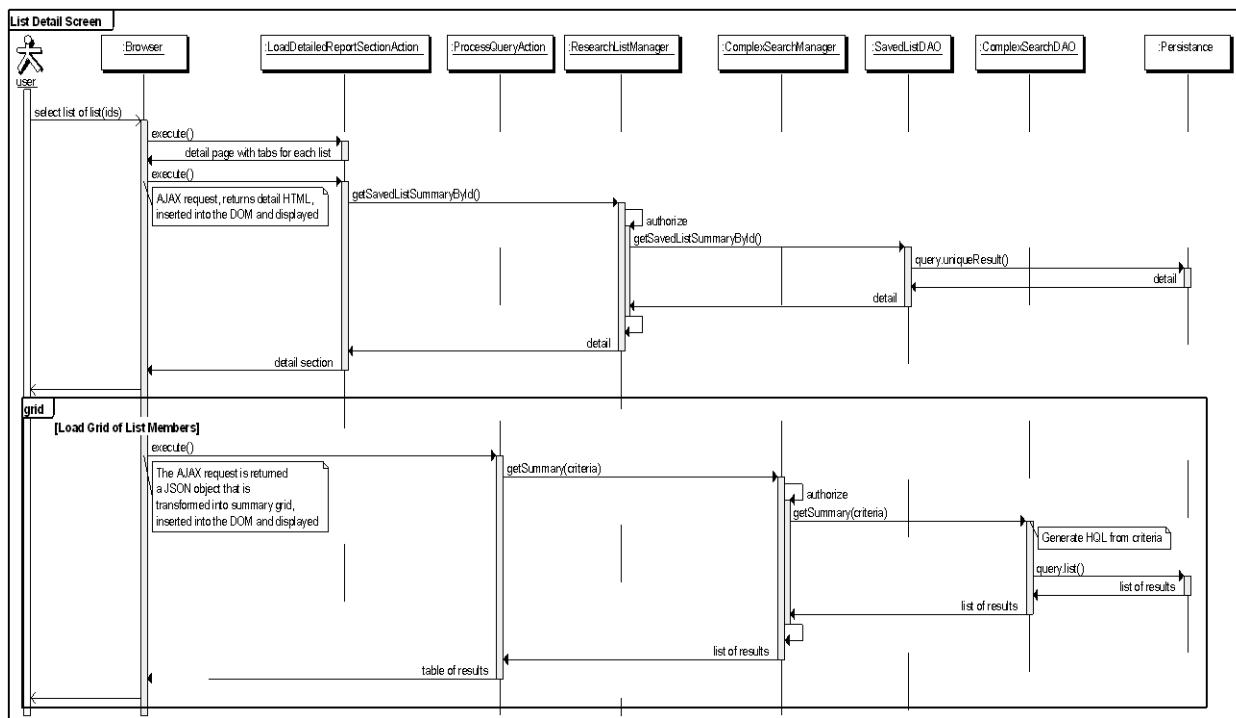
Below the summary table is a section titled "List Items" containing a grid view of data. The grid has columns: Exp Sample Acc, Exp Acc, Exp User-Def ID, Data Format, and Project Title. The data shows 25 rows of entries, all of which are "ITN019AD - Clinical Trial".

Exp Sample Acc	Exp Acc	Exp User-Def ID	Data Format	Project Title
ES82266	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82267	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82268	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82269	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82270	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82271	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82272	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82273	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82274	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82275	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82276	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82277	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial
ES82278	EXP3727	RPCIFLOW_ITN019AD	FCM_report_file	ITN019AD - Clinical Trial

15.3.1 List Detail Class Diagram



15.3.2 List Detail Sequence Diagram



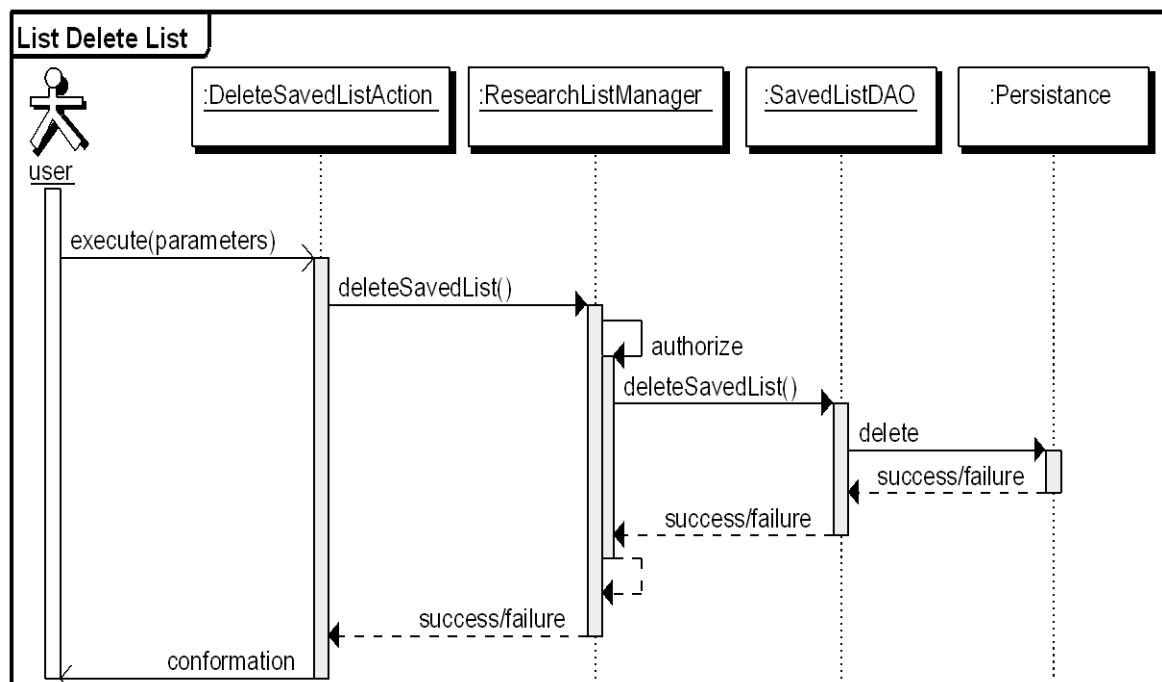
15.4 LIST DELETE

When the user is viewing the detail screen and they are the owner of the list they may choose to delete the list.

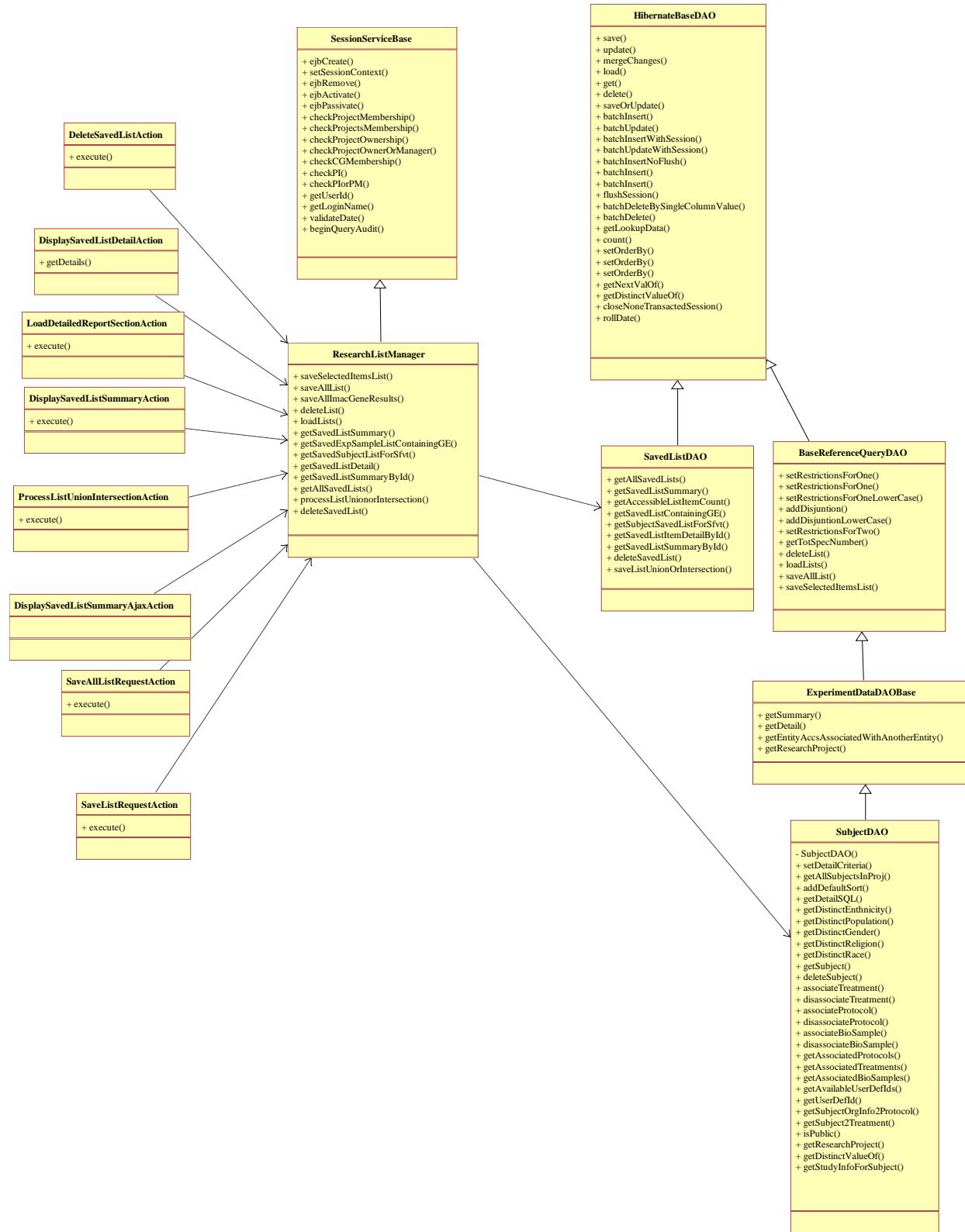
15.4.1 List Delete Class Diagram



15.4.2 List Delete Sequence Diagram



15.5 LIST MANAGEMENT CLASS DIAGRAM



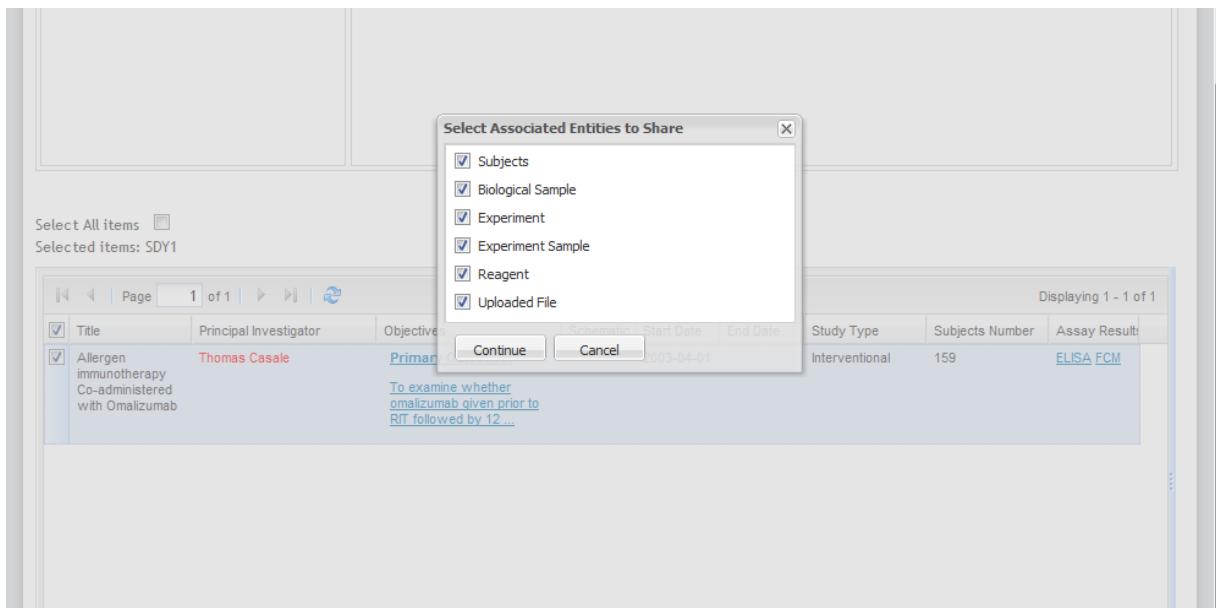
16.0 RESEARCH AND CLINICAL DATA SHARING DESIGN PACKAGES

To facilitate the sharing of information between collaborators users can create a collaborative project. The project is initially empty, but by using the sharing data module the user can link entities from other research projects to the collaborative project. Using this functionality, 2 researchers may both share data from their individual research projects to the collaborative project so both researchers may use the same shared data.

The first step in the process is to pick a source research project and destination collaborative project. The user is then presented with an advanced query screen so they can search for either a study, subject or experiment entity. The results of the query are used as the base criteria for all associated entity queries. The user may select all the results from the query or choose a subset of rows.

The screenshot shows a software application window titled "Advanced Search - Selected Return Data Type : Study". At the top, there is a search bar with the placeholder "Please Choose Data Type to Find : Study", a dropdown for "Results per page" set to 25, a "Display Columns" dropdown, and "Submit" and "Cancel" buttons. Below the search bar, it displays "Source Research Project: ITN019AD - Clinical Trial" and "Destination Collaborative Project: Regression Test CP for v2.5.1". A "Change Project" button is also present. On the left, there is a tree view titled "Advanced Search Attribute Tree" with nodes for Study, Subject, Biological Sample, Experiment, Experiment Sample, Reagent, Protocol, and Uploaded File. At the bottom left, there is a "Select All items" checkbox. The main right pane shows a table titled "Displaying 1 - 1 of 1" with one row of data. The columns are: Title, Principal Investigator, Objectives, Schematic, Start Date, End Date, Study Type, Subjects Number, and Assay Result. The data row is: Allergen immunotherapy Co-administered with Omalizumab, Thomas Casale, Primary Objective: To examine whether omalizumab given prior to RIT followed by 12 ..., 2003-04-01, Interventional, 159, ELISA FCM. There are navigation buttons at the bottom left of the table, and a "Done" button at the bottom left of the entire window.

After this selection the user is presented with a screen to pick what associated entities to the root entity would they also like to share.



For example, the user has chosen Study as the root entity and then they query for all experiments linked to Study1. The results grid displays 1 study and the user selects it. Next they are presented with a screen that allows them to choose additional entities they would like to share, with the restriction that these entities will be associated with that study.

Source Research Project: ITN019AD - Clinical Trial
Destination Collaborative Project: Regression Test CP for v2.5.1

Selected Study to be shared to the Collaborative Project

Title	Principal Investigator	Objectives	Schematic	Start Date	End Date	Study Type	Subjects Number	Assay Results
Allergen immunotherapy Co-administered with Omalizumab	Thomas Casale	Primary Objective: To examine whether omalizumab given prior to RIT followed by 12 ...	View	2003-04-01		Interventional	159	ELISA FCM

Subject to be shared to the Collaborative Project

Select All Subjects 159 items Selected

Sub Org Accession ▲	Species	Race/Strain	Gender	Description	Project Title
SUB73366	Homo sapiens	Asian	Male	Omalizumab + Ragweed IT	ITN019AD - Clinical Trial
SUB73367	Homo sapiens	White	Female	Omalizumab + Placebo IT	ITN019AD - Clinical Trial
SUB73368	Homo sapiens	White	Female	Omalizumab + Placebo IT	ITN019AD - Clinical Trial
SUB73369	Homo sapiens	White	Female	Placebo + Placebo IT	ITN019AD - Clinical Trial
SUB73370	Homo sapiens	White	Male	Placebo + Ragweed IT	ITN019AD - Clinical Trial
SUB73371	Homo sapiens	White	Male	Omalizumab + Ragweed IT	ITN019AD - Clinical Trial

Biological Sample to be shared to the Collaborative Project

Select All Biological Samples 4211 items Selected

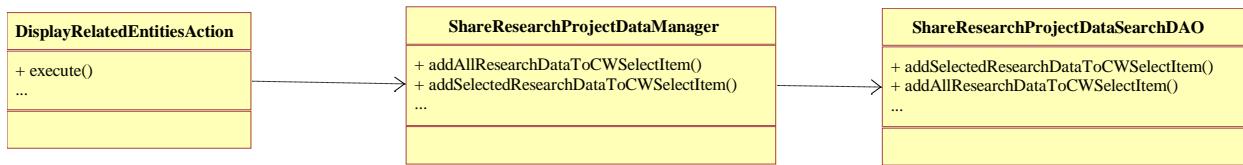
Accession Number	Study Day Collected	Subject Phenotype	Species	Type	Sub Type	Study Title	Arm
BS119963	.7	Ragweed-induced	Homo sapiens	Body Fluid	Whole Blood	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with anti-IgE
PC440004	7	Delayed-induced	Homo sapiens	Body Fluid	Whole Blood	Allergen immunotherapy Co-administered with Omalizumab	Immunotherapy with anti-IgE

The types of associated entities are: Experiment, Subject, Biological Sample, etc. When the user clicks Finish, the system is queried for all the chosen associated entities and the results are displayed in individual result grids. The user now has the option of selecting all, or some of the rows in each grid to share to the collaborative project. When they confirm their selection, rows are added to the database that link the collaborative project to the selected entities.

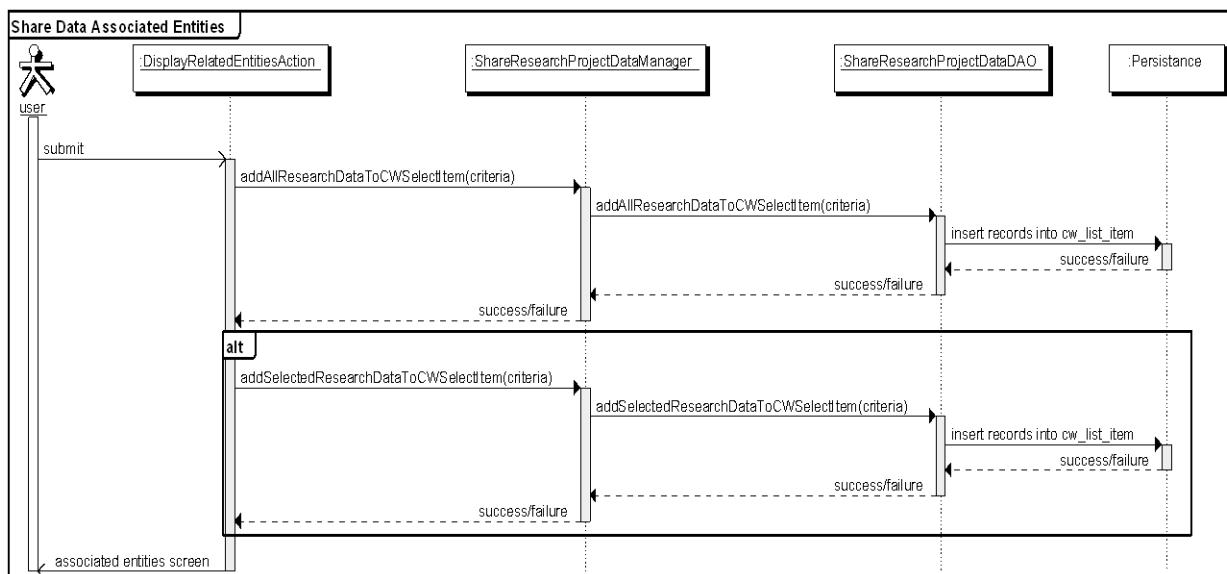
16.1 ASSOCIATE ENTITIES

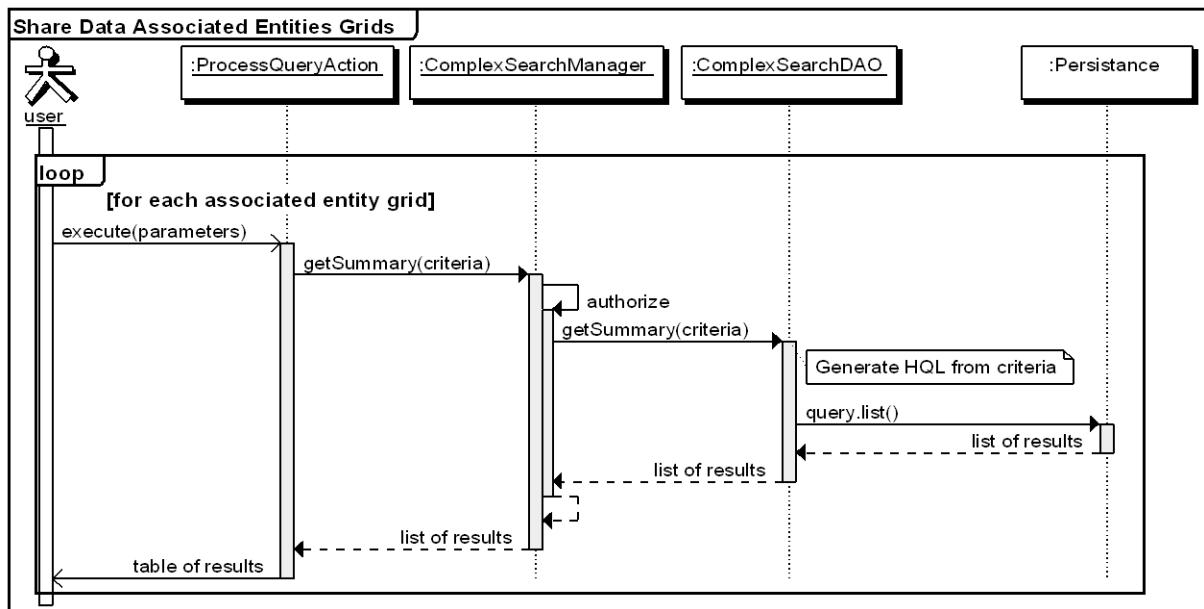
After the associated entities are selected, the execute method adds the selected root entities to the cw_list_item table, using either the addAllResearchDataToSWSelectItem or addSelectedResearchDataToCWSelectItem methods. Then for each associated entity grid on the screen and AJAX call is made to retrieve the associated entities using the advanced search technology.

16.1.1 Display Associated Entities Class Diagram



16.1.2 Display Associated Entities Sequence Diagram



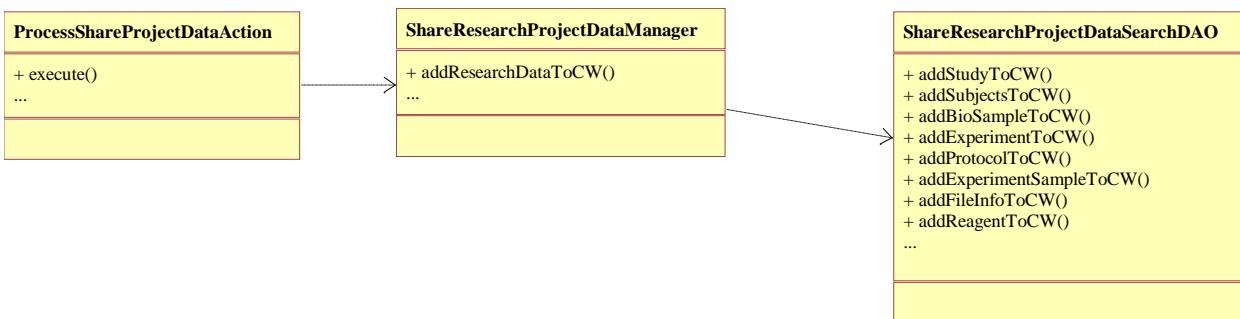


16.2 PROCESS

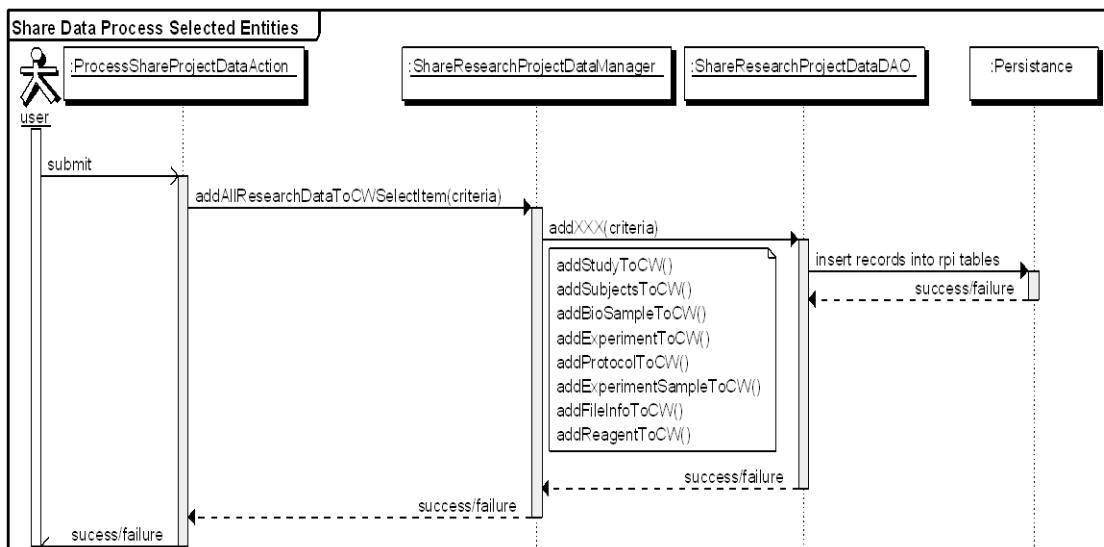
During this step the records previously selected are linked to the collaborative project by inserting record into the RPI tables. There is a RPI table for each major entity like subject, experiment, study, etc. that is used to link entities to a project. The project can be a research or collaborative project.

Summary of shared data.			
All the entities are now shared to Collaborative project "Regression Test CP for v2.5.1".			
Entity Type	Number of rows selected	Actual rows shared	Remarks
Study	1	1	
Subject	159	159	
Biological Sample	4211	4211	
Protocol	5	5	

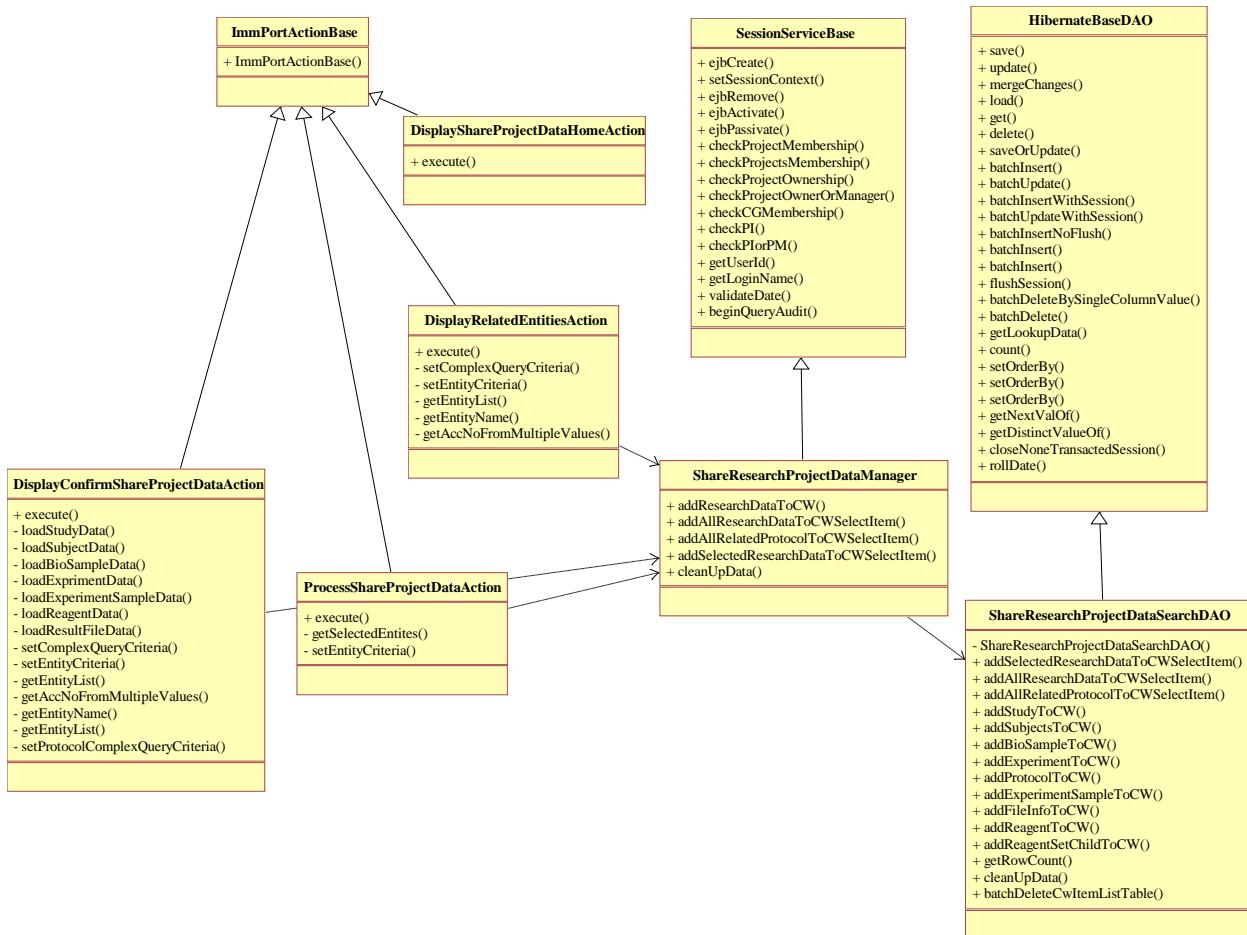
16.2.1 Process Class Diagram



16.2.2 Process Sequence Diagram



16.3 SHARE PROJECT CLASS DIAGRAM



17.0 REFERENCE DATA DESIGN PACKAGES

The ImmPort system supports the search and display of commonly used and biologically relevant data sources including NCBI's Entrez Gene and dbSNP, UniProt's SwissProt, KEGG pathways, protein interaction data sources (e.g. BIND, InnateDB), and human MHC alleles. There are two goals for the reference data searches: offer a convenient starting point to find reference content and integrate content between sources and with ImmPort analysis results. The reference data searches offer two search interfaces: a basic search that uses commonly used search attributes and an advanced search that allows the user to apply search attributes from several data sources to refine their search results. In both interfaces, the system displays a summary of the search results. In the basic search, the summary display is a HTML page and in the advanced search, the summary is an AJAX grid. This section focuses on the basic search interface module.

The user's interaction with the system using simple search is outlined in the steps below:

1. Present the user with a query screen where they can enter a few query terms, or select query terms from drop down choice lists.
2. Submit the query, build a search criteria object and run a Hibernate criteria query to return a set of results.
3. The user can then choose one or more entries from the results screen, to view details.
4. The detail screen displays one or more tabs across the screen to match the number of entries chosen in step 3.
5. For the first tab, the system is queried using an AJAX call to retrieve the detailed information for the first entry. And a series of collapsed panels is displayed representing additional information that is linked to the primary entity. As each panel is expanded an AJAX call is used to query for the additional information.

The display of the detailed information works the same, regardless of the search interface module that was used to execute the initial query.

17.1 DATA SOURCES FOR REFERENCE DATA

ImmPort aggregates data from several sources into a unified visualization format that seeks to bring together relevant information for a queried biological entity.

Table 1.2.1: Research Data Detail Pages

Page Name	Panel	Panel Information	Primary DB Table(s)
Gene Details	Gene Summary	HTML	ENTREZ.ENTREZ_GENE_INFO
Gene Details	Protein	HTML	UNIPROT.SWISSPROT
Gene Details	Gene Ontology	HTML	ONTOLOGY.GO
Gene Details	Protein Interactions	HTML	INNATEDB.INTERACTION_LIST
Gene Details	Pathways	HTML	KEGG.GENE_PATHWAY
Gene Details	Phenotypes and Diseases	HTML	ENTREZ.ENTREZ_GENE2MIM

Page Name	Panel	Panel information	Primary DB Table(s)
Gene Details	Polymorphism	HTML	DBSNP.DBSNP_FLAT
Gene Details	Homologs	HTML	HOMOLOGENE.HOMOLOGENE
Gene Details	Related Sequences	HTML	ENTREZ.ENTREZ_GENE_SYNONYMS
Gene Details	References	HTML	ENTREZ.ENTREZ_GENE2PUBMED
Protein Detail	Protein Summary	HTML	UNIPROT.SWISSPROT
Protein Detail	References	HTML	SWISSPROT.SWISSPROT_CITATIONS
Protein Detail	Features	HTML	SWISSPROT.SWISSPROT_FEATURE
Protein Detail	Protein Sequence	HTML	SWISSPROT.SWISSPROT
Protein Detail	Functions	HTML	ONTOLOGY.GO
Protein Detail	Pathways	HTML	KEGG.GENE_PATHWAY
MHC Alleles	Allele Summary	HTML	MHC_SEQ_VAR.ALLELE
MHC Alleles	Allele Frequencies	HTML	MHC_SEQ_VAR.MHC_HLA_ALLELE_FREQ
MHC Alleles	PubMed Publications	HTML	MHC_SEQ_VAR.PUBMED_REFERENCE
MHC Alleles	Sequences	HTML	MHC_SEQ_VAR.ALIGNMENT_SEQ
MHC Sequence Feature Structure Viewer	3-D Image	JMol	
MHC Sequence Feature Structure Viewer	Information	HTML	MHC_SEQ_VAR.PDB_MAPPING
MHC Sequence Feature Structure Viewer	Allele to PDB Mapping	HTML	MHC_SEQ_VAR.PDB_MAPPING
MHC Sequence Feature Structure Viewer	Sequence Features	HTML	MHC_SEQ_VAR.FEATURE
Pathways	Pathway Summary	HTML	KEGG.GENE_PATHWAY
Pathways	Genes In Pathway	HTML	KEGG.GENE_ID_4M_NCBI
Protein Networks	Summary	HTML	BIND.BIND_COMPLEX2SUBUNITS
Protein Networks	Interactions	HTML	BIND.BIND_INTS
SNPs	SNPs Summary	HTML	DBSNP.DBSNP_FLAT
SNPs	Genotype Information (HapMap)	HTML	DBSNP.DBSNP_FLAT
ImmPort Gene Lists	Cytokines	HTML	Flat files
ImmPort Gene Lists	Cytokine	HTML	Flat files

Page Name	Panel	Panel information	Primary DB Table(s)
	Receptors		
ImmPort Gene Lists	TCR Signalling Pathway	HTML	Flat files
ImmPort Gene Lists	BCR Signalling Pathway	HTML	Flat files
ImmPort Gene Lists	Natural Killer Cell Cytotoxicity	HTML	Flat files
ImmPort Gene Lists	Antigen Processing and Presentation	HTML	Flat files
ImmPort Gene Lists	Antimicrobials	HTML	Flat files
ImmPort Gene Lists	B Cell Stimualtion	HTML	Flat files
ImmPort Gene Lists	B Cell Development	HTML	Flat files

17.2 GENE

The gene is a basic unit of biological understanding and has many facets including location, genetic unit, region that codes for biological functions, etc. The ImmPort Gene query user interface prompts the user to provide commonly used search attributes.

Reference Data / Gene Query

[Home](#) | [Genes](#) | [Proteins](#) | [MHC Alleles](#) | [MHC Sequence Feature](#) | [Pathways](#) | [Protein Networks](#) | [SNPs](#) | [Imr Structure Viewer](#)

Fields marked with an asterisk * are required.

Search Options

Search Type*	<input type="text" value="Gene Symbol"/>
Search Option*	<input type="text" value="Like"/>
Search Text*	(Comma delimited, 256 max chars)
Search Species*	<input type="text" value="Homo sapiens"/> <input type="text" value="Mus musculus"/> <input type="text" value="Rattus rattus"/> <input type="text" value="Rattus norvegicus"/> <input type="text" value="Gallus gallus"/> <input type="text" value="Drosophila melanogaster"/> <input type="text" value="Macaca mulatta"/>

Results Per Page:

Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose to more details about selected records.

Reference Data / Gene Query Results ⓘ

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists ▾ | Data History ▾ | Reference Advanced Search | Structure Viewer

Search Criteria ⓘ (Modify Search)

Type: Gene Symbol Text: adh
Species: Homo sapiens Option: Like Genome Build: hg18

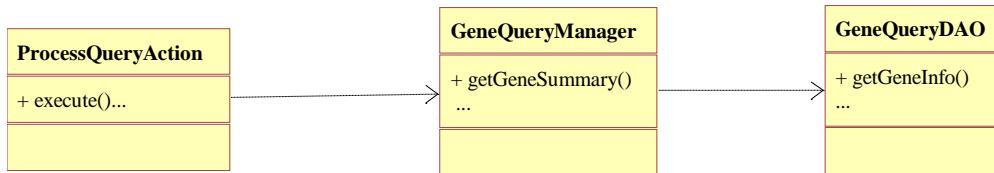
Select up to 10 search results to view in a detailed report. ⓘ indicates that the gene is in the ImmPort Comprehensive List of Immune-Related Genes. ([download](#))

View Detailed Report 0 item(s) selected.

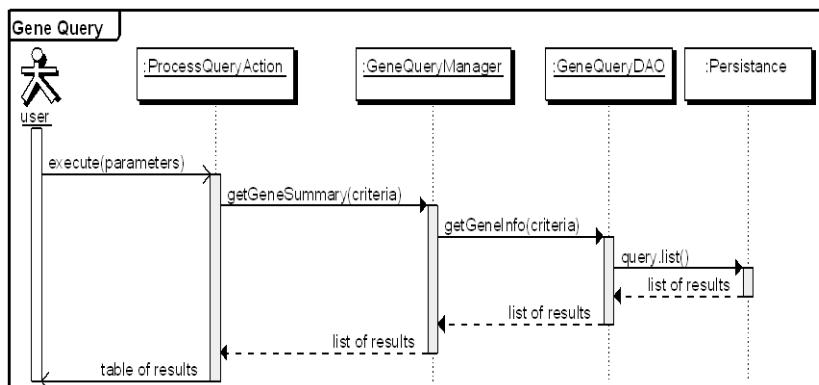
53 items found, displaying 1 to 25. [First/Prev] 1, 2, 3 [[Next/Last](#)]

	Entrez Gene ID	Symbol & Alias	Organism	Chromosome	Name	Proteins	Transcripts
<input type="checkbox"/>	34	ACADM, ACAD1; FLJ18227; FLJ93013; FLJ99884; MCAD; MCADH	Homo sapiens	1	acyl-CoA dehydrogenase, C-4 to C-12 straight chain	NP_001120800 NP_000007	NM_001127328 NM_000016
<input type="checkbox"/>	124	ADH1A, ADH1	Homo sapiens	4	alcohol dehydrogenase 1A (class I), alpha polypeptide	NP_000658	NM_000667
<input type="checkbox"/>	125	ADH1B, ADH2; DKFZp686C06125	Homo sapiens	4	alcohol dehydrogenase 1B (class I), beta polypeptide	NP_000659	NM_000668
<input type="checkbox"/>	126	ADH1C, ADH3	Homo sapiens	4	alcohol dehydrogenase 1C (class I), gamma polypeptide	NP_000660	NM_000669

17.2.1 Gene Query Class Diagram



17.2.2 Gene Query Sequence Diagram



The Gene detail page aggregates data from many sources that are pertinent to the gene record. The first section of the detail page is expanded by default and displays information provided in the NCBI gene summary section. The user may expand one or all blocks of additional information.

Reference Data / Gene Detailed Report

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

[Back to Advanced Search](#)

Adh5

Gene Summary

Entrez Gene ID	100145871	UniGene ID	Rn.222115
Entrez Gene Symbol	Adh5	Entrez Gene Name	alcohol dehydrogenase 5
Also known as	-	Organism (Taxon ID)	Rattus norvegicus (10116)
Chromosome Location	2q44	Genome Build	
Gene Expression	GEO		
Transcripts			

Protein

Gene Ontology [Ontology Browser]

Protein Interactions

Pathways

Phenotypes and Diseases

Polymorphism

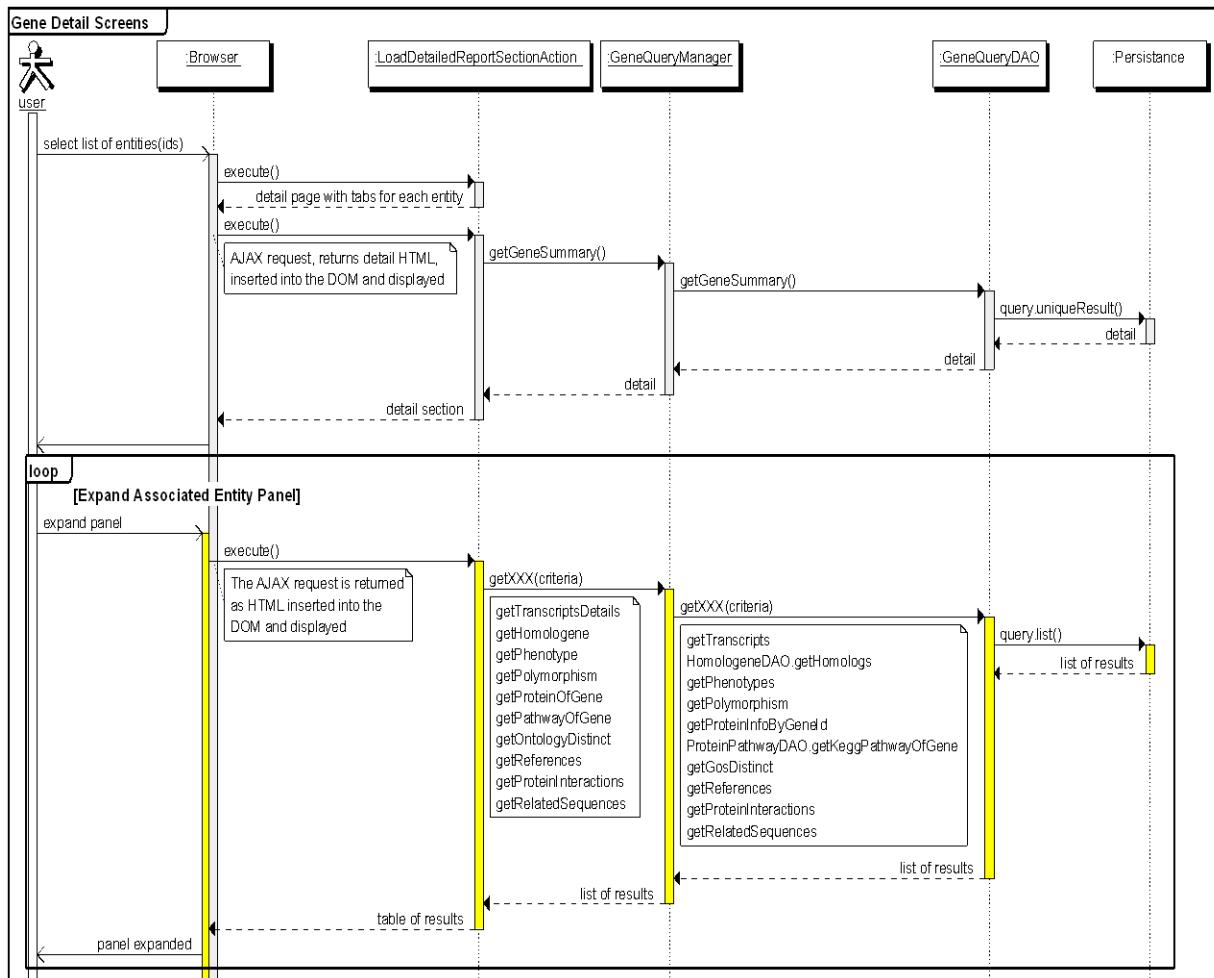
Homologs

Related Sequences

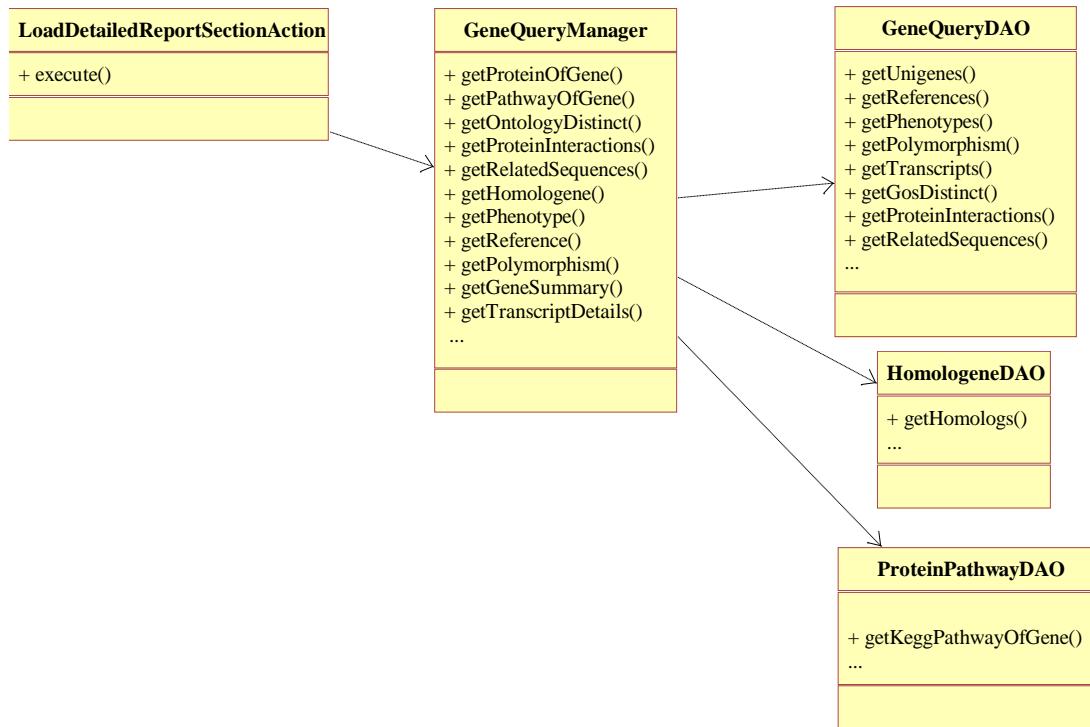
References

Export

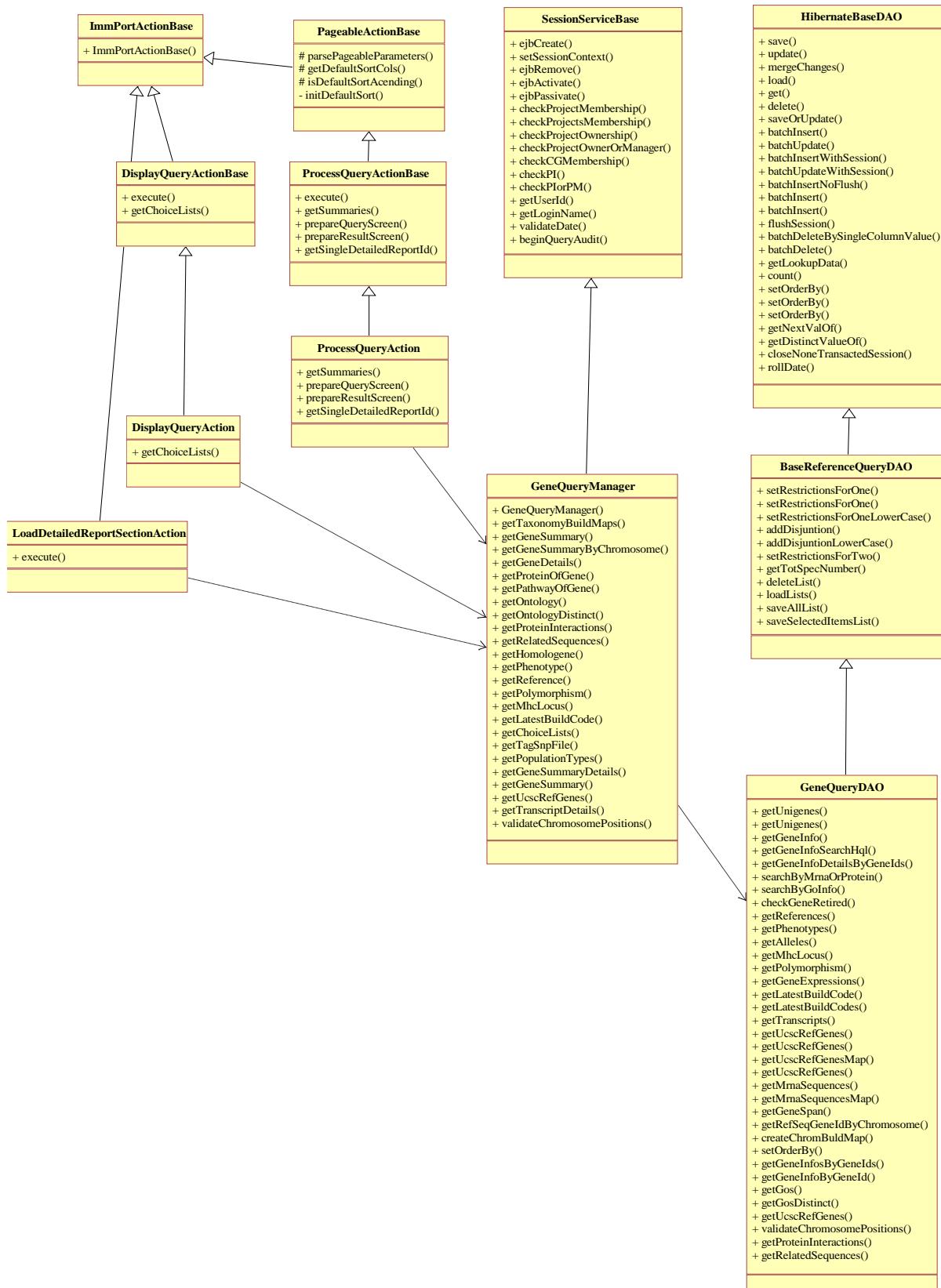
17.2.3 Gene Detail Sequence Diagram



17.2.4 Gene Detail Class Diagram



17.2.5 Gene Class Diagram



17.3 PROTEIN

Proteins are one of the most common functional biological units at the cellular and sub-cellular level and they are coded for by genes. The ImmPort Protein query user interface prompts the user to provide commonly used search attributes.

Reference Data / Protein Query 

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNI Structure Viewer

Fields marked with an asterisk * are required.

Search Options

Search Type*	Gene Symbol
Search Option*	Like
Search Text*	(Comma delimited, 256 max chars)
Search Species*	Homo sapiens Mus musculus Rattus rattus Rattus norvegicus Gallus gallus Drosophila melanogaster Macaca mulatta
<input type="button" value="Submit"/> <input type="button" value="Cancel"/> Results Per Page: 25 <input type="button" value="▼"/>	

Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose to more details about selected records.

Reference Data / Protein Query Results 

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists ▾ | Data History ▾ | Reference Advanced Search | Structure Viewer

Search Criteria [\(Modify Search\)](#)

Type: Gene Symbol Text: adh
Species: Homo sapiens Option: Like

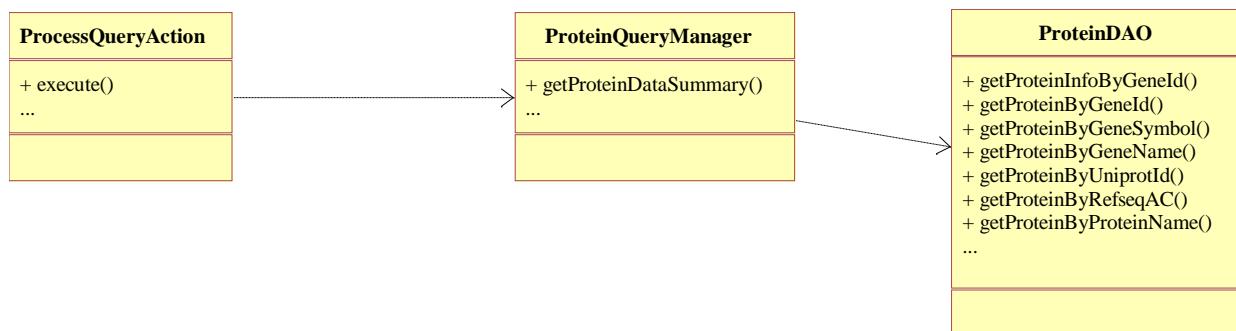
Select up to 10 search results to view in a detailed report.

[View Detailed Report](#) 0 item(s) selected.

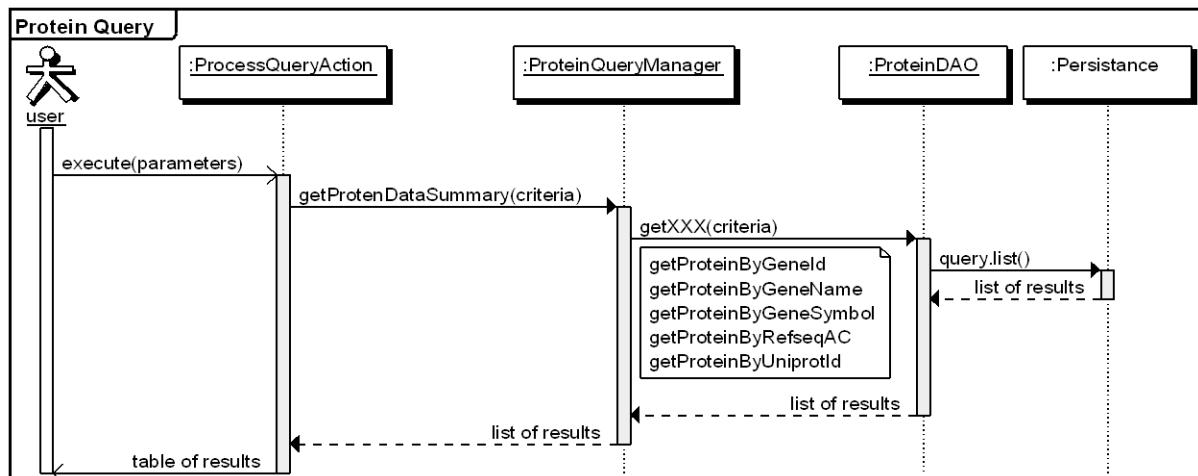
15 items found, displaying all items.

Swiss-Prot/KB Accession	Protein Name	Swiss-Prot/KB Entry Name	Entrez Gene ID	Gene Symbol	Organism
<input type="checkbox"/> P07327	Alcohol dehydrogenase 1A	ADH1A_HUMAN	124	ADH1A	Human
<input type="checkbox"/> P00325	Alcohol dehydrogenase 1B	ADH1B_HUMAN	125	ADH1B	Human
<input type="checkbox"/> P00326	Alcohol dehydrogenase 1C	ADH1G_HUMAN	126	ADH1C	Human
<input type="checkbox"/> P08319	Alcohol dehydrogenase 4	ADH4_HUMAN	127	ADH4	Human
<input type="checkbox"/> P11766	Alcohol dehydrogenase class-3	ADHX_HUMAN	128	ADH5	Human

17.3.1 Protein Query Class Diagram



17.3.2 Protein Query Sequence Diagram



The Protein detail page aggregates data from several sources that are pertinent to the protein record. The first section of the detail page is expanded by default and displays information provided in the UniProt protein summary section. The user may expand one or all blocks of additional information.

Reference Data / Protein Detailed Report

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

[Back To Search Results](#)

ADH4_HUMAN

Protein Summary

Protein Name:	Alcohol dehydrogenase 4		
Swiss-Prot/KB Accession:	P08319		
Swiss-Prot/KB Entry Name:	ADH4_HUMAN		
Organism & Taxon ID:	Human 9606		
Entrez Gene:	Gene ID	Gene Symbol	Gene Synonyms
127	Details Build 36.1, hg18 Details Build 35, hg17	ADH4	ADH-2

RefSeq Peptide

AAH22319 / BAG58459 / AAA51595 / P08319 / - / - / NP_000661 /	Protein gi	18490173 / 194380612 / 178121 / 83286923 / - / - / 71565152 /
CAA39813 / BAF83524 / EAQ06088 / EAQ06089 / - / AAX59034 /		825623 / 158255106 / 119626493 / 119626494 / - / 62003165 /

References

Features

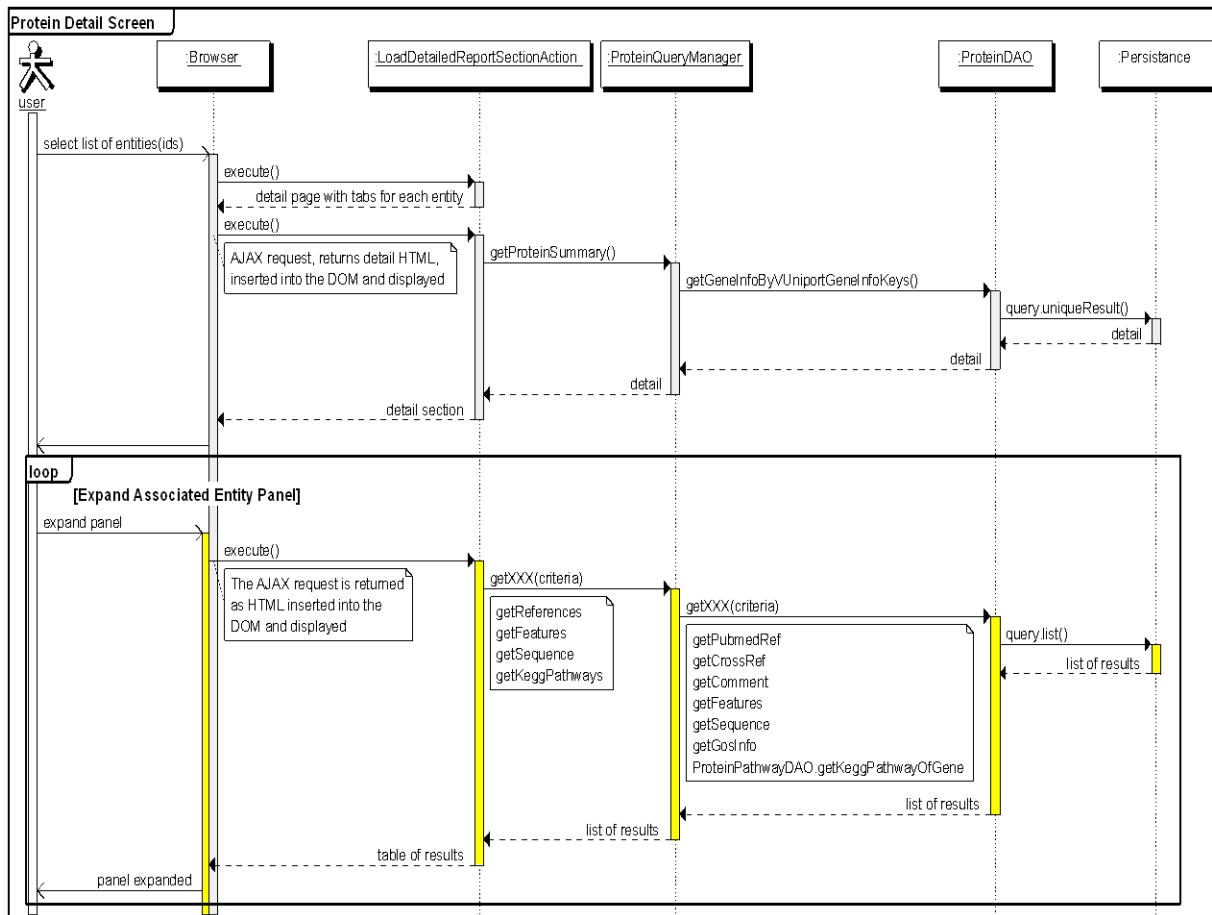
Protein Sequence

Functions

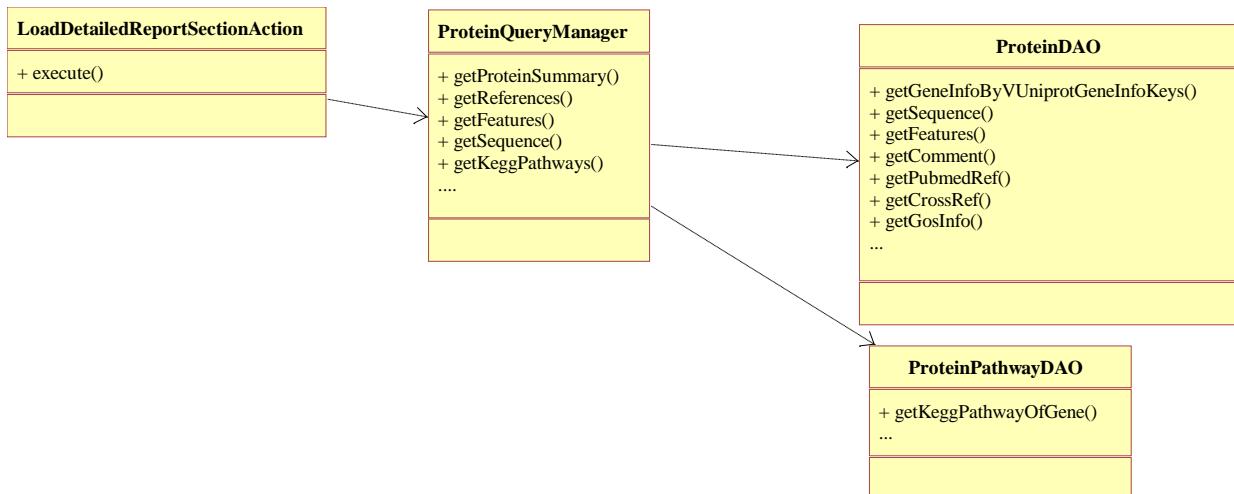
Pathways

Export

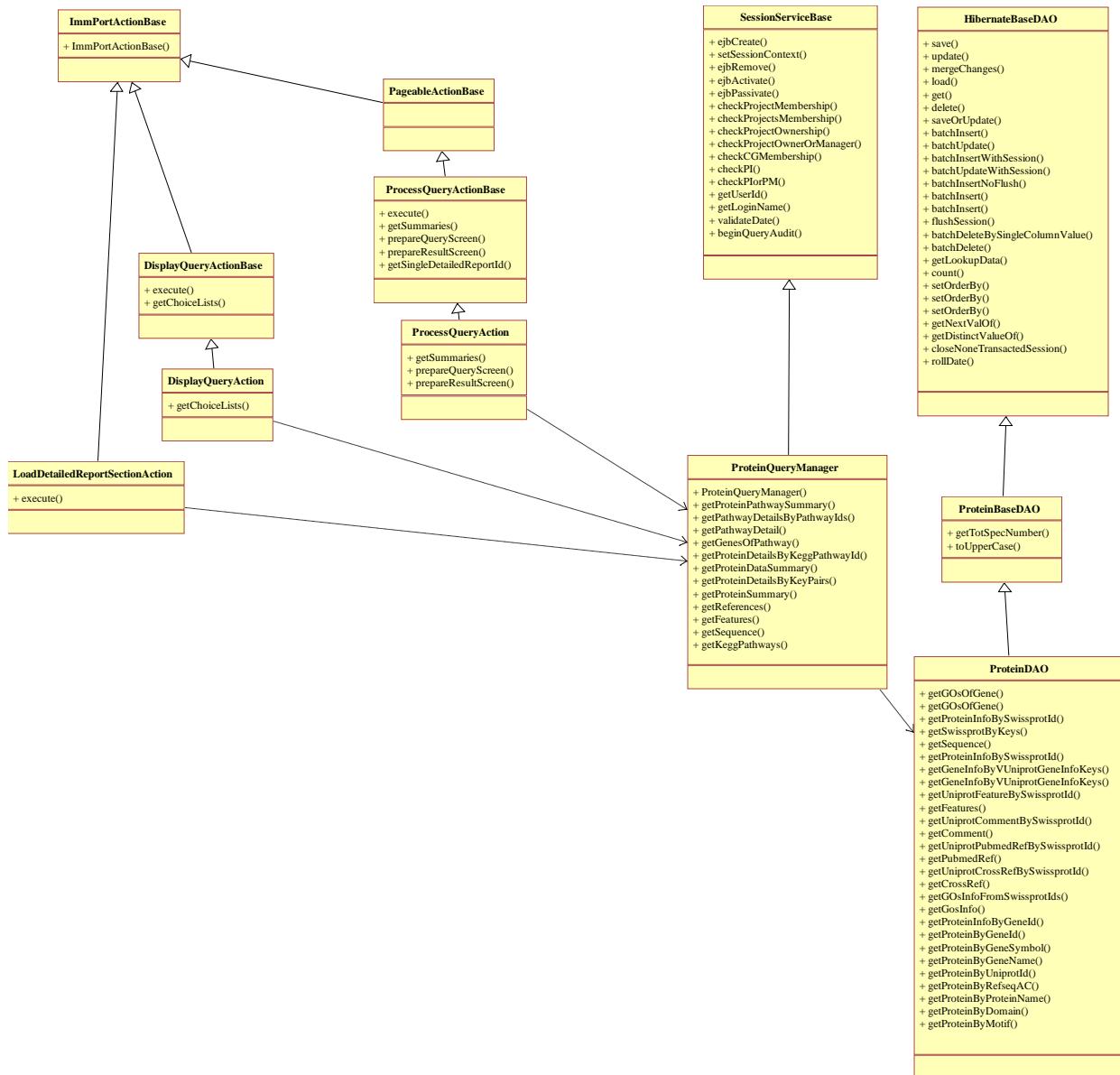
17.3.3 Protein Detail Sequence Diagram



17.3.4 Protein Detail Class Diagram



17.3.5 Protein Class Diagram



17.4 SNP

Single Nucleotide Polymorphisms are a common class of variants within genomic sequence that may have an impact on the regulation and function of genes and proteins.

The ImmPort SNP query user interface prompts the user to provide commonly used search attributes.

The screenshot shows the 'Reference Data / SNP Query' interface. At the top, there is a navigation bar with links to Home, Genes, Proteins, MHC Alleles, MHC Sequence Feature, Pathways, Protein Networks, SNPs, and Structure Viewer. Below the navigation bar, a message states: 'Fields marked with an asterisk * are required.' The main area is titled 'Search Options'. It includes fields for 'Search Type*' (set to 'Gene Symbol'), 'Search Option*' (set to 'Like'), 'Search Text*' (containing 'adh'), 'Search Species*' (set to 'Homo sapiens', with a dropdown menu showing other species like 'Mus musculus' and 'Rattus norvegicus'), 'Search dbSNP Version*' (set to 'dbSNP 126'), and 'SNP Function' (listing 'coding-nonsynonymous', 'coding-synonymous', 'exception', 'intron', and 'locus-region'). At the bottom, there are 'Submit' and 'Cancel' buttons, and a 'Results Per Page:' dropdown set to '25'.

Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose to more details about selected records.

Reference Data / SNP Query Results ⓘ

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists ▾ | Data History ▾ | Reference Advanced Search | Structure Viewer

Search Criteria [\(Modify Search\)](#)

Type:	Gene Symbol	Text:	adh
Species:	Homo sapiens	Option:	Like
Functions:		Genome Build:	dbSNP 126 - hg18

Select up to 10 search results to view in a detailed report.

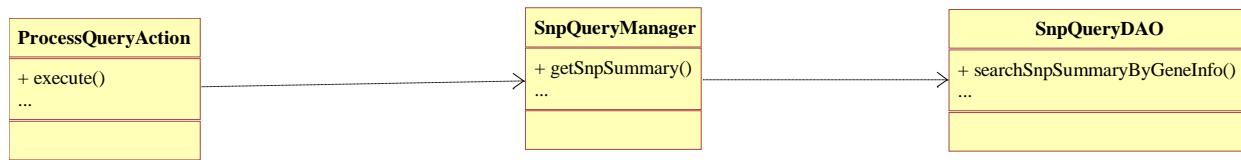
SNP IDs marked with ⓘ have HapMap genotype data available.

[View Detailed Report](#) 0 item(s) selected.

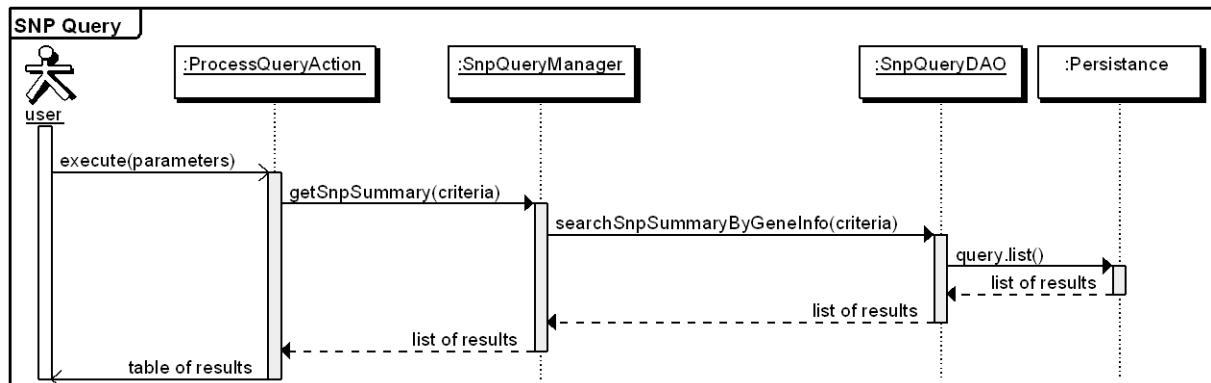
9,107 items found, displaying 1 to 25. [First/Prev] [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#) [Next/Last]

	SNP ID	Gene Symbol	Location	Allele	Gene Name	Function	Organism
<input type="checkbox"/>	rs17097429	ACADM	chr1 (+) 75968774	C/T	acyl-CoA dehydrogenase, C-4 to C-12 straight chain	coding-synonymous	human
<input type="checkbox"/>	rs11549022	ACADM	chr1 (+) 75999610	A/G	acyl-CoA dehydrogenase, C-4 to C-12 straight chain	coding-synonymous	human
<input type="checkbox"/>	rs2066702	ADH1B	chr4 (-) 100448040	C/T	alcohol dehydrogenase 1B (class I), beta polypeptide	coding-nonsynonymous	human
<input type="checkbox"/>	rs1126441	ADH1B	chr4 (-) 100456148	A/C	alcohol dehydrogenase 1B (class I), beta polypeptide	coding-nonsynonymous	human
<input type="checkbox"/>	rs41275699	ADH1B	chr4 (+) 100458291	A/G	alcohol dehydrogenase 1B (class I), beta polypeptide	coding-nonsynonymous	human

17.4.1 SNP Query Class Diagram

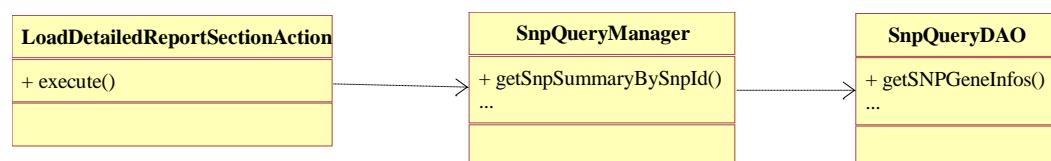


17.4.2 SNP Query Sequence Diagram

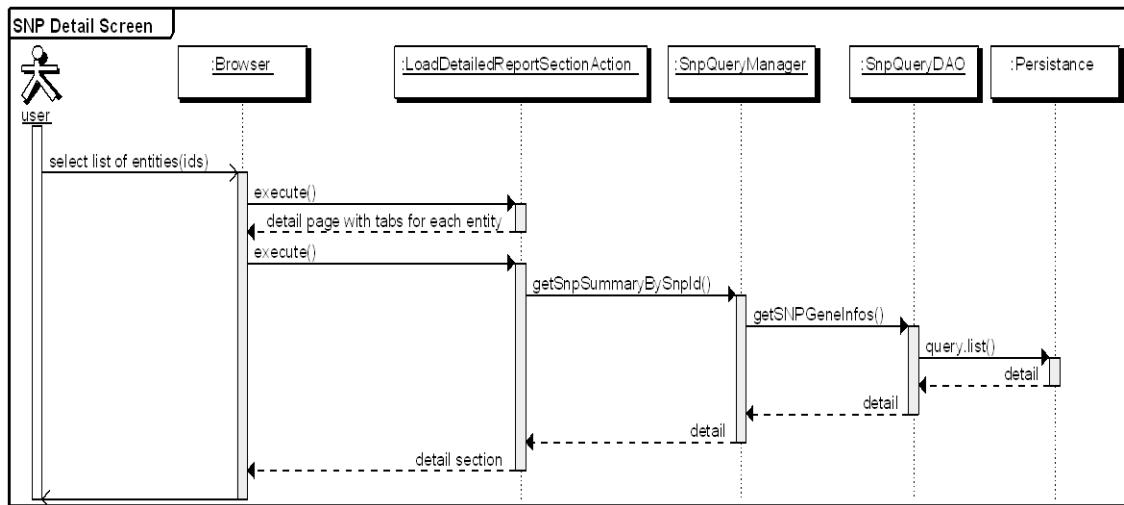


The SNP detail page aggregates data from several sources that are pertinent to the SNP record. The first section of the detail page is expanded by default and displays information provided in the NCBI dbSNP summary section. The user may expand one or all blocks of additional information.

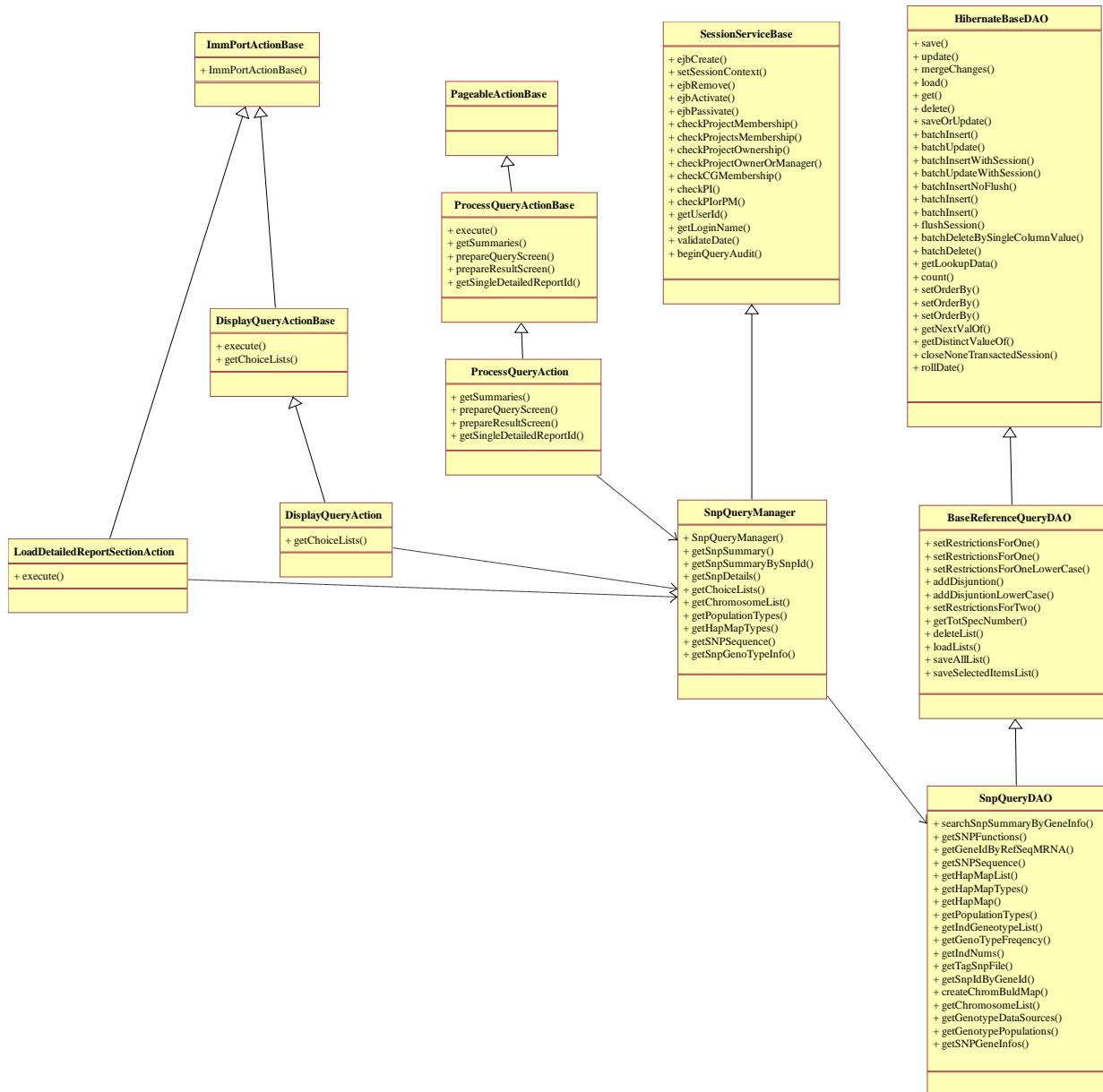
17.4.3 SNP Detail Class Diagram



17.4.4 SNP Detail Sequence Diagram



17.4.5 SNP Class Diagram



17.5 PATHWAY

Proteins and other biological entities interact with each other in pathways and networks to perform their functions.

The ImmPort Pathway query user interface prompts the user to provide commonly used search attributes.

Reference Data / Pathway Query [?](#)

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | Structure Viewer

Fields marked with an asterisk * are required.

Search Options

Search Type* Gene Symbol

Search Option* Like

Search Text* adh
(Comma delimited, 256 max chars)

Search Species* (Multiple) **Homo sapiens**
Mus musculus
Rattus rattus
Rattus norvegicus
Gallus gallus
Drosophila melanogaster
Macaca mulatta

Submit Cancel Results Per Page: 25

Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose to more details about selected records.

Reference Data / Pathway Query Results [?](#)

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search | Structure Viewer

Search Criteria ([Modify Search](#))

Type: Gene Symbol Text: adh
Species: Homo sapiens Option: Like

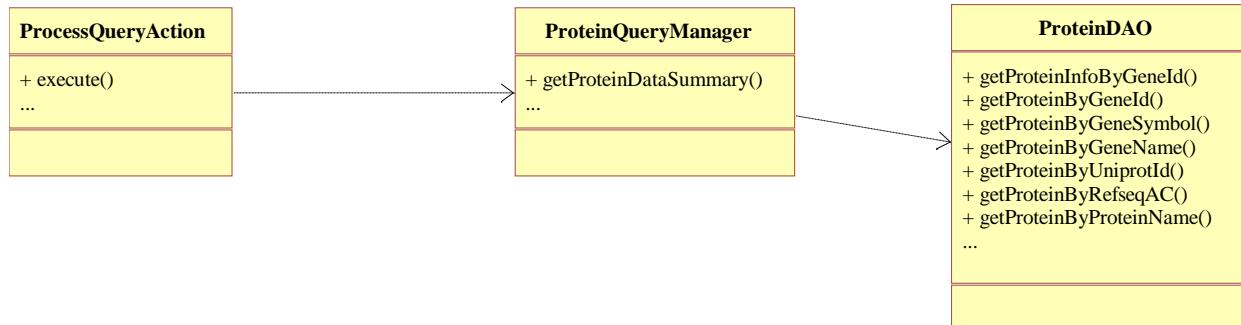
Select up to 10 search results to view in a detailed report.

View Detailed Report 0 item(s) selected.

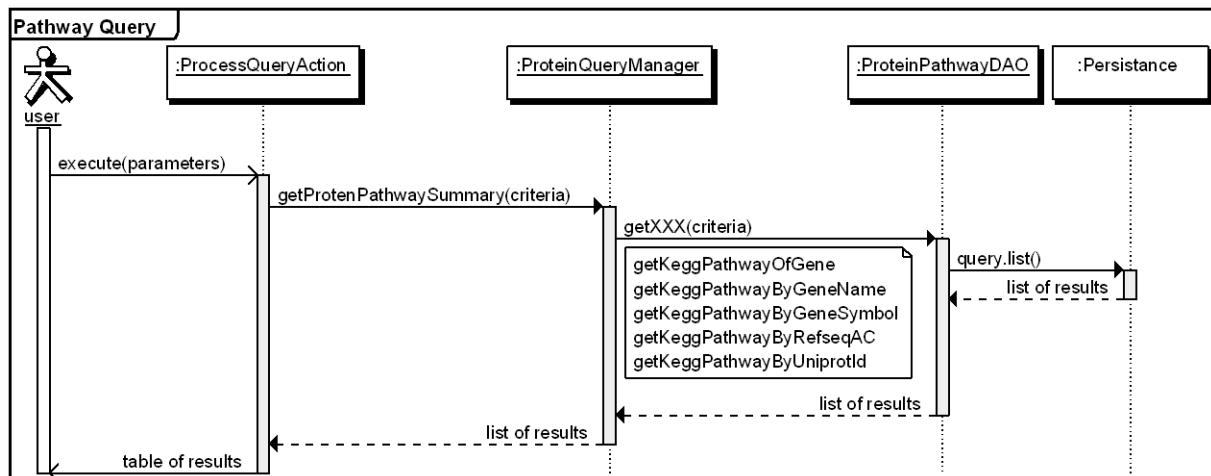
182 items found, displaying 1 to 25. [First/Prev] 1, 2, 3, 4, 5, 6, 7, 8 [Next/Last]

	Pathway ID	Kegg ID	Pathway Name	Enzref Gene ID	Gene Symbol	Gene Name
<input type="checkbox"/>	path:hsa00071	hsa:34	Fatty acid metabolism	34	ACADM	acyl-CoA dehydrogenase, C-4 to C-12 straight chain
<input type="checkbox"/>	path:hsa00280	hsa:34	Valine, leucine and isoleucine degradation	34	ACADM	acyl-CoA dehydrogenase, C-4 to C-12 straight chain
<input type="checkbox"/>	path:hsa00410	hsa:34	beta-Alanine metabolism	34	ACADM	acyl-CoA dehydrogenase, C-4 to C-12 straight chain
<input type="checkbox"/>	path:hsa00640	hsa:34	Propanoate metabolism	34	ACADM	acyl-CoA dehydrogenase, C-4 to C-12 straight chain
<input type="checkbox"/>	path:hsa01100	hsa:34	Metabolic pathways	34	ACADM	acyl-CoA dehydrogenase, C-4 to C-12 straight chain

17.5.1 Pathway Query Class Diagram



17.5.2 Pathway Query Sequence Diagram

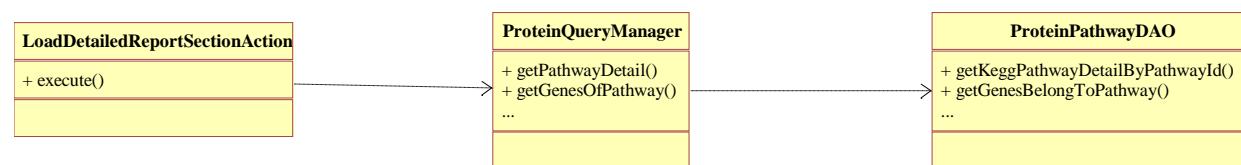


The Pathway detail page displays data from KEGG. The user may expand one or all blocks of additional information.

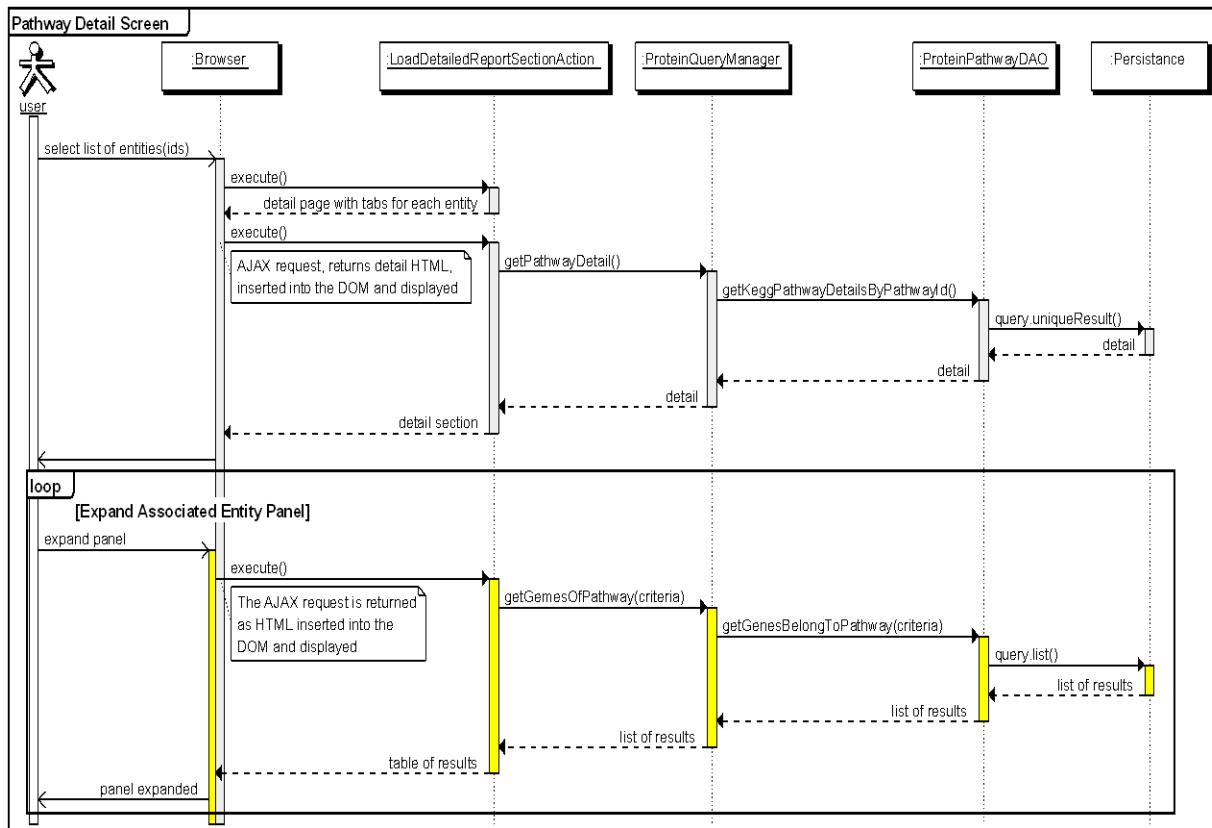
The screenshot shows a 'Reference Data / Pathway Detailed Report' page. At the top, there is a navigation bar with links like Home, Genes, Proteins, MHC Alleles, MHC Sequence Feature, Pathways, Protein Networks, SNPs, ImmPort Gene Lists, Data History, Reference Advanced Search, and a Back To Search Results link. Below the navigation, the pathway ID 'hsa00071' is displayed. The main content area is titled 'Pathway Summary' and contains details about the pathway: Pathway ID: path:hsa00071, Pathway Name: Fatty acid metabolism, Organism: Homo sapiens, and Data Source: KEGG. Below this is a table titled 'Genes In Pathway' with two rows:

Entrez Gene ID	Gene Symbol	Gene Name
130	ADH6	alcohol dehydrogenase 6 (class V)
1374	CPT1A	carnitine palmitoyltransferase 1A (liver)

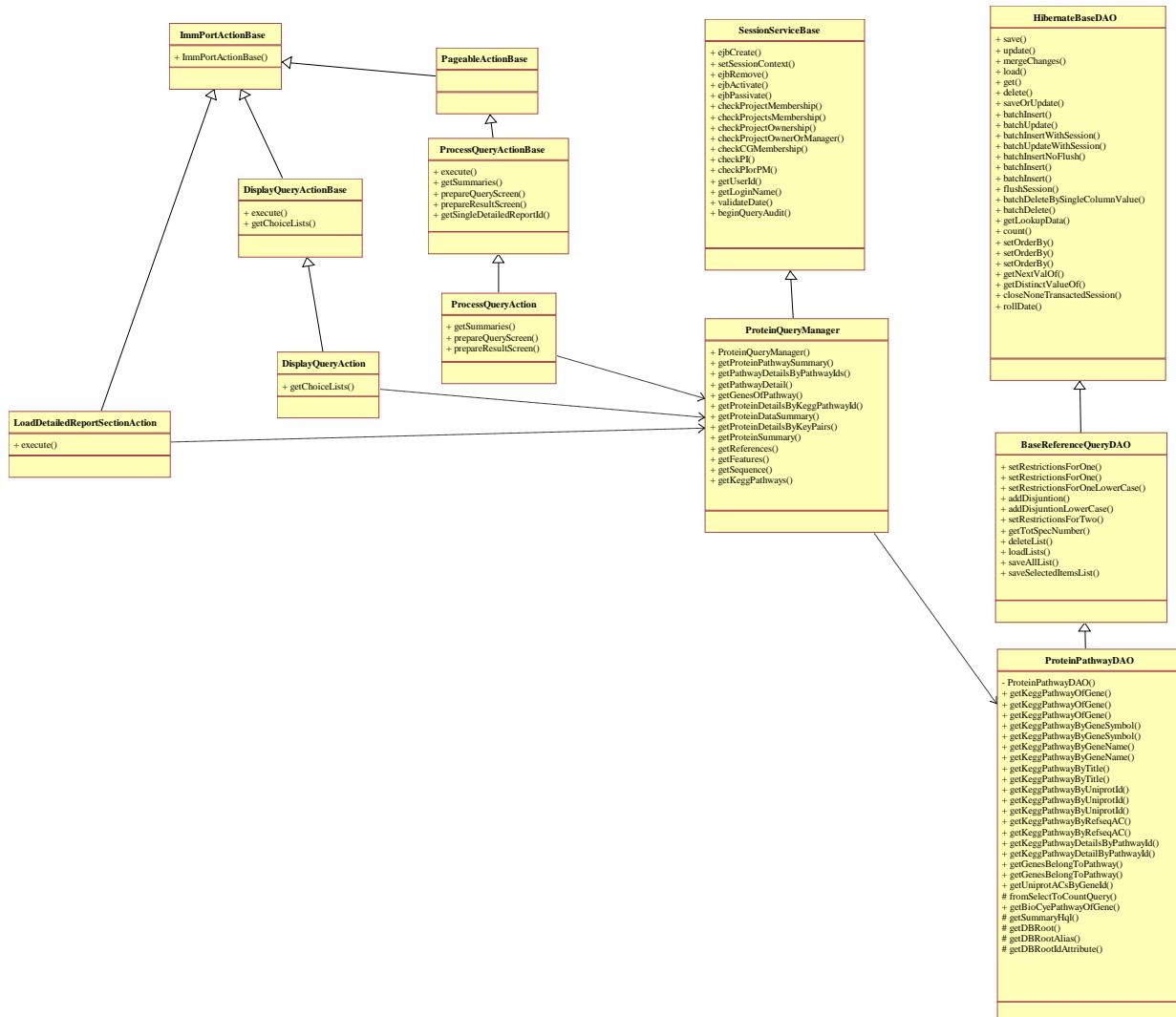
17.5.3 Pathway Detail Class Diagram



17.5.4 Pathway Detail Sequence Diagram



17.5.5 Pathway Class Diagram



17.6 PROTEIN NETWORKS

Proteins and other biological entities interact with each other in networks to perform their functions.

The ImmPort Network query user interface prompts the user to provide commonly used search attributes.

The screenshot shows a web-based search interface titled "Reference Data / Protein Network Query". At the top, there is a navigation bar with links to Home, Genes, Proteins, MHC Alleles, MHC Sequence Feature, Pathways, Protein Networks, and Structure Viewer. Below the navigation bar, a message states "Fields marked with an asterisk * are required." The main area is titled "Search Options". It contains several input fields and dropdown menus:

- "Search Type*" dropdown set to "Gene Symbol".
- "Search Option*" dropdown set to "Like".
- "Search Text*" text input field with placeholder "(Comma delimited, 256 max chars)".
- "Search Species*" dropdown menu showing species names:
 - Homo sapiens
 - Mus musculus
 - Rattus rattus
 - Rattus norvegicus
 - Gallus gallus
 - Drosophila melanogaster
 - Macaca mulattaA "Select All" link is located at the bottom right of this list.

At the bottom of the search options panel are "Submit" and "Cancel" buttons, and a "Results Per Page: 25" dropdown menu.

Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose to more details about selected records.

Reference Data / Protein Network Query Results ⓘ

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

Search Criteria ([Modify Search](#))

Type: Gene Symbol
Species: Homo sapiens

Text: adh
Option: Like

Interaction **Complex**

Select up to 10 search results to view in a detailed report.

View Detailed Report 0 item(s) selected.

7,254 items found, displaying 1 to 25. [[First/Prev](#)] [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#) [[Next/Last](#)]

Interaction ID	Data Source	Input Molecule	Interactor Molecule
24496	BIND	SMAD family member 3	nucleoporin 214kDa
54907	BIND	SMAD family member 3	v-ski sarcoma viral oncogene homolog (avian)
54911	BIND	SMAD family member 2	v-ski sarcoma viral oncogene homolog (avian)
55603	BIND	SMAD family member 3	SKI-like oncogene
55604	BIND	SMAD family member 2	SKI-like oncogene
55605	BIND	SMAD family member 1	SKI-like oncogene
57612	BIND	SMAD family member 4	v-ski sarcoma viral oncogene homolog (avian)
44759	BIND	standard database	standard database

The detail page for a given protein network would appear as shown below.

Reference Data / Protein Network Detailed Report ⓘ

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

[Back To Search Results](#)

66353

Summary

Interaction ID: 66353
Data Source: BIND
Detection Method: three-dimensional-structure

Collapse All: **Expand All:**

PubMed ID **PubMed Title**

[1896463](#) Structure of human beta 1 beta 1 alcohol dehydrogenase: catalytic effects of non-active-site substitutions.

Interactions

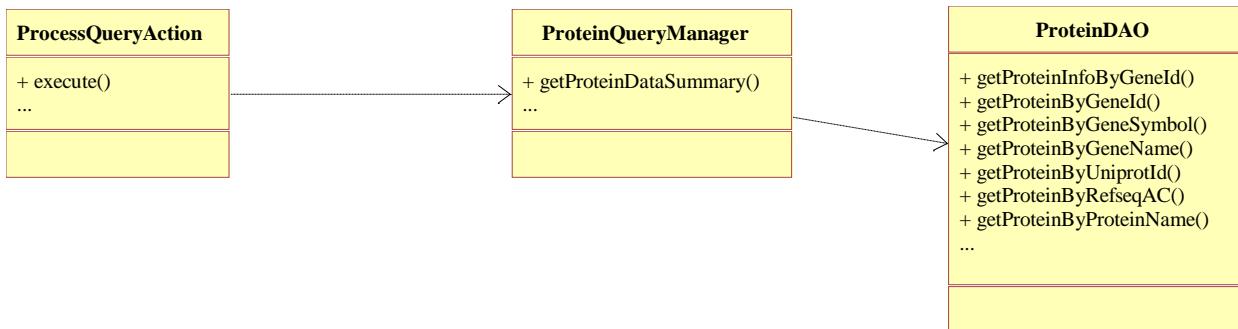
Input Molecule

Entrez Gene ID: 126 [[Details Build 36.1, hg18](#)] [[Details Build 35, hg17](#)]
Gene Name: alcohol dehydrogenase 1C (class I), gamma polypeptide
Gene Symbol: ADH1C , ADH3
Organism: Homo sapiens
Molecule Type: protein

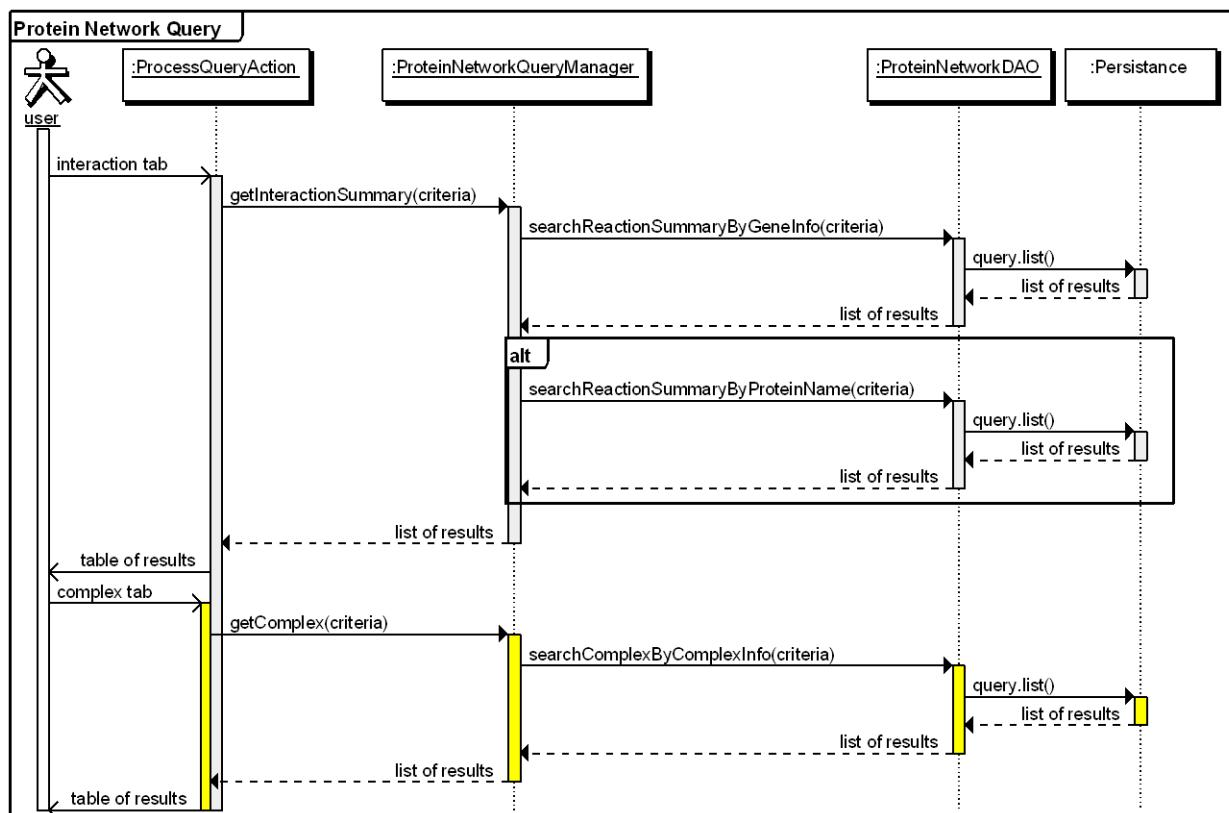
Protein Accession **Protein GI** **Protein Name**

AAA52278	178117	Protein Name not available.
AAC05739	2961522	Protein Name not available.
Q71V	74712853	Protein Name not available.
P00326	112398	Alcohol dehydrogenase 1C

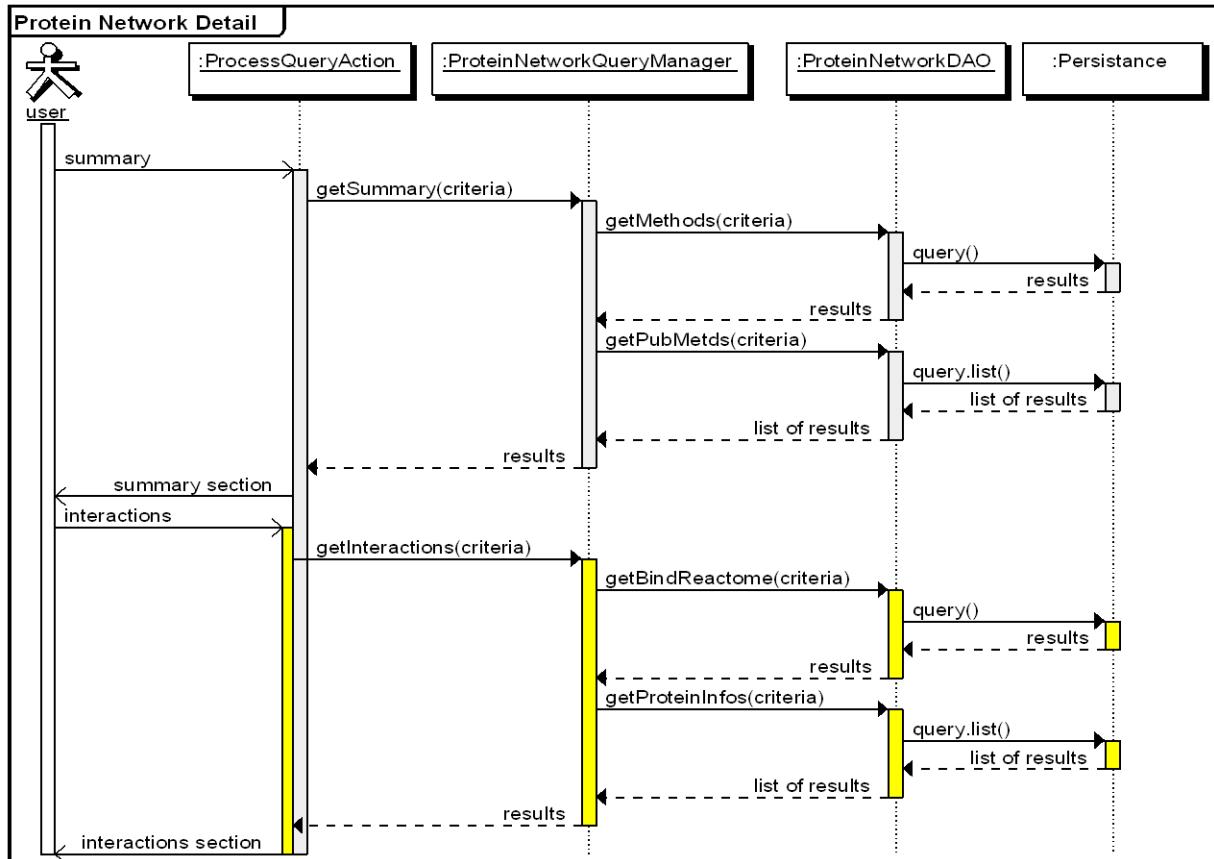
17.6.1 Protein Network Class Diagram



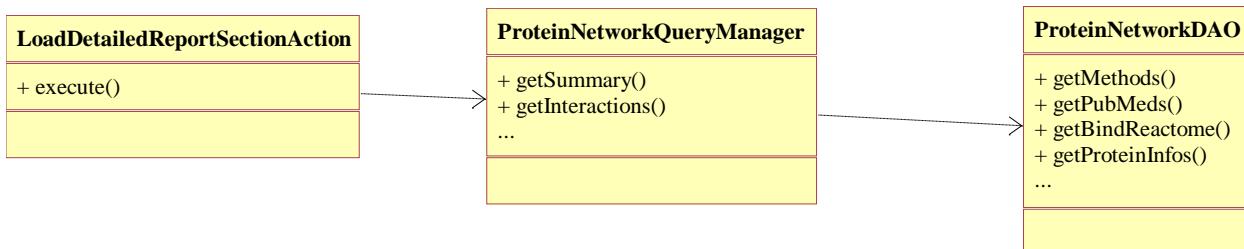
17.6.2 Protein Network Query Sequence Diagram



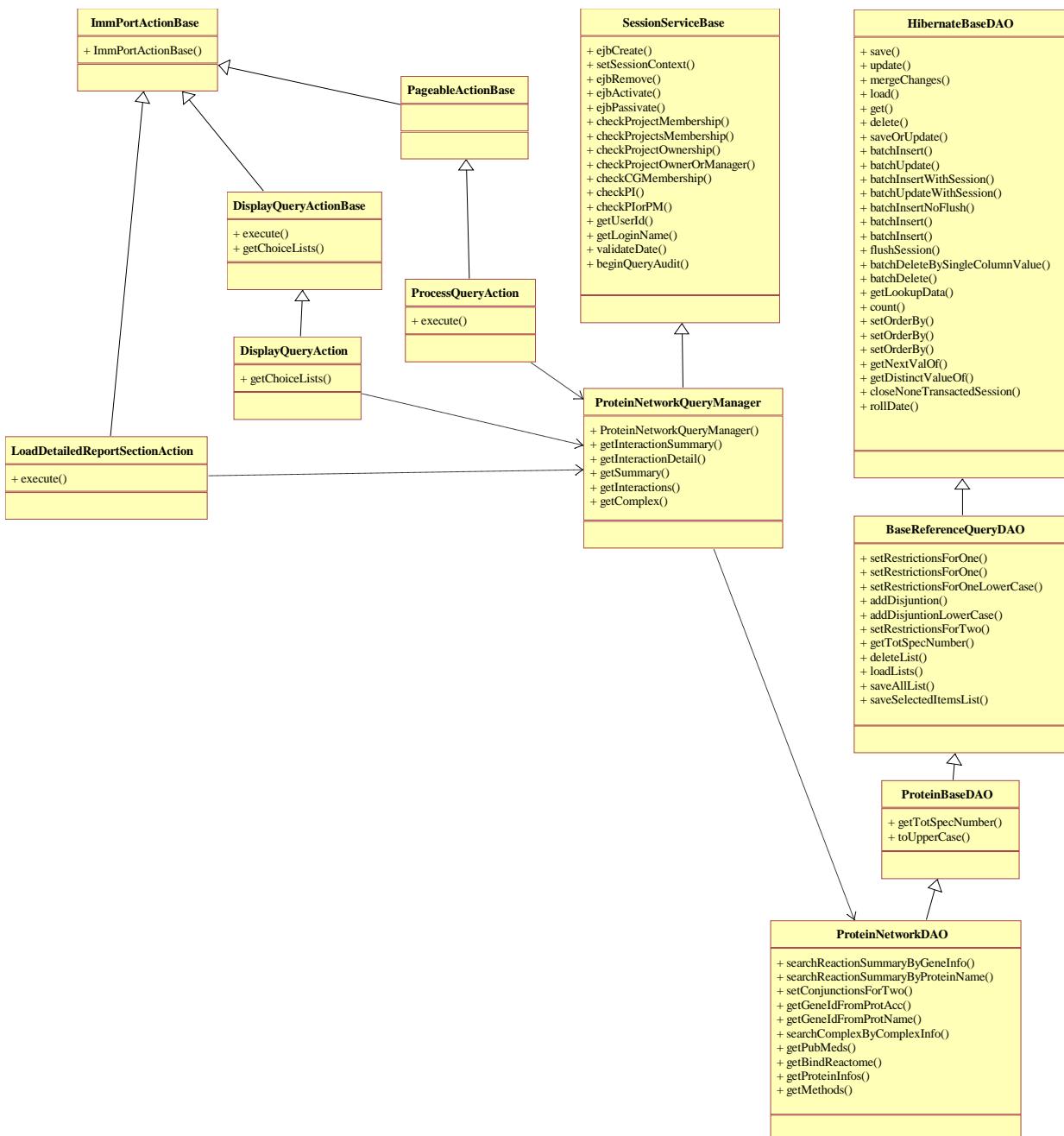
17.6.3 Protein Network Detail Sequence Diagram



17.6.4 Protein Network Detail Class Diagram



17.6.5 Protein Network Class Diagram



17.7 MHC ALLELE

The Major Histocompatibility Complex is of utmost interest to immunologists as it is a region of the genome that includes many of the immunologically related genes. The MHC data content is sufficiently unique to warrant a custom query and results display paradigm that distinguishes it from other reference data. ImmPort offers three primary query approaches to MHC data: allele names, sequence alignment, and sequence features.

17.7.1 MHC Allele Name Query

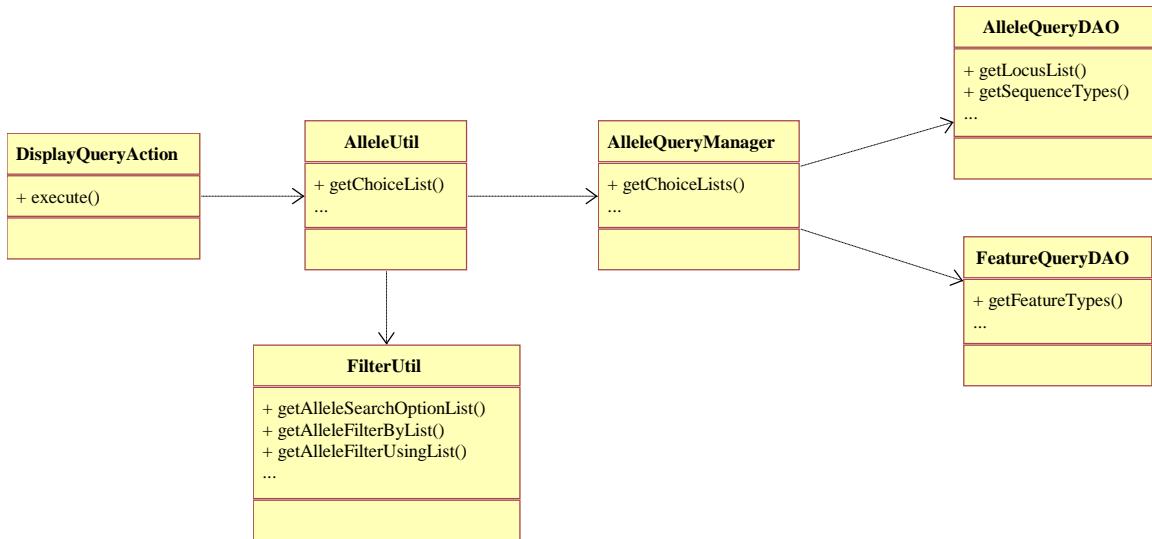
The MHC allele name query user interface displays commonly used query attributes.

The screenshot shows the ImmPort Reference Data / MHC Allele Search interface. At the top, there is a navigation bar with links to Home, Genes, Proteins, MHC Alleles, MHC Sequence Feature, Pathways, Protein Networks, SNPs, ImmPort Gene Lists, Data History, and Reference Advar Search. Below the navigation bar, a note states that HLA alleles and alignments are obtained from the IMGT/HLA database (version 3.0.0, April 2010) and provides references to the IMGT/HLA database and hla.allele.org database. A section titled "For more information on the sources of data for the MHC component in ImmPort, please see below." is present. The main search form is titled "MHC Allele Search - Selected Return Data Type : Allele". It includes a dropdown for "Please Choose Query Type" set to "Allele". Under "General Criteria", "Species" is set to "Homo sapiens" and "Locus" is set to "HLA-A". Under "Allele Criteria", "Search For" is set to "All Alleles", "Search Using" is set to "Allele Name", "Search Option" is set to "Like", and "Search Text" is a text input field containing "(Comma delimited)". At the bottom of the form are "Submit" and "Clear" buttons, and a "Results per page" dropdown set to 25.

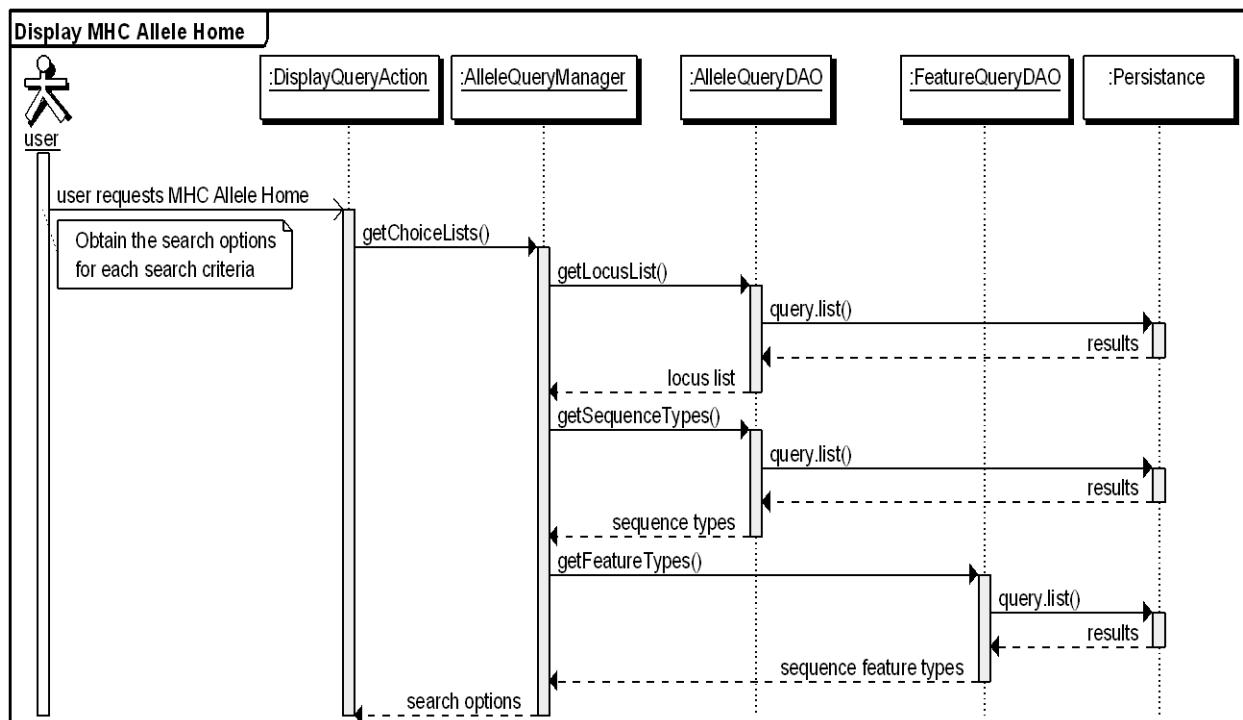
Upon completion of a query, ImmPort displays a tabular summary report of the query results. The user may choose from three options to see more information about a record: an allele detail report, a sequence feature report and an allele frequency report.

Old Allele Name	Sequence Features	Allele Frequency
A*01:01:01:CWD	[link]	[link]
A*01:01:01:01	[link]	[link]
A*01:01:01:02N	[link]	[link]
A*01:01:02	[link]	[link]
A*01:01:03	[link]	[link]
A*01:01:04	[link]	[link]
A*01:01:05	[link]	[link]

17.7.2 MHC Allele Name Query Class Diagrams



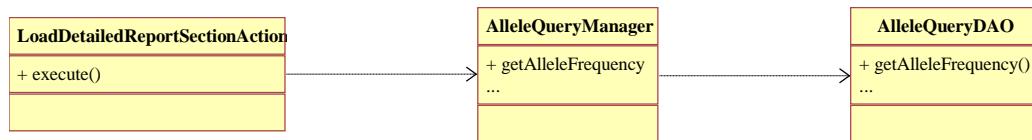
17.7.3 MHC Allele Name Query Sequence Diagrams



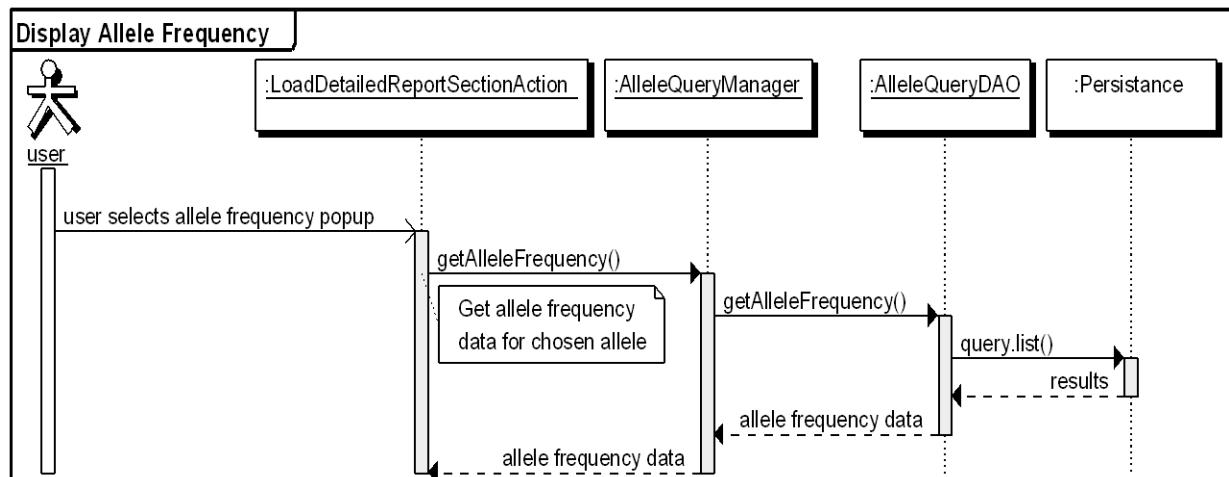
The allele frequency report displays data for a selected MHC allele in a static html table.

Allele Frequencies for A*01:01:01:01					
Allele Name	Data Type	Population Area	Allele Freq	Allele Count	Total Alleles Examined
A*01:01	HLA protein	Australia	0.0222	18	810
A*01:01	HLA protein	Europe	0.1642	600	3655
A*01:01	HLA protein	North Africa	0.1367	38	278
A*01:01	HLA protein	North America	0.0352	47	1337
A*01:01	HLA protein	North-East Asia	0.0587	62	1057
A*01:01	HLA protein	Oceania	0.0101	10	994
A*01:01	HLA protein	Other	0.0719	78	1085
A*01:01	HLA protein	South America	0.0021	1	479
A*01:01	HLA protein	South-East Asia	0.0072	42	5847

17.7.4 MHC Allele Frequency Popup Class Diagrams



17.7.5 MHC Allele Frequency Popup Sequence Diagrams

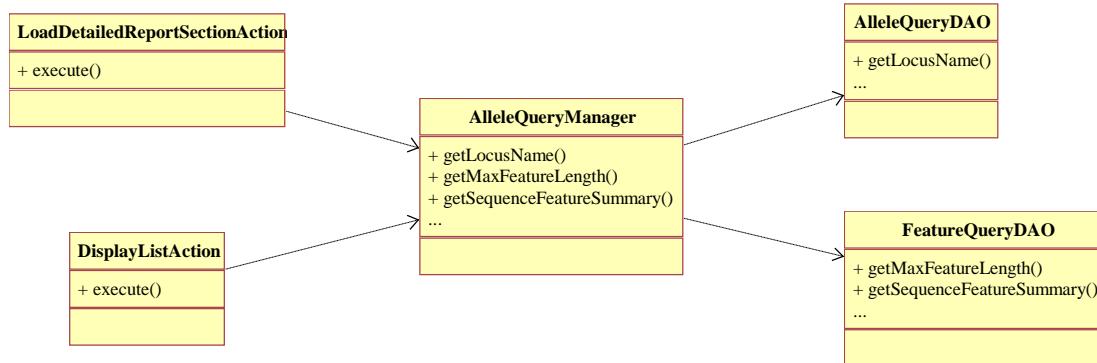


The sequence feature report displays as an interactive grid in a pop up window. The user may link to sequence feature variant types.

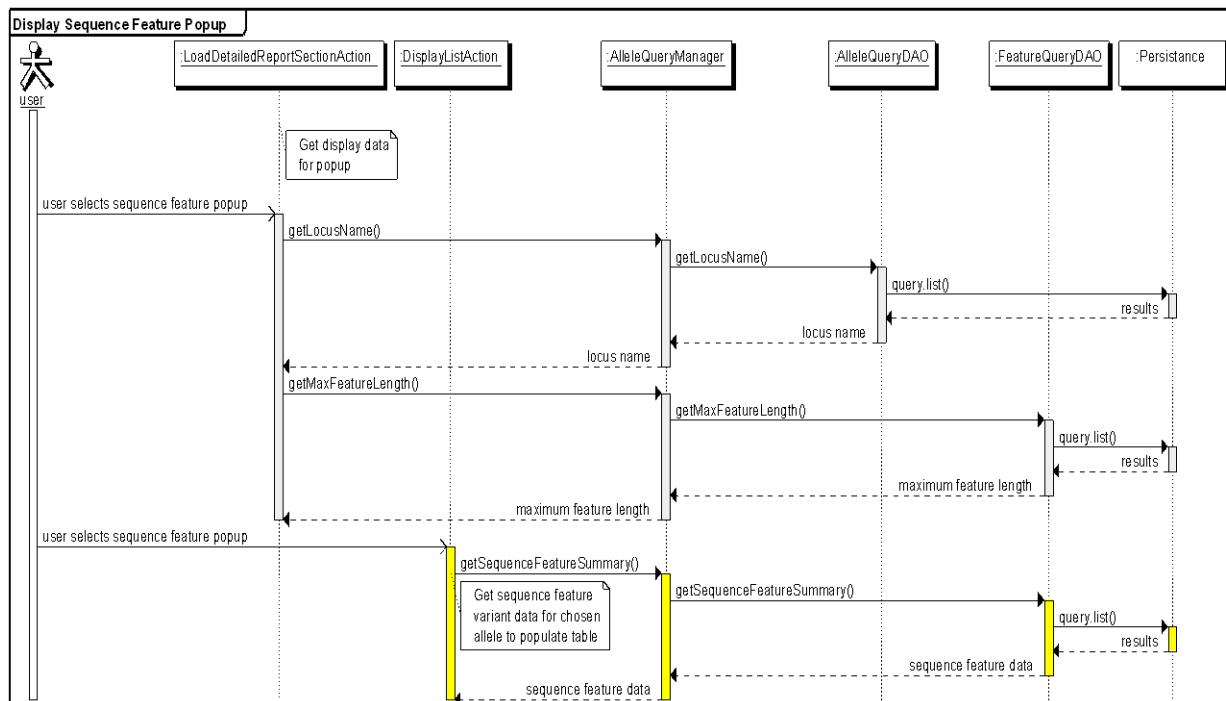
Allele Variant Sequence Features for A*01:01:01:01**Locus Name:** HLA-A

Variant Type	Sequence Feature Names	Sequence Feature Types	Positions	Sequence Motif
A*01:01	Hsa_HLA-A_allele	Structural - Complete protein		-24M_-23A_-22V_-21M_-20A
Hsa_HLA-A_SF2_VT1	Hsa_HLA-A_full-length protein	Structural - Complete protein	-24..341	-24M_-23A_-22V_-21M_-20A
Hsa_HLA-A_SF3_VT1	Hsa_HLA-A_signal peptide	Structural - Cleaved peptide	-24..-1	1G_2S_3H_4S_5M_6R_7Y_8F
Hsa_HLA-A_SF4_VT1	Hsa_HLA-A_mature protein	Structural - Complete protein	1..341	1G_2S_3H_4S_5M_6R_7Y_8F
Hsa_HLA-A_SF5_VT1	Hsa_HLA-A_alpha 1 domain	Structural - Domain	1..90	91G_92S_93H_94T_95I_96Q
Hsa_HLA-A_SF6_VT1	Hsa_HLA-A_alpha 2 domain	Structural - Domain	91..182	183D_184P_185P_186K_187T
Hsa_HLA-A_SF7_VT1	Hsa_HLA-A_alpha 3 domain	Structural - Domain	183..274	275E_276L_277S_278S_279Q
Hsa_HLA-A_SF8_VT1	Hsa_HLA-A_alpha 3 domain	Structural - Domain	275..284	285V_286G_287I_288I_289A
Hsa_HLA-A_SF9_VT1	Hsa_HLA-A_putative transmembrane	Structural - Domain	285..308	309R_310R_311K_312S_313S
Hsa_HLA-A_SF10_VT1	Hsa_HLA-A_putative cytoplasmic	Structural - Domain	309..341	

17.7.6 MHC Sequence Feature Popup Class Diagrams



17.7.7 MHC Sequence Feature Popup Sequence Diagrams



The allele detail report is the third option from the query summary report. This page displays data from several sources organized into expandable/collapsible blocks.

Reference Data / Allele Detailed Report

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

[Back To Search Results](#)

A*01:01:01:02N **A*01:01:01:01**

Allele Summary

Allele Name:	A*01:01:01:01
Old Allele Name:	A*01010101
IMGT/HLA Acc:	HLA00001
Sequence Features:	+
CWD Allele:	+
Entrez Gene ID:	3105 [Details Build 36.1, hg18] [Details Build 35, hg17]

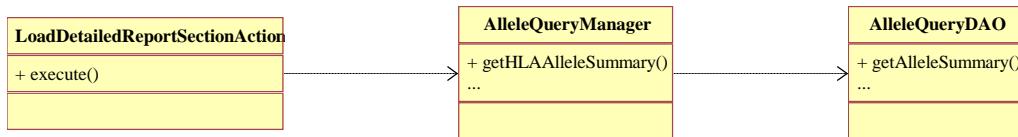
Assigned Date: 1989-08-01
Last Update: 1998-12-16
Reference Sequence Sources: M24043, Z93949, X55710, EU445470, AL645935, AJ278305, CR759913,
Reference Cell Lines: PP, B4702, APD, LCL721, COX, MOLT-4,
Aliases: A*01:01:01G (G-Code), A*01:01P (P-Code), A*01010101 (changed), A*0101 (changed), A*01011 (changed), [hla.allele.org](#) references for [HLA-A](#)

Allele Frequencies

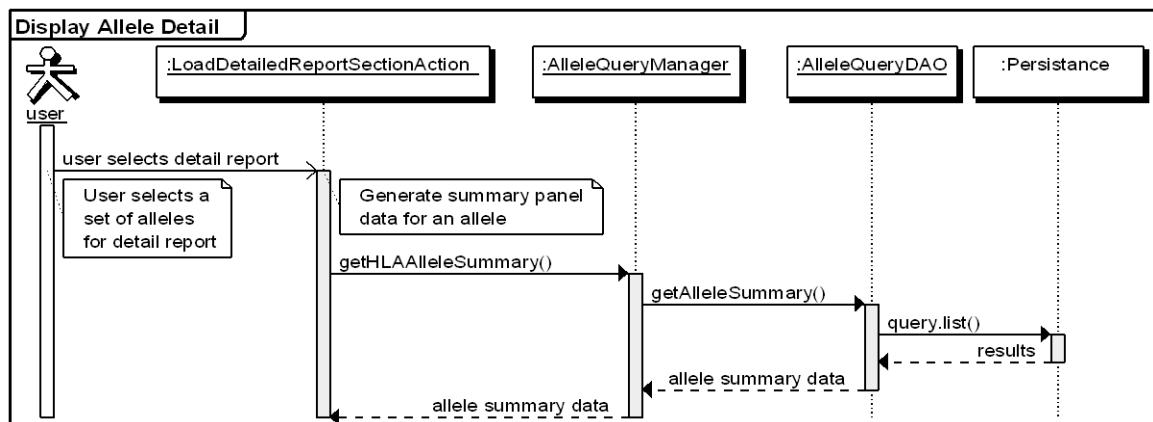
PubMed Publications

Sequences

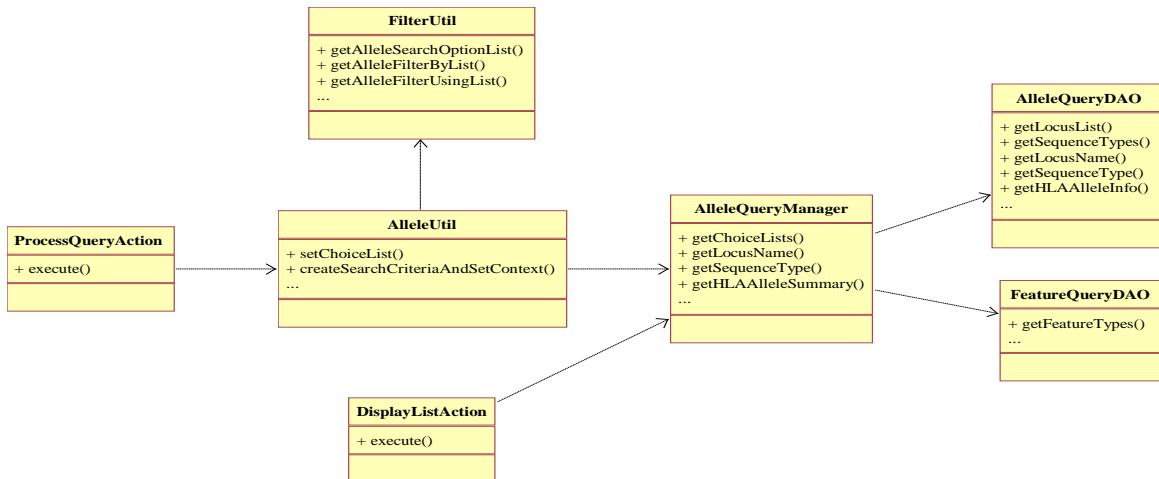
17.7.8 MHC Allele Name Detail Class Diagrams

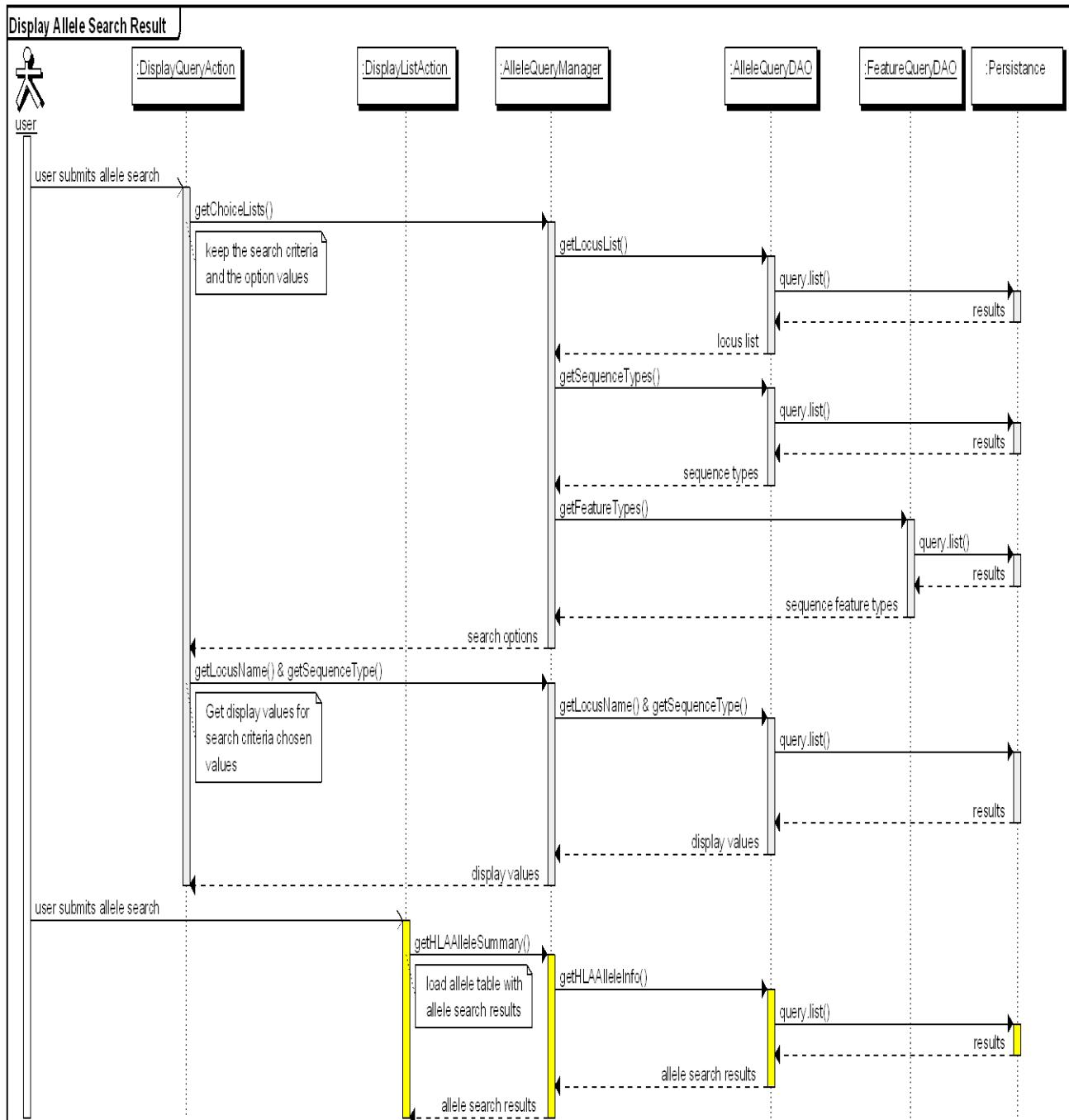


17.7.9 MHC Allele Detail Sequence Diagrams



17.7.10 MHC Allele Name Result Class Diagrams



17.7.11**MHC Allele Name Result Sequence Diagrams****17.7.12 MHC Allele Alignment Query Mode**

The sequence alignment query user interface displays the allele name search attributes and additional attributes that apply to gene, transcript, and protein sequences.

The HLA alleles and the allele alignments data are obtained from IMGT/HLA database (IMGT/HLA release version 3.0.0, April, 2010). Please publications for information regarding the [IMGT/HLA database](#) and the [hla.allele.org database](#):

- ▶ Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SG: The IMGT/HLA database, Nucleic Acids Research (2009), 37:D1013-D1017.
- ▶ Robinson J, Malik A, Parham P, Bodmer JG, Marsh SG: IMGT/HLA - a sequence database for the human major histocompatibility complex, Tissue Antigen (2004), 64:101-105.

For more information on the sources of data for the MHC component in ImmPort, please see below.

Fields marked with an asterisk * are required.

MHC Allele Search - Selected Return Data Type : Alignment

Please Choose Query Type **Alignment**

General Criteria:

Species: **Homo sapiens**
Locus *: **HLA-A**

Allele Criteria:

Search For *: **All Alleles**
Search Using *: **Allele Name**
Search Option*: **Like**
Search Text **(Comma delimited)**

Alignment Criteria:

Sequence Type *: **genomic**
Alignment Span *:

All
 Positions **Start** _____ **End** _____
 Features **Exon**
Intron
5UTR
3UTR

Submit **Clear** Results per page: **25**

The alignment search result page provides an alignment view of the sequence type (genomic, protein, or transcript) requested. The user may link to an MHC allele.

Reference Data / Allele Alignment Results

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | F S

([Modify Search](#))

Allele Search Criteria

Locus Name: **HLA-A**
Search For: **All Alleles**
Search Using: **Allele Name**
Search Option: **Like** Text:
Text:

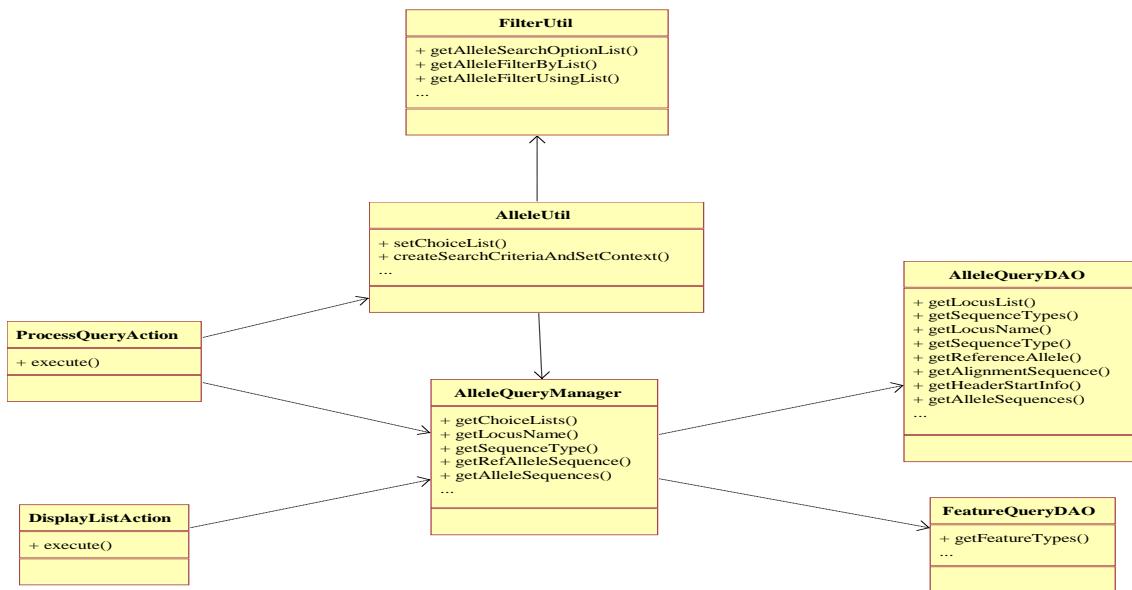
Alignment Specific Criteria

Sequence Type: **genomic**
Alignment Span: **ALL** Types:

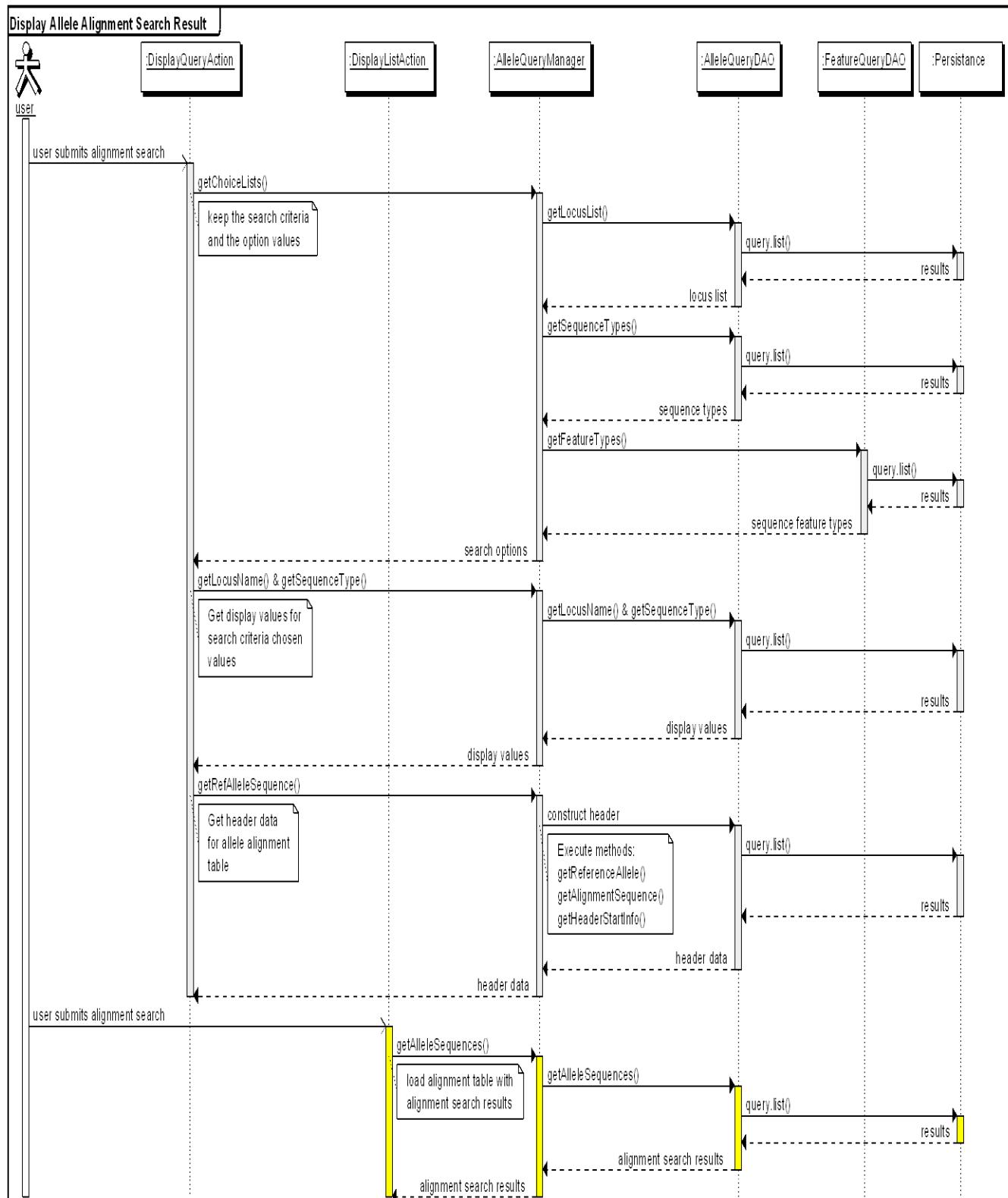
Allele ID	Sequence
A*01:01:01:01 (Ref)	[1 10 20 30 40 50 60 70 CAGGAGCAGA GGGGTCAGGG CGAAGTCCCA GGGCCCCAGG CGTGGCTCTC AGGGTCTAG GCCCCGAAGG CGGTGTA ***** A*01:01:01:01 A*01:01:01:02N A*01:03 A*01:04N A*01:09 A*01:11N A*01:14 A*01:16N]

Page 1 of 4 | Export | Displaying 1 - 25 of 82

17.7.13 MHC Allele Alignment Result Class Diagrams



17.7.14 MHC Allele Alignment Result Sequence Diagrams



17.7.15 MHC Allele Sequence Feature Query Mode

The MHC allele sequence feature query user interface displays the allele name search attributes and additional attributes that apply to structural and functional annotations of the MHC allele sequences.

MHC Allele Search - Selected Return Data Type : Sequence Feature

Please Choose Query Type **Sequence Feature**

General Criteria:

Species: Homo sapiens
Locus *: HLA-A

Allele Criteria:

Search For *: All Alleles
Search Using *: Allele Name
Search Option *: Like
Search Text (Comma delimited):

Sequence Feature Criteria:

Feature Type: All Types

- Structural
- Functional
- Sequence Alteration
- Structural - Complete protein
- Structural - Domain
- Structural - Secondary structure motif
- Structural - Cleaved peptide region
- Sequence Alteration - Single amino acid variation
- Sequence Alteration - Insertions and Deletions
- Structural_Functional Combination
- Structural_Sequence Alteration Combination
- Functional_Sequence Alteration Combination

Feature Names (Comma delimited):

Feature Locations (Comma delimited):

Results are returned in a summary table, where up to ten results may be selected to view in a more detailed page.

Reference Data / Sequence Feature Results

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data Structure Viewer

(Modify Search)

Allele Search Criteria

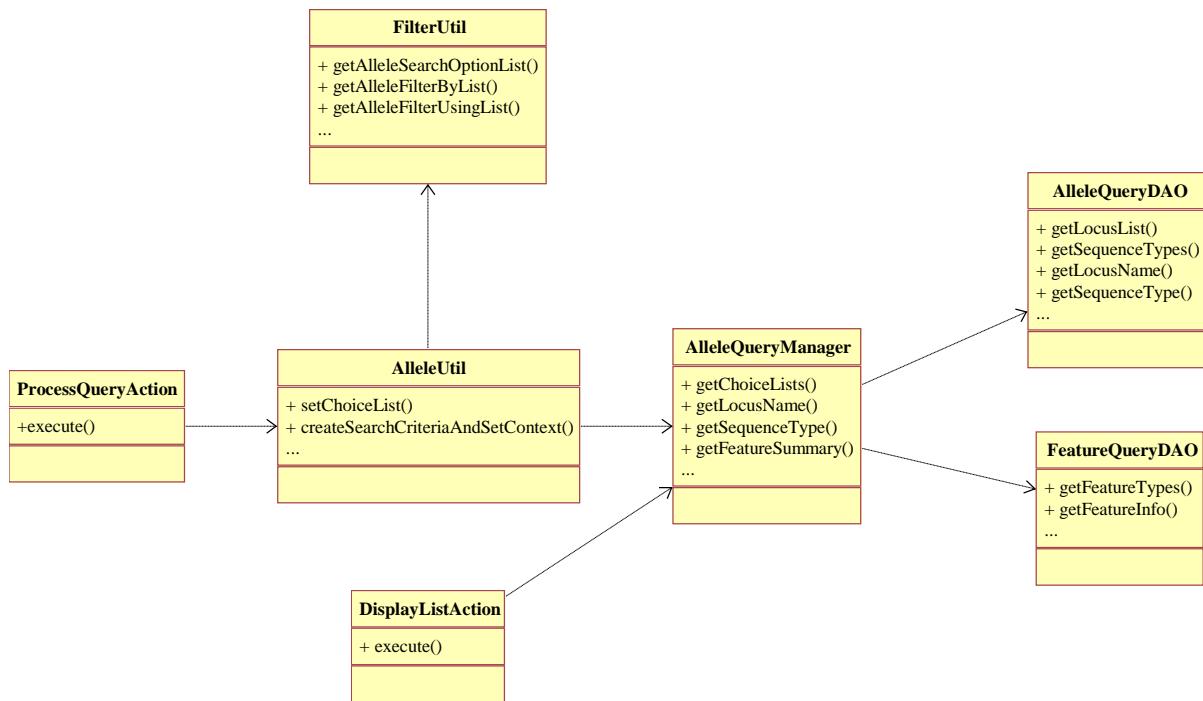
Locus Name: HLA-A
Search For: All Alleles
Search Using: Allele Name
Search Option: Like

Sequence Feature Specific Criteria

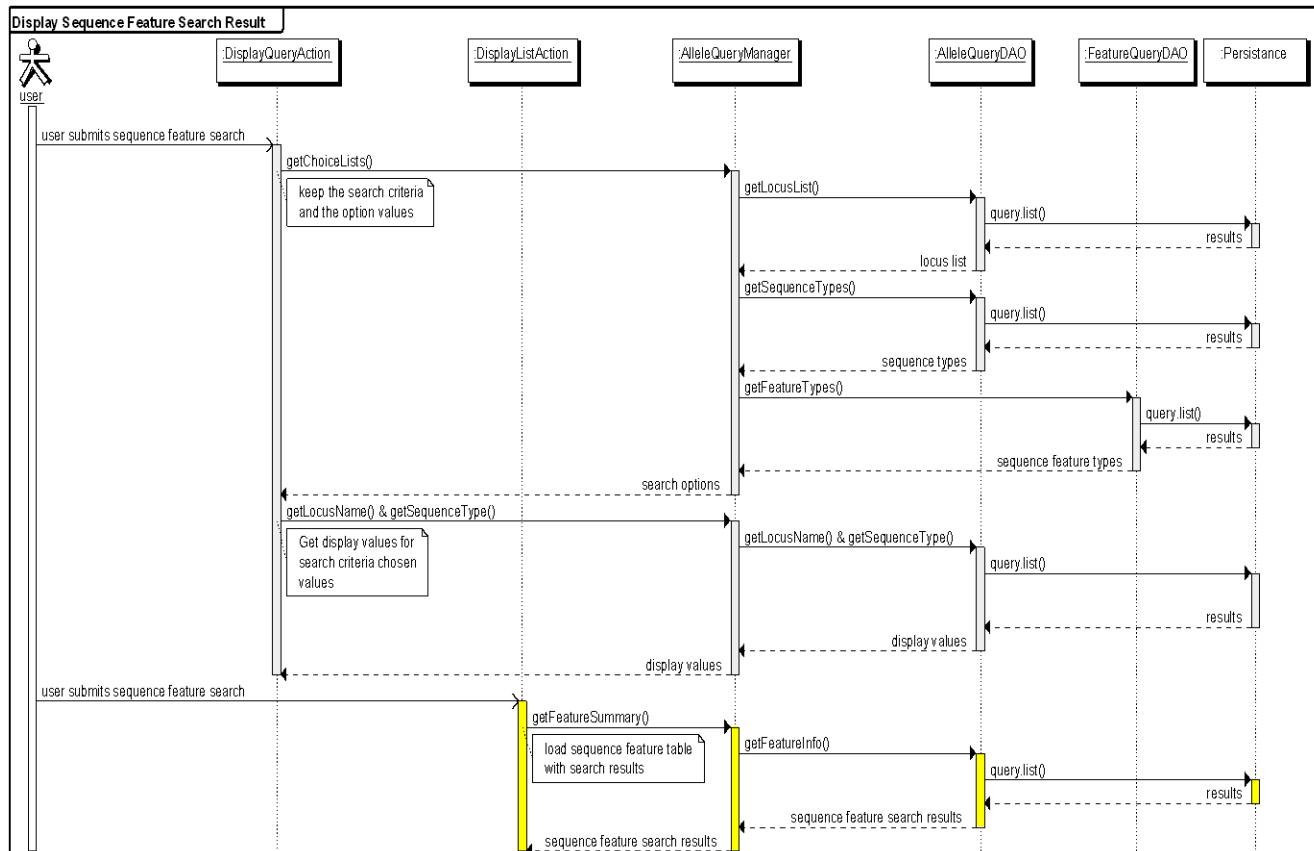
Feature Types: ALL,
Feature Names:
Feature Locations:

Page	1 of 14	View Details	Export
Sequence Feature Number	Sequence Feature Names	Sequence Feature Types	Positions
Hsa_HLA-A_SF1	Hsa_HLA-A_allele	Structural - Complete protein	
Hsa_HLA-A_SF2	Hsa_HLA-A_full-length protein	Structural - Complete protein	-24..341
Hsa_HLA-A_SF3	Hsa_HLA-A_signal peptide	Structural - Cleaved peptide region	-24..-1
Hsa_HLA-A_SF4	Hsa_HLA-A_mature protein	Structural - Complete protein	1..341
Hsa_HLA-A_SF5	Hsa_HLA-A_alpha 1 domain	Structural - Domain	1..90

17.7.16 MHC Allele Sequence Feature Result Class Diagrams



17.7.17 MHC Allele Sequence Feature Result Sequence Diagrams



The MHC allele sequence feature detail page displays number of alleles that share the feature, the variant type and the variant motif sequence.

Reference Data / Sequence Feature Detailed Report [?](#)

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | RefSeq

Back To Feature Search Results

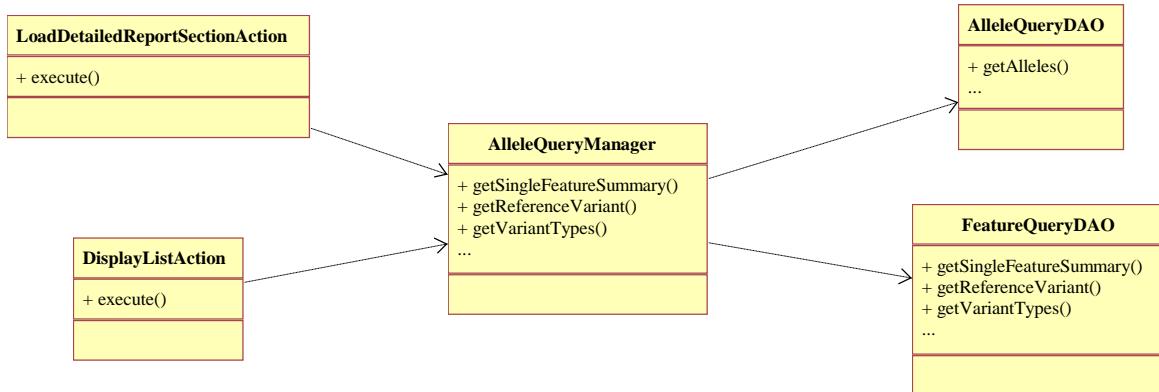
Hsa HLA A SF5 Hsa HLA A SF6

Locus Name:	HLA-A
Sequence Feature Names:	Hsa_HLA-A_alpha 1 domain
Positions:	1..90
Allele Alignment:	...
Sequence Feature Types:	Structural - Domain
MHC Sequence Feature Structure Viewer:	3H9S

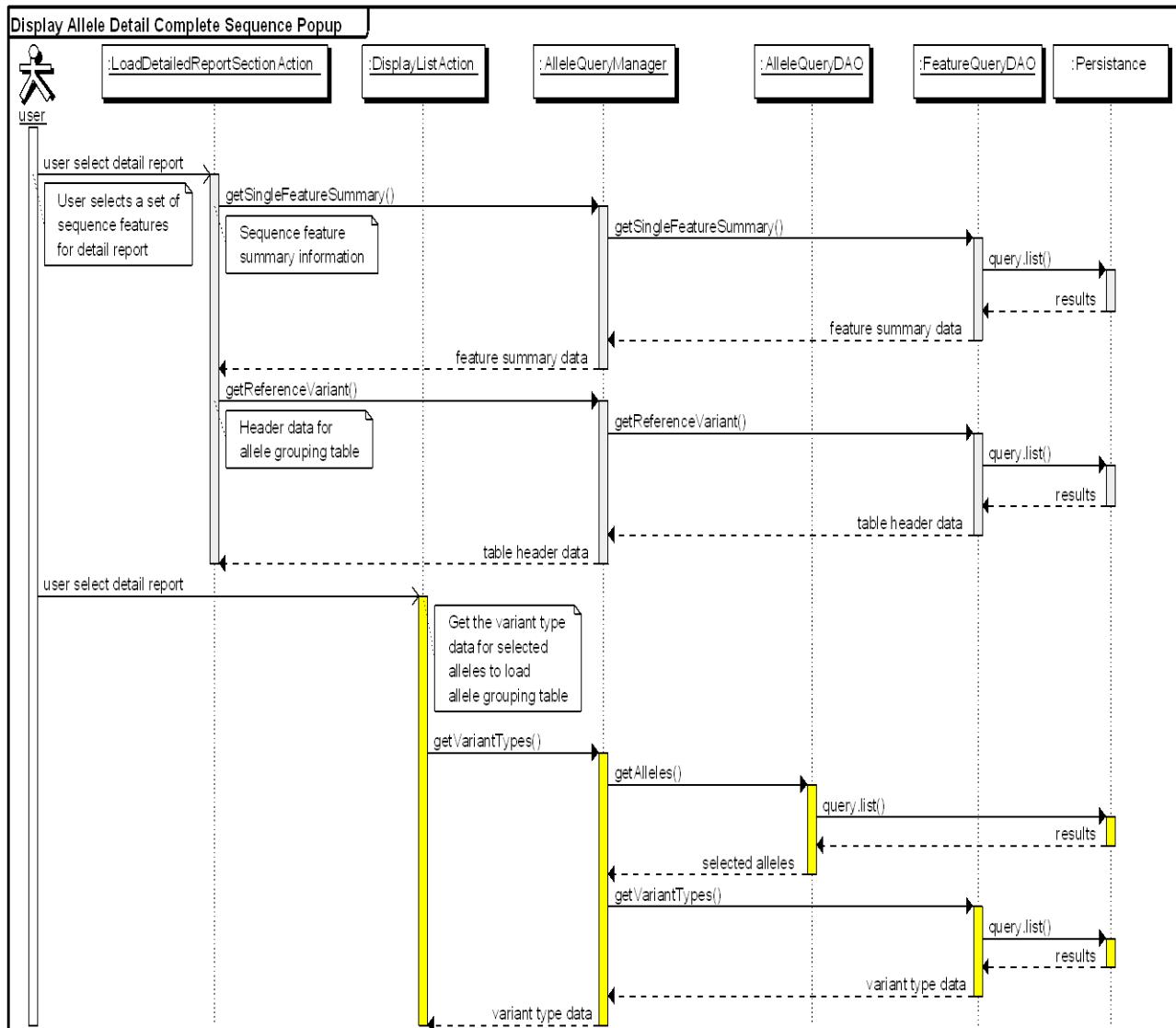
Allele Grouping

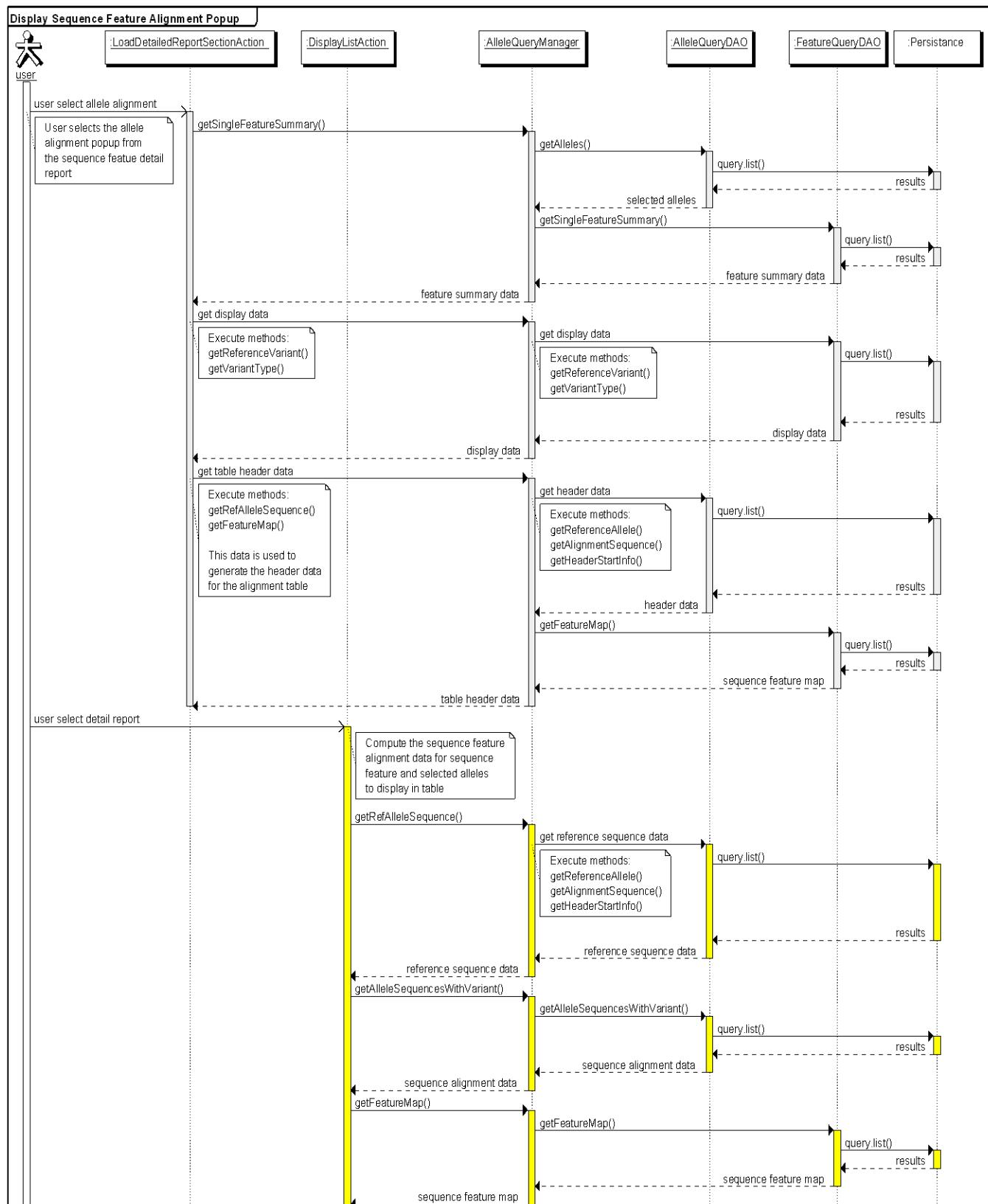
No of Alleles (All Alleles for Locus)	Variant Type	Variant Motif
926	Hsa_HLA-A_SF5_VT1	GSHSMRYFFTTSVRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIE
19	Hsa_HLA-A_SF5_VT1	-
2	Hsa_HLA-A_SF5_VT2	-S-
1	Hsa_HLA-A_SF5_VT3	-S-
70	Hsa_HLA-A_SF5_VT4	-L-
...		-R-

17.7.18 Sequence Feature Detail Class Diagrams



17.7.19 Sequence Feature Detail Sequence Diagrams





17.8 MHC SEQUENCE FEATURE STRUCTURE VIEWER (MSFSV)

The MHC Sequence Feature Structure Viewer (MSFSV) enables viewing 3 dimensional molecular structures of MHC proteins and their sequence feature variant types. This module implements the Jmol viewer for chemical structures in 3D with MHC sequence feature variant types developed by University of Texas Southwestern and the MHC community. The 3-dimensional molecular structures for the HLA loci were obtained from the Protein Database (PDB).

The MSFSV query user interface displays search attributes for MHA alleles.

The screenshot shows the ImmPort's MHC Sequence Feature Structure Viewer. At the top, there is a navigation bar with links to Home, Genes, Proteins, MHC Alleles, MHC Sequence Feature, Pathways, Protein Networks, SNPs, and ImmStructure Viewer. Below the navigation bar, a section titled "ImmPort's MHC Sequence Feature Structure Viewer highlights the locations of sequence features on crystal Features of the ImmPort MHC Sequence Feature Structure Viewer include:" lists various features: Selection/highlighting of any defined sequence feature on a reference structure for each locus, Selection/highlighting of other residues of interest on the MHC molecule, peptide or associated receptor, Various display modes, including ball and stick, space fill and ribbons, Coloring by charge, residue, structure type, temperature factor or composition, Zoom and spin, Jmol command line for use of advanced Jmol features, Image saving (.jpeg), and Save/Restore of the image viewer (zoom, highlighting, orientation, etc). Below this, a note says "To run the Jmol viewer, you must first select a Locus and then a PDB ID for that locus". It also states that fields marked with an asterisk * are required. A "Jmol Selection Criteria" form is shown with dropdown menus for Species (Homo sapiens), Locus (HLA-A), and PDB ID (3H9S), and a "Submit" button. At the bottom, a note says "Jmol is an open-source Java viewer for chemical structures in 3D".

The MSFSV query result page provides the 3-dimensional molecular structure for a protein matching the query. The following panels require information by the action classes servicing this page: Information, Allele to PDB Mapping, Sequence Features, and Highlight/Label Features. The other panels (Display Options, Save Restore, and Jmol Command Line) are static in that they do not require information from the server.

Reference Data / MHC Sequence Feature Structure Viewer [?](#)

Home | Genes | Proteins | MHC Alleles | MHC Sequence Feature | Pathways | Protein Networks | SNPs | ImmPort Gene Lists | Data History | Reference Advanced Search

Display Options

Collapse All: Expand All:

These options control the general appearance of the protein structure viewer.

Display Type:

Zoom:

Spin:

Sequence Features [?](#)

Highlight sequence features on the structure in red First select a sequence feature type(s) from a list. Then check the sequence features to highlight.

All Types Load

Structural
Functional
Sequence Alteration
Structural - Complete protein
Structural - Domain
Structural - Secondary structure motif
Structural - Cleaved peptide region
Sequence Alteration - Single amino acid variation
Sequence Alteration - Insertions and Deletions
Structural_Functional Combination
Structural_Sequence Alteration Combination
Functional_Sequence Alteration Combination

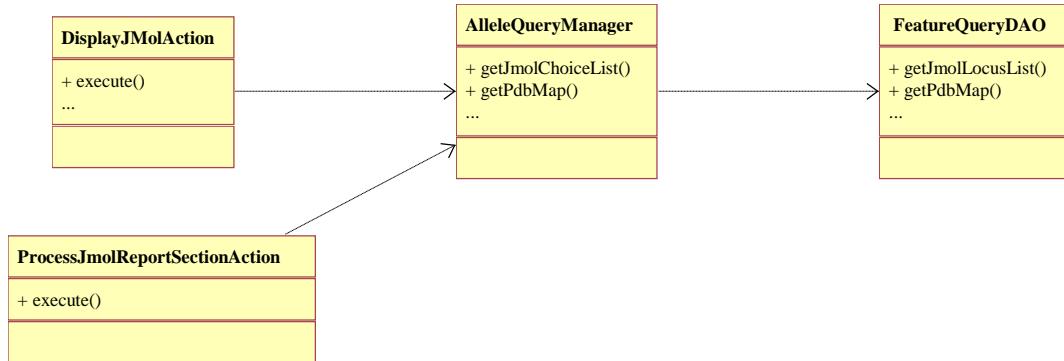
Display Display selected features from table into PDB protein structure

Clear Clear selected features from PDB protein structure and sequence feature table

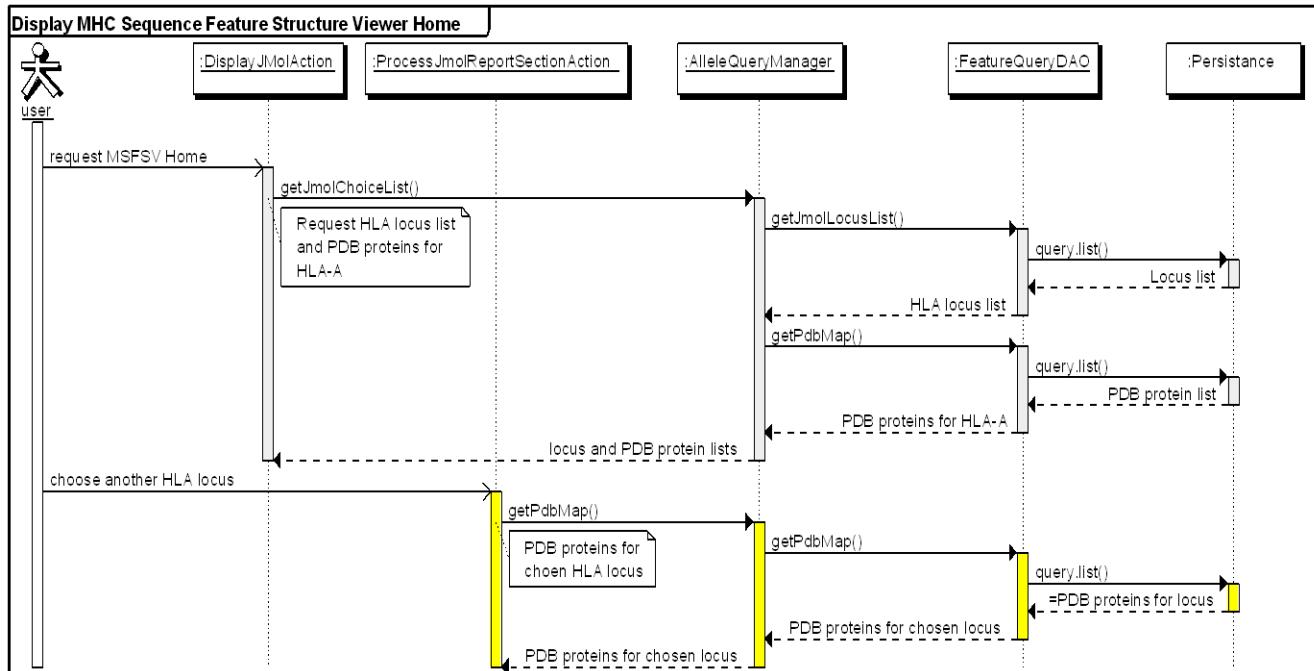
Jmol

Reset View Save View As Image

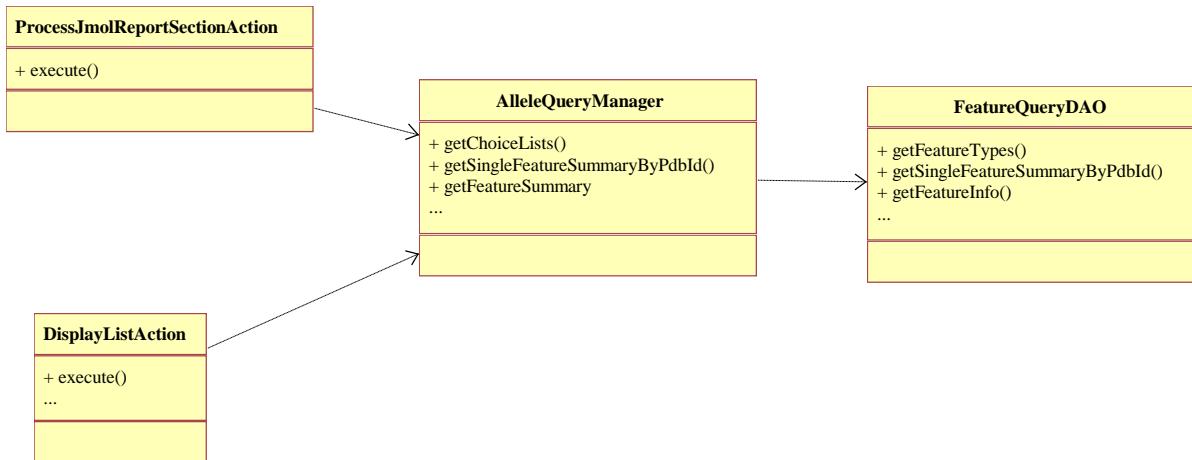
17.8.1 MSFSV Query Class Diagrams



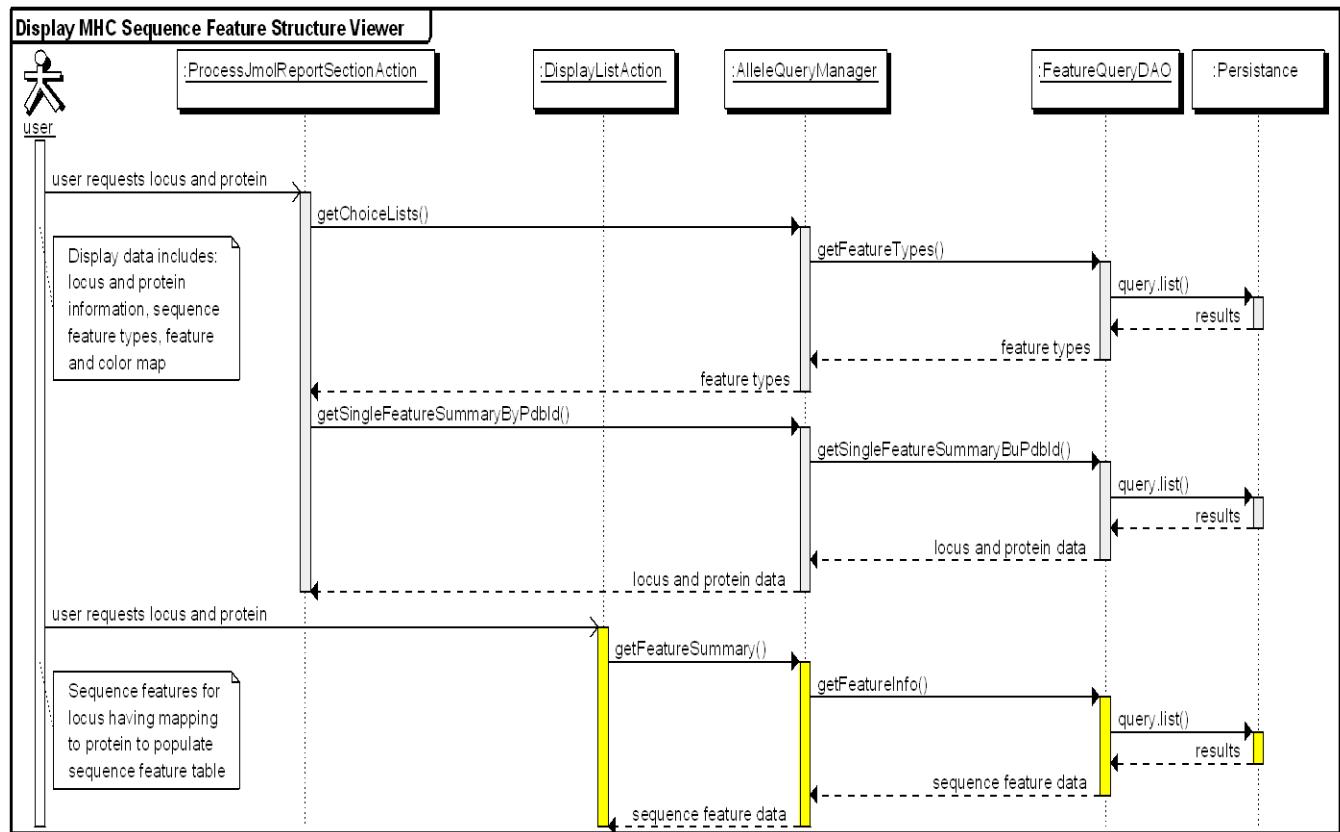
17.8.2 MSFSV Query Sequence Diagrams



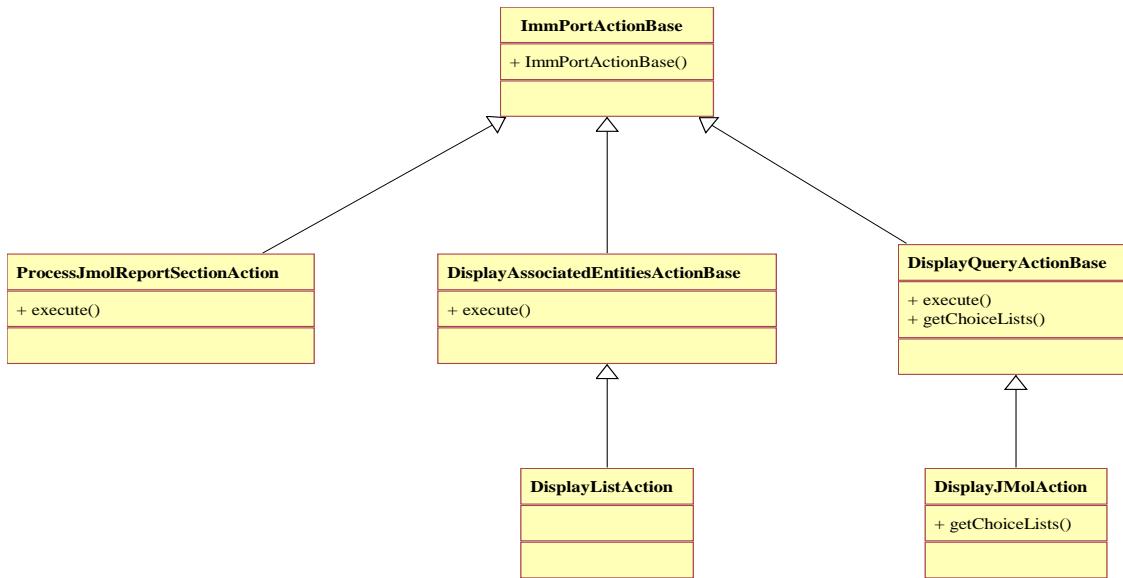
17.8.3 MSFSV Class Diagrams



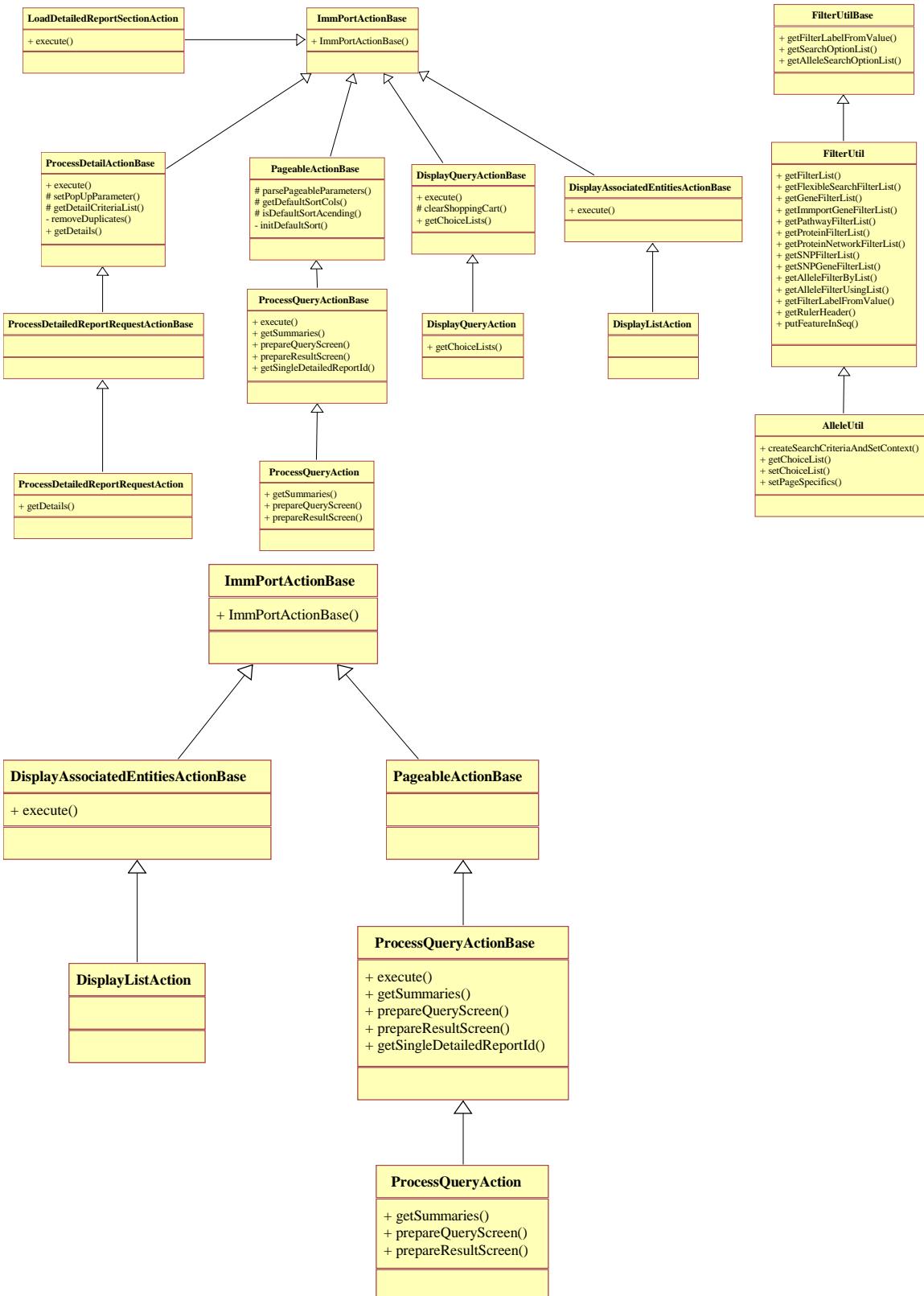
17.8.4 MSFSV Sequence Diagrams



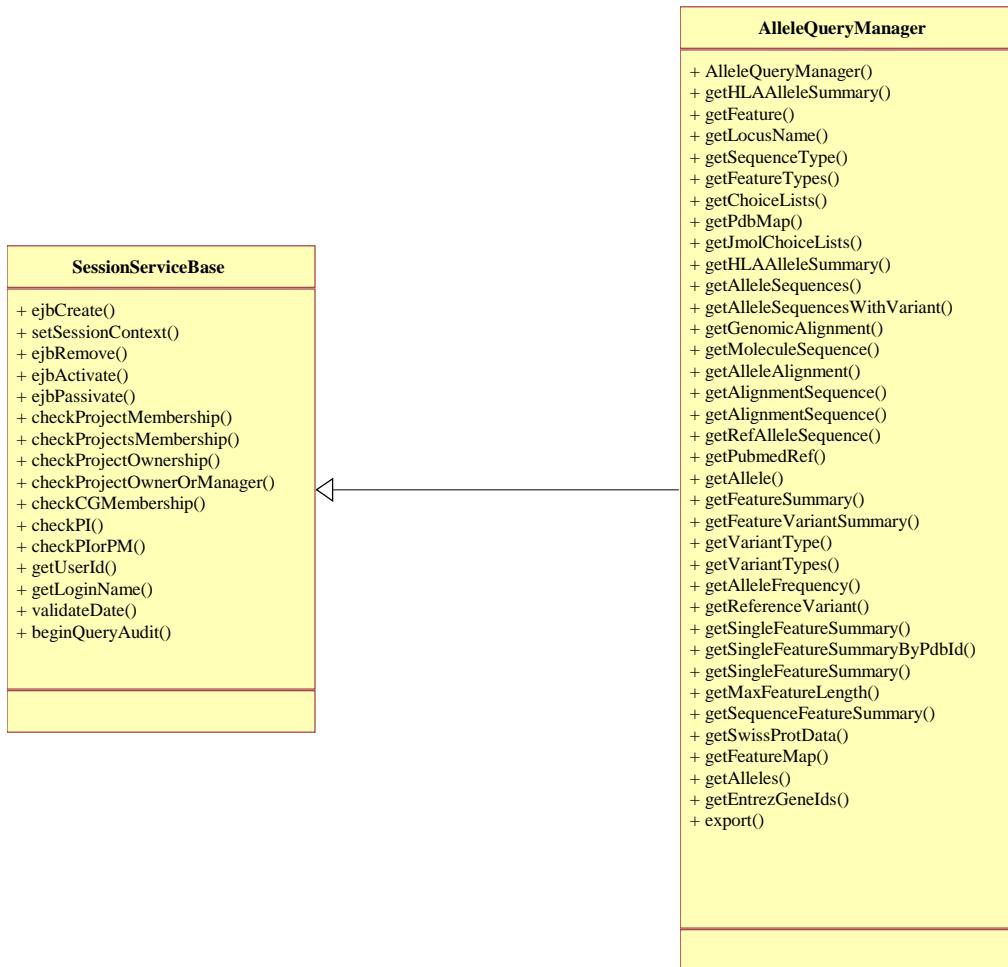
17.8.5 Actions



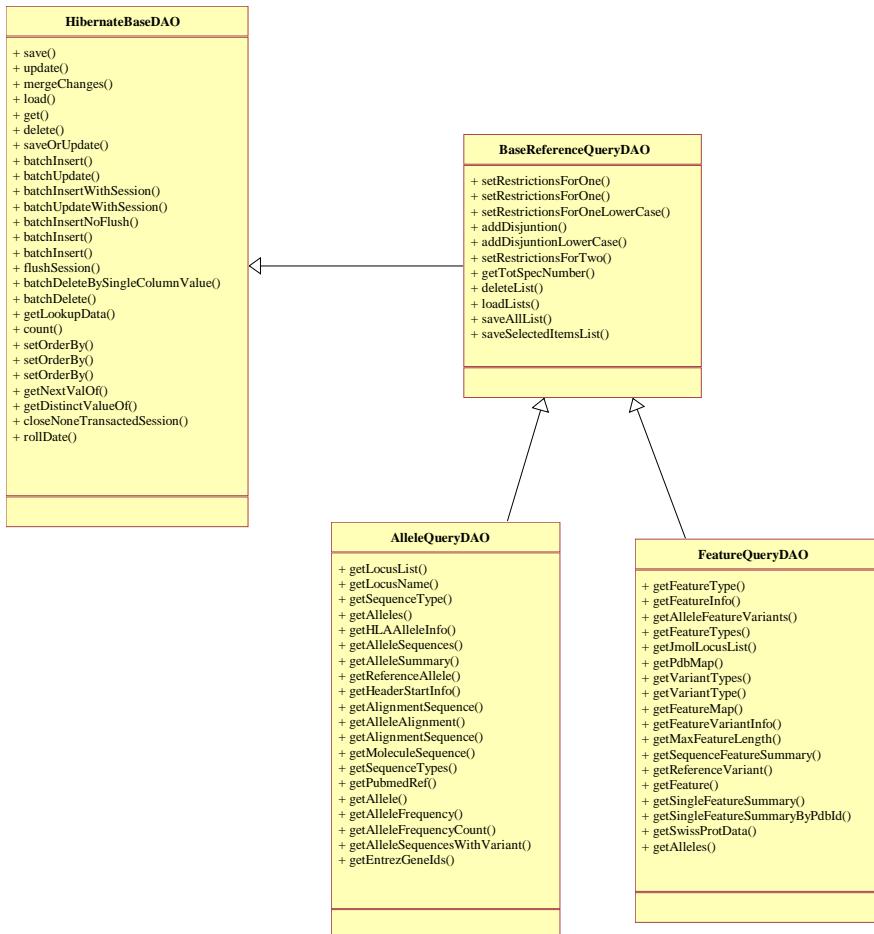
17.8.6 MSFSV Actions



17.8.7 MSFSV EJB



17.8.8 MSFSV DAO



18.0 FLOW ANALYSIS DESIGN PACKAGES

Flow cytometry analysis defines and quantifies cellular features making it an essential tool employed in immunology research. Challenges inherent in the management and interpretation of flow cytometry data led to the development of the ImmPort Flow Cytometry Analysis tool and its four supporting modules:

1. Data Management—single/multiple file upload, data set creation/editing
2. FCS file analysis using the FLOCK algorithm
3. Cross Sample analysis
4. Viewing/editing analysis results

The Data Management module allows the user to upload individual or multiple flow cytometry files, edit the descriptive content of the files and create/edit sets of uploaded files. To further facilitate data management and data set creation ImmPort employs components of the search module to find files of interest.

The FLOCK (FLOW Clustering without K) algorithm was developed as a novel method to identify unique populations within a single flow cytometry file. Analysis output includes two dimensional images of FLOCK-defined populations enhanced by the inclusion of the following tables for viewing or download: results summary, centroid, mean fluorescence intensity. (FLOCK publication: <http://www.ncbi.nlm.nih.gov/pubmed/20839340>).

Cross Sample analysis automatically maps populations of cells across multiple flow cytometry files utilizing centroid values saved from previous FLOCK results. Analysis results are displayed in two-dimensional format for each file in the comparison allowing the user to select which population/fluorescent marker to view across all files. Population proportions are available for all files in the comparison analysis as are results for individual files including centroid and mean fluorescence intensities.

Viewing and editing analysis results applies to FLOCK and individual results in Cross Sample Comparison. The Result Adjustment System (RAS) allows the user to edit two-dimensional images via moving, splitting or deleting a centroid for improved representation of the analyzed results. Supporting statistical output is recalculated to support the analysis edits.

The ImmPort Flow Cytometry analysis workflow is depicted below:

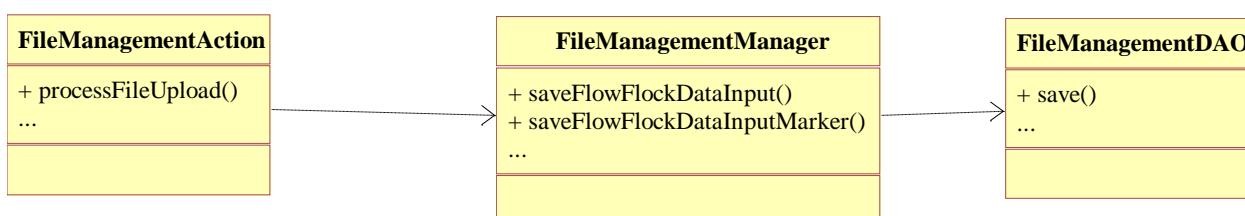


18.1 DATA MANAGEMENT: FILE UPLOAD

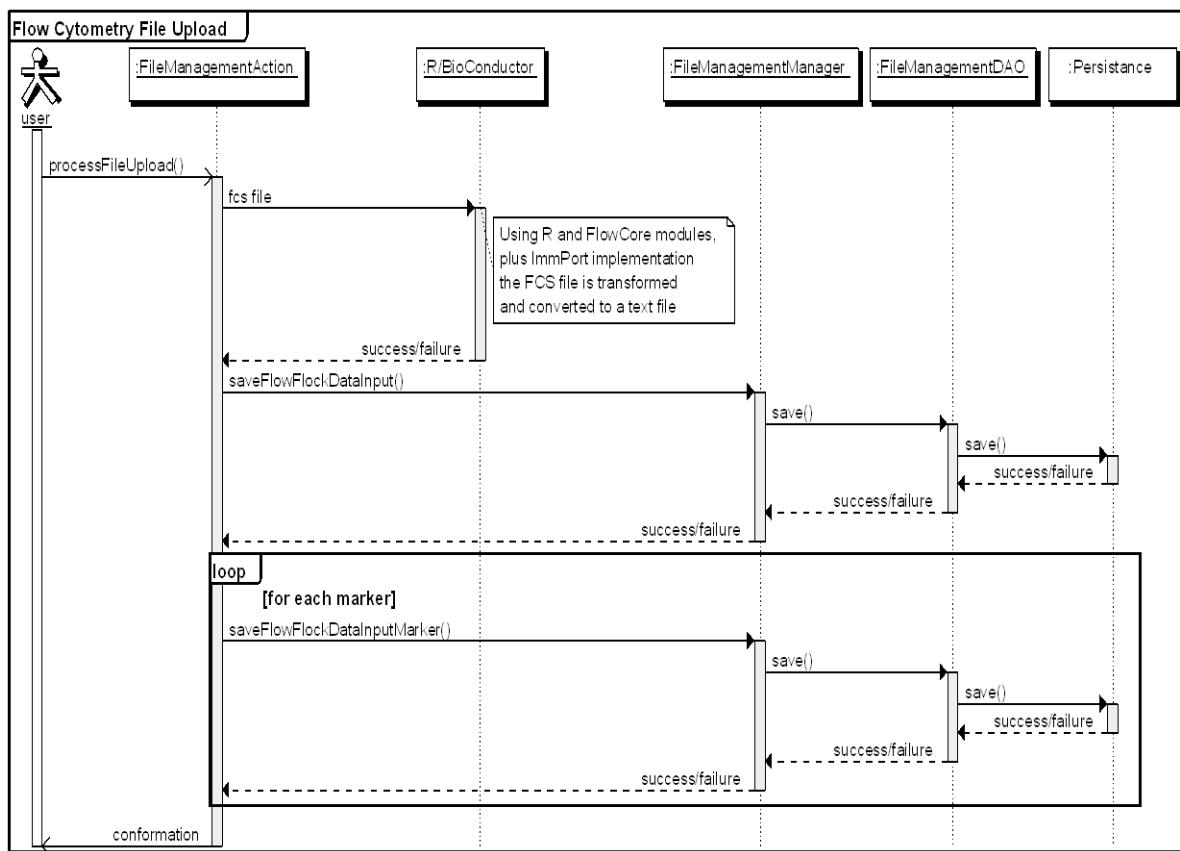
The output from a flow cytometer is typically a binary file that conforms to the Flow Cytometry Standard (FCS). Most analysis algorithms including FLOCK require the FCS binary file be converted and possibly transformed into a text file representation of the fluorescence expression values. The ImmPort Data Management File Upload module support both single and multiple FCS file upload. Uploaded files are automatically converted, transformed and stored in the ImmPort database. The conversion algorithm, FCSTrans, was written using R/BioConductor modules by the BISC team as described in <https://www.immport.net/help/ImmPort.FCS.conversion.pdf>. The ImmPort system supports linking of FCS files to metadata such as subject, study and experimental information, previously loaded into the ImmPort system.

The screenshot shows the ImmPort Data Management File Upload page. At the top, there is a navigation bar with links for 'Edit Profile' and 'Sign Out'. Below the navigation bar, the page title is 'Flow Cytometry Analysis (Beta) / Data Management / File Upload'. A sub-navigation bar includes 'Flow Cytometry', 'Data Management', 'FLOCK', 'View/Edit Results', 'Cross Sample', and 'Help'. A message at the top states: 'A project is required to upload data. Please see the [User Guide](#) or contact the [help desk](#) for more information.' Below this, a dropdown menu allows selecting a project to store uploaded files, with 'Test research project for v2.8' selected. A note below the dropdown explains file formats: 'The uploaded data file can be in either .fcs or .txt format. The filename can't contain the following characters: " / \ and :. ImmPort automatically converts .fcs files to .txt files when the upload includes only .fcs files. The .txt files can be created using third party tools including Tree Star FlowJo™ on Mac OS. For conversion details, please see the [help section](#).' There are fields for 'Name' and 'Description', and browse buttons for 'Data File (.fcs)', 'Data File (.txt)', and 'Marker File (.info)'. To the right of these fields are download links for example files: 'example Data File (.fcs)', 'example Data File (.txt)', and 'example Marker File (.info)'. A note at the bottom states: 'NOTE: Clicking "Upload" will upload the above files into the private project workspace of your chosen project and create the data set.' An 'Upload' button is located at the bottom left.

18.1.1 Upload Class Diagram



18.1.2 Upload Sequence Diagram



18.2 DATA MANAGEMENT: CREATE ANALYSIS SET

The Data Management module facilitates the creation of analysis sets by employing existing search module code and technologies outlined in the search module portion of this document. Created analysis sets may be used as input to either FLOCK or Cross Sample Analysis while existing data sets may be edited to add or remove FCS files. The UI consists of 3 panels:

- Search—used to query for available FCS files
- Results—displays search results
- Create Set—files of interest are moved to this panel to save the set to the database.

The meta information for the analysis sets is stored in the ImmPort data base.

Flow Cytometry Analysis (Beta) / Data Management / Create Analysis Set

Flow Cytometry | Data Management | FLOCK | View/Edit Results | Cross Sample | Help

Use the 'Uploaded Flow Files Search' section to find data files for dataset creation.
The number of items in a dataset is limited to 300.

Select project: Test research project for v2.8 Include Semi-Public files in the Search

Uploaded Flow Files Search

Search for Uploaded Files : Results per page : 25 Submit Cancel

Flow: Name Like

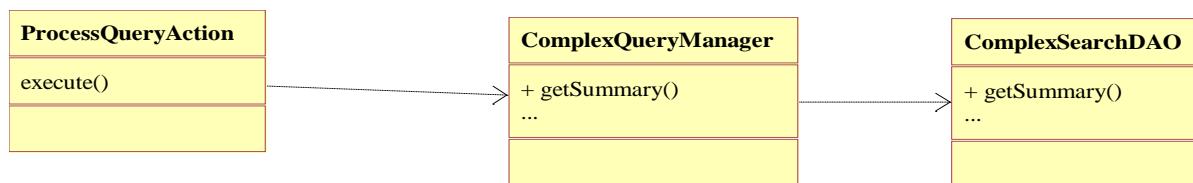
Name	Description	Status	FCS Text File	Date Create
test tcs upload		loaded	FCS002.txt	2010-07-19
20100719_Flock xxx		loaded	FLOCK_Data_Fi	2010-07-19
FCS001		loaded	FCS001.txt	2010-07-07
Test upload with		loaded	MarkerNameChe	2010-07-07
Test new convr to compare to ol		loaded	FCS001.txt	2010-07-06
Test2	single file FCS u	loaded	FCS001.txt	2010-07-02
Test1	single file fcs up	loaded	FCS001.txt	2010-07-02
test upload1	image generatio	loaded	FCS001.txt	2010-07-01

Use arrow button to move files in or out of analysis set.
Mouse-over column titles, click down arrow to hide or show columns.

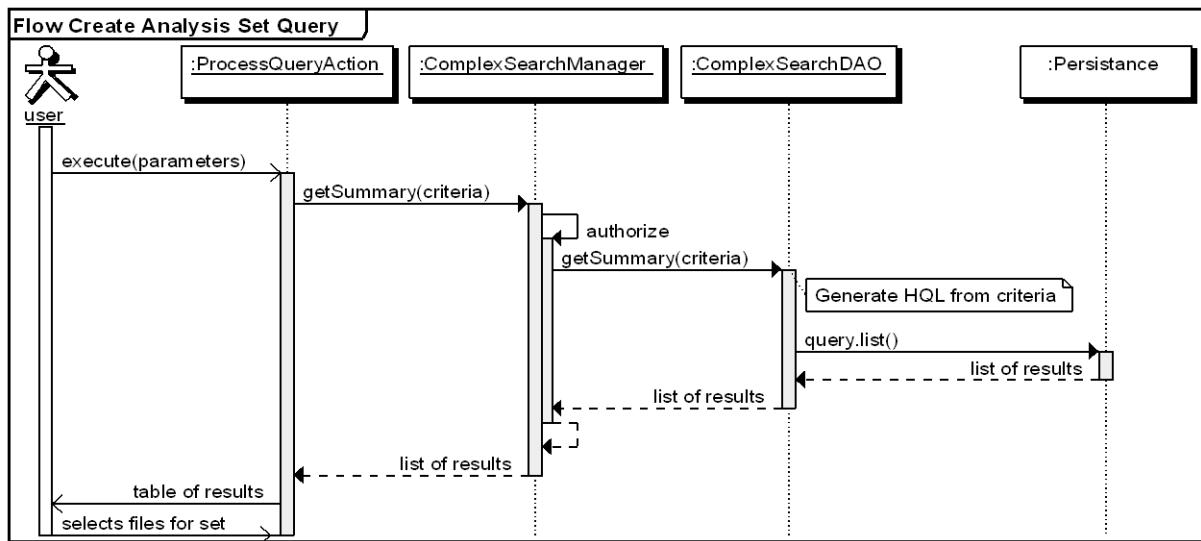
Create Set

Name	FCS Text File
test tcs upload	FCS002.txt

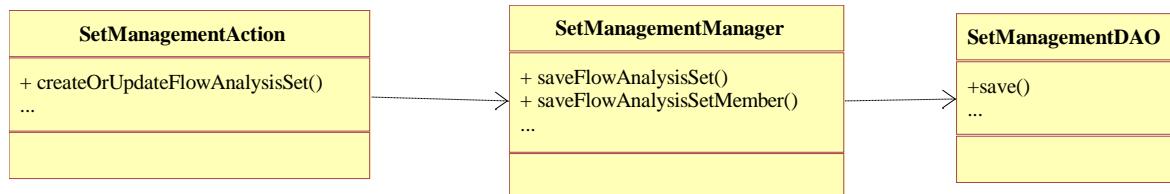
18.2.1 Create Analysis Set Query Class Diagram



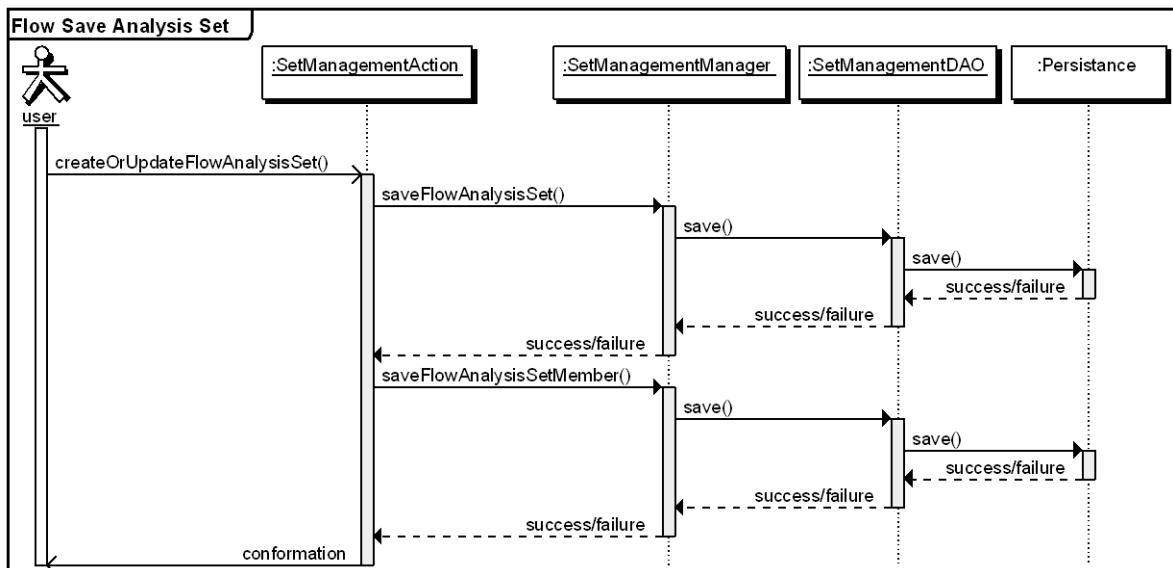
18.2.2 Create Analysis Set Query Sequence Diagram



18.2.3 Create Analysis Set Save Class Diagram



18.2.4 Create Analysis Set Save Sequence Diagram



18.3 FLOCK ANALYSIS

FLOCK analysis uses a rapid binning approach to determine the number of unique populations in high dimensional flow cytometry data. The user interface provides two means to initiate a FLOCK analysis:

- Individual file selection—single or multiple files can be selected for FLOCK analysis from the Uploaded Flow Files Search screen accessed via FLOCK/Analyze Individual Files dropdown menu
- Data set selection—previously created data sets can be selected for FLOCK analysis via FLOCK/Analyze Data Sets dropdown menu

From either analysis initiation point the process for running FLOCK requires the same steps: create an analysis task with the data set or selected individual files, select analysis settings, place the task in a queue, the FlockWorker retrieves the task and a Java program is called for the following:

1. Extracting the text version of the FCS file from the database and storing it in a task folder.
2. The FLOCK program written in ‘C’ reads the extracted FCS file and generates detailed and summary results representing the cell populations.
3. The result files are processed by the Java program generating overview images.

The FlockWorker process marks the task as completed making the results available for user review and subsequent editing.

18.3.1 FLOCK analysis: individual file selection

Name	Description	Exp Sample Acc	Status	FCS Text File	Date Created	Created By
Sub2Sample1			loaded	Sub2Sample1.txt	2010-09-21	Thomson
Sub2Sample1			loaded	Sub2Sample1.txt	2010-09-21	Thomson
Treg pure			loaded	DD082009_017.txt	2010-09-21	Thomson
Teff pure			loaded	DD082009_015.txt	2010-09-21	Thomson
CD4 purified			loaded	DD082009_012.txt	2010-09-21	Thomson
Spleen-pLN			loaded	DD082009_009.txt	2010-09-21	Thomson
Single file upload test IE8			loaded	Sub1Sample1.txt	2010-09-21	Thomson

18.3.2 FLOCK analysis: data set selection

Name	Description	Status	Project	Date Created
20100923 save all from spw 2	xxxxxx	loaded	Test research project for v2.9.1	2010-09-23
Dataset4	4files 7col	loaded	Test research project for v2.9.1	2010-09-21
Dataset3	save_all_7items	loaded	Test research project for v2.9.1	2010-09-21
Dataset2	4 items 7 col	loaded	Test research project for v2.9.1	2010-09-21
Dataset1	2 items 6 col	loaded	Test research project for v2.9.1	2010-09-21
20100921 SPW set	xxxxxx	loaded	Test research project for v2.9.1	2010-09-21

18.3.3 FLOCK Analysis Setting

Name	Marker #	1	2	3	4	5	6
Sample 001	6	Forward	Side	FITC	PE	Cy5	Cy7
Sample 003	6	Forward	Side	FITC	PE	Cy5	Cy7
Sample 002	6	Forward	Side	FITC	PE	Cy5	Cy7

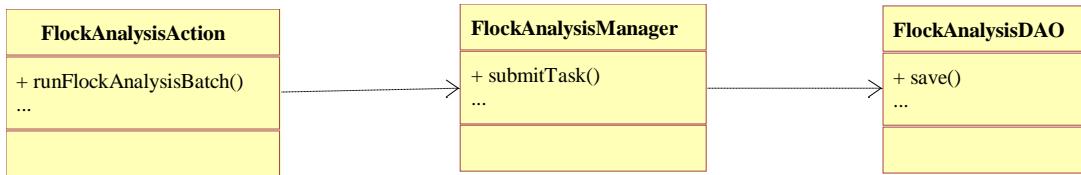
18.3.4 FLOCK: Analysis Status

Your FLOCK Analysis request was submitted successfully. Your task ID is **2619**.
You will be sent an email when your task has been completed.

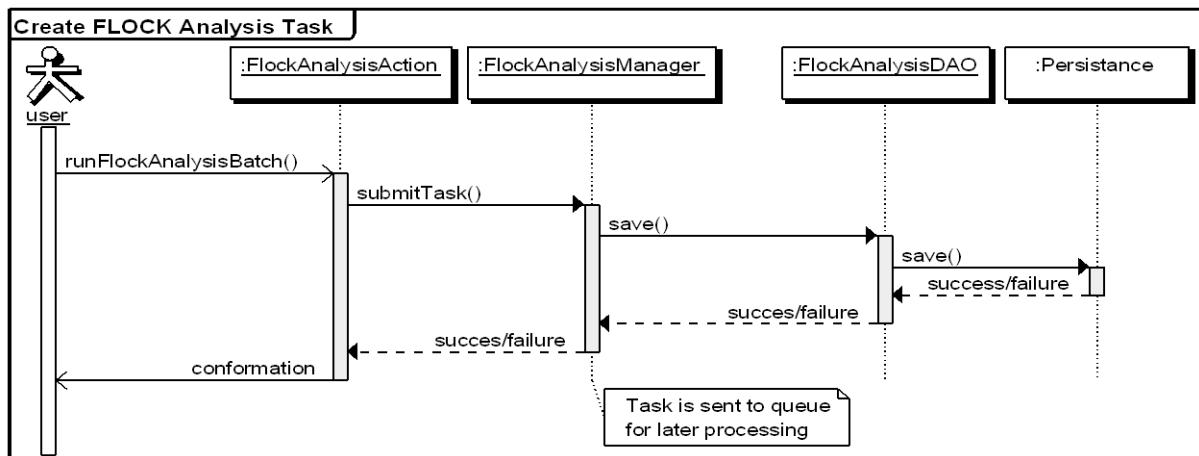
Click the "Show Analysis Results" button to proceed to the FLOCK Analysis History page.

Show Analysis Results

18.3.5 Submit FLOCK Analysis Task Class Diagram



18.3.6 Submit FLOCK Analysis Task Sequence Diagram



18.4 CROSS SAMPLE ANALYSIS

Cross Sample analysis automatically maps populations of cells across multiple flow cytometry samples and computes the summary statistics for downstream analysis. The analysis requires a previously created data set and saved centroid file which are presented for user selection in the Cross Sample Analysis UI. The initiating step for Cross Sample Analysis is the creation of an analysis task comprised of the data set and centroid which is followed by a review of the analysis components via Cross Sample Analysis Settings display. The task is placed in a queue where the CrossSampleWorker process picks up the task and calls a Java program for the following:

1. Extracts the text version of the FCS files from the database and stores it in a task folder.
2. The centroid adjustment program written in 'C' reads the extracted FCS file and generates detailed and summary results representing cell populations.
3. The result files are processed by the Java program generating overview images.

At the end of the CrossSampleWorker process the task is marked as completed providing the user with access to results which include summary statistics and marker by population two-dimensional displays.

18.4.1 Cross Sample Analysis: Data set and centroid selection

Flow Cytometry Analysis (Beta) / Cross Sample / Run Analysis

Flow Cytometry | Data Management | FLOCK | View/Edit Results | Cross Sample | Help

Select project: Test research project for v2.9.1

Choose Centroid file

Name	Description	Status	Project	Date Created
Cent_16_pop		loaded	Test research project for v2.9.1	2010-09-30
Centroid3		loaded	Test research project for v2.9.1	2010-09-21
Centroid2		loaded	Test research project for v2.9.1	2010-09-21
Centroid1		loaded	Test research project for v2.9.1	2010-09-21

Choose One or More Analysis Sets

Name	Description	Status	Project	Date Created
20100923 save all from spw 2	xxxxxx	loaded	Test research project for v2.9.1	2010-09-23
Dataset4	4files 7col	loaded	Test research project for v2.9.1	2010-09-21
Dataset3	save_all_7items	loaded	Test research project for v2.9.1	2010-09-21
Dataset2	4 items 7 col	loaded	Test research project for v2.9.1	2010-09-21
Dataset1	2 items 6 col	loaded	Test research project for v2.9.1	2010-09-21
20100921 SPW set	xxxxx	loaded	Test research project for v2.9.1	2010-09-21

Continue

18.4.2 Cross Sample Analysis: Analysis Setting

Flow Cytometry Analysis (Beta) / Cross Sample / Cross Sample Analysis Setting

Flow Cytometry | Data Management | FLOCK | View/Edit Results | Cross Sample | Help

Cross Sample

Project: Test research project for v2.9.1

Name:

Description:

Marker Selection

Show Assigned Name | Show Original Name | Help

Type	Name	Marker #	1	2	3	4	5	6	7
Centroid	20101021 test1	7	FSC-H	SSC-H	FL1-H	FL2-H	FL3-H	FL2-A	FL4-H
Flow File	Spleen-pLN	7	FSC-H	SSC-H	FL1-H	FL2-H	FL3-H	FL2-A	FL4-H
Flow File	Teff pure	7	FSC-H	SSC-H	FL1-H	FL2-H	FL3-H	FL2-A	FL4-H
Flow File	CD4 purified	7	FSC-H	SSC-H	FL1-H	FL2-H	FL3-H	FL2-A	FL4-H
Flow File	Treg pure	7	FSC-H	SSC-H	FL1-H	FL2-H	FL3-H	FL2-A	FL4-H

18.4.3 Cross Sample: Analysis Status

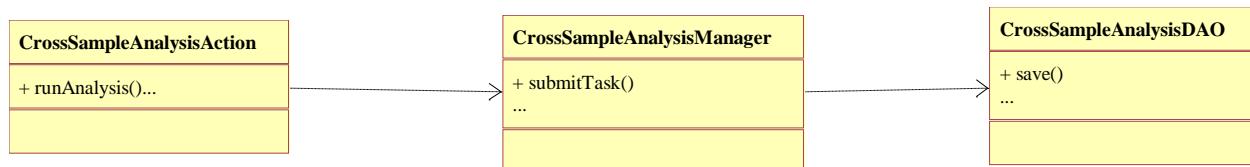
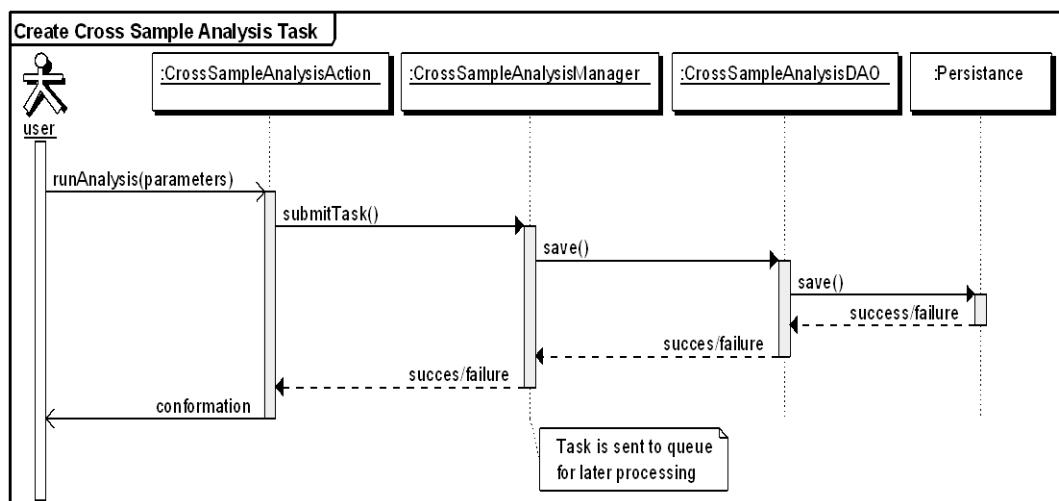
Flow Cytometry Analysis (Beta) / Cross Sample / Analysis Status

Flow Cytometry | Data Management | FLOCK | View/Edit Results | Cross Sample | Help

Your Cross Sample Analysis request was submitted successfully. Your task ID is **2618**.

You will be sent an email when your task has been completed.

Click the "Show Analysis Results" button to proceed to the Cross Sample Analysis Detail page.

Show Analysis Results**18.4.4 Submit Cross Sample Analysis task Class Diagram****18.4.5 Submit Cross Sample Analysis task Sequence Diagram**

18.5 FLOCK ANALYSIS RESULTS

Results of FLOCK analysis tasks are displayed in a tabular grid. Each row in the grid represents one task and contains summary information such as name, description, date completed, etc.

Selecting Detail provides greater task detail displayed in two panels

- FLOCK Analysis Task Detail—task ID, name, description, etc.
- FLOCK Analysis Result Files—individual FCS file results with a link to individual results detail

The individual file results contain several panels enabling the user to review both statistics and two dimensional images generated by the analysis.

18.5.1 FLOCK Analysis Results: tabular view

The screenshot shows a web-based application interface for 'Flow Cytometry Analysis (Beta) / FLOCK / Analysis History'. At the top, there's a navigation bar with links for 'Flow Cytometry', 'Data Management', 'FLOCK', 'View/Edit Results', 'Cross Sample', and 'Help'. Below the navigation is a dropdown menu labeled 'Select project:' with the value 'Test research project for v2.8'. A note says 'Mouse-over column titles, click down arrow to hide or show columns.' The main area is titled 'FLOCK Analysis History' and displays a table with 8 columns: Details, Name, Description, Task Type, Version, Status, Run By, Start Date, and End Date. There are 20 rows of data, each with a 'Detail' link under the 'Details' column. The data includes various test names like '20100831 test 2', '3 column test', and 'test TT3454', along with their respective details and run dates.

Details	Name	Description	Task Type	Version	Status	Run By	Start Date	End Date
Detail	20100831 test 2		flock	2	completed	Yuen	2010-08-31	2010-08-31
Detail	3 column test		flock	1	completed	Thomson	2010-07-07	2010-07-07
Detail	3 column test		flock	1	completed	Thomson	2010-07-07	2010-07-07
Detail	test TT3454	xxxx	flock	1	completed	Yuen	2010-07-07	2010-07-07
Detail	test edited marker r		flock	1	completed	Thomson	2010-07-07	2010-07-07
Detail	test overview displ		flock	1	completed	Thomson	2010-07-07	2010-07-07
Detail	20100707 TT2073t run run testapp2		flock	1	completed	Yuen	2010-07-07	2010-07-07

18.5.2 FLOCK Analysis Results: Detail display

The screenshot shows the ImmPort software interface. At the top, there is a navigation bar with links for 'Edit Profile' and 'Sign Out'. Below the navigation bar, the main content area displays 'Flow Cytometry Analysis (Beta) / FLOCK / Analysis Result'. The page includes a breadcrumb trail: 'Flow Cytometry' > 'Data Management' > 'FLOCK' > 'View/Edit Results' > 'Cross Sample' > 'Help'. A section titled 'FLOCK Analysis Task Detail' provides specific task information:

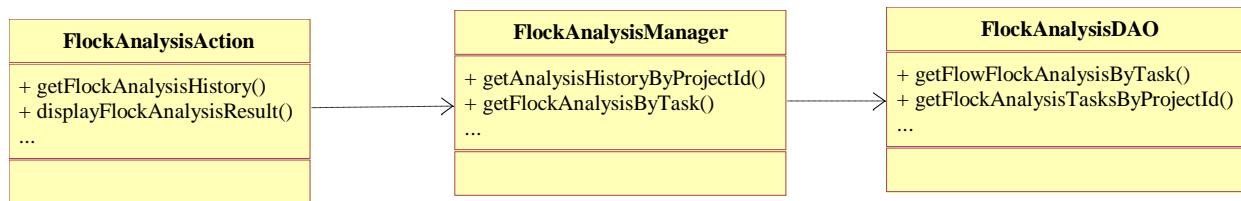
Task ID	1719
Name	test TT3454
Description	xxxx
Algorithm Version	1.0
Run By	Yuen
Start Date	07/07/2010 15:21:01
End Date	07/07/2010 15:21:28
Status	Completed

Below this, a section titled 'FLOCK Analysis Result Files' contains a table listing three samples:

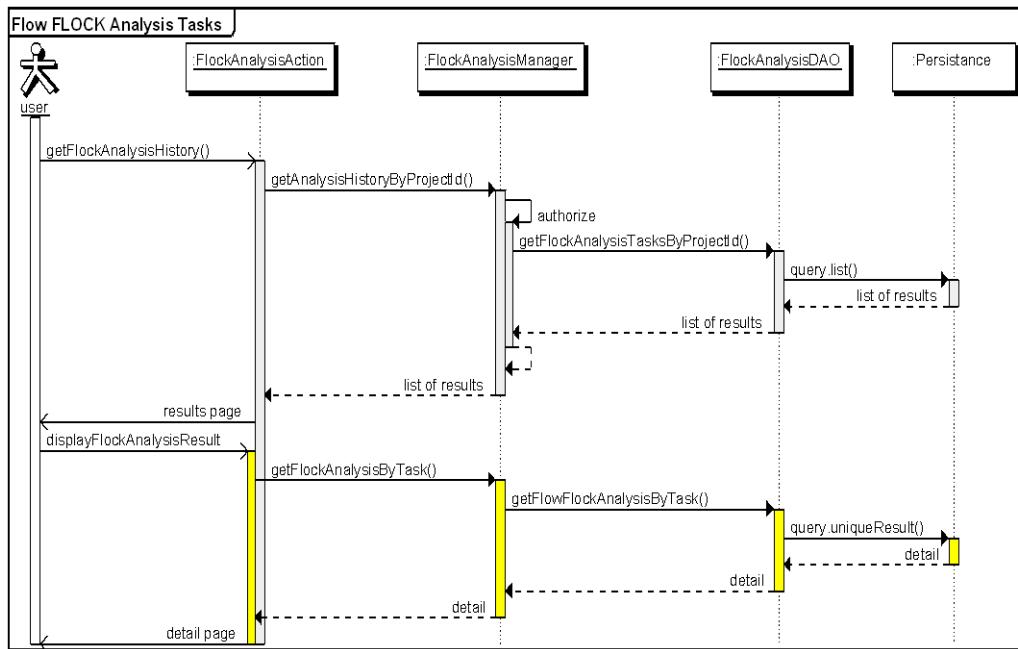
Name	Description	Exp Sample Acc	FCS Text File	Bins	Density	Populations	Status
Sample 001	Multiple file upload example file 1 ES118544		FCS001.txt	52	3	30	Results
Sample 002	Multiple file upload example file 2 ES118545		FCS002.txt	42	3	28	Results
Sample 003	Multiple file upload example file 3 ES118546		FCS003.txt	38	3	27	Results

At the bottom of the table, it says 'Displaying 1 - 3 of 3'.

18.5.3 FLOCK Analysis Results Class Diagram



18.5.4 FLOCK Analysis Results Sequence Diagram



18.6 CROSS SAMPLE ANALYSIS TASK RESULTS

The results of Cross Sample Analysis tasks are displayed in a tabular grid similar to that for FLOCK. Each row in the grid represents one task and contains summary information such as name, description, date completed, etc. Selecting Detail provides greater task detail displayed in three panels:

- Task detail—task ID, name, description, etc.
- Flow Analysis Sets—displays all sets in the analysis
- Cross Sample Analysis Result Files—displays all FCS files in the analysis

Selection of Analysis Results in the task detail panel links to the population proportion table and the marker-by-population two dimensional displays. The Results link within the Cross Sample Analysis Result Files panel links to individual file results.

18.6.1 Cross Sample Analysis: tabular view

A screenshot of a software interface titled "Flow Cytometry Analysis (Beta) / Cross Sample / Analysis History". The top navigation bar includes links for Flow Cytometry, Data Management, FLOCK, View/Edit Results, Cross Sample, and Help. A dropdown menu shows "Test research project for v2.9.1". Below the navigation is a message: "Mouse-over column titles, click down arrow to hide or show columns." The main area is titled "Cross Sample Analysis History" and displays a table with the following data:

Details	Name	Description	Task Type	Status	Run By	Start Date	End Date
Detail_2598			cross_sample_analys	completed	Thomson	2010-09-30	2010-09-30
Detail_2533			cross_sample_analys	completed	Thomson	2010-09-21	2010-09-21

18.6.2 Cross Sample Analysis: Detail display

Flow Cytometry Analysis (Beta) / Cross Sample / Analysis Detail

Flow Cytometry | Data Management | FLOCK | View/Edit Results | Cross Sample | Help

Cross Sample Analysis

Task ID	2533
Name	CrossSample1
Description	4 files 7 col
Run By	Thomson
Start Time	09/21/2010 15:17:33
End Time	09/21/2010 15:18:13
Status	Analysis Results

Centroid Information Detail

Name	Centroid3
Description	

Flow Analysis Sets

Displaying 1 - 1 of 1		
Name	Description	Status
Dataset4	4files 7col	loaded

Cross Sample Analysis Result Files

Displaying 1 - 4 of 4			
Name	Description	FCS Text File	Status
CD4 purified		DD082009_012.txt	Results
Spleen-pLN		DD082009_009.txt	Results
Teff pure		DD082009_015.txt	Results
Treg pure		DD082009_017.txt	Results

18.6.3 Cross Sample Analysis: Marker by Marker display

Flow Cytometry Analysis (Beta) / Cross Sample / Analysis Result

Percentage Table | Marker by Population | Download |

Cross Sample Analysis - Task Name: CrossSample.

Select a population and markers to view Cross Sample analysis.

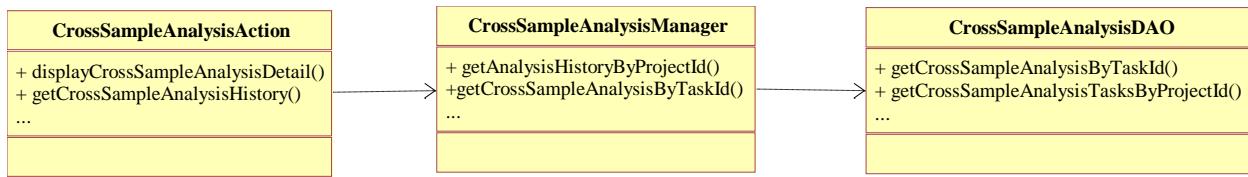
Population: 5 Marker 1: Side Marker 2: PE

Display

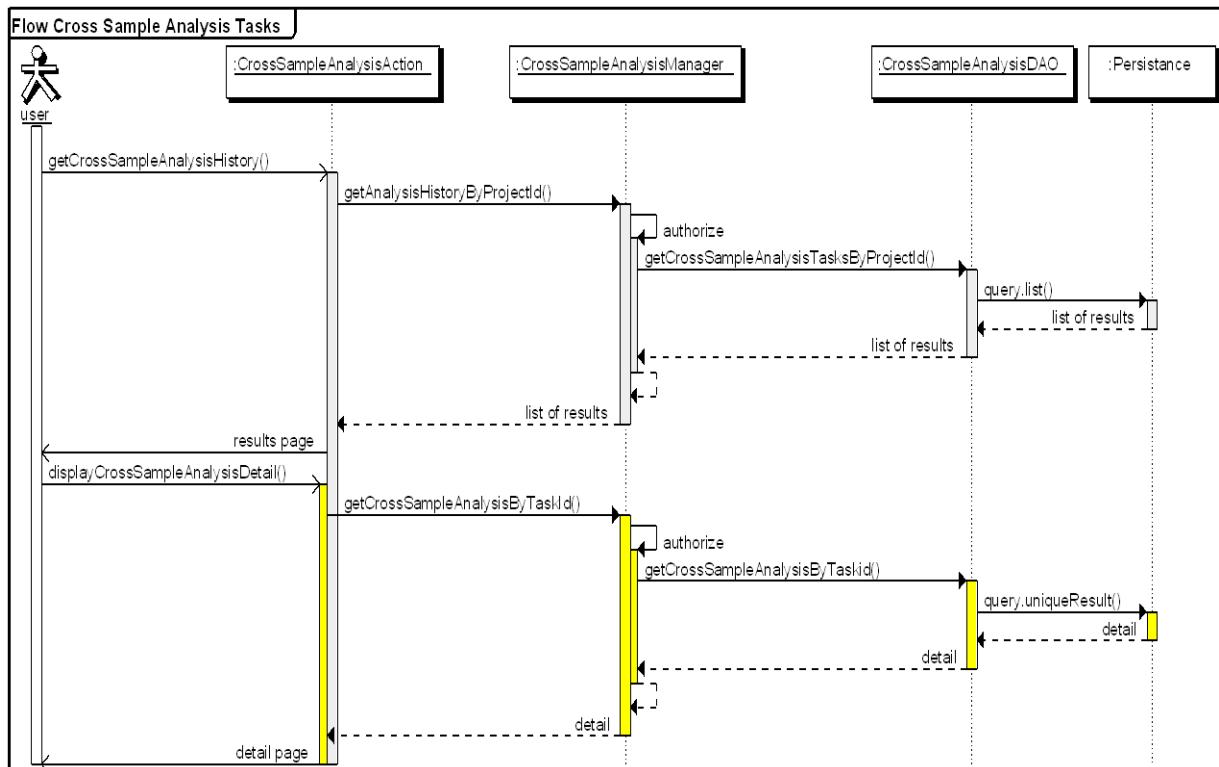
Population: 5 Marker 1: Side Marker 2: PE

File: FCS(002)	File: FCS-002	File: FCS,002	File: FCS-002	File: FCS,002
File: FCS002, 003, 004@@	File: FCSfile001	File: FCSfile002	File: FCSfile003	File: FCSfile004

18.6.4 Cross Sample Analysis Task Results Class Diagram



18.6.5 Cross Sample Analysis Task Results Sequence Diagram



18.7 FLOCK RESULTS VISUALIZATION

The FLOCK analysis Detail result screen includes 6 displays:

- Overview
- Summary Tables
- Download Tables
- Download Results
- Centroid
- 3D View

Each of the six sections will be reviewed but given the importance of image generation for the Overview and Result Adjustment sections a brief image generation overview is required.

FLOCK analysis produces a tab-delimited text file where rows are events and columns are markers—an exception is the last column which represents the population assigned to the event. The output file is transformed into binary format to facilitate real time image generation. UI requests for images results in a JAVA module reading the binary file, computing the coordinates and drawing dot-plot images in PNG format. Images are written to the file system under the analysis task folder where they are retrieved by the UI.

18.7.1 Overview

FLOCK results Overview displays a table of images where each cell represents the dot-plot of the populations assigned based on the expression values for two markers. On the right hand side of the screen there is a panel for selecting which populations to display, and summary information for each population. The user can select/deselect populations using the check box and then click on the Update button to change the display of the images to only include the selected populations. Each image may be selected—this action takes the user to the Result Adjustment screen.



18.7.2 Result Adjustment

Result Adjustment supports editing results as a means to fine-tune the centroid locations. Images will be regenerated when Save Changes is selected. In addition to moving a centroid the system supports splitting and deleting centroids.



18.7.3 Summary Tables

Three summary tables are available for each FLOCK analysis and include:

- Results—populations appear as table rows, channels/markers as columns with the population proportion presented in the last column
- Centroid—each cell of this table represents the center for the population/marker combination
- Mean Fluorescent Intensity (MFI)—each cell represents the mean fluorescence value.

18.7.4 Download Tables

The content of the download tables is the same content that is displayed in the Summary Tables mentioned above. In this case the content can be downloaded in text file format for further processing.

18.7.5 Download Results

Result files available for download include:

- Results table including event number and population assignment

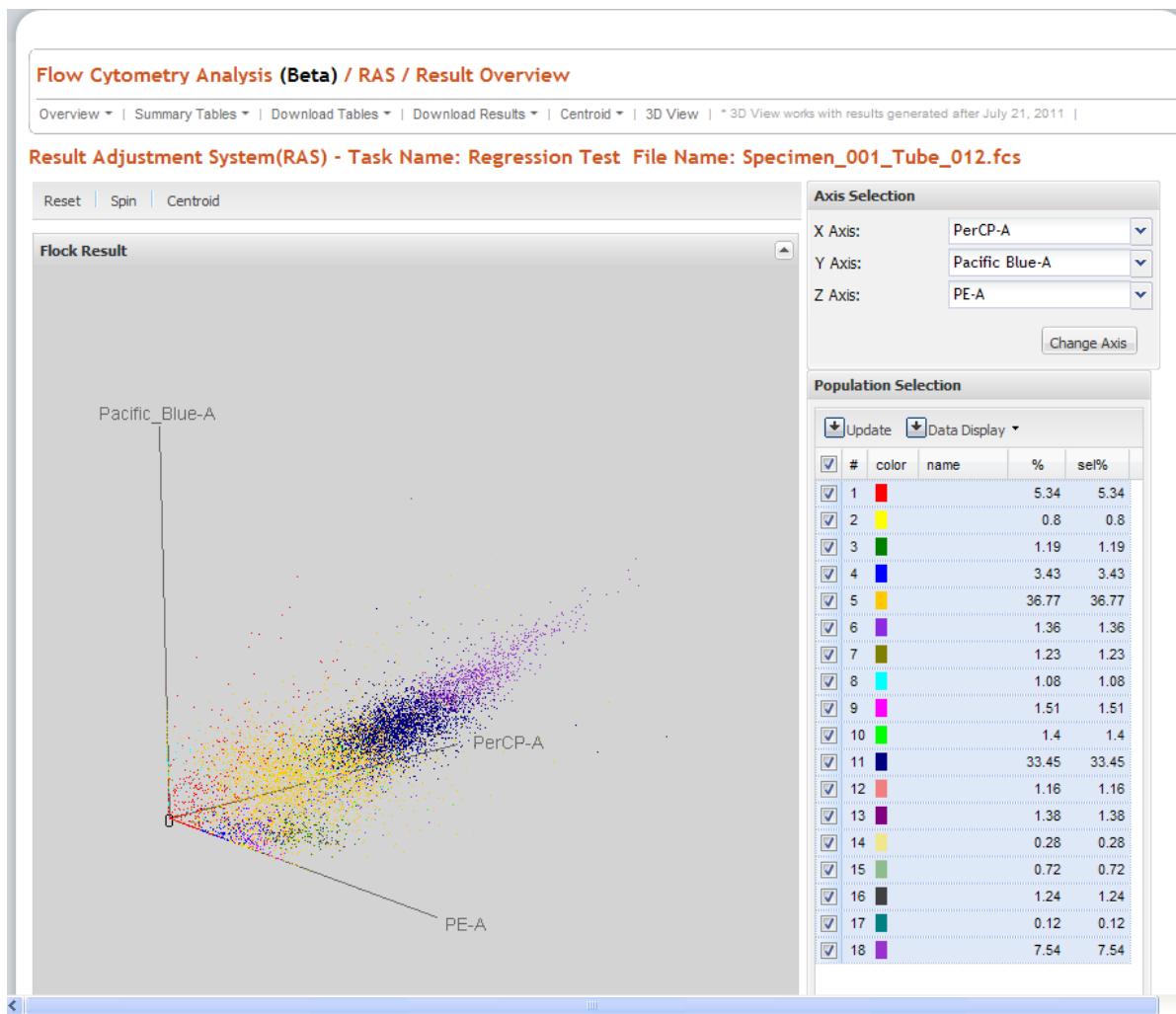
- A package of files representing all the result files.

18.7.6 Centroid

From this panel the user can select to save the results editing from Result Adjustment or undo any previous adjustments.

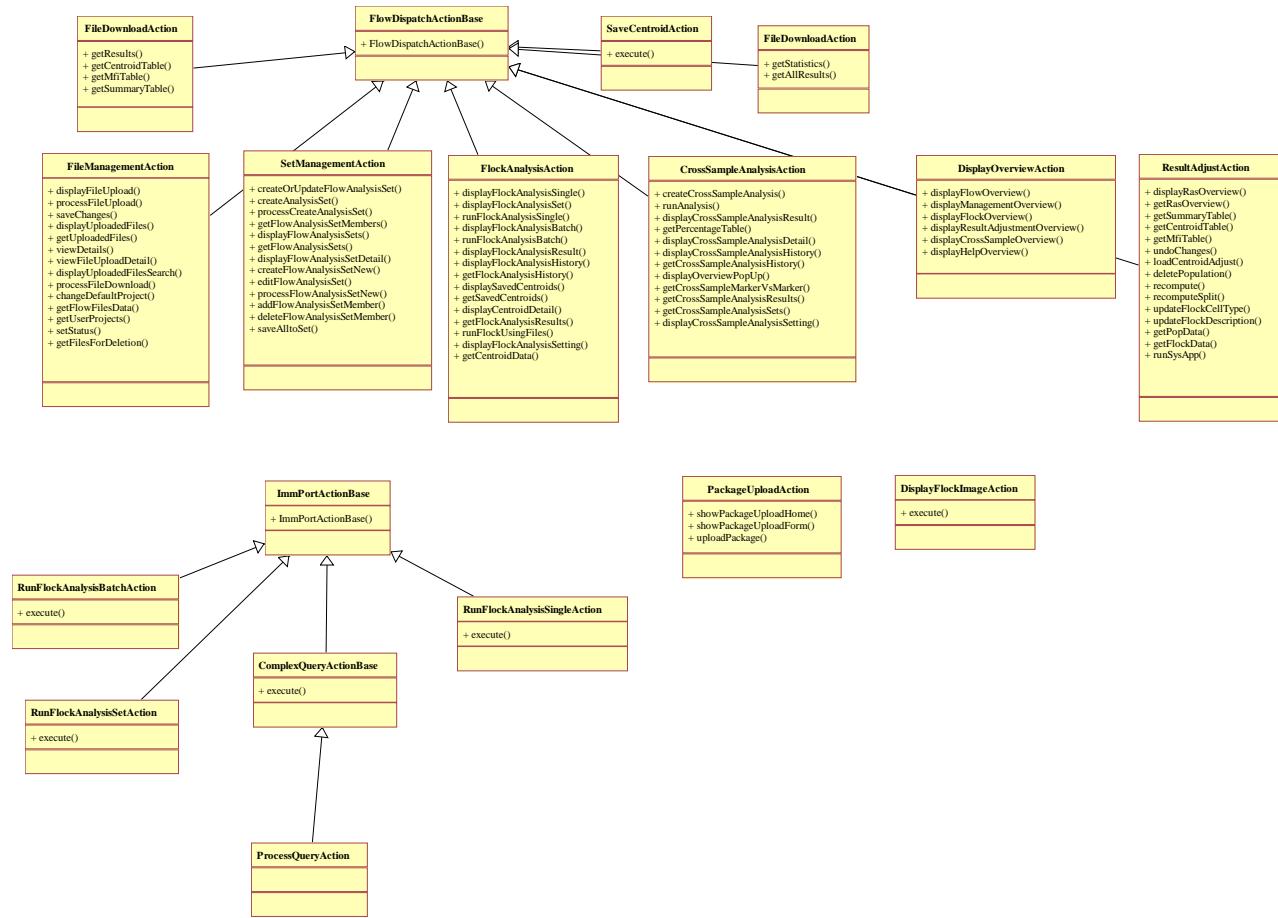
18.7.7 3D View

In this view the results for 3 of the markers can be displayed in a 3D view using the JMOL applet. On the right side panel, the user can select which markers to display and the user can choose which populations to display.

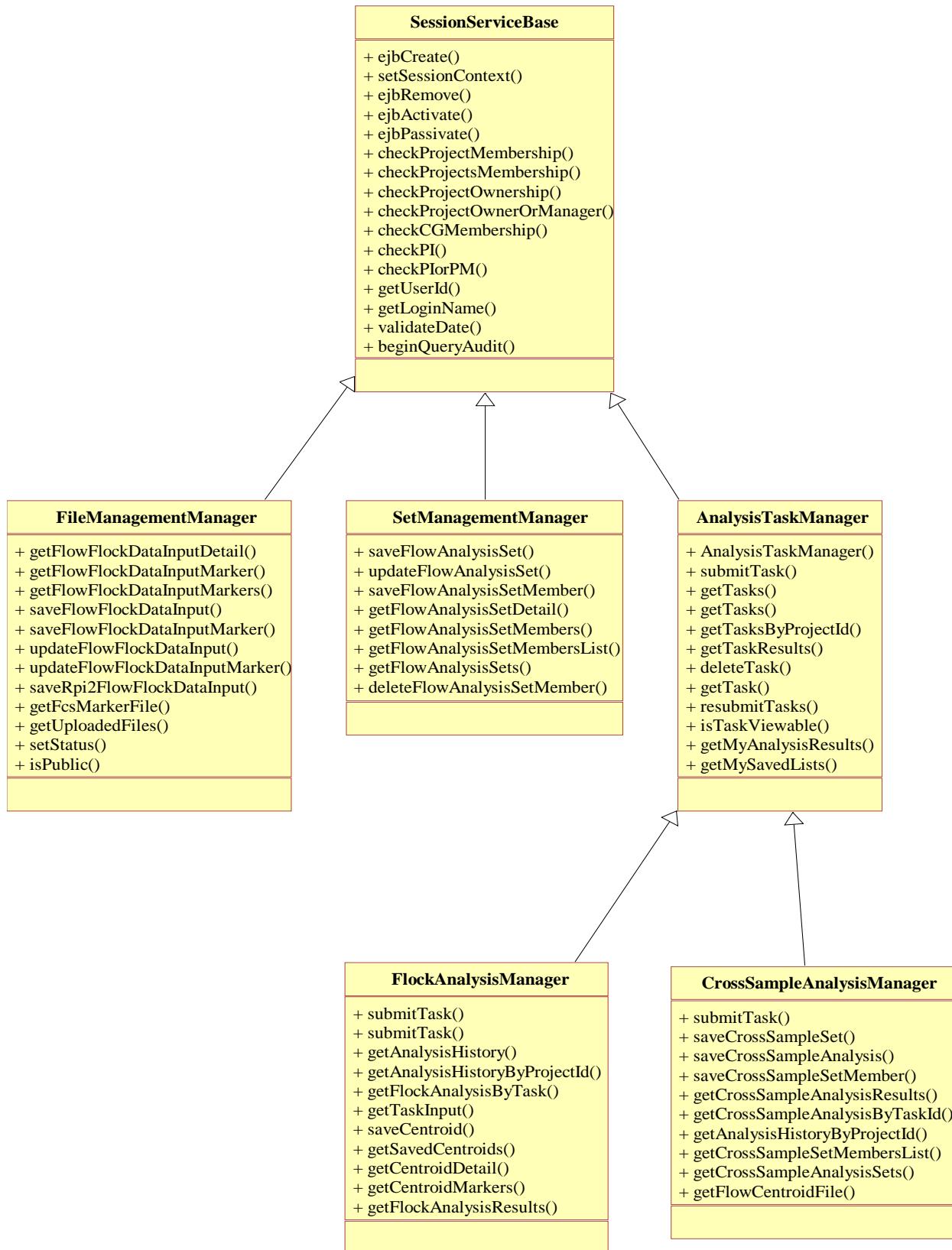


18.8 FLOW ANALYSIS CLASS DIAGRAMS

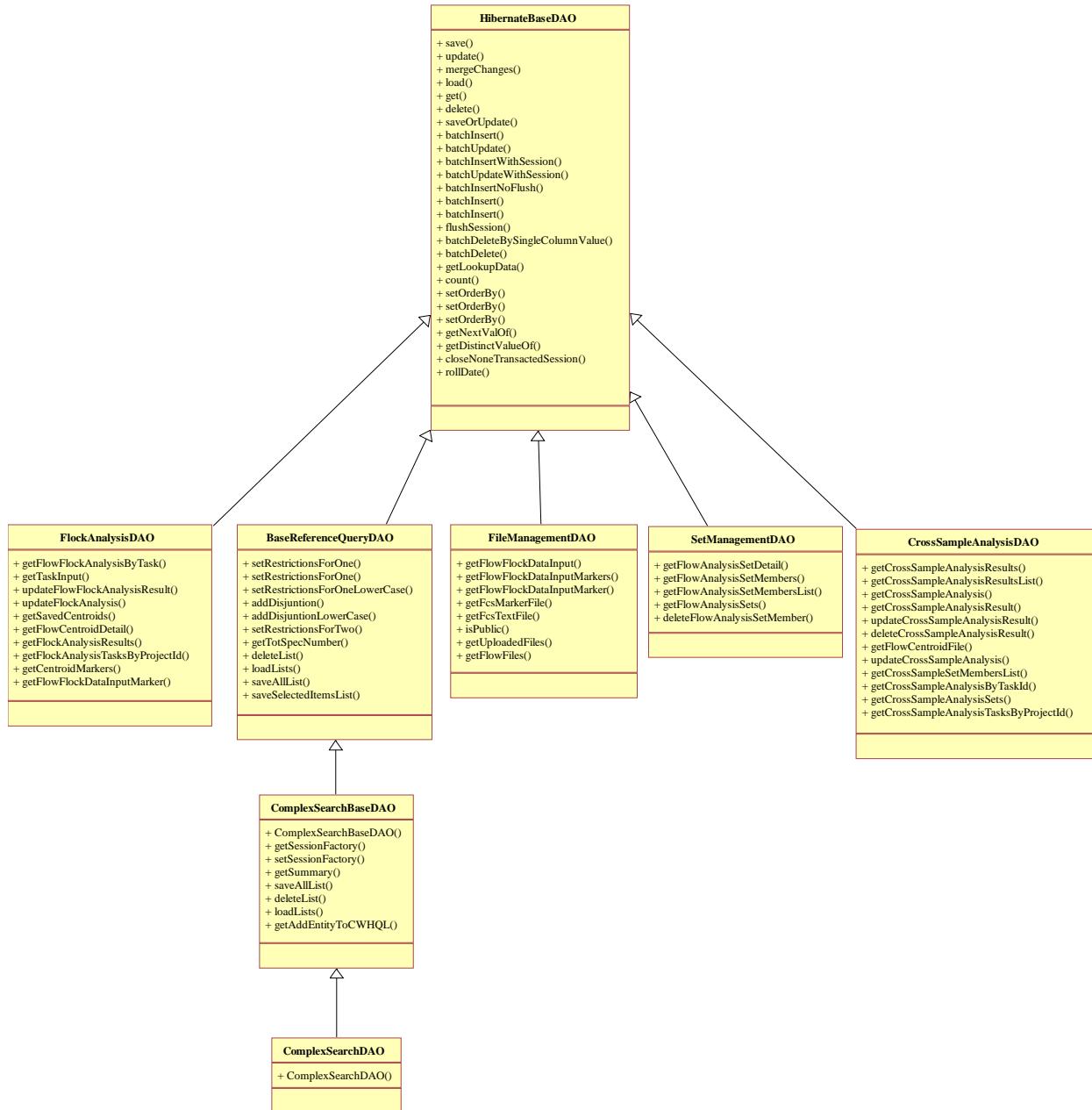
18.8.1 Actions



18.8.2 EJB



18.8.3 DAO



19.0 HLA ANALYSIS DESIGN PACKAGES

The MHC validation and sequence feature vector type (SFVT) analysis support system in ImmPort is separated into 3 modules:

1. Upload a HLA typing file from your computer
2. Create a an HLA data set from ImmPort Data
3. Re-analyze a HLA data set previously created in ImmPort
4. Run SFVT analysis

Each module is intended to provide a mechanism for specifying the input file to the MCH validation and analysis support system. Each module provides the following set of actions (workflows) for a specified input file:

1. Validation of alleles in the file
2. Validation and generation of SFVT vector files
3. Validation and allelic and genotypic ambiguity reduction
4. Run the SFVT analysis module
5. Validation and run the Pypop quality control tool.

When an action is requested by the user, the ImmPort system queues a hlavt pipeline job to run in the background on an ImmPort application server. When the job completes, it stores its results into the ImmPort Oracle database and the user is notified by e-mail that the job has completed and ready for viewing. Section 6 provides an outline of the hlavt pipeline job. Below is the screen shot illustrating the home page for the three modules.

Analysis / MHC Validation and SFVT Analysis / Home - Beta Release ⓘ

Home | Upload Data | Create Data Set | Analyze Data Set | SFVT Analysis | Analysis Results | User Guide | Ambiguity Reduction Slides / Logic

MHC SFVT Analysis (Sequence Feature Variant Type Analysis), is a tool to generate the sequence feature variant type vector for one or more subjects across any of the classical HLA loci.

Select data source:

Upload a HLA Typing File From Your Computer

Create a HLA Data Set From ImmPort Data

Re-analyze a HLA Data Set Previously Created in ImmPort

Run SFVT Analysis

Next

Sequence features (SF) have been defined for each classical HLA protein based on structural information (e.g. beta-sheet 1), functional information (e.g. peptide antigen binding), and polymorphism. These sequence features can be overlapping and continuous or discontinuous in the linear sequence. The extent of sequence variation among HLA alleles was then assessed for each HLA sequence feature to define all variant types (VT) found in the human population, [D. R. Karp et. al.](#) The SFVT Generation Tool takes HLA typing data and converts the identified HLA alleles into their component SFVTs. Please refer to the [User Guide](#) for detailed instruction on MHC SFVT Analysis use.

As this is the v2.0 Beta release of ImmPort MHC SFVT Analysis, we would very much appreciate feedback on any problems you may encounter and any advice you might have regarding future improvement. Please contact the help desk for questions and comments regarding the [MHC SFVT Analysis v2.0 Beta Release](#). Thank you.

19.1 UPLOAD AN HLA FILE FROM YOUR COMPUTER

This module provides the mechanism for specifying an HLA typing file on your computer ('select HLA typing file') to be used in the chosen action. This file needs to be in HLA typing result format either as a tab-separated text file or 2003 Excel worksheet file. This file will be uploaded into the ImmPort Oracle database and used by the hlavt pipeline job that is queued for the requested action. The screen shot shows the upload file page and illustrates the possible actions to be performed on the uploaded file and the set of parameters to be provided by the user. The upload display class and sequence diagrams illustrate how the imgt/release information and the list of projects are acquired. The acquisition of the imgt/hla result information and the list of projects is the same in all modules. The class and sequence diagrams illustrate how a jobs is submitted for background processing.

The screenshot shows a web-based application window titled "Upload & Generate files". At the top, there are buttons for "Display submitted request", "Previous", "Next", and "Cancel". A message at the top states: "The current IMGT/HLA Release is IMGT/HLA Release 3.00 (2010-04-01). The actions below accept both IMGT/HLA version 2.* or version 3.* allele formats and NMDP version 2 and 3 NMDP-code formats." Below this, a section titled "Select the action *" contains several radio button options:

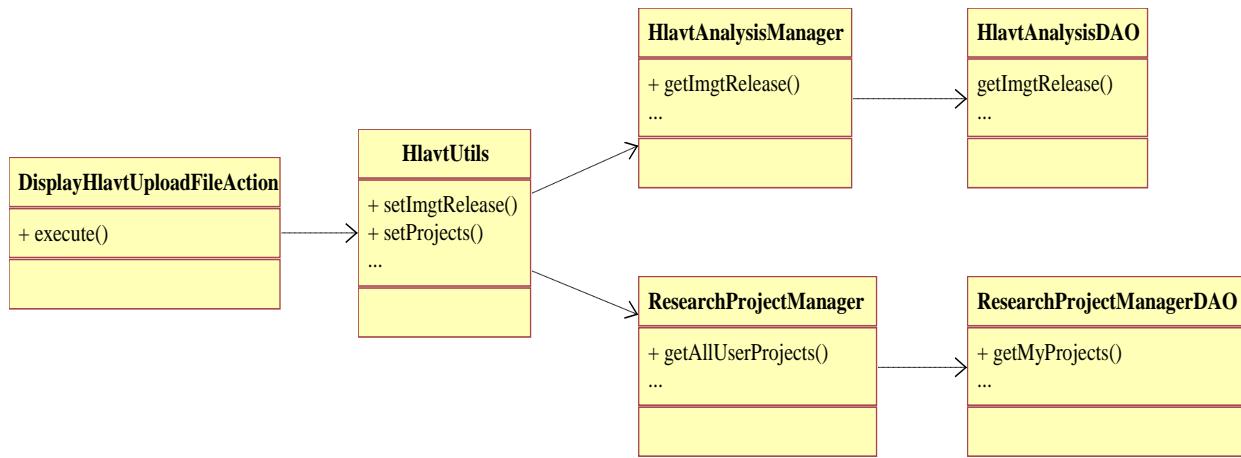
- Validate alleles
Validation uses the [ANTT](#) Tool to validate conformance to [IMGT/HLA version 2 & 3 nomenclature format and G- and P-Codes](#). Also, Validation converts the input file into IMGT/HLA version 3.* format since all options operate on that format.
- Validate and Generate SFVT Vector Files
- Validate and Reduce allele ambiguity
This [tool](#) was designed by Steven J. Mack et. al. and was developed with his invaluable co-operation. The input requires the column "Population Area" that specifies the population area ([Help](#)) associated with each row of data.
- Run [Pypop](#) HLA QC Pipeline
This pipeline validates the input file first and then runs Pypop. Please check out the [Pypop Timings](#) ([Help](#)).

Below this, a section titled "Select the IMGT/HLA Output Version *" contains two radio button options:

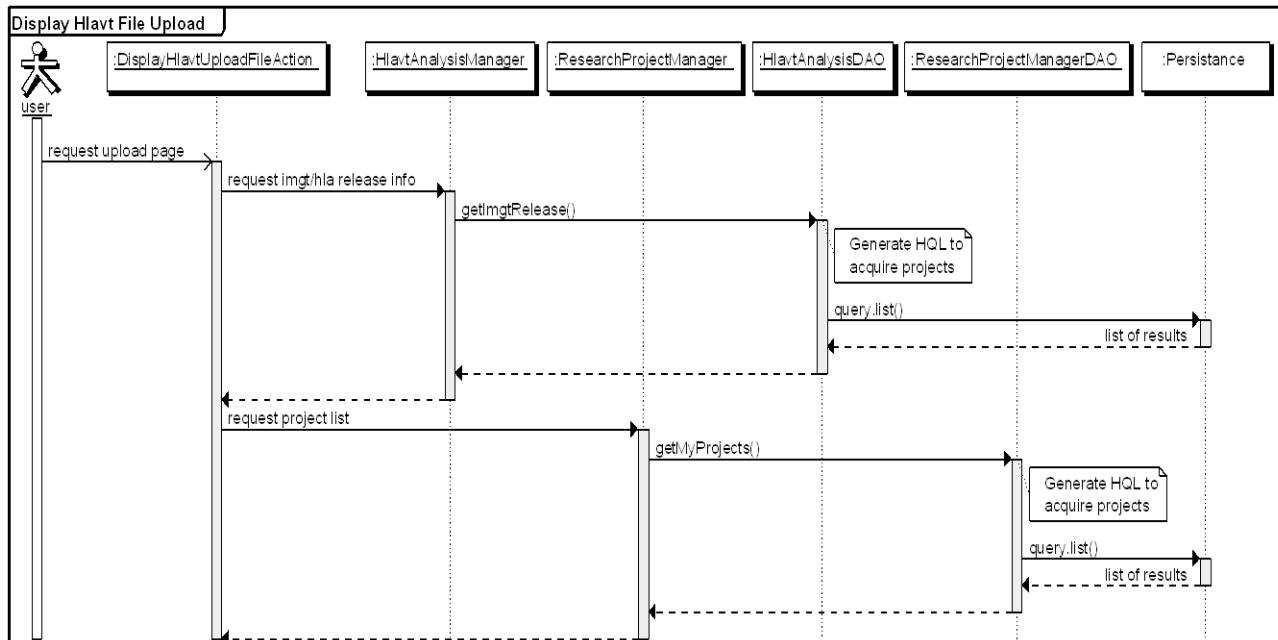
- Generated files will be formatted in [IMGT/HLA version 3.* format](#)
All the options above operate on IMGT/HLA version 3.* format and output all files in that format
- Generated files will be formatted in [IMGT/HLA version 2.* format](#) except for IMGT/HLA G- and P-codes
Conversion from IMGT/HLA version 3.* to version 2.* format uses the [ANTT](#) Tool.

Further down, there are fields for "Select a project to store input and generated files *": "03099: HLA Typing Result upload Error" and a "Browse..." button for "HLA Typing File (.txt or .xls)". Below these are download links for example files: "Download an example HLA Typing Data File (.txt) containing instructive errors", "Download an example HLA Typing Data File (Excel.xls) containing instructive errors", and "Download the HLA Typing Data File Template (Excel.xls)". There are also fields for "Dataset Name *" and "Dataset Descriptor". A note at the bottom states: "NOTE: Clicking "Next" will upload the above file into the private project workspace of your chosen project, execute the action, and generate files." At the very bottom, there is a "Done" button and some small icons.

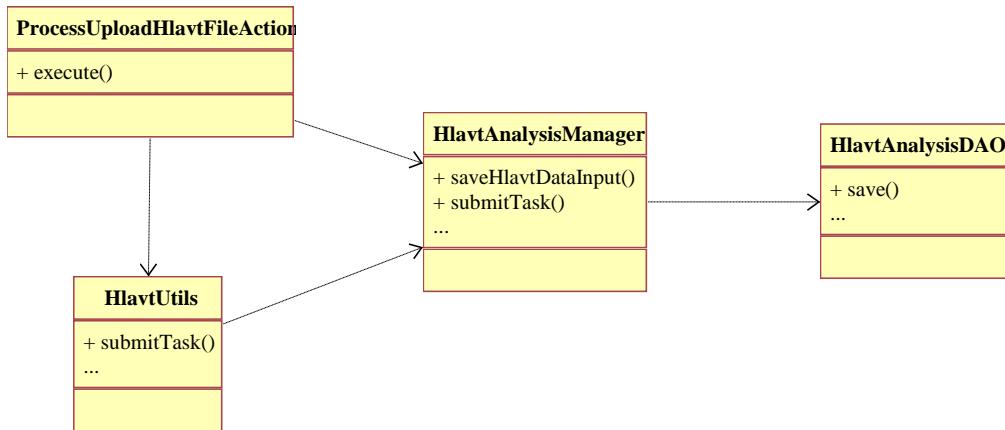
19.1.1 Upload Display Class Diagrams



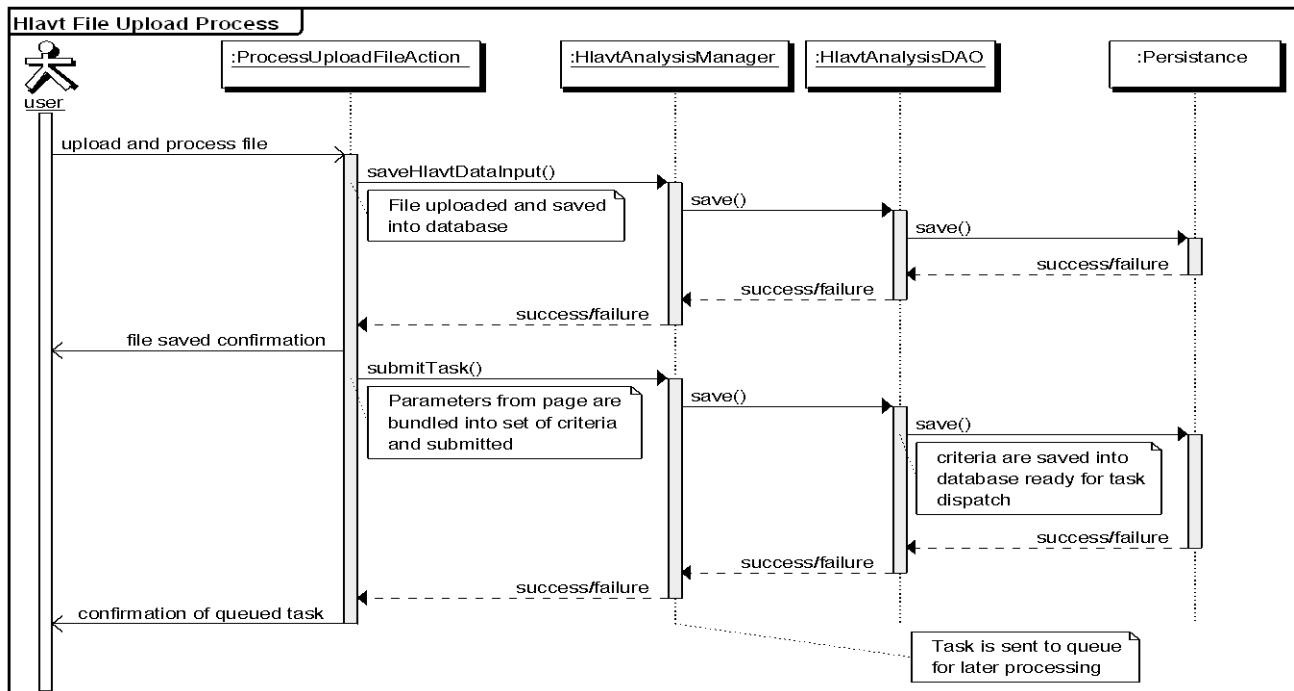
19.1.2 Upload Display Sequence Diagrams



19.1.3 Upload Process Class Diagrams



19.1.4 Upload Process Sequence Diagrams



19.2 CREATE A HLA DATA SET FROM IMMPORT DATA

This module provides the mechanism for using HLA typing results that have been previously loaded into the ImmPort Oracle database using the ImmPort research data batch upload process that is specified elsewhere. The selection page initially has only accessible projects and saved subject lists tables are populated. Once a selection of a set of projects or subject lists is made, then the subject and experiment sample table is populated with results having research data HLA typing results that are stored in ImmPort for the chosen projects or subjects lists. The selection screen shot illustrates the case where a set of subject lists have been chosen.

The screenshot shows a software interface for selecting subjects from ImmPort. At the top, there are three buttons: 'Select Subjects from ImmPort' (highlighted), 'Select HLA Loci', and 'Display submitted request'. Below these are 'Previous', 'Next', and 'Cancel' buttons. The main area is divided into two tables:

Accessible Projects with HLA Typing Results

ProjectID	Project Title	Project Type
3	TESTING: Bioinformatics Integration S	RP
4	Patrick_Dunn_Workspace	RP
17	SFVT	RP
26	example data packages	RP
48	Tom	RP

Saved Subject List having HLA Typing Result

ListID	List Name	List Description	Project ID
113	tsmith test sfvt :	Testing save subjects from pr	3
114	tsmith test sfvt :		3
115	tsmith test sfvt :	Testing save subjects from su	3
653	snp subjects 33	Subjects with SNP 3204 subject	109
654	1500 good subj	These subjects should be all go	109

Subject and Experimental Sample Data Having HLA Typing Results

SubjectOrgAccNum	ExpSampleAccNum	Gender	AffectionStatus	Race	Ethnicity	AffectionPhenotype	OriginalProjTitle
SUB23048	ES27224	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23052	ES27225	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23054	ES27226	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23056	ES27227	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23058	ES27228	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23060	ES27229	Female	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23071	ES27230	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23076	ES27231	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23089	ES27232	Male	Unknown	Unknown			1500 Subjects
SUB23091	ES27233	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects
SUB23097	ES27234	Male	Unknown	Caucasian	Non-Hispa		1500 Subjects

Once the subject and experiment sample table is populated, the user can select a set of subjects/experiment samples or all subjects from this table and proceed to the next page (display page). On the display page, the user sets the parameters requested. The extra parameter on the display page is the selection of the set of HLA loci to use in generating the HLA typing file. Once the set of parameters are chosen in the display page, then the user proceeds to request the action be processed. In this case, the process will generate an HLA typing file from the HLA typing results stored for the set of subjects/experiment samples chosen.

Select the action *

- Validate alleles
Validation uses the [ANTT](#) Tool to validate conformance to IMGT/HLA version 2 & 3 nomenclature format and G- and P-Codes. Also, Validation converts the input file into IMGT/HLA version 3.* format since all options operate on that format.
- Validate and Generate SFVt Vector Files
- Validate and Reduce allele ambiguity
This [tool](#) was designed by Steven J. Mack et. al. and was developed with his invaluable co-operation.
The input requires the column 'Population Area' that specifies the population area ([?](#)) associated with each row of data.
- Run [Pypop](#) HLA QC Pipeline
This pipeline validates the input file first and then runs Pypop. Please check out the '[Pypop Timings](#)' ([?](#)).

Select the IMGT/HLA Output Version *

- Generated files will be formatted in **IMGT/HLA version 3.* format**
All the options above operate on IMGT/HLA version 3.* format and output all files in that format
- Generated files will be formatted in **IMGT/HLA version 2.* format** except for IMGT/HLA G- and P-codes
Conversion from IMGT/HLA version 3.* to version 2.* format uses the [ANTT](#) Tool.

The MHC Validation and SFVt Analysis tool requires a **project: "workspace"**
The **Dataset Name** is used to name the HLA Typing File generated in this workflow (with a ".txt" suffix).
Please see the [User Guide](#) or contact the [help desk](#) for more information.

Select a project to store generated files *

Dataset Name *

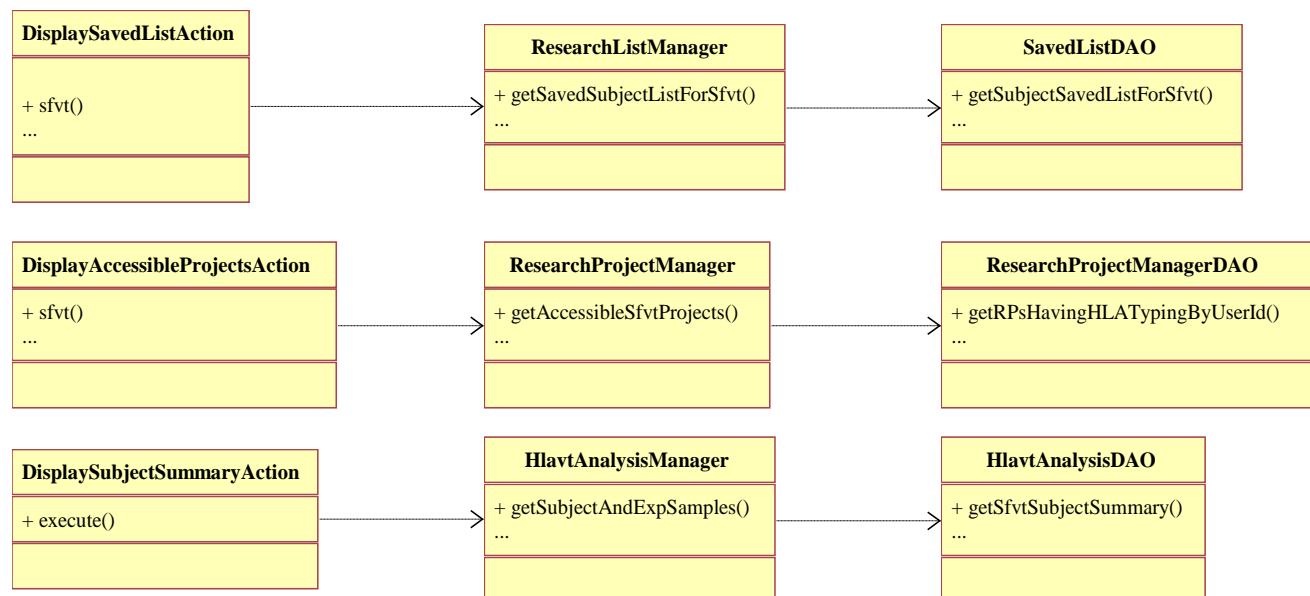
Dataset Descriptor

Select Loci: *

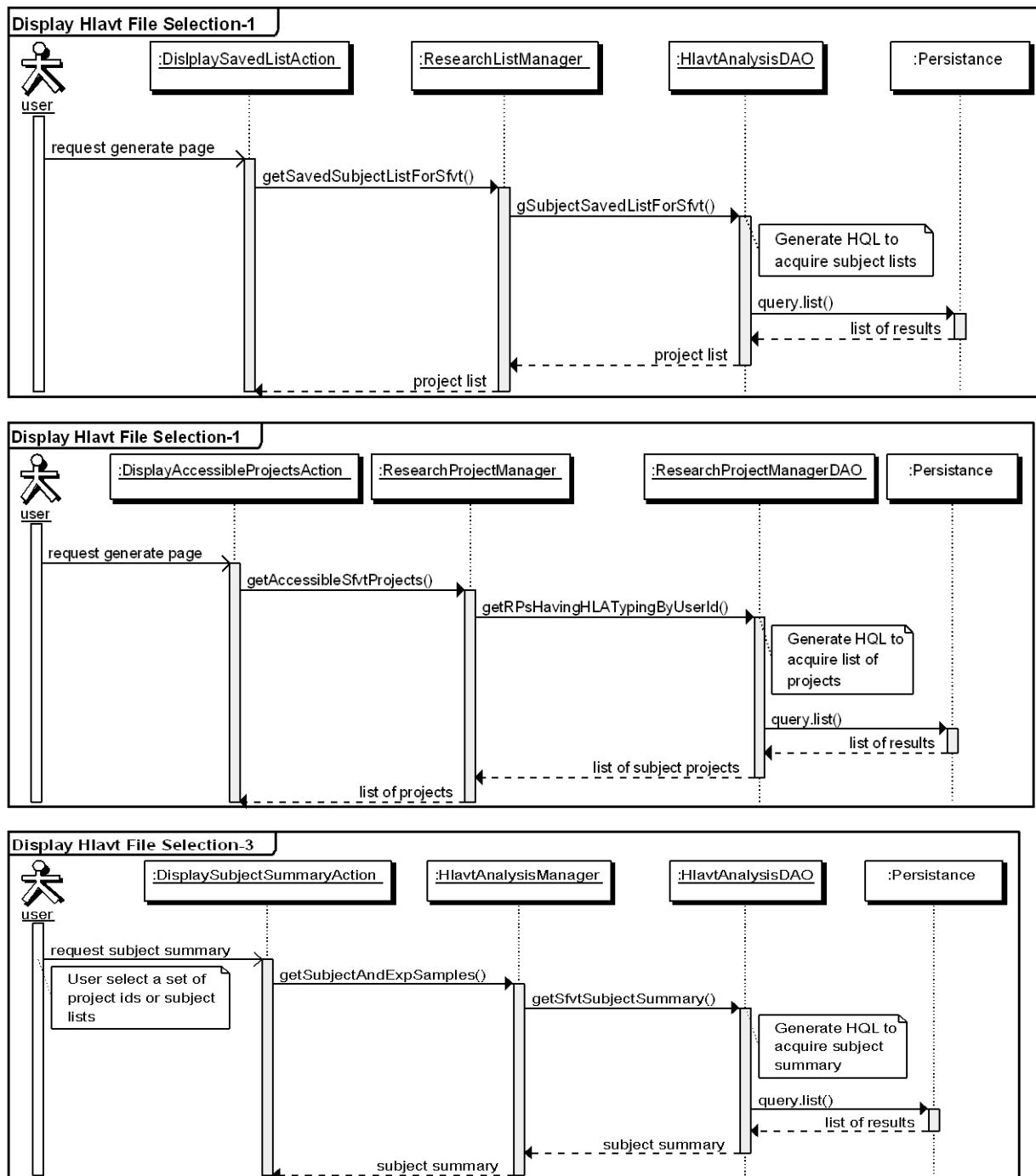
HLA Locus Name	Number of Subjects with Locus
HLA-A	50
HLA-B	50
HLA-C	50
HLA-DPB1	50
HLA-DQB1	50
HLA-DRB1	50

Done

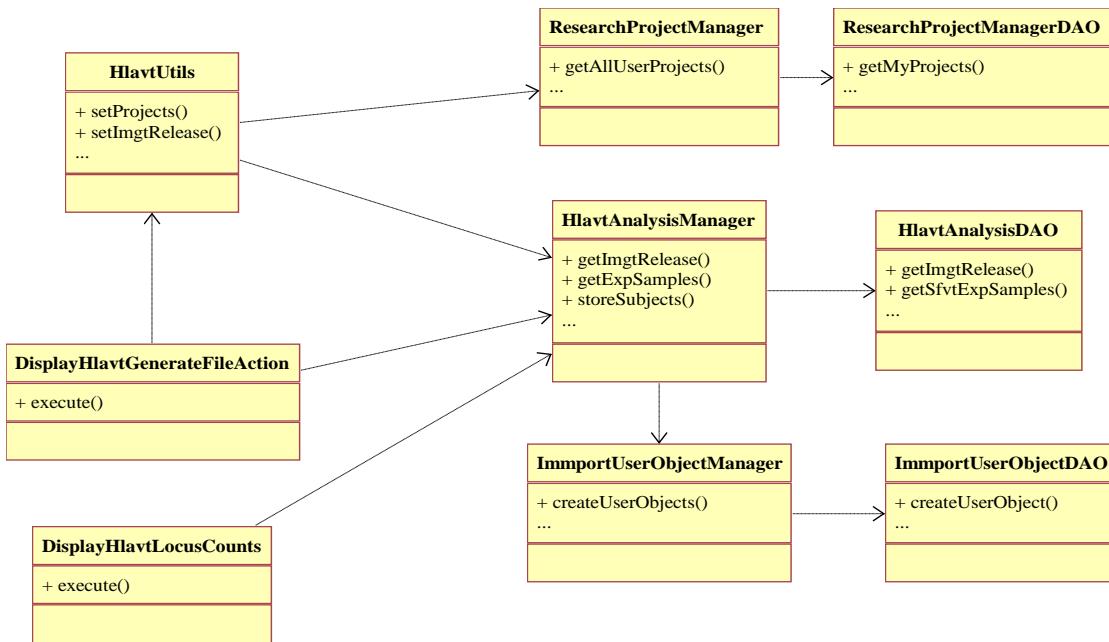
19.2.1 Create HLA Data Set Selection Class Diagram



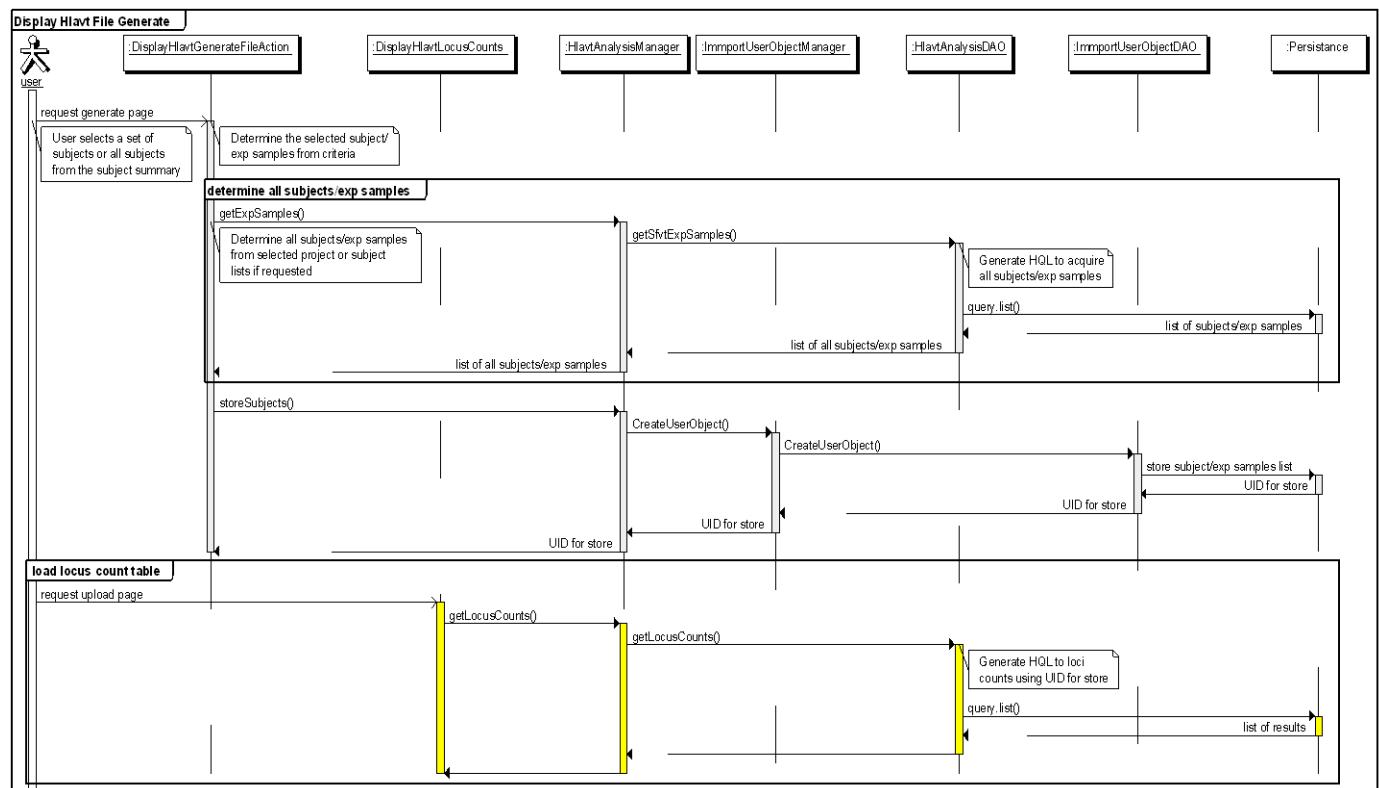
19.2.2 Create HLA Data Set Selection Sequence Diagram



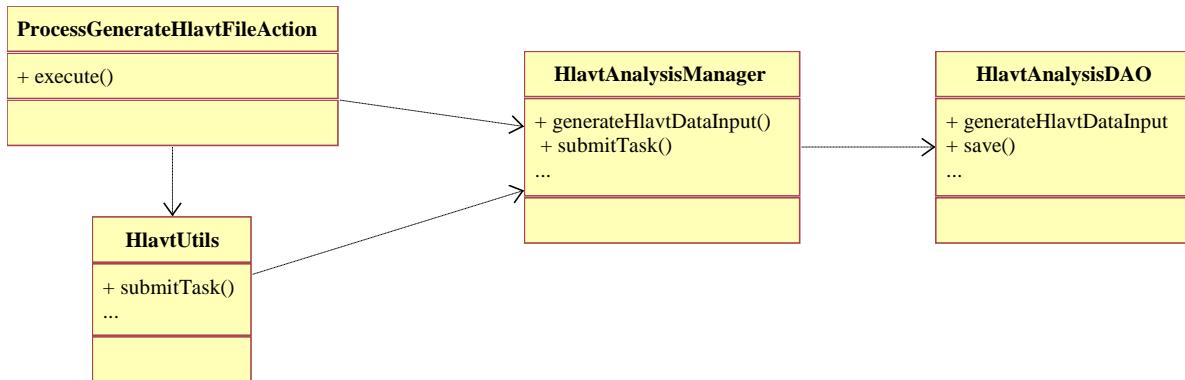
19.2.3 Create An HLA Data Set Display Class Diagram



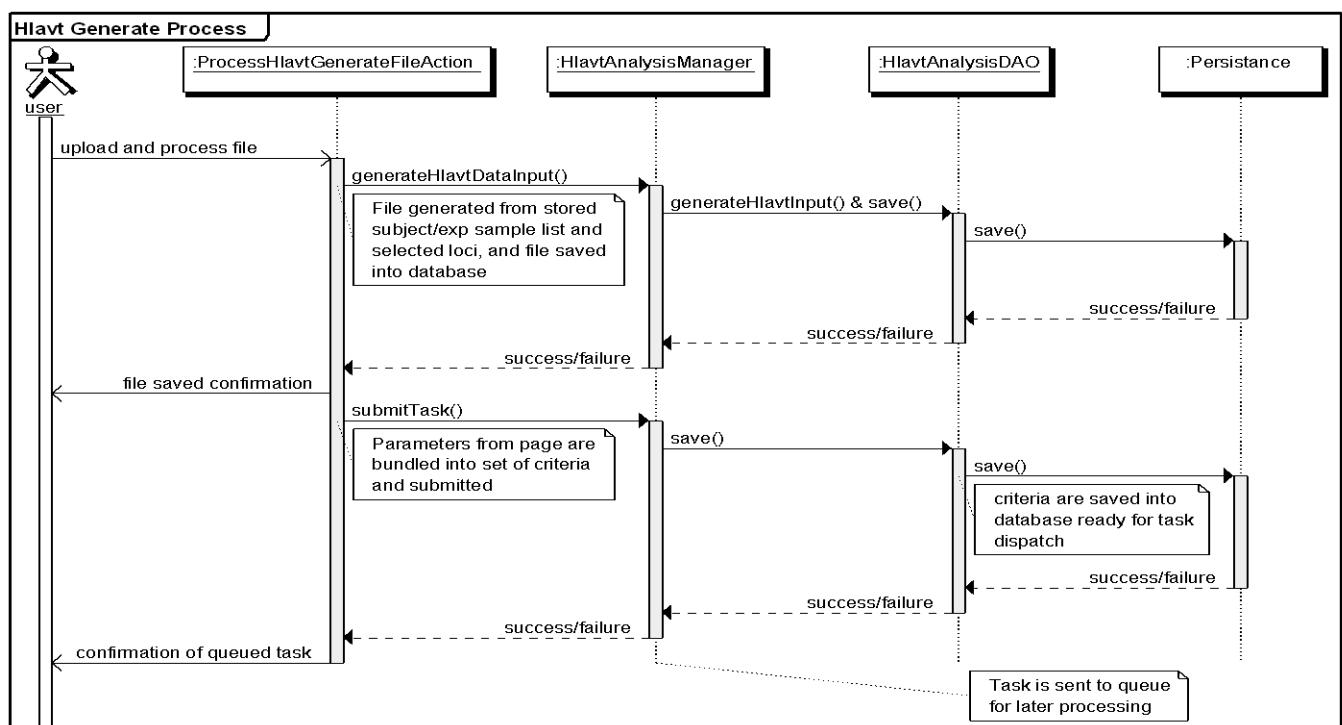
19.2.4 Create An HLA Data Set Display Sequence Diagram



19.2.5 Create HLA Data Set Process Class Diagram



19.2.6 Create HLA Data Set Process Sequence Diagram



19.3 RE-ANALYZE HLA DATA SET PREVIOUSLY CREATED IN IMMPORT

After a user has created a HLA typing file and annotated it with a name and a description, the data set may be retrieved and re-used

The current IMGT/HLA Release is [IMGT/HLA Release 3.00 \(2010-04-01\)](#).
The actions below accept both IMGT/HLA version 2.* or version 3.* allele formats and NMDP version 2 and 3 NMDP-code formats.

Fields marked with an asterisk * are required.

Select the action *

- Validate alleles
Validation uses the [ANTT](#) Tool to validate conformance to IMGT/HLA version 2 & 3 nomenclature format and G- and P-Codes.
Also, Validation converts the input file into IMGT/HLA version 3.* format since all options operate on that format.
- Validate and Generate SFVT Vector Files
- Validate and Reduce allele ambiguity
This tool was designed by Steven J. Mack et. al. and was developed with his invaluable co-operation.
The input requires the column 'Population Area' that specifies the population area ([Help](#)) associated with each row of data.
- Run [Pypop](#) HLA QC Pipeline
This pipeline validates the input file first and then runs Pypop. Please check out the [Pypop Timings](#) ([Help](#)).

Select the IMGT/HLA Output Version *

- Generated files will be formatted in **IMGT/HLA version 3.* format**
All the options above operate on IMGT/HLA version 3.* format and output all files in that format
- Generated files will be formatted in **IMGT/HLA version 2.* format** except for IMGT/HLA G- and P-codes
Conversion from IMGT/HLA version 3.* to version 2.* format uses the [ANTT](#) Tool

The MHC Validation and SFVT Analysis tool requires a **project: "workspace"**, and an **HLA Typing File**.
Please see the [User Guide](#) or contact the [help desk](#) for more information.

Select a project to store generated file * 03099: HLA Typing Result upload Error

Dataset Name *

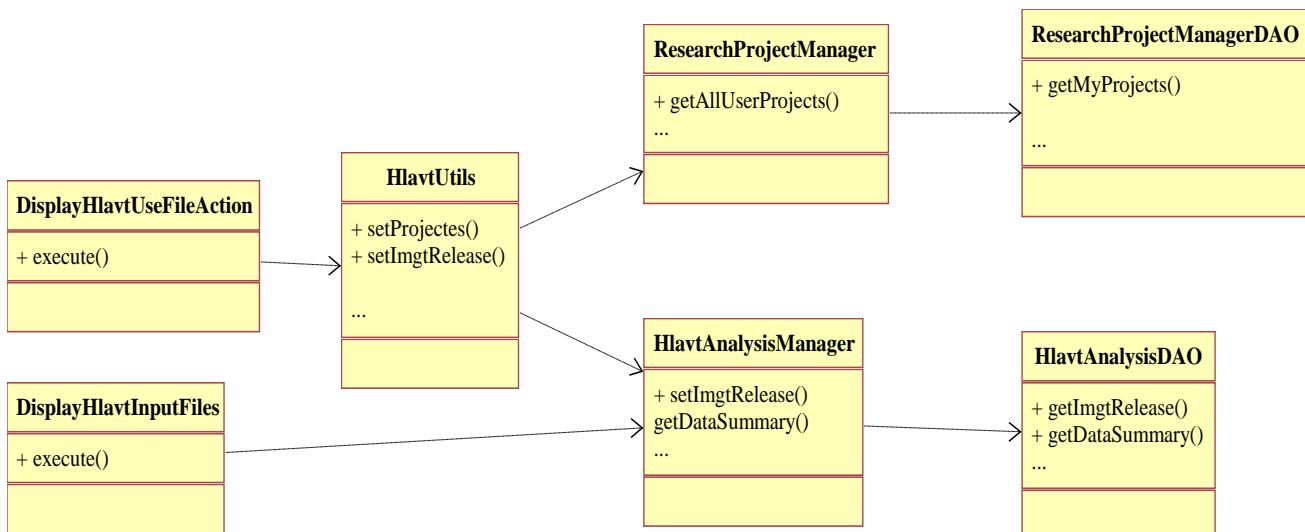
Dataset Descriptor *

Select an existing HLA Typing File below and then click Next *

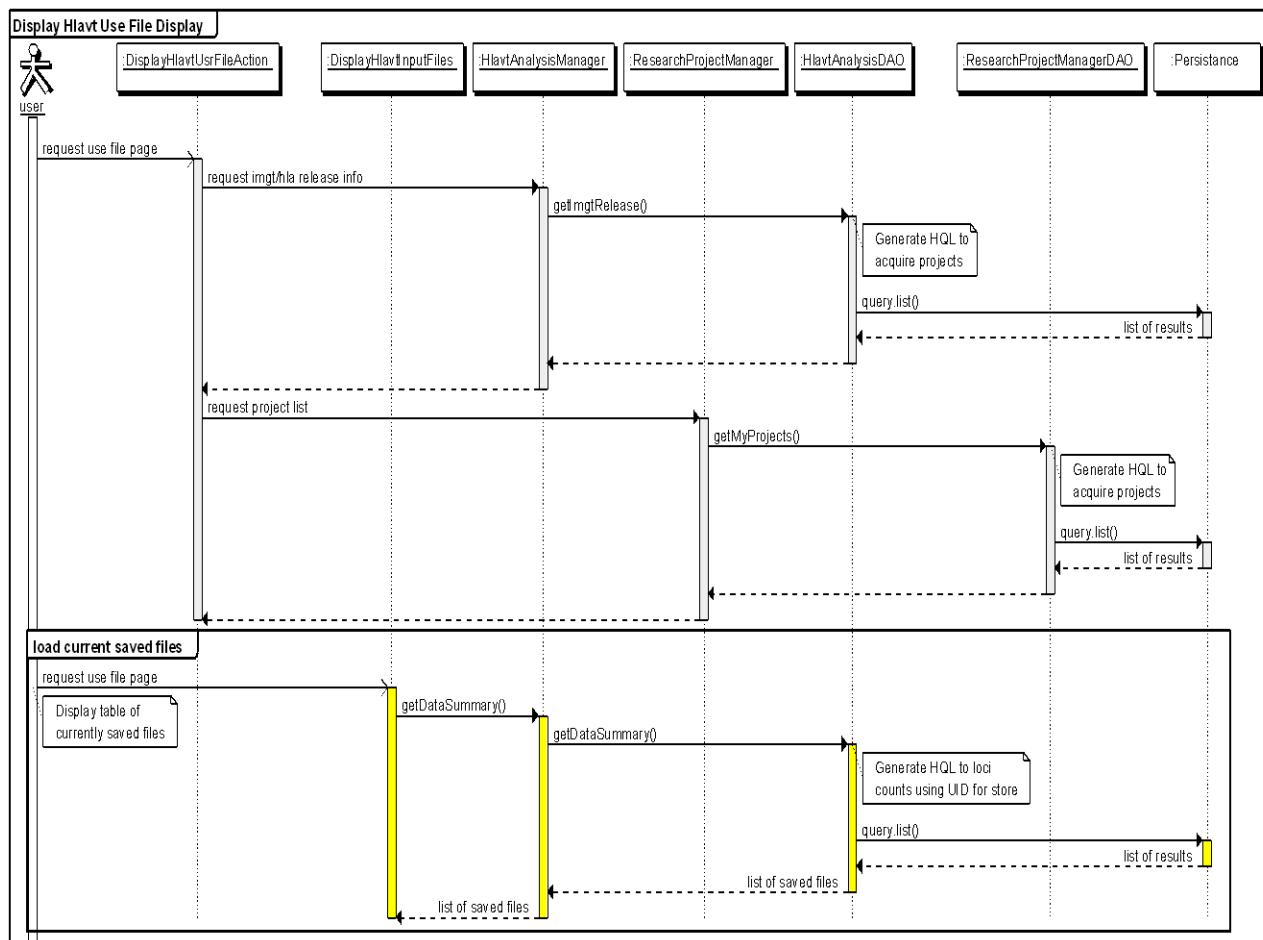
HLA Typing Result File Name	Source Typ	Project Nam	Creation Da	Source File	Last IMGT Version	Last Action Request	Last Workflow Use
10_subjects_with_all_loci.txt	generated fi	1500 Subj	2010-01-15	HLA Typing	IMGT Release 2.26 (2005	Allele Validation	Generate File
1500_data.txt	generated fi	SFVT	2010-07-07		IMGT Release 3.00 (2010	HLA QC Pipeline	Generate File
1500_data.txt	generated fi	03099: HLA	2010-07-07	HLA Typing	IMGT Release 3.00 (2010	HLA QC Pipeline	Generate File
1500_randomized_6_loci.txt	generated fi	1500 Subj	2010-01-15	HLA Typing	IMGT Release 3.00 (2010	HLA QC Pipeline	Generate File
1500_randomized_6_loci.txt	generated fi	1500 Subj	2010-01-15	HLA Typing	IMGT Release 2.26 (2005	Allele Validation	Generate File

Done

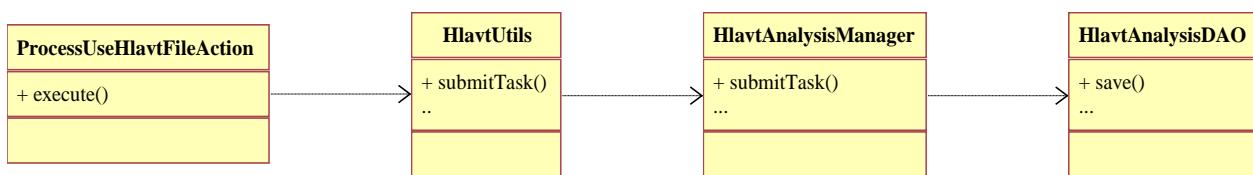
19.3.1 Re-Analyze HLA Data Set Previous Created Display Class Diagram



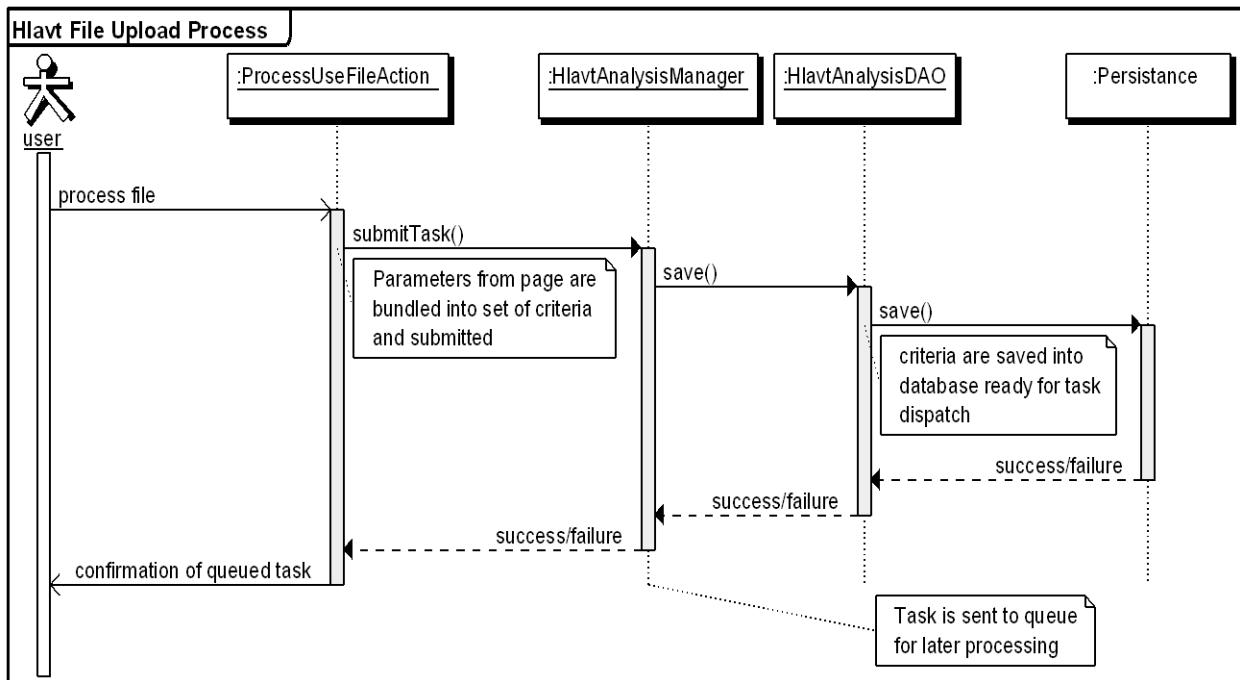
19.3.2 Re-Analyze HLA Data Set Previously Created Display Sequence Diagram



19.3.3 Re-Analyze a HLA Data Set Previously Created Process Class Diagram



19.3.4 Re-Analyze a HLA Data Set Previously Created Process Sequence Diagram



19.4 RUN SFVT ANALYSIS

After a user has generated sequence feature vector files, the data can be analyzed using statistical tests including chi-square test. The workflow requires the user to first choose the set of sequence feature vector files that have been generated. Then the user proceeds to the next step in the workflow. This step requires the user to choose the phenotype, the statistical tests, and the loci over which to run the tests.

The screenshot shows the ImmPort website interface for running SFVT analysis. At the top, there is a navigation bar with links for 'Edit Profile | Sign Out', 'About ImmPort', 'Admin', 'Access Data', 'Tools', 'Resources', 'News & Events', and a home icon. Below the navigation bar, a breadcrumb trail indicates the current page: 'Analysis / MHC Validation and SFVT Analysis / SFVT Analysis / Select SFVT File - Beta Release'. A horizontal navigation bar below the breadcrumb trail includes links for 'Home', 'Upload Data', 'Create Data Set', 'Analyze Data Set', 'SFVT Analysis', 'Analysis Results', 'User Guide', and 'Ambiguity Reduction Slides / Logic'. Below this, a series of buttons are shown: 'Select Vector File Result' (highlighted in grey), 'Select Parameters', and 'Display submitted request'. Below these buttons are links for 'Previous', 'Next', and 'Cancel'. A note states: 'Fields marked with an asterisk * are required.' Below this, a message says: 'The HLAVT Analysis requires a vector file result. See [User Guide](#) for more details.' A note also says: 'Select an existing vector file generation result and then click Next*'. A list of selected items is shown: '[5338, Generate VF, uploaded file, Steve Mack unambiguous data containing phenotype, Steve Mack ambiguity reduction test dataset with phenotype, Completed, IMGT Release 3.00 (2010-04-01)]'. At the bottom, a table displays a list of tasks:

Task ID	Start Date	End Date	Task Type	Source Type	Input File	General Name	Description	Status	IMGT Version
5338	2011-07-25	2011-07-25	Generate VF	uploaded file	Task5337.Genotype_a	IMGTH/ Steve Mack ur	Steve Mack ar	Completed	Release 3.00 (2010-04-01)
4523	2011-03-14	2011-03-14	Generate VF	uploaded file	HLA1_ver2_Orig_mod1	IMGTH/ nishanth full da	testing pipeline	Completed with Vi	Release 3.00 (2010-04-01)
4480	2011-03-10	2011-03-10	Generate VF	uploaded file (reused)	HLA1_ver2_Orig_mod1	IMGTH/ nishanth partia	testing pipeline	Completed with Vi	Release 3.00 (2010-04-01)
4479	2011-03-10	2011-03-10	Generate VF	uploaded file	HLA1_ver2_Orig_mod1	IMGTH/ nishanth partia	testing pipeline	Completed with Vi	Release 3.00 (2010-04-01)

Analysis / MHC Validation and SFVT Analysis / SFVT Analysis / Select Parameters - Beta Release

Home | Upload Data | Create Data Set | Analyze Data Set | SFVT Analysis | Analysis Results | User Guide | Ambiguity Reduction Slides / Logic

Select Vector File Result → Select Parameters → Display submitted request

Previous Next Cancel

You have chosen the following vector file generation task for SFVT Analysis:

Task ID: [5338](#)
 Task Info: [5338
 Generate VF
 uploaded file
 Steve Mack unambiguous data containing phenotype
 Steve Mack ambiguity reduction test dataset with phenotype
 Completed
 IMGT Release 3.00 (2010-04-01)]

Fields marked with an asterisk * are required.

- ▶ The SFVT Analysis tool requires a **project**.
- ▶ The **Dataset Name** is used to name the parameterization file generated in this workflow.
- ▶ Please see the [User Guide](#) or contact the [help desk](#) for more information.

Select a project to store generated files *

Silver-Standard_Development

Dataset Name *

Steve Mack unambiguous data containing phenotype

Dataset Descriptor

Steve Mack ambiguity reduction test dataset with phenotype

Select a phenotype column *

Subject Phenotype

Select a statistical test *

Chi-Square Test

Select Locus:*

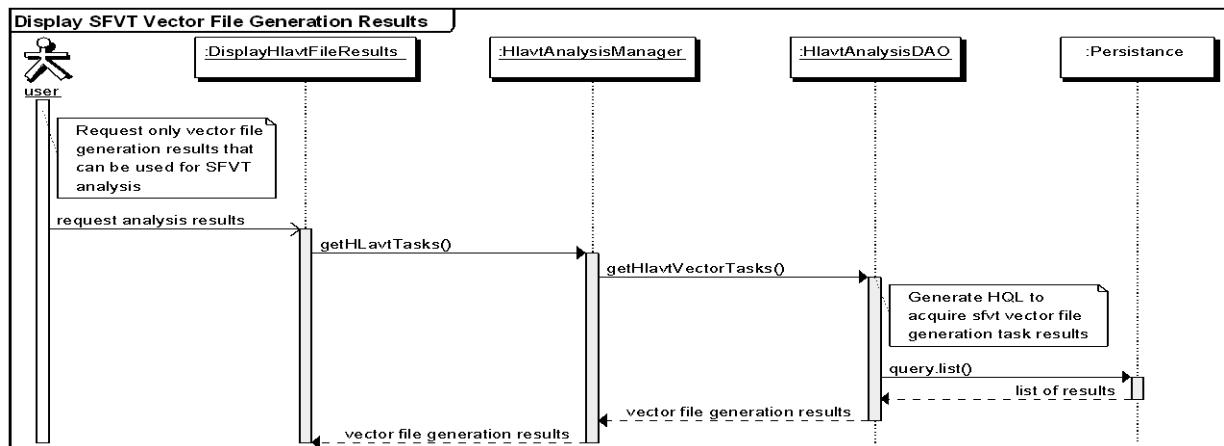
Selected items: HLA-A,HLA-B,HLA-C

Page 1 of 1		Displaying 1 - 3 of 3
<input checked="" type="checkbox"/>	HLA Locus Name ▲	
<input checked="" type="checkbox"/>	HLA-A	
<input checked="" type="checkbox"/>	HLA-B	
<input checked="" type="checkbox"/>	HLA-C	

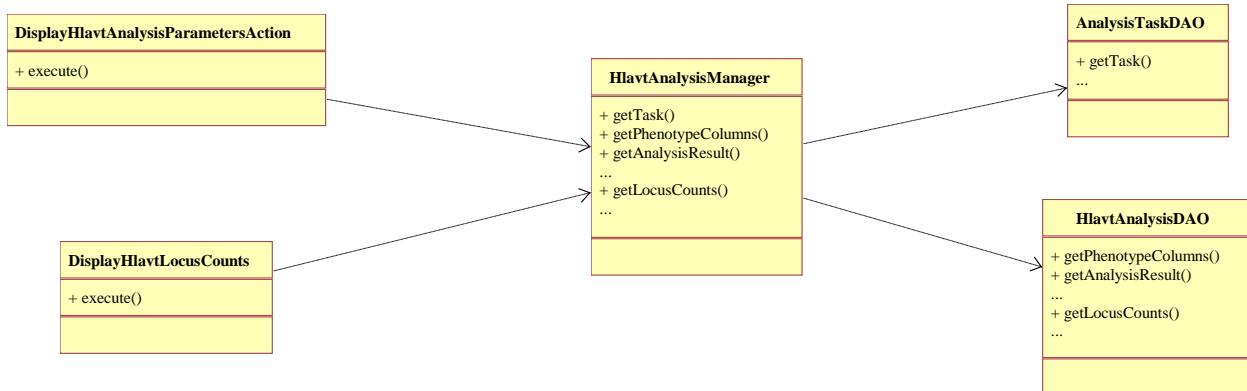
19.4.1 SFVT Analysis Variant Type Selection Class Diagram



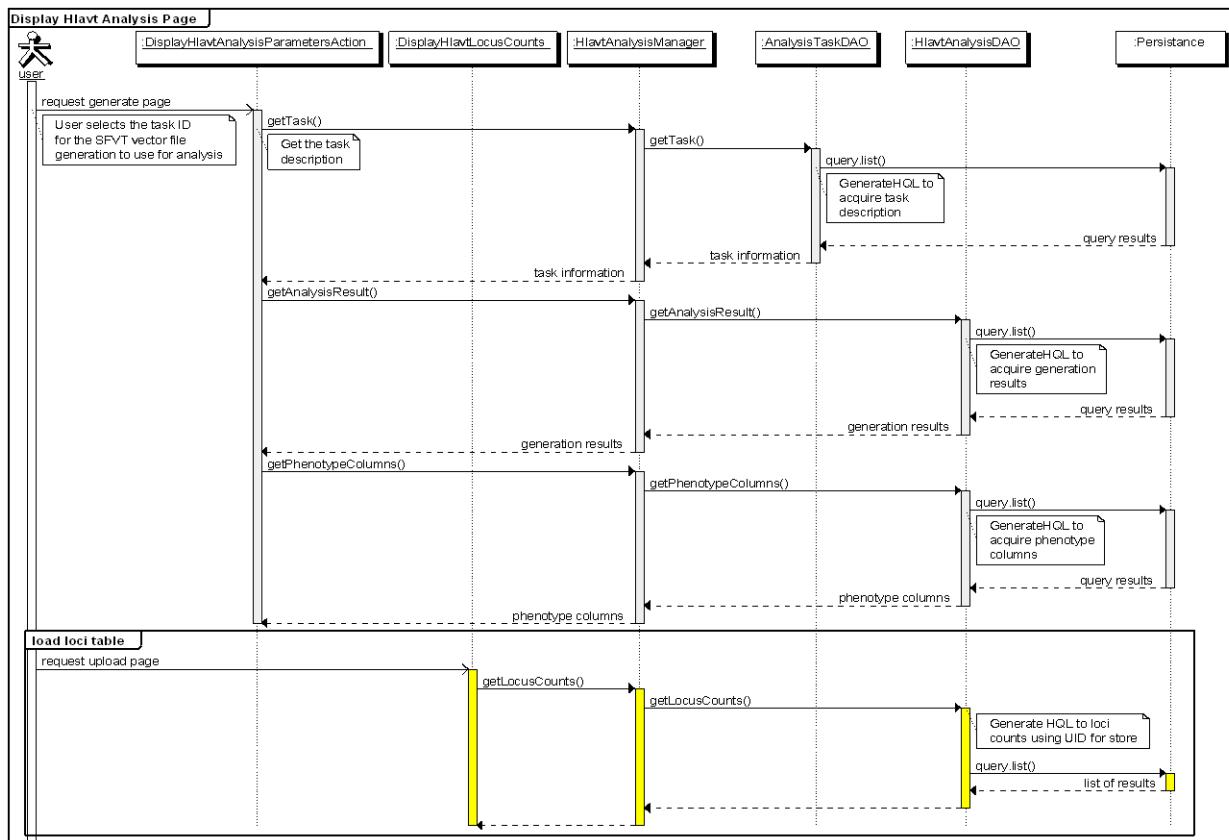
19.4.2 SFVT Analysis Variant Type Selection Sequence Diagram



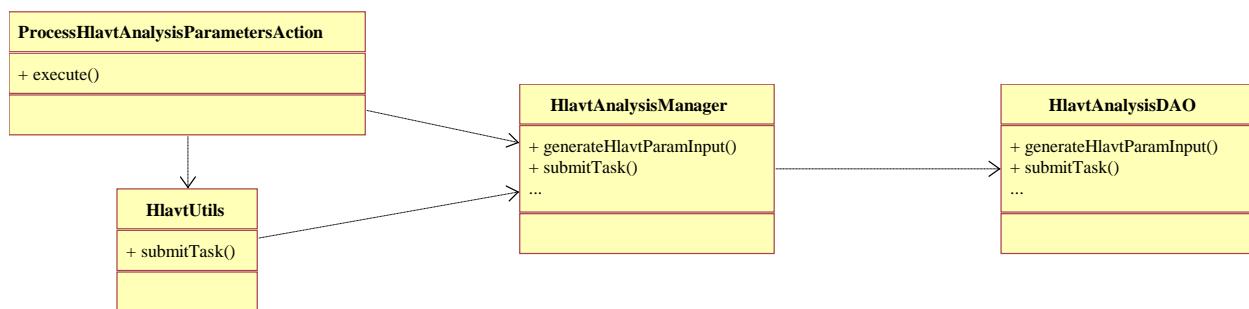
19.4.3 SFVT Analysis Loci Selection Class Diagram



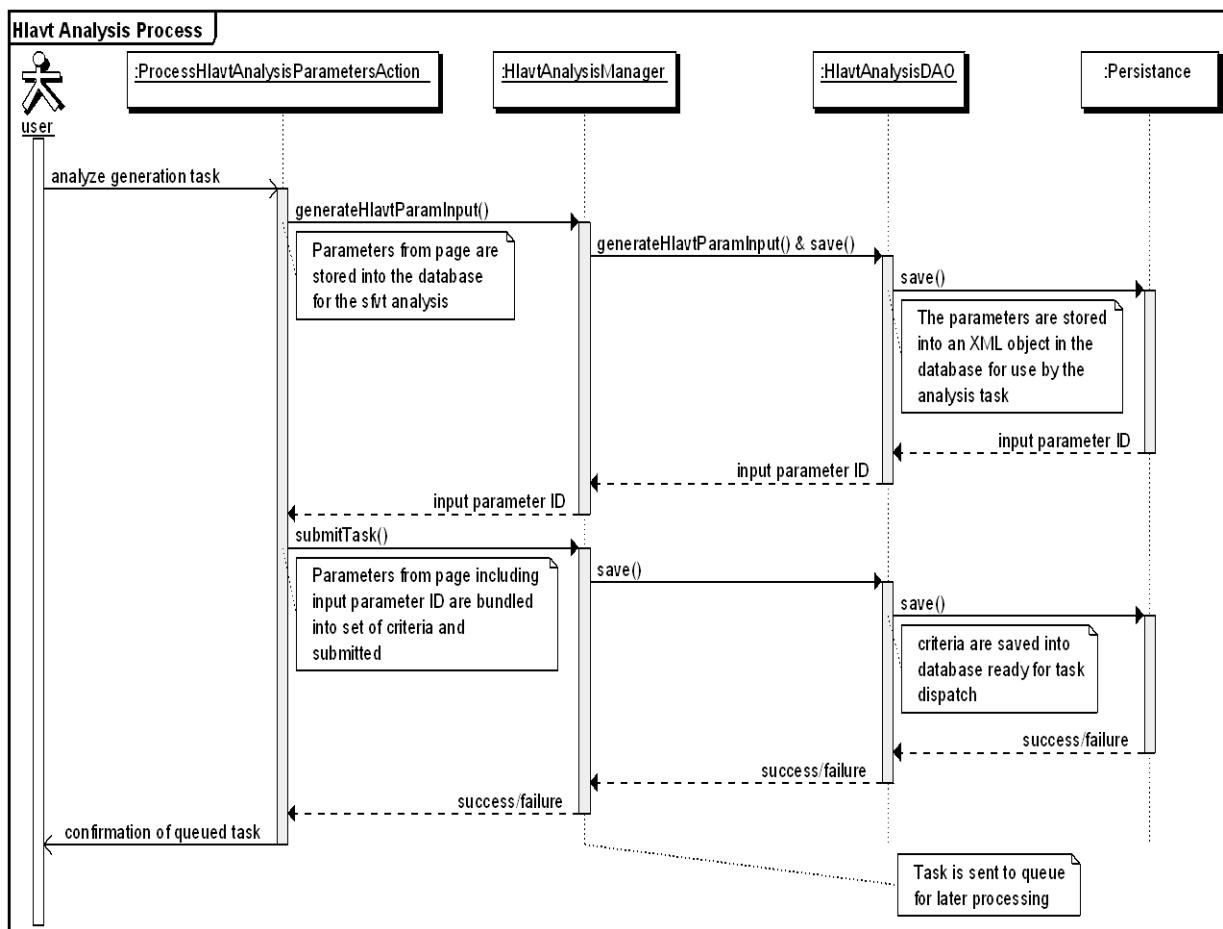
19.4.4 SFVT Analysis Loci Selection Sequence Diagram



19.4.5 SFVT Analysis Process Class Diagram



19.4.6 SFVT Analysis Process Sequence Diagram



19.5 VIEWING PIPELINE RESULTS

After an analysis task is completed, the results are displayed in a summary table showing all results for that user.



[Edit Profile](#) | [Sign Out](#)



[About ImmPort](#) [Admin](#) [Access Data](#) [Tools](#) [Resources](#) [News & Events](#)

Analysis / MHC Validation and SFVT Analysis / Analysis Results - Beta Release

[Home](#) | [Upload Data](#) | [Create Data Set](#) | [Analyze Data Set](#) | [SFVT Analysis](#) | [Analysis Results](#) | [User Guide](#) | [Ambiguity Reduction Slides / Logic](#)

This page is a list of the tasks you have submitted to the SFVT File Generation Tools. Click on a task ID to view detailed information about the task or to see the results. Please contact the help desk if you have any questions about the status of your analysis runs ([Contact Us](#))

Displaying 1 - 20 of 63										
Task ID	Start Date	End Date	Task Type	Source Type	Input File	Generat	Name	Description	Status	IMGT Version
5647	2011-10-14	2011-10-14	Analyze VF	uploaded file	Task5337.Genotype_a	IMGT/H	Steve Mack un	Steve Mack an	Completed	Release 3.00 (2010-04-01)
5339	2011-07-25	2011-07-25	Analyze VF	uploaded file	Task5337.Genotype_a	IMGT/H	Steve Mack un	Steve Mack an	Completed	Release 3.00 (2010-04-01)
5338	2011-07-25	2011-07-25	Generate VF	uploaded file	Task5337.Genotype_a	IMGT/H	Steve Mack un	Steve Mack an	Completed	Release 3.00 (2010-04-01)
5337	2011-07-25	2011-07-25	HLA Ambiguity	uploaded file	SJM-testPop1_ambig.p	IMGT/H	Steve Mack an	Steve Mack an	Completed	Release 3.00 (2010-04-01)
5336	2011-07-25	2011-07-25	Analyze VF	uploaded file	HLA1_ver2_Orig_modf	IMGT/H	nishanth full da	testing pipeline	Completed	Release 3.00 (2010-04-01)
4525	2011-03-14	2011-03-14	Analyze VF	uploaded file	HLA1_ver2_Orig_modf	IMGT/H	nishanth full da	testing pipeline	Completed	Release 3.00 (2010-04-01)
4523	2011-03-14	2011-03-14	Generate VF	uploaded file	HLA1_ver2_Orig_modf	IMGT/H	nishanth full da	testing pipeline	Completed with V	Release 3.00 (2010-04-01)
4522	2011-03-14	2011-03-14	Analyze VF	uploaded file	HLA1_ver2_Orig_modf	IMGT/H	nishanth partial	testing pipeline	Completed	Release 3.00 (2010-04-01)
4480	2011-03-10	2011-03-10	Generate VF	uploaded file (reused)	HLA1_ver2_Orig_modf	IMGT/H	nishanth partial	testing pipeline	Completed with V	Release 3.00 (2010-04-01)
4479	2011-03-10	2011-03-10	Generate VF	uploaded file	HLA1_ver2_Orig_modf	IMGT/H	nishanth partial	testing pipeline	Completed with V	Release 3.00 (2010-04-01)
4451	2011-03-03	2011-03-03	HLA Ambiguity	uploaded file (reused)	SJM-testPop1_ambig.t	IMGT/H	steve mack an	testing pipeline	Completed	Release 3.00 (2010-04-01)

The user may click on the Task_ID value to retrieve the results and details about that task.



[Edit Profile](#) | [Sign Out](#)

[About ImmPort](#) [Admin](#) [Access Data](#) [Tools](#) [Resources](#) [News & Events](#)

Analysis / MHC Validation and SFVT Analysis / Analysis Results - Beta Release

[Home](#) | [Upload Data](#) | [Create Data Set](#) | [Analyze Data Set](#) | [SFVT Analysis](#) | [Analysis Results](#) | [User Guide](#) | [Ambiguity Reduction Slides / Logic](#)

Data Set Details

Task ID	5338
Generation Status	Completed
Data Set Name	Steve Mack unambiguous data containing phenotype
Description	Steve Mack ambiguity reduction test dataset with phenotype
Research Project	Silver-Standard_Development
Source File	Task5337.Genotype_ambiguity_reduction_file.txt
Source Type	uploaded file
Source File Type	HLA Typing
Alleles generated in format	IMGT/HLA Version 3 (the ANTT generated file will be in the IMGT/HLA version different from that of the input file)
IMGT Release	Release 3.00 (2010-04-01)
Tool Information	Vector File Generator (3.0)

The files associated with the SFVT vector file generation results are:

[Download All Results](#)

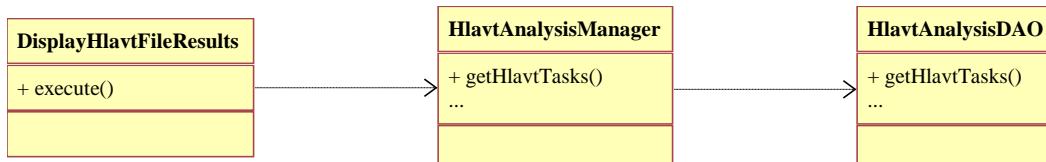
File Name	Description
Input Files	
Task5337.Genotype_ambiguity_reduction_file.txt	The input HLA typing data that was submitted by the user
Metadata Files	
Task5338.Allele_validation_summary.txt	Summary of validation results and of any errors and exceptions
Task5338.ANTT_validation_summary.txt	Summary of ANTT validation results and of any errors and exceptions encountered by the ANTT tool
Task5338.SFVT_generation_summary.txt	Generation Summary
Task5338.README.txt	Descriptive details about the input files, the analysis parameters and the output files
Result Files	
Task5338.Allele_validation_file.txt	Validated alleles are presented in this result file
Task5338.ANTT_validation_file.txt	The ANTT tools validation results are written to this file
Task5338.HLA-A.SFVT_vectors_file.xls	Generated SFVT vector file for locus HLA-A
Task5338.HLA-B.SFVT_vectors_file.xls	Generated SFVT vector file for locus HLA-B
Task5338.HLA-C.SFVT_vectors_file.xls	Generated SFVT vector file for locus HLA-C

[Previous](#)

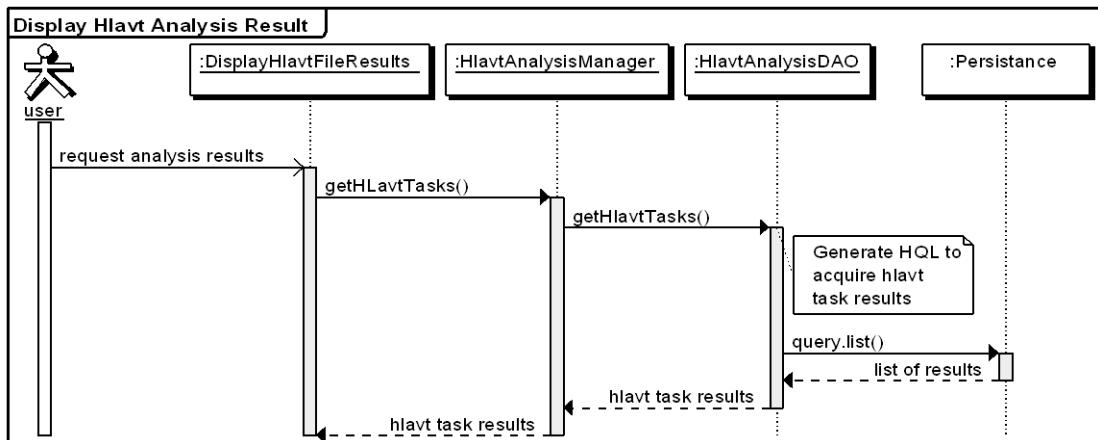
[Cancel](#)

[Run SFVT Analysis](#)

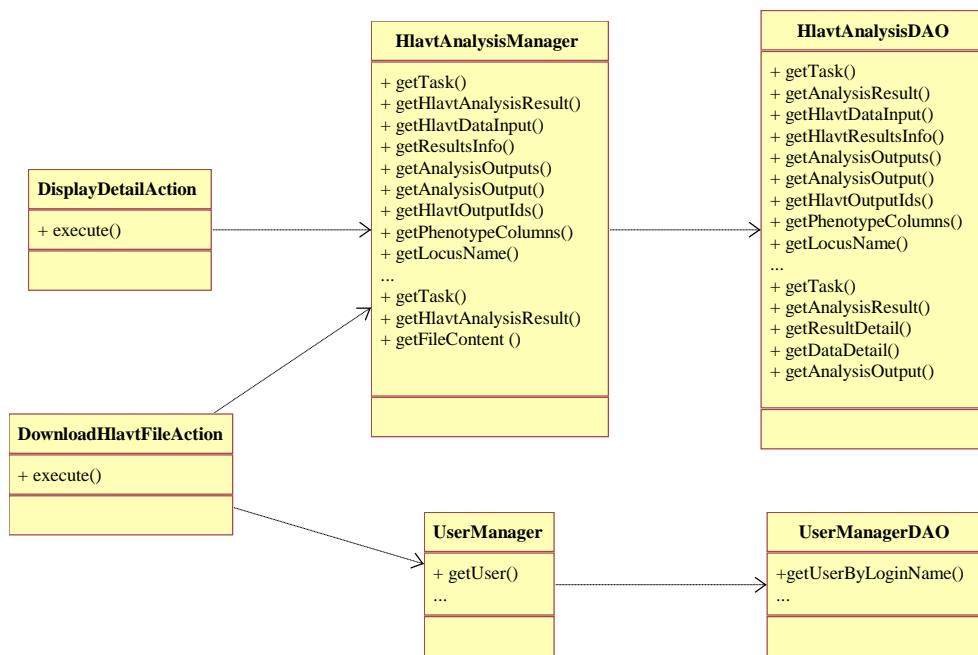
19.5.1 Pipeline Analysis Results Class Diagram



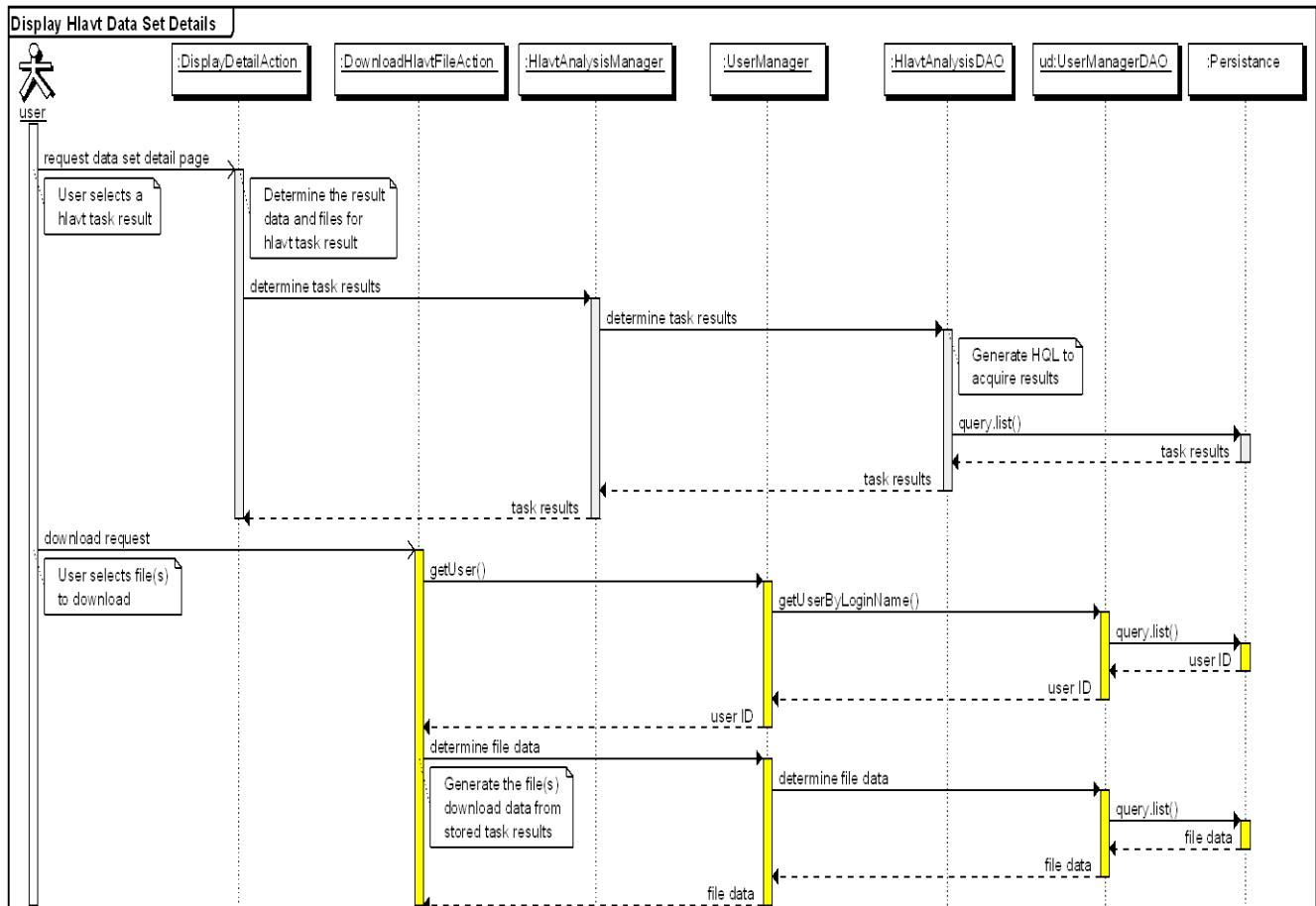
19.5.2 Pipeline Analysis Results Sequence Diagram



19.5.3 Pipeline Data Set Details Class Diagram



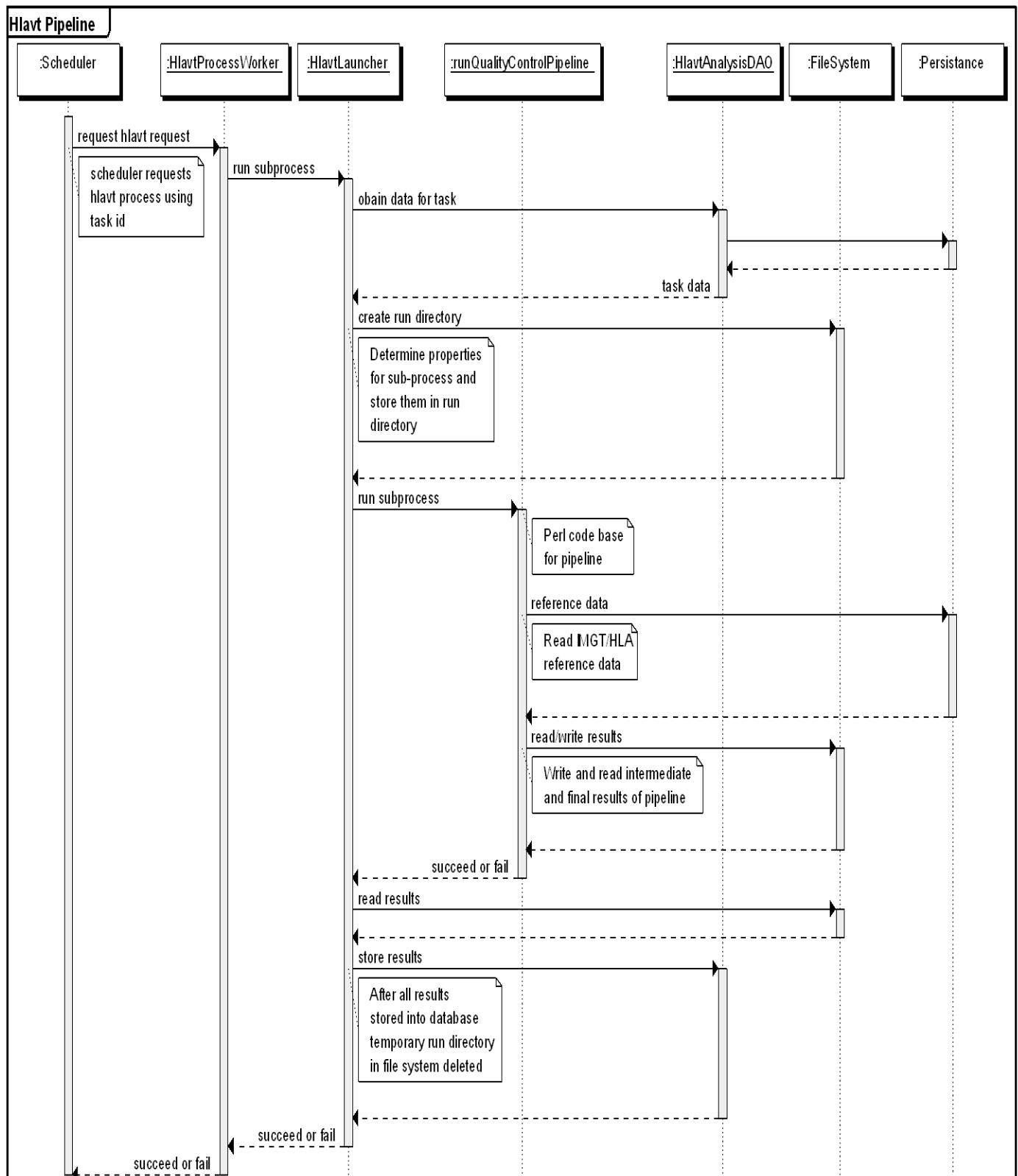
19.5.4 Pipeline Data Set Details Sequence Diagram



19.6 HLAVT PIPELINE

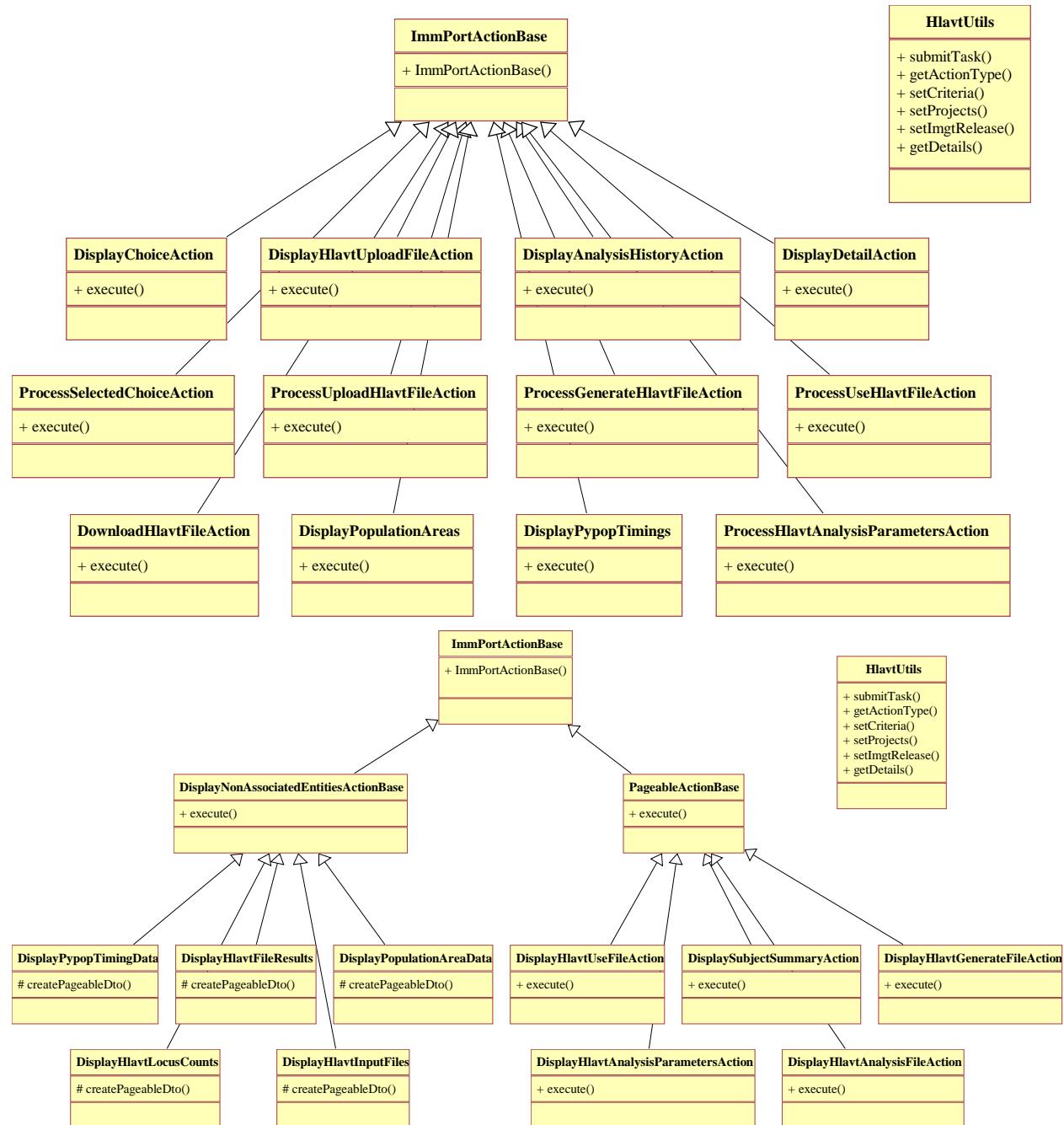
After a task is initiated from the front-end, a system call is made and the process is executed by a series of Perl scripts wrapped around the relevant tools that implement the algorithms.

19.6.1 HLAVT Pipeline Sequence Diagram

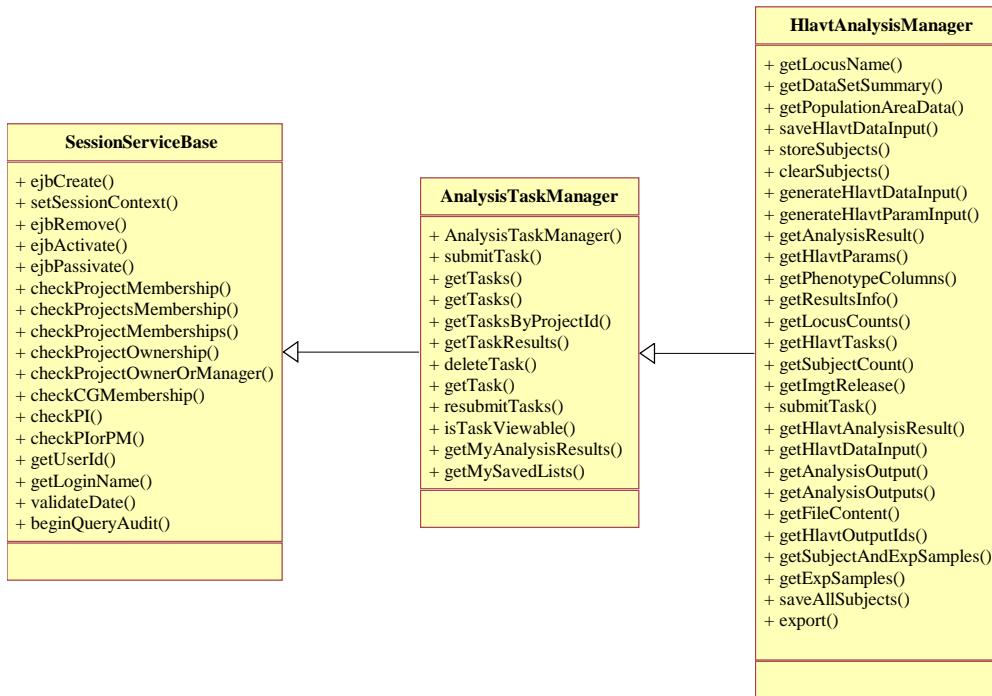


19.7 HLA ANALYSIS CLASS DIAGRAMS

19.7.1 Actions



19.7.2 EJB



19.7.3 DAO



19.8 HLA VT PIPELINE CODE BASE DESCRIPTION

The HLAVT pipeline executes the Perl code base that performs the requested action. The Perl software base provides the implementation of the programs specified above using a Perl class structure. An action is identified by an analysis task name (hlavf, hlavt, hlavd, hlava, and hlaqc). The controller for the pipeline is the Perl program, runQualityControlPipeline.pl. The controller will serially execute a number of Perl programs that define the particular action as specified below:

Program	Description
runQualityControlPipeline.pl	This program is the controller program. It manages execution of the programs defined by an action. If the action has requested that the generated output files be in IMGT/HLA version 2.* format, then this program converts all output files into version 2.* format after all results are generated.
validateAlleleNames.pl	This program performs validation of HLA locus alleles and codes that appear in the input file to the action. The alleles and codes (IMGT/HLA G- and P-codes) need to conform to IMGT/HLA format standards. Also, NMDP codes need to satisfy NMDP standards and be replaced by their corresponding group of alleles. This program executes the third-party tool ANTT.exec to perform IMGT/HLA format standard validation
generateVariantTypes.pl	This program takes the output of the validateAlleleNames.pl program and generates the sequence feature vector files associated with the loci that have data in the input file.
allelicAmbiguityReduction.pl	This program performs the allelic ambiguity reduction developed by a team of experts. This tool takes the output of the validateAlleleNames.pl tool and generates a file containing the results of allelic ambiguity reduction
genotypeAmbiguityReduction.pl	This program performs the genotypic ambiguity reduction developed by the same teams of experts as allelic ambiguity reduction. This tool takes the output of allelicAmbiguityReduction.pl and generates a file containing genotypic ambiguity reduction.

Program	Description
analyzeVariantTypes.pl	This program performs statistical tests including chi-square test on sequence feature variant types. The tool was developed by a team of experts. This tool takes the output of generateVariantTypes.pl and generates results of the analysis.
runPypop.pl	This program runs a population genomics analysis using the pypop tool on the output of the tool validateAlleleNames.pl. pypop tool is a third-party tool and is configured to run Hardy-Weinberg and Hardy-Weinberg-Guo-Thompson analyses.

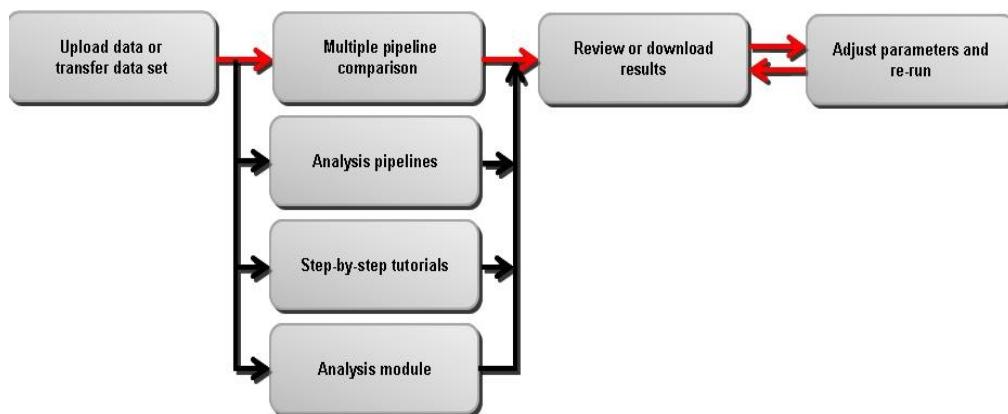
The following third-party tools are used in the hlavt pipeline.

Third-Party Tools	Description
ANTT.exe	This program is implemented in visual basic and performs allele validation using IMGT/HLA format standards. This tool is available from IDAWG (http://igdawg.org/) and is described there.
Pypop	This program is implemented in python and performs population genomics analysis. It is available from the website http://www.pypop.org/ and is described there.

20.0 GENE PATTERN

The ImmPort system provides a set of tools that enable scientists to perform reproducible *in silico* research for gene expression analysis. To do this, the ImmPort team has integrated the open source GenePattern platform by the Broad Institute at

<http://www.broadinstitute.org/cancer/software/genepattern/>. GenePattern is a powerful genomic analysis platform providing a number of useful state of the art analysis tools. The tools can run stand-alone or can be combined in a pipeline fashion where the output of a previous step is used as the input to the next step in a pipeline. Figure 13-1 GenePattern Workflow shows a typical workflow scenario:



The GenePattern component includes 57 categories of modules noted in the table below for gene expression analysis:

GenePattern Module Category	Total Number Avail. Modules
Annotation	2
Clustering	3
Data Format Conversion	1
Data Management	2
Gene List Selection	9
Image Creators	2
Pathway Analysis	2
Pipeline	11
Prediction	3
Preprocess & utilities	10
Projections	1
Visualizer	11
Total :	57

20.1 GENE PATTERN ARCHITECTURE

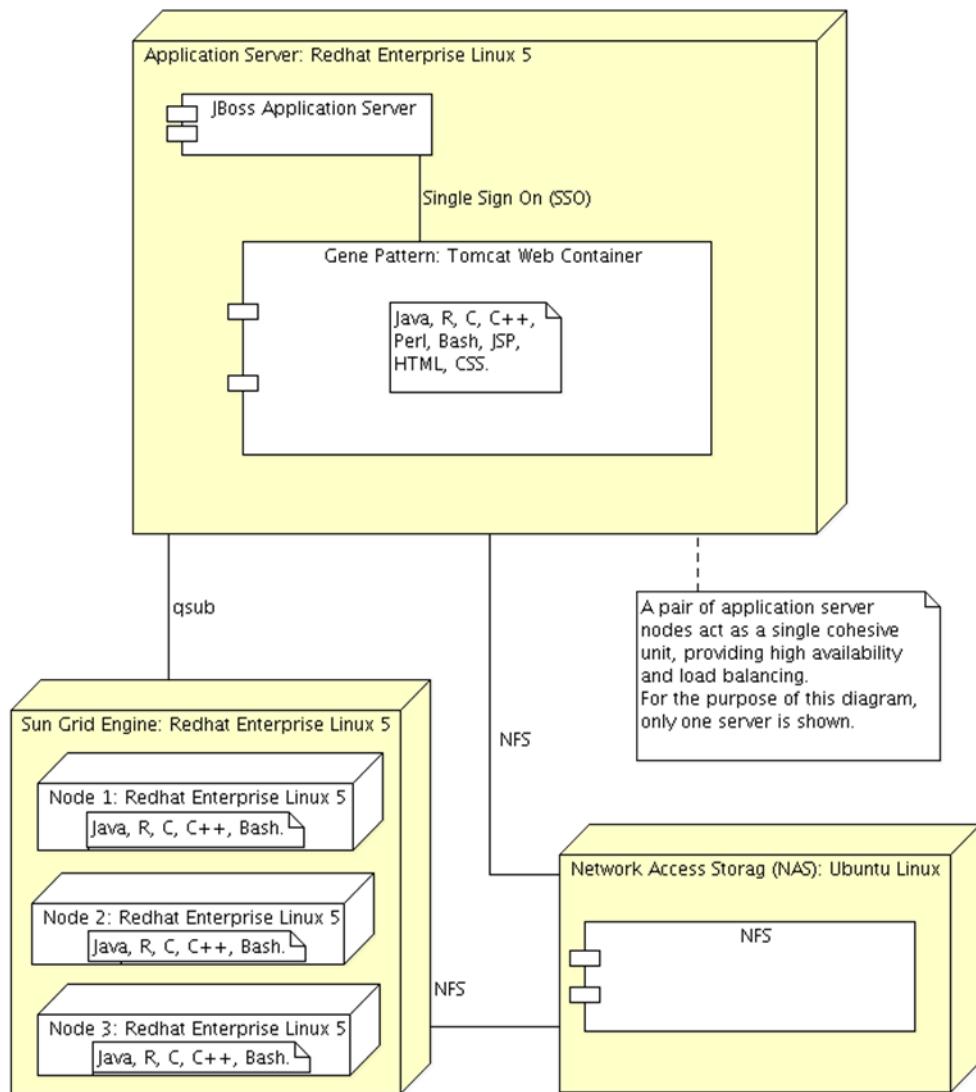
As depicted in Figure 20-1 GenePattern Software Architecture, an instance of the GenePattern platform is hosted on each application server. A GenePattern instance is a self-contained component, executing within its own process space and Tomcat Web Container outside the JBoss environment. Multiple GenePattern instances can be started on different hosts across the network. But only one installation of the GenePattern software exists. The GenePattern core components, including the bundled and home grown modules are installed on the Network Attached Storage (NAS) server. This configuration allows for a flexible environment, making it easy for software upgrades.

Authentication and authorization to GenePattern modules, job history, and results are based on Single Sign On using GenePattern's IauthenticationPlugin mechanism. The IauthenticationPlugin is a software interface component that facilitates access to remote GenePattern clients via standard HTTP requests or web services APIs.

The ImmPort Sun Grid Engine (SGE) Component is a 3-node grid computing environment where GenePattern tasks are queued for batch processing. SGE is powerful open source grid computing platform that supports state of the art features such as multi-job clustering, parallelism, job persistence, load distribution, job monitoring, and advanced scheduling algorithms. Since GenePattern is programming language agnostic, the executing nodes on the SGE are configured to support multiple programming languages, including R, Perl, C, C++, Java, and Bash. Jobs are sent off for processing via the SGE qsub utility.

In addition to hosting the GenePattern installation, the NAS server functions as central repository for storing intermediate files and analysis results. The SGE nodes and application servers are securely configured to access the NAS server. NAS communication is carried out via NFS v4 protocol.

Figure 20-1. GenePattern Software Architecture



20.2 GENE PATTERN ANALYSIS SUBMISSION

The GenePattern analysis task submission consists of 6 major steps:

- Authentication & Authorization
- Display the analysis modules the user has access to
- Select a module and enter parameters
- Schedule the task for execution
- Process and store the results
- Display the results

First, the user is authenticated and authorized with the appropriate level of access control. The user is then presented with all available modules and available options. After selecting the corresponding module, the system presents the user with a consistent GenePattern UI screen to collect metadata and required inputs. Upon submission of the task, the system sends a message via the qsub utility to SGE. SGE schedules the corresponding job for the next available node for job execution. The results are stored on the NAS server and the job status is updated accordingly. After job completion, the system presents the user with options to view or download results. At any point, the user can decide to cancel, view or delete any of the submitted jobs to which the user has access to.

This screen provides a summary overview of various types of Gene Expression analysis modules, both single analysis and pipeline. The search option is used to find the module of interest.

The screenshot shows the ImmPort website interface for the iGenePattern Analysis Modules. The top navigation bar includes links for About ImmPort, Admin, Access Data, Tools, Resources, and News & Events. A search bar is located at the top right. The main content area displays a search form titled "Search Options" with fields for Module: Category (Equal dropdown, value Preprocess-utilities), Module: Name (Like dropdown, empty input), and Module: Description (Like dropdown, empty input). Below the search form is a table listing analysis modules. The table has columns for "Module / Pipeline", "Core Function", and "Category". The table shows the following data:

Module / Pipeline	Core Function	Category
AffyCelProcessor	ImmPort version of ExpressionFileCreator to support more preprocessing methods. It creates the GenePattern formatted files (a GCT file and a CLS file) by preprocessing Affymetrix CEL files.	Preprocess-utilities
DataTransformation	Data scale transformation, row-wise (genes) or column-wise (samples) data normalization.	Preprocess-utilities
ExpressionFileCreator	Creates a RES or GCT file from a set of Affymetrix CEL files	Preprocess-utilities
IntersectLists	Create overlapped ID list from two individual ID lists.	Preprocess-utilities
Multi-pipelineComparison	Identify optimal microarray data workflow automatically using multiple pipeline comparison module	Preprocess-utilities
MultiplotPreprocess	Create derived data from an expression dataset for use in the Multiplot Visualizer module	Preprocess-utilities
PreprocessDataset	Provide several preprocessing options, including normalization, floor and ceiling thresholding, and variation filtering.	Preprocess-utilities

After a module is selected, the user is presented with the GenePattern analysis parameter entry screen (customized in the ImmPort stylesheets) in the central frame with the surrounding ImmPort menus and presentation.

iGenePattern (Gene expression analysis) (Beta) / Run Module

iGenePattern Home | Data Management | Analysis Pipelines | Analysis Modules | View Results | Help

CMS_ARACNE_NetworkAnalysis version 9

Show parameter descriptions

* required field

Run Reset properties | export | help

input filename* Specify URL Upload File
input filename - .res, .gct, .odf

filter flag Variation filter and thresholding flag

preprocessing flag Discretization or normalization flag

cls file* Specify URL Upload File
The class file - .cls

test direction* The test to perform

test statistic* The statistic to use

field The field to filter features on

min Select features with field >= min

max Select features with field <= max

field The field to filter features on

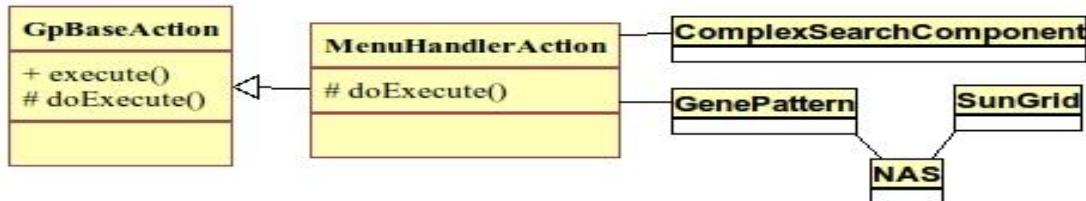
Recent Jobs

AffyCelProcessor (591) Oct 28 05:14:28 PM
EXP1.gct
EXP1.ds
stdout.txt
gp_execution_log.txt
EXP1.gct
EXP1.ds
stdout.txt
gp_execution_log.txt

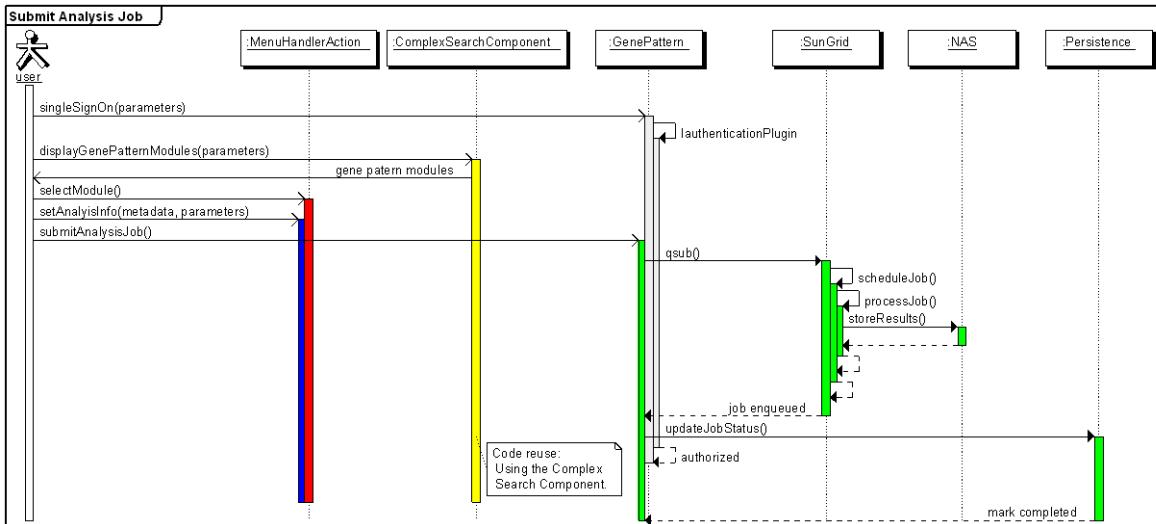
After an analysis task is completed, the GenePattern results interface is presented again within the context of the ImmPort menu structure.

Job Results								
Status	Job ▾	delete	Module Name	Total Size	Submit Date	Complete Date	Job Owner	Your Access
✓	636	□	ACP_SAM_2StepKMC_Classifi	7.5 MB	Sep 20 10:58:36 AM	Sep 20 11:21:11 AM	gp	
		□	↳ ACP_SAM_2StepKMC_Classifi_execution_log.html	16.0 kB		Sep 20 11:21:11 AM		
✓	637		↳ 1. AffyCellProcessor		Sep 20 11:58:44 AM	Sep 20 11:12:17 AM		
		□	↳ gp_execution_log.txt	1.0 kB		Sep 20 11:12:18 AM		
		□	↳ GSE5681_ACPdefault_Affy.cls	1.0 kB		Sep 20 11:12:17 AM		
		□	↳ GSE5681_ACPdefault_Affy.gct	4.6 MB		Sep 20 11:12:17 AM		
		□	↳ stdout.txt	3.0 kB		Sep 20 11:12:17 AM		
✓	638		↳ 2. ImmPort_SAMr		Sep 20 11:12:32 AM	Sep 20 11:13:41 AM		
		□	↳ gp_execution_log.txt	1.0 kB		Sep 20 11:13:41 AM		
		□	↳ GSE5681_ACPdefault_Affy_SAMdt28.class_indicator4sig.txt	8.0 kB		Sep 20 11:13:37 AM		
		□	↳ GSE5681_ACPdefault_Affy_SAMdt28.delta.txt	4.0 kB		Sep 20 11:13:37 AM		
		□	↳ GSE5681_ACPdefault_Affy_SAMdt28.gct	161.0 kB		Sep 20 11:13:37 AM		
		□	↳ GSE5681_ACPdefault_Affy_SAMdt28.list.txt	9.0 kB		Sep 20 11:13:37 AM		
		□	↳ GSE5681_ACPdefault_Affy_SAMdt28.sig.txt	129.0 kB		Sep 20 11:13:37 AM		
		□	↳ sam.png	19.0 kB		Sep 20 11:13:40 AM		
		□	↳ stdout.txt	2.0 kB		Sep 20 11:13:41 AM		
✓	639		↳ 3. ImmPort_2StepKmeans		Sep 20 11:13:42 AM	Sep 20 11:13:48 AM		
		□	↳ 2StepKMean_cluster.txt	19.0 kB		Sep 20 11:13:47 AM		
		□	↳ gp_execution_log.txt	1.0 kB		Sep 20 11:13:48 AM		
		□	↳ stdout.txt	1.0 kB		Sep 20 11:13:47 AM		
✓	640		↳ 4. ImmPort_Classifi		Sep 20 11:13:48 AM	Sep 20 11:18:41 AM		
		□	↳ Annotations.txt	618.0 kB		Sep 20 11:18:39 AM		
		□	↳ gp_execution_log.txt	1.0 kB		Sep 20 11:18:41 AM		
		□	↳ Results.txt	1.3 MB		Sep 20 11:18:39 AM		
		□	↳ stdout.txt	1.0 kB		Sep 20 11:18:40 AM		
		□	↳ Summary.txt	2.0 kB		Sep 20 11:18:39 AM		

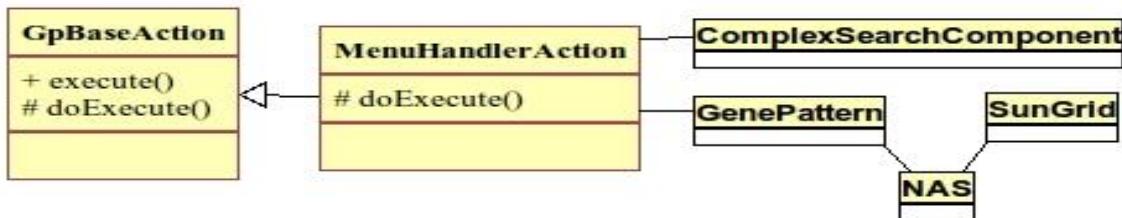
20.2.1 Gene Pattern Analysis Submission Class Diagram



20.2.2 Gene Pattern Analysis Submission Sequence Diagram

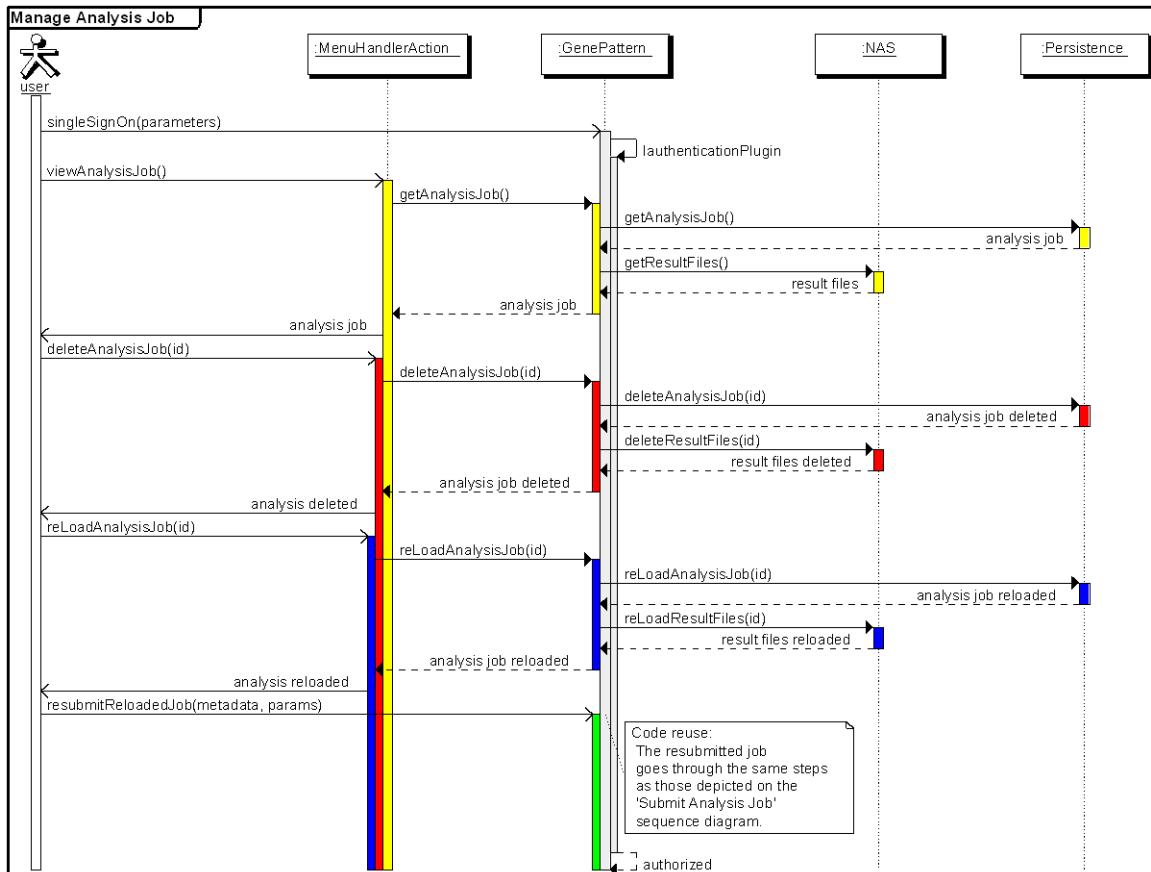


20.2.3 Gene Pattern Analysis Job Management Class diagram

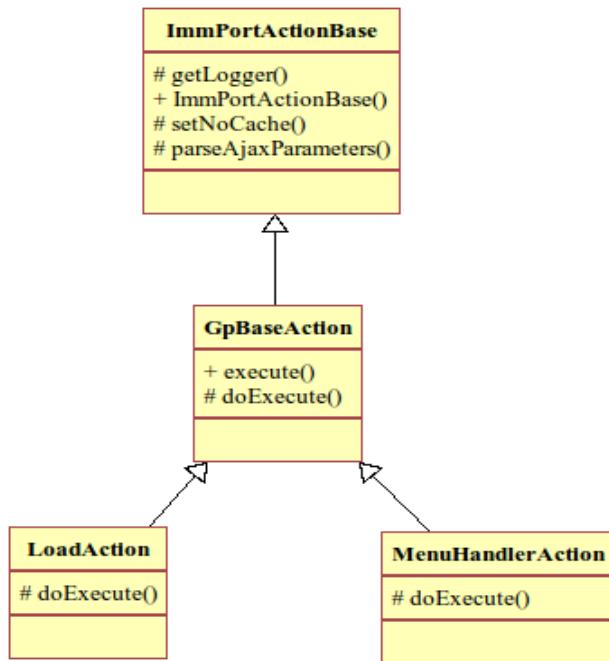


20.2.4 Gene Pattern Analysis Job Management Sequence diagram

Out of the box, Gene Pattern provides integrated analysis job management, allowing scientists to view the job history and individual results, delete results, and reload previously executed tasks. The diagrams below depict the interaction for viewing, deleting and reloading tasks. The reload function is particularly useful because it enables scientists to re-execute a previously analysis run, while modifying one or more parameters.



20.2.5 Gene Pattern Presentation Tier Class Diagram



20.2.6 Gene Pattern Business Tier Class Diagram

The GenePattern component re-utilizes the Advanced Search component for database access. Please, see the Advanced Search component class diagram.

20.2.7 Gene Pattern DAO Tier Class Diagram

The GenePattern component re-utilizes the Advanced Search component for database access. Please, see the Advanced Search component class diagram.

21.0 PED FILE GENERATION DESIGN PACKAGES

The ImmPort genetic analysis data set generation utility (PED Generation) allows you to create the .ped (pedigree) and .info (marker information) files for the ‘pre MAKEPED’ linkage file format commonly used in genetic analysis algorithms. The following screen shot displays the home page for PED Generation.

The screenshot shows the ImmPort PED Generation home page. At the top, there is a horizontal navigation bar with four steps: "Assign Subjects to Affected and Unaffected" (highlighted in grey), "Validate Subjects for Linkage File Generation", "Choose SNPs to include", and "Review Generated Linkage File Results". Below this is a section titled "Selected Projects" and "Change Projects". A note states: "The ImmPort genetic analysis data set generation utility (PED Generation) allows you to create the .ped (pedigree) and .info (marker information) files for the 'pre MAKEPED' linkage file format commonly used in genetic analysis algorithms." It also says: "You will be creating a .ped and .info file set from genotype data stored in ImmPort. The PED generation process has the following three steps:"

- 1. Choose Subject Lists & Assign Subjects To Phenotype Group:**

The subjects whose genotyping results that you wish to include in the linkage file should be defined in a user defined ImmPort SUBJECT lists. You can generate SUBJECT lists from the [Research Data Advanced Search](#) results. More details on [how to make an ImmPort list](#) are available.

The .ped file format supports three (3) phenotypic groups for association analysis:

 - Affected (case)
 - Unaffected (control)
 - Unknown

You may assign a SUBJECT list to a phenotype group by selecting a radio button in the appropriate column. One or more SUBJECT lists may be assigned to a phenotype group.
- 2. Validate Subjects:**

When you have assigned all of the lists you wish to use to their appropriate group, please click the "Validate" button. This will evaluate whether any of the subjects in the lists is assigned to more than one phenotype group. If no subjects are assigned to more than one phenotype group, then the PED generation tool evaluates which subjects in each group meet the [PED file generation criteria](#) and which do not.
- 3. Choose SNPs to be included in the dataset. There are three options:**
 - SNPs in common between all selected results.
 - All SNPs present in the result sets.
 - Filter SNPs in the result sets using a designated SNP list.

Notice:

- For PED generation, a user can determine the number of subjects and SNPs to be used to generate a PED file. For the current system, we have successfully tested up to 50 millions data-points. Data-points are defined by the formula: Data-points = #subjects x #SNPs.
- ImmPort currently supports generating non-family based PED files.
- Newly submitted data will be available for PED generation on the next day after submission.

Below this is a table titled "Subject List Name" showing five rows of subject lists and their status (Case, Control, Unknown). The table includes a "Validate" and "Clear" button at the top.

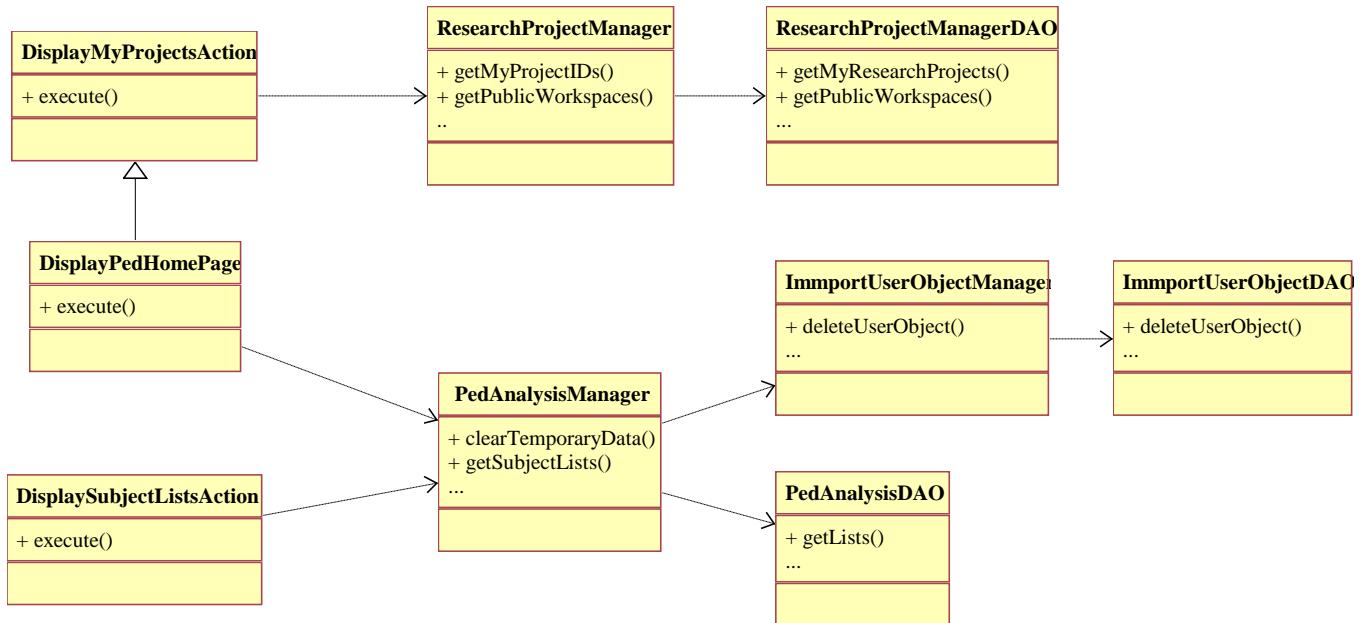
Subject List Name	Description	Case	Control	Unknown
1500 good subjects	These subject should be al...	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
2010 AS subject list1	1st 5	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
2010 AS subject list1-5	total - 5	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
2010 AS subject list116-120	total - 5	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
2010 AS subject list116-1	total - 5	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

21.1 CHOOSE SUBJECT LISTS

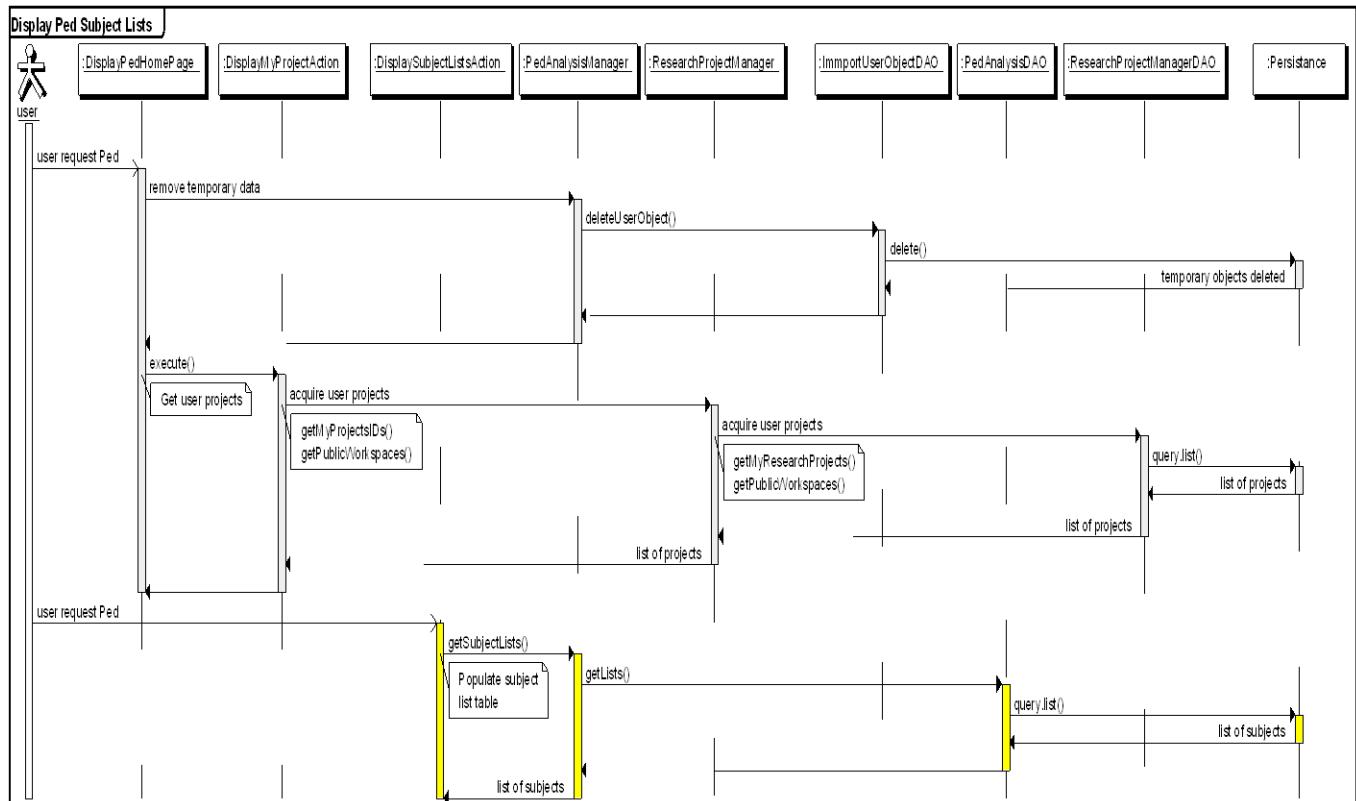
On the PED Generation home page, the user selects the subject lists that will be assigned to the case, control, and unknown groups. Not all groups need have subject lists selected. Each group must have a set of subjects that do not appear in the other two groups. If a two of the groups have common subjects, then a group validation error occurs and a validation error table appears with the subjects in validation showing the subject lists from the two groups containing the common subjects. If this validation succeeds (no common subjects among the groups), then the results proceeds to the subject validation page,

After subject selection, if there is a validation error the user is presented with the interface below.

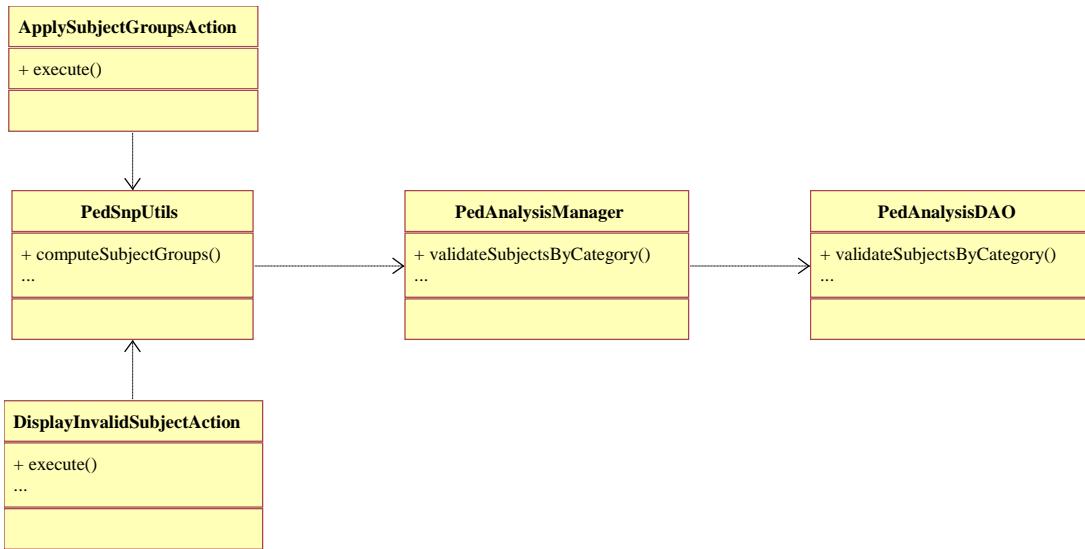
21.1.1 Choose Subject Lists Class Diagrams



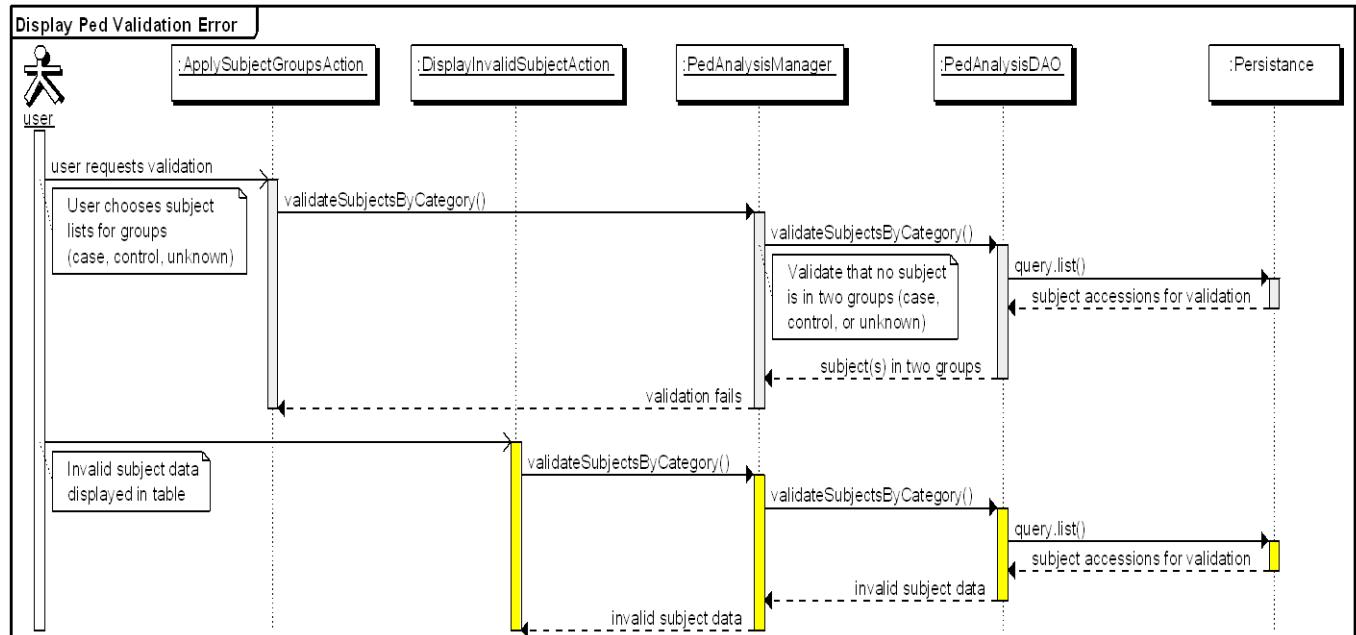
21.1.2 Choose Subject Lists Sequence Diagrams



21.1.3 Validation Error Class Diagrams



21.1.4 Validation Error Sequence Diagrams



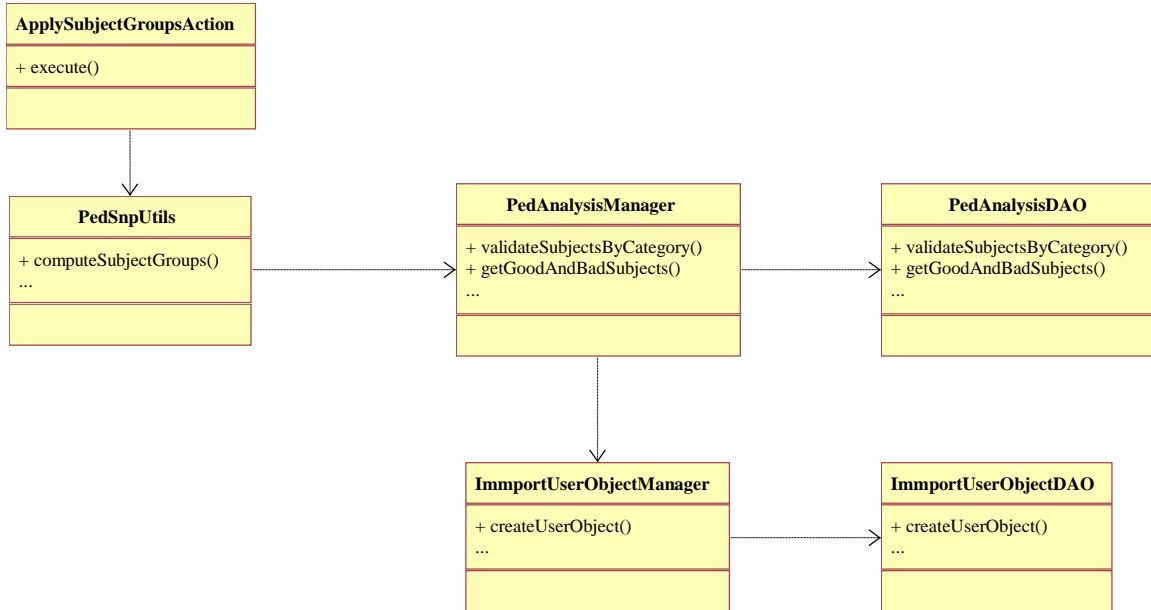
21.2 SUBJECT VALIDATION RESULTS

The subject validation results page displays the subjects in each group (case, control, and unknown) binned into subjects that meet all requirements for PED Generation and those subjects that do not. Each binned set of subject can be viewed on subject lists page. This page provides the details why subjects meet or do not meet PED Generation requirements. When a user proceeds to the next step from the subject validation page, he goes to the SNP selection page.

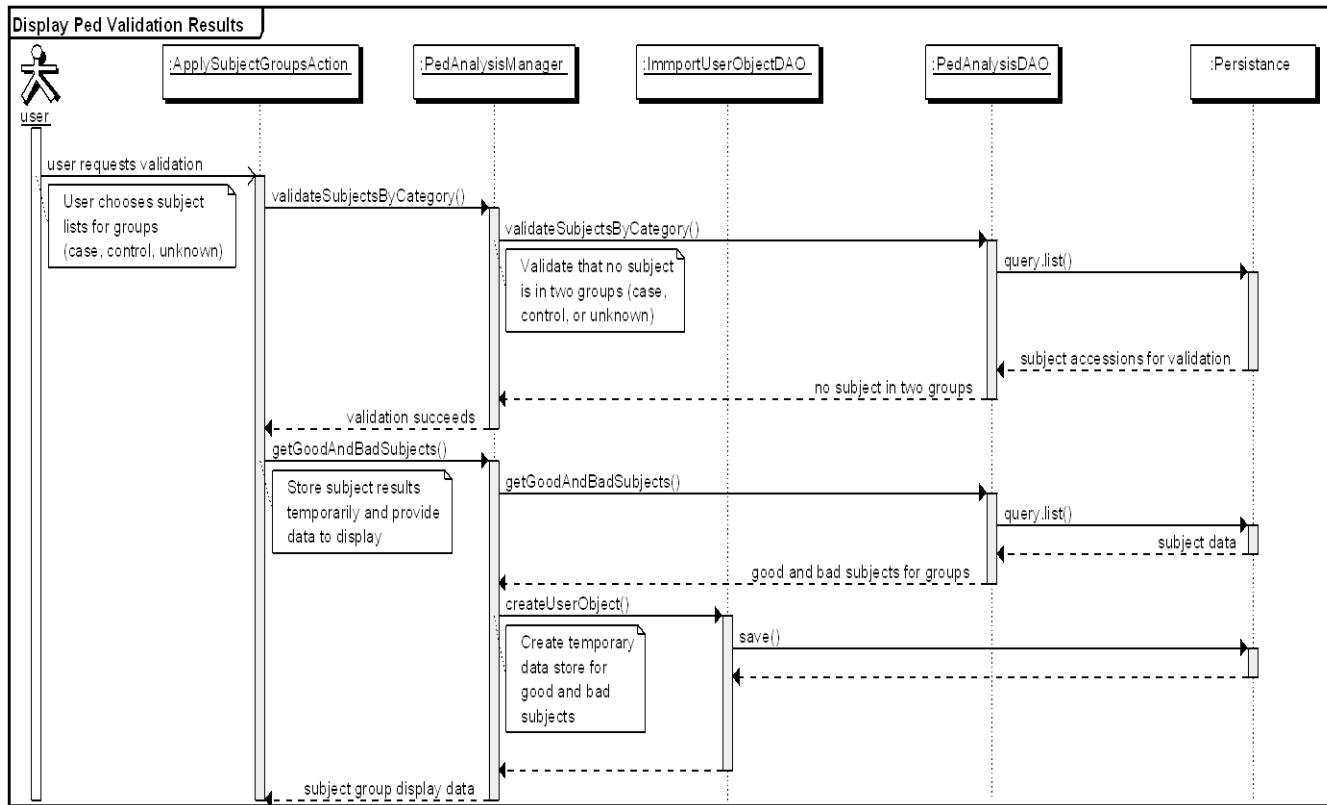
The screenshot shows the ImmPort interface for subject validation. At the top, there's a navigation bar with links for 'Edit Profile' and 'Sign Out'. Below the header, a breadcrumb trail shows 'Analysis / Subject Validation Results for Linkage File'. A horizontal flowchart indicates the process: 'Assign Subjects to Affected and Unaffected' → 'Validate Subjects for Linkage File Generation' → 'Choose SNPs to include' → 'Review Generated Linkage File Results'. Under the first step, it says 'Affected / Case' with 1500 subjects meeting criteria and 0 not meeting. Under 'Unaffected / Control', it shows 4 subjects meeting criteria and 0 not meeting. Under 'Unknown', it shows 0 subjects meeting criteria and 0 not meeting. At the bottom are 'Previous', 'Next', and 'Cancel' buttons.

If a user clicks on show details, he is presented with a standard summary grid for subject data as shown below.

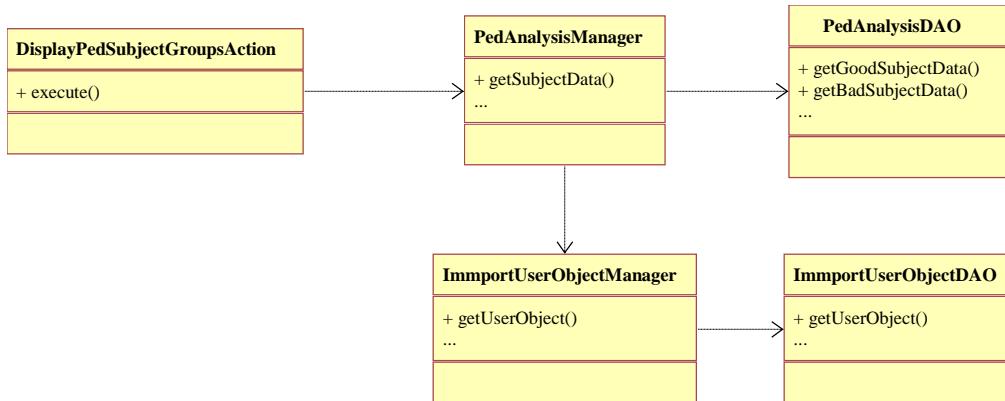
21.2.1 Subject Validation Results Class Diagrams



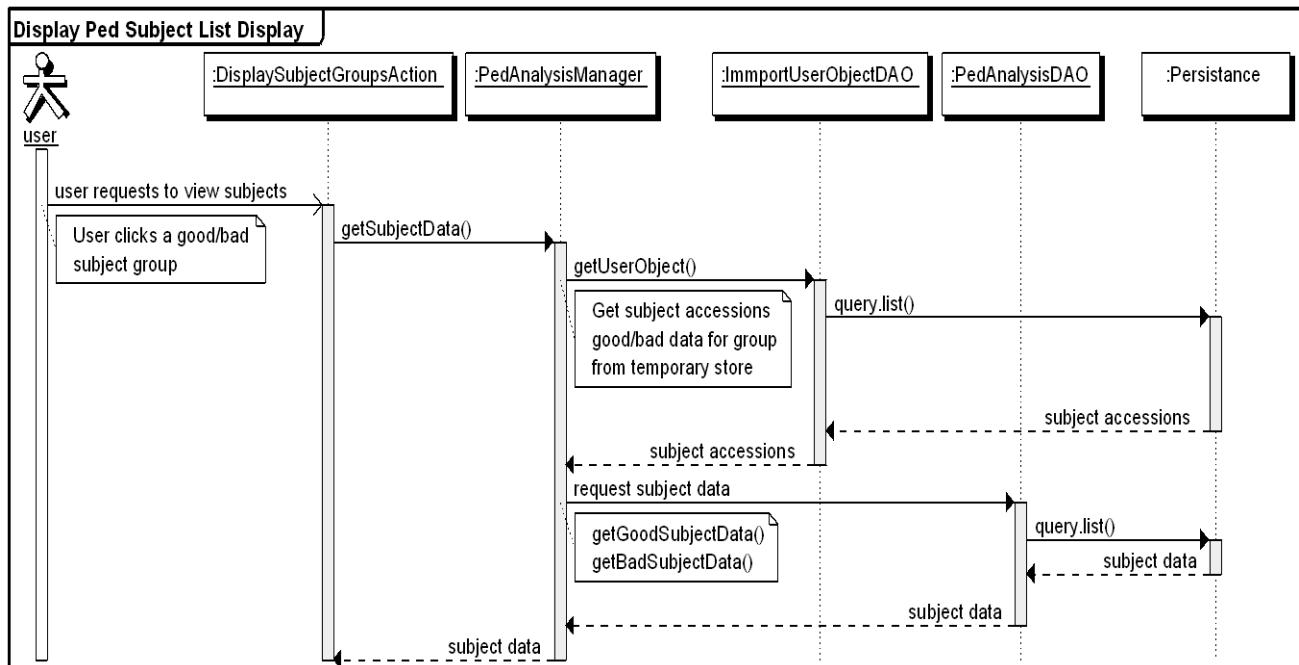
21.2.2 Subject Validation Results Sequence Diagrams



21.2.3 Subject List Class Diagrams



21.2.4 Subject List Sequence Diagrams



21.3 SNP SELECTION

The SNP selection page displays the allocation of SNP markers to the various reagents used in the SNP assay results for the set of all subjects chosen (all cases, controls, and unknowns). To generate the PED files (.ped and .info), the user can use all SNPs determined from all the SNP assays displayed, or only the common SNPs, or use a SNP list previously stored, or an uploaded file containing a SNP list. Once the parameters have been chosen on the page, the user proceeds to PED Generation.

Assign Subjects to Affected and Unaffected → Validate Subjects for Linkage File Generation → Choose SNPs to include → Review Generated Linkage File Results

The subjects you selected and that meet all .ped file generation criteria are linked to the result sets(experiment samples) listed below. The genotype assay platforms linked to the result sets are also listed. Please select one or more result sets to include in the .ped file.

Exp UserID (Affy)/ Exp Sample UserID (Illumina)	Genotype Platform	Number of SNPs with Genotype Results *
Illumina_FHCRC_NMDP_USA ...	Illumina MHC-Exon-Centric Panel, Illumina MHC-Exon-Centric Panel	1206
E-Case-Control	Affymetrix Mapping 500K Array Set - Combined 250K Nsp and 2...	20

* The values in the column 'Number of SNPs with Genotype Results' are computed using the total record counts for the experiment samples in an experiment. As a result, the numbers may not match the dbSNP rs accession counts provided below.

The number of dbSNP rs accessions that are found in common between all of the result sets is : 0
The total number of dbSNP rs accessions that are found in all the result sets is : 1246

Fields marked with an asterisk * are required.

Generate the .ped file from:

SNPs in common between all selected results
 All SNPs present in the result sets

If you want to use an Import SNP list to further select the SNPs to include in the .ped file, please choose an Import SNP list to select the SNPs. If you need to create the desired SNP list, please go to [Advanced Reference Search](#). On return from the advance reference query page you will need to refresh this page so that the SNP list will appear in the SNP lists menu below:

NO SNP LIST
OR
 If you want to use your own SNP list to further select the SNPs to include in the .ped file, please select your file:
 No file chosen

Select a research project below to save the PED results to:*

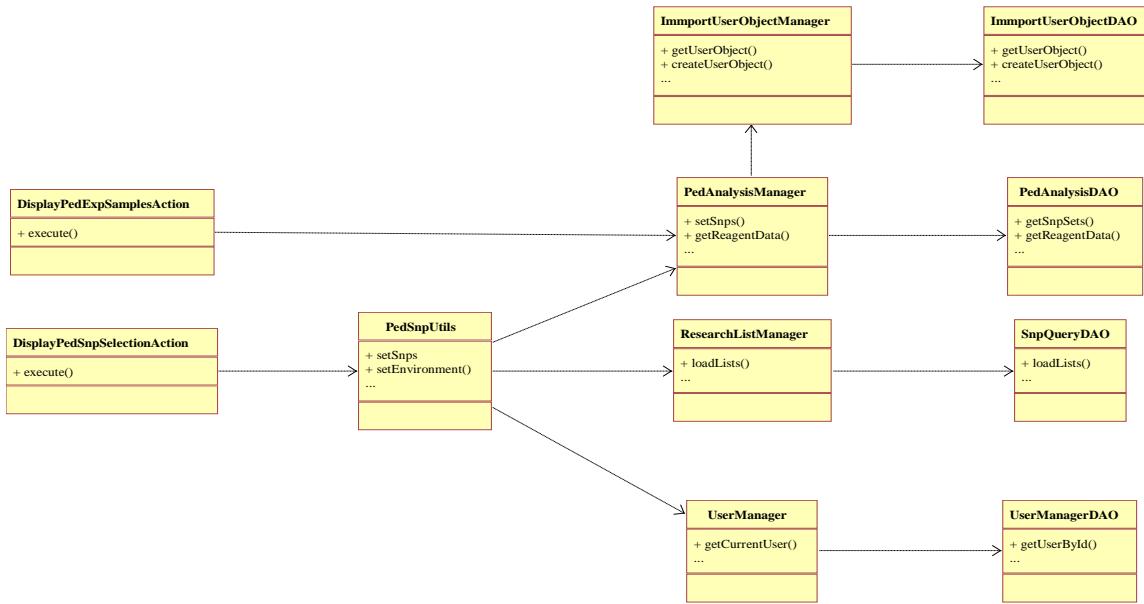
03099: HLA Typing Result upload Error

Dataset Name: *

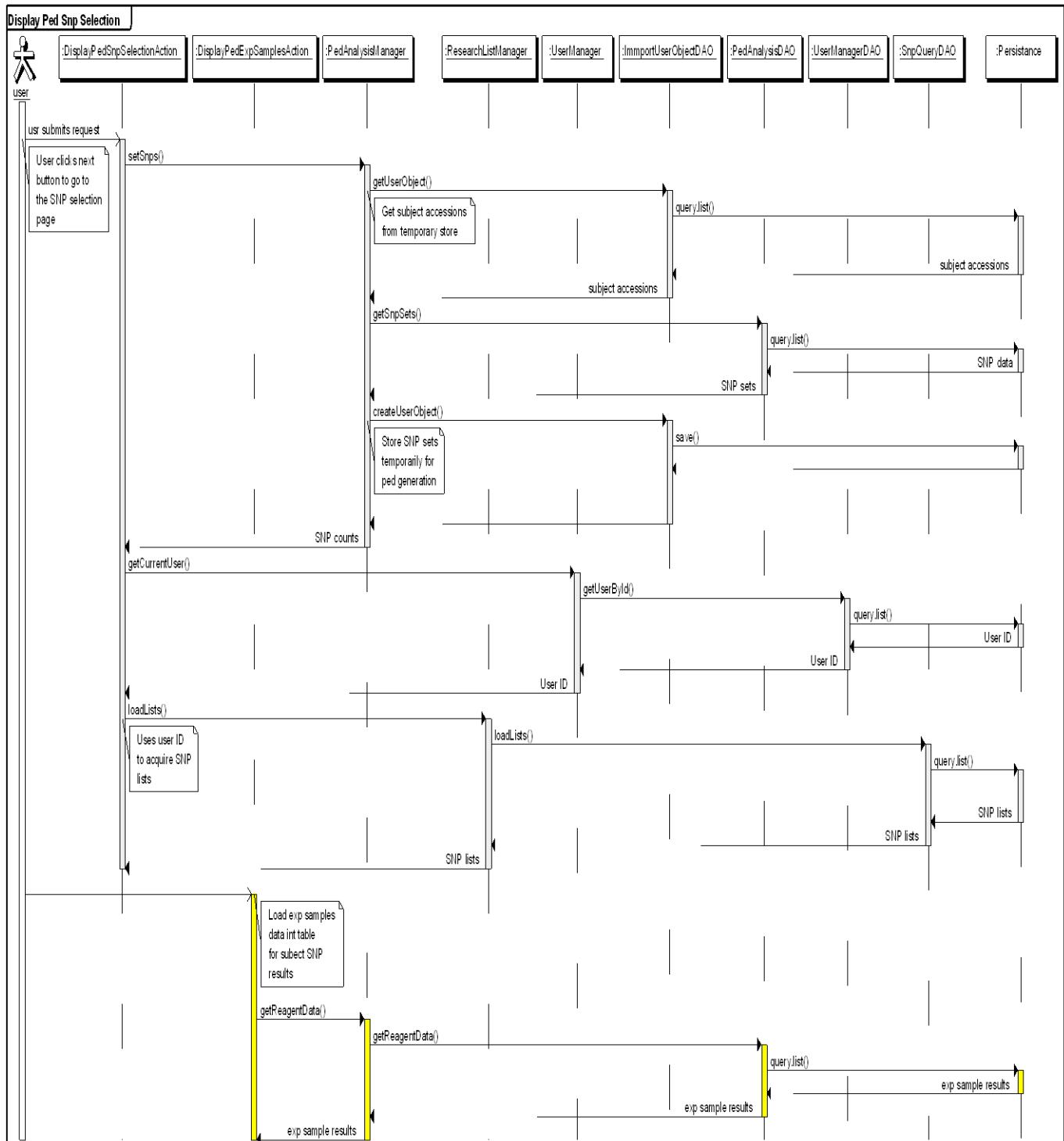
Dataset Description: *

Analysis Type:*

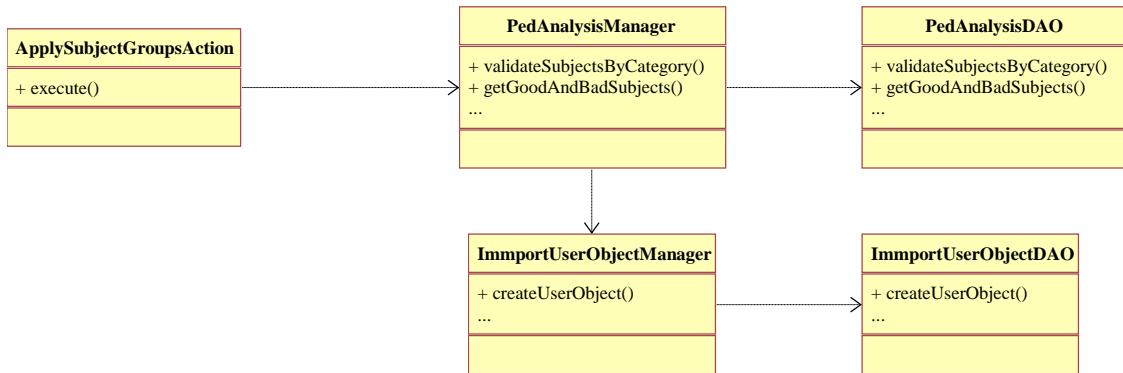
21.3.1 SNP Selection Class Diagrams



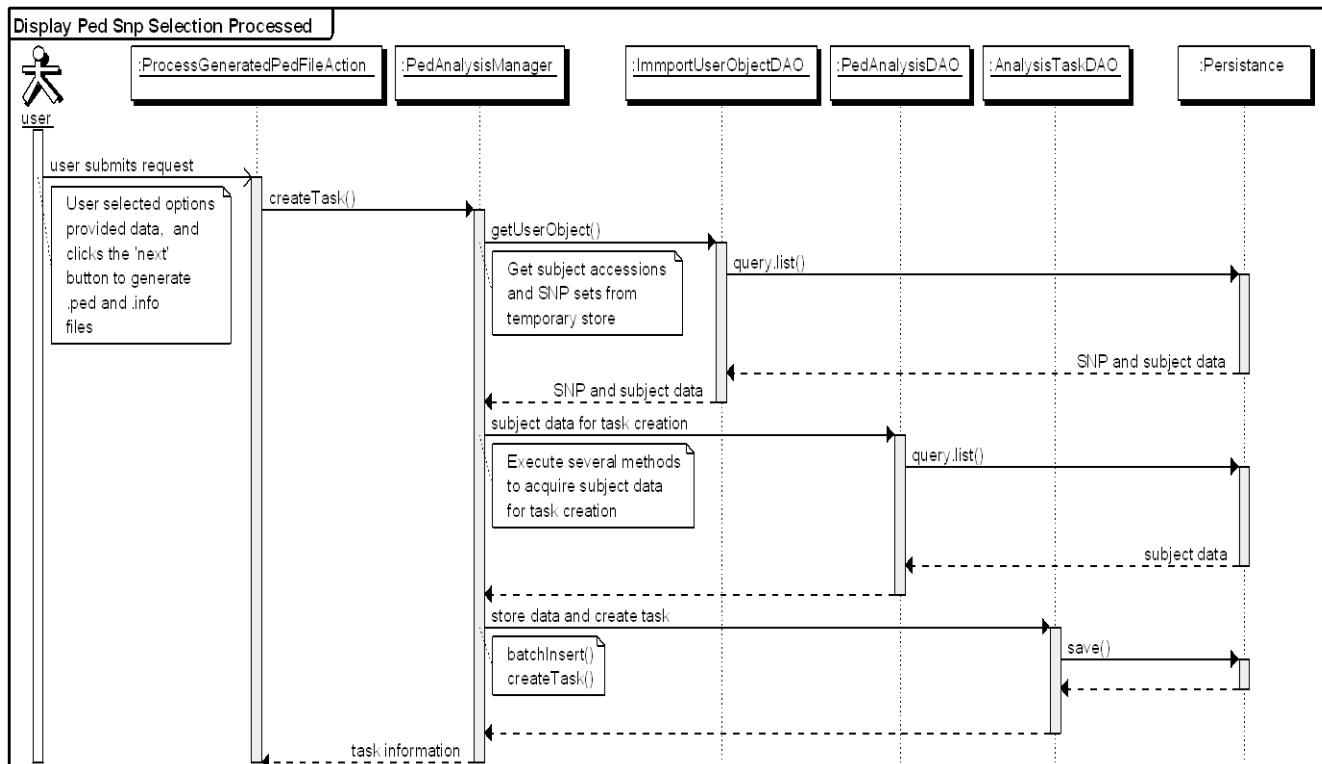
21.3.2 SNP Selection Sequence Diagrams



21.3.3 SNP Selection Processed Class Diagrams



21.3.4 SNP Selection Processed Sequence Diagrams



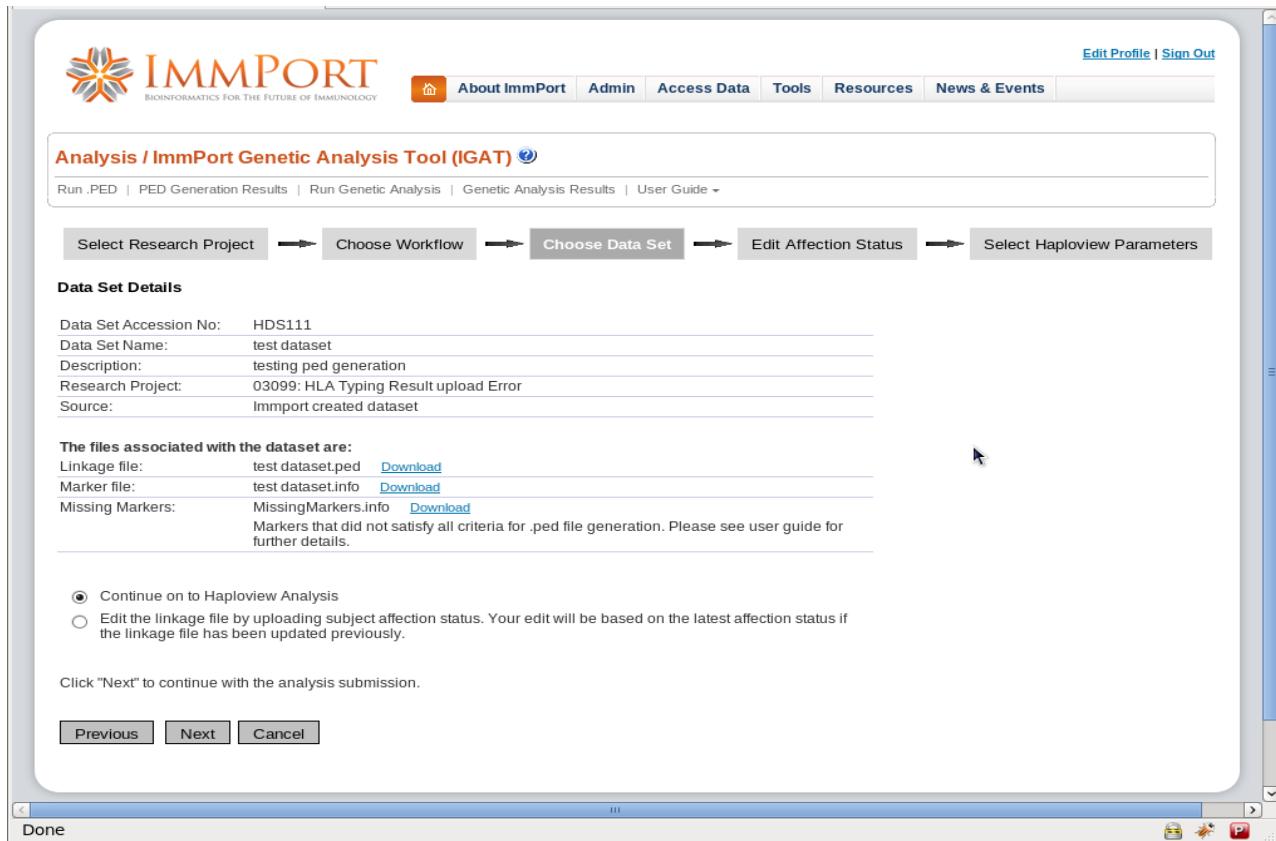
21.4 PED GENERATION RESULTS

The PED Generation results page displays the results of PED Generation and genotype analysis results (e.g., Haploview). The user can click on the PED Generation result and get the PED Generation results detail page.

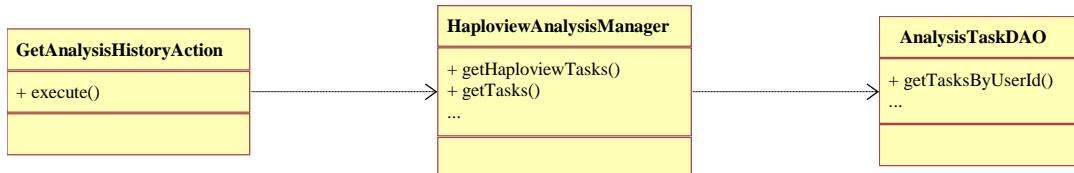
This screenshot shows the ImmPort Analysis / IGAT Analysis History page. The page header includes the ImmPort logo and navigation links for About ImmPort, Admin, Access Data, Tools, Resources, News & Events, Edit Profile, and Sign Out. Below the header, a message states: "This page is a list of the tasks you have submitted to the ImmPort Genetic Analysis Toolset. The history list can show up to 50 tasks. If the number exceeds 50, the oldest tasks will be dropped off the list. Click on a task ID to view more information about the task or to see the results when the task has completed execution." A table below lists 14 items found, displaying all items. The table columns are Task Id, Start Date, End Date, Task Type, Data Set Accession/Name, Dataset Source, Algorithm, and Status. The tasks listed include various operations such as Create_PED_File, Run_Haploview, and Update_Subject_Affection_Status, mostly completed successfully.

Task Id	Start Date	End Date	Task Type	Data Set Accession/Name	Dataset Source	Algorithm	Status
2613	10/20/2010 16:12:27	10/20/2010 16:35:20	Create_PED_File	HDS111 / test dataset	import_created	n/a	Completed
860	11/13/2009 08:00:07	11/13/2009 08:00:20	Run_Haploview	HDS44 / a	import_created	Haploview	Completed
859	11/13/2009 07:57:51	11/13/2009 07:57:54	Create_PED_File	HDS44 / a	import_created	n/a	Completed
826			Update_Subject_Affection_Status			n/a	n/a
825	10/29/2009 15:50:23	10/29/2009 15:50:41	Run_Haploview	HDS41 / a test file	custom_upload	Haploview	Completed
822	10/29/2009 14:18:28	10/29/2009 14:18:48	Run_Haploview	HDS40 / a test dataset to check updates	import_created	Haploview	Completed
820	10/29/2009 14:17:00	10/29/2009 14:17:02	Create_PED_File	HDS40 / a test dataset to check updates	import_created	n/a	Completed
806	10/27/2009 13:38:42	10/27/2009 13:38:52	Run_Haploview	HDS39 / subject names with spaces	import_created	Haploview	Completed
805	10/27/2009 13:32:15	10/27/2009 13:32:18	Create_PED_File	HDS39 / subject names with spaces	import_created	n/a	Completed
804	10/26/2009 15:57:51	10/26/2009 15:57:53	Create_PED_File	HDS38 / spaces and null family	import_created	n/a	Completed
742	10/22/2009 09:29:29	10/22/2009 09:29:37	Run_Haploview	HDS37 / zipping dataset updated (fixed subject names)	custom_upload	Haploview	Completed
741	10/22/2009 09:26:49	10/22/2009 09:26:55	Run_Haploview	HDS36 / from zipping	custom_upload	Haploview	Completed with error
716	10/21/2009 15:23:21	10/21/2009 15:23:41	Run_Haploview	HDS11 / non-family dataset	import_created	Haploview	Completed
375	09/16/2009 10:22:15	09/16/2009 10:22:20	Create_PED_File	HDS11 / non-family dataset	import_created	n/a	Completed

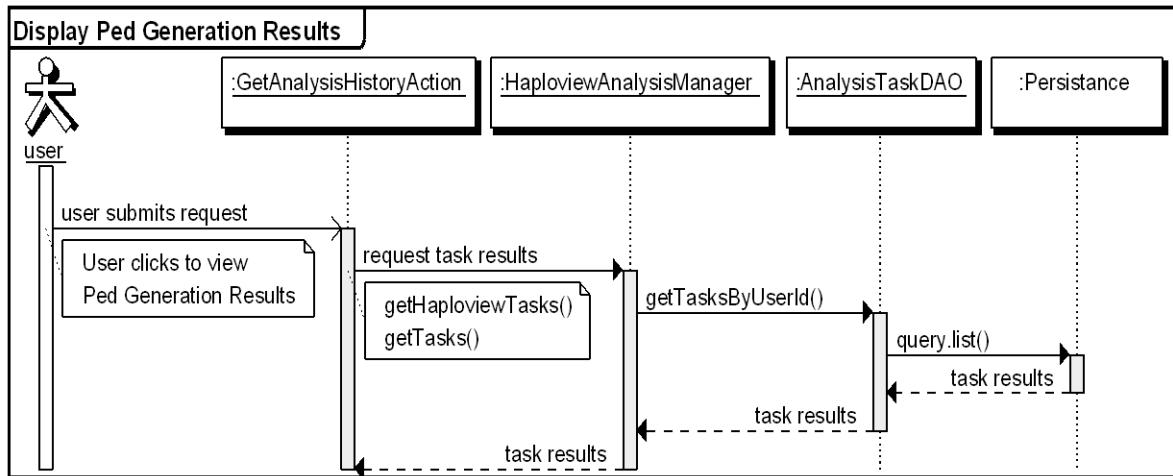
From the detail page, the user can proceed to Haploview analysis.



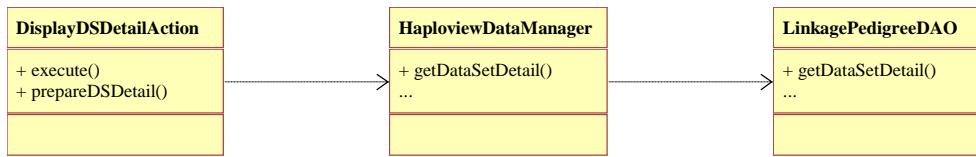
21.4.1 PED Generation Results Class Diagrams



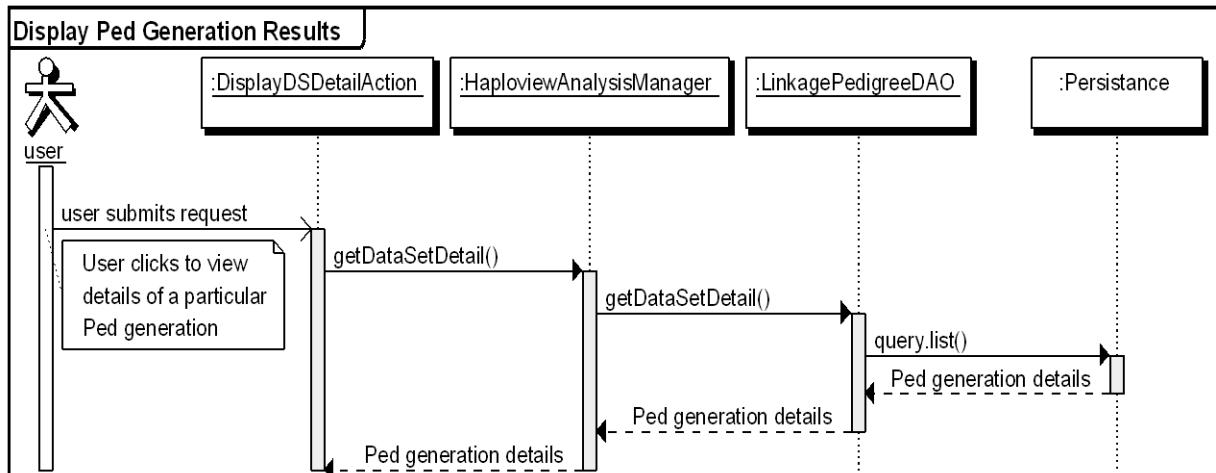
21.4.2 PED Generation Results Sequence Diagrams



21.4.3 PED Generation Results Detail Class Diagrams

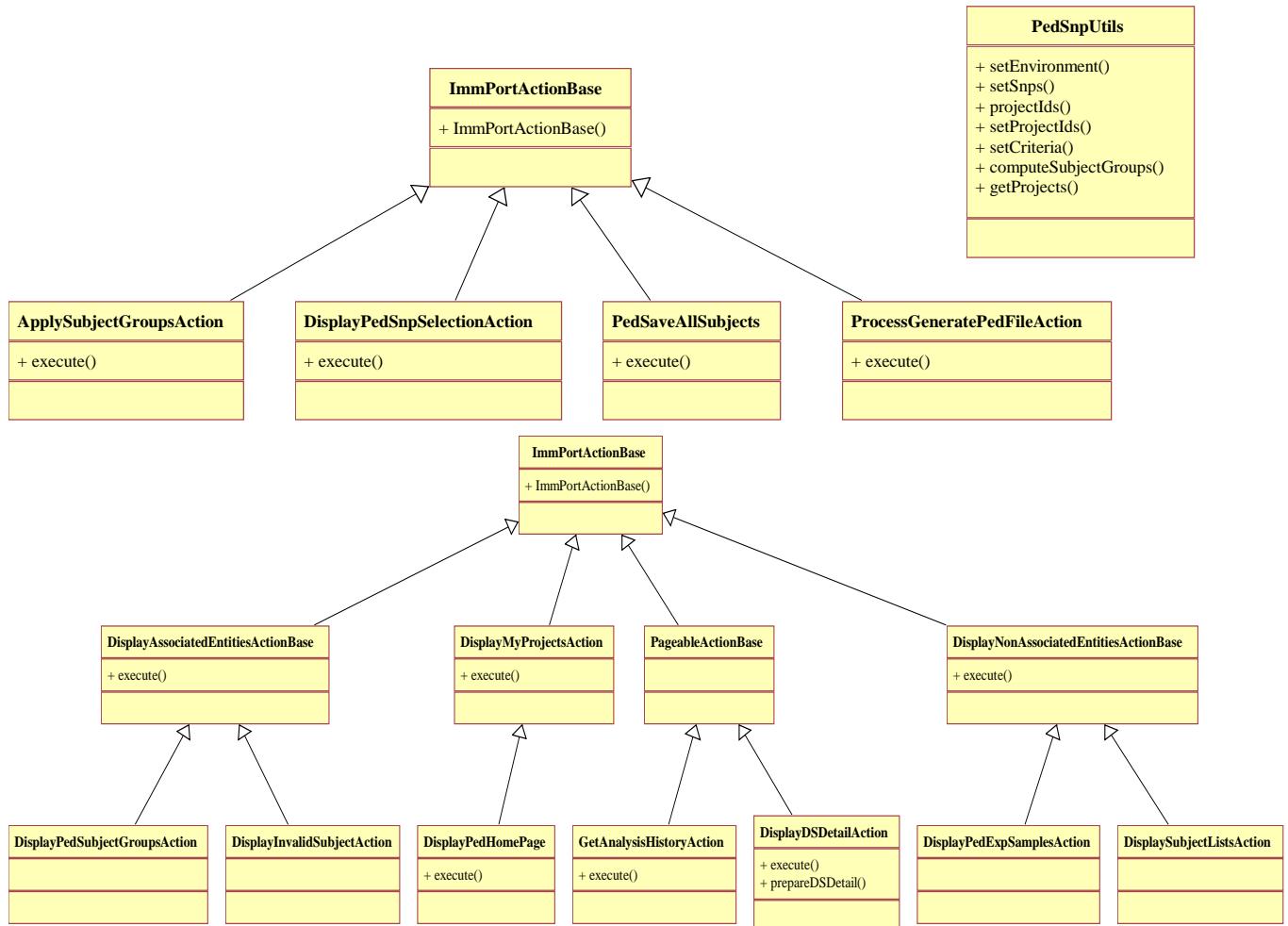


21.4.4 PED Generation Results Detail Sequence Diagrams

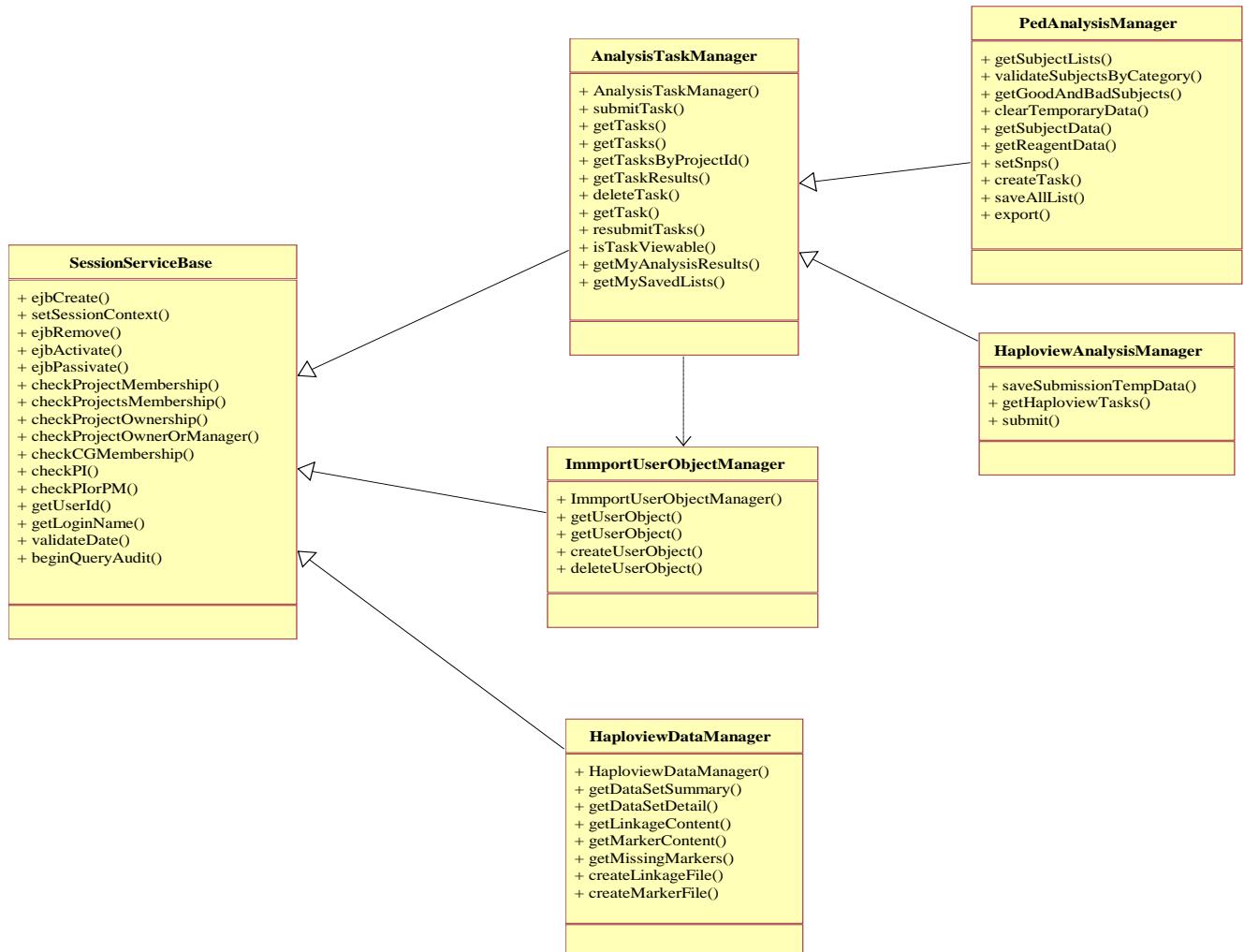


21.5 PED FILE GENERATION CLASS DIAGRAMS

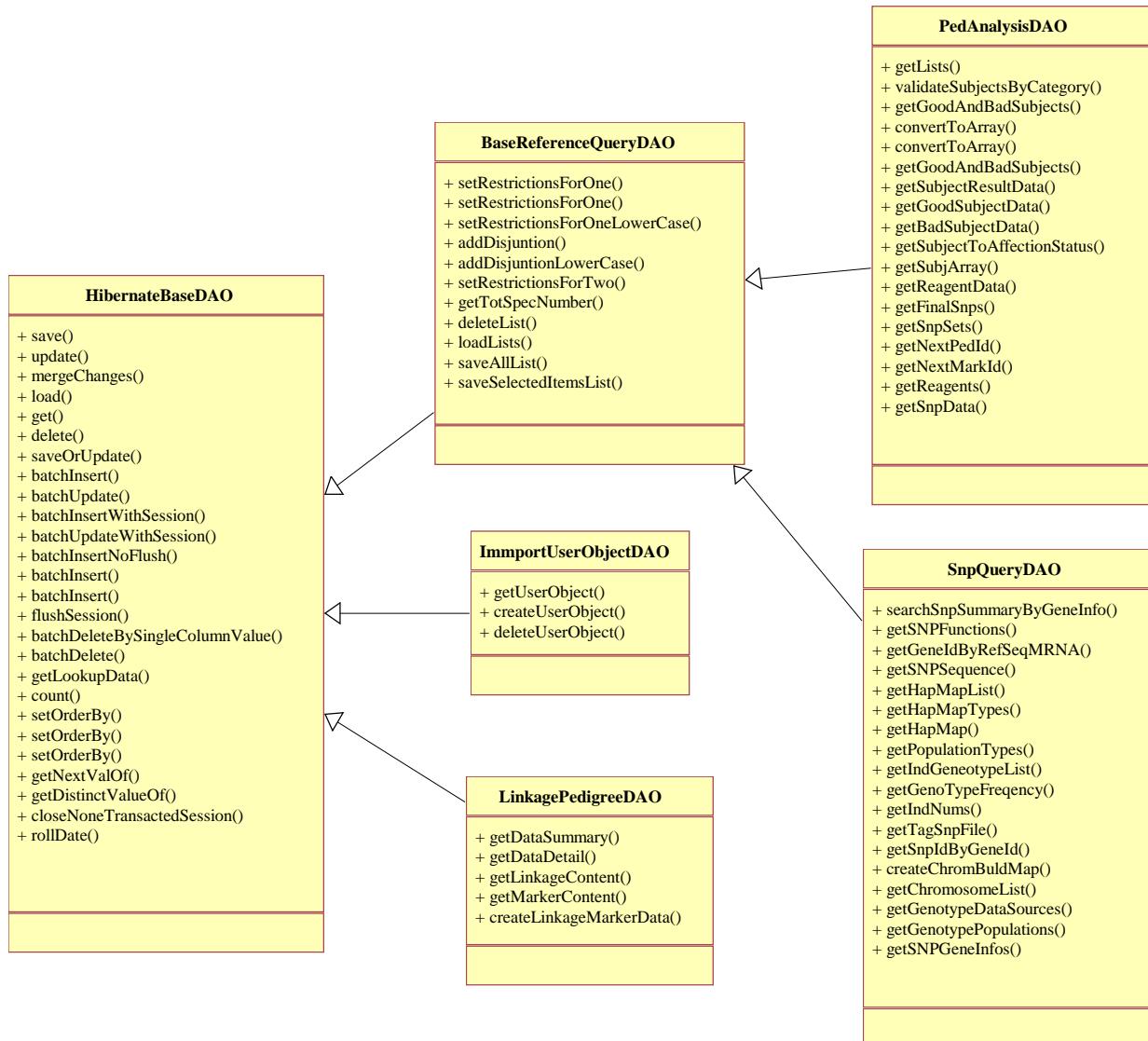
21.5.1 Actions



21.5.2 EJB



21.5.3 DAO



22.0 HAPLOVIEW DESIGN PACKAGES

22.1 SUBMIT HAPLOVIEW ANALYSIS

The ImmPort Genetic Analysis Tool (IGAT) allows users to analyze their genotype data. The algorithms implemented in IGAT are adapted from the Haploview algorithm developed by Mark Daly's laboratory at the Broad Institute (<http://www.broadinstitute.org/sections/science/programs/medical-and-population-genetics/haploview>).

ImmPort currently supports the following functionalities adapted from Haploview:

- Linkage Disequilibrium (LD) and haplotype block analysis
- Haplotype population frequency estimation
- Single SNP and haplotype association tests on population-based case/control studies
- Permutation testing for association significance

The desktop Haploview application can be run on dataset of more than 20,000 markers, but it cannot export the LD plot if the dataset is more than ~1000 markers. The ImmPort implementation of Haploview exports LD plots to the ImmPort web interface. To generate these very large LD plots, the entire image is created in memory and is sliced into several smaller pieces which are exported. The ImmPort enterprise version of Haploview can handle datasets consisting up to about 3000 markers.

The Haploview analysis task submission consists of 6 major steps:

- Authentication & Authorization
- Selecting a research project for accessing data and storing results
- Use an existing dataset or upload a new dataset
- Schedule the task for execution
- Process and store the results
- Display the results

First, the user is authenticated and authorized and granted the appropriate level of access control. The user is then presented with the option to use an existing dataset or upload a new dataset.

The ImmPort implementation of Haplovew accepts [linkage files \(.ped\)](#) and [marker info files \(.info\)](#) as Input Data Sets. Click [here](#) to view a detailed description of the two formats.

Please make a selection from the options below. One option is to use an existing ImmPort dataset for Haplovew analysis. Another is to upload a dataset from your computer. You can create Haplovew input dataset (linkage and marker info files) from ImmPort geno-typing results by using [PED file generation tool](#). Please note that currently only genotype data in the format of Illumina BeadStudio Final report or Affymetrix genotype .txt file are available for conversion to linkage and marker files.

Workflow Selection:

Run Haplovew using existing Linkage File and Marker Info file stored in ImmPort.
 Upload Linkage File and Marker Info File and Run Haplovew analysis.

[Previous](#) [Next](#) [Cancel](#)

After making the dataset choice, the system presents the user with a consistent Haplovew UI screen to collect the analysis metadata and required inputs.

Input File

Ignore pair wise comparison of markers >: 500 Kb apart
Exclude individuals with > : 50 % missing genotypes

Blocks Definitions

Confidence Intervals (Gabriel et al.)
 Four Gamete Rule
 Solid Spine of LD

LD Plot Color Scheme

4th Gamete
 Confidence Bounds
 R Squared
 Standard D' / LOD
 Alt D' / LOD
 GOLD

Alleles Display Mode

Alleles as colored squares
 Alleles as letters
 Alleles as numbers

Association Tests

Association Tests (Single Marker)
 Association Tests (Haplotypes)

Permutation Tests

Perform Permutations for Association Tests Single Marker
 Perform Permutations for Association Tests Single Markers and Haplotypes in Blocks
Number of permutation tests: 1000 (a large number may take a long time to run)

Blocks

Block Definition

Gabriel et al.

Upper confidence interval value: 0.98
Lower confidence interval value: 0.7
Upper confidence interval max for strong recombination value: 0.9
Min fraction of strong Id in info comparison: 0.95
Exclude marks below: 0.05 MAF

4th gamete rule

Fourth gamete frequency min: 0.01

Strong SD Spine

Upon submission of the task, the job is scheduled and processed accordingly. The results are stored on the NAS server and the job status is updated. After job completion, the system presents the user with options to view or download results.

IGAT Analysis Result Detail

Task Id	Start Date	End Date	Status
2371	08/21/2009 18:54:24	08/21/2009 18:54:49	Completed
Algorithm:	Haploview		
Data Set Accession:	HDS152		
Data Set Name:	ZP test PED for 2.5, SNP upload		
Data Set Description:	comm and SNP upload		

Result: Displayed below. The ImmPort enterprise version 2.3 of Haploview can handle datasets consisting up to about 100,000 markers; The viewable LD plot size limit (browse and png file displays) is ~1000 markers. There is a 4 Gbyte limit for memory allocation to the algorithm.

Haploview Result

[Expand All Nodes](#)
[Collapse All Nodes](#)

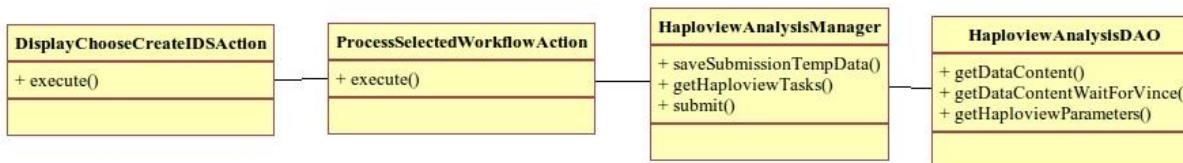
Parameters

No Block	false
Assoc Test Type	CC
Exclude Individuals Missing Genotypes	50
Ignore Pairwise Comp Markers Threshold	500
Number Of Permutation Tests	1000
Allele Thick Lines Threshold	10
Allele Thin Lines Threshold	1
Exclude Marks Below	0.05
Extend Spine Id D Prime	0.8
Forth Gamete Freq Min	0.01
Lower Confidence Interval Value	0.7
Min Fraction Of Strong L D In Info Comparison	0.95
Upper Confidence Interval Max For Strong Recomb	0.9
Upper Confidence Interval Value	0.98
Use Alleles As Colors	true
Use Alleles As Letters	false
Use Alleles As Numbers	false
Use Assoc Tests Single	false
Use Assoc Tests Single Plus Blocks	false
Use Block As Fourth Gamete	true
Use Block As Gabriel	true
Use Block As Solid Spine	true
Use L D Plot As Fourth Gamete	true
Use L D Plot As Confidence Bounds	true
Use L D Plot As R Squared	true
Use L D Plot As Std	true
Use L D Plot As Std Alt	true
Use Assoc Tests Single Permutations	false

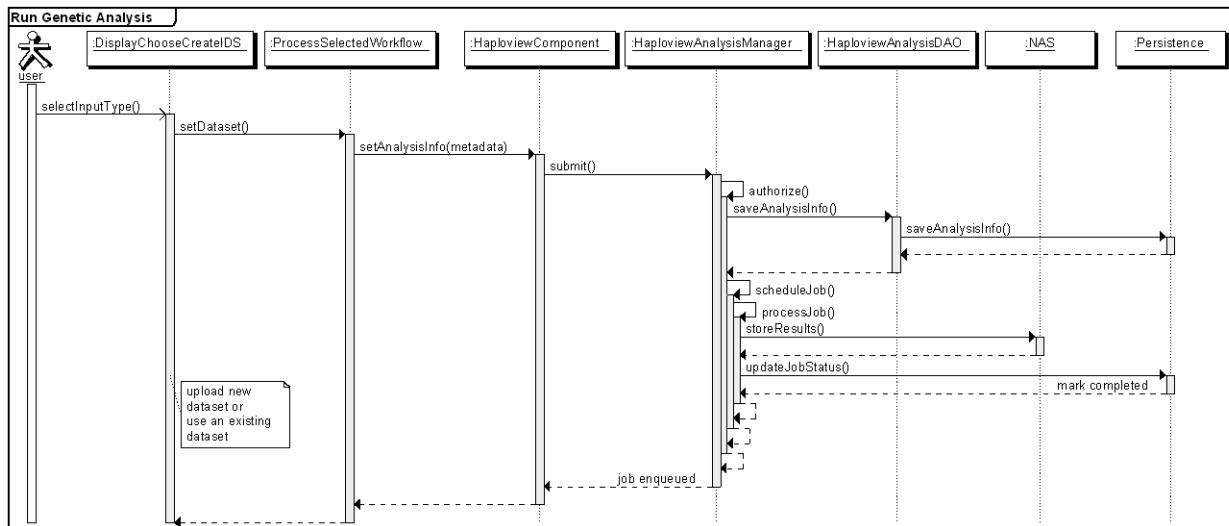
!!!

Done

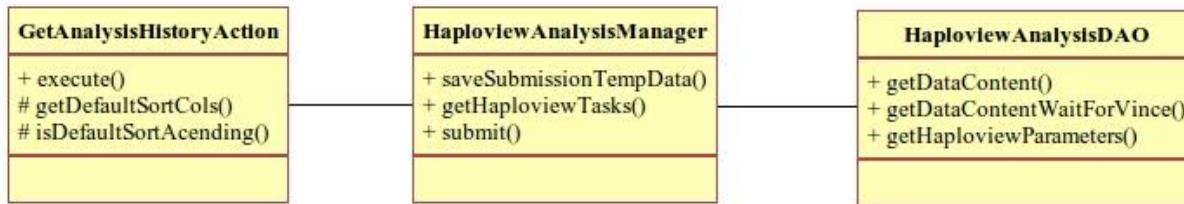
22.1.1 Haploview Analysis Submission Class Diagram



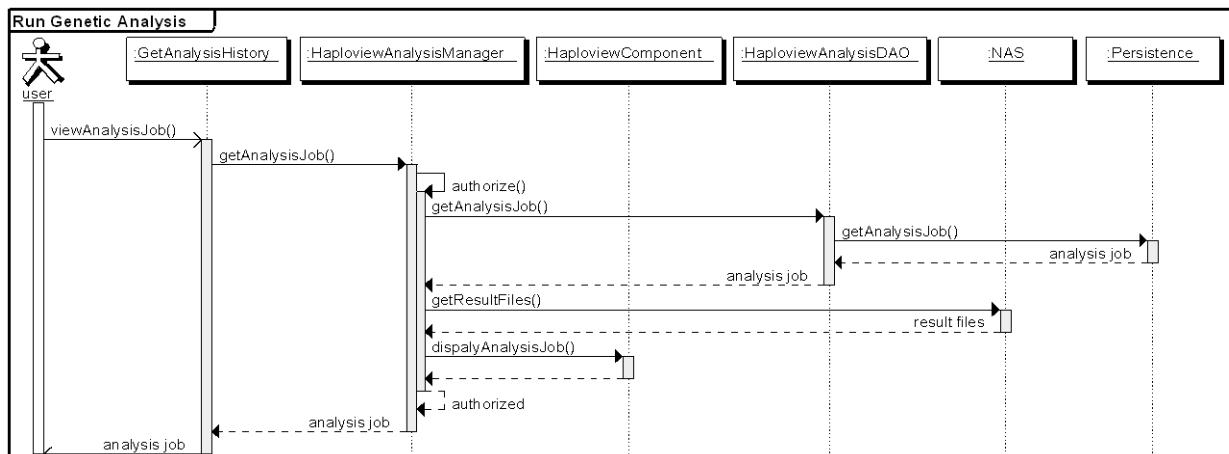
22.1.2 Haploview Analysis Submission Sequence Diagram



22.1.3 Haplovieview View Analysis Result Class Diagram

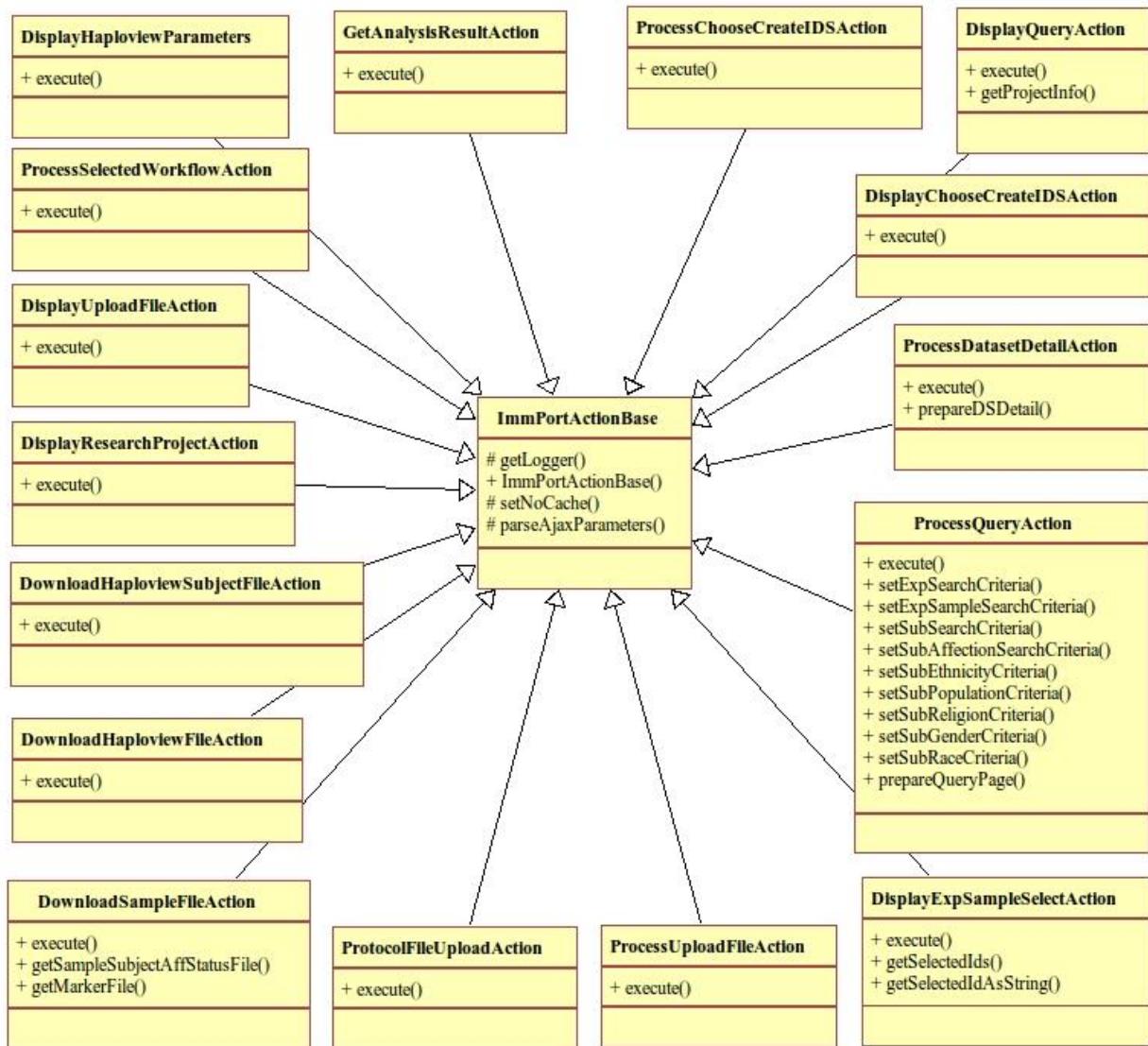


22.1.4 Haplovieview View Analysis Result Sequence Diagram

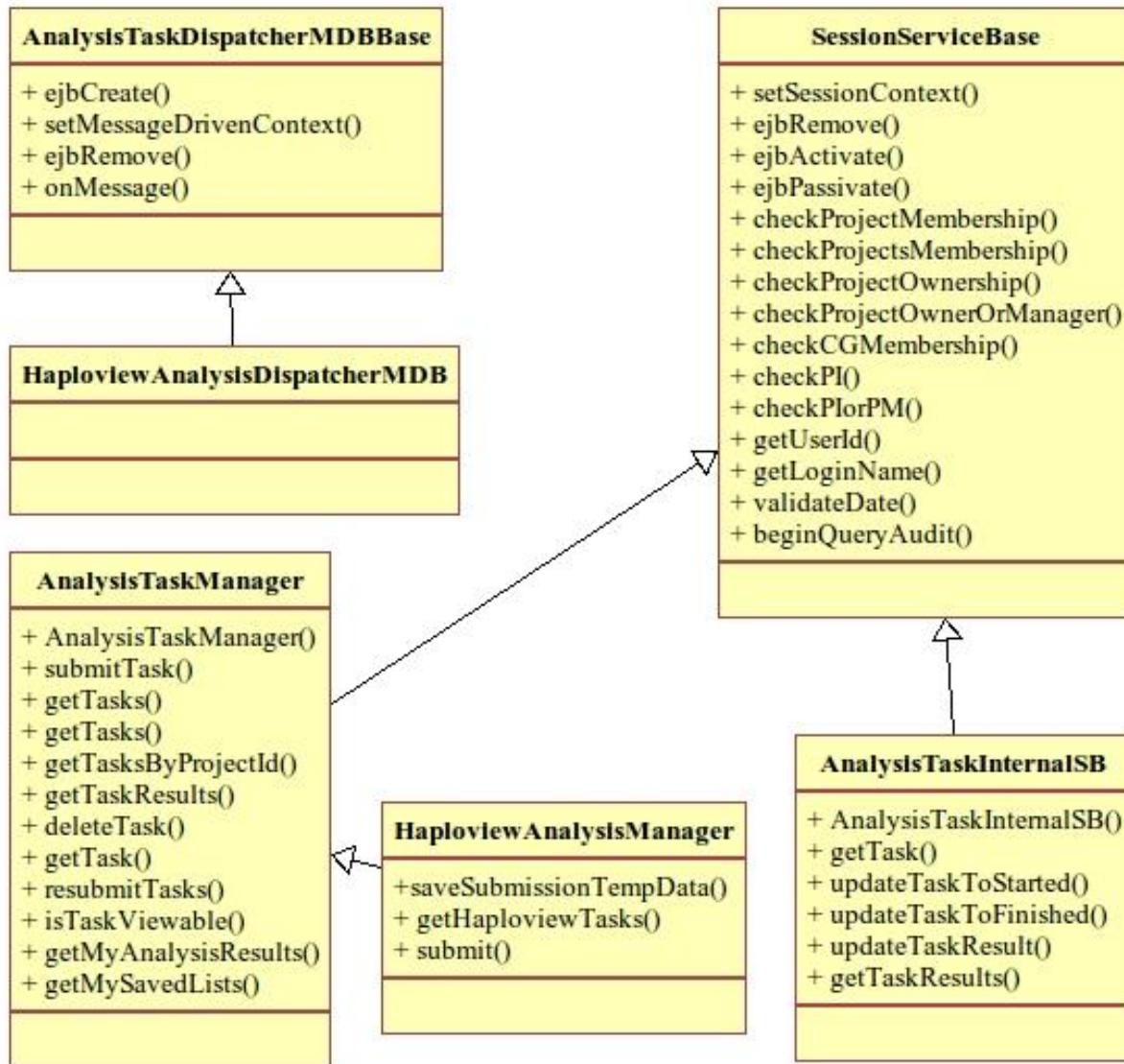


The View Analysis components enable scientist to view the job history and individual results.

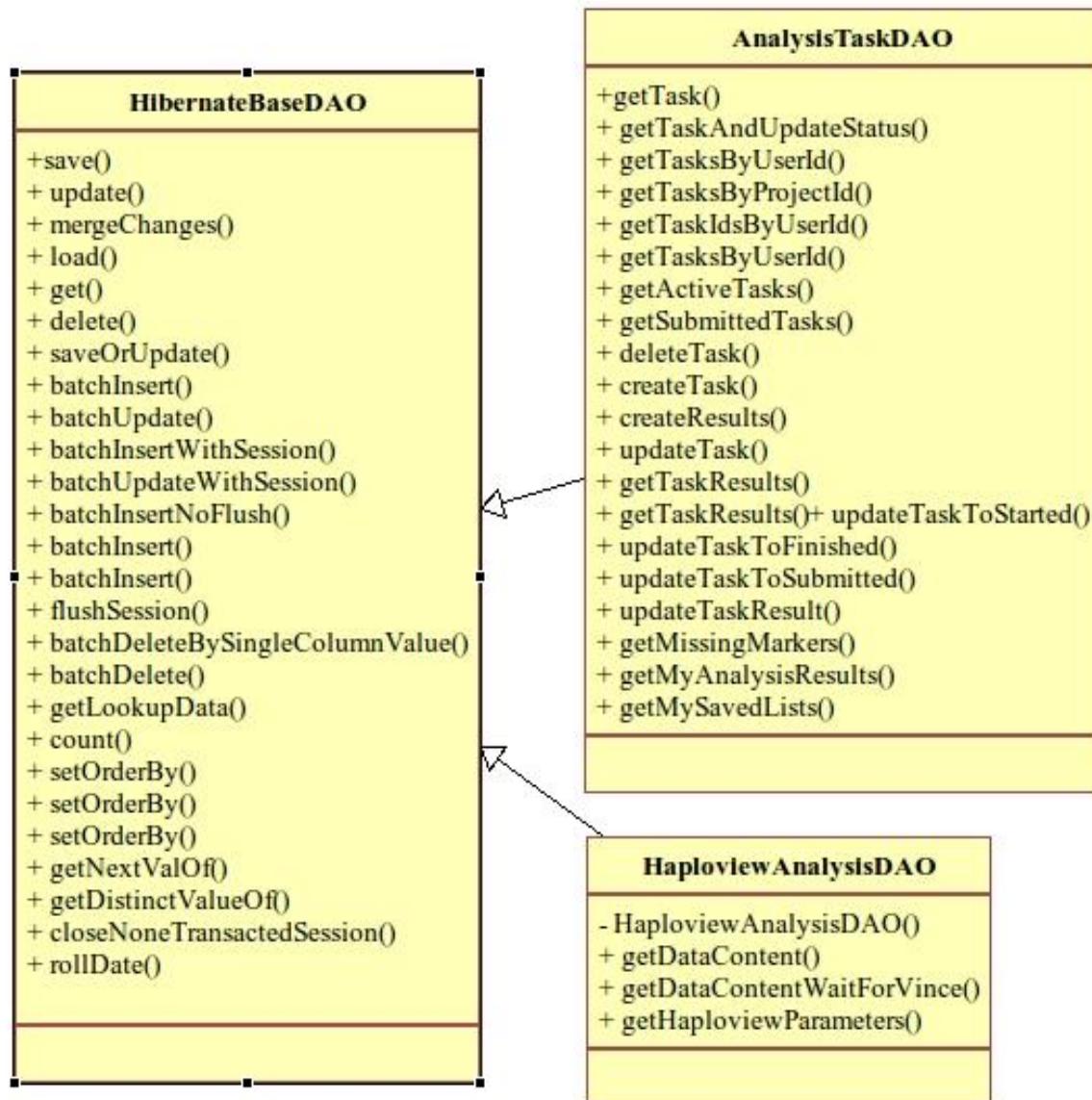
22.1.5 Haplovieview Presentation Tier Class Diagram



22.1.6 Haplovew Business Tier Class Diagram



22.1.7 Haplovie DAO Tier Class Diagram



23.0 TAGSNP DESIGN PACKAGES

Single Nucleotide Polymorphisms (SNPs) are nucleotide positions that vary between individuals in a population. Studies have shown that SNP site genotypes are significantly correlated; a small SNP subset, tag SNPs, are maximally informative because they capture most of the haplotype structure of the human genome and are therefore valued in genetic and disease association studies. ImmPort employs the LDSelect algorithm developed by a team headed by Dr. Chris Carlson (Carlson et al, Am J Hum Genet. 2004 Jan;74(1):106-20 using the human genome data from Phase II of the HapMap project to identify tagSNPs.

tagSNP / tagSNP Selection ⓘ

[Run tagSNP](#) | [Analysis Results](#)

The ImmPort system will use the LD Select algorithm (developed in the laboratory of Dr. Chris Carlson) and human genotype data from Phase II of the HapMap project to identify tagSNPs for one or more specified human genes or chromosome regions. The candidate tagSNPs are evaluated to confirm that the HapMap location is the same as the location in the selected Genome Build Version. Note that the LD Select algorithm may not provide reliable results when more than 1000 SNPs are included in a calculation. [Future ImmPort releases will provide additional tagSNP selection algorithms.]

Select one of the options below and then specify parameters on the following page(s).

Select tagSNP by

Genes
 Chromosome regions

[Next](#)

23.1 TAGSNP ANALYSIS GENE

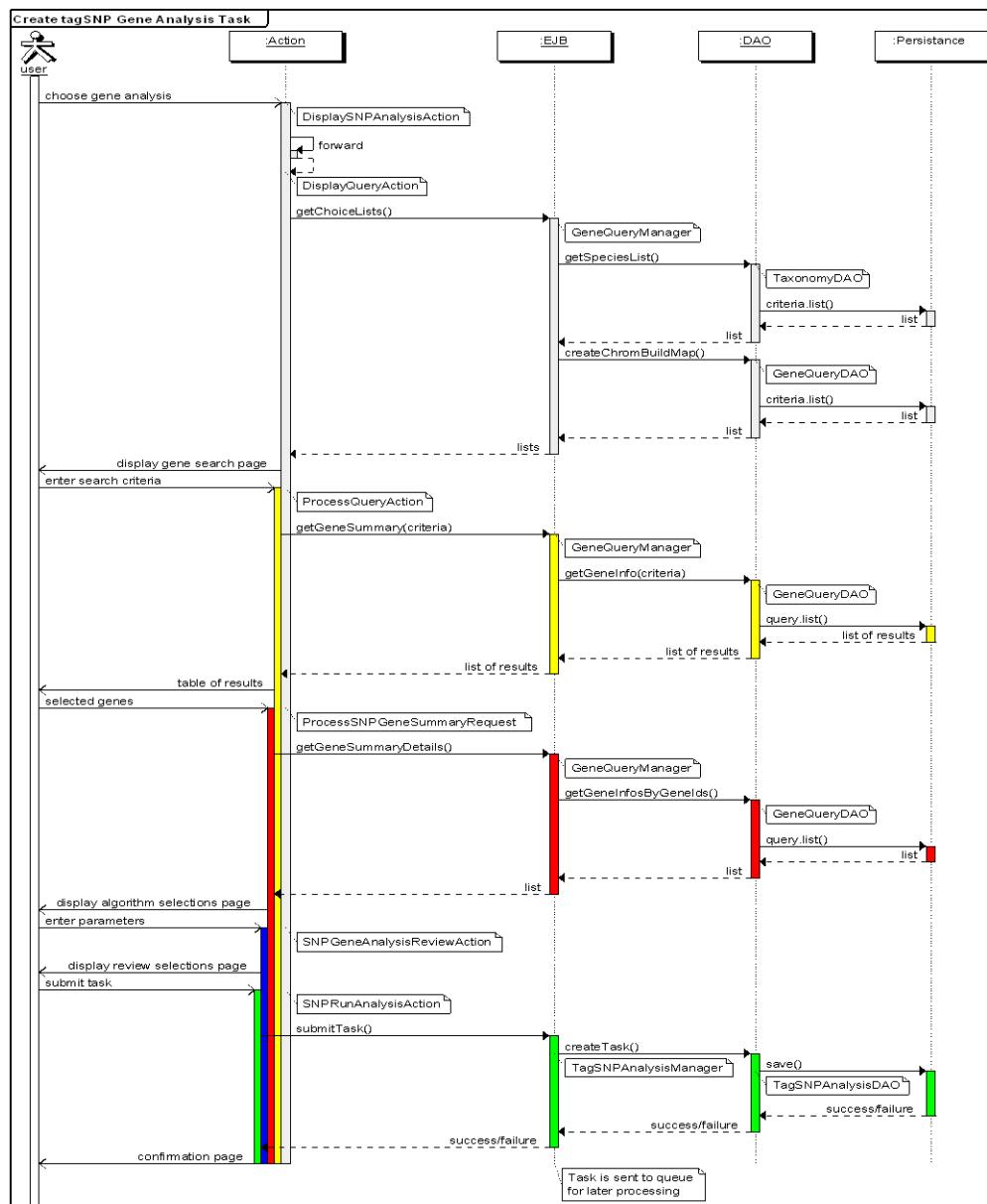
TagSNP analysis by gene is supported via the TagSNP Selection/Gene Search screen. Details of this screen include defining search options by type (gene symbol, name or ID), text, species, and genome build version.

The screenshot shows a web-based application titled "tagSNP / Gene Search". At the top, there are links for "Run tagSNP" and "Analysis Results". A note states: "Fields marked with an asterisk * are required." Below this is a "Search Options" section with the following fields:

- Search Type: Gene Symbol (dropdown menu)
- Search Option: Like (dropdown menu)
- Search Text: (Comma delimited, 1000 max chars) (text input field)
- Search Species: Homo sapiens (dropdown menu)
- Search Genome Build Version: Build 36.1, hg18 (dropdown menu)

At the bottom of the form are "Submit" and "Cancel" buttons, and a "Results Per Page" dropdown set to 25.

23.1.1 TagSNP Analysis Gene Sequence Diagram



23.2 TAGSNP ANALYSIS CHROMOSOME

TagSNP analysis by chromosome region is supported via the TagSNP Selection/Specify Chromosome Regions screen. Details of this screen include defining chromosome regions of interest, population, and LDSelect parameters.

tagSNP / Specify Chromosome Regions 

Run tagSNP | Analysis Results

Fields marked with an asterisk * are required.

Species* Build*

Chromosome Regions*
Maximum of 10 regions.
Example : chr22:28048224..28109123,
chrX:124754264..124824264
[Click here to view size of chromosomes.](#)

Choose a tagSNP selection algorithm and specify any additional parameters

Algorithm LDSelect Tagger (coming soon) HapBlock (coming soon)

Genotype Data Source Population *

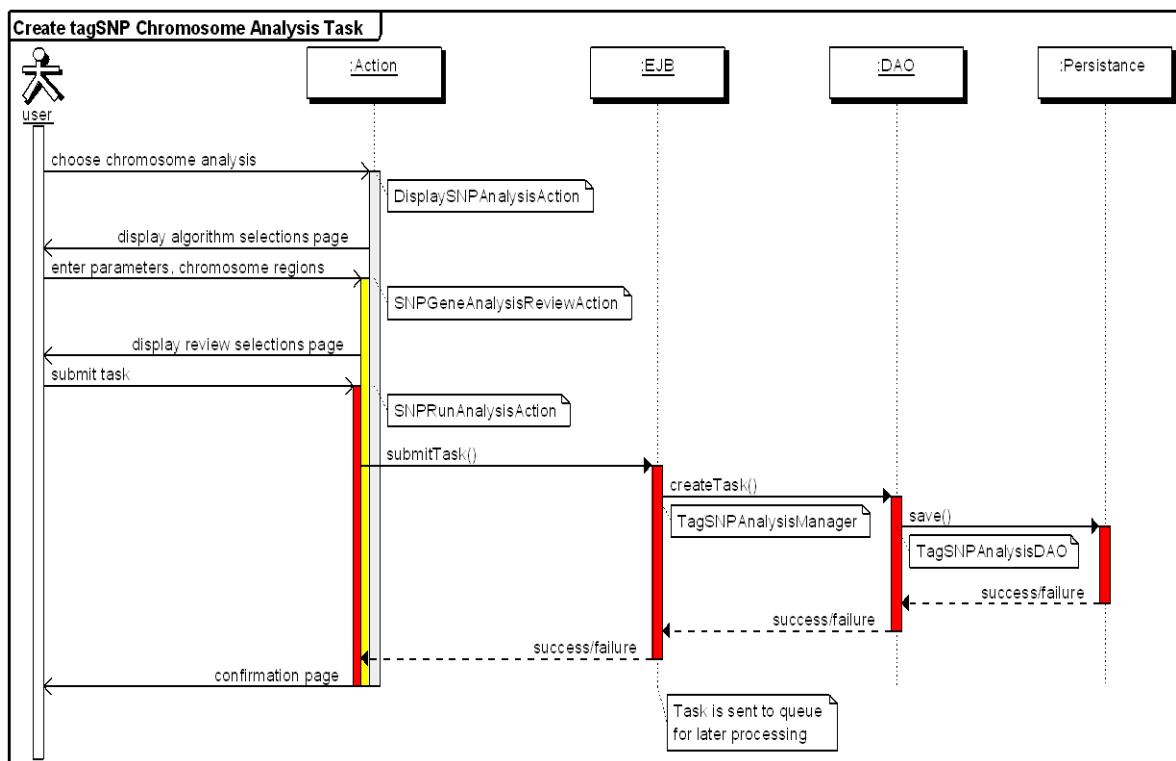
* for HapMap populations, only un-related individuals will be used for your tagSNP analysis

LDSelect Parameters

R² cutoff* (≥0.0 and ≤ 1)
Minor Allele Frequency Cutoff* (>0 and ≤ 0.5)

Next

23.2.1 TagSNP Analysis Chromosome Sequence Diagram



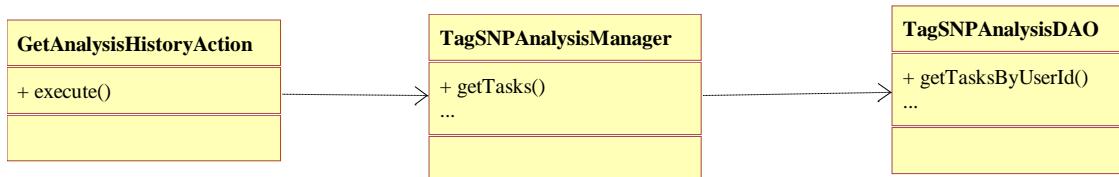
23.3 ANALYSIS HISTORY

Display of TagSNP analysis history includes a listing analysis tasks submitted, date, and time and task status.

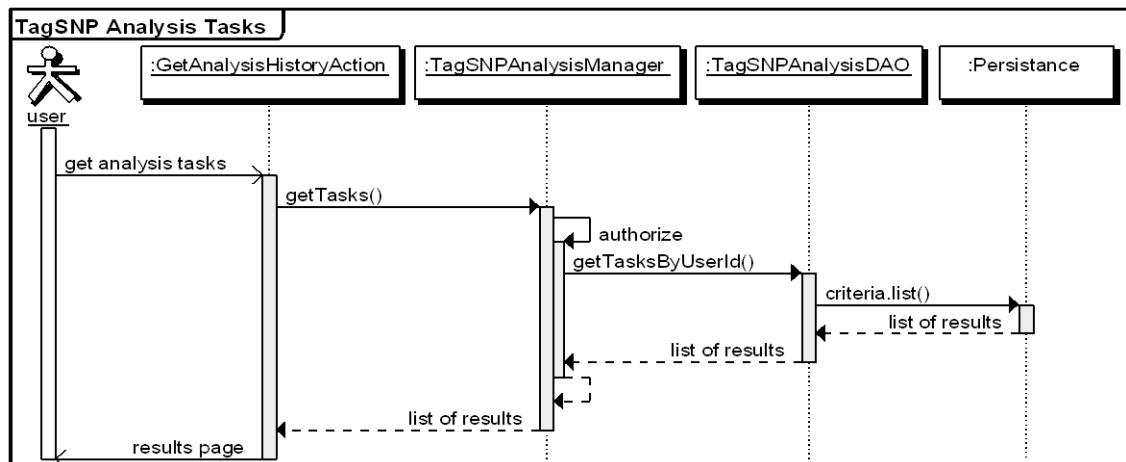
This screenshot shows the 'Analysis History' section of the tagSNP application. It displays a table of analysis tasks with the following columns: Task ID, Start Date, End Date, Algorithm, Genes, Chromosome Locations, and Status. There are two items listed:

Task Id	Start Date	End Date	Algorithm	Genes	Chromosome Locations	Status
79	10/27/2010 15:11:12		LDSelect		chr22:28048224..28109123	In Process
78	10/27/2010 14:58:09	10/27/2010 15:04:08	LDSelect	CA10		Completed

23.3.1 Analysis History Class Diagram



23.3.2 Analysis History Sequence Diagram



23.4 ANALYSIS RESULT

The Analysis Result display provides a summary of the analysis parameters with the returned SNPs presented in table format. The SNP table allows for selection of SNPs to be saved as a list.

Task Id	Start Date	End Date	Status
78	10/27/2010 14:58:09	10/27/2010 15:04:08	Completed

Algorithm Chosen: LDSelect

Warning: This algorithm may not give optimal results for ranges that contain greater than 1000 SNPs, for the statistical correlations derived from such long ranges may not have biological significance. Selection of a very large chromosome region may result in the analysis aborting prematurely with no results returned. If this occurs, select a shorter region and resubmit.

Genotype Data Source: HapMap
Population: (un-related individuals) HapMap CEU
Parameters:
R² Cutoff: 0.64
Minor Allele Frequency Cutoff: 0.1
Upstream Flanking Sequence: 10 kb
Downstream Flanking Sequence: 10 kb

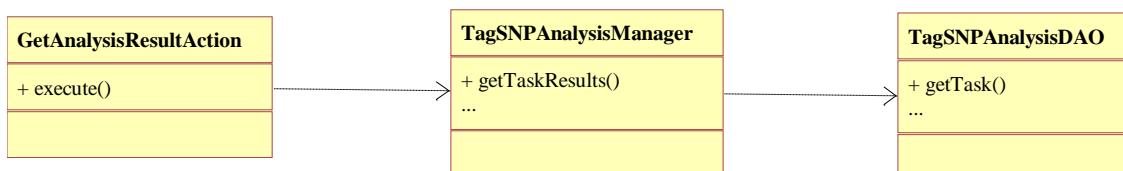
tagSNP Selection output (1 Result sets)

Collapse All: Expand All:

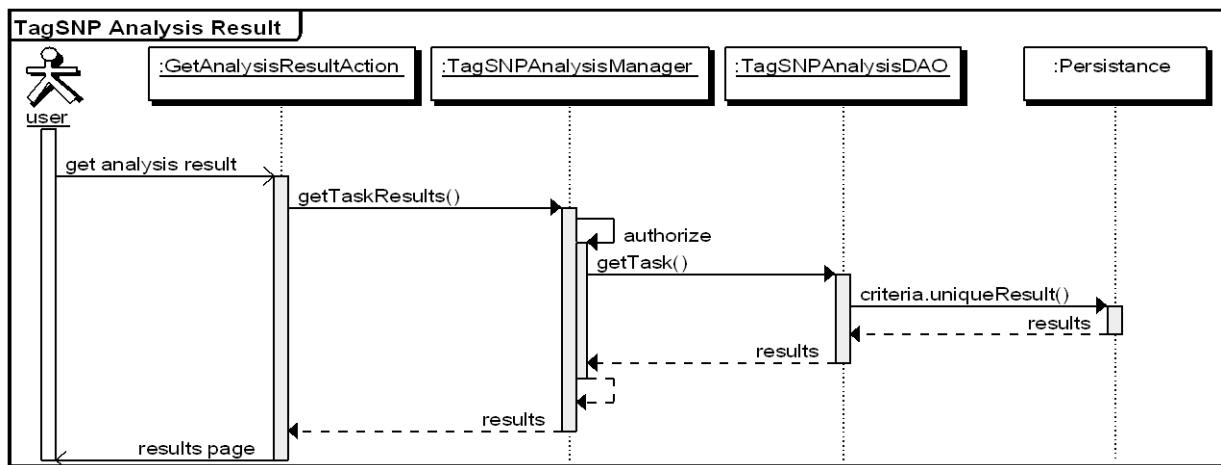
CA10 | Gene Id: 56934 | Result Id: 99

Gene: CA10 Gene ID: 56934 Chromosome: 17 Build Code: hg18 Synonyms: CA-RPX CARPX HUCEP-15 Method: LDSelect [Carlson et al.] Parameters: r2 threshold = 0.64, MAF cutoff = 0.1 Flanking sequences: upstream = 10kb, downstream = 10kb Population**: CEU							
Select a SNP Panel	Bin	SNP	Site	Function	Remarks**	Evolutionary Conservation Point Score***	Evolutionary Conservation Maximum Score***
<input type="checkbox"/>	1	rs203032	47422960	intron	tagSNP	0.0	0.0030
<input type="checkbox"/>	1	rs9910676	47520331	intron	tagSNP	0.04	0.212
<input type="checkbox"/>	1	rs2938141	47527352	intron	tagSNP	0.0	0.0010

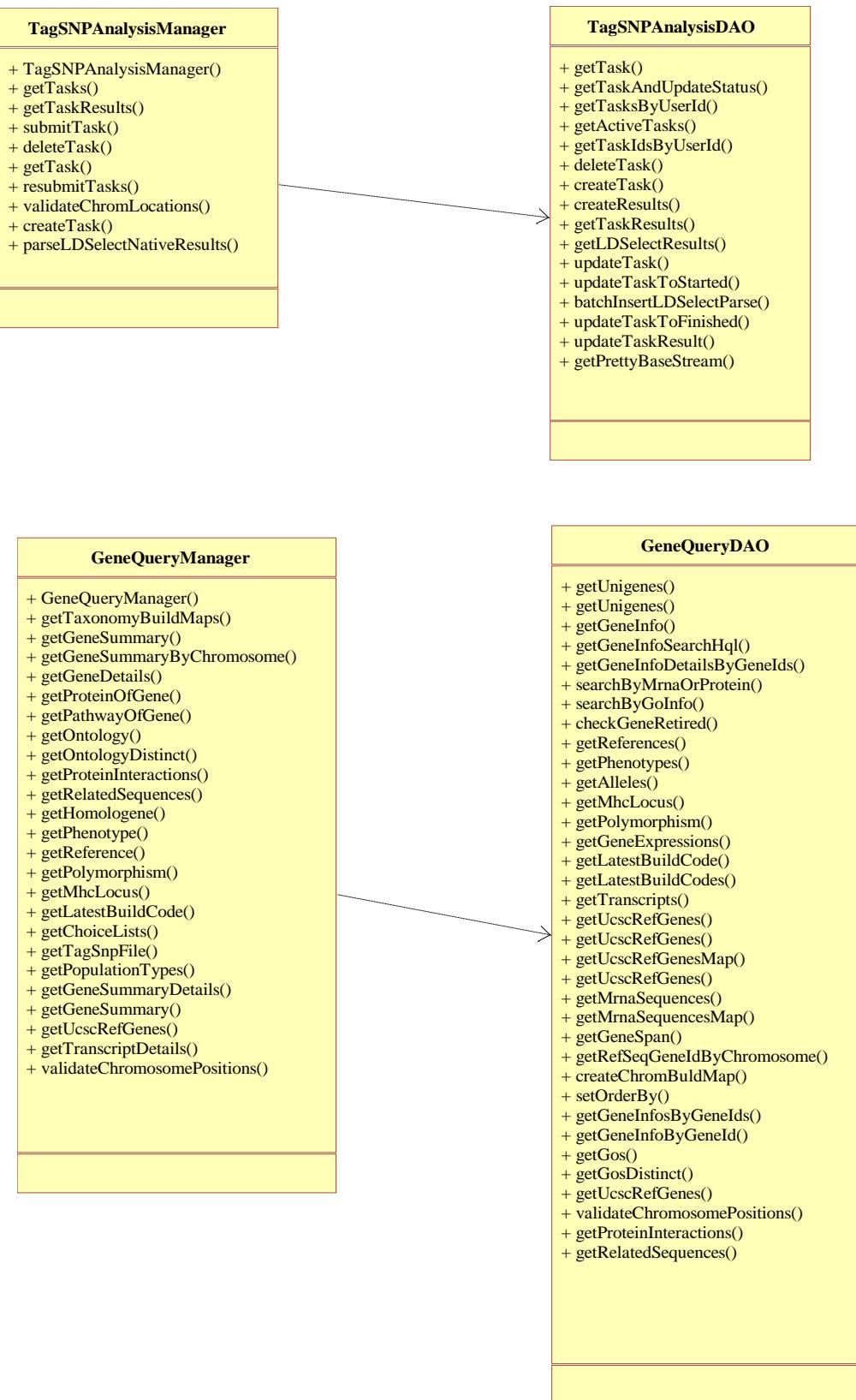
23.4.1 Analysis Result Class Diagram



23.4.2 Analysis Result Sequence Diagram



23.5 TAGSNP CLASS DIAGRAM

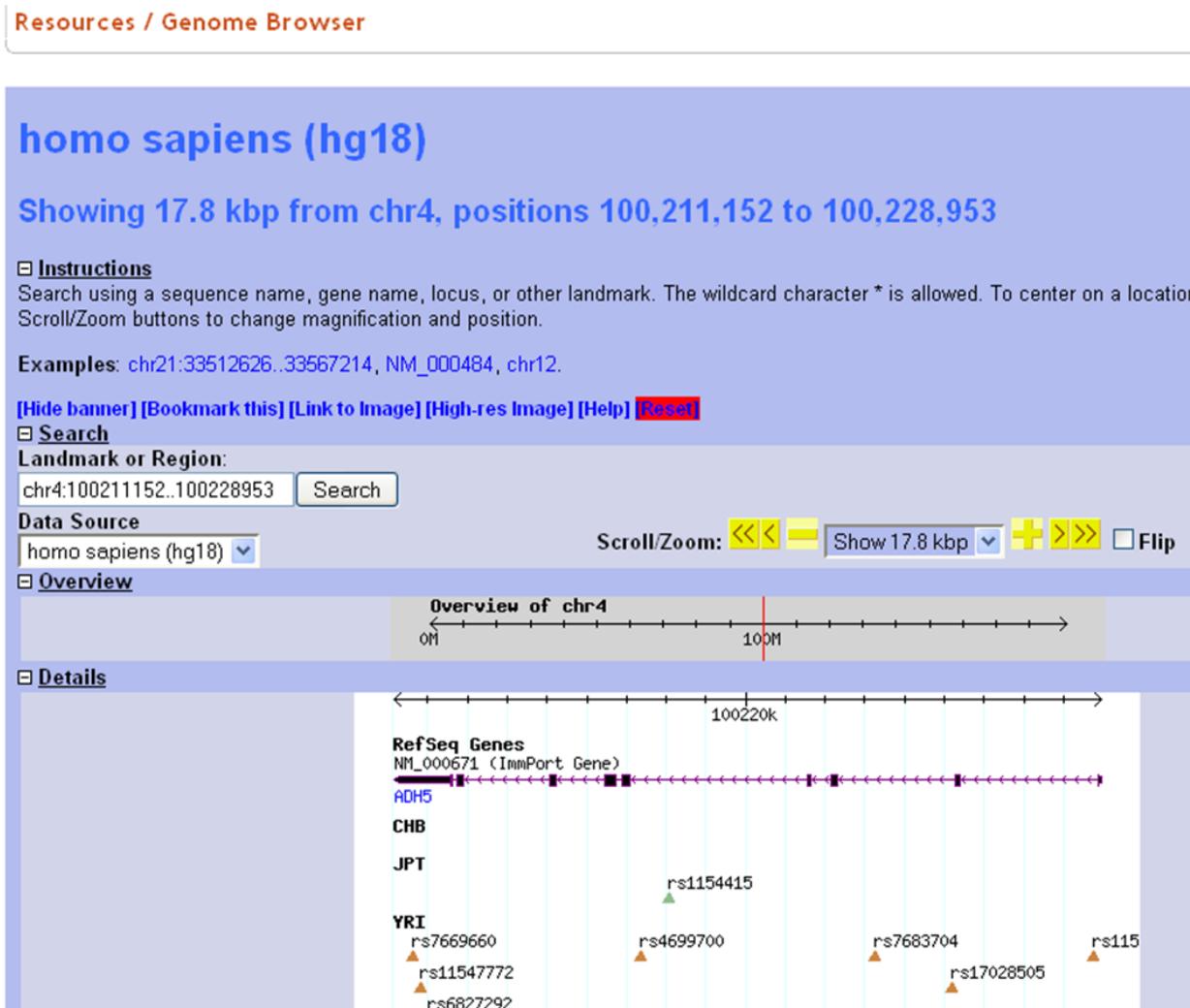


24.0 GBROWSE

The ImmPort system integrates the Generic Model Organism Database (GMOD) Gbrowse product as a genome sequence and feature visualization tool. database connectivity, multi-language support, and embedded images for use on internal web pages.

GBrowse is a Perl-based tool that relies on other key technologies such as the Apache web server for hosting web pages, Perl's Database Interface (DBI) for database connectivity, the Common Gateway Interface (CGI) perl library, and BioPerl. The GBrowse web application is hosted on Apache web server instances distributed across the network. This provides a flexible integration because the work is offloaded from the main web application engine. More GBrowse instances can be added as the workload volume increases in the future. The main ImmPort web application running on the JBoss application container communicates with GBrowse using the HTTP protocol.

The ImmPort GBrowse implementation supports the visualization of default genomic features including NCBI RefSeq and UCSC Known gene transcripts and NCBI dbSNP single nucleotide polymorphisms. Additional custom tracks can be added by the user.



25.0 ONTOLOGY BROWSER

The ImmPort system supports browsing and searching of three ontologies: Gene Ontology, Disease Ontology and Cell Type ontology. The Cell Type ontology will be used as an example to illustrate the process for supporting ontology data in ImmPort. The Cell Type OBO file is downloaded from the obo.cvs.sourceforge.net web site. The information in the OBO file is parsed into 3 formats suitable for loading into 3 database tables. The tables are: CELL_OBO, CELL_TERM and CELL_GRAPH_PATH. Additional information curated by the UT staff is merged with the parsed OBO output before loading this information into the database. In addition, we index the OBO content using the Lucene text searching software to support the free text searching option in the UI.

25.1 UI OVERVIEW

On the UI, the ontology tree is constructed from the data base using Oracle Connect-By queries and is displayed in the left panel using the Ext-JS tree component. The right panel consists of 2 tab panels: one tab for is used to initiate text based searching and displaying the results of a search, the second tab displays the summary page for a term.

Resources / Cell Type Ontology Browser

Cell Type Detail	
ID:	CL:0000988
Name:	hematopoietic cell
Definition:	A cell of a hematopoietic lineage. [GO_REF:0000031, GOC:add]
Xref:	BTO:0000574
Xref:	FMA:70366
Xref:	FMA:83598
Synonym:	haematopoietic cell
Synonym:	haemopoietic cell
Synonym:	hemopoietic cell
Is A:	CL:0000548 - animal cell
Is A:	CL:0002371 - somatic cell

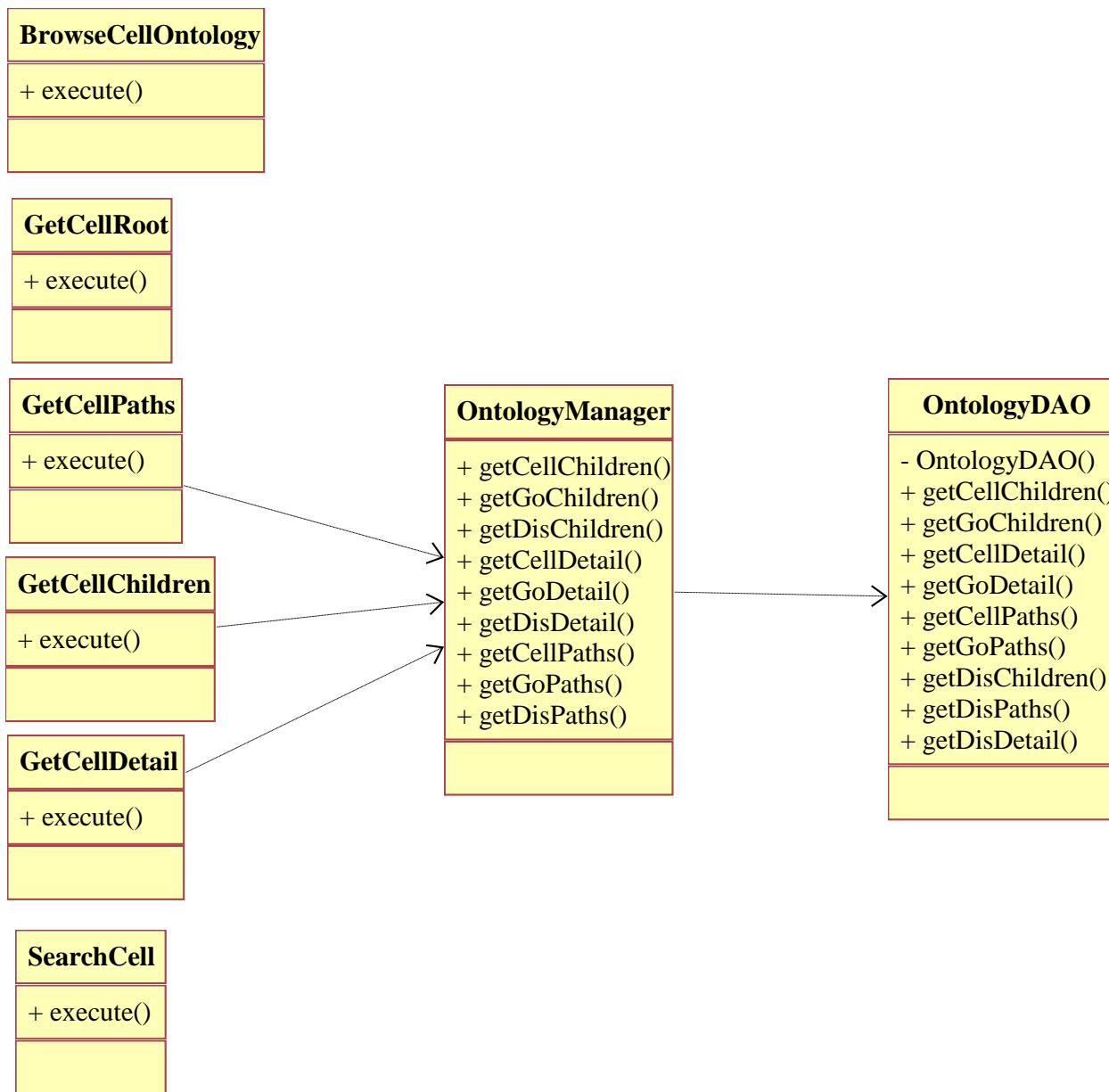
The user can then click on the term identifier to display a detail window for the term.

The screenshot shows a software application window titled "Cell Type Detail" for the term "T cell" (CL:0000084). The left sidebar has a tree view under "Cell Type Onto" with categories like Cell Type, cell in, ce, etc. The main panel displays detailed information about the T cell, including:

T cell	
Name	T cell
Cell Type Id	CL:0000084
Definition	A type of lymphocyte whose defining characteristic is the expression of a T cell receptor complex. [GOC:add, GOC:tfm, ISBN:0781735149]
Comment	
Synonyms	immature T cell; mature T cell; T lymphocyte; T-cell; T-lymphocyte
Is A	CL:000042: lymphocyte
Relationship	develops_from CL:0000827: pro-T cell
Disjoint From	CL:0000945: lymphocyte of B lineage
Union Of	
Cross Reference	BTO:0000782 FMA:62870
Links	
Definitional Phenotype	ID Symbol Term Expression
Surface Markers	
Cytokines	
Transcription Factors	
Gene Ontology	GO:0001667 ameboidal cell migration GO:0000792 heterochromatin GO:0030141 stored secretory granule GO:0002456 T cell mediated immunity
Additional Phenotype	ID Symbol Term Expression
	PR:000001288 CD22 B-cell receptor CD22 negative PR:000001289 CD20 B-cell surface antigen CD20 negative PR:000001002 CD19 CD19 negative PR:000001300 CD209 CD209 antigen negative PR:000001003 CD24 CD24 positive

At the bottom, there are navigation links for "mononuclear cell" and "myeloid leukocyte". The status bar shows "Trusted sites" and a lock icon.

25.2 ONTOLOGY CLASS DIAGRAM



25.3 SEQUENCE DIAGRAM

