

PharmGKB Database Documentation

1.0	Introduction.....	2
2.0	Background.....	2
3.0	Architecture.....	2
3.1	Schemata.....	2
3.2	Tables.....	3
3.3	Views.....	3
3.4	Triggers.....	3
3.5	Procedures.....	3
4.0	Data Overview.....	4
5.0	PGKBCOMM Schema.....	5
6.0	LIVE Schema.....	8
7.0	PREVIEW Schema.....	10
8.0	ARCHIVE Schema.....	10
9.0	PharmGKB Search.....	11
10.0	PharmGKB Archive System.....	11
11.0	PharmGKB TimeStamps.....	11

1.0 Introduction

This document describes the database architecture, tables, views, triggers, and procedures created to support the storage and management of data submitted to PharmGKB

2.0 Background

One purpose of PharmGKB is to be a repository for data about associations between genotypic and phenotypic data. Functional considerations for the database include the capability to accept data submissions from the user community, provide a preview of submitted data to the submitters for inspection and approval, provide access to all approved submissions to all registered users, provide a search capability to find major PharmGKB objects from any text field associated with the object, and track changes to the data and be able to retrieve a date dependent view of the data in PharmGKB. The following database architecture is designed to provide these capabilities in a secure and robust fashion.

3.0 Architecture

The PharmGKB database is built upon ORACLE 9i RDBMS. It is composed of several schemata to provide separation between common and curated, preview, live, and archived data. Table names, views, and synonyms have been created in each schema to allow the PharmGKB application to access the data in the same manner whether it is looking at preview, live, or archived data.

3.1 Schemata

The following schemata have been created in Oracle.

- PGKBCOMM: Contains tables that are used by both preview and live data. This includes administrative tables such as SUBMISSIONS, USERS and PROJECTS, tables of shared data objects such as GENES, DRUGS, and DISEASES, look-up tables for data integrity such as METHODTYPES and TEMPLATETYPES, all literature annotation submissions, and the PharmGKB curated entries of such objects as ReferenceSequences, FeatureSets, and GeneStructures.
- PREVIEW: Contains all current submitted genotypic and phenotypic data, both approved and not approved.
- LIVE: Contains all current approved genotypic and phenotypic data.

- ARCHIVE: Contains all past approved and common data. If a data record in either PGKBCOMM or LIVE has been modified or deleted, the original record is stored in the ARCHIVE schema prior to modification or deletion, along with timestamp values showing the data range for which this record was valid.

3.2 Tables

Tables have been created to store the data of PharmGKB. This includes genotypic data, phenotypic data, administrative data, and quality assurance/integrity data. Foreign key relationships have been created between tables to provide declarative referential integrity.

3.3 Views

Numerous views have been created on top of the tables to provide easier access to the data from the application. Most of these views have the effect of providing a de-normalized look at the data.

3.4 Triggers

Triggers have been created on the tables and views. Oracle ‘Instead of’ triggers on views were created to allow data manipulation – inserts, updates, and deletes – through the views. Tables and Views may have BEFORE and/or AFTER data manipulation triggers. The primary purpose of the BEFORE triggers is to manage the archiving of changed and deleted data.

3.5 Procedures

Stored procedures, written in PL/SQL, have been created in the various schemata to facilitate data processing. Most of the procedures are stored in sharable Oracle PL/SQL Packages. The primary packages are:

- PGKBCOMM.PharmGKBProcs – contains general utility procedures and archiving procedures.
- LIVE.ManageSubmissionData_PKG – contains procedures used to promote and remove submitted data upon submitter actions of approval, rejections, ...
- LIVE.PhenoGenoReport_PKG – contains procedures to build download files
- LIVE.UpdateDataFlags_PKG – contains procedures to maintain various data flags for PharmGKB objects.
- PREVIEW.CreateSystemDefaultObjects – contains procedures to create certain objects (currently only subjects and samples) directly without submission of XML file.

In addition to these packages, there are also a handful of individual procedures in the various schemata.

4.0 Data Overview

The primary data stored in PharmGKB are the PharmGKB Objects. There is one controlling table (PharmGKBObjects) which stores header information for each object in PharmGKB. There are numerous object types stored in PharmGKB, and each type of object is stored in its own table or set of tables. The PharmGKB object types include:

1. Gene
2. Reference Sequence
3. Sample Set
4. Feature Set
5. Feature
6. Experiment
7. Disease
8. Subject
9. Genotypes in Samples
10. PCR Result
11. Variant
12. Pooled Genotype
13. Genotyping Result
14. DHPLC Result
15. Pharmacogenetic Significance
16. Haplotype
17. Drug
18. Publication
19. Literature Annotation
20. DHPLC Assay
21. PCR Assay
22. Pyrosequencing Assay
23. RFLP Assay
24. TAQMAN Assay
25. Phenotype Data Set
26. PCR Sizing Assay
27. Generic Genotyping Assay
28. Generic Sequencing Assay
29. Artificial Construct
30. Assay
31. Chromosome
32. RNA
33. Protein
34. Polymorphism
35. Haplotype Set

36. Gene Structure
37. Study Group
38. Sample
39. Splice Set
40. Named Allele
41. Aggregated Genotype Variant
42. Aggregated Genotype
43. Pathway

All PharmGKB Objects are entered into PharmGKB via a submission process. Each submission is stored in PharmGKB in the Submissions table, and the objects entered by that submission are listed in the ObjectSubmissionAssociations table.

PharmGKB has recognized projects and authorized users in those projects from which it will accept submissions. For user submitted genotypic and phenotypic data, the data is first submitted to the PREVIEW schema, which is limited to only authorized users in the recognized projects. The PREVIEW schema allows the data to be reviewed for quality assurance. When the data has passed this review, it can be approved for publication to the LIVE schema by a user in the submitting project who has approval privilege. Once approved, the data is promoted to the LIVE schema where it is viewable by any registered user.

Most of the objects may be submitted by authorized users. Some objects, such as Genes, Drugs, and Diseases, may only be submitted, or are curated by PharmGKB. Other objects, such as Reference Sequences, Features, Feature Sets, and Gene Structures, may be either curated by PharmGKB and represent “default” instances, or submitted by users to provide specific object instances to be used by other objects in the project. For example, a curated reference sequence exists for each gene which stores the golden path sequence for that gene, and each submitted experiment must be associated with some reference sequence, which may be either a curated reference sequence or a submitted reference sequence.

5.0 PGKBCOMM Schema

The PGKCommon Schema contains the data that is shared by both the LIVE and PREVIEW schemata. This includes tables which hold administrative information, such as Users; tables which hold the PharmGKB objects which are entered and maintained only by PharmGKB, such as Genes, Drugs, and Diseases; tables which hold look-up information for other tables, such as Organism types and Feature types; tables holding data which need not be subjected to a quality assurance review, such as PharmacogeneticKnowledge which stores literature submissions; and tables which store the PharmGKB curated records for objects to support Genes, such as ReferenceSequences_CUR. Specifically, PGKBCOMM contains the following tables:

- AminoAcidTypes

- ChromosomeCLOBs
- Chromosomes
- CoordinateStrandTypes
- DiscussionForums
- DiseaseMESHTrees
- Diseases
- DiseaseSynonyms
- DrugIngredients
- DrugMethodOfActions
- DrugPharmacokinetics
- DrugPharmEffects
- Drugs
- DrugSynonyms
- DrugTreatedConditions
- EvidenceLocations
- EvidenceLocationTypes
- FeaureSets_CUR
- Features_CUR
- FeatureTypes
- FrequencyReportCellParents
- FrequencyReportCells
- FrequencyReportTypes
- FundingSources
- GenderTypes
- GeneOMIMPhenotypes
- GeneProducts
- Genes
- GeneStructureFeatureSets_CUR
- GeneStructures_CUR
- GeneSymbols
- GeneSynonyms
- GoldenPathPositions
- IngredientsForDrugs
- Keywords
- MapRegionTypes
- MethodOfActionForDrugs
- MethodTypes
- NamedAllelePolymorphisms
- NamedAlleles
- NIH_Ethnicities
- NIH_Races
- ObjectKnowledgeAssociations
- Ontologies
- OntologyTerms

- OrganismSpecies
- OrganismTypes
- PGKBPeopleRefs
- PharmacogeneticKnowledge
- PharmacogeneticKnowledgeHdr
- PharmacokineticsForDrugs
- PharmacoKnowledgeTempStore
- PharmEffectsForDrugs
- PharmGKBEvents
- PharmGKBNews
- PharmGKBObjectSyncQueue
- PharmGKBObjectTypes
- PharmGKBRoles
- PharmGKBXrefs
- PhenotypeColumnDataTypes
- PhenotypeColumnTypes
- PRN_Registration
- Programs
- ProjectOntologyAssociations
- ProjectPrivileges
- Projects
- ProjectUserPrivAssociations
- ProjectUsers
- PublicationAuthorAssoc
- Publications_CUR
- ReferenceSequenceCLOBs_CUR
- ReferenceSequenceSources
- ReferenceSequences_CUR
- RNATypes
- SNP_Report
- SNP_Sequence
- SNP_Subsnp
- SNP_Time
- StatusMsg
- SubmissionEditor
- SubmissionEditorCLOBs
- SubmissionManagementQueue
- SubmissionOntologyAssoc
- Submissions
- SubmissionStatusTypes
- TemplateTypes
- TrackingEventBLOBs
- TrackingEvents
- TrackingEventTypes

- TwoObjectKnowledgeRpt
- TwoObjectKnowledgeRptAssoc
- TwoObjRptKnowledgeAssoc
- UnitOfMeasureTypes
- UnitsOfMeasure
- UserGeneWatchlists
- UserOntologyAssociations
- UserPrivAssociations
- UserPrivileges
- Users
- VariantContextExperAssoc
- VariantContexts
- VariantPositions
- VariantsBySamples
- VariantSummarySamples
- XrefResources

6.0 LIVE Schema

The LIVE schema contains all the genotypic and phenotypic data which has been submitted to PharmGKB and has been approved for public access. It also contains report tables which are used to summarize the genotypic data or facilitate processing data to outside data repositories such as DBSNP. The specific tables include:

- AggregatedGenotypeAssayAssoc
- AggregateGenotypes
- AggregatedGenotypeSampleAssoc
- AggregatedGenotypeVariantAssoc
- ArtConstPolymorphisms
- ArtificialConstructRefSeqAssoc
- ArtificialConstructRefSeqs
- ArtificialConstructs
- AssayResults
- AssayResultVariantAssoc
- EthnicClasses
- EthnicClassNIHAssociations
- ExperimentAssays
- Experiments
- ExperimentSampleSetAssoc
- FeatureSet_SUB
- Features_SUB
- GeneStructureFeatureSets_SUB
- GeneStructures_SUB
- GenotypesInSamples

- HaplotypAnalyses
- HaplotypeExperimentAssoc
- Haplotypes
- HaplotypeSetPositions
- HaplotypeSets
- HaplotypeSetStudyGroups
- HaplotypeSubjects
- ObjectKeywordAssociations
- ObjectOntologyAssociations
- ObjectSubmissionAssociations
- ObjectXrefAssociations
- PCRResultInterrogatedRanges
- PharmacogeneticSignificance
- PharmGKBObjects
- PhenoGenoQuerysetColumns
- PhenoGenoQuerysetObjects
- PhenoGenoQuerysets
- PhenotypeDatasetAuthorAssoc
- PhenotypeDatasetCells
- PhenotypeDatasetColHeadings
- PhenotypeDatasetColumns
- PhenotypeDatasetContactAssoc
- PhenotypeDatasetDiseaseAssoc
- PhenotypeDatasetDrugAssoc
- PhenotypeDatasetGeneAssoc
- PhenotypeDatasetMethodAssoc
- PhenotypeDatasetRows
- PhenotypeDatasets
- PhenotypeNormalizedColumns
- Polymorphisms
- PooledGenoInterrogatedRanges
- PooledGenotypes
- PooledGenotypeSampleAssoc
- PooledGenotypeVariantAssoc
- ProteinChanges
- ProteinCLOBs
- Proteins
- Publications_SUB
- RacialClasses
- RacialClassNIHassociations
- ReferenceSequenceCLOBs_SUB
- ReferenceSequences_SUB
- RefSeqMapRegionsAssoc
- RNAs

- Samples
- SampleSets
- SamplesetSamples
- SequencedObjMapRegions
- SpliceFeatureAssociations
- Splices
- SpliceSets
- StatusMsg
- StudyGroupCompositions
- StudyGroups
- StudyGroupSubjects
- Subjects
- VariantAlleleSets
- Variants

7.0 PREVIEW Schema

The PREVIEW schema contains all the genotypic and phenotypic data which has been submitted to PharmGKB, both approved and awaiting approval. It has the same tables as are in the LIVE Schema with some additional tables that are used in submission management:

- SubmissionDependencies
- SubmissionDependencyDate
- UpdGeneStatus

8.0 ARCHIVE Schema

The Archive schema contains a copy of every record from the LIVE and PGKBCOMM schemata which has either been changed or deleted. For each table in LIVE and PGKBCOMM, there is a corresponding table in the ARCHIVE schema. The table in ARCHIVE has the same name as the table in LIVE or PGKBCOMM with “A_” prepended. For example, the corresponding table for LIVE.PharmGKBObjects is ARCHIVE.A_PharmGKBObjects, and the corresponding table for PGKBCOMM.Genes is ARCHIVE.A_Genes. Each ARCHIVE table which corresponds to a table in LIVE or PGKBCOMM also has an additional column, “ACTIONFLG”, to store the indicator of the action, update or delete, which caused the record to be archived.

There are two additional tables in the ARCHIVE schema: ArchiveErrorLog and ArchiveErrorDetails. These tables capture information about errors during the Archiving process along with the data of the record being archived when the error occurred, so that the archive information may be recovered.

9.0 PharmGKB Search

The PharmGKB application provides a search facility which allows the user to enter any text string for searching and have PharmGKB objects Associated with the search string returned. The search facility is provided via a 3rd party product, Lucene.

10.0 PharmGKB Archive System

Whenever a change is made to a record in LIVE or PGKBCOMM (update or delete), a copy of the original record is made in the ARCHIVE schema. The copy is placed into the Associated archive table which has the same name as the data table, prepended with “A_” (eg A_PharmGKBObjects is archive for PharmGKBObjects). This copying is done automatically in the BEFORE trigger on the table. First the data record to be archived is deconstructed into a data structure containing the fieldname, fieldtype, data value, and primary key indicator for each field in the record. This data structure is then sent to a procedure (PharmGKBProcs.ArchiveData) which reconstructs the record and inserts it into the correct archive table. If there is an error during the insert to the archive table, an error log record is placed in the ArchiveErrorLog table and the deconstructed information for the record is placed in the ArchiveErrorDetails table. The error log can later be analyzed and the archive data recovered and placed in the appropriate archive table.

11.0 PharmGKB TimeStamps

Each data table in PharmGKB has two Date fields: ValidFrom and ValidTo. These fields define the time frame in which the data in the record is valid. In the PREVIEW, LIVE, and PGKBCOMM schemata, the ValidTo field is NULL to indicate that it is currently a valid record. In the ARCHIVE schema both ValidFrom and ValidTo are populated to indicate the time at which that data was valid. The BEFORE trigger on the table sets these timestamp values. On INSERT, ValidFrom is set to SYSDATE. On UPDATE or DELETE, ValidTo is set to SYSDATE. On UPDATE, ValidFrom is Set to 1 second after the value ValidTo was just given. In this way, the PharmGKB data can be reconstructed to the values it contained at any past time.