# dbsnpload Java Classes

Author: sc

Created: June 12, 2006

Last Modified: September 6, 2006 09:46

## 1 Purpose of Document

Describe dbsnpload Java Classes

## 2 Related Documents

- SNP Class Diagram

## 3 Java Classes

### 3.1 Parsing and Raw Data Objects

#### 3.1.1 Parsing the dbSNP Genotypoe Files

##### 3.1.1.1 Parsing dbSNP Individuals (Strains and their strain IDs)

1. DBSNPGenotypeIndividualInput
   - Is: a data object representing an 'individual' from the dbsnp genotype input file
   - Has: a strain name (which might be a JAX registry id), and a dbSNP strain ID.
   - Does: provides getters and setter methods
2. DBSNPGenotypeIndividualInputFile
   - Is: a Representation of the 'Individuals' (strains and their ids) from the DBSNP Genotype input file
   - Has: a pointer to the input file
   - Does: provides an iterator to iterate over dbSNP genotype Individual records and returns a DBSNPGenotypoeIndividualInput object

##### 3.1.1.2 Parsing dbSNP Genotypes (SubSNP strain alleles by RefSNP).

1. DBSNPGenotypeRefSNPInputFile
   - Is: a Representation of RefSnp strain alleles from the Genotype XML input file
   - Has: a pointer to the input file
   - Does: provides an iterator to iterate over dbSNP genotype RefSnp records and returns a DBSNPGenotypeRefSNPInput object
2. DBSNPGenotypeRefSNPInput

- Is: a data object representing the populations, by SubSnp, for a RefSnp
- Has: an rs id and a mapping of SS IDs to populations
- Does: provides getters and setter methods

3. DBSNPGenotypePopulation
    - Is: a data object representing a dbsnp Population
    - Has: a population id and a mapping of strain/alleles for the population
    - Does: provides getters and setter methods
4. Allele
    - Is: a data object representing a dbsnp allele.
    - Has an allele typing and its orientation to the RS flanking sequence, forward or reverse.
    - Does:provides getters and setter methods

3.1.2 Parsing the dbSNP Main XML Files (Build 125 called these 'NSE', thus the naming convention)

1. DBSNPNseInputFile
    - Is: a representation of RefSnp data from a dbsnp XML file
    - Has: a pointer to the input file
    - Does: provides an iterator to iterate over dbSNP records in the input file which returns a DBSNPNseInput object
2. DBSNPNseInput
    - Is: an object that represents the set of raw data for a dbSNP RefSnp
    - Has:
        - A DBSNPNseRS
        - An rsId
        - A set of DBSNPNseSS objects representing SS belonging to this RS
        - A set of 5' DBSNPNseFlank objects
        - A set of 3' DBSNPNseFlank objects
        - A set of DBSNPNseContigHit objects
    - Does: provides getters and setter methods
3. DBSNPNseRS
    - Is: an object that represents basic raw RS data for a dbSNP RS from the XML input file
    - Has:

- an rsId
- a variation class
- the build in which this snp was first created
- the build in which this snp was last modified

- Does: provides getters and setter methods

4. DBSNPNseSS
   - Is: an object that represents raw data for a dbSNP SS in the XML input file
   - Has:
     - an ssId
     - a submitter SNP id
     - a submitter handle that identifies the submitter
     - a variation class
     - the orientation of the SS to the RS
     - whether this is the 'exemplar' SS (SS from which the flaning sequence is taken)
     - the set of alleles observed in this SS.
   - Does: provides getters and setter methods
5. DBSNPNseFlank
   - Is: an object that represents flanking sequence data.
   - Has: a nucleotide sequence, a sequence number (if there are > 255 chars in a flanking sequence, the sequence is split into multiple pieces).
   - Does: provides getters and setter methods
6. DBSNPNseContigHit
   - Is: a data object that represents raw 'contig hit' data.
   - Has: a chromosome name, an assembly (e.g. C57BL/6J), and a set of map locations (hits) - See DBSNPNseMapLoc.java
   - Does: provides getters and setter methods
7. DBSNPNseMapLoc
   - Is: an object that represents maplocation from a dbsnp input NSE file
   - Has:
     - a the starting coordinate
     - the orientation of the snp to the chromosome
     - a set of DBSNPNseFxnSets for this coordinate
   - Does: provides getters and setter methods
8. DBSNPNseFxnSet

- Is: an object that represents raw Function Set data
- Has:
  - locusId - entrezgene id of a Marker associated with this snp
  - fxnClass - functional class of the location of the snp in the Marker
  - contigAllele - allele on the contig where this snp lies
  - aaResidue - the amino acid where this snp lies on the contig
  - aaPosition - amino acid position
  - readingFrame - reading frame where this snp lies on the contig
  - nuclId - RefSeq nucleotide id for the marker
  - protId - RefSeq protein id for the marker
- provides getters and setter methods

## 3.2 Resolving Raw Data to MGI Values

### 3.2.1 DBSNPInputProcessor

- Is: an object that processes a DBSNPInput object. It resolves and/or translates raw values of the DBSNPInput into MGI values to create DAOs which write to bcp files.
- Has: A DBSNPInput object to resolve, Lookups to resolve attributes, SQLStream to write to bcp files, and a running list of RS ids already looked at, to avoid dups
- Does: Resolves snp attributes, writing them to bcp files a

### 3.2.2 AlleleOrderer

- Is: an object that orders a '/' delimited string of alleles
- Has: a Comparator which encapsulates the ordering rules.
- Does: orders a string of alleles

### 3.2.3 IUPACResolver

- Is: an object that converts a '/' delimeted string of alleles to an IUPAC code
- Has: a mapping of allele strings to IUPAC codes
- Does: provides a resolve method that, given an allele summary string, resolves it to an IUPAC code

### 3.2.4 Lookups

1. ChrSeqNumLookup
   - Is: An object that looks up the sequenceNum of a chromosome given a _Chromosome_Key
   - Has: A query to get _Chromosome_Key/sequenceNum pairs
   - Does: Provides a method to look up sequenceNum given a _Chromosome_Key
2. HandleNameByPopIdLookup

- Is: An object that knows how to look up Handle Name given a Population Id
- Has: A query to get all handle name/population id pairs
- Does: Provides a lookup method to look up and return a Handle Name given a Population Id

3. PopNameByPopIdLookup
    - Is: An object that knows how to look up a population name given a population ID.
    - Has: A query to get all population name/population ID pairs
    - Does: Provides a method to look up, and return, a population name given a population id.
4. StrainNameLookup
    - Is: an object that looks up a strain name given a strain key
    - Has: a query to get _Strain_key/strain pairs
    - Does: provides a method to look up, and return, a strain name given a strain key
5. MGI_dbinfoLookup
    - Is: an object that knows how to look up the MGI_dbinfo record for a given database.
    - Has: A query to get the MGI_dbinfo record
    - Does: Provides a lookup method to look up the MGI_dbinfo record and returns a MGI_dbinfoDAO
6. MGI_TablesLookup
    - Is: an object that knows how to look up an MGI_Tables record for a given database and table.
    - Has:
    - a database name
    - a query to get all MGI_Tables records
    - a mapping of the table name to a MGI_TablesDAO for each table in MGI_Tables for this database
    - Does: provides a lookup method to look up the MGI_Tables record, given a table name, and return a MGI_TablesDAO

## 3.3 Objects that manage DAOs (Resolved Data Objects)

1. SNPSNP
    - Is: an object that represents all resolved SNP information, ready to be inserted into a database.
    - Has:

- SNP_ConsensusSnpDAO
- a set of SNP_SubSnpDAOs
- a set of SNP_AccessionDAOs for all SNP accession ids
- a set of SNP_FlankDAOs
- a set of SNP_ConsensusSnp_StrainAlleleDAOs representing this consensusSnp's consensus alleles a
- a set of SNP_SubSnp_StrainAlleleDAOs representing the strain alleles of each SNP_SubSnpDAO
- a set of SNP_Coord_CacheDAOs representing coordinates for this consensusSnp
- a set of DP_SNP_MarkerDAOs representing markers associated with this consensusSnp
- a set of SNP_StrainDAOs representing any dbsnp strains that do not yet have database objects

- Does: provides a sendToStream method which passes each DAO to the SQLStream insert method

2. MGI_dbinfoUpdater
   - Is: An object that updates the snp_data_version and modification_date of the MGI_dbinfoDAO object
   - Has:
     - MGI_dbinfoLookup to lookup the existing record in a database and return a MGI_dbinfoDAO
     - a database name
     - date with which to update modification_date column
     - SNP Build Number with which to update snp_data_version column
   - Does: Provides an update method to update snp_data_version and modification_date of an MGI_dbinfoDAO object; returns the DAO.
3. MGI_TablesUpdater
   - Is: An object that updates the loaded date, modification date, and loaded by of an MGI_TablesDAO object
   - Has:
     - database name
     - name of the process, 'loaded by', using this Updater
     - current timeStamp for loaded and modification date
     - MGI_TablesLookup to lookup a table record in a database
     - MGIUserKeyLookup to resolve 'loaded by'

- Does: Provides an update method to update loaded date, modification date, and loaded by of an MGI_TablesDAO object, given a table name; returns the DAO.

4. The Loader and Support Classes
5. DBSNPLoader
    - Is: an DLALoader that parses DBSNP input files, resolves dbsnp attributes to MGI values, and reloads the snp database.
    - Has:
        - mgd SQLStream - for creating VocabLookups
        - snp SQLStream for writing and executing bcp files against the snp database
        - a set of chromosomes to determine which files to process
        - an XMLDataIterator to iterate through input files by chromosome
        - DBSNPInputProcessor to resolve attributes and write to bcp files
    - Does:
        - deletes snp accessions from a database
        - truncates snp tables
        - parses DBSNP input files
        - resolves DBSNP attributes to MGI values
        - creates snp database objects
        - counts and logs numbers of snps processed in various categories
6. DBSNPLoaderCfg
    - Is: an object that gets DBSNPLoader Configuration values
    - Has: a configuration manager
    - Does: provides methods to retrieve Configuration parameters that are specific to the DBSNPLoader
7. SNPLoaderConstants
    - Is: An object that contains constant definitions for SNP loaders.
    - Has: Constant definitions for:
        - dbSNP accession id prefixes
        - dbSNP assembly terms
        - dbSNP orientation terms
        - dbSNP variation class terms

## 3.4 Exceptions

1. SNPLoaderException

- Is: an MGIException which represents fatal exceptions occuring while processing SNPs

2. SNPLoaderExceptionFactory
   - Is: an ExceptionFactory for SNPLoaderExceptions
   - Has: a hashmap of predefined SNPLoaderExceptions stored by a name key
   - Does: looks up SNPLoaderExceptions by name
3. SNPMultiBL6ChrException
   - Is: a non-fatal MGIException thrown when a C57BL/6J SNP is on multiple chromosomes
4. SNPNoBL6Exception
   - Is: a non-fatal MGIException thrown when a SNP has no C57BL/6J coordinates
5. SNPNoConsensusAlleleSummaryException
   - Is: a non-fatal MGIException thrown when a Consensus Snp allele summary is empty
6. SNPNoStrainAlleleException
   - Is: a non-fatal MGIException thrown when there are no strain alleles for an RS
7. SNPRepeatException
   - Is: a non-fatal MGIException thrown when there are repeated SNPs in the input
8. SNPUnresolvedStrainException
   - Is: a non-fatal MGIException thrown when a strain cannot be resolved
9. SNPVocabResolverException
   - Is: a non-fatal MGIException thrown when cannot resolve a SNP vocabulary term