dbSNP Build 126 Data (Mouse Genome Build 34)

Author: sc

Created: August 30, 2006

Last Modified: September 6, 2006 09:46

1 Purpose of Document

Describe the dbSNP build 126 data

2 Introduction

The dbSNP XML files described below have vast amounts of information, of which MGI only loads a small subset. This document describes only the dbSNP data we load.

3 Definitions

- SubSNP or SS or SNP Assay or Submitted SNP - a dbSNP term for a submitted assay that turned up a polymorphism at a certain coordinate on the genome. A submission consists of 3' and 5' flanking sequence and SNP alleles found across a set of strains.
- RefSNP or RS or Reference SNP - a dbSNP term for a cluster of nucleotide polymor
- Genotype Input Files - dbSNP XML-format files, one for each chromosome, containing the strain alleles by SubSNP for each RefSNP
- Main Input Files - dbSNP XML-format files, one for each chromosome, containing most RefSNP and SubSNP data excluding the strain alleles.
- XML tag - a piece of text that describes a unit of data in XML. The tag is distinguisable as markup, as opposed to data, because it is surrounded by angle brackets (< and >). e.g. '<Rs' is the tag in the following line of XML:

  <Rs **rsId="3022833"** snpClass="snp" snpType="notwithdrawn" molType="genomic" genotype="true">

  XML tags may be nested within other tags.
- XML attribute- A qualifier on an XML tag that provides additional information. For example, in the above example, rsId is an attribute, and "3022833" is its value.
- XML text - A string of text between XML tags, e.g. flanking sequence:

  <Seq5>TTTTC...</Seq5>

  The XML parser handles text differently than attribute=value.

4 dbSNP Data Represented by MGI

High level description of the data, including structure and cardinality, we parse from dbSNP. See Section 6, "XML Data by File Type" on page 6

4.1 RefSNP

For each RefSNP we parse the following data:

1. RefSNP ID, one
   - A dbSNP assigned ID for a set of SubSNPs which cluster to the same locations on various assemblies.
   - Section 6.1, "Genotype XML File" on page 6, row 6 and See Section 6.2, "Main XML File" on page 8, row 1
2. A set of SubSNPs, one or more.
   - All SS that cluster to this location. See Section 4.2, "SubSNP" on page 2
3. SubSNP exemplar ID, one.
   - The SS whose flanking sequence is considered the best (typically longest). It is this flanking sequence that is used to determine RS orientation with respect to the chromosome, and to determine the gene(s) the RS lies within.
   - See Section 6.2, "Main XML File" on page 8, row 4.
4. RefSNP flanking sequences, one each 3' and 5'.
   - See Section 6.2, "Main XML File" on page 8, row 5.
5. Orientation of the RefSNP flanking sequence to the chromosome (genome), one.
   - See Section 6.2, "Main XML File" on page 8, row 14
6. A set of coordinates on the genome, one or more.
   - See Section 4.4, "Population" on page 4
   - This is the start coordinate for the polymorphism on a particular assembly. (The flanking sequence is blasted against various mouse strain assemblies).
7. dbSNP build number in which this RS was created, one.
   - See Section 6.2, "Main XML File" on page 8, row 2.
8. dbSNP build number in which this RS was updated, one.
   - See Section 6.2, "Main XML File" on page 8, row 3.

4.2 SubSNP

For each SubSNP we parse the following data:

1. A SubSNP ID, one.

- See Section , “” on page 7, row 6 and Section 6.1, “Genotype XML File” on page 6, row 7.

2. Submitter Handle, one.
    - This is a code that dbSNP gives to each submitter. See comprehensive list of handles in
    - See Section 6.2, “Main XML File” on page 8, row 7.
3. Submitter SNP ID, one.
    - This is what the submitter calls the assay.
    - See Section 7.1, “Population ID, Population Name, and Submitter Handle” on page 10 for a comprehensive list of submitter handles.
    - See Section 6.2, “Main XML File” on page 8, row 8.
4. Orientation of the SubSNP flanking sequence to theRefSNP flanking sequence, one.
    - Section 6.2, “Main XML File” on page 8, row 10.
5. Variation Class, one.
    - The type of polymorphism.
    - See Section 7.2, “SubSNP Variation Class” on page 12 for a comprehensive list of dbSNP variation classes
6. Set of Strain Alleles, zero or more.

    Some SubSNPs are submitted without strain alleles (don’t ask me why!).
    - Alleles of the polymorphism by strain.
    - See Section 4.3, “Strain Alleles” on page 3
7. Observed alleles, one.
    - Each SS has an observed allele string representing the set of alleles observed across all strains.
    - See Section 6.2, “Main XML File” on page 8, row 11.

4.3 Strain Alleles

For each strain allele we parse the following information:

Section 6.1.2, “Genotype File body” on page 7, rows 10 and 11.

1. A population, one or more.
    - See Section 4.4, “Population” on page 4
2. A strain, one.
3. An allele, one.
    - The allele of the polymorphism on that strain

- Note the alleles are given in pairs e.g. allele = ‘T/T’; with inbred mouse strains both pair members are always the same.
- If a strain was assayed, but failed, the allele is designated as “`N/N`”.

4.4 Population

For each Population we parse the following data:

1. Population Name
2. Population ID
3. Population Submitter Handle

See Section 7.1, “Population ID, Population Name, and Submitter Handle” on page 10

See Section 6.1.1, “Genotype File Header” on page 6, rows 1,2, and 3.

There can be multiple populations in one assay. Each population has a set of strain alleles. e.g. the submitter `ERO` has two populations, `PANEL1` and `PANEL2.`

Not all population are owned by a submitter, e.g. the submitter AEOMICA performs assays on other’s populations.

4.5 Genome Coordinate

1. Assembly, one
    - Name of assembly on which the coordinate is located.
    - See comprehensive list of assemblies in Section 7.4, “Assembly” on page 13
    - See Section 6.2, “Main XML File” on page 8, row 12.
2. The chromosome on this assembly, one or more.
    - See Section 6.2, “Main XML File” on page 8, row 13.
3. The coordinate on a given chromosome, one or more.
    - See Section 6.2, “Main XML File” on page 8, row 15.
4. Markers in which this assembly/chromosome/coordinate lie, one or more.
    - See Section 4.6, “Snp/Marker Relationship” on page 4

4.6 Snp/Marker Relationship

1. EntrezGene ID of the Marker, one
    - See Section 6.2, “Main XML File” on page 8, row 16.
2. Functional Class, one
    - Functional class of the location of the SNP on a particular assembly.
    - See comprehensive list Section 7.3, “Functional Class” on page 12
    - See Section 6.2, “Main XML File” on page 8, row 17.

3. Position of the variation in reading frame of transcript, 0 or 1.
    - See Section 6.2, “Main XML File” on page 8, row 18.
4. The allele, at this coordinate, observed in the contig sequence, 0 or 1.
    - See Section 6.2, “Main XML File” on page 8, row 19.
5. Translated amino acid residue for allele, 0 or 1
    - See Section 6.2, “Main XML File” on page 8, row 20.
6. Position of the variant residue in peptide sequence, 0 or 1.
    - See Section 6.2, “Main XML File” on page 8, row 24.
7. mRNA RefSeq ID, 0 or 1.
    - See Section 6.2, “Main XML File” on page 8, row 25.
8. Protein RefSeq ID, 0 or 1.
    - See Section 6.2, “Main XML File” on page 8, row 26.

5 Data Files

There is one Main Data File and one Genotype File each per chromosome, in XML format.

The Main Data File contains all RS and SS information, other than strain allele information. The Genotype File contains the strain alleles. Not all RS represented in the Main file have strain alleles in the Genotype file

The XML format of the Main file is described in:

- ftp://ftp.ncbi.nih.gov/snp/specs/docsum_2005.xsd

The XMLformat of the Genotype file is described in:

- http://ncbi.nlm.nih.gov/projects/SNP/specs/genoex_1_4_documentation.pdf

Documentation of the dbSNP ftp site:

- Getting Started Using the dbSNP FTP site - http://www.ncbi.nlm.nih.gov/About/outreach/gettingstarted/snpftp/index.html

The XML files can be downloaded from:

- ftp://ftp.ncbi.nih.gov/snp/organisms/mouse_10090/XML/ds_ch*.xml.gz
- ftp://ftp.ncbi.nih.gov/snp/organisms/mouse_10090/genotype/gt_chr*.xml.gz

The unzipped files are located locally at:

- /data/downloads/ftp.ncbi.nih.gov/snp/mouse/XML/ds_ch*.xml
- /data/downloads/ftp.ncbi.nih.gov/snp/mouse/genotype/gt_chr*.xml

6 XML Data by File Type

Columns in the table below:

- The '#' column simply numbers the data attributes
- The 'Data ' column describes the data to be parsed
- The 'XML Example' column gives XML tag examples with the data to be parsed in bold.
- The 'Required' column indicates whether a given data attribute is required to be listed in the file.
- The Store in MGI column indicates that the attribute will be stored in MGI database. Those attributes parsed that are not stored are used to either eliminate a RefSnp or used to calculate other values stored in the database.

See Section 7, "More Info by Attribute" on page 10 for detailed description of each attribute and its relationship to other attributes.

6.1 Genotype XML File

6.1.1 Genotype File Header

We parse data attributes 1 - 3 below to create SNP database population objects.

We parse data attributes 4 - 5 below in order to map the dbSNP strain ID, used to identify the strain in the strain alleles in the body of this file, to the strain itself. Note that strain could be either a strain name provided by the submitter, or a JAX registry ID as in the example.

Attributes 6-11 are the strain allele data.

#	Data Attribute	XML Example	Required in file	Store in MGI
1	Population ID	<Population **popId="1219"** handle="ROCHEBIO" locPopId="MM_PANEL">	yes	yes
2	Submitter Handle	<Population popId="1219" **handle="ROCHEBIO"** locPopId="MM_PANEL"> This is the handle of the **population** provider	yes	yes
3	Population Name	<Population popId="1219" handle="ROCHEBIO" **locPopId="MM_PANEL">**	yes	yes
4	dbSNP Strain ID	<Individual **indId="4567"** taxId="10090" sex="?">	yes	no
5	Strain ID of Source	<SourceInfo source="The Jackson Laboratory" sourceType="repository" ncbi-PedId="3592" pedId="3592" **indId="000677"** maId="0" paId="0"/>	yes	no

6.1.2 Genotype File body

#	Data	XML Example	Required in file	Store in MGI
6	RefSNP ID	<SnpInfo **rsId="3022833"**	yes	yes
7	SubSNP ID	<SsInfo **ssId="24767327"** locSnpId="Ren1_AN6747_1" ssOrientToRs="rev"**>**	yes	yes
8	SS Orientation To Chromosome	<SsInfo ssId="24767327" locSnpId="Ren1_AN6747_1" **ssOrientToRs="rev"**>	yes	yes
9	Population ID	<ByPop **popId="1182"** hwProb="0.001" hwChi2="42" hwDf="1" sampleSize="84">	yes	yes
10	Strain	<GTypeByInd gtype="G/G" **indId="2917"/**>	yes	yes
11	Allele	<GTypeByInd **gtype="G/G"** indId="2917"/>	yes	yes

6.2 Main XML File

#	**Data**	**XML Example**	**Required in file**	**Store in MGI**
1	RefSNP ID	**<Rs rsId="3022833"** snpClass="snp" snpType="notwithdrawn" molType="genomic" genotype="true">	yes	yes
2	Create Build	**<Create build="101"** date="2001-12-05 14:21"/>	yes	yes
3	Update Build	**<Update build="123"** date="2004-12-14 18:12"/>	yes	yes
4	SS Exemplar ID	**<Sequence exemplarSs="24767327">**	yes	yes
5	RS Flanking Sequences	**<Sequence** exemplarSs="24767327"> **<Seq5>TTTTC...</Seq5>** **<Seq5>ATCG...</Seq5>** **<Seq3>TCCT ...A</Seq3>** **</Sequence>** Note that flanking sequence is a XML text field and there are multiple Seq5 tags in this example.	yes	yes
6	SubSNP ID	**<Ss ssId="4319130"** handle="ROCHEBIO" batchId="4824" locSnpId="M-04294_1" subSnpClass="snp" orient="forward" strand="bottom" molType="genomic" buildId="103" methodClass="unknown" linkoutUrl="+M-04294_1">	yes	yes
7	Submitter Handle	**<Ss** ssId="4319130" **handle="ROCHEBIO"** batchId="4824" locSnpId="M-04294_1" subSnpClass="snp" orient="forward" strand="bottom" molType="genomic" buildId="103" methodClass="unknown" linkoutUrl="+M-04294_1"> This is the handle of the **submitter**, as opposed to the population provider.	yes	yes
8	Submitter SNP ID	**<Ss** ssId="4319130" handle="ROCHEBIO" batchId="4824" **locSnpId="M-04294_1"** subSnpClass="snp" orient="forward" strand="bottom" molType="genomic" buildId="103" methodClass="unknown" linkoutUrl="+M-04294_1">	yes	yes
9	SubSnp Variation Class	**<Ss** ssId="4319129" handle="ROCHEBIO" batchId="4824" locSnpId="M32352_1" **subSnpClass="snp"** orient="forward" strand="bottom" molType="genomic" buildId="10 1" methodClass="unknown" linkoutUrl="+M32352_1">	yes	yes
10	SS Orientation To RS	**<Ss** ssId="4319129" handle="ROCHEBIO" batchId="4824" locSnpId="M32352_1" subSnpClass="snp" **orient="forward"** strand="bottom" molType="genomic" buildId="101" methodClass="unknown" linkoutUrl="+M32352_1">	yes	yes
11	SS Observed Alleles	**<Ss** ... <**Sequence**> <Seq5> ... </Seq5> **<Observed>C/G</Observed>** <Seq3> ... </Seq3> **</Sequence>** </Ss> Note that observed is an XML text field	yes	yes

#	Data	XML Example	Required in file	Store in MGI
12	Assembly	**<Assembly** dbSnpBuild="125" genomeBuild="36_1" **groupLabel="C57BL/6J"** current="true"> Note: parsed only to determine whether if we want the RS.	yes	no
13	Genome Build	**<Assembly** dbSnpBuild="125" **genomeBuild="36_1"** groupLabel="C57BL/6J" current="true"> Note: parsed to determine the genome build. Build 126 XML files have coordinates for both build 35 and build 36.	yes	no
14	Chromosome	**<Assembly ...** **<Component** componentType="contig" ctgId="1009000016" accession="NT_039180.4" name="Mm1_39220_34" **chromosome="1"** start="117896164" end="139564857" orientation="fwd" gi="63477702" groupTerm="ref_strain" contigLabel="C57BL/6J">	yes	yes
15	RS Orientation To Genome	**<Assembly ...** **<Component ...** **<MapLoc** asnFrom="15325361" asnTo="15325361" locType="exact" alnQuality="0.96" **orient="forward"** physMapStr="133221526" physMapInt="133221525" leftFlankNeighborPos="552" rightFlankNeighborPos="554”leftContigNeighborPos="15325360"rightContigNeighborPos="15325362" numberOfMismatches="12" numberOfDeletions="0" numberOfInsertions="0"> ... **</MapLoc>**	yes	yes
16	Chromosome Start Coordinate	**<Assembly ...** **<Component ...** **<MapLoc** asnFrom="15325361" asnTo="15325361" locType="exact" alnQuality="0.96" orient="forward" physMap Str="133221526" **physMapInt="133221525"** leftFlankNeighborPos="552" rightFlankNeighborPos="554" leftContigNeighborPos="15325360" rightContigNeighborPos="15325362" numberOfMismatches="12" numberOfDeletions="0" numberOfInsertions="0"> ... </MapLoc> Note: In Build 125 some MapLoc are missing physMapInt so we can’t load them. PhysMapStr can indicate a coordinate range e.g. 34-36. PhysMapInt is always a single integer and it is **ZERO** based for Build 125 (It was 1 based for Build 124).	no	yes
17	EntrezGene ID	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet geneId="19701"** symbol="Ren1" mrnaAcc="NM_031192" mrnaVer="1" fxnClass="mrna-utr"/> **<FxnSet geneId="19702"** symbol="Ren2" mrnaAcc="NM_031193" mrnaVer="1" fxnClass="mrna-utr"/> **</MapLoc>**	no	yes

#	Data	XML Example	Required in file	Store in MGI
18	Functional Class	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet** geneId="19702" symbol="Ren2" mrnaAcc="NM_031193" mrnaVer="1" **fxnClass="mrna-utr"**/>	no	yes
19	Reading Frame	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet ... readingFrame="3" ...**	no	yes
20	Allele observed in contig	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet ... allele="C" ...**	no	yes
21	Amino Acid Residue	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet ... residue="P" ...**	no	yes
22	Amino Acid Position	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet ... aaPosition="13" ...**	no	yes
23	RefSeq mrna ID	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet** geneId="19702" symbol="Ren2" **mrnaAcc="NM_031193"** mrnaVer="1" fxnClass="mrna-utr"/>	no	yes
24	RefSeq Protein ID	**<Assembly ...** **<Component ...** **<MapLoc ...** **<FxnSet ... protAcc="XP_619638"**	no	yes

7 More Info by Attribute

7.1 Population ID, Population Name, and Submitter Handle

An Assay (an SS) may consist of one or more populations; each population with a set of results (strain alleles). A lab may submit data representing one of its own populations, another lab's populations, or both. Consider AEOMICA in the table below. This submitter uses other lab's populations; they have no data submitted from populations of their own.

A list of handles, population names and population ids (used in a given file) are found in the **<Population** section at the top of each genotype file. This refers to the population supplied by the

lab represented by the handle. We create a lookup from the population info in the header of the genotype file in order to resolve the PopId found in the records in the body of the file to a handle and a population name.

Current set of mouse population and submitter handles, popids, and population names:

Table 1:

PopId	Handle	PopName
1637	ABI	CRAMUS_MOUSE
None	ABI 164	None
None	AEOMICA	None
None	CGAP_GAI	None
836	CRAMUS	CRAMUS_MOUSE
513	ERO	PANEL_1
542	ERO	PANEL_2
1364	KMRGMT	KMRGMT-Mouse
None	KOKORO_BIOLOGY_GROUP	None
None	LPG-NCI	None
617	MITCCR	Mouse_(BJ)
1065	NCBI-GENBANK-VAR	MOUSE-POP-NC-005089
1219	PERLEGEN	MM_PANEL
1445	PERLEGEN	MM_PANEL2
1064	ROCHEBIO	RPAMM
None	SC_JCM	None
1780	TIZUMI	TB001
1182	TWGNF	TWGNF-Mouse-2
837	TWGNF	TWGNF-Mouse
735	WI	MOUSE
786	WIMOUSESNPS	MITROCHE_POP

7.2 SubSNP Variation Class

The type of polymorphism. List and definitions from dbSNP

- `snp` - true single nucleotide polymorphism;alleles of length=1 and from set of {A,T,C,G}
- `in-del` - insertion deletion polymorphism, deletions represented by '-' in allele string. This class is assigned to a variation if the variation alleles are of different lengths or if one of the alleles is deleted ('-').
- `het` - variation has unknown sequence composition, but is observed to be heterozygous
- `microsat` - microsatellite / simple sequence repeat
- `named-locus` - allele sequences defined by name tag instead of raw sequence, e.g. (Alu)/- .
- `no-variation` - submission reports invariant region in surveyed sequence.
- `mixed` - mixed class.
- `multinucleotide-polymorphism` - Not defined.

7.3 Functional Class

The position of a polymorphism with respect to identifiable features of a specific transcript for a gene. List and definitions from dbSNP. Note these are the subset of Functional classes found on mouse snps for Build 126.

- `Coding` - variation in coding region of gene, assigned if allele-specific class unknown
- `coding-nonsynonymous` - change in peptide with respect to contig sequence. There is an additional <FxnSet section that specifies the reference for determining this class.
- `coding-synononymous` - no change in peptide for allele with respect to contig seq. There is an additional <FxnSet section that specifies the reference for determining this class.
- `exception` - variation in coding region with exception raised on alignment. This occurs when protein with gap in sequence is aligned back to contig sequence. variations 3' of the gap have undefined functional inference.
- `intron` - variation in intron, but not in first 2 or last 2 bases of intron
- `locus-region` - variation in region of gene, but not in transcript.
- `mrna-utr` - variation in transcript, but not in coding region interval.
- `reference` - appears in conjunction with coding-synon' and 'coding-nonsynon' and specifies the reference allele observed in the reference contig sequence - See chr5 RS6277917.

- `splice-site` - variation in first 2 or last to bases of intron

7.4 Assembly

List from dbSNP. The assembly name is prepended with "assemblyVersion:". Build 126 assemblies are prepended "36:"

- 129_substrain
- A/J
- B6_CBAF1J
- Balb/c
- C3H
- C57BL/6J
- Celera
- MGSCv3
- NOD
- unknown