Applied Biosystems

TaqMan® Gene Expression and SNP Genotyping Assays

Flat File Catalog

Applied Biosystems

Table of Contents

Introduction.......... 4

- Data Type Definition.......... 4
- Data Model.......... 5

Gene.......... 6
Gene-OMIM.......... 7
OMIM.......... 8
Transcript.......... 9
GeX Assay-Transcript.......... 10
GeX Assay.......... 11
Transcript-Process.......... 12
Transcript-Function.......... 13
Set-GeX Assay.......... 14
EST-Transcript.......... 15
GT Assay-Gene.......... 16
GT Assay.......... 17
GTAssay-Human.......... 18
GTAssay-Non-Human.......... 20
Set-GT Assay.......... 21

Introduction

This document contains the formats for the flat files that comprise the Applied Biosystems TaqMan® Gene Expression and SNP Genotyping Assays Flat File Catalog. These files are compressed into a single file, created using Unix tar. The file name is:

`Assays_Catalog_mm_dd_yy_release.tar.gz`

The Gene Expression Assay Catalog files are:

- Gene_*mm_dd_yy*_release.txt
- Transcript_*mm_dd_yy*_release.txt
- GexAssay-Transcript_*mm_dd_yy*_release.txt
- GexAssay_*mm_dd_yy*_release.txt
- Set-GexAssay_*mm_dd_yy*_release.txt
- EST-Transcript_ *mm_dd_yy*_release.txt
- Transcript-Process_ *mm_dd_yy*_release.txt
- Transcript-Function_ *mm_dd_yy*_release.txt
- Gene-OMIM_ *mm_dd_yy*_release.txt
- OMIM_ mm_*dd_yy*_release.txt

The SNP Genotyping Assay Catalog files are:

- Gene_ *mm_dd_yy*_release.txt
- Transcript_ *mm_dd_yy*_release.txt
- GTAssay_*mm_dd_yy*_release.txt
- GTAssay-Human_*mm_dd_yy*_release.txt
- GTAssay-Non-Human_*mm_dd_yy*_release.txt
- GTAssay-Gene_*mm_dd_yy*_release.txt
- Set-GTAssay_*mm_dd_yy*_release.txt

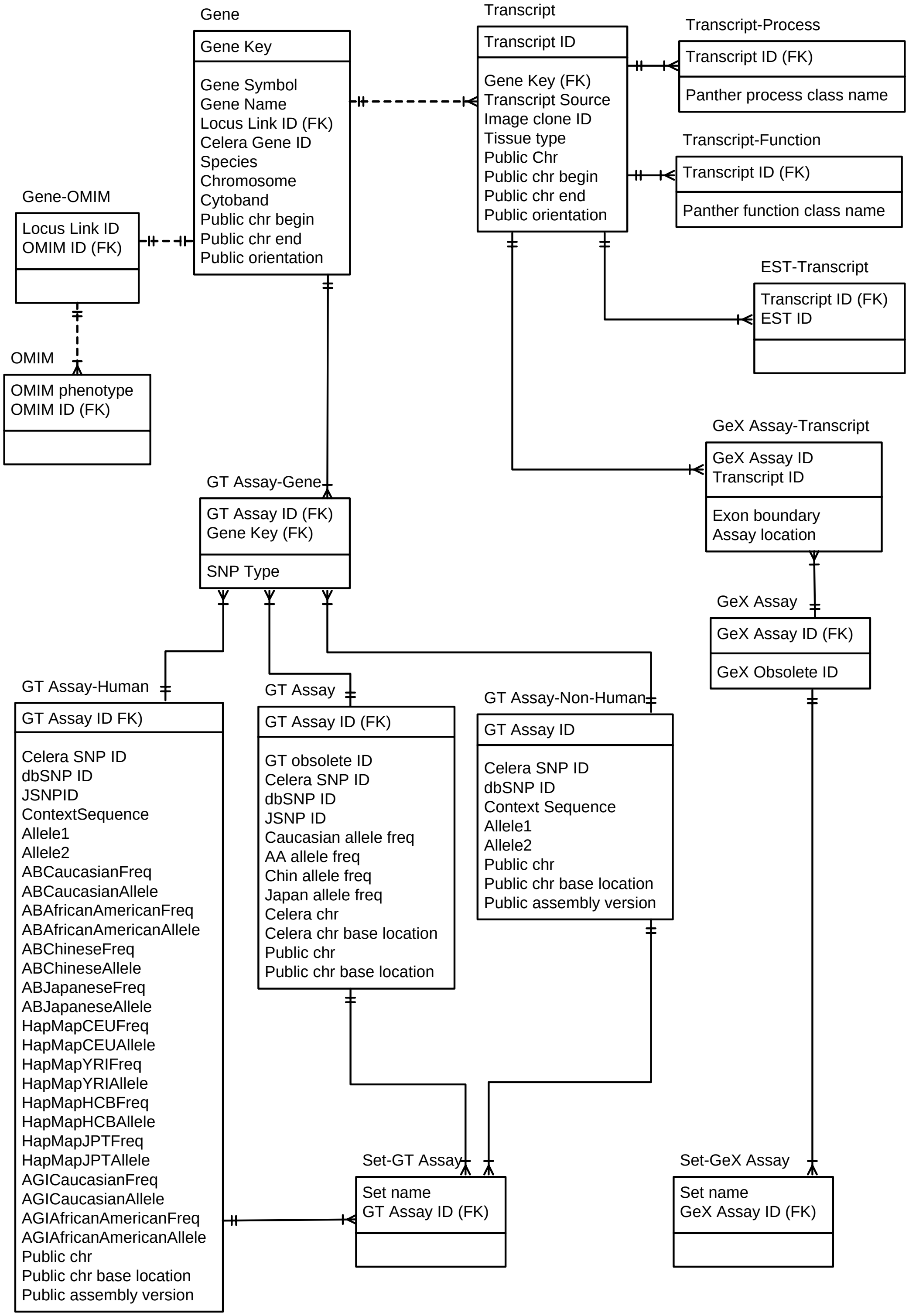
Data Type Definition

The following table defines the possible data types used in these files:

Data Type	**Definition**
NUMERIC (size)	Number string with up to size digits.
FLOAT (size)	Floating point number.
VARCHAR (size)	Variable length character string with up to size characters.

Data Model

In the following diagram, primary keys appear in the top of the file boxes.



Gene

This is a tab-delimited file that contains all gene attributes.

The file name is:

Gene_*mm_dd_yy*_release.txt

The format for each entry is:

```
Gene Key^Gene Symbol^Gene Name^LocusLink ID^Celera Gene
ID^Species^Chromosome^Cytoband^Public Chr Begin^Public Chr End^Public
Orientation
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
H_7184_21864^TRA1^tumor rejection antigen (gp96) 1^
7184^hCG21864^Homo sapiens^12^12q24.2-q24.3   104257107^104280369^1
```

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
Gene Key	VARCHAR (53)	Primary Key, and Foreign Key to GT ASSAY-GENE and TRANSCRIPT. A unique key to enable table joins. You should use this field only for table joins.
Gene Symbol	VARCHAR (32)	The LocusLink gene symbol (e.g., CALCB).
Gene Name	VARCHAR (256)	The LocusLink gene name (e.g., calcitonin-related polypeptide, beta).
LocusLink ID	VARCHAR (32)	Foreign Key to GENE-OMIM. The unique LocusLink ID for the corresponding gene.
Celera Gene ID	VARCHAR (32)	The corresponding Celera gene (hCG or mCG) ID; for example, hCG12345.
Species	VARCHAR (32)	One of the following: • Homo sapiens • Mus musculus • Rattus norvegicus
Chromosome	VARCHAR (20)	The chromosome number.
Cytoband	VARCHAR (256)	The cytoband location, if available.
Public Chr Begin	NUMERIC (20)	The begin position of the gene on the chromosome.
Public Chr End	NUMERIC (20)	The begin position of the gene on the chromosome.
Public Orientation	VARCHAR (4)	The orientation. One of the following appears: • 1: forward • -1: reverse

Gene-OMIM

This is a tab-delimited file that contains all Locus Link Gene - OMIM ID associations.

The file name is:

```
Gene-OMIM_mm_dd_yy_release.txt
```

The format for each entry is:

```
Locus Link ID^OMIM ID
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
196385^605884
```

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
Locus Link ID	VARCHAR (32)	Part of Primary Key. The unique LocusLink ID for the corresponding gene.
OMIM ID	NUMERIC (20)	Part of Primary Key, and Foreign Key to OMIM. The OMIM ID.

OMIM

This is a tab-delimited file that contains all OMIM ID-phenotype associations.

The file name is:

```
OMIM_mm_dd_yy_release.txt
```

The format for each entry is:

```
OMIM ID^OMIM Phenotype
```

A caret (^) indicates a tab. If a field is empty, no data appear.
For example:

```
607273^Birt-Hogg-Dube syndrome
```

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
OMIM ID	NUMERIC (20)	Part of Primary Key. The OMIM ID.
OMIM Phenotype	VARCHAR (256)	Part of Primary Key. The OMIM phenotype. NOTE: It is possible to have multiple phenotypes for the same OMIM ID. In this case, there will be multiple rows for the OMIM ID.

Transcript

This is a tab-delimited file that contains all gene-transcript associations.

The file name is:

```
Transcript_mm_dd_yy_release.txt
```

The format for each entry is:

```
Gene Key^Transcript Source^Transcript ID^Image Clone ID^Tissue Type^Public
Chr^Public Chr Begin^Public Chr End^Public Orientation
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
R_245925_0^RefSeq^NM_139093^hippocampus^1^201103717^
201136571^1
```

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
Gene Key	VARCHAR (53)	Primary Key and Foreign key to GENE. A unique key to enable table joins. Note: You should use this field only for table joins.
Transcript Source	VARCHAR (32)	One of the following: • RefSeq • GenBank • Celera
Transcript ID	VARCHAR (32)	Foreign Key to GEX ASSAY-TRANSCRIPT, TRANSCRIPT-PROCESS, TRANSCRIPT-FUNCTION, and ESTTRANSCRIPT. The transcript identifier or accession number as appropriate for the Transcript Source.
Image Clone ID	VARCHAR (32)	The Mammalian Clone Collection (MCG) ID.
Tissue Type	VARCHAR (256)	The tissue type.
Public Chr	VARCHAR (20)	The chromosome number on the public genome.
Public Chr Begin	NUMERIC (20)	The begin position of the transcript on the chromosome.
Public Chr End	NUMERIC (20)	The end position of the transcript on the chromosome.
Public Orientation	VARCHAR (4)	The orientation. One of the following appears: 1: forward -1: reverse

GeX Assay-Transcript

This is a tab-delimited file that contains all transcript-gene expression assay associations.

The file name is:

`GexAssay-Transcript_mm_dd_yy_release.txt`

The format for each entry is:

`GeX Assay ID^Transcript ID^Exon Boundary^Assay Location`

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

`Hs00378029_m1^hCT2340436^2^372`

The following table describes the fields that appear in this file:

Field Descriptions

Field	**Data Type**	**Definition**
GeX Assay ID	VARCHAR (16)	Part of Primary Key. The AB Gene Expression Assay ID. Note: The second to last position in the Assay ID contains a letter that indicates the assay placement. An “m” indicates an assay whose probe spans an exon junction and will not detect genomic DNA. An “s” indicates an assay whose primers and probes are designed within a single exon, such assays will, by definition detect genomic DNA. A “g” indicates an assay that may detect genomic DNA. RNA samples should be free from genomic DNA when using assays with either an “s” or a “g.”
Transcript ID	VARCHAR (32)	Part of primary key and foreign key to TRANSCRIPT. The transcript identifier or accession number as appropriate for the Transcript Source.
Exon Boundary	NUMERIC (38)	The number of the exon on the 5’-side of the exon pair spanned by the assay, as displayed on the myScience website.
Assay Location	NUMERIC (38)	The base location of the midpoint of context sequence, as displayed on the myScience website.

GeX Assay

This tab-delimited file maps current Gene Expression Assay IDs to previous versions of the Assay IDs.

The file name is:

`GexAssay_mm_dd_yy_release.txt`

A caret (^) indicates a tab. If a field is empty, no data appear.

The format for each entry is:

`GeX Assay ID^GeX Obsolete ID`

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
GeX Assay ID	VARCHAR (16)	Primary key, and Foreign Key to GEX ASSAYTRANSCRIPT and SET-GEX ASSAY. The AB Gene Expression Assay ID. Note: The second to last position in the Assay ID contains a letter that indicates the assay placement. An “m” indicates an assay whose probe spans an exon junction and will not detect genomic DNA. An “s” indicates an assay whose primers and probes are designed within a single exon, such assays will, by definition detect genomic DNA. A “g” indicates an assay that may detect genomic DNA. RNA samples should be free from genomic DNA when using assays with either an “s” or a “g.”
GeX Obsolete ID	VARCHAR (16)	A previous version for this assay ID, if applicable.

Transcript-Process

This is a tab-delimited file that contains Panther Protein biological process assignments for transcripts.

The file name is:

`Transcript-Process_mm_dd_yy_release.txt`

The format for each entry is:

`Transcript ID^Panther process class name`

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

`NM_000016^Acyl-CoA metabolism`

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
Transcript ID	VARCHAR (32)	Primary Key. The transcript identifier or accession number as appropriate for the Transcript Source.
Panther process class name	VARCHAR (256)	The Panther protein biological process name assigned to the corresponding transcript.

Transcript-Function

This is a tab-delimited file that contains Panther Protein molecular function assignments for transcripts.

The file name is:

```
Transcript-Function_mm_dd_yy_release.txt
```

The format for each entry is:

```
Transcript ID^Panther function class name
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
NM_000014^Protease inhibitor
```

The following table describes the fields that appear in this file:

Field	Data Type	Definition
Transcript ID	VARCHAR (32)	Primary Key. The transcript identifier or accession number as appropriate for the Transcript Source.
Panther function class name	VARCHAR (256)	The Panther protein molecular function name assigned to the corresponding transcript.

Set-GeX Assay

This is a tab-delimited file that indicates the assay set to which the gene expression assay belongs.

The file name is:

`Set-GexAssay_mm_dd_yy_release.txt`

The format for each entry is:

`Set Name^GeX Assay ID`

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

`MGC^Hs00152844_m1`

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
Set name	VARCHAR (20)	Part of Primary key. One of the following: • *Homo sapiens* • *Mus musculus* • *Rattus norvegicus* • ITN: assay is associated with a gene related to immune tolerance • MGC: assay interrogates a GenBank mRNA sequenced by NIH's Mammalian Gene Collection project • P450: assay is designed to P450 genes • Endogenous Controls: assay is associated with a set of housekeeping genes that can be used as experimental controls • Inventoried: a TaqMan® product that is kept in stock and will generally be shipped in 1 - 2 business days • Ambion siRNA available: assay has corresponding Silencer™ Validated and/or Pre-designed siRNAs from Ambion
GeX Assay ID	VARCHAR (16)	Part of Primary Key, and Foreign Key to GEX ASSAY. The AB Gene Expression Assay ID.

EST-Transcript

This is a tab-delimited file that associates transcripts to GenBank EST IDs.

The file name is:

```
EST-Transcript_mm_dd_yy_release.txt
```

The format for each entry is:

```
Transcript ID^EST ID
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
hCT2324646^AW327245
```

The following table describes the fields that appear in this file:

Field	Data Type	Definition
Transcript ID	VARCHAR (32)	Part of primary key. The transcript identifier or accession number as appropriate for the Transcript Source.
EST ID	VARCHAR (20)	Part of primary key. The GenBank EST ID.

GT Assay-Gene

This is a tab-delimited file that contains all gene-SNP genotyping assay associations.

The file name is:

```
GTAssay-Gene_mm_dd_yy_release.txt
```

The format for each entry is:

```
GT Assay ID^Gene Key^SNP Type
```

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

```
C___1838474_1_^H_94031_17387^Intron
```

The following table describes the fields that appear in this file:

Field	Data Type	Definition
GT Assay ID	VARCHAR (16)	Part of Primary Key. The SNP Genotyping AB Assay ID.
Gene Key	VARCHAR (53)	Part of Primary Key. A unique key to enable table joins. Note: You should use this field only for table joins.
SNP Type	VARCHAR (131)	One or more of the following appears: • acceptor splice site • coding region • donor splice site • exon • intergenic/unknown • intron • mis-sense mutation • nonsense mutation • putative utr 3 • putative utr 5 • silent mutation • utr 3 • utr 5

GT Assay

This is a tab-delimited file that contains all SNP genotyping assays' properties.

The file name is:

```
GTAssay_mm_dd_yy_release.txt
```

The format for each entry is:

```
GT Assay ID^GT Obsolete ID^Celera SNP ID^dbSNP ID^JSNP ID^
Caucasian allele freq^AA allele freq^Chin allele freq^Japan allele
freq^Celera chr^Celera chr base location^Public chr^Public chr base location
```

A caret (^) indicates a tab. If a field is empty, no data appear.
For example:

```
C___2517344_1_^hCV2517344^.44^.29^.22^.37^22^26458963^2239255149
```

The following table describes the fields that appear in this file:

Field	Data Type	Definition
GT Assay ID	VARCHAR (16)	Primary Key, and Foreign Key to GT ASSAYGENE and SET-GT ASSAY. The SNP Genotyping AB Assay ID.
GT Obsolete ID	VARCHAR (16)	A previous version for this assay ID, if applicable. This field is currently not used.
Celera SNP ID	VARCHAR (16)	The corresponding Celera RefSNP ID; for example, hCV12345.
dbSNP ID	VARCHAR (16)	The NCBI dbSNP cluster ID (rs #).
JSNP ID	VARCHAR (20)	The JSNP ID.
Caucasian allele freq	FLOAT (8)	The minor allele frequency in Caucasian populations.
AA allele freq	FLOAT (8)	The minor allele frequency in African American populations.
Chin allele freq	FLOAT (8)	The minor allele frequency in Chinese populations.
Japan allele freq	FLOAT (8)	The minor allele frequency in Japanese populations.
Celera chr	VARCHAR (20)	The chromosome number on the Celera reference genome.
Celera chr base location	NUMERIC (20)	The position on the Celera genome reference axis.
Public chr	VARCHAR (20)	The chromosome number on the public genome.
Public chr base location	NUMERIC (20)	The position on the public genome reference axis.

GTAssay-Human

This is a tab-delimited file that contains all SNP genotyping assays' properties.

The file name is:

GTAssay-Human_*mm_dd_yy*_release.txt

The format for each entry is:

```
GT Assay_ID^Celera SNP ID^dbSNP ID^JSNPID^ContextSequence^Allele1^Allele2^
ABCaucasianFreq^ABCaucasianAllele^ABAfricanAmericanFreq^
ABAfricanAmericanAllele^ABChineseFreq^ABChineseAllele^ABJapaneseFreq^
ABJapaneseAllele^HapMapCEUFreq^HapMapCEUAllele^HapMapYRIFreq^HapMapYRIAllele^
HapMapHCBFreq^HapMapHCBAllele^HapMapJPTFreq^HapMapJPTAllele^AGICaucasianFreq^
AGICaucasianAllele^AGIAfricanAmericanFreq^AGIAfricanAmericanAllele^
Public chr^Public chr base location^Public assembly version
```

A caret (^) indicates a tab. If a field is empty, no data appears.

For example:

```
C___8739217_10^hCV8739217^rs2011404^^TTTTAACAGACCCCGTTAACCTCTG[C/T]GGGGCGGTGC
TGGCTAAGTACCTGT^C^T^0.17^T^0.04^T^0.02^T^0.01^T^0.15^T^0^T^0^T^0^T^0.18^T^0.0
9^T^2^234292676^NCBI Build 36.2 Human
```

The following table describes the fields that appear in this file:

Field	Data Type	Definition
GT Assay ID	VARCHAR (16)	Primary Key, and Foreign Key to GT ASSAY-GENE and SET-GT ASSAY. The SNP Genotyping AB Assay ID.
Celera SNP ID	VARCHAR (16)	The corresponding Celera RefSNP ID (hCV #)
dbSNP ID	VARCHAR (16)	The NCBI dbSNP cluster ID (rs #).
JSNP ID	VARCHAR (32)	The JSNP ID (IMS-JST #).
ContextSequence	VARCHAR (55)	The SNP site, with up to 25 nucleotides flanking sequence on each side, is provided in the (+) genome strand orientation of the reference genome. The order of the bracketed SNP alleles specifies their probe reporter dye association, where [Allele 1 = VIC / Allele 2 = FAM].
Allele1	VARCHAR (16)	SNP allele reported by VIC probe
Allele2	VARCHAR (16)	SNP allele reported by FAM probe
ABCaucasianFreq	FLOAT (8)	The minor allele frequency in AB Caucasian population panel.
ABCaucasianAllele	FLOAT (8)	The minor allele in AB Caucasian population panel.
ABAfricanAmericanFreq	FLOAT (8)	The minor allele frequency in AB African American population panel.
ABAfricanAmericanAllele	FLOAT (8)	The minor allele in AB African American population panel.
ABChineseFreq	FLOAT (8)	The minor allele in AB Chinese population panel.
ABChineseAllele	FLOAT (8)	The minor allele frequency in AB Chinese population panel.
ABJapaneseFreq	FLOAT (8)	The minor allele frequency in AB Japanese population panel.
ABJapaneseAllele	FLOAT (8)	The minor allele in AB Japanese population panel.

Field	Data Type	Definition
HapMapCEUFreq	FLOAT (8)	The minor allele frequency in HapMap CEU population panel.
HapMapCEUAllele	FLOAT (8)	The minor allele in HapMap CEU population panel.
HapMapYRIFreq	FLOAT (8)	The minor allele frequency in HapMap YRI population panel.
HapMapYRIAllele	FLOAT (8)	The minor allele in HapMap YRI population panel.
HapMapHCBFreq	FLOAT (8)	The minor allele frequency in HapMap HCB population panel.
HapMapHCBAllele	FLOAT (8)	The minor allele in HapMap HCB population panel.
HapMapJPTFreq	FLOAT (8)	The minor allele frequency in HapMap JPT population panel.
HapMapJPTAllele	FLOAT (8)	The minor allele in HapMap JPT population panel.
AGICaucasianFreq	FLOAT (8)	The minor allele frequency in AGI Caucasian population panel.
AGICaucasianAllele	FLOAT (8)	The minor allele in AGI Caucasian population panel.
AGIAfricanAmericanFreq	FLOAT (8)	The minor allele frequency in AGI AfricanAmerican population panel.
AGIAfricanAmericanAllele	FLOAT (8)	The minor allele in AGI AfricanAmerican population panel.
Public chr	VARCHAR (20)	The chromosome number on the Public reference genome.
Public chr base location	NUMERIC (20)	The position on the Public genome reference axis.
Public assembly version	VARCHAR (32)	The build version and species of the Public reference genome.

GTAssay-Non-Human

This is a tab-delimited file that contains all SNP genotyping assays' properties.

The file name is:

GTAssay-Non-Human_*mm_dd_yy*_release.txt

The format for each entry is:

```
GT Assay_ID^Celera SNP ID^dbSNP ID^ ContextSequence^Allele1^Allele2 ^
Public chr^Public chr base location^Public assembly version
```

A caret (^) indicates a tab. If a field is empty, no data appears.

For example:

```
M__45282052_20^mCV45282052^rs4135869^TAGAACAGAGCAGTTGGACTGCCCA[A/G]CACCTTATGA
TGGTTTACCCTAAGA^A^G^17^24166142^NCBI Build 36.1 Mouse
```

The following table describes the fields that appear in this file:

Field	**Data Type**	**Definition**
GT Assay ID	VARCHAR (16)	Primary Key, and Foreign Key to GT ASSAY-GENE and SET-GT ASSAY. The SNP Genotyping AB Assay ID.
Celera SNP ID	VARCHAR (16)	The corresponding Celera RefSNP ID (mCV #)
dbSNP ID	VARCHAR (16)	The NCBI dbSNP cluster ID (rs #).
ContextSequence	VARCHAR (55)	The SNP site, with up to 25 nucleotides flanking sequence on each side, is provided in the (+) genome strand orientation of the reference genome. The order of the bracketed SNP alleles specifies their probe reporter dye association, where [Allele 1 = VIC / Allele 2 = FAM].
Allele1	VARCHAR (16)	SNP allele reported by VIC probe
Allele2	VARCHAR (16)	SNP allele reported by FAM probe
Public chr	VARCHAR (20)	The chromosome number on the Public reference genome.
Public chr base location	NUMERIC (20)	The position on the Public genome reference axis.
Public assembly version	VARCHAR (32)	The build version and species of the Public reference genome.

Set-GT Assay

This is a tab-delimited file that indicates the assay set to which the SNP genotyping assay belongs.

The file name is:

`Set-GTAssay_mm_dd_yy_release.txt`

The format for each entry is:

`Set Name^GT Assay ID`

A caret (^) indicates a tab. If a field is empty, no data appear.

For example:

`JSNP^C__10082020_1_`

The following table describes the fields that appear in this file:

Field	Data Type	Definition
Set name	VARCHAR (20)	Part of Primary Key. One of the following: • Homo sapiens • JSNP: assay is associated with a JSNP ID • Inventoried: a TaqMan® product that is kept in stock and will generally be shipped in 1 to 2 business days • Validated: Validated assays are TaqMan® SNP Genotyping Assays tested against four ethnic populations, consisting of approximately 45 individuals for each population. The data are used to establish that assays meet stringent performance criteria. Minor allele frequencies are determined and published for each individual population tested.
GT Assay ID	VARCHAR (16)	Part of Primary Key, and Foreign Key to GT-ASSAY. The AB SNP Genotyping Assay ID.