Phenotypic Mouse Allele Sequence Variant Annotation at Mouse Genome Informatics

Laurens G Wilming • Meiyee Law • Cynthia L Smith • Carol J Bult • MGI Team

The Jackson Laboratory • Bar Harbor • Maine • USA

The Jackson Laboratory

Leading the search for tomorrow's cures

ALLIANCE of GENOME RESOURCES

ABSTRACT

The power of the mouse as a model for human disease can only be fully exploited if researchers are able to find suitable mouse models for their human disease of interest. Many human diseases are ultimately caused by simple genomic mutations (single or multiple nucleotide variations (SNVs, MNVs) small insertions or deletions (indels)). However, the large number of genetic variants uncovered from individual patients presents challenges in identifying the causal gene or genomic regions. Although the Mouse Genome Informatics (MGI) database provides gene and genotype connections to phenotype annotations, the sequence context of the genome variants for phenotypic alleles was not yet available.

To provide researchers with a searchable, structured dataset of mouse mutations for comparative analysis, we have started annotating mouse variants, concentrating on SNVs, MNVs and small indels. These variants, characterized by their genomic position and sequence changes, are associated with engineered and spontaneous phenotypic alleles in the MGI database (www.informatics.jax.org). Variant attributes include variant type (insertion, point mutation, etc.) and molecular consequence (frameshift, stop gain, etc.). Data will be available in Human Genome Variation Society (HGVS) notation to provide transcript and protein contexts. Additional variants from large sequencing and mouse mutagenesis projects will be added to complement the manually curated data. By tying variant data and associated phenotype data to the genome, researchers will, in the near future, be able to search using human variants and find models with variants/mutations that result in the same amino acid change, have the same variant effect (missense, etc.), have the same functional impact (pathogenic, etc.), occur in the same protein domain(s), or have the same m of inheritance (recessive, dominant, etc.), resulting in phenotypes similar to a patient.

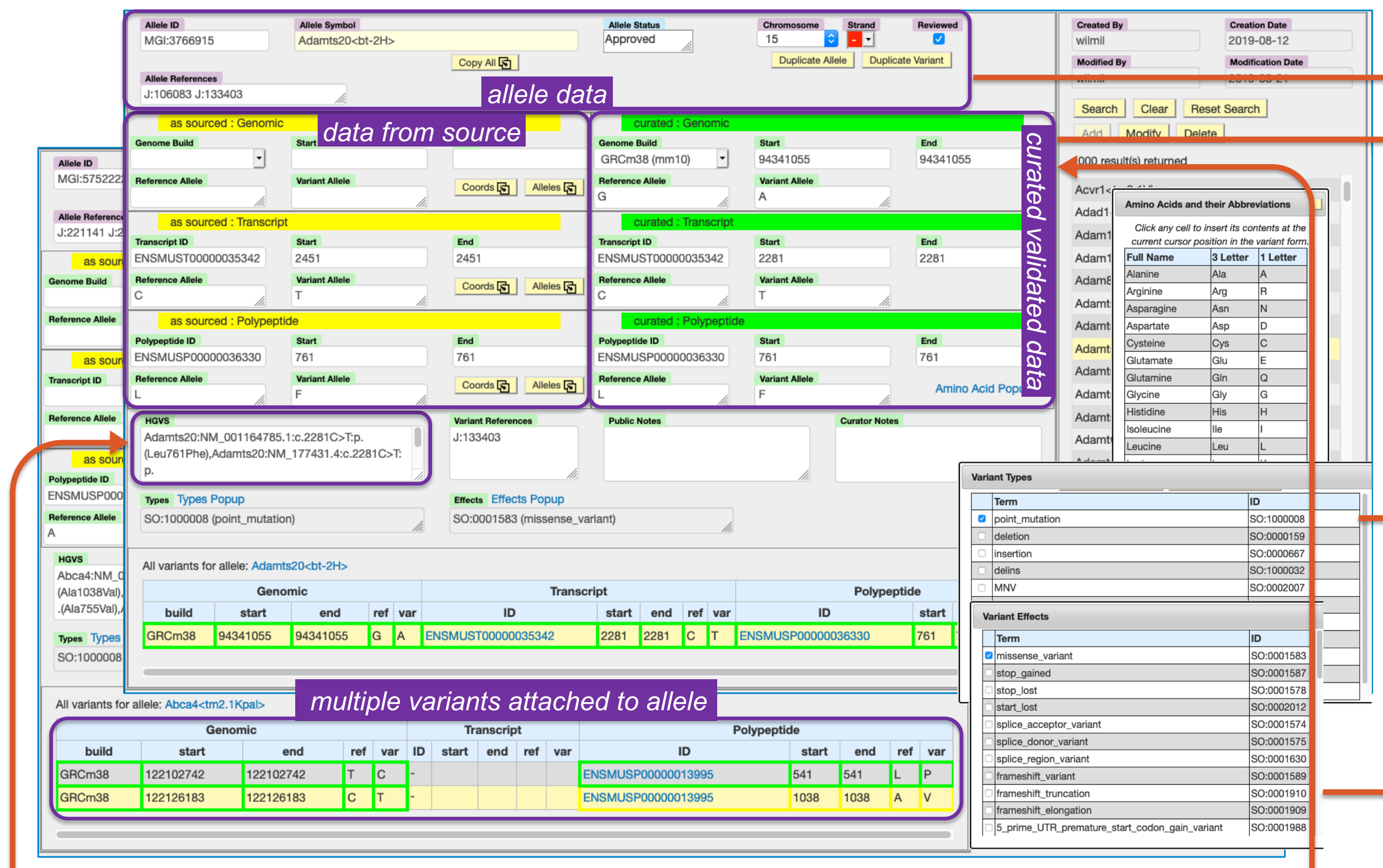
Variant data and associated phenotypic data is accessible from the Alliance of Genome Resources (www.alliancegenome.org) gene pages.

VARIANT ANNOTATION

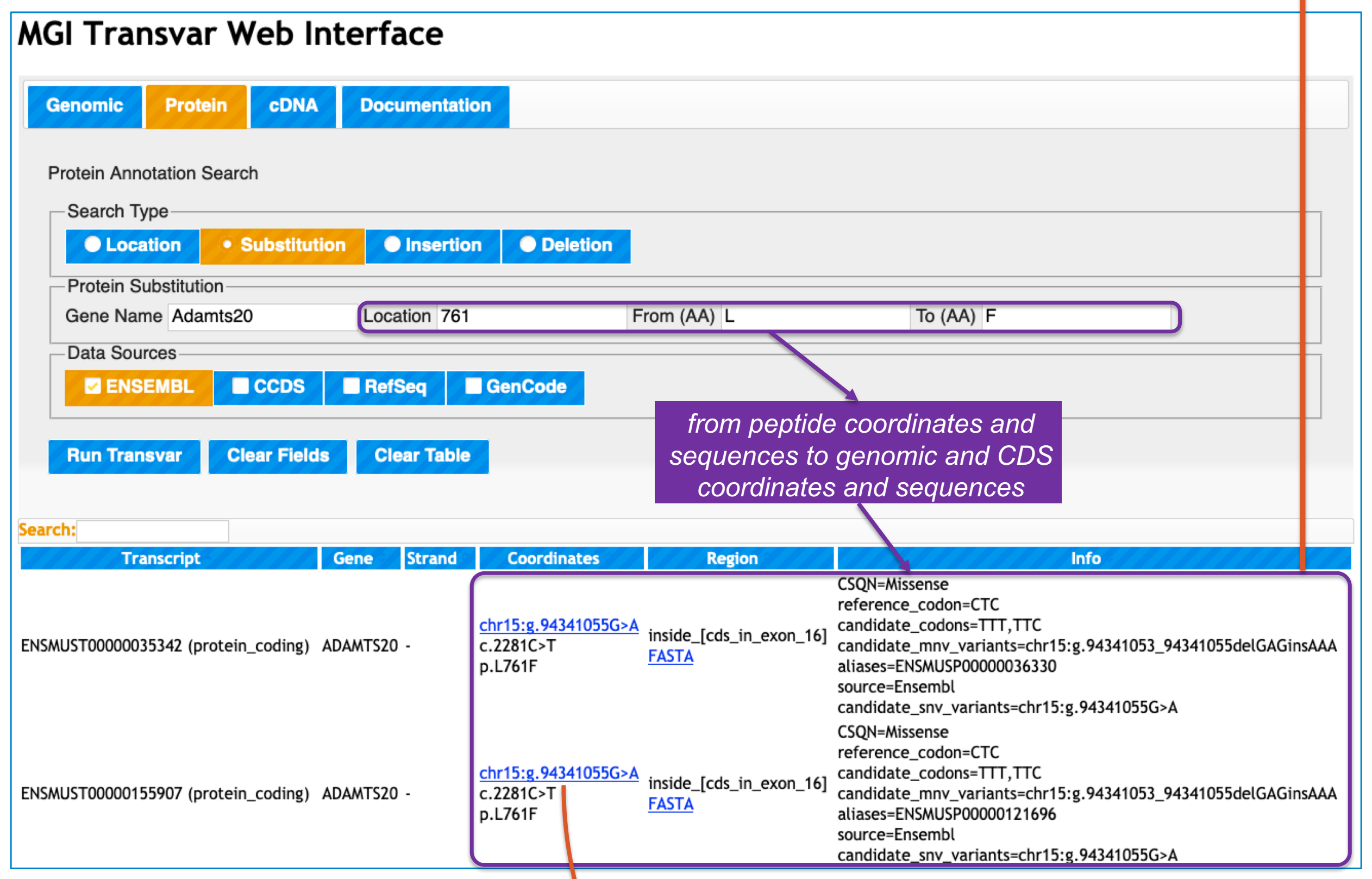
- Sequence variant data is mined from publications describing phenotypic alleles, and also provided by large-scale mutagenesis projects.
- Sources most often provide amino-acid changes, and often also transcript-level nucleotide changes.
- Sequence variants are defined by genomic coordinates and reference and variant sequences.
- Source information is entered into the web form, even if information is incorrect.
- Using various tools, source information is validated and genomic coordinates and sequence changes are determined.
- Validation of source data serves as quality control for published data:
  - a common error is using incorrect single-letter amino-acid codes (e.g. L is leucine, not lysine; lysine is K)
  - another common error is using c.466A>G type notation but using absolute transcript coordinates in stead of the proper CDS coordinates

ANNOTATION TOOLS

VARIANT ANNOTATION WEB FORM to create sequence variant records for phenotypic alleles



TRANSVAR web interface to translate between genomic, cDNA/CDS and peptide coordinates and sequence


MGI Transvar Web Interface

Protein Annotation Search

Search Type: ○ Location ● Substitution ○ Insertion ○ Deletion

Protein Substitution: Gene Name Adamts20 Location 761 From (AA) L To (AA) F

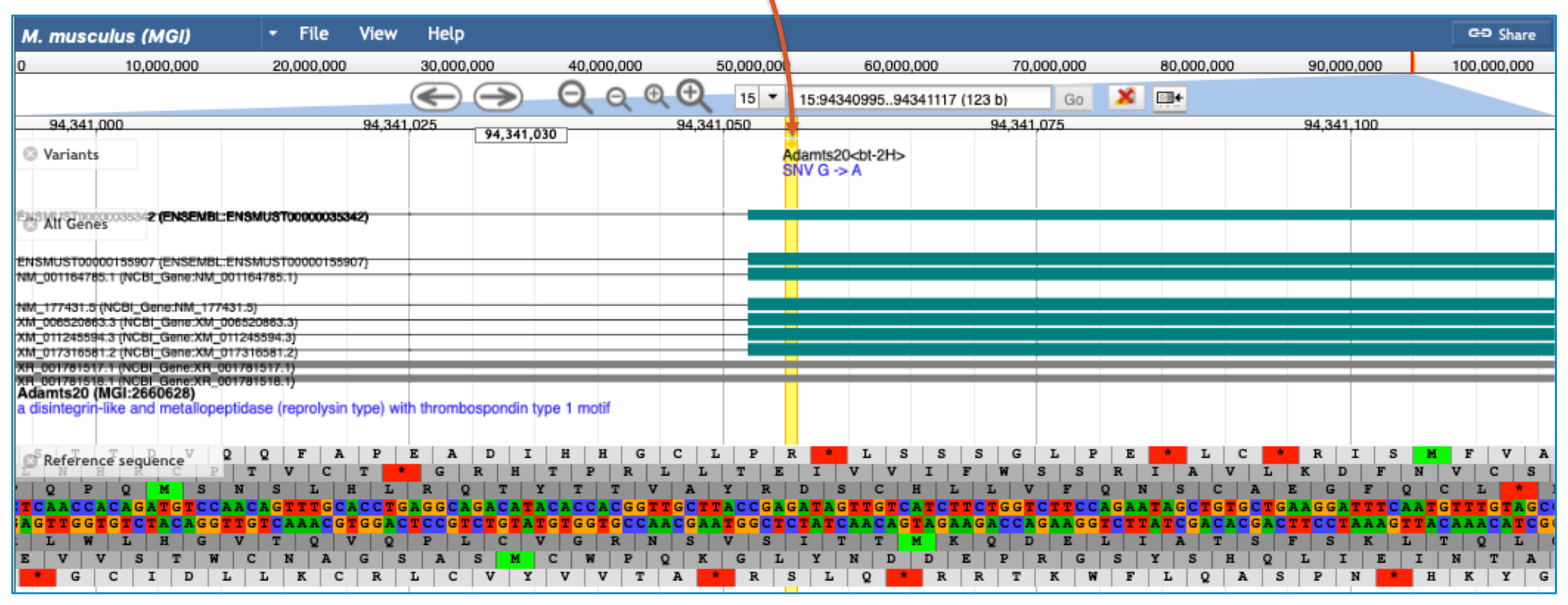
Data Sources: ☑ ENSEMBL ☐ CCDS ☐ RefSeq ☐ GenCode

from peptide coordinates and sequences to genomic and CDS coordinates and sequences

Search:

Transcript	Gene	Strand	Coordinates	Region	Info
ENSMUST00000035342 (protein_coding)	ADAMTS20	-	chr15:g.94341055G>A c.2281C>T p.L761F	inside_[cds_in_exon_16] FASTA	CSQN=Missense reference_codon=CTC candidate_codons=TTT,TTC candidate_mnv_variants=chr15:g.94341053_94341055delGAGinsAAA aliases=ENSMUSP00000036330 source=Ensembl candidate_snv_variants=chr15:g.94341055G>A
ENSMUST00000155907 (protein_coding)	ADAMTS20	-	chr15:g.94341055G>A c.2281C>T p.L761F	inside_[cds_in_exon_16] FASTA	CSQN=Missense reference_codon=CTC candidate_codons=TTT,TTC candidate_mnv_variants=chr15:g.94341053_94341055delGAGinsAAA aliases=ENSMUSP00000121696 source=Ensembl candidate_snv_variants=chr15:g.94341055G>A

JBROWSE genome browser

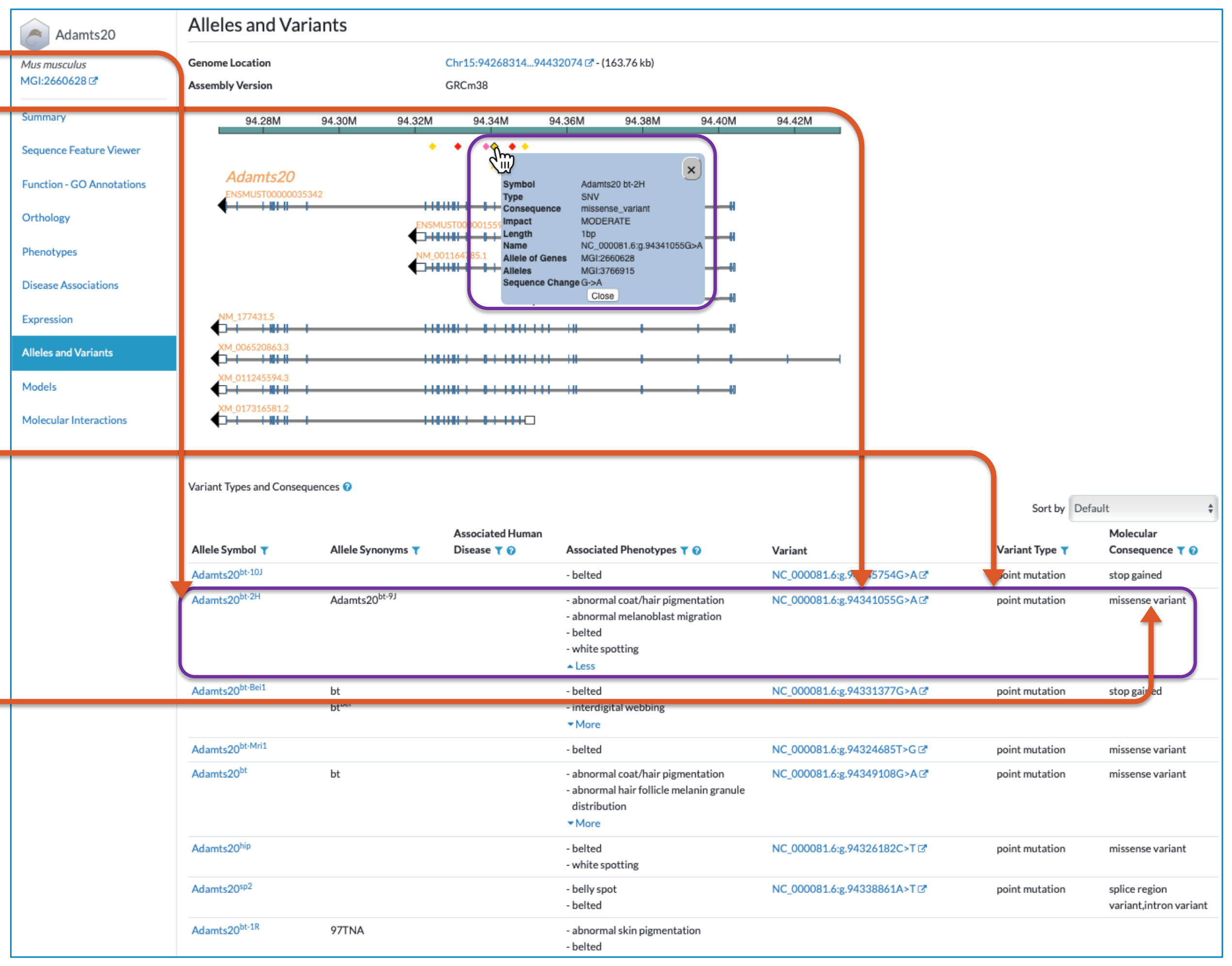


JANNOVAR script to create HGVS strings

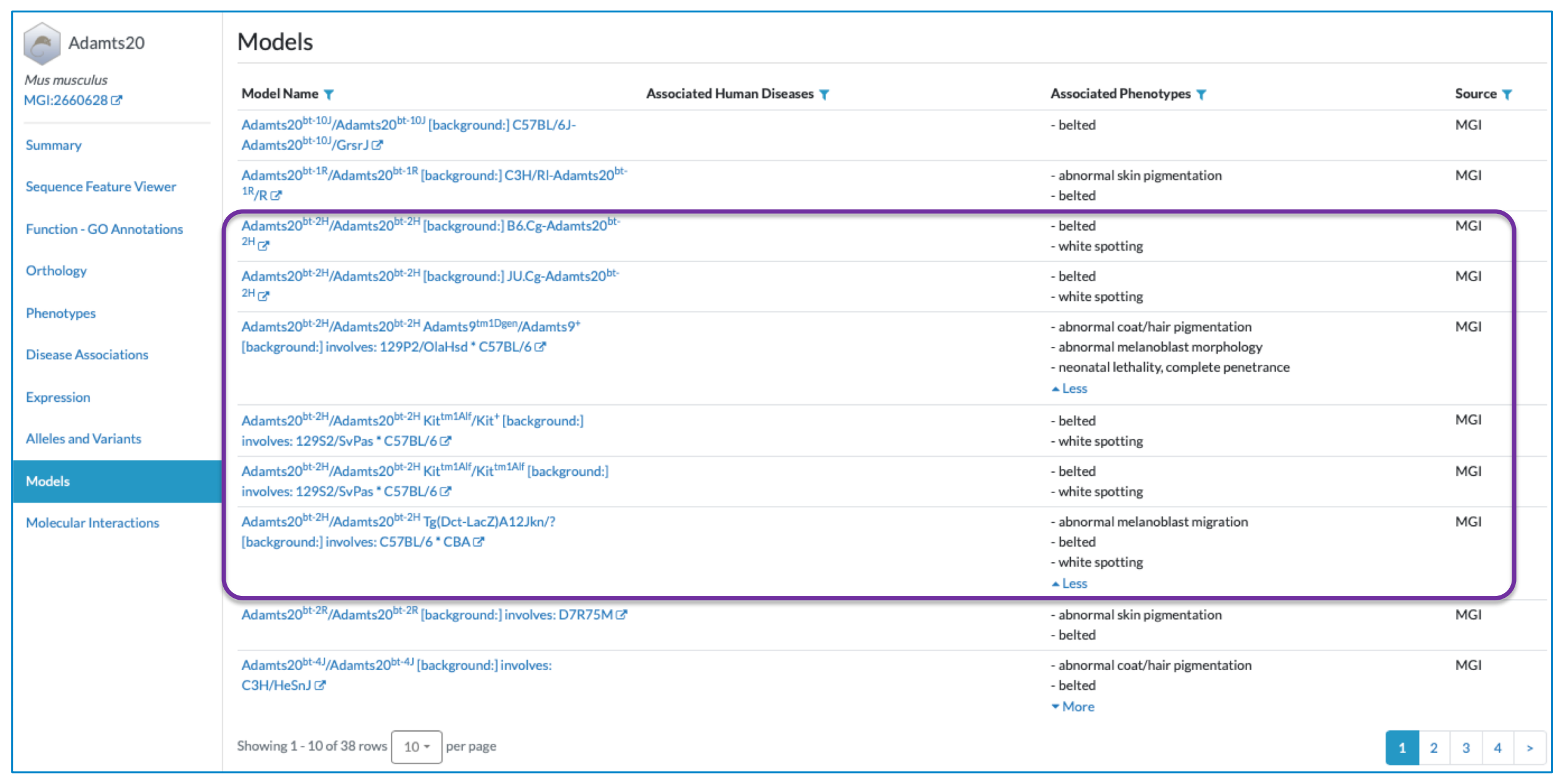
```
java -jar jannovar-cli-0.25.jar annotate-pos -d data/mm10_refseq.ser --3-letter-amino-acids --show-all -c 'chr15:94341055G>A'
MISSENSE_VARIANT,MISSENSE_VARIANT,MISSENSE_VARIANT,MISSENSE_VARIANT,MISSENSE_VARIANT,NON_CODING_TRANSCRIPT_EXON_VARIANT,NON_CODING_TRANSCRIPT_EXON_VARIANT
Adamts20:NM_001164785.1:c.2281C>T:p.(Leu761Phe),Adamts20:NM_177431.4:c.2281C>T:p.(Leu761Phe),Adamts20:XM_006520863.3:c.2308C>T:p.(Leu770Phe),Adamts20:XM_011245594.2:c.2281C>T:p.(Leu761Phe),Adamts20:XM_017316581.1:c.547C>T:p.(Leu183Phe),Adamts20:XR_001781517.1:n.1037C>T:,Adamts20:XR_001781518.1:n.2350C>T:
```

ALLIANCE GENE PAGE

AGR gene page Alleles & Variants section shows alleles and associated sequence variants and high level phenotypes



AGR gene page Models section shows mouse genotypes or strains and associated high level phenotypes and human diseases they model



QUALITY CONTROL

Annotation is subject to QC, both at the time of entry and after entry. Only entries that pass QC appear on GRC web page.

At time of entry, for example:

- Coordinates entered? ➞ sequence ID required
- Coordinates entered? ➞ End ≥ Start
- Sequence entered? ➞ ref-var pairs

After entry, for example:

- Type point_mutation? ➞ ref seq length = var seq length = 1
- Type insertion? ➞ ref seq length = 1, var seq length > 1
- Type deletion? ➞ ref seq length > 1, var seq length = 1
- Type MNV? ➞ ref seq length = var seq length > 1
- Ref seq different from var seq?
- Sequence ID associated with allele parent locus?
- Coordinates within genomic span of allele parent locus?

ACKNOWLEDGMENTS

Funding: NIH U41HG000330 (MGD) + NIH U24HG010859 (Alliance)

Software: Paul Hale

www.informatics.jax.org **www.alliancegenome.org** **www.jax.org**