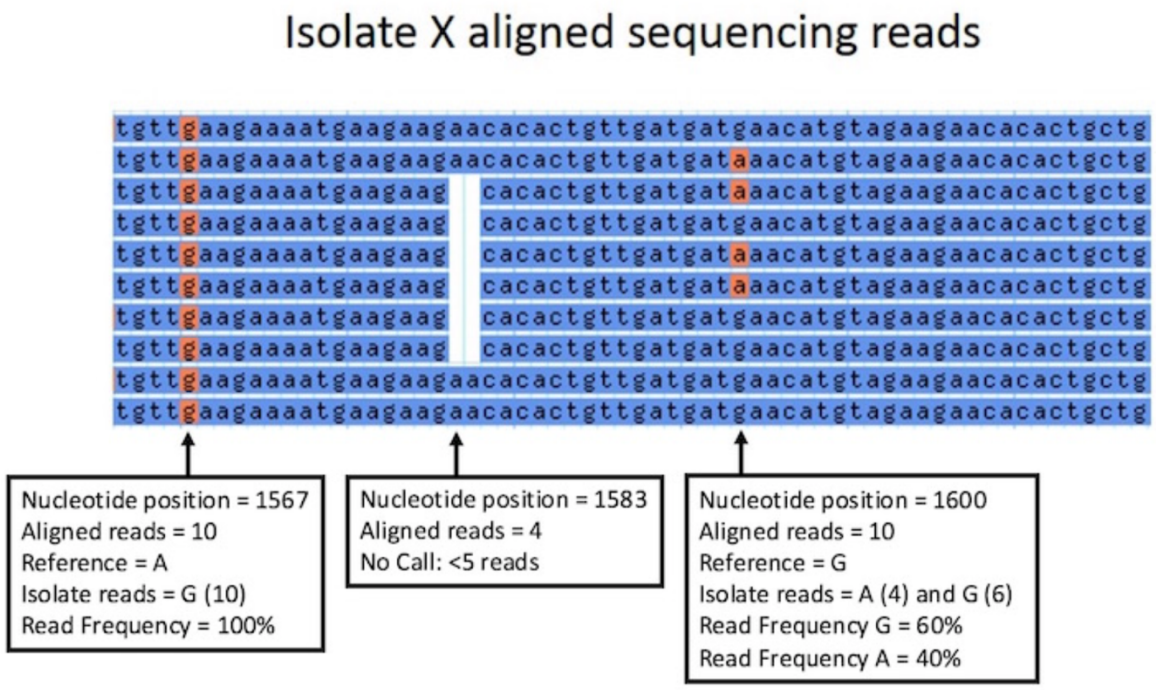
FungiDB: SNPs and Population Genetics

Single Nucleotide Polymorphisms (SNPs) can be used to characterize similarities and differences within a group of isolates or between two groups of isolates. They can also be used to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection).

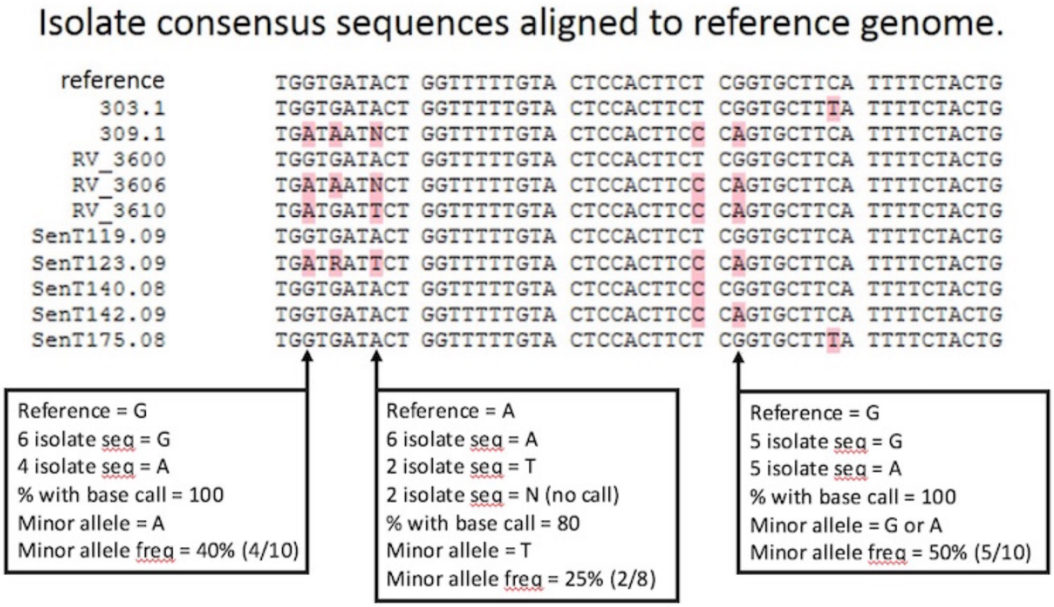
Isolates are assayed for SNPs in EuPathDB by two basic methods: re-sequencing and the alignment of sequence reads to a reference genome or DNA hybridization to a SNP-chip array.

Read Frequency Threshold: Calling SNPs for each isolate in your group.
Each isolate's sequencing reads are aligned to a reference genome (Organism) and then each nucleotide position with 5 or more aligned reads is examined. A base call is made if the aligned reads meet your Read Frequency Threshold. For example, Isolate X has 10 aligned reads at nucleotide position 1600. If 6 reads are G and 4 reads are A, the read frequency is 60% for the G call and 40% for A. Running this search with the Read Frequency Threshold set to 80% will prevent a base call and consequently exclude Isolate X when returning SNPs for nucleotide position 1600. Running the search with the Read Frequency Threshold set to 60% will bring back a G for this isolate and a 40% threshold will return two calls (both G and A) at this position. The parameter lets you control the quality of the sequencing data and the confidence of the SNP calls. Read Frequency Threshold is a particularly important parameter when dealing with diploid (or aneuploid) organisms since a read frequency of ~50% is expected for heterozygous SNPs.



Minor allele frequency: Parameter for calling SNPs across your isolate group.
The minor allele frequency refers to the least common base call for a single nucleotide position across all isolates. The default setting for this parameter is 0% and returns all SNPs - instances where at least one isolate has a base call that differs from reference. Increase the

Minor allele frequency to ensure that SNPs returned by the search are shared by a larger percentage of isolates in your group.



Percent isolates with a base call: Parameter for calling SNPs across your isolate group Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, a SNP can be returned by the search based on a subset of your isolate group. The 'Percent isolates with a base call' parameter defines the fraction of isolates that must have a base call before a SNP is returned for that nucleotide position. The default setting for this parameter is 80% or 8 out of 10 isolates in your group must have a base call for a SNP to be returned by the search. The higher this parameter, the more likely the SNP is to be high quality as regions difficult to align or difficult to sequence will tend to have a lower percentage of calls since the coverage and/or quality will be lower in that region.

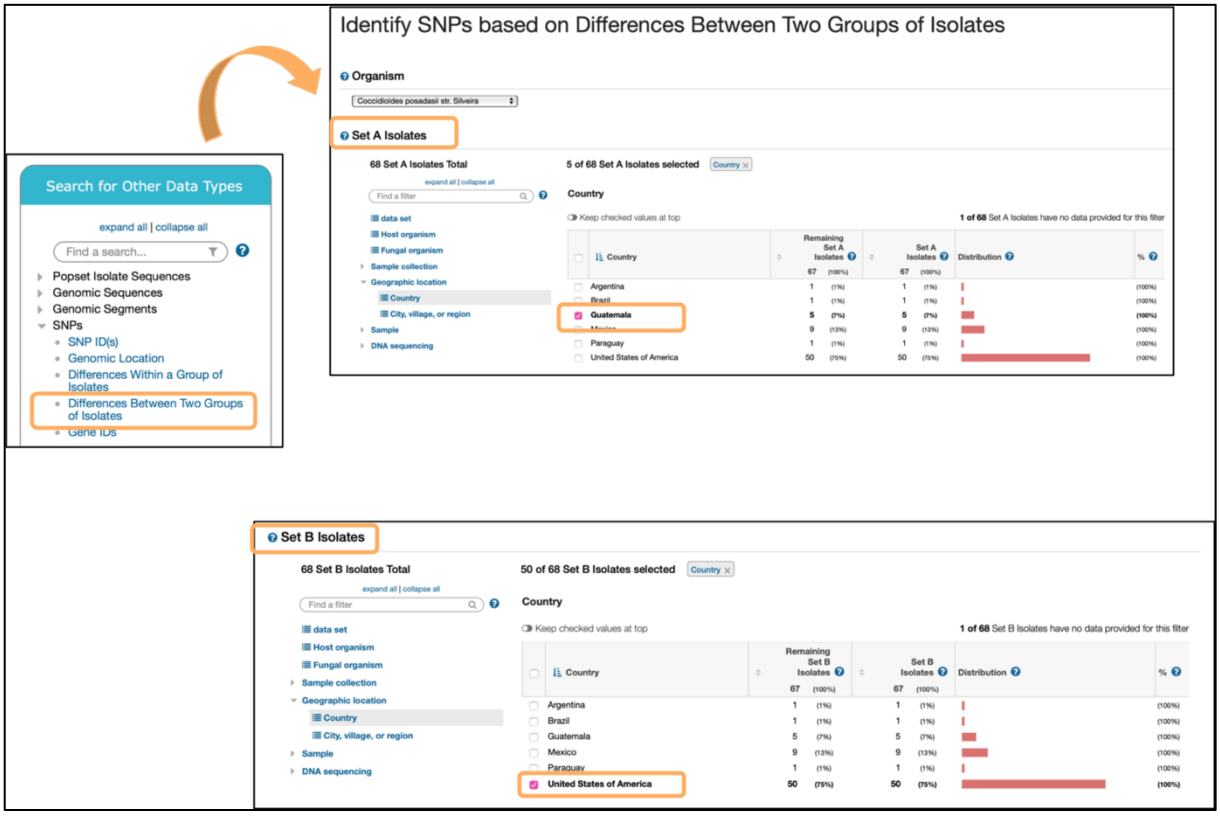
1. Identifying SNPs between fungal isolates collected in various geographical areas

The example described below identifies SNPs in *Coccidioides posadasii* (*C. posadasii*) str. Silveira isolates collected from patients with Coccidioidomycosis in the US and Latin America. Coccidioidomycosis, also known as Valley fever, is a fungal disease caused by two closely related species – *C. immitis* and *C. posadasii*. The disease is associated with high morbidity and mortality rates that affects tens of thousands of people each year. The two fungal species are endemic to several regions in the Western Hemisphere, but recent epidemiological and population studies suggest that the geographic range of these fungal species is becoming wider.

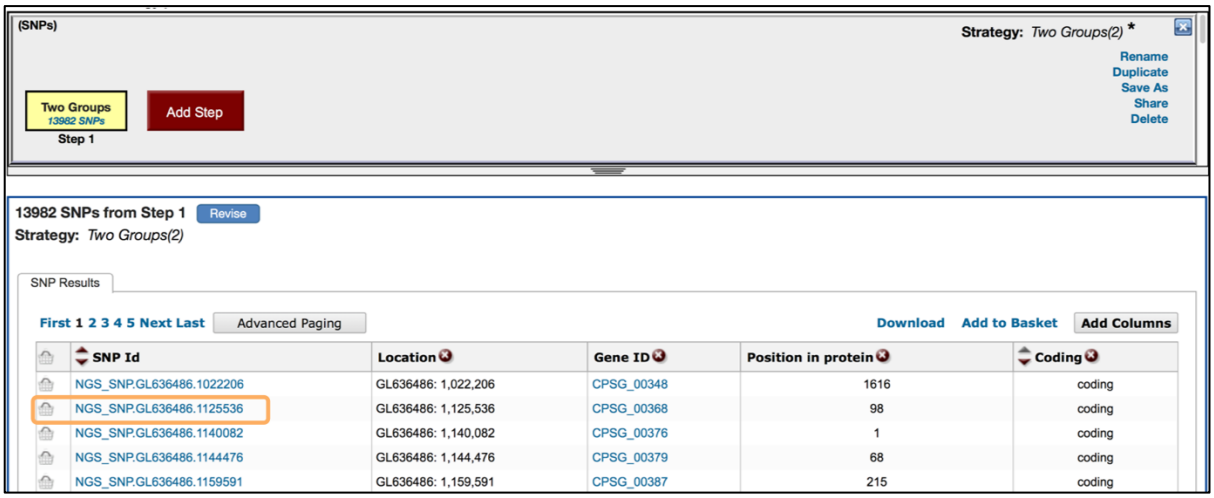
a) Identify SNPs based on differences between isolates collected in Guatemala and the US.

- From the *Search for Other Data Types* panel, navigate to the *Identify SNPs based on Differences Between Two Groups of Isolates*.

- In the resulting window first select the target organism '*C. posadasii* str. Silveira' then scroll through the metadata options on the left and make appropriate *Geographic Location* selections from the *Host* section of Characteristic separately for set A and set B isolates. *Set A isolates* should be set to *Guatemala* and *Set B* to the *United States of America*. All other parameters for both sets should be left as *default* (read frequency threshold – 80%, major allele frequency – 80, percent isolates with base call – 50).



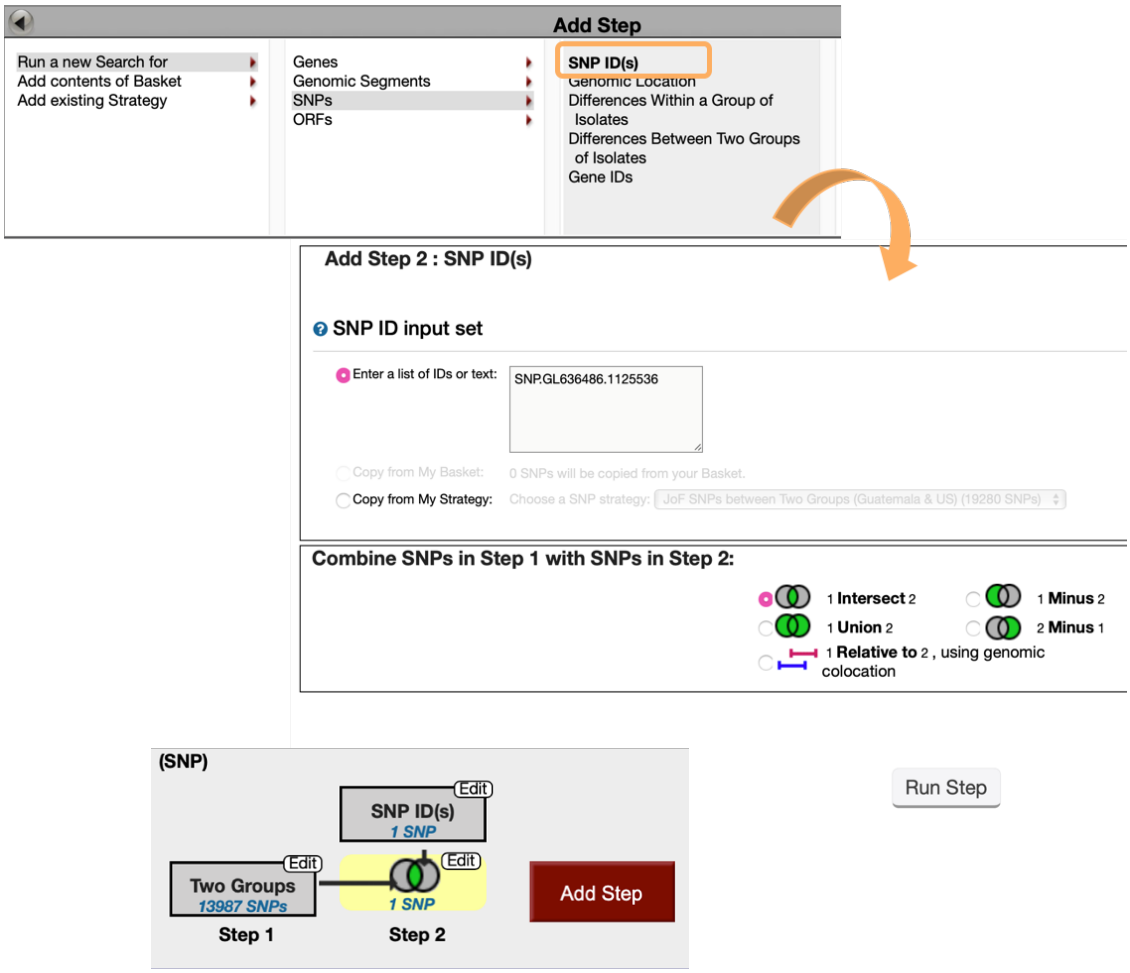
The search strategy returns SNPs rather than genes, which are classified by genomic location within the results table. When individual SNPs fall within a gene, its corresponding Gene ID



is listed next to the SNP record.

- To examine a SNP record page, click on the *SNP.GL636486.1125536* in the CPSG_00368 gene. *Note, you might have to scroll down to find the SNP or you can follow the next step.*
    - If your results table looks somewhat different and you cannot easily locate the SNP mentioned above – can you think of other ways to locate this SNP within your results?

Hint: Click Add Step and look up the SNP by its ID: SNP.GL636486.1125536



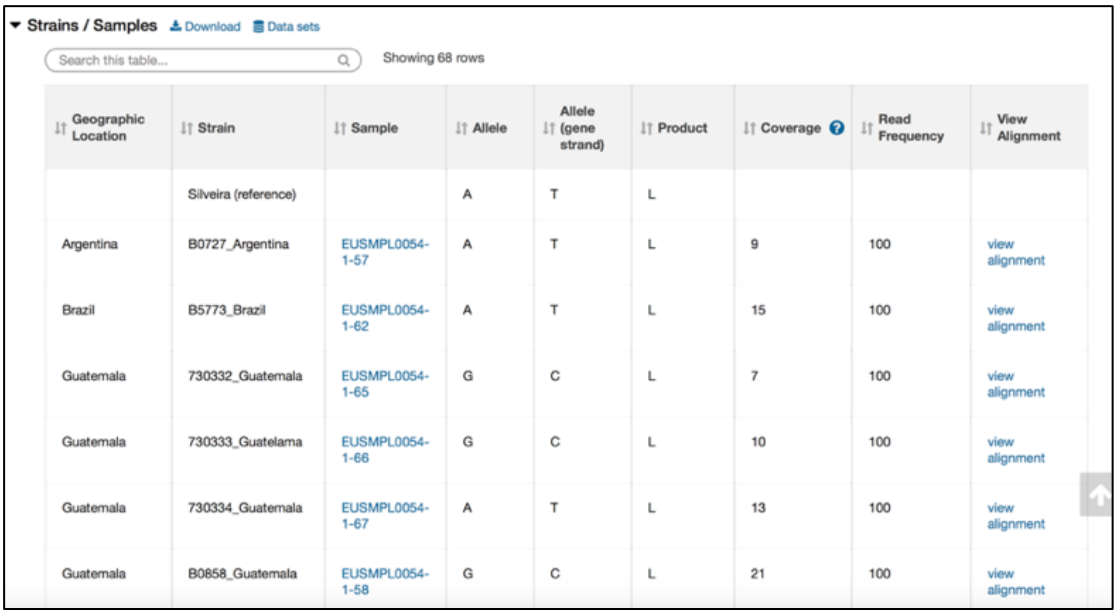
SNP location, allele summary, associated GeneID, major and minor allele records can be found at the top of the page, followed by DNA polymorphism summary and SNP records table that is searchable by isolates.

SNP: NGS_SNP.GL636486.1125536

Organism: Coccidioides posadasii str. Silveira
Location: GL636486: 1,125,536
Type: coding
Number of Strains: 68
Gene ID: CPSG_00368
Gene Strand: reverse
Major Allele: A (0.84)
Minor Allele: G (0.16)
Distinct Allele Count: 2
Reference Allele: A
Reference Product: L 98
Allele (gene strand): T
SNP context: CGTCCATCCTCTCACTCCCTGTCCCTGCCAAATCGGTGTCGAAGTGTGTGGCTGAGATCTC
SNP context (gene strand): GAGATCTCAGCCACACACTTCGACACCGATTTGGCAGGGACAGGGAGTGAGAGGATGGACG

Genomic location, SNP type and aligned reads can be displayed in GBrowse by clicking on the *View in genome browser* button. SNP tracks can be activated from the *Select Tracks* tab by selecting *SNPs by coding potential* under *DNA polymorphism* in the *Genetic variation* section. Hover over SNPs labeled as red diamonds (nonsense SNPs) to get more information.

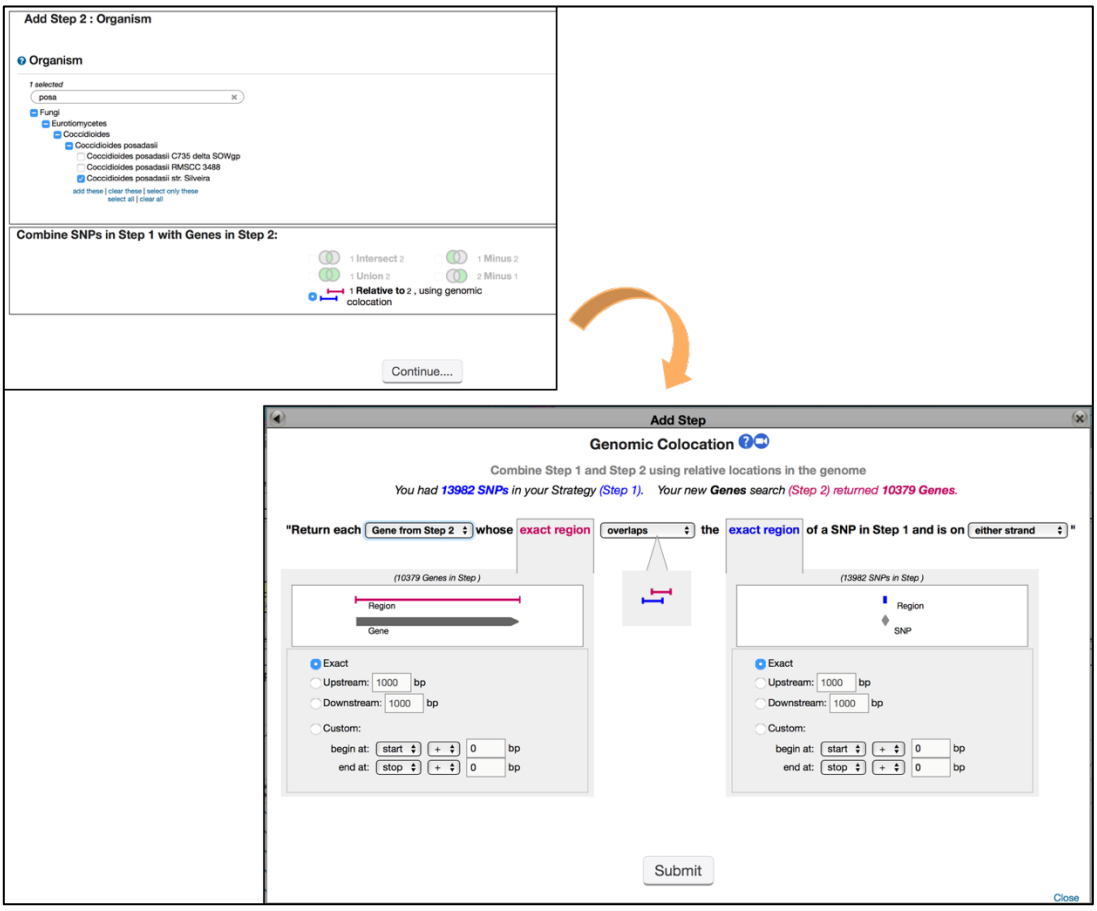
- Examine SNP record page further. Note that in addition to US and Guatemala SNP records it also contains information for other isolates collected elsewhere, where individual reads can be activated by clicking on the *view alignment* link from within the table. This action will re-direct you to the GBrowse where you can select either all or specific isolates listed under the *Aligned Genomic Sequence Reads for C. posadasii str. Silveira* to view specific tracks.



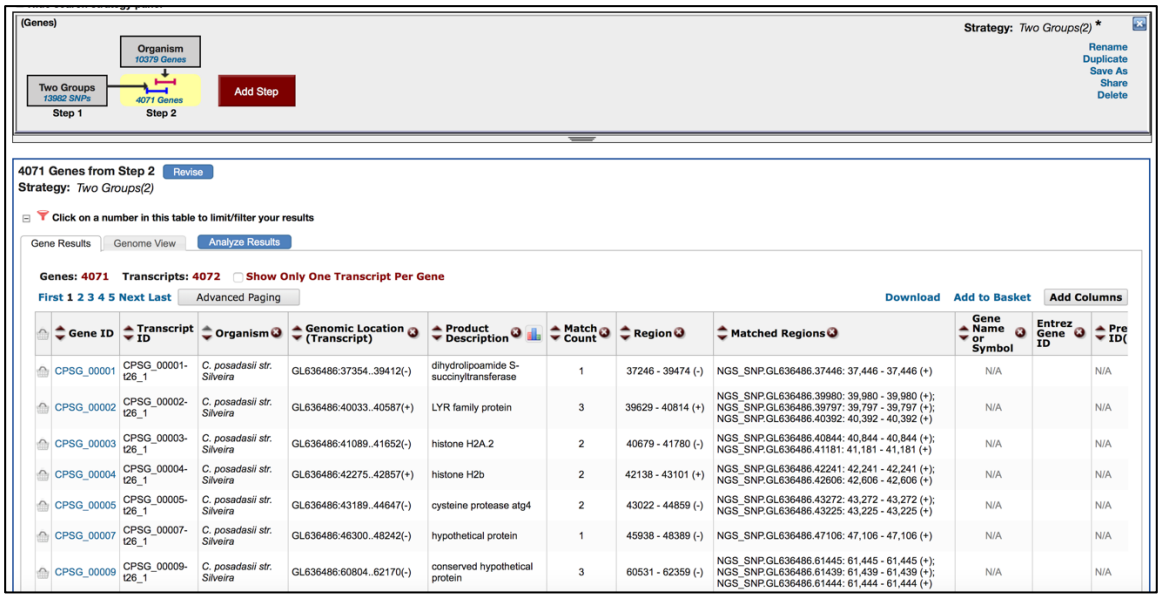
Geographic Location	Strain	Sample	Allele	Allele (gene strand)	Product	Coverage	Read Frequency	View Alignment
	Silveira (reference)		A	T	L			
Argentina	B0727_Argentina	EUSMPL0054-1-57	A	T	L	9	100	view alignment
Brazil	B5773_Brazil	EUSMPL0054-1-62	A	T	L	15	100	view alignment
Guatemala	730332_Guatemala	EUSMPL0054-1-65	G	C	L	7	100	view alignment
Guatemala	730333_Guatelama	EUSMPL0054-1-66	G	C	L	10	100	view alignment
Guatemala	730334_Guatemala	EUSMPL0054-1-67	A	T	L	13	100	view alignment
Guatemala	B0858_Guatemala	EUSMPL0054-1-58	G	C	L	21	100	view alignment

b) Determine genes that map to each of the SNPs identified in Step 1.

- *Add Step, Run a new Search for, Genes, Taxonomy, and choose C. posadasii str. Silveira*



- Next window will bring up a colocation tool where you will be able to set parameters of your gene search.
- Choose to Return each *Gene from Step 2* whose exact region *overlaps* the exact region of a SNP in Step 1 and is on *either strand*
- Click *Submit*
- Examine gene list returned

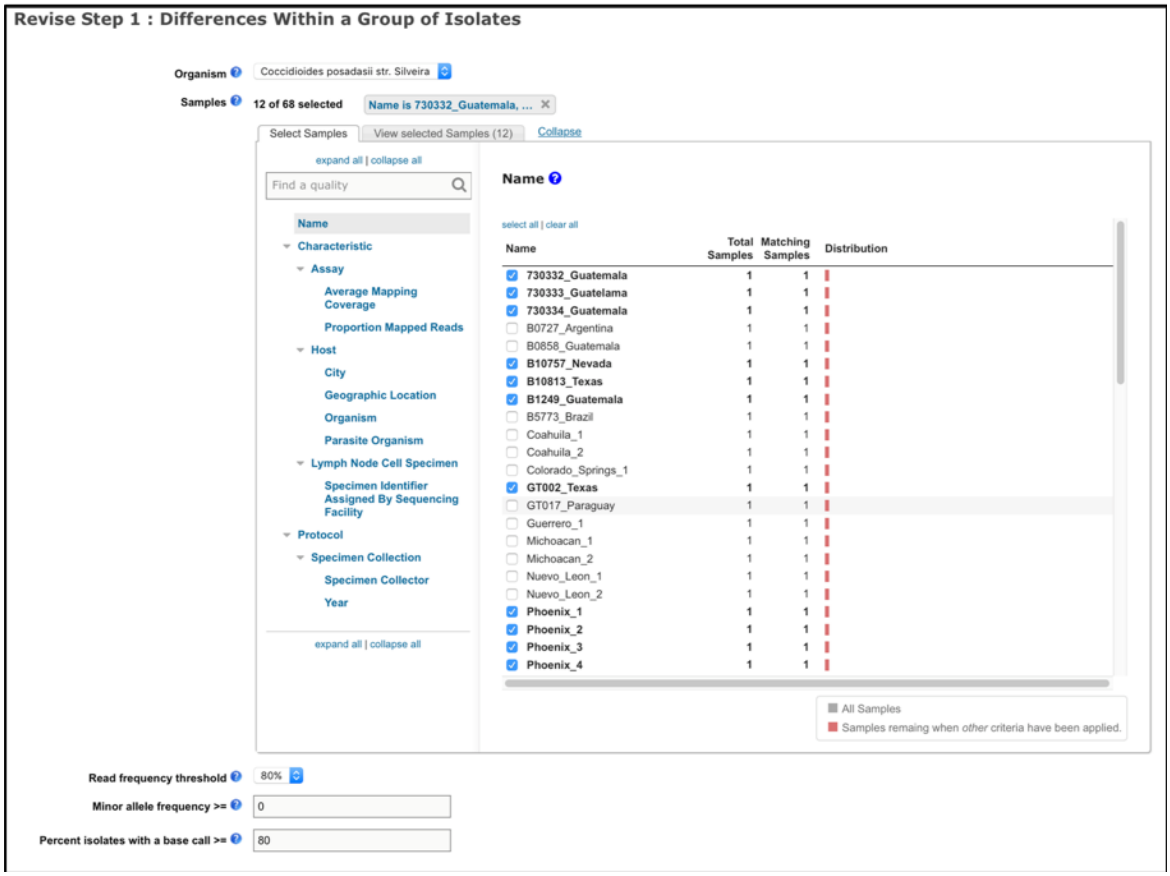


- Think about how can you analyze this data further?

Hint: you can extract genes that have *hypothetical* in the product description via the *Text* search. *You can also perform GO enrichment or identify orthologs in other species, or map to metabolic pathways etc., or you can take use other resources as shown previously to cross reference the integrated data.*

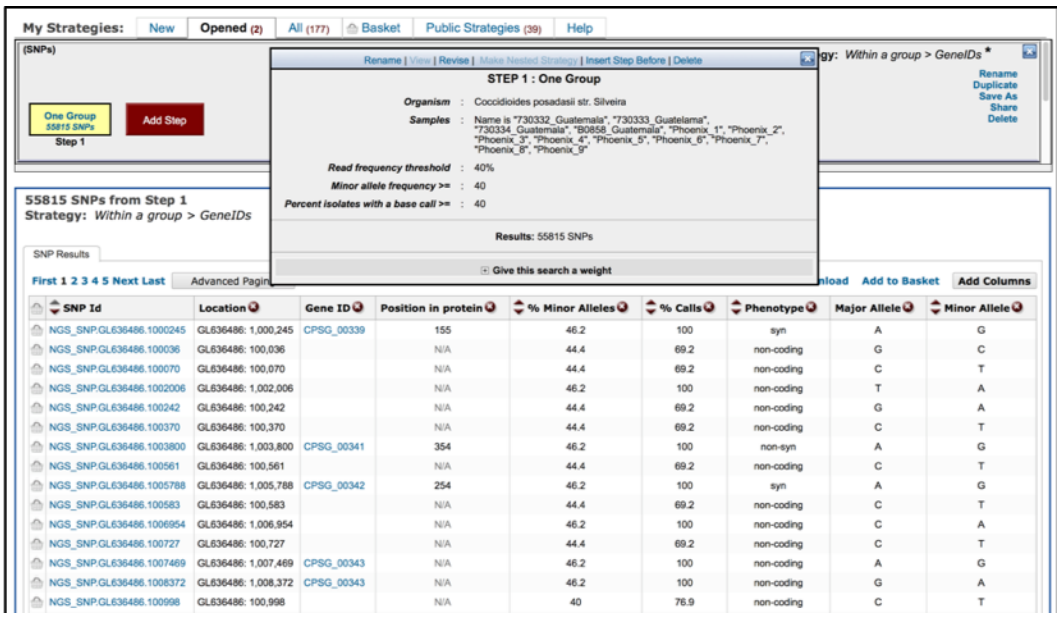
2. Identify SNPs within a group of isolates

- Navigate to the *Identify other Data Types* panel and select the *Differences Within a Group of Isolates* search from *SNPs* menu
- Let's look at the SNPs from Coccidioides species.
- Select *Guatemala*, *Texas*, *Phoenix*, and *Nevada* isolates.



- To set stringent control for quality and confidence of the SNP calls select 80% Read frequency threshold, leave minor allele frequency at default and percent isolates parameter at 80.
- How many SNPs were returned?
- How would you identify heterozygous SNPs?

Note: Create a new search or revise an existing search strategy. Modify *a read frequency threshold of 40% and revise this search and increase the minor allele frequency threshold (try 20 and 40 and compare results).*



Note: Read frequency threshold *applies to the sequencing reads of individual isolates and defines a stringency for data supporting a SNP call between an isolate and the reference genome (Organism). Each nucleotide position of each isolate is compared to the reference genome and a SNP call is made if the portion of the isolate's aligned reads that support the SNP is above the Read Frequency Threshold (RFT). Find high quality haploid SNPs with 80% RFT or heterozygous diploid/aneuploid SNPs with 40%.*

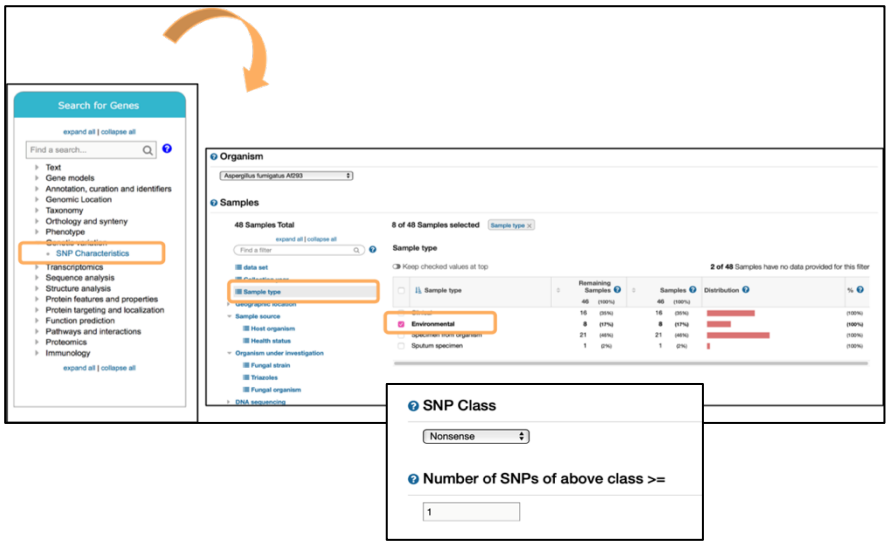
- How many SNPs did you identify?
- Why might you want to increase the minor allele threshold when you run SNP searches?

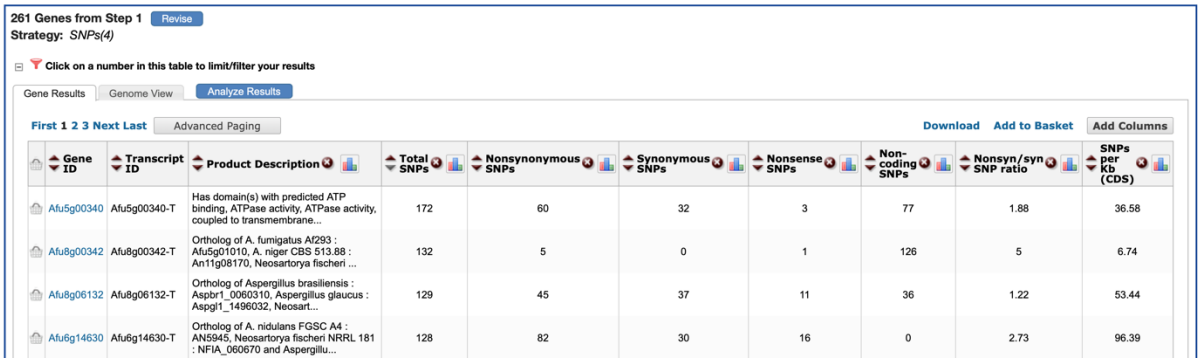
Note: Minor Allele Frequency parameter applies to your group of isolates. A SNP can occur in any number of isolates in your group and the least frequent SNP call across all isolates is the Minor Allele Frequency. A SNP will be returned by the search if the frequency of the minor allele is equal to or greater than your Minor Allele Frequency.

1. Identify genes with nonsense SNPs

- Navigate to the Genes by SNP characteristics search, which can be found under the Genetic Variation category in the gene searches section.
- Select *Aspergillus fumigatus* from the list of organisms and configure the search to identify SNPs in isolates originating from ‘environmental’ sample types.

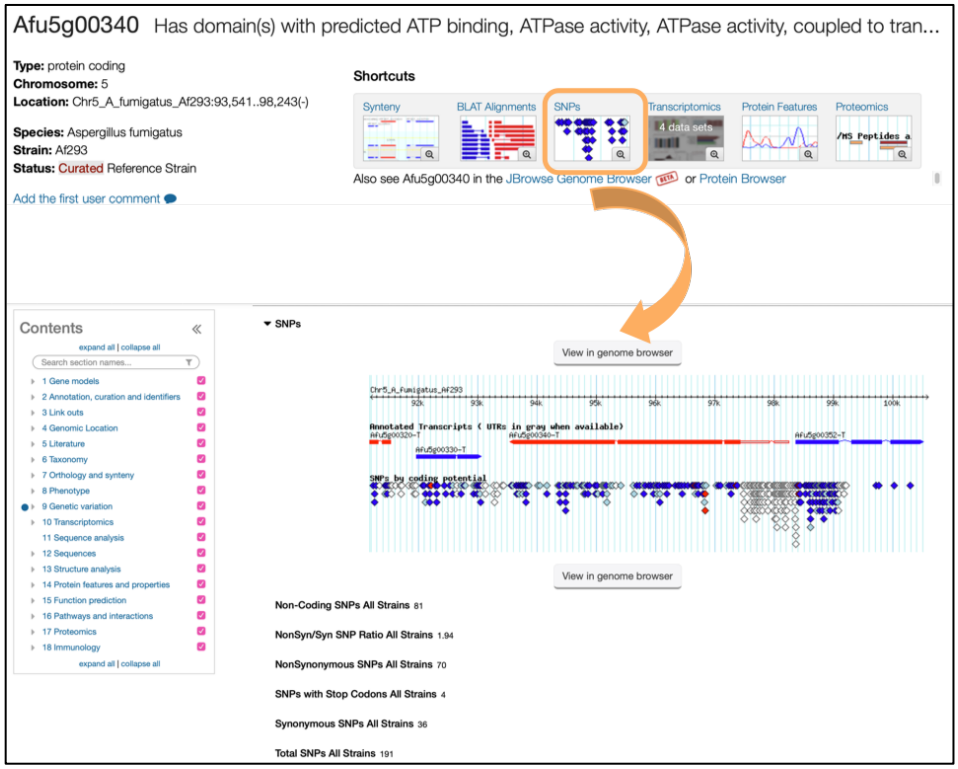
- Examine your results. How many genes were identified in your search?



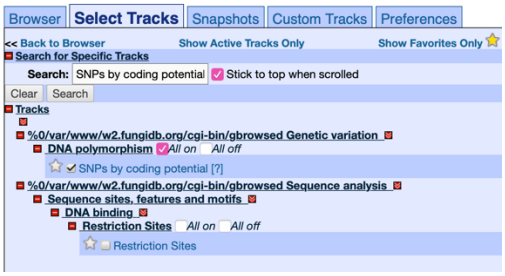


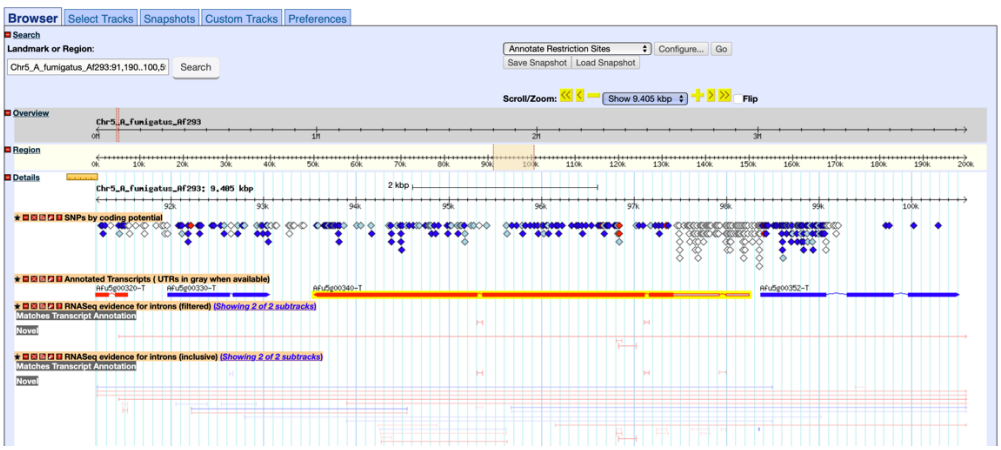
Gene ID	Transcript ID	Product Description	Total SNPs	Nonsynonymous SNPs	Synonymous SNPs	Nonsense SNPs	Non-coding SNPs	Nonsyn/syn SNP ratio	SNPs per Kb (CDS)
Afu5g00340	Afu5g00340-T	Has domain(s) with predicted ATP binding, ATPase activity, ATPase activity, coupled to transmembrane...	172	60	32	3	77	1.88	36.58
Afu8g00342	Afu8g00342-T	Ortholog of A. fumigatus Af293 : Afu5g01010, A. niger CBS 513.88 : An11g08170, Neosartorya fischeri ...	132	5	0	1	126	5	6.74
Afu8g06132	Afu8g06132-T	Ortholog of Aspergillus brasiliensis : Aspbr1_0060310, Aspergillus glaucus : Aspgl1_1496032, Neosart...	129	45	37	11	36	1.22	53.44
Afu6g14630	Afu6g14630-T	Ortholog of A. nidulans FGSC A4 : AN5945, Neosartorya fischeri NRRL 181 : NFIA_060670 and Aspergillu...	128	82	30	16	0	2.73	96.39

- How do these results change if you modify the stringency of the selection criteria?
- Afu5g00340 is one of the genes with several nonsense SNPs. Navigate to its gene record page in FungiDB and click on the **SNPs** data shortcut to be redirected to the Genetic variation section of the page.

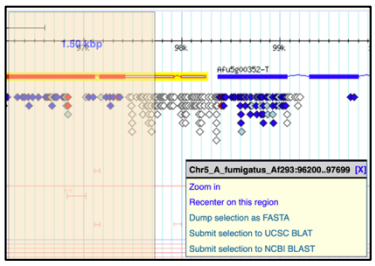


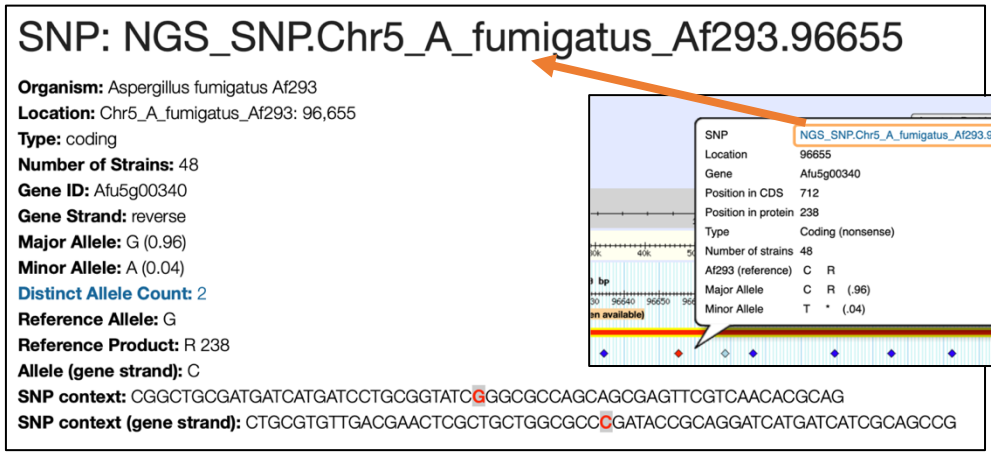
- Navigate to GBrowse by clicking on the *View in genome browser* button. Once in GBrowse, activate *SNPs by coding potential* track
(*Hint: navigate to the Select Tracks tab to select*)





- Zoom in to region with several nonsense SNPs.
- Identify isolates that contain nonsense mutations.
- Click on the link to be redirected to the SNP record page.
    - Examine other records on the isolate record page.
- How many alleles are reported for this locus?





- Can you identify the specific isolates that contained a nonsense mutation?

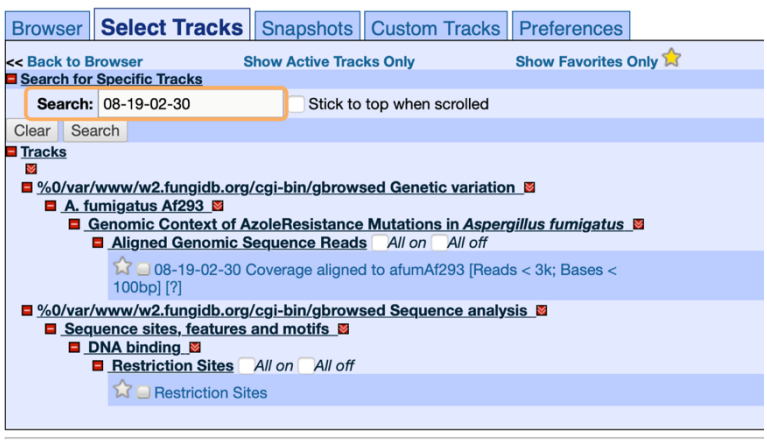
Hint: Look in Strains/Sample table.

Strains / Samples Download Data Sets

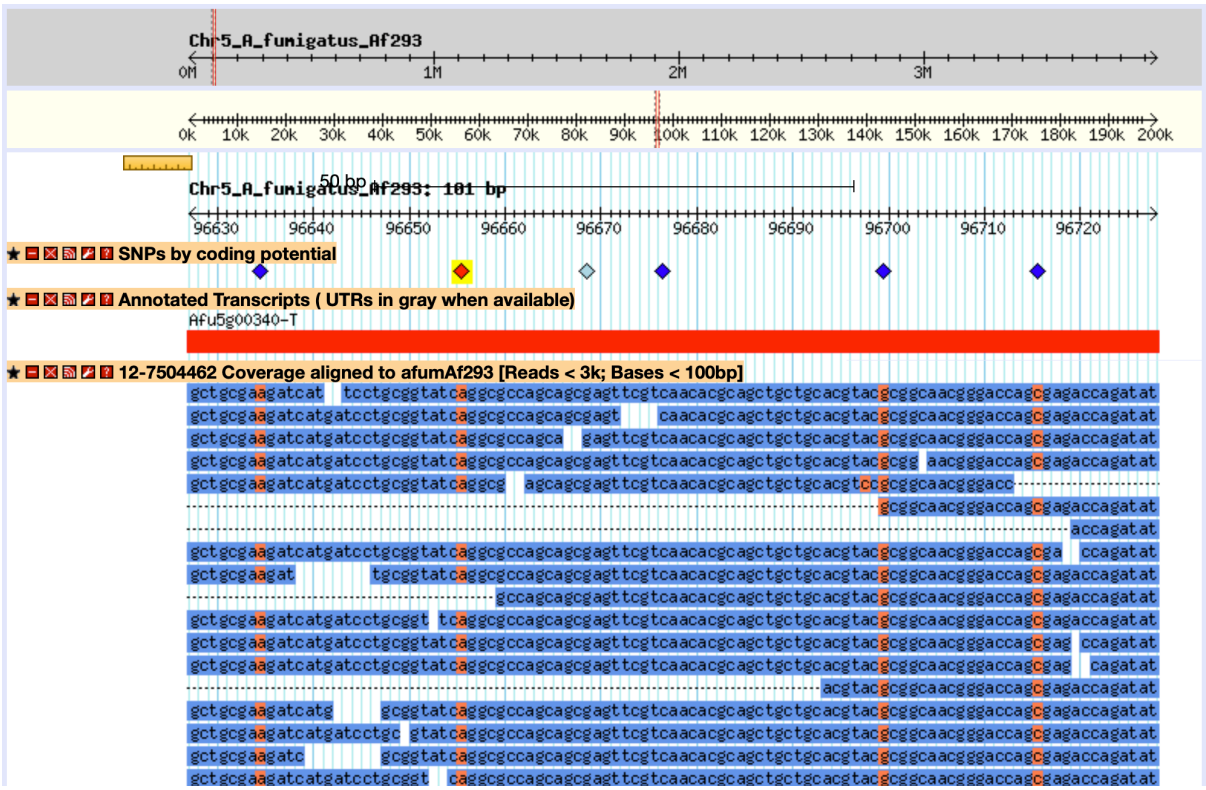
Search this table... Showing 48 rows

Geographic Location	Strain	Sample	Allele	Allele (gene strand)	Product	Coverage	Read Frequency	DNA-seq reads for strain
	Af293 (reference)		G	C	R			
India	Afu_1042-09	EUSMPL0067-1-16	G	C	R	60	100	view DNA-seq reads
India	Afu_124-E11	EUSMPL0067-1-19	G	C	R	66	100	view DNA-seq reads
India	Afu_166-E11	EUSMPL0067-1-20	G	C	R	54	100	view DNA-seq reads
India	Afu_218-E11	EUSMPL0067-1-22	G	C	R	55	100	view DNA-seq reads
India	Afu_257-E11	EUSMPL0067-1-21	G	C	R	49	100	view DNA-seq reads
India	Afu_343-P-11	EUSMPL0067-1-17	G	C	R	62	98.41	view DNA-seq reads
India	Afu_591-12	EUSMPL0067-1-18	G	C	R	44	100	view DNA-seq reads
India	Afu_942-09	EUSMPL0067-1-15	G	C	R	74	100	view DNA-seq reads
Kingdom of the Netherlands	08-12-12-13	EUSMPL0067-1-7	G	C	R	92	98.92	view DNA-seq reads
Kingdom of the Netherlands	08-19-02-10	EUSMPL0067-1-14	G	C	R	120	100	view DNA-seq reads
Kingdom of the Netherlands	08-19-02-30	EUSMPL0067-1-11	A	T	*	45	100	view DNA-seq reads

- Navigate to GBrowse by clicking on View in genome browser button and activate coverage tracks for *08-19-02-30 and 12-7504462 isolates* from the *Aligned genome sequence reads* menu.



- Zoom in to 100bp to visualize reads.
    - Does sequence analysis support the reported SNP data?



Note: A specific SNP record can be also selected directly from the gene record page by hovering over the SNP of interest to bring up a pop-up window with a direct link to this SNP record page: