

Add Annotation Tracks from External Databases to NextGENe

April 2014

John McGuigan, Kevin LeVan, Megan Manion, Jonathan Liu, Shouyong Ni

Introduction

The Track Manager tool in NextGENe allows for import of variant annotation from external databases that can be automatically queried for use in both the mutation report and the variant comparison report (figure 1). Tracks such as dbSNP, COSMIC, 1000 genomes, tracks with causative prediction from dbNSFP, and NHLBI GO Exome Sequencing Project (ESP) can be added to your projects, as well as custom databases. Once a track is imported for a given preloaded reference, it can be automatically queried for every new project run against that reference or it can be queried manually in the NextGENe Viewer.

SNP db_xref	COSMIC	Mutation Call	Amino Acid Change	INFO_AF	SIFT_	PolyPhen2	LRT_	Mutati	Mutati	PhyloP	INFO_MAF
rs9428083		IVS420+6A>G		0.6600							12.1977
rs10925391	COSM146836	IVS464-8A>C		0.3900							22.9323
rs10754602		IVS677-11T>AT		0.5900							44.0270
rs3765097		c.1359C>CT	453S>SS	0.5500							39.9490
rs2045955		IVS1612+14T>CT		0.5400							45.5936
rs2253273		c.2973A>G	991S>S	0.8300							4.1546
		c.6281G>AG	2094G>DG		D	D	D	D	M	2.5710	
rs707189		c.6906T>C	2302L>L	0.9600							0.0609
rs72549416		c.7365C>CT	2455D>DD	0.0018							0.6663
rs34967813		c.8873A>AG	2958Q>QR	0.1400	T	B	D	P	L	2.1640	30.4163
rs2797436		c.9318T>G	3106S>S	0.9700							0.0486

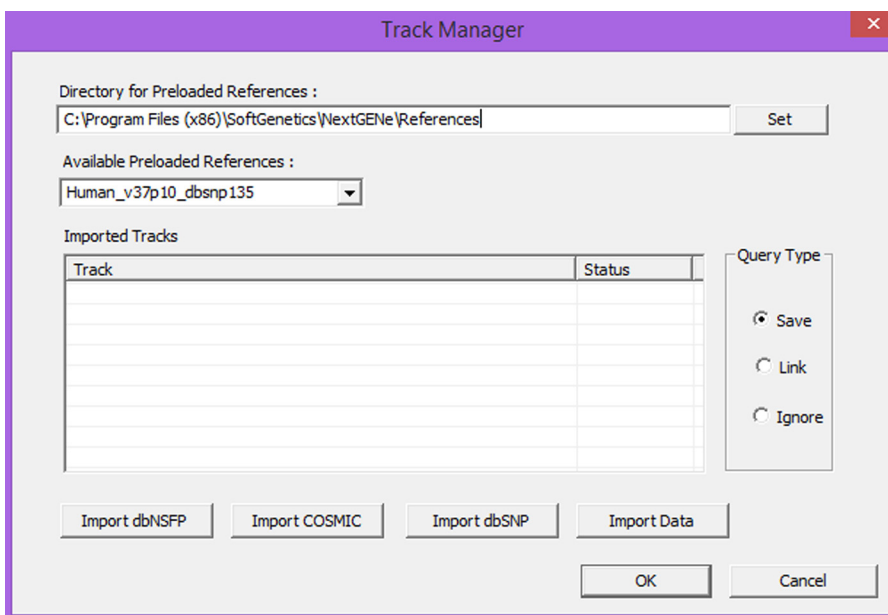
Figure 1: The NextGENe Mutation Report with annotation from multiple tracks

Procedure

Importing a Track

The Track Manager can be opened from the NextGENe tools menu (Figure 2). When importing a track, NextGENe should be run with administrator permissions if the reference directory is protected (including the default reference installation directory). Any previously imported tracks are shown for the selected reference.

Figure 2: The Track Manager Tool



There are three databases with specialized support. More information on some of these can be found in the ‘References’ dialog of the NextGENe ‘Help’ menu.

- **dbNSFP (Database of Non-Synonymous Functional Prediction)**

- The compressed database can be downloaded from the dbNSFP website which is linked in the dbNSFP import dialog. This file can then be loaded into the tool where it will be automatically extracted and imported (requires approximately 80 GB of free space).
- Functional Prediction, Conservation, and (when available) population frequency information is available for all non-synonymous SNPs. Included in the latest supported version (2.1) :

Functional Prediction	Conservation	Population Frequency
PolyPhen2 HDIV	GERP++ NR	1000 Genomes
PolyPhen2 HVAR	GERP++ RS	NHLBI Go Exome Seq Project
LRT	PhyloP	
Mutation Taster	SiPhy 29-way log odds	
Mutation Assessor	LRT	
SIFT		
FATHMM		

- **COSMIC (Catalog of Somatic Mutations in Cancer)**

- The COSMIC database can be downloaded in VCF format and imported- Coding Variants, Non-Coding Variants, or both. The reports will link the COSMIC ID to the COSMIC website for further annotation.
- The compressed vcf files (vcf.gz) listed on the COSMIC website are not actually compressed VCF files, and will have to be extracted manually and renamed before being imported.

- **dbSNP**

- A version of dbSNP is already built into NextGENe’s preloaded whole-genome references, but newer builds can be imported in the Track Manager tool.
- Several different VCF files are available for download from NCBI, each with a subset of the entire database related to the clinical status of the variants.
- The custom track import dialog can be used to import additional information from the dbSNP VCF files.

There is also a general “Import Track” tool which can be used to import data from a variety of formats with custom selection of annotation for display and/or filtering (figure 3).

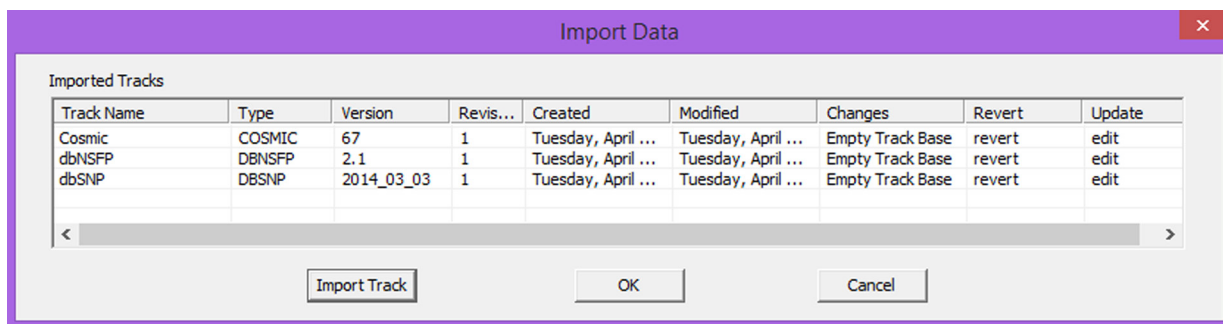


Figure 3: Custom Track Import Dialog

When importing a custom track, the fields are automatically categorized. Some fields (Chr, ChrPos, WT_Seq, etc) are needed in order to match the variants. Others can be set to “Display Only”, “Display and Filtering”, or “Skip”. When importing data for filtering, the data type can be set to “String” (text-based with filters for matching the text), “Integer”, or “Decimal”. The last two include filters comparing the values (>, =, etc). Categories of custom track files include:

- COSMIC files
- COSMIC VCF files
- dbNSFP
- ESP VCF files
- dbSNP (Including MAF and clinical information as seen in figure 4)
- VCF (any valid VCF file)
- ESP (NHLBI GO Exome Sequencing Project)

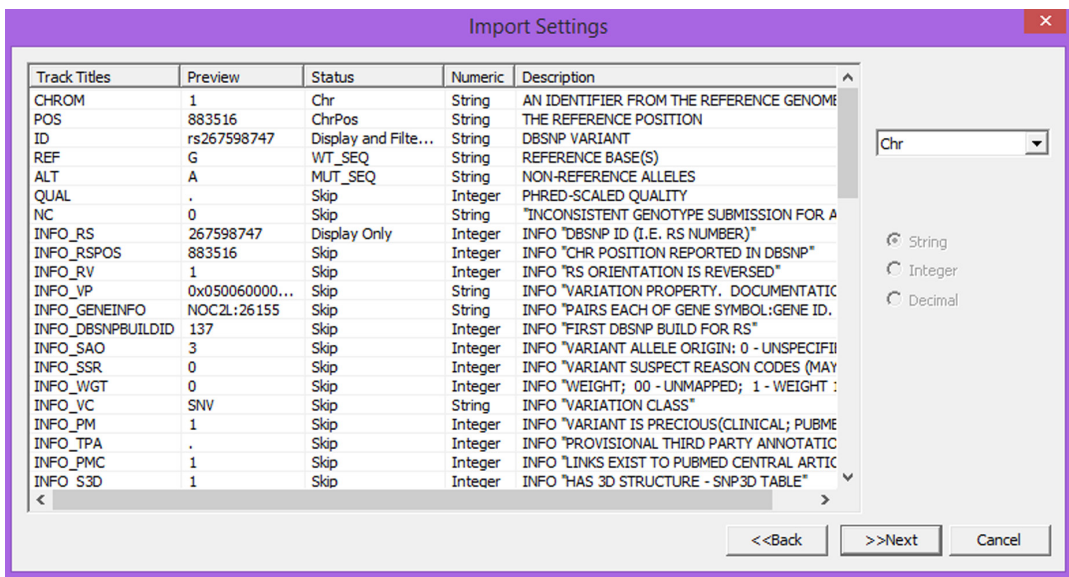


Figure 4: Custom Import of a dbSNP VCF file

Querying Tracks in the Viewer

The “Query Reference Tracks” dialog can be used to import or update track information in an alignment project. If all of the projects in the Variant Comparison Tool have the same tracks queried, the information is also available in that tool.

The available tracks for the current reference are listed (see figure 5), along with any other tracks already saved in the project. Each track is “Saved”, “Linked” or “Not Used”. Tracks that are linked will be available to NextGENe Viewer as long as the project can access the tracks in the reference folder. Saved tracks will continue to work even if the original track is edited or the project is moved to another computer.

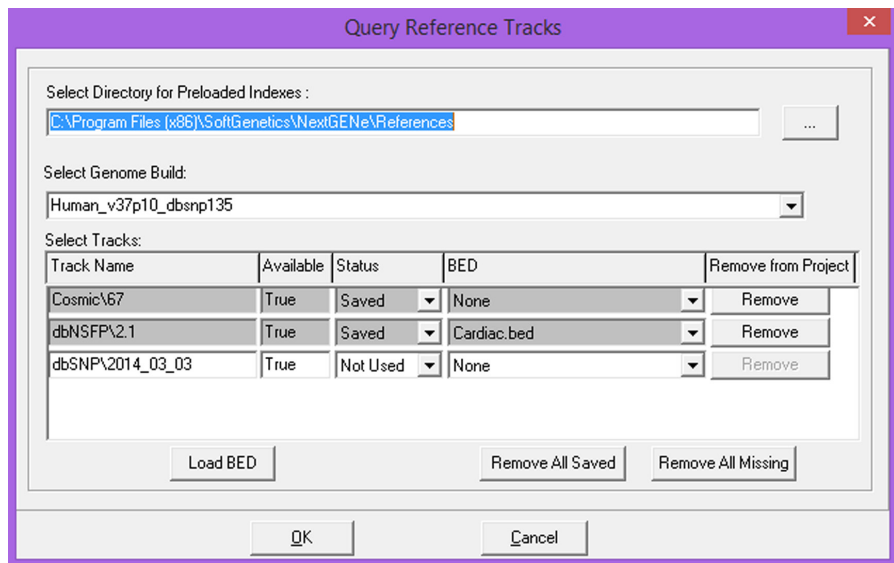


Figure 5: The “Query Reference Tracks” dialog. Here there are three reference tracks available with two of them (COSMIC and dbNSFP) already saved in the project.

Filtering using Tracks

dbSNP and COSMIC tracks have basic display and filtering functions related to the existence of variants in the databases. dbNSFP has one display page and several filter pages- one for each category of annotation, such as the functional prediction scores shown in figure 6. Each individual score can be filtered based on the numerical value and/or classification, and a minimum number of passing scores can be set.

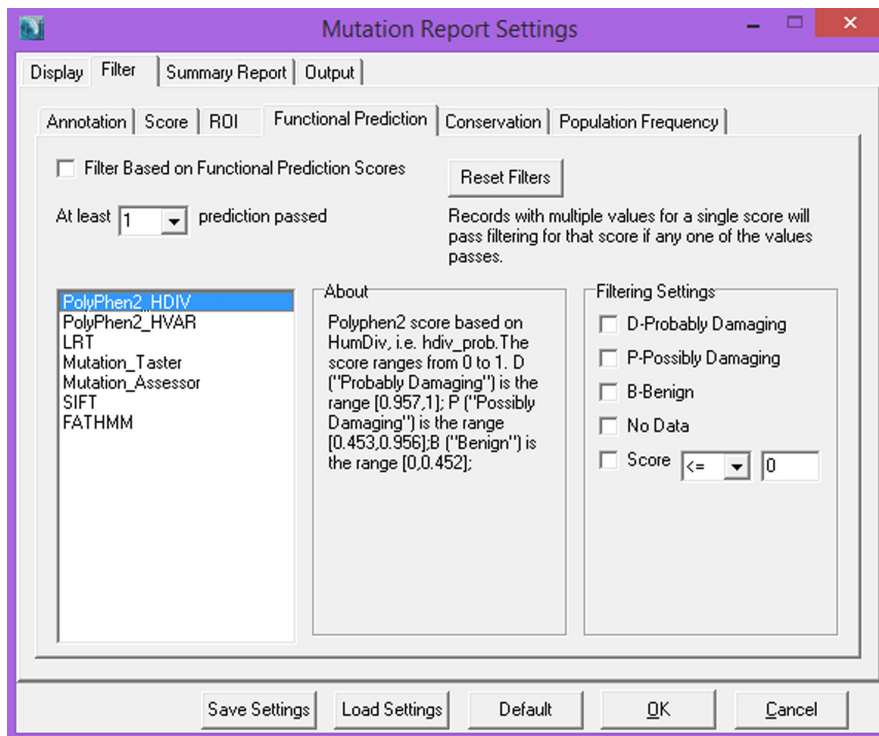


Figure 6: Filtering using Functional Prediction scores from the dbNSFP database

Custom tracks will have a single display page and a single filter page. The available data will depend on which fields were selected during import. Figure 7 shows the filtering page for a 1000 genomes custom track. The “ID” category is a string-based filter, allowing for inclusion or exclusion of exact matches for a specific variant ID. The other categories were set as numeric filters, allowing value comparison (>, ≥, =, ≤, <, ≠). For example, common variants with a population frequency above 5% (0.05) could be excluded.

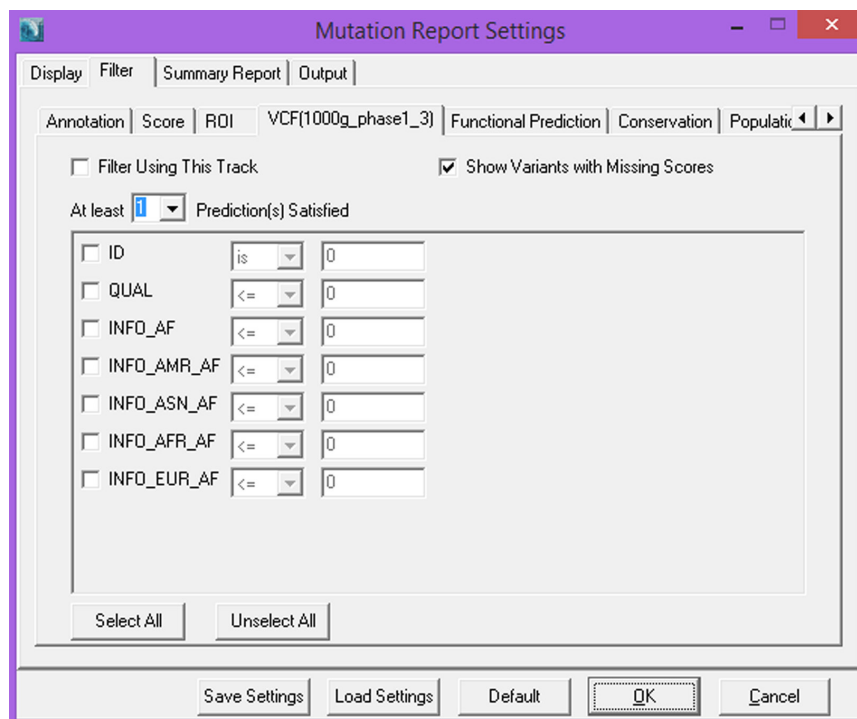


Figure 7: The filter page for a 1000 genomes custom track.