Population Variation

Accounting for Population Heterogeneity in Assay Design

Developed by Spencer Smith, Grant Fujimoto, Matt Monroe, Larissa Rodriguez, and Samuel Payne

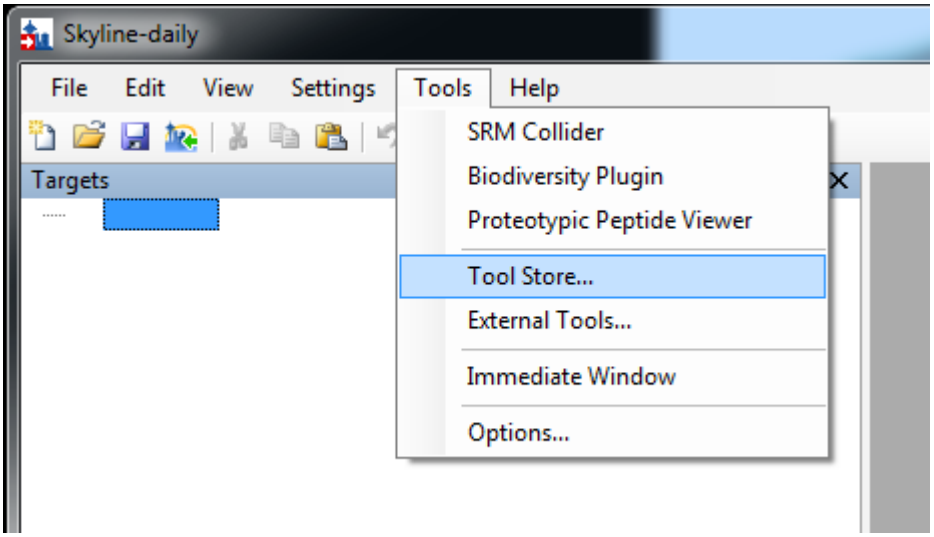
This tutorial is to help users install and use the Population Variation plug-in within Skyline. Frequently Asked Questions (FAQs) follow at the end of the tutorial.

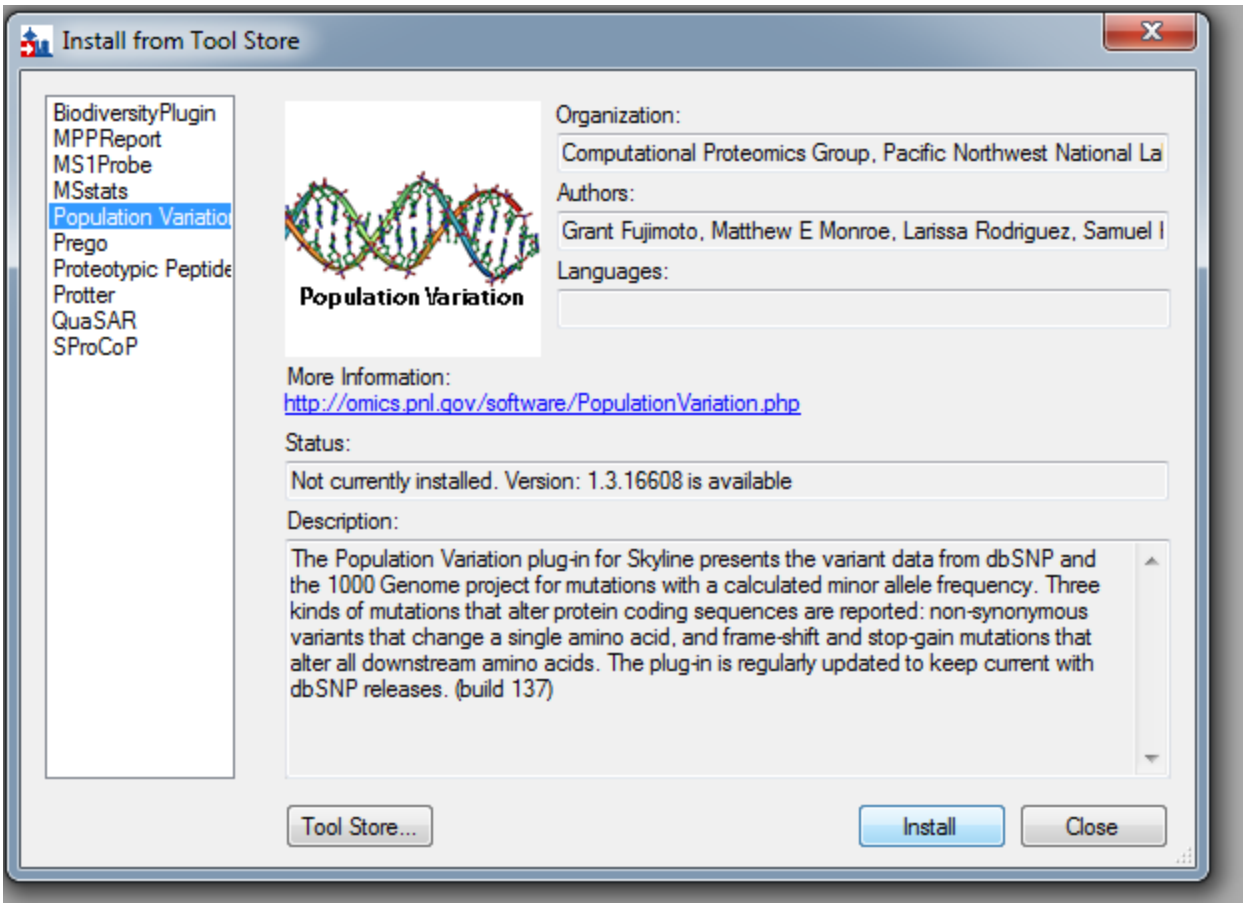
UPDATE: this is the manual for version 2. The major update from version 1 was the inclusion in the output of minor allele frequency for specific subpopulations and ethnicities.

Installation

The plug-in is available as a zip file from two web locations: the skyline website (https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/tools/begin.view) or the PNNL website (http://omics.pnl.gov/software/PopulationVariation.php). After downloading the zip file to your computer, it can be installed within Skyline as follows.

1. Open the Tools menu to import from the Skyline Store

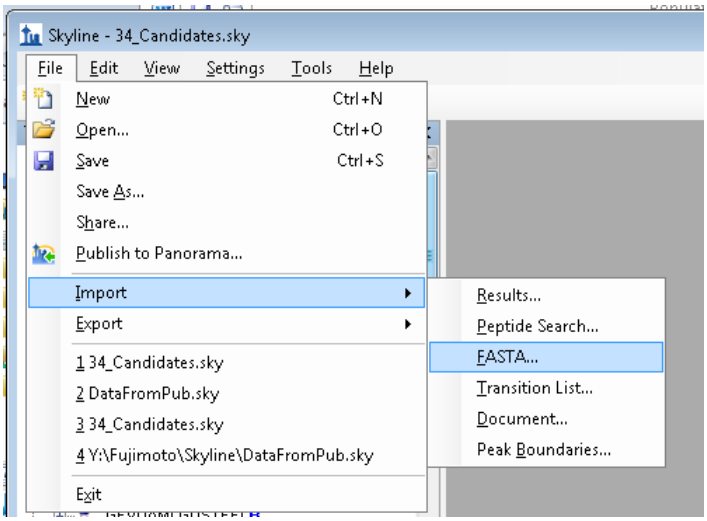


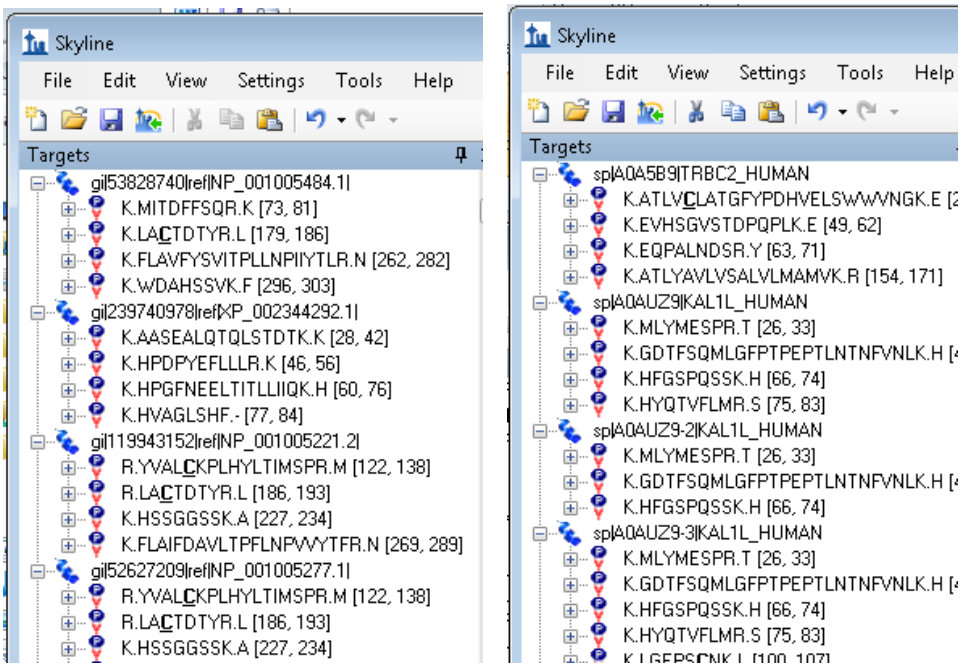


The file should install with no additional effort.

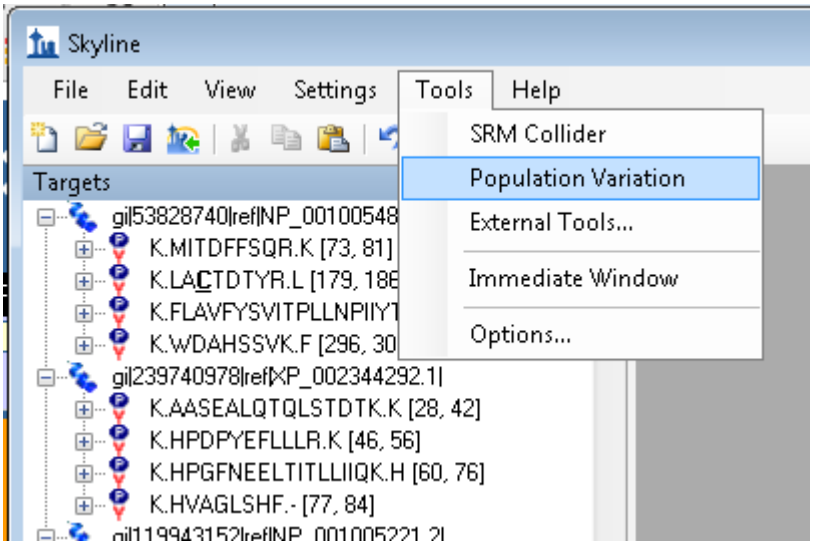
Use of the Population Variation Tool

To start using the tool in Assay design, you must have peptides and proteins as part of a skyline file. It is important to note that proteins must be properly named with accessions. The easiest way to get this is to start with a protein sequence fasta file that is from either NCBI's RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/protein.fa.gz) or Uniprot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/HUMAN.fasta.gz). This is done within Skyline by using the File->Import->Fasta menu.

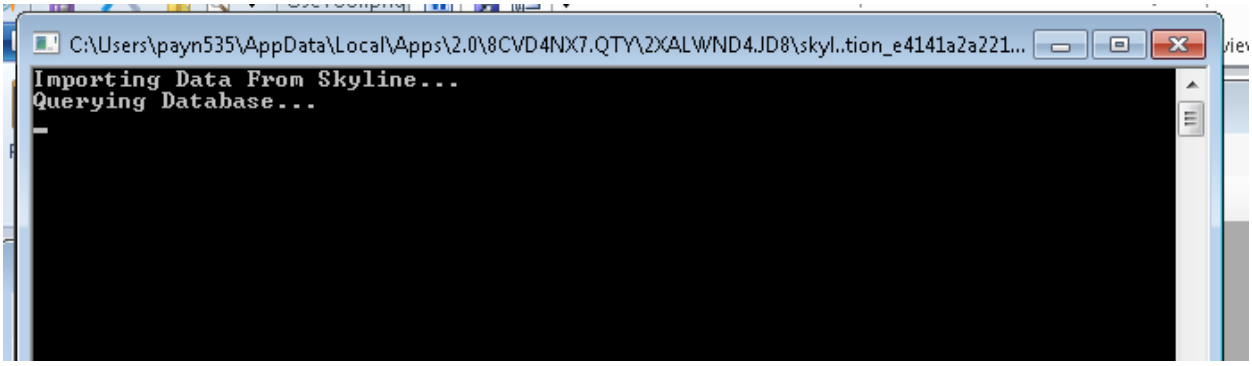




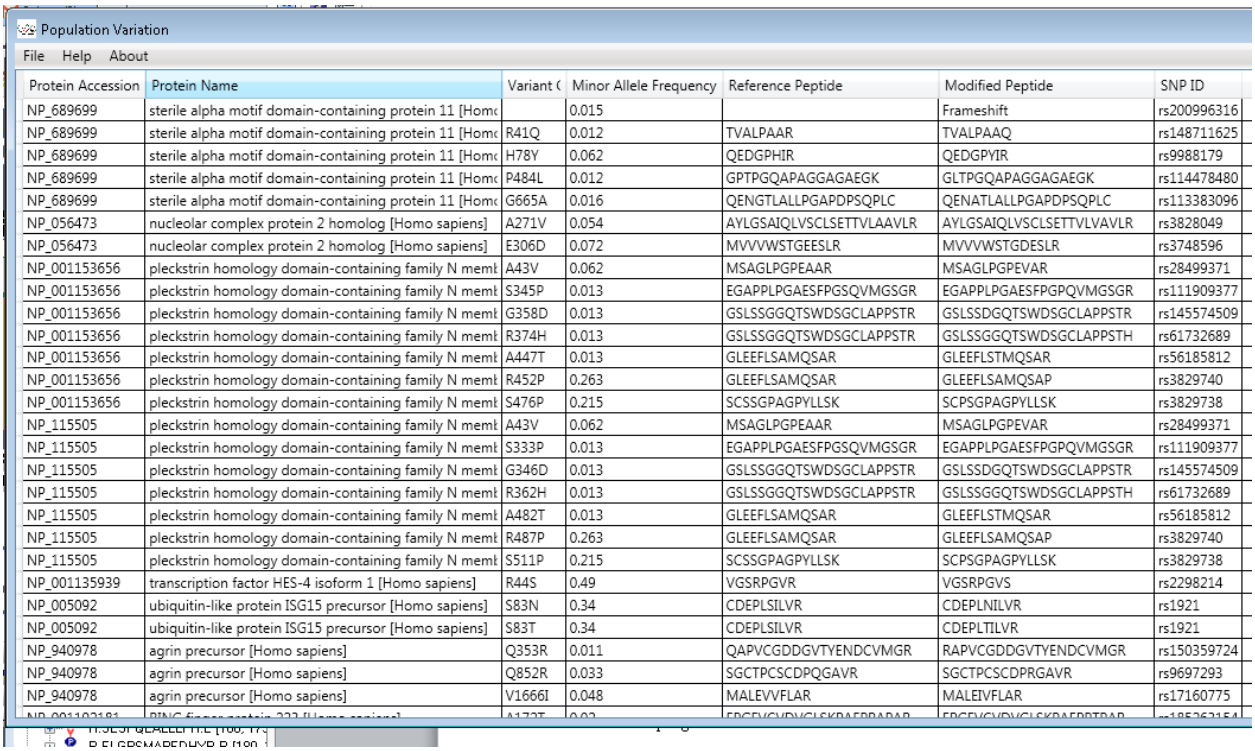
With properly named proteins, you are ready to use the Population Variation tool. Simply choose the tool from the Tools->PopulationVariation



While data is sent from Skyline to the plug-in, a command prompt window will pop-up to show the progress.



The final output is a table showing proteins, and the known variants. This file can be saved (File->Save) for later use in programs like Excell.



Protein Accession	Protein Name	Variant (	Minor Allele Frequency	Reference Peptide	Modified Peptide	SNP ID
NP_689699	sterile alpha motif domain-containing protein 11 [Hom(		0.015		Frameshift	rs200996316
NP_689699	sterile alpha motif domain-containing protein 11 [Hom(	R41Q	0.012	TVALPAAR	TVALPAAQ	rs148711625
NP_689699	sterile alpha motif domain-containing protein 11 [Hom(	H78Y	0.062	QEDGPHIR	QEDGPYIR	rs9988179
NP_689699	sterile alpha motif domain-containing protein 11 [Hom(	P484L	0.012	GPTPGQAPAGGAGAEGK	GLTPGQAPAGGAGAEGK	rs114478480
NP_689699	sterile alpha motif domain-containing protein 11 [Hom(	G665A	0.016	QENGTLALLPGAPDPSQPLC	QENATLALLPGAPDPSQPLC	rs113383096
NP_056473	nucleolar complex protein 2 homolog [Homo sapiens]	A271V	0.054	AYLGSAIQLVSCLSETTVLAAVLR	AYLGSAIQLVSCLSETTVLVAVLR	rs3828049
NP_056473	nucleolar complex protein 2 homolog [Homo sapiens]	E306D	0.072	MVVVWSTGEESLR	MVVVWSTGDESLR	rs3748596
NP_001153656	pleckstrin homology domain-containing family N memt	A43V	0.062	MSAGLPGPEAAR	MSAGLPGPEVAR	rs28499371
NP_001153656	pleckstrin homology domain-containing family N memt	S345P	0.013	EGAPPLPGAESFPGSQVMGSGR	EGAPPLPGAESFPGPQVMGSGR	rs111909377
NP_001153656	pleckstrin homology domain-containing family N memt	G358D	0.013	GSLSSGGQTSWDSGCLAPPSTR	GSLSSDGQTSWDSGCLAPPSTR	rs145574509
NP_001153656	pleckstrin homology domain-containing family N memt	R374H	0.013	GSLSSGGQTSWDSGCLAPPSTR	GSLSSGGQTSWDSGCLAPPSTH	rs61732689
NP_001153656	pleckstrin homology domain-containing family N memt	A447T	0.013	GLEEFLSAMQSAR	GLEEFLSTMQSAR	rs56185812
NP_001153656	pleckstrin homology domain-containing family N memt	R452P	0.263	GLEEFLSAMQSAR	GLEEFLSAMQSAP	rs3829740
NP_001153656	pleckstrin homology domain-containing family N memt	S476P	0.215	SCSSGPAGPYLLSK	SCPSGPAGPYLLSK	rs3829738
NP_115505	pleckstrin homology domain-containing family N memt	A43V	0.062	MSAGLPGPEAAR	MSAGLPGPEVAR	rs28499371
NP_115505	pleckstrin homology domain-containing family N memt	S333P	0.013	EGAPPLPGAESFPGSQVMGSGR	EGAPPLPGAESFPGPQVMGSGR	rs111909377
NP_115505	pleckstrin homology domain-containing family N memt	G346D	0.013	GSLSSGGQTSWDSGCLAPPSTR	GSLSSDGQTSWDSGCLAPPSTR	rs145574509
NP_115505	pleckstrin homology domain-containing family N memt	R362H	0.013	GSLSSGGQTSWDSGCLAPPSTR	GSLSSGGQTSWDSGCLAPPSTH	rs61732689
NP_115505	pleckstrin homology domain-containing family N memt	A482T	0.013	GLEEFLSAMQSAR	GLEEFLSTMQSAR	rs56185812
NP_115505	pleckstrin homology domain-containing family N memt	R487P	0.263	GLEEFLSAMQSAR	GLEEFLSAMQSAP	rs3829740
NP_115505	pleckstrin homology domain-containing family N memt	S511P	0.215	SCSSGPAGPYLLSK	SCPSGPAGPYLLSK	rs3829738
NP_001135939	transcription factor HES-4 isoform 1 [Homo sapiens]	R44S	0.49	VGSRPGVR	VGSRPGVS	rs2298214
NP_005092	ubiquitin-like protein ISG15 precursor [Homo sapiens]	S83N	0.34	CDEPLSILVR	CDEPLNILVR	rs1921
NP_005092	ubiquitin-like protein ISG15 precursor [Homo sapiens]	S83T	0.34	CDEPLSILVR	CDEPLTILVR	rs1921
NP_940978	agrin precursor [Homo sapiens]	Q353R	0.011	QAPVCGDDGVTYENDCVMGR	RAPVCGDDGVTYENDCVMGR	rs150359724
NP_940978	agrin precursor [Homo sapiens]	Q852R	0.033	SGCTPCSCDPQGAVR	SGCTPCSCDPRGAVR	rs9697293
NP_940978	agrin precursor [Homo sapiens]	V1666I	0.048	MALEVVFLAR	MALEIVFLAR	rs17160775

We note that the plug-in uses sequence variant data as collected by dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), which indexes proteins using the NCBI RefSeq accession set. Thus if using a SwissProt/Uniprot accession set, the program converts these using the codex supplied by Uniprot.

In the updated v2.0 of the plugin, we now have minor allele frequency for each of the 1000Genome sub-populations, i.e. East Asian, European, African, American, and South Asian.

East Asian	European	African	American	South Asian	SNP ID
2.18%	19.58%	60.74%	11.53%	15.03%	rs6647
0%	5.67%	0.08%	5.76%	0%	rs17580
31.65%	24.85%	9.23%	35.73%	47.55%	rs1303
32.64%	6.06%	6.28%	23.78%	15.44%	rs2231137
29.07%	9.44%	1.29%	14.12%	9.71%	rs2231142
21.73%	19.98%	12.1%	15.27%	11.96%	rs3731608
21.73%	19.98%	12.1%	16.43%	11.96%	rs3731607

Frequently Asked Questions

1. Where can I get accessions for my fasta files?

Accepted accessions are NCBI RefSeq or UniProt.
ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/protein.fa.gz,
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/HUMAN.fasta.gz

2. What is the source data used for SNP variants?

Data is downloaded from dbSNP. The primary data source is the 1000 Genome project and the HapMap Consortium

3. What types of mutations are reported?

Three mutations are reported: non-synonymous missense mutations that change a single amino acid, stop-gain nonsense mutations that terminate a protein translation, and frameshift indel mutations which alter all subsequent residues.

4. I get an error message "Unable to process accession" or "Accession not found in fasta". What does that mean?

The Population Variation tool searches for SNPs in dbSNP via the protein accession. Therefore, if the protein name listed in Skyline is not properly formatted, the tool will not be able to index the database. We strongly encourage users to create proteins within Skyline using properly formatted fasta files (either under File->Import->FASTA, or Edit->Insert->FASTA).

If using these options and the NCBI or Uniprot derived fasta files, the accessions will be in the proper format (See FAQ #1). An alternative is to simply list the accession at the beginning of the protein name in Skyline.

5. Why do I get several Accessions for the same SNP?

The Population Variation tool searches for SNPs in dbSNP via the protein accession. dbSNP natively uses the RefSeq accession set from NCBI. If the protein name given from Skyline uses Uniprot accessions, the program must convert those to NCBI. Conversions are taken from Uniprot's website (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/). Note that Uniprot and NCBI do not always have a strict one-to-one mapping, especially in the case of alternative isoforms at a single genetic locus. See http://www.uniprot.org/uniprot/P47710 for an example.