SNVerGUI Manual

A Desktop Tool for Variant Analysis of Next Generation Sequencing Data

8/19/2012

Content

1. Introduction

SNVer is among the very few existing tools that are capable of detecting variants (both single nucleotide variation and small indel) from both pooled and individual next generation sequencing (NGS) data (1). It is the only existing tool that can handle single-pool NGS data, to our knowledge. Most existing NGS variant calling tools (2-5) including SNVer, however, are based on a command-line interface. Users need to execute non-interactive commands for running these programs followed by additional customized parsing or filtering steps. These sophisticated pipelines executed by command lines discourage less-programming-trained users, such as biologists, geneticists and clinicians, who, however, are the main end users of NGS data. It motivates us to implement SNVerGUI, a graphical user interface (GUI) for running SNVer. With SNVerGUI, the users can run the entire variant calling pipeline after simply configuring several parameters via the user-friendly GUI. Moreover, SNVerGUI displays the results as a typical table view supporting various interactive filtering and sorting. It also generates an automatic summary report for further analysis.

2. Downloads and Requirements

2.1 JAVA

SNVerGUI is based on Rich Client Platform (RCP: http://www.eclipse.org/community/rcp.php and www.eclipse.org/rcp/), which **requires strict operating system (OS) and Java Runtime Environment (JRE)**. For example, **if OS is Windows 64-bit, then we need to download Windows 64-bit JRE**. If OS is Linux x64, then we need to download JRE Linux x64. For different versions of JRE, they can be downloaded from: http://java.com/en/download/manual.jsp.

2.2 SNVerGUI

SNVerGUI can be downloaded from http://snver.sourceforge.net/snvergui. Depending on OS and the version of JAVA installed, users can use the suitable version of SNVerGUI which matches to the installed JAVA.

It should be noted that our SNVerGUI was compiled based on JavaSE-1.6 (JRE6). To make sure suitable version of Java that has been installed, user can check this by do the following operations:

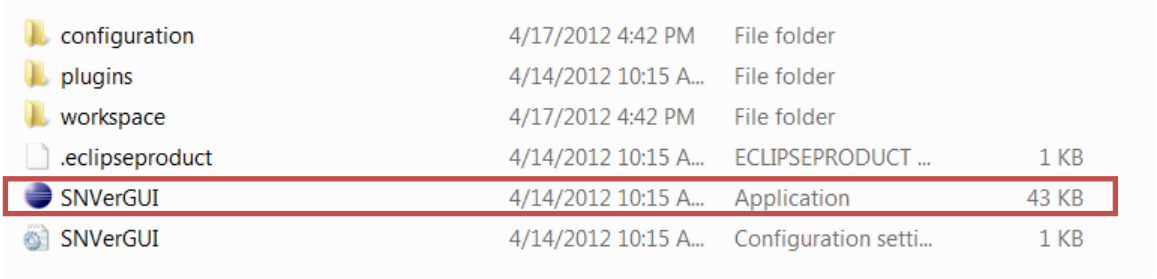
Take Windows for an example, Start-> Search program "Command" -> In the "Command Prompt"-> type "java -version". The java version in the computer would be,

```
#######################################################
C:\Users\Wei>java -version
java version "1.6.0_30"
Java(TM) SE Runtime Environment (build 1.6.0_30-b12)
Java HotSpot(TM) 64-Bit Server VM (build 20.5-b03, mixed mode)
#######################################################
```

3. Installation and Start

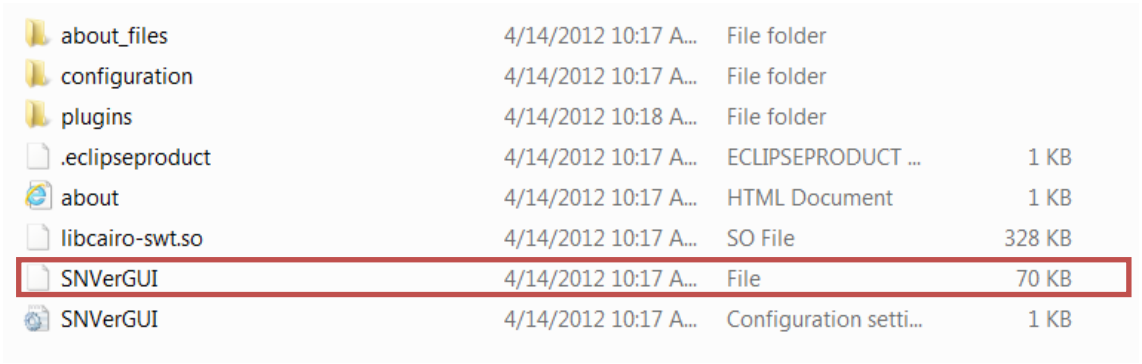
To install SNVerGUI from package file, put the downloaded '.zip' file into the desired directory on your computer. Unzip the file (usually by double-clicking it). To start SNVerGUI,

Windows OS users can simply double-click "**SNVerGUI.exe**" file under SNVerGUI folder. This will open SNVer GUI.



Mac OS users can double-click "SNVerGUI.app" file in the SNVerGUI folder to open SNVerGUI.

Linux OS users can simply double-click "**SNVerGUI**" file under SNVerGUI folder, which will open SNVerGUI.



4. Supported Input Data

4.1 Required Input Data

4.1.1 Bam/SAM files (Aligned data)

The aligned sequencing data can be output in either BAM or SAM format. Please refer to SAMTools manual for details: http://samtools.sourceforge.net/samtools.shtml. It can be produced by popular alignment tools, such as BWA(6) and BOWTIE(7) for Illumina platform, BFAST(8) and SHRIMP2(9) for SOLiD platform, and BWA-SW(10) for 454 platform.

4.1.2 Reference Genome File

The files can be downloaded from UCSC: http://hgdownload.cse.ucsc.edu/downloads.html#human.

If the sequenced data is from a specific region on a specific chromosome, say X, then we just need to download **chrX.fa.gz** from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes. The file can be uncompressed by double-clicking it.

This directory (chromosomes) contains the Feb. 2009 assembly of the human genome (hg19, GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1)) in one gzip-compressed FASTA file per chromosome.
If the sequenced data is from whole genome, then we need to download **hg19.2bit** from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/. The hg19.2bit contains the complete hg19 Human Genome in the 2bit format. To transfer the .2bit compressed file back to ordinary fasta.fa file, users can use a utility program **twoBitToFa** to extract .fa file(s) from this file, which can be downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/ for Linux OS and http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.i386/ for Mac OS.

Alternatively, users can download the whole genome .fa file from here, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz

It should be noted that the reference file should be matched the aligned data file, that is, the format of reference file should be the same as the one used in aligning raw sequencing data! Otherwise, "ReferenceNotMatchException" would be thrown from PICARD API, from which SNVer calls for piling up. For example, users can check the SAM/BAM header through SAMTools:

"samtools view –H test1.bam", which will give

@HD VN:1.0 GO:none SO:coordinate

@SQ SN:chrX LN:155270560

If there is no header information in aligned data or the reference length and name are not correct, it would prevent the program from running. If that is the case, we suggest the user redo the mapping to make sure that the reference file is consistent with the header information in aligned data.

4.1.3 Pool info configuration file/ Configuration for pool data

This file is only used in pooling design option when the number of samples is different in different pools. This is a tab-delim file containing five columns, which are file names, the number of haploids in the pool, the number of samples, base quality score and mapping quality score, respectively. For example, if assuming diploid individual, the number of haploids should be 2 * no. of samples in a pool. Here each bam file is a pool.

The following is an example of the pool info file (**test.ini**), the line starting with # will be omitted automatically.

#names	no.haploids	[no.samples]	[bq]	[mq]
test1.bam	2	1	17	30
test2.bam	2	1	20	20

4.2 Optional Input Data

4.2.1 Target regions (bed format) file

With the help of sequence capture technology, people are able to study part of selected genomic regions from entire complex human genomic DNA at a relatively low cost. There are three major platforms: Agilent, Illumina and Nimblegen, providing kits for the whole exome or customized regions. The target regions file is usually in .bed format and can be found from their website. Take whole-exome sequencing for an example, the target regions can be download from http://www.nimblegen.com/products/seqcap/ez/v2/.

Here is a short example of the .bed file for a target region on X chromosome:

chrX	140991641	140991764
chrX	140992502	140992600
chrX	140992755	140992862
chrX	140993194	140997187

4.2.2 dbSNP database file

The genome build version of dbSNP should be matched to the genome build of reference file. You may get this from UCSC or ANNOVAR website (http://www.openbioinformatics.org/annovar/annovar_download.html).

5. SNV Detection

5.1 SNVerGUI for Individual Sequencing

5.1.1 Load data

Aligned Data: This must be a single .bam file

Reference Genome: This is a reference file, which must have the same formats as used in the .bam file. Inconsistent referencefile, such as different chromosome names and different chromosome length, will prevent SNVer from running.

Target Regions: This is a target region file with bed file format. If there is no file specified, SNVer will pileup for the entire reference genome and call variants for the entire reference genome. Otherwise, SNVer will call variants from the target region specified in the target file.

dbSNP Path: This is a dbSNP file, which must have columns including chromosome, position and SNP ID. The build version of the file should be the same as genome build of reference file. The format is "path for dbSNP, column number of chromosome, position and snp_id". The default is null, meaning that no such query needed.

5.1.2 Parameter Setting (More Options)

Base Quality Threshold: Only consider bases with base quality above the cutoff. The default is 17 (>=0).

Mapping Quality Threshold: Only consider reads with mapping quality above the cutoff. The default is 20 (>=0).

Heterozygosity:	the prior for computing posterior probability of genotypes. The default is 0.001 [0-1].
Strand Bias Threshold:	Aiming to remove potential false positives due to strand bias issue. SNVer uses a one-sided binomial test for alternative forward count, and alternative reverse count. The default p-value cutoff is 0.0001 [0-1].
Fisher's Exact Threshold:	Aim to remove potential false positives due to allele imbalance issue. SNVer uses a one-sided Fisher's exact test for contingency table of alternative forward count, alternative reverse count, reference forward count, reference reverse count. The default p-value cutoff is 0.0001 [0-1].
P-value Threshold:	The SNVer p-value threshold for testing significant variants. The default p-value cutoff is based on Bonferroni correction, the definition is 0.05/the number of tests. If specify a p-value cutoff, say 0.5,the loci with p-value greater than the cutoff would be filtered out. P-value range is [0-1].
Reads Supporting Threshold:	Require at least this number of reads supporting each strand For alternative allele. For example, if alternative forward count is 0 and alternative reverse count is 10. The loci would be discarded. The default is at least 1 supported read (>=1).
Ratio Discarded Threshold:	Require the read ratio of alt/ref above the threshold, aiming to filter out loci with reference bias problem. The default ratio is 0.25 [0-1].
Error Tolerated Threshold:	The default is 30 (>=0), which means if observing 30 or more alterative count, SNVer will not conduct such tests (one-sided binomial test for Strand Bias Threshold and Fisher's Exact Test Threshold). Decreasing the threshold will increase the sensitivity but lower the specificity.
Haploids:	The number of haploids. the default is 2.

5.1.3 Output

SNVerGUI generates the following major output files:

VCF file for SNV and indel detection: prefix.all.filter.vcf will be outputted, according to the p-value cutoff used. The default is input_file.all.filter.vcf (input_file is the input bam file name). More details about VCF format can be found at **http://www.1000genomes.org/node/101**

Here is an example VCF file:

```
##fileformat=VCFv4.0
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AC,Number=1,Type=Integer,Description="Alternative Allele Count">
##INFO=<ID=SP,Number=.,Type=Float,Description="Strand Bias Pvalue">
##INFO=<ID=FS,Number=.,Type=Float,Description="Fisher's Exact Pvalue">
##INFO=<ID=PV,Number=.,Type=Float,Description="Pvalue generated by SNVer">
##FORMAT=<ID=AC1,Number=1,Type=Integer,Description="Alternative Allele Count Forward">
##FORMAT=<ID=AC2,Number=1,Type=Integer,Description="Alternative Allele Count Reverse">
##FORMAT=<ID=RC1,Number=1,Type=Integer,Description="Reference Allele Count Forward">
##FORMAT=<ID=RC2,Number=1,Type=Integer,Description="Reference Allele Count Reverse">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=1,Type=Float,Description="Phred Scaled Posterior Probablity of AA,AB,BB">
#CHROM      POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT      test1.bam
chrX    140993642       .       C       T       .       .
        DP=192;AC=101;FS=0.384;SP=0.036;PV=1.6860120965662937E-175      AC1:AC2:RC1:RC2:GT:PL
        60:41:48:43:1/0:1075.62,-0,1700.73
chrX    140993877       .       C       A       .       .       DP=168;AC=38;FS=0.005;SP=0.072;PV=0.0
        AC1:AC2:RC1:RC2:GT:PL14:24:83:47:1/0:1834.67,-0,331.93
```

The VCF files have the following features:

INFO	Meaning
DP	total depth above base quality threshold
AC	alternative allele count above base quality threshold
SP	strand bias test p-value
FS	fisher's exact test p-value
PV	p-value generated by SNVerIndividual
FORMAT	**Meaning**
AC1	alternative allele forward count above base quality threshold
AC2	alternative allele reverse count above base quality threshold
RC1	reference allele forward count above base quality threshold
RC2	reference allele reverse count above base quality threshold
GT	genotype: 1 for alt, 0 for ref 1/1: homozygous alternate 1/0: heterozygous 0/0: homozygous reference
PL	Phred posterior probablity of 1/1,1/0,0/0

CSV file: This is the same as prefix.all.filter.vcf, but it is easy to be opened in Excel.

Two log files: one is the prefix.console.log, the file includes the input file names and parameter values used in the calculation. Another is prefix.failed.log. It contains the explanations that those sites could not be called for some reasons, such as failing strand bias test and not supported for each strand. SNVerGUI will also print the ACGT counts in the failed loci for users' further investigation.

5.2 SNVerGUI for Pooled Sequencing

5.2.1 Load data

Aligned Data: This must be a folder, where one bam file is corresponding to one pool. For example, if ten bam files in the folder, it means that there are ten pools for analysis.

Reference Genome: This is a reference file, which must have the same formats as used in the .bam file. Inconsistent reference file, such as different chromosome names and different chromosome length, will prevent SNVer from running.

Target Regions: This is a target region file with bed file format. If there is no file specified, SNVer will pileup for the entire reference genome and call variants for the entire reference genome. Otherwise, SNVer will call variants from the target region specified in the target file.

Configuration: This file is used to calculate the number of haploids in each pool.

dbSNP Path: This is a dbSNP file, which must have columns including chromosome, position and SNP ID. The build version of the file should be the same as genome build of reference file.

5.2.2 Parameter Setting (More Options)

Base Quality Threshold: Only consider bases with base quality above the cutoff. The default is 17 (>=0).

Mapping Quality Threshold: Only consider reads with mapping quality above the cutoff. The default is 20 (>=0).

Strand Bias Threshold: Aiming to remove potential false positives due to strand bias issue. SNVer uses a one-sided binomial test for alternative forward count, and alternative reverse count. The default p-value cutoff is 0.0001 [0-1].

Fisher's Exact Threshold: Aim to remove potential false positives due to allele imbalance issue. SNVer uses a one-sided Fisher's exact test for contingency table of alternative forward count, alternative reverse count, reference forward count, reference reverse count. The default p-value cutoff is 0.0001 [0-1].

P-value threshold: The SNVer p-value threshold for testing significant variants. The default p-value cutoff is based on Bonferroni correction, the definition is 0.05/the number of tests. If specify a p-value cutoff, say 0.5,the loci with p-value greater than the cutoff would be filtered out. P-value range is [0-1].

Reads Supporting Threshold: Require at least this number of reads supporting each strand For alternative allele. For example, if alternative forward count is 0 and alternative reverse count is 10. The loci would be discarded. The default is at least 1 supported read (>=1).

Error Tolerated Threshold: The default is 30 (>=0), which means if observing 30 or more alterative count, SNVer will not conduct such tests (one-sided binomial test for Strand Bias Threshold and Fisher's Exact Test Threshold). Decreasing the threshold will increase the sensitivity but lower the specificity.

Allele Frequency Threshold: if you want to test all variants (both rare and common variants),this should be 0. The default is 0.01. Allele frequency range is [0-1].

Haploids: The number of haploids. If assuming diploid individual, the number should be 2 * no. of individuals in a pool. The actual used values are shown in Configuration file mentioned in above Load data section

5.2.3 Output

SNVerGUI generates the following major output files:

VCF file for SNV and indel detection: prefix.all.filter.vcf will be outputted, according to the p-value cutoff used. The default is input_directory.all.filter.vcf (input_directory/prefix is the directory contains all bam files of pooled align data). More details about VCF format can be found at **http://www.1000genomes.org/node/101**

The VCF files have the following features:

INFO	Meaning
DP	total depth above base quality threshold, among all the pools
NP	number of pools without no coverage or strand bias
AF	estimated alternative allele frequency
PV	p-value generated by SNVerPool
FORMAT	**Meaning**
AC	alternative allele count above base quality threshold in this pool
DP	total depth above base quality threshold in this pool

CSV file: This is the same as prefix.all.filter.vcf, but it is easy to be opened in Excel.

Two log files: one is the prefix.console.log, the file includes the input file names and parameter values used in the calculation. Another is prefix.failed.log. It contains the explanations that those sites could not be called for some reasons, such as failing strand bias test and not supported for each strand. SNVerGUI will also print the ACGT counts in the failed loci for users' further investigation.

6. Tutorial

6.1 SNVerGUI for Individual Sequencing

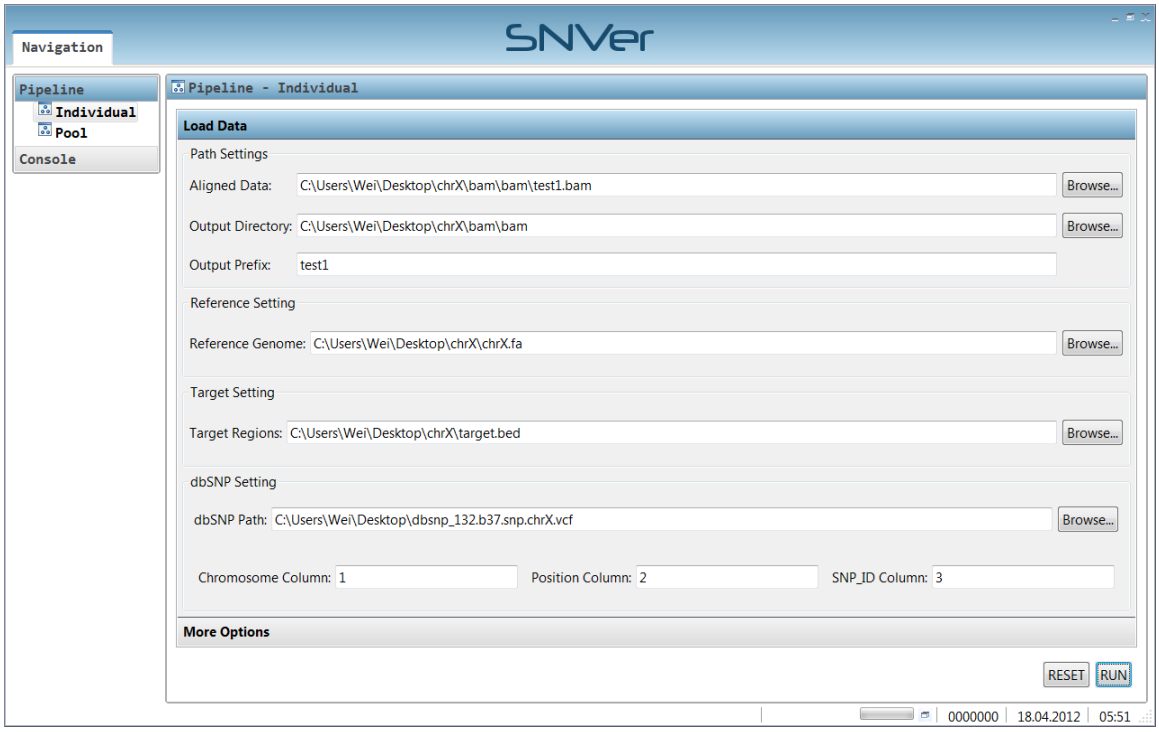
6.1.1 Download the example data

We can download the example data from http://snver.sourceforge.net/data.html . These files should be downloaded:

Reference for ChrX (fasta format): chrX.fa.gz
Input for individual data (bam/sam format): test1.bam or test2.bam
Target regions (bed format): target.bed
dbSNP file: dbsnp_132.b37.snp.chrX.vcf

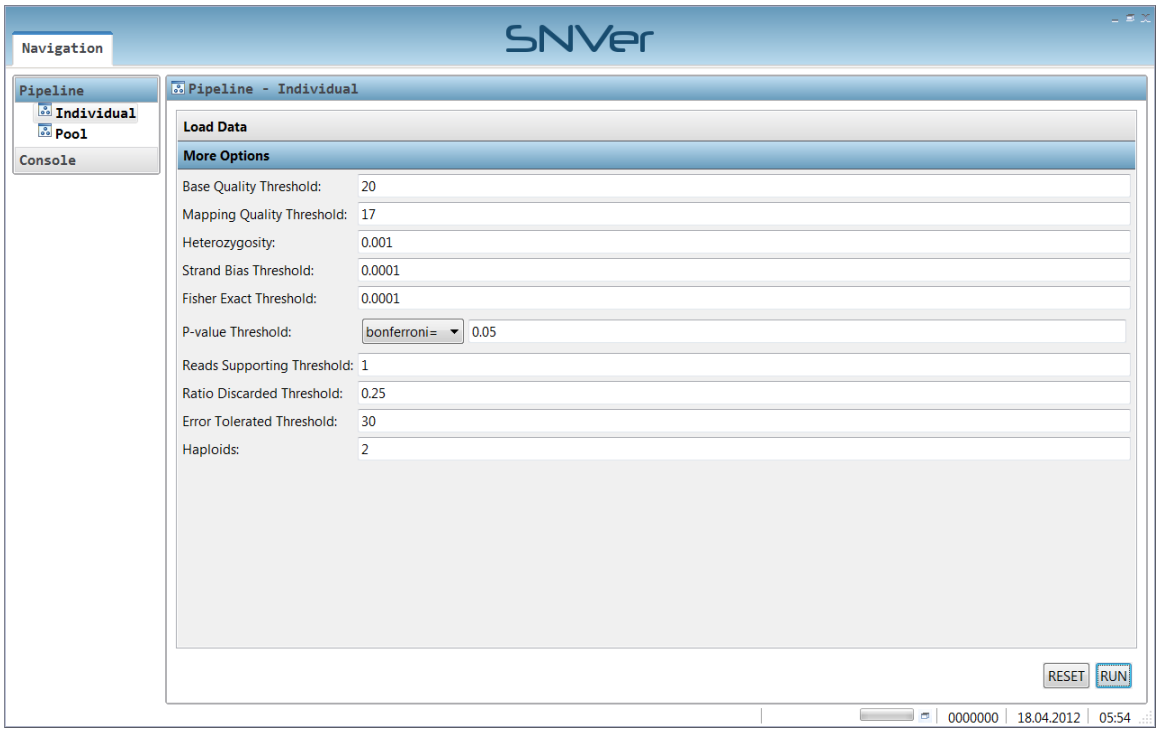
6.1.2 Load example data

When we start SNVerGUI successfully, click "Individual". Then we can specify the downloaded data:
A: specify aligned data file test1.bam at Align data;
B: Specify chrX.fa at Reference Genome. Note we need to first uncompress chrX.fa.gz by double-clicking it;
C: Specify target.bed at Target Regions
D: specify dbsnp_132.b37.snp.chrX.vcf at dbSNP Path. We also need to specify which columns in the file include information for chromosome number, physical position and SNP ID, respectively.



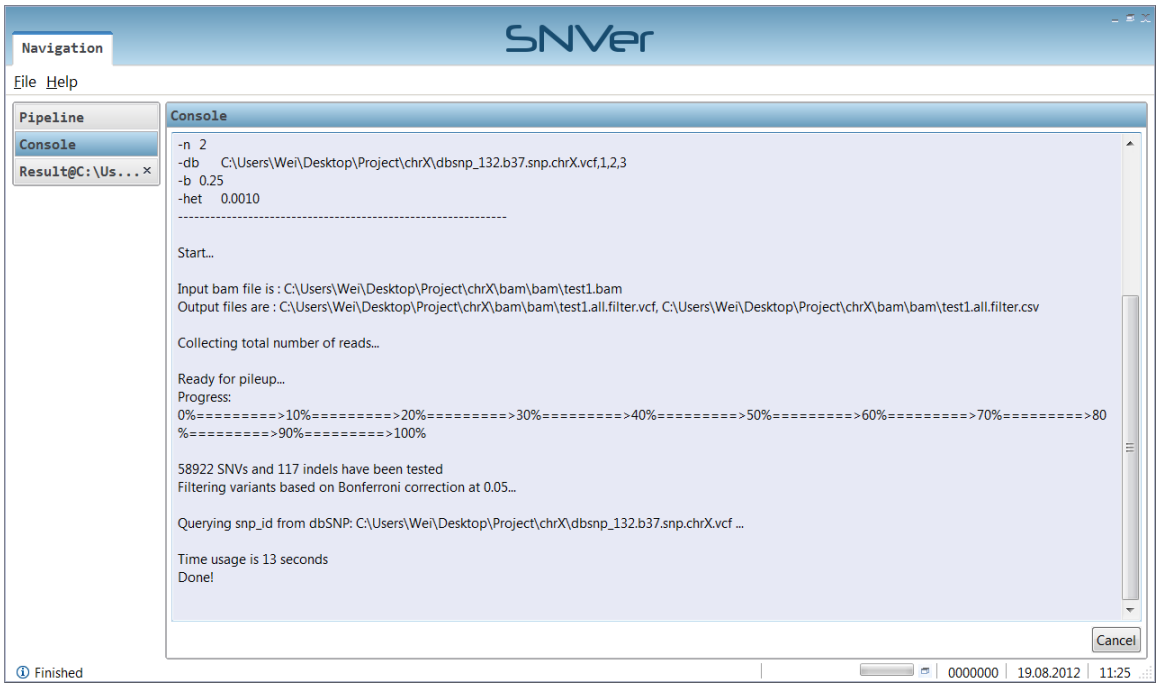
6.1.3 More Options

Here we set the default parameter values used in SNVer. The explanations of these parameters are discussed in Section 5.1. If users do not want to change the default values, then you can ignore these options. If users want to change some of these values, then they can modify the default values.



6.1.4 RunSNVer

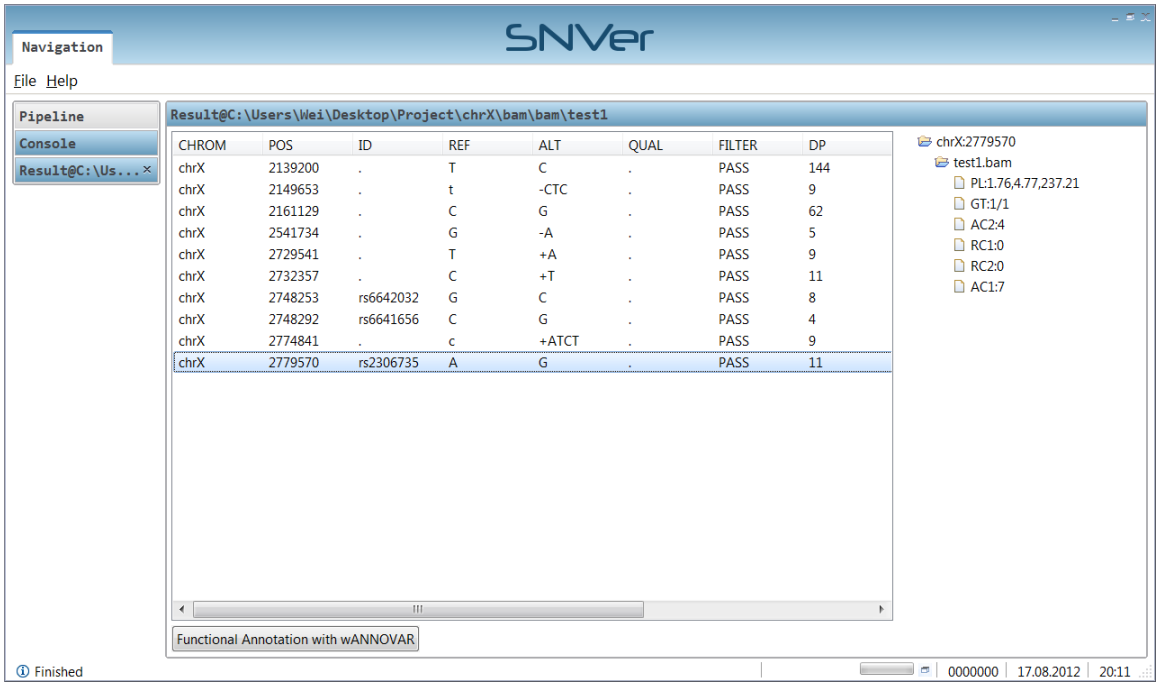
Just click "Run" button. The progress bar will monitor the running status. When the computing is completed, a bunch of files will be generated: test1.failed.log, test1.console.log , test1.all.filter.vcf and test1.all.filter.csv. The following information will be outputted in "Console" panel as follows, which is the same as test1.console.log:



6.1.5 Results

When Section 6.1.4 run successfully, the variant detection results (based on filtered results in vcf file) will be output in a table format in "Result" panel. The same results will be outputted in a .CSV file, where users are

able to rank the variants based on p-values, chromosomes, positions and so on. The meaning of each column corresponds to what is in VCF the result. For details, please refer to 5. SNV Detection.



6.1.6 Functional Annotations

By default, SNVerGUI works with wANNOVAR (11) for functional annotations after detecting variants. User can simply click the button to go to wANNOVAR website. Sample identifier and email must be entered to receive results, and the input file is the prefix.all.filter.vcf, which is generated by SNVerGUI. Note that the reference genome build need to be consistent with the build used in aligning the sequencing. Then user can submit the annotation job through wANNOVAR website. For more information, please see wANNOVAR tutorial (http://wannovar.usc.edu/tutorial.html).

6.2 SNVerGUI for Pooled Sequencing

6.2.1 Download the example data

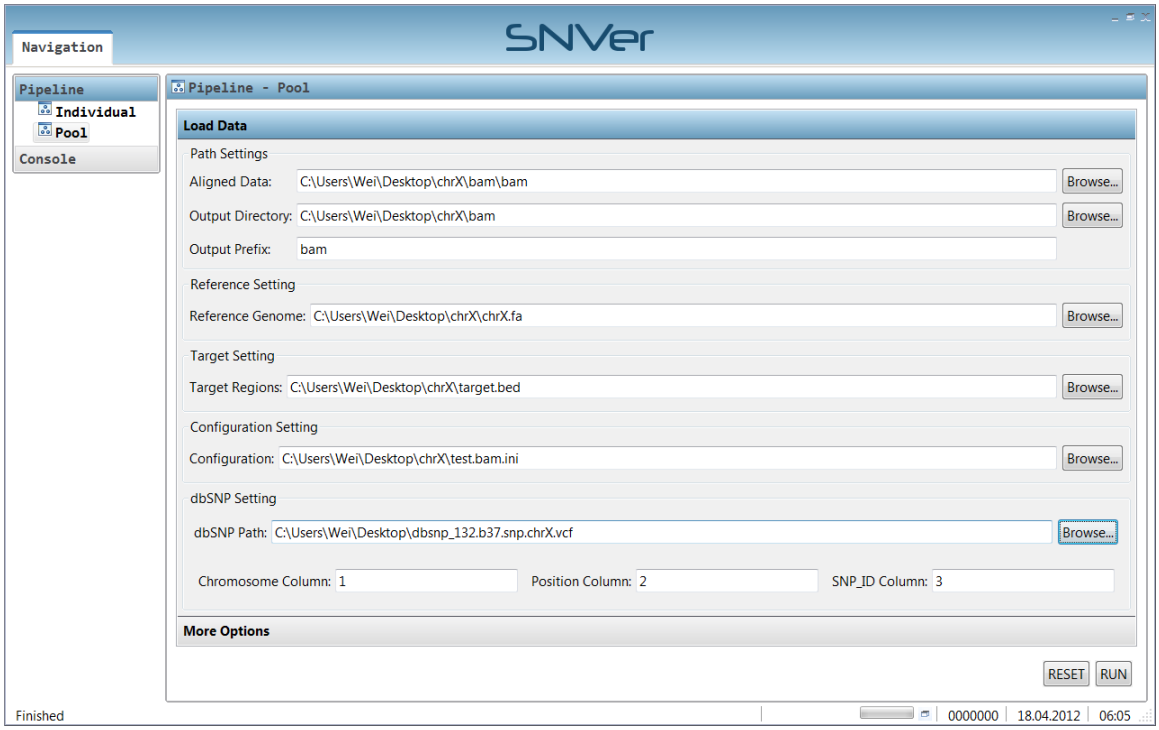
We can download the example data from http://snver.sourceforge.net/data.html . These files should be downloaded:

Reference for ChrX (fasta format): chrX.fa.gz
Input for pool data (bam/sam format): bam.zip (unzip it to **bam folder** by double-clicking it)
Target regions (bed format): target.bed
Configuration for pool data: test.ini
dbSNP file: dbsnp_132.b37.snp.chrX.vcf

6.2.2 Load example data

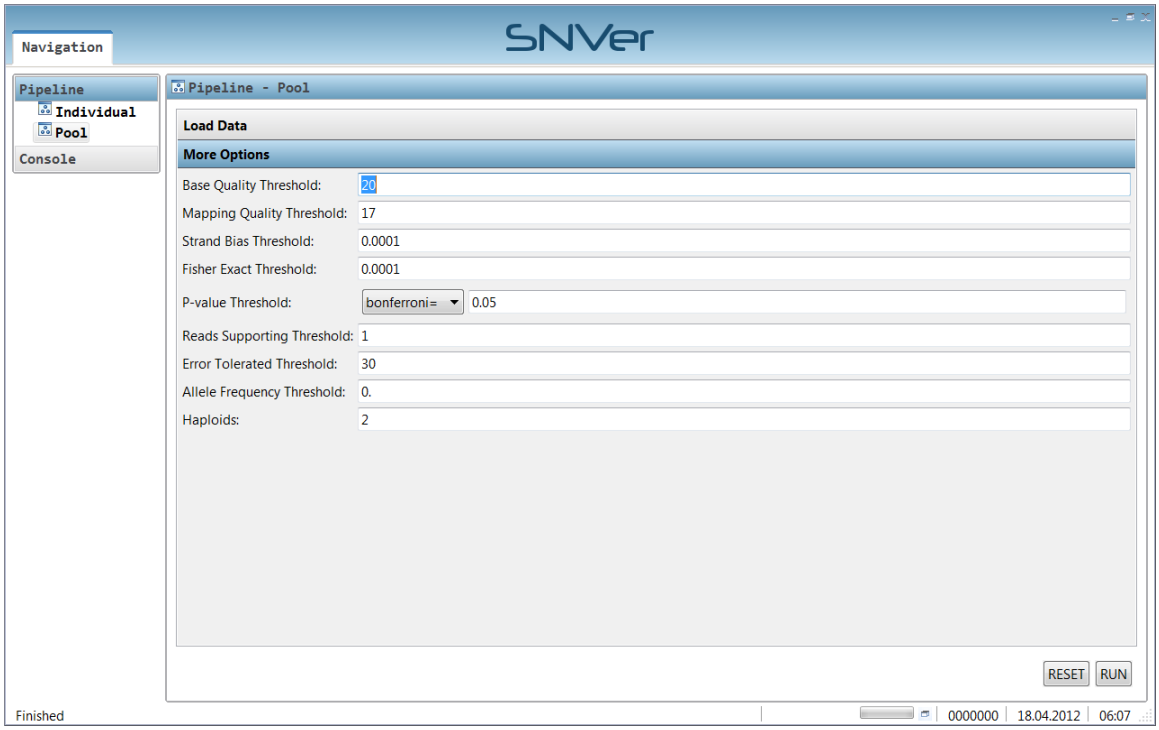
When we start SNVerGUI successfully, click "Pool". Then we can specify the downloaded data:
A: specify the folder with bam aligned data Align data;
B: Specify chrX.fa at Reference Genome. Note we need to first uncompress chrX.fa.gz by double-clicking it;
C: Specify target.bed at Target Regions

D: Specify pool information configuration file test.ini at Configuration.
E: specify dbsnp_132.b37.snp.chrX.vcf at dbSNP Path. We also need to specify which columns in the file include information for chromosome number, physical position and SNP ID, respectively.



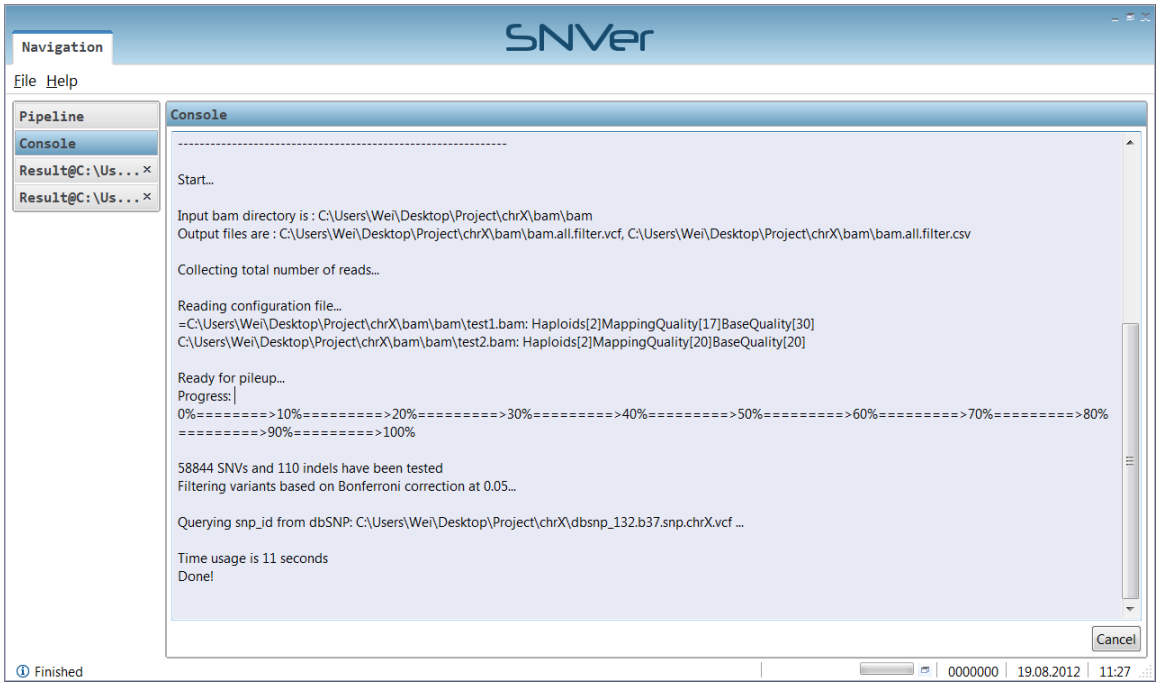
6.2.3 More Options

Here we set the default parameter values used in SNVer. The explanations of these parameters are discussed in Section 5.2. If users do not want to change the default values, then you can ignore these options. If users want to change some of these values, then they can modify the default values.



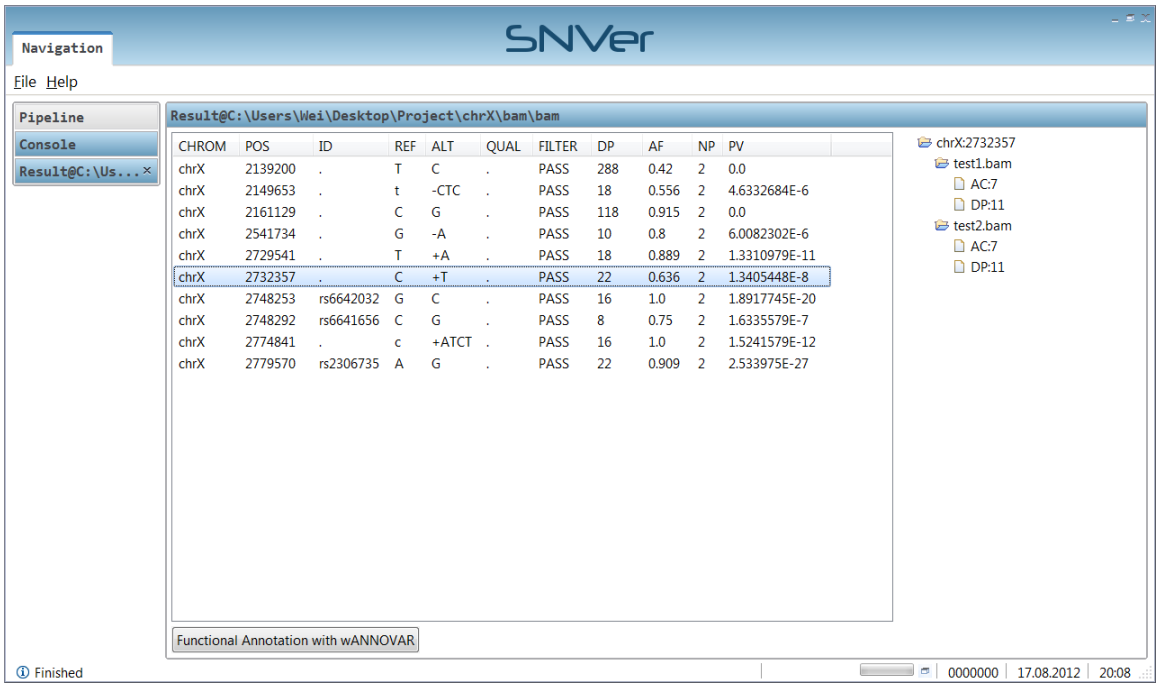
6.2.4 RunSNVer

Just click “Run” button. The progress bar will monitor the running status. When the computing is completed, a bunch of files will be generated: bam.failed.log, bam.console.log , bam.all.filter.vcf and bam.all.filter.csv. The following information will be output in “Console” panel as follows, which is the same as bam.console.log:



6.2.5 Results

When Section 6.2.4 run successfully, the variant detection results (based on filtered results in vcf file) will be outputted in a table format, where users are able to rank the variants based on p-values, chromosomes, positions and so on. Meanwhile, such results are also output to a .csv file, which can be also opened by Excel for further manipulation. The meaning of each column is corresponding to what is in VCF the result. For details, please refer to 5. SNV Detection. Note that if the results panel has no output, that means there are no sites meeting the criterion (e.g. MAF>=0.01), so smaller “Allele Frequency Threshold” should be set.



6.2.6 Functional Annotations

By default, SNVerGUI works with wANNOVAR (11) for functional annotations after detecting variants. User can simply click the button to go to wANNOVAR website. Sample identifier and email must be entered to receive results, and the input file is the prefix.all.filter.vcf, which is generated by SNVerGUI. Note that the reference genome build need to be consistent with the build used in aligning the sequencing. Then user can submit the annotation job through wANNOVAR website. For more information, please see wANNOVAR tutorial (http://wannovar.usc.edu/tutorial.html).

7. References

1. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic acids research. 2011;39(19):e132. Epub 2011/08/05.
2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10.
3. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 2010;26(12):i318-24. Epub 2010/06/10.
4. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283-5. Epub 2009/06/23.
5. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491-8. Epub 2011/04/12.
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20.
7. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25. Epub 2009/03/06.
8. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PloS one. 2009;4(11):e7767. Epub 2009/11/13.

9. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics. 2011;27(7):1011-2. Epub 2011/02/01.
10. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. Epub 2010/01/19.
11. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. Journal of medical genetics. 2012;49(7):433-6. Epub 2012/06/22.