

CusVarDB Manual

CusVarDB is a windows based tool for creating a variant protein database from Next-generation sequencing datasets. The program supports variant calling for Genome, RNA-Seq and ExomeSeq datasets. The program performs mainly 4 modules

1. Align the datasets with reference database
2. Perform the variant calling using Genome Analysis Toolkit (GATK)
3. Annotate the variant using ANNOVAR
4. Create the variant protein database

Apart from the main modules, the program also supports additional functions such as

1. Download the SRA
2. Convert the SRA file to fastq file format
3. Download the annotation (ANNOVAR) database and Dry-run concept to customize the commands

Availability of executables

1. <https://sourceforge.net/projects/cusvardb/>
2. <http://bioinfo-tools.com/Downloads/CusVarDB/V1.0.0/>

Test dataset

Test dataset is available at https://sourceforge.net/projects/cusvardb/files/Test_datasets.zip/download or http://bioinfo-tools.com/Downloads/CusVarDB/V1.0.0/test_dataset.rar

System requirements

- Windows 10 or above
- Minimum system requirements include Intel i5 or i7 having at least 4 cores with 8 GB of RAM and 1 TB hard drive

(**Note:** High performance processors such as Intel i9 or Xeon and large quantity of RAM can enable faster execution of tasks.)

Prerequisites

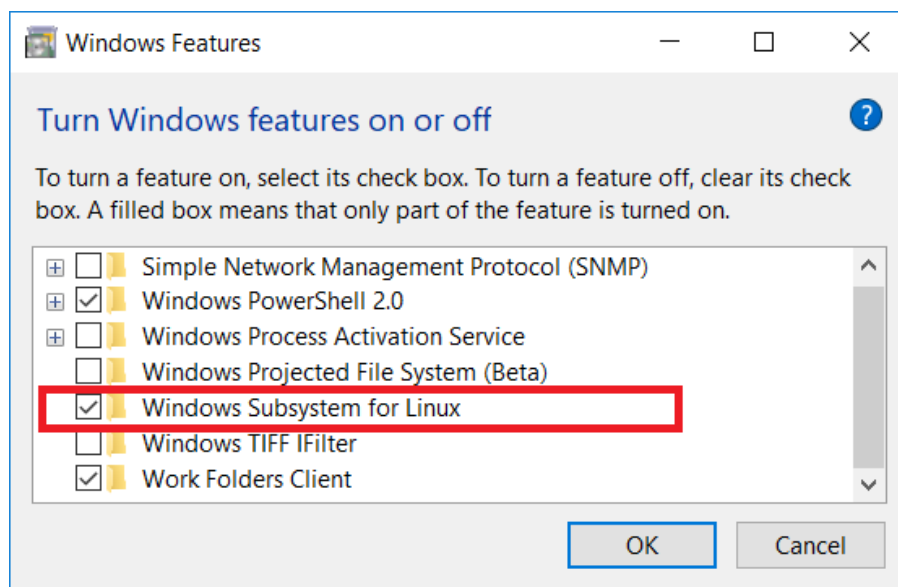
The tool works on Windows 10 operating system. It requires Ubuntu 18.04 LTS and ANNOVAR to be downloaded and installed.

Install Ubuntu 18.04 LTS on Windows OS

Settings->Update & Security->For Developers, enable the "Developer mode" by clicking the radio button. This will install the packages requires to Run the Linux environment.

Enable bash on Linux

Control panel->Uninstall programs->Turn windows feature on or off-> click on "Windows Subsystems for Linux". Restart the Operating System to take the effect. Refer below image to enable “Windows Subsystem for Linux”.



Restart the Operating System to take the effect.

Installing the Ubuntu 18.04 LTS

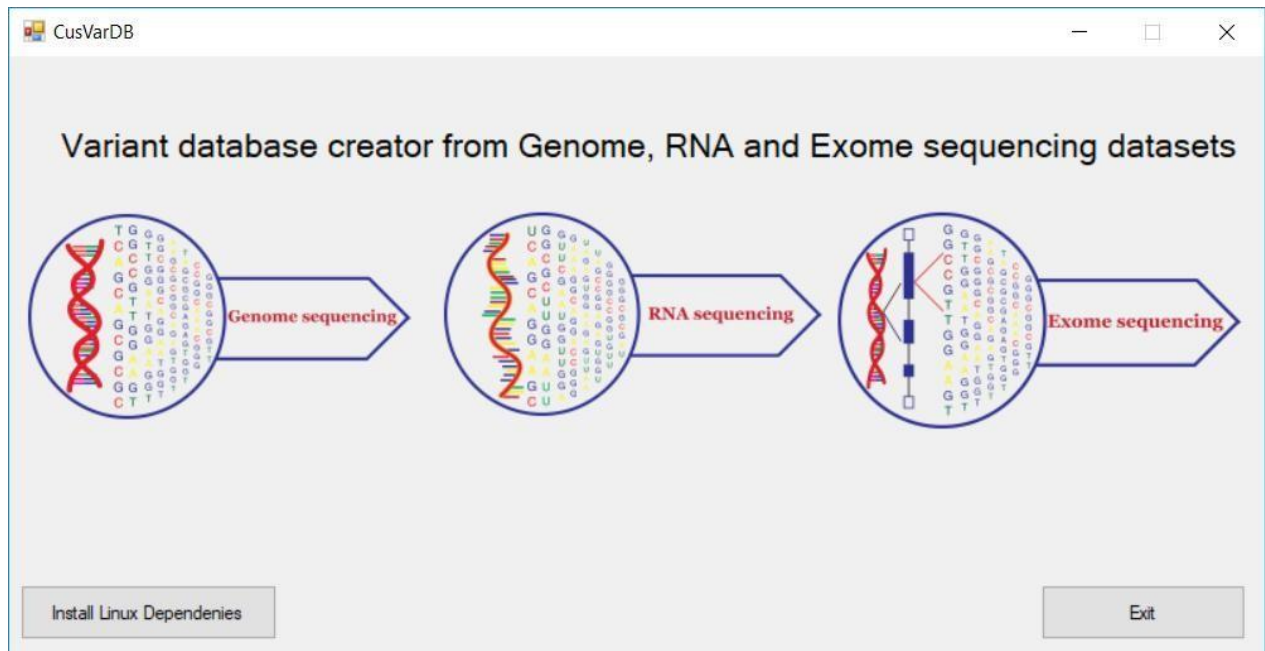
Go to Microsoft stores->Search->

Type “Ubuntu 18.04 LTS” and download the Ubuntu 18.04 LTS. (If problem in installing, please follow the YouTube or the web link

<https://www.youtube.com/watch?v=Cvrqmq9A3tA> or <https://www.windowscentral.com/how-install-bash-shell-command-line-windows-10>). After the installation of Ubuntu, necessary packages has to be installed such as

- Java 8
- BWA
- Samtools
- Unzip

These tools and packages can be downloaded by the CusVarDB tool. Below image will provide the brief of downloading the Linux dependencies and packages.



Install Linux dependencies option will download the tools and packages. After the each tool / package installation the terminal will asks for user to press any key to continue and download the next tool / package.

In case, if user will find any problem in installing the Linux dependencies, below command will help in installing the packages / tools manually.

```
sudo add-apt-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install openjdk-8-jre
sudo apt-get install bwa
sudo apt-get install Samtools
sudo apt install unzip
```

(**Note:** Make sure that you have followed the order of downloading the tools / packages. In case any error occurred during the alignment process then make sure that you have installed the Samtools version 1.7 (using htlib 1.7-2) and BWA version (0.7.17-r1188).)

The tool ANNOVAR needs to be downloaded by the user.

1. ANNOVAR can be downloaded at:

<http://annovar.openbioinformatics.org/en/latest/user-guide/download/>

All the downloaded files are kept in a folder named “tools” as mentioned in the below image.

CusVarDB > tools				
Name	Date modified	Type	Size	
annovar	2/18/2020 11:32 PM	File folder		
DB_making	2/18/2020 11:32 PM	File folder		
FastQC	2/18/2020 11:32 PM	File folder		
gatk	2/18/2020 11:32 PM	File folder		
perl	2/18/2020 11:33 PM	File folder		

(**Note:** Make sure that the Folder and the file names of downloaded tool are same as above mentioned. Otherwise the tool will end up giving error message “No such file or directory”)

Additional features

User can download the SRA files from the SRR, ERR or DRR ids. The configuration panel provides the download option for downloading the SRA dataset using the wget command and fastq-dump command will convert them into single or paired end datasets.

CusVarDB (Exome analysis)

File Tools

Configuration Quality control Alignment and Variant calling Annotation Dry-run

Set the Reference fasta file path Browse

Set the dbSNP file path Browse

C:\Users\new\Desktop\Test_Path Browse

Set the BED interval file path Browse

Or

Enter the chromosome (Empty will consider all the chromosomes) Enable

Set the Annovar directory Browse

SRA file download

ERR486857 Download

Convert SRA file to Fastq

Select layout

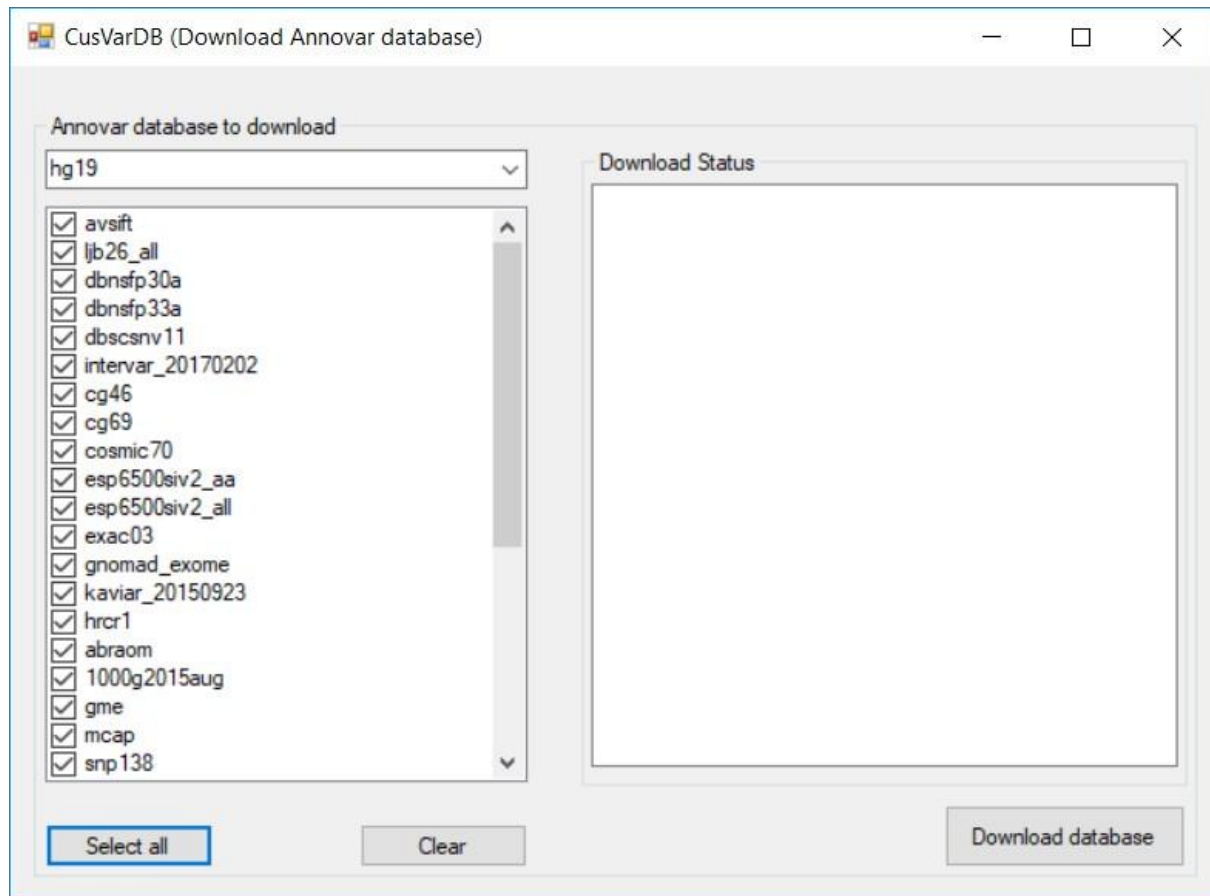
Download SRA file. Eg: SRR102145 Browse

Convert

[Click here to download the database](#)

The variant annotation is performed using ANNOVAR tool. The “download database” option will allow the user to download the annotation database provided by ANNOVAR. The user can select the database by clicking the individual “check box” or by clicking the “select all” button to download the listed database.

(**Note:** During the variant annotation step, the user needs to select the downloaded database. In other case “database not exist” error will occur)



Dry-run




The dry run option helps the advanced or professional user to customize the command. This panel generates the command ready to run as bash script. The bash scripts generates raw VCF file. Further, the annotation and variant protein database can be generated from the “**Annotation tab**”. Dry run generated commands are stored in a dry_run.sh file. This file can be executed in terminal by typing below command

```
>bash dry_run.sh
```

(**Note:** The bash script needs to be run on Linux Environment.)

Running the GUI

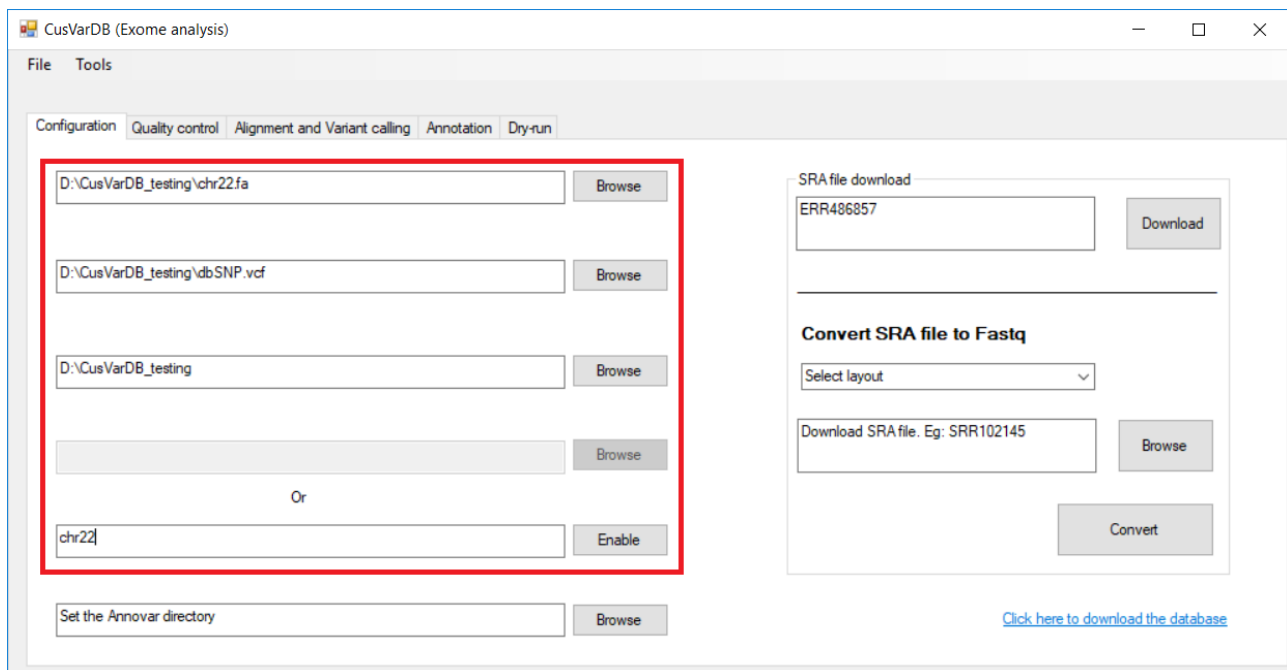
To run the application, extract the CusVarDB.rar file and navigate to the folder CusVarDB. You will find the executable called CusVarDB.exe or CusVarDB Refer the below figure.

 CusVarDB.exe	3/18/2019 8:58 PM	Application	1,678 KB
 CusVarDB.exe.config	7/25/2018 10:17 A...	XML Configuration...	1 KB
 CusVarDB.pdb	3/18/2019 8:58 PM	Program Debug D...	172 KB

Working with test datasets

The test data set a sub set from the large whole exome data, where we have extracted the chromosome 22 reads and made the dataset (read1.fastq and read2.fastq). The reads were extracted from the Samtools view command. From whole exome data we extracted the chr22 bam file and converted the BAM file to fastq file using the picard samtofastq command. Apart from the Fastq datasets, you will find dbSNP and the reference genome database used for the analysis.

Make sure that before starting any task **set the working directory**. Keep all the datasets and required prerequisites in the working directory. It is good to have different directories for different studies to avoid conflicts. All the analysis files and the results files are stored in the working directory. User always has the option to go back to the necessary files to get to know whether the analysis is completely performed or not. Below figure shows uploading the required files for the analysis.



Quality control (QC)

Before we proceed to analysis, it is good to know about the quality of data. Generally the variant calling required high quality data which gives the true positive variants. We have used FastQC tool to run the quality for raw reads. This operation will provide the basic statistics and per base sequence quality in the main panel. The detailed summary option will be provided by the “Show full summary” option.

Alignment and variant calling

Alignment and variant calling is performed with multiple steps, we have used tools such as BWA, Samtools, picard and GATK to perform these steps. The current panel provides the user to select the GATK algorithms (UnifiedGenotyper or HaplotypeCaller) and the data layout (Single or Paired-end). User can customize the threads and the memory, default thread is calculated based on total number of threads / 2 and the memory is calculated by dividing the total memory (in GB) by 2 (total memory / 2) . Read group ID, Platform unit, sample and platform are the labels, detailed information is provided by the GATK (<https://gatkforums.broadinstitute.org/gatk/discussion/6472/read-groups>)

In addition, break after each command will pause the execution of command. The next command will be resumed only if the user enters any key. This option is helpful in understanding the backend process.

Annotation and variant database generation

This panel performs the variant annotation using ANNOVAR tool. Make sure that the database which has been selected for the annotation is already been downloaded. After the annotation step, variant protein database is generated from the raw variants generated by the GATK tool.

The variant protein database is generated from the RefSeq version 81. User has the option to update the database to new version, the RefSeq database will be found at directory “DB-making\homo_sapiens_RS81.fasta”.

(Note: The latest version of RefSeq database can be downloaded at NCBI. Make sure that the header of the database must be in specific format (as mentioned below).

```
>gi|53828740|ref|NP_001005484.1| NP_001005484.1#OR4F5#79501#olfactory receptor 4F5  
[Homo sapiens]
```

(Note: Any change in the header will lead to index error.)