DNAscan2 Graphical User Interface (GUI) Manual

Contents

1. Introduction

DNAscan2 is an end-to-end next generation sequencing analysis pipeline which uses popular bioinformatics tools to identify a wide range of genomic variants, including SNVs, small indels, transposable elements, short tandem repeats and large structural variants. 30-40x whole genome sequencing data (gold standard for clinical diagnostics) can be optimally processed with DNAscan2 using approximately 4 CPUs and 16GB RAM.

As next generation sequencing is becoming more accessible in the clinical and biomedical genetics community, we appreciate that researchers with limited bioinformatics proficiency and/or access to high performance computing facilities may struggle when using DNAscan2 in its original command-line tool format. Therefore, we have developed a graphical user interface to allow those individuals to perform next generation sequencing analysis.

2. Installation

a. Minimum Requirements

Before downloading DNAscan2, you will require the following:

1. A computer which can run Linux (Ubuntu >=14.04). To use DNAscan2 on MacOS or Windows computers please see the following:

   MacOS

   You can install and use Ubuntu alongside your current operating system by either:

   i. Creating a bootable Ubuntu USB stick
   The following tutorial (https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview) will guide you through the process. It is up to you whether you want Ubuntu to run alongside macOS (dual booting) or replace macOS (for more information, see https://ubuntu.com/tutorials/install-ubuntu-desktop#6-drive-management).

   ii. Running an Ubuntu virtual machine by installing and configuring VirtualBox
   A virtual machine can be thought of as 'a computer inside a computer,' running software alongside and separately from your computer (a.k.a the 'host') operating system. To install a virtual machine that runs Ubuntu, follow this guide (https://ubuntu.com/tutorials/how-to-run-ubuntu-desktop-on-a-virtual-machine-using-virtualbox#1-overview).
   Note: this option can be trickier when transferring files between systems, as the virtual machine needs to communicate with the filesystem on your computer.

   Windows

   The best way to run Ubuntu on Windows is to install Windows Subsystem for Linux (WSL), which allows you to run commands without having to dual-boot or install a virtual machine. To do so, follow this guide to install WSL (https://learn.microsoft.com/en-us/windows/wsl/install).

2. Git, Python3 and PySimpleGUI installed on the system
   Both of these should already be installed in Ubuntu. To check if Git is installed, copy and paste the following when you open up the Ubuntu terminal:

   $ git –version

   If you get output like the following, Git is installed:

git version 2.25.1

If not, install Git using the following commands and check again:

$ sudo apt install git
$ git --version

Similarly, to check if python3 is installed, insert the following command in the terminal:
$ python --version

The terminal should display a version of python3, i.e. Python 3.x.x like the following:
Python 3.8.5

If not, e.g. the version is < 3, install python using the following command and check again:
$ sudo apt install python3.8
$ python --version

Next, ensure that pip is installed on your system by running the command:

$ python3 -m pip --version

If you get a command similar to the following, you do not have to do anything:

pip 21.2.4 from /path/to/pip

Otherwise, you can install pip by running the following command and checking the pip version again:

$ python3 -m ensurepip --default

If this does not work, you can download a helper script via the link https://bootstrap.pypa.io/get-pip.py, and run:

$ python get-pip.py

Next, install PySimpleGUI:

$ pip install pysimplegui==4.60.4

3. Necessary storage space
   We recommend at least 110Gb of free space for the whole DNAscan2 deployment, which includes alignment and annotation dependencies.
   - Alignment: 13.5Gb is required for the reference human genome and the bwa and hisat2 indexes.
   - Annotation: For SNV and indel annotation, the size of the Annovar databases can range from tens of megabytes to hundreds of gigabytes (e.g. CADD database). If the user wishes to perform the annotation step they must take this into account. Without the CADD database, the other default Annovar databases require ~ 90Gb free space. If you wish to perform SV and MEI annotation, AnnotSV human annotations occupy ~ 4Gb.

   We also recommend that you have some additional storage if you are:
   - Performing alignment: we recommend using at least 3 times the size of your input data e.g. if your fastq.gz files are 100Gb, you would need 300Gb of free space.
   - Detecting transposable elements: MELT requires an additional 20GB overhead storage.
   - If you don't wish to perform these steps, a proportion of data-to-analyse: free-space of 1:1 would be enough e.g. if your input data is a 50Gb bam file, you would need only 50Gb of free space.

b. Downloading DNAscan2 **and starting the GUI**

Assuming Git is installed, DNAscan2 is available to download from GitHub by navigating to the directory you want the tool to be downloaded in and executing the following command:

$ git clone https://github.com/KHP-Informatics/DNAscanv2

If Git is not installed, you can directly download the repository by entering its URL and then unzipping the directory:

$ wget https://github.com/KHP-Informatics/DNAscanv2/archive/refs/heads/master.zip

$ unzip DNAscanv2-master.zip

Once DNAscan2 is downloaded, navigate to the newly created DNAscanv2 directory by entering the following:

$ cd DNAscanv2

Once you are in the DNAscanv2 directory, enter:

$ python3 DNAscan_gui.py

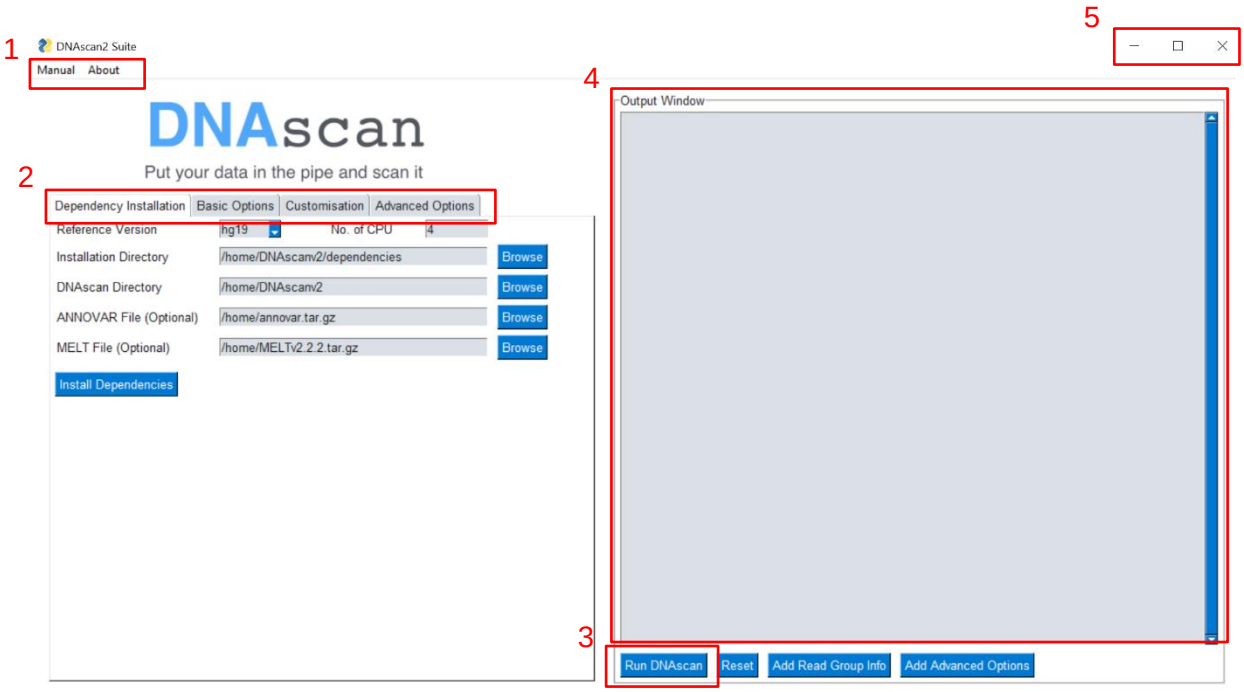
The GUI should now be up and running!

Note: Before running the GUI, you must register for and download Annovar at the following link (https://www.openbioinformatics.org/annovar/annovar_download_form.php) if you want to annotate SNVs and indels. If you want to perform mobile/transposable element detection, you can download MELT here (https://melt.igs.umaryland.edu/downloads.php).

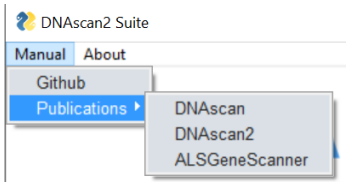
3. GUI Documentation

a. Overview

The basic layout of the DNAscan2 user interface is shown below. Before you analyse your own data, we suggest that you familiarise with the interface, especially if using DNAscan2 for the first time.



1. There are two menu options – the Manual tab is where you will find the DNAscan2 manual for the command line ('Github'), as well as links to our publications outlining the previous version of DNAscan, DNAscan2 and ALSGeneScanner, a component of DNAscan. The About tab gives basic information about the developers of DNAscan2. The full menu is below:

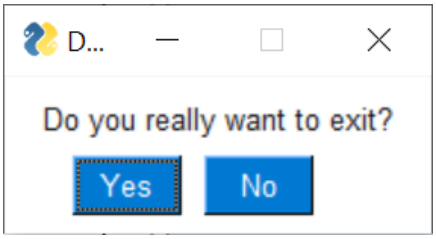


2. Before running DNAscan2, you must specify several options for your analysis. For first time users who may have limited bioinformatics proficiency, the ‘Install Dependencies’ tab allows you to install all of the necessary software and resources to carry out your analysis. The ‘Basic Options’ and ‘Customisation’ tab allows you to use DNAscan2 in its simplest form, just like the command line tool version, whilst the parameters on the ‘Advanced Options’ tab are displayed based on your input data and desired analysis and are purely optional.
3. After you are happy with your options, you can press the ‘Run DNAscan’ button and they will be converted into command-line format for DNAscan2 to perform analysis.

4. This is where you will see the running of your analysis, just like you would if you ran DNAscan2 via the command line. You will know when DNAscan2 has finished running as the following message will be displayed in the output window:

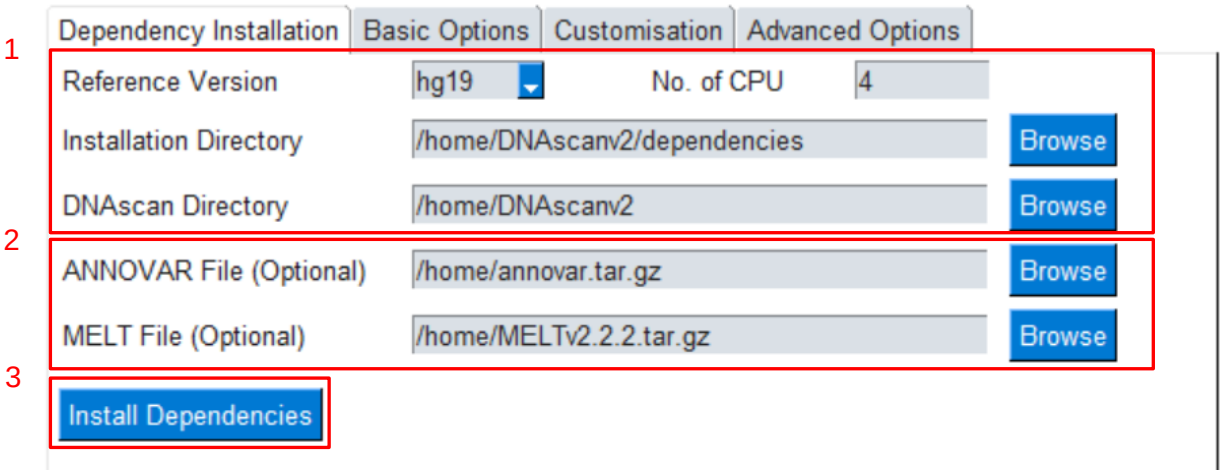
   **********DNAscan has finished running**********
   **********Your results are available to view from the results/ directory**********

5. When you have finished your analysis, simply press on the cross button in the top right corner. A popup window will appear to confirm that you want to exit the app:



b. Installing Dependencies

This section of the app is only for first time users of DNAscan2 who have limited bioinformatics proficiency so may struggle to install dependencies via the command line, as documented in the DNAscan2 GitHub.



1. There are several compulsory options for installing dependencies:
    - Reference Version: This is required to download the correct reference genome and indexes for analysis. Options are either hg19 or hg38. Default is hg19.
    - No. of CPU: Specifies the number of CPUs used to download and install dependencies. Default = 4.
    - Installation Directory: This is a folder where you want all of the dependencies to be downloaded into. We recommend the dependencies folder inside the DNAscanv2 directory, but you can download them wherever provided you specify the path.
    - DNAscan Directory: This is the location of the downloaded DNAscanv2 directory (see Section 2b).
2. Optional: If you want to perform SNV/indel annotation, provide the path to the annovar.tar.gz file in the 'ANNOVAR File' field. Similarly, if you want to detect transposable elements, the path to the MELTv2.2.2.tar.gz file in the 'MELT File' field. **Note: if you do not have the paths to these files and do not want to perform these steps, there will be warnings telling you that the directories cannot be found. This is normal.**
3. Click the 'Install Dependencies' button and the relevant software and resources i.e. reference genome and indexes will be downloaded and will automatically be entered into the paths_configs.py file, which is the file containing the paths to all of the tools necessary for DNAscan2 to run. This will take approximately 1 hour. When this process is complete, the output window will say the following:

    **********Dependencies have been installed**********

    Note: Depending on your Ubuntu system permissions, there will be a couple of instances where your terminal will ask for your password. Go onto the same terminal you are running the GUI from, input your password (which is not shown) and press enter. The output will then continue to be displayed in the GUI's output window. Furthermore, there will be one step (installation of perl cpanminus) which requires authorisation in all cases. Again, go to the terminal and type 'yes' and press enter.

c. Basic Usage

When you have installed dependencies and are ready to analyse your genomic data, this is the first tab to which you should navigate. It contains all of the basic options needed for DNAscan2 to run.

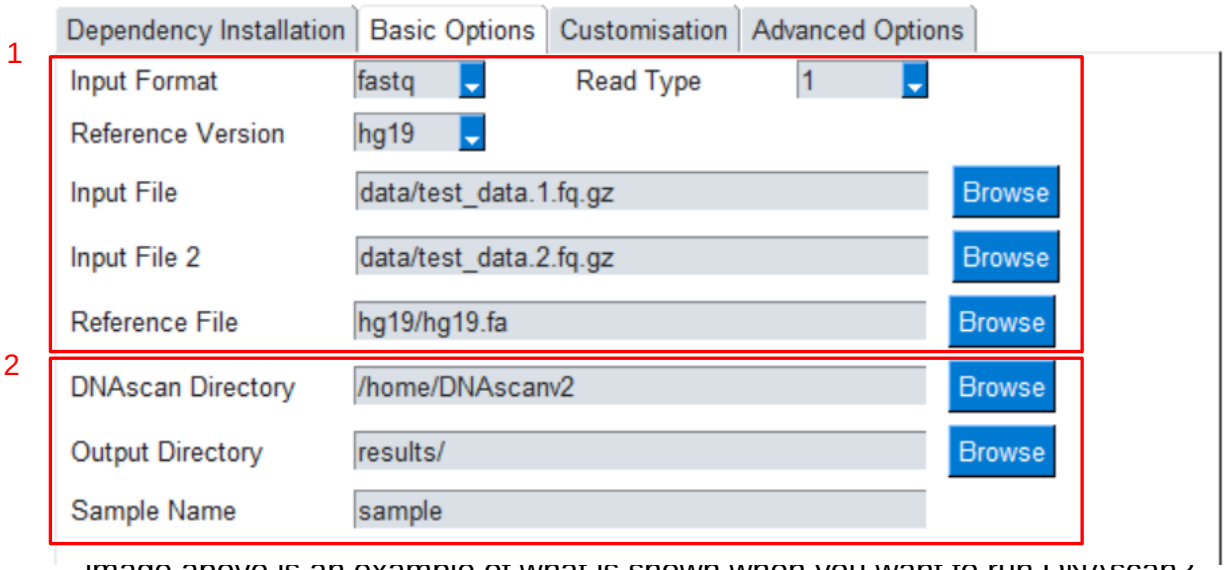


Image above is an example of what is shown when you want to run DNAscan2 using paired fastq files.

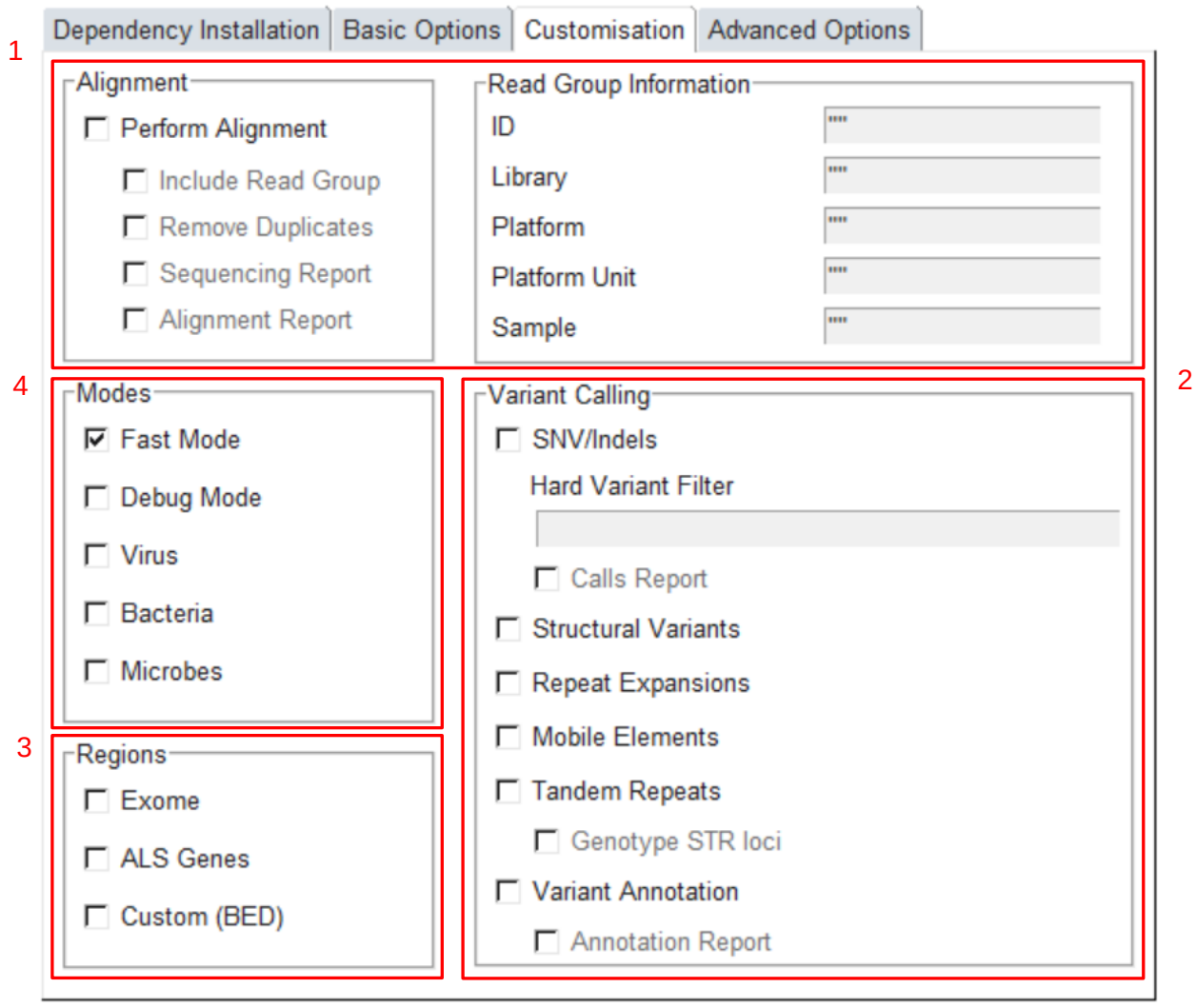
- Input Format: The type of data you want DNAscan2 to analyse. Options are either fastq, sam, bam or vcf. Default = fastq.
- Read Type: This is only displayed when 'Input Format' = fastq. Options are 0 (single-end) and 1 (paired-end). Default = 1.
- Reference Version: This is used for alignment, variant calling and annotation and should be the reference version that you have installed. Options are hg19, hg38, grch37 and grch38. Default = hg19.
- Input File: This should contain a path to the file you want DNAscan2 to analyse. It is always displayed regardless of the file type you want to analyse.
- Input File 2: This is only displayed when 'Input Format' = fastq and 'Read Type' = 1, and should be a path to the second paired fastq file (usually contains a 2 in the name).
- Reference File: Path to the installed reference file, which should be in fasta/fa format. Default = hg19/hg19.fa. **Note: If your genome is hg19 or hg38, your hg19.fa/hg38.fa files and indexes should be in the DNAscan2/hg19 or DNAscan2/hg38 folders.**

2. Once you have entered the format and locations of the input files, you should enter:
    - DNAscan2 Directory: This is the location of the downloaded DNAscanv2 directory (see Section 2b).

- Output Directory: This is a folder where all of the results files and reports will be available after DNAscan2 performs your analysis. We recommend that this is the DNAscan2/results folder, however, you can specify any location. Default = results/.
- Sample Name: This is the name of your sample undergoing analysis. Usually, this can be found in the name of the input file. Default = sample.

d. Customisation

Once your basic options have been entered, you can move to the customisation tab, which allows you to tailor the analysis to your requirements. There are several steps in this tab which you should follow carefully. Note that some of the options (in lighter font and boxes) cannot be specified until certain options are clicked.



1. This tab section is only required if you want DNAscan2 to perform alignment with input data in fastq format.
    - Perform Alignment: If you want to call only SNVs and indels (see point 2), only HISAT2 will be used to align all reads, otherwise BWA-MEM will additionally be used to realign any clipped and unaligned reads to improve detection of more difficult to detect variants such as mobile elements and structural variants. If ticked, the four alignment-based options will be enabled.
        - Include Read Group: When performing alignment, the reads will include the read group information in the aligned bam file. This is used to identify sets of reads which come from the same run of a sequencing experiment. Default = False.
            - Read Group Information: These fields will only be enabled if 'Include Read Group' = True. There are several components which make up the read group identifier. When entering the read group information, make sure to enclose them in the quotation marks. Once you have added this information, you can click 'Add Read Group Info' at the bottom of the output window to add to the analysis.
                - ID: Read Group Identifier. Default = "" .
                - Library: DNA library preparation identifier. Default = ""
                - Platform: This is the technology used to produce the read e.g. ILLUMINA. Default = "" .
                - Platform Unit: Flowcell and lane information in the format FLOWCELL_BARCODE:LANE:SAMPLE_BARCODE. Default = "" .
                - Sample: The name of the sample you want to analyse. Default = "" .
        - Remove Duplicates: This will remove duplicate reads from the aligned bam file. Default = False.
        - Sequencing Report: This will generate a quality report of the input sequencing data in fastq format with FastQC. Default = False.
        - Alignment Report: This will generate a report describing alignment metrics and characteristics with samtools flagstat. Default = False.

2. This is the main analysis customisation panel of DNAscan2. It allows you to specify the types of variants that you want to detect in your sample. The various options and descriptions are listed below.
    - SNV/Indels: Germline SNVs and indels are called with Strelka2. Default = False. If ticked, two SNV/indel calling-related options are enabled.
        - Hard Variant Filter: This refers to the hard variant filtering criteria applied to the Strelka2 generated variant file. Default = 'FORMAT/FT == "PASS" && FORMAT/DP > 10 && MQ > 40 &&

GQ > 20 && ID/SB < 2 && ADF > 0 && ADR > 0'. Currently this cannot be changed as this filter ensures retained variants are of the highest reliability and quality.
    - Calls Report: An SNV/indel variant report is generated using bcftools stats. Default = False.
- Structural Variants: Manta and Delly are used to call deletions, insertions, inversions and duplications greater than 50bp in length. Default = False. **Note: If 'Fast Mode' is selected (see point 4), only Manta will be used to detect structural variants.**
- Repeat Expansions: ExpansionHunter is used to scan for and genotype known pathogenic repeat expansions from a predefined repeat catalogue. Default = False.
- Mobile Elements: MELT is used to call mobile insertion elements - i.e. Alu, SVA and LINE1 transposable elements. Default = False.
- Tandem Repeats: A genome-wide short tandem repeat (STR) loci profile will be generated with ExpansionHunter Denovo. Default = False.
    - Genotype STR loci: If 'Tandem Repeats' = True, the STR loci identified will be genotyped with ExpansionHunter. Default = False. **Note: If 'Fast Mode' is selected (see point 4), this step will not be performed.**
- Variant Annotation: If 'SNV/Indels' = True, ANNOVAR is used to annotate small variant calls using several default databases (specified in the scripts/paths_configs.py file). If 'Structural Variants' and/or 'Mobile Elements' = True, AnnotSV annotates variant calls using their integrated databases. Default = False.
    - Annotation Report: If 'Variant Annotation' = True, variant reports will be generated. If 'SNV/Indels' = True, ANNOVAR results will be converted into a variant report in tab-delimited format. If 'Structural Variants' and/or 'Mobile Elements' = True, an HTML results report will be generated with knotAnnotSV. In addition, a concise report describing the basic characteristics of called simple and structural variants will be generated.

3. You can restrict variant calling to specific regions of the genome, including:
    - Exome: Analysis will be restricted to exonic regions. Default = False.
    - ALS Genes: Analysis will be restricted to ~170 amyotrophic lateral sclerosis (ALS) genes as part of ALSGeneScanner. Default = False.
    - Custom (BED): Analysis will be restricted to custom regions specified in a BED file. Default = False.

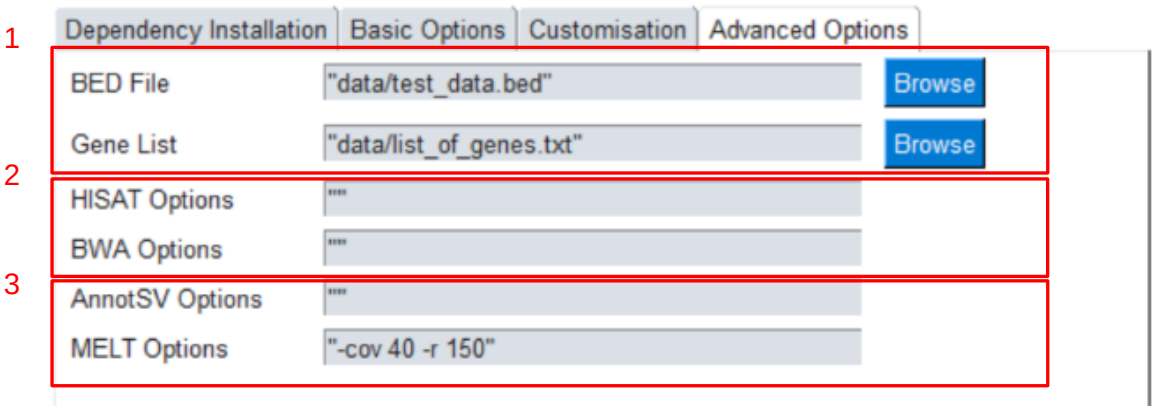
4. There are several optional modes which DNAscan2 offers, including:
    - Fast Mode: DNAscan2 will not call SVs with Delly and genotype STRs identified with ExpansionHunter Denovo. This is because Delly is time

intensive (~24 hours per genome) and genotyping requires a lot of RAM if performed on a genome-wide basis. Default = True.

- Debug Mode: All temporary and intermediate files will be kept after DNAscan2 has finished running. Default = False.
- Virus: The genome will be scanned for the presence of viral DNA. Default = False.
- Bacteria: The genome will be scanned for the presence of bacterial DNA. Default = False.
- Microbes: The genome will be scanned for the presence of custom microbes. Default = False.

e. Advanced Usage

There are some options in the customisation tab that, once selected, have additional parameters that the user can choose to set. These are listed in the 'Advanced Options' tab and only appear when the respective customisation-related option is clicked. The image below shows the tab when all advanced options are listed but no advanced options are set. **Note: These parameters do not have to be set, and if they are, they must be listed in-between quotation marks like below.**



1. These options appear when you want to restrict your analysis to custom regions (when 'Custom (BED)' = True). These can be specified using either a BED file, which describes the regions' location (chromosome, start position, end position) or a gene list with the names of one gene per line. The example above contains the locations of the bed file and gene lists of the test data. **Note: When both fields contain the file paths, the analysis will run using the regions located in the bed file.**

2. These options appear when you want to perform alignment (‘Perform Alignment = True’), with ‘HISAT Options’ becoming visible when ‘SNV/Indels’ = True, and ‘BWA Options’ becoming visible when you want other variant types to be called.

3. These options appear when you want to annotate your variants (‘Variant Annotation’ = True), with ‘AnnotSV Options’ becoming visible when ‘Structural Variants’ = True, and ‘MELT Options’ becoming visible when ‘Mobile Elements’ = True.

4. When you have entered your advanced parameters for analysis, press the ‘Add Advanced Options’ button in the bottom of the output window to provide this to DNAscan2.