PARSER: Reference Sequence Generator

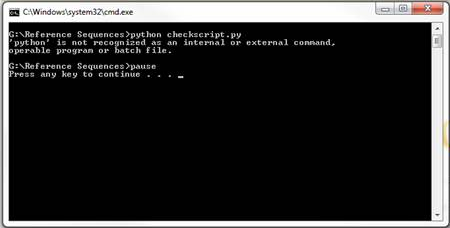
Introduction

Each genetic test designed and executed by the laboratory aims to compare the genetic sequence of a patient with an established wild type gene sequence. Most clinically actionable genes have an accepted reference against which all known variants can be mapped using the standardised HGVS nomenclature. When variants are identified using either Sanger sequencing or NGS technologies, they are double checked against the appropriate reference before reports are written to ensure that the variant numbering and nomenclature have been written correctly.

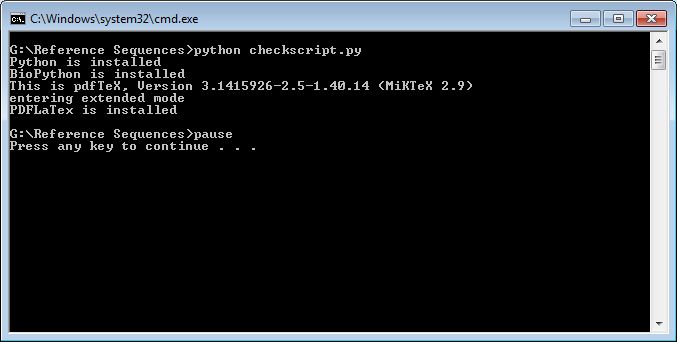
This program, PARSER (Python Automated Reference Sequencer) greatly speeds up the production of these documents, and removes human error by publishing references which are correctly typeset and numbered. This requires an input file containing the reference sequence in Genbank or LRG file formats, and outputs a PDF document. This program is capable of automatically annotating a sequence with in-house primers, and additional useful features (such as MLPA probe binding sites) can be added to the PDF without compromising the basic content.

# Access to the program

To use the program, your computer must have a few pieces of software installed. Installation of these should only take a few minutes, but will have to be performed by an IT Admin (Chris Kotara or Gareth Masson). If you are unsure if your computer has these already installed, double click the file called **Check** in **G:\Reference Sequences.**



The screen above will appear if Python is not installed



If Python is installed, messages will appear to state if the other dependencies are installed or not. If the result of this is that any items are not installed, please contact IT before attempting to use the program.

## Location of the Program

The program files are located in the folder: **G:\Reference Sequences**. This directory contains a range of files and folders, with key items described in the table below:

|  |  |
| --- | --- |
| **FOLDERS** |  |
| **Name** | **Purpose** |
| Input | Stores files to be used as input |
| output | Stores folders containing program output |
| primers | Location to store primer details when required |
| **FILES** |  |
| **Name** | **Purpose** |
| README | Instructions & guide |
| Run Main | Double click to run main program |
| Check | Run to see if dependencies are installed |

Table 1 key contents of the Reference Sequences directory and what each is used for

The input for this program can be provided in LRG or GenBank file formats, both types are stored in the same *input* folder.

# LRG

* All LRG files are available at [lrg-sequence.org](http://www.lrg-sequence.org/)
* These can be downloaded one-by-one, using the search bar to find specific gene references
* These can all be downloaded at once as a zipped folder; unzip into *input*

# GenBank

* Most can be found at <http://www.ncbi.nlm.nih.gov/nucleotide/>
* Use the search bar to enter a gene name, and follow the link to Genomic references

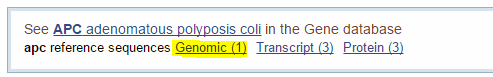


Figure 1 An example of the appearance of the link to genomic reference sequences

* This program requires NG\_ accession references, as NM\_ files will not contain the intronic sequence required for the final files

## Creating References

1. Download any files to be used as input, make sure they have an identifiable name (such as <Genename>\_Genbank), and put them into the ‘**input**’ directory
2. Double-click on the file called ‘Run Main’ to start the process. This will create the window shown in Fig. 2. By default the selected file name will be ‘**input/LRG\_292.xml**’ (BRCA1), this can be changed by using the ‘**Browse…**’ button.

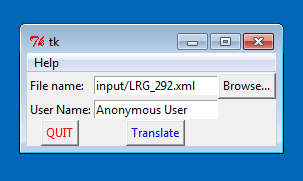


Figure 2 the interface to the PARSER

1. Press the **Browse** button to display the window in Fig. 3 which will default to the **input**folder (Fig. 3). Click on any file which you wish to create a reference sequence for and press **Open**, the name will be added to the ‘**File name:**’ bar on the main window.

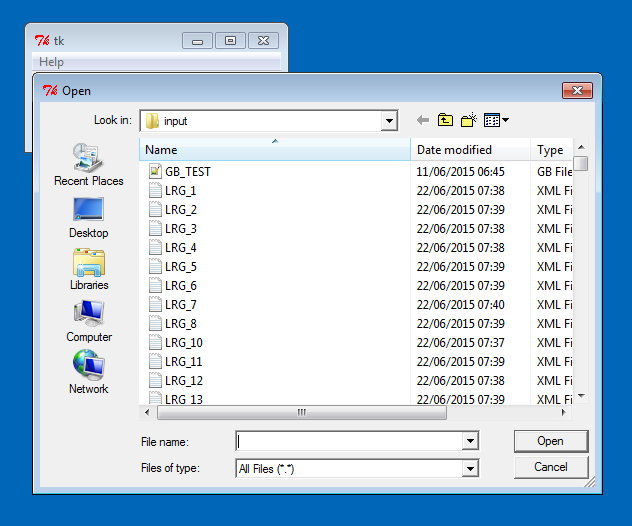


Figure 3 the 'input' directory displayed when using 'Browse...'

1. Add your username to the ‘**User Name’** field of the main window and press ‘**Translate**’.
   1. After pressing ‘**Translate**’, a number of lines will be printed to the terminal (an example is shown in Fig. 4). This is just a progress indicator which will print out as the program runs. This shows the exons identified for each transcript, should complete with the “**Process has completed successfully**” message, followed by a prompt to exit. Press Enter to close this window.

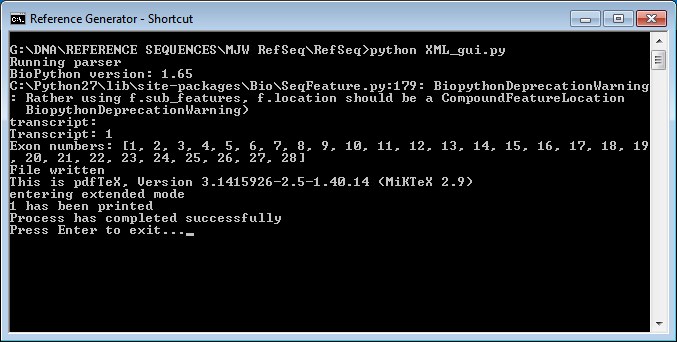


Figure 4 the text output of a successful completion on the terminal

## Accessing output

1. After running the program, go to the ‘**output**’ folder. There is a folder called **‘tex files’** which can be ignored, and a list of PDF files which are the completed references. The names of these are designed to be self-explanatory, and follow the pattern:

<Gene name>\_<NM/LRG number+transcript number>\_<Date>\_<Time>

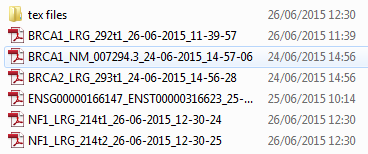


Figure an example of the output directory folder names

1. Figure 6 shows an example of this directory containing references created from different sources.
   1. The two BRCA1 files were created from an LRG file and an NG\_ file containing the same transcript, and are named accordingly. Although the contents are the same, the file name will be different.
   2. ‘ENSG…’ is a file created from an Ensembl transcript in a GenBank file format. Due to some formatting differences, files gathered from Ensembl may require some changes to be made before being accepted by the program (Contact Matt W.).
   3. The two NF1 files each contain a different form of the transcript, identified by ‘**t1**’ and ‘**t2**’ in the file names, and both of these files were created from running the single LRG file.

**Important Note: If a reference is run more than once it will overwrite the previous version of the document. This has been done to ensure that only the most recent version of a file is kept in the output folder. A reference from an LRG and a GenBank source file will not overwrite each other (as shown in Fig. 6). When the file is overwritten, the text file which was used to create it can still be found in the ‘tex files’ folder, so it can be recreated if required.**

# Errors

The program may not run successfully every time. This can result in a few easy to fix error messages:

* ‘**This program only works for GenBank and LRG files’**
  + indicates that the file extension of the selected input was not ‘.xml’, ‘.gbk’, or ‘.gb’. Usually this will be due to an accidental wrong choice of input file, and will likely be resolved by running the program on the correct input. If the intended input is not in one of these file types, either re-download the input file or contact one of the staff members listed in this document.
* ‘**The specified file cannot be located:**’
  + usually indicates that the file selected for input does not exist. This is usually due to spelling errors on manually entering the file name. Choosing the input by using the Browse button will usually solve this issue.

A final type of error may occur when the Base sequence does not match the Amino Acid sequence being printed (Fig. 4: the program will isolate each codon and translate them using a codon lookup table, comparing the output to the AA to be printed). This indicates an issue with the file contents and will freeze the program. Use print screen to record a copy of the message and try re-running the program (close the command console window to close the program). If the error persists try re-downloading the input file. If the file is definitely creating problems, contact a member of staff (for program help, Matthew Welland).

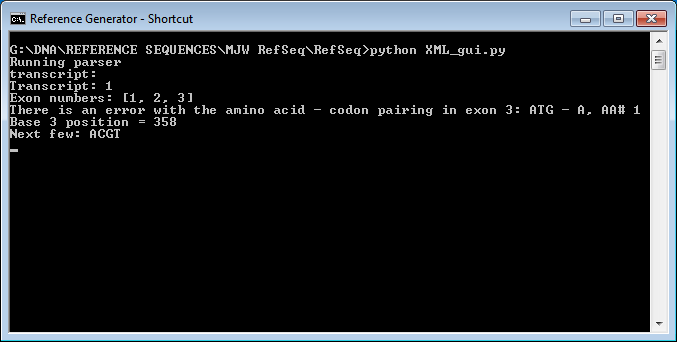


Figure 5 the output provided when codons and AAs do not match up

# Adding Primers

Although this program is typically used for creating a plain reference of the input genetic sequences, it is capable of performing automatic primer annotation when provided with a document in an appropriate format. This feature may be used for checking overlapping Sanger sequence fragments, or choosing the most appropriate primers for a re-sequencing check:

1. Open the spread sheet containing the current primer records for your gene of interest. An example location is ‘G:\DNA\Cancer section\BRCA\Primer sequences and batch records’. Within this folder the will be a spread sheet containing details of primers and current batch numbers. Make sure to navigate to the tab of the document (along the bottom) which contains the list of primers named ‘**<date> - Present**’

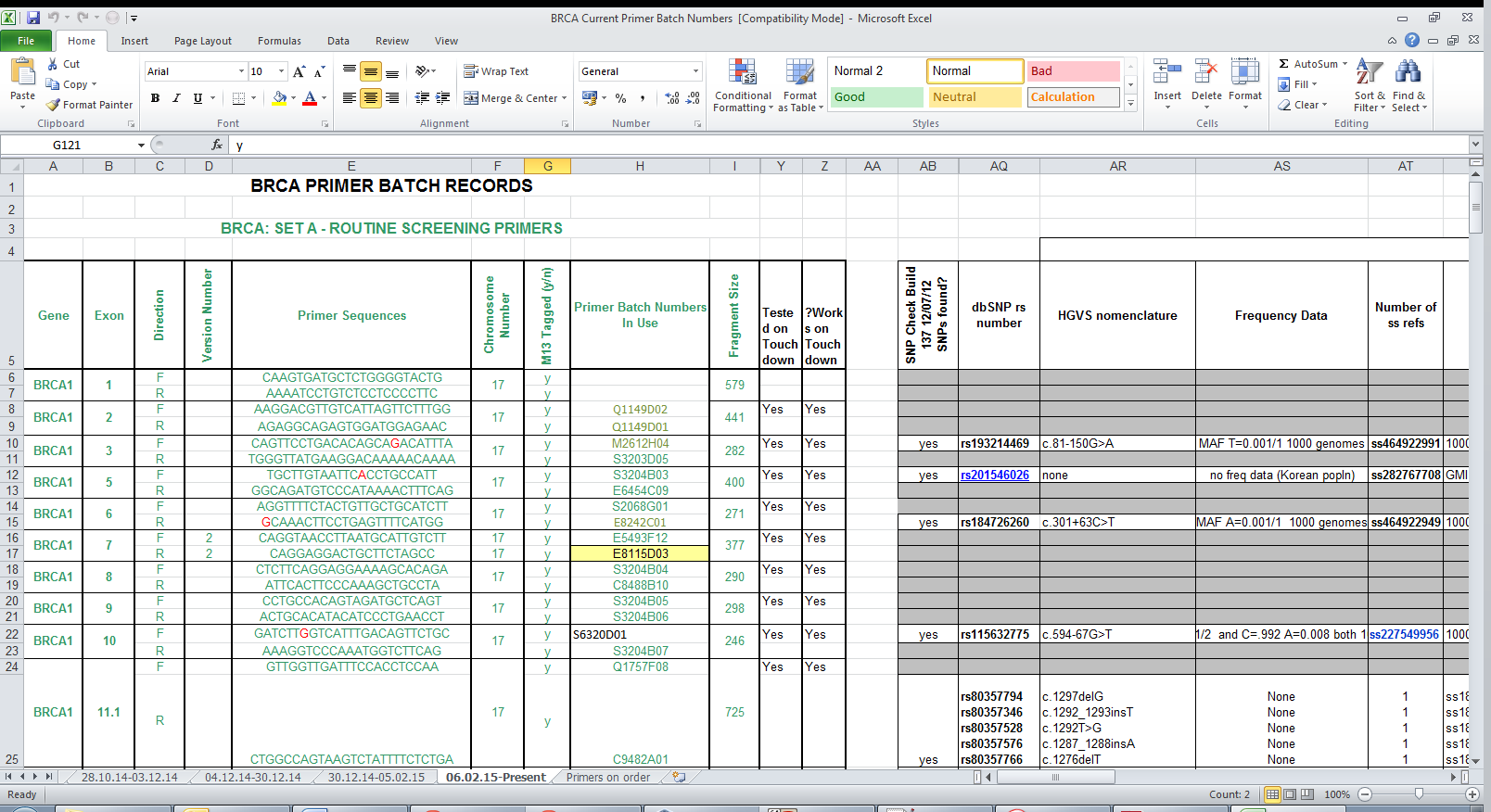


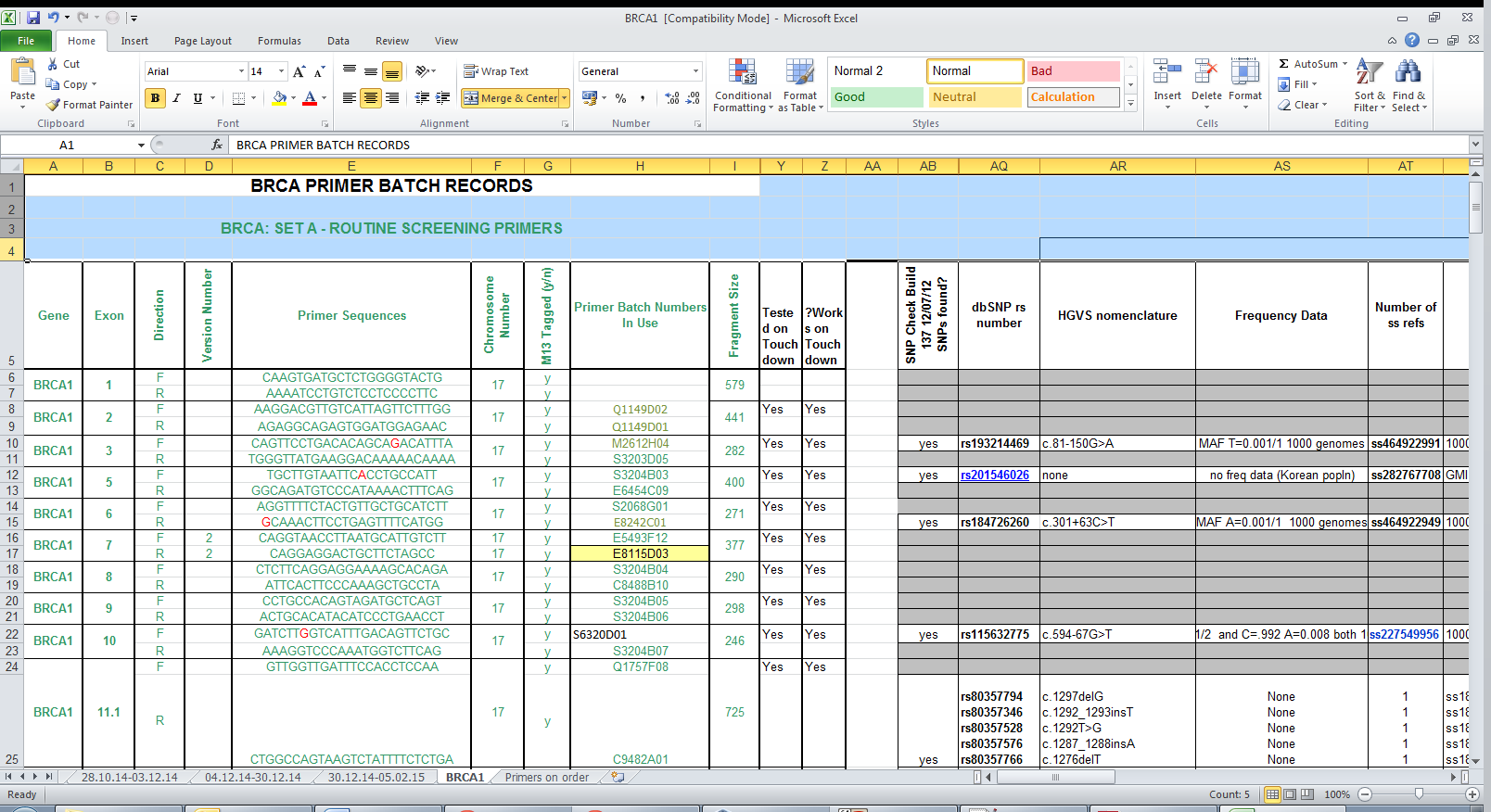
Figure 7 an example of a primer sequences and batch records document

1. Save a ‘.CSV’ version of this document into the folder ‘**G\:Reference Sequences\primers**’ under the name of the gene (Block Capitals and numbers). To do this go to the file menu and choose ‘**Save As…**’, navigate to the primers folder and change the document name to the name of the Gene, and below the file name, use the drop-down ‘**Save as type**’ menu to ‘**CSV (Comma Delimited)**’.
2. Two notifications will pop up, the first confirms that only the active sheet will be saved. Check that the open tab is marked as ‘**<date> - Present**’ and click **Yes**. The second will notify you that some spreadsheet features will be incompatible with ‘.csv’, click **OK**.

**NOTE: Make sure you have saved the document before making any changes, as the original file should not be edited.**

**After the file has been re-saved**

1. The top line of the file should contain the table headers (e.g. Gene, Exon, Direction…). To make sure this is the top row, highlight all rows above this by highlighting the row numbers on the left-hand side, right clicking and pressing delete (these have to be removed, not just cleared)



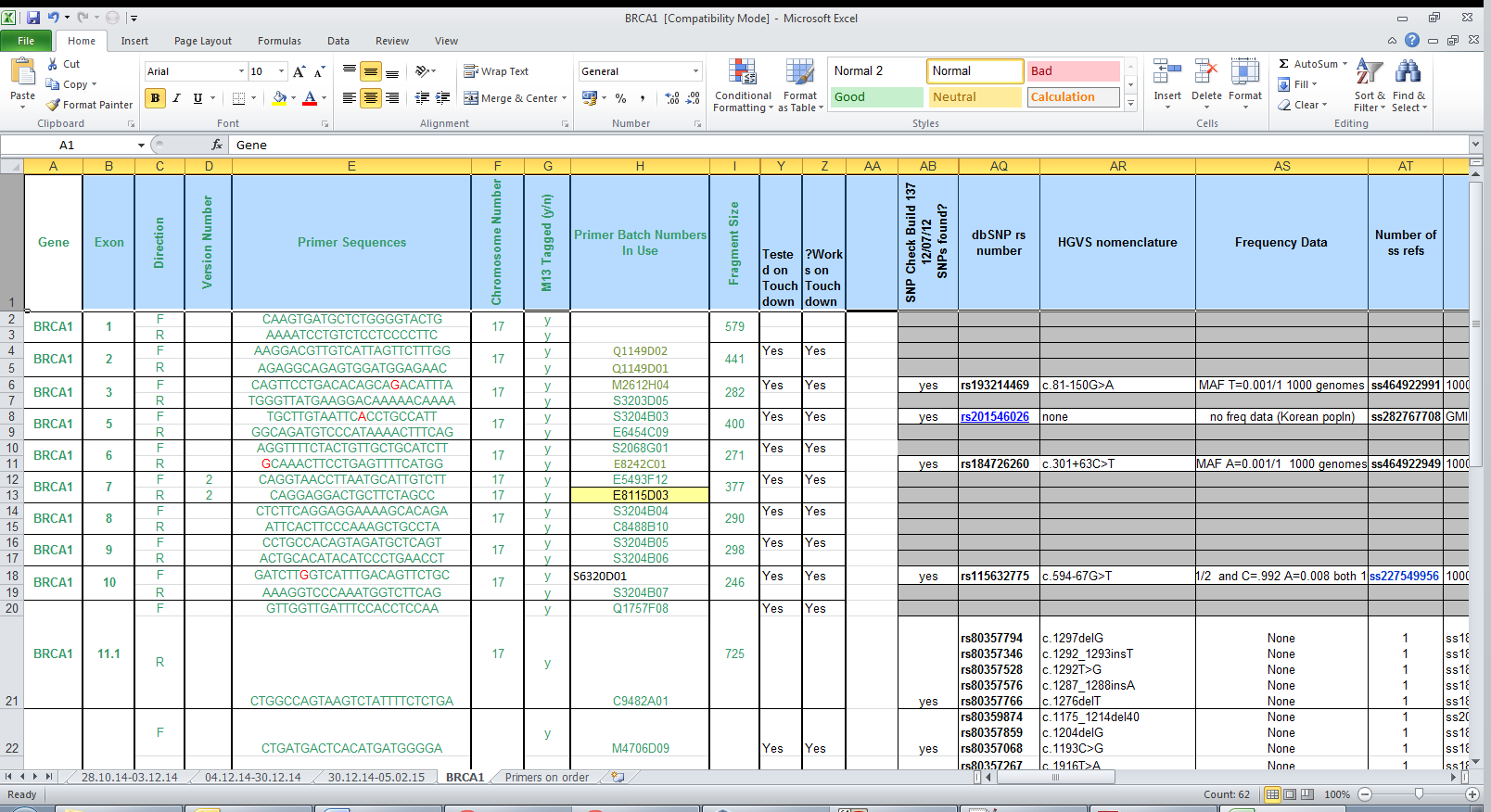


Figure 8 An example of a primer sequence document with headings (Top), and after the headings have been removed (Bottom)

1. If the document contains primers sequences for multiple different genes (for example when a document is grouped by condition rather than by individual gene), delete any rows which relate to the wrong gene name. Again, delete these rows by selecting the row numbers and using **Right Click > delete.** At the same time, any separate section involving B set primers should be removed, and any rows which containing only blank cells should be deleted.
   1. To use B set primers, the reference should be recreated with only the B set present in the primer document. A and B set primers may overlap, which will break the system used for annotation, so ability to apply both sets to each document has been removed
2. Once the Primer file has been set up, run the sequence for the selected gene as required and the primers should be annotated onto the final document, appearing as sections of highlighted text with a small speech bubble next to it, which will show details of the associated primer when the cursor scrolls over it.

**NOTE: If there are any problems created by using the primer annotation module (e.g. the program freezing or wrong output) please try:**

* **Opening the CSV file in excel and deleting any rows which are completely blank**
* **Opening the CSV file in a text editor (e.g. notepad) and removing and lines from the *end* of the document which are only commas with no content**

**If neither of these methods fixes the problem, please contact Matt Welland**