Primer Design Quest

ABOUT:

This program takes one of four types of input for primer design: VCF, CHROMPOS, FASTA, GTF. Primers are designed using Primer3 [1]. The input sequence and resulting amplicon sequence are aligned to the user-selected reference genome using BLAST [2] and the number of significant alignments found from each is reported in an output file containing the primer sequences and other information. Additional outputs include a table of any sequences failing primer design, all BLAST information for the input sequences and all BLAST information for the amplicons. The program works with single or multiple sequences. Each input type is described in detail below.

INSTALLATION:

This program was built to work on Linux systems and tested on Ubuntu 22.04. You may potentially get this to work in Mac or on Windows with a bash emulator, but this has not been tested.

First, download the PDQ_Dependencies .sh file matching the version of the program you are using. From a terminal window, give the script permission to run (chmod +x) and then run (./). This program will test if everything you need to run PDQ is installed on your system. The program will create a new directory containing the name DependencyTester_PDQ with the version and date added. Inside this directory you will find a log file that tells you if the program or command tool was found or if it needs to be installed, along with suggestions on how to install things.

RUNNING:

Once you have all the dependencies installed, give permission and launch the PDQ program from the terminal window (using chmod +x and ./). A graphical window will appear where you can enter the various parameters.

Figure 1. PDQ input page. Information on input files and primer design parameters are added by the user. 1) Clicking this link provides access to this document. 2) Initials provided by the user are recorded in the log file. 3) A directory name is supplied where the outputs of the program are stored. Directory names should not have spaces or any symbols other than an underscore (_). 4) The mode of the program determines the input files required and how the program runs. There are 4 types of input files the user can supply: VCF file, FASTA file, GTF file with accompanying gene name list, or chromosome position list. 5) The optimal Tm for primer3 primer design. 6) The optimal primer size for primer3. 7) The minimum primer size. 8) The maximum primer size. 9) PCR product size lower limit. 10) PCR product size upper limit (clicking the box for all the primer3 parameters allows a user to type in their own values). 11) The BLAST percent identity (pident) of identical nucleotide matches in an alignment. For example, if the user selects 85, then BLAST hits where the percentage of matches between subject and query are less than 85 will not be counted and reported in the primer design output table. 12) The reference genome. This is required for all program modes. It should be a .fa, .fna., or .fasta file. 13) The GTF file associated with the reference genome. Note that GFF files can be converted to GTF format using the tool gffread. GTF file is only required when selecting GTF mode. 14) If GTF mode is selected, a single column list of gene names (that exactly match the name in the GTF file is supplied. 15) If ChromPos mode is selected, then a 6 column file that includes the chromosome and position for primer design is supplied (see details below). 16) If FASTA mode is selected, provide the FASTA file. 17) If VCF mode is selected, provide the VCF file (can be .vcf or .vcf.gz). 18) Notes are added to the log file. 19) Click OK to start the program.
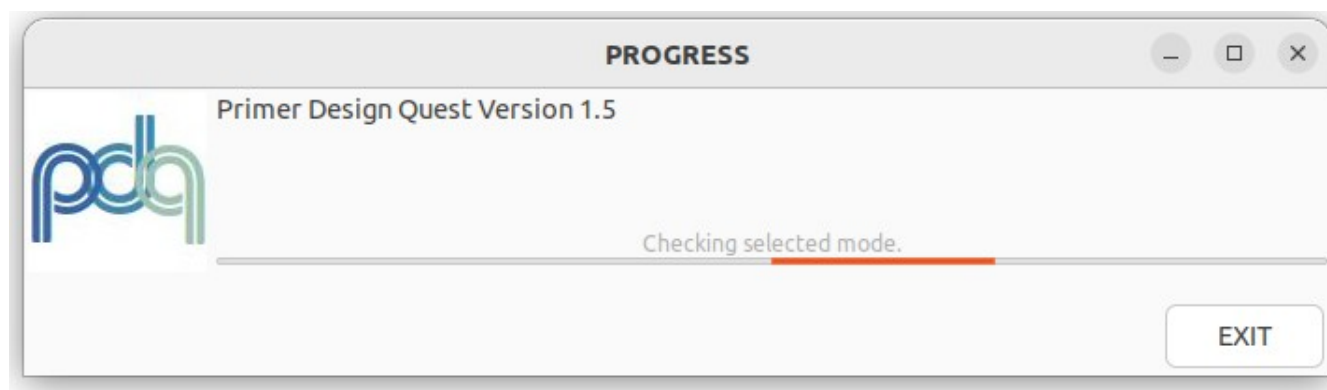
Figure 2. PDQ progress window. Light gray text above the orange bar indicates the program actions.

OUTPUT FILES:

Five output files are produced by PDQ. BLAST output information (in format 6) is provided for the sequence used as input for Primer3 and also for the amplicon sequence itself. These files contain all BLAST hits, regardless of the pident threshold used in PDQ. Both input and amplicon information is collected to aid in prioritizing primer choice with the assumption that primers designed from input sequences whereby both input and amplicon have single significant alignments to the reference genome may be more desirable to ones where the input sequence aligns to many regions of the reference genome. In most cases primers that amplify amplicons that align to multiple regions of a reference genome are undesirable. A log file and a file listing any sequences that resulted in no suitable primers from Primer3 is also produced. The data containing primer sequences, input, amplicon and BLAST alignment information is fond in a .csv file beginning with PDQ_Primers. Information in this output file will vary slightly depending on the mode chosen. Specific examples are provided below.
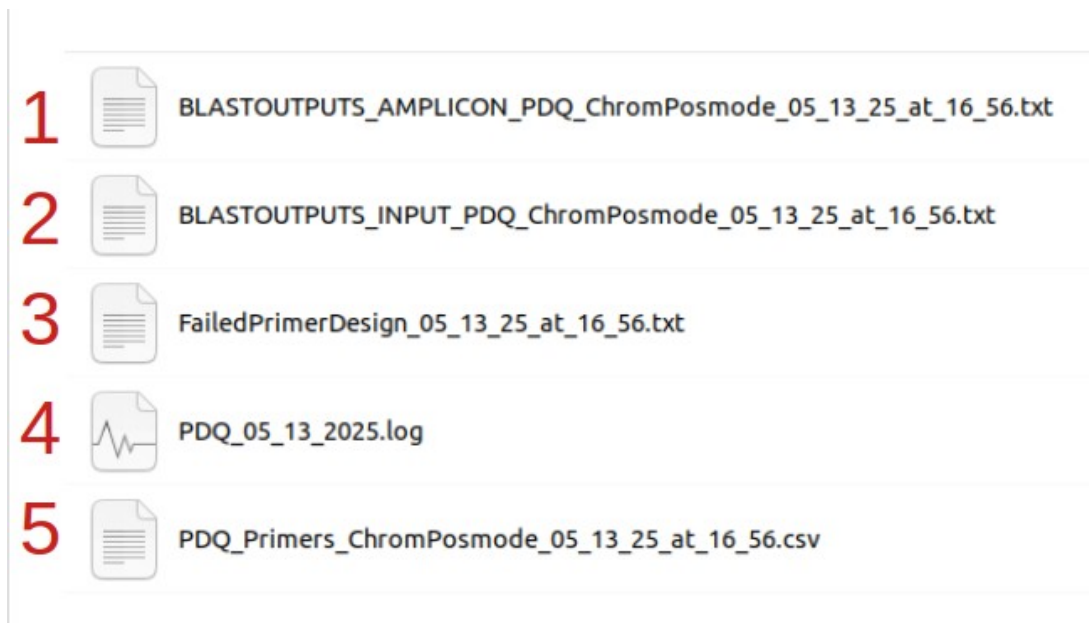
Figure 3. Output files from PDQ. 1) All BLAST alignments for the amplicon sequence from the designed primers. 2) All BLAST alignments from input sequences. 3) A list of any sequences for which Primer3 produced no suitable primer pairs. 4) PDQ log file. 5) Output file containing primer sequences, melting temperatures, input sequence, amplicon sequence, BLAST hits and other information (see MODES section for more details).

MODES:

VCF

This mode takes a VCF file (compressed or not compressed) as input. The program takes the chromosome name and position as the target name and then extracts nucleotide sequence surrounding the position of the variant (2x the size of the maximum PCR product chosen by the user, with the variant position at its center). This is used as input sequence for BLAST and Primer3. PDQ will search for a .fai file associated with the reference genome and make one if one is not present. The .fai file is used to correct for cases where the variant is close to the end of the chromosome. If the calculated position for extracting sequence exceeds the end of the chromosome, the position of the chromosome end is used. Likewise, if the variant is too close to the "front" end of the chromosome such that calculated starting position is a negative number, then the position is set to zero. The output .csv file contains the following columns:

Name,LeftPrimer,LeftPrimer_Tm,RightPrimer,RightPrimer_Tm,Product_Size,InputSequence, AmpliconSequence,Number_BlastHits_Input_Pident,Number_BlastHits_Amplicon_Pident,Chr om,InputStart,InputEnd,SeqTarget,Length

PDQ searches for a BLAST database in the same directory as the reference genome. If one does not exist, the program will generate a database. The number of BLAST hits are the number of alignments that meet or exceed the percent sequence identity as set by the user. SeqTarget is the Primer3 SEQUENCE_TARGET (https://primer3.ut.ee/primer3web_help.htm#SEQUENCE_INCLUDED_REGION) with Length being the length of the target. For this mode, the length is set to 1/6th of the maximum amplicon length. This is to ensure that the selected variant is not too close to the primer sequence as to interfere with validation by Sanger sequencing.

NOTES: PDQ will attempt to design primers for all variants. This means that for cases where two variants are in close distance, the same primer pairs (or very similar pairs in terms of position) may be generated multiple times.

The reference genome used to make the VCF should be selected. Chromosome names much match for the program to work.

CHROMPOS

This mode takes a six column file as input. It is essentially a manual version of VCF mode. The columns of the input file are:

Name Chromosome
Input_Starting_Position,Target_Sequence_Position,Target_Sequence_Length

Unlike VCF mode, it is up to the user to ensure that the starting and ending positions exist on the reference genome sequence. Chromosome names much match exactly to the reference genome supplied. The output .csv is the same format as in VCF mode.

FASTA

The input for this mode is a single or multiple sequence FASTA file. FASTA files should contain all sequences on a single line. When possible sequence headers should not contain spaces (PDQ will replace spaces with _). For example, an acceptable file is:

>Sequence1
GGTTATTGGCCGTATGGC
>Sequenc2
GCCTTGACCAATTGCAGG

The following is not acceptable

>Sequence1
GGTTATTGG
CCGTATGGC
>Sequenc2
GCCTTGACC
AATTGCAGG

If your sequence is split into different lines, you can change this using the Fasta-Manipulator tool (https://github.com/bjtill/Fasta-Manipulator-FM-GUI).

In this mode the supplied sequences are directly used as input for BLAST and Primer3. PDQ calculates the length of each input sequences and sets the target region in the center with the length at 1/6 of the max amplicon size. Chromosome positions are not used. As such, the PDQ output .csv contains less information than it does from the VCF or CHROMPOS modes. The columns in the output file are:

Name, LeftPrimer, LeftPrimer_Tm, RightPrimer, RightPrimer_Tm, Product_Size, InputSequence, AmpliconSequence, Number_BlastHits_Input_Pident, Number_BlastHits_Amplicon_Pident

If you want chromosome position information, please see the BLAST output file for the amplicons.

GTF:

This mode aims to facilitate the design of primers from a list of candidate genes where coding sequence is favored over non-coding sequences. The inputs are the GTF file associated with the reference genome and a list of candidate genes that exactly match gene names in the GTF (case sensitive). The program extracts gene model information for each provided gene name and then searches for either the largest annotated exon, or if none are annotated, the CDS, for primer design. The output .csv file is similar to that from VCF mode or CHROMPOS mode with added information about region used for primer design. The columns are:

Name, LeftPrimer, LeftPrimer_Tm, RightPrimer, RightPrimer_Tm, Product_Size, InputSequence, AmpliconSequence, Number_BlastHits_Input_Pident, Number_BlastHits_Amplicon_Pident, Chrom, InputStart, InputEnd, GeneModelInfo

GeneModelInfo is take from the GTF file (for example, exon_number_2).

REFERENCES:

1. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012 Aug;40(15):e115.

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.