Using Bioinformatics to Identify Promoters in Genome Sequences

Carlos M. Hernandez-Garcia and Gabriel Abud

Department of Horticulture and Crop Science, OARDC/The Ohio State University, 1680 Madison Ave., Wooster, OH 44691, USA

hernandez-garcia.1@buckeyemail.osu.edu

THE OHIO STATE UNIVERSITY

OARDC

Purpose of this tutorial

- Provide step-by-step instructions to automatically identify DNA sequences in a genome sequence using a BioPerl script

Rationale

- With the continual release of plant genome sequences, the accumulation of genomics information has been exponentially increasing

- Automated analysis and mining of genome databases is becoming essential and routine due to the volume of data

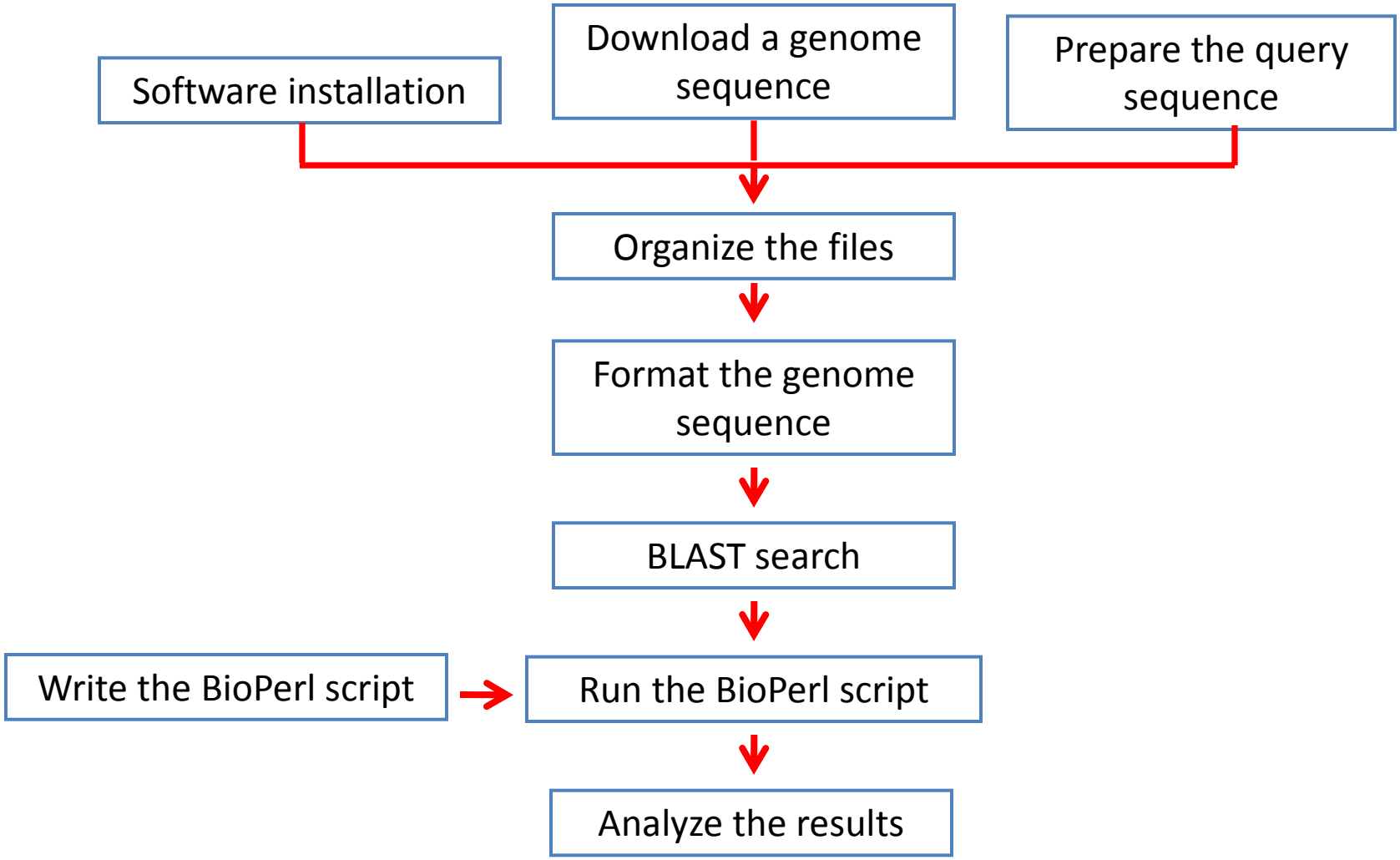
Biological rationale for targeting promoters

- Promoters are DNA sequences located in front of gene coding sequences
- Due to their influence on gene expression, promoters are of interest as tools to regulate Genetically Modified Organism (GMO) products and as markers for natural variation
- Promoters with different functionality in terms of inducibility, tissue specificity, strength etc. are needed
- In this tutorial, we will focus on the identification of ubiquitin promoters of tomato which drive strong gene expression in important crops such as maize (Christensen and Quail 1996), rice (Sivamani and Qu 2006), potato (Garbarino et al. 1995), tomato (Rollfinke et al. 1998) and soybean (Hernandez-Garcia et al. 2009)

Specific objective:

To identify *ubiquitin* genes and their promoter sequences in a draft genome of tomato using a BioPerl script in a Windows operating system

Methodology



1. Software installation

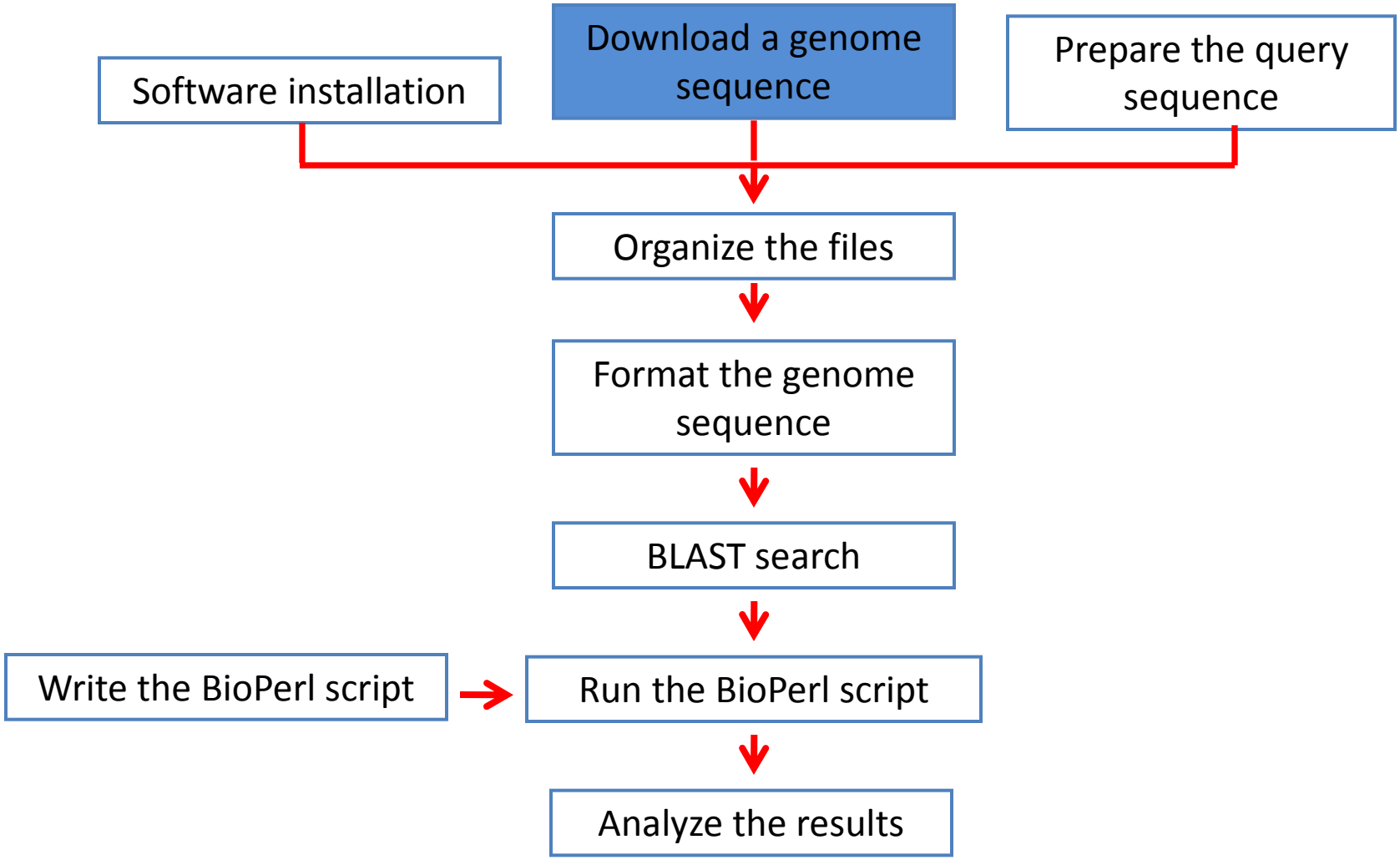
- ❑ Cygwin, UNIX emulator for Windows OS http://www.cygwin.com/

- ❑ BioPerl http://www.bioperl.org/wiki/Main_Page

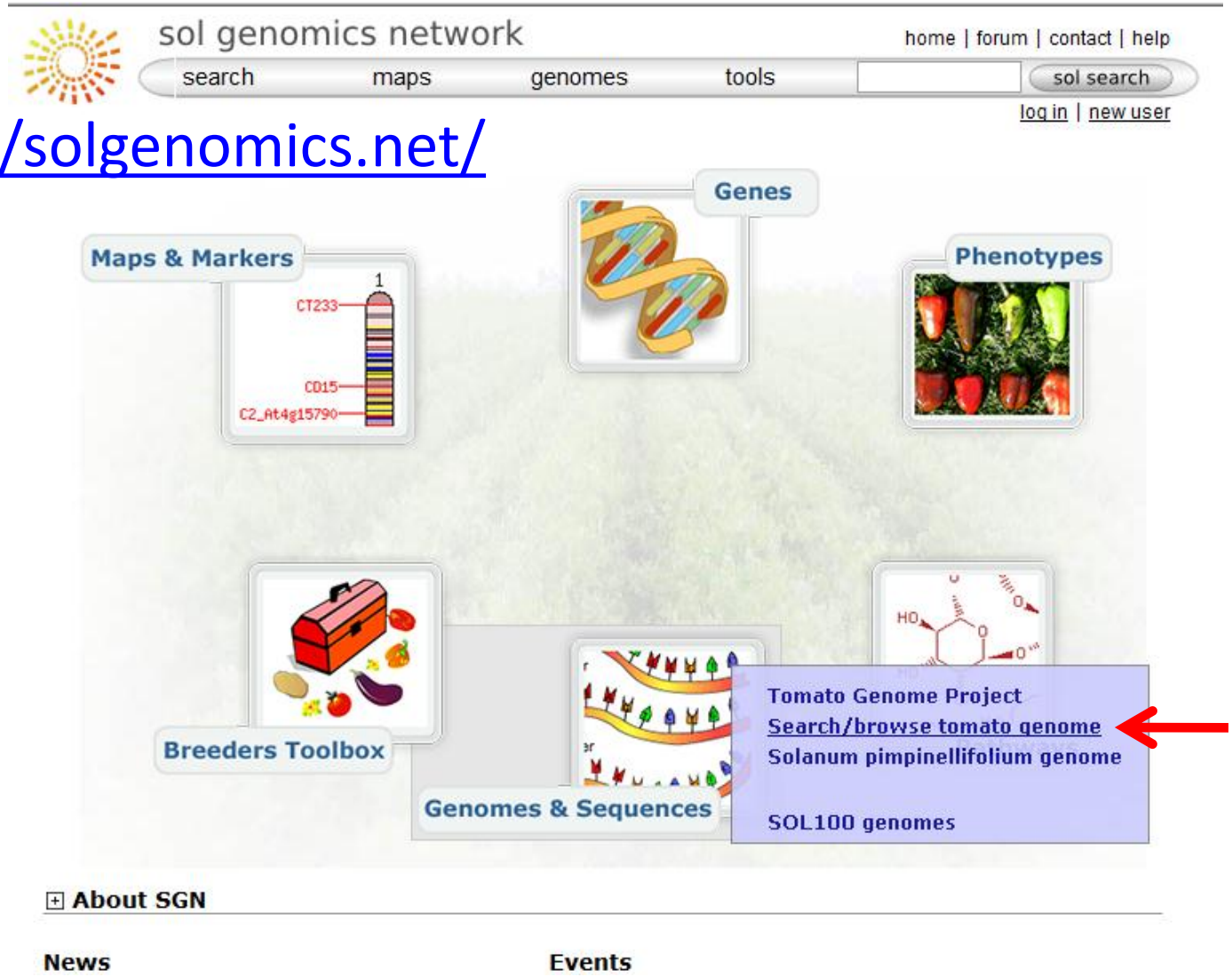
- ❑ StandAlone BLAST http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/pc_setup.html

- ❑7-ZIP compressor software http://www.7-zip.org/

Methodology



2. Download a draft genome sequence of tomato



http://solgenomics.net/

2. Download a draft genome sequence of tomato

http://solgenomics.net/

sol genomics network

home | forum | contact | help

search maps genomes tools sol search

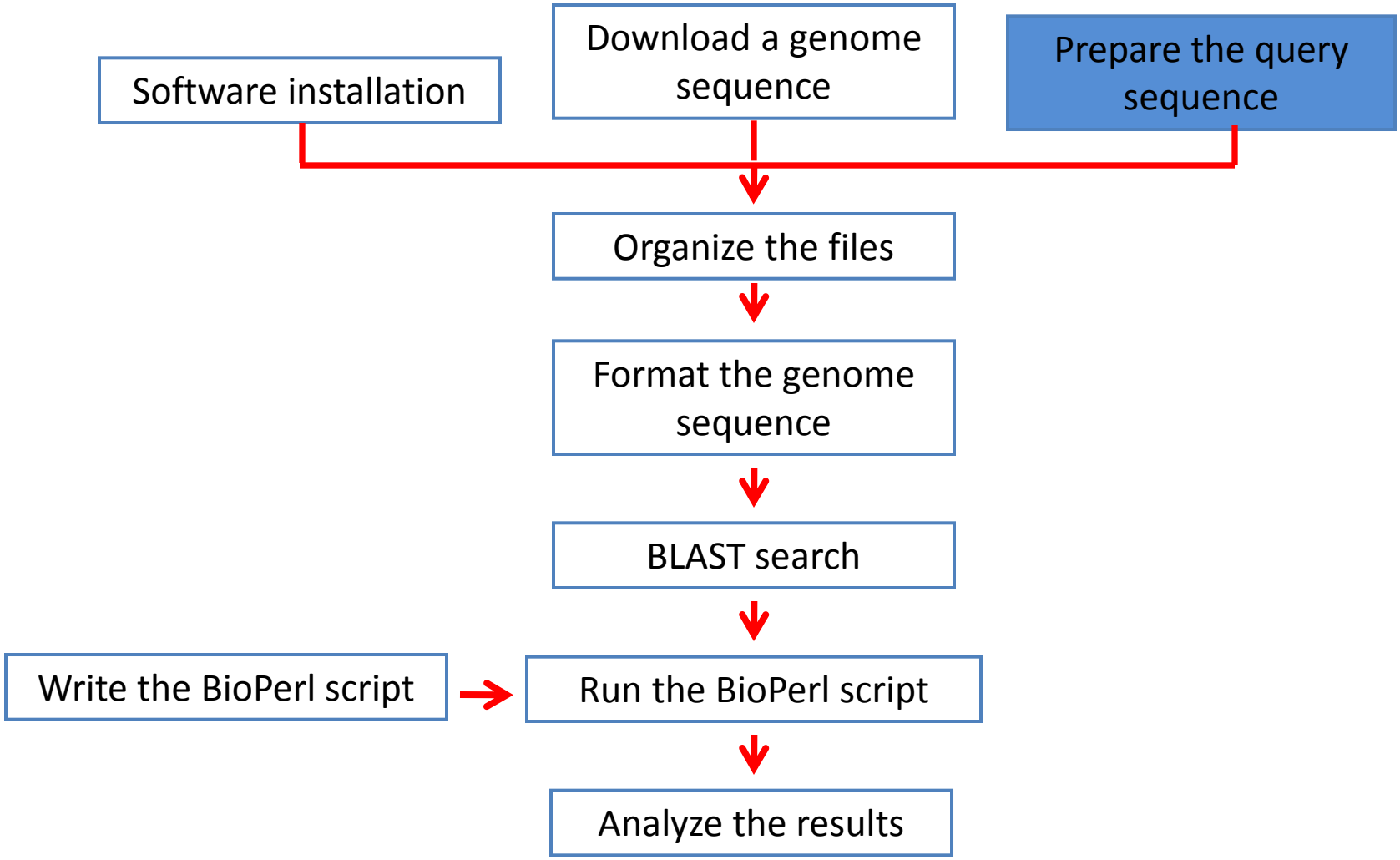
log in | new user

Tomato Genome Data

Tomato genome sequence builds

Release	Date	Description	Annotation	Download
1.00	Dec 2009	initial build, based on the Newbler assembler and containing only 454 sequencing data	ITAG1	scaffolds proteins cds
1.03	Jan 2010	like 1.00, but with additional 454 runs and improved contamination screen	Not annotated	scaffolds
cabog1.00	Mar 2010	All 454 data, bac end and fosmid end data, assembled using the CABOG assembler	Not annotated	scaffolds
1.50	Apr 2010	Includes all 454 data, bac ends, fosmid ends, polishing with Solexa and SOLiD data	Not annotated	scaffolds
2.00	Jun 2010	Release withdrawn.	Not annotated	-
2.10	Jun 2010	Additional scaffold merging using clone end sequences. Scaffolds placed and oriented using multiple physical maps, first release to include chromosome pseudomolecule sequences.	Not annotated	scaffolds, chromosomes

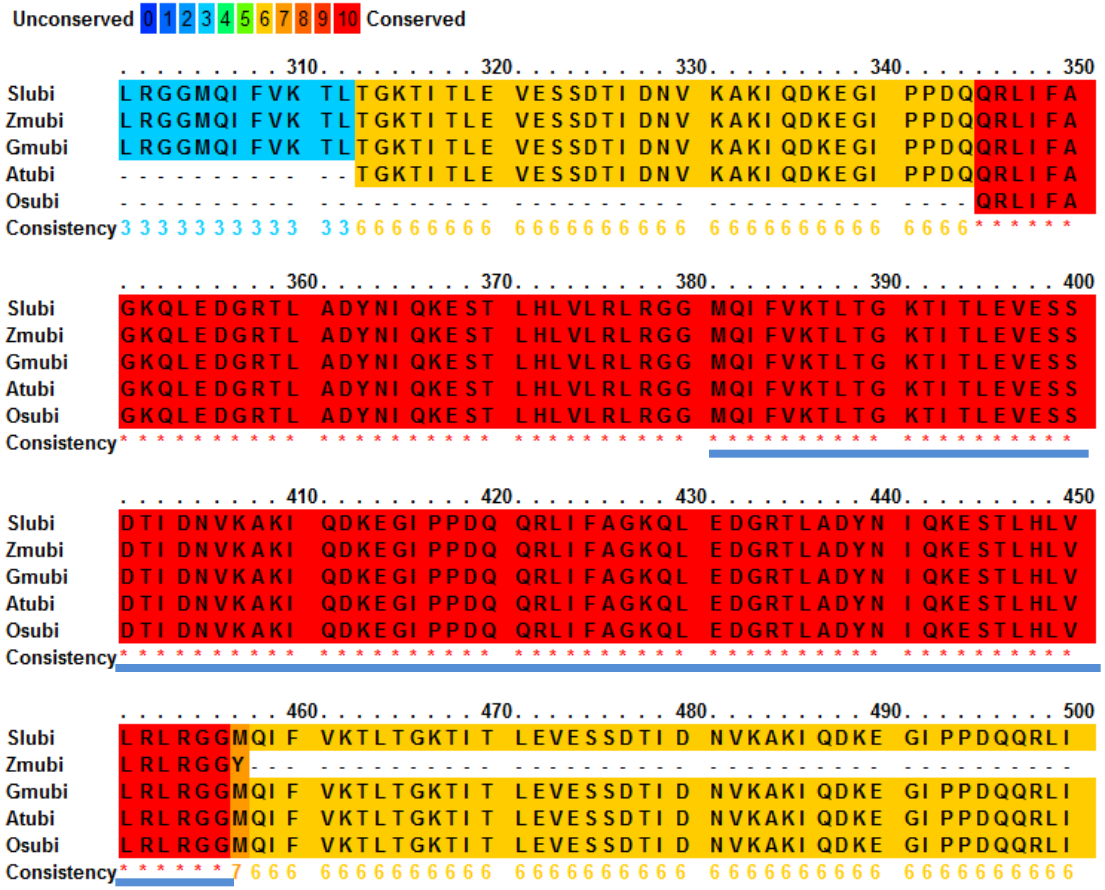
Methodology



3. Preparing the query sequence

The query sequence used in this tutorial corresponds to an ubiquitin domain present in ubiquitin proteins. Note that ubiquitin proteins are highly conserved in plants and contain single or multiple ubiquitin domains. This is what allows us to look for gene sequences in tomato using information from other plant species. The figure below depicts an alignment of ubiquitin proteins from tomato (Slubi, GenBank: CAA51679.1), maize (Zmubi, GenBank: AAC49014.1), rice (Osubi,GenBank: BAA02241.1), *Arabidopsis* (Atubi, GenBank: ABH08755.1) and soybean (Gmubi, GenBank: BAA05085.1).

Sequences were aligned using the PRALINE multiple sequence alignment tool (http://www.ibi.vu.nl/programs/pralinewww/). A complete ubiquitin domain (75 amino acid) is underlined.



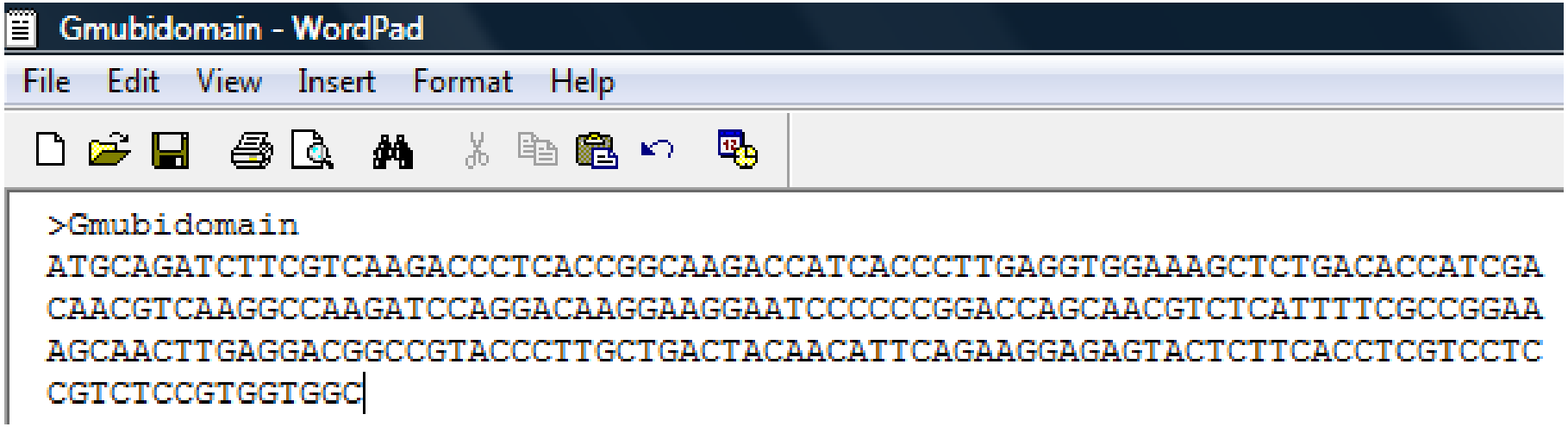
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved

```
            .........310.........320.........330.........340.........350
Slubi       LRGGMQIFVK TLTGKTITLE VESSDTIDNV KAKIQDKEGI PPDQQRLIFA
Zmubi       LRGGMQIFVK TLTGKTITLE VESSDTIDNV KAKIQDKEGI PPDQQRLIFA
Gmubi       LRGGMQIFVK TLTGKTITLE VESSDTIDNV KAKIQDKEGI PPDQQRLIFA
Atubi       ---------- --TGKTITLE VESSDTIDNV KAKIQDKEGI PPDQQRLIFA
Osubi       ---------- ---------- ---------- ---------- ----QRLIFA
Consistency 3333333333 3366666666 6666666666 6666666666 6666******

            .........360.........370.........380.........390.........400
Slubi       GKQLEDGRTL ADYNIQKEST LHLVLRLRGG MQIFVKTLTG KTITLEVESS
Zmubi       GKQLEDGRTL ADYNIQKEST LHLVLRLRGG MQIFVKTLTG KTITLEVESS
Gmubi       GKQLEDGRTL ADYNIQKEST LHLVLRLRGG MQIFVKTLTG KTITLEVESS
Atubi       GKQLEDGRTL ADYNIQKEST LHLVLRLRGG MQIFVKTLTG KTITLEVESS
Osubi       GKQLEDGRTL ADYNIQKEST LHLVLRLRGG MQIFVKTLTG KTITLEVESS
Consistency ********** ********** ********** ********** **********

            .........410.........420.........430.........440.........450
Slubi       DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV
Zmubi       DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV
Gmubi       DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV
Atubi       DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV
Osubi       DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV
Consistency ********** ********** ********** ********** **********

            .........460.........470.........480.........490.........500
Slubi       LRLRGGMQIF VKTLTGKTIT LEVESSDTID NVKAKIQDKE GIPPDQQRLI
Zmubi       LRLRGGY--- ---------- ---------- ---------- ----------
Gmubi       LRLRGGMQIF VKTLTGKTIT LEVESSDTID NVKAKIQDKE GIPPDQQRLI
Atubi       LRLRGGMQIF VKTLTGKTIT LEVESSDTID NVKAKIQDKE GIPPDQQRLI
Osubi       LRLRGGMQIF VKTLTGKTIT LEVESSDTID NVKAKIQDKE GIPPDQQRLI
Consistency ******7666 6666666666 6666666666 6666666666 6666666666
```

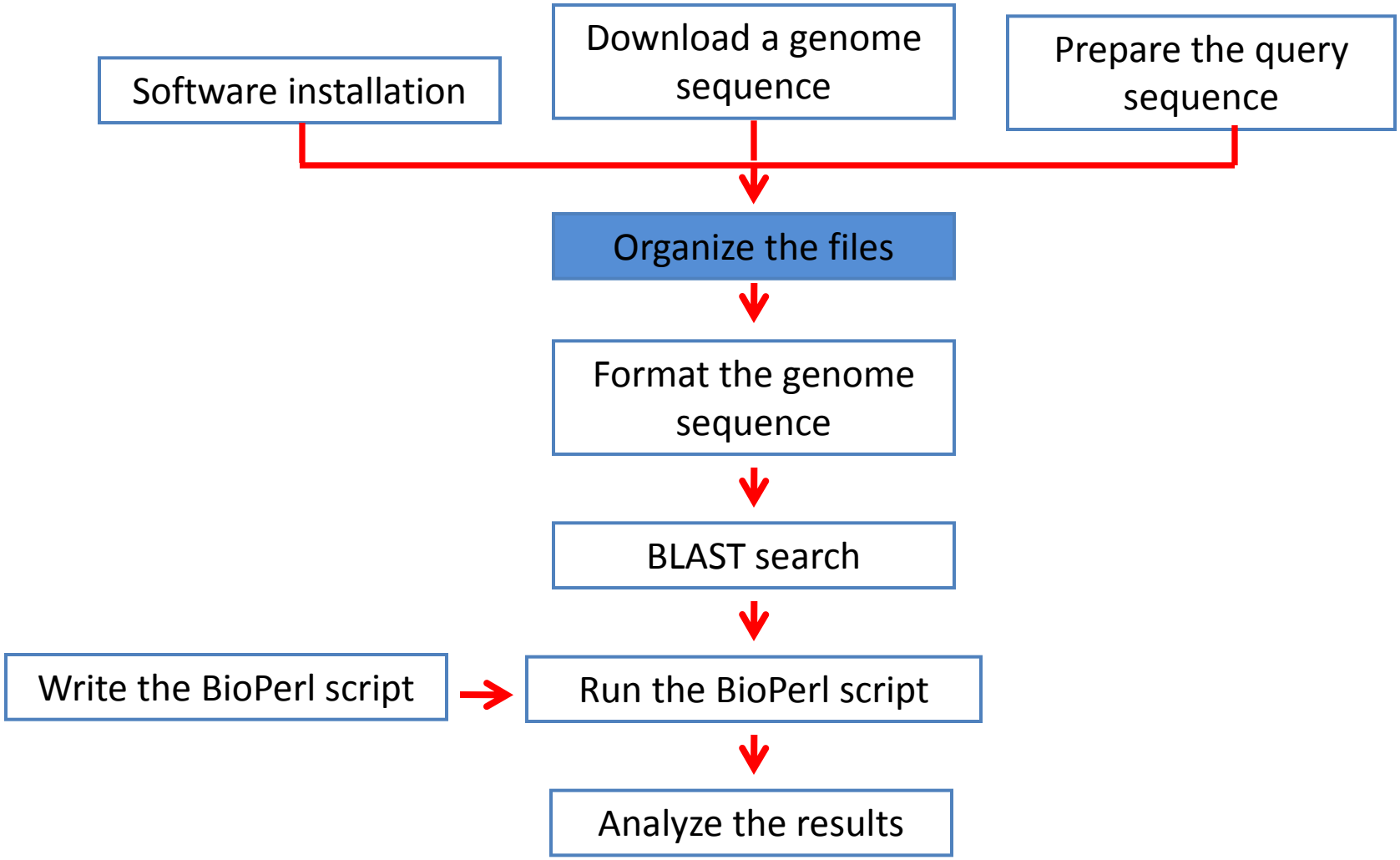
3. Preparing the query sequence

For simplicity, the nucleotide sequence from soybean corresponding to the 76-amino acid domain from ubiquitin was used as the query sequence.

The nucleotide sequence, in the fasta format, was placed in a ".txt" file using a text editor.



Methodology



4. Organizing the files

Decompress the genome sequence file with any compressor software (e.g. 7-ZIP compressor)

Place your decompressed genome sequence file (***SL2.10sc***) and the query sequence file (***Gmubidomain***) in a new created folder (e.g C:\cygwin\home\Carlos***Solanum_Lycopersicum***)

5. Format the genome sequence before using BLAST

Open UNIX (In our case, we are using Cygwin emulator for PC)
Go to **Carlos@Carlos-PC ~/Solanum_Lycopersicum**
Type ***formatdb.exe –i SL2.10sc.fasta –p F***
–i indicates the input file to format into a searchable database
–p asks if the input data is protein sequence
F indicates "false", which specifies a nucleotide database

```
Carlos@Carlos-PC ~
$ cd Solanum_Lycopersicum

Carlos@Carlos-PC ~/Solanum_Lycopersicum
$ formatdb.exe -i SL2.10sc.fasta -p F
```

5. Format the genome sequence before using BLAST

Output files

Name	Date modified	Type	Size	Ta
SL2.10sc	6/25/2010 8:53 AM	FASTA File	772,567 KB	
formatdb	8/15/2010 9:20 PM	Text Document	1 KB	
SL2.10sc.fasta.nhr	8/15/2010 9:20 PM	NHR File	259 KB	
SL2.10sc.fasta.nin	8/15/2010 9:20 PM	NIN File	41 KB	
SL2.10sc.fasta.nsq	8/15/2010 9:20 PM	NSQ File	197,554 KB	
Gmubidomain	8/5/2010 3:38 PM	File	1 KB	
SL_Gmubi	8/5/2010 3:40 PM	OUT File	40 KB	
find_blast_matches	8/5/2010 4:26 PM	PL File	8 KB	

A correct format will create three files (NHR, NIN and NSQ files). In addition, you will see a report for completion in a TXT format.

6. BLAST the query sequence against the genome sequence

Open Cygwin
Go to **Carlos@Carlos-PC ~/Solanum_Lycopersicum**
Type ***blastall.exe –p blastn –d SL2.10sc.fasta –i gmubidomain –o SL_Gmubi.out***

-p is the name of the program to use (***blastn***)
-d specifies the file for the genome sequence (***SL2.10sc.fasta***)
-i indicates the query sequence file (***gmubidomain***)
-o indicates the name of the output file that will be created (***Gmubi.out***)

7. Checking the output file

In this file, we expect to see a list of sequences with significant alignments and their e-values, and the actual alignments between the query sequence and the subject genome

SL_Gmubi - Notepad

File Edit Format View Help

```
BLASTN 2.2.20 [Feb-08-2009]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= Gmubidomain
         (228 letters)

Database: SL2.10sc.fasta
           3433 sequences; 781,288,557 total letters

Searching..................................................done


                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

SL2.10sc05010                                                       180   6e-044
SL2.10sc04626                                                       180   6e-044
SL2.10sc05925                                                       172   1e-041
SL2.10sc05380                                                       125   3e-027
SL2.10sc03748                                                       103   1e-020
SL2.10sc04323                                                        72   4e-011
SL2.10sc03806                                                        64   9e-009
SL2.10sc06101                                                        60   1e-007
SL2.10sc05732                                                        52   4e-005
SL2.10sc03923                                                        50   1e-004
SL2.10sc04813                                                        46   0.002
SL2.10sc06214                                                        40   0.13
SL2.10sc05632                                                        36   2.1
SL2.10sc04126                                                        36   2.1
SL2.10sc03796                                                        36   2.1
SL2.10sc04777                                                        34   8.3
SL2.10sc03714                                                        34   8.3
SL2.10sc03685                                                        34   8.3

>SL2.10sc05010
          Length = 20453289

 Score =  180 bits (91), Expect = 6e-044
 Identities = 193/227 (85%)
 Strand = Plus / Minus

Query: 1        atgcagatcttcgtcaagaccctcaccggcaagaccatcacccttgaggtggaaagctct 60
                ||||||||||||| || ||||| || || ||||| |||||||| ||||| ||    ||
Sbjct: 13588607 atgcagatcttcgtgaaaaccctaacggggaagacgatcaccctagaggttgagtcttcc 13588548


Query: 61       gacaccatcgacaacgtcaaggccaagatccaggacaaggaaggaatcccccggaccag 120
```

8. BioPerl script

The BioPerl script used in this tutorial (provided as a .txt file, do not forget to change the file extension to .pl):

-Parses the output blast file against the genome sequence file to identify the sequences with the highest similarities with the query sequence

-Extracts the promoter sequences for those genes

-Allows the adjustment of e-values and the length and the number of extracted promoters

find_blast_matches - WordPad

File Edit View Insert Format Help

```
#!/usr/bin/env perl
#-----------------------------------------------------------------------
--------------------------------#
# Made by Gabriel Abud
#
# This script takes a blast output file and a contig or genome file as
arguments (in that order).          #
# It then parses through the blast file to check for good matches
                          #
# Those matches are then used back into the genome file, and the
sequence in the genome              #
# is displayed (to STDOUT)
                  #
# To redirect the output to a file, use the Unix '>' operator
                                  #
# Note: If you want to see the output in a Unix shell, please be sure
to maximize the terminal      #
# before hand.
                  #
#
                  #
# Changes:
            #
# Optional flags for maximum e-value, amount of hits shown, and
promoter region were added              #
# Use --help for more information
#
#-----------------------------------------------------------------------
--------------------------------#
use warnings;
# Modules
use lib '/cygdrive/c/Perl/site/lib';
use lib '/cygdrive/c/Perl/site/lib/Bio';
use Bio::SearchIO;
use Bio::Seq;
# use File::Basename;

# Variables
my $line;
my @scaffolds;
my $inputFile;
my $scaffoldFind;
```

Main components of the BioPerl script used

The command --***help*** displays a brief description of the main parameters that can be adjusted:

-e to modify the e-value (the default is 0.01)

-p to specify the promoter length

-n to specify the number of promoters (the default is one promoter at a time)

The desired values are entered after the specification of each parameter

```
52  # Help screen
53  if( defined($ARGV[0]) && "$ARGV[0]" =~ /^-?-?help$/i ) {
54      print "\n$baseProg:\n\n\n";
55      print "Syntax:\n\t$baseProg  blast_output_file  contig_or_genome_file  [OPTIONS]\n\n";
56      print "Options:\n";
57      print "\t-e, maximum e-value for matches (0.01 by default)\n\n";
58      print "\t-p, base pairs of promoter region to be included (should only be used in DNA sequences)\n\n";
59      print "\t-n, number of top hits to display, starting with the highest hit (1 by default)\n\n";
60      exit
```

Main components of the BioPerl script used

Commands to enter to execute the script

```
# Checks for correct amount of arguments
if( @ARGV < 2 ) {
    print STDERR "USAGE: $baseProg  blast_output_file  contig_or_genome_file  [OPTIONS]\n";
    print STDERR "Type '$baseProg --help' for more information\n";
    exit;
}
```

The database

The blast file

The parameters: -e –p -n

The script's name

Main components of the BioPerl script used

Parameters specified by default

E-value

Number of promoters extracted

```
69
70  # Defaults
71  $e_value = 0.01;                    # Default e-value (if none specified)
72  $matches = 1;                       # Default # of matches printed
73  # $promoter is undefined by default
74
```

Users must specify
the promoter length

Main components of the BioPerl script used

This script is able to extract promoters in the + and – strands. This is done by getting the reverse complement of - strands

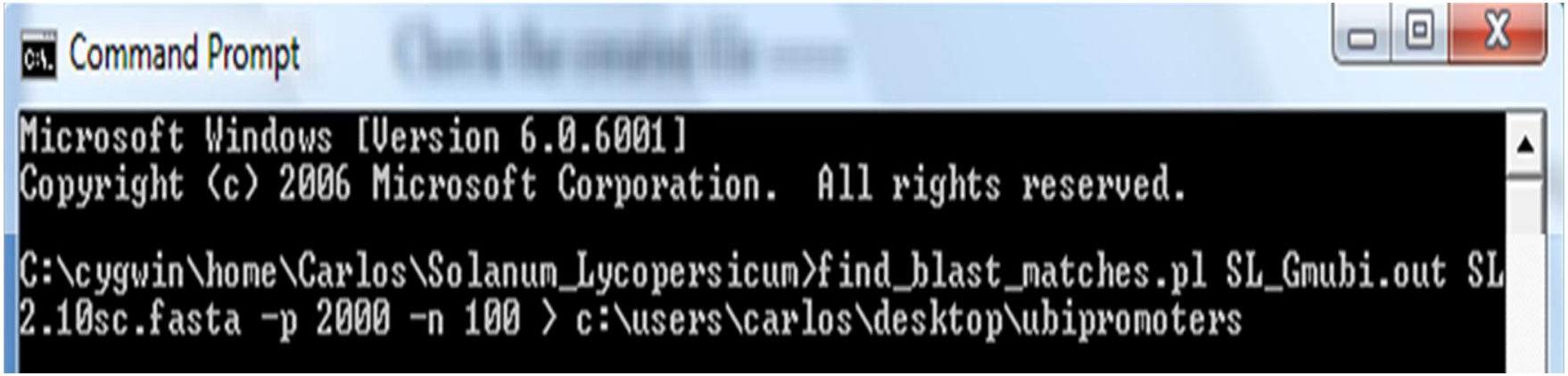
```
# Checks to see if the hit sequence is the reverse compliment
# If it is, it changes it so that it matches the query sequence
if( $strand[$m] == -1 && defined($promoter) ) {                # If a promoter region was specified AND the strand is +/-
    print " Promoter = $promoter";
    print " (showing the reverse compliment)";             # Lets you know that this is the reverse compliment of the real sequence
    $total_size = length($seq_obj->seq);
    if( ( $promoter + $end[$m] ) > $total_size ) {             # If promoter is bigger than the contig, print an error and don't display promot
        print " !!Promoter is too big for seq!";
        print STDERR "ERROR: Promoter is too big (Max promoter = ", $total_size - $end[$m] , " )\n";
        print STDERR "Showing sequence without promoter region...\n";
        $baseList = $seq_obj->subseq($start[$m], $end[$m]);
    }
```

9. Run the BioPerl script

Now we need to extract the upstream sequences (from our genome sequence) for genes with high similarities identified during the previous Blast procedure.

-Go to Command Prompt
-Locate the main folder containing all the files (Solanum_Lycopersicum folder)
C:\cygwin\home\Carlos\Solanum_Lycopersicum>
-Type: ***find_blast_matches.pl SL_Gmubi.out SL2.10sc.fasta –p 2000 –n 100 > c:\users\Carlos|Desktop\ubipromoters***

-p is promoter length (2000 bp)
-n 100 pulls out up to 100 promoters if present (the default is 1 promoter)
The default e-value is 0.01
">" indicates the creation of an output file (txt) at any specified location



10. Analyzing the output file

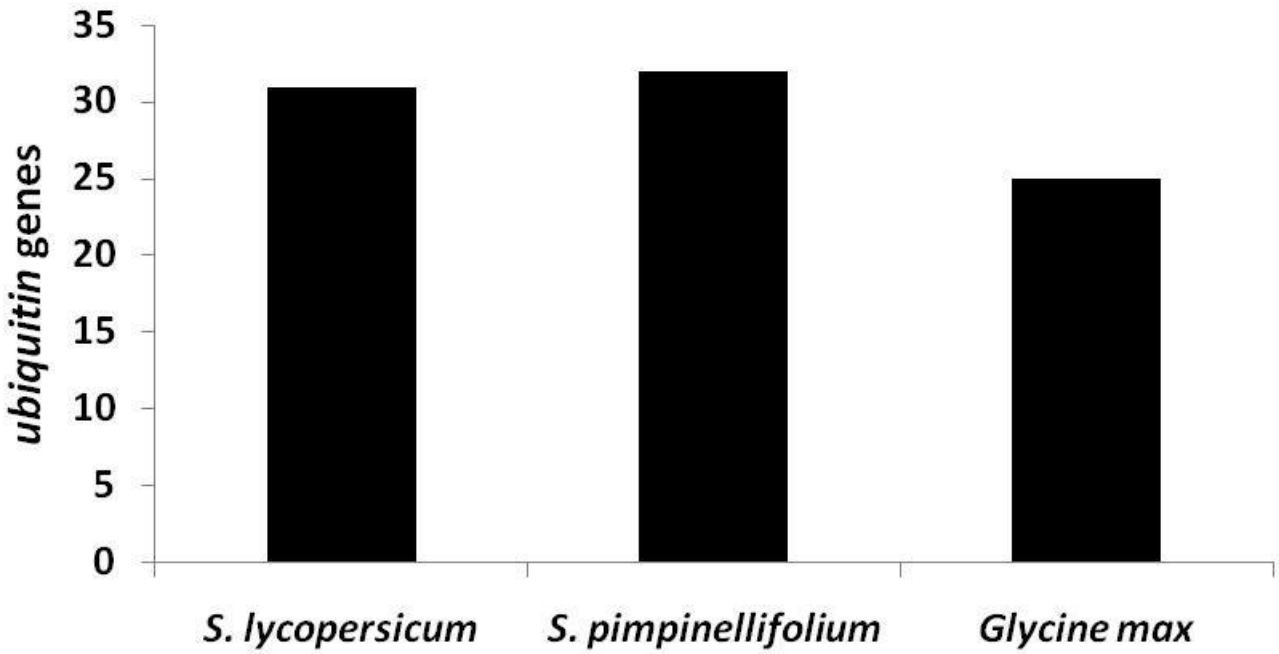
- The output file contains the promoter sequences identified in our target genome
- 31 ubiquitin promoters were found in the draft genome sequence of tomato
- Each promoter is extracted from an *ubiquitin* gene with at least one ubiquitin domain, therefore, 31 *ubiquitin* genes were identified in this draft genome of tomato

ubipromoters - Notepad

File Edit Format View Help

```
>SL2.10sc05925  (SL2.10sc.fasta) at 4246810 - 4247037 Promoter = 2000
TCAAGAGTTTTTTCATACACAACACTTTAGGCTCAAAATCAAGATCTTTGATTAAGAAAATAATAGTTTC
GTCTACAACCCACTATCATTTAGTGCTCAAATATTGCTTCTTTTTCTTCTTATCAGTTGAAGTTAGATTT
TTCAGTTTAAAATTTTAGAACATCGTGCAATTGAGTCCATTTTATTCAATTCAAAGTTTAAGATTTTTGA
CTATAGCATCGTAAAGTTTAATTTTTTAAGTTTATACCTTAGAGATATTGTATCCTAAAATCGAGTTGTG
TGATTCGAACTTACTATTGGTAGAATAATTATTTTGATTAACCATCAGATGAAGTTGCCCCTTTCATAAC
ACATATCACGATTAAGTACGAAAAATATTTCCTAACATAATTTAAATTAGTCGAATTCTAACGAATATCA
CACCTAAATAGATCTCACGTGATAGAAATTTGACTTAGTCAGACTCCAATACTCGCTCAAAAATAAAATA
AAATAAAATAAATTGATAATTATGTCTCAACTACTTTGGTTGGATTTGAATATTGTTGTTGTTGGATTAT
TATATAGGAAGTTGCCATCCAAACGGCACCTAAATTTTCTCCACCGATCAAAGAAAAGGAGACGTGTCAT
ACTATTATTGGTTACTTCATTGGCAATTTCTTCACAAATTGGCCATTAACAACTATAAAAGGAGGCTCTC
TGTAACCCCAATTCATTCATTGATTTCTCATCTCTTCAAATTCTTCAGAAAAAAAACCTCTCAAGGTATA
CTTTTCCTTCTTTTTAGTTTTTTTGTTTATTTTTTCTATATTTTGTGCTATTTGTTTTGTTGTGTCTGTG
TTTAGAATCAGAATTCTGTTTTAAAAATTCAATTTTTATGTTGTTTGAGTGCTGTGTATGCTTTAATTTT
CTGTATTTTTTATTGTGTCTGTGTTTAGAATCAGAAAGGGTGTTTGAATCTGTTTTAAAAATTCAATGTT
TATGTTGTTTTAGTGTTCTGTATGCTTTAATTTTCTGTATTTTTTTGCTATTTGTTTATTGTGTCTGTGT
TTGGAATCAGAAAGGGTGTTTGAATCTGTTTTAAAAATTCAATCTTTATGTTGTTTGAGTGTCTGTATGA
TTTAATTTTCTGTATTTTGTGCTATTTGGTTTATTGGGTATCTAAAAAAGTTGGATCTTGATTTTGTTTT
CAATATTCTGTATGCTTTAGTTTTCTGTTAATTGTGTTATTTGTTGTGTTTGGGTCTGGTTTTGAATCAG
AAAGGGCTTTTTAATTTTGTTCTAAAAGTTAAATCTTGATTTTGTTTCAACCTGGTGTGTGCCTTAGGTT
TCTGTATTTATGGACTTGGAAACCTTTTAATGATCTAAAAAGATTGTATTTTAGTTTTGTTTATGAGCTG
AAAAACCATACAAGTATGATCTGAAAAGTTGAGTCTTGATTTTGACTCAACTTCTGTATGTGTTGTTTTG
GTTATATGATAGCCGCTGCTTCATTTCGTTTTCTGTTTTTTTTTTGTTTGTGTGAGCTGAGTTTCTATAG
CAATTTCGCATACACCCTCTCCGTACTCCACTTGTGGGATCACTTGCACAATCTATATTGTTGTTGTTGT
ATTTTTGACCTGTGTGTTTCTGTTGTTTGTTTGTTTATGATATAAAGAACCTTCCTATTTCTATTTGTAA
GAACTCAATCTTGATTTTGGTTTCAATTTTTTGTTGATGCACCAAGTTGAATTTTGTGTGGTTTTTAAAA
TAGCCTCTGCTTATCCCTTTAAATTTCTGTTATTTGTGTTGTTTGGTCTGTGTTTATGATCTAAAGCCTG
TCAATTATTTTGTTTTGAAAAAATTGGAAGCCTTCTTATTGGATGATGATTATGAATGGTGTTCTGTTGG
GTTTTGATTTGTGATATGAAATATCAACTTTAATCTGTCATCTCGTTTTATAGAGTCATGTCTGTGTCTA
TTCGAGTATGTGTGCTTATTTTGTTGAATTTTGTGTACAGATGCAGATATTTGTTAAAACCCTCACTGGA
AAGACTATCACCCTTGAGGTGGAAAGCTCAGACACCATTGACAATGTTAAAGCCAAGATCCAGGACAAGG
AAGGCATTCCCCCAGACCAGCAGAGGCTGATCTTTGCAGGAAAGCAGCTTGAAGACGGTCGTACTCTAGC
TGATTACAACATTCAGAAAGAGTCAACTCTCCACTTGGTGCTCCGTCTTCGTGGTGGC

>SL2.10sc05925  (SL2.10sc.fasta) at 4247068 - 4247228 Promoter = 2000
TGTATCCTAAAATCGAGTTGTGTGATTCGAACTTACTATTGGTAGAATAATTATTTTGATTAACCATCAG
ATGAAGTTGCCCCTTTCATAACACATATCACGATTAAGTACGAAAAATATTTCCTAACATAATTTAAATT
AGTCGAATTCTAACGAATATCACACCTAAATAGATCTCACGTGATAGAAATTTGACTTAGTCAGACTCCA
ATACTCGCTCAAAAATAAAATAAAATAAAATAAATTGATAATTATGTCTCAACTACTTTGGTTGGATTTG
AATATTGTTGTTGTTGGATTATTATATAGGAAGTTGCCATCCAAACGGCACCTAAATTTTCTCCACCGAT
CAAAGAAAAGGAGACGTGTCATACTATTATTGGTTACTTCATTGGCAATTTCTTCACAAATTGGCCATTA
ACAACTATAAAAGGAGGCTCTCTGTAACCCCAATTCATTCATTGATTTCTCATCTCTTCAAATTCTTCAG
AAAAAAAACCTCTCAAGGTATACTTTTCCTTCTTTTTAGTTTTTTTGTTTATTTTTTCTATATTTTGTGC
TATTTGTTTTGTTGTGTCTGTGTTTAGAATCAGAATTCTGTTTTAAAAATTCAATTTTTATGTTGTTTGA
GTGCTGTGTATGCTTTAATTTTCTGTATTTTTTATTGTGTCTGTGTTTAGAATCAGAAAGGGTGTTTGAA
TCTGTTTTAAAAATTCAATGTTTATGTTGTTTTAGTGTTCTGTATGCTTTAATTTTCTGTATTTTTTTGC
TATTTGTTTATTGTGTCTGTGTTTGGAATCAGAAAGGGTGTTTGAATCTGTTTTAAAAATTCAATCTTTA
TGTTGTTTGAGTGTCTGTATGATTTAATTTTCTGTATTTTGTGCTATTTGGTTTATTGGGTATCTAAAAA
AGTTGGATCTTGATTTTGTTTTCAATATTCTGTATGCTTTAGTTTTCTGTTAATTGTGTTATTTGTTGTG
TTTGGGTCTGGTTTTGAATCAGAAAGGGCTTTTTAATTTTGTTCTAAAAGTTAAATCTTGATTTTGTTTC
AACCTGGTGTGTGCCTTAGGTTTCTGTATTTATGGACTTGGAAACCTTTTAATGATCTAAAAAGATTGTA
TTTTAGTTTTGTTTATGAGCTGAAAAACCATACAAGTATGATCTGAAAAGTTGAGTCTTGATTTTGACTC
AACTTCTGTATGTGTTGTTTTGGTTATATGATAGCCGCTGCTTCATTTCGTTTTCTGTTTTTTTTTTGTT
```

Comparison of *ubiquitin* genes from other plants



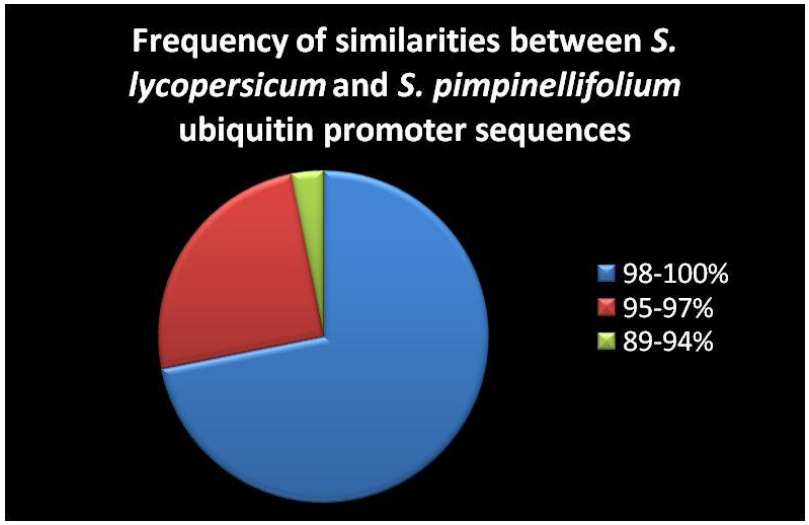
An additional analysis of a draft genome of *S. pimpinellifolium* using the same methodology revealed 32 *ubiquitin* genes. A previous study in soybean showed 25 *ubiquitin* genes (Hernandez-Garcia et al. 2010b). These results confirm that our methodology is robust and likely extracted most of the *ubiquitin* genes present in the genome of tomato.

Homology between ubiquitin promoters from *S. lycopersicum* and *S. pimpinellifolium*

- ❑ Similarities/differences between these two promoter groups would give insights into differential gene expression regulation

- ❑ The identified ubiquitin promoters from *S. lycopersicum* were Blasted against the ubiquitin promoters of *S. pimpinellifolium*

Sequence similarities of promoters from both species ranged from 100% to minor bp substitutions, including short INDELs




```
Query: 855   gctagagttgcggttgagttgaagacgatggcacgtttgctgcactccacaaataacaaa 914
             |||||| || ||||||||||||||||   | ||||||||| || ||||||||||||||||
Sbjct: 181   gctagatttacggttgagttgaagacgtcagtacgtttgctacagtccacaaataacaaa 240

Query: 915   gaagaaaacataaaagtagggggtcagtacaaacacaggtactgagtaggtatcatcggc 974
             |||  ||||||| || |||||| | ||||||||||||| ||||||||||| ||||||||||
Sbjct: 241   gaaagaaacatacaactaggggtt-agtacaaacacaagtactgagtagatatcatcggc 299

Query: 975   taactcaaaatagaaaacagtatatatcagataacatcataaaatcaactaaaatactca 1034
              ||||||||||||||||||||||||||||||||||||||||||| |||| || || |
Sbjct: 300   caactcaaaatagaaaacagtatatatcagataacatcataaaatcagctaatatcctta 359

Query: 1035  acatgcagcattttcaattaccataacccttggtcataacaccaagctcatcaacgagga 1094
              ||||||||||||| | |||||||| ||||||||||| |||||||||| |||||| |||||
Sbjct: 360   gcatgcagcatttaccattaccatcacccttggtcacaacaccaagcacatcaatgagga 419

Query: 1095  ctcacgcctcctcatcatactcatttgggaattaggttcattagattgaatatattaaca 1154
             |||||||||||||||||||||||||||||| ||||||||||| ||||| ||||||||||
Sbjct: 420   ctcacgcctcctcatcatactcatttgggaactaggttcattaaattgagtatattaaca 479

Query: 1155  tctttcaagattcattttctttattcctctcatgtcggtacgtgacactccgctcctcaa 1214
             ||| ||||||||||||| |||||||||||||||||| ||||||||||||||| ||||||| |
Sbjct: 480   tctctcaagattcattatctttattcctctcatgttggtacgtgacactctgctcctc-a 538

Query: 1215  tatactatcctcgtgtcagaacgtgacactctgatcctcattctatcctggtgtcgaaat 1274
             |||||||| || ||||| |||||||||||||| |||||||||||||||||||||||||| ||
Sbjct: 539   tatactattctggtgtcggaacgtgacactccgatcctcattctatcctggtgtcggaac 598

Query: 1275  gtgacacccgatccatattctatcatggtaccggaacgtggcacccgatctatatactat 1334
             ||||||| ||||||||||||||||| |||||||||||| |||||||| ||| |||||||||
Sbjct: 599   gtgacactcgatccatattctatcttggtaccggaacatggcacccaatccatatactat 658

Query: 1335  cctggtgtcgaaacgtgacactccgatcctc--attctatcctggtgtcggaacgtgaca 1392
             ||||||||| || || |||||||||||||||  ||  |||||||||||||||||||||||
Sbjct: 659   cctggtgtcggaaggtaacactccgatcctcatatattatcctggtgtcggaacgtgaca 718

Query: 1393  c-ccgatc--catattctatcctggtaccggaatgtggcacctgatccgtatactatcct 1449
             | | ||||  ||||| ||||||||||||| |||| |||||||| ||||| |||| ||||||
Sbjct: 719   ctcggatcctcatatactatcctggtactggaacgtggcacccgatccatatattatcct 778

Query: 1450  ggtgtcggaacgtgacac 1467
             ||||| ||||||||||||
Sbjct: 779   ggtgttggaacgtgacac 796
```

Conclusions and biological interpretations

- Our methodology is robust to identify promoter sequences of interest in genome sequences
- *ubiquitin* genes and their promoters are very conserved among plant species
- Identified promoters can be used for further analysis of natural variation, or be cloned and their expression characterized for potential application in the development of GMOs. A complete tutorial on rapid characterization of plant promoters is freely accessible at http://www.jove.com/index/details.stp?id=1733 (Hernandez-Garcia et al. 2010a)

References Cited

Christensen A. H. and P. H. Quail. 1996. Ubiquitin promoter-based vectors for high levels of selectable and/or screenable marker genes in monocotyledonous plants. Transgenic Research 5:213–218.

Garbarino, J. E., T. Oosumi, and W. R. Belknap. 1995. Isolation of a polyubiquitin promoter and its expression in transgenic potato plants. Plant Physiology 109:1371–1378.

Hernandez-Garcia, C. M., R. A. Bouchard, P. J. Rushton, M. L. Jones, X. Chen, M. P. Timko and J. J. Finer. 2010b. High level transgenic expression of soybean (Glycine max) GmERF and Gmubi gene promoters isolated by a novel promoter analysis pipeline. BMC Plant Biology 10:237.

Hernandez-Garcia, C. M., J. M. Chiera, and J. J. Finer. 2010a. Robotics and dynamic image analysis for studies of gene expression in plant tissues. Journal of Visualized Experiments 39. (Available online at: www.jove.com/index/details.stp) (verified 10 Dec 2010).

Hernandez-Garcia, C. M., A. P. Martinelli, R. A. Bouchard, and J. J. Finer. 2009. A soybean (Glycine max) polyubiquitin promoter gives strong constitutive expression in transgenic soybean. Plant Cell Reports 28:837-849.

Rollfinke, I. K., M. V. Silber, and U. M. Pfitzner. 1998. Characterization and expression of a heptaubiquitin gene from tomato. Gene 211:267–276.

Sivamani, E. and R. Qu. 2006. Expression enhancement of a rice polyubiquitin gene promoter. Plant Molecular Biology 60:225–239.

External Links

BioPerl [Online]. Available at: www.bioperl.org/wiki/Main_Page (verified 10 Dec 2010).

Cygwin [Online]. Red Hat. Available at: www.cygwin.com/ (verified 10 Dec 2010).

National Center for Biotechnology Information. BLAST+ Setup: Procedures for Windows PC [Online]. U. S. National Library of Medicine, National Institutes of Health. Available at: www.ncbi.nlm.nih.gov/staff/tao/URLAPI/pc_setup.html (verified 10 Dec 2010).

Praline Multiple Sequence Alignment [Online]. The Centre for Integrative Bioinformatics VU. Available at: www.ibi.vu.nl/programs/pralinewww/ (verified 10 Dec 2010).

Sol Genomics Network [Online]. Available at: solgenomics.net/ (verified 10 Dec 2010).

7-Zip [Online]. Igor Pavlov. Available at: www.7-zip.org/ (verified 10 Dec 2010).

Acknowledgements

The authors would like to thank Drs David Francis and Heather Merk for reviewing this tutorial and providing helpful suggestions. Input from classmates enrolled in HCS806 course (Summer, 2010) is also highly appreciated. The work on promoter analysis mentioned here has being developed in the laboratory of Dr. John J. finer (Department of Horticulture and Crop Science, The Ohio State University/OARDC). CMHG is funded by a Graduate Associateship from the Department of Horticulture and Crop Science, The Ohio State University, and partial support from CONACYT, Mexico.