Spoon

Taste your PCR primers

Spoon software simulates a PCR reaction by running an approximate-match searching analysis on user-entered primer pairs against the provided sequences and compare between *in vitro* and *in silico* PCR results. Spoon reports amplimers close to or adjoining genes/SNPs and those that are shared between *in vitro* and *in silico* PCR results in order to select the most appropriate amplimers for gene discovery. Spoon could be used for comparing physical and genetic maps, studying the primer set genome coverage for PCR-walking and NGS gaps-filling techniques. It also reports chromosomal anchored markers, which could be used for linkage and association mapping. Spoon could be used for KASP assay primer design in order to, introduce different QTL loci acquired by classical molecular markers into high-throughput genotyping to provide a major SNP markers resource for the dissection of genotypic and phenotypic variation. In addition to human-readable output files, Spoon creates Circos configurations that illustrate different *in silico* results, which will give the user the ability to merge results from different bioinformatics tools with/without slight reformatting procedures.

LICENSE and CITATION

Spoon is free open source software, both source and windows executable version can be downloaded from ############

Please cite:

Alsamman M.A, Ibrahim S.D and Hamwieh A . ######### (2018). ######.

CONTRIBUTORS

Alsamman Mahmoud Alsamman: smahmoud@ageri.sci.eg
Shafik Darwish Ibrahim: shafikdarwish@ageri.sci.eg
Aladdin Hamwieh: a.hamwieh@cgiar.org

INSTLATION

Spoon has a simple excitable exe for windows installation. For Linux and OS Spoon can be run by simple PERL compilation:

Perl Spoon.pl

GRAPHICAL USER INTERFACE

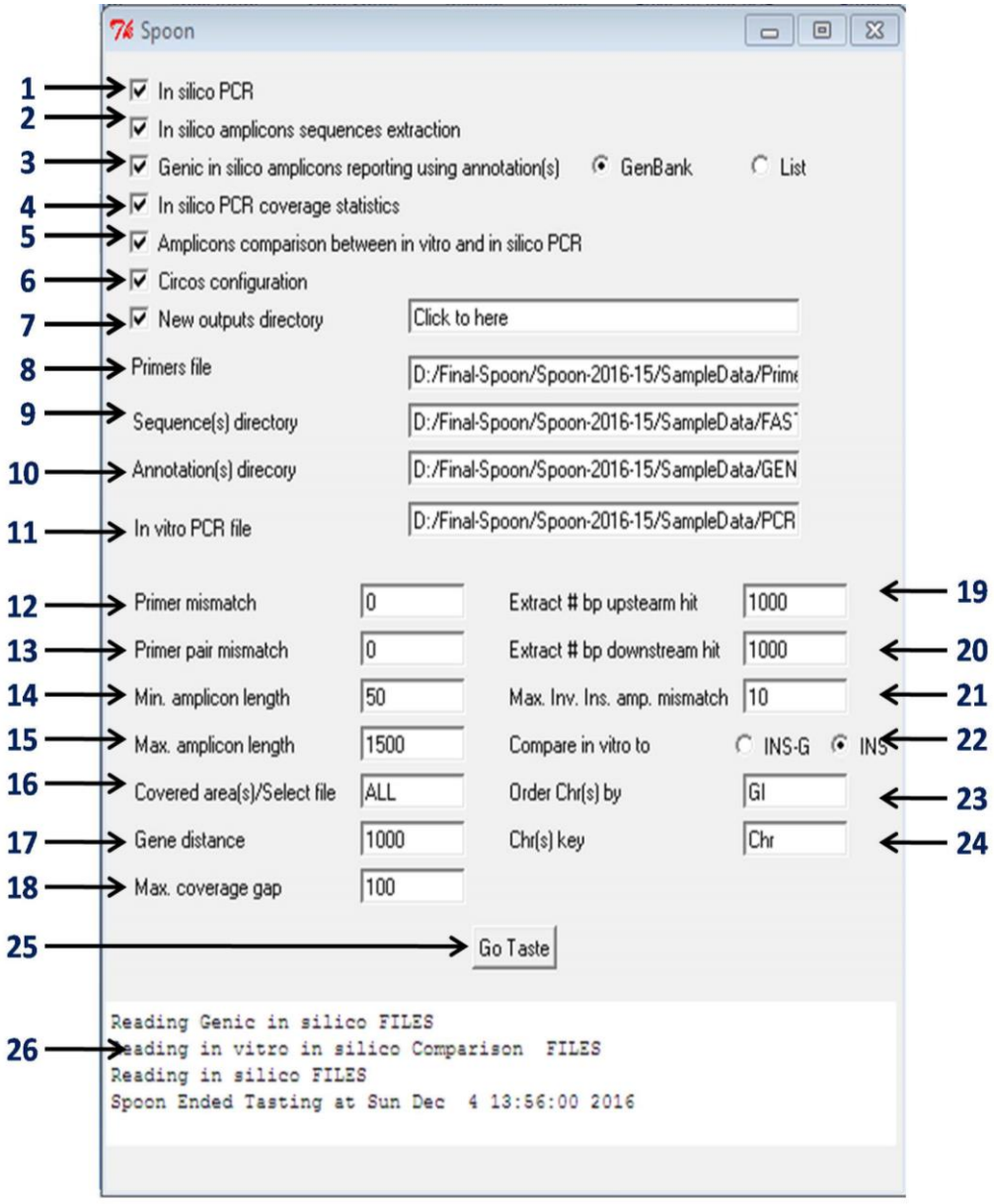


Fig. 1: Spoon graphical user interface, numbers refers exe s spoon tools that are explained in following discussion

Spoon PCR analysis

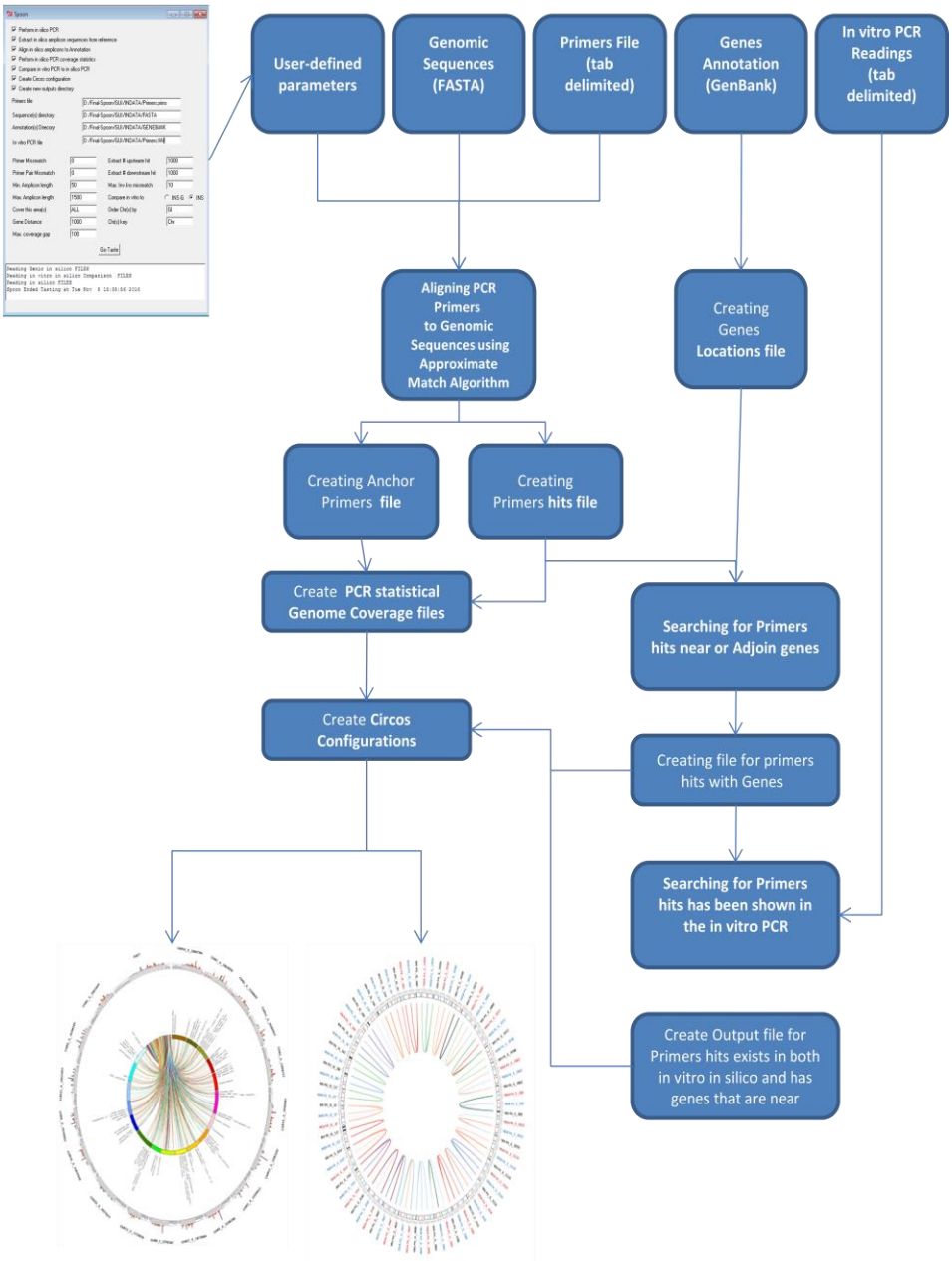


Figure 2: Flowchart illustration for Spoon *in silico* and *in vitro* analysis procedure.

Spoon can perform different kinds of PCR primers analysis procedures. It can perform *in silico* PCR procedure (1) using string approximate matching algorithm between PCR primer sequences (forward/reverse) (8) against the provided genomic sequences (9). The primer/hit mismatch values can be defined for the single primer (12), additionally the primer pair mismatch (forward mismatch + reverse mismatch) (13) can be also defined. The produced amplimer minimum (14) and maximum length (15) can be defined.

On the other hand Spoon will produce all possible amplimers even that, where one primers act as forward and reverse (single primer amplification) (**Figure 3**).

Using in silico PCR results Spoon can extract the in silico PCR amplicons sequences (2) from defined genomic sequences (9). Additionally user can specify a region in the upstream (19) and /or downstream (20) the PCR primer hit.

Additionally, Spoon can search the in silico PCR results for amplimers that are adjoin/near genes (3) by specifying the annotation files directory (GenBank format (**GB**)/gene list format **(List)**) and the distance between amplimer position on the genome and the gene (17).

Spoon can offer the ability to perform primers vs. genomic area coverage statistical analysis (4).This analysis can be performed using the overlap layout consensus algorithm with a user defined gap was (18) used to report the primer covered area compared to area covered by the primers set/primer**.** Additionally user can change the area/areas by specify an areas coverage file (specific format) (16) to specify genomic regions for this analysis (can be used for walking PCR analysis). The user defined gap is gap between two different amplimers can be collectively assumed. **Figure 4** illustrates how spoon deals with primer coverage analysis.

On the other hand user can compare the *in silico* PCR to *in vitro* PCR results (5). The in vitro PCR results must be supplied using specific format (11).Because the length of the in vitro PCR length (molecular weight) is approximately, user can define the maximum length mismatch between the *in silico* and *in vitro* amplimers (21). Additionally, user can *in vitro* results (22) to all *in silico* PCR result (**INS**) or only these amplimers that are near/adjoin genes.

Spoon uses Circos layout (Krzywinski *et al*., 2009) to visualize both *in silico* and *in vitro* PCR results (6). Chromosomes in the Circos configuration can be ordered by their **(GI)** number or by other user common number (23). On the other hand, user can change chromosomes key names to another one (24).

User can create new output result folder or select one (7). The difference between the two options that if the selected folder contains in silico PCR results spoon

can continue to perform other procedures such as PCR primers coverage but the process that produced these files must not be selected again unless it will be restarted again and if necessary all files required for next process must be selected. This option gives the user the ability to edit Spoon *in silico* PCR results and continue analysis.

By clicking on "go taste" button (25) spoon will start performing all selected processes. Spoon analysis progress can be seen using progress outputs area (26).

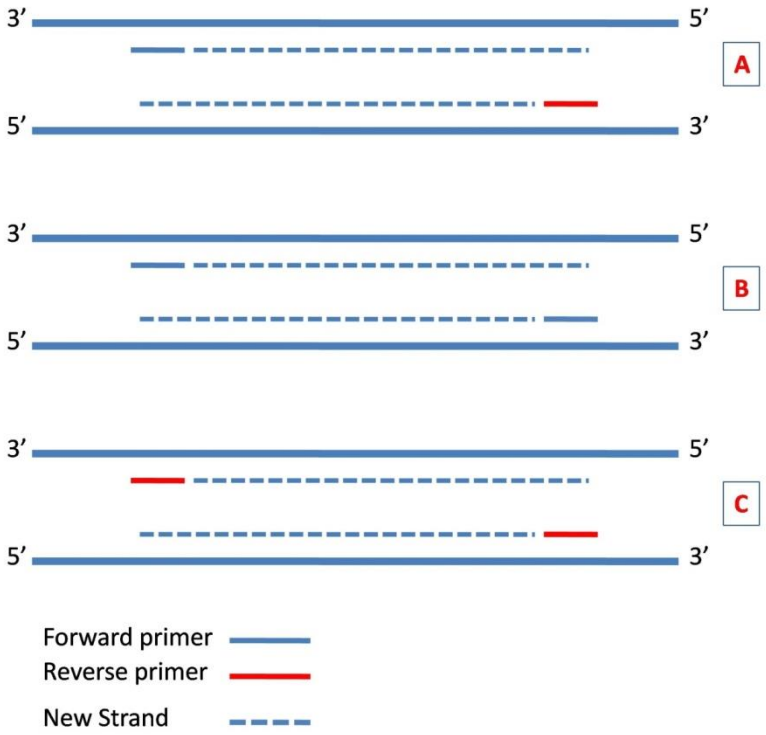
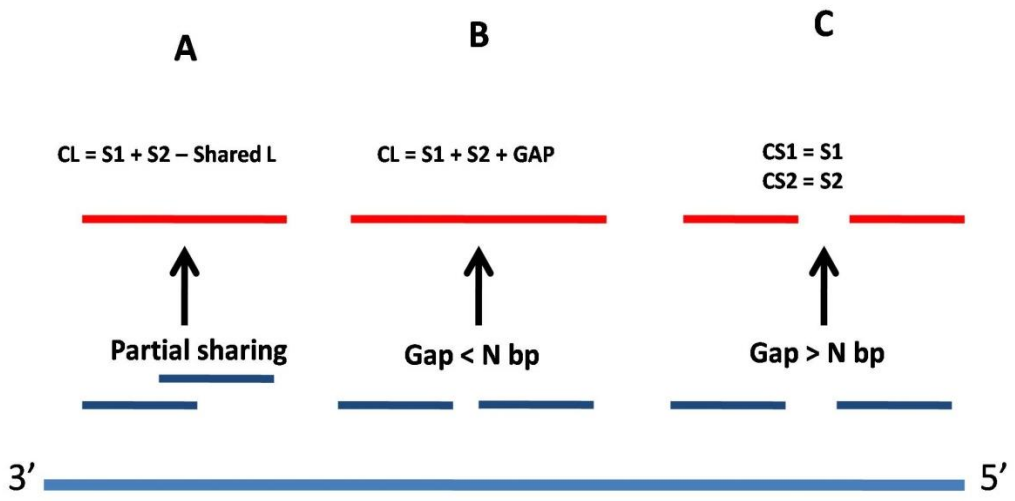


Figure 3: The different primers amplification probabilities that Spoon consider in the *in silico* PCR analysis. (A) A normal PCR amplification where both forward and reverse primers are integrated. (B) and (C) probable amplification where one primer (forward/reverse) is integrated.



CL = primer covered area length.
N = any number of nucleotides.
S2, S1 = in silico amplimers covers specific genome area.
CS1, CS1= the length of two different covered areas.

Figure 4: The different primers genomic region coverage probabilities that Spoon consider in the statistical analysis for the *in silico* PCR result. (A) The length of the covered area is an assumption for two different amplicons shares a common genomic area. (B) The length of the covered area is an assumption for two different amplicons shares does not share a common area but the gap between is smaller than the coverage gap specified by the user. (C) Two different areas are reported.

INPUT

(8) **Primers file** (**necessary**): Primer name, Reverse sequence and forward sequence in tab delimited format

Primer file format:

ptest1	TGATTGGTTAGGCACTTT	TAAAATACGAAATTGATG
ptest2	CCTCGATACTAGTGATCT	ATCAAAATTTTCCTCATG
ptest3	AATCATCCCTCTCAACCT	CTGCTACAGTGTCGCCTA
ptest4	ACGTAATTTAGGCTTAGT	GTTGAAAAAATTTGAGTC

(9) Sequence format (**necessary**): Spoon accept single sequence FASTA file format and cannot deal with multiple sequence FASTA format. Where the chromosome name, gene start location, gene end location, gene name and item type gene/SNP/other must be defined.

(10) Annotation file (**optional**): can be provided using GenBank format or gene list format.

Gene list file format:

Chr19	218114	221413	hypothetical_protein	Gene
Chr19	226515	228528	hypothetical_protein	Gene
Chr19	249047	255958	MtN30_family_protein	Gene
Chr19	257049	257492	hypothetical_protein	Gene
Chr19	258983	259816	hypothetical_protein	Gene
Chr19	295055	303570	hypothetical_protein	Gene
Chr19	306524	307215	hypothetical_protein	Gene
Chr19	310712	311538	HyPRP_family_protein	Gene

(11) In vitro PCR results file (**optional**): Primer name and amplimer molecular weight.

In vitro PCR format:

ptest27	767
ptest25	1035
ptest23	327
ptest16	1468
ptest22	335
ptest26	1246
ptest21	205
ptest27	769
ptest24	584

(16) Specific genomic area coverage file (**optional**): User can defined specific area to be statistical analysis by clicking on (16) text box or leave it with default “ALL” option. The file format must contains FASTA accession name (chromosome/GI/other) , area start location and area end location , separated by “:” sign.

Coverage file format:

```
Chr1:10000:154207
Chr2:15244:175627
Chr3:15247:195715
Chr4:18771:124903
Chr5:0:121591
Chr6:1254:153487
Chr7:129000:129799
Chr8:160100:163603
```

OUTPUT

A) In silico PCR Results file (CSV format) : Primer name, forward primer sequence, reverse primer sequence, amplimer length, forward primer vs. hit mismatch, and primer vs. hit mismatch, sequence Name, forward hit Pos and reverse hit pos.

Ptest14	CATAAAGCATTTAGAGAG	GGGTTAGGGTTTAGGGTT	1107	0	0	Chr1	352	1459
ptest22	CGATAGTGATTGGAATCC	TTTAGGGTTTAGGGTTTA	943	0	0	Chr2	420	1363
ptest26	TAGGGTTTAGGGTTTAGG	GATGTACCAAATTTAATG	1122	0	0	Chr3	384	1506
ptest27	TAGGGTTTAGGGTTTAGG	GATGTACCAAATTTAATG	157	0	0	Chr4	436	593
ptest29	TCTGCCTTTCATCACCTT	CCCACCATAAAGGAATGG	1029	0	0	Chr5	91	1120
ptest25	CCATCTCAGGTAACGTGG	AATGTACGTCTCGGGACA	1323	0	0	Chr6	136	1459
ptest25	CCATCTCAGGTAACGTGG	AATGTACGTCTCGGGACA	1037	0	0	Chr7	136	1173
ptest16	GCTGCTCATTGGAAAAGA	TCCTCGATAATATGTATG	600	0	0	Chr8	176	776
ptest21	ACGCGCCCCGAGGCGTAC	ACTATAGCGCGTGCGATA	1269	0	0	Chr9	206	1475
ptest29	TCTGCCTTTCATCACCTT	CCCACCATAAAGGAATGG	502	0	0	Chr10	236	738

B) In silico PCR Results near or adjoin genes (CSV format): same as the in silico result format in addition to a gene name row.

C) Amplicons with genes (CSV format): Amplimers that adjoin/near genes .Primer name, chromosome name, amplimer location start and amplimer location end.

D) Covered genes file: genes that are covered by one or more primer .Chromosome name, gene location start and gene location end.

E) Covered areas (CSV format): areas covered by one or more PCR amplicons (after amplifying the primer coverage algorithm: Primer name, area covered specification, area start location and location end. This file shows if one area (specified early by user) if it was continuously covered by linked amplimers (the gap between amplimers is less than (18)) or it was covered with discontinuously amplicons areas which means that user must cover this areas with additional primers.

SCoT-06	chr20:0:28666397	4546450	4548658
SCoT-01	chr41:0:28666367	20666402	20668300
SAMPL-09_1	chr44:0:28666391	4058478	4059033
OP-G13	chr20:0:28666397	19337298	19338220
SAMPL-11	chr41:0:28666367	4397315	4397942
SCoT-47	chr20:0:28666397	13211975	13212911
SCoT-37	chr41:0:28666367	2643191	2643781

F) Coverage statistical analysis output (by base pair) (CSV format): primer name, all primer coverage across genome and the coverage across every chromosome separately. The last row for all primers set coverage. This coverage values are calculated using the primer coverage algorithm for example all primer set coverage could be not equal to the assumption of the area covered by every primer separately, as one area could be covered more than once by different primers.

PrimerName	PrimerALLCoverage	chr44:0:28666391	chr14:0:28666387	chr8:0:28666369
SCoT-06	148366	3918	5132	7150
SCoT-01	60880	670		
SAMPL-09_1	126242	3935	2994	5125
OP-G13	922			
SAMPL-11	39276		1794	925
SCoT-47	37880	4360	564	1442
SCoT-37	188742	7478	3508	6810
PrimerSet	15926082	512697	488371	486290

G) Coverage statistical analysis output (by percentage) (CSV format): In this file the coverage is transformed to the percentage covered compared to the chromosome length (area length) or the total genome length in the case of all primer/primer set coverage.

H) Anchor primers (CSV format): primers that have one PCR amplicons in one chromosome/area. These primers are useful for linkage mapping analysis. it shows primer name , area specification, primer amplicons start location and primer amplicons end location.

I) Report table (tab delimited format): these table summaries all the Spoon PCR results for every primer. It shows: primer-name, forward-miss-forms, reverse-miss-forms, band-count, band-length-count, Chr-count, genome-coverage, genes-count, actual-genes-count and inv-ins-bands.

SCoT-06	171	171	216	91	31	148366	22	15	5
SCoT-01	40	40	118	56	24	60880	2	1	8
SAMPL-09_1	2	2	242	221	33	126242	43	26	5
OP-G13	1	1	2	1	1	922	0	0	7
SAMPL-11	2	2	87	87	26	39276	18	12	1
SCoT-47	66	66	86	40	21	37880	24	9	2
SCoT-37	225	225	366	125	31	188742	28	13	8

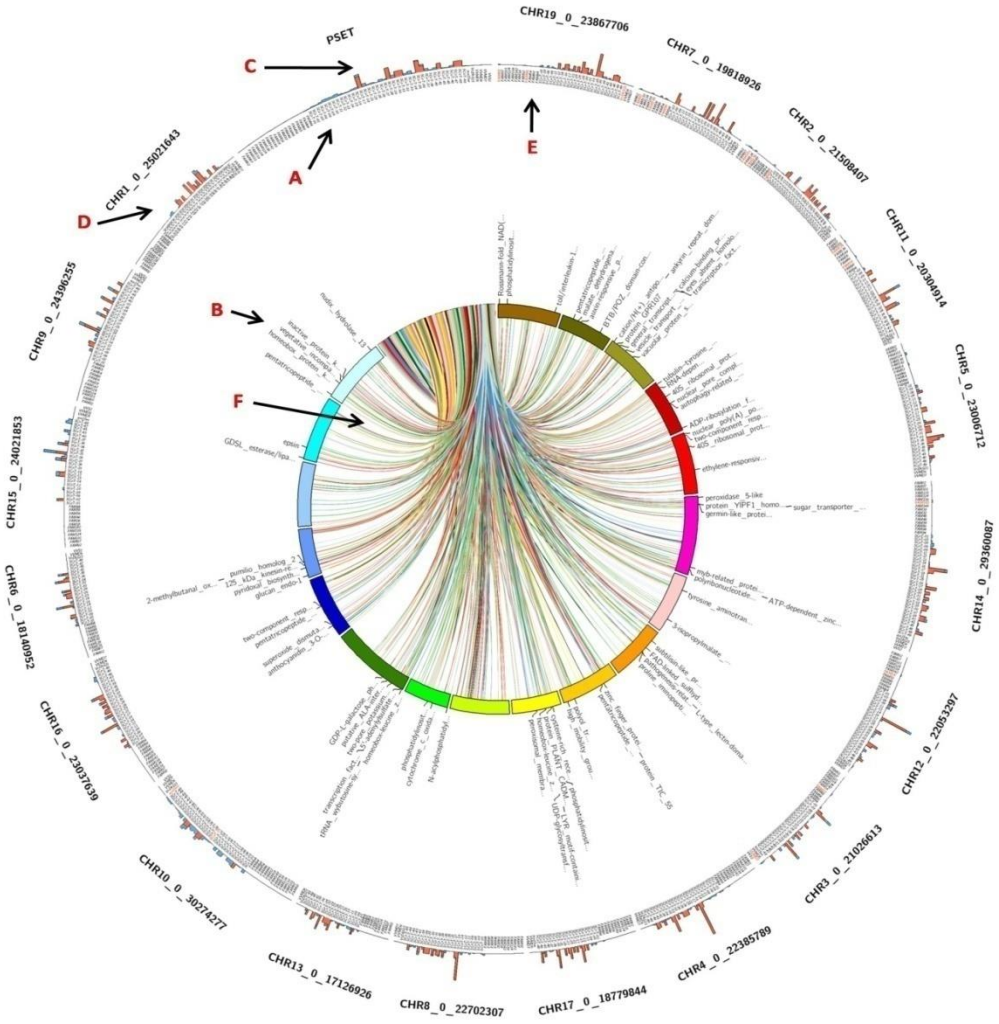
J) In silico amplimers sequence: the sequence of in silico amplimers in FASTA format.

K) In silico vs. in vitro PCR comparison output file (tab delimited format): In silico amplicons that are approximately match other *in vitro* amplicons (according to user defined mismatch (21)) under the same primer. it shows primers name , insilico amplicons length, chromosome name , *in vitro* amplicons length and if the user chosen the to compare in silico amplicons with genes (INS-G) (22) an addition row will appear with the gene name. This file important for selecting in vitro PCR amplicons for gene sequencing in gene discoveries protocols.

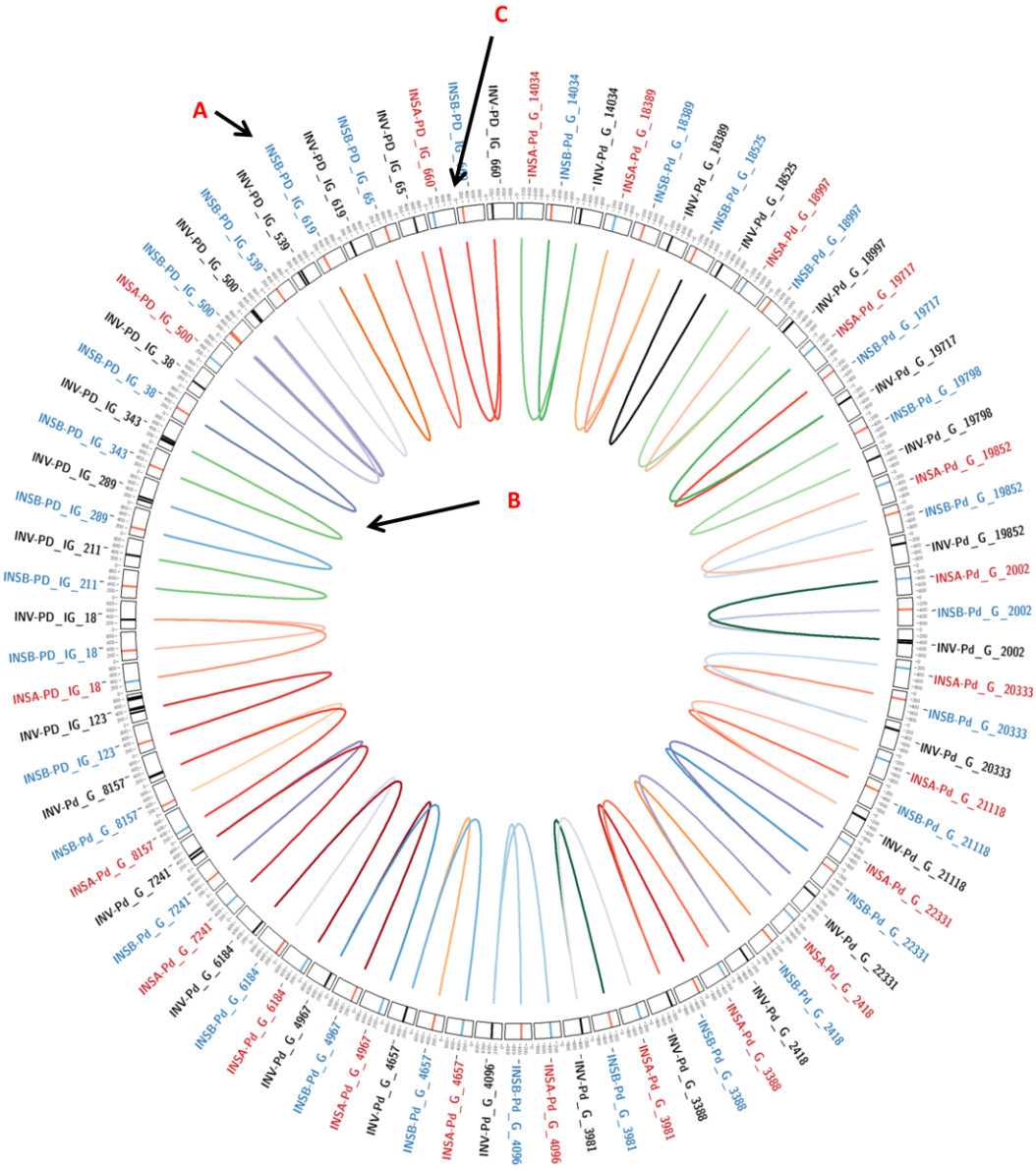
L) Report (text format): It summaries all Spoon results for all primer set. It shows different information as shown below.

```
Total number of primers :       29              primer(s)
Total number of primers with hits :             29              primer(s)
Total number of insilico bands :                10610           band(s)
Total number of genic insilico bands :          383             gene(s)
Genome Coverage :               1883770bp       which about 0.6273008% of the
total area
Primers exists in one Chr. (anchor) :           4               primer(s)
```

M) Circos configuration for in silico PCR analysis: The *in silico* PCR results for the SCoT and SSR set (A) against *V.vinifera* genome, revealing the possible adjoined genes (B), primer total genome coverage percentages statics (C), primer total chromosome coverage percentages statics (D) and the position of possible PCR amplimers with genes (extended lines) or without (short lines) (E), anchor primers with red font.



N) Spoon configuration for the comparison between *in silico* PCR (**INS**) analysis and the PCR band sizes retrieved by wet lab PCR (**INV**). **(A)** *In silico* or *in vitro* band where INS/INV + primer name. (B) Different amplicons highlighted along with their band size. (B) physical link between amplimer existed in both *in silico* and *in vitro* PCR.



When the marker's linkage map position is provided, Spoon can provide a comprehensive comparison between *in silico* and linkage maps and the assignment report for every linkage map to chromosome(s).Spoon will generate different Circos configurations for *in silico* PCR statistics results (Figure 1), the comparison between the *in silico* and *in vitro* PCR data (Figure 3) and linkage and *in silico* (physical) maps (Figure 4). If SNP database is provided, Spoon could be used to design KASP assay primers for SNP genotyping in order to target QTL and genic chromosomal regions. These KASP primers are designed to target all SNPs that are nearby or adjoin *in silico* amplimers. A report files will provided explain KASP targeted genes and loci.