KASPspoon

About KASPspoon

The final and main goal of KASPspoon is converting the classically acquired QTL information into more comprehensive, routine, and accurate molecular marker technologies such as a KASP assay. In order to reach this goal, KASPspoon can be used efficiently to compare in vitro (laboratory observed) and *in silico* (predicted) PCR results, and reports SNPs that are close to or adjoined by PCR amplimers by integrating the database for known SNPs.

KASPspoon uses this information to design KASP assay primers in order to convert QTL markers into high-throughput genotyping to provide major SNP markers for the dissection of genotypic and phenotypic variation. It also reports chromosomal-specific primers (anchored) and compares between physical and genetic maps, which could be useful for linkage and association mapping analysis. Additionally, primer set genome coverage can support genome-walking PCR procedures that cover gene-rich genomic regions. Also, by processing simple sequence repeat (SSR) markers, KASPspoon can report amplimers that adjoin SSRs in order to analyze comparisons between observed and predicted motifs and to use their abundance in the genome for more accurate primer selectivity.

Furthermore, KASPspoon produces different Circos configurations to illustrate in silico results, which can be easily handled through the Circos software package . This helps users to visualize results, unify biological data results for multiple analysis tools, magnify and exclude any part of the results, and develop further analysis techniques using simple programming (i.e. Circos scripting) without handling the original source code.

KASPspoon applications

Creating KASP primers targeting SNPs nearby SSR markers is the final and main aim of using **Spoon**. On the other hand, there are many applications this tool could be used for:

1- Extracting more information about QTL markers revealed by classical molecular markers such as SSR markers through in silico PCR. These information includes :
    a- Genomic positions.
    b- Genes nearby.
    c- Simple sequence repeats.
    d- Single nucleotide variations.
2- Comparing *in silico* and *in vitro* PCR through Position-depend mispairing weight, this could provide more information about polymorphic bands and those that are highly linked to mophigical traits.
3- Comparing genetic and physical linkage maps, this could provide more information about QTL markers and help in fine mapping.
4- Searching for chromosomal anchored markers.
5- Testing primer set whole genome coverage.
6- SNPs nearby QTLs reported through Spoon, could be used to study SNP-gene effect studies.

LICENSE and CITATION

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

http://www.ageri.sci.eg/index.php/facilities-services/ageri-softwares/spoon

OR

https://github.com/AlsammanAlsamman/Spoon

CONTRIBUTORS

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

INSTLATION

FOR WINDOWS: Spoon has a simple excitable **exe** for windows installation.

FOR LINUX and OS: You need to install perl 5, version 18, subversion 2 (v5.18.2) and TK Module and install Spoon through:

> sudo sh Install.sh

> perl Spoon.pl

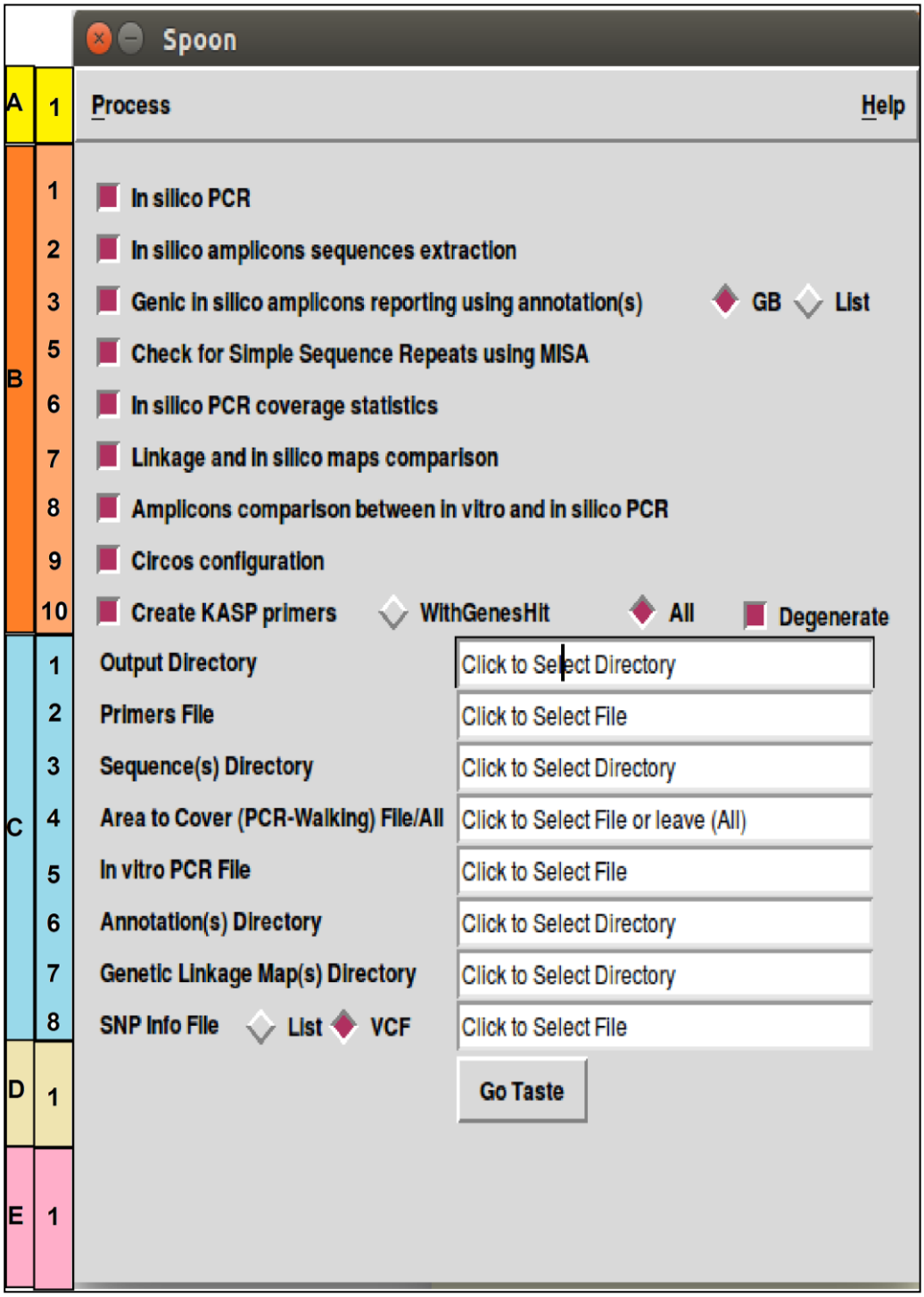
Or You can run spoon with "Spoon-without-GUI.pl" which does not need to install Tk from CPAN but you need to configure your run through config.info using :

> Spoon-without-GUI.pl CONFIGURATION-FILE

FOR TUTORIAL PLEASE WATCH THIS VEDIO:

https://youtu.be/hLPcLNej4z4

KASPspoon Main Window



Main Window

A1 The main menu item: process includes parameters and Help includes information.

B1 in silico PCR: Performing *in silico* PCR analysis using PCR primers and genomic sequences. (Minimum data: **(C1-3)**).

B2 In silico amplicons sequences extraction: Extracting amplicon sequences in multiple FASTA file (Minimum data: **(C1-3)**).

B3 Genic in silico amplicons reporting using annotation (s): Reporting in silico amplicons close to or adjoin genes, gene annotation in GenBank format or list (tab delimited) (Minimum data: **(C1-3 and 6)**).

B5 Check for simple sequence repeats using MISA: Reporting simple sequence repeats adjoin PCR amplimers (Minimum data: **(C1-3)**)..

B6 In silico PCR coverage statistics: reporting the statistics of PCR primer set genome coverage (Minimum data: **(C1-3)**).

B7 linkage and in silico maps comparison: comparing physical and genetic position of different PCR markers (Minimum data: **(C1-3 and C7)**).

B8 Amplicon comparison between in vitro and in silico PCR: comparison in silico amplicons acquired through Spoon and in vitro amplicons required through wet PCR through band size weighting (Minimum data: **(C1-3 and 5)**).

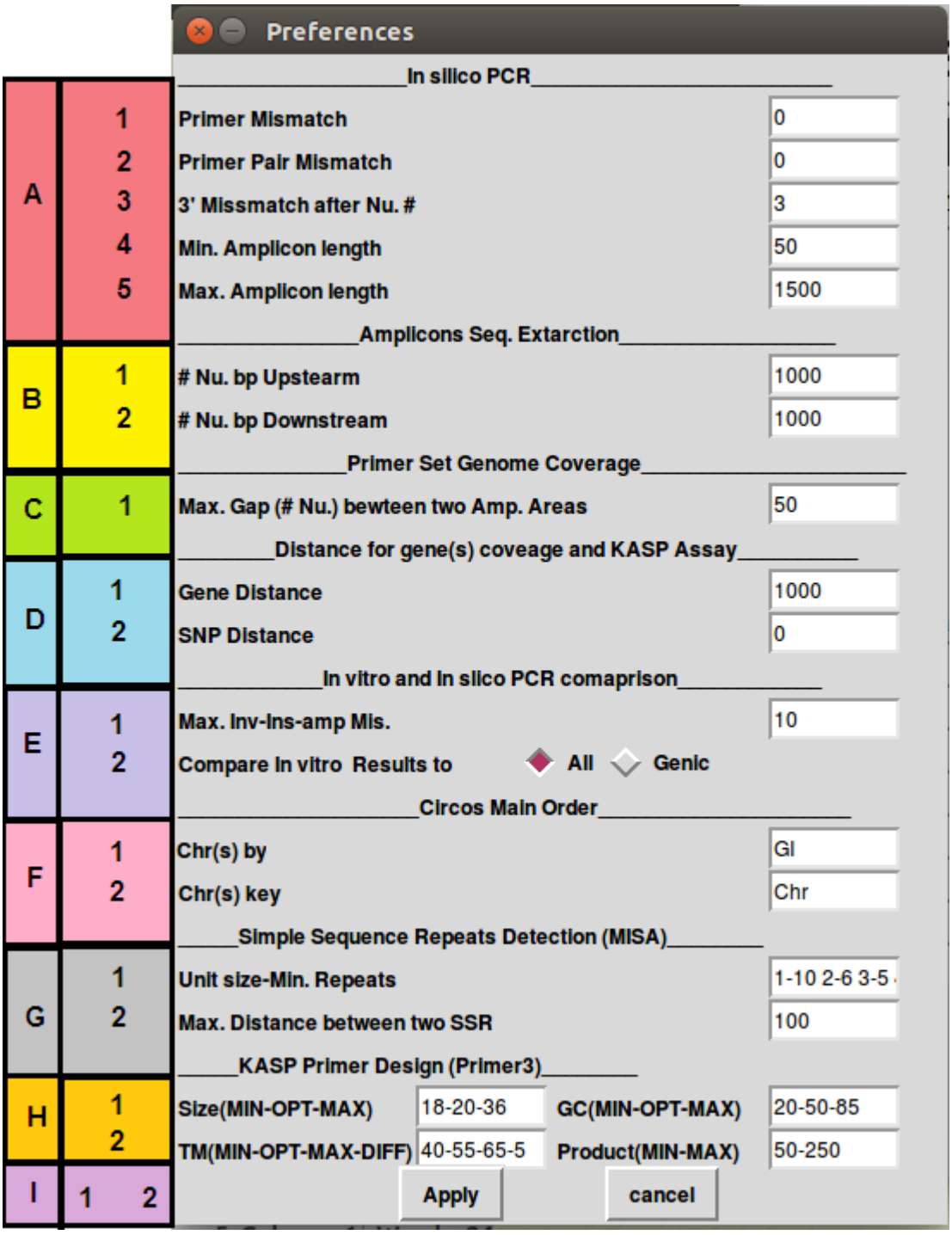
B9 Circos configuration: creating Circos configuration for *in silico* PCR amplicon position in the genome, in vitro and in silico PCR comparing and genetic and physical maps comparison(Minimum data: depending on analysis).

B10 Create KASP primers: developing KASP primers through amplicon amplimers and SNP database. If user choose with genes hit, KASP primers will be created for SNPs near genes and if degenerated is checked if there are SNPs unwanted near genes it will be masked through IUPAC (Minimum data: **(C1-3 and 8)**).

D1 Start analysis button.

E1 The progress area.

KASPspoon Parameters



PARAMTERS

In silico PCR

A1 Primer mismatch : The maximum mismatch between primer and DNA template sequence using base-pair complementation by base pair.

A2 Primer pair mismacth: the maximum summation of mismatch of forward and reverse primers (by base pair).

A3 3' mismatch after nucleotide number : if the mismatch between PCR primer and DNA template is in the range of the first # nucleotide in the 3' direction this PCR amplification will not be reported.

A4 Minimum amplicon length: minimum PCR amplicon length.

A4 Minimum amplicon length: maximum PCR amplicon length.

Amplicon Sequence Extraction

B1 # nucleotide base pair upstream: extracting the sequence of the PCR amplimers starting from # nucleotide in the upstream direction.

B2# nucleotide base pair downstream: extracting the sequence of the PCR amplimers starting from # nucleotide in the downstream direction.

Primer Set Genome Coverage

C1 Max Gap (# nucleotide) between two amplimers areas: the maximum gap between two different PCR amplimers located on the same chromosome reported as one fragment in the PCR primer set genome coverage.

Distance for gene (s) coverage and KASP assay

D1 Gene distance: The maximum distance between PCR amplimer position in the genome and nearest gene.

D2 SNP distance: The maximum distance between PCR amplimer position in the genome and nearest SNP.

In vitro and in silico PCR comparison

E1 maximum in vitro – in silico amplicons mismatch: the maximum difference between in silico and in vitro PCR amplicon to be considered as one amplicon.

E2 compare in vitro amplicons results to: in silico PCR amplicons have genes nearby or compare to all PCR amplicons.

Circos Main order

F1 Chr(s) by: Chromosome ordering in Circos figures using GI number

F2 Chr(s) key: The chromosome key name in the Circos configuration.

Simple Sequence Repeats Detection (MISA)

G1 Unit size- Minimum repeats: The Motif Sequence minimum size and and the minimum number of repeats

G2 Maximum distance between two repeats : the maximum distance between two repeats to be considered as one repeat.

KASP primer Design (PRIMER3)

H1 Size (MIN-OPT-MAX): The length of primer (minimum, optimum and maximum).
H2 GC (MIN-OPT-MAX): The GC content of primer sequence (minimum, optimum and maximum).
H3 TM (MIN-OPT-MAX): The annealing temperature of primer sequence (minimum, optimum and maximum).
H4 Product (MIN-MAX): The length the of PCR primer product (minimum and maximum).

Apply and Cancel buttons

I1 APPLY: Apply these new parameters.
I2 CANCEL: forget and work with previous parameters.

How Does KASPspoon works?

KASPspoon handles a different experimental genetic data type, correlates between the genetic and physical position of loci, and generates ultimate positioning markers for KASP assay. Snapshots of KASPspoon outputs are shown in Supplementary 1.

KASPspoon was developed as a standalone package for PCR primer analysis using both C and Perl programming languages. The Boyer–Moore–Horspool (Horspool, 1980) and Baeza-Yates–Perleberg (Baeza-Yates and Perleberg, 1992) string approximate-matching algorithms were used through C to search genomic sequences provided PCR primer-pair sequences were used as queries. For primer genome coverage statistics and PCR-walking procedure, the overlap layout consensus algorithm with a user-defined gap between amplimers is used to report the primer(s') covered area (Supplementary 1).

KASPspoon uses common biological data formats as an input and only amplimers that do not exceed the user-defined maximum primer mismatch or the maximum total mismatch (forward primer mismatch + reverse primer mismatch), or those that do not have a mismatch in the first user-defined 3′ nucleotides are reported (Supplementary 1). KASPspoon can compare the *in silico* (predict PCR product size) and *in vitro* PCR results (observed PCR product size) by defining the maximum molecular weight mismatch between the *in silico* and *in vitro* amplimers if the *in vitro* PCR product length (in base pairs) is provided (approximately). For comparing *in silico* and *in vitro* PCR amplimers, KASPspoon generates a text file containing amplimers that exist in both and that contain (or do not) genes. Additionally, MISA Perl script (pgrc.ipk-gatersleben.de/misa/) is integrated inside the KASPspoon tool in order to report all SSRs that lie between the PCR-amplified regions.

When the marker's linkage map position is provided, KASPspoon can provide a comprehensive comparison between *in silico* and linkage map(s), and the assignment report for every linkage map to chromosome(s) where KASPspoon can compare between the physical (bp) and genetic (cM) positions for provided molecular markers.

If a list of SNP variations is provided, KASPspoon can generate KASP assay primers that can be used for SNP genotyping. These KASP primers are designed to target all SNPs that are nearby PCR-amplified chromosomal regions. Report files that contain all KASP-targeted genes and marker loci are generated. The KASP sequences are designed according to a KASP primer design manual published by LGC (www.lgcgroup.com).The Primer3 tool (Untergasser *et al*., 2012) was used to design two allele-specific forward primers, and a common reverse primer for allele-specific assays such as KASP assay. These primers designed by KASPspoon use a

user-provided SNP database to create degenerate PCR primers in order to provide primers with minimal mismatches, where the target nucleotide is marked by "[]"and untargeted nucleotides are masked according to IUPAC codes.

KASPspoon will generate different Circos configurations for *in silico* PCR statistical results, the comparison between *in silico* and *in vitro* PCR data and linkages, and *in silico* (physical) maps.

The search returns a sequence output file in FASTA format, containing all sequences in the database that lie between, and include, the primer pair. The FASTA header describes the region in the database and the primer names. Comma-separated output files generated by KASPspoon include:

(1) *in silico* PCR-generated amplimer information
(2) *in silico* amplimers near/adjoining genes
(3) SNPs near *in silico* amplimers in VCF format
(4) *in silico* amplimers adjoining SSRs
(5) *in vitro* and *in silico* amplimers acquired by the same PCR primer that share the same approximate band size
(6) location of genes adjoining or close to PCR primer regions
(7) genomic areas that are covered using this primer set
(8) primer set coverage statistics with both base-pair and percentage scales (compared to length of the total genome sequence covered and chromosomal sequence length)
(9) chromosomal assignment for linkage genetic groups
(10) KASP primer(s) sequence and information
(11) final report files containing different information about this run in an abbreviated form.

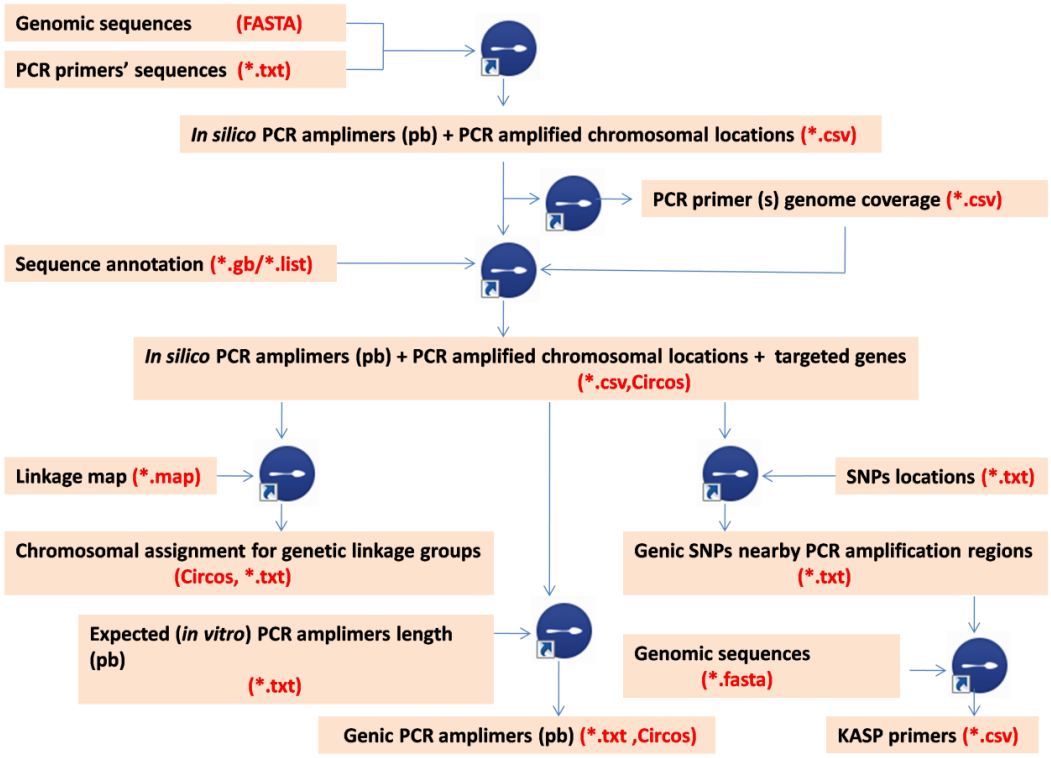


Figure 3 : Flowchart illustration for Spoon *in silico* and *in vitro* analysis procedure.

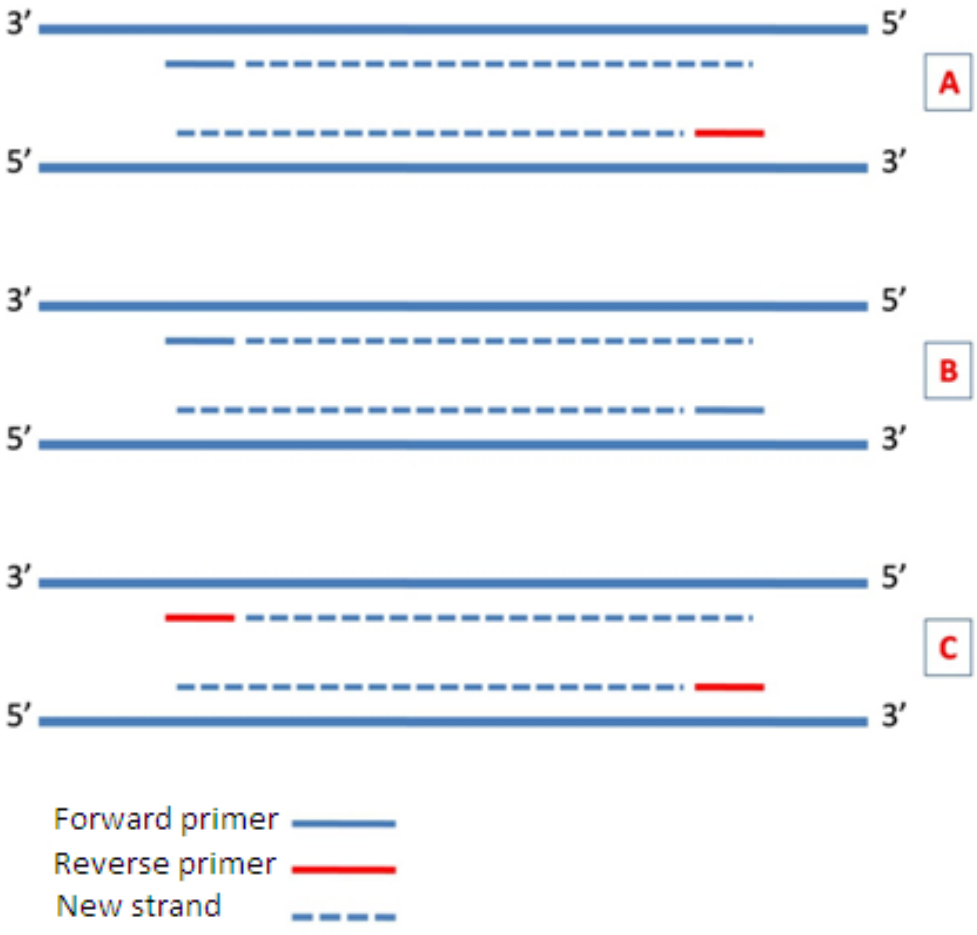


Figure 4: The different primers'amplification probabilities that Spoon considers during the *in silico* PCR analysis. (A) A normal PCR amplification where both forward and reverse primers are integrated. (B) and (C) probable amplification where one primer (forward/reverse) is integrated.

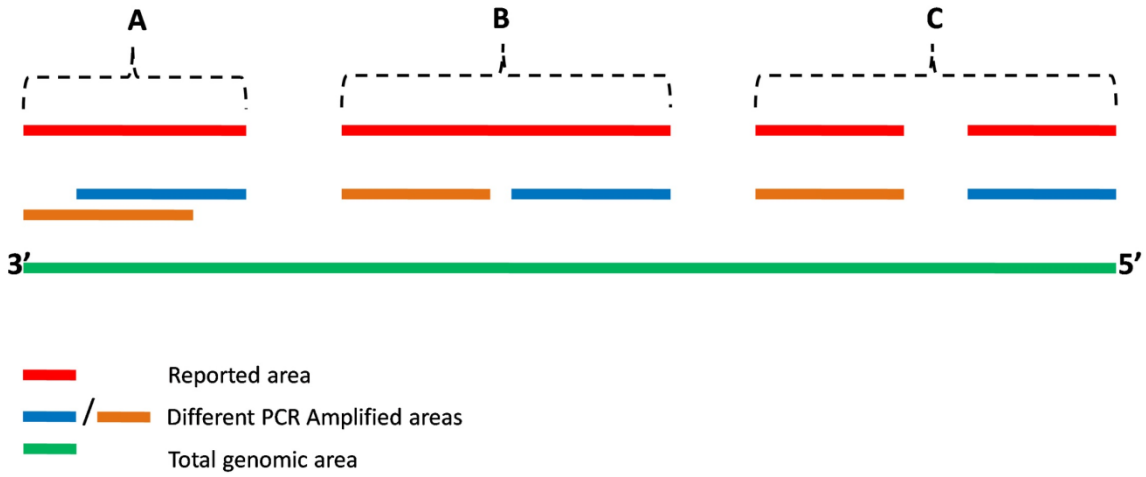


Figure 5: The different primer's genomic region coverage probabilities, which Spoon considers during the amplimers-genome coverage statistical analysis for the *in silico* PCR results. (A) The length of the covered area is an assumption for two different amplicons sharing a common genomic area. (B) The length of the covered area is an assumption for two different amplicons with a gap less than the user specified coverage gap. (C) Two different areas are reported with a gap more than the user specified coverage gap.

INPUT

In main window section C, the minimum spoon input data spoon can work with are the output directory (C1)**, primers file** (C2) and the sequence directory (C3). On the other hand Spoon is a continuous pipeline once the required output files are created, these output will be an inputs for the next procedures if selected.

(C1) The output folder selection (**necessary**): The output folder where output files will be stored.

(C2) **Primers file** (**Required for B1** *in silico PCR*): Primer name, Reverse sequence and forward sequence in tab delimited format

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

(C3) Sequence Directory (**Required for B1-10**): Spoon accepts sequences in FASTA file format. Multiple sequences could be retrieved from one directory where FASTA files are extended with ".fasta".

(C4) Specific genomic area coverage file (**optional for specific genome coverage statistics**): User can be defined specific area to be statistical analysis by clicking on 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.

```
Chr1:10000:154207
Chr2:15244:175627
Chr3:15247:195715
Chr4:18771:124903
Chr5:0:121591
```

```
Chr6:1254:153487
Chr7:129000:129799
Chr8:160100:163603
```

(C5) In vitro PCR results file (**Required for B8**): Primer name and amplimer molecular weight.

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

(C6) Annotation file (**Required for B3**): can be provided using GenBank format or gene list format. Where the chromosome name, gene start location, gene end location, gene name and item type gene/SNP/other must be defined.

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

(C7) Genetic linkage maps directory folder (**Required for B7**): The directory of genetic linkage maps could be used for comparing physical and genetic marker positions. Every map file corresponds to one chromosome/linkage map. The marker centimorgans position and marker name. Files must be extended by ".map".

```
0   CaM1135
1.2 OPU18
2.1 ISSR8603
3.3 AAMCTT07
3.9 H1A12
5.5 CaM2064
6.2 PR5
7.4 CaM1648
8.1 OPD3-4
10.5    ICCM0166a
```

(C8) SNP file information (**Required for B10**): the SNP file information holds the nucleotide variations. The SNP information could be provided in two different formats (LIST or VCF).

LIST FROMAT

```
Ca172 T   G
Ca1164A   T
```

Ca1	164	A T
Ca1	218	T C
Ca1	240	C T
Ca1	247	T G
Ca1	288	C A
Ca1	331	A G
Ca1	331	A G
Ca1	341	C A

VCF FORMAT

Ca1	72	Ca1_T-G	T	G	.	. .
Ca1	164	Ca1_A-T	A	T	.	. .
Ca1	164	Ca1_A-T	A	T	.	. .
Ca1	218	Ca1_T-C	T	C	.	. .
Ca1	240	Ca1_C-T	C	T	.	. .
Ca1	247	Ca1_T-G	T	G	.	. .
Ca1	288	Ca1_C-A	C	A	.	. .
Ca1	331	Ca1_A-G	A	G	.	. .
Ca1	331	Ca1_A-G	A	G	.	. .
Ca1	341	Ca1_C-A	C	A	.	. .

OUPUT

A) In silico PCR Results file (**In_silico_PCR_Results.csv**): 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 adjoin SSR (**In_silico_PCR_Results_With_SSR.csv**): same as the in silico result format in addition to a SSR motif row.

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

D) Amplicons with genes (**Amplicons_with_genes.csv**): Amplimers that adjoin/near genes .Primer name, chromosome name, amplimer location start and amplimer location end.

E) Covered genes file (**Covered_genes_file.txt**): genes that are covered by one or more primer .Chromosome name, gene location start and gene location end.

F) location of genes adjoining or close to PCR primer regions (**Seq_Genes_Loc.txt**).

G) Covered areas (**CoveredAreas.csv**): 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 provided gap or it was covered with discontinuously amplicons areas which means that user must cover this areas with additional primers.

SSR-06	chr20:0:28666397	4546450	4548658
SSR-01	chr41:0:28666367	20666402	20668300
SSR-47	chr20:0:28666397	13211975	13212911
SSR-37	chr41:0:28666367	2643191	2643781

H) Coverage statistical analysis output (by base pair) (**Coverage_statistical_analysis_output_by_base_pair.csv**): 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
SSR-06	148366	3918	5132	7150
SSR-01	60880	670		
SSR-47	37880	4360	564	1442
SSR-37	188742	7478	3508	6810
PrimerSet	15926082	512697	488371	486290

I) Coverage statistical analysis output (by percentage) (**Coverage_statistical_analysis_output_by_percentage.csv**): 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.

J) Anchor primers (**Anchor_primers.csv**): 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.

SSR-01	40	40	118	56	24	60880	2	1	8
SSR-47	66	66	86	40	21	37880	24	9	2
SSR-37	225	225	366	125	31	188742	28	13	8

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

L) In silico amplimers sequence: the sequence of in silico amplimers in FASTA format contains SSR (**In_silico_amplimers_sequence-for-SSR.fasta**).

M) MISA output for amplimers contain SSR motif (**In_silico_amplimers_sequence-for-SSR.fasta.misa**).

N) SNPs near in silico amplimers in VCF format (**SNPs-NearBy.vcf**) and text format (**SNPs-NearBy.txt**).

O) The SNP report nearby genes **(SNPs-NearBy.txt.report.txt).**

P) The motif report for every primer (**PRIMER-MOTIF-report.txt**), if the primer has more than one motif.

Q) In silico vs. in vitro PCR comparison output file (tab delimited format) (**INS_INV_Amps.txt**): In silico amplicons that are approximately match other in vitro amplicons (according to user defined mismatch) 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) 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.

R) Report table (tab delimited format) (**Report_table.csv**): 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.

S) Report (**Report_text.txt**): 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)

T) Circos configuration for in silico PCR analysis: The in silico PCR results for the SSR 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.

U) 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.

V) The chromosomal assignment for linkage genetic groups (**LG-ASSIGNMENT-Report.txt**).

W) KASP primer(s) sequence and information (**RESULTSKASP-PRIMERS.txt**).

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.

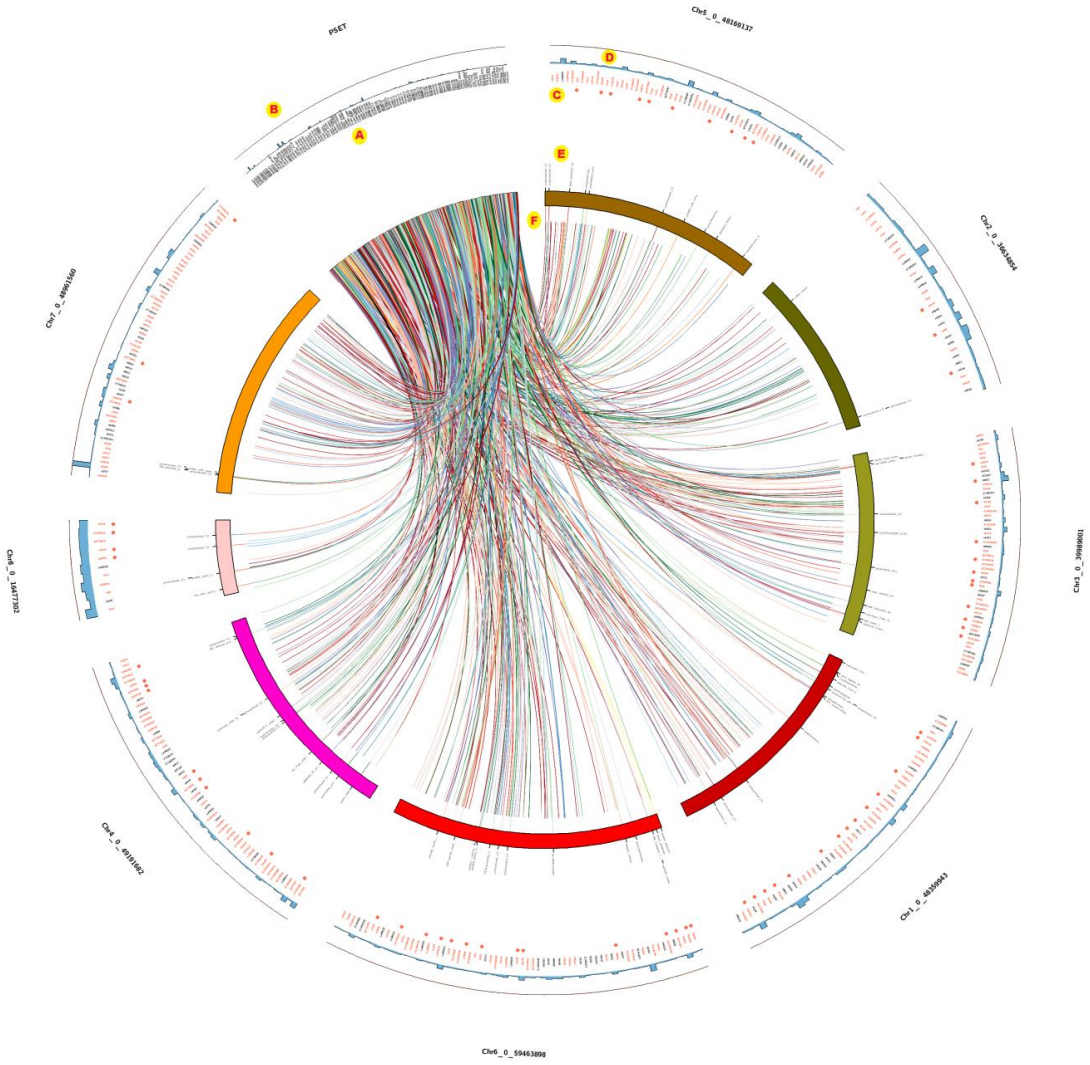


Figure 3: Circos configuration for the SSR primer set *in silico* PCR analysis (A) against the chickpea genome, revealing the primers' total genome coverage percentage statistics (B) where some primers acquired genes on this chromosome (dotted) or are anchored (red font) (C) and their chromosomal coverage percentage statistics (D), the possible adjoined genes (E) and the position of possible PCR amplimers with genes (extended lines) or without (short lines) (F).

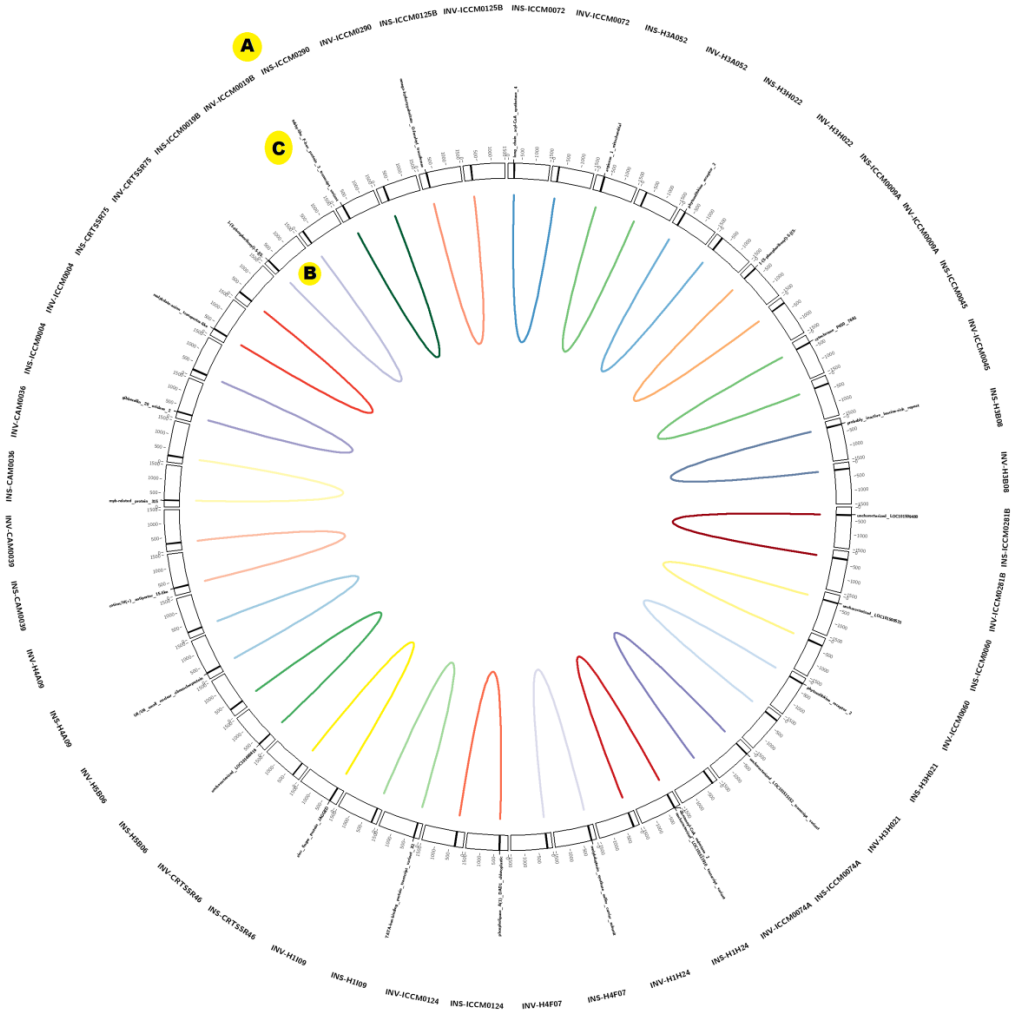


Figure 4: Spoon configuration for the comparison between *in silico* PCR (INS) analysis and the PCR band sizes retrieved by wet lab PCR (INV) for the chickpea SSR as a virtual gel electrophoresis.(A)*In silico* or *in vitro* band where INS/INV + primer name. (B) Different amplicons highlighted along with their band size and the physical link between amplimer shared between *in silico* and *in vitro* PCR data. (C) Genes located near *in silico* amplimers.

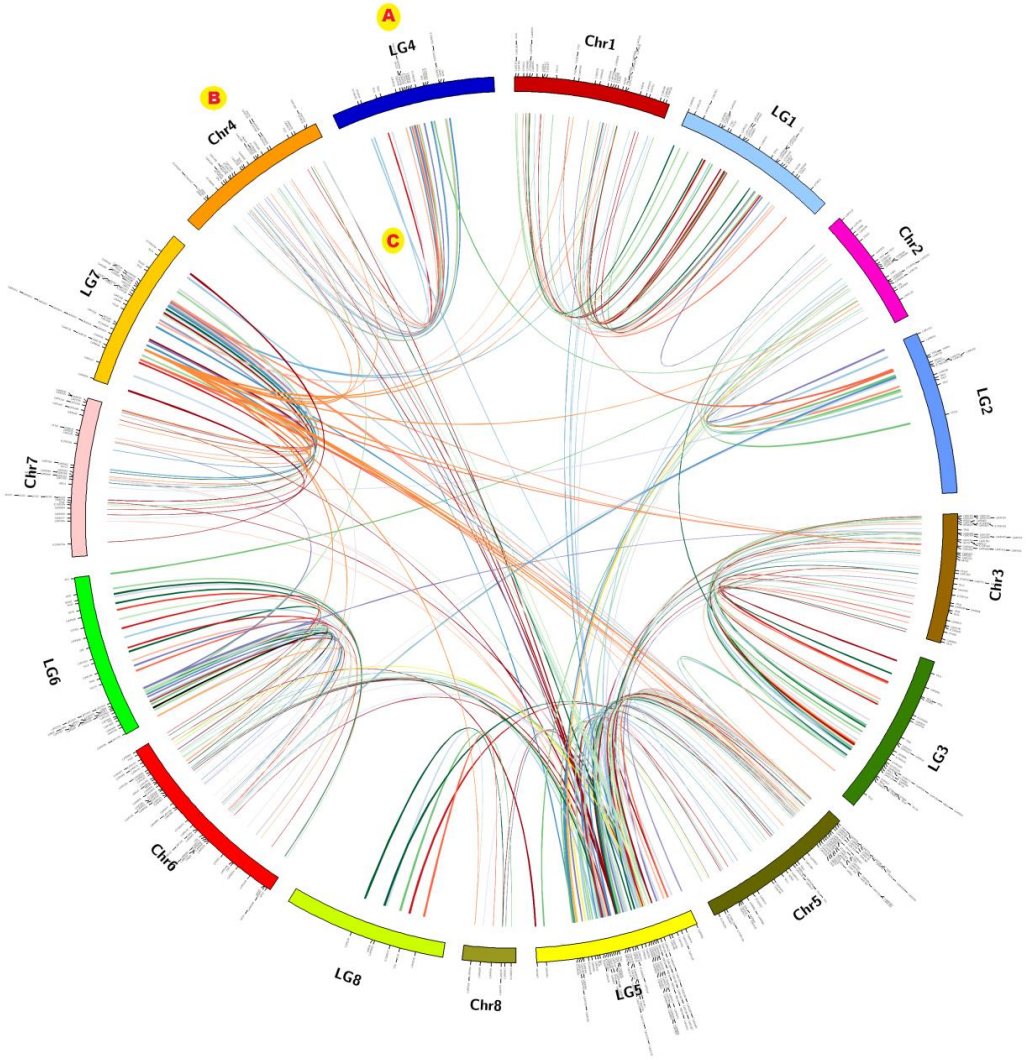


Figure 5: Spoon configuration for the comparison between linkage maps **(A)** and *in silico* maps (**B**) analysis, where shared primers are linked through different linkage groups and physical chromosomes**(C)** for chickpea.