The shortest path to CSSL selection

Sarah Ayling[1] and Mathias Lorieux[1,2]

1. Agrobiodiversity and Biotechnology Project, International Center for Tropical Agriculture (CIAT), A.A. 6713, Cali, Colombia. s.ayling@cgiar.org
2. Institut de Recherche pour le Développement (IRD), Plant Genome and Development Laboratory, UMR 5096 IRD-CNRS-Perpignan University, 911 Av. Agropolis, 34394 Montpellier Cedex 5, France. mathias.lorieux@ird.fr

Chromosome Segment Substitution Lines (CSSLs)

CSSLs are plant lines developed with traditional breeding approaches that ideally contain a single segment of donor genome within a recipient genome background [1]. CSSLs are generated by backcrossing the offspring from a cross between the recipient and the donor repeatedly with the recipient species (Figure 1). Marker analysis can be performed to determine the extent of donor genome within the offspring. The goal is to identify a set of lines which each contain a single segment of the donor genome, and which together cover the entire donor genome. This can be achieved through the development of many introgression lines, with the selection being performed at the final stage, or through the development of fewer lines, where marker-aided selection is applied at each generation. CSSL populations can then be phenotyped to identify QTLs (Quantitative Trait Loci) associated with desirable features such as yield increase and drought tolerance.

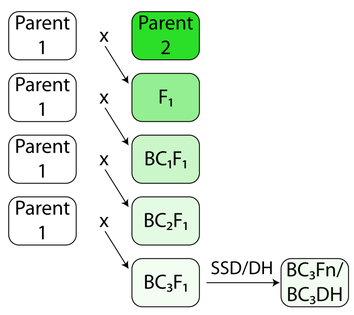


Figure 1. Parent 1 and 2 are crossed initially to produce the F1 generation. The offspring are repeatedly crossed with parent 1. After multiple rounds of backcross, the offspring contain few segments of the donor genome in a genome that predominantly resembles the recipient parent. The BC_3F_1 can be fixed by single seed descent (SSD) or doubled haploid (DH) in order to produce BC_3F_n or BC_3DH.

CSSL Finder http://mapdisto.free.fr/CSSLFinder/

CSSL Finder (Figure 2) is software designed to aid the selection of a set of lines that cover the donor genome, whilst minimising the presence of donor genome background. Developed as an Excel-VBA application, CSSL Finder reads in a matrix of markers and gives users the option to select a subset of those markers for use in the analysis (automatically or manually). There is also the opportunity to infer missing data points, provided flanking markers are sufficiently close and unambiguous. A greedy algorithm for line selection selects the optimal line for the segment covering the first markers in linkage group 1, and then continues along each linkage group, selecting lines without replacement, until all markers are covered. The main limitation of this heuristic is that markers in the first linkage group can be chosen from all lines, whereas markers in the last group have a smaller set of lines remaining to choose from, biasing the selection.

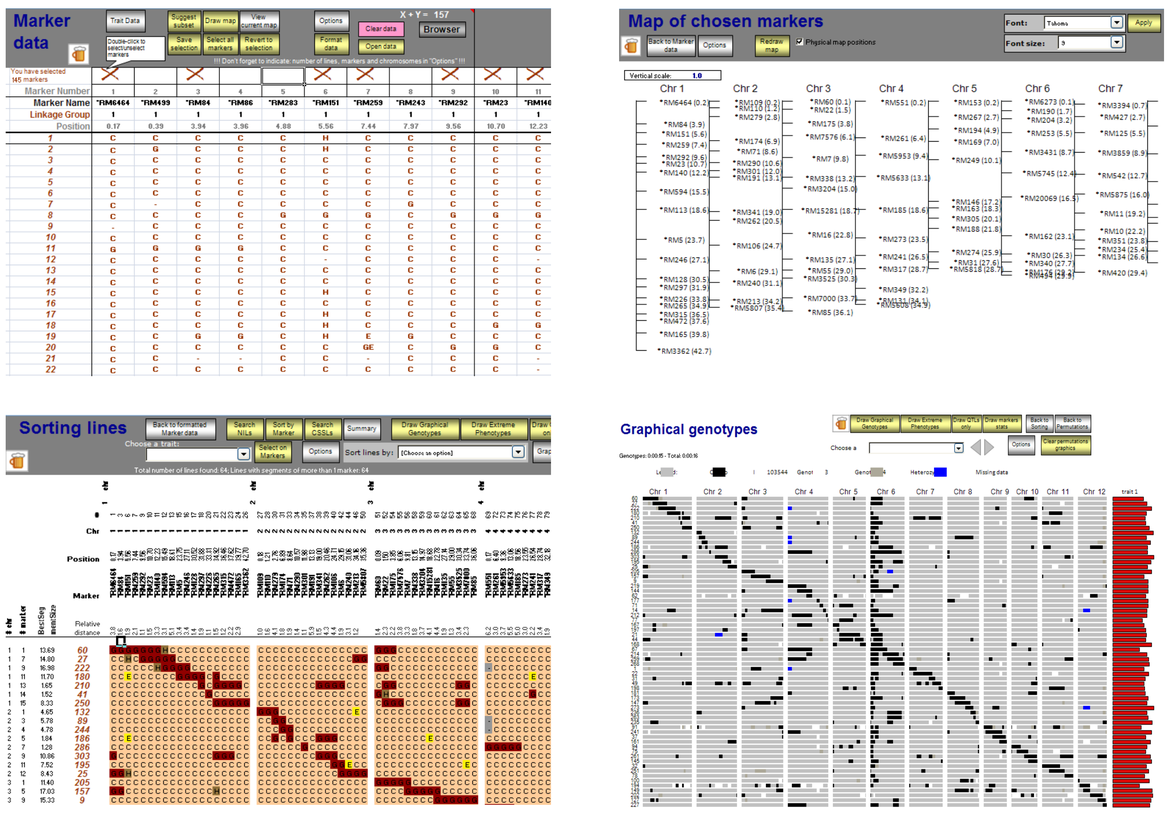


Figure 2. CSSL Finder screenshots. Top left: Input data matrix, with markers automatically selected to give a more even distribution. Top right: Genetic map showing marker positions. Bottom left: Selected lines in spreadsheet view. Bottom right: Selected lines in graphical genotype view.

Which lines do we want to select?

For a given segment, an optimal line is one which has little background coverage of the donor genome, in terms of number of donor segments and genomic extent of those segments. The chosen segments should ideally be of a uniform size, and overlap neighbouring segments in other lines by one (or more) markers.

Single Source Shortest Path Approach

Here, we propose a graph theoretic algorithm to improve the selection of CSSLs (Figure 3). Scores are generated for each donor segment in each line (Figure 3.2). A directed graph is constructed for each linkage group, where donor markers are represented by nodes, and marker nodes within a line are connected by weighted edges. Dijkstra's Single Source Shortest Path (SSSP) algorithm [2,3] is then used to select the path with least weight, representing the optimal path through the graph (Figure 2.4 and 2.5).

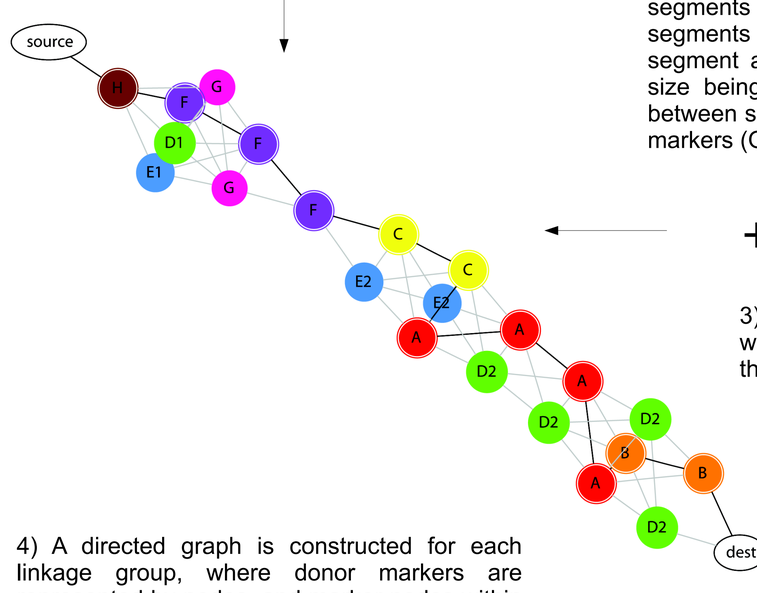
Figure 3. SSSP method

Pos	2.43	3.15	5.09	8.83	16.7	21.9	25.7	26.6	26.9	27.3
A	C	C	C	C	C	G	G	G	G	C
B	C	C	C	C	C	C	C	C	G	G
C	C	C	C	C	G	G	C	C	C	C
D	C	G	C	C	C	C	G	G	G	G
E	C	G	C	C	G	G	C	C	C	C
F	C	G	G	G	C	C	C	C	C	C
G	C	G	G	C	C	C	C	C	C	C
H	G	C	C	C	C	C	C	C	C	C

1) The data matrix is used to generate four scores

Line	Seg	Bgsegs	Bgcov	Sizediff	Overlap
A	1	0	0	-0.2	B:1 C:1 D2:3 E2:1
B	1	0	0	-7.1	A:1 D2:2
C	1	0	0	3.0	A:1 E2:2
D	1	1	3.9	-6.7	E1:1 F:1 G:1
	2	1	1.3	-4.1	A:3 B:2
E	1	1	11	-6.7	D1:1 F:1 G:1
	2	1	1.3	3.0	A:1 C:2
F	1	0	0	2.0	D1:1 E1:1 G:2
G	1	0	0	-3.8	D1:1 E1:1 F:2
H	1	0	0	-5.2	--

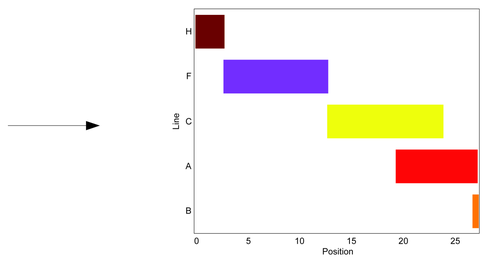
2) Scores are: number of background donor segments (Bgsegs); extent of those background segments (Bgcov); size difference between that segment and the ideal size (Sizediff), the ideal size being user defined; and the overlap size between segments in different lines, measured in markers (Overlap)



Parameters	Range
Bgsegs	0-1
Bgcov	0-1
Sizediff	0-1
Overlap	0-1

3) The scores are normalised, and weighted by the importance placed on them by the user

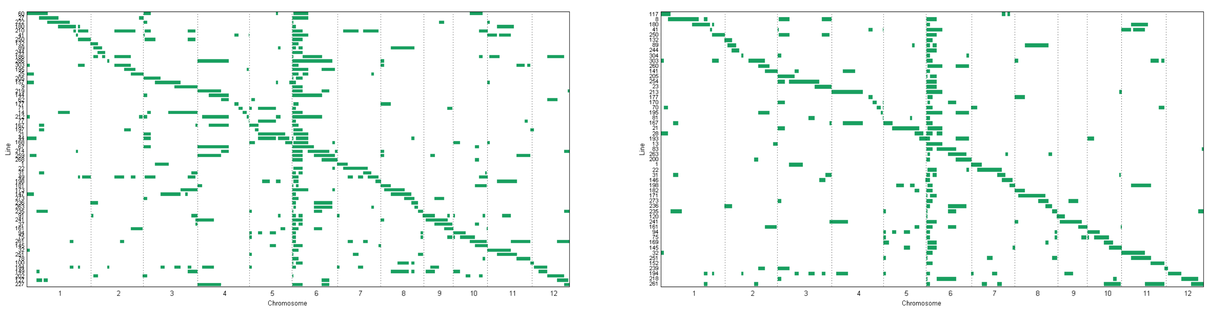
4) A directed graph is constructed for each linkage group, where donor markers are represented by nodes, and marker nodes within a line are connected by edges. Edges between neighbouring markers receive a weight of '0', whilst discontinuous markers receive a large weight. Edges are added between lines where donor segments overlap or are adjacent. These intra-line edges are weighted by the combination of Bgsegs, Bgcov, Sizediff and Overlap scores defined by the user. 'Source' and 'destination' nodes are also added to represent the start and end of the linkage group. Dijkstra's SSSP algorithm is used to select the path with least weight, shown here with black edges and circled nodes



5) The chosen lines are displayed graphically

Evaluation

A set of 312 introgression lines generated from a cross between the recipient *Oryza sativa L.* and the donor *Oryza Glaberrima Steud*, were genotyped at 200 marker loci [4]. The resulting data matrix was read into CSSL Finder and 145 loci were selected for further use. Missing data was inferred, reducing the percentage of missing data points from 4.7% to 0.2%. Line selection with the original greedy algorithm and the proposed SSSP algorithm was performed.



Property	Greedy algorithm	SSSP algorithm
Number of lines	64	53
Number of background segments	263	170
Background coverage (Mb)	1338.4	716.7
Average segment size difference from ideal (8Mb)	4.04	3.96
Overlapping segments	76%	66%
Average overlap (Mb)	1.2	0.78
Avoidable gaps	3	0

Figure 4. Evaluation. Top left: Selected lines from greedy algorithm. Top right: Selected lines from SSSP algorithm. Bottom: Table comparing properties of selected sets from both algorithms.

References

[1] Ebitani, T. *et al.* Construction and Evaluation of Chromosome Segment Substitution Lines Carrying Overlapping Chromosome Segments of *indica* Rice Cultivar 'Kasalath' in a Genetic Background of *japonica* Elite Cultivar 'Koshihikari'.(2005) Breeding Science 55: 65-73
[2] Dijkstra, E. W. A note on two problems in connexion with graphs. (1959) Numerische Mathematik 1: 269–271
[3] http://search.cpan.org/CPAN/authors/id/J/JH/JHI/Graph-0.94.tar.gz
[4] Gutierrez, A. *et al.* Identification of a Rice Stripe Necrosis Virus resistance locus and yield component QTLs using Oryza sativa x O. glaberrima introgression lines. (2010) BMC Plant Biology 10:6

Conclusions

The SSSP algorithm outperformed the original greedy heuristic, reducing the number of lines selected whilst covering more markers. Background donor segments were also reduced both in terms of number of segments and genomic extent.
We plan to implement this approach in a future release of CSSL Finder.