

International Center for Tropical Agriculture Consultative Group on International Agricultural Research

The shortest path to CSSL selection



Institut de recherche pour le développement

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

Sarah Ayling¹ and Mathias Lorieux^{1,2}

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.

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



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 BC₃F₁ 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.

Pos	2.43	3.15	5.09	8.83	16.7	21.9	25.7	26.6	26.9	27.3
А	С	С	С	С	С	G	G	G	G	С
В	С	С	С	С	С	С	С	С	G	G
С	С	С	С	С	G	G	С	С	С	С
D	С	G	С	С	С	С	G	G	G	G
Ε	С	G	С	С	G	G	С	С	С	С
F	С	G	G	G	С	С	С	С	С	С
G	С	G	G	С	С	С	С	С	С	С
Η	G	С	С	С	С	С	С	С	С	С

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
В	1	0	0	-7.1	A:1 D2:2
С	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
Ε	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
Н	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

dest

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

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

Marker	Trait Da	ita S	uggest subset Drav	v map Vie curren	w t map	Options	Cleard	X + Y =	157		
data 🧃	Double-cil select/uns markers	ick to elect	Save Sele election ma	ect all Reve rkers select	ert to stion	Format data	Open da	ata romosomes in '	'Options" !!!		
You have selected 145 markers	X		\mathbf{X}			X	X		\mathbf{X}	X	\mathbf{X}
Marker Number	1	2	3	4	5	6	7	8	9	10	11
Marker Name	'RM6464	"RM499	"RM84	"RM86	"RM283	'RM151	'RM259	'RM243	'RM292	"RM23	"BM14
Linkage Group	1	1	1	1	1	1	1	1	1	1	1
Position	0.17	0.39	3.94	3.96	4.88	5.56	7.44	7.97	9.56	10.70	12.23
1	С	С	С	С	С	н	С	С	С	C	С
2	С	G	С	С	С	н	С	С	С	C	С
3	С	С	С	С	С	С	С	С	С	С	С
4	С	С	С	С	С	С	С	С	С	C	C
5	С	С	С	С	С	С	С	С	С	C	C
6	С	С	С	С	С	C	С	С	C	C	C
7	С	-	С	С	С	C	С	G	C	C	C
8	С	С	С	С	G	G	G	С	G	G	G
9	-	C	С	C	С	C	C	C	C	C	C
10	С	C	C	C	C	C	C	C	C	C	C
11	G	G	G	G	C	C	C	C	C	C	C
12	С	C	C	C	C	-	C	C	C	C	-
13	С	C	C	C	C	C	C	C	C	C	C
14	С	C	C	C	C	C	C	C	C	C	C
15	С	C	C	C	C	H	C	C	C	C	C
16	С	C	C	C	C	C	C	C	C	C	C
17	С	C	C	C	C	н	C	C	C	C	C
18	С	C	C	C	C	н	C	C	C	G	G
19	С	C	G	G	C	н	E	G	C	C	C
20	C	C	C	C	C	C	GE	C	G	G	C
21	C	C	-	-	C	C	-	C	C	C	-
22	C	C	C	C	C	C	C	C	C	C	

Position Bit States Options Sort lines by: [Choose as option] Image: State St	S	orti	ing lin	Back to formatted Sea Marker data NI	rch Sort by Ls Marker	Search CSSLs	Summary	Draw Graphical Genotypes	Draw Extreme D Phenotypes	raw (on
Image: State	\sim		-	Choose a trait:	Select on	Options	Sort line:	s by: [Choose an option]		Grap
Indefinitive of lines found b4 Lines with segments of more than 1 marker b4 Image: State in the state					Markers					_
Image: Point of the second s				Total number of lines found: 64; Lines	with segments of	more than 1 m	iarker: 64			
N N			÷.		S.		÷.		ż	
			-	-	N				+	
Chr NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN						000000	1468 2			ထတ
Position ************************************			-		4 446966	000044				r- r-
Position 1<			Chr		- ~~~~~	~~~~~	1000	*****	~~ +++++++	**
Position Link Relative distance 1 1 1 1					P	~~~~~	യയയം .		≠wwww.ast:	
Marker No.2000 Strand Str			Position		2 8578886 8578886 86588886 86588886 86588888 86588888 86588888 86588888 86588888 86588888 86588888 86588888 865888888 865888888 86588888 865888888 865888888 86588888888	22222XX	18778 S	198955559225282		28
Marker Relative distance Relative Relative distance Relative Relative Relative distance Relative Rela								≈	8	
Marker ####################################					3995668			3223 233 434 55 3223 434 55 323 435 323 435 32 32 32 32 32 32 32 32 32 32 32 32 32		20
Image: Second condition Relative distance			Marker	*****************	t éfféff	ŔŔŔŔŔŔ	ÊÊÊÊ Â	1£¢¢¢¢¢¢¢¢¢¢¢	ÊÊ ÊÊÊÊÊÊÊÊ	řř.
Image: State and State	ł۳.									
Image: Solution of the second state in the	Ť	1Si2								
1 1 13.69 60	Ē	Ses Nen	Relative	39975-9355 5 5+9-9933	ဝယ္ကထ္ကန	-36330	100	*338385555555	N956999	20
1 1 13.69 60 <t< td=""><td></td><td></td><td>distance</td><td></td><td></td><td></td><td></td><td>-000002044220</td><td></td><td>() _</td></t<>			distance					-000002044220		() _
1 7 14.80 27 CCHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	1	13.69	60	GGGGGGGHCCCCCCCCCC	c ccccc	cccccc	cccc (GGGCCCCCCCCC	cc cccccc	СС
1 9 16.98 222 CCCCCCCHARCECCCCCCCCCCCCCCCCCCCCCCCCCCCC	7	14.80	27	CCHC <mark>GGGGGGCC</mark> CCCCCCC	cccccc	cccccc	CGGG	<u>cc</u> ccccccccc	cc <u>c</u> cccccc	CC
1 11 11.70 180 CCECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	9	16.98	222	CCCCCCCHGGGGCCCCCCCCC	cccccc	ccccc	ccc I	GG CCCCCCCCCC	cc cccccc	CC
1 13 1.65 210 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	11	11.70	180	CCECCCCCC <mark>GGGG</mark> C <mark>GCCC</mark>	cccccc	.cc <u>cccc</u>	ccc (c <u>cc</u> ccccccc <u>c</u>	<u>c</u> c cccccc <mark>e</mark>	CC
1 14 152 41 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	13	1.65	210	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	c ccccc	CCGGGGG	CCC (C <mark>GG</mark> CCCCCCCC <mark>G</mark>	CCCCCCC	CC
1 15 8.33 250 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	14	1.52	41	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	cccccc	ccccc	CCCC I	<u>Hcc</u> cccccc <u>c</u>	cc cccccc	CC
2 1 4.65 132 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	15	8.33	250	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCCCCC	CCCCCC	CCCC (C <mark>GGG</mark> CCCCCCC <mark>G</mark>	CCCCCCC	CC
2 3 5.78 89 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	1	4.65	132	CCCCCCCCCCCCCCCCCCC	GGGCCC	ccccc	CC <mark>E</mark> C (CCCCCCCCCCCC	CC CCCCCC	CC
2 4 4.78 244 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	3	5.78	89	CCCCCCCCCCCCCCCCCCC	C CC <mark>GG</mark> CC	CCCCCC	CCCC (ccccccccccc	cc -cccccc	CC
2 5 1.84 1.86 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	4	4.78	244	CCCCCCCCCCCCCCCCCCC	CCCGGC	ccccc	CCCC (ccccccccccc	cc cccccc	CC
2 7 1.28 286 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5	1.84	186	CCECCCCCCCCCCCCCC	C C C C C C C C	CCGGGG	CCCC (CCCCCCC <mark>E</mark> CCCC	cc cccccc	CC
2 9 10.86 30.3 BCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	7	1.28	286	CCCCCCCCCCCCCCCCCC	CCCCCC	GCCCCC	CCCC (ccccccccccc	CC GGGGGCC	CC
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	9	10.86	303	GCCCCCCCCCCCCCGGGGC	000000	CCGGGG	CCC	ccccccccccc	00 0000000	CC
2 12 8.43 25 GGHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	11	7.52	195	000000000000000000000000000000000000000	000000	CCCCC	ECC		EC CCCCCCC	CC
3 1 11.40 205 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	12	8.43	25	66H0000000000000000	000000	00000	S G G G	000000000000000000000000000000000000000	222222 22	C C
3 5 17.03 157 DECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	1	11 40	205	000000000000000000000000000000000000000		000000	2000		2222222 22	0.0
	5	17.02	157			000000				00
- X N 1533 - V - TELEVISIONISTICICUUM DISCONSIGNISTICUUM ENVERENCE CONCORDE COOPER	ğ	15.32	10			000000			0000000 00	C C









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.





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.

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] <u>http://search.cpan.org/CPAN/authors/id/J/JH/JHI/Graph-0.94.tar.gz</u>
- [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

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.

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.