PHÁT TRIỂN CÁC ĐẦU DÒ ĐẢO NGƯỢC PHÂN TỬ CHO KIỂU GEN CHỌN LỌC BỘ GEN CỦA CÂY ĐẬU TƯƠNG

Haichuan Wang[1], Benjamin Campbell[2], Mary Happ[1], Samantha McConaughy[1], Aaron Lorenz[2], Keenan Amundsen[1], Qijian Song[3], Vincent Pantalone[4], David Hyten[1]

1. Khoa Nông học và Làm vườn, Đại học Nebraska–Lincoln, Lincoln, NE, Mỹ
2. Khoa Nông học và Di truyền thực vật, Đại học Minnesota,St.Paul,MN,USA
3. USDA–ARS, Phòng thí nghiệm cải tiến và gen đậu tương, Beltsville, MD, Mỹ
4. Khoa Khoa học Thực vật, Đại học Tennessee, Knoxville, TN, Mỹ

TÓM TẮT

Việc tăng tỷ lệ di truyền cho các đặc điểm nông học quan trọng thông qua chọn lọc bộ gen đòi hỏi phải phát triển các phương pháp phân tử mới để chạy các đơn nucleotide đa hình trên toàn bộ bộ gen (SNPs). Hạn chế chính của các phương pháp hiện tại là chi phí quá cao để sàng lọc các quần thể chọn giống. Các đầu dò đảo ngược phân tử (MIP) là một phương pháp xác định kiểu gen theo trình tự (GBS) được nhắm mục tiêu có thể được sử dụng cho cây đậu tương [*Glycine max* (**L.**) **Merr.**] vừa tiết kiệm chi phí, thông lượng cao và cung cấp chất lượng dữ liệu cao sàng lọc tế bào mầm của nhà chọn giống để chọn lọc bộ gen. Một bộ 1K MIP SNP đã được phát triển cho cây đậu tương với các marker phân bố đồng đều trên toàn bộ bộ gen. Các SNP đã được chọn để tối đa hóa số lượng các marker thông tin trong tế bào mầm đang được thử nghiệm trong các chương trình chọn giống đậu tương ở các vùng trung tâm phía Bắc và trung tâm phía Nam của Mỹ. Bộ 1K SNP MIP đã được thử nghiệm trên các nguồn tế bào mầm đa dạng và quần thể dòng lai tái tổ hợp (RIL). Trình tự được nhắm mục tiêu với MIP đã thu được mức độ làm giàu 85% cho SNP được nhắm mục tiêu. Độ chính xác của kiểu gen MIP là 93% về tổng thể, trong khi độ chính xác của đồng hợp tử là 98% với dữ liệu bị thiếu < 10%. Độ chính xác của MIP kết hợp với chi phí cho mỗi mẫu thấp làm cho nó trở thành một công cụ mạnh mẽ để cho phép lựa chọn bộ gen trong các chương trình chọn giống đậu tương.

1. GIỚI THIỆU

Việc tạo kiểu gen trên toàn bộ bộ gen của các dòng tế bào mầm của đậu tương [*Glycine max* (**L.**) **Merr.**] đã được sử dụng cho nhiều ứng dụng, chẳng hạn như lập bản đồ QTL trong quần thể lưỡng cực, chọn lọc nền cho tính trạng xâm nhập và chọn lọc bộ gen trên các dòng thế hệ con cháu. Các ứng dụng này thường sử dụng hơn 100–6.000 marker phân tử (Basnet và ctv, 2022; Hyten và ctv, 2004; Kim và ctv, 2008; Sebastian và ctv, 2012). Từ đầu những năm 2000, các chỉ thị đơn nucleotide đa hình (SNP) bắt đầu được sử dụng trong các ứng dụng chọn giống đậu tương (Lee và ctv, 2004; Song và ctv, 2004; Zhu và ctv, 2003). Mặc dù loại marker vẫn nhất quán, nhưng công nghệ tạo kiểu gen SNPs không ngừng phát triển để giảm chi phí tạo kiểu gen cho từng mẫu riêng lẻ với một bộ marker phân tử trên toàn bộ gen (Elshire và ctv, 2011; Hyten và ctv, 2008; Lee và ctv, 2004; Song và ctv, 2013).

Một trong những công nghệ đầu tiên cho phép chạy một tập hợp lớn các chỉ thị phân tử trên chất mầm đậu tương là xét nghiệm GoldenGate (Hyten và ctv, 2008). Sự phát triển của Universal SNP Linkage Panel (USLP) 1.0 đã cho phép các quần thể đậu tương lập bản đồ được xác định kiểu gen với 1.536 marker trong khoảng thời gian 3 ngày (Hyten và ctv, 2010). Bảng điều khiển được thiết kế sao cho một tỷ lệ cao trong số 1.536

marker đó sẽ là đa hình trong bất kỳ quần thể lập bản đồ lưỡng cực nào. Trong các quần thể lưỡng cực với các giống đậu tương ưu tú là bố mẹ, USLP 1.0 có trung bình 458 chỉ thị đa hình (cung cấp thông tin) phân tách trong quần thể (Hyten và ctv, 2010). Số lượng chỉ thị đa hình này cho phép đủ độ bao phủ trong các quần thể đậu tương khác nhau để lập bản đồ QTL hiệu quả (Kim và ctv, 2012; Phansak và ctv, 2016; Warrington và ctv, 2015). Mặc dù hầu hết các marker là đơn hình, nhưng việc tiết kiệm chi phí cho mỗi mẫu để thu được 458 marker đa hình rẻ hơn so với chạy từng marker riêng lẻ.

Khi chi phí giải trình tự giảm do các kỹ thuật giải trình tự thế hệ thứ hai, các phương pháp như giải trình tự kiểu gen theo trình tự enzyme hạn chế (reGBS) đã được phát triển để thu thập thông tin về kiểu gen (Elshire và ctv, 2011). Enzyme giới hạn GBS có cả ưu điểm và nhược điểm khi so sánh với USLP 1.0. Lợi ích lớn nhất của reGBS là giảm đáng kể chi phí cho mỗi mẫu. Điều này cho phép sử dụng reGBS cho nhiều ứng dụng bộ gen như lập bản đồ QTL, lựa chọn bộ gen và phân tích tính đa dạng (Ravelombola và ctv, 2021). Mặc dù chi phí của reGBS rẻ hơn so với mảng kiểu gen, nhưng chất lượng dữ liệu đã trở thành một vấn đề. Enzyme hạn chế GBS có dữ liệu bị thiếu đáng kể do mức độ bao phủ thấp của cùng một SNP trên các mẫu. Để giảm thiểu dữ liệu bị thiếu này, các phương pháp quy nạp đã được áp dụng để điền vào dữ liệu còn thiếu (Wickland và ctv, 2017). Trong khi đó, mức độ bao phủ đọc thấp của mỗi SNP cũng khiến các alen dị hợp tử không được gọi trong các bộ dữ liệu cuối cùng ngay cả sau khi cắt bỏ (Nazzicari và ctv, 2016). Chọn giống đậu tương thường tạo ra các dòng ở thế hệ F4, trung bình sẽ có 12% dị hợp trong các dòng lai tạo của chúng. Tính không đồng nhất này có thể sẽ được xác định bằng cách sử dụng reGBS.

Do reGBS có tỷ lệ dữ liệu bị thiếu cao và gọi là alen dị hợp tử nên chip 6K Illumina Infinium (BARCSoySNP6K) vẫn được cộng đồng đậu tương tích cực sử dụng (Beche và ctv, 2021; Li và ctv, 2021; Song và ctv, 2020). Con chip này cho phép các nhà nghiên cứu thu được dữ liệu bị thiếu thấp với độ chính xác cao khi gọi các dị hợp tử và có quy trình làm việc dễ dàng để xử lý mẫu và phân tích dữ liệu. Mặc dù BARCSoySNP6K có chi phí cao hơn reGBS, nhiều nhà nghiên cứu và nhà chọn giống đậu tương vẫn sẵn sàng sử dụng phương pháp này để tăng chất lượng dữ liệu và phân tích dữ liệu đơn giản.

Ý tưởng cốt lõi

Đầu dò đảo ngược phân tử (MIP) cung cấp tùy chọn kiểu gen thông lượng cao, chi phí thấp để tạo kiểu gen mật độ thấp cho đậu tương.

Một bộ 1K SNP đã được chọn để tối đa hóa tính thông tin trong các chương trình chọn giống đậu tương.

MIPs là một công nghệ tạo kiểu gen lý tưởng cho việc lựa chọn bộ gen.

Dữ liệu chất lượng cao với khả năng tiết kiệm chi phí của giải trình tự thế hệ thứ hai có thể thu được bằng các phương pháp GBS được nhắm mục tiêu (Teer và ctv, 2010). Các phương pháp GBS được nhắm mục tiêu có thể làm phong phú thêm cho một số đoạn nhỏ được biết là có chứa SNP. Điều này cho phép sắp xếp có chọn lọc các đoạn được làm giàu đến độ sâu đọc cao, tạo ra chất lượng dữ liệu đủ cao cho các ứng dụng của nhà tạo giống (Mamanova và ctv, 2010). Độ sâu trình tự cao này tại các locus được nhắm mục tiêu tạo ra đủ số lần đọc để giảm đáng kể dữ liệu bị thiếu đến mức không cần cắt bỏ và có thể xác định chính xác các alen dị hợp tử (Teer và ctv, 2010).

Một phương pháp GBS được nhắm mục tiêu đã được áp dụng thành công ở nhiều loài là các đầu dò đảo ngược phân tử (MIP) (Niedzicka và ctv, 2016; Porreca và ctv, 2007; Turner và ctv, 2009). Nói chung, một MIP chứa một nhánh mở rộng và nhánh thắt bổ sung cho chuỗi sườn xung quanh SNP được nhắm mục tiêu. Phần mở rộng và vòng thắt được kết nối với trình tự xương sống được sử dụng cho bước phản ứng chuỗi polymerase vạn năng (PCR). Thông qua bước lai–mở rộng–thắt, phần mở rộng và nhánh thắt của mẫu thăm dò với trình tự mục tiêu. Phần mở rộng xảy ra từ nhánh mở rộng trên chuỗi được nhắm mục tiêu. Polymerase được sử dụng không có hoạt tính exonuclease 5′ đến 3′, do đó phần mở rộng dừng lại ở nhánh thắt. Quá trình thắt diễn ra giữa DNA mở rộng và nhánh thắt MIP để tạo vòng tròn cho MIP. Sau đó, các đoạn mồi PCR phổ quát bổ sung cho đường trục MIP sẽ khuếch đại MIP được tuần hoàn hóa dưới dạng một bộ khuếch đại tuyến tính. PCR phổ quát sẽ khuếch đại nhiều MIP được tuần hoàn hóa có chứa cùng một đường trục nhưng nhắm mục tiêu vào các SNP khác nhau. Đối với nhiều mẫu, PCR này được sử dụng để thêm mã vạch kép để nhận dạng mẫu và trình tự Illumina cần thiết để giải trình tự trên nền tảng Illumina. Phương pháp này đã được chứng minh là có thể nhắm mục tiêu SNP từ một vài đến 55.000 trong một phản ứng MIP duy nhất với độ chính xác và hiệu quả chi phí cao (Turner và ctv, 2009; Waalkes và ctv, 2018).

Lựa chọn bộ gen trong các dòng thế hệ con cháu yêu cầu chi phí cho mỗi mẫu cực thấp để có được các marker toàn bộ bộ gen thông tin trong quần thể được chọn. Những dòng con cháu này sẽ đến từ nhiều quần thể với các tổ hợp bố mẹ khác nhau. Khi bố mẹ của các quần thể lưỡng cực đã được tạo kiểu gen với mật độ chỉ thị cao thì một bộ SNP phổ quát giống như bộ được phát triển cho GoldenGate USLP 1.0 có thể được sử dụng trong chọn lọc bộ gen và có hiệu quả trên nhiều quần thể có nguồn gốc từ các cặp bố mẹ khác nhau. Bằng cách phát triển một bộ SNP phổ quát gồm 1.000 marker SNP, các nhà lai tạo có thể sử dụng nó với phương pháp GBS được nhắm mục tiêu, chẳng hạn như MIP, để thực hiện chọn lọc bộ gen trong các dòng đậu tương. Số lượng các marker thông tin thu được từ bộ 1.000 SNP sẽ đủ để quy ra mật độ cao hơn của các marker từ bố mẹ sang con cái (Gonen và ctv, 2018). Điều này sẽ cho phép các nhà lai tạo giảm chi phí tạo kiểu gen bằng cách chạy cùng một bộ SNP trên nhiều quần thể sinh sản. Mục tiêu của chúng tôi là chứng minh rằng MIPs có thể được sử dụng như một phương pháp GBS được nhắm mục tiêu cho kiểu gen SNP mật độ thấp và phát triển một bộ MIPs SNP phổ quát có thể được sử dụng cho các chương trình chọn giống đậu tương tập trung vào chọn giống cho các khu vực trung tâm phía bắc và phía nam nước Mỹ.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

Tổng cộng có 96 mẫu giống từ Bộ sưu tập mầm đậu tương của USDA (USDA, Trạm nghiên cứu nông nghiệp, Đại học Illinois, Urbana, IL) và 96 dòng lai tái tổ hợp (RIL) từ phép lai của Essex × Williams82 (ExW82) đã được sử dụng trong nghiên cứu này (Bảng bổ sung S1). Tất cả các dòng trước đây đã được tạo kiểu gen với chip SoySNP50K Infinium (Song và ctv, 2015, 2016). Mô lá được thu thập từ một cây từ mỗi trong số 96 mẫu cấy và 20 đĩa lá được lấy từ năm cây từ mỗi dòng tự nhiên đối với các RIL ExW82. Mô lá được đông khô sau đó tách chiết DNA từ các mô lá bằng phương pháp CTAB như được mô tả bởi Happ và ctv (2019). Các mẫu DNA được định lượng bằng hệ thống Quantifluor DSDNA (Promega) trên đầu đọc đĩa Synergy 2 (BioTek) bằng cách làm theo hướng dẫn của nhà sản xuất.

Các marker SNP có trong bộ SNP 1K Soy MIPs là một tập hợp con của chip BARCSoySNP6K hiện có (Song và ctv, 2020). Mục tiêu là chọn 1.000 SNP vừa đạt được độ bao phủ bộ gen đồng nhất vừa tối đa hóa số lượng SNP đa hình trong bất kỳ quần thể

chọn giống lưỡng cực nào có thể được hình thành bằng cách lai các dòng lai ưu tú từ các chương trình chọn giống đậu tương tham gia ở Northern Unform Soybean Tests. Một số SNP đa hình đã được đánh giá bằng cách sử dụng một tập hợp con đại diện gồm 1.285 dòng lai ưu tú được đưa vào Northern Unform Soybean Tests của USDA trong 10 năm qua. Mỗi dòng được tạo kiểu gen với BARCSoySNP6K. Tám mươi bốn SNP đã bị xóa vì thiếu > 20% dữ liệu hoặc có tần số alen phụ <5%. Bốn mươi mốt SNP đã bị loại bỏ vì chúng không lập bản đồ tới nhiễm sắc thể trong tổ hợp bộ gen Phiên bản 2. Tổng cộng có 5.875 SNP còn lại để phân tích xuôi dòng và là những ứng cử viên để đưa vào bộ SNP 1K Soy MIP. Vị trí di truyền của SNP trên BARCSoySNP6K được thực hiện theo bản đồ di truyền ExW82 do Song và ctv (2020). Các vị trí di truyền của SNP không được định vị trên bản đồ di truyền ExW82 được nội suy bằng cách sử dụng làm mịn biểu đồ phân tán được ước tính cục bộ trong R liên quan đến các vị trí vật lý với các vị trí di truyền.

Chương trình đại học và số dòng được tạo kiểu gen cho mỗi chương trình được hiển thị trong Bảng 1. Tất cả các phép lai lưỡng cực có thể có giữa tất cả các dòng, cũng như giữa tất cả các dòng trong một chương trình, đã được mô phỏng và số lượng SNP đa hình cho mỗi phép lai có thể được liệt kê cho từng phương pháp . Sáu phương pháp khác nhau để lựa chọn SNP đã được sử dụng và so sánh: Random, Haploview_BIN, Haploview_Tagger, BIN_PIC, Song_PIC_MIN_0.1 và Song_PIC_MIN_0.2.

1. Random: Một bộ gồm 1.000 SNP được chọn ngẫu nhiên từ BARCSoySNP6K. Việc lấy mẫu SNP ngẫu nhiên được lặp lại 10 lần.

2. Haploview_BIN: Mỗi nhiễm sắc thể được chia thành các ô có chiều dài 5 cM để có tổng số 501 ô. Phần mềm Haploview (Barrett và ctv, 2005) đã được sử dụng để phân tích SNP trong mỗi thùng và chọn SNP gắn thẻ các alen haplotype nhất. Cuối cùng, số lượng thùng có SNP được chọn là 492, với 453 thùng có hai SNP và 39 thùng chỉ có một SNP.

3. Haploview_Tagger: Số marker thẻ được chỉ định cho mỗi nhiễm sắc thể dựa trên chiều dài di truyền của mỗi nhiễm sắc thể. Chương trình Haploview 4.2 Tagger đã được sử dụng để chọn SNP bằng cách sử dụng logarit của ngưỡng tỷ lệ chênh lệch là 3.0 (Barrett, 2009).

4. BIN_PIC: Mỗi nhiễm sắc thể được chia thành các ô có chiều dài 5 cM để có tổng số 501 ô. hàm lượng thông tin đa hình (PIC) của mỗi SNP trong mỗi chương trình trong số 14 chương trình chọn giống được tính toán là $PIC = 1 - [(p^2 + q^2) - 2p^2q^2]$ (Botstein và ctv, 1980), trong đó p là tần số alen chính và q là tần số alen phụ. Điểm PIC trung bình có trọng số được tính trên tất cả các chương trình chọn giống, với trọng số là số dòng trong mỗi chương trình chọn giống (Bảng 1). Hai SNP trong mỗi thùng có PIC có trọng số cao nhất đã được chọn. Bốn thùng chỉ chứa một SNP cho tổng số 998 SNP được chọn.

5. Song_PIC_MIN_0.1 và Song _PIC_MIN_0.2: Song và ctv. (2013) đã triển khai một thuật toán để chọn SNP vừa có điểm chất lượng kiểu gen cao vừa bao phủ đồng nhất bộ gen với mật độ khác nhau cho các vùng đồng nhất và dị sắc. Phương pháp này đã tính điểm chỉ số kết hợp điểm thiết kế Illumina và "điểm ưu tiên". Thuật toán lặp được phát triển để chọn một số SNP nhất định có khoảng cách vật lý cụ thể giữa chúng, với khoảng cách vật lý được xác định bằng tỷ lệ giữa khoảng cách bản đồ di truyền với khoảng cách bản đồ vật lý trong các vùng đồng sắc và dị sắc. Mật độ cần thiết của các SNP đã chọn được xác định là lớn hơn năm lần ở các vùng đồng màu so với các vùng dị sắc. Khoảng cách giữa các SNP đã được xác

định và đối với khoảng cách nhỏ nhất trên bất kỳ nhiễm sắc thể nhất định nào, "điểm chỉ số lựa chọn SNP" đã được tính cho các marker sườn của khoảng cách. SNP có điểm chỉ số nhỏ nhất đã bị xóa. Quy trình này được lặp đi lặp lại cho đến khi số lượng SNP mong muốn được chọn cho từng vùng đồng nhất và dị sắc trên mỗi nhiễm sắc thể. Thông tin chi tiết về quy trình này được cung cấp trong Song và ctv (2013). Nghiên cứu này đã triển khai thuật toán tương tự với ngoại lệ là "điểm số chỉ số lựa chọn SNP" đã được thay thế bằng PIC. Giá trị PIC tối thiểu là 0,1 và 0,2 đã được sử dụng cho các phương thức Song_PIC_MIN_0.1 và Song_PIC_MIN_0.2 tương ứng.

Bảng 1. Số lượng trung bình các SNP trong bộ MIP 1K có thể được cho là đa hình đối với các thể lưỡng cực liên quan đến các tổ hợp khác nhau cho từng nguồn gen của chương trình chọn giống đậu tương

Chương trình	Số dòng	Ngẫu nhiên	Haploview_BIN	Haploview_Tagger	BIN_PIC	Song_PIC_MIN_0.1	Song_PIC_MIN_0.2
Đại học Guelph	31	293	288	293	383	375	387
Đại học Illinois	328	293	286	300	421	405	417
Đại học IowaState	97	280	289	298	421	407	422
Đại học KansasState	36	253	269	274	386	369	380
Đại học State Michigan	51	293	288	302	412	400	414
Đại học Minnesota	354	322	323	332	443	434	448
Đại học Nebraska	51	297	294	305	412	400	414
Đại học Ohio State	38	242	249	257	351	342	355
Đại học Purrdue	25	269	269	277	405	387	400
Đại học Guelph-Ridgetown	13	286	308	312	421	409	426
Đại học South Dakota State	33	292	294	304	409	402	418
Đại học Southern Illinois	9	240	244	259	338	335	347
Đại học Tennessce	3	240	229	235	352	343	353

MIPgen (Boyle và ctv, 2014) đã được sử dụng để thiết kế một bộ gồm 998 MIP trên toàn bộ bộ gen đậu tương. Marker 998 (được gọi là 1K) là một tập con của Chip Infinium BARCSoySNP6K và Chip Infinium SoySNP50K (Song và ctv, 2020; Song và ctv, 2013). 998 SNP đã được chọn bằng phương pháp BIN_PIC được mô tả ở trên. Các MIP được thiết kế để ghi lại chuỗi 200 bp với SNP được nhắm mục tiêu ở giữa chuỗi và được lọc dựa trên điểm hậu cần. Chức năng địa lý trong R được sử dụng để ước tính vị trí centimorgan cho các vị trí BARKSoySNP6K dựa trên vị trí centimorgan từ quần thể ExW82 (Song và ctv, 2016).

998 MIP được tổng hợp riêng lẻ bởi Công nghệ DNA tích hợp (Coralville, IA, Mỹ) ở đầu dò 0,2 nmole/probe (Bảng bổ sung S2). Các mẫu dò được pha loãng theo tỷ lệ 1:10 với nước không có nuclease. Các đầu dò sau đó được gộp lại với nhau ở phân tử bằng nhau. Để phosphoryl hóa các đầu dò gộp lại, một bộ T4 polynucleotide kinase (ThermoFisher) đã được sử dụng bằng cách làm theo hướng dẫn của nhà sản xuất. Tóm lại, một phản ứng phosphoryl hóa 20µl bao gồm 15µl mẫu dò gộp pha loãng, bộ đệm ligase DNA 1× T4, 1mM ATP và 0,5 U/µl của T4 polynucleotide kinase. Phản ứng được ủ ở 37^0C trong 30 phút, sau đó là vô hiệu hóa kinase ở 75^0C trong 10 phút.

Việc lai tạo, mở rộng đầu dò và thắt được thực hiện trong một bước duy nhất. Phản ứng 6,34-µl mỗi giếng được thực hiện trong đĩa PCR 384 giếng. Nó bao gồm các pipet riêng lẻ 3,34µl mẫu DNA 12 ng/µl (~ 40 ng bộ gen DNA) vào các giếng riêng lẻ của tấm PCR 384 giếng. Hỗn hợp chính 3µl bao gồm 0,88µl mẫu dò phosphoryl hóa gộp (~1.000:1 DNA so với tỷ lệ mẫu dò), bộ đệm ligase DNA 1× Ampligase (Epicentre Technologies), 0,5µl Betain 5 M (Sigma Aldrich), DNA 3,25 U AmpliTaq polymerase (ThermoFisher), 16 U của DNA Ampligase (Epicentre Technologies), 0,016mM dNTP và 0,08µl nước. Hỗn hợp chính được phân phối vào từng giếng trong số 384 giếng bằng cách sử dụng Bộ xử lý chất lỏng Mantis (Formulatrix). Phản ứng được làm nóng ở 98^0C trong 10 phút, sau đó ủ liên tục ở 60^0C trong ~ 24 giờ. Nắp gia nhiệt cho máy PCR được đặt ở 105^0C trong 30 phút đầu tiên và sau đó thay đổi thành 80^0C sau 30 phút ủ đầu tiên.

Sau bước lai–mở rộng–thắt, hỗn hợp 2µl gồm 6,4 U của Exo I (New England Biolabs), 32U của Exo III (New England Biolabs), bộ đệm ligase DNA 1× Ampligase và 1,16µl nước được tự động phân phối vào từng giếng trong số 384 giếng bằng cách sử dụng Bộ xử lý chất lỏng Mantis (Formulatrix) để tiêu hóa các MIP và mẫu DNA không tuần hoàn. Phản ứng được ủ ở 37^0C trong 30 phút sau đó là 95^0C trong 2 phút.

Quá trình khuếch đại PCR của các MIP được tuần hoàn hóa diễn ra riêng lẻ cho từng mẫu bằng cách sử dụng 1,25µl ảnh chụp MIP được tuần hoàn hóa được lấy từ bước trước đó, 0,5µl của 10-µM đoạn mồi tiến và lùi được lập chỉ mục (Bảng S3 bổ sung) và 3 µl của 2× iProof HF Master Mix (BioRad). 3µl hỗn hợp chính 2 × được phân phối bằng cách sử dụng Bộ xử lý chất lỏng Mantis (Formulatrix). PCR được thực hiện với các điều kiện sau trong đĩa 384 giếng: 98^0C trong 30 giây; 21 chu kỳ 98^0C trong 10 giây, 60^0C trong 30 giây, 72^0C trong 30 giây; và 72^0C trong 60 giây, giữ ở 12^0C.

Tất cả các sản phẩm PCR từ bước trước được gộp lại bằng cách trộn 1 µl/mẫu/ giếng, sau đó các thư viện hỗn hợp được phân tách trên gel agarose 1,5% trong 45 phút ở 95 V cùng với thang DNA 50-bp (Fisher) và 100-bp (GoldBio). Dải sắc nét ở ~ 286 bp đã được cắt bỏ và tinh chế bằng bộ chiết gel (Qiagen) bằng cách làm theo hướng dẫn của nhà sản xuất. Ngoài ra, các sản phẩm PCR gộp được tinh sạch bằng bộ dụng cụ làm sạch PCR sbeadex (LGC, Biosearch Technologies) bằng cách làm theo hướng dẫn của nhà sản xuất để tinh sạch dải PCR ở 286bp. Sản phẩm PCR tinh khiết được định lượng trên Qubit (ThermoFisher) với Bộ xét nghiệm Qubit RNA HS (ThermoFisher) trước, sau đó được định lượng chính xác hơn nữa trên QuantStudio 6 plex (ThermoFisher) bằng cách sử dụng bộ định lượng thư viện KAPA nền tảng Illumina (Roche).

Dựa trên số lượng từ PCR định lượng, các thư viện gộp tinh khiết ở 1,5pM đã được nạp vào thiết bị Illumina NextSeq 500/550 với bộ công cụ Mid Output v2.5 (150 chu kỳ) (Illumina) bằng cách sử dụng đoạn mồi trình tự tùy chỉnh (cả đoạn mồi chỉ số và đoạn mồi khuếch đại) (Bảng bổ sung S4). Lần chạy giải trình tự bao gồm lần đọc đơn 150bp cho chỉ số 5 và lần đọc 8 bp cho chỉ số 7. Các đầu ra giải trình tự được xử lý trên Trung tâm Điện toán Hà Lan, UNL với phần mềm chuyển đổi bcl2fastq2 (v2.17, Illumina), kết hợp đầu ra giải trình tự chính (tệp BCL) từ một lần chạy và chuyển đổi cũng như phân tách chúng thành các tệp FASTQ dựa trên các chỉ số được liên kết với từng mẫu.

Trước khi căn chỉnh các lần đọc theo trình tự tham chiếu, các lần đọc thô được lọc bằng Trimmomatic để loại bỏ ô nhiễm trình tự bộ điều hợp, các lần đọc bị cắt bớt và các lần đọc với chất lượng cơ bản tổng thể thấp (Bolger và ctv, 2014). Các lần đọc được lọc sau đó được căn chỉnh theo bộ gen tham chiếu tùy chỉnh, bao gồm các vùng sườn 200bp

xung quanh mỗi SNP được nhắm mục tiêu cho mỗi trong số 998 SNP sử dụng Bowtie 2 (Langmead & Salzberg, 2012) ở chế độ end-to-end nhạy cảm thiết lập (Niedzicka và ctv, 2016). Sau khi căn chỉnh, các lần đọc có điểm chất lượng bản đồ <5 đã bị loại bỏ bằng Samtools. Các kiểu gen SNP sau đó được gọi bằng các công cụ GATK4 HaplotypeCaller và GenotypeGVCF, trong đó các cuộc gọi có điểm tin cậy <20 đã bị xóa.

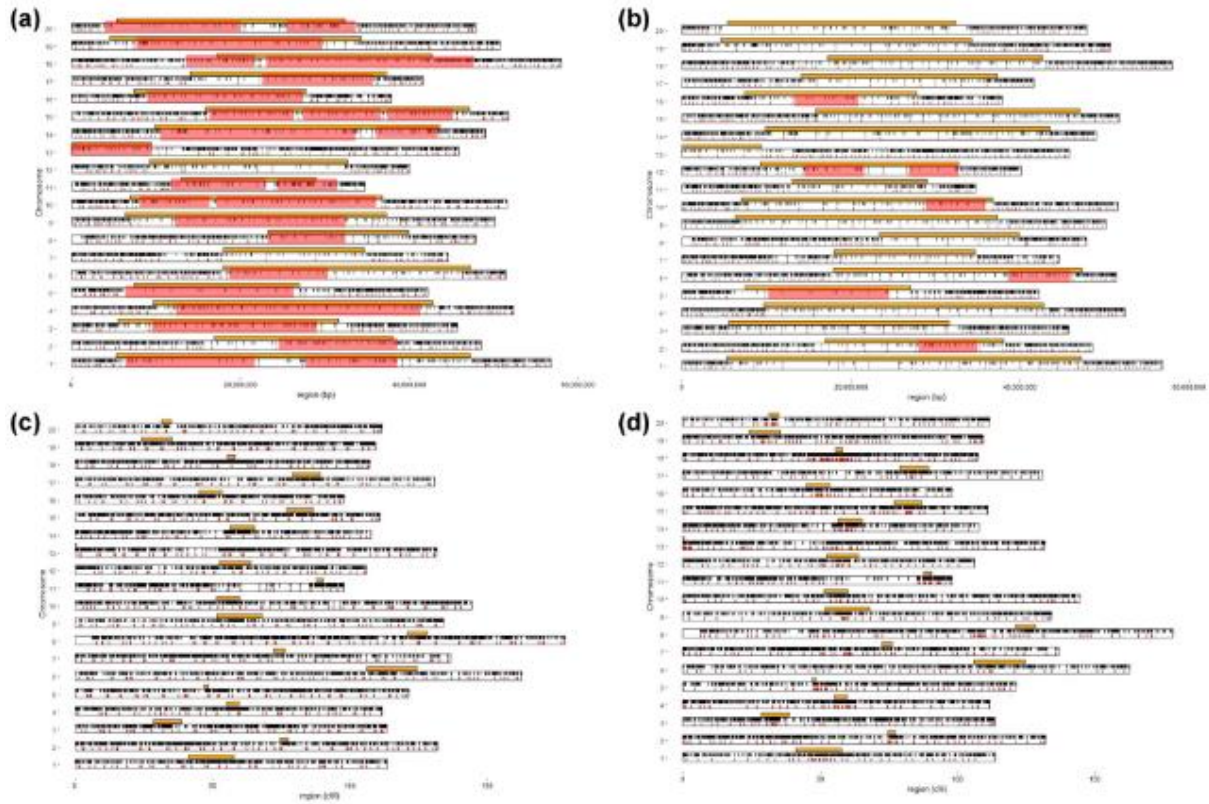
3. KẾT QUẢ

3.1 Lựa chọn SNP cho bộ đầu dò

Nhiều phương pháp đã được khám phá để lựa chọn một bộ 1.000 SNP tối ưu có thể tạo nên bộ MIP SNP để lựa chọn bộ gen. Vật liệu mầm mục tiêu cho bộ 1K SNP là các dòng lai ưu tú đến từ các chương trình chọn giống đậu tương trên khắp các vùng sản xuất đậu tương trung tâm phía Bắc và Nam của Mỹ. Lý tưởng nhất là bộ thăm dò 1K này sẽ tối đa hóa số lượng chỉ thị đa hình trong mỗi phép lai tạo giống tiềm năng và thậm chí có phạm vi bao phủ trên toàn bộ bộ gen. Điều này sẽ cung cấp đủ các marker thông tin cho từng cá thể lai mà không cần phải tạo các bảng marker tùy chỉnh cho từng quần thể trong một chương trình chọn giống.

Sáu phương pháp khác nhau đã được thử nghiệm để xác định phương pháp nào có khả năng có nhiều marker đa hình nhất trong các chương trình chọn giống và có các marker phân bố đồng đều trên bộ gen. Từ sáu phương pháp khác nhau được thử nghiệm, ba phương pháp cho mức độ marker đa hình tương tự trên tất cả các tế bào mầm mục tiêu. Phương pháp BIN_PIC tạo ra trung bình 443 chỉ thị đa hình cho quần thể lưỡng cực ngẫu nhiên có bố mẹ đậu tương ưu tú, Song_PIC_MIN_2 tạo ra 445 chỉ thị đa hình và Song_PIC_MIN_0.1 tạo ra 430 chỉ thị đa hình. Ba phương pháp còn lại có ít marker đa hình ước tính hơn, với Haploview Tagger có 323 marker đa hình, Haploview_BIN có 312 và Random có 314.

Mặc dù số lượng trung bình của các marker đa hình là quan trọng, nhưng việc phân phối các marker này cũng rất quan trọng để đạt được độ bao phủ hiệu quả trên toàn bộ bộ gen. Phương pháp Song_PIC_MIN_2 có ít khoảng trống lớn hơn so với phương pháp BIN_PIC khi xem xét sự phân bố của các marker trên không gian vật lý của bộ gen (Hình 1a, b). Tất cả các khoảng cách lớn về khoảng cách vật lý từ phương pháp BIN_PIC đều rơi vào vùng quanh tâm động. Vùng này có tính dị nhiễm sắc cao và ức chế nghiêm trọng sự tái tổ hợp di truyền. BIN_PIC đã sử dụng các thùng dựa trên khoảng cách di truyền để giúp phân phối các marker đồng đều trên không gian di truyền của bộ gen. Điều này dẫn đến ít marker được nhóm lại ở vùng màng ngoài tim khi so sánh với phương pháp Song_PIC_MIN_2 khi xem xét sự phân bố của các marker trên không gian di truyền của bộ gen (Hình 1c, d).



Hình 1. Phân phối đầu dò đảo ngược phân tử (MIP) 1K đơn nucleotide đa hình được chọn (SNP) so với BARCSoy6k SNP. MIP 1K SNP được biểu thị bằng dấu tick màu đỏ, BARCSoy6k SNP được biểu thị bằng dấu tick màu đen. Các hộp màu đỏ xác định các khoảng trống lớn không có SNP nào được chọn cho 1K MIP. Các hộp màu cam biểu thị các vùng màng ngoài tim của đậu tương. (a) Phân phối MIP 1K SNP được chọn bằng phương pháp BIN_PIC trên bộ gen vật lý của đậu tương. (b) Phân phối MIP 1k SNP được chọn bằng phương pháp Song_PIC_MIN_0.2 trên bộ gen vật lý của đậu tương. (c) Phân phối MIP 1K SNP được chọn bằng phương pháp BIN_PIC trên bộ gen khoảng cách di truyền của đậu tương. (d) Phân phối MIP 1k SNP được chọn bằng phương pháp Song_PIC_MIN_0.2 trên bộ gen khoảng cách di truyền của đậu tương.

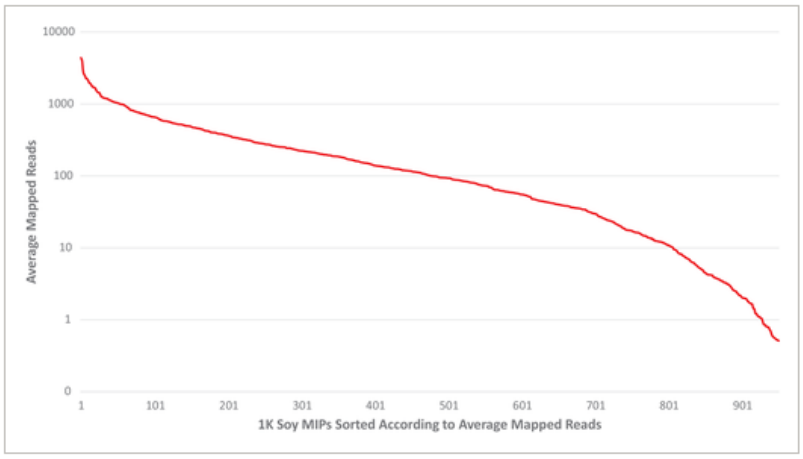
Số lượng trung bình của các marker đa hình vẫn ở mức cao đối với phương pháp BIN_PIC ngay cả khi tế bào mầm được sử dụng để chọn các marker được phân chia theo chương trình chọn giống tương ứng của chúng (Bảng 1). Số lượng trung bình của các marker đa hình có phạm vi ước tính là 338–443 marker đa hình. Các chương trình chọn giống với số lượng chỉ thị đa hình ước tính ít nhất trong các quần thể là Nebraska, Ohio, miền nam Illinois và Tennessee. Số lượng các marker đa hình từ phương pháp BIN_PIC rất giống với hai phương pháp Song và vượt trội hơn đáng kể so với các phương pháp Haploview và Random (Bảng 1).

3.2 Đọc căn chỉnh của MIP

Bộ 1K Soy MIP SNP đã được sử dụng để sàng lọc 96 lần bổ sung vùng đất đã được sàng lọc trước đó bằng chip SoySNP50K. Tổng số 26.631.895 lượt đọc 150bp thô đã thu được cho 93 trên 96 mẫu. Ba mẫu không tạo ra đủ dữ liệu giải trình tự có thể là do chất lượng DNA thấp. Sau khi lọc kiểm soát chất lượng, 25.343.687 lượt đọc được giữ lại (95%), trong đó, 22.811.487 lượt đọc được căn chỉnh theo các vùng SNP được nhắm mục tiêu. Điều này tương đương với mức làm giàu 85% cho các chuỗi được nhắm mục tiêu từ tổng số lần đọc thô.

Sự phong phú của trình tự được nhắm mục tiêu này đã dẫn đến phạm vi đọc tốt trên các SNP được nhắm mục tiêu. Phạm vi đọc trung bình trên mỗi SNP trong mỗi lần

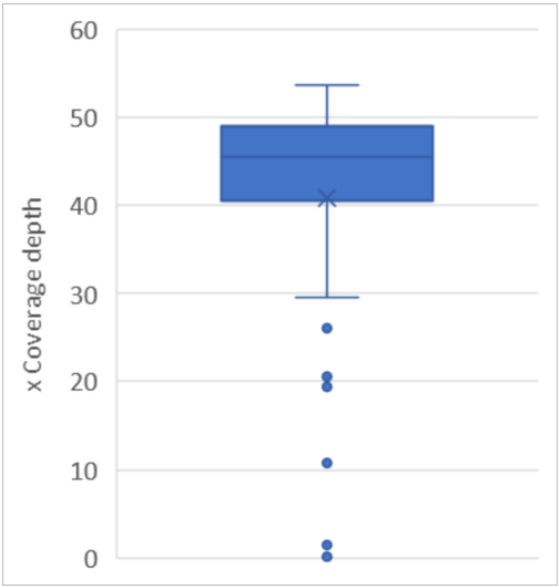
gia nhập là 245 ×. Tổng cộng có 808 đầu dò có trung bình từ 10 lần đọc trở lên đối với tất cả các lần bổ sung đậu tương, trong khi trung bình 53 đầu dò có hơn 1.000 lượt đọc trở lên (Hình 2). Sau khi loại bỏ 192 đầu dò với trung bình <10 lần đọc, SNP được gọi trên các đầu dò còn lại. Sự phù hợp của các cuộc gọi SNP từ dữ liệu 1K đến chip SoySNP50K là 98% đối với mẫu 93 PI tham gia. Các mẫu PI có tính lai tạo cao với rất ít dị hợp tử. Độ chính xác 98% thu được từ các lần nhập này phản ánh hầu hết các kiểu gen đồng hợp tử được gọi.



Hình 2. Hiệu suất của các marker thăm dò đảo ngược phân tử (MIP) được sử dụng để tạo kiểu gen cho 96 mẫu đậu tương PI tham gia. Độ lệch chuẩn của các đầu dò MIP trên 96 dòng được biểu thị bằng các thanh lỗi.

3.3 Độ chính xác của kiểu gen

Để tiếp tục truy cập độ chính xác của kiểu gen và lượng dữ liệu bị thiếu của kiểu gen MIP đối với kiểu gen đồng hợp tử và dị hợp tử, một quần thể RIL, ExW82, đã được tạo kiểu gen bằng bộ 1K MIP SNP. Các RIL này trước đây đã được tạo kiểu gen với chip SoySNP50K (Song và ctv, 2016). Chip SoySNP50K bao gồm tất cả 1K MIP SNP. 96 dòng được giải trình tự theo độ sâu đọc MIP trung bình là 47 × (Hình 3). Tổng cộng có 74 đầu dò được căn chỉnh sai mục tiêu và 199 đầu dò bị thiếu >30% dữ liệu. 273 đầu dò này đã bị loại khỏi phân tích phù hợp. Sáu trong số các kiểu gen có dữ liệu bị thiếu > 30% trong tổng số 292 đầu dò trong số 725 mẫu đậu tương còn lại là đa hình (40% đa hình) trong quần thể RIL.



Hình 3. Độ sâu bao phủ trình tự trung bình đối với kiểu gen của đầu dò đảo ngược phân tử (MIP) 1K của các dòng lai tái tổ hợp Essex × Williams82.

Sự phù hợp của kiểu gen MIP được xác định bằng cách so sánh kiểu gen SNP của RIL được thực hiện bởi Song và ctv (2016) với kiểu gen 1K MIP. Điều này cho phép xác

định độ chính xác về kiểu gen của các kiểu gen SNP đồng hợp tử và dị hợp tử được gọi bằng MIP. Ngoài ra, nhiều tham số gọi SNP của GATK đã được khám phá để xác định ảnh hưởng của chúng đối với độ chính xác của kiểu gen (Bảng 2). Nhìn chung, độ chính xác là > 93% đối với kiểu gen với MIP. Độ nhạy căn chỉnh, chất lượng bản đồ, độ sâu đọc và lệnh độ tin cậy GATK đã hạn chế ảnh hưởng đến độ chính xác của kiểu gen tổng thể và độ chính xác của kiểu gen SNP đồng hợp tử. Các tham số gọi SNP khác nhau đã ảnh hưởng đến các kiểu gen dị hợp tử và dữ liệu bị thiếu. Giá trị trung bình cho độ chính xác của kiểu gen dị hợp tử tăng lên > 90% khi sử dụng tham số chất lượng bản đồ là 30. Độ sâu đọc là 5 so với 10 đã tăng độ chính xác trung bình từ 90% lên 94%. Tuy nhiên, độ sâu đọc cao hơn làm tăng độ chính xác trung bình cho các dị hợp tử chỉ 2% và tăng dữ liệu bị thiếu lên 13,2%.

Bảng 2. Các tham số đơn nucleotide đa hình (SNP) và ảnh hưởng đến độ chính xác bằng cách sử dụng sự phù hợp giữa kiểu gen của đầu dò đảo ngược phân tử (MIP) và cùng kiểu gen SNP được tạo bằng chip SoySNP50k Infinium trên 90 dòng lai tái tổ hợp Essex × Williams 82. Các MIP bị xóa với các mục tiêu sai (sai tham chiếu/thay thế hoặc phù hợp <80), các MIP bị xóa >30 dữ liệu bị thiếu và các kiểu gen bị xóa >30 dữ liệu bị thiếu

Tham số	Cài đặt	Tổng thể			Homozygous			Heterozygous			Dữ liệu khuyết		
		TB	Trung vị	SD	TB	Trung vị	SD	TB	Trung vị	SD	TB	Trung vị	SD
Alignment sensitivity	Very sensitive, local	93	97	9	98	98	2	72	86	30	9	4	16
	Very sensitive, end to end	94	97	9	98	99	2	69	82	29	7	1	17
	Very sensitive, local	93	97	9	98	98	2	73	89	30	10	4	17
	Very sensitive, end to end	94	98	9	98	99	2	71	83	29	7	1	17
Mapping quality	10	94	98	9	98	99	2	70	83	29	7	1	17
	20	94	98	9	98	99	2	71	83	29	7	1	17
	30[(a)]	94	97	9	98	99	2	74	90	31	6	1	18
Read depth	5[(a)]	94	98	9	98	98	3	74	90	31	9	1	21
	10	92	97	11	96	98	7	76	94	32	13	4	23
GATK confidence call	5	94	98	9	98	98	2	74	90	31	9	2	21
	10	94	98	9	98	99	3	74	90	31	9	2	21
	20[(a)]	94	98	9	98	99	2	75	91	31	9	2	21

(a) tham số cuối cùng được chọn cho đường ống gọi MIPs SNP.

4. THẢO LUẬN

Chúng tôi đã chỉ ra rằng MIP có thể được sử dụng để tạo thành công SNPs nhắm mục tiêu kiểu gen trong đậu tương. Các đầu dò nhắm mục tiêu SNP có thể làm giàu trình tự thu được cho các SNP đã chọn thêm 85%. Mức độ làm giàu cao này cao hơn mức độ làm giàu 77,3% thu được ở các sinh vật phi mô hình trơn (*Lissotriton Vulgaris Vulgaris*) và sa giông Carpathian (*Lissotriton montandoni*) (Niedzicka và ctv, 2016). Sa giông có 80% trong số 248 MIP với độ bao phủ trung bình trong một đơn hàng của nhau (Niedzicka và ctv, 2016). Đây là phạm vi bao phủ MIP đồng đều hơn so với thu được với bộ thăm dò 1K đậu tương. Sự khác biệt chính về mức độ bao phủ trung bình có thể liên quan đến việc tái cân bằng các MIP trong kiểu gen của sa giông. Cân bằng lại các đầu dò riêng lẻ liên quan đến việc tăng nồng độ của các MIP tạo ra số lần đọc thấp hơn mức

trung bình và giảm nồng độ của các MIP tạo ra số lần đọc cao hơn mức trung bình. Việc tái cân bằng này sẽ đạt được sự phân phối đồng đều hơn về phạm vi đọc trung bình cho mỗi MIP (Niedzicka và ctv, 2016). Khi sử dụng bộ MIP không cân bằng ở người, tính đồng nhất của phạm vi đọc thăm dò tương tự như kết quả thu được với bộ MIP đậu tương (Mamanova và ctv, 2010). Việc cân bằng lại nồng độ ban đầu của các MIP riêng lẻ có thể dễ dàng được thực hiện khi các mẫu dò được tổng hợp riêng lẻ. Thông qua tinh chỉnh thêm, phạm vi đọc trung bình cho bộ 1K MIP SNP có thể được cải thiện cho đậu tương. Điều này sẽ cho phép ghép nhiều mẫu hơn ở độ sâu trình tự trung bình thấp hơn trên mỗi lần chạy trình tự.

Độ chính xác 98% của các đồng hợp tử về kiểu gen đối với bộ 1K MIP SNP đậu tương tương đương với độ chính xác 99% được báo cáo ở các loài khác (Niedzicka và ctv, 2016; Teer và ctv, 2010). Tuy nhiên, độ chính xác của dị hợp tử thấp hơn mong đợi trong nghiên cứu này. Một lý do tiềm năng là kiểu gen Illumina ban đầu của các RIL này được thực hiện với DNA được chiết xuất từ một cây F5 duy nhất tạo ra RIL. Kiểu gen của MIP đã sử dụng mô lá từ tổ hợp của năm cây được trồng từ RIL. Mặc dù việc lấy mẫu chỉ năm cây từ trong dòng có thể thực tế hơn đối với các nhà chọn giống khi lấy mẫu một dòng lai, nhưng nó có thể không đủ để nắm bắt được tất cả tính không đồng nhất có trong dòng. Ngoài ra, bằng cách chỉ lấy mẫu năm cây, có thể có khả năng lấy mẫu không đồng đều của hai alen có trong RIL không đồng nhất đối với SNP. Điều này có thể ảnh hưởng đến độ chính xác của việc gọi SNP dị hợp tử. Do đó, độ sâu trình tự sâu hơn hoặc lấy mẫu nhiều cây hơn trong một dòng có khả năng tăng độ chính xác các dị hợp tử trong các dòng không đồng nhất khi cần độ chính xác cao hơn. Độ chính xác kiểu gen tổng thể của các dòng tự nhiên là > 93%. Điều này nằm trong phạm vi độ chính xác kiểu gen 92–98% của reGBS khi giải trình tự các dòng ở mức trung bình có độ sâu 9 × (Torkamaneh và ctv, 2016).

Để đạt được số lượng lớn nhất các marker thông tin trong quần thể chọn giống đậu tương, nhiều phương pháp đã được khám phá khi chọn bộ 1K MIP SNP. Các phương pháp Bin_PIC và Song có cùng số lượng SNP đa hình trong tất cả các quần thể lưỡng cực có thể và vượt trội hơn cả hai phương pháp Haploview và chọn ngẫu nhiên SNP. Trung bình, các phương pháp Bin_PIC và Song có tỷ lệ đa hình là 43–44%, trong khi ba phương pháp chọn SNP còn lại có tỷ lệ đa hình là 31–32%. Tỷ lệ đa hình với các phương thức Bin_PIC và Song cao hơn so với 1536 USLP 1.0 ban đầu. 1536 USLP 1.0 có 30% đa hình của các marker giữa hai bố mẹ ưu tú ngẫu nhiên bất kỳ (Hyten và ctv, 2010). Tỷ lệ đa hình cao hơn này với bộ MIP có thể đến từ bộ SNP 1K MIP được chọn từ 50.000 marker SNP, trong khi 1536 USLP 1.0 gốc chỉ được chọn từ 3.000 SNP. Điều này chỉ ra rằng mặc dù bộ 1K MIP SNP có ít hơn 35% marker SNP, nhưng nó vẫn tạo ra số lượng marker đa hình tương tự như 1536 USLP 1.0.

Mặc dù bộ 1K MIPs SNP được thiết kế để chọn giống đậu tương ở Bắc Mỹ, nhưng quá trình tổng hợp riêng lẻ của từng mẫu dò mang lại sự linh hoạt trong việc thêm hoặc loại bỏ các marker có thể không có tỷ lệ đa hình cao trong các nhóm tế bào mầm khác. Điều này có thể cho phép các nhà nghiên cứu làm việc với nhiều nguồn tế bào mầm sử dụng một bộ chỉ thị cốt lõi hoạt động trên các nhóm tế bào mầm và bổ sung cho mỗi lõi các chỉ thị cụ thể của tế bào mầm để làm cho bảng cụ thể hơn đối với loại tế bào mầm cụ thể đang được thử nghiệm. Ngoài ra, việc tổng hợp các đầu dò riêng lẻ giúp dễ dàng loại bỏ các đầu dò hoạt động kém và thêm các đầu dò SNP được nhắm mục tiêu trong tình trạng mất cân bằng liên kết cao với các đầu dò hoạt động kém đó. Các marker SNP dành

riêng cho đặc điểm nhắm mục tiêu thăm dò cũng có thể được thêm vào bộ 1K MIP cho lựa chọn được hỗ trợ bởi marker.

Sự khác biệt chính giữa phương pháp BIN_PIC và Song là cách SNP được phân phối trên các nhiễm sắc thể. Bằng cách buộc một số SNP cụ thể trong mỗi ngăn, phương pháp BIN_PIC đảm bảo rằng các SNP được phân phối đồng đều trên không gian di truyền của bộ gen. Các phương pháp Song có sự phân cụm khác biệt trong toàn bộ bộ gen. Điều này chủ yếu xảy ra xung quanh các vùng màng ngoài tim. Các vùng quanh tâm động này có sự mất cân bằng liên kết rất cao trong quần thể đậu tương và không yêu cầu số lượng lớn các marker trên toàn vùng vật lý của nó để bao hàm biến thể haplotype (Song và ctv, 2015). Phương pháp Haploview không hoạt động tốt trong việc tạo ra các marker đa hình thông tin như phương pháp Bin_PIC và Song. Điều này có thể là do Haploview tập trung vào việc xác định các SNP gắn thẻ nhiều alen nhất. Điều này không nhất thiết giống như việc nắm bắt các marker đa hình nhất và khiến phương pháp hoạt động kém hiệu quả trong việc tạo ra một số lượng lớn các marker thông tin.

Với số lượng lớn các marker thông tin, MIP có khả năng chạy trên một số lượng lớn các dòng lai cần thiết để triển khai hiệu quả quá trình chọn lọc bộ gen hoặc các ứng dụng khác yêu cầu mật độ marker thấp để chạy trên tế bào mầm. Việc triển khai mã vạch kép cho MIP cung cấp khả năng ghép kênh >1.000 mẫu trong một lần chạy. Hiện tại, chúng tôi đã tạo kiểu gen thành công cho 1.152 mẫu trong một lần chạy giải trình tự trên Illumina NextSeq 500 với bộ SNP 1K MIPs đậu tương. Điều này vừa làm tăng thông lượng vừa giảm chi phí thuốc thử xuống còn $4,10 cho mỗi mẫu. Chi phí thuốc thử cho việc xây dựng thư viện là $3,08, đây là chi phí cố định cho mỗi mẫu, trong khi chi phí giải trình tự sẽ tăng nếu có ít hơn 1.152 mẫu được giải trình tự trong một lần chạy. Ngoài ra, giao thức cho MIP rất đơn giản và có thể sửa đổi để tự động hóa. Sử dụng thiết bị xử lý chất lỏng tự động, chẳng hạn như Mantis, đã cho phép một người xử lý nhiều đĩa 384 giếng trong một ngày.

Khả năng của MIP đối với thông lượng mẫu cao, độ chính xác kiểu gen cao và tỷ lệ marker đa hình cao trong bộ 1K MIPs SNP làm cho nó trở thành công nghệ tạo kiểu gen khả thi để thực hiện mật độ marker thấp để chọn lọc bộ gen trong quần thể dòng lai đậu tương.

5. TÀI LIỆU THAM KHẢO

- Barrett, J. C. (2009). Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor Protocols*, 2009, pdb.ip71. https://doi.org/10.1101/pdb.ip71
    - View
    - PubMedGoogle Scholar
- Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21, 263– 265.
    - View
    - CASPubMedWeb of Science®Google Scholar
- Basnet, P., Meinhardt, C. G., Usovsky, M., Gillman, J. D., Joshi, T., Song, Q., Diers, B., Mitchum, M. G., & Scaboo, A. M. (2022). Epistatic interaction between *Rhg1-a* and *Rhg2* in PI 90763 confers resistance to virulent soybean cyst nematode populations. *Theoretical and Applied Genetics*, 135, 2025– 2039. https://doi.org/10.1007/s00122-022-04091-2
    - View
    - CASPubMedWeb of Science®Google Scholar

- Beche, E., Gillman, J. D., Song, Q., Nelson, R., Beissinger, T., Decker, J., Shannon, G., & Scaboo, A. M. (2021). Genomic prediction using training population design in interspecific soybean populations. *Molecular Breeding*, 41, 15. https://doi.org/10.1007/s11032-021-01203-6
    - View
    - CASWeb of Science®Google Scholar
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114– 2120. https://doi.org/10.1093/bioinformatics/btu170
    - View
    - CASPubMedWeb of Science®Google Scholar
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32, 314– 331.
    - CASPubMedWeb of Science®Google Scholar
- Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A., & Shendure, J. (2014). MIPgen: Optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*, 30, 2670– 2672. https://doi.org/10.1093/bioinformatics/btu353
    - View
    - CASPubMedWeb of Science®Google Scholar
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6, e19379. https://doi.org/10.1371/journal.pone.0019379
    - View
    - CASPubMedWeb of Science®Google Scholar
- Gonen, S., Wimmer, V., Gaynor, R. C., Byrne, E., Gorjanc, G., & Hickey, J. M. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in bi-parental plant populations. *Theoretical and Applied Genetics*, 131, 2345– 2357. https://doi.org/10.1007/s00122-018-3156-9
    - View
    - CASPubMedWeb of Science®Google Scholar
- Happ, M., Wang, H., Graef, G., & Hyten, D. (2019). Generating high density, low cost genotype data in soybean [*Glycine max* (L.) Merr.]. *G3 Genes, Genomes, Genetics*, 91, 2153– 2160. https://doi.org/10.1534/g3.119.400093
    - View
    - Web of Science®Google Scholar
- Hyten, D. L., Choi, I. Y., Song, Q. J., Specht, J. E., Carter, T. E., Shoemaker, R. C., Hwang, E.-Y., Matukumalli, L. K., & Cregan, P. B. (2010). A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. *Crop Science*, 50, 960– 968. https://doi.org/10.2135/cropsci2009.06.0360
    - View
    - CASWeb of Science®Google Scholar
- Hyten, D. L., Pantalone, V. R., Sams, C. E., Saxton, A. M., Landau-Ellis, D., Stefaniak, T. R., & Schmidt, M. E. (2004). Seed quality QTL in a prominent soybean population. *Theoretical and Applied Genetics*, 109, 552– 561. https://doi.org/10.1007/s00122-004-1661-5
    - View
    - CASPubMedWeb of Science®Google Scholar

- Hyten, D. L., Song, Q., Choi, I. Y., Yoon, M. S., Specht, J. E., Matukumalli, L. K., Nelson, R. L., Shoemaker, R. C., Young, N. D., & Cregan, P. B. (2008). High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics*, 116, 945– 952. https://doi.org/10.1007/s00122-008-0726-2
    - View
    - CASPubMedWeb of Science®Google Scholar
- Kim, K.-H., Kim, M.-Y., Van, K.-J., Moon, J.-K., Kim, D.-H., & Lee, S.-H. (2008). Marker-assisted foreground and background selection of near isogenic lines for bacterial leaf pustule resistant gene in soybean. *Journal of Crop Science and Biotechnology*, 11, 263– 268.
    - Google Scholar
- Kim, K. S., Diers, B. W., Hyten, D. L., Mian, M. A. R., Shannon, J. G., & Nelson, R. L. (2012). Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. *Theoretical and Applied Genetics*, 125, 1353– 1369. https://doi.org/10.1007/s00122-012-1944-1
    - View
    - PubMedWeb of Science®Google Scholar
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9, 357– 359. https://doi.org/10.1038/nmeth.1923
    - View
    - CASPubMedWeb of Science®Google Scholar
- Lee, S. H., Walker, D. R., Cregan, P. B., & Boerma, H. R. (2004). Comparison of four flow cytometric SNP detection assays and their use in plant improvement. *Theoretical and Applied Genetics*, 110, 167– 174.
    - View
    - CASPubMedWeb of Science®Google Scholar
- Li, Y., Ye, H., Song, L., Vuong, T. D., Song, Q., Zhao, L., Shannon, J. G., Li, Y., & Nguyen, H. T. (2021). Identification and characterization of novel QTL conferring internal detoxification of aluminium in soybean. *Journal of Experimental Botany*, 72, 4993– 5009. https://doi.org/10.1093/jxb/erab168
    - View
    - CASPubMedWeb of Science®Google Scholar
- Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7, 111– 118. https://doi.org/10.1038/nmeth.1419
    - View
    - CASPubMedWeb of Science®Google Scholar
- Nazzicari, N., Biscarini, F., Cozzi, P., Brummer, E. C., & Annicchiarico, P. (2016). Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). *Molecular Breeding*, 36, 69. https://doi.org/10.1007/s11032-016-0490-y
    - View
    - Web of Science®Google Scholar
- Niedzicka, M., Fijarczyk, A., Dudek, K., Stuglik, M., & Babik, W. (2016). Molecular inversion probes for targeted resequencing in non-model organisms. *Science Reports*, 6, 24051. https://doi.org/10.1038/srep24051
    - View
    - CASPubMedWeb of Science®Google Scholar
- Phansak, P., Soonsuwon, W., Hyten, D. L., Song, Q., Cregan, P. B., Graef, G. L., & Specht, J. E. (2016). Multi-population selective genotyping to identify soybean [*Glycine*

max (L.) Merr.] seed protein and oil QTLs. *G3 Genes, Genomes, Genetics*, 6, 1635–1648. https://doi.org/10.1534/g3.116.027656

- View
- CASWeb of Science®Google Scholar

- Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., LeProust, E. M., Peck, B. J., Emig, C. J., Dahl, F., Gao, Y., Church, G. M., & Shendure, J. (2007). Multiplex amplification of large sets of human exons. *Nature methods*, 4, 931– 936. https://doi.org/10.1038/nmeth1110
  - View
  - CASPubMedWeb of Science®Google Scholar
- Ravelombola, W., Qin, J., Shi, A., Song, Q., Yuan, J., Wang, F., Chen, P., Yan, L., Feng, Y., Zhao, T., Meng, Y., Guan, K., Yang, C., & Zhang, M. (2021). Genome-wide association study and genomic selection for yield and related traits in soybean. *PLoS One*, 16, e0255761. https://doi.org/10.1371/journal.pone.0255761

  - View
  - CASPubMedWeb of Science®Google Scholar
- Sebastian, S. A., Feng, L., & Kuhlman, L. C. (2012). Accelerated Yield Technology™: A platform for marker assisted selection of simple and complex traits. In R. F. Wilson (Ed.), Designing soybeans for 21st century markets (pp. 297– 305). AOCS Press. https://doi.org/10.1016/B978-0-9830791-0-1.50020-0
  - View
  - Google Scholar
- Song, Q. J., Hyten, D. L., Jia, G. F., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One*, 8, e54985. https://doi.org/10.1371/journal.pone.0054985

  - View
  - CASPubMedWeb of Science®Google Scholar
- Song, Q. J., Hyten, D. L., Jia, G. F., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *G3 Genes, Genomes, Genetics*, 5, 1999– 2006. https://doi.org/10.1534/g3.115.019000

  - View
  - Web of Science®Google Scholar
- Song, Q. J., Marek, L. F., Shoemaker, R. C., Lark, K. G., Concibido, V. C., & Delannay, X. (2004). A new integrated genetic linkage map of the soybean. *Theoretical and Applied Genetics*, 109, 122– 128. https://doi.org/10.1007/s00122-004-1602-3
  - View
  - CASPubMedWeb of Science®Google Scholar
- Song, Q., Jenkins, J., Jia, G., Hyten, D. L., Pantalone, V., Jackson, S. A., Schmutz, J., & Cregan, P. B. (2016). Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *BMC Genomics*, 17, 33. https://doi.org/10.1186/s12864-015-2344-0
  - View
  - PubMedWeb of Science®Google Scholar
- Song, Q., Yan, L., Quigley, C., Fickus, E., Wei, H., Chen, L., Dong, F., Araya, S., Liu, J., Hyten, D., Pantalone, V., & Nelson, R. L. (2020). Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. *The Plant Journal*, 104, 800– 811. https://doi.org/10.1111/tpj.14960

    - View
    - CASPubMedWeb of Science®Google Scholar
- Teer, J. K., Bonnycastle, L. L., Chines, P. S., Hansen, N. F., Aoyama, N., Swift, A. J., Abaan, H. O., Albert, T. J., Margulies, E. H., Green, E. D., Collins, F. S., Mullikin, J. C., & Biesecker, L. G., NISC Comparative Sequencing Program. (2010). Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Research*, 20, 1420– 1431. https://doi.org/10.1101/gr.106716.110
    - View
    - CASPubMedWeb of Science®Google Scholar
- Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS One*, 11, e0161333. https://doi.org/10.1371/journal.pone.0161333
    - View
    - PubMedWeb of Science®Google Scholar
- Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature methods*, 6, 315– 316. https://doi.org/10.1038/nmeth.f.248
    - View
    - CASPubMedWeb of Science®Google Scholar
- Waalkes, A., Smith, N., Penewit, K., Hempelmann, J., Konnick, E. Q., Hause, R. J., Pritchard, C. C., & Salipante, S. J. (2018). Accurate pan-cancer molecular diagnosis of microsatellite instability by single-molecule molecular inversion probe capture and high-throughput sequencing. *Clinical Chemistry*, 64, 950– 958. https://doi.org/10.1373/clinchem.2017.285981
    - View
    - CASPubMedWeb of Science®Google Scholar
- Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., Bajjalieh, N., Li, Z., & Boerma, H. R. (2015). QTL for seed protein and amino acids in the Benning x Danbaekkong soybean population. *Theoretical and Applied Genetics*, 128, 839– 850. https://doi.org/10.1007/s00122-015-2474-4
    - View
    - CASPubMedWeb of Science®Google Scholar
- Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, 18, 586. https://doi.org/10.1186/s12859-017-2000-6
    - View
    - PubMedWeb of Science®Google Scholar
- Zhu, Y. L., Song, Q. J., Hyten, D. L., Van Tassell, C. P., Matukumalli, L. K., Grimm, D. R., Hyatt, S. M., Fickus, E. W., Young, N. D., & Cregan, P. B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics*, 163, 1123– 1134.
    - View
    - CASPubMedWeb of Science®Google Scholar