ĐẶC ĐIỂM TÁC ĐỘNG CỦA DÒNG ĐẬU TƯƠNG NGOẠI LAI ĐỐI VỚI SỰ PHÁT TRIỂN GIỐNG ƯU TÚ

Benjamin B. Stewart-Brown[12], Justin N. Vaughn[3], Thomas E. Carter Jr[1], Zenglu LiID[1]

Võ Như Cầm biên dịch.

1. Khoa Cây trồng và Khoa học Đất, Viện chọn giống cây trồng, Di truyền và Gen, Đại học Georgia, Athens, GA, Mỹ.

2 Đơn vị nghiên cứu về Gen và Tin sinh học, USDA-ARS, Athens, GA, Mỹ.

3 Đơn vị nghiên cứu về đậu tương & cố định nitơ, USDA-ARS, Raleigh, NC, Mỹ.

TÓM TẮT

Sự đa dạng di truyền của các giống đậu tương Bắc Mỹ phần lớn bị ảnh hưởng bởi một số ít tổ tiên. Các dòng lai năng suất cao sở hữu phả hệ ngoại lai đã được phát triển, nhưng việc xác định các alen ngoại lai có lợi rất khó khăn do sự tương tác phức tạp của các alen năng suất với nguồn gốc và môi trường di truyền cũng như bản chất định lượng cao của năng suất. PI 416937 đã được sử dụng để phát triển nhiều dòng năng suất cao đã được đưa vào các Thí nghiệm đồng nhất các bang miền Nam của USDA trong hơn 20 năm qua. Mục tiêu chính của nghiên cứu này là xác định các vùng gen được chọn lọc từ PI 416937 và giới thiệu phương pháp luận để xác định và có khả năng sử dụng sự đa dạng có lợi từ các dòng phổ biến trong tổ tiên của các giống cây trồng ưu tú. Sử dụng SoySNP50K Infinium BeadChips, 52 dòng năng suất cao có nguồn gốc PI 416937 cũng như bố mẹ của chúng đã được định kiểu gen để xác định các alen PI 416937 trong quá trình chọn giống. Chín vùng gen trên ba nhiễm sắc thể và 17 vùng gen trên bảy nhiễm sắc thể đã được xác định nơi các alen PI 416937 được chọn lọc tích cực hoặc tiêu cực. Các mối liên hệ có ý nghĩa tối thiểu giữa các alen PI 416937 và năng suất đã được quan sát thấy trong các thí nghiệm năng suất lặp lại của năm quần thể RIL, làm nổi bật khó khăn của việc phát hiện các mối liên hệ năng suất một cách nhất quán.

GIỚI THIỆU

Nói chung, sự đa dạng di truyền có trong các giống cây trồng cải tiến chỉ thể hiện một phần nhỏ trong tổng số sự đa dạng có trong loài mà từ đó cây trồng được tạo ra [1]. Sự suy giảm tính đa dạng này được minh chứng bằng đậu tương, trong đó gần 75% các giống ở Bắc Mỹ được phóng thích từ năm 1947–1988 có nguồn gốc từ 17 tổ tiên và gần 50% chỉ có nguồn gốc từ sáu dòng tổ tiên [2]. Để tăng hơn nữa tỷ lệ di truyền thu được trong ứng dụng chọn giống vượt quá mức quan sát hiện nay, bắt buộc phải khai thác các nguồn gen toàn cầu để tìm các alen có lợi, chẳng hạn như các alen mới để kháng sâu bệnh [3]. Sau đó, các alen này có thể được xâm nhập vào các giống thông qua chọn giống thông thường hoặc phương pháp tiếp cận chọn lọc có sự hỗ trợ của marker [4–6]. Về mặt lịch sử, sự xâm nhập của các tính trạng từ các tế bào mầm của tổ tiên hoang dã cũng như các chủng tộc nói chung chỉ giới hạn ở các tính trạng do các gen chính kiểm soát. Những đặc điểm như vậy dễ xác định hơn với sự tự tin, ít phụ thuộc vào nền tảng di truyền và dễ theo dõi hơn trong quá trình sinh sản. Mặc dù các alen hoang dã cho các đặc điểm phức tạp như năng suất đã được xác định thành công bằng cách sử dụng các dòng gần đồng gen, nhưng các phương pháp này sử dụng nhiều tài nguyên và thường bỏ sót các alen có liên quan [7].

Ude và cs [8] chỉ ra rằng các giống đậu Nhật Bản đã cung cấp một nguồn nguyên liệu khác biệt về mặt di truyền để cải tiến các giống đậu tương Bắc Mỹ. PI 416937 là giống cây Nhật Bản hiện diện trong phả hệ của nhiều dòng/giống ưu tú ở miền Đông Nam Mỹ, đáng chú ý nhất là 'Woodruff' [9]. Woodruff có 25% đóng góp di truyền từ PI 416937 theo phả hệ và năng suất đạt 111, 122 và 111% của giống ưu tú, 'Benning' [10] trong các Thí nghiệm đồng nhất Nam Mỹ của Bộ Nông nghiệp Mỹ (USDA) từ năm 2003–2005 [11– 13]. Người ta đã báo cáo rằng PI 416937 sở hữu một số đặc điểm phân biệt bao gồm héo tán chậm [14–17], rễ sợi phát triển nhanh [18–20], chống chịu nhôm [21–22], phản ứng thay đổi thâm hụt áp suất hơi và các đặc điểm sinh lý khác được kiểm soát liên quan hạn hán [23–26]. 'N7002' (PI 647085) [27], 'N8001' (PI 647086) [28], và 'USDA-N8002' (PI 676972) [29] là các giống bổ sung có > 12,5% đóng góp di truyền từ PI 416937 theo phả hệ và đã vượt trội hơn các giống kiểm tra về năng suất trong các Thí nghiệm đồng nhất Nam Mỹ của USDA [11–13, 30–37]. Do đó, không giống như trường hợp phổ biến trong đó nguồn gen ngoại lai được sử dụng làm nguồn cho một gen cụ thể, những đóng góp của PI 416937 có vẻ phức tạp và các dòng xuất phát của nó là ví dụ về việc kết hợp nguồn gen ngoại lai trong việc phát triển các giống với năng suất cao hơn và cung cấp sự đa dạng di truyền có lợi trong thời gian dài.

Trong nghiên cứu này, các dòng có thông tin phả hệ đã biết liên quan đến PI 416937 được khai thác bằng cách sử nucleotide đơn đa hình (SNP) trên toàn bộ bộ gen để theo dõi các vùng gen ngoại lai mới đã được lựa chọn và lặp lại trong khoảng 20 năm qua. Ý tưởng khai thác phả hệ chọn giống để phát hiện các locus chọn lọc đã được sử dụng trước đây trong nỗ lực phát hiện các locus quan trọng về mặt nông học ở đậu tương [38–40]. Phân tích tương tự cũng đã được thực hiện trên cây lạc tiên (*Arachis hypogaea*) [41]. Cách tiếp cận này tương tự như các xét nghiệm mất cân bằng lây truyền (TDTs) được tiên phong trong di truyền học động vật [42]. Các giống cây trồng được phóng thích được cho là sản phẩm của nhiều giai đoạn chọn lọc và do đó, các alen mang lại tính trạng ưu việt được cho là sẽ lệch khỏi sự truyền tải ngẫu nhiên (50%) [42]. Trong khi các phiên bản gốc của TDT chủ yếu được thực hiện trong di truyền động vật, Jannink và cs [43] đề xuất rằng TDT có thể được thích nghi với cây trồng tự thụ phấn bằng cách kiểm tra các dòng lai và cây trồng trong nhiều thập kỷ để xác định các alen truyền tải ưu tiên được cho là có liên kết bất cân bằng (LD) với các locus tính trạng số lượng (QTL) thuận lợi. Mặc dù cách tiếp cận này về mặt lý thuyết là rất mạnh mẽ, nhưng các nghiên cứu trước đây đã gặp phải tình trạng mật độ marker thấp [38, 44] hoặc các phả hệ bị che khuất khiến cho việc suy luận thống kê chặt chẽ trở nên có vấn đề [40]. Mật độ marker cao hơn cho phép suy luận tin cậy về các dạng đơn bội được chia sẻ trong các tổ hợp con của bố mẹ, và do đó, nâng cao khả năng xác định và đếm chính xác số lượng phép lai được kiểm tra locus về ảnh hưởng của chọn lọc. Trái ngược với việc lập bản đồ năng suất hai bố mẹ đơn giản, phương pháp phả hệ sâu này có ưu điểm là phân biệt các vùng gen dưới sự chọn lọc lai tạo trên nhiều nền tảng và môi trường di truyền và mạnh mẽ đối với các hiện tượng biến dạng phân ly có thể được quan sát thấy đối với một tổ hợp bố mẹ đơn giản.

Mục tiêu của nghiên cứu này là xác định các kiểu đơn bội dưới sự chọn lọc của nhà chọn giống từ PI 416937. Các kiểu đơn bội được định nghĩa là các SNP liên tiếp được thừa hưởng từ một kiểu gen tiền thân của cá thể. Phương pháp luận này có ứng dụng rộng rãi cho bất kỳ dòng lai nào phổ biến trong tổ tiên của các giống năng suất cao. Chúng tôi cũng đã thảo luận về ý nghĩa của các kiểu đơn bội được xác định thông qua phương pháp luận này để nhắm mục tiêu các vùng gen thiếu sự đa dạng trong các giống mới.

VẬT LIỆU VÀ PHƯƠNG PHÁP

Vật liệu cây trồng và phát triển quần thể

Dòng có nguồn gốc từ PI 416937

Các dòng năng suất cao có nguồn gốc PI 416937 được chọn dựa trên việc đưa vào các Thí nghiệm đồng nhất Nam Mỹ của USDA, cho thấy rằng các dòng này có tiềm năng năng suất tuyệt vời dựa trên các Thí nghiệm năng suất lặp lại trước đó. Các dòng có nguồn gốc từ PI 416937 kéo dài các nhóm trưởng thành (MG) V-VIII. Sự kết hợp giữa dòng có nguồn gốc PI 416937 và dòng gốc trực tiếp của nó được xác định là một bộ ba. Có tổng cộng 52 bộ ba được biên soạn và mỗi thành viên của bộ ba được định kiểu gen để phân tích phả hệ. Đối với bảy trong số các bộ ba này, cả bố và mẹ đều có nguồn gốc từ PI 416937. Có một số dòng được bắt nguồn từ các sự kết hợp bố mẹ giống nhau. 13 trong số 29 sự kết hợp bố mẹ duy nhất có nhiều thế hệ con cháu, mỗi sự kết hợp được coi là bộ ba độc lập. Các dòng này được phát triển bằng phương pháp chọn giống một hạt (SSD), trong đó một quả đơn được chọn từ mỗi cây trong các thế hệ tự phối ban đầu [45] vì vậy mỗi dòng có khả năng cao là một cây F2 duy nhất.

Các dòng được chọn để phân tích đã có mặt trong Thí nghiệm đồng nhất Nam Mỹ của USDA từ năm 1994 đến 2015, ngoại trừ giống N93-110-6 được lai tạo để có năng suất hạt vượt trội trong điều kiện khô hạn [46]. Bốn mươi bốn dòng được đưa vào phân tích đã được lai tạo trong Chương trình chọn giống đậu nành Raleigh của Cơ quan Nghiên cứu Nông nghiệp (ARS) của USDA, trong khi tám dòng còn lại được lai tạo trong Chương trình chọn giống đậu nành của Đại học Georgia (UGA). 52 bộ ba bao gồm tổng cộng 76 dòng lai độc lập. Dữ liệu kiểu gen cho mỗi dòng hoặc được thu thập từ SoyBase (http://soybase.org) [47] hoặc được tạo ra từ hạt giống thu được từ các tổ chức có nguồn gốc.

Phần mềm Helium [48] được sử dụng để tạo một cây phả hệ gồm 232 dòng và 387 mối quan hệ giữa các dòng này (Hình S1). Sáu thế hệ con cháu [N7001, N90-7202, N90-7241, N93-110-6, N91-7254 và N93-1264] có PI 416937 là bố mẹ trực tiếp. Tất cả các vùng gen của PI 416937 được thừa hưởng ở các dòng khác trong phân tích này đều có nguồn gốc từ 6 dòng này. N7001 có ảnh hưởng mạnh nhất đến phân tích của chúng tôi với 12 con cháu trực tiếp và 37 con cháu gián tiếp được sử dụng trong phân tích phả hệ. Thế hệ con cháu gián tiếp được định nghĩa là hậu duệ của thế hệ con cháu trực tiếp. N90-7202 có ảnh hưởng mạnh tiếp theo với sáu thế hệ con cháu trực tiếp và bảy thế hệ con cháu gián tiếp. N90-7241 chỉ có hai con cháu trực tiếp được sử dụng trong phân tích nhưng 11 con cháu gián tiếp. N91-7254 có một thế hệ con cháu và một thế hệ con cháu gián tiếp duy nhất. N93-110-6 có một con cháu trong khi N93-1264 không có con cháu nào được sử dụng trong phân tích. Mặc dù các tổ tiên này thu hẹp phạm vi của các vùng có thể tích lũy một số lượng lớn các thử nghiệm trong phân tích phả hệ, chúng không làm tăng tầm quan trọng của các vùng được tìm thấy được lựa chọn vì mỗi bộ ba là một thử nghiệm độc lập của một vùng.

Sự phát triển của quần thể RIL

Năm quần thể dòng lai tái tổ hợp (RIL) có nguồn gốc F5 được phát triển với mục đích phát triển các dòng để tăng cường tế bào mầm hoặc phóng thích giống. Các quần thể RIL này được sử dụng để đánh giá các alen PI 416937 đang được chọn lọc về ảnh hưởng của chúng đối với năng suất và điều tra sự khác biệt tiềm tàng trong các vùng gen liên quan đến năng suất so với sự biến dạng phân ly. Những quần thể này được lai tạo bằng phương

pháp SSD. Bốn trong số các quần thể RIL (RIL-1, 2, 3, 4) gồm 84 dòng, quần thể RIL thứ năm (RIL-5) có 150 dòng (Bảng S1). Mỗi quần thể RIL có PI 416937 trong phả hệ của nó. Thuật ngữ "bộ ba" được sử dụng trong trường hợp này để chỉ một RIL riêng lẻ và cả hai dòng bố mẹ.

Kiểu gen SNP

20 hạt từ mỗi dòng được trồng trong các cốc xốp trong nhà kính UGA. Sau 3 tuần, mô từ 15–20 cây của mỗi dòng được tạo khối trong các ống Falcon 50ml (Fisher Scientific, Waltham, MA, US), sau đó đông khô và nghiền thành bột mịn bằng GenoGrinder (SPEX US, Metuchen, NJ, CHÚNG TA). DNA được tách chiết bằng cách tuân theo quy trình của Keim và cs [49], với một số điều chỉnh để cải thiện độ tinh khiết của DNA. Các điều chỉnh chính bao gồm thêm đệm chiết xuất Edwards, NaCl, polyvinypyrrolidone và Proteinase K vào đệm CTAB 2ME trong khi thực hiện bước thứ hai 24:1 chloroform: isoamyl alcohol để loại bỏ protein và polysaccharide. Rửa thêm 75% ethanol cũng được thực hiện.

Đối với các dòng được đưa vào phân tích phả hệ, dữ liệu kiểu gen SNP của 10 dòng được lấy từ SoyBase (http://soybase.org) [47] trong khi 66 dòng còn lại được định kiểu gen tại Đại học Bang Michigan hoặc USDA-ARS, (Beltsville, MD) sử dụng SoySNP50K Infinium BeadChips [50]. Các locus SNP không có vị trí tương ứng trong bộ gen tham chiếu Gmax2.0 đã bị loại trừ và tập hợp cuối cùng gồm 41.934 SNP được sử dụng để phân tích. Dữ liệu kiểu gen cho các dòng bổ sung được sử dụng để kiểm tra cấu trúc quần thể cũng được thu thập từ SoyBase (Bảng S2). Năm quần thể RIL được định kiểu gen bằng cách sử dụng SoySNP6K Infinium BeadChips tại USDA-ARS (Beltsville, MD). SoySNP6K Infinium BeadChip là một tập hợp con chứa nhiều thông tin của SoySNP50K Infinium BeadChip giúp cho việc định kiểu gen một số lượng lớn các cá thể như quần thể RIL hiệu quả hơn về chi phí [51]. Vị trí vật lý của SNP, ban đầu dựa trên bộ gen tham chiếu Glyma.Wm82.a1 (Gmax1.01) [52] đã được chuyển đổi thành phiên bản Gmax2.0 để phân tích.

Phân tích phả hệ sử dụng dữ liệu SNP toàn bộ gen

Các phả hệ của các dòng có nguồn gốc PI 416937 này được truy tìm từ các dòng tiền thân có thể phát hiện sớm nhất (Hình S1 và Bảng S3). Phần mềm Helium được sử dụng để hiển thị thông tin phả hệ [48]. Sử dụng dữ liệu SoySNP6K Infinium BeadChip, đóng góp bộ gen của PI 416937 so với các tổ tiên chính ở phía Nam của Bắc Mỹ [53] cho mỗi thế hệ con cháu có năng suất cao có nguồn gốc từ PI 416937 đã được đo lường. Phép thử chi bình phương của các xác suất đã cho được thực hiện trong R, sử dụng chức năng chisq.test trong gói “stats” [54] để kiểm tra có bao nhiêu dòng sai lệch so với tỷ lệ phần trăm đóng góp bộ gen từ PI 416937. Đối với mỗi dòng có nguồn gốc từ PI 416937, tỷ lệ phần trăm quan sát được về đóng góp bộ gen từ PI 416937 dựa trên các chỉ thị phân tử từ SoySNP50K Infinium BeadChip cho phép chúng tôi theo dõi sự thừa kế của các alen trở lại PI 416937 hoặc các kiểu gen tổ tiên khác. Quy trình truy tìm nguồn gốc thừa kế này được giải thích chi tiết hơn trong quy trình phân tích phả hệ. Tỷ lệ phần trăm dự kiến của đóng góp bộ gen từ PI 416937 được thu thập từ thông tin phả hệ.

Để cung cấp bối cảnh của vật liệu di truyền được sử dụng trong phân tích này so với bảng đại diện của mầm đậu nành, các biểu đồ được tạo ra để kiểm tra cấu trúc quần thể của các dòng có nguồn gốc PI 416937 so với các dòng tổ tiên ở Bắc Mỹ, cũng như các dòng công khai sẵn có/mầm đậu tương tư nhân/giống được phát hành trước năm 2016 (Hình S2 và S3). Những vật liệu này bao gồm PI 416937, 95 dòng có nguồn gốc từ PI 416937, 32

dòng phía Nam được chọn lọc, 38 dòng tổ tiên của Bắc Mỹ, 464 giống công và 70 giống tư nhân (Bảng S2). Mười chín trong số 95 dòng có nguồn gốc PI 416937 không được đưa vào phân tích phả hệ vì chúng không đóng góp vào bộ ba hoàn chỉnh với dữ liệu kiểu gen. 19 dòng này được giữ lại để phân tích cấu trúc quần thể nhằm cung cấp bối cảnh đầy đủ hơn về nơi cư trú của các dòng có nguồn gốc PI 416937 so với các dòng tổ tiên và các giống mới. 32 dòng phía Nam được chọn bởi vì chúng là các dòng gốc không phải PI 416937 từ các bộ ba cũng như các tiền thân phổ biến trong các phả hệ phía Nam đã được xác định trong khi truy tìm lịch sử phả hệ của bộ ba. Các dòng công cộng và tư nhân có sẵn được chia thành các nhóm dựa trên thập kỷ phát hành, từ năm 1940 đến năm 2000. Cladogram được tạo từ dữ liệu SoySNP50K Infinium BeadChip thông qua phương pháp phân nhóm huyết thống trong Tassel 5 [55] và được vẽ bằng ggtree [56] được triển khai trong R [54]. Phân tích cấu trúc quần thể này cũng hỗ trợ việc phân nhóm MG để so sánh giữa các kiểu đơn bội được chọn lọc từ PI 416937 và các vùng gen có mức độ đa dạng thấp ở các giống đậu tương hiện đại.

Phân tích phả hệ được thực hiện trên các dòng có nguồn gốc PI 416937 và năm quần thể RIL. Trong phần này, các dòng và RIL có nguồn gốc từ PI 416937 được gọi là con cháu. Bước đầu tiên là xác định các alen nào được thừa hưởng từ mỗi dòng bố mẹ bằng cách so khớp tuần tự các alen của mỗi dòng con với mỗi dòng bố mẹ. Để phù hợp với việc phân tích các dòng năng suất cao có nguồn gốc PI 416937, không có bản đồ di truyền nào được lập cho các quần thể RIL. Mục tiêu là để tính toán điểm số mở rộng trận đấu, xác định điểm số bố mẹ đã đóng góp một khu vực cụ thể của kiểu con lai cận huyết. Mỗi alen phù hợp có giá trị một điểm. Nếu một locus ở dòng bố mẹ hoặc dòng con là dị hợp tử, thì nó được gọi là một nửa điểm. Các giá trị bị thiếu có giá trị bằng không. Sự kết hợp này tiếp tục cho đến khi sự kết hợp bị phá vỡ bởi một alen đối diện có mặt ở bố mẹ. Sau đó, đối với một phân đoạn phù hợp nhất định, điểm được tính dựa trên chất lượng của trận đấu và trận đấu có điểm cao hơn được gọi là gốc của điểm gốc. Đối sánh gốc phải có ít nhất hai marker dài hơn trong một bố mẹ để được coi là một bố mẹ so với đối tượng khác. Về lý thuyết, một vùng gen được gọi cho một bố mẹ có thể là từ một trận đấu liên tiếp ngắn hơn bởi vì trận đấu dài hơn với bố mẹ kia có thừa locus dị hợp tử hoặc thiếu điểm dữ liệu. Chiến lược này tương tự như chiến lược đối sánh haplotype được thực hiện trong Vaughn và Li [53]. Nếu một vùng nhất định có cùng điểm ở hai vùng bố mẹ hoặc trận đấu không còn hai marker ở một vùng cha mẹ, thì vùng đó chắc chắn không thể được gọi cho một trong hai vùng cha mẹ và nó được gọi là không rõ ràng. Sau khi các vùng bố mẹ được xác định, nguồn gốc của mỗi vùng được xác định có liên quan đến PI 416937 cũng như các dòng tổ tiên chủ yếu của vật liệu ưu tú miền Nam của Bắc Mỹ theo Vaughn và Li [53]. Tổ tiên miền Nam của Bắc Mỹ được chọn so với tất cả tổ tiên Bắc Mỹ vì các kiểu gen được sử dụng trong phân tích chủ yếu bao gồm tế bào mầm miền Nam và điều này cũng dẫn đến ít mơ hồ hơn trong việc xác định nguồn gốc tổ tiên cho các vùng gen. Bằng cách này, nguồn gốc bộ gen của các dòng con cháu này có thể được truy ngược từ nguồn gốc của cha mẹ chúng và sau đó là nguồn gốc của tổ tiên ban đầu của chúng. Phân tích tập trung vào các vùng trong đó một bên bố mẹ chứa alen từ PI 416937 tại một vị trí cụ thể và bố mẹ còn lại thì không. Với số lượng đủ các tình huống như vậy, xác suất quỹ tích là trung tính có thể được đánh giá thống kê bằng cách sử dụng phép thử nhị thức chính xác hai mặt. Ngay cả khi một vùng không thể được chỉ định dứt khoát cho tổ tiên phía Nam, nó vẫn được coi là một phép thử nếu có thể xác định được alen đó không phải từ PI 416937 (giả sử alen thay thế là từ PI 416937).

Để có được giá trị P cho các alen đang được lựa chọn, một phép thử nhị thức chính xác hai mặt [57] được thực hiện trong R, sử dụng chức năng binom.test trong gói "stats" [54]. Nếu chúng ta giả định có 50% cơ hội thừa kế alen từ kiểu gen của bố mẹ và phân phối kết quả theo kiểu nhị thức khi không có áp lực chọn lọc nào được áp dụng, thì một phép thử nhị thức chính xác hai mặt là phù hợp để đánh giá độ lệch thống kê so với một giả thuyết vô hiệu về không có áp lực lựa chọn. Số lần thành công được nhập là số lần di truyền alen PI 416937. Số lần thử nghiệm được nhập là số bộ ba trong đó alen PI 416937 được thử nghiệm ở bố mẹ. Không có lựa chọn nào, tỷ lệ thành công được mong đợi là 0,5. Một haplotype trước đây liên quan đến năng suất hạt từ PI 416937 được gọi là Yld1 [58] đã được sử dụng để đặt ngưỡng cho bằng chứng chọn lọc quan trọng. Sử dụng 66 $F_{4:6}$ RIL có nguồn gốc từ N07-14221 × Clifford (PI 596414) đang phân ly cho alen PI 416937 tại locus Yld1, Eickholt [58] trước đây đã chỉ ra rằng alen PI 416937 dẫn đến một kết quả có ý nghĩa ($P < 0,05$) tăng năng suất hạt 76 kg/ha. Đây là mức trung bình trên ba địa điểm trong năm 2015. Trong phân tích phả hệ của chúng tôi, haplotype Yld1 từ PI 416937 cho thấy một tín hiệu mạnh mẽ để lựa chọn so với các haplotype có nguồn gốc PI 416937 khác, cung cấp thêm hỗ trợ cho việc sử dụng giá trị P của vùng gen này như một ngưỡng thống kê. Đối với các quần thể RIL, các vùng gen dường như chịu áp lực chọn lọc lớn hơn đáng kể nên chúng được xác định là có ý nghĩa dựa trên một ngưỡng đã hiệu chỉnh nhiều thử nghiệm. Ngưỡng ý nghĩa được đặt ở mức alpha là 0,05 và hiệu chỉnh nhiều phép thử được thực hiện dựa trên sự mất cân bằng liên kết (LD) giữa các marker. Để điều chỉnh ngưỡng ý nghĩa cho LD, số lượng marker có bình phương R nhỏ hơn 0,8 đã được tính toán và những marker này được gắn nhãn là SNP của thẻ. SNP của thẻ được xác định bằng cách sử dụng chức năng gắn thẻ trong Haploview [59–60]. Sau đó, ngưỡng giá trị P là 0,05 được chia cho số SNP của thẻ để có được ngưỡng ý nghĩa. Số SNP của thẻ nằm trong khoảng từ 423 đến 616 tùy thuộc vào quần thể.

Đối với cả phân tích dòng có nguồn gốc PI 416937 và phả hệ RIL, các vùng được lựa chọn được xác định là bất kỳ loạt marker nào liên tiếp vượt qua ngưỡng ý nghĩa thống kê đã chọn của chúng tôi. Có một số tình huống trong đó các marker cho thấy các mức độ ý nghĩa khác nhau trong một loạt các marker liên tiếp. Các marker có mức ý nghĩa cao nhất trong những lần chạy các marker liên tiếp này được gọi là vùng đỉnh. Các marker trong kết quả của chúng tôi không được kiểm tra trong ít nhất 10 bộ ba đã bị loại trừ.

Thí nghiệm năng suất và phân tích RILs

Năm quần thể RIL có PI 416937 hiện diện trong phả hệ của chúng đã được kiểm tra năng suất để xác nhận các vùng gen được xác định là đang được chọn lọc từ phân tích phả hệ có nguồn gốc PI 416937. Đánh giá năng suất của RIL-1 được thực hiện vào năm 2014 và 2016 tại hai địa điểm ở Georgia (Athens và Plains). RIL-2 được đánh giá trong cùng năm nhưng chỉ ở Athens vì thiếu hạt giống. Đánh giá năng suất của RIL-3 và RIL-4 đã được thực hiện vào năm 2015 và 2016 tại các địa điểm giống như RIL-1. Mỗi quần thể bao gồm 84 $F_{5:6}$ RIL, được chia theo thời gian sinh trưởng thành hai tập hợp con, mỗi tập hợp 42 RIL để đánh giá năng suất. Hai giống ưu tú được đưa vào làm kiểm tra trong mỗi tập hợp con. Các thí nghiệm này được thực hiện trong một thiết kế khối hoàn chỉnh ngẫu nhiên với hai lần lặp lại cho mỗi địa điểm vào năm 2014 và 2015 và ba lần lặp lại cho mỗi địa điểm vào năm 2016. Đối với cả hai địa điểm, các dòng được trồng trong các ô hai hàng, dài 4,9m và cách nhau 76cm. Các ô được cắt còn 3,7m ở giai đoạn R5-6 và được thu hoạch để có năng suất. Dữ liệu năng suất được chuẩn hóa trên cơ sở độ ẩm 13%. Thời gian sinh trưởng được ghi nhận là số ngày đến ngày chín kể từ ngày 1 tháng 9.

Đánh giá năng suất của RIL-5 đã diễn ra trên 5 môi trường vào năm 2014 và 2015 ở Georgia và Louisiana. Các thí nghiệm này cũng được thực hiện trong một thiết kế khối hoàn chỉnh ngẫu nhiên với hai lần lặp lại trên mỗi môi trường. Tập hợp này bao gồm 150 RIL có nguồn gốc từ F5 được tách thành ba tập hợp con gồm 50 RIL/tập hợp dựa trên thời gian sinh trưởng. Hai giống kiểm tra ưu tú được đưa vào hai lần trong mỗi tập hợp con. Đối với môi trường Georgia (Athens và Plains), các ô giống như RIL1-4. Tại Thành phố Bossier, LA, RIL được trồng trên các ô hai hàng dài 4,9m và cách nhau 102cm và cả hai hàng đều cho thu hoạch. Các ghi chú về độ trưởng thành cho tất cả năm quần thể RIL đã được ghi lại ở Athens, GA trên tất cả các lần lặp lại trong mỗi năm.

Các quần thể RIL này được định kiểu gen với SoySNP6K Infinium BeadChip và đóng góp bộ gen của tổ tiên được xác định bằng cách sử dụng cùng một phương pháp như đối với các dòng có nguồn gốc PI 416937. PI 416937 các khu vực tách biệt trong các quần thể RIL này đã được xác định. Một số vùng PI 416937 được phát hiện là đang được chọn lọc trong phân tích phả hệ của các dòng có nguồn gốc PI 416937 dường như đang tách biệt trong các quần thể RIL, vì vậy các quần thể RIL này đóng vai trò là nguồn xác thực tiềm năng về tác động của các vùng này đối với năng suất. Đối với mỗi quần thể RIL, các mô hình hỗn hợp được sử dụng để ước tính năng suất trên các môi trường với vùng có nguồn gốc PI 416937 được lựa chọn dưới dạng hiệu ứng cố định và môi trường, vùng gen × môi trường và tập hợp con trong môi trường là ảnh hưởng ngẫu nhiên. Phép thử so sánh nhiều nghĩa là Tukey HSD được thực hiện trên mỗi vùng phân tách để biết năng suất hạt.

Một phân tích tương tự đã được thực hiện đối với sự trưởng thành để kiểm tra xem có sự khác biệt đáng kể về sự trưởng thành liên quan đến các vùng gen từ PI 416937 trong quần thể RIL hay không. Để hoàn thiện, mô hình hỗn hợp phải được đơn giản hóa bằng cách loại bỏ thuật ngữ tương tác để phát hiện sự khác biệt thống kê. Các phân tích thống kê cho tất cả các phân tích mô hình hỗn hợp đã đề cập trước đây được thực hiện bằng cách sử dụng JMP® Pro 13.0.0 [61].

KẾT QUẢ

PI 416937 đóng góp bộ gen trong các dòng có năng suất cao

Một thử nghiệm chi-square được thực hiện để kiểm tra dòng nào có sự khác biệt đáng kể về tỷ lệ phần trăm của PI 416937 và tổ tiên phía Nam được ước tính bằng các marker so với những gì được mong đợi bởi phả hệ (Bảng 1). Các khu vực không rõ ràng không được đưa vào phân tích vì vậy tỷ lệ phần trăm bộ gen được thừa hưởng từ PI 416937 và tổ tiên phía Nam được chuẩn hóa cho mỗi dòng.

Theo phả hệ, tỷ lệ phần trăm bộ gen PI 416937 dao động từ 12,5 đến 50,0% ở thế hệ con cháu có năng suất cao (Bảng 1). Sử dụng dữ liệu marker, phần trăm bộ gen PI 416937 dao động từ 6,8 đến 51,0% trên tất cả các bộ ba. N09-12455 chứa nhiều hơn 2,7 lần bộ gen PI 416937 theo marker so với dự đoán theo phả hệ (33,7% thực tế so với 12,5% dự đoán). N93-1264 chứa bộ gen PI 416937 theo marker ít hơn 1,6 lần so với dự đoán theo phả hệ (30,9% thực tế so với 50,0% dự đoán). Đây là những khác biệt lớn nhất được quan sát thấy trong dự đoán và đo lường. N96-6755 chứa phần lớn nhất của bộ gen PI 416937 với 51,0% dựa trên dữ liệu marker không khác biệt ($P < 0,05$) so với 50% đã được dự đoán. N05-7375 chứa phần nhỏ nhất của bộ gen PI 416937 với 6,8% dựa trên dữ liệu marker. Điều này khác biệt có ý nghĩa ($P < 0,05$) so với 25% được dự đoán.

13 trong số 52 dòng năng suất cao thu được từ PI 416937 có tỷ lệ PI 416937 với tổ tiên phía Nam khác nhau có ý nghĩa so với những gì đã được dự đoán (P <0,05). Sáu dòng chứa nhiều hơn đáng kể bộ gen PI 416937 và bảy dòng chứa bộ gen PI 416937 ít hơn đáng kể so với dự đoán (P <0,05). Dường như không có lợi thế chọn lọc để thừa hưởng một phần lớn hơn của bộ gen PI 416937.

Bảng 1. Các bộ ba được sử dụng trong phân tích phả hệ PI 416937 và kiểm định chi-square để kiểm tra độ lệch của sự đóng góp bộ gen dự kiến so với quan sát được của PI 416937 và tổ tiên phía Nam trong mỗi dòng có năng suất cao PI 416937. Được đặt hàng theo% PI 416937 bởi các marker.

Name[a]	MG	% PI 416937 by pedigree	% PI 416937 by markers	% Southern ancestor by pedigree	% Southern ancestor by markers	% Ambiguous by markers	Female parent	Male parent	Year(s) entered in USDA Uniform Test
N05-7375***	VI	25.0	6.8	75.0	84.5	8.7	N7002[b]	N98-7265	2009, 2010, 2011
G08-3279RR	VIII	12.5	8.0	87.5	80.3	11.6	Woodruff[b]	G03-952RR	2011, 2012, 2013, 2014
N06-7564	VII	12.5	8.1	87.5	80.2	11.6	NC-Roy	N8001[b]	2008, 2009, 2010
N07-14221	V	12.5	10.7	87.5	83.5	5.7	N7002[b]	Clifford	2012
G08-3282RR	VIII	12.5	11.1	87.5	76.1	12.7	Woodruff[b]	G03-952RR	2011, 2012
G10-3896RR	VIII	12.5	11.9	87.5	82.2	5.9	G03-825RR	G00-3213[b]	2013
N05-7452	VII	12.5	12.2	87.5	78.2	9.5	N7002[b]	5601T	2007, 2008, 2009, 2010, 2011
N05-7353***	VI	25.0	12.2	75.0	83.0	4.8	N7002[b]	N98-7265	2009, 2010, 2011
N07-15546	VI	12.5	12.4	87.5	79.4	8.1	N7002[b]	PI 221717	2012
N06-7535	VII	12.5	12.5	87.5	77.3	10.1	NC-Roy	N8001[b]	2009, 2010
N05-7396**	VII	25.0	13.0	75.0	76.3	10.6	N7002[b]	N98-7265	2007, 2008 2009, 2010
N8002**	VIII	25.0	13.5	75.0	80.7	5.7	N7002[b]	N98-7265	2007, 2008 2009, 2010, 2011, 2012, 2013, 2014, 2015
N01-11118*	VII	25.0	13.8	75.0	73.6	12.6	NTCPR94-5157	N96-7031[b]	2005
N07-15529	VII	12.5	14.0	87.5	73.0	12.9	N7002[b]	PI 221717	2014, 2015
N01-11777**	VII	25.0	14.5	75.0	78.9	6.5	Graham	N96-7031[b]	2004, 2005, 2006, 2007, 2008, 2009
G08-2869RR	VIII	12.5	14.9	87.5	74.6	10.5	Woodruff[b]	G03-364RR	2011, 2012
G10-3833RR	VII	12.5	15.3	87.5	75.0	9.7	G03-825RR	G00-3213[b]	2013, 2014
N06-7280*	VI	25.0	15.5	75.0	76.6	7.9	N98-7265	N7002[b]	2009, 2010
N8001	VIII	25.0	16.7	75.0	73.2	10.1	N7001[b]	Cook	2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007
N07-14182	VI	12.5	16.8	87.5	73.9	9.2	N7002[b]	Clifford	2011, 2012
G00-3083	VIII	25.0	17.0	75.0	76.3	6.7	N7001[b]	Benning	2003
N05-7380	VII	25.0	18.0	75.0	73.7	8.2	N7002[b]	N98-7265	2012
N01-11136	VII	25.0	18.9	75.0	70.7	10.3	NTCPR94-5157	N96-7031[b]	2004, 2005, 2006, 2007, 2008, 2009
N01-11491	VII	25.0	19.7	75.0	64.6	15.6	NTCPR94-5157	N96-7031[b]	2005, 2006, 2007
N01-11771	VII	25.0	21.1	75.0	71.3	7.5	Graham	N96-7031[b]	2006, 2007, 2008, 2009
N05-7281	VII	25.0	21.3	75.0	71.7	6.9	N96-6809[b]	N98-7265	2007, 2008, 2009, 2010, 2011
N01-11424	VIII	25.0	21.4	75.0	69.0	9.5	NTCPR94-5157	N96-6767[b]	2006, 2007, 2008
N01-11884	VII	25.0	22.6	75.0	70.1	7.2	Graham	N96-6767[b]	2006, 2007
Woodruff	VII	25.0	23.2	75.0	70.3	6.4	N7001[b]	Boggs	2003, 2004, 2005, 2006

N05-7462	VII	25.0	23.7	75.0	69.6	6.6	5601T	N96-6809[b]	2007, 2008, 2009, 2010, 2011
N01-11791	VII	25.0	23.8	75.0	70.6	5.6	Graham	N96-7031[b]	2005
N01-11832	VIII	25.0	25.4	75.0	69.0	5.5	Graham	N96-7031[b]	2005
TCWN23-507	VI	25.0	25.6	75.0	60.7	13.7	N77-114	N96-6809[b]	2007
N06-7187	VIII	25.0	25.6	75.0	69.6	4.7	N98-7265	N93-110-6[b]	2009, 2010, 2011
N05-316	VI	25.0	25.6	75.0	65.2	9.1	NC-Roy	N96-6752[b]	2013, 2014, 2015
N7002	VII	25.0	25.9	75.0	68.6	5.4	N7001[b]	Cook	2000, 2001, 2003, 2004, 2005, 2006, 2007
G07-3557RR***	VIII	12.5	26.2	87.5	64.1	9.6	G00-3213[b]	P97M50	2010, 2011, 2012
N09-12414***	VII	12.5	27.5	87.5	60.6	11.8	N7002[b]	Misuzu Daizu	2011, 2012
N09-12441***	VII	12.5	29.8	87.5	58.3	11.8	N7002[b]	Misuzu Daizu	2013
N93-1264**	V	50.0	30.9	50.0	59.0	10.0	Brim	PI 416937	1998
N09-12455***	VII	12.5	33.7	87.5	52.9	13.3	N7002[b]	Misuzu Daizu	2013, 2014
N99-8141*	V	25.0	34.3	75.0	60.5	5.1	N7001[b]	Graham	2002, 2003
N04-8947	VII	50.0	40.2	50.0	55.3	4.4	N96-6894[b]	N97-9812[b]	2008, 2009, 2010
N05-7229	VII	50.0	41.8	50.0	42.4	15.7	N96-6809[b]	N96-7031[b]	2007
N93-110-6	VI	50.0	42.0	50.0	55.3	2.6	Young	PI 416937[b]	Devi et al., 2014[c]
N05-7260	VII	50.0	43.6	50.0	51.1	5.2	N96-6809[b]	N96-7031[b]	2007, 2008, 2009, 2010
N90-7202	VII	50.0	46.7	50.0	49.3	4.0	N77-114	PI 416937[b]	1994
N7001	VII	50.0	48.7	50.0	47.6	3.7	N77-114	PI 416937[b]	1994, 1995, 1996, 1997
N96-6751	VII	50.0	49.7	50.0	42.6	7.7	N90-7202[b]	N7001[b]	1998
N96-6752	VIII	50.0	49.9	50.0	45.6	4.4	N90-7202[b]	N7001[b]	1999, 2000, 2001, 2002, 2003, 2004, 2005
N96-6809	VII	50.0	50.5	50.0	44.6	4.8	N90-7202[b]	N7001[b]	1998, 1999, 2000, 2001, 2002
N96-6755	VI	50.0	51.0	50.0	43.3	5.6	N90-7202[b]	N7001[b]	2001, 2002, 2003

, ** và * biểu thị sự khác biệt đáng kể về sự thừa kế ước tính theo phả hệ so với sự thừa kế được đo lường bằng marker theo kiểm tra chi-square cho các xác suất đã cho tương ứng với alpha là 0,05, 0,01 và 0,001.*

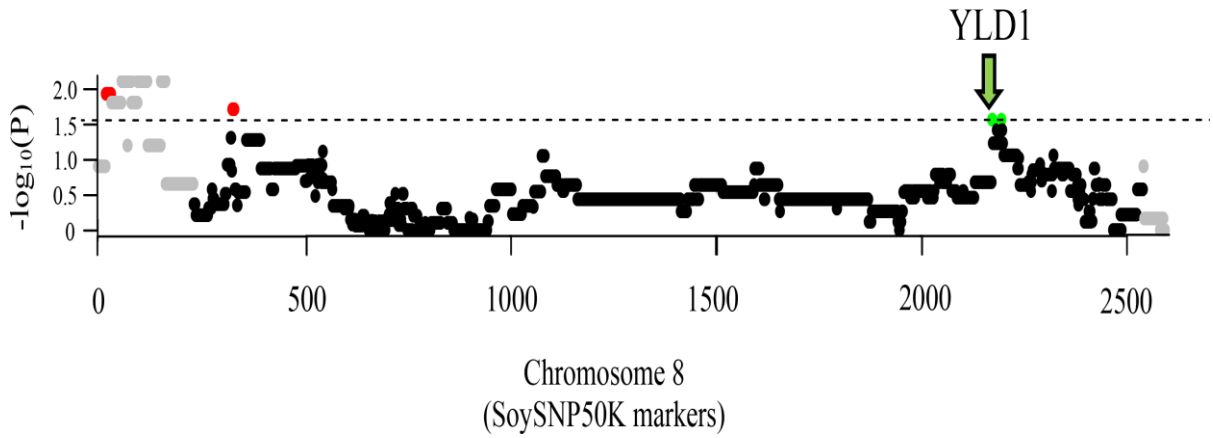
[a] Tiền tố G và N trong tên biểu thị các dòng được phát triển tương ứng tại Đại học Georgia và tại USDA Raleigh. Woodruff được phát triển tại Đại học Georgia trong khi TCWN23-507 được phát triển tại USDA Raleigh.

[b] Bố mẹ có PI 416937 trong phả hệ.

[c] Dòng không được đánh giá trong Thí nghiệm đồng nhất của USDA, nhưng đã được lai tạo để có năng suất hạt vượt trội trong điều kiện khô hạn.

Khám phá các vùng gen trong cả lựa chọn tích cực và tiêu cực

Mặc dù không có lợi thế chọn lọc đối với PI 416937 về đóng góp tổng thể của bộ gen, người ta đã giả thuyết rằng có sự tương đồng giữa các vùng gen cụ thể từ PI 416937, một số có lợi, một số có hại, được di truyền cho thế hệ con cháu có năng suất cao. Sử dụng 52 bộ ba, các vùng gen từ PI 416937 được xác định theo cả lựa chọn tích cực và tiêu cực. Cấu trúc của phân tích phả hệ cho phép xác định trực tiếp mô hình rỗng và đánh giá các giá trị P. Ý nghĩa thống kê của vùng PI 416937 với bằng chứng trước đây về mối liên quan với năng suất được đặt làm ngưỡng thực nghiệm. Vùng này (Yld1) được xác định là vùng quan trọng thứ tám của chúng tôi về sự chọn lọc dương tính vì nó đã được thử nghiệm trong 41 bộ ba và được di truyền ở 28 thế hệ con cháu có năng suất cao ($P = 2,75 \times 10^{-2}$) (Hình 1, tệp bổ sung S1). Bất kỳ marker hoặc haplotype nào đáp ứng mức ý nghĩa thống kê này hoặc cao hơn đều được xác định là các vùng có bằng chứng lựa chọn ủng hộ hoặc chống lại. Tổng cộng, chín vùng gen được chọn lọc dương tính và 17 vùng gen được chọn lọc âm tính từ PI 416937 đã được xác định trên bảy nhiễm sắc thể (Bảng 2 và Hình S4). Kết quả đầy đủ từ phân tích phả hệ PI 416937 được trình bày trong Tệp S2.



Hình 1. Phân tích phả hệ dựa trên toàn bộ bộ gen PI 416937 của nhiễm sắc thể số 8.
Phần trên cùng chỉ ra các vùng gen trên nhiễm sắc thể số 8 trong sự chọn lọc dương tính (xanh lá cây) so với tiêu cực (đỏ). Ngưỡng thống kê được đặt ở - $\log_{10}$ (P) là 1,56 (Yld1). Trục x hiển thị vị trí marker từ SoySNP50K Infinium BeadChips. Màu xám chỉ ra một locus có ít hơn 10 bộ ba kiểm tra alen PI 416937 so với alen thay thế. Màu đen cho biết quỹ tích có 10 bộ ba trở lên nhưng đã giảm xuống dưới ngưỡng ý nghĩa của chúng ta.

Các vùng được lựa chọn tích cực có độ dài từ một marker đến 76 marker liên tiếp. Các vùng lần lượt được tìm thấy trên Chrs 8 (2 vùng), 13 (3 vùng) và 17 (4 vùng) (Bảng 2 và Hình S4). Khoảng cách vật lý của vùng lớn nhất được chọn lọc tích cực là 985,307bp trên Chr 17. Có ba vùng gen khác ở Chrs 13 và 17 có bằng chứng lớn nhất về chọn lọc tích cực ($P = 2{,}49 \times 10^{-3}$). Vùng đầu tiên là vùng 99.528bp trên Chr 13. Vùng thứ hai là vùng 9.223bp trên Chr 17 nằm trong vùng quan trọng lớn hơn là 206.570bp. Vùng gen thứ ba có bằng chứng lớn nhất về sự chọn lọc tích cực nằm trong khoảng thời gian 985.307bp và cũng trên Chr 17.

Các khu vực được xác định theo lựa chọn phủ định có độ dài dao động từ một marker đến 137 marker liên tiếp. Các khu vực được tìm thấy tương ứng trên Chrs 5 (2 khu vực), 8 (2 khu vực), 9 (2 khu vực), 12 (2 khu vực), 13 (6 khu vực), 16 (1 khu vực), 17 (1 khu vực) và 19 (1 vùng) (Bảng 2 và Hình S4). Khoảng cách vật lý của vùng lớn nhất được chọn lọc âm là 4.011.395bp trên Chr 12. Một vùng đỉnh được xác định có ý nghĩa cao hơn trong vùng lớn hơn này, dài 2.748.460bp. Vùng gen có bằng chứng chọn lọc âm tính lớn nhất ($P = 1{,}95 \times 10^{-3}$) dài 82.263bp trên Chr 13.

Đánh giá các vùng có PI 416937 đang được lựa chọn bằng cách sử dụng các quần thể RIL

Sử dụng năm quần thể RIL có PI 416937 trong phả hệ của chúng, chúng tôi đã kiểm tra lợi thế năng suất liên quan đến bất kỳ vùng gen nào được chọn lọc. Phương pháp lập bản đồ QTL tiêu chuẩn không được sử dụng do có độ chính xác thống kê cao hơn thu được từ các phép so sánh trực giao được lập kế hoạch trước [62] do phân tích phả hệ của các vùng gen thu được từ PI 416937.

Bảy vùng được chọn lọc tích cực được xác định từ phân tích phả hệ của các dòng có nguồn gốc PI 416937 đã được tách biệt trong ít nhất một trong các quần thể RIL này. Các vùng 5 và 6 trên Chr 8 được tách biệt thành ba quần thể RIL (RIL-1, 3 và 5); vùng 19 trên Chr 13 trong bốn quần thể RIL (RIL-1, 2, 4 và 5); và các vùng 21, 22, 23 và 24 trên Chr 17 trong ba quần thể RIL (RIL-1, 2 và 5). Phân tích năng suất được thực hiện trong các quần thể riêng lẻ trên các môi trường và ảnh hưởng năng suất có ý nghĩa duy nhất là loại đơn bội PI 416937 cho vùng 19 có hiệu quả năng suất dương là 27 kg/ha ($P < 0{,}05$)

trong RIL-4 và trưởng thành 0,8 ngày sau đó. Loại haplotype PI 416937 này dẫn đến hiệu quả năng suất dương lớn hơn ở RIL-2 (41 kg/ha) và RIL-5 (76 kg/ha) mặc dù những tác động này không có ý nghĩa (P <0,05). Cũng có sự khác biệt không đáng kể về số ngày đến ngày chín giữa các haplotype trong RIL-2 và RIL-5 vì RILs với haplotype PI 416937 chín sớm hơn 0,4 ngày và 0,1 ngày sau đó. Loại haplotype PI 416937 có ảnh hưởng tiêu cực đến 105 kg/ha so với haplotype thay thế trong RIL-1 nhưng không có ý nghĩa (P <0,05). Hai trong số bảy vùng cách ly dao động giữa việc có tác động tích cực và tiêu cực và năm vùng còn lại có tác động tiêu cực đến các quần thể cách ly, mặc dù các tác động này không đáng kể (P <0,05).

Bảng 2. Tóm tắt phân tích phả hệ PI 416937

Genomic region	Chr	Direction of selection	Physical start position (bp)	Physical stop position (bp)	SNPs	Trios tested	No. of times inherited	*P*-value
1	5	Negative	488551	523725	3	36	10	0.011
			558763	-	1	41	11	0.004
			562322	593685	3	39	10	0.003
			595812	-	1	41	11	0.004
2	5	Negative	1496131	-	1	38	11	0.014
3	8	Negative	1017668	1373179	13	11	1	0.012
4	8	Negative	5750622	5968621	6	19	4	0.019
5	8	Positive	41792467	41796167	2	41	28	0.028
6	8	Positive	42070881	-	1	41	28	0.028
7	9	Negative	45849012	45885099	3	10	1	0.022
			45913326	45941083	2	11	1	0.012
8	9	Negative	46044100	46082968	3	11	1	0.012
9	12	Negative	14014843	14384675	7	20	4	0.012
10	12	Negative	17662053	20410513	34	28	7	0.013
			20474981	21673448	15	27	7	0.019
11	13	Positive	26986028	27085556	14	26	21	0.003
12	13	Negative	27971359	-	1	20	3	0.003
			27979190	27991927	3	19	3	0.004
13	13	Negative	28203902	28286301	19	22	5	0.017
14	13	Negative	28346050	28351526	2	20	3	0.003
15	13	Negative	28475417	29111990	83	22	5	0.017
16	13	Negative	29128801	29533624	48	22	5	0.017
17	13	Negative	30683322	30765585	5	10	0	0.002
18	13	Positive	36166900	36175098	2	32	23	0.020
19	13	Positive	37465322	37527009	18	33	24	0.014
			37529648	37550786	2	35	26	0.006
			37551787	37553062	3	41	29	0.012
20	16	Negative	6871009	6896309	6	12	1	0.006
			6914854	6948002	6	15	2	0.007
21	17	Positive	796471	1309414	26	17	14	0.013
22	17	Positive	2202411	-	1	26	20	0.009
23	17	Positive	2246668	2408798	13	26	20	0.009
			2409261	2418484	3	26	21	0.003
			2419489	2438284	7	26	20	0.009
			2452744	2453238	2	10	9	0.022
24	17	Positive	2510699	3496006	76	26	21	0.003
25	17	Negative	38511430	38757540	26	13	2	0.023
26	19	Negative	1743312	3274996	111	13	2	0.023

Phân tích phả hệ tương tự được thực hiện trên các dòng có nguồn gốc PI 416937 cũng được thực hiện trong mỗi nhóm trong số năm quần thể RIL để xác định xem có bất kỳ vùng nào được lựa chọn tích cực hay tiêu cực trong các quần thể RIL này hay không. Cácquần thể RIL này được phát triển cho mục đích chọn giống và trải qua phương pháp SSD trong ba thế hệ, tiếp theo là cây đơn và sau đó là chọn lọc hàng cây đơn lẻ thông qua đánh giá trực quan trước khi được trồng trong các thử nghiệm nâng cao năng suất. Do đó, các vùng được tìm thấy dưới sự lựa chọn tích cực hoặc tiêu cực đáng kể sẽ là các vùng chịu sự biến dạng phân ly. Các vùng này không phát sinh từ các thử nghiệm năng suất rộng rãi như với các vùng từ phân tích phả hệ PI 416937 được thực hiện ở trên. Các vùng gen chịu áp lực chọn lọc trong các vòng tuyển chọn ban đầu cho năm quần thể RIL này được so sánh với các vùng trong phân tích phả hệ dòng có nguồn gốc PI 416937 về khả

năng chồng chéo tiềm ẩn. Nếu vậy, đây sẽ là bằng chứng cho thấy những vùng này được chọn dựa trên sức sống rõ ràng có thể thấy rõ bằng mắt thường hoặc các vùng từ PI 416937 liên quan đến thể trạng của thế hệ con cháu.

18 vùng được lựa chọn từ RIL-1 đã được xác định trên 10 Chrs (Hình 2 và Bảng S4). Ba khu vực được lựa chọn tích cực đối với Nhóm 1 (1 khu vực), 11 (1 khu vực) và 18 (1 khu vực). Mười lăm vùng được tìm thấy dưới lựa chọn tiêu cực trên Nhóm 1 (1 vùng), 2 (1 vùng), 5 (1 vùng), 6 (2 vùng), 7 (3 vùng), 10 (1 vùng), 11 (3 vùng) , 12 (2 vùng) và 19 (1 vùng). RIL-2 có một vùng duy nhất được lựa chọn tích cực trên Chr 1 (Hình 2 và Bảng S4). RIL-3 không có vùng nào được lựa chọn (Hình 2). RIL-4 có 2 vùng được lựa chọn tích cực (Hình 2 và Bảng S4). Một vùng nằm trên Chr 8 trong khi vùng còn lại ở Chr 18. RIL-5 có một vùng duy nhất được lựa chọn tiêu cực trên Chr 6 (Hình 2 và Bảng S4).

Vùng có bằng chứng quan trọng nhất về chọn lọc tích cực ($P = 1{,}39 \times 10^{-11}$) là vùng 45.711 bp trong RIL-2 nằm trên Chr 1. Mức ý nghĩa này vượt xa vùng có ý nghĩa nhất trong chọn lọc tích cực trong PI 416937 phân tích phả hệ ($P = 2{,}49 \times 10^{-3}$). Vùng quan trọng nhất được lựa chọn tiêu cực trong phân tích phả hệ RIL là vùng 2.227.287bp nằm trên Chr 6 và được phát hiện trong quần thể RIL-5. Vùng này là đỉnh trong một vùng rộng lớn 4.143.902bp. Sự khác biệt có ý nghĩa thậm chí còn lớn hơn đối với các vùng được lựa chọn tiêu cực khi so sánh với phân tích phả hệ PI 416937 ($P = 9{,}19 \times 10^{-26}$ so với $1{,}95 \times 10^{-3}$). Kết quả đầy đủ từ phân tích phả hệ RIL được trình bày trong Tệp S3.

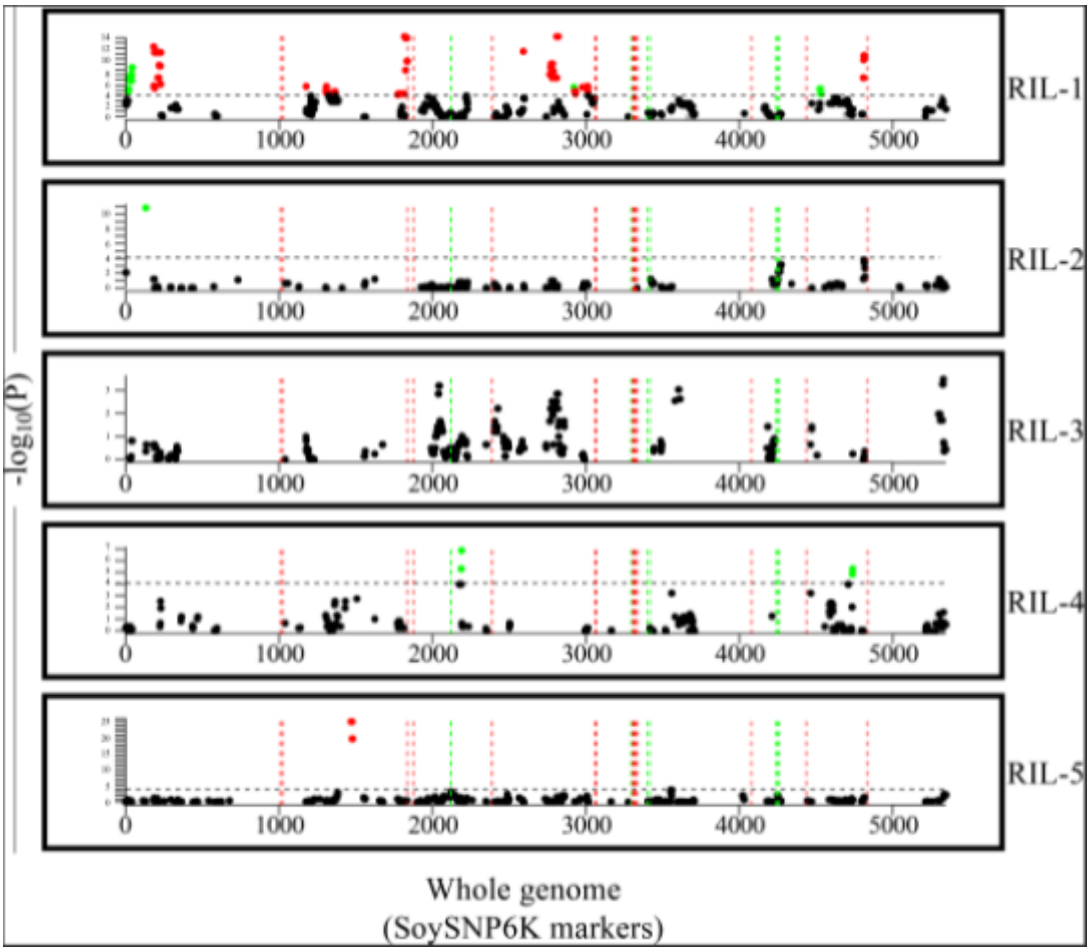
THẢO LUẬN

Trong nghiên cứu này, chúng tôi sử dụng thông tin phả hệ và kiểu gen mở rộng để xác định các vùng gen từ PI 416937 với bằng chứng về sự chọn lọc tích cực và tiêu cực. Năm quần thể RIL sau đó đã được sử dụng trong một nỗ lực để xác nhận các khu vực quan tâm này trong các thử nghiệm năng suất lặp lại, nhưng các mối liên quan về năng suất đáng kể tối thiểu đã được phát hiện. Các vùng gen từ PI 416937 cũng được chọn lọc mạnh mẽ trong quần thể RIL nhưng có vẻ là quần thể cụ thể và không trùng lặp với kết quả từ phân tích phả hệ PI 416937.

Xem xét mức độ phổ biến của PI 416937 trong phả hệ của các dòng năng suất cao trong các Thí nghiệm đồng nhất của USDA, có thể có sự lựa chọn thuận lợi cho toàn bộ bộ gen PI 416937, nhưng chúng tôi không tìm thấy xu hướng nhất quán để thừa kế bộ gen PI 416937 nhiều hơn được mong đợi bởi phả hệ giữa các dòng có nguồn gốc PI 416937. Xem xét PI 416937 là một cây trồng giới thiệu ban đầu được Hệ thống mầm thực vật quốc gia của Mỹ tiếp nhận vào năm 1974 [63], nó không có khả năng cạnh tranh với các loại mầm ưu tú hiện đại hơn về năng suất hạt mỗi giây [64] và khả năng cao là các dạng đơn bội trên toàn bộ bộ gen là trung tính hoặc bất lợi cho tiềm năng năng suất. Các nhà nghiên cứu cũng đã tìm thấy các đặc điểm nông học không thuận lợi liên quan đến PI 416937 như tính mẫn cảm với tuyến trùng đậu tương (*Heterodera glycines*) chủng 3,5 và 14 [65], khiến các nhà chọn giống khó có thể chọn PI 416937 ở cấp độ toàn bộ bộ gen. Phân tích phả hệ đã cung cấp thêm độ phân giải để xác định các kiểu đơn bội cụ thể hơn từ PI 416937 có bằng chứng về sự chọn lọc trong các nỗ lực lai tạo.

Sau khi các vùng gen từ PI 416937 dưới sự chọn lọc thuận lợi được xác định, một so sánh được thực hiện để xác định xem các vùng gen này có trùng lặp với QTL đã tìm thấy trước đây về hiệu quả sử dụng nước hay không [25], khả năng kháng nhôm [21], hình thái rễ [20], khả năng chịu hạn [66], và héo tán cây [15]. Không có sự chồng chéo nào được

phát hiện. Các thí nghiệm nâng cao về năng suất do các nhà chọn giống thực hiện có xu hướng được quản lý chặt chẽ hơn để giảm các tác nhân gây căng thẳng như hạn hán. Nhiều QTL được lập bản đồ từ PI 416937 đã kiểm tra khả năng chống chịu với các điều kiện liên quan đến hạn hán, vì vậy có thể không ngạc nhiên khi QTL được lập bản đồ từ các phép lai liên quan đến PI 416937 không trùng lặp nhiều với các vùng được tìm thấy có liên quan đến năng suất hạt. Ngoài ra, nghiên cứu này đang tìm cách xác định các khu vực đang được lựa chọn trên các môi trường đa dạng chủ yếu ở Bắc Carolina và Georgia cũng như trên các nền tảng di truyền đa dạng từ hai chương trình chọn giống này.



Hình 2. Các vùng gen từ PI 416937 dưới sự lựa chọn tích cực (xanh lá cây) và tiêu cực (đỏ) trong quần thể RIL ở cấp độ toàn bộ bộ gen.

Trục x hiển thị vị trí marker từ SoySNP6K Infinium BeadChips. Ngưỡng thống kê được đặt cho từng quần thể dựa trên hiệu chỉnh nhiều lần kiểm tra được điều chỉnh LD. Màu xám chỉ ra một locus có ít hơn 10 bộ ba kiểm tra alen PI 416937 so với alen thay thế. Màu đen cho biết quỹ tích có 10 bộ ba trở lên nhưng đã giảm xuống dưới ngưỡng ý nghĩa của chúng ta. Các đường nét đứt dọc chỉ ra các vùng được xác định trong quá trình chọn giống từ phân tích phả hệ ban đầu được thực hiện trên các dòng năng suất cao có nguồn gốc từ PI 416937.

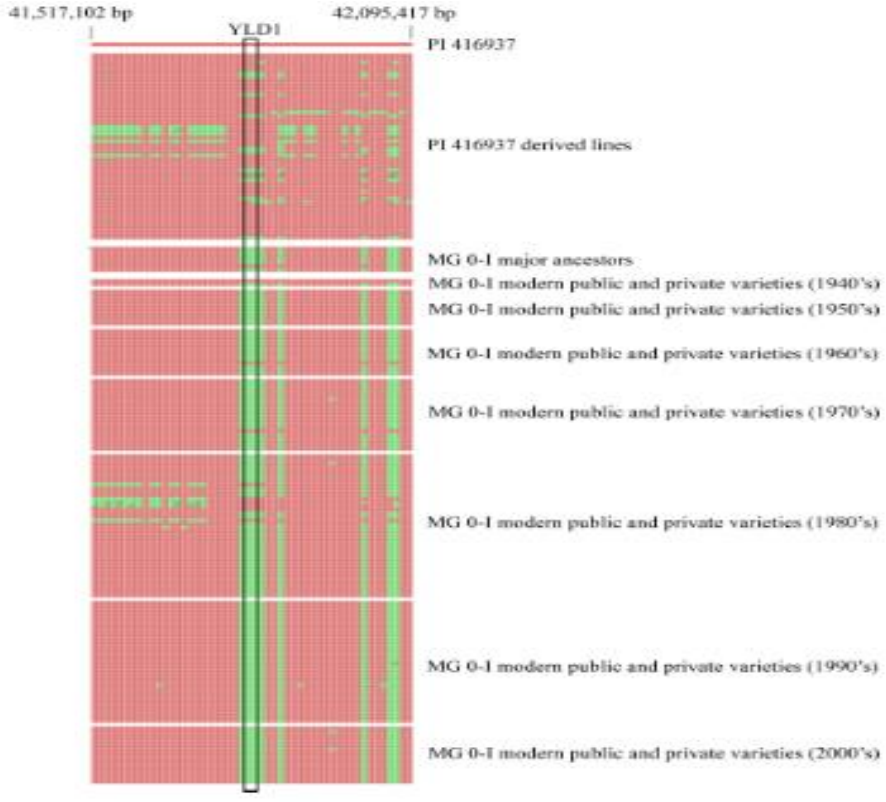
Chỉ một trong số các kiểu đơn bội PI 416937 phân ly trong quần thể RIL được thử nghiệm cho thấy ảnh hưởng có thể phát hiện được đối với năng suất ($P < 0,05$). Khác với các vùng gen đã được xác định do lỗi loại I, một giải thích có thể xảy ra là hiệu ứng haplotype bị nhầm lẫn bởi hiệu ứng nền di truyền khi được thử nghiệm trong quần thể RIL, cung cấp thêm bằng chứng về sự khó khăn trong việc phát hiện ra QTL năng suất hạt đáng kể vượt qua nền tảng di truyền và ảnh hưởng môi trường. Concibido và cs. [7]

đã báo cáo một số thành công trong việc đưa QTL năng suất từ *Glycine soja* (PI 407305) vào dòng đậu tương HS-I. Khi QTL được thâm nhập vào các nền ưu tú khác, sự mâu thuẫn nảy sinh trong các hiệu ứng năng suất được báo cáo. QTL dường như cho thấy khả năng thích ứng hạn chế trên tất cả các nền tảng di truyền. Rất khó để nắm bắt được tác động thực sự của một vùng riêng lẻ đối với năng suất hạt giống vì một số lý do. Năng suất hạt giống là một đặc điểm có số lượng cao, chịu tác động lớn của môi trường và dễ bị sai số kiểu hình (ví dụ, lỗi kết hợp, không tính đến các tác động đồng ruộng). Mặc dù nghiên cứu này đang tìm cách khám phá các vùng có thể vượt qua sự không đồng nhất về di truyền và môi trường, nhưng cũng có thể là các haplotype PI 416937 đang cạnh tranh với các haplotype tương đương hoặc cao cấp hơn. Một yếu tố khác có khả năng góp phần vào việc thiếu tác động thống kê đối với năng suất liên quan đến các loại đơn bội PI 416937 này là G × E. Bốn mươi bốn trong số 52 dòng có nguồn gốc PI 416937 đã được phát triển trong chương trình chọn giống đậu tương Raleigh USDA và do đó đã trải qua giai đoạn phát triển ban đầu và kiểm tra năng suất chủ yếu trong môi trường Bắc Carolina, trong khi quần thể RIL được sử dụng để kiểm tra năng suất chủ yếu được trồng trong môi trường Georgia.

Vùng Yld1 là vùng được xác định rõ ràng (3,7kb) trong phân tích phả hệ PI 416937. Xem xét rằng vùng Yld1 đã được xác định quá hẹp, được hỗ trợ bởi các đánh giá năng suất trong một nghiên cứu trước đây và cho thấy bằng chứng mạnh mẽ về sự chọn lọc của nhà chọn giống được chứng minh là vùng gen cao thứ tám được chọn lọc tích cực trên toàn bộ bộ gen, chúng tôi cảm thấy vùng này xứng đáng khám phá thêm. Trong khoảng thời gian này, có một mô hình gen duy nhất hiện diện cho *Glyma.08g299800*, là mô tả của ATG24090.1, một chitinase A được tìm thấy trong Arabidopsis. *Glyma.08g299800* nằm từ 41,795,912 đến 41,796,546bp (http://soybase.org) [47], một phần trùng lặp với vùng Yld1 nằm từ 41,792,467 đến 41,796,167bp. Chitinase thường liên quan đến khả năng bảo vệ thực vật chống lại nấm bệnh hoặc côn trùng vì chitin là thành phần phổ biến của thành tế bào nấm và bộ xương ngoài của côn trùng [67]. Có một số QTL cho các đặc điểm khác nhau đã được lập bản đồ cho vùng này, một trong số đó là "Sclero 9–2", một QTL liên quan đến tính kháng nấm [68]. "Sclero 9–2" là một QTL liên quan đến khả năng kháng *Sclerotinia sclerotiorum* và được lập bản đồ từ một phép lai của PI 391589B × IA2053. Alen thuận lợi cho QTL này được di truyền từ IA2053, là bố mẹ nhiễm trung bình trong phép lai. Do điều kiện khí hậu của miền Đông Nam Mỹ, có thể cho rằng áp lực nấm là mối quan tâm thường xuyên. Ở Georgia từ năm 2005–2013, trung bình hàng năm chi 2,7 triệu đô la cho việc kiểm soát bệnh hại cây trồng từ các tác nhân gây stress sinh học, chủ yếu là bệnh gỉ sắt đậu tương châu Á (*Phakopsora pachrhizi*) (UGA CAES, 2017) [69]. Nếu PI 416937 haplotype tại Yld1 cung cấp khả năng kháng nấm bệnh ở mức độ trung bình, thì có thể hiểu được tại sao nó cho thấy mối liên quan với năng suất hạt ở các dòng lai ở miền Nam nước Mỹ, mặc dù áp lực này thay đổi đáng kể theo năm, dẫn đến G × E đáng kể. Điều thú vị là Eickholt [58] đã báo cáo sự hiện diện của tương tác môi trường × QTL liên quan đến Yld1 ở ba địa điểm, cũng cho thấy loại haplotype PI 416937 có thể truyền đạt một số mức độ kháng bệnh nấm. Địa điểm duy nhất không phát hiện được sự gia tăng đáng kể về năng suất cũng là môi trường duy nhất mà thuốc trừ nấm toàn thân được áp dụng, có thể làm giảm tác dụng có lợi của Yld1. Mối liên hệ của Yld1 với khả năng kháng nấm bệnh cần được xác minh trong các thí nghiệm tiếp theo.

Mặc dù Yld1 không cho thấy tác động thống kê đến năng suất trong các đánh giá năng suất của chúng tôi đối với một số quần thể RIL hai bố mẹ, nhưng thường rất khó phát hiện ảnh hưởng năng suất của quỹ đạo đơn lẻ theo cách này khi xem xét các yếu tố gây

nhiễu nói trên và các bằng chứng khác. Kết quả cho thấy Yld1 là một ứng cử viên khả dĩ cho sự xâm nhập của sự đa dạng có lợi, đặc biệt là vào các vùng có mức độ đa dạng thấp trong các giống đậu tương mới. Vaughn và Li [53] đã quan sát thấy cấu trúc quần thể mạnh mẽ ở các giống đậu tương mới chịu ảnh hưởng nặng nề của MG. Họ chỉ ra rằng các giống đậu tương Bắc Mỹ có xu hướng phân thành ba nhóm chính dựa trên sự tương đồng về mặt di truyền. Các nhóm này trải dài trong phạm vi trưởng thành MG0-I, MG III-IV và MG V +. Trong các nhóm khác nhau này, Vaughn và Li [53] sau đó xác định các tổ tiên sáng lập và đánh giá các vùng giảm đa dạng gây ra bởi sự giống nhau trong các tổ tiên sáng lập hoặc có thể là chọn lọc sớm. Kết quả phân tích phả hệ PI 416937 được so sánh với các kết quả đó để quan sát xem có vùng nào tương ứng với các vùng có sự đa dạng hạn chế về các giống đậu tương Bắc Mỹ mới hay không. Vùng duy nhất thể hiện sự chồng chéo là locus Yld1 chồng lên vùng có độ đa dạng thấp mà Vaughn và Li [53] đã xác định trên Chr 8. Vùng có độ đa dạng thấp được phát hiện trong quần thể MG 0-I và nằm giữa 41,517,102 và 42,095,417bp (Gmax2.0) (Hình 3). Đa số các tổ tiên sáng lập của quần thể này chỉ có một haplotype duy nhất và do đó có sự đa dạng hạn chế trước nhiều thập kỷ chọn giống. Khu vực này nhanh chóng mất đi sự đa dạng này trong giai đoạn chọn giống ban đầu. Từ những kết quả này, khu vực này có thể được hiểu là có sự đa dạng khá hạn chế giữa các dòng tổ tiên ở Bắc Mỹ và nhanh chóng mất đi sự đa dạng đó trong quần thể MG 0-I. Gizlice và cs. [2] cũng xác định rằng 80% cơ sở di truyền phía Bắc có thể được tính bằng 10 dòng tổ tiên theo phả hệ.



Hình 3. Sự đa dạng đơn bội của locus Yld1 trên nhiễm sắc thể số 8 giữa các dòng tổ tiên và các giống mới từ MG 0-I.

Dòng trên cùng cho biết loại haplotype PI 416937. Phần thứ hai hiển thị các kiểu haplotypes cho tất cả 52 dòng có nguồn gốc PI 416937 năng suất cao được sử dụng trong phân tích phả hệ của chúng tôi. Phần thứ ba hiển thị haplotypes cho các tổ tiên chính của MG 0-I theo Vaughn và Li [53]. Phần dưới cùng hiển thị các dạng haplotypes cho các giống công cộng và tư nhân mới được chọn tạo cho MG 0-I theo thập kỷ phát hành. Các khối màu đỏ là các alen giống với PI 416937 trong khi các khối màu xanh lá cây là các alen thay thế cho mỗi locus. Vùng Yld1 được đánh dấu

bằng đường viền màu đen. Hình ảnh hóa được thực hiện bằng cách sử dụng trực quan hóa kiểu gen đồ họa của Flapjack [70].

Chúng tôi đã sử dụng một bảng gồm 135 dòng lai đậu tương mới (MG 0-I), được phóng thích mỗi thập kỷ từ những năm 1940 đến những năm 2000, để hình dung sự đa dạng haplotype của khu vực này trong quá khứ thông qua các vật liệu chọn giống mới hơn. Loại haplotype chính phổ biến trong các dòng tổ tiên MG 0-I là haplotype chủ yếu được chia sẻ trong vật liệu chọn giống MG 0-I mỗi thập kỷ. PI 548572 "Harly" (những năm 1940), PI 548550 "Disoy" (những năm 1960), PI 548536 "Coles" (những năm 1970) và PI 548640 "AC Scorpio" (những năm 1980) là bốn dòng duy nhất trong số 135 có chứa PI 416937 haplotype cho khu vực này. Mặc dù các marker là phù hợp về kiểu gen với PI 416937, vẫn không thể xác định được vùng này có hoàn toàn giống nhau ở tất cả các locus SNP hay không trừ khi thực hiện giải trình tự trên tất cả các dòng. Bất chấp điều đó, vùng PI 416937 có mặt thưa thớt. Phát hiện này cho thấy có cơ hội cho sự xâm nhập có mục tiêu của đa dạng có lợi vào một vùng có mức đa dạng thấp đối với vật liệu chọn giống MG 0-I. Vùng này không chứa bất kỳ QTL trưởng thành nào được ghi nhận trong tài liệu, vì vậy không có khả năng vùng này đã trở nên cố định do cố định các gen liên quan đến sự trưởng thành. Có thể có các yếu tố môi trường khác đối với các nhà lai tạo MG 0-I đã dẫn đến sự cố định vùng này cho loại haplotype cụ thể này. Cũng có thể là trường hợp vùng MG 0-I xuất hiện đặc hữu của locus này đã thiếu sự cạnh tranh từ các dạng đơn bội tiềm năng cao cấp khác.

Vaughn và Li [53] không báo cáo rằng vùng gen nói trên là vùng có tính đa dạng thấp ở MG III-IV hoặc V +. Trong nghiên cứu này, khi điều tra sự đa dạng haplotype trong vùng này đối với MG III-IV, các dòng tổ tiên cũng như các giống mới dường như bị chi phối bởi cùng một haplotype phổ biến ở MG 0-I mặc dù không đến mức nghiêm trọng. (Hình S5). Loại haplotype ưu thế này cũng phổ biến giữa các dòng tổ tiên MG V+ và các giống mới, nhưng dường như có sự đa dạng về haplotype hơn ở các giống miền Nam nước Mỹ này (Hình S6). Loại đơn bội PI 416937 không có trong các dòng tổ tiên của MG III-IV và MG V+ nhưng được tìm thấy ở chín trong số 198 giống MG III-IV (Hình S5) và năm trong số 114 giống MG V+ (Hình S6). Mặc dù đa dạng haplotype hơn xuất hiện trong MG III-IV và MG V+ so với MG0-I, sự hiếm hoi của haplotype PI 416937 và sự thống trị dường như của các haplotype khác cho thấy tiềm năng cho các nhóm trưởng thành này được hưởng lợi từ sự xâm nhập của haplotype PI 416937.

Các vùng gen như Yld1 cho thấy bằng chứng về sự chọn lọc giữa các dòng có nguồn gốc PI 416937 năng suất cao, nhưng có thể các vùng có thể tăng tần số do sự biến dạng phân ly, vì vậy một phân tích phả hệ tương tự đã được thực hiện trong các quần thể RIL. Mỗi quần thể RIL chia sẻ từ 88 đến 97 phần trăm các alen PI 416937 được thí nghiệm trên các quần thể RIL khác, do đó, có sự sao chép trong đó các alen PI 416937 được thử nghiệm trên các quần thể. Không có sự chồng chéo giữa các vùng được lựa chọn trong các quần thể RIL (Hình 2 và Bảng S4). Do đó, các khu vực này dường như là đặc thù về quần thể và có thể có môi trường. McMullen và cs. [71] đã kiểm tra sự biến dạng phân ly trên diện rộng và trong 25 họ của quần thể lập bản đồ liên kết lồng nhau của ngô. Mặc dù sự biến dạng đáng kể đã được phát hiện, các vùng gen bị biến dạng thường là đặc trưng của quần thể và các alen được chọn lọc thuận lợi cũng có thể khác nhau tùy thuộc vào quần thể. Các yếu tố gây căng thẳng môi trường cũng có thể phát sinh trong quá trình thí nghiệm giai đoạn đầu, có thể ảnh hưởng đến các tín hiệu lựa chọn này. Các yếu tố gây căng thẳng môi trường tiềm ẩn có thể bao gồm từ các yếu tố gây căng thẳng sinh học như mầm bệnh hoặc sâu bệnh đến các yếu tố gây căng thẳng phi sinh học như hạn hán hoặc lũ lụt. Các

nhà lai tạo cũng có xu hướng lựa chọn các đặc điểm nông học thuận lợi như khả năng chịu đựng. Không có sự chồng chéo của các vùng được lựa chọn trong phân tích phả hệ RIL và kết quả lập bản đồ QTL từ các nghiên cứu liên quan đến PI 416937 để chỉ ra một nguồn có khả năng ảnh hưởng đến môi trường đối với các tín hiệu lựa chọn.

Cũng không có sự trùng lặp giữa phân tích phả hệ RIL và phân tích phả hệ PI 416937 (Hình 2 và Bảng S4). Chúng tôi đã kiểm tra liệu sự khác biệt giữa mật độ marker trong SoySNP50K và SoySNP6K Infinium BeadChips có thể giải thích sự thiếu chồng chéo trong kết quả hay không. Chúng tôi kết luận điều này là khó xảy ra vì 18 trong số 26 khu vực được lựa chọn trong phân tích phả hệ PI 416937 chứa các marker từ 6K BeadChip trong khu vực được lựa chọn. 8 khu vực còn lại có marker 6K BeadChip bên sườn không quá 42 Kb và trung bình là 19 Kb, cách xa. Một yếu tố hạn chế phần lớn giải thích sự thiếu trùng lặp với phân tích phả hệ PI 416937 là chỉ có bảy (5, 6, 19, 21, 22, 23, 24) trong số 26 khu vực được xác định trong phân tích phả hệ PI 416937 là tách biệt cho một PI 416937 haplotype trong quần thể RIL. Mặc dù chúng tôi không thể nắm bắt tất cả bộ gen PI 416937 trong số các bố mẹ của các quần thể RIL này, chúng tôi có thể kiểm tra sự chồng chéo tiềm năng để lấy mẫu các vùng này.

Ý nghĩa của việc lựa chọn đối với nhiều vùng trong số này đã vượt xa mức độ lựa chọn trong phân tích phả hệ PI 416937. Các khu vực được xác định trong phân tích phả hệ RIL có thể không chịu trách nhiệm về những thay đổi đáng kể trong năng suất hạt giống như chúng chịu trách nhiệm về thể chất hoặc các dấu hiệu trực quan về sức sống mà các nhà chọn giống quan sát được khi chọn dòng để đưa vào các thí nghiệm tăng năng. Mặc dù không phải tất cả các vùng có nguồn gốc PI 416937 có bằng chứng về sự chọn lọc đều tách biệt trong quần thể RIL, nhưng kết quả này làm tăng niềm tin của chúng tôi rằng các vùng được chọn lọc trong phân tích phả hệ PI 416937 không phải là kết quả của sự biến dạng phân ly, mà là sự lựa chọn trong quá trình đánh giá năng suất rộng rãi. Các khu vực từ PI 416937 giảm thể lực sẽ chỉ tồn tại trong một thế hệ duy nhất và sẽ bị loại bỏ trong bộ ba tổ tiên. Những vùng này sẽ có một lần kiểm tra trong phân tích phả hệ PI 416937 và do đó sẽ không bao giờ được phát hiện có ý nghĩa trong việc chọn lọc âm tính. Đối với các vùng từ PI 416937 dưới sự lựa chọn tích cực trong phân tích phả hệ RIL, những vùng này sẽ phát sinh trong một phép lai duy nhất nhưng khi PI 416937 được lai với các vật liệu khác, những vùng này sẽ không còn được chọn lọc mạnh mẽ trừ khi các alen thay thế ở mọi bố mẹ của mỗi phép lai giảm thể lực, điều này rất khó xảy ra. Phân tích phả hệ được thực hiện cho PI 416937 xác định các tín hiệu chọn lọc là kết quả của nỗ lực chọn giống từ sàng lọc giai đoạn đầu đến kiểm tra năng suất lặp lại trên các nền tảng di truyền và môi trường khác với phân tích phả hệ RIL, xác định các tín hiệu cụ thể về quần thể/ môi trường liên quan đến sự biến dạng phân ly hoặc sức sống giai đoạn đầu.

KẾT LUẬN

Các mối quan tâm đã được bày tỏ rằng áp lực chọn lọc hiện tại và các phép lai liên tục giữa các loài ưu tú sẽ dẫn đến một "cao nguyên chọn giống" cho lợi ích di truyền cũng như tính nhạy cảm với bệnh và áp lực côn trùng cao hơn [72].

Khi quá trình lai tạo đã thu hẹp sự đa dạng di truyền của các vật liệu ưu tú, ngày càng có áp lực để phát hiện ra các alen ngoại lai hoặc hoang dã có lợi và có thể được kết hợp một cách chọn lọc vào vật liệu chọn giống ưu tú đồng thời tránh các alen khác kém thuận lợi hơn về mặt nông học. Các nhà chọn giống do dự trong việc phá vỡ các khối liên kết thuận lợi [40] và hy sinh năng suất và sức sống để tăng tính đa dạng từ các mầm ngoại lai [73]. Các chiến lược được nêu ở đây kết hợp kiến thức phả hệ chi tiết và xác định kiểu gen mật

độ cao để cho phép phát hiện ra các alen có lợi trên các nền tảng và môi trường di truyền. Nghiên cứu này tập trung vào việc xác định các kiểu đơn bội có lợi từ PI giữa nền gen có thể kém thuận lợi hơn về mặt nông học và các alen này là ứng cử viên chính để thách thức các vùng có tính đa dạng thấp trong các giống đậu tương mới. Đặc biệt đối với các vùng gen thiếu sự đa dạng trong tổ tiên thành lập của giống đậu tương Bắc Mỹ, các alen mới có khả năng làm tăng năng suất. Ngay cả các vùng gen đã có lịch sử đa dạng cũng cần được thử thách với các alen mới trong nỗ lực đạt được lợi ích liên tục. Mặc dù không thể xác thực các vùng gen được phát hiện bằng các phương pháp theo ý của chúng tôi, nhưng phương pháp được hiển thị trong nghiên cứu này sẽ mở ra cánh cửa cho những cách thức mới để tiếp cận việc tìm kiếm và khai thác sự đa dạng có lợi nhằm vượt qua sự hạn chế trong tương lai về lợi ích di truyền cho các nhà chọn giống đậu tương liên kết với những giống ưu tú tiếp tục bằng cách lai với giống ưu tú để phát triển quần thể chọn giống. Cregan [74] đã dự đoán một tương lai trong đó hàng chục nghìn giống đậu tương sẽ được đặc trưng về kiểu gen và một câu hỏi trong tương lai mà các nhà di truyền học đậu tương phải đối mặt là làm thế nào để khai thác dữ liệu này để cải tiến đậu tương. Chúng tôi tin rằng chúng tôi đã đề xuất một chiến lược hiệu quả để sử dụng dữ liệu bộ gen để cải thiện đậu tương thông qua việc xác định sự đa dạng có lợi tiềm năng.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0235434#pone-0235434-t002

Supplementary information

S1 File. A collection of files depicting results from the PI 416937 pedigree analysis for each individual chromosome.

https://doi.org/10.1371/journal.pone.0235434.s001 (DOCX)

S2 File. Full results for the PI 416937 pedigree analysis.

https://doi.org/10.1371/journal.pone.0235434.s002 (XLSX)

S3 File. Full results for the RIL pedigree analysis.

https://doi.org/10.1371/journal.pone.0235434.s003 (XLSX)

S1 Fig. Pedigree tree of PI 416937-derived lines along with the ancestors traced back to earliest decipherable ancestors.

https://doi.org/10.1371/journal.pone.0235434.s004 (TIF)

S2 Fig. Cladogram showing population structure of high yielding PI 416937-derived lines relative to soybean ancestors and modern varieties colored by MG.

https://doi.org/10.1371/journal.pone.0235434.s005 (TIF)

S3 Fig. Cladogram showing population structure of high yielding PI 416937-derived lines relative to ancestors and modern varieties colored by descriptors.

https://doi.org/10.1371/journal.pone.0235434.s006 (TIF)

S4 Fig. Genomic regions from PI 416937 under positive (green) and negative (red) selection across the whole genome.

https://doi.org/10.1371/journal.pone.0235434.s007 (TIF)

S5 Fig. Haplotype diversity of the Yld1 locus on chromosome 8 among ancestral lines and modern varieties from MG III-IV.

https://doi.org/10.1371/journal.pone.0235434.s008 (TIF)

S6 Fig. Haplotype diversity of the Yld1 locus on chromosome 8 among ancestral lines and modern varieties from MG V+.

https://doi.org/10.1371/journal.pone.0235434.s009 (TIF)

S1 Table. Description of RIL populations.

https://doi.org/10.1371/journal.pone.0235434.s010 (DOCX)

S2 Table. List of genotypes used in cladograms.

https://doi.org/10.1371/journal.pone.0235434.s011 (DOCX)

S3 Table. List of genotypes included in S1 Fig.

https://doi.org/10.1371/journal.pone.0235434.s012 (DOCX)

S4 Table. Results of whole-genome based RIL pedigree analysis.

https://doi.org/10.1371/journal.pone.0235434.s013 (DOCX)