

Multiple QTL Mapping in Autopolyploids: A Random-Effect Model Approach with Application in a Hexaploid Sweetpotato Full-Sib Population

Guilherme da Silva Pereira,^{*,†,1,2} Dorcus C. Gemenet,^{*,1} Marcelo Mollinari,^{*,†} Bode A. Olukolu,[§]
Joshua C. Wood,^{**} Federico Diaz,^{††} Veronica Mosquera,^{††} Wolfgang J. Gruneberg,^{††} Awais Khan,^{**}
C. Robin Buell,^{**} G. Craig Yencho,[†] and Zhao-Bang Zeng^{*,†}

^{*}Bioinformatics Research Center and [†]Department of Horticultural Science, North Carolina State University, Raleigh, North Carolina 27695, [‡]International Potato Center, ILRI Campus, Nairobi, Kenya 25171-00603, [§]Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, Tennessee 37996, ^{**}Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, ^{††}International Potato Center, Peru, Lima 1558, and ^{†††}Plant Pathology and Plant-Microbe Biology Section, Cornell University, Geneva Campus, New York 14456

ORCID IDs: 0000-0002-7106-8630 (G.d.S.P.); 0000-0003-4901-1694 (D.C.G.); 0000-0002-7001-8498 (M.M.); 0000-0003-4143-8909 (B.A.O.); 0000-0001-6583-0628 (G.C.Y.); 0000-0002-3115-1149 (Z.-B.Z.)

ABSTRACT In developing countries, the sweetpotato, *Ipomoea batatas* (L.) Lam. ($2n = 6x = 90$), is an important autopolyploid species, both socially and economically. However, quantitative trait loci (QTL) mapping has remained limited due to its genetic complexity. Current fixed-effect models can fit only a single QTL and are generally hard to interpret. Here, we report the use of a random-effect model approach to map multiple QTL based on score statistics in a sweetpotato biparental population ('Beauregard' × 'Tanzania') with 315 full-sibs. Phenotypic data were collected for eight yield component traits in six environments in Peru, and jointly adjusted means were obtained using mixed-effect models. An integrated linkage map consisting of 30,684 markers distributed along 15 linkage groups (LGs) was used to obtain the genotype conditional probabilities of putative QTL at every centiMorgan position. Multiple interval mapping was performed using our R package QTL_{POLY} and detected a total of 13 QTL, ranging from none to four QTL per trait, which explained up to 55% of the total variance. Some regions, such as those on LGs 3 and 15, were consistently detected among root number and yield traits, and provided a basis for candidate gene search. In addition, some QTL were found to affect commercial and noncommercial root traits distinctly. Further best linear unbiased predictions were decomposed into additive allele effects and were used to compute multiple QTL-based breeding values for selection. Together with quantitative genotyping and its appropriate usage in linkage analyses, this QTL mapping methodology will facilitate the use of genomic tools in sweetpotato breeding as well as in other autopolyploids.

KEYWORDS multiple interval mapping; polyploid QTL model; restricted maximum likelihood; variance components; yield components; heritability

GENETIC analyses in polyploid species pose extra challenges in comparison with diploid species, in spite of the

evolutionary benefits that duplication of whole sets of chromosomes may have brought (Comai 2005; Van de Peer *et al.* 2009). When it comes to molecular markers, a codominant, biallelic single nucleotide polymorphism (SNP) directly informs on the genotypes of a diploid locus, but the best it can do alone in a polyploid locus is to inform on its allele dosage. In diploid species, molecular markers are usually scored qualitatively, and there are several methodologies and tools for performing linkage (*e.g.*, Stam 1993; Margarido *et al.* 2007) and quantitative trait loci (QTL) analyses (*e.g.*, Broman *et al.* 2003; Da Costa E Silva *et al.* 2012a). In allopolyploid species, such as cotton (Wu *et al.* 2015) and wheat (Hulse-Kemp *et al.* 2015), where preferential pairing dictates meiotic chromosome behavior much

Copyright © 2020 da Silva Pereira *et al.*

doi: <https://doi.org/10.1534/genetics.120.303080>

Manuscript received December 30, 2019; accepted for publication April 26, 2020; published Early Online May 5, 2020.

Available freely online through the author-supported open access option.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at figshare: <https://doi.org/10.25386/genetics.12246134>.

¹These authors contributed equally to this work.

²Corresponding author: North Carolina State University, 1 Lampe Dr., 368 Ricks Hall, Raleigh, NC 27695-7566. E-mail: gdasilv@ncsu.edu

like diploids, existing approaches can be readily applied. However, despite many successful studies in diploids and allopolyploids, QTL mapping in autopolyploids remains difficult. In fact, unlike diploid mapping populations, which can have two to four segregating QTL genotypes (in case of inbred or outbred species, respectively), autopolyploid mapping populations can have a much wider range of possible genotypes per locus. For example, there are up to 36, 400, or 4900 possible genotypes from crosses between two tetra-, hexa-, or octoploid outbred parents, respectively.

Single-dose markers, segregating in a 1:1, 3:1, or 1:2:1 fashion, have limited information for building integrated genetic maps in autopolyploids, and generally result in either separate parental maps (e.g., Shirasawa *et al.* 2017) or limited map integration (e.g., Balsalobre *et al.* 2017). In order to make use of multiple-dose markers, the first step is to perform dosage or quantitative SNP calling. Although most methods were designed for tetraploid species (e.g., Voorrips *et al.* 2011; Schmitz Carley *et al.* 2017), additional studies have produced methods that can analyze data for higher ploidy levels (Serang *et al.* 2012; Gerard *et al.* 2018). For building integrated genetic maps in tetraploid species, one can use the well-established TETRAPLOIDSNPMAP (Hackett *et al.* 2016) as well as TETRAORIGIN (Zheng *et al.* 2016), which also considers multivalent pairing. For higher ploidy species, MAPPOLY (Mollinari and Garcia 2019) is a better option than POLYMAPR (Bourke *et al.* 2018), because the former has implemented a hidden Markov model (HMM) general enough to analyze higher ploidy levels, whereas the latter is limited to tetra- and hexaploid species, and lacks HMM implementation to robustly map all multiple-dose markers (see Mollinari *et al.* 2020). With an integrated map, one can calculate the genotype conditional probabilities of putative QTL, ideally using appropriate HMM (Hackett *et al.* 2016; Mollinari and Garcia 2019). Based on a polyploid model in Kempthorne (1955), a single-QTL model, hereinafter referred to as fixed-effect interval mapping (FEIM), was proposed for autotetraploids (Hackett *et al.* 2001, 2014), and later extended for autohexaploids (van Geest *et al.* 2017).

For an even ploidy level m , the FEIM model can be written as

$$y_i = \mu' + \sum_{j=2}^m \alpha_j X_{ij} + \sum_{j=m+2}^{2m} \alpha_j X_{ij} + \varepsilon_i \quad (1)$$

where y_i is the phenotypic value of individual i , μ' is the intercept, α_j is the additive effect of allele j , X_{ij} is the conditional probability of allele j in individual i , and ε_i is the residual error. The constraints $\alpha_1 = 0$ and $\alpha_{m+1} = 0$ are naturally imposed to satisfy the conditions $\sum_{j=1}^m X_j = m/2$ and $\sum_{j=m+1}^{2m} X_j = m/2$, so that μ' is a constant that is hard to interpret due to these constraints (Hackett *et al.* 2001, 2014). Note that $2m-2$ additive allele effects need to be estimated, i.e., tetra-, hexa-, or octoploid models will have 6, 10, or

14 main effects, respectively. In order to test whether the additive allele effects are different from zero (the null hypothesis), likelihood-ratio tests (LRT) are performed along positions on a genetic map. Commonly, the tests are presented as “logarithm of the odds” (LOD scores), where $LOD = LRT/[2 \times \ln(10)]$. In order to declare a QTL, empirical LOD thresholds are computed for each trait using permutations (Churchill and Doerge 1994). This approach has been used widely so far (e.g., Schumann *et al.* 2017; van Geest *et al.* 2017; Massa *et al.* 2018). However, limitations in fitting multiple-QTL models have been presented, due mostly to the possibility of over-parameterization or the lack of optimized algorithms for model selection (Mengist *et al.* 2018; Klaassen *et al.* 2019).

Variance component methods have been used for performing QTL mapping in related individuals of complex population structures or families in humans (Lippert *et al.* 2014), animals (Druet *et al.* 2008), and plants (Crepieux *et al.* 2005). In common, these approaches take into account the flexibility of mixed models in dealing with the correlated QTL genotype effects among individuals due to shared alleles identical-by-descent (IBD) by each relative pair at a particular location in the genome. Since a higher ploidy level leads to a much larger number of allele combinations, genotypic effects may be very hard to assess from the small population sizes usually available. In this case, the integrated genetic map provides key information on the inheritance of sets of chromosomal segments from parents to progeny (Mollinari and Garcia 2019), making up the basis for IBD-based additive relationship estimation. If a locus is linked to a region underlying the variation of a trait of interest, more shared alleles IBD for that locus are expected among individuals with similar phenotypic values (Almasy and Blangero 2010). Thus, the key parameters in this model are the variance components attributable to putative QTL, which determine the presence of linkage. Because only one parameter per QTL (the variance component) needs to be estimated, one could try to build a multiple-QTL model for polyploids, inspired by the corresponding multiple interval mapping (MIM) for diploid mapping populations (Kao *et al.* 1999), without the risk of model over-parameterization.

A multiple-QTL mapping approach is expected to increase detection power, enable separation of linked QTL, and provide the basis for studying QTL interaction (epistasis). Thus, such a model may benefit several autopolyploid horticultural (e.g., potato, blueberry, kiwifruit, strawberry), ornamental (e.g., rose, chrysanthemum), forage (e.g., alfalfa, guinea grass), and field (e.g., sugarcane) crops. The sweetpotato [*Ipomoea batatas* (L.) Lam. ($2n = 6x = 90$)] is a staple food in several developing countries, with a production of 112 million tons worldwide in 2017 (FAO 2019). Particularly, it has attracted growing interest due to its characteristics for food and nutrition security (Mwanga *et al.* 2017). In addition to

carbohydrates, dietary fiber, vitamins, and minerals, orange-fleshed sweetpotatoes provide high levels of β -carotene to fight vitamin A deficiency in vulnerable populations, such as those in sub-Saharan Africa (Low *et al.* 2017). In order to increase production and meet farmer's and market needs, it is imperative to make molecular-assisted selection an effective part of sweetpotato breeding programs. Toward this end, one of the first steps is to characterize the genetic architecture of traits of interest, such as those related to storage root yield and quality, and resistance to biotic and abiotic stresses (Khan *et al.* 2016). In spite of being considered an "orphan" crop, there have been recent advances in building genome references from its wild diploid relatives (Wu *et al.* 2018), optimizing a genotyping-by-sequencing protocol (GBSpoly) for high-throughput SNP genotyping (Wadl *et al.* 2018), and building a high-density integrated genetic map (Mollinari *et al.* 2020). In this paper, we introduce a random-effect multiple interval mapping (REMIM) model for autopolyploids. Using a genome-assisted, GBSpoly-based integrated genetic map from a sweetpotato biparental population, we map QTL for yield-related traits with our open-source software, QTL_{POLY}.

Materials and Methods

Full-sib population

A bi-parental mapping population (named BT) comprising 315 individuals was developed by crossing an orange-fleshed American variety, 'Beauregard' (CIP440132), and a nonorange-fleshed African landrace, 'Tanzania' (CIP440166), as male and female parents, respectively. The parents show contrasting phenotypes for several traits such as dry matter, β -carotene and sugar content, and susceptibility to biotic (*e.g.*, virus disease) and abiotic (*e.g.*, drought) stresses. 'Beauregard' is known as to have higher yield than 'Tanzania', and the current QTL mapping study will focus on yield components.

Phenotypic analyses

Field trials: In addition to the 315 full-sibs, parents (each replicated twice) and another variety, 'Daga' (CIP199062.1), were used as checks, making up a total of 320 individuals per replication in an 80×4 alpha-lattice design. Virus-free planting material derived from tissue culture was obtained from the CIP-Peru Genebank in La Molina. The clones were grown in a screen house in CIP substation San Ramon, and the planting material multiplied under low-disease pressure field conditions in Satipo, where cuttings for the six experiments were obtained. Four experiments were conducted in Ica ($14^{\circ}01' S$ and $75^{\circ}44' W$, 420 m), with two independent trials over two seasons, and one experiment each was conducted in San Ramon ($11^{\circ}07' S$ and $75^{\circ}21' W$, 828 m) and Pucallpa ($8^{\circ}23' S$ and $74^{\circ}31' W$, 154 m). The number of replications were two at Ica and three at San Ramon and

Pucallpa. In all trials, 1 m and 0.3 m of inter- and intra-row spacing was used, respectively. In the first season at Ica (from 25 February to 29 June 2016), the plot size was $6 m^2$ of 16 plants arranged in four rows (4 plants per row) with one empty row between plots. In the second season at Ica (from 15 November 2016 to 17 March 2017), the plot size was $4.8 m^2$ of 16 plants arranged in two rows (8 plants per row) with no empty row between plots. In San Ramon (from 14 May to 15 September 2016) and Pucallpa (from 1 July to 4 November 2016), the plot size was $9 m^2$ of 30 plants arranged in three rows (10 plants per row) with no empty row between plots.

Phenotypic data: Eight yield-related phenotypes (see File S1) were collected per plot at harvest, ~ 120 days after transplanting. For analysis purposes, foliage and root yield data were standardized by plot size (relative to the largest) and converted to tons per hectare ($t ha^{-1}$) to allow comparisons across trials. Number of roots was divided by the number of plants in the plot. The total number of storage roots per plant (TNR) and total root yield (RYTHA) considered all storage roots from the whole plot regardless of their individual weight. Number of commercial roots per plant (NOCR) and commercial root yield (CYTHA) considered only storage roots of marketable size (≥ 100 g for African market). Number of noncommercial roots per plant (NONC) and noncommercial root weight (NCYTHA) were obtained from the difference between total and commercial roots. Foliage yield (FYTHA) was measured by weighing all above-ground biomass per plot. Finally, commercial index (CI) was calculated as the ratio between CYTHA and total biomass (*i.e.*, the sum of RYTHA and FYTHA).

Multi-environment phenotypic model: We considered each one of the six field trials as an environment. Jointly adjusted means for each individual were obtained by using the following mixed-effect model

$$Y_{ijkl} = \mu + e_l + r_{k(l)} + b_{j(kl)} + g_i + ge_{il} + \varepsilon_{ijkl} \quad (2)$$

where Y_{ijkl} is the phenotypic observation of the i^{th} genotype in the j^{th} block within the k^{th} replicate at the l^{th} environment, μ is the overall mean, e_l is the random effect of the l^{th} environment ($l = 1, \dots, L; L = 6$) with $e_l \sim \mathcal{N}(0, \sigma_e^2)$, $r_{k(l)}$ is the random effect of the k^{th} replicate ($k = 1, \dots, K; K = 2$ or 3 depending on the environment) at the l^{th} environment with $r_{k(l)} \sim \mathcal{N}(0, \sigma_r^2)$, $b_{j(kl)}$ is the random effect of the j^{th} block ($j = 1, \dots, J; J = 80$) within the k^{th} replicate at the l^{th} environment with $b_{j(kl)} \sim \mathcal{N}(0, \sigma_b^2)$, g_i is the fixed effect of the i^{th} genotype ($i = 1, \dots, I; I = 318$), ge_{il} is the random effect of genotype-by-environment ($G \times E$) interaction with $ge_{il} \sim \mathcal{N}(0, \sigma_{ge}^2)$, and ε_{ijkl} is the random residual error with $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_l^2)$ (*i.e.*, environment specific variances). Variance components were estimated by restricted maximum likelihood (REML) using ASREML-R (v4.1; <https://www.vsni.co.uk/software/asreml-r>).

Mean-basis broad-sense heritabilities (H^2) were approximate as the ratio between genotypic and phenotypic variances as

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{L} + \frac{\sigma_e^2}{KL}}$$

where σ_g^2 is the variance component associated with the g_i term from Equation 2 when treated as a random effect, i.e., $g_i \sim \mathcal{N}(0, \sigma_g^2)$, σ_e^2 is the variance component associated with the residual error but with a common variance for all environments, i.e., $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_e^2)$, and $\bar{K} = 2.25$ is the harmonic mean of the number of replicates across environments.

Finally, Pearson's correlations (significance $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$) were computed among the individual adjusted means.

Genotypic analyses

GBSpoly and dosage calling: A modified GBS protocol called GBSpoly was carried out according to Wadl *et al.* (2018) and described in detail for the BT population by Mollinari *et al.* (2020). In brief, total DNA was extracted and double restricted using a CviAII-TseI enzyme combination for all full-sibs and parents (each parent replicated 10 times). Restriction fragments were ligated to adapters, size selected, and amplified. Adapters contained an 8-bp buffer sequence in addition to sample-specific variable length barcodes (6–9 bp). Each 64-plex library was sequenced using eight lanes of an Illumina HiSeq 2500 system in order to ensure optimal read depth for dosage calling. We trimmed the 8-bp buffer sequence from the reads using the FASTX-TOOLKIT (available at hannonlab.cshl.edu/fastx_toolkit/). A modified version of TASSEL-GBS pipeline (v4.3.8), called TASSEL4-POLY (Pereira *et al.* 2018, available at <https://github.com/gramarga/tassel4-poly>) was used to demultiplex and to count and store the actual read depth for all loci in variant call format (VCF) files. We used BOWTIE2 (Langmead and Salzberg 2012) to align 64-bp tags against the *I. trifida* and *I. triloba* genomes, two sweetpotato wild relative diploid species (Wu *et al.* 2018, available at <http://sweetpotato.plantbiology.msu.edu>). Finally, the software SUPERMASSA (Serang *et al.* 2012, available at <https://bitbucket.org/orserang/supermassa>) was used to perform multi-threading dosage call through a wrapper function named VCF2SM (Pereira *et al.* 2018, available at <https://github.com/gramarga/vcf2sm>).

Linkage mapping: A linkage map was constructed by Mollinari *et al.* (2020) using the R package MAPPOLY (Mollinari and Garcia 2019, available at <https://github.com/mmollina/mappoly>; see File S2). In brief, we computed two-point recombination fractions between all 38,701 nonredundant, high quality GBSpoly-based markers, and sorted the most likely linkage phase between each marker pair. Markers were then grouped into 15 linkage groups (LGs) by using the Unweighted Pair Group Method

with Arithmetic Mean (UPGMA) hierarchical clustering method. For each LG, large-scale ordering was obtained using multidimensional scaling as implemented in the R package MDSMAP (Preedy and Hackett 2016), and then small-scale ordering was refined based on the reference genomes (see details in Mollinari *et al.* 2020). Map distances, computed using Haldane's map function, were re-estimated based on the individual posterior probabilities from SUPERMASSA dosage calls. The final integrated, completely phased map was composed of 30,684 markers distributed along 15 LGs with a total length of 2708.3 centiMorgans (cM) and no major gaps between markers (11.3 markers every cM, on average). Multi-point genotype conditional probabilities of putative QTL were estimated for every individual given the final map using an HMM algorithm (Lander and Green 1987; Jiang and Zeng 1997) adapted for polyploids (Mollinari *et al.* 2020) as implemented in MAPPOLY. Since 17 full-sibs were filtered out along the map construction (Mollinari *et al.* 2020), only the remaining 298 individuals were ultimately used for QTL mapping.

QTL mapping analyses

Under random bivalent pairing, an autopolyploid individual of a species with an even ploidy level m can produce $\binom{m}{m/2}$, or “ m choose $m/2$ ”, different gametes with the same probability, and a cross between two individuals can generate $p = \binom{m}{m/2}^2$ different genotypes. As an example, consider two contrasting parents, A and B, of a hexaploid species (such as sweetpotato) and their respective genotypes for a QTL as *abcdef* and *ghijkl*, each one with potentially six different alleles. As each parent can produce 20 different gametes, the cross A \times B would generate $p = 20^2 = 400$ possible different genotypes. The model detailed next can be adapted easily to any polyploid species with an even ploidy level by simply changing p accordingly, e.g., $p = 6^2 = 36$ for autotetraploid and $p = 70^2 = 4,900$ for autooctoploid full-sib families.

REMIM model and hypothesis testing: Taking a full-sib population with n individuals derived from a cross between two hexaploid parents, A and B, the multiple-QTL mapping model is expressed by

$$\mathbf{y} = \mathbf{1}\mu + \sum_{q=1}^Q \mathbf{Z}_q \mathbf{u}_q + \boldsymbol{\varepsilon} \quad (3)$$

where \mathbf{y} is the $n \times 1$ vector of phenotypic values (in our case, the jointly adjusted means from the phenotypic analysis), μ is the fixed effect of population mean, \mathbf{u}_q is the $p \times 1$ random vector of additive genetic values of QTL q ($q = 1, \dots, Q$) with $\mathbf{u}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{\Pi}\sigma_q^2)$, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ random vector of residual error with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$. $\mathbf{1}$ and \mathbf{I} are an $n \times 1$ vector of 1's and an $n \times n$ identity matrix, respectively, \mathbf{Z}_q is the $n \times p$ incidence matrix of genotype conditional probabilities of QTL q , and $\mathbf{\Pi}$ is a $p \times p$ additive relationship matrix between the p possible

QTL genotypes. This matrix is fixed for a given ploidy level, and, for a hexaploid species, $\mathbf{\Pi}$ cells assume one out of seven different values ($\mathbf{\Pi}_{jj'} = \{0/6, 1/6, 2/6, 3/6, 4/6, 5/6, 6/6\}$ with $j, j' = \{1, \dots, p\}$), depending on how many alleles IBD are shared between two genotypes. For example, the genotype pair *abcghi-defjkl* shares 0 out of 6 alleles, hence $\mathbf{\Pi}_{jj'} = 0/6$, whereas *abcghi-abcghi* shares 6 out of 6 alleles, hence $\mathbf{\Pi}_{jj'} = 6/6$.

Assuming that the random-effect QTL are uncorrelated, each with expectation zero, the expectation of the vector of phenotypic values \mathbf{y} is

$$E(\mathbf{y}) = \mathbf{1}\mu$$

and its variance-covariance matrix is

$$\text{Var}(\mathbf{y}) = \mathbf{V} = \sum_{q=1}^Q \mathbf{G}_q \sigma_q^2 + \mathbf{I} \sigma^2$$

where $\mathbf{G}_q = \mathbf{Z}_q \mathbf{\Pi} \mathbf{Z}_q'$ is the $n \times n$ additive relationship matrix between the n individuals on the putative QTL q , and σ_q^2 and σ^2 are the respective variance components associated with QTL q and the residual error. In other words, each \mathbf{G}_q cell $\mathbf{G}_{ii'}^{(q)}$ ranges from 0 to 1, as the result of the sum over the products of QTL genotype probabilities, $\mathbf{Z}_{ij}^{(q)} = \Pr(\mathcal{G}_{ij}^{(q)} | \text{map})$ and $\mathbf{Z}_{i'j'}^{(q)} = \Pr(\mathcal{G}_{i'j'}^{(q)} | \text{map})$, from \mathbf{Z}_q of a sib-pair i and i' ($i, i' = \{1, \dots, n\}$), weighted by the proportion of shared alleles IBD $\mathbf{\Pi}_{jj'}$ from $\mathbf{\Pi}$ between each pair of possible genotypes j and j' ($j, j' = \{1, \dots, p\}$), *i.e.*

$$\begin{aligned} \mathbf{G}_{ii'}^{(q)} &= \sum_{jj'=1}^p \Pr(\mathcal{G}_{ij}^{(q)} | \text{map}) \Pr(\mathcal{G}_{i'j'}^{(q)} | \text{map}) \mathbf{\Pi}_{jj'} \\ &= \sum_{jj'=1}^p \mathbf{Z}_{ij}^{(q)} \mathbf{Z}_{i'j'}^{(q)} \mathbf{\Pi}_{jj'} \end{aligned}$$

If the respective genotypes j and j' of two sibs i and i' , $\mathcal{G}_{ij}^{(q)}$ and $\mathcal{G}_{i'j'}^{(q)}$, are observed with certainty, computing pairwise additive relationship is straightforward, *e.g.*, $\Pr(\mathcal{G}_{i1}^{(q)} = \text{abcghi} | \text{map}) = 1$ and $\Pr(\mathcal{G}_{i2}^{(q)} = \text{abcghj} | \text{map}) = 1$, hence $\mathbf{G}_{ii'}^{(q)} = 5/6$. However, if the probability of the genotype of i is split in two, *e.g.*, $\Pr(\mathcal{G}_{i1}^{(q)} = \text{abcghi} | \text{map}) = 1/2$ and $\Pr(\mathcal{G}_{i2}^{(q)} = \text{abcghj} | \text{map}) = 1/2$, then $\mathbf{G}_{ii'}^{(q)} = 1/2 \cdot 1 \cdot 5/6 + 1/2 \cdot 1 \cdot 6/6 = 5.5/6$. It is interesting to note that, if \mathbf{G}_q is averaged out for all map positions, $\mathbf{G}_{ii'} \approx \mathbf{1} \forall i = i'$ (an individual with itself), and $\mathbf{G}_{ii'} \approx 1/2 \forall i \neq i'$ (any sib pair), consisting of the realized additive relationship matrix.

Here, our interest is in testing

$$H_0 : \sigma_q^2 = 0 \text{ vs. } H_a : \sigma_q^2 > 0$$

i.e., whether QTL q contributes to the variation in \mathbf{y} or not, so that several tests have to be performed along the genome. As part of the algorithm described next, we test for the presence of multiple QTL in consecutive rounds. In practice, we compute and store a \mathbf{G}_q matrix for every putative QTL q ,

representing genomic positions at a certain step size (*e.g.*, every 1 cM). In this case, Equation 3 can be rewritten as

$$\mathbf{y} = \mathbf{1}\mu + \sum_{q=1}^Q \mathbf{g}_q + \varepsilon \quad (4)$$

where \mathbf{g}_q is an $n \times 1$ random vector of the individual breeding values on the QTL q with $\mathbf{g}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_q \sigma_q^2)$. Notice how Equation 3 and 4 are closely connected. At the QTL peak q , \mathbf{u}_q (Equation 3) are the best linear unbiased prediction (BLUP) values of the p QTL genotype effects, whereas \mathbf{g}_q (Equation 4) are the BLUP values for the n full-sibs, *i.e.*, the individual breeding values based on the additive allele effects. Therefore, the QTL-based breeding values $\hat{\mathbf{g}}_q$ are equivalent to those from $\mathbf{Z}_q \hat{\mathbf{u}}_q$. These alternatives forms are conveniently used in different contexts of QTL detection and characterization, as described next. While Equation 4 is preferred in order to make the multiple-QTL model selection less computational intensive because the stored \mathbf{G}_q can be used recursively, Equation 3 is used to describe the QTL genotype effects, which are ultimately used to compute additive allele effects.

We computed linear score statistics according to Qu *et al.* (2013) at every position and compared its P -value with a prescribed critical value, as part of the algorithm used to declare QTL as described in the next section. In order to compute the test statistics, we assume that \mathbf{y} has normal distribution, *i.e.*, $\mathbf{y} \sim \mathcal{N}(\mathbf{1}\mu, \mathbf{V}(\boldsymbol{\tau})\sigma^2)$, with the variance components of QTL q rescaled in relation to σ^2 as $\sigma_q^2 = \sigma^2 \tau_q$, so that

$$\mathbf{V}(\boldsymbol{\tau}) = \mathbf{G}_r \tau_r + \sum_{q \neq r} \mathbf{G}_q \tau_q + \mathbf{I}$$

where r is the putative QTL being tested in a model with $q = 1, \dots, r, \dots, Q$ QTL and $\boldsymbol{\tau} = (\tau_r, \tau_{(Q-1)})'$. Using this form, the hypothesis testing becomes $H_0 : \tau_r = 0$ vs. $H_a : \tau_r > 0$, and $\tau_{(Q-1)}$ components are regarded as nuisance parameters. The other nuisance parameters μ and σ^2 are removed when writing the log REML profiled function

$$\ell(\boldsymbol{\tau}) \propto -\frac{1}{2} \left\{ \tilde{\mathbf{y}}' \tilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \tilde{\mathbf{y}} + \log \left[\left| \tilde{\mathbf{V}}(\boldsymbol{\tau}) \right| \right] \right\}$$

from the restricted parameterization of $\tilde{\mathbf{y}} = \mathbf{A}\mathbf{y}$, in which \mathbf{A} is an $\tilde{n} \times n$ matrix ($\tilde{n} = n - 1$), such that $\mathbf{A}\mathbf{1} = \mathbf{0}$ and $\mathbf{A}\mathbf{A}' = \mathbf{I}_{\tilde{n}}$ (Qu *et al.* 2013). Accordingly, it follows that $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}(\boldsymbol{\tau})\sigma^2)$, with $\tilde{\mathbf{V}}(\boldsymbol{\tau}) = \tilde{\mathbf{G}}_r \tau_r + \sum_{q \neq r} \tilde{\mathbf{G}}_q \tau_q + \mathbf{I}_{\tilde{n}}$ and $\tilde{\mathbf{G}}_q = \mathbf{A}\mathbf{G}_q\mathbf{A}'$. Based on $\ell(\boldsymbol{\tau})$, the score function of τ_r is given by $S_r(\boldsymbol{\tau}) = \partial \ell / \partial \tau_r$, *i.e.*

$$S_r(\boldsymbol{\tau}) \equiv \frac{\tilde{n}}{2} \frac{\tilde{\mathbf{y}}' \tilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \tilde{\mathbf{G}}_r \tilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \tilde{\mathbf{y}}}{\tilde{\mathbf{y}}' \tilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \tilde{\mathbf{y}}} - \frac{1}{2} \text{tr} \left[\tilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \tilde{\mathbf{G}}_r \right]$$

(Qu *et al.* 2013).

On the one hand, when there is only one QTL ($r = Q = 1$), *i.e.*, $\hat{\tau}_0 = \hat{\tau}_r = 0$ under H_0 , $S_r(0)$ is the exact (nonasymptotic) test statistic. On the other hand, a moment-based approximation

to the null distribution is used when two or more QTL are present in the model ($Q > 1$), i.e., $\hat{\tau}_0 = (0, \hat{\tau}_{(Q-1)})'$ under H_0 (Qu *et al.* 2013). The validity of the moment-based approximation was assessed through simulations, as suggested by Qu *et al.* (2013). In any case, large score value indicates departure from H_0 . The P -values associated with the linear score test are continuous over the unit interval as a result of weighted sums of the scores from the profiled likelihood. Herein, we conveniently take the “logarithm of P ” as $LOP = -\log_{10}(P)$ for graphic representation and supporting interval calculation purposes. Support intervals are defined as the QTL peak neighboring region with LOP greater than, or equal to, $LOP-d$, where d is a constant that subtracts the highest LOP (thus from the QTL peak) in that region, as similarly proposed for the LOD score statistics (Lander and Green 1987).

QTL detection and characterization: In order to select QTL, we adapted the MIM methodology described by Kao *et al.* (1999) to a random-effect model framework as follows:

1. *Forward search* adds one QTL at a time to the model at the position with the highest score statistic if the P -value is smaller than a genome-wide significance threshold level (e.g., $\alpha=0.20$), and fits it into the model. Consecutive rounds of search for a new QTL are carried out conditioning to the one(s) in the model until no more QTL positions can reach the threshold. A window size (e.g., of 20 cM) is avoided on either side of QTL already in the model when searching for a new QTL;
2. *Model optimization* follows rounds of position refinement and backward elimination when no more QTL can be added in the forward search step. In turn, a QTL position is updated conditional to all the other QTL in the model, and its score statistic is re-evaluated at a more stringent significance threshold level (e.g., $\alpha=0.05$), when the QTL may be dropped. The final set of QTL is defined when all selected positions are significant, and, thus, no more positions change or QTL are dropped;
3. *Forward search* (now with a threshold value as stringent as the one used for backward elimination, e.g., $\alpha=0.05$) as well as model optimization procedures are repeated until no more QTL are added (via forward search) or dropped (via backward elimination). Finally, *QTL profiling* is performed with the remaining significant QTL after the last round of model optimization has been carried out. The score statistics and their associated P -values are computed for all genomic positions conditional to the final set of QTL.

Notice that, as part of the strategy for selecting QTL, we were less stringent during the first step of *forward search*, so that we were able to allow more positions to be tested again during *model optimization*. In fact, power for detecting significant positions is expected to increase when conditioning the forward search as well as the backward elimination to other QTL already in the model (Da Costa E Silva *et al.*

2012a). For the forward search performed after the first backward elimination, we used the last threshold set from the backward elimination in order to avoid false positives. Here, we adopted the score-based resampling method to assess the genome-wide significance level proposed by Zou *et al.* (2004). In brief, n independent samples from a standard normal distribution were used to obtain the score statistic for every map position under evaluation (e.g., every 1 cM) and the P -value of the highest score was stored. After repeating this $N = 1000$ times (resampling), the score-based threshold for a significance α level was then defined from the $100(1-\alpha)$ percentile of ascending ordered P -values from the N samples.

Once the QTL were selected, we were able to estimate their variance components and compute QTL heritabilities, h_q^2 , as the ratio between the QTL variance component and total variance. Given the parameter estimates, QTL-based breeding values are directly obtained as the BLUPs of the QTL genotypes (i.e., the vector \hat{u}_q) from Equation 3. BLUPs of the $p = 400$ possible genotypes were further decomposed in order to compute the additive effects as the average of \hat{u}_q (BLUPs) containing the respective alleles. Note that, in an F_1 population, the allele substitution effects (the very definition of additive effects) can be assessed only among alleles within each parent. Due to the model assumptions of zero mean for random effects, additive allele effects sum up to zero. These effects should be interpreted as the heritable contributions from parent to offspring, hence providing straightforward estimation of QTL-based breeding values to be used for selection.

Simulations

We conducted the following simulation study to examine the performance of REMIM (Equation 4) and compare it with FEIM (Equation 1). We simulated quantitative traits with three QTL each ($q = 1, \dots, Q; Q = 3$) positioned along the BT linkage map (Mollinari *et al.* 2020, $n = 298$; see File S2) according to three scenarios (1000 simulations each): (i) unlinked, where all QTL were positioned in different LGs each; (ii) random, where each QTL was positioned randomly, but no closer than 20 cM from each other in case of being assigned to the same LG; and (iii) linked, where at least two QTL were positioned in the same LG, but no closer than 50 cM from each other. The QTL heritabilities were simulated as $h_q^2 = \{0.3, 0.2, 0.1\}$ following their respective QTL genotype effect distributions as $g_q \sim \mathcal{N}(0, \mathbf{G}_q \sigma_q^2)$, where $\sigma_q^2 = \{0.75, 0.50, 0.25\}$. The environmental error was simulated from a standard normal distribution ($\sigma^2 = 1$), while the population mean was simulated as zero ($\mu = 0$).

One round of *forward search* followed by *model optimization* (steps 1 and 2 from the algorithm described above) was carried out for each simulated trait using different genome-wide significance forward ($\alpha = 0.20$) and backward ($\alpha = \{0.15, 0.10, 0.05, 0.01\}$) P -value thresholds based on the score-based resampling method (Zou *et al.* 2004). For comparison, we ran FEIM (Equation 1) with the same

simulated traits using different genome-wide significance LOD thresholds ($\alpha = \{0.20, 0.15, 0.10, 0.05, 0.01\}$) based on 1000 permutations (Churchill and Doerge 1994). We also stored the error vectors used to add noise to each simulated phenotype, and ran FEIM and REMIM models again, now using the respective error vectors as offset variables, which simply subtract the error from its respective phenotype. In this case, we expected that both FEIM and REMIM models would perform similarly, since all the noise had been controlled, so that the only variation left was thus due to the QTL. We used the same step size of 2 cM as well as the same window size of 20 cM for both models. $LOP-d$ (from REMIM) and $LOD-d$ (from FEIM) support intervals were calculated for three different d values ($d = \{1.0, 1.5, 2.0\}$).

Following the definitions and summary statistics from Da Costa E Silva *et al.* (2012b), all QTL kept after model optimization were considered “mapped.” A mapped QTL was considered “paired” if <20 cM apart from the simulated position, and a paired QTL was considered “matched” (true QTL) if its support interval included a simulated QTL. Finally, a mapped QTL was considered “mismatched” (false QTL) if it was not matched. We summarized detection power and empirical false discovery rate (FDR) for each support interval. Power was calculated as the ratio between the number of matched QTL over the total number of simulated QTL. FDR was estimated as the ratio between the number of mismatched QTL over the total number of mapped QTL. The absolute distance differences between simulated and mapped positions of paired QTL (precision) were averaged out. The proportion of matched QTL over the total number of paired QTL as an approximation of support intervals (coverage) was provided for each d value.

Software implementation

We implemented the algorithm for detection and characterization of multiple QTL based on the REMIM model in an R package called QTLPOLY (available at <https://github.com/guilherme-pereira/qtlpoly>). We integrated functions from the R package VARCOMP (v0.2-0; Qu *et al.* 2013) to compute the score statistics. The rounds of QTL search and model optimization use the variance components estimated in the previous round, so that the new estimates iterate faster. In addition, calculations for different genomic positions were paralleled in order to speed up the process by using the R base package PARALLEL (v3.5.2; R Core Team 2019). Final models were fitted using the R package SOMMER (v3.6; Covarrubias-Pazaran 2016), from which BLUPs were extracted and used for the computation of additive allele effects and QTL-based breeding values. Both VARCOMP and SOMMER packages use REML estimation to compute the variance components from the random-effect QTL model. Functions for plotting QTL profiles, effects and support intervals were based on GGLOT2 (v3.1.0; Wickham 2016). Additional functions for running FEIM model and multi-threaded permutations were included in QTLPOLY and were based on the `lm()` function from R base package STATS (v3.5.2; R Core Team 2019).

Gene expression profiling

A developmental time-course gene expression profiling data of ‘Beauregard’ (NCBI BioProject PRJNA491292) was reported previously (Wu *et al.* 2018), and a parallel time-series of development with ‘Tanzania’ (NCBI BioProject PRJNA549660) roots was recently analyzed, including identification of differentially expressed genes (Gemenet *et al.* 2020). In brief, ‘Beauregard’ and ‘Tanzania’ roots were harvested from four biological replicates at 10, 20, 30, 40, and 50 days after transplanting (DAT), and 30, 40, and 50 DAT roots were classified into fibrous and storage roots based on diameter as described by Wu *et al.* (2018). RNA-sequencing (RNA-seq) datasets from both genotypes were generated and processed as described previously (Lau *et al.* 2018), with the one exception that the ‘Tanzania’ 30 DAT storage root sample was subsampled for 30 million reads. Differentially expressed genes were determined as described in Gemenet *et al.* (2020) using DESEQ2 (v1.22.2; Love *et al.* 2014) with a log₂ fold-change (lfc) threshold of 2 and an adjusted P -value cutoff of 0.01, based on the fragments per kilobase exon model per million mapped reads (FPKM). To provide a comparison of expression abundances in the roots to leaves, ‘Beauregard’ and ‘Tanzania’ plants were grown as described in Lau *et al.* (2018) for control conditions and RNA-seq libraries from leaves processed as described above.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Raw RNA-seq reads are available at NCBI under BioProject numbers PRJNA491292 and PRJNA549660. Raw GBSpoly reads and VCF files are available via Mollinari *et al.* (2020). File S1 contains phenotypic data. File S2 contains genetic map information used in this study (also available via Mollinari *et al.* 2020). File S3 is the raw expression abundance matrix (also available via Gemenet *et al.* 2020). File S4 is the log₂ FPKM expression matrix (also available via Gemenet *et al.* 2020). File S5 contains the differentially expressed genes associated with this study. MAPPOLY software used for linkage mapping analyses is available at GitHub (<https://github.com/mmolina/mappoly>). QTLPOLY software used for QTL mapping analyses and simulations is available at GitHub (<https://github.com/guilherme-pereira/qtlpoly>). Supplemental material available at figshare: <https://doi.org/10.25386/genetics.12246134>.

Results

Trait heritabilities and correlations

Each one of the eight yield-related traits from six environments were analyzed using a multi-environment mixed-effect model, from which we were able to obtain jointly predicted means for each full-sib and variance component estimates (Table 1). Parents showed contrasting means for all traits, with ‘Beauregard’ presenting higher means for number of

Table 1 Phenotypic analysis summary of eight yield-related traits from ‘Beauregard’ × ‘Tanzania’ (BT) full-sib family

	NOCR	NONC	TNR	CYTHA	NCYTHA	RYTHA	FYTHA	CI
\bar{B}	2.881	1.934	4.834	12.567	2.208	15.318	15.521	0.472
\bar{T}	0.990	0.763	1.598	4.793	0.869	5.730	41.144	0.107
\bar{F}_1	2.840	1.971	4.795	17.739	2.247	19.980	22.994	0.420
min(F_1)	1.388	0.513	1.822	6.000	0.572	6.658	13.801	0.140
max(F_1)	4.494	4.184	7.947	34.226	4.817	37.106	36.880	0.605
σ_g^2	0.386	0.277	1.117	27.611	0.313	31.568	23.677	5.88×10^{-3}
σ_{ge}^2	0.272	0.213	0.571	17.028	0.310	18.538	34.451	2.30×10^{-3}
σ^2	0.686	0.559	1.462	32.271	1.082	35.098	50.836	5.23×10^{-3}
H^2 (%)	80.07	78.31	84.59	84.08	70.39	84.73	71.35	88.42

Parental (\bar{B} and \bar{T}) and progeny (\bar{F}_1) means, minimum, and maximum F_1 means, and genetic (σ_g^2), genotype-by-environment interaction (σ_{ge}^2) and residual (σ^2) variance components and heritability (H^2) estimates are shown for eight traits: number of commercial (NOCR), noncommercial (NONC) and total (TNR) roots per plant, commercial (CYTHA), noncommercial (NCYTHA) and total (RYTHA) root yield in t ha⁻¹, foliage yield (FYTHA) in t ha⁻¹, and commercial index (CI)

roots and root yield (both commercial and noncommercial) and commercial index when compared to ‘Tanzania’, which surpassed ‘Beauregard’ only for foliage yield. Interestingly, transgressive segregation was observed among the full-sibs for all traits, with emphasis on several individuals, with CYTHA higher than the most productive parent. Broad-sense heritabilities (H^2) were generally high, ranging from 70.39% (NCYTHA) to 88.42% (CI). Correlations between the predicted means were also estimated (Figure 1). Low to non-significant correlations (from -0.04 to 0.16^{**}) were observed between FYTHA and root yield traits. The highest correlation (0.99^{***}) was between CYTHA and RYTHA, which was expected, since most RYTHA is derived from CYTHA. Among the traits used for CI calculation, CYTHA also had the highest correlation with CI (0.78^{***}), likely because it is its main component. TNR components were also highly correlated with TNR, namely NOCR (0.89^{***}) and NONC (0.86^{***}). Finally, NOCR and NONC turned out to be highly correlated with CYTHA (0.81^{***}) and NCYTHA (0.86^{***}), respectively.

Mapping QTL in the BT population

Simulations: The BT linkage map based on 298 F_1 progenies was used to simulate 1000 quantitative traits with three QTL each. We ran FEIM (Equation 1) and REMIM (Equation 4) for each simulated trait in order to assess their power and empirical FDR in three increasingly difficult scenarios. We notice that the proportions of mapped QTL paired to the simulated positions with the highest heritability ($h_1^2 = 0.3$) were similar regardless of the method and criterion. However, higher proportions of simulated QTL with low heritability ($h_3^2 = 0.1$) were consistently mapped under the multiple-QTL mapping approach (see Supplemental Material, Table S1). In general, the average absolute difference between the simulated and mapped QTL peak location did not differ whatsoever when comparing models or thresholds. For REMIM, different genome-wide α level forward thresholds did not impact power or FDR (results not shown), but varying α level backward thresholds was critical. From testing different d values for $LOD-d$ and $LOP-d$, we learned that $d = 1.5$ was a good approximation of 95% support interval for both FEIM and

REMIM (see Table S2). Based on results for such a support interval, Figure 2 compares different threshold criteria for declaring a QTL during FEIM and REMIM (for forward $\alpha=0.20$ threshold). On the one hand, both FEIM and REMIM have shown a relative control of FDR, with $<15\%$ of false discoveries for most α levels, regardless of scenario. On the other hand, power differed in up to $\sim 18\%$ when comparing different α levels that delivered $<10\%$ FDR. Such a drop in detection power is more noticeable when FEIM is dealing with linked QTL, whereas power barely changes for REMIM across scenarios. Interestingly, even with the most stringent criteria of $\alpha=0.01$ backward threshold for REMIM, we were able to map as many QTL as using $\alpha=0.20$ for FEIM, but with better FDR control ($\sim 7\%$ for REMIM in comparison to $>15\%$ for FEIM). By fitting the model without the simulated residual error in the “unlinked” scenario, we observed $\sim 92\%$ power for FEIM. The small difference in comparison to $\sim 100\%$ power for REMIM is likely due to the fact that the QTL genotype effects were simulated based on the REMIM additive relationship model (Equation 4). Despite the relative bias, FEIM failure in separating linked QTL became more evident in the “random” and “linked” scenarios. As a consequence, detection power plateaus $\sim 87\%$ and 66% were observed, while REMIM exhibited $\sim 100\%$ power (Figure 2).

Yield-related traits: We adopted genome-wide α levels of 0.20 and 0.05 as the respective forward and backward thresholds for detecting QTL in eight yield-related traits in the BT population using REMIM (Equation 4), such that respective resampling-based P -value thresholds were defined as 5.83×10^{-4} and 1.42×10^{-4} . In total, 13 QTL were identified (Figure 3), with P -values ranging from 1.42×10^{-7} (QTL 2 for TNR) to 1.37×10^{-4} (QTL 2 for NOCR) (Table 2). The number of QTL per trait ranged from one (CYTHA, RYTHA and FYTHA) to four (NONC and TNR); NOCR had two QTL, and no QTL were found for NCYTHA and CI. Four LG harbored QTL regions: LGs 1, 3, and 10 harbored three QTL each, and LG 15 harbored four QTL. Approximate 95% support intervals computed as $LOP-1.5$ (see Figure S1) showed that QTL were colocalized mostly within each LG. QTL peaks for NOCR, NONC, and TNR can be found from 137.60 to

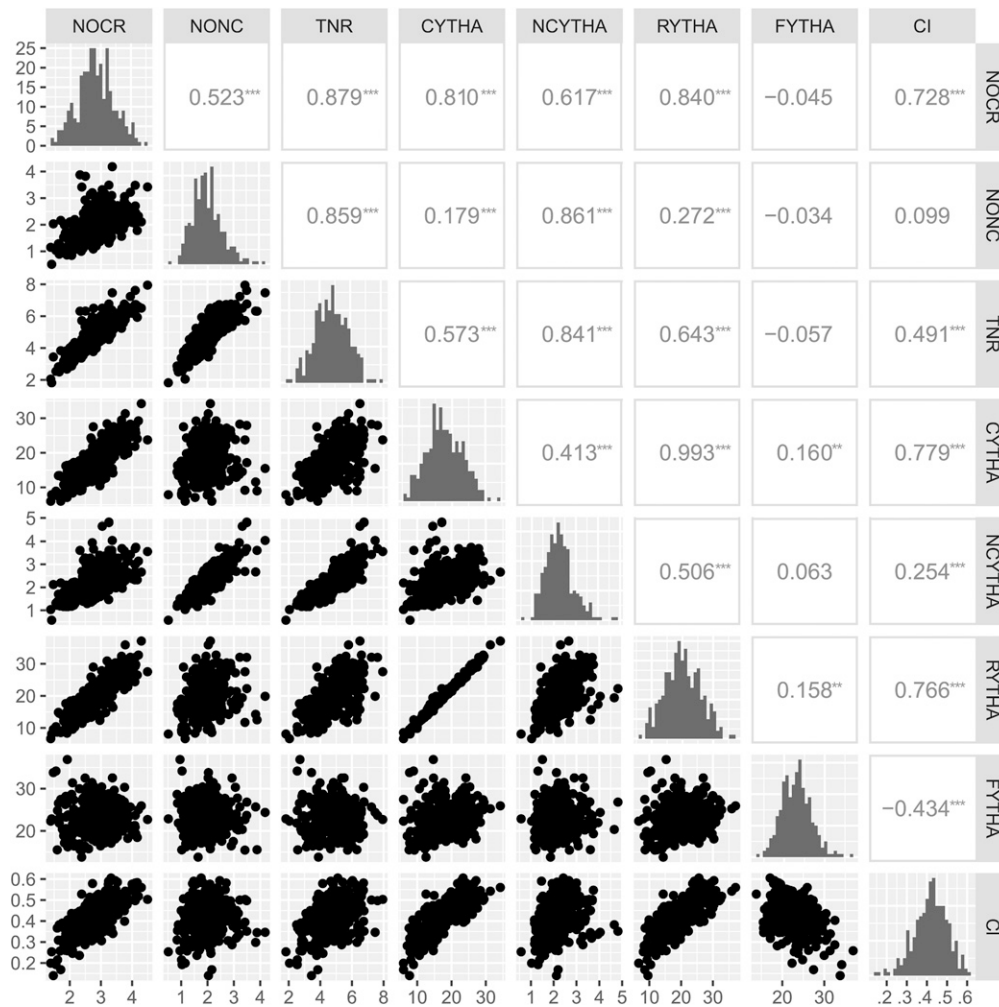


Figure 1 Pearson's correlations (** $P < 0.01$, *** $P < 0.001$) among predicted means of eight yield-related traits from 'Beauregard' × 'Tanzania' (BT) full-sib family. Trait abbreviations: number of commercial (NOCR), noncommercial (NONC) and total (TNR) roots per plant, commercial (CYTHA), noncommercial (NCYTHA), and total (RYTHA) root yield in $t\ ha^{-1}$, foliage yield (FYTHA) in $t\ ha^{-1}$, and commercial index (CI).

142.07 cM on LG 1, and from 13.11 to 20.18 on LG 3. On LG 15, QTL peaks were localized either at 67.20 and 78.04 cM for NONC and TNR, respectively, or at 5.34 cM for both CYTHA and RYTHA. QTL variance (σ_q^2) and heritability (h_q^2) estimates from Equation 3 are shown in Table 2, where the subscript q denotes the QTL number for a specific trait. QTL heritabilities ranged from 8.99 (QTL 3 for TNR) to 22.04% (QTL 2 for TNR), representing the proportion of the total variance explained by that QTL, conditional to all the other QTL in the model. Out of 13 QTL, 4 were considered major QTL ($h_q^2 > 15\%$), which happened as pairs of colocalized QTL at the beginning of LGs 3 for NOCR/TNR and 15 for RYTHA/CYTHA (see Figure S2). Altogether, multiple QTL explained as much as 35.63%, 49.19%, and 55.06% of the total variance for NOCR, NONC, and TNR, respectively. In order to compare these QTL detection results with FEIM, we adopted an $\alpha=0.05$, so that permutation-based LOD score thresholds ranged from 7.63 to 7.85, depending on the trait (see Figure S4). A total of 12 QTL were mapped (see Table S3): one for each CYTHA, RYTHA, and FYTHA; two for NOCR; three for TNR; and four for NONC. No QTL were found for NCYTHA and CI either. The same four LGs harbored QTL: LGs 1 and 10 had two QTL each, LG 3 had three and LG 15 had five QTL,

with the most significant QTL (adjusted $R^2 > 11$) found on LGs 1, 3, and 15. In comparison with REMIM, FEIM did not detect QTL for NONC or TNR on LGs 1 and 10, respectively. Instead, it allowed two QTL on LG 15 for NONC (one at 56.86 cM and another at 119.08 cM).

From REMIM, additive allele effects (see Table S4) were derived from the QTL genotype BLUPs (Equation 3). In general, although both parents have shown allele contributing to either decreasing or increasing the trait means, 'Beauregard' seemed to contribute more, increasing the number of roots, while 'Tanzania' exhibited major alleles increasing root yield. These effects represent the parental contributions to the population mean, *i.e.*, how much one adds to, or subtracts from, the mean, given 1 of the 400 possible genotypes. For instance, Figure 4 shows the additive allele effects of QTL 1 for CYTHA. Inferences on which alleles contribute more to the mean as well as which ought to be selected for breeding purposes are straightforward. For example, individuals with the haplotypes *b* from 'Beauregard' and *i* from 'Tanzania', and without the haplotypes *c* and *j* through *l* from the respective parents will have the highest QTL-based breeding value estimates for CYTHA. By computing QTL-based breeding values, we could hypothesize on the genetic basis of

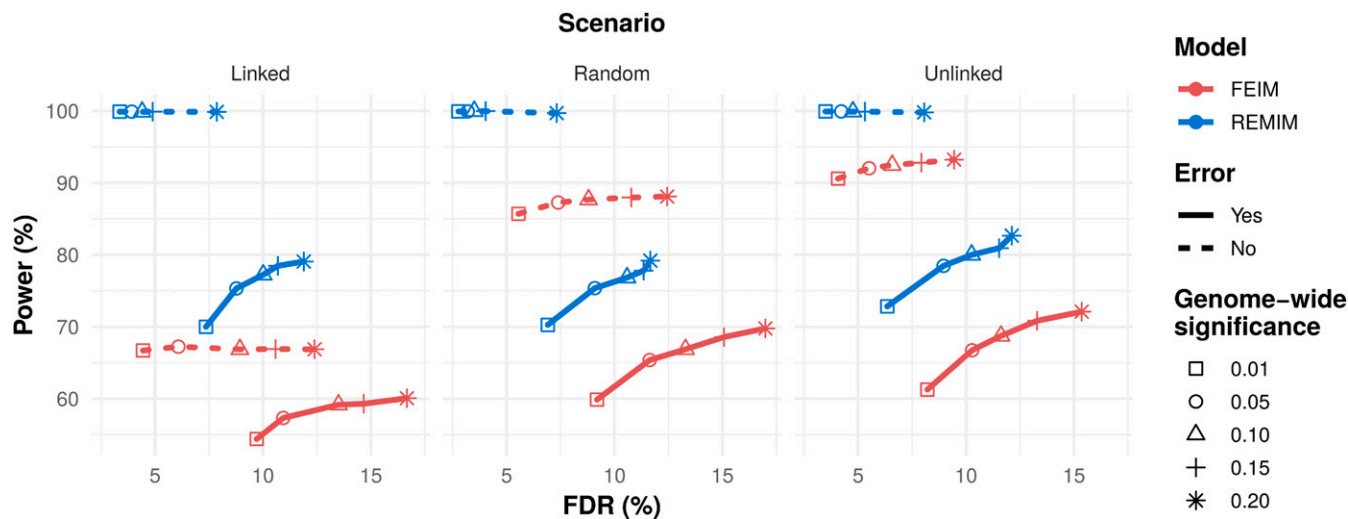


Figure 2 Detection power (in percentage) vs. empirical false discovery rate (FDR, in percentage) from QTL mapping analyses of simulated traits in 'Beauregard' × 'Tanzania' (BT) full-sib family. Each trait was simulated with three QTL ($q = \{1, 2, 3\}$) with different heritabilities ($h_q^2 = \{0.3, 0.2, 0.1\}$) positioned along the BT linkage map ($n = 298$). At least two out of three QTL were linked or not depending on three scenarios (linked, random, and unlinked), with 1000 simulations each scenario. Fixed-effect interval mapping (FEIM, red) and random-effect multiple interval mapping (REMIM, blue) were carried out with (solid lines) and without (dotted lines) the simulated error. FEIM and REMIM used different genome-wide significance thresholds ($\alpha = \{0.20, 0.15, 0.10, 0.05, 0.01\}$, symbols) based on permutation tests or resampling method, respectively. For a ~95% support interval coverage, power was computed as the proportion of true QTL over the total number of simulated QTL, and FDR as the proportion of false QTL over the total number of mapped QTL.

trait correlation (see Figure S3). For instance, NOCR and NONC were highly correlated (0.51^{***}) given a couple of colocalized QTL. While TNR is significantly correlated to NOCR/NONC ($>0.76^{***}$), smaller correlation was observed between TNR and CYTHA/RYTHA ($<0.15^*$). Interestingly, QTL-based breeding values for FYTHA did not seem to correlate to any other root-related trait. Finally, the absolute positive correlation of 1.00^{***} between predicted means from CYTHA and RYTHA (Figure 1) could be explained by a single colocalized QTL, since the correlation between QTL-based breeding values was also very high (0.99^{***}).

Candidate genes underlying major QTL

We elected to examine putative candidate genes under two QTL with the highest heritabilities: the QTL for TNR on LG 3 (colocalized for NOCR and NONC) and the QTL for CYTHA on LG 15 (colocalized for RYTHA). The QTL peak on LG 3 was positioned at 1,591,872 bp relative to *I. trifida* genome (see Table S5), and 75 genes were found within a ~500-kb window around this peak. Examination of functional annotation of these genes, coupled with expression profiles in leaves, as well as a time course of developing roots in both 'Beauregard' and 'Tanzania' (Gemenet *et al.* 2020) (see File S3, File S4), revealed three candidate genes of interest (see Figure S5). The first *I. trifida* gene, *itf03g02930*, encodes a homolog of SKU5, a glycosyl phosphatidylinositol modified protein in *Arabidopsis thaliana*; the second candidate gene, *itf03g03280*, encodes a protein with sequence similarity to annexin (ANN1 and ANN2); and the last candidate gene, *itf03g03460*, encodes a protein

similar to the WUSCHEL homeobox family protein (AtWOX13). In general, these genes were expressed at a low level in leaves, but highly in roots. Six additional genes were found to be differentially expressed in 'Beauregard' storage roots relative to fibrous roots across the time course, while only a single gene was differentially expressed in 'Tanzania' (see File S5).

On LG 15, a major QTL for CYTHA, with the peak at 477,772 bp, spanned positions from 21,822 to 1,915,814 bp (see Table S5) and over 300 genes. As this was too large a distance to manually curate candidate genes responsible for the trait, we restricted our manual review to 25 genes distal and proximal to the most significant marker. Within this region, two genes encoded functions that may be associated with storage root development and had expression profiles that supported a role in storage root development (see Figure S5, File S3, File S4). The hormone ethylene has diverse roles in cell proliferation and elongation, and the *itf15g01020* gene encodes a protein with similarity to the *A. thaliana* CONSTITUTIVE TRIPLE RESPONSE 1 gene (*CTR1*) with functions in the ethylene signaling pathway. This gene was expressed in leaves but expressed at twice the levels in developing roots. Storage roots are grown for their high starch content, and *itf15g01120* encodes a protein with similarity to starch branching enzyme 2.2. This gene was expressed in leaves and roots with the highest expression levels detected in storage, not fibrous nor developing roots. Analysis of differentially expressed genes in storage vs. fibrous roots of 'Beauregard' and 'Tanzania' within this QTL region revealed 39 unique differentially expressed genes in 'Beauregard' across the time course, and 12 unique differentially expressed genes in 'Tanzania' (see File S5).

Table 2 Random-effect multiple interval mapping (REMIM) of yield-related traits from 'Beauregard' × 'Tanzania' (BT) full-sib family

Trait	QTL	LG	Position (cM)	Score	P-value	σ_q^2	h_q^2 (%)
NOCR	1	1	137.60 (99.43–152.87)	222.89	9.35×10^{-6}	0.0622	13.70
	2	3	20.18 (0.00–49.27)	172.52	1.37×10^{-4}	0.0996	21.93
NONC	1	1	142.07 (128.08–159.30)	207.83	1.42×10^{-5}	0.0447	10.11
	2	3	13.11 (0.00–51.33)	165.75	1.07×10^{-4}	0.0420	9.50
	3	10	102.26 (96.50–113.55)	267.28	3.40×10^{-7}	0.0647	14.63
	4	15	67.20 (39.10–78.04)	247.92	1.33×10^{-6}	0.0661	14.95
TNR	1	1	140.43 (128.08–152.87)	251.13	7.36×10^{-7}	0.1789	10.97
	2	3	20.18 (13.11–43.69)	279.18	1.42×10^{-7}	0.3595	22.04
	3	10	165.43 (102.26–187.27)	192.09	2.69×10^{-5}	0.1467	8.99
	4	15	78.04 (35.50–119.08)	207.07	1.34×10^{-7}	0.2131	13.06
CYTHA	1	15	5.34 (0.00–34.27)	242.24	5.62×10^{-6}	6.3128	19.93
RYTHA	1	15	5.34 (0.00–35.50)	226.56	1.32×10^{-5}	6.9177	18.97
FYTHA	1	10	29.09 (16.12–134.37)	203.16	4.4×10^{-5}	2.1077	14.78

Linkage group (LG), map position (in centiMorgans) and its ~95% support interval (within parenthesis), score statistic and its corresponding P-value, variance (σ_q^2), and heritability (h_q^2 , in percentage) of mapped QTL using resampling-based genome-wide significance P-value threshold of 0.05 (backward elimination)

Trait abbreviations: number of commercial (NOCR), noncommercial (NONC) and total (TNR) roots per plant, commercial (CYTHA), noncommercial (NCYTHA) and total (RYTHA) root yield in t ha⁻¹, foliage yield (FYTHA) in t ha⁻¹, and commercial index (CI).

Discussion

Polyloid single- vs. multiple-QTL models

QTL mapping in autopolyploid species has been limited to a fixed-effect interval mapping (FEIM) model proposed for tetraploids (Hackett *et al.* 2001, 2014) and also expanded for hexaploids (van Geest *et al.* 2017). Consisting of a single-QTL model, $2m - 2$ main effects are fitted (m is the ploidy level), and this model is compared to a null model (with no QTL) using LRT, ultimately expressed as LOD scores. Permutation-based genome-wide significance LOD thresholds are then used to declare a QTL. Trying to fit additional QTL into FEIM model could rapidly lead it to over-parameterization, since each QTL requires as many as 6 (for tetraploids), 10 (for hexaploids), or 14 (for octoploids) main effects to be tested and estimated in such a fixed-effect multiple-QTL model. Furthermore, new rounds of permutation tests, based on a model with QTL, would need to be carried out in order to provide an updated LOD score threshold (Klaassen *et al.* 2019). In contrast, the random-effect multiple interval mapping (REMIM) model presented here is designed to fit multiple random-effect QTL by estimating only one single parameter (σ_q^2) per QTL. Score statistic tests are performed in order to assess whether a QTL variance component is zero or not, conditional to other QTL in the model. These tests provide an approach for comparing two nested models with the reduced model having a random effect excluded, similar to what restricted LRT (RLRT) would do. However, (R)LRT is more prone to numerical errors because the null hypothesis ($H_0 : \sigma_q^2 = 0$) falls on the boundary of the parameter space, whereas score-based methods can be robust to eventual misspecification of the distribution of random effects (Verbeke and Molenberghs 2003). A score-based resampling method (Zou *et al.* 2004) was used for setting genome-wide significance thresholds, which facilitates a forward–backward search to identify an optimal multiple QTL-model in a computationally tractable manner.

Here, we used the BT population genetic map to simulate quantitative traits based on multiple QTL with different heritabilities each, in order to compare FEIM and REMIM performances under three increasingly difficult scenarios (Figure 2). Both approaches would potentially detect similar number of QTL in case they were all unlinked. However, despite the small bias created by the way QTL were simulated (based on REMIM model), FEIM showed a relative loss of power. Multiple-QTL model approaches have proven to provide greater power and better FDR control than single-QTL models for both univariate (Zeng *et al.* 1999; Laurie *et al.* 2014) and multivariate (Da Costa E Silva *et al.* 2012b) models, due mostly to the differences in detecting QTL with smaller effects. In fact, this is rather expected as a multiple-QTL model has a smaller residual variance, which helps to detect additional QTL. Multiple-QTL models are also supposed to improve detection of more than one QTL on the same LG (Mayer 2005), as they are usually hard to separate from each other due to the large extension of linkage disequilibrium in mapping populations. For polyploids, a nonoptimized approach of using residuals from a fitted single-QTL model as phenotypic data to find a second linked QTL has been proposed (Mengist *et al.* 2018), as it requires additional manual steps. In contrast, the forward–backward search employed here has been shown to be optimized to detect linked QTL. The consistently superior results in comparison to FEIM pointed out that the linear score statistics behaved well as part of our algorithm, and the impact of using different α level thresholds for QTL detection was also assessed here. In QTL mapping analysis, it is important to have a reasonable balance between detection power and FDR, as we are interested in mapping as many true QTL as possible. When deciding on which α level to adopt, one should consider the goals of the study, *i.e.*, whether it is intended to use a few very reliable QTL for marker-assisted breeding, or to discover as many QTL-related putative genes as possible for further validation. Although one could use a more relaxed criteria in

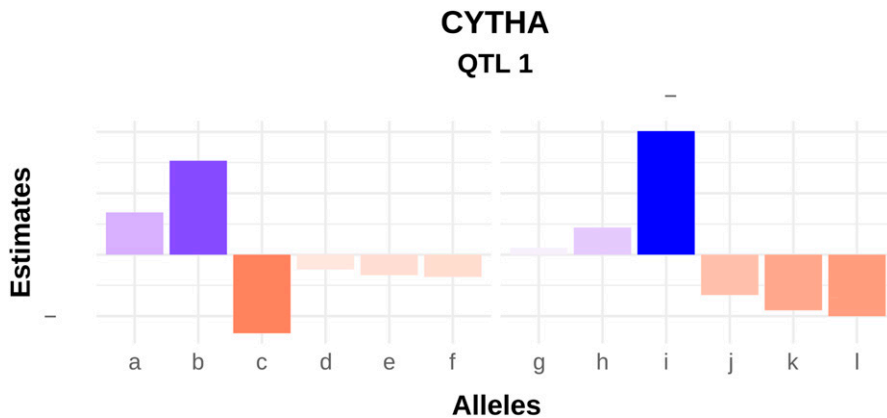


Figure 4 Additive allele effects from the decomposed best linear unbiased predictions (BLUPs) for the QTL 1 (on linkage group 15 at 5.27 cM) of commercial root yield in $t\ ha^{-1}$ (CYTHA) in a hexaploid sweetpotato full-sib family ('Beauregard' \times 'Tanzania'). Letters represent each of the six haplotypes from each parent.

order to increase the power of detection while still maintaining an acceptable level of FDR, REMIM with respective forward and backward α levels of 0.20 and 0.05 seemed reasonable.

QTL mapping for yield traits in sweetpotato

Most of the linkage and QTL mapping work done for sweetpotato so far has relied on strategies based on a double pseudotestcross approach for diploid species (Grattapaglia and Sederoff 1994). For example, separate parental maps have been built based on this diploid-based simplification, using qualitative marker systems such as randomly amplified polymorphic DNA (RAPD; Ukoskit and Thompson 1997), amplified fragment length polymorphism (AFLP; Kriegner *et al.* 2003; Cervantes-Flores *et al.* 2008a; Nakayama *et al.* 2012), retrotransposon insertion polymorphisms (Monden *et al.* 2015), and simple sequence repeats (SSR; Kim *et al.* 2017). A recent map was developed from a selfing population and used only single-dose SNPs, resulting in higher marker saturation in comparison to the previous maps (Shirasawa *et al.* 2017), though the map was still not integrated. In some of these cases, QTL mapping analyses were performed for several traits, related mostly to quality (Cervantes-Flores *et al.* 2011; Zhao *et al.* 2013; Yu *et al.* 2014; Kim *et al.* 2017) and resistance to biotic stresses (Cervantes-Flores *et al.* 2008b; Yada *et al.* 2017a). For yield-related traits, only two studies have been reported to date (Chang *et al.* 2009; Li *et al.* 2014). The use of DNA markers with unknown DNA sequence limited our ability to compare their results with *I. trifida* and *I. triloba* genomes (Wu *et al.* 2018), and, ultimately, with our present QTL study (see Table S5). Moreover, although these diploid-based strategies were the state-of-the-art at that time for qualitative marker-based, low density genetic maps, they imposed significant restrictions on statistical power for QTL detection and its genetic interpretation.

Recently, more improved methods and computational tools that take into account autopolyploid complexity for dosage SNP calling (Voorrips *et al.* 2011; Serang *et al.* 2012; Schmitz Carley *et al.* 2017; Gerard *et al.* 2018) and integrated linkage map construction (Hackett *et al.* 2016; Bourke *et al.* 2018; Mollinari and Garcia 2019) have become available, mostly

dedicated to tetraploids. Taking advantage of the newly developed MAPPOLY package, Mollinari *et al.* (2020) built the first integrated genetic map for sweetpotato, from the BT population used here. For a hexaploid species, this has opened up new opportunities for more interpretable QTL genetic models due to MAPPOLY implementation of a HMM that delivers QTL genotype conditional probabilities along a fully integrated genetic map (Mollinari and Garcia 2019). Based on this map, we detected 13 QTL (Figure 3) using REMIM, with QTL heritabilities ranging from 8.99 to 22.05% (Table 2, see Figure S2). Most of these QTL were also mapped among the 12 QTL using FEIM (see Figure S4), with proportion of variance explained (PVE) ranging from 8.42 to 12.43% (see Table S3). Based on the double pseudotestcross approach, previous studies found nine QTL for storage root yield ($17.7\% \leq PVE \leq 59.3\%$) (Li *et al.* 2014), seven QTL for root and top (foliage) weight ($16.0\% \leq PVE \leq 29.5\%$), plus one QTL for root number ($PVE = 14.8\%$) (Chang *et al.* 2009). Because of likely estimation bias due to reduced population sizes ($n < 200$), and the use of not very informative markers and linkage maps, these previous PVE findings are hard to compare with our results.

Although number of roots seemed to be as heritable as root yield (Table 1), only one colocalized QTL was detected for CYTHA/RYPHA. These traits are likely more complex in terms of their genetic architecture, though. That is, not only number of roots contributes to yield, but also size and composition, so we can expect that more regions are involved in root yield, in addition to those involved in number of roots. Nevertheless, Yada *et al.* (2017b) found a rather low trait heritability (likely individual-basis) for commercial root yield ($H^2 = 24\%$) among 278 full-sibs of a cross between 'New Kawogo', a Ugandan landrace, and 'Beauregard', possibly due to stronger $G \times E$ interaction, which adds to the trait complexity. Here, $G \times E$ interaction seemed important for all traits, and its consequences to QTL mapping and breeding will be explored in future studies. As QTL mapping targets major QTL, usually stable across environments, most of the minor ones must have gone undetected. Moreover, additional genetic variation could be due to higher-order allele interactions and genetic epistasis, which the current models do not account for.

Colocalized QTL among number of roots and yield traits explain some of the correlations among QTL-based breeding values (see Figure S3), partially explaining the correlations among the predicted means for these traits (Figure 1). Based on the correlation between QTL-based breeding values, FYTHA does not seem to be useful in indirect selection for CYTHA, as suggested previously (Chang *et al.* 2009).

'Beauregard' and 'Tanzania' contributed more importantly with positive and negative major effects, respectively. However, the presence of both favorable and unfavorable QTL alleles were observed in either parents (see Table S4), which possibly explains the presence of transgressive segregants for all traits. Transgression in polyploids seems to be due to cumulative complementary alleles not only at different loci (Tanksley 1993), but also from the same QTL. In fact, increased heterozygosity has been suggested as one of the major forces of polyploid evolutionary success, as a broader allele repertoire may result in the variation of gene expression and regulation needed to thrive in more diverse environmental conditions (Van de Peer *et al.* 2009). As an example, 'Tanzania' exhibited alleles contributing to increase CYTHA from a major QTL (Figure 4), although this landrace was not very productive in our environments overall. The additive effects are the most important from a breeding point-of-view, and their estimation provides straightforward direction on which alleles to select. Simpler biallelic-based models proposed previously (Hackett *et al.* 2014; Chen *et al.* 2018) may be used to estimate other interactions. The effective use of these allele interactions in QTL detection and breeding remains limited, though. As noted by Gallais (2003), estimating multi-allelic interactions reliably would require larger populations.

Several studies have looked at genes involved in storage root initiation and development in sweetpotato as reviewed by Khan *et al.* (2016). The storage roots differentiate from lateral roots by development of cambia around the protoxylem and secondary xylem, while lignification of the steles of some lateral roots inhibits this transformation (Villordon *et al.* 2012). Using the expression profile of the parents of the current mapping population, we found genes in leaves and roots (see Figure S5) related to root directional growth (*e.g.*, SKU5, *itf03g02930* homolog; Sedbrook *et al.* 2002) and lateral development (*e.g.*, AtWOX13, *itf03g03460* homolog; Deveaux *et al.* 2008) as well as with sugar transport to the root tip (*e.g.*, ANN1 and ANN2, *itf03g03280* homolog; Wang *et al.* 2018) within the QTL region on LG 3 associated with number of storage roots. Thus, both root restructuring and carbon supply is likely involved in the number of lateral root that transform to storage root. Other genes such as MADS-box (*e.g.*, *itf03g02230* homolog; Kim 2002), expansin (EXP, *itf03g05010* homolog), and BEL1-like homeodomain (*e.g.*, *itf03g02670* homolog; Ponniah *et al.* 2017) have been strongly implicated in storage root formation and development in sweetpotato, and were found within the QTL region on LG 3. On the QTL related to storage root weight on LG 15, we found genes related with the hormonal control of cell proliferation (*e.g.*, *CTR1*, *itf15g01020* homolog; Street *et al.*

2015) and with starch biosynthesis (*e.g.*, starch branching enzyme 2.2, *itf15g01120* homolog; Li *et al.* 2014). The association between these and other differentially expressed genes listed in this study (see File S5) is yet to be defined, and suggests the complex nature of storage root formation and development.

Final considerations

Here, we present a stepwise-based algorithm for multiple-QTL model selection in full-sib populations of autopolyploid species with a fully integrated map, from which QTL genotype conditional probabilities can be calculated. The use of score statistics is a key component of this new method, which depends on a dynamic and fast-computing test for model selection during the QTL detection process. Simulations were performed in order to assess the impact of using different threshold criteria and to provide some empirical sense on how to use the method in practice. REMIM has been carried out in a hexaploid sweetpotato population to detect major loci contributing to the variation of yield-related traits that may be targeted in molecular-assisted breeding. The use of random-effect models has created the context for fitting multiple QTL, providing straightforward information on variance components, important for computing QTL heritabilities. Finally, QTL genotype predictions (BLUPs) allowed us to estimate additive effects for characterizing major allele contributions, and compute QTL-based breeding values that can be used for performing selection. This novel approach may enable more complex models, such as those accounting for interaction among QTL as well as multiple traits or multiple environments in order to study shared genetic control in different traits/environments and $G \times E$ interaction at QTL level. Such a model can also be expanded in order to consider multi-parental designs and double reduction or preferential pairing as long as the reconstructed haplotypes can be used to inform on shared alleles IBD. For these more complex models, one could expect additional computational cost, for which further investigation is needed. Understanding the genetic architecture of root yield and other traits related to quality and resistance to biotic and abiotic stresses represents great opportunity for improving characteristics of interest in sweetpotato and other polyploids. Most of these important traits are polygenic in nature and only assessed later in a breeding program, where marker-assisted selection could help to speed up the process.

Acknowledgments

We would like to thank *ad hoc* reviewers and Associate Editor for their constructive suggestions on this manuscript. We acknowledge CIP sweetpotato breeding technical team in Peru for running experiments and collecting phenotypic data. This work was supported by the Bill & Melinda Gates Foundation [OPP1052983] as part of the Genomic Tools for Sweetpotato (GT4SP) Project. Research at CIP was undertaken as part of the CGIAR Research Program on Roots,

Tubers and Bananas (RTB), which is supported by CGIAR Fund Donors (<http://www.cgiar.org/about-us/our-funders/>).

Author contributions: D.C.G., W.J.G., A.K., G.C.Y. and Z.-B.Z. conceived and designed the experiments. B.A.O. performed DNA sequencing. D.C.G., F.D., V.M., W.J.G. and A.K. carried out field experiments. G.S.P., M.M. and Z.B.Z. developed tools and analyzed the data. C.R.B., J.C.W. and D.C.G. carried out candidate gene expression profiling. G.S.P. wrote the manuscript. All authors read and approved the manuscript.

Literature Cited

- Almasy, L., and J. Blangero, 2010 Variance component methods for analysis of complex phenotypes. *Cold Spring Harb. Protoc.* 2010: pdb.top77. <https://doi.org/10.1101/pdb.top77>
- Balsalobre, T. W. A., G. S. Pereira, G. R. A. Margarido, R. Gazaffi, F. Z. Barreto *et al.*, 2017 GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics* 18: 72. <https://doi.org/10.1186/s12864-016-3383-x>
- Bourke, P. M., G. Van Geest, R. E. Voorrips, J. Jansen, T. Kranenburg *et al.*, 2018 polypmapR - linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics* 34: 3496–3502 [corrigenda: *Bioinformatics* 35: 540 (2019)]. <https://doi.org/10.1093/bioinformatics/bty371>
- Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19: 889–890. <https://doi.org/10.1093/bioinformatics/btg112>
- Cervantes-Flores, J. C., G. C. Yencho, A. Kriegner, K. V. Pecota, M. A. Faulk *et al.*, 2008a Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose AFLP markers. *Mol. Breed.* 21: 511–532. <https://doi.org/10.1007/s11032-007-9150-6>
- Cervantes-Flores, J. C., G. C. Yencho, K. V. Pecota, B. Sosinski, and R. O. M. Mwangi, 2008b Detection of quantitative trait loci and inheritance of root-knot nematode resistance in sweetpotato. *Journal of American Horticultural Science* 133: 844–851. <https://doi.org/10.21273/JASHS.133.6.844>
- Cervantes-Flores, J. C., B. Sosinski, K. V. Pecota, R. O. Mwangi, G. L. Catignani *et al.*, 2011 Identification of quantitative trait loci for dry-matter, starch, and β -carotene content in sweetpotato. *Mol. Breed.* 28: 201–216. <https://doi.org/10.1007/s11032-010-9474-5>
- Chang, K. Y., H. F. Lo, Y. C. Lai, P. J. Yao, K. H. Lin *et al.*, 2009 Identification of quantitative trait loci associated with yield-related traits in sweet potato (*Ipomoea batatas*). *Bot. Stud. (Taipei, Taiwan)* 50: 43–55.
- Chen, J., F. Zhang, L. Wang, L. Leach, and Z. Luo, 2018 Orthogonal contrast based models for quantitative genetic analysis in autotetraploid species. *New Phytol.* 220: 332–346. <https://doi.org/10.1111/nph.15284>
- Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963–971.
- Comai, L., 2005 The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6: 836–846. <https://doi.org/10.1038/nrg1711>
- Covarrubias-Pazarán, G., 2016 Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11: e0156744. <https://doi.org/10.1371/journal.pone.0156744>
- Crepieux, S., C. Lebreton, P. Flament, and G. Charmet, 2005 Application of a new IBD-based QTL mapping method to common wheat breeding population: Analysis of kernel hardness and dough strength. *Theor. Appl. Genet.* 111: 1409–1419. <https://doi.org/10.1007/s00122-005-0073-5>
- Da Costa E Silva, L., S. Wang, and Z.-B. Zeng, 2012a Composite Interval Mapping and Multiple Interval Mapping: Procedures and Guidelines for Using Windows QTL Cartographer. In *Quantitative Trait Loci (QTL): Methods and Protocols*, edited by S. A. Rifkin, *Methods in Molecular Biology*, chapter 6, pp. 75–119, Humana Press, Totowa, NJ.
- Da Costa E Silva, L., S. Wang, and Z.-B. Zeng, 2012b Multiple trait multiple interval mapping of quantitative trait loci from inbred line crosses. *BMC Genet.* 13: 67. <https://doi.org/10.1186/1471-2156>
- Deveaux Y., C. Toffano-Nioche, G. Claisse, V. Thureau, H. Morin *et al.*, 2008 Genes of the most conserved WOX clade in plants affect root and flower development in Arabidopsis. *BMC Evol. Biol.* 8: 291. <https://doi.org/10.1186/1471-2148-8-291>
- Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume *et al.*, 2008 Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. *Genetics* 178: 2227–2235. <https://doi.org/10.1534/genetics.107.085035>
- FAO, 2019 FAOSTAT Crops. Available at <http://www.fao.org/faostat/>. Accessed on Feb-18–2019.
- Gallais, A., 2003 *Quantitative genetics and breeding methods in autopolyploids plants*, Ed. 1. INRA, Paris.
- Gemenet, D. C., G. da Silva Pereira, B. De Boeck, J. C. Wood, M. Mollinari *et al.*, 2020 Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β -carotene and starch content in hexaploid sweetpotato [*Ipomoea batatas* (L.) Lam.]. *Theor. Appl. Genet.* 133: 23–36. <https://doi.org/10.1007/s00122-019-03437-7>
- Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping polyploids from messy sequencing data. *Genetics* 210: 789–807. <https://doi.org/10.1534/genetics.118.301468>
- Grattapaglia, D., and R. Sederoff, 1994 Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudotestcross: mapping strategy and RAPD markers. *Genetics* 1137: 1121–1137.
- Hackett, C. A., J. E. Bradshaw, and J. W. McNicol, 2001 Interval mapping of quantitative trait loci in autotetraploid species. *Genetics* 159: 1819–1832.
- Hackett, C. A., J. E. Bradshaw, and G. J. Bryan, 2014 QTL mapping in autotetraploids using SNP dosage information. *Theor. Appl. Genet.* 127: 1885–1904. <https://doi.org/10.1007/s00122-014-2347-2>
- Hackett, C. A., B. Boskamp, A. Vogogias, K. F. Freedy, I. Milne *et al.*, 2016 TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *J. Hered.* 108: 438–442. <https://doi.org/10.1093/jhered/esx022>
- Hulse-Kemp, A. M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu, *et al.*, 2015 Development of a 63K SNP Array for Cotton and High-Density Mapping of Intraspecific and Interspecific Populations of *Gossypium* spp. G3 (Bethesda) 5: 1187–1209. <https://doi.org/10.1534/g3.115.018416>
- Jiang, C., and Z.-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* 101: 47–58. <https://doi.org/10.1023/A:1018394410659>
- Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203–1216.
- Kempthorne, O., 1955 The correlation between relatives in a simple autotetraploid population. *Genetics* 40: 168–174.
- Khan, M. A., D. C. Gemenet, and A. Villordon, 2016 Root system architecture and abiotic stress tolerance: current knowledge in root and tuber crops. *Front. Plant Sci.* 7: 1584. <https://doi.org/10.3389/fpls.2016.01584>

- Kim, S.-H., 2002 Isolation of MADS-box genes from sweet potato (*Ipomoea batatas* (L.) Lam.) expressed specifically in vegetative tissues. *Plant Cell Physiol.* 43: 314–322. <https://doi.org/10.1093/pcp/pcf043>
- Kim, J.-H., I. K. Chung, and K.-M. Kim, 2017 Construction of a genetic map using EST-SSR markers and QTL analysis of major agronomic characters in hexaploid sweet potato (*Ipomoea batatas* (L.) Lam.). *PLoS One* 12: e0185073. <https://doi.org/10.1371/journal.pone.0185073>
- Klaassen, M. T., P. M. Bourke, C. Maliepaard, and L. M. Trindade, 2019 Multi-allelic QTL analysis of protein content in a biparental population of cultivated tetraploid potato. *Euphytica* 215: 14. <https://doi.org/10.1007/s10681-018-2331-z>
- Kriegner, A., J. C. Cervantes, K. Burg, R. O. M. Mwanga, and D. Zhang, 2003 A genetic linkage map of sweetpotato [*Ipomoea batatas* (L.) Lam.] based on AFLP markers. *Mol. Breed.* 11: 169–185. <https://doi.org/10.1023/A:1022870917230>
- Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* 84: 2363–2367. <https://doi.org/10.1073/pnas.84.8.2363>
- Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9: 357–359. <https://doi.org/10.1038/nmeth.1923>
- Lau, K. H., M. del Rosario Herrera, E. Crisovan, S. Wu, Z. Fei *et al.*, 2018 Transcriptomic analysis of sweet potato under dehydration stress identifies candidate genes for drought tolerance. *Plant Direct* 2: e00092. <https://doi.org/10.1002/pld3.92>
- Laurie, C., S. Wang, L. A. Carlini-Garcia, and Z.-B. B. Zeng, 2014 Mapping epistatic quantitative trait loci. *BMC Genet.* 15: 112. <https://doi.org/10.1186/s12863-014-0112-9>
- Li, H., N. Zhao, X. Yu, Y. Liu, H. Zhai *et al.*, 2014 Identification of QTLs for storage root yield in sweetpotato. *Sci. Hortic. (Amsterdam)* 170: 182–188. <https://doi.org/10.1016/j.scienta.2014.03.017>
- Lippert, C., J. Xiang, D. Horta, C. Widmer, C. Kadie *et al.*, 2014 Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* 30: 3206–3214. <https://doi.org/10.1093/bioinformatics/btu504>
- Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15: 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Low, J. W., R. O. Mwanga, M. Andrade, E. Carey, and A. M. Ball, 2017 Tackling vitamin A deficiency with biofortified sweetpotato in sub-Saharan Africa. *Glob. Food Secur.* 14: 23–30. <https://doi.org/10.1016/j.gfs.2017.01.004>
- Margarido, G. R. A., A. P. Souza, and A. A. F. Garcia, 2007 OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144: 78–79. <https://doi.org/10.1111/j.2007.0018-0661.02000.x>
- Massa, A. N., N. C. Manrique-Carpintero, J. Coombs, K. G. Haynes, P. C. Bethke *et al.*, 2018 Linkage analysis and QTL mapping in a tetraploid russet mapping population of potato. *BMC Genet.* 19: 87. <https://doi.org/10.1186/s12863-018-0672-1>
- Mayer, M., 2005 A comparison of regression interval mapping and multiple interval mapping for linked QTL. *Heredity* 94: 599–605. <https://doi.org/10.1038/sj.hdy.6800667>
- Mengist, M. F., S. Alves, D. Griffin, J. Creedon, M. J. McLaughlin *et al.*, 2018 Genetic mapping of quantitative trait loci for tuber-cadmium and zinc concentration in potato reveals associations with maturity and both overlapping and independent components of genetic control. *Theor. Appl. Genet.* 131: 929–945. <https://doi.org/10.1007/s00122-017-3048-4>
- Mollinari, M. and A. A. F. Garcia, 2019 Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden markov models. *G3 (Bethesda)* 9: 3297–3314. <https://doi.org/10.1534/g3.119.400378>
- Mollinari, M., B. A. Olukolu, G. S. Pereira, A. Khan, D. Gemenet, *et al.*, 2020 Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. *G3 (Bethesda)* 10: 281–292. <https://doi.org/10.1534/g3.119.400620>
- Monden, Y., T. Hara, Y. Okada, O. Jahana, A. Kobayashi *et al.*, 2015 Construction of a linkage map based on retrotransposon insertion polymorphisms in sweetpotato via high-throughput sequencing. *Breed. Sci.* 65: 145–153. <https://doi.org/10.1270/jsbbs.65.145>
- Mwanga, R. O., M. I. Andrade, E. E. Carey, J. W. Low, G. C. Yencho *et al.*, 2017 Sweetpotato (*Ipomoea batatas* L.), pp. 181–218 in *Genetic Improvement of Tropical Crops*, chap 6., edited by H., Campos, and P. D. Caligari. Springer Nature, Berlin, Germany. https://doi.org/10.1007/978-3-319-59819-2_6
- Nakayama, H., M. Tanaka, Y. Takahata, K. Matsui, H. Iwahori *et al.*, 2012 Development of AFLP-derived SCAR markers associated with resistance to two races of southern root-knot nematode in sweetpotato. *Euphytica* 188: 175–185. <https://doi.org/10.1007/s10681-012-0678-0>
- Pereira, G. S., A. A. F. Garcia, and G. R. A. Margarido, 2018 A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinformatics* 19: 398. <https://doi.org/10.1186/s12859-018-2433-6>
- Ponniiah, S. K., J. Thimmapuram, K. Bhide, V. K. Kalavacharla, and M. Manoharan, 2017 Comparative analysis of the root transcriptomes of cultivated sweetpotato (*Ipomoea batatas* [L.] Lam) and its wild ancestor (*Ipomoea trifida* [Kunth] G. Don). *BMC Plant Biol.* 17: 9. <https://doi.org/10.1186/s12870-016-0950-x>
- Preedy, K. F., and C. A. Hackett, 2016 A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theor. Appl. Genet.* 129: 2117–2132. <https://doi.org/10.1007/s00122-016-2761-8>
- Qu, L., T. Guennel, and S. L. Marshall, 2013 Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics* 69: 883–892. <https://doi.org/10.1111/biom.12095>
- R Core Team, 2019 R: A Language and Environment for Statistical Computing. R Foundation for statistical computing, Vienna, Austria.
- Schmitz Carley, C. A., J. J. Coombs, D. S. Douches, P. C. Bethke, J. P. Palta *et al.*, 2017 Automated tetraploid genotype calling by hierarchical clustering. *Theor. Appl. Genet.* 130: 717–726. <https://doi.org/10.1007/s00122-016-2845-5>
- Schumann, M. J., Z. B. Zeng, M. E. Clough, and G. C. Yencho, 2017 Linkage map construction and QTL analysis for internal heat necrosis in autotetraploid potato. *Theor. Appl. Genet.* 130: 2045–2056. <https://doi.org/10.1007/s00122-017-2941-1>
- Sedbrook, J. C., K. L. Carroll, K. F. Hung, P. H. Masson, and C. R. Somerville, 2002 The Arabidopsis SKU5 gene encodes an extracellular glycosyl phosphatidylinositol-anchored glycoprotein involved in directional root growth. *Plant Cell* 14: 1635–1648. <https://doi.org/10.1105/tpc.002360>
- Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS One* 7: e30906. <https://doi.org/10.1371/journal.pone.0030906>
- Shirasawa, K., M. Tanaka, Y. Takahata, D. Ma, Q. Cao *et al.*, 2017 A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). *Sci. Rep.* 7: 44207. <https://doi.org/10.1038/srep44207>
- Stam, P., 1993 Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* 3: 739–744. <https://doi.org/10.1111/j.1365-3113.1993.00739.x>
- Street I. H., S. Aman, Y. Zubo, A. Ramzan, X. Wang *et al.*, 2015 Ethylene inhibits cell proliferation of the arabidopsis

- root meristem. *Plant Physiol.* 169: 338–350. <https://doi.org/10.1104/pp.15.00415>
- Tanksley, S. D., 1993 Mapping polygenes. *Annu. Rev. Genet.* 27: 205–233. <https://doi.org/10.1146/annurev.ge.27.120193.001225>
- Ukoskit, K., and P. G. Thompson, 1997 Autopolyploidy vs. Allopolyploidy and low-density randomly amplified polymorphic DNA linkage maps of sweetpotato. *J. Am. Soc. Hortic. Sci.* 122: 822–828. <https://doi.org/10.21273/JASHS.122.6.822>
- Van de Peer, Y., S. Maere, and A. Meyer, 2009 The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10: 725–732. <https://doi.org/10.1038/nrg2600>
- van Geest, G., P. M. Bourke, R. E. Voorrips, A. Marasek-Ciolakowska, Y. Liao *et al.*, 2017 An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theor. Appl. Genet.* 130: 2527–2541. <https://doi.org/10.1007/s00122-017-2974-5>
- Verbeke, G., and G. Molenberghs, 2003 The use of score tests for inference on variance components. *Biometrics* 59: 254–262. <https://doi.org/10.1111/1541-0420.00032>
- Villordon, A., D. LaBonte, J. Solis, and N. Firon, 2012 Characterization of lateral root development at the onset of storage root initiation in ‘Beauregard’ sweetpotato adventitious roots. *HortScience* 47: 961–968. <https://doi.org/10.21273/HORTSCI.47.7.961>
- Voorrips, R. E., G. Gort, and B. Vosman, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12: 172. <https://doi.org/10.1186/1471-2105-12-172>
- Wadl, P. A., B. A. Olukolu, S. E. Branham, R. L. Jarret, G. C. Yencho *et al.*, 2018 Genetic diversity and population structure of the USDA sweetpotato (*Ipomoea batatas*) germplasm collections using GBSpoly. *Front. Plant Sci.* 9: 1166. <https://doi.org/10.3389/fpls.2018.01166>
- Wang, J., J. Song, G. Clark, and S. J. Roux, 2018 ANN1 and ANN2 function in post-phloem sugar transport in root tips to affect primary root growth. *Plant Physiol.* 178: 390–401. <https://doi.org/10.1104/pp.18.00713>
- Wickham, H., 2016 *ggplot2: elegant graphics for data analysis*. Springer, Berlin, Germany.
- Wu, Q. H., Y. X. Chen, S. H. Zhou, L. Fu, J. J. Chen *et al.*, 2015 High-density genetic linkage map construction and QTL mapping of grain shape and size in the wheat population Yanda1817 x Beinaong6. *PLoS One* 10: e0118144. <https://doi.org/10.1371/journal.pone.0118144>
- Wu, S., K. H. Lau, Q. Cao, J. P. Hamilton, H. Sun *et al.*, 2018 Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nat. Commun.* 9: 4580. <https://doi.org/10.1038/s41467-018-06983-8>
- Yada, B., A. Alajo, G. N. Ssemakula, R. O. Mwanga, G. Brown-Guedira *et al.*, 2017a Selection of simple sequence repeat markers associated with inheritance of sweetpotato virus disease resistance in sweetpotato. *Crop Sci.* 57: 1421–1430. <https://doi.org/10.2135/cropsci2016.08.0695>
- Yada, B., G. Brown-Guedira, A. Alajo, G. N. Ssemakula, E. Owusu-Mensah *et al.*, 2017b Genetic analysis and association of simple sequence repeat markers with storage root yield, dry matter, starch and β -carotene content in sweetpotato. *Breed. Sci.* 67: 140–150. <https://doi.org/10.1270/jsbbs.16089>
- Yu, X. X., N. Zhao, H. Li, Q. Jie, H. Zhai *et al.*, 2014 Identification of QTLs for starch content in sweetpotato (*Ipomoea batatas* (L.) Lam.). *J. Integr. Agric.* 13: 310–315. [https://doi.org/10.1016/S2095-3119\(13\)60357-3](https://doi.org/10.1016/S2095-3119(13)60357-3)
- Zeng, Z.-B. B., C.-H. H. Kao, and C. J. Basten, 1999 Estimating the genetic architecture of quantitative traits. *Genet. Res.* 74: 279–289. <https://doi.org/10.1017/S0016672399004255>
- Zhao, N., X. Yu, Q. Jie, H. Li, H. Li *et al.*, 2013 A genetic linkage map based on AFLP and SSR markers and mapping of QTL for dry-matter content in sweetpotato. *Mol. Breed.* 32: 807–820. <https://doi.org/10.1007/s11032-013-9908-y>
- Zheng, C., R. E. Voorrips, J. Jansen, C. A. Hackett, J. Ho *et al.*, 2016 Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics* 203: 119–131. <https://doi.org/10.1534/genetics.115.185579>
- Zou, F., J. P. Fine, J. Hu, and D. Y. Lin, 2004 An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* 168: 2307–2316. <https://doi.org/10.1534/genetics.104.031427>

Communicating editor: J. Endelman