**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Desenvolvimento de modelos de causalidade com informações de QTLs para
estudo do relacionamento de caracteres fenotípicos relativos à absorção de
fósforo em milho**

Adriana Cheavegatti Gianotto

Tese apresentada para obtenção do título de Doutora
em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2015**

Adriana Cheavegatti Gianotto
Engenheira Agrônoma

**Desenvolvimento de modelos de causalidade com informações de QTLs para estudo do
relacionamento de caracteres fenotípicos relativos à absorção de fósforo em milho**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Tese apresentada para obtenção do título de Doutora
em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2015**

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Gianotto, Adriana Cheavegatti

Desenvolvimento de modelos de causalidade com informações de QTLs para estudo do relacionamento de caracteres fenotípicos relativos à absorção de fósforo em milho / Adriana Cheavegatti Gianotto. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2015.

94 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelo de equações estruturais 2. Modelos gráficos 3. Mapeamento de QTLs
4. Causalidade 5. Análise multivariada I. Título

CDD 633.15
G434d

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Aos meus pais, **Vergílio e Eutália**,
Pelo exemplo de uma vida honrada e feliz;
Aos meus amados **Valgerio e Davi**,
Por serem o sentido da minha vida,
dedico.

*Would that I could discover truth
as easily as I can uncover falsehood.*

Cícero (44 B.C.)

Ninguém pode servir a dois senhores;
pois ou há de aborrecer a um e amar ao outro, ou há de unir-se a um e desprezar ao outro.
Não podeis servir a Deus e às riquezas.

Por isso vos digo:

Não andeis cuidadosos da vossa vida pelo que haveis de comer ou beber,
nem do vosso corpo pelo que haveis de vestir;
não é a vida mais que o alimento, e o corpo mais que o vestido?

Olhai para as aves do céu,

que não semeiam, nem ceifam, nem ajuntam em celeiros, e vosso Pai celestial as alimenta;
não valeis vós muito mais do que elas?

Qual de vós, por mais ansioso que esteja, pode acrescentar um cúbito à sua estatura?

Por que andais ansiosos pelo que haveis de vestir?

Considerai como crescem os lírios do campo:

eles não trabalham nem fiam, contudo vos digo que nem Salomão em toda a sua glória se
vestiu como um deles.

Se Deus, pois, assim veste a erva do campo, que hoje existe, e amanhã é lançada no forno,
quanto mais a vós, homens de pouca fé?

Assim não andeis ansiosos, dizendo:

Que havemos de comer? ou:

Que havemos de beber? ou:

Com que nos havemos de vestir?

(Pois os gentios é que procuram todas estas coisas);

porque vosso Pai celestial sabe que precisais de todas elas.

Mas buscai primeiramente o seu reino e a sua justiça, e todas estas coisas vos serão acrescentadas.

Mateus 6:24-33

AGRADECIMENTOS

A Deus.

Ao CNPq, pela concessão da bolsa de Doutorado que me permitiu realizar este trabalho.

À Universidade de São Paulo, à Escola Superior de Agricultura “Luiz de Queiroz” e, especialmente, ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pela oportunidade de realizar este trabalho.

Ao meu orientador Augusto, a quem muito admiro, pela orientação e por ter me aceitado em seu seletto grupo.

À pesquisadora Claudia Teixeira Guimarães e ao meu amigo e estudante de doutorado Gabriel Corradi, da Embrapa Milho e Sorgo, pelo compartilhamento dos dados e pelas discussões extremamente produtivas.

A toda equipe da Embrapa Milho e Sorgo que colaboraram com geração dos dados deste trabalho, especialmente a: Bárbara Negri, Bárbara Hufnagel, Jurandir V Magalhaes, Uiraci Lana e Sylvia Morais de Sousa.

Aos amigos do Laboratório de Genética Estatística:

À “velha guarda”, Rodrigo, Maria Marta, Gabriel, Marcelo, Edjane, Renato, Graciela, Kateto e Luciano, por terem desbravado muitos assuntos realmente difíceis e terem a generosidade de compartilhá-los conosco; aos meus “contemporâneos”, Carina, João Ricardo, Guilherme, Rafael e Izabel, pela amizade sólida e pelo compartilhamento de conhecimento e experiências; aos “novatos”, Luiz Felipe, Fernando, Letícia, Marianella e Amanda, por trazerem uma nova energia ao laboratório e pela contribuição às nossas produtivas discussões; a “todos”, pelo ótimo ambiente de trabalho. Vocês são uma equipe realmente excepcional!

Especialmente, ao querido Luciano da Costa e Silva, por fornecer o pacote para mapeamento de QTLs, oneqtl, ainda em desenvolvimento e por sua paciência em responder todas minhas questões específicas sobre o uso do pacote e sobre mapeamento por múltiplos intervalos.

Especialmente a Guilherme, Carina e João Ricardo, pela amizade verdadeira, pelos momentos de convívio deliciosos e pelos debates científicos produtivos. Vocês sempre serão muito especiais para mim!

À minha família:

Ao meu marido Valgerio e ao meu filho Davi por aceitarem, sem nunca reclamar, esta “aventura” científica, com todas suas delícias e dificuldades; aos meus pais, que me criaram com o espírito livre e sempre apoiaram todas minhas decisões; aos meus irmãos e cunhados, Denise,

Rafael, Francisco (Robocop) e Mariana, simplesmente por serem a família que eu sempre quis ter; ao querido Joaquim que está chegando para nos dar ainda mais alegrias; aos meus sogros, Waldemar e Otília, e à minha tia Téo, que me ajudaram muito durante todo este processo.

Às minhas amigas, Tatiane, Elaine, Hellen, Kelly, Roberta e Márcia, pelo apoio e por tornarem este desafio mais fácil para mim.

Aos funcionários da genética Berdan, Sr. Antônio, Waldir, Léia e demais colegas de trabalho pelo auxílio durante o trabalho.

SUMÁRIO

RESUMO	11
ABSTRACT	13
1 INTRODUÇÃO	15
2 REVISÃO BIBLIOGRÁFICA	19
2.1 Causalidade	19
2.2 Randomização Mendeliana	20
2.3 Abordagens multivariadas para mapeamento de QTLs	21
2.4 Modelos de equações estruturais	23
2.4.1 Histórico	24
2.4.2 Terminologia	25
2.4.3 Estimação	31
2.4.4 Aplicações de MEEs na área biológica	32
2.5 Modelos gráficos	34
2.5.1 Terminologia	34
2.5.2 Algoritmos de descoberta de causalidade	36
2.5.3 Algoritmos de descoberta de causalidade auxiliados por QTLs	37
2.5.4 Aplicações dos modelos gráficos na área biológica	41
3 MATERIAL E MÉTODOS	43
3.1 Material	43
3.1.1 Dados fenotípicos	43
3.1.2 Dados genotípicos	44
3.2 Métodos	44
3.2.1 Mapa genético	44
3.2.2 Análises fenotípicas	45
3.2.3 Mapeamento de QTLs	46
3.2.4 Modelo de equações estruturais	48
3.2.5 Algoritmo PC de descoberta de causalidade	50
3.2.6 Algoritmo QDG de descoberta de causalidade com informação de QTL	52
3.2.7 Algoritmo QPSO de descoberta de causalidade com informação de QTL	53
4 RESULTADOS	55
4.1 Análises fenotípicas	55

4.2	Mapa genético	58
4.3	Mapeamento de QTLs	58
4.4	Modelo de equações estruturais	65
4.5	Modelos gráficos	67
4.5.1	Algoritmos de descoberta de causalidade (PC e Esqueleto de PC)	67
4.5.2	Algoritmos de descoberta de causalidade com auxílio de informações de QTLs (QDG e QSPO)	69
5	DISCUSSÃO	77
6	CONCLUSÃO	85
	REFERÊNCIAS	87

RESUMO

Desenvolvimento de modelos de causalidade com informações de QTLs para estudo do relacionamento de caracteres fenotípicos relativos à absorção de fósforo em milho

Metodologias de mapeamento de QTLs modernas empregam abordagem multivariada e se beneficiam da matriz de covariâncias fenotípicas para melhorar as estimativas de localização e efeitos de QTLs. No entanto, a correlação fenotípica pode ser em parte atribuída às relações de causalidade entre os fenótipos e mesmo as abordagens de mapeamento de QTLs multivariadas atuais têm desconsiderado tais relacionamentos. Dentre as metodologias científicas desenvolvidas para o estudo da causalidade em dados observacionais, destacam-se os modelos de equações estruturais e os modelos gráficos. Neste trabalho, foi estudado um conjunto de caracteres fenotípicos relacionados à morfologia de raízes, absorção de fósforo e acúmulo de biomassa em uma população composta de 145 linhagens endogâmicas recombinantes (RILs) do programa de melhoramento de milho da EMBRAPA Milho e Sorgo. O mapeamento de QTLs para os caracteres fenotípicos foi realizado utilizando mapeamento de múltiplos intervalos univariado (MIM) e multivariado (MT-MIM). A análise MIM revelou QTLs afetando diâmetro de raízes, área de superfície de raízes finas, peso seco da parte aérea e concentração de fósforo na parte aérea e nas raízes. A análise MT-MIM revelou 12 QTLs, com diferentes padrões de pleiotropia, com efeitos marginais para as sete variáveis analisadas. Um modelo de relacionamento causal entre os caracteres fenotípicos foi desenvolvido utilizando conhecimento prévio e modelagem de equações estruturais. O modelo de equações estruturais apresentou fluxo unidirecional de causalidade entre as variáveis, com as variáveis de morfologia de raízes exercendo efeito sobre as variáveis de acúmulo de biomassa, que por sua vez, têm efeito sobre as variáveis de absorção de fósforo. A aplicação do algoritmo PC para a descoberta de causalidade automatizada baseada nos padrões de independências condicionais não foi capaz de orientar todas as relações de causalidade descobertas, porém revelou um relacionamento mais complexo que o modelo de equações estruturais, com potenciais ciclos de retroalimentação causais. O emprego de algoritmos de descoberta de causalidade baseados em informações de QTLs, chamados QDG e QPSO, permitiu a orientação de todos os relacionamentos de causalidade encontrados pelo algoritmo PC e confirmou a existência de dois ciclos vizinhos de relacionamento causais entre as variáveis estudadas. Como regra geral, os QTLs pleiotrópicos detectados pela metodologia MT-MIM apresentaram efeitos sobre caracteres fenotípicos alinhados causalmente nos modelos propostos pelos algoritmos PC e QDG, sugerindo que alguns dos QTLs detectados são na realidade efeitos indiretos de QTLs situados em posição mais elevada no modelo causal. O emprego da abordagem MT-MIM aliada à análise de causalidade permitiu melhor compreensão da arquitetura genética dos caracteres de morfologia de raiz, acumulação de biomassa e aquisição de fósforo em milho.

Palavras-chave: Modelo de equações estruturais; Modelos gráficos; Mapeamento de QTLs; Causalidade; Análise multivariada

ABSTRACT

Development of causal models with QTL information to the study of relationship among traits associated with phosphorus uptake in maize

Modern QTL mapping approaches are multivariate and take advantage of the phenotypic covariance matrix to improve estimates of QTL positions and effects. However, phenotypic correlation can also be assigned to the causal relationship among phenotypes, and even modern multivariate QTL analysis does not take these relationships into account. Structural equation models and graphical models are the main methodologies to study causality from observational data. We studied a set of phenotypes related to root morphology, biomass accumulation and phosphorus acquisition in maize. These phenotypes were measured in a maize population from the EMBRAPA breeding program composed of 145 recombinant inbred lines (RILs) derived from the crossing of two divergent lines for phosphorus acquisition efficiency. QTL mapping for the traits was performed using univariate (MIM) and multivariate (MT-MIM) multiple interval mapping. MIM analysis revealed QTL affecting root diameter, fine root surface area, shoot dry weight and root dry weight. MT-MIM analysis revealed 12 QTL with different pleiotropy patterns and QTL with marginal effects affecting all seven studied characters. A causal model for phenotype characters was developed using *a priori* knowledge and structural equation model techniques. The structural equation model presented an unidirectional causal flow among the variables, with root morphological traits exerting causal effects over biomass traits, which in turn cause phosphorus acquisition traits. Using PC algorithm for an automatic search of causal models based on conditional independence was not able to orient all discovered causal relationships among traits but revealed a more intricate relationship than the structural equation model, with potential causal feedback loops among the traits. Employing causal search algorithms based on QTL information (named QDG and QPSO) allowed the orientation of all causal relationships detected by PC algorithm and it has also confirmed the presence of two neighbor causal cycles among the studied traits. As a general rule, pleiotropic QTL detected by MT-MIM approach exerted effects over traits according to the causal model discovered by PC and QDG algorithms, suggesting that some of the QTL detected effects were indirect effects of QTL located upstream at the proposed causal model. Employing MT-MIM approach and causal analysis has allowed a better comprehension of genetic architecture underlying root morphology, biomass accumulation and phosphorus acquisition traits in maize.

Keywords: Structural equation models; Graphical models; QTL mapping; Causality; Multivariate analysis

1 INTRODUÇÃO

Em experimentos de mapeamento de QTLs é comum a coleta de dados sobre caracteres fenotípicos correlacionados, com análise estatística dos mesmos sendo feita separadamente. Metodologias de mapeamento de QTLs mais modernas empregam a abordagem multivariada e se beneficiam da utilização da matriz de covariâncias fenotípicas para melhorar as estimativas de localização e efeitos de QTLs (JIANG; ZENG, 1995; KNOTT; HALEY, 2000; SILVA; ZENG, 2010; SILVA; WANG; ZENG, 2012). No entanto, em sistemas biológicos, os caracteres fenotípicos podem apresentar relacionamentos de causalidade entre si e os modelos multivariados baseados apenas em covariâncias não são capazes de levar tais relacionamentos em consideração (ROSA et al., 2011).

Ao contrário das covariâncias e correlações, que são medidas de associação simétricas entre variáveis, os relacionamentos de causalidade são direcionados, e o estudo dos mesmos permite prever como um sistema biológico irá se comportar em caso de perturbações externas. Esta aplicabilidade do estudo da causalidade é de grande interesse para programas de melhoramento (ROSA et al., 2011). Uma outra grande vantagem do estudo da causalidade em experimentos de mapeamento de QTLs é a possibilidade de uma melhor compreensão da arquitetura genética subjacente a um conjunto de caracteres fenotípicos correlacionados (ROSA et al., 2011). Por exemplo, a detecção de um QTL pleiotrópico com efeito sobre dois fenótipos distintos pode levantar as seguintes hipóteses: (i) o QTL tem efeito sobre as duas características fenotípicas simultaneamente; (ii) o QTL tem efeito direto sobre o primeiro fenótipo que por sua vez exerceria um relacionamento causal sobre o segundo fenótipo; (iii) a relação de causalidade é inversa, ou seja, o QTL afeta o segundo fenótipo que exerce uma relação causal sobre o primeiro (SCHADT et al., 2005). Li et al. (2006) demonstraram que, na realidade, o problema é ainda mais complexo do que as hipóteses levantadas acima e que com apenas um trio de variáveis, um QTL e dois fenótipos, é possível hipotetizar até dez modelos causais distintos. Neste cenário, análises de QTL multivariadas não auxiliam na distinção das diferentes hipóteses e apenas modelos que levem em consideração a causalidade entre os caracteres fenotípicos poderiam ajudar na compreensão da arquitetura genética de um QTL pleiotrópico.

Durante muito tempo, o estudo da causalidade em dados observacionais foi renegado pela metodologia científica e, ainda hoje, muita discussão persiste a respeito deste assunto (SHIPLEY, 2002). No entanto, o estudo de causalidade em dados observacionais é inevitável em cenários nos quais a experimentação é impedida devido a restrições de várias naturezas (éticas,

logísticas, orçamentárias) ou mesmo quando a experimentação é impossível porque as variáveis de interesse são mensuradas concomitantemente na mesma unidade experimental. Devido à necessidade e ao interesse no estudo da causalidade, metodologias científicas específicas para esta finalidade foram desenvolvidas e, ainda que com certas restrições, estão disponíveis atualmente. As principais metodologias existentes atualmente são os modelos de equações estruturais (JÖRESKOG, 1970; BOLLEN, 1989) e os modelos gráficos (PEARL, 2000; SPIRTES; GLYMOUR; SCHEINES, 2000).

Na área biológica, o estudo de causalidade é facilitado pela randomização Mendeliana (THOMAS; CONTI, 2004), que afirma que os fatores genéticos (QTLs ou marcadores, por exemplo) são aleatorizados durante a meiose, e pelo dogma central da biologia, que afirma que o genótipo “causa” o fenótipo. Assim, a disponibilidade de informações genotípicas em populações experimentais é de grande auxílio para o estudo da causalidade (ROSA et al., 2011). A partir dos anos 2000, com a grande quantidade de dados genotípicos e fenotípicos disponíveis, os modelos de equações estruturais e os modelos gráficos passaram a ser muito utilizados para estudo de relacionamentos causais em sistemas biológicos complexos (JANSEN; NAP, 2001; STEIN et al., 2003; MAGWENE; KIM, 2004; LI et al., 2005; SCHADT et al., 2005; KULP; JAGALUR, 2006; LI et al., 2006; CHEN; EMMERT-STREIB; STOREY, 2007; ZHU et al., 2007; CHAIBUB-NETO et al., 2008; LIU; HOESCHELE, 2008; JANSEN et al., 2009; CHAIBUB-NETO; KELLER; YANDELL, 2010; DUARTE; ZENG, 2011; BHADRA; MALLICK, 2013; CAI; BAZERQUE; GIANNAKIS, 2013; WANG et al., 2013; ZHANG et al., 2013; CHAIBUB-NETO; YANDELL, 2014; TUR; ROVERATO; CASTELO, 2014; WANG; EEUWIJK, 2014).

Particularmente, os experimentos de mapeamento de QTLs para fenótipos correlacionados fornecem uma situação análoga à experimentação. Para este caso específico, uma abordagem multivariada de mapeamento de QTLs, que leva em consideração a estrutura causal fenotípica, foi proposta por Mi et al. (2010a, 2010b). Considerando que a estrutura de relacionamento causal é conhecida, os autores desenvolveram uma metodologia de mapeamento de QTLs por intervalos compostos (CIM) para múltiplos caracteres utilizando modelos de equações estruturais em linhagens recombinantes endogâmicas (RILs). Os autores utilizaram uma rede de fenótipos relacionados à produtividade de grãos de trigo bem estabelecida para estruturar uma matriz de relacionamento causal e aplicaram esta matriz ao modelo de mapeamento CIM. Os autores demonstraram que esta metodologia permite a estimação dos efeitos diretos, indiretos e totais dos QTLs identificados e aumenta o poder de detecção de QTLs que apresentam efeitos diretos e indiretos com sinais (direções) opostos.

A pressuposição de que o modelo causal é conhecido, requerida pela metodologia de mape-

amento de Mi et al. (2010a, 2010b), raramente é verdadeira em sistemas biológicos complexos. A descoberta de modelos de causalidade costuma ser, na realidade, o passo mais crítico para este tipo de estudo. Teoricamente, o modelo de causalidade poderia ser descoberto utilizando conhecimento prévio dos pesquisadores. Na prática, nem sempre todo o conhecimento necessário para propor um modelo causal para as variáveis de estudo está disponível, principalmente quando o estudo é exploratório e envolve um grande número de variáveis. Outra maneira de realizar a descoberta de modelos de causalidade é propor modelos causais aleatórios ou baseados no conhecimento prévio existente e utilizar técnicas de modelos de equações estruturais e/ou modelos causais para testá-los, numa abordagem de tentativa e erro.

No entanto, este tipo de estratégia não é aplicável mesmo para sistemas causais com um número modesto de variáveis, uma vez que o número de modelos potenciais a serem testados cresce exponencialmente com o número de variáveis estudadas. Por este motivo, algoritmos de descoberta de causalidade têm sido propostos na literatura científica (PEARL, 2000; SPIRTES; GLYMOUR; SCHEINES, 2000). Basicamente, estes algoritmos realizam testes de independência condicionais para descobrir relações causais entre as variáveis, seguidos de busca por padrões de distribuições de probabilidade que permitam a orientação dos relacionamentos causais encontrados (SHIPLEY, 2002; COLOMBO; MAATHUIS, 2014; SHALIZI, 2014). Por ser computacionalmente muito eficiente, o algoritmo de Peter-Clark (PC) (SPIRTES; GLYMOUR; SCHEINES, 2000) se destaca dentre os algoritmos de descoberta de causalidade disponíveis e é atualmente considerado o algoritmo padrão para a realização deste tipo de análise (SHALIZI, 2014).

Muito embora o algoritmo de PC não requeira informações genotípicas, podendo ser aplicado para qualquer conjunto de variáveis correlacionadas, ele apresenta a desvantagem de frequentemente produzir modelos causais apenas parcialmente direcionados. A inclusão de informações de QTLs em algoritmos de descoberta de causalidade auxilia no direcionamento dos relacionamentos causais, pois “quebra” a equivalência de verossimilhança existente entre os dois possíveis relacionamentos causais entre um par de variáveis (CHAIBUB-NETO et al., 2008). Dentre os algoritmos de descoberta de causalidade que utilizam informações genotípicas disponíveis na literatura (ATEN et al., 2008; CHAIBUB-NETO; KELLER; YANDELL, 2010; LOGSDON; MEZEY, 2010; CAI; BAZERQUE; GIANNAKIS, 2013; WANG; EEUWIJK, 2014), o algoritmo QDG (*QTL-directed dependency graph*) (CHAIBUB-NETO et al., 2008) é atualmente considerado o estado da arte para este tipo de análise (LOGSDON; MEZEY, 2010; WANG; EEUWIJK, 2014). Recentemente, Wang e Eeuwijk (2014) propuseram o algoritmo QPSO (*QTL+phenotype supervised orientation*), como uma alternativa mais eficiente

ao algoritmo QDG.

Atualmente, mesmo em trabalhos que abordam a questão de causalidade fenotípica, o mapeamento de QTLs é realizado num passo separado da análise de causalidade. Chaibub-Neto, Keller e Yandell (2010) afirmam que a análise da causalidade e o mapeamento de QTLs deveriam ser realizados conjuntamente, pois o relacionamento causal pode melhorar as estimativas dos parâmetros dos QTLs (número, posição e efeito) e vice-versa. Com esta motivação os autores propuseram o algoritmo QTLnet (*QTL-driven phenotype network method*). Este algoritmo se baseia em estatística bayesiana e integra o mapeamento de QTL e a busca de modelos de causalidade utilizando MCMC (*Markov Chain Monte Carlo*), numa abordagem complexa que requer alto poder computacional. Por este motivo, este algoritmo ainda não pode ser aplicado a modelos causais com grande número de variáveis. Logsdon e Mezey (2010) demonstraram em estudos de simulação que o algoritmo QDG apresenta melhor performance que o algoritmo QTLnet mesmo em redes com pequeno número de variáveis. Assim, uma abordagem eficiente para a descoberta conjunta de modelos de causalidade e de QTL que afetam as características estudadas ainda precisa ser desenvolvida.

O milho é o cereal mais produzido mundialmente e um alimento básico para a segurança alimentar de países tropicais tais como os da América Latina (AWIKA, 2011), onde os solos frequentemente apresentam baixos níveis de fósforo. Assim, a obtenção de plantas com sistema radicular bem desenvolvido, morfologia radicular adequada e, portanto, mais eficientes na absorção de fósforo, é de grande interesse do melhoramento vegetal (SHEN et al., 2011).

A morfologia radicular pode ser aferida por meio da mensuração de caracteres fenotípicos correlacionados tais como diâmetro, comprimento, volume e área de raízes, proporção de raízes de diâmetros específicos, ramificação, etc (SOUSA et al., 2011). Sabe-se também que a morfologia radicular é um caráter complexo controlado por vários genes em milho (HOCHHOLDINGER; TUBEROSA, 2009; RAMAEKERS et al., 2010). Assim, o estudo do relacionamento causal entre os caracteres fenotípicos relacionados à morfologia de raízes, aliado a informações de QTLs controlando tais caracteres, pode fornecer informações muito úteis para os melhoristas, permitindo uma melhor compreensão da arquitetura genética destes fenótipos e de como os mesmos desenvolvem conjuntamente.

Este trabalho teve como objetivo analisar um conjunto de fenótipos relacionados à morfologia de raízes, acumulação de biomassa e concentração de fósforo em milho empregando mapeamento multivariado de QTLs e técnicas de modelagem de causalidade para a melhor compreensão da arquitetura genética subjacente a este conjunto de variáveis.

2 REVISÃO BIBLIOGRÁFICA

2.1 Causalidade

O conhecimento da estrutura de relação causal entre fenótipos permite entender sistemas biológicos complexos e prever como os mesmos se comportarão sob intervenção externa. No entanto, associações entre fenótipos são usualmente estudadas por meio da avaliação dos parâmetros genético-estatísticos covariância e correlação. Estes parâmetros estão relacionados com a proporção da variância que dois fenótipos compartilham devido a causas genéticas e quanto desta variância é comum a dois fenótipos, permitindo prever sua variação conjunta (ROSA et al., 2011).

Covariância e correlação são medidas de associações simétricas e podem ter explicações diversas: devido ao efeito de genes pleiotrópicos comuns aos dois fenótipos estudados; devido a genes distintos em desequilíbrio de ligação; devido à presença de uma série temporal ou a erros amostrais; ou, ainda, devido a um relacionamento de causalidade existente entre os fenótipos estudados. Dessa forma, a detecção de correlação entre características fenotípicas não indica necessariamente existência de relacionamento causal entre os fenótipos, muito menos revela a natureza deste relacionamento causal (SPIRITES; GLYMOUR; SCHEINES, 2000). Por estes motivos, a expressão “*Correlação não implica causalidade*” é frequentemente evocada em cursos e livros de Genética e de Estatística.

A literatura científica é rica em trabalhos a respeito das dificuldades de inferência de causalidade em experimentos observacionais. Talvez o exemplo mais notório seja o trabalho de Ronald A. Fisher com críticas contundentes aos estudos epidemiológicos que atribuíam maior probabilidade de desenvolvimento de câncer no pulmão ao hábito de fumar (FISHER, 1960; STOLLEY, 1991). Fisher (1960) corretamente afirmou que a constatação de existência de correlação entre duas variáveis em dados observacionais não distingue entre possível relação de causalidade entre (i) ambas variáveis; ou (ii) a existência de uma variável latente com efeito sobre as variáveis, e que apenas a experimentação e o emprego da aleatorização poderia distinguir entre estas duas possibilidades. Este exemplo notório demonstra como a dificuldade de inferência de causalidade em dados observacionais fez com que este tema fosse ignorado pelo método científico por um longo período (SHIPLEY, 2002). No entanto, a dificuldade de se tratar uma questão pela metodologia científica não faz com que sua importância diminua. Dessa forma, a partir do século XX, diversas tentativas de desenvolvimento de métodos aptos para lidar com

esta questão começaram a ser feitas.

Antes de discutir possíveis metodologias existentes para lidar com a causalidade, primeiramente, uma definição para este conceito é requerida. Para o escopo deste trabalho, a causalidade não será explorada em sua forma mais ampla, que envolve discussões milenares entre cientistas e filósofos. Ao contrário, o conceito será restrito à sua aplicabilidade contrafactual (SHALIZI, 2014): afirmar que X causa Y implica que a distribuição de Y é alterada caso haja uma modificação na distribuição de X , sendo o reverso não verdadeiro. Este conceito simples, aliado às seguintes pressuposições, é suficiente para o desenvolvimento de métodos estatísticos para o estudo de causalidade (SHIPLEY, 2002; SHALIZI, 2014; SPIRITES; GLYMOUR; SCHEINES, 2000):

- **Irreflexividade:** uma variável não pode causar a si própria. Ou seja, se X causa Y , Y não pode causar X , simultaneamente e simetricamente. Esta pressuposição leva à conclusão de que existe um Gráfico Acíclico Direcionado (DAG - *Directed Acyclic Graph*) G que representa o relacionamento causal entre as variáveis estudadas. É importante ressaltar que esta pressuposição não impede o estudo de alças de retroalimentações causais comuns em sistemas biológicos. Entretanto, para estudos deste tipo, informações de séries temporais que possibilitem conceber modelos causais onde X causa Y (no tempo t) e Y causa X (no tempo $t + \Delta_t$) são necessárias;
- **Transitividade:** se X causa Y e Y causa Z , então X deve causar Z ;
- **Condição Markoviana:** uma variável é causada apenas por sua causa mais próxima. Ou seja, se X causa Y e Y causa Z , a influência causal de X em Z é bloqueada caso Y seja impedido de variar;
- **Fidelidade:** as variáveis apresentam distribuição conjunta de todas as suas dependências condicionais em concordância com o gráfico G . Ou seja, combinações de valores específicos de parâmetros em G não mascaram os padrões de dependências preditos por G .

2.2 Randomização Mendeliana

A variação genética que ocorre naturalmente em populações segregantes é uma grande fonte de informação para estudos de causalidade do genótipo sobre o fenótipo. Cada variante genética (alelos, marcadores genéticos, etc) pode ser interpretado como uma intervenção exógena causal num sistema biológico. Além disso, os cruzamentos genéticos, por meio da recombinação e segregação, aleatorizam a distribuição destes variantes na população ou na progênie de

um cruzamento (ROCKMAN, 2008). Este fenômeno foi primeiramente estudado por Mendel que verificou heranças independentes para pares de genes (Segunda Lei de Mendel). Posteriormente, ficou comprovado que a Segunda Lei de Mendel é apenas válida para genes que estão em equilíbrio de ligação.

Randomização Mendeliana é o termo utilizado para estudos que empregam a independência da segregação de variantes genéticas como **variáveis instrumentais** para fazer inferências de causalidade em experimentos observacionais nos quais a aleatorização é impedida (LAWLOR et al., 2008). A independência na distribuição dos variantes genéticos que ocorre durante a meiose faz com que experimentos que se baseiam nestes variantes não sejam afetados por fatores de confundimento normalmente observado em experimentos observacionais (LAWLOR et al., 2008).

O conceito de randomização Mendeliana apareceu pela primeira vez em 1986, quando foi proposto que a relação de causalidade entre baixas taxas de colesterol e câncer seria melhor estudada pela avaliação entre câncer e o gene *Apo E*, com conhecido efeito sobre taxas de colesterol (KATAN, 2004). A ideia simples era que o gene *Apo E*, assim como outros fatores genéticos, seriam aleatorizados naturalmente em populações segregantes, e que tal associação estaria livre de efeitos de confundimento comuns em dados observacionais. Adicionalmente, devido ao dogma central da biologia, é assumido que genótipo tem efeito causal sobre o fenótipo (THOMAS; CONTI, 2004). Assim, a randomização de alelos durante a meiose pode ser vista como uma situação similar à experimentação requerida por Fisher para inferir causalidade (ROSA et al., 2011).

A disponibilidade em escala genômica de dados sobre variantes genéticas (genes, transcritos, proteínas e metabólitos) em populações segregantes despertou o interesse de utilizar estas informações para tentar modelar redes biológicas complexas (JANSEN; NAP, 2001). Este tipo de estudo, no qual se pretende inferir redes causais de observações biológicas, tem sido chamado de engenharia reversa, pois o objetivo vai além da identificação dos componentes do sistema para o entendimento de como o sistema biológico interage e funciona como um todo (ROCKMAN, 2008).

2.3 Abordagens multivariadas para mapeamento de QTLs

Os experimentos de mapeamento de QTLs possuem ligação imediata com o conceito de segregação Mendeliana e podem representar uma ótima oportunidade para o estudo de causalidade entre fenótipos correlacionados. No entanto, a maioria dos métodos de mapeamento de

QTLs disponíveis na literatura científica se refere à abordagem univariada de QTLs (LANDER; BOTSTEIN, 1989; HALEY; KNOTT, 1992; JANSEN, 1993; JANSEN; STAM, 1994; ZENG, 1993, 1994) e é empregada para mapeamento de QTLs para cada fenótipo individualmente, mesmo que o estudo tenha coletado dados sobre um conjunto de fenótipos correlacionados (MI et al., 2010a).

Devido à complexidade inerente às estratégias multivariadas de mapeamento de QTLs, alguns autores propuseram a redução da dimensionalidade das variáveis pela construção de variáveis canônicas utilizando metodologias de análise multivariada, principalmente a análise discriminante e a análise de componentes principais, seguidas de mapeamento de QTL univariado (WELLER et al., 1996; MANGIN; THOQUET; GRIMSLEY, 1998; CALINSKI et al., 2000; GILBERT; Le Roy, 2003). No entanto, este tipo de abordagem pode levar a uma simplificação artificial da questão genética e a resultados de difícil interpretação biológica (MI et al., 2010b; SILVA; ZENG, 2010).

Jiang e Zeng (1995) propuseram uma extensão do Mapeamento por Intervalos Compostos (CIM - *Composite Interval Mapping*) de Zeng (1993) para o caso multivariado. Silva, Wang e Zeng (2012) estenderam o Mapeamento de Intervalos Múltiplos (MIM - *Multiple Interval Mapping*) de Kao, Zeng e Teasdale (1999) para a análise conjunta de fenótipos correlacionados. Ambas as metodologias multivariadas de mapeamento de QTLs utilizam-se dos padrões de correlação representados na matriz de variância/covariância (vcov) para aumento do poder de detecção de regiões genômicas com efeito sobre os fenótipos estudados. Dessa forma, estes métodos apresentam a vantagem de, sob certas circunstâncias, apresentar maior poder de detecção de QTLs e precisão das estimativas dos parâmetros genéticos associados, além de permitirem o estudo do fenômeno de epistasia (SILVA; ZENG, 2010). Os modelos mistos também têm sido empregados para investigar as bases genéticas subjacentes às correlações entre caracteres fenotípicos correlacionados, especialmente para distinção entre as hipóteses de pleiotropia versus QTLs ligados (MARGARIDO, 2011). Malosetti et al. (2008), por exemplo, propuseram uma metodologia que emprega modelos mistos para mapeamento de QTLs para múltiplos caracteres e múltiplos locais (MTME - *multi trait multi environment*), que permite investigar, além das hipóteses citadas, a interação genótipo versus ambiente. No entanto, todos estes métodos desconsideram os possíveis relacionamentos de causalidade existentes entre os fenótipos estudados (MI et al., 2010b).

Mais recentemente, tem havido esforços para tentar incorporar informações de causalidade entre fenótipos em modelos multivariados de mapeamento de QTLs. Mi et al. (2010a, 2010b) desenvolveram metodologia de mapeamento de QTLs multivariada, aliando o CIM com modelos

de equações estruturais, num cenário em que é assumido que o relacionamento entre os fenótipos estudados é previamente conhecido. Outros trabalhos (CHAIBUB-NETO et al., 2008; ATEN et al., 2008; LOGSDON; MEZEY, 2010; WANG; EEUWIJK, 2014) propuseram a utilização de informações de mapeamento de QTLs para auxiliar a construção de modelos de relacionamento causais fenotípicos mais realistas. Chaibub-Neto, Keller e Yandell (2010), por sua vez, propuseram um algoritmo para inferir conjuntamente os parâmetros de mapeamento de QTLs e a rede causal entre os fenótipos estudados. Todas estas estratégias podem vir a permitir um melhor entendimento da arquitetura genética de fenótipos correlacionados.

2.4 Modelos de equações estruturais

Os Modelos de Equações Estruturais (MEEs) são um conjunto de técnicas estatísticas multivariadas que permite aos pesquisadores explicitarem um modelo de relacionamento causal, estimar e testar simultaneamente os efeitos diretos, indiretos e totais entre uma ou mais variáveis preditoras e uma ou mais variáveis respostas, bem como verificar a plausibilidade global do modelo proposto. Tanto as variáveis preditoras quanto as respostas podem ser contínuas ou discretas, observáveis ou latentes. No contexto dos MEEs, as variáveis também podem ser, simultaneamente, preditoras e respostas, de acordo com o modelo hipotetizado (LI et al., 2006; ULLMAN, 2006; ROSA et al., 2011). As várias técnicas de MEEs são chamadas na literatura de Modelagem da Causalidade (“*Causal Modeling*”), Análise de Causalidade (“*Causal Analysis*”), Modelagem Simultânea de Equações (“*Simultaneous Equation Modeling*”), Análise das Estruturas de Covariâncias (“*Analysis of Covariance Structures*”), dentre outras denominações.

Os MEEs se baseiam na minimização das diferenças entre a matriz de covariância observada e a matriz de covariância predita pelo modelo causal hipotetizado. O fundamento dos MEEs é que o relacionamento causal entre variáveis determina um padrão esperado de correlação. Os passos básicos dos MEEs são (MARUYAMA, 1997; SHIPLEY, 2002):

- Hipotetizar um modelo causal;
- Traduzir o modelo em equações estruturais;
- Derivar as variâncias preditas pelo modelo para cada par de variável;
- Estimar os parâmetros livres do modelo, minimizando as diferenças entre as matrizes de covariância predita e observada;
- Calcular a probabilidade de observação das diferenças minimizadas, assumindo que, se

o modelo causal proposto estiver correto, as mesmas são devido tão somente à variação amostral;

- Se a probabilidade calculada for menor que um valor p previamente estabelecido, concluir que os dados observados são consistentes com o modelo causal proposto e que o mesmo pode ser provisoriamente aceito.

O uso dos MEEs não pode aspirar o impossível, ou seja, provar causalidade na ausência de experimentação. O objetivo é verificar a plausibilidade de modelos causais hipotetizados. Além disto, é comum a descoberta de um conjunto de possíveis modelos causais compatíveis com as observações. Este conjunto é comumente chamado de modelos equivalentes e demonstram as limitações inerentes à estimação de causalidade em dados observacionais (SHIPLEY, 2002; MARUYAMA, 1997).

2.4.1 Histórico

A origem dos MEEs é atribuída a Sewall Wright e a Charles Spearman, no início do século XX (SCHUMACKER; LOMAX, 2010). Spearman (1904) propôs a metodologia de análise fatorial exploratória, que consiste na análise de padrões de correlações para inferência de fatores não observados que agregam as variáveis estudadas. Na década de 1920, Wright propôs a metodologia de análise de trilha (WRIGHT, 1921), também conhecida como "análise de caminhos" em alguns contextos. Wright trabalhou em sistemas em que os agentes causais (genes dos genitores) e suas consequências (fenótipos da prole) eram conhecidos e possuíam sentido unidirecional de causalidade (pais $\rightarrow$ prole). A presença de unidirecionalidade e ausência de variáveis latentes é a única situação apropriadamente chamada de análise de trilha (MARUYAMA, 1997).

Na análise de trilha, análise de correlações e técnicas de regressão são utilizadas para propor relacionamentos mais complexos que os modelados ordinariamente por regressão múltipla, permitindo a inferência de efeitos diretos e indiretos entre variáveis estudadas. A análise de trilha foi proposta no mesmo tempo histórico que a análise de variância (ANOVA - "*Analysis of Variance*") (FISHER, 1925), imediatamente reconhecida como uma técnica estatística superior por testar modelos causais num arcabouço experimental, com a atribuição de nível de significância. A análise de trilha, por sua vez, permite apenas estimar efeitos causais diretos e indiretos com a pressuposição de que o modelo causal proposto está correto (SHIPLEY, 2002). Além disto, o pensamento científico predominante na época, do qual Fisher era grande defensor, preconizava que a metodologia científica seria incapaz de inferir relações de causalidade em dados observacionais. Por estes motivos, apesar de o próprio Sewall Wright considerar a Análise de Trilha como

uma de suas mais relevantes contribuições à Ciência, ela teve sua aplicação bastante limitada nas Ciências Biológicas e Agrárias quando comparada à ANOVA (SHIPLEY, 2002).

Desde o seu início, a análise fatorial exploratória foi aplicada às Ciências Sociais, nas quais a randomização requerida pela ANOVA nem sempre é possível. No entanto, a metodologia tinha a limitação de não permitir o teste de modelos causais pré-concebidos. A dificuldade de realização de experimentos randomizados nas Ciências Sociais fez com que a análise de trilha fosse resgatada pelos cientistas sociais na década de 1960. Porém, as limitações da análise de trilha para estas Ciências, nas quais sistemas causais compostos de variáveis latentes, com relacionamentos complexos que incluem retroalimentação e erros de mensuração são comuns, constituíram um empecilho para sua adoção ampla. Em 1969, Jöreskog (1969) propôs o método da análise fatorial confirmatória, que permite verificar se um conjunto de variáveis correlacionadas são capazes de definir uma variável latente hipotetizada (conhecidas como “construtos” pela Psicologia). Jöreskog (1970) propôs a metodologia dos modelos de equações estruturais (MEEs) baseada na estimação dos parâmetros e teste do modelo causal proposto em relação aos dados observados por meio de Máxima Verossimilhança (ML), solucionando as maiores limitações da análise de trilha e da análise fatorial. O mesmo autor desenvolveu o programa computacional LISREL (“*Linear Structural RELations*”) facilitando a adoção de sua metodologia.

Assim, os MEEs tomaram sua forma atual, representando uma modelagem que combina análise de trilha e análise fatorial confirmatória, solucionadas por meio de ML (MARUYAMA, 1997).

2.4.2 Terminologia

Os MEEs podem ser representados tanto por meio de gráficos quanto por meio de equações estruturais. Ambas as representações são completas por tornarem explícitas as hipóteses de relacionamento entre as variáveis estudadas (MARUYAMA, 1997).

Em sua representação gráfica, os MEE são ilustrados por diagramas contendo variáveis conectadas por setas. Setas unidirecionais ($\rightarrow$) são empregadas para representar relação de causalidade, com a variável do início da seta causando diretamente a variável para a qual a seta é direcionada. A omissão de seta entre duas variáveis no gráfico implica na ausência de relação de causalidade entre as mesmas no modelo hipotetizado.

A Figura 1 representa um modelo estrutural clássico (BOLLEN, 1989). Este exemplo será utilizado para ilustrar a terminologia envolvida na estimação dos modelos de equações estruturais. A notação apresentada a seguir também é a sugerida por Bollen (1989).

Figura 1 – Modelo estrutural de variáveis latentes: democracia política em 1960 (η_1), democracia política em 1965 (η_2) e industrialização em 1960 (ξ_1) em países em desenvolvimento. Adaptado de Bollen (1989)

No modelo de equações estruturais da Figura 1, três variáveis são relacionadas: nível de industrialização (ξ_1), nível de democracia em 1960 (η_1) e em 1965 (η_2), em países em desenvolvimento. As variáveis estão representadas dentro de elipses, o que na linguagem dos MEEs indicam que as mesmas não foram mensuradas diretamente, ou seja, representam **variáveis latentes**. A variável que não recebe nenhuma seta é chamada de **exógena**, tal qual ξ_1 no modelo da Figura 1. A variável que recebe pelo menos uma seta é chamada de variável **endógena**. As variáveis η_1 e η_2 são variáveis endógenas e, por isso, possuem resíduos associados (ζ_1 e ζ_2), que representam as variâncias não explicadas pelo modelo. A variável que recebe uma seta é chamada de **variável resposta** ou **dependente**, e a variável da qual uma seta emana é chamada de **variável explanatória** ou **independente** (BOLLEN, 1989). No MEEs, uma variável pode ser simultaneamente dependente e independente tal qual a variável η_1 no modelo da Figura 1.

A magnitude do efeito de causalidade entre as variáveis é dada pelos **coeficientes de trilha**. Na Figura 1, o coeficiente de trilha β_{12} representa o **efeito direto** de η_1 sobre η_2 . A variável ξ_1 tem efeito direto sobre η_1 e η_2 , representados pelos coeficientes γ_{21} e γ_{11} e **efeito indireto** sobre η_2 representado pelo produto dos dois coeficientes de trilha correspondentes ($\gamma_{11} \times \beta_{12}$) (BOLLEN, 1989). O modelo também pode ser representado pelo seguinte sistema de equações:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{12} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} \begin{bmatrix} \xi_1 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \quad (1)$$

Ou mais compactamente como:

$$\eta = B\eta + \Gamma\xi + \zeta \quad (2)$$

em que η representa o vetor de m variáveis endógenas ($m \times 1$) e B representa a matriz de coeficientes de trilha das relações de causalidade entre as variáveis endógenas ($m \times m$); ξ é o vetor de n variáveis exógenas ($n \times 1$), e Γ é a matriz de coeficientes de trilha das relações de causalidade das variáveis exógenas sobre as variáveis endógenas ($m \times n$); e ζ é o vetor dos resíduos ($m \times 1$) (BOLLEN, 1989).

O modelo proposto por Bollen (1989) é chamado de **recursivo**, por não conter alças de retro-alimentação, ou seja, a causalidade segue sentido unidirecional. O termo recursivo implica que as equações do modelo estrutural podem ser ordenadas e resolvidas sequencialmente para cada variável dependente e o resultado pode ser retornado *recursivamente* ao sistema de equações até a completa solução do mesmo (MARUYAMA, 1997). Em modelos **não-recursivos**, este tipo de abordagem não é possível e a solução do sistema tem que ser realizada iterativamente.

O modelo da Figura 1 representa o relacionamento hipotético de variáveis latentes, não mensuradas diretamente. Variáveis latentes são comumente estimadas via análise fatorial confirmatória, uma metodologia de análise multivariada que testa o quão bem as variáveis medidas, também conhecidas como **variáveis indicadoras**, que representam um número menor de variáveis latentes. Por este motivo, para que os parâmetros presentes na Figura 1 possam ser estimados, é necessário um modelo de mensuração adicional, representado na Figura 2.

O modelo da Figura 2 indica que as variáveis latentes democracia em 1960 (η_1) e 1965 (η_2) foram inferidas por meio das variáveis indicadoras liberdade de imprensa (y_1 e y_5), liberdade de oposição (y_2 e y_6), eleições livres (y_3 e y_7) e atuação do poder legislativo (y_4 e y_8), mensuradas em 1960 e 1965, respectivamente. Como as variáveis indicadoras são efetivamente mensuradas, elas são representadas dentro de retângulos para serem diferenciadas das variáveis latentes do modelo.

O modelo assume que todas as variáveis indicadoras são mensuradas com erro (ε_i) e que os erros das variáveis medidas em tempos diferentes são correlacionados. Os erros dos pares de variáveis (y_4 e y_2) e (y_8 e y_6) também são correlacionados por virem do mesmo banco de dados (BOLLEN, 1989). Os relacionamentos entre as variáveis latentes e as variáveis indicadoras são medidos pelos **fatores de carga** λ_i , sendo que o primeiro fator de carga é fixado em 1 para fixar a escala de variância da variável latente igual à variância da primeira variável indicadora (BOLLEN, 1989). Dessa forma, temos as equações do modelo de mensuração para as variáveis endógenas η_1 e η_2 :

Figura 2 – Modelo de mensuração de variáveis latentes: democracia política em 1960 (η_1), democracia política em 1965 (η_2) e industrialização em 1960 (ξ_1) em países em desenvolvimento. Adaptado de Bollen (1989)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ \lambda_4 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \\ 0 & \lambda_3 \\ 0 & \lambda_4 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix} \quad (3)$$

Ou mais compactamente como:

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (4)$$

em que $\mathbf{y}$ é o vetor de p variáveis indicadoras de m variáveis endógenas ($p \times 1$); $\mathbf{\Lambda}_y$ é a matriz de coeficientes dos fatores de carga entre as variáveis indicadoras e latentes $p \times m$; $\boldsymbol{\eta}$ representa o vetor de m variáveis endógenas ($m \times 1$); e $\boldsymbol{\varepsilon}$ é o vetor dos erros de mensuração das variáveis indicadoras ($p \times 1$) (BOLLEN, 1989).

A variável latente industrialização (ξ_1) foi estimada indiretamente por meio da mensuração das variáveis indicadoras produto interno bruto *per capita* (x_1), consumo de energia *per capita* (x_2) e porcentagem da força de trabalho ocupada no setor industrial (x_3), que também apresentam erros de mensuração (δ_i), porém não correlacionados. As equações deste submodelo de mensuração são apresentadas a seguir:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_6 \\ \lambda_7 \end{bmatrix} \begin{bmatrix} \xi_1 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \quad (5)$$

Ou mais compactamente como:

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \quad (6)$$

em que $\mathbf{x}$ é o vetor de q variáveis indicadoras de n variáveis exógenas ($q \times 1$); $\mathbf{\Lambda}_x$ é a matriz de coeficientes dos fatores de carga entre as variáveis indicadoras e latentes exógenas $q \times n$; $\boldsymbol{\xi}$ representa o vetor de n variáveis endógenas ($n \times 1$); e $\boldsymbol{\delta}$ é o vetor dos erros de mensuração das

Figura 3 – Modelo de Equações Estruturais para o estudo das variáveis latentes: democracia política em 1960 (η_1), democracia política em 1965 (η_2) e industrialização em 1960 (ξ_1) em países em desenvolvimento (BOLLEN, 1989)

variáveis indicadoras ($p \times 1$) (BOLLEN, 1989).

A combinação do modelo estrutural representado pela Figura 1 e do modelo de mensuração representado pela Figura 2 leva ao modelo de equações estruturais na sua forma mais geral representado pela Figura 3. Assim, os MEEs representam a resolução simultânea do seguinte sistema de equações:

$$\begin{cases} \eta = B\eta + \Gamma\xi + \varsigma \\ y = \Lambda_y\eta + \varepsilon \\ x = \Lambda_x\xi + \delta \end{cases} \quad (7)$$

2.4.3 Estimação

Nos MEEs, assume-se que as variáveis endógenas seguem distribuição normal multivariada, enquanto que as variáveis exógenas podem ser contínuas ou discretas. Intuitivamente, a estimação dos parâmetros dos MEEs envolve a escolha de valores paramétricos que respeitem as restrições impostas pelo modelo causal proposto e que tornem a matriz de variância/covariância (vcov) predita o mais similar possível à matriz vcov observada (LI et al., 2006). Estatisticamente, a hipótese de nulidade testada pelos MEEs é de que, sendo o modelo causal proposto verdadeiro, a matriz de variância populacional (Σ) pode ser escrita como uma função dos parâmetros (θ) do modelo (BOLLEN, 1989):

$$\Sigma = \Sigma(\theta)$$

As matrizes de variâncias implicadas pela Equação 7 são descritas a seguir. A matriz vcov residual das variáveis endógenas η_i é denominada Ψ , enquanto a matriz vcov das variáveis exógenas ξ_i é representada por Φ . A matriz vcov residual das variáveis indicadoras y_i é representada por Θ_ϵ e das variáveis x_i é representada por Θ_δ . Os valores de B , Γ , Θ e Φ reunidos no vetor θ são os parâmetros a serem estimados de forma a minimizar as diferenças entre as matrizes vcov predita ($\hat{\Sigma}$) e observada (S). As Equações 8 e 9 apresentam as matrizes de variâncias $\Sigma(\theta)$ das variáveis indicadoras x_i e y_i (BOLLEN, 1989):

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} \quad (8)$$

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I - B)^{-1}]'\Gamma'_y + \Theta_\epsilon & \Gamma_y(I - B)^{-1}\Gamma\Phi\Gamma'_x \\ \Gamma_x\Phi\Gamma'[(I - B)^{-1}]'\Gamma'_y & \Gamma_x\Phi\Gamma'_x + \Theta_\delta \end{bmatrix} \quad (9)$$

O método mais utilizado para estimar os parâmetros dos MEEs é o método da Máxima Verossimilhança. A função F_{ML} , apresentada na Equação 10, é minimizada em relação aos parâmetros contidos em θ . Bollen (1989) apresenta a derivação desta função assumindo que y e x apresentam distribuição normal multivariada.

$$F_{ML} = \log|\Sigma(\theta)| + tr\{S\Sigma^{-1}(\theta)\} - \log|S| - (p + q) \quad (10)$$

Se os dados observados, representados pela matriz S , tiverem sido gerados pelo modelo

causal proposto, a minimização de F_{ML} faz com que as diferenças restantes entre S e $\Sigma(\theta)$ sejam devido à variação aleatória amostral. Nesta situação, $(N - 1)F_{ML}$ possui distribuição assintótica de χ^2 , com $[v(v + 1) - (p + q)]$ graus de liberdade, em que v representa o número de variáveis observáveis na matrix S (SHIPLEY, 2002).

É importante ressaltar que os MEEs são flexíveis e que o modelo de Bollen (1989) é bastante geral. Assim, é possível derivar modelos mais simples, que representam casos específicos, escolhendo matrizes vcov apropriadas. Por exemplo, MEEs sem variáveis latentes são obtidos da Equação 7 considerando $\Theta_\epsilon = 0$, $\Theta_\delta = 0$, $\Lambda_y = I_p$ e $\Lambda_x = I_q$.

Para que o modelo estrutural seja *identificável*, isto é, para que seja possível realizar a estimação de seus parâmetros, é necessário que o número de covariâncias observadas seja maior que o número de parâmetros a serem estimados (BOLLEN, 1989). Infelizmente, o cumprimento desta condição não é garantia de identificação pois, os modelos de equações estruturais requerem condições adicionais para serem identificáveis. O tópico "identificabilidade dos modelos de equações estruturais" é complexo e, por fugir do escopo desta revisão não será aqui tratado. Shipley (2002) traz uma revisão interessante sobre este assunto para o leitor interessado no tópico.

2.4.4 Aplicações de MEEs na área biológica

A vantagem dos MEEs em sistemas biológicos é a possibilidade de estudo de alças de retroalimentação comumente encontradas em rotas metabólicas. No entanto, os MEEs pressupõem relacionamentos lineares entre as variáveis, o que nem sempre é realista em sistemas biológicos (ROCKMAN, 2008). Apesar desta limitação, os MEEs vêm sendo empregados com sucesso para a modelagem de dados biológicos. Estes modelos se mostraram particularmente interessantes para delineamentos ancorados em gêmeos ("*twin designs*"), comuns na Medicina. Os MEEs representam uma abordagem mais poderosa que as análises de correlações par a par até então usualmente feitas, pois baseiam-se na solução simultânea de toda informação intrapares (NEALE; CARDON, 1992). Utilizando os MEEs neste tipo de delineamento, Stunkard et al. (1990) e Hong et al. (1997) verificaram a importância relativa de efeitos genéticos e ambientais de caracteres de massa corporal e síndrome metabólica.

Com a profusão de dados genômicos em sistemas biológicos complexos multivariados (Biologia de Sistemas), os MEEs foram redescobertos e atualmente têm sido bastante utilizados como uma ferramenta para auxiliar na interpretação de dados fenotípicos e moleculares (JANSEN; NAP, 2001).

Oord (2000) propôs metodologia de mapeamento de QTLs (“*Quantitative Trait Loci*”) usando os MEEs em estudos de mapeamento associativo de síndromes metabólicas complexas, tais como hipertensão arterial. Stein et al. (2003) empregaram os MEEs para estudos de posições genômicas relacionadas a subfenótipos da síndrome metabólica (obesidade, hipertensão, diabetes e níveis sanguíneos de colesterol e triglicérides) em humanos. Os autores foram capazes de concluir que obesidade, hipertensão e diabetes estavam sob o controle de vários locos pleiotrópicos.

Li et al. (2006) utilizaram os MEEs em sistemas causais mais complexos que o causal, reativo e independente propostos inicialmente por Schadt et al. (2005) e avaliados por meio de modelos gráficos. Na metodologia proposta por Li et al. (2006), a estrutura de causalidade é inferida *a posteriori*, após duas rodadas de buscas por QTLs ao longo do genoma, sendo que a primeira busca é realizada para os fenótipos individuais e a segunda utiliza fenótipos intermediários como cofatores na análise dos fenótipos principais. Utilizando esta metodologia, os autores foram capazes de distinguir QTLs que controlam adiposidade daqueles que controlam o tamanho corporal em ratos. Outro exemplo de aplicação da metodologia proposta por Li et al. (2006) pode ser encontrado em Scheffer (2008).

Liu e Hoeschele (2008) utilizaram os MEEs com o objetivo de reconstruir redes de regulação em experimentos de escala genômica contendo dados de expressão de 4.589 genes de leveduras. Os pesquisadores utilizaram modelos estruturais locais para identificar pares de eQTLs relacionados (regulador/alvo). Com esta restrição inicial do espaço amostral, a rede total foi inferida utilizando os MEEs. Baseados nesta metodologia, os autores foram capazes de inferir uma rede contendo várias centenas de genes e eQTLs.

No melhoramento animal, os MEEs têm sido empregados para a inferência de redes de causalidade entre fenótipos relacionados à produção na ausência de dados genotípicos (VALENTE et al., 2010, 2011; MATURANA et al., 2009; BOWMAN et al., 2014). Em plantas, Mi et al. (2010a, 2010b) desenvolveram metodologia de mapeamento de QTLs multivariada, aliando o CIM com os MEEs. Os autores empregaram estrutura causal de fenótipos relacionados à produtividade de grãos de trigo previamente estabelecida no modelo de mapeamento. A comparação dessa metodologia com CIM tradicional em dados de simulação revelou que, além de os MEEs permitirem a estimação dos efeitos diretos, indiretos e totais dos QTLs, os MEEs apresentam maior poder de detecção de QTLs. Este aumento foi maior em QTLs que apresentavam efeitos diretos e indiretos em direções opostas. Ao aplicar a metodologia em dados experimentais para o mapeamento de QTLs em trigo, os autores foram capazes de identificar QTLs não relatados anteriormente.

2.5 Modelos gráficos

Os modelos gráficos, ou redes bayesianas, são também uma técnica de modelagem de sistemas causais. Em comum com os MEEs, os modelos gráficos representam as hipóteses de relacionamento causal entre as variáveis com um gráfico direcionado contendo variáveis mensuradas, chamadas de **nós**, e relações de causalidade representadas como setas direcionadas entre os nós. No entanto, enquanto os MEEs enfatizam a minimização das diferenças entre os padrões de correlação observados e previstos pelo modelo causal proposto, os modelos gráficos focam na comparação entre as distribuições de probabilidade implicadas pelo modelo causal proposto e as observadas entre as variáveis estudadas.

Comparados ao MEEs, os modelos gráficos têm histórico recente. Judea Pearl lançou as pedras fundamentais desta metodologia no final dos anos 1980 (PEARL, 2000). Pearl desenvolveu o conceito de d-separação e estabeleceu uma série de algoritmos para extrair informações de causalidade a respeito de dados observacionais (PEARL, 2000). Imediatamente após o desenvolvimento do conceito de d-separação, outros grupos de pesquisadores com interesse na área desenvolveram seus próprios algoritmos (SPIRITES; GLYMOUR; SCHEINES, 2000).

2.5.1 Terminologia

As redes bayesianas são uma representação de distribuição de probabilidade conjunta de um grupo de variáveis. Assim, se X é um conjunto de variáveis aleatórias $X_1, X_2, \dots, X_n$, a rede bayesiana que descreve o relacionamento destas variáveis é composta pelo Gráfico Direcionado Acíclico (DAG) G , cujos **nós** ou **vértices** são compostos pelos elementos de X . Setas direcionadas ($\rightarrow$) indicam a influência direta de um nó **pai** em um nó **filho**. Os vértices com efeitos indiretos sobre uma variável são chamados de **ancestrais**. Um gráfico é dito **não-direcionado** ou **parcialmente direcionado** quando apresenta todas ou algumas ligações entre os nós não direcionadas por meio de setas. Neste caso, as setas ($\rightarrow$) são substituídas na representação gráfica por um traço ($-$). Quando um gráfico apresenta alças de retroalimentação, ele é chamado de **cíclico** (FRIEDMAN et al., 2000).

Dados três vértices X_1, X_2 e X_3 adjacentes em G , as orientações $X_1 \rightarrow X_2 \rightarrow X_3$ e $X_1 \leftarrow X_2 \leftarrow X_3$ são chamadas **cadeias**, enquanto a orientação $X_1 \leftarrow X_2 \rightarrow X_3$ é chamada de **forquilha** e a orientação $X_1 \rightarrow X_2 \leftarrow X_3$ é chamada de **colisor** ou **estrutura em V** (PEARL, 2000; SHALIZI, 2014). A Figura 4 é um exemplo simples de DAG.

A propriedade mais importante de uma rede Bayesiana é a condição Markoviana, ou seja, o estado de uma variável é dependente apenas do estado de seus nós parentais, ou em outra formu-

Figura 4 – Exemplo de modelo gráfico simples, com cinco nós (FRIEDMAN et al., 2000)

lação, uma variável é independente de todas as suas causas ancestrais dado seus nós parentais (SHIPLEY, 2002).

Se um gráfico não possui alças de retroalimentação, ou seja, é um DAG, a condição Markoviana é diretamente traduzida para o conceito de **d-separação**:

Considere um DAG G , que contenha as variáveis X e Y e um subconjunto de variáveis S (que não contenha X e Y). Então, X e Y são d-separadas se houver S , tal que $\{X \perp\!\!\!\perp Y | S\}$. Se X e Y são d-separadas, então não pode haver uma ligação causal entre estas variáveis, representada por $\rightarrow$ ou $-$, em G .

Além da representação gráfica, os modelos causais implicam numa distribuição de probabilidade conjunta. Devido à condição Markoviana e à d-separação, o gráfico da Figura 4 implica em várias afirmações de dependência condicionais: $\{E \perp\!\!\!\perp A, D\}$, $\{B \perp\!\!\!\perp D | A\}$, $\{C \perp\!\!\!\perp E, A, D | B\}$, $\{D \perp\!\!\!\perp E, B, C | A\}$. Esta rede também implica na seguinte distribuição de probabilidade conjunta (FRIEDMAN et al., 2000):

$$P(A, B, C, D, E) = P(A)P(E)P(B|A, E)P(D|A)P(C|B)$$

Embora os modelos gráficos estejam sendo atualmente bastante utilizados na área biológica, principalmente para a reconstrução de redes biológicas, eles também têm aplicação em várias outras áreas do conhecimento nas quais a modelagem de incertezas é requerida, dentre as quais pode-se citar a Ciência da Computação, principalmente em Inteligência Artificial (KOLLER; FRIEDMAN, 2009).

2.5.2 Algoritmos de descoberta de causalidade

A definição de modelos de causalidade que devem ser testados (**Especificação de Modelo**), é um passo crucial na análise de causalidade. Os modelos a serem testados podem ser definidos por meio de conhecimento prévio, por tentativa e erro ou por algoritmos de descoberta de causalidade (SHIPLEY, 2002).

O conhecimento prévio dos pesquisadores sempre deve ser levado em consideração no momento de especificação do modelo. Porém, muitas vezes o conhecimento pré-existente não é suficiente ou exatamente acurado, principalmente ao se lidar com um grande número de variáveis. Além disso, o espaço amostral de modelos possíveis cresce exponencialmente com a adição de características a serem estudadas, o que torna a análise exploratória de modelos por tentativa e erro bastante limitada. Assim, algoritmos de descoberta de causalidade são de grande utilidade para a geração de modelos de causalidade hipotéticos e têm sido empregados cada vez mais frequentemente (SHIPLEY, 2002). No caso da Genética, o emprego destes algoritmos tem se tornado muito comum em experimentos envolvendo a expressão gênica.

O conceito de d-separação permitiu o desenvolvimento destes algoritmos (SHIPLEY, 2002). Basicamente, os algoritmos realizam testes sequenciais de d-separação para verificar a existência de setas entre as variáveis (passo de descoberta de setas) e, posteriormente, tentam orientar as setas identificadas pela identificação de trio de variáveis que constituem um colisor no modelo (SHALIZI, 2014). Dado um DAG G , em adição às pressuposições de irreflexibilidade, transitividade, fidelidade e condição Markoviana comuns às análises de causalidade, as pressuposições dos algoritmos de descoberta causal incluem (SHALIZI, 2014):

1. Todos os membros da população têm a mesma distribuição;
2. Todas as variáveis relevantes foram mensuradas. Esta pressuposição é também conhecida como **suficiência causal**;
3. Existe apenas um gráfico G para o qual a distribuição é fiel.

O algoritmo SGS (Spirtes-Glymour-Scheines) (SPIRITES; GLYMOUR; SCHEINES, 2000) foi um dos primeiros algoritmos de descoberta de causalidade a ser desenvolvido. Os passos do algoritmo SGS são os seguintes:

1. Iniciar com um gráfico G com setas não direcionadas entre todas as variáveis;
2. Para cada par de variáveis entre as X_1 e X_2 e cada subconjunto S de outras variáveis, verificar se $\{X_1 \perp\!\!\!\perp X_2 | S\}$; caso verdadeiro, retirar a seta entre X_1 e X_2 ;

3. Para cada trio de variáveis adjacentes em G , verificar se o mesmo constitui um colisor pelo padrão de correlações parciais; caso verdadeiro, orientar as setas adequadamente;
4. Tentar orientar as demais setas do gráfico por meio da consistência com os colisores já identificados.

O algoritmo PC (Peter-Clark) (SPIRITES; GLYMOUR; SCHEINES, 2000) tem as mesmas pressuposições do algoritmo SGS, porém é computacionalmente mais eficiente, pois diminui o espaço de buscas ao realizar os testes de independência condicional no subconjunto S' , em que S' representa um subconjunto de variáveis adjacentes a X_1 e X_2 (KALISCH; COLOMBO; MAATHUIS, 2012). Shalizi (2014) afirma que o algoritmo PC deve ser o algoritmo de escolha padrão para testes de descoberta causal.

Uma grande limitação dos algoritmos de descoberta de causalidade é o passo de orientação de setas, pois apenas trios colisores apresentam padrão de correlação identificável. Devido a este fato, é comum que os algoritmos forneçam um gráfico apenas parcialmente orientado (CPDAG - *Completely Partially Directed Acyclic Graph*), que representa um conjunto de gráficos causais com distribuições de probabilidade equivalentes (SHIPLEY, 2002).

2.5.3 Algoritmos de descoberta de causalidade auxiliados por QTLs

Devido às propriedades da randomização Mendeliana, informações a respeito de QTLs são de grande auxílio na orientação de setas em algoritmos de descoberta de causalidade. Chaibub-Neto et al. (2008) desenvolveram um algoritmo chamado QDG (*“QTL-directed Dependency Graph”*) para esta finalidade. O algoritmo QDG é composto de dois passos. Na fase de descoberta de setas, uma rede não direcionada é construída utilizando os algoritmos UDG (*“Undirected Dependency Graph”*) (FUENTE et al., 2004) ou o passo de descoberta de setas do algoritmo PC (Esqueleto de PC) (SPIRITES; GLYMOUR; SCHEINES, 2000). O passo de orientação das setas identificadas é realizado por meio de testes de *LOD score* condicionados em QTLs conectados aos fenótipos estudados. Os criadores do algoritmo QDG (CHAIBUB-NETO et al., 2008) frisam que os fenótipos estudados devem possuir pelo menos um QTL não compartilhado para garantir que a função densidade de probabilidade entre os dois modelos possíveis (causal e reativo) possuam verossimilhança distintas.

Considerando-se duas variáveis, y_1 e y_2 , os relacionamentos causais $y_1 \rightarrow y_2$ e $y_1 \leftarrow y_2$ produzem a mesma distribuição de probabilidade, pois: $f(y_1)f(y_2|y_1) = f(y_1, y_2) = f(y_2)f(y_1|y_2)$. O algoritmo QDG utiliza informação de causalidade oriunda de mapeamento de QTL como ilustrada na Figura 5 para desfazer a equivalência de verossimilhança entre $y_1 \rightarrow y_2$ e $y_1 \leftarrow y_2$ como

descrito na Equação 11, permitindo a orientação da causalidade entre y_1 e y_2 . Dessa forma, o algoritmo QDG permite a obtenção de um gráfico direcionado único, ao invés de uma classe de gráficos com equivalência Markoviana representado numa CPDAG (*completed partially directed acyclic graph*) comumente encontrados pelo algoritmo de PC.

Figura 5 – Relacionamento causal entre duas variáveis y_1 e y_2 e seus respectivos QTLs. Adaptado de Chaibub-Neto et al. (2008)

$$f(\mathbf{q}_1)f(y_1|\mathbf{q}_1)f(y_2|y_1, \mathbf{q}_2)f(\mathbf{q}_2) \neq f(\mathbf{q}_2)f(y_2|\mathbf{q}_2)f(y_1|y_2, \mathbf{q}_1)f(\mathbf{q}_1) \quad (11)$$

em que, $\mathbf{q}_1$ e $\mathbf{q}_2$ representam os conjunto de i e j QTLs afetando as características y_1 e y_2 , respectivamente.

Para o passo de descoberta de setas, o algoritmo QDG utiliza o algoritmo de esqueleto de PC (SPIRITES; GLYMOUR; SCHEINES, 2000). Para a orientação de setas, o algoritmo QDG avalia a direção da causalidade de cada seta conectando um par de fenótipos, usando a razão de LOD condicional aos QTLs que afetam estes fenótipos. O algoritmo pode acomodar alças de retroalimentação no gráfico global de causalidade pois a adição de informação de QTLs também quebra a equivalência na distribuição conjunta de probabilidades de gráficos cíclicos. Assim é possível distinguir entre $y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow y_1$ e $y_1 \leftarrow y_2 \leftarrow y_3 \leftarrow y_1$ pois estes gráficos possuem verossimilhanças distintas, graças à presença dos QTLs (Equação 2.5.3). Assim:

$$f(y_1|y_3, \mathbf{q}_1)f(y_2|y_1, \mathbf{q}_2)f(y_3|y_2, \mathbf{q}_3) \neq f(y_1|y_2, \mathbf{q}_1)f(y_2|y_3, \mathbf{q}_2)f(y_3|y_1, \mathbf{q}_3)$$

Os autores verificaram a eficácia de sua metodologia em estudos de simulação e foram capazes de prever uma rede de genes e metabólitos ligados a obesidade e suscetibilidade à diabetes

em ratos. Posteriormente, experimentos *in vitro* com o sistema estudado confirmaram alto nível de acertos das predições do algoritmo (CHAIBUB-NETO et al., 2008). Wang e Eeuwijk (2014) consideram este algoritmo o “estado da arte” nos estudos de causalidade envolvendo informações de QTLs.

Aten et al. (2008) propuseram um algoritmo de descoberta de causalidade (NEO - *Network Edge Orienting*) similar ao QDG. No entanto, ao invés de realizar testes de LOD, o algoritmo NEO utiliza os MEEs para orientar as setas, fazendo testes de ajuste de modelos competidores localmente para cada par de variáveis e globalmente para o modelo completo.

O algoritmo QPSO (*QTL+Phenotype Supervised Orientation*) propõe melhoria ao algoritmo QDG, ao utilizar toda a informação presente no modelo causal para orientar as setas entre o par de características analisadas (WANG; EEUWIJK, 2014). Wang e Eeuwijk (2014) argumentam que, adicionalmente às informações de QTL, os relacionamentos causais entre os fenótipos do modelo causal também devem ser considerados na orientação da seta ligando os fenótipos y_1 e y_2 (Figura 6). A Figura 6 é uma extensão do relacionamento causal entre duas variáveis y_1 e y_2 e seus respectivos QTLs, representado na Figura 5.

O algoritmo QPSO considera toda a informação adjacente ao relacionamento causal entre y_1 e y_2 , representada numa rede local fenotípica generalizada (LGPN - *Local Generalized Phenotype Network*). A LGPN é constituída por: (i) QTLs e fenótipos causais identificados para y_1 e/ou y_2 (nós parentais); e, (ii) fenótipos conectados a y_1 e/ou y_2 por setas não direcionadas (nós vizinhos) (WANG; EEUWIJK, 2014). A LGPN representa 2^n DAGs distintos, em que n representa o número de setas não direcionadas presentes na LGPN. Para cada par de fenótipos conectados por uma seta não-direcionada, o algoritmo QPSO identifica o DAG com melhor valor de LOD entre aqueles representados pela respectiva LGPN. Este DAG é chamado de gráfico direcionado ótimo local (LODG - *Locally Optimal Directed Graph*). Ao realizar esta operação várias vezes, partindo de diferentes pares de fenótipos conectados por setas não direcionadas, o algoritmo QPSO é capaz de chegar a um único modelo de causalidade para explicar o relacionamento causal entre todos os fenótipos estudados (WANG; EEUWIJK, 2014).

Figura 6 – Representação de uma rede fenotípica local generalizada (LGPN) entre dois fenótipos y_1 e y_2 ligados por um relacionamento de causalidade não direcionado (seta verde). $\mathbf{P}_1 = \{P_{11}, \dots, P_{1k}\}$ e $\mathbf{P}_2 = \{P_{21}, \dots, P_{2l}\}$ são os nós parentais exclusivos de y_1 e y_2 , respectivamente; $\mathbf{P}_{12} = \{P_1, \dots, P_s\}$ são os nós parentais compartilhados entre y_1 e y_2 . $\mathbf{P}_1$, $\mathbf{P}_2$ e $\mathbf{P}_{12}$ podem ser constituídos por QTLs e/ou fenótipos causais de y_1 e y_2 . $\mathbf{C}_1 = \{C_{11}, \dots, C_{1u}\}$ e $\mathbf{C}_2 = \{C_{21}, \dots, C_{2v}\}$ são os nós vizinhos exclusivos de y_1 e y_2 , respectivamente; $\mathbf{C}_{12} = \{C_1, \dots, C_t\}$ são os nós vizinhos compartilhados entre y_1 e y_2 . $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_{12}$ somente podem ser constituídos por fenótipos com relação de causalidade ainda não direcionada sobre y_1 e y_2 . Adaptado de Wang e Eeuwijk (2014)

O algoritmo QPSO apresenta a vantagem de não requerer QTLs específicos para cada fenótipo presente no gráfico não direcionado e levar em consideração as interações de causalidade fenotípicas para a orientação de setas. Percebe-se que esta é uma maneira de incluir mais informações no modelo, além daquelas consideradas pelo algoritmo QDG. Wang e Eeuwijk (2014) apresentaram resultados de simulação que demonstraram que o algoritmo QPSO tem uma aplicabilidade mais ampla e fornece orientação de setas mais acuradas que o algoritmo QDG.

Uma questão inerente a todos os algoritmos listados é que o mapeamento de QTL e a inferência do gráfico não direcionado são realizados em passos distintos e a informação de QTLs é utilizada posteriormente apenas para a orientação de setas. No entanto, o mapeamento de QTLs pode ser enviesado pela descoberta de QTLs com efeitos indiretos devido ao desconhecimento do relacionamento causal entre os fenótipos durante o passo de mapeamento de QTLs. Por outro lado, nos algoritmos de descoberta de causalidade, ao não se levar em consideração a existência de QTLs que afetam as características, estes QTLs se comportam como variáveis latentes e há uma clara violação da pressuposição de suficiência causal na inferência do modelo. Por estes motivos, Chaibub-Neto, Keller e Yandell (2010) propuseram o algoritmo QTLnet (*QTL-driven*

phenotype Network) para inferir conjuntamente os parâmetros de mapeamento de QTLs e a rede causal entre os fenótipos. O algoritmo QTLnet condiciona o mapeamento de QTLs na estrutura causal fenotípica, porém como ambos os parâmetros são desconhecidos, o algoritmo itera entre a busca de QTLs e a de redes utilizando metodologia MCMC (*Markov Chain Monte Carlo*) e a rede causal é inferida pela média dos modelos Bayesianos (CHAIBUB-NETO; KELLER; YANDELL, 2010). Claramente, tal abordagem é bastante complexa do ponto de vista estatístico e computacional.

2.5.4 Aplicações dos modelos gráficos na área biológica

Uma limitação das redes Bayesianas é a pressuposição de ausência de alças de retroalimentação, comuns em sistemas biológicos (ROCKMAN, 2008). Apesar desta desvantagem, as redes Bayesianas têm sido empregadas com sucesso na modelagem de sistemas biológicos complexos.

Friedman et al. (2000) foram os primeiros autores a propor o uso de redes Bayesianas para o estudo de dados de expressão gênica. Utilizando a propriedade de independência condicional, os autores conseguiram inferir rede causal entre genes relacionados com a evolução do ciclo celular em leveduras. Pe'er et al. (2001) utilizaram a abordagem proposta por Friedman et al. (2000) em um experimento de expressão gênica com mutantes de leveduras e foram capazes de recuperar redes de interação gênica em rotas metabólicas conhecidas. Esta mesma abordagem foi utilizada por Magwene e Kim (2004) para analisar dados de expressão de, aproximadamente, 5.000 genes de levedura. Neste caso, o espaço amostral de possíveis redes causais foi severamente restrito ao se utilizar apenas independência condicional de primeira ordem (MAGWENE; KIM, 2004). Kulp e Jagalur (2006) estenderam esta abordagem ao incorporar modelos causais que contemplavam a interação entre genes expressos e fenótipos na análise de pares de genes reguladores/alvo. Os autores foram capazes de distinguir entre polimorfismos que afetam a expressão em *cis* (polimorfismo do mesmo gene) e *trans* (polimorfismo em outro gene) (KULP; JAGALUR, 2006).

Schadt et al. (2005) utilizaram o conceito de independência condicional com o objetivo de identificar genes relacionados à obesidade em ratos. Os autores desenvolveram uma metodologia de seleção de modelos de causalidade baseada em ML. Os pesquisadores aplicaram esta metodologia num experimento genômico com dados de eQTLs ("*Expressed QTLs*") e foram capazes de identificar genes com conhecida atuação sobre a obesidade e novos genes provavelmente relacionados ao desenvolvimento da doença (SCHADT et al., 2005). Outros autores aplicaram abordagens semelhantes em diferentes sistemas biológicos (SEGAL et al., 2003; LI

et al., 2005; CHEN; EMMERT-STREIB; STOREY, 2007; ZHU et al., 2007; JANSEN et al., 2009).

Uma característica comum dos sistemas biológicos é a existência de nós mais influentes, que apresentam muitas conexões com os demais integrantes da rede. Duarte e Zeng (2011) propuseram que a descoberta destes genes reguladores chaves fosse realizada para auxiliar na inferência de redes de regulação. Os autores propõe iterar entre a descoberta destes genes reguladores chaves e a construção de redes de eQTLs por meio de testes de independência condicional com os genes regulados. Os autores argumentam que esta metodologia aumenta o poder e diminui o custo computacional na identificação de redes biológicas (DUARTE; ZENG, 2011).

Recentemente, as redes Bayesianas têm se tornado o método padrão de inferência de redes em biologia de sistemas (CHAIBUB-NETO et al., 2013; ZHANG et al., 2013; CAI; BAZERQUE; GIANNAKIS, 2013; BHADRA; MALLICK, 2013; WANG et al., 2013; TUR; ROVERATO; CASTELO, 2014). Excetuando o trabalho de Wang e Eeuwijk (2014), há uma falta de exemplos de aplicação de modelos gráficos na genética e no melhoramento vegetal na literatura científica. No entanto, com o avanço de pesquisas nesta área, espera-se que o uso dos modelos gráficos possa aumentar nossa compreensão da arquitetura genética das características quantitativas.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Dados fenotípicos

Os dados utilizados neste trabalho foram gentilmente cedidos pelo programa de melhoramento genético de milho da Embrapa Milho e Sorgo, Sete Lagoas, Minas Gerais.

Os dados fenotípicos foram obtidos pela avaliação de 145 linhagens recombinantes endogâmicas (RILs - *Recombinant Inbred Lines*) de milho provenientes do cruzamento entre as linhagens L3 e L22 do programa de melhoramento genético de milho da Embrapa Milho e Sorgo. As linhagens L3 e L22 são contrastantes para a eficiência no uso de fósforo (P), sendo a linhagem L3 considerada eficiente e a linhagem L22 ineficiente na utilização deste mineral (PARENTONI; SOUZA-JUNIOR, 2008). As linhagens também são contrastantes para diversos fenótipos relacionados à morfologia do sistema radicular (SOUSA et al., 2011).

O experimento foi conduzido no delineamento em blocos casualizados (quatro dias consecutivos de estabelecimento e avaliação). A parcela foi constituída por três plantas germinadas conjuntamente em papel de germinação. As sementes das 145 RILs e das linhagens L3 e L22 foram germinadas e, após quatro dias, três plântulas uniformes para cada repetição foram transferidas para papel de germinação, após a retirada do endosperma de cada plântula para eliminação das reservas das sementes. O sistema de germinação foi mantido em solução nutritiva com concentração limitante de P (2, $5\mu M$), por 13 dias adicionais. O sistema radicular foi então separado da parte aérea e as imagens das raízes de cada plântula foram digitalizadas e analisadas com programas específicos *RootReader2D* (<http://www.plantmineralnutrition.net/rootreader.htm>) e *WinRHIZO* (http://www.regent.qc.ca/assets/winrhizo_about.html). Todo procedimento foi realizado de acordo com Sousa et al. (2011). As características avaliadas foram:

- Diâmetro médio das raízes (mm) - **RD**;
- Área de superfície obtida pela projeção do sistema radicular (cm^2) - **SA**;
- Área de superfície obtida pela projeção das raízes finas (diâmetro entre 1,0 e 2,0 mm) do sistema radicular (cm^2) - **SA2**.

O sistema radicular e a parte aérea das três plântulas correspondentes a uma repetição de cada uma das RILs foram secos a 65 °C até atingirem peso constante, para determinação das seguintes características:

- Peso seco da parte aérea (mg) - **SDW**;
- Peso seco do sistema radicular (mg) - **RDW**.

O sistema radicular e a parte aérea das três plântulas correspondentes a uma repetição de cada uma das RILs foram então submetidos a análises laboratoriais de acordo com Sousa et al. (2011) para a determinação da concentração de P, fornecendo os dados das seguintes características:

- Concentração de P da parte aérea (mg P / g planta) - **ShootP**;
- Concentração de P do sistema radicular (mg P / g planta) - **RootP**.

3.1.2 Dados genotípicos

A população de RILs foi genotipada com 60 marcadores microssatélites (SSR - *Single Sequence Repeats*) de acordo com Ninamango-Cárdenas et al. (2003) e 332 marcadores SNPs (*Single Nucleotide Polymorphism*). As sequências e as posições genômicas dos marcadores SSR e SNPs estão disponíveis no *MaizeGBD locus* (http://www.maizegdb.org/data_center/locus).

Devido a interesse específico em mecanismos de aquisição de P em milho, as RILs foram genotipadas com sequências iniciadoras (*primers*) específicas para seis genes de milho que apresentaram homologia com o gene *OsPSTOL1*, reconhecido por estar relacionado ao crescimento de raízes, produtividade e eficiência de aquisição de P em arroz (GAMUYAO et al., 2012).

As RILs também foram genotipadas com marcadores STS (*Sequence Tagged Site*) e CAPS (*Cleaved Amplified Polymorphic Sequences*) desenvolvidos para mapear genes relacionados com a morfologia do sistema radicular de milho: *rth1* (WEN et al., 2005), *Bk2l3* (BRADY et al., 2007) e *Rtcs* (TARAMINO et al., 2007).

3.2 Métodos

3.2.1 Mapa genético

Os marcadores genéticos foram testados em relação à segregação esperada (1:1), usando teste de qui-quadrado ($p < 0,05$), corrigido para múltiplos testes pelo método de Bonferroni.

O mapa de ligação foi construído utilizando o programa MAPMAKER/EXP 3.0 (LANDER et al., 1987), considerando LOD mínimo de 3,0 e frequência de recombinação de 0,4 e a função de mapeamento de Kosambi (KOSAMBI, 1954) para conversão das frequências de recombinação em distâncias genéticas (medidas em cM). A posição dos marcadores genéticos no mapa final obtido foi comparada com a sequência física disponível para o genoma do milho. Adicionalmente, as posições dos marcadores foram confirmadas pela utilização da função `ripple` do pacote `onemap` (MARGARIDO; SOUZA; GARCIA, 2007), com tamanho de janela (*window size*) de 5 marcadores.

3.2.2 Análises fenotípicas

Os valores fenotípicos ajustados para cada característica foram obtidos pelo seguinte modelo estatístico:

$$y_{ij} = \mu + B_i + G_j + \varepsilon_{ij} \quad (12)$$

em que,

y_{ij} é a i -ésima observação fenotípica ($i = 1, 2, \dots, r = 4$) da j -ésima RIL ($j = 1, 2, \dots, 145$);

μ é o intercepto;

B_i é o efeito do i -ésimo bloco;

G_j é o efeito específico da j -ésima RIL;

ε_{ij} é o resíduo associado a y_{ij} .

O ajuste do modelo da Equação 12 foi realizado considerando resíduos independentes e identicamente distribuídos ($\varepsilon_{ij} \sim N(0, \sigma^2)$) e permitindo heterocedasticidade ($\varepsilon_{ij} \sim N(0, \sigma_j^2)$). Ambos modelos foram ajustados usando a função `glm` e comparados usando a função `ANOVA`, ambas do pacote `nlme` do programa R (PINHEIRO et al., 2014). Para cada característica, os valores fenotípicos ajustados foram extraídos do modelo com melhor ajuste.

As herdabilidades no sentido amplo (h^2) ao nível das médias das linhagens foram estimadas de acordo com a Equação 13.

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \frac{\hat{\sigma}_e^2}{r}} \quad (13)$$

em que $\hat{\sigma}_e^2$ é a variância residual estimada pelo quadrado médio dos resíduos (QM_ε) na análise de variância; e $\hat{\sigma}_g^2$ é a variância genética estimada por $(QM_g - QM_\varepsilon)/r$, sendo QM_g o quadrado médio dos genótipos na análise de variância e r o número de repetições.

Foram calculadas correlações de Pearson entre as variáveis estudadas, tanto por meio de seus valores brutos (correlação fenotípica) quanto dos valores ajustados obtidos da Equação 12 (correlações genotípicas).

Devido à grande variabilidade dos valores absolutos das características estudadas, os valores fenotípicos ajustados para cada uma das características avaliadas foram estandardizados:

$$z_i = \frac{y_i - \bar{x}}{s} \quad (14)$$

em que,

z_i é o valor padronizado do i -ésimo valor fenotípico;

y_i é o i -ésimo valor fenotípico ajustado após a realização da ANOVA;

$\bar{x}$ é o valor fenotípico médio populacional;

s é o desvio padrão populacional dos valores fenotípicos.

3.2.3 Mapeamento de QTLs

O mapeamento de QTLs foi realizado utilizando modelo de mapeamento de múltiplos intervalos univariado (MIM - *Multiple Interval Mapping*) e multivariado (MT-MIM - *Multiple Trait Multiple Interval Mapping*). Na abordagem MIM (KAO; ZENG, 1997; KAO; ZENG; TEASDALE, 1999; ZENG; KAO; BASTEN, 1999), o seguinte modelo de mapeamento foi empregado para cada uma das variáveis analisadas:

$$z_i = \mu + \sum_{r=1}^m \beta_r x_{ir} + \varepsilon_i \quad (15)$$

em que,

z_i é o valor fenotípico ajustado do i -ésimo indivíduo ($i = 1, 2, \dots, 145$);

μ é o intercepto;

β_r é o coeficiente da regressão que mede o efeito aditivo do r -ésimo QTL incluído no modelo ($r = 1, 2, \dots, m$);

x_{ir} é uma variável explanatória, definida de acordo com o modelo genético de Cockerham (KAO; ZENG; TEASDALE, 1999; KAO; ZENG, 2002). Para cada indivíduo i da população de RILs, a variável x_{ir} assume valor de 1 ou -1 se o r -ésimo QTL tem genótipo homozigoto dominante (QQ) ou homozigoto recessivo (qq), respectivamente;

ε_i é o resíduo do i -ésimo indivíduo. Assume-se que os resíduos são independentes e identicamente distribuídos ($\varepsilon_i \sim N(0, \sigma^2)$).

O mapeamento de QTLs conjunto para as sete características estudadas, MT-MIM, foi realizado empregando o seguinte modelo de mapeamento (SILVA; ZENG, 2010; SILVA; WANG; ZENG, 2012):

$$z_{ti} = \mu_t + \sum_{r=1}^m \beta_{tr} x_{ir} + \varepsilon_{ti} \quad (16)$$

em que,

z_{ti} é o valor fenotípico do i -ésimo indivíduo ($i = 1, 2, \dots, 145$) para a t -ésima característica ($t = 1, 2, \dots, T$);

μ_t é o intercepto para a t -ésima característica;

β_{tr} é o coeficiente da regressão que mede o efeito aditivo do r -ésimo QTL incluído no modelo ($r = 1, 2, \dots, m$) para a t -ésima característica;

x_{ir} é uma variável explanatória, definida de acordo com o modelo genético de Cockerham (KAO; ZENG; TEASDALE, 1999; KAO; ZENG, 2002). Para cada indivíduo i da população de RILs, a variável x_{ir} assume valor de 1 ou -1 se o r -ésimo QTL tem genótipo homozigoto dominante (QQ) ou homozigoto recessivo (qq), respectivamente;

ε_{ti} é o resíduo do i -ésimo indivíduo na t -ésima característica. Assume-se que os resíduos são independentes e identicamente distribuídos ($\varepsilon_{ti} \sim NMV(0, \Sigma_\varepsilon)$).

Todas as análises de QTL, incluindo as funções `fwdmain` e `backmain` e seus argumentos, foram realizadas utilizando um pacote `Oneqtl` do programa R (versão 3.1.2), que está em desenvolvimento no Laboratório de Genética Estatística do Departamento de Genética da ESALQ e deve ser disponibilizado em breve.

Tanto o mapeamento uni quanto o multivariado foi iniciado com uma busca *forward* para os efeitos principais (aditivos), utilizando a função `fwdmain`, numa grade de posições genômicas de 1 cM (`sparse.grid = 1`) e uma janela de exclusão de busca por QTLs à direita e à esquerda de cada QTL presente no modelo de 5 cM (`wsize = 5`). Os modelos das Equações 15 e 16 foram ajustados usando a aproximação de Haley-Knott (HALEY; KNOTT, 1992) e a significância para inclusão de QTLs no modelo a nível genômico (*genome-wide*) foi estabelecida em 20% (`sig.score = 0.2`) pela metodologia de reamostragem *score* (ZOU et al., 2004; SILVA; ZENG, 2010; LAURIE et al., 2014). As posições e efeitos dos QTLs inclusos no modelo foram reestimadas com um passo de refinamento do modelo a cada três ciclos de busca por QTLs ao longo do genoma (`freq.refine = 3`).

Utilizando esta abordagem, um modelo final foi selecionado e utilizado para calcular a magnitude dos efeitos aditivos para cada QTL, a proporção da variância fenotípica explicada pelos

QTLs encontrados, dada pela razão da soma das variâncias e covariâncias do efeito do QTL sobre a variância fenotípica vezes 100 (%VarCov), o coeficiente de determinação (R^2) dado pela soma das %VarCov de cada QTL e o perfil de LOD ao longo dos cromossomos. Para a obtenção do modelo de mapeamento final, este procedimento foi repetido até que o próximo QTL candidato a entrar no modelo não excedesse o nível de significância pré-estabelecido. Os intervalos de confiança dos QTLs (*LOD support interval*) foram obtidos pelo método de subtração de 1,5LOD, resultando num nível de confiança aproximado de 0,95 (LYNCH; WALSH, 1998).

Para as análises dos algoritmos de causalidade utilizando informações de QTLs, apenas os efeitos marginalmente significantes para cada característica estudada, no modelo de mapeamento de QTLs multivariado, foram considerados. Para tanto, o modelo da Equação 16 obtido pelo procedimento anteriormente descrito foi submetido a um procedimento de eliminação *backward* utilizando a função `backmain`, com um valor de significância a nível genômico de 1% (`sig.score = 0.01`). Como todos os efeitos presentes no modelo já haviam sido previamente testados com proteção para múltiplos testes na função `fwdmain` e o nível de significância para permanência no modelo foi estridente, a eliminação *backward* não considerou nenhuma medida de controle de múltiplos testes (`multi.protect = FALSE`).

3.2.4 Modelo de equações estruturais

O ajuste do modelo de equações estruturais para explicar as relações de causalidade entre as variáveis diâmetro de raízes (RD), área projetada de raízes (SA), área projetada de raízes finas (SA2), peso seco da parte aérea (SDW), peso seco de raízes (RDW), concentração de fósforo na parte aérea (ShootP) e concentração de fósforo nas raízes (RootP) foi realizado levando-se em consideração conhecimento *a priori* a respeito do relacionamento destas variáveis e o auxílio do pacote `sem` (FOX; NIE; BYRNES, 2014).

O conhecimento *a priori* foi obtido junto aos pesquisadores da EMBRAPA Milho e Sorgo, que hipotetizaram que os fenótipos relacionados à morfologia de raiz (RD, SA e SA2) deveriam influenciar os parâmetros de acúmulo de biomassa (SDW e RDW), que por sua vez devem ter um relacionamento com os parâmetros de concentração de fósforo (ShootP e RootP). Também é esperado que os caracteres de raízes (RDW e RootP) influenciem os caracteres de parte aérea (SDW e ShootP).

Não foi considerada a existência de erros de mensuração das variáveis estudadas, pois se utilizou os valores ajustados obtidos por meio das Equações 12 e 14. Portanto, os modelos estruturais propostos não possuíam variáveis latentes. Levando-se em consideração a pressupo-

sição de ausência de variáveis latentes no sistema causal e as informações *a priori*, uma série de modelos estruturais tentativos foram testados utilizando as funções `sem` e `modIndices` do pacote `sem` (FOX; NIE; BYRNES, 2014).

O pacote `sem` usa a notação RAM (*Reticular Action Model*) para especificação dos modelos estruturais. A notação RAM é uma simplificação da notação LISREL mais comumente utilizada e descrita nas Seções 2.4.2 e 2.4.3. O modelo RAM é apresentado na Equação 17:

$$v = Av + u \quad (17)$$

em que,

A representa a matriz dos coeficientes de regressão do modelo estrutural: parâmetros estruturais e fatores de carga (em modelos com erros de mensuração);

v representa o vetor de variáveis exógenas e endógenas e suas variáveis indicadoras (em modelos com erros de mensuração);

u é o vetor que contém variáveis exógenas e suas variáveis indicadoras e os erros do modelo estrutural (FOX, 2006).

Os vetores v e u são sobrepostos para as variáveis exógenas medidas sem erro. A matriz vcov de u é chamada de P . Considerando que as primeiras n variáveis das m variáveis contidas no vetor u são variáveis exógenas observáveis, J é a matriz de seleção:

$$J_{m \times m} = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}$$

A matriz vcov observada, ou seja calculada a partir dos dados, é chamada de S . A matriz vcov implicada pelo modelo estrutural da Equação 17 é chamada de C e pode ser obtida por meio da Equação 18:

$$C = E(Jvv'J') = J(I_m - A)^{-1}P(I_m - A)^{-1'}J' \quad (18)$$

A função de máxima verossimilhança a ser minimizada é dada pela Equação 19 (FOX, 2006) que é análoga à Equação 10 apresentada na Seção 2.4.3.

$$F_{ML} = \log|C| + tr\{SC^{-1}\} - \log|S| - n \quad (19)$$

O ajuste do modelo de equação estrutural obtido por meio da aplicação da função `sem` foi verificado por meio do valor de χ^2 , considerando um nível de significância de 0,05. Adicional-

mente, o pacote `sem` apresenta os seguintes índices que medem a adequabilidade de ajuste do modelo:

- **GFI** (*Goodness-of-fit*) e **AGFI** (*Adjusted Goodness-of-fit*) que são medidas descritivas *ad hoc* comumente empregadas na metodologia dos MEEs. Valores de GFI e AGFI próximos a 1 indicam um bom ajuste de modelo (FOX, 2002).

$$GFI = 1 - \frac{\text{trace}\{[\mathbf{C}^{-1}(\mathbf{S} - \mathbf{C})]^2\}}{\text{trace}[(\mathbf{C}^{-1}\mathbf{S})^2]}$$

$$AGFI = 1 - \frac{n(n+1)}{2 \times df}(1 - GFI)$$

em que df é o número de graus de liberdade do MEE.

- **RMSEA** (*Root mean-squared error approximation*): medida de comparação entre o modelo proposto e um modelo completamente saturado que reproduz fielmente a estrutura *vcov* observada. Valores de RMSEA menores que 0,05 indicam um bom ajuste de modelo (FOX, 2002).

$$RMSEA = \sqrt{\max\left(\frac{F_{ML}}{df} - \frac{1}{N-1}, 0\right)}$$

- **BIC** (*Bayesian Information Criterion*): critério mais utilizado para comparação de modelos por possuir sustentação estatística e, por isso, foi preferido aos critérios *ad hoc* anteriormente apresentados. O BIC ajusta a estatística χ^2 da razão de verossimilhanças (L^2) para o número de parâmetros, o número de variáveis observadas (n) e o tamanho amostral (N) (Equação 3.2.4). Considerou-se que valores negativos de BIC indicam modelos com maior suporte dos dados que o modelo totalmente identificado ($df = 0$), para o qual o BIC é igual a zero (FOX, 2002).

$$BIC = L^2 - df \times \log(n \times N)$$

3.2.5 Algoritmo PC de descoberta de causalidade

O relacionamento causal das 7 variáveis também foi analisado em relação aos padrões de independência condicional, utilizando o conceito de d-separação (SPIRITES; GLYMOUR; SCHEINES, 2000). Para tanto, a matriz de correlação foi submetida ao algoritmo de Peter-Clark (PC) com o auxílio da função `pc` do pacote `pcalg` (KALISCH; COLOMBO; MAATHUIS, 2012).

Como todas as variáveis ajustadas e padronizadas apresentaram distribuição normal, os testes de verificação de independência condicional (ou de correlação parcial diferente de 0) para cada variável foram realizados com o auxílio da função `gaussCItest` como parâmetro do argumento `indepTest` da função `pc`.

O número de testes de independência condicional realizados pelo algoritmo PC depende do nível de significância empregado nos testes de independência condicionais individuais, tornando difícil fixar um nível global de significância *a priori* para o controle de múltiplos testes do algoritmo. Kalisch, Colombo e Maathuis (2012) sugerem que o parâmetro `alpha` seja tratado como parâmetro de ajuste do algoritmo. Assim, para o estabelecimento do nível global de significância do algoritmo PC, os dados foram analisados variando o parâmetro `alpha` da função `pc` entre 0,05 e 0,0002.

O resultado do algoritmo PC pode ser dependente da ordem de entrada dos dados nos passos de descoberta de setas. Por este motivo, Colombo e Maathuis (2014) desenvolveram modificações no algoritmo, tornando-o mais estável (*PC-stable*). Kalisch, Colombo e Maathuis (2012) afirmam que a função `pc` do pacote `pcalg` já possui esta modificação implementada. Para verificar se o resultado da função era independente de ordem de entrada, a ordem das colunas da matriz de correlação utilizada como argumento da função `pc` foi permutada com o auxílio da função `sample` do pacote básico do programa R e os resultados obtidos com as diferentes ordens de entradas foram comparados posteriormente para confirmar a estabilidade do algoritmo.

O resultado do passo de orientação de setas também pode ser dependente de ordem de entrada dos dados. Para lidar com esta situação, foi utilizada a versão conservativa da função `pc`, com o argumento `conservative = TRUE`. Esta versão verifica a orientação das setas de um trio de variáveis adjacentes seguindo a regra de orientação dos trios estritamente. Um determinado trio de variáveis adjacentes somente é identificado como uma estrutura em *v* caso os testes de independência condicional não apresentem nenhuma ambiguidade. Caso contrário, o trio é marcado como *ambíguo* e nenhuma ação é tomada em relação ao mesmo. Kalisch, Colombo e Maathuis (2012) argumentam que esta versão do algoritmo PC é bastante rigorosa com a identificação de estruturas em *v*.

Erros aleatórios ou a presença de variáveis não mensuradas podem levar a informações conflitantes em relação à orientação das setas. Para que o resultado do algoritmo PC não fosse dependente da ordem em que os trios adjacentes fossem visitados, foram utilizados os argumentos `solve.confl = TRUE` e `u2pd = "relaxed"` da função `pc`. Dessa forma, todas as informações conflitantes de trios de variáveis adjacentes não ambíguos são refletidas no resultado final do algoritmo PC (KALISCH; COLOMBO; MAATHUIS, 2012) e algumas setas do

gráfico podem não ser direcionadas.

Utilizando-se todas as definições descritas para a função `pc`, o algoritmo PC pode resultar numa CPDAG (*Completely Partially Directed Acyclic Graph*). Adicionalmente, erros aleatórios, presença de variáveis relevantes não-mensuradas ou ausência de fidedignidade nos dados podem levar a uma CPDAG inválida, o que significa que não existe uma DAG com o mesmo esqueleto e estruturas em v representadas pela DAG resultante do algoritmo. Por exemplo, modelos de causalidade com relacionamentos de retroalimentação não podem ser representados por DAGs. Por este motivo, os dados também foram analisados com o algoritmo PC sem os passos de orientação de setas (Algoritmo Esqueleto de PC - *PC Skeleton Algorithm*).

O algoritmo esqueleto de PC foi implementado com o auxílio da função `skeleton` do pacote `pcalg` (KALISCH; COLOMBO; MAATHUIS, 2012). O algoritmo de esqueleto de PC foi implementado utilizando as mesmas definições descritas acima para o algoritmo PC, exceto para os passos de orientação de setas que não são realizados pelo algoritmo esqueleto de PC.

3.2.6 Algoritmo QDG de descoberta de causalidade com informação de QTL

As informações de causalidade fornecidas pelo mapeamento de QTLs da Seção 3.2.3 foram utilizadas para tentar orientar os relacionamentos causais entre os caracteres fenotípicos utilizando o algoritmo QDG (*QTL-directed Dependency Graph*) (CHAIBUB-NETO et al., 2008).

A implementação do algoritmo QDG foi automatizada utilizando-se a função `qdg` do pacote `qtlnet` (CHAIBUB-NETO; YANDELL, 2014) do programa R. O algoritmo foi executado seguindo as recomendações de Chaibub-Neto et al. (2008):

1. Uma rede não direcionada (UDG) foi construída utilizando o algoritmo esqueleto de PC (argumento `skel.method = "pcskel"` da função `qdg`). Foi utilizado o nível de significância de 0,0025 para a realização dos testes de independência condicionais (argumento `alpha = 0.0025`);
2. Os genótipos dos QTLs encontrados com a aplicação do modelo MT-MIM foram utilizados para direcionar cada seta presente na UDG. Somente os QTLs com efeitos marginais significantes após o passo de eliminação *backward* foram considerados nesta etapa (Seção 3.2.3). O gráfico resultante foi chamado de gráfico direcionado (DG);
3. Uma ordem aleatória de todas as setas presentes na DG foi escolhida. Esta ordem foi chamada de ordem do gráfico direcionado (ODG);

4. Para cada seta na ODG, a ordem da causalidade foi re-calculada utilizando testes de LOD com todos os fenótipos pais e informações dos genótipos dos QTLs para ambos os nós. Em caso de mudança de direção, o DG era atualizado e passava-se à próxima seta na ODG;
5. O passo 4 foi realizado até a estabilização da direção das setas. O resultado foi gravado como uma solução possível. Em caso de ausência de convergência para gráfico resultante único, todos os gráficos resultantes foram gravados como possíveis soluções;
6. Os passos 3, 4 e 5 foram repetidos 1000 vezes (argumento `n.qdg.random.starts = 1000` da função `qdg`). Todos os resultados foram gravados;
7. Todas as soluções encontradas no passo 6 foram ranqueadas utilizando verossimilhança e o gráfico melhor ranqueado foi escolhido como a solução mais provável.

Os valores dos coeficientes estruturais do modelo gráfico resultante do algoritmo QDG foram obtidos com o auxílio da função `qdg.sem`

3.2.7 Algoritmo QPSO de descoberta de causalidade com informação de QTL

As informações de causalidade fornecidas pelo mapeamento de QTLs também foram utilizadas para tentar orientar os relacionamentos causais entre os caracteres fenotípicos utilizando o algoritmo QPSO (*QTL+Phenotype Supervised Orientation*) (WANG; EEUWIJK, 2014). Similarmente ao algoritmo QDG, o algoritmo QPSO baseia-se na quebra de equivalência na distribuição de probabilidade entre as causalidades de um par de variáveis com a adição de informações de QTL. No entanto, o algoritmo QPSO adicionalmente considera as demais relações causais existentes entre os nós parentais e vizinhos ao par de variáveis analisadas.

A implementação do algoritmo QPSO foi automatizada utilizando-se a função `result` de um código fonte do programa MATLAB (MATLAB, 2013) gentilmente fornecido pelo pesquisador Fred A. van Eeuwijk, do Departamento de Ciências Agrárias da Universidade de Wageningen, Holanda.

A função `result` requer uma matriz de adjacências (`init_G`) pré-estimada contendo informações a respeito do relacionamento causal entre os fenótipos e QTLs identificados. O relacionamento fenotípico entre os fenótipos da matriz `init_G` foi estimado com o algoritmo esqueleto de PC. As informações de causalidade de QTLs foram fornecidas pelo mapeamento de QTLs. Adicionalmente, considerou-se que os dados fenotípicos e que os valores genotípicos dos QTLs apresentaram distribuição contínua.

O algoritmo foi executado seguindo as recomendações de Wang e Eeuwijk (2014):

1. Um par de fenótipos conectados por uma seta não direcionada foi escolhido aleatoriamente, respeitando os critérios de que ambos fenótipos possuíam nós parentais e pelo menos um deles apresentava nós parentais exclusivos;
2. A LGPN (*Local Generalized Phenotype Network*) (Figura 6) referente a este par de fenótipos foi extraída;
3. A LODG (*Locally Optimal Directed Graph*) dentre todos os possíveis gráficos direcionados derivados desta LGPN foi identificada com base no LOD. Todas as setas desta LGPN foram reorientadas de acordo com a LODG identificada;
4. Os passos 1, 2 e 3 foram repetidos até esgotar os pares de fenótipos que atendiam as condições requeridas pelo passo 1;
5. Caso a rede fenotípica ainda apresentasse setas não direcionadas, um par de fenótipos conectados por seta não direcionada foi escolhido aleatoriamente desde que pelo menos um dos fenótipos tivesse nós parentais;
6. A LGPN referente a este par de fenótipos foi extraída;
7. A LODG dentre todos os possíveis gráficos direcionados derivados desta LGPN foi selecionada. Todas as setas desta LGPN foram reorientadas de acordo com a LODG identificada;
8. Os passos 5, 6 e 7 foram repetidos até não restar nenhuma seta adicional que pudesse ser orientada. O resultado da orientação global da rede foi guardado;
9. Os passos anteriores foram repetidos 1000 vezes (`niteration = 1000`);
10. Todas as redes resultantes foram avaliadas com base no BIC. A rede com melhor BIC foi considerada como o gráfico causal estimado pelo algoritmo QPSO.

A função `init_G` foi utilizada para calcular o valor de BIC do gráfico causal estimado pelo algoritmo QPSO. Desde que existam nós parentais para pelo menos um dos fenótipos de todos os pares fenotípicos relacionados por relações de causalidade não direcionadas, o algoritmo QPSO é capaz de retornar uma rede causal completamente direcionada.

4 RESULTADOS

4.1 Análises fenotípicas

Todos os fenótipos de morfologia de raiz, acúmulo de biomassa e concentração de fósforo apresentaram variação genotípica significativa na população endogâmica recombinante avaliada (Tabela 1). As herdabilidades estimadas foram altas, variando de 0,64 para a concentração de fósforo na parte aérea (ShootP) a 0,82 para área de superfície radicular (SA) e peso seco da parte aérea (SDW).

Tabela 1 – Estimativas das médias fenotípicas, coeficientes de variação (CV), variância genotípica ($\hat{\sigma}_g^2$), variância residual ($\hat{\sigma}_e^2$) e herdabilidade (h^2) das características **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

Fenótipo	Média	CV (%)	Variância Genotípica ($\hat{\sigma}_g^2$)	Variância Residual ($\hat{\sigma}_e^2$)	Herdabilidade (h^2)
RD - cm ³	0,89	7,44	0,004 13	0,004 36	0,79
SA - cm ²	38,48	23,43	90,395 46	81,321 97	0,82
SA2 - cm ²	14,49	24,39	13,293 19	12,491 06	0,81
SDW - g	0,06	22,23	0,000 17	0,000 15	0,82
RDW - g	0,04	19,51	0,000 06	0,000 06	0,78
ShootP - mg/g	7,71	20,36	1,090 21	2,466 84	0,64
RootP - mg/g	6,76	23,73	2,306 69	2,573 61	0,78

As características apresentaram uma grande variação de escala em seus valores fenotípicos médios e em suas variâncias (Tabela 1). Por este motivo e para facilitar a interpretação das análises multivariadas, as variáveis foram padronizadas de forma a apresentarem valores fenotípicos como desvios da média e variância unitária (Equação 14).

Tanto os coeficientes de variação quanto as herdabilidades encontradas para as variáveis deste trabalho (Tabela 1) são coerentes com os resultados encontrados por Sousa et al. (2011). Altas herdabilidades também foram relatadas por Chen et al. (2008, 2009) e Zhu, Kaeppler e Lynch (2005b) para caracteres fenotípicos relacionados à morfologia de raízes e à absorção de fósforo em milho.

As análises fenotípicas revelaram um padrão concordante entre as correlações fenotípicas e genotípicas estimadas entre as variáveis. Os valores estimados de correlações genotípicas

tenderam a apresentar maiores magnitudes, porém com o mesmo sentido que as correlações fenotípicas (Figuras 7 e 8).

As únicas discordâncias entre as correlações fenotípicas e genotípicas foram entre as correlações estimadas entre diâmetro de raízes (RD) e área projetada de raízes finas (SA2), que apresentaram correlação fenotípica baixa ($r_f = 0,28^{**}$) e correlação genotípica nula ($r_g = 0,03^{ns}$); entre as correlações estimadas entre RD e concentração de P na parte aérea (ShootP), que apresentaram correlação fenotípica nula ($r_f = 0,06^{ns}$) e genotípica moderada ($r_g = 0,50^{**}$); e entre as correlações estimadas entre RD e concentração de P nas raízes (RootP), que apresentaram correlação fenotípica nula ($r_f = 0,09^{ns}$) e genotípica moderada ($r_g = 0,32^{**}$).

Figura 7 – Correlação fenotípica entre as variáveis estudadas: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

Os histogramas e as densidades das distribuições dos valores ajustados das características avaliadas são apresentados na diagonal do correlograma da Figura 8. A inspeção visual dos histogramas e das densidades das distribuições sugerem que não há uma clara violação da pressuposição de distribuição Gaussiana para as características estudadas. Adicionalmente, verificou-se que uma linhagem endogâmica apresentou valor discrepante de concentração de fósforo. A aná-

lise dos dados brutos desta linhagem (RIL 19) revelou que esta linhagem apresentou valores consistentemente altos de concentração de fósforo nas três repetições e, por isto, este valor foi mantido nas análises posteriores.

A correlação genotípica estimada entre as variáveis de morfologia de raízes SA e SA2 foi alta e significativa ($r_g = 0,80^{**}$). Estas variáveis também apresentaram correlações estimadas altas e positivas com características de acumulação de biomassa RDW e SDW, que por sua vez, também apresentaram correlação alta entre si ($r_g = 0,82^{**}$). As variáveis de morfologia de raiz (SA e SA2) e de acumulação de biomassa (RDW e SDW) apresentaram correlações moderadas e negativas com as características de concentração de fósforo (ShootP e RootP), refletindo um efeito de diluição da concentração de fósforo com o aumento da biomassa. As variáveis de biomassa apresentaram correlação positiva mediana para alta ($r_g = 0,70^{**}$) entre si (Figura 8).

Figura 8 – Correlação genotípica entre as variáveis estudadas: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

A variável RD apresentou padrão de correlação distinto das demais variáveis de morfologia de raízes. A correlação estimada de RD com SA foi negativa e mediana ($r_g = -0,49^{**}$) e com SA2 foi inexistente ($r_g = 0,03^{ns}$). As variáveis de acumulação de biomassa apresentaram

correlações negativas e baixas ou inexistentes com RD, enquanto as variáveis de concentração de fósforo apresentaram correlação positiva e mediana com RD (Figura 8).

4.2 Mapa genético

O mapa genético final foi constituído de 10 grupos de ligação que corresponderam aos 10 cromossomos do milho. O maior grupo de ligação correspondeu ao cromossomo 1, com 53 marcadores e 316,19 cM e o menor grupo de ligação correspondeu ao cromossomo 2 com 15 marcadores e 111,99 cM.

O mapa genético total apresentou 293 marcadores que cobriram 1.787,46 cM do genoma do milho, com um intervalo médio de 6,1 cM entre marcadores adjacentes (Figura 9). Este mapa é comparável a outros mapas já publicados em trabalhos de mapeamento de QTLs em milho. Usando 567 marcadores, Chen et al. (2008, 2009) obtiveram um mapa de tamanho similar (1.681,3 cM), porém com menores intervalos médios (3,84 cM). Por outro lado, Zhu, Kaeppler e Lynch (2005a) trabalharam com um mapa genético com menos marcadores (196), e também obtiveram tamanho de genoma similar (1.703 cM) e espaçamento médio de 8,73 cM entre os marcadores.

4.3 Mapeamento de QTLs

Os QTLs identificados neste trabalho foram identificados seguindo a seguinte notação: *T@C_P*, em que T representa o fenótipo no modelo de mapeamento univariado ou é substituído por "multi" no mapeamento multivariado; C representa o cromossomo onde o QTL se localiza; e P representa a posição (em cM) do QTL.

O mapeamento de múltiplos intervalos (MIM) para os fenótipos individuais foi capaz de detectar 3 QTLs afetando o diâmetro de raiz (*RD@1_94,44*, *RD@4_25,00* e *RD@7_76,00*) e 1 QTL afetando cada uma dos seguintes fenótipos: área de superfície de raízes finas (*SA2@10_34,18*), peso seco da parte aérea (*SDW@1_217,00*), concentração de fósforo na parte aérea (*ShootP@1_159,00*) e concentração de fósforo nas raízes (*RootP@1_188,00*) (Tabela 2 e Figura 9).

O QTL *SA2@10_34,18* apresentou o maior efeito aditivo sobre os fenótipos estudados individualmente, com o alelo da linhagem L22 contribuindo com um aumento de 0,393 desvio padrão da média de área de superfície de raízes finas, o que correspondeu a 15,12% da variância deste fenótipo (Tabela 2).

Somente foi possível incluir mais de um QTL no modelo de mapeamento de múltiplos intervalos para o fenótipo diâmetro de raízes (RD). Conjuntamente, os três QTLs com efeitos

sobre RD foram capazes de explicar 25,64% da variação deste fenótipo (Tabela 2).

Figura 9 – Perfis de LOD ao longo dos 10 cromossomos de milho para os mapeamentos de QTL uni e multivariado. **Multi**: Mapeamento multivariado (MT-MIM). Mapeamento univariado (MIM) para os fenótipos: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes. O mapa genético estimado é representado no eixo *x* em cM, pela posição dos marcadores genéticos ilustrada por traços verticais. Triângulos invertidos e segmentos de retas acima do mapa indicam a posição e os intervalos de confiança, respectivamente, dos QTLs incluídos no modelo final selecionado

Pelo nosso conhecimento, este é o primeiro trabalho de mapeamento de caracteres fenotípi-

cos de raízes de milho que emprega o modelo de mapeamento por múltiplos intervalos. Assim, não é possível comparar diretamente os valores de r^2 deste trabalho com valores de r^2 disponíveis na literatura científica. Entretanto, trabalhos de mapeamento de QTLs para características de raízes de milho que utilizaram modelos de mapeamento por intervalo composto (CIM) também relatam QTLs que explicam valores não superiores que 15% da variação fenotípica (ZHU; KAEPLER; LYNCH, 2005b; CHEN et al., 2009, 2008).

De modo geral, os valores de r^2 dos modelos MIM (Tabela 2) foram inferiores aos valores de herdabilidade (h^2) dos caracteres fenotípicos estudados (Tabela 1). Mais uma vez, estes resultados parecem estar em consonância com dados da literatura que também apresentam altas herdabilidades aliadas a baixos coeficientes r^2 para caracteres fenotípicos relacionados ao sistema radicular de milho (ZHU; KAEPLER; LYNCH, 2005b; CHEN et al., 2009, 2008).

O modelo de mapeamento multivariado (MT-MIM) foi capaz de detectar 12 posições genômicas com efeito em, pelo menos, um dos 7 fenótipos avaliados (Tabela 3 e Figura 9). A proporção da variância fenotípica explicada pelo modelo MT-MIM selecionado variou de 9,25% para peso seco de raízes (RDW) a 32,29% para diâmetro de raízes (RD). O QTL *multi@10_35,00* apresentou o efeito aditivo de maior magnitude detectado ($-0,402$ desvios padrões da média) da superfície de raízes finas (SA2), representando 15,29% da variação deste fenótipo. O modelo MT-MIM apresentou valores de r^2 marginais variando de 9,25% (RDW) a 32,29% (RD) (Tabela 3), que foram superiores aos modelos MIMs para os caracteres fenotípicos individualmente. No entanto, assim como a abordagem univariada, o modelo MT-MIM também apresentou valores de r^2 inferiores aos valores de herdabilidade (h^2) dos caracteres fenotípicos estudados.

Como regra geral, as estimativas de LOD foram superiores e os intervalos de confiança foram menores para os QTLs detectados utilizando o abordagem MT-MIM quando comparados aos QTL detectados com o modelo de mapeamento univariado, corroborando a tendência de o modelo MT-MIM apresentar maior poder estatístico quando comparado aos modelos MIM aplicados a cada fenótipo separadamente. Ambas abordagens foram capazes de detectar QTLs em 4 posições genômicas coincidentes, identificada pela sobreposição dos intervalos de suporte dos QTLs: *multi@1_94,44* e *RD@1_94,44*; *multi@1_183,80*, *ShootP@1_159,00* e *RootP@1_188,00*; *multi@7_77,27* e *RD@7_76,00*; e *multi@10_35,00* e *SA2@10_34,18*. Em todas estas regiões coincidentes, as estimativas da posição do QTL encontrado foram mais precisas com o modelo MT-MIM, no sentido de que o modelo MT-MIM forneceu menores intervalos de suporte de LODs.

Oito regiões com efeito em pelo menos uma das características estudadas somente foram reveladas com o emprego da abordagem multivariada: *multi@1_263,00*, *multi@2_72,67*, *multi@3*

Tabela 2 – QTLs detectados usando análise MIM para os caracteres: **RD**: Diâmetro de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

Caráter	QTL ¹	Intervalo ²	LOD	Efeito Aditivo ³	% Var ⁴	R ²
RD	RD@1_94,44	88 - 112	3,80	0,310***	9,60	25,64
	RD@4_25,00	18 - 68	2,74	0,270**	6,83	
	RD@7_76,00	65 - 81	3,95	-0,331***	10,01	
SA2	SA2@10_34,18	26 - 39	5,16	-0,393***	15,12	15,12
SDW	SDW@1_217,00	195,5 - 232	2,58	-0,336***	7,86	7,86
ShootP	ShootP@1_159,00	153 - 191	3,65	0,346***	10,93	10,93
RootP	RootP@1_188,00	173 - 195	3,15	0,331***	9,50	9,50

¹Fenótipo@cromossomo_posição (cM). ²Intervalo de confiança baseado em 1,5LOD. ³Efeito aditivo medido como desvios padrões da média; valores positivos indicam que a linhagem endogâmica L3 carrega o alelo que aumenta a característica; Significância dos efeitos baseada em *p*-valor estimado por reamostragem baseada na estatística *score* (***p* < 0,01; *** *p* < 0,001). ⁴Razão da variância do efeito do QTL sobre a variância fenotípica, vezes 100

_0,00, *multi*@3_83,00, *multi*@3_149,00, *multi*@6_130,54, *multi*@8_56,00 e *multi*@9_40,00, corroborando mais uma vez a superioridade dos modelos MT-MIM na detecção de QTLs.

O perfil de LOD do modelo MT-MIM foi consistentemente superior, no sentido de apresentar maiores valores, em relação aos modelos MIM individuais para os sete fenótipos estudados ao longo de todo o genoma, com exceção de pequenos trechos no início dos cromossomos 4, 5 e 6 (Figura 9). Um QTL para o fenótipo RD, que somente foi identificado com a abordagem MIM, situa-se em um destes trechos: *RD*@4_25,00. Usando simulação, Silva, Wang e Zeng (2012) demonstraram que os modelos MT-MIM podem apresentar menor poder de detecção em regiões com efeito restrito a apenas um subconjunto pequeno das características avaliadas. Esta pode ser uma explicação para a falha do modelo MT-MIM detectar o QTL *RD*@4_25,00, que parece ser específico para o fenótipo diâmetro de raízes.

Um QTL situado no cromossomo 1 (Figura 9), também só foi revelado com a abordagem univariada: *multi*@1_263,00. Diferentemente dos QTLs específicos para RD, o perfil de LOD do modelo MT-MIM foi mais alto que os modelos univariados. No entanto, este pico não foi estatisticamente significativo de acordo com critério baseado na estatística *score*. Embora a explicação proposta por Silva, Wang e Zeng (2012) não possa ser descartada, é importante ressaltar que, neste caso, o modelo MT-MIM acomodou 3 QTLs neste mesmo grupo de ligação: *multi*@1_94,44, *multi*@1_183,80 e *multi*@_263,00 (Figura 9), o que pode ter dificultado a detecção de um quarto QTL neste mesmo grupo de ligação pela abordagem multivariada. De qualquer forma, a estratégia de aliar o mapeamento MT-MIM com os mapeamentos MIM, adotada neste trabalho, permitiu a identificação das regiões referentes aos QTLs *SDW*@1_217,00 e

RD@4_25,00.

Para as análises dos algoritmos de causalidade com informação de QTLs, apenas os efeitos marginais significantes para cada variável analisada foram considerados. A Tabela 4 apresenta os efeitos aditivos marginalmente significantes após o procedimento de eliminação *backward* empregando um nível de significância de 1% ($p = 0,01$). Optou-se pela adoção de um nível de significância restritivo na eliminação *backward* para garantir que os QTLs selecionados tivessem realmente efeitos sobre as variáveis do modelo causal.

O modelo resultante após o procedimento de eliminação *backward* representado na Tabela 4 reteve os mesmos 12 QTLs identificados pelo procedimento MT-MIM ilustrado na Figura 9. No entanto, algumas estimativas de posição e intervalo de confiança dos QTLs apresentaram certa variação (Tabelas 3 e 4). Estas diferenças podem ser explicadas pela re-estimação dos parâmetros do modelo após a eliminação dos efeitos marginalmente não significantes e novo ajuste nas posições. Obviamente, o modelo MT-MIM após a eliminação *backward* (Tabela 4) apresentou valores de r^2 menores que o modelo contendo todos os efeitos marginais, mesmo que não significantes (Tabela 3). As maiores diferenças entre os valores de r^2 foram observadas para os fenótipos RDW e RootP, que retiveram apenas 3 QTLs com efeitos marginalmente significantes e tiveram, grosseiramente, seus valores de r^2 reduzidos à metade do modelo antes da eliminação *backward*. Por outro lado, a característica fenotípica área de superfície apresentou pequena diferença de r^2 entre os dois modelos (redução de 17,87% para 15,35%).

O QTL *multi@1_94,44* foi o que apresentou maior número de efeitos marginalmente significantes sobre as características avaliadas, com efeito marginalmente não significantes apenas para área de superfície de raízes finas (SA2). Por outro lado, o QTL *multi@3_10,00* apresentou efeito marginal específico para concentração de fósforo na parte aérea (ShootP). Dentre as características avaliadas, diâmetro de raízes (RD) foi a que teve a maior variância fenotípica explicada pelos QTLs detectados pelo modelo MT-MIM (25,37%), enquanto o fenótipo matéria seca de raízes (RDW) foi o que apresentou menor variância fenotípica (4,34%) explicada após a eliminação dos efeitos marginalmente não significantes (Tabela 4).

Tabela 3 – QTLs detectados usando análise MT-MIM para as características: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

QTL ¹	Intervalo ²	LOD	Efeito Aditivo ³ (% Var ⁴)						
			RD	SA	SA2	SDW	RDW	ShootP	RootP
multi@1_94,44	89 – 97	6,63	0,258*** (6,59)	-0,147* (2,15)	0,056 (0,32)	-0,204** (4,14)	-0,124 (1,52)	0,215** (4,60)	0,152* (2,31)
multi@1_183,80	178 – 184	8,76	-0,017 (0,03)	0,013 (0,02)	0,052 (0,027)	-0,068 (0,46)	-0,058 (0,33)	0,239** (5,62)	0,295*** (8,57)
multi@1_263,00	260,9 – 268	6,73	0,118 (1,24)	-0,157 (2,22)	-0,132 (1,58)	-0,152 (2,08)	-0,011 (0,01)	-0,031 (0,09)	-0,092 (0,77)
multi@2_72,67	69 – 82	7,18	-0,147* (2,07)	0,146 (2,01)	0,000 (0,00)	0,083 (0,65)	-0,074 (0,52)	-0,077 (0,57)	-0,146* (2,01)
multi@3_0,00	0 – 15	5,59	-0,046 (0,20)	-0,002 (0,00)	-0,097 (0,87)	-0,004 (0,00)	-0,020 (0,04)	0,207** (3,98)	0,141 (1,85)
multi@3_83,00	81 – 84	10,58	-0,164* (2,44)	0,135 (1,66)	0,106 (1,02)	-0,128 (1,49)	0,012 (0,01)	-0,010 (0,01)	-0,128 (1,50)
multi@3_149,00	136 – 165	5,91	-0,115 (1,04)	0,022 (0,04)	0,003 (0,00)	-0,092 (0,66)	0,103 (0,83)	0,063 (0,32)	0,118 (1,11)
multi@6_130,54	127 – 133	8,05	0,175* (3,87)	-0,020 (0,04)	0,185* (3,25)	0,101 (0,96)	0,098 (0,92)	0,011 (0,01)	-0,127 (1,55)
multi@7_77,27	70 – 79	7,38	-0,304*** (9,11)	0,133 (1,74)	0,034 (0,11)	-0,015 (0,02)	0,018 (0,03)	-0,141 (1,96)	-0,034 (0,12)
multi@8_56,00	46 – 65	6,80	0,061 (0,28)	-0,171 (2,16)	-0,156 (1,80)	0,032 (0,07)	-0,159 (1,87)	-0,043 (0,14)	-0,066 (0,32)
multi@9_40,00	33 – 46	7,27	0,000 (0,00)	0,009 (0,01)	-0,130 (1,36)	-0,004 (0,00)	-0,069 (0,38)	0,095 (0,72)	0,158 (1,98)
multi@10_35,00	33 – 39	9,31	-0,10 (0,91)	-0,242** (5,55)	-0,402*** (15,29)	-0,134 (1,71)	-0,181* (3,11)	0,121 (1,39)	0,201* (3,86)
R² ⁵			32,29	17,87	24,49	11,34	9,25	19,03	23,51

¹"multi"@cromossomo_posição (cM). ²Intervalo de confiança baseado em 1,5LOD. ³Efeito aditivo medido como desvio padrões da média; significância dos efeitos baseadas em p-valor estimado por reamostragem baseada na estatística *score* (* $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$). ⁴Razão entre a variância do efeito do QTL sobre a variância fenotípica, vezes 100. ⁵R² calculado com o modelo contendo todos os efeitos marginais, mesmo que não significativos

Tabela 4 – QTLs detectados após eliminação *backward* dos efeitos marginais não significativos do modelo MT-MIM para as características: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes.

QTL ¹	Intervalo ²	LOD	Efeito Aditivo ³ (% Var ⁴)						
			RD	SA	SA2	SDW	RDW	ShootP	RootP
multi@1_94,44	91 – 97	6,60	0,261*** (6,75)	-0,203*** (4,09)	<i>sur</i>	-0,264*** (6,95)	-0,185*** (3,42)	0,232*** (5,34)	0,175** (3,04)
multi@1_183,80	180 – 184	6,76	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>	0,221*** (4,81)	0,318*** (9,94)
multi@1_263,00	260 – 268	4,81	<i>sur</i>	-0,133*** (1,60)	-0,171*** (2,62)	-0,151*** (2,06)	<i>sur</i>	<i>sur</i>	<i>sur</i>
multi@2_74,00	68 – 83	4,71	<i>sur</i>	0,054* (0,27)	<i>sur</i>	<i>sur</i>	-0,121** (1,35)	<i>sur</i>	<i>sur</i>
multi@3_10,00	0 – 23	3,37	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>	0,231*** (3,55)	<i>sur</i>
multi@3_83,00	80 – 84	8,93	-0,197** (3,51)	0,090** (0,70)	<i>sur</i>	-0,158*** (2,62)	<i>sur</i>	<i>sur</i>	<i>sur</i>
multi@3_147,00	128 – 154	5,54	-0,147* (1,81)	<i>sur</i>	<i>sur</i>	-0,129*** (1,38)	<i>sur</i>	<i>sur</i>	<i>sur</i>
multi@6_133,00	124 – 142	4,75	<i>sur</i>	<i>sur</i>	0,114*** (1,15)	<i>sur</i>	<i>sur</i>	<i>sur</i>	-0,204*** (3,70)
multi@7_77,27	69 – 79	7,44	-0,345*** (11,71)	0,138*** (1,87)	<i>sur</i>	<i>sur</i>	<i>sur</i>	-0,150** (2,22)	<i>sur</i>
multi@8_55,00	47 – 63,32	7,01	<i>sur</i>	-0,201*** (2,89)	-0,220*** (3,45)	<i>sur</i>	-0,221*** (3,50)	<i>sur</i>	<i>sur</i>
multi@9_52,78	50 – 61	5,14	<i>sur</i>	<i>sur</i>	-0,128*** (1,59)	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>
multi@10_35,00	30 – 39	6,03	<i>sur</i>	-0,123*** (1,43)	-0,216*** (4,43)	<i>sur</i>	<i>sur</i>	<i>sur</i>	<i>sur</i>
R²			25,37	15,35	18,87	7,61	4,34	17,72	14,13

¹"multi"@cromossomo_posição (cM); ²Intervalo de confiança baseado em 1,5LOD. ³Efeito aditivo medido como desvios padrões da média; Significância dos efeitos baseadas em p-valor estimado por reamostragem baseada na estatística *score* (* $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$). ⁴Razão entre a variância do efeito do QTL sobre a variância fenotípica, vezes 100. *sur*: Efeito não relacionado à regressão (*seemingly unrelated regression*)

4.4 Modelo de equações estruturais

O modelo de equações estruturais, ajustado via máxima verossilhança (Equação 19), com o auxílio de conhecimento prévio e das funções `sem` e `modIndices` do pacote `sem` do R, pode ser visto na Figura 10. O modelo proposto é recursivo, ou seja, não apresentou alças de retroalimentação de relações causais e os parâmetros foram estimados sem necessidade de iteração. A causalidade flui no sentido unidirecional: variáveis de morfologia de raízes (RD, SA e SA2) → variáveis de acumulação de biomassa (SDW e RDW) → variáveis de concentração de fósforo (ShootP e RootP). Não foi possível ajustar um modelo estrutural com causalidade no sentido: variáveis de raízes (RDW e RootP) → variáveis de parte aérea (SDW e ShootP), como anteriormente hipotetizado de acordo com o conhecimento prévio dos pesquisadores. Na realidade, o modelo proposto apresenta sentido de causalidade inverso ao anteriormente hipotetizado pelos especialistas.

Figura 10 – Modelo de equação estrutural relacionando os caracteres: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes. γ 's representam os coeficientes de regressão estruturais relacionando variáveis exógenas a variáveis endógenas; β 's representam os coeficientes de regressão estruturais relacionando duas variáveis endógenas; os subscritos dos coeficientes estruturais referem-se à posição da coluna referente às variáveis relacionadas causalmente na matriz de correlação

$$\left\{ \begin{array}{l} SA = \gamma_{13}RD + \gamma_{32}SA2 + \theta_{22} \\ SDW = \gamma_{14}RD + \beta_{24}SA + \theta_{44} \\ RDW = \gamma_{15}RD + \beta_{25}SA + \beta_{45}SDW + \theta_{55} \\ ShootP = \gamma_{16}RD + \beta_{46}SDW + \theta_{66} \\ RootP = \gamma_{37}SA2 + \beta_{57}RDW + \beta_{67}ShootP + \theta_{77} \end{array} \right. \quad (20)$$

O modelo de equações estruturais proposto na Figura 10 e no sistema de equações estruturais 20 apresentou valor do teste de χ^2 , com 8 graus de liberdade, igual a 4,13 (p -valor = 0,84), indicando um bom ajuste aos dados observados.

Os valores dos índices que medem o ajuste do modelo da Figura 10 são apresentados na Tabela 5. Os valores dos parâmetros *ad hoc* de ajuste do modelo GFI e AGFI foram próximos a 1, indicando bom ajuste do modelo proposto. O valor de RMSEA foi igual a 0, ou seja, o valor mínimo da Equação 3.2.4. Valores baixos de RMSEA (menores que 0,05) indicam que o modelo proposto se ajusta aos dados tão bem quanto o modelo saturado que reproduz a matriz vcov observada. Mais importante, o valor de BIC do modelo proposto foi negativo. Valores negativos de BIC indicam modelos com maior suporte dos dados que o modelo totalmente identificado (com 0 graus de liberdade), para o qual o BIC é igual a zero (FOX, 2002). Raftery (1993) afirma que diferenças em valores de BIC podem ser utilizadas para selecionar MEEs, sendo que uma diferença entre BICs da magnitude de 10 ou mais unidades deve ser considerada como uma “evidência conclusiva” de superioridade de um MEE em comparação a outro.

Tabela 5 – Valores dos índices que medem o ajuste do modelo de equação estrutural. **GFI**: *Goodness-of-Fit Index*; **AGFI**: *Adjusted Goodness-of-Fit Index*; **RMSEA**: *Root Mean-Squared Error Approximation*; **BIC**: *Bayesian Information Criterion*

	GFI	AGFI	RMSEA	BIC
Estimativa	0,992	0,973	0 (NA - 0,06)	-35,68

Tomados em conjunto, os valores de χ^2 e dos parâmetros de ajuste de modelo da Tabela 5 indicam que o modelo estrutural da Figura 10 apresenta bom ajuste ao padrão de correlação observado entre os fenótipos estudados.

As variáveis endógenas do modelo estrutural apresentado na Figura 10 são SA, SDW, RDW, ShootP e RootP. A variação destas variáveis explicadas pelo modelo de equações estruturais proposto é apresentada na Tabela 6. Verifica-se que o modelo estrutural é capaz de explicar proporção considerável da variabilidade observada para todas as variáveis endógenas.

Tabela 6 – Estimativas de r^2 (%) para as variáveis endógenas do modelo de equação estrutural. **SA**: Área projetada de raízes; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes.

	SA	SDW	RDW	ShootP	RootP
r^2	83,90	71,05	75,33	54,31	55,97

As estimativas dos parâmetros do modelo de equações estruturais apresentado na Figura 10 podem ser vistas na Tabela 7. Todos os coeficientes de regressão estruturais ilustrados na Figura 10 são estatisticamente significativos. A magnitude dos coeficientes de regressão estrutural variou de 0,142 (RD $\rightarrow$ SDW) a 0,902 (SA $\rightarrow$ SDW). Três coeficientes estruturais apresentaram sinais negativos: RD $\rightarrow$ SA (-0,504), SA2 $\rightarrow$ RootP (-0,380) e SDW $\rightarrow$ ShootP (-0,564), significando que o aumento de uma unidade na variável independente leva a um decréscimo do valor correspondente (em desvios padrões da média) na variável dependente. Os demais coeficientes apresentaram valores positivos.

Os valores de θ 's da Tabela 7 representam a variância residual das variáveis endógenas no modelo estrutural. Os valores de θ da Tabela 7 refletem os valores de r^2 da Tabela 6, sendo que o fenótipo SA foi o que representou menor valor de variância residual e o fenótipo ShootP foi o que apresentou o maior valor de variância residual.

4.5 Modelos gráficos

4.5.1 Algoritmos de descoberta de causalidade (PC e Esqueleto de PC)

A estrutura causal estimada por meio da aplicação dos algoritmos de Peter-Clark (PC) e esqueleto de PC (SPIRITES; GLYMOUR; SCHEINES, 2000) é apresentada na Figura 11. Ambas as estruturas foram estimadas utilizando-se o nível de significância global = 0,0025. Este nível de significância foi selecionado dentre os testados (0,0002 a 0,05), pois considerou-se que os níveis de significâncias muito restritivos produziram gráficos muito esparsos, sem ligação entre variáveis claramente conectadas. Além disto, o gráfico obtido se manteve estável com o nível de significância variando de 0,0025 a 0,05.

Para o nível de significância de 0,0025, os algoritmos realizaram 156 testes de independências condicionais no passo de descoberta de setas. Como os testes de independências condicionais utilizados pelos algoritmos não podem ser considerados independentes, o nível de significância global para o algoritmo não pode ser estimado trivialmente.

Tabela 7 – Estimativas do parâmetros do modelo de equações estruturais para as variáveis: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

Parâmetro	Relação Estrutural	Efeito	Erro Padrão	Estatística z	p-valor
γ_{13}	$RD \rightarrow SA$	-0,506	0,027	-18,561	$6,70 \times 10^{-77}$
γ_{16}	$RD \rightarrow ShootP$	0,336	0,059	1,277	$1,28 \times 10^{-08}$
γ_{15}	$RD \rightarrow RDW$	0,229	0,049	4,719	$2,37 \times 10^{-06}$
γ_{14}	$RD \rightarrow SDW$	0,142	0,051	2,760	$5,79 \times 10^{-03}$
γ_{32}	$SA2 \rightarrow SA$	0,811	0,027	29,75	$1,76 \times 10^{-194}$
γ_{37}	$SA2 \rightarrow RootP$	-0,380	0,085	-4,478	$7,53 \times 10^{-06}$
β_{24}	$SA \rightarrow SDW$	0,902	0,051	17,599	$2,50 \times 10^{-69}$
β_{25}	$SA \rightarrow RDW$	0,568	0,084	6,760	$1,38 \times 10^{-11}$
β_{45}	$SDW \rightarrow RDW$	0,412	0,077	5,358	$8,39 \times 10^{-08}$
β_{46}	$SDW \rightarrow ShootP$	-0,564	0,059	-9,562	$1,15 \times 10^{-21}$
β_{57}	$RDW \rightarrow RootP$	0,159	0,091	1,745	$8,10 \times 10^{-02}$
β_{67}	$ShootP \rightarrow RootP$	0,628	0,065	9,711	$2,70 \times 10^{-22}$
θ_{22}	$SA \leftrightarrow SA$	0,107	0,013	8,485	$2,15 \times 10^{-17}$
θ_{44}	$SDW \leftrightarrow SDW$	0,289	0,034	8,485	$2,15 \times 10^{-17}$
θ_{55}	$RDW \leftrightarrow RDW$	0,246	0,029	8,485	$2,15 \times 10^{-17}$
θ_{66}	$ShootP \leftrightarrow ShootP$	0,457	0,054	8,485	$2,15 \times 10^{-17}$
θ_{77}	$RootP \leftrightarrow RootP$	0,441	0,052	8,485	$2,15 \times 10^{-17}$

γ 's representam os coeficientes de regressão estruturais entre variáveis exógenas sobre as variáveis endógenas; β 's representam os coeficientes de regressão estruturais entre variáveis endógenas; θ 's indicam os valores estimados de variância das variáveis endógenas

Figura 11 – Modelos gráficos obtidos pela utilização dos algoritmos de **A**: Peter-Clark (PC) e **B**: Esqueleto de PC. Variáveis: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

O resultado apresentado na Figura 11 foi consistente para as diversas ordens de entrada nos algoritmos PC e esqueleto de PC (dados não apresentados). Entretanto, o algoritmo de PC não foi capaz de estimar uma DAG única para os dados analisados. Na realidade, dentre os oito relacionamentos de causalidade encontrados apenas quatro setas foram orientadas, resultando numa CPDAG (*Completely Partially Directed Acyclic Graphic*) (Figura 11 A). Este resultado não é incomum quando o algoritmo PC é aplicado, uma vez que é raro conseguir estimar uma DAG única utilizando-se dados observacionais e, normalmente, apenas uma classe de equivalência de modelos causais representada pela CPDAG é passível de ser inferida.

Observando a CPDAG da Figura 11 A, é possível hipotetizar que a violação de ausência de retroalimentação causal no sistema estudado tenha sido violada pois é possível visualizar dois ciclos de causalidade potencialmente presentes na CPDAG: $SA2 \rightarrow SA \rightarrow SDW \rightarrow RDW \rightarrow SA2$ e $RD \rightarrow SA \rightarrow SDW \rightarrow ShootP \rightarrow RD$.

Embora a presença de ciclos de retroalimentação no sistema não possa ser descartada, erros aleatórios e a presença de variáveis latentes não mensuradas, que levam à violação de pressuposição de suficiência causal, também podem ser motivos para a não definição de uma DAG única pelo algoritmo PC. Adicionalmente, se o verdadeiro modelo causal estudado apresentar equivalência Markoviana com outros modelos causais, o algoritmo PC não será capaz de retornar um modelo causal único, mesmo na ausência de qualquer violação de pressuposições.

O uso do argumento `conservative = TRUE` permitiu a identificação de trios de variáveis adjacentes ambíguos que justificam a dificuldade do algoritmo PC em orientar todas as relações de causalidade encontradas. No caso do modelo gráfico da Figura 11 A, os seguintes trios foram marcados como ambíguos: (ShootP – RD – SA2), (SA2 – SDW – ShootP), (SA2 – RDW – SDW), (RD – ShootP – SDW), (RD – ShootP – RootP) e (SDW – ShootP – RootP).

Apesar da dificuldade de orientação de setas nos trios de variáveis descritos acima, o algoritmo PC foi capaz de orientar 4 setas que identificam 2 trios de variáveis adjacentes ligadas pelas mesmas relações de causalidade também identificados pela abordagem de Modelos de Equações Estruturais: $RD \rightarrow SA \leftarrow SA2$ e $SA \rightarrow SDW \rightarrow RDW$ (Figuras 10 e 11 A).

4.5.2 Algoritmos de descoberta de causalidade com auxílio de informações de QTLs (QDG e QSP0)

A aplicação do algoritmo QDG (CHAIBUB-NETO et al., 2008) resultou em duas possíveis soluções diferentes (gráficos direcionados) representadas na Figura 12. O algoritmo QPSO (WANG; EEUWIJK, 2014) resultou em apenas um gráfico direcionado, representado na Figura

12 **A.** A Figura 12 não apresenta QTLs, para facilitar a visualização do relacionamento causal entre os fenótipos.

Interessantemente, o Gráfico Direcionado **A**, encontrado pela aplicação do algoritmo QPSO, foi revelado pelo algoritmo QDG como a segunda melhor solução para os dados, ou seja, os dois algoritmos não foram concordantes no resultado da busca de melhor gráfico direcionado para os dados, muito embora os resultados sejam próximos. Os índices de ajuste de modelo baseados em máxima verossimilhança (LogLike e BIC) indicam que o Gráfico Direcionado **B** apresenta melhor ajuste aos dados dentre os dois possíveis modelos de causalidade. No entanto, as técnicas de modelagem dos modelos gráficos empregadas por ambos os algoritmos pressupõem que os modelos causais estudados possam ser representados por uma DAG, o que não pode ser considerado verdadeiro para ambos os Gráficos Direcionados encontrados pelos algoritmos QDG e QPSO (Figura 12). Por este motivo, os dois gráficos foram avaliados com técnicas de Modelos de Equações Estruturais, que permitem a existência de ciclos de retroalimentação nos sistemas causais. Esta análise forneceu valores de BIC dos Modelos **A** e **B** de 114,21 e 7,22, respectivamente, indicando mais uma vez a superioridade do modelo identificado pelo algoritmo QDG.

A análise das setas direcionadas presentes no Gráfico **B** revelou que este modelo de causalidade mantém as mesmas direções das quatro setas direcionadas encontrado pelo algoritmo de PC: $RD \rightarrow SA \leftarrow SA2$ e $SA \rightarrow SDW \rightarrow RDW$ (Figuras 11 **A** e 12 **B**). Adicionalmente, o modelo **B** manteve o padrão de colisores encontrados pelo algoritmo PC, sendo que apenas o trio $RD \rightarrow SA \leftarrow SA2$ foi identificado como colisor no modelo causal encontrado por ambos os algoritmos (Figuras 11 **A** e 12 **B**). Assim, o modelo o Gráfico Direcionado **B** mostrou-se concordante com o padrão de independências condicionais encontrado pelo algoritmo de PC. Estas razões justificam a seleção do Gráfico Direcionado **B** como o melhor gráfico direcionado para os fenótipos avaliados neste trabalho.

Como mencionado, o modelo **B** representa claramente uma violação da pressuposição de aciclicidade, pois apresenta dois ciclos vizinhos de retroalimentação causais: $RD \rightarrow SA \rightarrow SDW \rightarrow ShootP \rightarrow RD$ e $SA2 \rightarrow SA \rightarrow SDW \rightarrow RDW \rightarrow SA2$. Os coeficientes estruturais do Gráfico Direcionado **B** foram estimados via técnicas de Modelos de Equações Estruturais, pois esta metodologia apresenta capacidade de lidar com modelos não-recursivos (Tabela 8).

A relação de causalidade $SA \leftarrow SA2$ foi a que apresentou maior coeficiente estrutural, indicando que a área de superfície de raízes finas é um componente importante da área total de superfície de raízes. Além disto, a relação de causalidade $RD \rightarrow SA$ apresentou coeficiente estrutural negativo, refletindo o fato de que raízes finas constribuem mais com a área de superfície

de raízes. A relação $SDW \rightarrow ShootP$ também apresentou valor negativo, refletindo o efeito de diluição, ou seja, a tendência da concentração de fósforo ($ShootP$) diminuir com o incremento do peso seco nas raízes (SDW).

Figura 12 – Modelos gráficos obtidos pela utilização do algoritmo QPSO (**A**) e QDG (**A** e **B**) para representar os relacionamentos fenotípicos causais entre as variáveis: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes. Os QTLs utilizados para a orientação das setas não estão ilustrados para facilitar visualização

As Figuras 13, 14 e 15 ilustram o relacionamento dos QTLs encontrados com o modelo de mapeamento de QTL MT-MIM (Seção 4.3) utilizados pelo algoritmo QDG para a descoberta do Gráfico Direcionado **B**. Como regra geral, os QTLs apresentaram efeitos detectáveis sobre variáveis adjacentes e em sequência concordante com o Gráfico Direcionado **B**. Dos 12 QTLs encontrados, 7 QTLs seguiram este padrão: *multi@1_94,44*, *multi@1_183,00*, *multi@1_263,00*, *multi@3_83,00*, *multi@7_77,27*, *multi@8_55,00* e *multi@10_35,00*. Apenas dois QTLs apresentaram efeito sobre apenas uma variável: *multi@3_10,00* e *multi@9_52,78* (Figuras 14 e 15).

Somente três QTLs mostraram efeitos em apenas duas variáveis da extremidade de um relacionamento causal predito pelo Gráfico Direcionado **B**. Destes, dois QTLs apresentaram efeito em apenas 2 variáveis das extremidades de um trio de variáveis adjacentes em sequência concordante de causalidade. O QTL *multi@2_74,00* não apresentou efeito sobre a variável SDW na rota causal: $SA \rightarrow SDW \rightarrow RDW$ (Figura 13 D). O QTL *multi@3_147,00* não apresentou

Tabela 8 – Estimativas do parâmetros estruturais do gráfico direcionado obtido por meio do algoritmo QDG. Variáveis: **RD**: Diâmetro de raízes; **SA**: Área projetada de raízes; **SA2**: Área projetada de raízes finas; **SDW**: Peso seco da parte aérea; **RDW**: Peso seco de raízes; **ShootP**: Concentração de fósforo na parte aérea; **RootP**: Concentração de fósforo nas raízes

Parâmetro	Relação Estrutural	Efeito	Erro Padrão	Estatística z	p-valor
β_{12}	RD $\rightarrow$ SA	-0,488	0,030	-16,881	$6,19 \times 10^{-64}$
β_{32}	SA2 $\rightarrow$ SA	0,776	0,029	26,745	$1,40 \times 10^{-157}$
β_{24}	SA $\rightarrow$ SDW	0,676	0,049	13,690	$1,16 \times 10^{-42}$
β_{45}	SDW $\rightarrow$ RDW	0,695	0,051	13,573	$5,76 \times 10^{-42}$
β_{53}	RDW $\rightarrow$ SA2	0,512	0,069	7,402	$1,35 \times 10^{-13}$
β_{46}	SDW $\rightarrow$ ShootP	-0,590	0,063	-9,392	$5,87 \times 10^{-21}$
β_{61}	ShootP $\rightarrow$ RD	0,300	0,081	3,695	$2,20 \times 10^{-04}$
β_{67}	ShootP $\rightarrow$ RootP	0,696	0,059	11,856	$1,20 \times 10^{-32}$
θ_{11}	RD $\leftrightarrow$ RD	0,788	0,099	7,998	$1,27 \times 10^{-15}$
θ_{33}	SA2 $\leftrightarrow$ SA2	0,494	0,067	7,384	$1,53 \times 10^{-13}$
θ_{22}	SA $\leftrightarrow$ SA	0,108	0,013	8,385	$5,06 \times 10^{-17}$
θ_{44}	SDW $\leftrightarrow$ SDW	0,330	0,042	7,874	$3,44 \times 10^{-15}$
θ_{55}	RDW $\leftrightarrow$ RDW	0,346	0,043	8,108	$5,15 \times 10^{-16}$
θ_{66}	ShootP $\leftrightarrow$ ShootP	0,565	0,067	8,404	$4,32 \times 10^{-17}$
θ_{77}	RootP $\leftrightarrow$ RootP	0,515	0,061	8,485	$2,16 \times 10^{-17}$

β 's representam os coeficientes de regressão estruturais entre variáveis endógenas; θ 's indicam os valores estimados de variância das variáveis endógenas.

efeito sobre a variável SA na rota causal: RD $\rightarrow$ SA $\rightarrow$ SDW (Figura 14 C).

O QTL *multi@6_133,00* foi o único QTL a apresentar padrão bastante discordante do Gráfico Direcionado **B** (Figura 14 D). Este QTL apresentou efeitos sobre as variáveis SA2 e RootP, que não apresentam relações de causalidade diretas entre si e estão nos extremos do Gráfico Direcionado **B**. A análise do perfil de LOD dos mapeamentos MIM para as características SA2 e RootP revela que ambas características apresentam picos de LOD próximos, porém não concordantes, no final do cromossomo 6 (Figura 9). Não obstante, o modelo MT-MIM revelou apenas um QTL com efeito sobre ambas as características nesta região. Isto pode ser um artifício do mapeamento MT-MIM, pois foi fixado uma janela de exclusão de QTLs de 10cM entre os QTLs já presentes no modelo. Em análises onde a janela de exclusão foi fixada em 5 cM, o modelo MT-MIM detecta QTLs distintos para cada variável (resultados não mostrados). Assim, é possível que o QTL *multi@6_133,00* represente duas regiões genômicas diferentes.

Figura 13 – Representação dos QTLs **A:** *multi@1_94,44*, **B:** *multi@1_183,80*, **C:** *multi@1_263,00* e **D:** *multi@2_74,00* no gráfico direcionado revelado pelo algoritmo QDG. **RD:** Diâmetro de raízes; **SA:** Área projetada de raízes; **SA2:** Área projetada de raízes finas; **SDW:** Peso seco da parte aérea; **RDW:** Peso seco de raízes; **ShootP:** Concentração de fósforo na parte aérea; **RootP:** Concentração de fósforo nas raízes

Figura 14 – Representação dos QTLs **A:** *multi@3_10,00*, **B:** *multi@3_83,00*, **C:** *multi@3_147,00* e **D:** *multi@6_133,00* no gráfico direcionado revelado pelo algoritmo QDG. **RD:** Diâmetro de raízes; **SA:** Área projetada de raízes; **SA2:** Área projetada de raízes finas; **SDW:** Peso seco da parte aérea; **RDW:** Peso seco de raízes; **ShootP:** Concentração de fósforo na parte aérea; **RootP:** Concentração de fósforo nas raízes

Figura 15 – Representação dos QTLs A: *multi@7_77,27*, B: *multi@8_55,00*, C: *multi@9_52,78* e D: *multi@10_35,00* no gráfico direcionado revelado pelo algoritmo QDG. RD: Diâmetro de raízes; SA: Área projetada de raízes; SA2: Área projetada de raízes finas; SDW: Peso seco da parte aérea; RDW: Peso seco de raízes; ShootP: Concentração de fósforo na parte aérea; RootP: Concentração de fósforo nas raízes

5 DISCUSSÃO

Desde o surgimento dos métodos de genética quantitativa, a associação (correlação) entre caracteres fenotípicos tem sido alvo de pesquisas. No início, apenas medidas fenotípicas eram utilizadas e, posteriormente, informações de QTLs passaram a ser incorporadas nestes estudos.

Frequentemente, os estudos de mapeamento de QTLs são realizados para conjuntos de fenótipos que apresentam padrões de associação entre si. Recentemente, foram desenvolvidas metodologias de mapeamento de QTLs multivariadas que se beneficiam da utilização destes padrões de associações fenotípicas, representado pela matriz de covariâncias, para melhorar as estimativas das posições e efeitos dos QTLs detectados (SILVA; ZENG, 2010; SILVA; WANG; ZENG, 2012). No entanto, embora se saiba que estes padrões de associação possam refletir relações de causalidade entre os fenótipos, raramente estas relações de causalidade são levadas em consideração nas análises de QTLs (CHAIBUB-NETO et al., 2008; MI et al., 2010a; WANG; EEUWIJK, 2014).

O presente trabalho utilizou técnicas de mapeamento de QTLs e de modelagem de causalidade em um conjunto de fenótipos relacionados à morfologia de raízes, acumulação de biomassa e concentração de fósforo em milho, visando a melhor compreensão da arquitetura genética subjacente a estes fenótipos. Um melhor entendimento desta arquitetura é de grande utilidade para o melhoramento.

Boa parte da variância observada para as características deste trabalho pôde ser atribuída a fatores genéticos (Tabela 1), com herdabilidades de altas magnitudes. Além disso, as características também apresentaram claro padrão de correlação genética entre si (Figura 8), tornando o conjunto de dados estudados um ótimo cenário para análise com as técnicas de modelagem de causalidade empregadas neste trabalho. No entanto, os dados fenotípicos foram obtidos numa única mensuração, representando plântulas avaliadas aos 17 dias de germinação. A ausência de longitudinalidade (mensurações ao longo do tempo) representa uma complicação para a análise de causalidade pois pode levar ao aparecimento de gráficos cíclicos que, de outra forma, poderiam ser modelados como espirais no tempo (SHIPLEY, 2002). No entanto, pode-se considerar que, ao arbitrariamente mensurar num tempo específico uma variável que se desenvolve ao longo do tempo, têm-se um saldo líquido do conjunto de influências ao longo deste tempo e, conseqüentemente, a variável pode ser modelada como um somatório das alças de retroalimentação que a geraram (MARUYAMA, 1997).

Outra questão que merece ser enfatizada é que as variáveis deste estudo foram mensuradas

diretamente e, portanto, foram desconsiderados erros de mensuração que poderiam ter sido acomodados sem muito esforço nos modelos de equações estruturais (MEEs) aqui empregados. É importante ressaltar que a desconsideração de erros de mensuração pode aumentar a probabilidade de rejeição de modelos causais corretos (Erro Tipo I) (SHIPLEY, 2002).

Os resultados da análise de mapeamento de QTLs revelaram que as características estudadas claramente apresentam um controle genético complexo. Notavelmente, os QTLs encontrados, tanto pela abordagem univariada quanto pela abordagem multivariada, apresentaram pequenos efeitos sobre a variância das características (Tabelas 2 e 3 e Figura 9). O QTL de maior efeito foi encontrado tanto na abordagem univariada (*SA2@10_34,18*) quanto na multivariada (*multi@10_35,00*), explicando 15,12% e 15,29%, respectivamente, da variação de área de superfície de raízes finas.

Apesar das boas herdabilidades encontradas para os fenótipos estudados (Tabela 1), os modelos de mapeamento de múltiplos intervalos foram capazes de explicar apenas uma pequena parte da variação fenotípica observada, com valores de r^2 variando de 7,86 a 25,64 nos mapeamentos univariados e de 9,25 a 32,29 no mapeamento multivariado (Tabelas 2 e 3). Uma parte da diferença entre os valores estimados de h^2 e r^2 pode ser explicada pelo fato de este trabalho ter focado na busca de efeitos principais (aditivos) dos QTLs que poderiam ser utilizados subsequentemente na modelagem da causalidade. Assim, os efeitos de interação entre QTLs (epistasia) não foram incluídos no modelo de mapeamento final, o que fatalmente aumentaria os valores de r^2 relatados aqui.

Não foi possível encontrar na literatura estudos de mapeamento de QTLs em milho para caracteres semelhantes aos estudados neste trabalho que é, provavelmente, o primeiro mapeamento de múltiplos intervalos para estes fenótipos nesta cultura. Assim, os valores de r^2 estimados não puderam ser comparados com dados similares publicados. No entanto, estudos de mapeamento de QTLs que empregaram mapeamento por intervalo composto (CIM) para características de morfologia de raízes, aquisição de fósforo e acumulação de biomassa em milho revelaram QTLs com efeitos de magnitudes similares aos encontrados utilizando as abordagens de mapeamento deste trabalho (ZHU; KAEPLER; LYNCH, 2005b, 2005a; ZHU et al., 2006; CHEN et al., 2008, 2009; AZEVEDO et al., 2015). Estes mesmos trabalhos também relatam altas herdabilidades para os caracteres fenotípicos estudados. Dessa forma, é razoável assumir que as características estudadas de fato apresentam arquitetura genética complexa, com vários locos de pequeno efeito individual. Este tipo de arquitetura genética é comumente encontrada em características morfológicas que tendem a apresentar, concomitantemente, altas herdabilidades e herança poligênica (VISSCHER; HILL; WRAY, 2008).

As metodologias de análise de causalidade empregadas neste trabalho se dividiram em duas classes principais: as que não se basearam em informações genotípicas para a orientação de setas (modelagem de equações estruturais (MEEs) e o algoritmo de PC); e os algoritmos que utilizam das informações de QTL para orientar setas, como os algoritmos QDG (*QTL-directed dependency graph*) e QPSO (*QTL+phenotype supervised orientation*).

A modelagem de equações estruturais é bastante flexível, permitindo o estudo de modelos com variáveis latentes, com erros de mensuração e com ciclos de retroalimentação causal. Assim, uma vez que a estrutura causal é conhecida, o teste do modelo causal e as estimativas nos parâmetros do modelo é trivial. O grande desafio deste tipo de modelagem é a descoberta do modelo causal, ou de um conjunto reduzido de modelos causais para serem testados via MEEs, uma vez que o espaço de modelos possíveis cresce exponencialmente com o número de variáveis estudadas (SHIPLEY, 2002). Considerando que, entre um par de variáveis, quatro relações de causalidade são possíveis ($x \rightarrow y$, $x \leftarrow y$, $x \leftrightarrow y$ e ausência de relacionamento causal entre as variáveis x e y), no caso deste trabalho, com sete variáveis, o número de modelos causais possíveis é de $4,398047 \times 10^{12}$, o que torna a estratégia de teste e descarte de modelos causais aleatórios inexequível. Neste cenário, o conhecimento prévio dos pesquisadores da área é imprescindível para o descarte de modelos causais improváveis e a eleição de modelos que valem a pena ser testados. Ademais, isto permite a incorporação de informações *a priori*, algo cada vez mais comum na ciência moderna.

O modelo de equações estruturais proposto neste trabalho foi obtido utilizando conhecimento prévio e o auxílio da função `mod.indices` do pacote `sem`, num processo iterativo de tentativa e erro. A função `mod.indices` verifica as maiores distorções nas matrizes dos parâmetros estruturais (matriz A) e de covariâncias (matriz P) (FOX, 2006) e indica modificações no modelo causal proposto, de forma a minimizar as diferenças entre as correlações observadas e as previstas pelo modelo causal. Assim, iniciando-se com um modelo proposto por conhecimento prévio, foi possível melhorá-lo até a obtenção de um modelo causal plausível. É importante ressaltar que o modelo causal inicialmente proposto não continha alças de retroalimentações causais, pois embora fosse possível imaginar que os fenótipos estudados pudessem apresentar tal tipo de relacionamento, o conhecimento causal disponível era limitado; assim, optou-se arbitrariamente por trabalhar com modelos causais mais simples.

A metodologia dos MEEs verifica se as discrepâncias entre as correlações observadas e as previstas pelo modelo causal testado são pequenas e podem ser atribuídas à variação aleatória. Em caso afirmativo, o modelo proposto não é rejeitado. Assim, é possível que vários modelos causais sejam compatíveis com os dados observados e os MEEs não sejam capazes de distinguir

entre este conjunto de modelos equivalentes (SHIPLEY, 2002). Por este motivo, a despeito de seus bons índices de ajuste, o modelo causal proposto neste trabalho (Figura 10) deve ser visto com a cautela recomendada para a inferência de causalidade em dados observacionais. Ainda assim, o modelo causal proposto apresenta hipóteses para serem testadas por meio de posterior experimentação.

Vários algoritmos de descoberta de casualidade baseados no conceito de d-separação foram propostos para automatizar e tornar menos subjetiva a descoberta de modelos causais para serem testados durante a modelagem de causalidade. Dentre os algoritmos propostos, o algoritmo de PC se destaca por ser computacionalmente muito eficiente (SHIPLEY, 2002; COLOMBO; MAATHUIS, 2014; SHALIZI, 2014) e foi escolhido como algoritmo padrão para a descoberta de causalidade neste trabalho. A análise dos dados com este algoritmo apresentou um resultado bastante distinto da modelagem com os MEEs, revelando um padrão de relacionamento causal mais complexo, com dois potenciais ciclos de retroalimentação causal.

O algoritmo de PC falhou na orientação de metade das oito setas descobertas, resultando num gráfico parcialmente direcionado (CPDAG), representando um conjunto de 16 possíveis modelos causais. A falha na orientação de algumas setas pelo algoritmo PC não é incomum e vários motivos podem ter contribuído para esta dificuldade de orientação das setas encontradas. No gráfico proposto pelo algoritmo PC, foi descoberto apenas um trio de variáveis que constituem claramente uma estrutura em v: $RD \rightarrow SA \leftarrow SA2$. Neste caso, o algoritmo orientou as setas de forma a permitir a constituição deste colisor. Porém, outros trios de variáveis foram marcados como ambíguos e, neste caso, o algoritmo não pode orientar as setas. Outra razão para a manutenção de setas bidirecionais na CPDAG final é que o algoritmo não permite a constituição de ciclos de retroalimentação causais. Tomando por exemplo a seta bidirecionada $RDW \leftrightarrow SA2$ mantida na Figura 11 A, a orientação $RDW \rightarrow SA2$ levaria à formação do ciclo $SA2 \rightarrow SA \rightarrow SDW \rightarrow RDW \rightarrow SA2$, o que violaria a pressuposição de aciclicidade. Por outro lado, a orientação $RDW \leftarrow SA2$ levaria à constituição do colisor $SDW \rightarrow RDW \leftarrow SA2$, que não foi identificado sem ambiguidade pelo algoritmo. Dessa forma, o algoritmo deixou a dificuldade de orientação desta seta explícita na CPDAG final.

Embora o conceito de d-separação tenha sido desenvolvido para gráficos acíclicos (DAGs), Spirtes (1995) demonstrou que a d-separação em gráficos cíclicos ainda implica em independência na distribuição de probabilidade conjunta que o gráfico gera, desde que a relação entre as variáveis seja linear. Porém, o mesmo não ocorre quando o relacionamento é não linear e as variáveis apresentam distribuição contínua. Chaibub-Neto et al. (2008) demonstraram por simulações que o algoritmo PC é capaz de recuperar grande parte da topologia de gráficos causais

cíclicos, apresentando certa robustez à violação da pressuposição de aciclicidade. No entanto, é importante ressaltar que modelagem de causalidade em gráficos cíclicos é mais complicada e requer pressuposições adicionais que o estudo de DAGs. É também razoável pensar que estes ciclos de retroalimentação causais potenciais tenham sido gerados devido à falta de medidas repetidas no tempo para o conjunto de variáveis estudadas neste trabalho, conforme já mencionado.

Outro fator que pode levar à dificuldade de orientação de setas pelo algoritmo PC é a presença de variáveis latentes não consideradas no modelo de causalidade (violação da pressuposição de suficiência causal). No caso das variáveis estudadas aqui, a desconsideração dos efeitos genéticos compartilhados na modelagem do relacionamento causal entre estes fenótipos é claramente uma violação desta pressuposição. Este fato é uma das maiores motivações para a inclusão de informações de QTLs na descoberta de modelos causais. Outra vantagem da adição de informações de QTLs nas análises causais é a possibilidade de quebrar a equivalência na distribuição de probabilidade na causalidade entre duas variáveis (CHAIBUB-NETO et al., 2008), permitindo a orientação de setas que, de outra maneira, não seria realizada pelos algoritmos de descoberta que se baseiam exclusivamente na descoberta de trios colisores, tais como o algoritmo PC. Tal inclusão torna-se cada vez mais facilitada pelo surgimento de novas técnicas de genotipagem e modelos de mapeamento de QTLs.

Dentre os algoritmos de descoberta de causalidade baseados em informações de QTL, o algoritmo QDG se destaca como um dos mais utilizados, sendo considerado atualmente como o estado da arte deste tipo de análise (LOGSDON; MEZEY, 2010; WANG; EEUWIJK, 2014). Wang e Eeuwijk (2014) propuseram recentemente o algoritmo QPSO com a intenção de promover melhoria ao algoritmo QDG ao utilizar simultaneamente as relações de causalidade fenotípicas para a orientação de setas. Dessa forma, o algoritmo QPSO não requer, como o algoritmo QDG, que todos os fenótipos avaliados tenham QTLs identificados. Estes dois algoritmos foram utilizados para as análises de causalidade neste trabalho. Ambos algoritmos dependem dos primeiros passos do algoritmo PC para a descoberta de setas do modelo causal (esqueleto de PC), porém diferem nas etapas subsequentes de orientação das setas encontradas (CHAIBUB-NETO et al., 2008; WANG; EEUWIJK, 2014). Enquanto o algoritmo QDG realiza a orientação de setas individuais entre pares de fenótipos adjacentes utilizando inicialmente somente informações de QTL e posteriormente informações de causalidade dos fenótipos pais sobre as variáveis, o algoritmo QPSO baseia-se na construção de redes fenotípicas generalizadas locais (LGPNs) que modelam simultaneamente informações causais fenotípicas e de QTLs.

Os resultados dos dois algoritmos para os dados deste trabalho diferiram. O algoritmo QPSO convergiu para um modelo de causalidade que foi o segundo melhor modelo de causalidade en-

contrado pelo algoritmo QDG. Embora a determinação de um modelo causal a partir de dados observacionais seja atualmente impossível, considerou-se que o algoritmo QDG tenha de fato alcançado uma melhor solução que o algoritmo QPSO, pois o algoritmo QDG foi capaz de identificar claramente os dois modelos e compará-los via BIC. Embora o cálculo de BIC empregado pelo algoritmo QDG não seja exatamente acurado devido à violação de aciclicidade apresentada pelos dois modelos, a comparação dos mesmos via MEEs também indicou que o modelo encontrado pelo QDG é possivelmente superior. Finalmente, o modelo causal sugerido pelo algoritmo QDG manteve o padrão de independências condicionais encontrado pelo algoritmo PC.

Wang e Eeuwijk (2014) afirmam que o algoritmo QPSO apresenta duas vantagens sobre o algoritmo QDG: (i) não requerer QTLs para todos os caracteres fenotípicos estudados; e, (ii) levar em consideração os relacionamentos causais fenotípicos na orientação das setas descobertas. Talvez o algoritmo QDG tenha apresentado melhor desempenho para o conjunto de dados deste trabalho porque o cenário requerido para a aplicação da primeira vantagem não se aplica a este caso específico, uma vez que nossa abordagem foi capaz de identificar QTLs significantes para todas as características avaliadas. Além disto, entende-se que as iterações do algoritmo QDG também levam em consideração os relacionamentos causais entre os fenótipos e a alegada segunda vantagem do algoritmo QPSO pode não ser de grande importância.

Chaibub-Neto, Keller e Yandell (2010) argumentam que o mapeamento de QTLs e o estudo de causalidade devem ser realizados conjuntamente, pois modelos causais auxiliam na detecção de QTLs, que por sua vez auxiliam na orientação de setas de modelos causais. Embora os autores estejam corretos neste argumento, a metodologia proposta por eles (algoritmo QTLnet) é bastante complexa e computacionalmente intensiva. Logsdon e Mezey (2010) compararam os algoritmos QDG e QTLnet e verificaram que o algoritmo QDG ainda apresenta resultados superiores ao algoritmo QTLnet para a maioria dos cenários que envolvem o estudo de causalidade. Por estes motivos, a modelagem de causalidade e o mapeamento de QTLs foram realizados em passos distintos neste trabalho.

A análise do resultado gerado pelo algoritmo QDG revelou que o modelo proposto apresenta duas alças vizinhas de retroalimentação causal (Figura 12 **B**). A presença destas alças de retroalimentação causais é uma explicação para a dificuldade do algoritmo PC na orientação de setas deste conjunto de dados. Chaibub-Neto et al. (2008) demonstraram por estudo de simulações que o algoritmo QDG apresenta uma boa performance tanto na descoberta quanto na orientação de setas em modelos causais cíclicos. Esta pode ser outra explicação para a melhor performance do algoritmo QDG neste conjunto de dados quando comparado ao algoritmo QPSO.

A análise do modelo causal gerado pelo algoritmo QDG (Figuras 13, 14 e 15), revelou

que os QTLs detectados, no geral, apresentaram efeitos em fenótipos causalmente relacionados. Este padrão levanta a hipótese de que os efeitos dos QTLs encontrados a jusante do primeiro fenótipo, para o qual o QTL tem efeito direto, sejam, na realidade, reflexos dos efeitos indiretos do QTL, detectados devido ao relacionamento causal dos fenótipos.

Este trabalho revelou regiões genômicas com atuação direta nas variáveis concentração de fósforo na parte aérea e nas raízes, que são características fenotípicas de grande interesse para os melhoristas de milho. A metodologia de mapeamento univariada revelou dois QTLs sobrepostos no cromossomo 1 de milho com efeito sobre ambas variáveis (*ShootP@1_159,00* e *RootP@1_188,00*) (Tabela 2 e Figura 9). O resultado do mapeamento multivariado também detectou tal região (*multi@1_183,80*) (Tabela 3 e Figura 9) e a análise de causalidade com o algoritmo QDG confirmou a atuação do QTL *multi@1_183,80* sobre as variáveis que são relacionamente causais: $\text{ShootP} \rightarrow \text{RootP}$ (Figura 13 B).

O modelo MT-MIM detectou outros QTLs de interesse para o melhoramento, por apresentarem efeitos sobre as variáveis ShootP e RootP. O primeiro apresenta efeito direto apenas sobre ShootP: *multi@3_10,00* (Tabela 3 e Figura 9). Outros QTLs com efeitos sobre as variáveis de concentração de fósforo são: (i) *multi@1_94,44*, que além de ShootP e RootP apresenta efeitos também sobre as variáveis RD, SA, SDW e RDW; (ii) *multi@7_77,27*, que afeta ShootP e também apresenta efeito sobre RD e SA; e (iii) *multi@6_133,00*, que apresenta efeito sobre RootP e SA2. Neste caso, como discutido, há a possibilidade de este QTL representar duas regiões distintas (Tabela 2 e Figura 9).

Além dos QTLs identificados com efeitos sobre as variáveis de absorção de fósforo, a análise da causalidade também revelou um relacionamento complexo entre os caracteres estudados (Figura 12 B). No entanto, algumas informações úteis para o melhoramento podem ser retiradas deste gráfico. Primeiro, a variável causalmente mais próxima de ShootP é SDW, que apresenta efeito causal alto, porém negativo, sobre ShootP ($\text{SDW} \rightarrow \text{ShootP} = -0,590$) (Tabela 8). O sinal negativo deste parâmetro estrutural representa o efeito de diluição da concentração de fósforo com o aumento do peso seco da parte aérea. Por este motivo, uma medida interessante para o melhoramento é o produto $\text{SDW} \times \text{ShootP}$ que resulta na conteúdo total de fósforo da parte aérea (SOUSA et al., 2011). Assim, plantas com maiores valores de SDW tendem a ter maior quantidade (em unidade de massa) de fósforo, porém em menor concentração, o que é de interesse dos melhoristas.

O parâmetro estrutural de maior magnitude foi $\text{SA2} \rightarrow \text{SA}$ (Tabela 8), informando que a área de superfície de raízes finas é um importante constituinte da área total de superfícies de raízes. O diâmetro de raiz (RD) tem efeito causal negativo sobre a área de superfície de raiz,

refletindo sua correlação negativa com o comprimento (dados não mostrados). Assim, raízes grossas tendem a ser também curtas, diminuindo a área de superfície de raiz total. SA, por sua vez, tem efeito causal positivo direto sobre SDW ($SA \rightarrow SDW = 0,676$) (Tabela 8). Assim, os QTLs com efeito sobre estes dois caracteres fenotípicos (*multi@1_263,00* e *multi@3_83,00*) (Figuras 13 e 14) podem auxiliar indiretamente na seleção de plantas com maior conteúdo de fósforo.

Os resultados apresentados neste trabalho ajudam no entendimento da arquitetura genética subjacente aos caracteres fenotípicos de morfologia de raiz, acumulação de biomassa e concentração de fósforo em milho. Certamente, este conhecimento tem potencial para ser empregado em programas de melhoramento com o objetivo de obtenção de variedades de milho com melhor absorção deste mineral.

6 CONCLUSÃO

A análise de relações de causalidade aliada a metodologias modernas de mapeamento de QTLs multivariadas constitui uma abordagem promissora para a análise da arquitetura genética de caracteres fenotípicos correlacionados. Neste trabalho, um conjunto de caracteres fenotípicos relacionados à morfologia do sistema radicular, acumulação de biomassa e concentração de fósforo em milho foi analisado utilizando técnicas de modelagem de causalidade e mapeamento de QTLs.

Utilizando mapeamento de múltiplos intervalos uni e multivariado, foi possível detectar QTLs com efeitos para todas as características avaliadas. As análises de QTL também revelaram herança complexa para os caracteres, com QTLs de pequenos efeitos individuais.

O emprego de técnicas de modelagem com modelos de equações estruturais e modelos gráficos permitiu que modelos causais recursivos e não-recursivos fossem propostos para explicar o relacionamento causal entre os caracteres fenotípicos estudados. Algoritmos de descoberta de causalidade revelaram claramente um padrão de relacionamento causal complexo, com ciclos de retroalimentação causal, potencialmente descobertos devido à falta de longitudinalidade dos dados avaliados. Como regra geral, os QTLs descobertos foram coerentes com o modelo causal proposto.

Os resultados deste trabalho têm o potencial de ajudar na compreensão da arquitetura genética dos caracteres fenotípicos avaliados e levantam hipóteses interessantes para ser testadas por meio de experimentos posteriores.

REFERÊNCIAS

- ATEN, J. E.; FULLER, T. F.; LUSIS, A. J.; HORVATH, S. Using genetic markers to orient the edges in quantitative trait networks: The NEO software. **BMC Systems Biology**, London, v. 2, n. 1, p. 34, 2008.
- AWIKA, J. Major cereal grains production and use around the world. In: AWIKA, J.; PIIRONEN, V.; BEAN, S. (Ed.). **Advances in cereal science: implications to food processing and health promotion**. Washington: American Chemical Society, 2011. p. 1–13.
- AZEVEDO, G. C.; CHEAVEGATTI-GIANOTTO, A.; F NEGRI, B.; HUFNAGEL, B.; da COSTA e SILVA, L.; MAGALHÃES, J. V.; GARCIA, A. A. F.; LANA, U. G.; GUIMARÃES, C. T.; SOUSA, S. M. de. QTL mapping and gene expression analyses reveal that PSTOL candidate genes are associated with root morphology and phosphorus acquisition in maize. **BMC Plant Biology**, London, (no prelo).
- BHADRA, A.; MALLICK, B. K. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. **Biometrics**, Washington, v. 69, p. 447–457, 2013.
- BOLLEN, K. A. **Structural equations with latent variables**. New York: Wiley, 1989. 514 p.
- BOWMAN, A. C.; VALENTE, B. D.; JANSS, L. L. G.; BOVENHUIS, H.; ROSA, G. J. M. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. **Genetics Selection Evolution**, Paris, v. 46, p. 2, 2014.
- BRADY, S. M.; SONG, S.; DHUGGA, K. S.; RAFALSKI, J. A.; BENFEY, P. N. Combining expression and comparative evolutionary analysis. The COBRA gene family. **Plant Physiology**, Bethesda, v. 143, p. 172–187, 2007.
- CAI, X.; BAZERQUE, J. A.; GIANNAKIS, G. B. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. **PLoS Computational Biology**, Cambridge, v. 9, n. 5, p. e1003068, 2013.
- CALINSKI, T.; KACZMAREK, Z.; KRAJEWSKI, P.; FROVAS, C.; SARI-GORLA, M. A multivariate approach to the problem of QTL localization. **Heredity**, London, v. 84, n. November 1999, p. 303–310, 2000.
- CHAIBUB-NETO, E.; BROMAN, A. T.; KELLER, M. P.; ATTIE, A. D.; ZHANG, B.; ZHU, J.; YANDELL, B. S. Modeling causality for pairs of phenotypes in system genetics. **Genetics**, Austin, v. 193, p. 1003–1013, 2013.
- CHAIBUB-NETO, E.; FERRARA, C. T.; ATTIE, A. D.; YANDELL, B. S. Inferring causal phenotype networks from segregating populations. **Genetics**, Austin, v. 179, n. 2, p. 1089–100, 2008.
- CHAIBUB-NETO, E.; KELLER, M. P.; YANDELL, B. S. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes, Supplement. **The Annals of Applied Statistics**, Cleveland, v. 4, n. 1, p. 320–339, 2010.

CHAIBUB-NETO, E.; YANDELL, B. S. **qtlnet**: Causal Inference of QTL Networks. 2014. Disponível em: <<http://cran.r-project.org/package=qtlnet>>. Acesso em: 18 Out. 2014

CHEN, J.; XU, L.; CAI, Y.; XU, J. QTL mapping of phosphorus efficiency and relative biologic characteristics in maize (*Zea mays* L.) at two sites. **Plant and Soil**, The Hague, v. 313, p. 251–266, 2008.

_____. Identification of QTLs for phosphorus utilization efficiency in maize (*Zea mays* L.) across P levels. **Euphytica**, Wageningen, v. 167, p. 245–252, 2009.

CHEN, L. S.; EMMERT-STREIB, F.; STOREY, J. D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. **Genome Biology**, London, v. 8, n. 10, p. R219, 2007.

COLOMBO, D.; MAATHUIS, M. H. Order-Independent Constraint-Based Causal Structure Learning. **Journal of Machine Learning Research**, Massachusetts, v. 15, p. 3741–3782, 2014.

DUARTE, C. W.; ZENG, Z.-B. High-confidence discovery of genetic network regulators in expression quantitative trait loci data. **Genetics**, Austin, v. 187, p. 955–964, 2011.

FISHER, R. A. **Statistical methods for research workers**. Edinburgh: Oliver & Boyd, 1925. 237 p.

_____. Smoking: the cancer controversy. **The American Journal of the Medical Sciences**, Edinburgh, v. 239, n. 3, p. 395, 1960.

FOX, J. **An R and S-Plus companion to applied regression**. Thousand Oaks: Sage, 2002. 448 p.

_____. Teacher's corner: Structural equation modeling with the sem package in R. **Structural Equation Modeling**, Florence, v. 13, n. 3, p. 465–486, 2006.

FOX, J.; NIE, Z.; BYRNES, J. **sem**: Structural Equation Models. 2014. Disponível em: <<http://cran.r-project.org/package=sem>>. Acesso em: 12 Jan. 2014

FRIEDMAN, N.; LINIAL, M.; NACHMAN, I.; PE, D. Using Bayesian networks to analyze expression data. **Journal of Computational Biology**, New York, v. 7, n. 3-4, p. 601–620, 2000.

FUENTE, A. de la; BING, N.; HOESCHELE, I.; MENDES, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. **Bioinformatics**, Oxford, v. 20, n. 18, p. 3565–3574, 2004.

GAMUYAO, R.; CHIN, J. H.; PARIASCA-TANAKA, J.; PESARESI, P.; CATAUSAN, S.; DALID, C.; SLAMET-LOEDIN, I.; TECSON-MENDOZA, E. M.; WISSUWA, M.; HEUER, S. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. **Nature**, London, v. 488, p. 535–539, 2012.

GILBERT, H.; Le Roy, P. Comparison of three multitrait methods for QTL detection. **Genetics Selection Evolution**, Paris, v. 35, n. 3, p. 281–304, 2003.

HALEY, C. S.; KNOTT, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. **Heredity**, London, v. 69, n. 1992, p. 315–324, 1992.

- HOCHHOLDINGER, F.; TUBEROSA, R. Genetic and genomic dissection of maize root development and architecture. **Current Opinion in Plant Biology**, London, v. 12, p. 172–177, 2009.
- HONG, Y.; PEDERSEN, N. L.; KERSTIN, B.; FAIRE, U. de. Genetic and environmental architecture of the features of the insulin-resistance syndrome. **American Journal of Human Genetics**, Chicago, v. 60, p. 143–152, 1997.
- JANSEN, R. C. Interval Mapping of Multiple Quantitative Trait Loci. **Genetics**, Austin, v. 135, p. 205–211, 1993.
- JANSEN, R. C.; NAP, J.-P. Genetical genomics: the added value from segregation. **Trends in Genetics**, Amsterdam, v. 17, n. 7, p. 388–391, 2001.
- JANSEN, R. C.; STAM, P. High Resolution of Quantitative Traits Into Multiple Loci via Interval Mapping. **Genetics**, Austin, v. 136, p. 1447–1455, 1994.
- JANSEN, R. C.; TESSON, B. M.; FU, J.; YANG, Y.; MCINTYRE, L. M. Defining gene and QTL networks. **Current Opinion in Plant Biology**, London, v. 12, n. 2, p. 241–6, abr. 2009.
- JIANG, C.; ZENG, Z.-B. Multiple trait analysis of genetic mapping for quantitative trait loci. **Genetics**, Austin, v. 140, p. 1111–1127, 1995.
- JÖRESKOG, K. G. A general approach to confirmatory maximum likelihood factor analysis. **Psychometrika**, Williamsburg, v. 34, n. 2, p. 183–202, 1969.
- _____. A general method for analysis of covariance structures. **Biometrika**, London, v. 57, n. 2, p. 239–251, 1970.
- KALISCH, M.; COLOMBO, D.; MAATHUIS, M. H. Causal inference using graphical models with the R package pcalg. **Journal of Statistical Software**, Los Angeles, v. 47, n. 11, p. 1–26, 2012.
- KAO, C. H.; ZENG, Z.-B. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. **Biometrics**, Washington, v. 53, p. 653–665, 1997.
- KAO, C.-h.; ZENG, Z.-b. Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. **Genetics**, Austin, v. 160, p. 1243–1261, 2002.
- KAO, C. H.; ZENG, Z. B.; TEASDALE, R. D. Multiple interval mapping for quantitative trait loci. **Genetics**, Austin, v. 152, p. 1203–1216, 1999.
- KATAN, M. Apolipoprotein E isoforms, serum cholesterol, and cancer. **International Journal of Epidemiology**, Oxford, v. 33, p. 9, 2004.
- KNOTT, S. A.; HALEY, C. S. Multitrait Least Squares for Quantitative Trait Loci Detection. **Genetics**, Austin, v. 156, n. 1999, p. 899–911, 2000.
- KOLLER, D.; FRIEDMAN, N. **Probabilistic graphical models: principles and techniques**. Massachusetts: MIT Press, 2009. 1231 p.
- KOSAMBI, D. D. The estimation of map distances from recombination values. **Annals of Eugenics**, London, v. 12, n. 1, p. 172–175, 1954.

- KULP, D. C.; JAGALUR, M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. **BMC Genomics**, London, v. 7, p. 125, 2006.
- LANDER, E. S.; BOTSTEIN, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**, Austin, v. 121, p. 185–199, 1989.
- LANDER, E. S.; GREEN, P.; ABRAHAMSON, J.; BARLOW, A.; DALY, M. J.; LINCOLN, S. E.; NEWBURG, L. MAPMAKER: An Interactive Computer Package for Constructing Primary Genetic Linkage Maps of Experimental and Natural Populations. **Genomics**, San Diego, v. 1, p. 174–181, 1987.
- LAURIE, C.; WANG, S.; CARLINI-GARCIA, L. A.; ZENG, Z.-B. Mapping epistatic quantitative trait loci. **BMC Genetics**, London, v. 15, p. 112, 2014.
- LAWLOR, D. A.; HARBORD, R. M.; STERNE, J.A. C.; TIMPSON, N.; DAVEY SMITH, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. **Statistics in Medicine**, New York, v. 27, p. 1133–1163, 2008.
- LI, H.; LU, L.; MANLY, K. F.; CHESLER, E. J.; BAO, L.; WANG, J.; ZHOU, M.; WILLIAMS, R. W.; CUI, Y. Inferring gene transcriptional modulatory relations: a genetical genomics approach. **Human Molecular Genetics**, Oxford, v. 14, n. 9, p. 1119–1125, 2005.
- LI, R.; TSAIH, S.-W.; SHOCKLEY, K.; STYLIANOU, I. M.; WERGEDAL, J.; PAIGEN, B.; CHURCHILL, G. a. Structural model analysis of multiple quantitative traits. **PLoS genetics**, San Francisco, v. 2, n. 7, p. e114, 2006.
- LIU, B.; HOESCHELE, I. Gene network inference via structural equation modeling in genetical genomics experiments. **Genetics**, Austin, v. 178, p. 1763–1776, 2008.
- LOGSDON, B. A.; MEZEY, J. Gene Expression Network Reconstruction by Convex Feature Selection when Incorporating Genetic Perturbations. **PLoS Computational Biology**, Cambridge, v. 6, n. 12, p. e1001014, 2010.
- LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Massachusetts: Sinauer Sunderland, 1998. 980 p.
- MAGWENE, P. M.; KIM, J. Estimating genomic coexpression networks using first-order conditional independence. **Genome Biology**, London, v. 5, n. 12, p. R100, 2004.
- MALOSETTI, M.; RIBAUT, J. M.; VARGAS, M.; CROSSA, J.; EEUWIJK, F. A. van. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). **Euphytica**, Wageningen, v. 161, p. 241–257, 2008.
- MANGIN, B.; THOQUET, P.; GRIMSLEY, N. Pleiotropic QTL Analysis. **Biometrics**, Washington, v. 54, n. 1, p. 88–99, 1998.
- MARGARIDO, G. R. A. **Maapeamento de QTLs em múltiplos caracteres e ambientes em um cruzamento comercial de cana-de-açúcar usando modelos mistos**. 2011. 107 p. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz", 2011.
- MARGARIDO, G. R. A.; SOUZA, A. P.; GARCIA, A. A. F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, Lund, v. 144, p. 78–79, 2007.

- MARUYAMA, G. M. **Basics of structural equation modeling**. Thousand Oaks: Sage, 1997. 310 p.
- MATLAB. **MATLAB** Versão 8.1.0.604 (R2013a). Natick, Massachusetts: The MathWorks Inc., 2013.
- MATURANA, E. L.; WU, X. L.; GIANOLA, D.; WEIGEL, K. A.; LO, E.; ROSA, G. J. M. Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model. **Genetics**, Austin, v. 181, p. 277–287, 2009.
- MI, X.; ESKRIDGE, K.; WANG, D.; BAENZIGER, P. S.; CAMPBELL, B.; GILL, K. S.; DWEIKAT, I. Bayesian mixture structural equation modelling in multiple-trait QTL mapping. **Genetics Research**, Cambridge, v. 92, n. 3, p. 239–250, 2010.
- MI, X.; ESKRIDGE, K.; WANG, D.; BAENZIGER, P. S.; CAMPBELL, B. T.; GILL, K. S.; DWEIKAT, I.; BOVAIRD, J. Regression-based multi-trait QTL mapping using a structural equation model. **Statistical Applications in Genetics and Molecular Biology**, Berlin, v. 9, n. 1, jan. 2010.
- NEALE, M.; CARDON, L. **Methodology for genetic studies of twins and families**. Dordrecht: Springer, 1992. 496 p.
- NINAMANGO-CÁRDENAS, F. H.; GUIMARÃES, C. T.; MARTINS, P. R.; PARENTONI, S. N.; CARNEIRO, N. P.; LOPES, M. A.; MORO, J. R.; PAIVA, E. Mapping QTLs for aluminum tolerance in maize. **Euphytica**, Wageningen, v. 130, p. 223–232, 2003.
- OORD, E. J. C. G. van den. Framework for identifying quantitative trait loci in association studies using structural equation modeling. **Genetic Epidemiology**, New York, v. 18, p. 341–359, 2000.
- PARENTONI, S. N.; SOUZA-JUNIOR, C. L. de. Phosphorus acquisition and internal utilization efficiency in tropical maize genotypes. **Pesquisa Agropecuária Brasileira**, Brasília, v. 43, n. 7, p. 893–901, 2008.
- PEARL, J. **Causality: models, reasoning and inference**. Cambridge: Cambridge Univ Press, 2000. 384 p.
- PE'ER, D.; REGEV, A.; ELIDAN, G.; FRIEDMAN, N. Inferring subnetworks from perturbed expression profiles. **Bioinformatics**, Oxford, v. 17, n. 1, p. S215–S224, 2001.
- PINHEIRO, J.; BATES, D.; DEBROY, S.; R-CORE-TEAM. **nlme: Linear and nonlinear mixed effects models**. 2014. Disponível em: <<http://cran.r-project.org/package=nlme>>. Acesso em: 10 Jan. 2014
- RAFTERY, A. E. Bayesian Model Selection in Structural Equation Models. In: BOLLEN, K. A.; LONG, J. S. (Ed.). **Testing structural equation models**. Thousand Oaks: Sage, 1993. v. 154. p 325.
- RAMAEKERS, L.; REMANS, R.; RAO, I. M.; BLAIR, M. W.; VANDERLEYDEN, J. Strategies for improving phosphorus acquisition efficiency of crop plants. **Field Crops Research**, Amsterdam, v. 117, p. 169–176, 2010.
- ROCKMAN, M. V. Reverse engineering the genotype-phenotype map with natural genetic variation. **Nature**, London, v. 456, n. 7223, p. 738–44, 2008.

ROSA, G. J. M.; VALENTE, B. D.; CAMPOS, G. de los; WU, X.-L.; GIANOLA, D.; SILVA, M. a. Inferring causal phenotype networks using structural equation models. **Genetics, Selection, Evolution**, Paris, v. 43, n. 1, p. 6, 2011.

SCHADT, E. E.; LAMB, J.; YANG, X.; ZHU, J.; EDWARDS, S.; GUHATHAKURTA, D.; SIEBERTS, S. K.; MONKS, S.; REITMAN, M.; ZHANG, C.; LUM, P. Y.; LEONARDSON, A.; THIERINGER, R.; METZGER, J. M.; YANG, L.; CASTLE, J.; ZHU, H.; KASH, S. F.; DRAKE, T. a.; SACHS, A.; LUSIS, A. J. An integrative genomics approach to infer causal associations between gene expression and disease. **Nature Genetics**, New York, v. 37, n. 7, p. 710–7, 2005.

SCHEFFER, D. K. **Modelos de equações estruturais aplicados ao mapeamento de genes em cruzamentos controlados**. 2008. 129 p. Tese (Dissertação de Mestrado) — Universidade de São Paulo, 2008.

SCHUMACKER, R. E.; LOMAX, R. G. **A beginner's guide to structural equation modeling**. New York: Routledge Academic, 2010. 481 p.

SEGAL, E.; SHAPIRA, M.; REGEV, A.; BOTSTEIN, D.; KOLLER, D.; FRIEDMAN, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. **Nature Genetics**, New York, v. 34, n. 2, p. 166–176, 2003.

SHALIZI, C. R. **Advanced data analysis from an elementary point of view**. Cambridge: Cambridge Univ Press, 2014. 584 p.

SHEN, J.; YUAN, L.; ZHANG, J.; LI, H.; BAI, Z.; CHEN, X.; ZHANG, W.; ZHANG, F. Phosphorus Dynamics: From Soil to Plant. **Plant Physiology**, Bethesda, v. 156, p. 997–1005, 2011.

SHIPLEY, B. **Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference**. Cambridge: Cambridge University Press, 2002. 317 p.

SILVA, L. d. C. e.; WANG, S.; ZENG, Z.-B. Multiple trait multiple interval mapping of quantitative trait loci from inbred line crosses. **BMC Genetics**, London, v. 13, p. 67, 2012.

SILVA, L. d. C. e.; ZENG, Z.-B. Current progress on statistical methods for mapping quantitative trait loci from inbred line crosses. **Journal of Biopharmaceutical Statistics**, Durham, v. 20, p. 454–481, 2010.

SOUSA, S. M. D.; CLARK, R. T.; MENDES, F. F.; OLIVEIRA, A. C. D.; VASCONCELOS, M. J. V.; PARENTONI, S. N.; KOCHIAN, L. V.; GUIMARAES, C. T.; MAGALHAES, J. V. A role for root morphology and related candidate genes in P acquisition efficiency in maize. **Functional Plant Biology**, Victoria, v. 39, n. 11, p. 925–935, 2011.

SPEARMAN, C. The proof and measurement of association between two things. **American Journal of Psychology**, Champaign, v. 15, n. 1, p. 72–101, 1904.

SPIRITES, P. Directed Cyclic Graphical Representations of Feedback Models. In: ELEVENTH CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE. **Proceedings...** Quebec: Morgan Kaufmann Publishers Inc., 1995. p. 491–498.

SPIRITES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, prediction, and search**. 81. ed. Massachusetts: MIT Press, 2000. 526 p.

- STEIN, C. M.; SONG, Y.; ELSTON, R. C.; JUN, G.; TIWARI, H. K.; IYENGAR, S. K. Structural equation model-based genome scan for the metabolic syndrome. **BMC Genetics**, London, v. 5, p. 1–5, 2003.
- STOLLEY, P. D. When genius errs: R. A. Fisher and the lung cancer controversy. **American Journal of Epidemiology**, Baltimore, v. 133, n. 5, p. 416–425, 1991.
- STUNKARD, A. J.; HARRIS, J. R.; PEDERSEN, N. L.; MCCLEARN, G. The body-mass index of twins who have been reared apart. **New England Journal of Medicine**, Waltham, v. 322, n. 21, p. 1483 – 1487, 1990.
- TARAMINO, G.; SAUER, M.; STAUFFER-JR, J. L.; MULTANI, D.; NIU, X.; SAKAI, H.; HOCHHOLDINGER, F. The maize (*Zea mays* L.) RTCS gene encodes a LOB domain protein that is a key regulator of embryonic seminal and post-embryonic shoot-borne root initiation. **The Plant Journal**, Malden, v. 50, p. 649–659, 2007.
- THOMAS, D. C.; CONTI, D. V. Commentary: the concept of ‘Mendelian randomization’. **International Journal of Epidemiology**, Oxford, v. 33, p. 21–25, 2004.
- TUR, I.; ROVERATO, A.; CASTELO, R. Mapping eQTL networks with mixed graphical Markov models. **Genetics**, Austin, v. 198, n. 4, p. 1377 – 1393, 2014.
- ULLMAN, J. B. Structural equation modeling: reviewing the basics and moving forward. **Journal of Personality Assessment**, Hillsdale, v. 87, n. 1, p. 35–50, 2006.
- VALENTE, B. D.; ROSA, G. J. M.; de Los Campos, G.; GIANOLA, D.; SILVA, M. A. Searching for recursive causal structures in multivariate quantitative genetics mixed models. **Genetics**, Austin, v. 185, n. 2, p. 633–44, 2010.
- VALENTE, B. D.; ROSA, G. J. M.; SILVA, M. A.; TEIXEIRA, R. B.; TORRES, R. A. Searching for phenotypic causal networks involving complex traits: an application to European quail. **Genetics Selection Evolution**, Paris, v. 43, n. 1, p. 37, 2011.
- VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era — concepts and misconceptions. **Nature Reviews Genetics**, New York, v. 9, p. 255–266, 2008.
- WANG, H.; EEUWIJK, F. a. van. A new method to infer causal phenotype networks using QTL and phenotypic information. **PloS One**, San Francisco, v. 9, n. 8, p. e103997, 2014.
- WANG, W.; BALADANDAYUTHAPANI, V.; HOLMES, C. C.; DO, K.-A. Integrative network-based Bayesian analysis of diverse genomics data. **BMC Bioinformatics**, London, v. 14, n. 13, p. S8, 2013.
- WELLER, J.; WIGGANS, G.; VANRADEN, P.; RON, M. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. **Theoretical and Applied Genetics**, Berlin, v. 92, n. 8, p. 998–1002, 1996.
- WEN, T.-J.; HOCHHOLDINGER, F.; SAUER, M.; BRUCE, W.; SCHNABLE, P. S. The roothairless1 gene of maize encodes a homolog of sec3, which is involved in polar exocytosis. **Plant Physiology**, Bethesda, v. 138, p. 1637–1643, 2005.
- WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, Washington, v. 4, n. 557, p. 557–586, 1921.

ZENG, Z. B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 90, p. 10972–10976, 1993.

ZENG, Z.-B. Precision Mapping of Quantitative Trait Loci. **Genetics**, Austin, v. 136, p. 1457–1468, 1994.

ZENG, Z.-B.; KAO, C. H.; BASTEN, C. J. Estimating the genetic architecture of quantitative traits. **Genetics Research**, Cambridge, v. 74, p. 279–289, 1999.

ZHANG, B.; GAITERI, C.; BODEA, L.-G.; WANG, Z.; MCELWEE, J.; PODTELEZHNIKOV, A. A.; ZHANG, C.; XIE, T.; TRAN, L.; DOBRIN, R.; FLUDER, E.; CLURMAN, B.; MELQUIST, S.; NARAYANAN, M.; SUVER, C.; SHAH, H.; MAHAJAN, M.; GILLIS, T.; MYSORE, J.; MACDONALD, M. E.; LAMB, J. R.; BENNETT, D. A.; MOLONY, C.; STONE, D. J.; GUDNASON, V.; MYERS, A. J.; SCHADT, E. E.; NEUMANN, H.; ZHU, J.; EMILSSON, V. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. **Cell**, Cambridge, v. 153, p. 707–720, 2013.

ZHU, J.; KAEPPPLER, S. M.; LYNCH, J. P. Mapping of QTL controlling root hair length in maize (*Zea mays* L.) under phosphorus deficiency. **Plant and Soil**, The Hague, v. 270, p. 299–310, 2005.

_____. Mapping of QTLs for lateral root branching and length in maize (*Zea mays* L.) under differential phosphorus supply. **Theoretical and Applied Genetics**, Berlin, v. 111, p. 688–695, 2005.

ZHU, J.; MICKELSON, S. M.; KAEPPPLER, S. M.; LYNCH, J. P. Detection of quantitative trait loci for seminal root traits in maize (*Zea mays* L.) seedlings grown under differential phosphorus levels. **Theoretical and Applied Genetics**, Berlin, v. 113, p. 1–10, 2006.

ZHU, J.; WIENER, M. C.; ZHANG, C.; FRIDMAN, A.; MINCH, E.; LUM, P. Y.; SACHS, J. R.; SCHADT, E. E. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. **PLoS Computational Biology**, Cambridge, v. 3, n. 4, p. e69, 2007.

ZOU, F.; FINE, J. P.; HU, J.; LIN, D. Y. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. **Genetics**, Austin, v. 168, p. 2307–2316, 2004.