UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Desenvolvimento e utilização de um *array* de genotipagem de SNPs na genética de populações e melhoramento genético de *Araucaria angustifolia (Bert.)* O. Ktze

PEDRO ITALO TANNO SILVA

BRASÍLIA, 2020

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Desenvolvimento e utilização de um *array* de genotipagem de SNPs na genética de populações e melhoramento genético de *Araucaria angustifolia (Bert.)* O. Kuntze

PEDRO ITALO TANNO SILVA

Orientador: Dario Grattapaglia

Tese apresentada ao Programa de Pós-Graduação em Biologia Molecular da Universidade de Brasília, como requisito para obtenção do título de Doutor em Biologia Molecular.

BRASÍLIA, 2020

PEDRO ITALO TANNO SILVA

Desenvolvimento e utilização de um *array* de genotipagem de SNPs na genética de populações e melhoramento genético de *Araucaria angustifolia (Bert.)* O. Kuntze

Tese apresentada ao Programa de Pós-Graduação em Biologia Molecular da Universidade de Brasília, como requisito para obtenção do título de Doutor em Biologia Molecular.

Banca Examinadora:

____

Prof. Dr. Dario Grattapaglia (Orientador) (CEL – UnB)

____

Prof. Dr. Robert Neil Gerard Miller (CEL – UnB)

____

Prof. Dr. Rafael Tassinari Resende (EA – UFG)

____

Dr. Orzenil Bonfim da Silva Junior (EMBRAPA)

APOIO FINANCEIRO

Projeto FAP-DF Pronex 2009/00106-8 “Nextree – Núcleo de Excelência em Genômica Florestal aplicada” e projeto Embrapa "Desenvolvimento de sistema de genotipagem de alto desempenho de SNPs para *Araucaria angustifolia*” 02.11.08.005.00.03

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Araucaria angustifolia (Bertol.) Kuntze é uma conífera subtropical endêmica no sul e sudeste do Brasil e áreas menores na Argentina e no Paraguai. Essa espécie icônica destaca-se como uma das duas únicas gimnospermas nativas brasileiras, atualmente em risco de extinção e ainda essencialmente inexplorada do ponto de vista do melhoramento genético apesar de seu alto valor madeireiro. Neste trabalho investigamos a variação genética para características de crescimento de *Araucaria angustifolia* em um estudo envolvendo 122 famílias de 15 procedências de duas regiões brasileiras. Medidas coletadas nas idades de 7, 24, 32, 33 e 35 anos foram usadas para ajustar curvas de crescimento contínuo com base em modelos de efeito misto não-lineares para 2.158 árvores, fornecendo estimativas para idades não medidas no intervalo de 7-35. Os valores estimados coincidiram com os observados e uma redução do coeficiente de variação residual foi observada nos dados estimados, tornando as curvas estimadas mais confiáveis para prever padrões de crescimento. Procedências com grande potencial de melhoramento e conservação genética foram identificadas com adaptação variável para apoiar as mudanças climáticas e paisagísticas globais. Os dados de crescimento indicam claramente o potencial de seleção precoce de 7 a 10 anos, com 85% de precisão na seleção aos 35 anos e a possibilidade de encurtar a idade de rotação para 15 a 20 anos, selecionando os melhores indivíduos e famílias. Um programa de melhoramento moderno depende da disponibilidade de ferramentas de genotipagem de alto rendimento, que se tornou um pré-requisito não apenas para análises mais sofisticadas da diversidade e estrutura genética em populações naturais, mas também para a prática de melhoramento genético envolvendo reconstrução de parentesco e implementação da seleção genômica. Até o momento, estudos genéticos de populações naturais de araucária foram realizados usando dezenas de marcadores microssatélites limitando assim a capacidade de evoluir para investigações mais sofisticadas, e não existe nenhum estudo associando análises a marcadores moleculares e programas de melhoramento. Neste estudo, desenvolvemos um catálogo de 44.318 SNPs anotados para *Araucaria angustifolia*, os primeiros SNPs para o gênero, descobertos a partir dos dados de sequenciamento de RNAseq e RAD. A partir do catálogo SNP, um array Axiom® SNP com ~3.000

SNPs validados foi desenvolvido e usado para fornecer uma visão abrangente da diversidade genética e estrutura de 15 populações em toda a faixa de ocorrência natural da espécie. Das 22.983 sequências utilizadas, 15.144 possuem homologia com sequências já publicadas, e destas, 5.301 possuem homologia com *Picea sitchensis*, conífera de grande importância industrial e ecológica em países temperados. Ao comparar dados de microssatélites e SNP no mesmo conjunto de indivíduos de *A. angustifolia*, mostramos que os SNPs refletem com maior precisão os padrões reais de diversidade e estrutura genética em todo o genoma, desafiando avaliações anteriores baseadas em microssatélites. Além disso, os SNPs corroboraram o principal gradiente genético norte-sul conhecido, e permitiram uma atribuição mais precisa à diferenciação regional versus entre populações, indicando o potencial de selecionar marcadores informativos de ancestralidade. Também. Combinando os SNPs desenvolvidos e as curvas de crescimento ajustadas, este trabalho também teve como objetivo propor uma metodologia baseada na seleção genômica (GS) para acelerar o melhoramento genético de *A. angustifolia*. Foi utilizado um conjunto de 1.710 SNPs combinados com 26 SSRs, 857 plantas foram genotipadas com ambas as plataformas de marcadores. Utilizando dados de crescimento de 35 anos, foram observados componentes de variância aditiva, de dominância e epistática, sendo o primeiro o mais prevalente na expressão volumétrica do crescimento das árvores. Nossas descobertas mostram que é possível treinar modelos a partir dos 12 anos de idade visando a seleção indireta bem-sucedida de indivíduos aos 35 anos. Dado o longo tempo necessário para alcançar o crescimento máximo das árvores e a boa qualidade dos modelos, a SG foi muito competitiva com a seleção fenotípica abrindo excelentes oportunidades para estabelecer um programa de melhoramento economicamente viável e eficiente dessa icônica conífera brasileira. Esses resultados destacam o enorme potencial de expansão dos investimentos em silvicultura de melhoramento e plantio de *A. angustifolia*, com um aprimoramento concomitante dos esforços de conservação.

Palavras-chave: Marcadores moleculares, SNP, SSR, Araucária, Fst, genética de populações, melhoramento florestal, Seleção Genômica.

ABSTRACT

Araucaria angustifolia (Bertol.) Kuntze is a subtropical coniferous tree endemic to southern and southeastern Brazil and smaller areas in Argentina and Paraguay. This iconic species stands out as one of the only two native Brazilian gymnosperms, currently at risk of extinction and still essentially unexplored from the point of view of genetic improvement despite its high logging value. In this study we investigated the genetic variation for growth traits of *Araucaria angustifolia* in a trial involving 122 families from 15 provenances from two Brazilian regions. Measurements at ages 7, 24, 32, 33 and 35 were used to adjust continuous growth curves based on nonlinear mixed-effect models for 2,158 trees, providing estimates for unmeasured ages in the 7-35 interval. Estimated values closely matched observed ones and a reduction of the coefficient of residual variation was observed in the estimated data, making the estimated curves more reliable to predict growth patterns. Provenances with great potential for breeding and genetic conservation were identified with variable adaptation to support global climate and landscape change. Growth data clearly indicates potential for early selection at age 7-10 with 85% accuracy of selection at age 35, and the possibility of shortening rotation age to 15-20 years by selecting the best individuals and families. A modern breeding program relies on the availability of high throughput genotyping tools, which became a prerequisite not only for more sophisticated analyses of diversity and genetic structure in natural populations but also for the practice of advanced breeding involving kinship reconstruction and implementation of genomic selection. To date, genetic studies of natural *Araucaria* populations have been performed using dozens of microsatellite markers limiting the ability to evolve to more sophisticated investigations, and no study exists associating analysis with molecular markers and breeding programs. In this study, we developed a 44,318 annotated SNP catalog for *Araucaria angustifolia*, the first SNPs for the genus, discovered from RNAseq and RAD-sequencing data. From the SNP catalog, an Axiom® SNP array with 3,038 validated SNPs was developed and used to provide a comprehensive look at the genetic diversity and structure of 15 populations across the natural range of the species. Of the 22,983 sequences used, 15,144 have homology with sequences already published, and of these, 5,301 have homology with *Picea sitchensis*, conifer of great industrial and ecological importance in temperate countries. By matching microsatellite and SNP data on the same set of *A. angustifolia* individuals, we show that SNPs reflect more precisely the actual genome-wide patterns of genetic diversity and structure, challenging previous microsatellite-based assessments. Moreover, SNPs corroborated the known major north-south genetic cline and allowed a more accurate attribution to regional versus among-population differentiation, indicating the potential to select ancestry-informative markers. Combining the SNPs developed and the adjusted growth curves, this work also aimed to propose a methodology based on genomic selection (GS) to accelerate the genetic improvement of *A. angustifolia*. A set of 1,710 SNPs combined with 26 SSRs was used, 857 plants were genotyped with both marker platforms. Using 35 years of growth data, components of additive, dominant and epistatic variance were observed, the first being the most prevalent in volumetric expression throughout the growth of the trees. Our findings show that one can train models as early as 12 years of age aiming at successful indirect selection of individuals at age 35 years. Given the long time necessary to reach maximal growth of the trees and the good quality of the models, GS was very competitive with phenotypic selection, opening outstanding opportunities to establish an economically viable and efficient breeding program of this iconic Brazilian conifer. These results underscore the huge potential of expanding investments in breeding and plantation forestry of *A. angustifolia* with a concomitant enhancement of conservation efforts.

Keywords: molecular markers, SNP, SSR, Araucaria, Fst, population genetics, forestry breeding, Genomic Selection.

INTRODUÇÃO

A Araucaria angustifolia (Bertol.) Kuntze é uma árvore conífera subtropical endêmica do sul e sudeste do Brasil e de áreas menores na Argentina e no Paraguai, ocorrendo em altitudes entre 500 e 1800 m. Esta espécie icônica se destaca como a única espécie de gimnosperma nativa no Brasil e, junto com *A. araucana*, as únicas duas coníferas nativas da América do Sul. *A. angustifolia* é também a espécie líder na Floresta Ombrófila mista (também conhecida como Floresta de Araucária), composta por uma mistura distinta de elementos florísticos temperados e tropicais (Klein, 1960).

Até o início do século passado, *A. angustifolia* ocupava uma área de aproximadamente 200.000 km^2 em todo o sul do Brasil, que foi rapidamente reduzida pela extração extensiva de madeira e conversão de terras através das várias ondas de colonização e expansão agrícola. A espécie integra hoje a Lista Vermelha de espécies ameaçadas de extinção da IUCN. A exploração desordenada principalmente durante o século 20 reduziu sua área de ocorrência natural a aproximadamente 3% do total inicial (GUERRA; SILVEIRA; DOS SANTOS; ASTARITA *et al.*, 2000). A alta qualidade da madeira da Araucária permitiu que fosse usada para diversos fins - construção, movelaria e celulose (CARVALHO, 1994). Sua exploração intensificou-se no começo do século 20 e teve seu auge entre as décadas de 1950 e 1970, quando o metro cúbico da espécie era o produto madeireiro mais importante do Brasil (SHIMIZU; OLIVEIRA, 1981). A coleta e comercialização de pinhão constituem uma importante fonte de renda para famílias dessa região até os dias de hoje. Em 2015 foram produzidas aproximadamente 8.393 toneladas de pinhão gerando uma renda de mais de 21 milhões de reais, sendo 80% desse montante oriundos dos estados do Paraná e de Santa Catarina (IBGE, 2015).

Estudos de diversidade genética em populações naturais compreendem a descrição dos níveis e dinâmica da variação genética existente e a forma como essa variação é estruturada dentro e entre populações. Informações sobre a diversidade e estrutura genética de populações naturais ajudam a definir estratégias de conservação bem como subsidiar ações de domesticação e melhoramento genético da espécie (PUTMAN; CARBONE, 2014). A caracterização da diversidade genética de *Araucaria angustifolia* foi por muito tempo realizada exclusivamente com base na mensuração de características fenotípicas no âmbito de programas iniciais de testes de procedências e progênies, com destaque para crescmento volumétrico. O desenvolvimento e uso de marcadores moleculares permitiu um importante avanço na caracterização de recursos genéticos da espécie. Inicialmente trabalhos usando marcadores isoenzimáticos foram utilizados na década de 80 revelando uma alta diversidade genética dentro de populações e baixa entre populações naturais (AULER; REIS; GUERRA; NODARI, 2002; MANTOVANI; MORELLATO; DOS REIS, 2006; SHIMIZU; JAEGER; SOPCHAKI, 2000). A partir do final da década de 90 trabalhos foram publicados caracterizando a diversidade genética em *A. angustifolia* com base em marcadores moleculares RAPD (MAZZA, 1997; MEDRI; RUAS; HIGA; MURAKAMI *et al.*, 2003) e AFLP (SOUZA; SALGUEIRO; CARNAVALE-BOTTINO; FÉLIX *et al.*, 2009). Marcadores microssatélites foram em seguida desenvolvidos e utilizados para análises genéticas de Araucaria. SCOTT; SHEPHERD e HENRY (2003) desenvolveram primers para 10 marcadores microssatélites em *A. cunninghamii*. ROBERTSON; HOLLINGSWORTH; KETTLE; ENNOS *et al.* (2004) desenvolveram primers para 5 marcadores microssatélites em *A. columnari*. SALGUEIRO; CARON; DE SOUZA; KREMER *et al.* (2005) desenvolveram primers para mais 6 microssatélites em *A. angustifolia* e *A. araucana*., SCHMIDT; CIAMPI; GUERRA e NODARI (2007) desenvolveram primers para 29 microssatélites em *A. angustifolia* e MARTIN; MATTIONI; LUSINI; DRAKE *et al.* (2012) para mais 10 marcadores microssatélites porém apenas 8 deles foram polimórficos em *A. angustifolia*. Este último trabalho foi o único feito utilizando dados de sequenciamento de próxima geração (Next generation sequencing), enquanto que todos os anteriores os microssatélites foram desenvolvidos a partir de bibliotecas enriquecidas.

Marcadores microssatélites foram utilizados para estimar e comparar a diversidade genética entre fragmentos florestais, florestas contínuas e plantações, investigar o sistema de acasalamento preferencial e determinar o parentesco e taxas de fluxo gênico entre populações (Bittencourt & Sebbenn, 2007, 2008, 2009; Medina-Macedo et al. , 2015; CM Patreze & SM Tsai, 2010; Sant'Anna et al., 2013; VM Stefenon et al., 2007a; VM Stefenon, Gailing, & Finkeldey, 2008). Devido ao seu dioecismo, taxas de cruzamento próximos à unidade foram relatadas (Bittencourt & Sebbenn, 2008; Ferreira et al., 2012) seja em povoamentos naturais, contínuos ou fragmentados, bem como em plantações comerciais. Dispersão de pólen a distâncias variáveis tem sido relatadas com melhor oportunidade de fluxo gênico para árvores em bordas de fragmentos florestais e predominância de meios irmãos em progênies devido à polinização eólica que favorece a mistura de pólen de diferentes árvores, geralmente não relacionadas (Medina-Macedo et al. al., 2016). Embora um número relativamente pequeno de populações tenha sido estudado até agora, limitado a uma faixa menor que a distribuição geográfica das espécies, geralmente todas as populações mostraram altos níveis de diversidade genética e com baixa diferença entre fragmentos isolados e florestas contínuas, embora fragmentos com muito poucas árvores mostram uma diversidade ligeiramente reduzida (de Souza et al., 2009). Os dados até agora indicaram que a distância a partir de florestas naturais extensas e antigas tem maior impacto na diversidade genética de uma população do que o seu tamanho relativo, de tal forma que mesmo pequenos fragmentos perturbados mantêm altos níveis de diversidade genética se conectados a populações maiores (Medina-Macedo et al., 2016; Medina-Macedo et al., 2015). Apesar das expectativas de uma redução geral drástica da diversidade genética devido à exploração insustentável, os dados genéticos coletados até agora parecem mostrar que A. angustifolia é eficiente em manter sua diversidade genética (Medina-Macedo et al., 2016; VM Stefenon, Steiner, 1998). Guerra, & Nodari, 2009).

No melhoramento genético a disponibilidade de marcadores de DNA acessíveis e informativos, juntamente com o desenvolvimento de métodos de reconstrução de pedigrees tem permitido a conversão de testes de progênies de pedigree incompleto em testes de pedigree completo, eliminando assim as limitações associadas a premissas muitas vezes não cumpridas na estimativa de parâmetros genéticos (Lambeth et al., 2001). A análise de dados quando não se tem a informação completa de pedigree normalmente requer partir de pressupostos quanto à constituição genética das famílias testadas e o número de parentais envolvidos na sua formação, bem como a contribuição de cada parental, ou ainda problemas mais simples como por exemplo a identificação dos materiais coletados. Como esses pressupostos podem não ser realistas na prática, os parâmetros genéticos resultantes e suas inferências são muitas vezes tendenciosas, levando em última análise a vários graus de imprecisão e ineficiência (Askew e El-Kassaby, 1994). A informação genômica derivada de marcadores moleculares por outro lado permite estimar com alta acurácia os parentescos realizados entre qualquer grupo de indivíduos independentemente da sua genealogia e construir uma matriz de parentesco realizado (matriz G) que pode substituir a matriz A nos algoritmos de estimação de componentes de variação e outros parâmetros genéticos (VANRADEN, 2008). Isso torna possível o emprego de metodologias e desenhos experimentais independentes de pedigree (EL-KASSABY; KLÁPŠTĚ; GUY, 2012; HAYES; VISSCHER; GODDARD, 2009; MUÑOZ; RESENDE; GEZAN; RESENDE *et al.*, 2014; THOMAS; COLTMAN; PEMBERTON, 2002; ZAPATA-VALENZUELA; WHETTEN; NEALE; MCKEAND *et al.*, 2013).

O objetivo deste trabalho é realizar uma descoberta de SNPs a partir de dados de sequenciamento e com base nesta informação desenvolver e utilizar um array (microarranjo) de genotipagem de SNPs para *Araucaria angustifolia*. Este *array* será aplicado para geração de dados genotípicos para responder diferentes perguntas relacionadas aos padrões de distribuição e estruturação da variabilidade genética de coleções de populações naturais de diferentes locais na área de ocorrência da espécie. Além disso os dados serão utilizados para estimar parâmetros genéticos a partir de um teste de procedências e progênies e explorar o potencial da predição genômica de características quantitativas de crescimento para, em última análise, acelerar o programa de melhoramento.

REVISÃO DE LITERATURA

A espécie *Araucaria angustifolia* (Bert.) O. Kuntze

Araucaria angustifolia (Bert.) Kuntze, comumente chamada de pinheiro-do-Paraná, é uma conífera endêmica, dióica, perenifólia e dominante com ocorrência localizada majoritariamente na região Sul do Brasil mas também sendo encontrada no estado de São Paulo e sul do estado de Minas Gerais (KLEIN, 1960). Em conjunto com outras espécies arbóreas compõe a formação florestal que recebe o nome de Floresta Ombrófila Mista ou ainda Floresta de Araucária devido à quantidade e ao porte da espécie, que imprime a fisionomia imponente e característica na regiao Sul do país. *A. angustifolia* é polinizada principalmente pelo vento e no sul do Brasil a deiscência dos grãos de pólen ocorre de agosto a novembro, sendo o pico de polinização no mês de setembro.

Apesar de já ter ocupado uma área de aproximadamente 200.000 km^2 a espécie hoje integra a Lista Vermelha de espécies ameaçadas de extinção da IUCN, uma vez que a exploração desordenada principalmente durante o século 20 reduziu sua área de ocorrência natural a aproximadamente 3% do total inicial devido (GUERRA; SILVEIRA; DOS SANTOS; ASTARITA *et al.*, 2000).

A. angustifolia ainda é uma das árvores mais importantes em sua região de ocorrência devido a seu papel ecológico, econômico e social (AULER; REIS; GUERRA; NODARI, 2002). Do ponto de vista ecológico é uma espécie secundária e dominante atuando como espécie berçário no avanço de outras espécies florestais lenhosas sobre os campos adjacentes, pois cria um microambiente para espécies tolerantes à sombra como bromélias, orquídeas e outras espécies epífitas (DUARTE; DOS-SANTOS; HARTZ; PILLAR, 2006). Ainda em relação ao aspecto ecológico, suas sementes, popularmente chamadas de pinhões, possuem alto valor nutricional, contendo aproximadamente 31% de amido e 3% de proteínas além de ser uma fonte de fibras, magnésio e cobre (CORDENUNSI; DE MENEZES WENZEL; GENOVESE; COLLI *et al.*, 2004). As sementes são produzidas em grandes quantidades e dispersadas em uma época de relativa escassez – entre os meses de abril e agosto, servindo de alimento principalmente para aves e pequenos mamíferos (IOB; VIEIRA, 2008).

A alta qualidade da madeira da Araucaria permitiu que fosse usada para diversos fins - construção, movelaria e celulose (CARVALHO, 1994). Sua exploração intensificou-se no começo do século 20 e teve seu auge entre as décadas de 1950 e 1970, quando o metro cúbico da espécie era o produto madeireiro mais importante do Brasil (SHIMIZU; OLIVEIRA, 1981). A coleta e comercialização de pinhão constituem uma importante fonte de renda para famílias dessa região até os dias de hoje. Em 2015 foram produzidas aproximadamente 8.393 toneladas de pinhão gerando uma renda de mais de 21 milhões de reais, sendo 80% desse montante oriundos dos estados do Paraná e de Santa Catarina (IBGE, 2015).

Marcadores moleculares e tecnologias de genotipagem

Por definição, marcadores moleculares são locos genéticos que podem ser detectados e qualificados em uma população e podem ou não estar associados a um gene ou característica de interesse (HAYWARD; TOLLENAERE; DALTON-MORGAN; BATLEY, 2015). Dessa forma surgiram os primeiros marcadores moleculares baseados em variantes alélicas de enzimas – as chamadas isozimas ou isoenzimas (TANKSLEY; ORTON, 1983). Capazes de detectar presença ou ausência de determinado alelo, com custo baixo e metodologia relativamente simples, as isoenzimas foram amplamente utilizadas nas décadas de 80 e 90 em estudos de mapeamento genético, melhoramento e também genética de populações (ANDRÉS; ORTIZ, 1995; COOKE, 1984; STENLID, 1985; STRAUSS; BOUSQUET; HIPKINS; HONG, 1992; TANKSLEY; ORTON, 1983; TANKSLEY; RICK, 1980; WOJNICKA-PÓŁTORAK, 1997).

Entretanto, as limitações quanto ao número de marcadores disponíveis e expressão restrita normalmente a um estágio de desenvolvimento ou tecido específico levaram à utilização cada vez maior de marcadores baseados na análise do DNA, à medida que tecnologias foram desenvolvidas para isso, tais como enzimas de restrição e a reação em cadeia da polimerase (PCR). Marcadores de DNA são detectáveis em qualquer tecido, independentemente do estágio de crescimento, diferenciação ou estágio de desenvolvimento e não são afetados por efeitos de ambiente ou interações entre marcadores (AGARWAL; SHRIVASTAVA; PADH, 2008).

O trabalho publicado por JEFFREYS e FLAVELL (1977) foi o primeiro a descrever a técnica de RFLP (*Restriction fragment length polymorphism*), enquanto o trabalho publicado por BOTSTEIN; WHITE; SKOLNICK e DAVIS (1980) foi o primeiro a usar a ferramenta para construir um mapa genético. Esse tipo de polimorfismo é detectado pela hibridização de uma sonda de DNA marcada com um isótopo e o *Southern Blot* de DNA digerido por enzimas de restrição, gerando então um perfil de fragmentos com tamanhos distintos (AGARWAL; SHRIVASTAVA; PADH, 2008). São marcadores com polimorfismo relativamente alto, co-dominantes e reprodutíveis. Apesar de ter sido usado em inúmeros trabalhos de mapeamento físico e construção de mapas genéticos em plantas, esse tipo de marcador molecular era caro, exigia conhecimento de sequência para a construção das sondas, laborioso, envolvia a utilização de reagentes radioativos e demandava uma grande quantidade de DNA. Essas limitações fizeram com que esse tipo de marcador fosse substituído nos anos 90 por marcadores baseados em PCR (*Polymerase Chain Reaction*), entre eles marcadores RAPD (*Random amplified polymorphic DNA*), marcadores AFLP (*Amplified Fragment Length Polymorphism*) e marcadores microssatélites ou SSR (*Simple sequence repeats*).

A metodologia de detecção de polimorfismo dos marcadores RAPD baseava-se no uso de um iniciador (primer) de sequência arbitrária, de em geral 10 nucleotídeos com conteúdo GC mínimo de 40%, em uma reação de PCR. Apesar do tamanho curto do iniciador fazendo com que ele pudesse se anelar em vários pontos no genoma, polimorfismos de sequência no sítio de anelamento do iniciador reduzia a complexidade do genoma efetivamente amplificado, de forma que era possível detectar marcadores discretos em géis de agarose. Marcadores RAPD são marcadores dominantes – ou seja, não é possível diferenciar os heterozigotos dos homozigotos para aquele loco, assim a genotipagem se dava por presença ou ausência do segmento amplificado em questão. Porém, apesar das vantagens desse tipo de marcador, isto é, baixa quantidade de DNA, metodologia simples, não necessidade de conhecimento genômico prévio da espécie de interesse, o próprio processo de amplificação com iniciadores de sequência arbitrária introduz dificuldade em se reproduzir os resultados em diferentes laboratórios ou até mesmo diferentes termocicladores.

A técnica de AFLP se baseia na amplificação de porções dos fragmentos de restrição usando PCR. O DNA é cortado com enzimas de restrição (uma de corte frequente e outra de corte raro) e os fragmentos são ligado a adaptadores para gerar o molde para a amplificação. Esses adaptadores servem como sitio de anelamento dos iniciadores da PCR os quais possuem algumas bases seletivas no terminal 3' de forma que apenas parte dos fragmentos que foram cortados por ambas as enzimas de restrição são amplificados (VOS; HOGERS; BLEEKER; REIJANS *et al.*, 1995). Essa técnica pode ser aplicada a qualquer espécie e diferentemente dos marcadores RAPD é altamente reprodutível, uma vez que combina a especificidade da digestão com enzimas de restrição com a capacidade de amplificação de DNA da PCR. Apesar de poder ter a genotipagem semi-automatizada com o uso de sequenciadores, necessita de uma maior quantidade de DNA e requer o emprego de mais técnicas e de mais equipamentos do que o RAPD. Foi utilizada em plantas em diversos estudos de diversidade e mapeamento porém uma vez que foi patenteada por uma empresa, sua utilização ficou relativamente restrita (LI; DING; CHU; ZHOU *et al.*, 2008; MBA; TOHME, 2005; RONIKIER, 2002; TATIKONDA; WANI; KANNAN; BEERELLI *et al.*, 2009; VAN EE; JELINSKI; BERRY; HIPP, 2006).

Microssatélites ou SSR (*Simple Sequence Repeats*) são locos polimórficos presentes no DNA que consistem de unidades de um a seis pares de base repetidas em tandem. São encontrados em genomas de procariotos e eucariotos, dispersos ao longo do genoma, podendo ser encontrados tanto em regiões gênicas quanto não-codantes. Microssatélites são amplificados por meio de PCR usando iniciadores complementares às sequencias que flanqueiam essas regiões. Os produtos da PCR são então separados em géis de poliacrilamida ou agarose de alta resolução ou por eletroforese capilar em sequenciadores automáticos, sendo o polimorfismo detectado baseado na diferença do número de repetições da sequência repetitiva. Os polimorfismos no número de repetições do microssatélite é resultado da adição ou deleção de bases causados pelo efeito do escorregamento da DNA polimerase durante a replicação ou ainda causados por erros na recombinação (KASHI; KING; SOLLER, 1997). Por ser um marcador de DNA altamente polimórfico, multi-alélico, co-dominante, experimentalmente reprodutível, podendo ter a genotipagem automatizada e em muitos casos transferível entre espécies relacionadas, marcadores SSR tem sido amplamente usados em estudos de genética de plantas. As mais variadas áreas da genética de plantas foram beneficiadas com o uso de marcadores microssatélites incluindo: genética de populações (MCCOUCH; TEYTELMAN; XU; LOBOS *et al.*, 2002; MENGONI; GORI; BAZZICALUPO, 2000; PATREZE; TSAI, 2010;

ROBERTSON; HOLLINGSWORTH; KETTLE; ENNOS *et al.*, 2004; RYYNANEN; TONTERI; VASEMAGI; PRIMMER, 2007; VAN INGHELANDT; MELCHINGER; LEBRETON; STICH, 2010), estrutura populacional (BELAJ; MUÑOZ-DIEZ; BALDONI; PORCEDDU *et al.*, 2007; DAYANANDAN; DOLE; BAWA; KESSELI, 1999; DU; WANG; WEI; ZHANG *et al.*, 2012; ECKERT; VAN HEERWAARDEN; WEGRZYN; NELSON *et al.*, 2010; ZHANG; LI; LI; LIU *et al.*, 2011), parentesco e análises de paternidade (ALI; RAJEWSKI; BAENZIGER; GILL *et al.*, 2008; ARANZANA; CARBÓ; ARÚS, 2003; BROWN-GUEDIRA; THOMPSON; NELSON; WARBURTON, 2000; CIPRIANI; SPADOTTO; JURMAN; DI GASPERO *et al.*, 2010; DREISIGACKER; ZHANG; WARBURTON; VAN GINKEL *et al.*, 2004; EVANS; PATOCCHI; REZZONICO; MATHIS *et al.*, 2011; SEFC; STEINKELLNER; GLÖSSL; KAMPFER *et al.*, 1998), construção de mapas genéticos (ADAM-BLONDON; ROUX; CLAUX; BUTTERLIN *et al.*, 2004; CREGAN; JARVIK; BUSH; SHOEMAKER *et al.*, 1999; GUPTA; BALYAN; EDWARDS; ISAAC *et al.*, 2002; HONG; CHEN; LIANG; LIU *et al.*, 2010; MCCOUCH; TEYTELMAN; XU; LOBOS *et al.*, 2002; SHEN; GUO; ZHU; YUAN *et al.*, 2005; ZHOU; KOLB; BAI; DOMIER *et al.*, 2003), seleção assistida por marcadores moleculares (BUERSTMAYR; BAN; ANDERSON, 2009; DIRLEWANGER; GRAZIANO; JOOBEUR; GARRIGA-CALDERÉ *et al.*, 2004; GUO; ZHANG; SHEN; YU *et al.*, 2003; JENA; JEUNG; LEE; CHOI *et al.*, 2006; STEELE; PRICE; SHASHIDHAR; WITCOMBE, 2006; ZHANG; YUAN; YU; GUO *et al.*, 2003).

O uso de marcadores microssatélites tem aumentado linearmente desde a sua detecção nos anos 1980 (TAUTZ; RENZ, 1984) e tem sido utilizados extensivamente em diversas aplicações. Porém, o desafio de se gerar e interpretar corretamente os resultados é normalmente subestimado, e a questão de saber se um número limitado de marcadores de microssatélites reflete com precisão a diversidade do genoma continua a ser uma questão importante.

SNPs como marcadores moleculares

Os marcadores SNP (*Single Nucleotide Polymorphism*) são conhecidos desde quando começaram os esforços de sequenciamento de DNA (SANGER; NICKLEN; COULSON, 1977) e em contraste aos microssatélites eram raramente utilizados em estudos de genética de populações até recentemente por causa das limitações de tecnologias que permitissem a descoberta e genotipagem em larga escala, principalmente em organismos não-modelo (HELYAR; HEMMER-HANSEN; BEKKEVOLD; TAYLOR *et al.*, 2011). Porém, nos últimos anos o seu uso tem crescido exponencialmente (GUICHOUX; LAGACHE; WAGNER; CHAUMEIL *et al.*, 2011), graças ao avanço técnico e redução de custos das novas tecnologias de sequenciamento e genotipagem. Essas novas tecnologias facilitaram o acesso a dados genômicos e permitem a descoberta e identificação de milhões de marcadores SNPs bem como a estimação simultânea das frequências alélicas entre indivíduos, populações e espécies (DAVEY; HOHENLOHE; ETTER; BOONE *et al.*, 2011; SCHLOTTERER; TOBLER; KOFLER; NOLTE, 2014).

Um marcador SNP (*Single Nucleotide Polymorphism*) pode ser definido como um sítio do DNA onde é observada a substituição de uma única base entre alelos a um mesmo loco ou entre indivíduos de uma mesma população (RISCH; MERIKANGAS, 1996). Um SNP é entendido ainda como uma posição no DNA onde diferentes sequências alternativas (alelos) co-existem em indivíduos em uma população, e o alelo menos frequente possui frequência maior que 1% (BROOKES, 1999). Ao considerar a frequência alélica na definição de um SNPs, distingue-se esta classe de marcador de meras mutações ao acaso que ocorrem no genoma ou mesmo de erros de sequenciamento que, pela baixa frequência, dificultariam o uso da tecnologia em análise genética.

Em teoria, é possível a existência de quatro alelos diferentes para cada nucleotídeo em um sítio SNP, uma vez que existem quatro bases nitrogenadas que compõem o DNA (A, C, T, G). Porém, na prática, o que se observa é a presença em maior frequência de apenas duas possíveis variações, fato que pode ser explicado pela ocorrência desigual de substituições de base do tipo transição (A ↔ G, T ↔ C) e transversão (A ↔ C, A ↔ T, G ↔ C, G ↔ T). Assim, apesar do número possível de transversões ser duas vezes maior do que o de transições, o que se observa na prática é que a ocorrência de transições é cerca de duas vezes maior do que de transversões (VIGNAL; MILAN; SANCRISTOBAL; EGGEN, 2002). Sendo assim, podemos considerar que em geral marcadores SNP normalmente possuem uma natureza bi-alélica. Além disso a maioria das metodologias de genotipagem baseadas em detecção de fluorescências alternativas permitem somente uma análise bi-alélica destes variantes.

Em princípio, a análise de polimorfismo bi-alélico definido pela variação de um sítio de DNA não é novidade, visto que marcadores RAPD também se baseiam no mesmo tipo de polimorfismo bi-alélico embora nesse caso os fenótipos detectados sejam presença e ausência de um segmento amplificado de DNA. Com isso a herança e segregação de marcadores RAPD é equivalente a um marcador morfológico que apresenta dominância, isto é, não é possível distinguir os genótipos heterozigotos (presença/ausência de banda) dos homozigotos (presença/presença de bandas) fazendo com que o conteúdo informativo seja limitado não permitindo por exemplo, estimativas corretas de endogamia, fluxo gênico e integração de mapas genéticos entre outras. SNPs por outro lado têm comportamento co-dominante, isto é, no genótipo heterozigoto ambos os alelos são identificados. Embora a natureza bi-alélica dos SNPs seja menos informativa do que a hipervariabilidade dos marcadores microssatélites, esta limitação e facilmente compensada pela abundância de SNPs ao longo do genoma e a facilidade de automação da genotipagem (KRUGLYAK, 1997). O SNP é o polimorfismo de DNA mais abundante no genoma e a sua ocorrência e distribuição em genomas de plantas têm sido amplamente estudadas. Em Arabidopsis, por exemplo, observa-se 1 SNP a cada 3.3 kilobases (DRENKARD; RICHTER; ROZEN; STUTIUS *et al.*, 2000); em soja, 1 SNP a cada 200 pb (WANG; GRAEF; PROCOPIUK; DIERS, 2004); em milho são ainda mais frequentes, podendo chegar a 1 SNP a cada 31pb em regiões não codificadoras e 1 SNP a cada 124 pb em regiões codificadoras (CHING; CALDWELL; JUNG; DOLAN *et al.*, 2002). Em coníferas esse número também varia: 1 SNP a cada 166pb em *Pinus contorta* (PARCHMAN; GEIST; GRAHNEN; BENKMAN *et al.*, 2010), 1 SNP a cada 63pb em *Pinus taeda* (BROWN; GILL; KUNTZ; LANGLEY *et al.*, 2004) e 1 SNP a cada 69pb em *Picea abies* (HEUERTZ; DE PAOLI; KÄLLMAN; LARSSON *et al.*, 2006).

A estimativa do número de SNPs observados em uma espécie depende, naturalmente, das relações de vínculo genético entre as amostras de acessos utilizadas na análise. Se as amostras apresentam grande diversidade genética, a tendência é se observar maior número de SNPs a cada kpb analisado.

Tendo em vista à sua abundância no genoma, baixa taxa de mutação e pela possibilidade automatizar o processo de genotipagem, os SNPs vêm sendo cada vez mais utilizados como marcadores moleculares (SALGOTRA; GUPTA; STEWART, 2014). Os marcadores SNPs, vem sendo utilizados em grande escala tanto em estudos de genética e melhoramento de plantas incluindo mapeamento de QTLs (BUERSTMAYR; BAN; ANDERSON, 2009; GRIMMER; KRAFT; FRANCIS; ASHER, 2008), Seleção Assistida por Marcadores (*Marker Assisted Selection* - MAS) (HA;

HUSSEY; BOERMA, 2007; JENA; JEUNG; LEE; CHOI *et al.*, 2006), Estudos de Associação genômica ampla (*Genome Wide Association Studies* - GWAS) (KORTE; FARLOW, 2013; KORTE; VILHJALMSSON; SEGURA; PLATT *et al.*, 2012; LI; YANG; XIAOJING; JIANGWEI *et al.*, 2015; MORRIS; RAMU; DESHPANDE; HASH *et al.*, 2013; ZHU; ZHANG; HU; BAKSHI *et al.*, 2016) e Seleção Genômica (*Genomic Selection* - GS) (GODDARD; HAYES; MEUWISSEN, 2011; HEFFNER; SORRELLS; JANNINK, 2009; MEUWISSEN; HAYES; GODDARD, 2013; RESENDE; RESENDE; SANSALONI; PETROLI *et al.*, 2012; ZHONG; DEKKERS; FERNANDO; JANNINK, 2009); como também vem sendo usados em análises de genética de populações e estudos para conservação de populações naturais.

Graças a uma maior disponibilidade e diminuição do preço das tecnologias de sequenciamento de próxima geração, dezenas de milhares de SNPs tem sido disponibilizado para as mais diversas espécies, embora em sua grande maioria espécies anuais de grandes culturas tais como soja, arroz, milho, trigo, sorgo, cevada ou espécies frutíferas e florestais de grande importância econômica. Para as principais espécies de plantas cultivadas diversos trabalhos de detecção e genotipagem de SNPs baseados em genotipagem de larga escala foram publicados nos últimos anos (GANAL; POLLEY; GRANER; PLIESKE *et al.*, 2012).

Nos últimos anos, os custos de tecnologias de genotipagem de SNPs em arranjos de micro contas (*BeadArrays*) também têm se tornado mais accessíveis e os *chips* de genotipagem foram desenvolvidos permitindo a genotipagem rápida de milhares de SNPs previamente identificados via trabalhos de re-sequenciamento em diversas espécies de plantas (CHAGNÉ; CROWHURST; TROGGIO; DAVEY *et al.*, 2012; CHEN; XIE; HE; YU *et al.*, 2014; GANAL; DURSTEWITZ; POLLEY; BÉRARD *et al.*, 2011; SILVA-JUNIOR; GRATTAPAGLIA, 2015; SONG; HYTEN; JIA; QUIGLEY *et al.*, 2013).

Uma vez identificados os SNPs existe a possibilidade de se construir um *chip* ou um *array* de genotipagem para aquela espécie de interesse. Esse conjunto de marcadores poderá ser usado por diferentes grupos e objetivos sem a necessidade de se descobrir os marcadores novamente. Hoje em dia duas plataformas se destacam quando a idéia é se genotipar com alta qualidade e reproducibilidade: Infinium® (Illumina) e Axiom® (Affymetrix).

A plataforma Infinium da Illumina é um sistema de genotipagem de SNP de alta performance que permite a genotipagem de até 2,5 milhões de SNPs por amostra de

DNA. Ao contrário do ensaio Golden Gate, também da Illumania mas agora descontinuado, que usava primers universais para amplificar os fragmentos de DNA, o ensaio Infinium depende da hibridação direta de alvos genômicos com sequências presas fisicamente ao *array*. A extensão de base única é seguida por coloração fluorescente, amplificação de sinal, digitalização e análise usando o software proprietário chamado Genome Studio. Devido ao elevado número de SNPs que podem ser analisados em um único chip Infinium, a simplicidade de preparação de amostras e relativa facilidade de análise de dados, esta plataforma foi amplamente utilizada nos mais diversos estudos tanto humano quanto animal ou ainda com plantas.

A Illumina também oferece a opção de produzir chips personalizados Infinium, ou também chamados de iSelect, que podem conter entre 3.000 e 200.000 SNPs a serem genotipados por amostra. Isso dá a opção ao usuário de ter análises genéticas personalizadas a um nível não disponível antes em outras plataformas.

A plataforma de genotipagem Axiom é um ensaio baseado em sondas de oligonucleótidos 30-mer, sintetizadas *in situ* em um substrato sólido em esquema de microarranjo, com processamento automatizado e paralelo de 96 ou 384 amostras. O chip se caracteriza por quadrados *features* de 3 μm, num campo de 5 μm de centro para centro, com um total de ~ 1,38 milhão de *features* disponíveis. Cada *feature* de SNP contém uma sequência oligonucleotídica única complementar à sequência genômica que flanqueia o SNP na cadeia *forward* ou *reverse*. As sondas na solução que suportam os locais de ligação para um dos dois corantes, dependendo da base 3' (A ou T, versus C ou G) hibridizam com o complexo alvo/sonda e são então ligadas. Os *features* normalmente são repetidos duas vezes na matriz, de modo que cada SNP é interrogado por pelo menos dois *features*. Um máximo de ~690,000 SNPs podem ser acomodado neste formato. Este número é reduzido se os SNPs A/T ou C/G estiverem incluídos, ou se *features* adicionais forem usados para melhorar a resolução de SNPs específicos.

Descoberta de SNPs

Atualmente existem diversas estratégias baseadas em NGS (Next Generation Sequencing) para descoberta em larga escala de SNPs se comparando sequências de DNA de diferentes indivíduos. Uma dessas estratégias compara o transcriptoma de diversos indivíduos após a transcriptase reversa de RNA mensageiro (BARBAZUK; EMRICH; CHEN; LI *et al.*, 2007; HASENEYER; SCHMUTZER; SEIDEL; ZHOU *et al.*, 2011; HIREMATH; FARMER; CANNON; WOODWARD *et al.*, 2011; NOVAES; DROST; FARMERIE; PAPPAS *et al.*, 2008).

A vantagem desse método é a identificação de SNPs localizados basicamente em genes, que podem ser genes de cópia única no genoma o que é uma vantagem ainda maior. Essa vantagem porém limita a amostragem do genoma onde serão detectados os polimorfismos o que pode não ser o ideal se há necessidade de se descobrir dezenas de milhares de polimorfismos.

Métodos que combinam NGS com redução de complexidade do genoma estão em alta nos últimos anos. Reduções de complexidade genômica se baseiam no seleção da fração de DNA que será sequenciada. Essa seleção pode ser derivada de digestão com enzimas de restrição sensíveis a metilação (DESCHAMPS; LA ROTA; RATASHAK; BIDDLE *et al.*, 2010; GORE; CHIA; ELSHIRE; SUN *et al.*, 2009; GORE; WRIGHT; ERSOZ; BOUFFARD *et al.*, 2009), pré-amplificação de combinações de *primers* AFLP específicos (VAN ORSOUW; HOGERS; JANSSEN; YALCIN *et al.*, 2007) ou ainda com o uso de RAD (*Restriction-site Associated DNA*) (DAVEY; HOHENLOHE; ETTER; BOONE *et al.*, 2011). Em teoria, todos os três métodos podem ser utilizados com qualquer organismo alvo independentemente do tamanho do genoma, resultando em milhares de marcadores distribuídos ao longo do genoma de interesse e amostrando mais regiões que não apenas regiões gênicas.

Quando falamos de NGS podemos dividir as etapas de bioinformática basicamente em três passos: alinhamento de sequências e detecção de SNPs (MIELCZAREK; SZYDA, 2016). O primeiro passo consiste em se alinhar as sequências curtas geradas pelo sequenciamento – *reads* - a um genoma de referência existente ou se montar uma referência *de novo* a qual os *reads* possam ser alinhados. Em geral, esse alinhamento começa com uma indexação das sequências – podendo ser tanto das sequências de referência e/ou dos reads. Alinhadores populares como o Bowtie2 (LANGMEAD; SALZBERG, 2012) e SOAP3-dp (LIU; WONG; WU; LUO *et al.*, 2012) indexam as sequências de referência, o que é vantajoso computacionalmente uma vez que só se faz necessario fazer uma única vez enquanto indexar os *reads* precisa ser feito para cada amostra separadamente. Ambos os alinhadores citados utilização um sistema de indexação índice-FM que combina o algoritmo de transformação de Burrows-Wheeler (BWT) com estruturas de dados auxiliares.

Quanto a chamada dos SNPs ainda podemos dividir em dois passos: chamada propriamente dita dos SNPs e a genotipagem em si. Quando se analisa uma única amostra a chamada dos SNPs e a genotipagem são similares já que locus em heterozigose ou homozigose para o alelo alternativo indicam a presença de um SNP, porém quando se analisam diversas amostras simultaneamente o SNP é identificado se pelo menos uma amostra é heterozigota ou homozigota para o alelo alternativo. Sendo assim, a chamada de SNP pode ser definida como o processo de se identificar sítios que diferem da sequência de referência enquanto a genotipagem se refere a estimação dos genótipos propriamente ditos de cada amostra (NIELSEN; PAUL; ALBRECHTSEN; SONG, 2011). Os softwares de chamada de SNP podem usar métodos heurísticos (baseados em múltiplas fontes de informação associadas a estrutura e qualidade dos dados) ou ainda probabilísticos (fornecem medidas de incerteza estatística para os genótipos chamados, possibilitando monitorar a precisão de chamada de genótipos).

A chamada de genótipos baseia-se em cálculos de verossimilhança de genótipos e utiliza o teorema de Bayes. Após as etapas de pré-processamento (realinhamento e a recalibração), o próximo passo calcula a verossimilhança para cada genótipo possível em cada base (um homozigoto para o alelo de referência, um homozigoto para o alelo alternativo ou um heterozigoto). Baseando-se nos dados de qualidade e contagens de alelos para cada SNP. Na estrutura bayesiana, a probabilidade calculada é combinada com uma probabilidade *a priori* do genótipo, o que leva a uma probabilidade *a posteriori* de um genótipo. Como resultado, o genótipo com maior probabilidade *a posteriori* é escolhido. A relação entre as probabilidades mais alta e a segunda maior pode ser usada como medida de confiança. Alguns dos softwares mais populares para chamada de SNP e genotipagem utilizam o método Bayesiano, como por exemplo: SAMtools (LI; HANDSAKER; WYSOKER; FENNELL *et al.*, 2009), bcftools (LI; LI; FANG; YANG *et al.*, 2009) e GATK (MCKENNA; HANNA; BANKS; SIVACHENKO *et al.*, 2010).

Marcadores moleculares para *Araucaria angustifolia*

A caracterização da diversidade genética de *Araucaria angustifolia* foi por muito tempo realizada exclusivamente com base na mensuração de caractéres morfológicos e quantitativos, os quais são afetados pelo ambiente e pela própria metodologia de coleta de dados utilizada. No começo da década de 80 diversos trabalhos avaliaram a diversidade genética em *A. angustifolia para características* como crescimento volumétrico. Conforme esperado, estes trabalhos encontraram diferenças genéticas importantes entre indivíduos dentro de procedências e entre diferentes procedências (GURGEL-FILHO, 1980; KAGEYAMA; JACOB, 1980; SHIMIZU; HIGA, 1980).

O desenvolvimento e uso de marcadores moleculares isentos de efeito ambiental na determinação da diversidade genética de populações permitiu um importante avanço na caracterização e conservação de recursos genéticos da espécie. Inicialmente trabalhos usando marcadores isoenzimáticos foram utilizados revelando uma alta diversidade genética dentro de populações e baixa entre populações naturais (AULER; REIS; GUERRA; NODARI, 2002; MANTOVANI; MORELLATO; DOS REIS, 2006; SHIMIZU; JAEGER; SOPCHAKI, 2000). AULER; REIS; GUERRA e NODARI (2002) relatou ainda uma variação nas taxas de endogamia nas diferentes populações e a presença de alelos raros apenas em algumas das populações analisadas no trabalho.

Com os avanços técnicos crescentes nesta área nas últimas décadas alguns marcadores baseados na análise de DNA se tornaram disponíveis para caracterizar populações naturais. A partir do final da década de 90 diversos trabalhos que visavam caracterizar a diversidade genética em *A. angustifolia* usando marcadores moleculares comecaram a ser publicados. MAZZA (1997) usando marcadores RAPD encontrou similaridade genética, mesmo que baixa, entre populações distantes geograficamente. Também usando marcadores RAPD, MEDRI; RUAS; HIGA; MURAKAMI *et al.* (2003) comparou a diversidade genética de uma população natural, uma população manejada e um teste de progênie, não encontrando diferença significativa na diversidade genética média de cada população e também encontrando um baixo indíce de diferenciação genética entre as mesmas.

Marcadores microssatélites foram em seguida desenvolvidos e utilizados para análises genéticas de Araucaria. SCOTT; SHEPHERD e HENRY (2003) desenvolveram primers para 10 marcadores microssatélites em *A. cunninghamii*. ROBERTSON; HOLLINGSWORTH; KETTLE; ENNOS *et al.* (2004) desenvolveram primers para 5 marcadores microssatélites em *A. columnari*. SALGUEIRO; CARON; DE SOUZA; KREMER *et al.* (2005) desenvolveram primers para mais 6 microssatélites em *A. angustifolia* e *A. araucana*., SCHMIDT; CIAMPI; GUERRA e NODARI (2007) desenvolveram primers para 29 microssatélites em *A. angustifolia* e MARTIN; MATTIONI; LUSINI; DRAKE *et al.* (2012) para mais 10 marcadores microssatélites porém apenas 8 deles foram polimórficos em *A. angustifolia*. Este último trabalho foi o único feito utilizando dados de sequenciamento de próxima geração (Next generation sequencing), enquanto que todos os anteriores os microssatélites foram desenvolvidos a partir de bibliotecas enriquecidas.

Estimativas de diferenciação genética entre populações

Um dos principais parâmetros de interesse ao se estudar a estrutura de populações naturais é a divergência genética estimada pelo índice de fixação Fst ou Gst. Este parâmetro é frequentemente utilizado para guiar proposições de conservação de populações por meio de coleta de germoplasma ou estabelecimento de reservas genéticas. Apresentado primeiramente por WRIGHT (1949) e MALÉCOT (1948), o Fst foi originalmente desenvolvido como coeficiente de endogamia. Ele se baseia nas diferenças nas frequências alélicas entre populações e na probabilidade de identidade por descendência. Além da descrição original feita por Wright, o Fst também foi definido por diversos outros autores (COCKERHAM, 1969; HUDSON; SLATKIN; MADDISON, 1992; NEI, 1973; SLATKIN, 1991) e múltiplos estimadores de Fst foram descritos na literatura (HOLSINGER, 1999; HUDSON; SLATKIN; MADDISON, 1992; NEI, 1973; 1986; WEIR; HILL, 2002).

Em resumo, o Fst estima o quanto da diversidade genética entre populações pode ser explicada pela estrutura. Seus valores variam de 0 a 1 onde um valor de Fst = 0 indica panmixia completa e que as subpopulações se intercruzam livremente (sem estrutura), e onde um valor de Fst = 1 indica que toda a diversidade genética se deve a estrutura genética e que as subpopulações não compartilham diversidade genética.

O Gst é um estimador semelhante ao Fst porém derivado explicitamente para lidar com alelos múltiplos por loco (NEI, 1973) e exceto por alguns detalhes no processo de estimação, o Gst é equivalente ao Fst, sendo definido nos termos de Hs (heterozigosidade dentro das subpopulações) e Ht (heterozigosidade total de todas as populações), sendo Gst = (Ht-Hs)/Ht (WHITLOCK, 2011).

Diversos trabalhos recentes apontaram a dificuldade de se comparar e interpretar estimativas de Gst (HEDRICK, 2005; JOST, 2008; JOST, 2009; MEIRMANS; HEDRICK, 2011). Considerando duas populações sendo avaliadas com base em dois alelos, o Gst varia entre 0 e 1, assim como o esperado, porém ao se usar mais de dois alelos o Gst não pode atingir o valor de 1 nem mesmo quando nenhum alelo é compartilhado entre as duas subpopulações uma vez se houver mutação sempre haverá heterozigosidade dentro das subpopulações (WHITLOCK, 2011). Isso dificulta a comparação entre estudos já que os marcadores utilizados podem ser diferentes e sendo assim teriam diferentes valores de máximo obtível para o Gst para cada marcador. Considerando que marcadores microssatélites em sua maioria possuem alta taxa de mutação pode-se dizer que para alguns loci microssatélites as estimativas de Gst serão muito menores quando comparadas as estimativas obtidas com outros tipos de marcadores, como por exemplo SNPs. Isso em si não indica um problema com o estimador propriamente dito, mas é uma indicação de cautela ao se interpretar Gst ou comparar as estimativas obtidas com diferentes tipos de marcadores moleculares.

Em casos onde se tem marcadores microssatélites com alta heterozigosidade o valor de Gst máximo obtido normalmente varia entre 0.1 e 0.2, o que pode levar a conclusões erradas se considerarmos as classes de valores sugerida por Wright onde um valor entre 0 e 0.05 indica pouca diferenciação genética, entre 0.05 e 0.15 indica uma diferenciação moderada e 0.15 a 0.25 indica uma alta diferenciação. Para tentar resolver esse problema HEDRICK (2005) propôs um estimador padronizado G'st, que pode ser calculado se dividindo o Gst para um dado marcador pelo valor máximo teórico do Gst baseado na heterozigosidade daquele marcador. Também tentando resolver esse problema, JOST (2008) introduziu outro estimador de diferenciação, D, que mede a fração da variação alélica entre populações. Em ambos os estimadores, G'st e D serão iguais a 1 em situações de diferenciação completa entre populações (mesmo em populações com alta variabilidade dentro da população) e igual a zero quando não houver diferenciação.

Estudos de genética de populações em *Araucaria angustifolia*

Na maioria dos organismos a diversidade genética pode ser observada tanto em niveis morfológicos quanto moleculares. A diversidade genética é o principal fator que impacta a manutenção e evolução de populações, espécies e em última análise de ecossistemas. A diversidade genética representa o potencial de determinada espécie sobreviver em um ambiente em constante mudança. Estudos de diversidade genética em populações naturais compreendem a descrição dos níveis de variação genética existente dentro e entre populações e a forma com que essa variação é estruturada entre populações. Informações sobre a diversidade genética e o conhecimento da estrutura genética das populações naturais ajudam a definir estratégias de conservação e gerenciamento responsável da espécie bem como subsidiar ações de domesticação e melhoramento genético (REIS; GRATTAPAGLIA, 2004).

Com a disponibilidade dos cerca de 60 marcadores microssatélites para espécies de Araucaria diversos trabalhos foram publicados estudando o fluxo gênico entre populações bem como realizando análises comparativas da diversidade genética entre populações e efeitos de fragmentação (BITTENCOURT; SEBBENN, 2007; BITTENCOURT; SEBBENN, 2008; BITTENCOURT; SEBBENN, 2009; DANNER; RIBEIRO; ZANETTE; BITTENCOURT *et al.*, 2013a; b; MEDINA-MACEDO; SEBBENN; LACERDA; RIBEIRO *et al.*, 2014; PATREZE; TSAI, 2010; SANT’ANNA; SEBBENN; KLABUNDE; BITTENCOURT *et al.*, 2013; STEFENON; BEHLING; GAILING; FINKELDEY, 2008; STEFENON; GAILING; FINKELDEY, 2007; STEFENON, VALDIR MARCOS; GAILING, OLIVER; FINKELDEY, REINER, 2008; STEFENON, V. M.; GAILING, O.; FINKELDEY, R., 2008).

Usando marcadores AFLP e SSR, STEFENON (2007) analisou a distribuição da diversidade genética em populções naturais de diferentes regiões do Brasil. A análise revelou uma alta taxa de diversidade gênica, uma diferenciação moderada porém uma divergência pronunciada entre as populações mais a norte, isoladas geograficamente. A distância genética entre essas populações aumentava de acordo com o aumento da distância geográfica. O autor ainda analisou a diversidade genética, diversidade gênica e riqueza alélica em florestas plantadas de *A. angustifolia* quando comparadas a populações naturais – os resultados sugerem que a diversidade gênica e a riqueza alélíca foram significativamente superiores em florestas plantadas em relação a populações naturais, enquanto o grau de endogamia não diferiu entre as mesmas quando usados marcadores microssatélites. O autor mostrou ainda que em geral a estrutura genética da população original não foi fortemente alterada nas florestas plantadas quando comparada às populações originais. O que pode ser explicado pelo fato que os efeitos da diminuição do número de indivíduos e a fragmentação das populações na diversidade genética só são percebidos após várias gerações. Uma vez que a exploração desenfreada ocorreu a menos de cem anos, e *A. angustifolia* é uma árvore com longo ciclo de vida, ainda não houve tempo suficiente para se detectar essa diminuição na diversidade genética.

Para analisar os efeitos da fragmentação florestal na diversidade e estrutura de *A. angustifolia* em populações no Sul do Brasil, BITTENCOURT e SEBBENN (2009) compararam dados genotípicos de oito marcadores SSR em quatro populações fragmentadas pequenas, quatro grupos de árvores encontradas em pastagens e em três parcelas em uma população contínua grande. O efeito mais claro da fragmentação encontrado foi a perda de alelos raros ($p \leq 0.05$) e de frequência intermediária ($0.05 < p \leq 0.25$) em populações fragmentadas e a perda de alelos raros em grupos de árvores em pastagens quando comparado à população contínua. Os autores também encontraram que populações fragmentadas possuem um índice de fixação significativamente maior que populações contínuas - Fis = 0.121 e 0.083 respectivamente, e também que a maior diferenciação genética foi detectada entre grupos de árvores em pastagens (G'st = 0.258, $P<0.01$), seguido das populações fragmentadas (G'st = 0.031, $P<0.05$) e por fim as populações contínuas (G'st = 0.026, $P<0.05$). A hipótese de que esses fragmentos florestais são resultados de um *bottleneck* recente (redução no tamanho efetivo da população) ao invés da hipótese de populações historicamente pequenas foi confirmada em dois dos quatro fragmentos estudados. Esses resultados corroboram os resultados também encontrados usando aloenzimas por AULER; REIS; GUERRA e NODARI (2002) – perda de alelos, *bottlenecks* e aumento na taxa de *inbreeding* e diferenciação entre fragmentos. Ambos os trabalhos também relatam que a maior parte da diversidade genética em *A. angustifolia* pode ser encontrada dentro de populações e não entre populações e que populações mais conservadas apresentam maior diversidade de alelos por loco, maior polimorfismo e heterozigosidade observada.

Conservação, domesticação e melhoramento genético de *Araucaria angustifolia*

Com a exploração predatória da espécie e a consequente introdução desta na lista das espécies ameaçadas de extinção, categoria vulnerável, o Governo Federal instituiu a proibição do corte de Araucária nativa (Resolução Nº 278 do CONAMA, em

24 de maio de 2001). Com essa restrição ao uso da espécie, regenerações tem sido eliminadas pois são consideradas como um empecilho para uso futuro das propriedades. Áreas de Preservação Permanente devem existir para assegurar o potencial de evolução das espécies existentes, mas somente essas áreas não garantirão a conservação de todo o recurso necessário para os programas futuros. De acordo com SOUSA e AGUIAR (2012) a conservação "*on farm*", ou conservação pela comunidade é a melhor opção para a conservação no momento.

Atualmente, grande parte dos esforços tanto de conservação quanto de melhoramento da espécie tem sido feito pela Embrapa Florestas que trabalha com essa espécie desde a década de 1970. Visando a conservação genética da espécie, a Embrapa Florestas implantou vários bancos de conservação de germoplasma. Inicialmente, esses bancos eram formados por uma mistura de sementes de várias procedências, visando estimular o cruzamento entre si, para promover as recombinações entre possíveis raças geográficas e preservar a variabilidade genética. O programa de conservação da Embrapa também visa a manutenção das variedades separadamente, bem como, ecótipos de regiões contrastantes; a manutenção das populações de conservação na região de origem (que serão material futuro para o melhoramento), para preservar genes de adaptação local, importantes para uma maior produção e resistência; e também a coleta de novos materiais para o enriquecimento da base genética e assegurar o avanço de programas de melhoramento genético em longo prazo SOUSA e AGUIAR (2012).

Apesar da importância econômica da espécie, ainda não se tem um programa de melhoramento em estágio avançado considerando o longo prazo e visando maior produtividade, qualidade da madeira e produção de pinhão. Embora algumas empresas florestais e instituições de pesquisa tenham investido muito nos plantios com essa espécie, nem sempre os mesmos foram baseados em programas de melhoramento próprios. Sendo assim, não existe disponível no mercado materiais melhorados provenientes de programas de melhoramento – seja público ou privado.

Visando identificar material geneticamente superior para plantio, instituições como a Embrapa e o Instituto Florestal de São Paulo implantaram no início da década de 1980 diversos testes de procedências/progênies isolados. Porém, estes não compunham um programa maior ou uma rede experimental. Diversas pesquisas mostraram diferenças significativas entre procedências para caracteres quantitativos e entre esses trabalhos destacam-se os de (BALDANZI; RITTERSHOFER;

REISSMAN, 1973; GURGEL-FILHO, 1980; GURGEL; GURGEL FILHO, 1973; KAGEYAMA; JACOB, 1980; SEBBENN, A.; PONTINHA, A.; GIANNOTTI, E.; KAGEYAMA, P., 2003; SEBBENN; PONTINHA; FREITAS; FREITAS, 2004; SEBBENN, A. M.; PONTINHA, A. D. A. S.; GIANNOTTI, E.; KAGEYAMA, P. Y., 2003; SHIMIZU; HIGA, 1980).

Apesar de existirem esforços nas áreas de conservação e também de melhoramento florestal, outras áreas como silvicultura, manejo, clonagem, legislação etc., dificultam o desenvolvimento de estudos com a araucária, que no momento, não consegue competir no que se refere à produtividade de madeira com gêneros introduzidos como por exemplo *Eucalyptus* e *Pinus*. Apesar disso tudo, mesmo sem nenhum grau de melhoramento, de acordo com (CARVALHO, 1994), a partir do terceiro ano de idade, em sítios adequados, a araucária pode apresentar um incremento médio anual em altura de 1 m e, a partir do quinto ano, de 1.5 m a 2.0 m, sendo que o incremento em volume pode atingir 30 m^3 ha^{-1} ano^{-1} o que evidencia o potencial da espécie em plantios comerciais.

Estimativa de parâmetros genéticos com marcadores moleculares

O melhoramento de plantas, incluídas as espécies arbóreas, normalmente segue o esquema clássico de seleção recorrente, que é caracterizado por ciclos repetitivos de cruzamento, teste e seleção (ALLARD, 1999; NAMKOONG; KANG; BROUARD, 2012). Esses programas em sua maioria lidam com múltiplas populações e um grande número de parentais e progênies, plantados em múltiplos locais e em diferentes anos. Para diminuir os esforços associados a geração de progênies com informação completa de pedigree, melhoristas tem adotado protocolos mais simplificados, variando entre alguns protocolos que não necessitam de informação de pedigree (ex.: testes de procedência) a alguns que necessitam apenas da informação incompleta de pedigree (ex.: populações de polinização aberta ou meios-irmãos).

Porém, a análise de dados quando não se tem a informação completa de pedigree normalmente requer aceitar pressupostos quanto a constituição genética das famílias testadas e o número de parentais envolvidos na sua formação, bem como a contribuição de cada parental, ou ainda problemas mais simples como por exemplo a identificação dos materiais coletados. Como esses pressupostos podem não ser realistas na prática, os parâmetros genéticos resultantes e suas inferências são muitas vezes tendenciosas, levando em última análise a vários graus de imprecisão e ineficiência (ASKEW; EL-KASSABY, 1994).

A disponibilidade de marcadores de DNA acessíveis e altamente informativos, juntamente com o desenvolvimento de métodos sofisticados de reconstrução de pedigree, aumentou a utilidade na conversão de ensaios de pedigree incompletos em testes efetivamente completos, eliminando assim as armadilhas associadas à utilização de pressupostos não cumpridos (LAMBETH; LEE; O'MALLEY; WHEELER, 2001).

Estimativas de parâmetros genéticos de características quantitativas, como por exemplo herdabilidade, são importantes porque dão uma indicação da habilidade da espécie responder à seleção e também do potencial da espécie evoluir (LANDE; SHANNON, 1996). Entretanto, o método tradicional de se estimar parâmetros genéticos requer conhecimento do relacionamento (ou parentesco) entre os indivíduos avaliados (WRIGHT, 1922), o que na maioria dos estudos com populações de melhoramento conduzidas sem controle paterno não é conhecido. Nesses casos, ferramentas baseadas em marcadores moleculares podem ajudar na inferência desses relacionamentos e na reconstrução do pedigree e na construção de uma matriz de parentesco realizado (EL-KASSABY; CAPPA; LIEWLAKSANEEYANAWIN; KLÁPŠTĚ *et al.*, 2011; GAMAL EL-DIEN; RATCLIFFE; KLAPSTE; PORTH *et al.*, 2016; QUELLER; GOODNIGHT, 1989; THOMAS; HILL, 2000).

Os coeficientes de parentescos entre indivíduos baseados em pedigree (matriz A) são comumente usados para se estimar componentes de variância genética se usando o método de Máxima Verossimilhança Restrita – REML [*Restricted Maximum Likelihood*] (GILMOUR; THOMPSON; CULLIS, 1995) e predizer o valor genético de cada indivíduo através de algoritmos de Melhor Predição Desenviesada Linear, ou BLUP [*Best Linear Unbiased Prediction*] (HENDERSON, 1976). Apesar de ser um método eficiente, a matriz de parentesco baseada em pedigree ignora a variação entre membros das mesmas famílias e assume o valor médio da família para todos os indivíduos da mesma – irmãos-completos diferem no grau parentesco (HILL; WEIR, 2011).

A informação genômica permite estimar com alta acurácia os parentescos realizados entre qualquer grupo de indivíduos independentemente da sua genealogia e construir uma matriz de parentesco realizado (matriz G) que pode substituir a matriz A nos algoritmos de estimação de componentes de variação e outros parâmetros genéticos (VANRADEN, 2008). Isso torna possível o emprego de metodologias e desenhos experimentais menos complicados e completamente independentes de pedigree (EL-KASSABY; KLÁPŠTĚ; GUY, 2012; HAYES; VISSCHER; GODDARD, 2009; MUÑOZ; RESENDE; GEZAN; RESENDE *et al.*, 2014; THOMAS; COLTMAN; PEMBERTON, 2002; ZAPATA-VALENZUELA; WHETTEN; NEALE; MCKEAND *et al.*, 2013)

As ferramentas baseadas em dados moleculares para se inferir o parentesco podem ser agrupadas em duas categorias: estimadores que usam o método dos momentos – usados para estimar relacionamento como uma medida contínua com base em alelos compartilhados (LYNCH, 1988; LYNCH; RITLAND, 1999; QUELLER; GOODNIGHT, 1989), e técnicas de verossimilhança – que determinam a verossimilhança de um par cair em uma determinada classe de relacionamento, exemplo: *full-sibs* ou *nonsibs*, dado a informação genotípica observada (MOUSSEAU; RITLAND; HEATH, 1998; THOMPSON; BROTHERSTONE; WHITE, 2005). RITLAND (1996) propôs uma abordagem de regressão para a estimação de parâmetros, em que as medidas de similaridade fenotípica são regredidas contra o parentesco par-a-par. Porém, se existe informação sobre a estrutura populacional, procedimentos baseados em verossimilhança devem ser usados, onde cada par é colocado em uma estrutura populacional predeterminada de acordo com a probabilidade de observar os seus genótipos e fenótipos (MOUSSEAU; RITLAND; HEATH, 1998; THOMAS; PEMBERTON; HILL, 2000).

De acordo com THOMAS; PEMBERTON e HILL (2000), há perda de informação quando se usa técnicas que utilizam comparações par-a-par. Por exemplo, se três indivíduos são amostrados em uma única geração e têm os cada um tem os seguintes genótipos a_ia_i, a_ja_j, a_ka_k (sendo a_i, a_j e a_k alelos mutualmente exclusivos) eles não podem ser irmãos-completos, mas pelo método par-a-par essa exclusão não é possível.

Melhoramento Florestal em Coníferas

O melhoramento florestal é parte integrante da silvicultura moderna, usado visando maximizar o rendimento econômico através da produção aprimorada de madeira (WHITE; ADAMS; NEALE, 2007). O melhoramento florestal aborda principalmente espécies de importância econômica e para as quais regeneração artificial, por plantação ou por semeadura direta de variedades melhoradas, é usado para florestação ou reflorestamento. Variedades melhoradas são usadas mais fortemente em florestas industriais mas também em agro-silvicultura e, em alguns casos, para enriquecimento de florestas locais (PÂQUES, 2013). Uma produção rápida (tempo de rotação curto) e de alto rendimento de madeira de boa qualidade é frequentemente a prioridade em um programa de melhoramento florestal. O melhoramento de árvores acelera a adaptação das árvores às mudanças de ambientes (REHFELDT; JAQUISH; SÁENZ-ROMERO; JOYCE *et al.*, 2014) modificando a composição genética das populações de determinada espécie para melhor atender às necessidades do homem.

Comparado ao melhoramento de espécies anuais, o melhoramento florestal apresenta desafios mais complexos como: ciclos de reprodução mais longos, floração tardia, correlações juvenil-adulto fracas (GRATTAPAGLIA; SILVA-JUNIOR; RESENDE; CAPPA *et al.*, 2018). Além do que espécies florestais geralmente apresentam fenologia reprodutiva, criando problemas adicionais relacionados ao planejamento de cruzamentos e produção das progênies necessárias para os testes a campo; e apresentam a expressão tardia das características de interesse, uma vez que a maioria das características econômicas, como a qualidade da madeira, são avaliadas em idades avançadas. Apesar dos desafios, existem muitos casos de sucesso de melhoramento para diferentes características fenotípicas em diversas espécies de coníferas – *Pinus pinaster* (BOUFFIER; CHARLOT; RAFFIN; ROZENBERG *et al.*, 2008; BUTCHER, 2007; BUTCHER; HOPKINS, 1993), *Pinus radiata* (CHAUHAN; SHARMA; THOMAS; APIOLAZA *et al.*, 2013; KUMAR; LEE, 2002; WU; POWELL; YANG; IVKOVIĆ *et al.*, 2007), *Pinus taeda* (ADAMS; LAND JR; BELLI; MATNEY, 2008; GWAZE, 2009; GWAZE; BRIDGWATER; WILLIAMS, 2002; GWAZE; BYRAM; LOWE; BRIDGWATER, 2001; XIANG; LI; ISIK, 2003), *Pinus sylvestris* (HAAPANEN; HYNYNEN; RUOTSALAINEN; SIIPILEHTO et al., 2016; JANSSON; HANSEN; HAAPANEN; KVAALEN et al., 2017; RIEKSTS-RIEKSTIŅŠ; ZELTIŅŠ; BALIUCKAS; BRŪNA et al., 2020), *Picea abies* (ERIKSSON, 2010; ROSVALL, 2011; SKRØPPA; STEFFENREM, 2016), *Pseudotsuga menziesii* (COPES, 1999; ISAAC-RENTON; STOEHR; BEALLE STATLAND; WOODS, 2020; STONECYPHER; PIESCH; HELLAND; CHAPMAN et al., 1996; YANCHUK, 1996; YE; JAYAWICKRAMA, 2012).

No melhoramento florestal o tempo é a maior restrição e tem consequências óbvias na avaliação fenotípica do material genético, portanto as estratégias de melhoramento visam encurtar o ciclo reprodutivo e otimizar os ganhos genéticos por unidade de tempo. O progresso na iniciação e estimulação de floração por meios mecânicos, como a cinta do caule e tratamentos químicos, como a injeção de giberelinas ou mesmo paclobutrazol, aceleraram a fase de recombinação em várias espécies naturalmente reprodutoras tardias (HASAN; REID, 1995). O desenvolvimento de preditores indiretos precoces, a seleção assistida por marcadores (MAS - Molecular Assisted Selection) por exemplo, tinham uma perspectiva boa de acelerar os programas de melhoramento florestal e diminuir o esforço de teste a campo, porém dado que uma das pressuposições da seleção assistida por marcadores é a existência de desequilíbrio de ligação entre o marcador e QTL de interesse, e dado o nível de heterozigosidade das populações de árvores florestais em equilíbrio de ligação, adicionado às interações entre QTL e background genético, o uso de MAS no melhoramento florestal se mostrou limitado (GRATTAPAGLIA, 2014). À medida que mais indivíduos por família e mais famílias são analisados, o poder de detecção aumenta, mais QTLs são descobertos, o efeito estimado de cada um diminui e a inconsistência desses efeitos se torna mais evidente, e com um número maior de QTLs controlando cada característica, e cada QTL com um efeito pequeno e imprevisível variável, a probabilidade de implementar o MAS para várias características simultaneamente é praticamente excluída (GRATTAPAGLIA, 2014). Nesse contexto, a seleção genômica surge com perspectivas promissoras no melhoramento de árvores para aumentar o ganho genético por unidade de tempo, através de uma melhor estimativa dos valores de melhoramento genético (seleção dos pais) e genotípicos (seleção de clones) e redução do tempo de geração.

Seleção Genômica (GS) no melhoramento florestal

Recentemente, a utilização de metodologias envolvendo o desenvolvimento de modelos de predição de fenótipos complexos com base na genotipagem ampla do genoma tem surgido, revolucionando a perspectiva da aplicação de informações genômicas na prática da seleção. Esta metodologia dispensa a necessidade de identificação prévia de QTLs/genes individuais, focando exclusivamente nos aspectos de eficiência operacional e ganho genético. Este tipo de abordagem, denominado seleção genômica (GS, *Genomic Selection*) ou seleção genômica ampla (GWS, *Genome-Wide Selection*), foi proposto alguns anos atrás (MEUWISSEN; HAYES; GODDARD, 2001) e vem ganhando interesse e aplicação crescente como uma nova abordagem do melhoramento de plantas cultivadas anuais (ANNICCHIARICO; NAZZICARI; LI; WEI *et al.*, 2015; CROSSA; PEREZ; HICKEY; BURGUEÑO *et al.*, 2014; MASSMAN; JUNG; BERNARDO, 2013; MORRELL; BUCKLER; ROSS-IBARRA, 2012; POLAND; ENDELMAN; DAWSON; RUTKOSKI *et al.*, 2012; YABE; YAMASAKI; EBANA; HAYASHI *et al.*, 2016), perenes florestais (BEAULIEU; DOERKSEN; CLÉMENT; MACKAY *et al.*, 2014; BEAULIEU; DOERKSEN; MACKAY; RAINVILLE *et al.*, 2014; GRATTAPAGLIA; RESENDE, 2011; IWATA; HAYASHI; TSUMURA, 2011; RESENDE; RESENDE; SANSALONI; PETROLI *et al.*, 2012; RESENDE; MUÑOZ; ACOSTA; PETER *et al.*, 2012; ZAPATA-VALENZUELA; LSIK; MALTECCA; WEGRZYN *et al.*, 2012; ZAPATA-VALENZUELA; WHETTEN; NEALE; MCKEAND *et al.*, 2013) e frutíferas (IWATA; MINAMIKAWA; KAJIYA-KANEGAE; ISHIMORI *et al.*, 2016; KUMAR; BINK; VOLZ; BUS *et al.*, 2012; KUMAR; CHAGNÉ; BINK; VOLZ *et al.*, 2012; KUMAR; MOLLOY; MUÑOZ; DAETWYLER *et al.*, 2015; MURANTY; TROGGIO; SADOK; RIFAÏ *et al.*, 2015). A GS pode ser definida como sendo a seleção simultânea para centenas ou milhares de marcadores, a depender do organismo e extensão do desequilíbrio de ligação, cobrindo todo o genoma. Desta forma, todos os alelos de interesse estarão em desequilíbrio de ligação com pelo menos um ou mais marcadores genotipados e, portanto, devidamente capturados nos modelos preditivos (GRATTAPAGLIA, 2014).

A GS, assim como a GWAS, utiliza uma genotipagem com grande número de marcadores cobrindo todo o genoma, mas difere por não se basear na aplicação de testes de significância. Logo, a GS estima simultaneamente o efeito de todos os marcadores sobre o fenótipo em uma população representativa dos indivíduos, na qual se pretende aplicar o processo de seleção. Logo, ao contrário da GWAS que foca na detecção de associações individuais, a GS utiliza todos ou uma grande proporção dos marcadores para predizer o fenótipo através de modelos preditivos (GRATTAPAGLIA, 2014). Por conseguinte, GS funciona segundo o princípio de que o DL fornecido por uma densa genotipagem é suficiente para capturar grande parte dos QTLs relevantes para a característica alvo. Ao evitar a seleção de marcadores e por estimar os efeitos de marcadores em uma população de treinamento ampla e representativa, a GS tende a capturar uma maior variância genética para a característica avaliada. Logo, a GS mitiga o dilema de como capturar a herdabilidade faltante de características complexas, explicada por um grande número de QTLs de pequenos efeitos (MAKOWSKY; PAJEWSKI; KLIMENTIDIS; VAZQUEZ *et al.*, 2011; MANOLIO; COLLINS; COX; GOLDSTEIN *et al.*, 2009). Pequenos efeitos estes que a seleção assistida com base nos poucos QTLs detectados via mapeamento de QTL e GWAS não conseguem capturar (GRATTAPAGLIA, 2014; GRATTAPAGLIA; PLOMION; KIRST; SEDEROFF, 2009).

A GS abre uma perspectiva concreta de acelerar significativamente o progresso do melhoramento de espécies florestais, devido ao longo ciclo de vida e características que apresentam controle genético complexo e expressão tardia (GRATTAPAGLIA, 2014). As metodologias preditivas da GS, dispensando a necessidade de mapear e localizar QTLs/genes, mas focando exclusivamente no aumento da eficiência com redução do ciclo de melhoramento e no aumento do ganho genético, podem ter uma maior probabilidade de sucesso (GRATTAPAGLIA; PLOMION; KIRST; SEDEROFF, 2009; GRATTAPAGLIA; RESENDE, 2011). Somente nos últimos 5 anos é que novas tecnologias de genotipagem em larga escala tem permitido alcançar densidades e coberturas de marcadores a custos muito accessíveis, o que rapidamente renovou o interesse por metodologias “caixa preta” (*black box*) de predição de fenótipo com base em genótipo (GODDARD, 2009; HABIER; FERNANDO; GARRICK, 2013). Resultados recentes na literatura, principalmente de genética e melhoramento animal são extremamente animadores (CASELLAS; PIEDRAFITA, 2015; FORNERIS; STEIBEL; LEGARRA; VITEZICA *et al.*, 2016; GODDARD, 2009; GODDARD; HAYES; MEUWISSEN, 2010; HAYES; GODDARD, 2010; HAYES; BOWMAN; CHAMBERLAIN; VERBYLA *et al.*, 2009; HAYES; BOWMAN; CHAMBERLAIN; GODDARD, 2009; HAYES; DAETWYLER; GODDARD, 2016; HAYES; LEWIN; GODDARD, 2013; MEUWISSEN; HAYES; GODDARD, 2013; MEUWISSEN; HAYES; GODDARD, 2016; PORTO-NETO; BARENDSE; HENSHALL; MCWILLIAM *et al.*, 2015; THOMASEN; EGGER-DANNER; WILLAM; GULDBRANDTSEN *et al.*, 2014; VAN BINSBERGEN; CALUS; BINK; VAN EEUWIJK *et al.*, 2015), pois indicam que esta abordagem é particularmente interessante para características de baixa herdabilidade e para organismos de ciclo de vida longo (HAYES; DAETWYLER; GODDARD, 2016; HAYES; DONOGHUE; REICH; MASON *et al.*, 2016; LEGARRA; ROBERT-GRANIÉ; MANFREDI; ELSEN, 2008; SCHAEFFER, 2006; THOMASEN; EGGER-DANNER; WILLAM; GULDBRANDTSEN *et al.*, 2014).

A GS atualmente já é uma realidade para o melhoramento animal, com diversos trabalhos que mostram os ganhos genéticos alcançados pela seleção precoce e vantagens em relação ao melhoramento convencional (CASELLAS; PIEDRAFITA, 2015; HAYES; GODDARD, 2010; HAYES; BOWMAN; CHAMBERLAIN; GODDARD, 2009; HAYES; DONOGHUE; REICH; MASON *et al.*, 2016; MEUWISSEN; HAYES; GODDARD, 2016; PORTO-NETO; BARENDSE; HENSHALL; MCWILLIAM *et al.*, 2015; THOMASEN; EGGER-DANNER; WILLAM; GULDBRANDTSEN *et al.*, 2014; VAN BINSBERGEN; CALUS; BINK; VAN EEUWIJK *et al.*, 2015). Experimentos comprovaram a excelente perspectiva de aplicação da GS no melhoramento de plantas anuais, como: milho (CROSSA; PEREZ; HICKEY; BURGUEÑO *et al.*, 2014; MASSMAN; GORDILLO; LORENZANA; BERNARDO, 2013; PACE; YU; LÜBBERSTEDT, 2015), trigo (CROSSA; PEREZ; HICKEY; BURGUEÑO *et al.*, 2014; THAVAMANIKUMAR; DOLFERUS; THUMMA, 2015), cevada (SCHMIDT; KOLLERS; MAASBERG-PRELLE; GROßER *et al.*, 2016), arroz (SPINDEL; BEGUM; AKDEMIR; VIRK *et al.*, 2015; SPINDEL; BEGUM; AKDEMIR; COLLARD *et al.*, 2016) e soja (CHANG; BROWN; LIPKA; DOMIER *et al.*, 2016).

Na área florestal, a GS começou a ser abordada através de alguns estudos com simulação (GRATTAPAGLIA; RESENDE, 2011; IWATA; HAYASHI; TSUMURA, 2011) e logo depois em dois trabalhos pioneiros com dados empíricos de *Pinus (RESENDE; MUÑOZ; RESENDE; GARRICK et al., 2012)* e de *Eucalyptus (RESENDE; RESENDE; SANSALONI; PETROLI et al., 2012)*. Subsequentemente, um estudo de simulação para testar a eficiência da GS, incluindo efeito de dominância no modelo, foi publicado em *Eucalyptus* (DENIS; BOUVET, 2013). Em seguida, foram publicados diversos outros trabalhos em espécies florestais de diversos gêneros de coníferas, como: *Pinus* (ISIK; BARTHOLOMÉ; FARJAT; CHANCEREL *et al.*, 2016; RESENDE; MUÑOZ; RESENDE; GARRICK *et al.*, 2012; ZAPATA-VALENZUELA; LSIK; MALTECCA; WEGRZYN *et al.*, 2012; ZAPATA-VALENZUELA; WHETTEN; NEALE; MCKEAND *et al.*, 2013), *Larix* (KLÁPŠTĚ; LSTIBŮREK; EL-KASSABY, 2014) e *Picea* (BEAULIEU; DOERKSEN; CLEMENT; MACKAY *et al.*, 2014; BEAULIEU; DOERKSEN; MACKAY; RAINVILLE *et al.*, 2014; GAMAL EL-DIEN; RATCLIFFE; KLÁPŠTĚ; CHEN *et al.*, 2015; GAMAL EL-DIEN; RATCLIFFE; KLAPSTE; PORTH *et al.*, 2016).

A GS poderá representar uma mudança radical de paradigma no melhoramento florestal, por permitir a seleção ultra-precoce de árvores elite ainda no estágio de mudas no viveiro para características de expressão tardia, tais como: crescimento volumétrico, qualidade da madeira e tolerância a estresses abióticos e bióticos (GRATTAPAGLIA, 2014). Indivíduos podem ser selecionados visando a instalação de testes clonais ou a sua utilização como genitores para a próxima geração de melhoramento ou ambos, como vem sendo feito em algumas empresas atualmente. Esta estratégia busca explorar a combinação de características favoráveis e identificar indivíduos excepcionais que consolidem diversas características desejáveis. Programas de melhoramento com esta configuração são plenamente adequados para a implementação da abordagem de GS.

OBJETIVOS

Objetivo Geral

O objetivo deste trabalho é apresentar e aplicar um método eficiente para estimar curvas de crescimento em *A. angustifolia* em idades não mensuradas e desenvolver e utilizar um *array* de genotipagem de SNPs para *Araucaria angustifolia*. Os dados genotípicos e fenotípicos obtidos serão utilizados para investigar a diversidade e estrutura genética de populações naturais, estimar parâmetros genéticos em um teste de progênie e explorar modelos de predição genômica para características quantitativas de crescimento.

Objetivos Específicos

1. Apresentar e aplicar um método para estimar curvas de crescimento em A. angustifolia em idades não medidas e examinar a distribuição da variação genética entre e dentro das regiões, procedências e famílias dentro das procedências.

2. Realizar a descoberta de SNPs a partir de dados de sequenciamento genômico e de RNA de *Araucaria angustifolia*.

3. Selecionar SNPs com propriedades adequadas para a genotipagem em array de genotipagem baseado na tecnologia Axiom (Affymetrix) e avaliar a qualidade de genotipagem dos SNPs desenvolvidos vai parâmetros de clusterização de classes genotípicas e repetibilidade de genótipos.

4. Estudar a diversidade e estrutura genética de populações naturais representadas em um teste de procedências e progênies coletadas em 15 localidades que contemplaram a região de ocorrência de *Araucaria angustifolia* no Brasil.

5. Comparar os cenários inferidos sobre a organização, distribuição e estrutura da diversidade genética dentro e entre populações de *Araucaria angustifolia* com base em marcadores bialélicos SNPs e marcadores multialélicos hipervariáveis microssatélites.

6. Estimar parâmetros genéticos de herdabilidade e correlações genéticas para características de crescimento volumétrico em diferentes idades de crescimento utilizando matriz de realizado com base em marcadores SNP (matriz G).

7. Desenvolver e avaliar a capacidade preditiva de modelos de predição genômica para seleção de indivíduos e famílias para características de crescimento em diferentes idades e estimar o ganho genético potencial da seleção genômica em relação ao melhoramento convencional via seleção recorrente.

Referências Bibliográficas

ADAM-BLONDON, A.-F.; ROUX, C.; CLAUX, D.; BUTTERLIN, G. *et al.* Mapping 245 SSR markers on the Vitis vinifera genome: a tool for grape genetics. **Theoretical and Applied Genetics**, 109, n. 5, p. 1017-1027, 2004.

ADAMS, J. P.; LAND JR, S. B.; BELLI, K. L.; MATNEY, T. G. Comparison of 17-year realized plot volume gains with selection for early traits for loblolly pine (Pinus taeda L.). **Forest ecology and management**, 255, n. 5-6, p. 1781-1788, 2008.

AGARWAL, M.; SHRIVASTAVA, N.; PADH, H. Advances in molecular marker techniques and their applications in plant sciences. **Plant Cell Reports**, 27, n. 4, p. 617-631, 2008. journal article.

ALI, M.; RAJEWSKI, J.; BAENZIGER, P.; GILL, K. *et al.* Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. **Molecular Breeding**, 21, n. 4, p. 497-509, 2008.

ALLARD, R. W. **Principles of plant breeding**. John Wiley & Sons, 1999. 0471023094.

ANDRÉS, F. G.; ORTIZ, J.-M. Isoenzymatic Characterization of Potentially Useful Forage Shrubs Belonging to the Genus Cytisus and Allies. **Biochemical Systematics and Ecology**, 23, n. 7/8, p. 813-824, 1995.

ANNICCHIARICO, P.; NAZZICARI, N.; LI, X.; WEI, Y. *et al.* Accuracy of genomic selection for alfalfa biomass yield in different reference populations. **BMC Genomics**, 16, p. 1020, Dec 01 2015.

ARANZANA, M.; CARBÓ, J.; ARÚS, P. Microsatellite variability in peach [Prunus persica (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. **Theoretical and Applied Genetics**, 106, n. 8, p. 1341-1352, 2003.

ASKEW, G. R.; EL-KASSABY, Y. A. Estimation of relationship coefficients among progeny derived from wind-pollinated orchard seeds. **Theoretical and Applied Genetics**, 88, n. 2, p. 267-272, May 01 1994. journal article.

AULER, N. M. F.; REIS, M. S. d.; GUERRA, M. P.; NODARI, R. O. The genetics and conservation of Araucaria angustifolia: I. Genetic structure and diversity of natural populations by means of non-adaptive variation in the state of Santa Catarina, Brazil. **Genetics and Molecular Biology**, 25, p. 329-338, 2002.

BALDANZI, G.; RITTERSHOFER, F.; REISSMAN, C., 1973, **Ensaio comparativo de procedências de Araucaria angustifolia (Bert.) O. Ktze**. 123-124.

BARBAZUK, W. B.; EMRICH, S. J.; CHEN, H. D.; LI, L. *et al.* SNP discovery via 454 transcriptome sequencing. **Plant J**, 51, n. 5, p. 910-918, Sep 2007.

BEAULIEU, J.; DOERKSEN, T.; CLEMENT, S.; MACKAY, J. *et al.* Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. **Heredity**, 113, n. 4, p. 343-352, 10//print 2014. Original Article.

BEAULIEU, J.; DOERKSEN, T.; CLÉMENT, S.; MACKAY, J. *et al.* Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. **Heredity**, 113, 2014.

BEAULIEU, J.; DOERKSEN, T. K.; MACKAY, J.; RAINVILLE, A. *et al.* Genomic selection accuracies within and between environments and small breeding groups in white spruce. **BMC Genomics**, 15, n. 1, p. 1048, 2014. journal article.

BELAJ, A.; MUÑOZ-DIEZ, C.; BALDONI, L.; PORCEDDU, A. *et al.* Genetic diversity and population structure of wild olives from the north-western Mediterranean assessed by SSR markers. **Annals of Botany**, 100, n. 3, p. 449-458, 2007.

BITTENCOURT, J. V. M.; SEBBENN, A. M. Patterns of pollen and seed dispersal in a small, fragmented population of the wind-pollinated tree Araucaria angustifolia in southern Brazil. **Heredity**, 99, n. 6, p. 580-591, 10/10/online 2007.

BITTENCOURT, J. V. M.; SEBBENN, A. M. Pollen movement within a continuous forest of wind-pollinated Araucaria angustifolia, inferred from paternity and TwoGener analysis. v. 9, 2008.

BITTENCOURT, J. V. M.; SEBBENN, A. M. Genetic effects of forest fragmentation in high-density Araucaria angustifolia populations in Southern Brazil. **Tree Genetics & Genomes**, 5, n. 4, p. 573-582, October 01 2009. journal article.

BOTSTEIN, D.; WHITE, R. L.; SKOLNICK, M.; DAVIS, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **Am J Hum Genet**, 32, n. 3, p. 314-331, May 1980.

BOUFFIER, L.; CHARLOT, C.; RAFFIN, A.; ROZENBERG, P. *et al.* Can wood density be efficiently selected at early stage in maritime pine (Pinus pinaster Ait.)? **Annals of Forest Science**, 65, n. 1, p. 106-106, 2008/01/01 2008.

BROOKES, A. J. The essence of SNPs. **Gene**, 234, n. 2, p. 177-186, 7/8/ 1999.

BROWN-GUEDIRA, G.; THOMPSON, J.; NELSON, R.; WARBURTON, M. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. **Crop Science**, 40, n. 3, p. 815-823, 2000.

BROWN, G. R.; GILL, G. P.; KUNTZ, R. J.; LANGLEY, C. H. *et al.* Nucleotide diversity and linkage disequilibrium in loblolly pine. **Proceedings of the National Academy of Sciences of the United States of America**, 101, n. 42, p. 15255-15260, October 19, 2004 2004.

BUERSTMAYR, H.; BAN, T.; ANDERSON, J. A. QTL mapping and marker ‐ assisted selection for Fusarium head blight resistance in wheat: a review. **Plant breeding**, 128, n. 1, p. 1-26, 2009.

BUTCHER, T. B. Achievements in forest tree genetic improvement in Australia and New Zealand 7: Maritime pine and Brutian pine tree improvement programs in Western Australia. **Australian Forestry**, 70, n. 3, p. 141-151, 2007/01/01 2007.

BUTCHER, T. B.; HOPKINS, E. R. Realised gains from breeding Pinus pinaster. **Forest Ecology and Management**, 58, n. 3, p. 211-231, 1993/05/01/ 1993.

CARVALHO, P. E. R. **Espécies florestais brasileiras: recomendações silviculturais, potencialidades e uso da madeira**. EMBRAPA-CNPF/SPI Brasil, 1994. 8585007338.

CASELLAS, J.; PIEDRAFITA, J. Accuracy and expected genetic gain under genetic or genomic evaluation in sheep flocks with different amounts of pedigree, genomic and phenotypic data. **Livestock Science**, 182, p. 58-63, 12// 2015.

CHAGNÉ, D.; CROWHURST, R. N.; TROGGIO, M.; DAVEY, M. W. *et al.* Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple. **PLOS ONE**, 7, n. 2, p. e31745, 2012.

CHANG, H.-X.; BROWN, P. J.; LIPKA, A. E.; DOMIER, L. L. *et al.* Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. **BMC Genomics**, 17, n. 1, p. 153, 2016. journal article.

CHAUHAN, S. S.; SHARMA, M.; THOMAS, J.; APIOLAZA, L. A. *et al.* Methods for the very early selection of Pinus radiata D. Don. for solid wood products. **Annals of Forest Science**, 70, n. 4, p. 439-449, 2013/06/01 2013.

CHEN, H.; XIE, W.; HE, H.; YU, H. *et al.* A high-density SNP genotyping array for rice biology and molecular breeding. **Mol Plant**, 7, n. 3, p. 541-553, Mar 2014.

CHING, A.; CALDWELL, K. S.; JUNG, M.; DOLAN, M. *et al.* SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, 3, n. 1, p. 19, 2002. journal article.

CIPRIANI, G.; SPADOTTO, A.; JURMAN, I.; DI GASPERO, G. *et al.* The SSR-based molecular profile of 1005 grapevine (Vitis vinifera L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. **Theoretical and Applied Genetics**, 121, n. 8, p. 1569-1585, 2010.

COCKERHAM, C. C. Variance of gene frequencies. **Evolution**, p. 72-84, 1969.

COOKE, R. J. The characterisation and identification of crop cultivars by electrophoresis. **ELECTROPHORESIS**, 5, n. 2, p. 59-72, 1984.

COPES, D. Breeding graft-compatible Douglas-fir rootstocks (Pseudotsuga menziesii (Mirb.) Franco). **Silvae genetica**, 48, p. 188-192, 1999.

CORDENUNSI, B. R.; DE MENEZES WENZEL, E.; GENOVESE, M. I.; COLLI, C. *et al.* Chemical composition and glycemic index of Brazilian pine (Araucaria angustifolia) seeds. **J Agric Food Chem**, 52, n. 11, p. 3412-3416, Jun 02 2004.

CREGAN, P.; JARVIK, T.; BUSH, A.; SHOEMAKER, R. *et al.* An integrated genetic linkage map of the soybean genome. **Crop Science**, 39, n. 5, p. 1464-1490, 1999.

CROSSA, J.; PEREZ, P.; HICKEY, J.; BURGUEÑO, J. *et al.* Genomic prediction in CIMMYT maize and wheat breeding programs. **Heredity**, 112, 2014.

DANNER, M. A.; RIBEIRO, J. Z.; ZANETTE, F.; BITTENCOURT, J. V. M. *et al.* Impact of monoecy in the genetic structure of a predominately dioecious conifer species, Araucaria angustifolia (Bert.) O. Kuntze. **Plant Systematics and Evolution**, 299, n. 5, p. 949-958, 2013a. journal article.

DANNER, M. A.; RIBEIRO, J. Z.; ZANETTE, F.; BITTENCOURT, J. V. M. *et al.* Mendelian segregation in eight microsatellite loci from hand- and open-pollinated progenies of Araucaria angustifolia (Bert.) O. Kuntze (Araucariaceae). **Silvae Genetica**, 62, n. 1-2, p. 18-25, 2013b.

DAVEY, J. W.; HOHENLOHE, P. A.; ETTER, P. D.; BOONE, J. Q. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nat Rev Genet**, 12, n. 7, p. 499-510, Jun 17 2011.

DAYANANDAN, S.; DOLE, J.; BAWA, K.; KESSELI, R. Population structure delineated with microsatellite markers in fragmented populations of a tropical tree, Carapa guianensis (Meliaceae). **Molecular Ecology**, 8, n. 10, p. 1585-1592, 1999.

DENIS, M.; BOUVET, J. M. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genet Genomes**, 9, 2013.

DESCHAMPS, S.; LA ROTA, M.; RATASHAK, J. P.; BIDDLE, P. *et al.* Rapid Genome-wide Single Nucleotide Polymorphism Discovery in Soybean and Rice via Deep Resequencing of Reduced Representation Libraries with the Illumina Genome Analyzer. **The Plant Genome**, 3, n. 1, p. 53-68, 2010.

DIRLEWANGER, E.; GRAZIANO, E.; JOOBEUR, T.; GARRIGA-CALDERÉ, F. *et al.* Comparative mapping and marker-assisted selection in Rosaceae fruit crops. **Proceedings of the National Academy of Sciences of the United States of America**, 101, n. 26, p. 9891-9896, 2004.

DREISIGACKER, S.; ZHANG, P.; WARBURTON, M.; VAN GINKEL, M. *et al.* SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments. **Crop science**, 44, n. 2, p. 381-388, 2004.

DRENKARD, E.; RICHTER, B. G.; ROZEN, S.; STUTIUS, L. M. *et al.* A Simple Procedure for the Analysis of Single Nucleotide Polymorphisms Facilitates Map-Based Cloning in Arabidopsis. **Plant Physiol**, 124, n. 4, p. 1483-1492, Dec 2000.

DU, Q.; WANG, B.; WEI, Z.; ZHANG, D. *et al.* Genetic diversity and population structure of Chinese white poplar (Populus tomentosa) revealed by SSR markers. **Journal of Heredity**, p. ess061, 2012.

DUARTE, L. D. S.; DOS-SANTOS, M. M. G.; HARTZ, S. M.; PILLAR, V. D. Role of nurse plants in Araucaria Forest expansion over grassland in south Brazil. **Austral Ecology**, 31, n. 4, p. 520-528, 2006.

ECKERT, A. J.; VAN HEERWAARDEN, J.; WEGRZYN, J. L.; NELSON, C. D. *et al.* Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). **Genetics**, 185, n. 3, p. 969-982, 2010.

EL-KASSABY, Y. A.; CAPPA, E. P.; LIEWLAKSANEEYANAWIN, C.; KLÁPŠTĚ, J. *et al.* Breeding without Breeding: Is a Complete Pedigree Necessary for Efficient Breeding? **PLOS ONE**, 6, n. 10, p. e25737, 2011.

EL-KASSABY, Y. A.; KLÁPŠTĚ, J.; GUY, R. D. Breeding without breeding: selection using the genomic best linear unbiased predictor method (GBLUP). **New Forests**, 43, n. 5, p. 631-637, 2012. journal article.

ERIKSSON, G. **Picea abies: Recent Genetic Research**. Department of Plant Biology and Forest Genetics, SLU, 2010. 9157690162.

EVANS, K.; PATOCCHI, A.; REZZONICO, F.; MATHIS, F. *et al.* Genotyping of pedigreed apple breeding material with a genome-covering set of SSRs: trueness-to-type of cultivars and their parentages. **Molecular breeding**, 28, n. 4, p. 535-547, 2011.

FORNERIS, N. S.; STEIBEL, J. P.; LEGARRA, A.; VITEZICA, Z. G. *et al.* A comparison of methods to estimate genomic relationships using pedigree and markers in livestock populations. **Journal of Animal Breeding and Genetics**, 133, n. 6, p. 452-462, 2016.

GAMAL EL-DIEN, O.; RATCLIFFE, B.; KLÁPŠTĚ, J.; CHEN, C. *et al.* Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. **BMC Genomics**, 16, n. 1, p. 370, 2015. journal article.

GAMAL EL-DIEN, O.; RATCLIFFE, B.; KLAPSTE, J.; PORTH, I. *et al.* Implementation of the Realized Genomic Relationship Matrix to Open-Pollinated White Spruce Family Testing for Disentangling Additive from Nonadditive Genetic Effects. **G3**, 6, n. 3, p. 743-753, 2016.

GANAL, M. W.; DURSTEWITZ, G.; POLLEY, A.; BÉRARD, A. *et al.* A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. **PLOS ONE**, 6, n. 12, p. e28334, 2011.

GANAL, M. W.; POLLEY, A.; GRANER, E.-M.; PLIESKE, J. *et al.* Large SNP arrays for genotyping in crop plants. **Journal of Biosciences**, 37, n. 5, p. 821-828, November 01 2012. journal article.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. **Biometrics**, 51, n. 4, p. 1440-1450, 1995.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, 136, n. 2, p. 245-257, Jun 2009.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. Genomic selection in livestock populations. **Genet Res (Camb)**, 92, n. 5-6, p. 413-421, Dec 2010.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. Using the genomic relationship matrix to predict the accuracy of genomic selection. **J Anim Breed Genet**, 128, n. 6, p. 409-421, Dec 2011.

GORE, M. A.; CHIA, J. M.; ELSHIRE, R. J.; SUN, Q. *et al.* A first-generation haplotype map of maize. **Science**, 326, n. 5956, p. 1115-1117, Nov 20 2009.

GORE, M. A.; WRIGHT, M. H.; ERSOZ, E. S.; BOUFFARD, P. *et al.* Large-Scale Discovery of Gene-Enriched SNPs. **The Plant Genome**, 2, n. 2, p. 121-133, 2009.

GRATTAPAGLIA, D. Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. *In*: TUBEROSA, R.;GRANER, A.*, et al* (Ed.). **Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources**. Dordrecht: Springer Netherlands, 2014. p. 651-682.

GRATTAPAGLIA, D.; PLOMION, C.; KIRST, M.; SEDEROFF, R. R. Genomics of growth traits in forest trees. **Curr Opin Plant Biol**, 12, n. 2, p. 148-156, Apr 2009.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genet Genomes**, 7, 2011.

GRATTAPAGLIA, D.; SILVA-JUNIOR, O. B.; RESENDE, R. T.; CAPPA, E. P. *et al.* Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. **Frontiers in Plant Science**, 9, n. 1693, 2018-November-22 2018. Mini Review.

GRIMMER, M. K.; KRAFT, T.; FRANCIS, S. A.; ASHER, M. J. C. QTL mapping of BNYVV resistance from the WB258 source in sugar beet. **Plant Breeding**, 127, n. 6, p. 650-652, 2008.

GUERRA, M. P.; SILVEIRA, V.; DOS SANTOS, A. L. W.; ASTARITA, L. V. *et al.* Somatic Embryogenesis in Araucaria angustifolia (Bert) O. Ktze. *In*: JAIN, S. M.;GUPTA, P. K.*, et al* (Ed.). **Somatic Embryogenesis in Woody Plants: Volume 6**. Dordrecht: Springer Netherlands, 2000. p. 457-478.

GUICHOUX, E.; LAGACHE, L.; WAGNER, S.; CHAUMEIL, P. *et al.* Current trends in microsatellite genotyping. **Molecular Ecology Resources**, 11, n. 4, p. 591-611, 2011.

GUO, W.; ZHANG, T.; SHEN, X.; YU, J. Z. *et al.* Development of SCAR marker linked to a major QTL for high fiber strength and its usage in molecular-marker assisted selection in upland cotton. **Crop Science**, 43, n. 6, p. 2252-2256, 2003.

GUPTA, P.; BALYAN, H.; EDWARDS, K.; ISAAC, P. *et al.* Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. **Theoretical and Applied Genetics**, 105, n. 2-3, p. 413-422, 2002.

GURGEL-FILHO, O. A., 1980, **Silvica da Araucaria angustifolia (Bert.) O. Ktze**. 29-68.

GURGEL, J.; GURGEL FILHO, O. d. A. Caracterização de ecótipos, em âmbito nacional para o pinheiro brasileiro Araucaria angustifolia (Bert.) O. Ktze. **Silvicultura em São Paulo**, 8, p. 127-134, 1973.

GWAZE, D. Optimum selection age for height in shortleaf pine. **New Forests**, 37, n. 1, p. 9-16, 2009/01/01 2009.

GWAZE, D.; BRIDGWATER, F.; WILLIAMS, C. Genetic analysis of growth curves for a woody perennial species, Pinus taeda L. **Theoretical and Applied Genetics**, 105, n. 4, p. 526-531, 2002/09/01 2002.

GWAZE, D.; BYRAM, T.; LOWE, W.; BRIDGWATER, F. Genetic parameter estimates for growth and wood density in loblolly pine (Pinus taeda L.). **International Journal of Forest Genetics**, 2001.

HA, B.-K.; HUSSEY, R. S.; BOERMA, H. R. Development of SNP Assays for Marker-Assisted Selection of Two Southern Root-Knot Nematode Resistance QTL in Soybean. **Crop Science**, 47, n. S2, p. S-73-S-82, 2007.

HAAPANEN, M.; HYNYNEN, J.; RUOTSALAINEN, S.; SIIPILEHTO, J. *et al.* Realised and projected gains in growth, quality and simulated yield of genetically improved Scots pine in southern Finland. **European Journal of Forest Research**, 135, n. 6, p. 997-1009, 2016/12/01 2016.

HABIER, D.; FERNANDO, R. L.; GARRICK, D. J. Genomic-BLUP decoded: a look into the black box of genomic prediction. **Genetics**, 194, 2013.

HASAN, O.; REID, J. Reduction of generation time in Eucalyptus globulus. **Plant Growth Regulation**, 17, n. 1, p. 53-60, 1995.

HASENEYER, G.; SCHMUTZER, T.; SEIDEL, M.; ZHOU, R. *et al.* From RNA-seq to large-scale genotyping - genomics resources for rye (Secale cereale L.). **BMC Plant Biology**, 11, n. 1, p. 131, September 28 2011. journal article.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, 53, n. 11, p. 876-883, Nov 2010.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. C.; VERBYLA, K. *et al.* Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genet Sel Evol**, 41, 2009.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, 92, n. 2, p. 433-443, 2// 2009.

HAYES, B. J.; DAETWYLER, H. D.; GODDARD, M. E. Models for Genome × Environment Interaction: Examples in Livestock. **Crop Science**, 56, n. 5, p. 2251-2259, 2016.

HAYES, B. J.; DONOGHUE, K. A.; REICH, C. M.; MASON, B. A. *et al.* Genomic heritabilities and genomic estimated breeding values for methane traits in Angus cattle1. **Journal of Animal Science**, 94, n. 3, p. 902-908, 2016.

HAYES, B. J.; LEWIN, H. A.; GODDARD, M. E. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. **Trends Genet**, 29, n. 4, p. 206-214, 2013.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genet Res (Camb)**, 91, n. 1, p. 47-60, Feb 2009.

HAYWARD, A. C.; TOLLENAERE, R.; DALTON-MORGAN, J.; BATLEY, J. Molecular marker applications in plants. **Methods Mol Biol**, 1245, p. 13-27, 2015.

HEDRICK, P. W. A standardized genetic differentiation measure. **Evolution**, 59, n. 8, p. 1633-1638, Aug 2005.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J.-L. Genomic Selection for Crop Improvement. **Crop Science**, 49, n. 1, p. 1-12, 2009.

HELYAR, S. J.; HEMMER-HANSEN, J.; BEKKEVOLD, D.; TAYLOR, M. I. *et al.* Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. **Molecular Ecology Resources**, 11, p. 123-136, 2011.

HENDERSON, C. R. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. **Biometrics**, 32, n. 1, p. 69-83, 1976.

HEUERTZ, M.; DE PAOLI, E.; KÄLLMAN, T.; LARSSON, H. *et al.* Multilocus Patterns of Nucleotide Diversity, Linkage Disequilibrium and Demographic History of Norway Spruce Picea abies (L.) Karst. **Genetics**, 174, n. 4, p. 2095-2105, 2006.

HILL, W. G.; WEIR, B. S. Variation in actual relationship as a consequence of Mendelian sampling and linkage. **Genetics Research**, 93, n. 1, p. 47-64, 2011.

HIREMATH, P. J.; FARMER, A.; CANNON, S. B.; WOODWARD, J. *et al.* Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. **Plant Biotechnol J**, 9, n. 8, p. 922-931, Oct 2011.

HOLSINGER, K. E. Analysis of Genetic Diversity in Geographically Structured Populations: A Bayesian Perspective. **Hereditas**, 130, n. 3, p. 245-255, 1999.

HONG, Y.; CHEN, X.; LIANG, X.; LIU, H. *et al.* A SSR-based composite genetic linkage map for the cultivated peanut (Arachis hypogaea L.) genome. **BMC Plant Biology**, 10, n. 1, p. 17, 2010.

HUDSON, R. R.; SLATKIN, M.; MADDISON, W. P. Estimation of Levels of Gene Flow from DNA Sequence Data. *In*: **Genetics**, 1992. v. 132, p. 583-589.

IBGE. **Produção da Extração Vegetal e da Silvicultura.**, 2015. Disponível em: http://www.ibge.gov.br/home/estatistica/economia/pevs/2015/default.shtm. Acesso em: 02/19/2017.

IOB, G.; VIEIRA, E. M. Seed predation of Araucaria angustifolia (Araucariaceae) in the Brazilian Araucaria Forest: influence of deposition site and comparative role of small and 'large' mammals. **Plant Ecology**, 198, n. 2, p. 185-196, 2008. journal article.

ISAAC-RENTON, M.; STOEHR, M.; BEALLE STATLAND, C.; WOODS, J. Tree breeding and silviculture: Douglas-fir volume gains with minimal wood quality loss under variable planting densities. **Forest Ecology and Management**, 465, p. 118094, 2020/06/01/ 2020.

ISIK, F.; BARTHOLOMÉ, J.; FARJAT, A.; CHANCEREL, E. *et al.* Genomic selection in maritime pine. **Plant Science**, 242, p. 108-119, 1// 2016.

IWATA, H.; HAYASHI, T.; TSUMURA, Y. Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. **Tree Genet Genomes**, 7, 2011.

IWATA, H.; MINAMIKAWA, M. F.; KAJIYA-KANEGAE, H.; ISHIMORI, M. *et al.* Genomics-assisted breeding in fruit trees. **Breeding Science**, 66, n. 1, p. 100-115, 2016.

JANSSON, G.; HANSEN, J. K.; HAAPANEN, M.; KVAALEN, H. *et al.* The genetic and economic gains from forest tree breeding programmes in Scandinavia and Finland. **Scandinavian Journal of Forest Research**, 32, n. 4, p. 273-286, 2017/05/19 2017.

JEFFREYS, A.; FLAVELL, R. A physical map of the DNA regions flanking the rabbit β-globin gene. **Cell**, 12, n. 2, p. 429-439, 1977.

JENA, K.; JEUNG, J.; LEE, J.; CHOI, H. *et al.* High-resolution mapping of a new brown planthopper (BPH) resistance gene, Bph18 (t), and marker-assisted selection for BPH resistance in rice (Oryza sativa L.). **Theoretical and Applied Genetics**, 112, n. 2, p. 288-297, 2006.

JOST, L. G(ST) and its relatives do not measure differentiation. **Mol Ecol**, 17, n. 18, p. 4015-4026, Sep 2008.

JOST, L. O. U. D vs. GST: Response to Heller and Siegismund (2009) and Ryman and Leimar (2009). **Molecular Ecology**, 18, n. 10, p. 2088-2091, 2009.

KAGEYAMA, P. Y.; JACOB, W. S. Variação genética entre e dentro de populações de Araucária angustifolia (Bert) O. Ktze. 1980.

KASHI, Y.; KING, D.; SOLLER, M. Simple sequence repeats as a source of quantitative genetic variation. **Trends Genet**, 13, n. 2, p. 74-78, Feb 1997.

KLÁPŠTĚ, J.; LSTIBŮREK, M.; EL-KASSABY, Y. A. Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. **Tree Genetics & Genomes**, 10, n. 2, p. 241-249, 2014. journal article.

KLEIN, R. M. O aspecto dinâmico do pinheiro brasileiro. **Sellowia**, 12, n. 12, p. 17-44, 1960.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with GWAS: a review. **Plant Methods**, 9, p. 29, 2013.

KORTE, A.; VILHJALMSSON, B. J.; SEGURA, V.; PLATT, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. **Nat Genet**, 44, n. 9, p. 1066-1071, 09//print 2012. 10.1038/ng.2376.

KRUGLYAK, L. The use of a genetic map of biallelic markers in linkage studies. **Nat Genet**, 17, n. 1, p. 21-24, Sep 1997.

KUMAR, S.; BINK, M. C. A. M.; VOLZ, R. K.; BUS, V. G. M. *et al.* Towards genomic selection in apple (Malus × domestica Borkh.) breeding programmes: Prospects, challenges and strategies. **Tree Genetics & Genomes**, 8, n. 1, p. 1-14, 2012. journal article.

KUMAR, S.; CHAGNÉ, D.; BINK, M. C. A. M.; VOLZ, R. K. *et al.* Genomic selection for fruit quality traits in apple (Malus × domestica Borkh.). **PLoS One**, 7, 2012.

KUMAR, S.; LEE, J. Age-age correlations and early selection for end-of-rotation wood density in radiata pine. **Forest Genetics**, 9, n. 4, p. 323-330, 2002.

KUMAR, S.; MOLLOY, C.; MUÑOZ, P.; DAETWYLER, H. *et al.* Genome-Enabled Estimates of Additive and Non-additive Genetic Variances and Prediction of Apple Phenotypes Across Environments. **G3: Genes|Genomes|Genetics**, 2015.

LAMBETH, C.; LEE, B.-C.; O'MALLEY, D.; WHEELER, N. Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. **Theoretical and Applied Genetics**, 103, n. 6, p. 930-943, November 01 2001. journal article.

LANDE, R.; SHANNON, S. The Role of Genetic Variation in Adaptation and Population Persistence in a Changing Environment. **Evolution**, 50, n. 1, p. 434-437, 1996.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods**, 9, n. 4, p. 357-359, Mar 04 2012.

LEGARRA, A.; ROBERT-GRANIÉ, C.; MANFREDI, E.; ELSEN, J. M. Performance of genomic selection in mice. **Genetics**, 180, 2008.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T. *et al.* The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, 25, n. 16, p. 2078-2079, Aug 15 2009.

LI, H. F.; YANG, W.; XIAOJING, Z.; JIANGWEI, X. *et al.* Pathway-Based Genome-Wide Association Studies for Two Meat Production Traits in Simmental Cattle.

Scientific Reports, Published online: 17 December 2015; | doi:10.1038/srep18389, 2015-11-17 2015.

LI, R.; LI, Y.; FANG, X.; YANG, H. *et al.* SNP detection for massively parallel whole-genome resequencing. **Genome Res**, 19, n. 6, p. 1124-1132, Jun 2009.

LI, X.; DING, X.; CHU, B.; ZHOU, Q. *et al.* Genetic diversity analysis and conservation of the endangered Chinese endemic herb Dendrobium officinale Kimura et Migo (Orchidaceae) based on AFLP. **Genetica**, 133, n. 2, p. 159-166, Jun 2008.

LIU, C. M.; WONG, T.; WU, E.; LUO, R. *et al.* SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. **Bioinformatics**, 28, n. 6, p. 878-879, Mar 15 2012.

LYNCH, M. Estimation of relatedness by DNA fingerprinting. **Molecular Biology and Evolution**, 5, n. 5, p. 584-599, 1988.

LYNCH, M.; RITLAND, K. Estimation of Pairwise Relatedness With Molecular Markers. **Genetics**, 152, n. 4, p. 1753-1766, 1999.

MAKOWSKY, R.; PAJEWSKI, N. M.; KLIMENTIDIS, Y. C.; VAZQUEZ, A. I. *et al.* Beyond missing heritability: prediction of complex traits. **PLoS Genet**, 7, 2011.

MALÉCOT, G. Les mathématiques de l'hérédité. Paris: Masson. 1948.

MANOLIO, T. A.; COLLINS, F. S.; COX, N. J.; GOLDSTEIN, D. B. *et al.* Finding the missing heritability of complex diseases. **Nature**, 461, n. 7265, p. 747-753, 10/08/print 2009. 10.1038/nature08494.

MANTOVANI, A.; MORELLATO, L. P.; DOS REIS, M. S. Internal genetic structure and outcrossing rate in a natural population of Araucaria angustifolia (Bert.) O. Kuntze. **J Hered**, 97, n. 5, p. 466-472, Sep-Oct 2006.

MARTIN, M. A.; MATTIONI, C.; LUSINI, I.; DRAKE, F. *et al.* Microsatellite development for the relictual conifer Araucaria araucana (Araucariaceae) using next-generation sequencing. **Am J Bot**, 99, n. 5, p. e213-215, May 2012.

MASSMAN, J. M.; GORDILLO, A.; LORENZANA, R. E.; BERNARDO, R. Genomewide predictions from maize single-cross data. **Theor Appl Genet**, 126, n. 1, p. 13-22, Jan 2013.

MASSMAN, J. M.; JUNG, H.-J. G.; BERNARDO, R. Genomewide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. **Crop Science**, 53, n. 1, p. 58-66, 2013.

MAZZA, M. Use of RAPD markers in the study of genetic diversity of Araucaria angustifolia (Bert.) populations in Brazil. **International Foundation for Science, Florianópolis**, 1997.

MBA, C.; TOHME, J. Use of AFLP markers in surveys of plant diversity. **Methods Enzymol**, 395, p. 177-201, 2005.

MCCOUCH, S. R.; TEYTELMAN, L.; XU, Y.; LOBOS, K. B. *et al.* Development and mapping of 2240 new SSR markers for rice (Oryza sativa L.). **DNA research**, 9, n. 6, p. 199-207, 2002.

MCKENNA, A.; HANNA, M.; BANKS, E.; SIVACHENKO, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Res**, 20, n. 9, p. 1297-1303, Sep 2010.

MEDINA-MACEDO, L.; SEBBENN, A. M.; LACERDA, A. E. B.; RIBEIRO, J. Z. *et al.* High levels of genetic diversity through pollen flow of the coniferous Araucaria angustifolia: a landscape level study in Southern Brazil. **Tree Genetics & Genomes**, 11, n. 1, p. 814, 2014. journal article.

MEDRI, C.; RUAS, P. M.; HIGA, A. R.; MURAKAMI, M. *et al.* Effects of forest management on the genetic diversity in a population of Araucaria angustifolia (bert.) O. Kuntze. **Silvae Genetica**, 52, n. 5-6, p. 202-205, 2003. Article.

MEIRMANS, P. G.; HEDRICK, P. W. Assessing population structure: FST and related measures. **Molecular Ecology Resources**, 11, n. 1, p. 5-18, 2011.

MENGONI, A.; GORI, A.; BAZZICALUPO, M. Use of RAPD and microsatellite (SSR) variation to assess genetic relationships among populations of tetraploid alfalfa, Medicago sativa. **Plant Breeding**, 119, n. 4, p. 311-317, 2000.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Accelerating improvement of livestock with genomic selection. **Annu Rev Anim Biosci**, 1, p. 221-237, Jan 2013.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, 6, n. 1, p. 6-14, 2016.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157, n. 4, p. 1819-1829, Apr 2001.

MIELCZAREK, M.; SZYDA, J. Review of alignment and SNP calling algorithms for next-generation sequencing data. **Journal of Applied Genetics**, 57, n. 1, p. 71-79, February 01 2016. journal article.

MORRELL, P. L.; BUCKLER, E. S.; ROSS-IBARRA, J. Crop genomics: advances and applications. **Nat Rev Genet**, 13, n. 2, p. 85-96, 02//print 2012. 10.1038/nrg3097.

MORRIS, G. P.; RAMU, P.; DESHPANDE, S. P.; HASH, C. T. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **Proc Natl Acad Sci U S A**, 110, n. 2, p. 453-458, Jan 08 2013.

MOUSSEAU, T. A.; RITLAND, K.; HEATH, D. D. A novel method for estimating heritability using molecular markers. **Heredity**, 80, n. 2, p. 218-224, 02/01/print 1998.

MUÑOZ, P. R.; RESENDE, M. F. R.; GEZAN, S. A.; RESENDE, M. D. V. *et al.* Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. **Genetics**, 198, n. 4, p. 1759-1768, 2014.

MURANTY, H.; TROGGIO, M.; SADOK, I. B.; RIFAÏ, M. A. *et al.* Accuracy and responses of genomic selection on key traits in apple breeding. **Horticulture Research**, 2, p. 15060, 12/23/online 2015. Article.

NAMKOONG, G.; KANG, H. C.; BROUARD, J. S. **Tree Breeding: Principles and Strategies: Principles and Strategies**. Springer Science & Business Media, 2012. 1461238927.

NEI, M. Analysis of Gene Diversity in Subdivided Populations. **Proceedings of the National Academy of Sciences of the United States of America**, 70, n. 12 Pt 1-2, p. 3321-3323, 1973.

NEI, M. Definition and estimation of fixation indices. **Evolution**, 40, n. 3, p. 643-645, 1986.

NIELSEN, R.; PAUL, J. S.; ALBRECHTSEN, A.; SONG, Y. S. Genotype and SNP calling from next-generation sequencing data. **Nat Rev Genet**, 12, n. 6, p. 443-451, Jun 2011.

NOVAES, E.; DROST, D. R.; FARMERIE, W. G.; PAPPAS, G. J., Jr. *et al.* High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. **BMC Genomics**, 9, p. 312, Jun 30 2008.

PACE, J.; YU, X.; LÜBBERSTEDT, T. Genomic prediction of seedling root length in maize (Zea mays L.). **The Plant Journal**, 83, n. 5, p. 903-912, 2015.

PÂQUES, L. E. **Forest tree breeding in Europe**. Springer, 2013.

PARCHMAN, T. L.; GEIST, K. S.; GRAHNEN, J. A.; BENKMAN, C. W. *et al.* Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. **BMC Genomics**, 11, n. 1, p. 180, March 16 2010. journal article.

PATREZE, C. M.; TSAI, S. M. Intrapopulational genetic diversity of Araucaria angustifolia (Bertol.) Kuntze is different when assessed on the basis of chloroplast or nuclear markers. **Plant Systematics and Evolution**, 284, n. 1, p. 111-122, 2010. journal article.

POLAND, J.; ENDELMAN, J.; DAWSON, J.; RUTKOSKI, J. *et al.* Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. **The Plant Genome**, 5, n. 3, p. 103-113, 2012.

PORTO-NETO, L. R.; BARENDSE, W.; HENSHALL, J. M.; MCWILLIAM, S. M. *et al.* Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. **Genetics Selection Evolution**, 47, n. 1, p. 84, 2015. journal article.

PUTMAN, A. I.; CARBONE, I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. **Ecology and Evolution**, 4, n. 22, p. 4399-4428, 2014.

QUELLER, D. C.; GOODNIGHT, K. F. Estimating Relatedness Using Genetic Markers. **Evolution**, 43, n. 2, p. 258-275, 1989.

REHFELDT, G. E.; JAQUISH, B. C.; SÁENZ-ROMERO, C.; JOYCE, D. G. *et al.* Comparative genetic responses to climate in the varieties of Pinus ponderosa and Pseudotsuga menziesii: Reforestation. **Forest Ecology and Management**, 324, p. 147-157, 2014/07/15/ 2014.

REIS, A. M. M.; GRATTAPAGLIA, D. RAPD variation in a germplasm collection of Myracrodruon urundeuva (Anacardiaceae), an endangered tropical tree: recommendations for conservation. **Genetic Resources and Crop Evolution**, 51, n. 5, p. 529-538, 2004.

RESENDE, M. D. V.; RESENDE, M. F. R.; SANSALONI, C. P.; PETROLI, C. D. *et al.* Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194, n. 1, p. 116-128, 2012.

RESENDE, M. F. R.; MUÑOZ, P.; ACOSTA, J. J.; PETER, G. F. *et al.* Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytol**, 193, 2012.

RESENDE, M. F. R.; MUÑOZ, P.; RESENDE, M. D. V.; GARRICK, D. J. *et al.* Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). **Genetics**, 190, 2012.

RIEKSTS-RIEKSTIŅŠ, R.; ZELTIŅŠ, P.; BALIUCKAS, V.; BRŪNA, L. *et al.* Pinus sylvestris Breeding for Resistance against Natural Infection of the Fungus Heterobasidion annosum. **Forests**, 11, n. 1, p. 23, 2020.

RISCH, N.; MERIKANGAS, K. The future of genetic studies of complex human diseases. **Science**, 273, n. 5281, p. 1516-1517, Sep 13 1996.

RITLAND, K. A marker-based method for inferences about quantitative inheritance in natural populations. **Evolution**, 50, p. 1062-1073, // 1996. 10.2307/2410647.

ROBERTSON, A.; HOLLINGSWORTH, P. M.; KETTLE, C. J.; ENNOS, R. A. *et al.* Characterization of nuclear microsatellites in New Caledonian Araucaria species. **Molecular Ecology Notes**, 4, n. 1, p. 62-63, 2004.

RONIKIER, M. The use of AFLP markers in conservation genetics--a case study on Pulsatilla vernalis in the Polish lowlands. **Cell Mol Biol Lett**, 7, n. 2b, p. 677-684, 2002.

ROSVALL, O. Review of the Swedish tree breeding program. **Skogforsk, Uppsala, Sweden**, 2011.

RYYNANEN, H. J.; TONTERI, A.; VASEMAGI, A.; PRIMMER, C. R. A comparison of biallelic markers and microsatellites for the estimation of population and conservation genetic parameters in Atlantic salmon (Salmo salar). **J Hered**, 98, n. 7, p. 692-704, Nov-Dec 2007.

SALGOTRA, R. K.; GUPTA, B. B.; STEWART, C. N. From genomics to functional markers in the era of next-generation sequencing. **Biotechnology Letters**, 36, n. 3, p. 417-426, 2014. journal article.

SALGUEIRO, F.; CARON, H.; DE SOUZA, M.; KREMER, A. *et al.* Characterization of nuclear microsatellite loci in South American Araucariaceae species. **Molecular Ecology Notes**, 5, n. 2, p. 256-258, 2005.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, 74, n. 12, p. 5463-5467, December 1, 1977 1977.

SANT'ANNA, C. S.; SEBBENN, A. M.; KLABUNDE, G. H. F.; BITTENCOURT, R. *et al.* Realized pollen and seed dispersal within a continuous population of the dioecious coniferous Brazilian pine [Araucaria angustifolia (Bertol.) Kuntze]. **Conservation Genetics**, 14, n. 3, p. 601-613, 2013. journal article.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, 123, n. 4, p. 218-223, 2006.

SCHLOTTERER, C.; TOBLER, R.; KOFLER, R.; NOLTE, V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. **Nat Rev Genet**, 15, n. 11, p. 749-763, Nov 2014.

SCHMIDT, A. B.; CIAMPI, A. Y.; GUERRA, M. P.; NODARI, R. O. Isolation and characterization of microsatellite markers for Araucaria angustifolia (Araucariaceae). **Molecular Ecology Notes**, 7, n. 2, p. 340-342, 2007.

SCHMIDT, M.; KOLLERS, S.; MAASBERG-PRELLE, A.; GROßER, J. *et al.* Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. **Theoretical and Applied Genetics**, 129, n. 2, p. 203-213, 2016. journal article.

SCOTT, L. J.; SHEPHERD, M.; HENRY, R. J. Characterization of highly conserved microsatellite loci in Araucaria cunninghamii and related species. **Plant Systematics and Evolution**, 236, n. 3, p. 115-123, 2003. journal article.

SEBBENN, A.; PONTINHA, A.; GIANNOTTI, E.; KAGEYAMA, P. Genetic variation in provenance-progeny test of Araucaria angustifolia (Bert.) O. Ktze. in São Paulo, Brazil. **Silvae genetica**, 52, n. 5-6, p. 181-184, 2003.

SEBBENN, A. M.; PONTINHA, A. d. A. S.; FREITAS, S. A. d.; FREITAS, J. A. d. Variação genética em cinco procedências de Araucaria angustifolia (Bert.) O. Ktze. no sul do Estado de São Paulo. **Revista do Instituto Florestal**, 16, n. 2, p. 91-99, 2004.

SEBBENN, A. M.; PONTINHA, A. d. A. S.; GIANNOTTI, E.; KAGEYAMA, P. Y. Variação genética entre e dentro de procedências e progênies de Araucaria angustifolia no sul do estado de São Paulo. **Revista do Instituto Florestal**, 15, n. 2, p. 109-124, 2003.

SEFC, K.; STEINKELLNER, H.; GLÖSSL, J.; KAMPFER, S. *et al.* Reconstruction of a grapevine pedigree by microsatellite analysis. **Theoretical and Applied Genetics**, 97, n. 1-2, p. 227-231, 1998.

SHEN, X.; GUO, W.; ZHU, X.; YUAN, Y. *et al.* Molecular mapping of QTLs for fiber qualities in three diverse lines in Upland cotton using SSR markers. **Molecular breeding**, 15, n. 2, p. 169-181, 2005.

SHIMIZU, J.; HIGA, A., 1980, **Variaçao genética entre procedências de Araucaria angustifolia (Bert.) O. Ktze na regiao de Itapeva-SP, estimada até 6° ano de idade.** 78-82.

SHIMIZU, J. Y.; JAEGER, P.; SOPCHAKI, S. A. Genetic variability in a remnant population of Araucaria in the Iguaçu National Park, Brazil. **Boletim de Pesquisa Florestal**, n. No. 41, p. 18-36, 2000.

SHIMIZU, J. Y.; OLIVEIRA, Y. M. M. d. Distribuição, variação e usos dos recursos genéticos da araucária no sul do Brasil. v. 4, 1981.

SILVA-JUNIOR, O. B.; GRATTAPAGLIA, D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. **New Phytol**, 208, n. 3, p. 830-845, Nov 2015.

SKRØPPA, T.; STEFFENREM, A. Selection in a provenance trial of Norway spruce (Picea abies L. Karst) produced a land race with desirable properties. **Scandinavian Journal of Forest Research**, 31, n. 5, p. 439-449, 2016/07/03 2016.

SLATKIN, M. Inbreeding coefficients and coalescence times. **Genet Res**, 58, n. 2, p. 167-175, Oct 1991.

SONG, Q.; HYTEN, D. L.; JIA, G.; QUIGLEY, C. V. *et al.* Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLOS ONE**, 8, n. 1, p. e54985, 2013.

SOUSA, V. A. d.; AGUIAR, A. V. d. Programa de melhoramento genético de araucária da Embrapa Florestas: situação atual e perspectivas. 2012-08 2012. Boletim Técnico.

SOUZA, M. I. F. d.; SALGUEIRO, F.; CARNAVALE-BOTTINO, M.; FÉLIX, D. B. *et al.* Patterns of genetic diversity in southern and southeastern Araucaria angustifolia (Bert.) O. Kuntze relict populations. **Genetics and Molecular Biology**, 32, p. 546-556, 2009.

SPINDEL, J.; BEGUM, H.; AKDEMIR, D.; VIRK, P. *et al.* Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. **PLOS Genetics**, 11, n. 2, p. e1004982, 2015.

SPINDEL, J. E.; BEGUM, H.; AKDEMIR, D.; COLLARD, B. *et al.* Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. **Heredity**, 116, n. 4, p. 395-408, 04//print 2016. Original Article.

STEELE, K.; PRICE, A.; SHASHIDHAR, H.; WITCOMBE, J. Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. **Theoretical and Applied Genetics**, 112, n. 2, p. 208-221, 2006.

STEFENON, V. M. **The distribution of the genetic diversity in Araucaria angustifolia (Bert.) O. Kuntze populations and its implication for the conservation of the species' genetic resources.** Orientador: FINKELDEY, P. D. R. 2007. 120 f. (PhD) - Faculty of Forestry and Forest Ecology, University of Göttingen Disponível em: http://resolver.sub.uni-goettingen.de/purl/?webdoc-1588.

STEFENON, V. M.; BEHLING, H.; GAILING, O.; FINKELDEY, R. Evidences of delayed size recovery in Araucaria angustifolia populations after post-glacial colonization of highlands in Southeastern Brazil. **Anais da Academia Brasileira de Ciências**, 80, p. 433-443, 2008.

STEFENON, V. M.; GAILING, O.; FINKELDEY, R. Genetic structure of Araucaria angustifolia (Araucariaceae) populations in Brazil: implications for the in situ conservation of genetic resources. **Plant Biol (Stuttg)**, 9, n. 4, p. 516-525, Jul 2007.

STEFENON, V. M.; GAILING, O.; FINKELDEY, R. Genetic structure of plantations and the conservation of genetic resources of Brazilian pine (Araucaria angustifolia). **Forest Ecology and Management**, 255, n. 7, p. 2718-2725, 4/20/ 2008.

STEFENON, V. M.; GAILING, O.; FINKELDEY, R. The role of gene flow in shaping genetic structures of the subtropical conifer species Araucaria angustifolia. **Plant Biol (Stuttg)**, 10, n. 3, p. 356-364, May 2008.

STENLID, J. Population structure of Heterobasidion annosum as determined by somatic incompatibility, sexual incompatibility, and isoenzyme patterns. **Canadian Journal of Botany**, 63, n. 12, p. 2268-2273, 1985/12/01 1985.

STONECYPHER, R. W.; PIESCH, R. F.; HELLAND, G. G.; CHAPMAN, J. G. *et al.* Results from Genetic Tests of Selected Parents of Douglas-Fir (Pseudotsuga menziesii [Mirb.] Franco) in an Applied Tree Improvement Program. **Forest Science**, 42, n. suppl_1, p. a0001-0035, 1996.

STRAUSS, S. H.; BOUSQUET, J.; HIPKINS, V. D.; HONG, Y.-P. Biochemical and molecular genetic markers in biosystematic studies of forest trees. **New Forests**, 6, n. 1, p. 125-158, 1992. journal article.

TANKSLEY, S. D.; ORTON, T. J. **Isozymes in Plant Genetics and Breeding**. Elsevier Science, 1983. 9780444600387.

TANKSLEY, S. D.; RICK, C. M. Isozymic gene linkage map of the tomato: Applications in genetics and breeding. **Theoretical and Applied Genetics**, 58, n. 2, p. 161-170, 1980. journal article.

TATIKONDA, L.; WANI, S. P.; KANNAN, S.; BEERELLI, N. *et al.* AFLP-based molecular characterization of an elite germplasm collection of Jatropha curcas L., a biofuel plant. **Plant Sci**, 176, n. 4, p. 505-513, Apr 2009.

TAUTZ, D.; RENZ, M. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. **Nucleic Acids Research**, 12, n. 10, p. 4127-4138, 1984.

THAVAMANIKUMAR, S.; DOLFERUS, R.; THUMMA, B. R. Comparison of Genomic Selection Models to Predict Flowering Time and Spike Grain Number in Two Hexaploid Wheat Doubled Haploid Populations. **G3: Genes|Genomes|Genetics**, 5, n. 10, p. 1991-1998, 2015.

THOMAS, S. C.; COLTMAN, D. W.; PEMBERTON, J. M. The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. **Journal of Evolutionary Biology**, 15, n. 1, p. 92-99, 2002.

THOMAS, S. C.; HILL, W. G. Estimating quantitative genetic parameters using sibships reconstructed from marker data. **Genetics**, 155, n. 4, p. 1961-1972, Aug 2000.

THOMAS, S. C.; PEMBERTON, J. M.; HILL, W. G. Estimating variance components in natural populations using inferred relationships. **Heredity**, 84, n. 4, p. 427-436, 2000.

THOMASEN, J. R.; EGGER-DANNER, C.; WILLAM, A.; GULDBRANDTSEN, B. *et al.* Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. **Journal of Dairy Science**, 97, n. 1, p. 458-470, 1// 2014.

THOMPSON, R.; BROTHERSTONE, S.; WHITE, I. M. S. Estimation of quantitative genetic parameters. **Philosophical Transactions of the Royal Society B: Biological Sciences**, 360, n. 1459, p. 1469-1477, 07/07 2005.

VAN BINSBERGEN, R.; CALUS, M. P. L.; BINK, M. C. A. M.; VAN EEUWIJK, F. A. *et al.* Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. **Genetics Selection Evolution**, 47, n. 1, p. 71, 2015. journal article.

VAN EE, B. W.; JELINSKI, N.; BERRY, P. E.; HIPP, A. L. Phylogeny and biogeography of Croton alabamensis (Euphorbiaceae), a rare shrub from Texas and Alabama, using DNA sequence and AFLP data. **Mol Ecol**, 15, n. 10, p. 2735-2751, Sep 2006.

VAN INGHELANDT, D.; MELCHINGER, A. E.; LEBRETON, C.; STICH, B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theoretical and Applied Genetics**, 120, n. 7, p. 1289-1299, 2010.

VAN ORSOUW, N. J.; HOGERS, R. C.; JANSSEN, A.; YALCIN, F. *et al.* Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. **PLoS One**, 2, n. 11, p. e1172, Nov 14 2007.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **J Dairy Sci**, 91, n. 11, p. 4414-4423, Nov 2008.

VIGNAL, A.; MILAN, D.; SANCRISTOBAL, M.; EGGEN, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genet Sel Evol**, 34, n. 3, p. 275-305, May-Jun 2002.

VOS, P.; HOGERS, R.; BLEEKER, M.; REIJANS, M. *et al.* AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Res**, 23, n. 21, p. 4407-4414, Nov 11 1995.

WANG, D.; GRAEF, G.; PROCOPIUK, A.; DIERS, B. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. **Theoretical and Applied Genetics**, 108, n. 3, p. 458-467, 2004.

WEIR, B. S.; HILL, W. G. Estimating F-statistics. **Annu Rev Genet**, 36, p. 721-750, 2002.

WHITE, T. L.; ADAMS, W. T.; NEALE, D. B. **Forest genetics**. Cabi, 2007. 1845932854.

WHITLOCK, M. C. G'st and D do not replace FST. **Molecular Ecology**, 20, n. 6, p. 1083-1091, 2011.

WOJNICKA-PÓŁTORAK, A. Changes of genetic structure of Pinus sylvestris L. populations exposed to industrial pollution. **Acta Societatis Botanicorum Poloniae**, 66, n. 1, p. 73-78, 1997.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, 56, n. 645, p. 330-338, 1922.

WRIGHT, S. The Genetical Structure of Populations. **Annals of Eugenics**, 15, n. 1, p. 323-354, 1949.

WU, H. X.; POWELL, M. B.; YANG, J. L.; IVKOVIĆ, M. *et al.* Efficiency of early selection for rotation-aged wood quality traits in radiata pine. **Annals of Forest Science**, 64, n. 1, p. 1-9, 2007.

XIANG, B.; LI, B.; ISIK, F. Time trend of genetic parameters in growth traits of Pinus taeda L. **Silvae genetica**, 52, n. 3-4, p. 114-120, 2003.

YABE, S.; YAMASAKI, M.; EBANA, K.; HAYASHI, T. *et al.* Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops. **PLOS ONE**, 11, n. 4, p. e0153945, 2016.

YANCHUK, A. General and specific combining ability from disconnected partial diallels of coastal Douglas-fir. **Silvae genetica**, 45, n. 1, p. 37-45, 1996.

YE, T.; JAYAWICKRAMA, K. Early selection for improving volume growth in coastal Douglas-fir breeding programs. **Silvae Genetica**, 61, n. 1-6, p. 186-198, 2012.

ZAPATA-VALENZUELA, J.; LSIK, F.; MALTECCA, C.; WEGRZYN, J. *et al.* SNP markers trace familial linkages in a cloned population of Pinus taeda – prospects for genomic selection. **Tree Genet Genomes**, 8, 2012.

ZAPATA-VALENZUELA, J.; WHETTEN, R. W.; NEALE, D.; MCKEAND, S. *et al.* Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. **G3 (Bethesda)**, 3, n. 5, p. 909-916, May 20 2013.

ZHANG, P.; LI, J.; LI, X.; LIU, X. *et al.* Population structure and genetic diversity in a rice core collection (Oryza sativa L.) investigated with SSR markers. **PloS one**, 6, n. 12, p. e27565, 2011.

ZHANG, T.; YUAN, Y.; YU, J.; GUO, W. *et al.* Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. **Theoretical and Applied Genetics**, 106, n. 2, p. 262-268, 2003.

ZHONG, S.; DEKKERS, J. C. M.; FERNANDO, R. L.; JANNINK, J.-L. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. **Genetics**, 182, n. 1, p. 355-364, 2009.

ZHOU, W. C.; KOLB, F.; BAI, G. H.; DOMIER, L. *et al.* Validation of a major QTL for scab resistance with SSR markers and use of marker‐assisted selection in wheat. **Plant breeding**, 122, n. 1, p. 40-46, 2003.

ZHU, Z.; ZHANG, F.; HU, H.; BAKSHI, A. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. **Nat Genet**, 48, n. 5, p. 481-487, May 2016.

Capítulo 1

Time trends in genetic parameters and growth curves across country-wide provenances of the iconic subtropical conifer tree Araucaria angustifolia

Rafael T. Resende[1], Pedro Italo T. Silva[2,3], Orzenil B. Silva-Junior[4,5], Evandro V. Tambarussi[6], Valderes A. Sousa[7], Ananda V. de Aguiar[7], Dario Grattapaglia[4,5,*]

[1] Federal University of Goiás, School of Agronomy/ Plant Breeding Sector, 74690-900, Goiânia – GO, Brazil;

[2] Corteva Agriscience™ Guarapuava Research Station, Guarapuava – PR, Brazil;

[3] University of Brasília, Cell Biology Department, Campus Universitário, 70910-900, Brasília – DF, Brazil;

[4] Plant Genetics Laboratory, Embrapa Genetic Resources and Biotechnology, 70770-910, Brasília – DF, Brazil;

[5] Graduate Program in Genomic Sciences, Universidade Católica de Brasília, 70790-160, Brasília – DF, Brazil;

[6] Midwestern State University, Department of Forestry Engineering, 84505-677, Irati – PR, Brazil;

[9] Embrapa Florestas, 83411-000, Colombo – PR, Brazil.

Keywords

Individual tree modeling; Random regression; Genetic parameters; Mixed models; Early selection; Conifer breeding

Abstract

Understanding the growth patterns of long-lived conifer tree species is important to devise early selection strategies, predict future biomass productivity and assess adaptive tree fitness for long term conservation efforts. We investigated the genetic variation for growth traits of Araucaria angustifolia, the grandiose renowned "Paraná pine" tree, in a trial involving 122 families across 15 provenances covering the entire natural range of the species in Brazil. Measurements at ages 7, 24, 32, 33 and 35 were used to adjust continuous growth curves based on nonlinear mixed-effect models for all 2,158 trees, providing estimates for unmeasured ages in the 7-to-35-year interval. Estimated values closely matched observed ones and a reduction of the coefficient of residual variation was observed in the estimated data, possibly due to removal of random error in the observed measurements, making the estimated curves more reliable to predict growth patterns. Genetic variation for growth within provenances was greater than between, with a trend of increasing heritabilities over time for most provenances. Substantial genetic variation found both within and between families could drive efficient early selection at both levels. All provenances included individual trees and families with good potential to be selected for shorter rotations. Growth curves show that trees invest first in height and later in diameter growth. Considerable variation was observed across provenances for the optimal age and optimal tree volume at which annual growth increment peaks, a tipping point that could be used as a predictor of the optimal rotation age and expected tree volume. The data clearly indicate potential for early selection for growth at age 7-10 with an 85% prediction accuracy of growth at age 35. Additionally, growth data indicate potential of shortening harvest age from 30-35 to 15-20 years by selecting the best individuals and families. These results underscore the potential of expanding investments in breeding and plantation forestry of A. angustifolia, which in parallel could contribute to enhancing conservation efforts of this iconic subtropical conifer.

1. Introduction

Paraná pine [*Araucaria angustifolia* (Bertol.) Kuntze.] is an iconic long-lived subtropical conifer distributed exclusively in South America, with most of its populations concentrated in southern and southeastern Brazil (Reis *et al.*, 2014), and some populations also found in Argentina and Paraguay. Other species in the genus are found elsewhere, such as *Araucaria araucana* (Molina) K. Koch in Chile, *Araucaria bidwillii* Hooker and *Araucaria cunninghamii* Aiton ex D.Don in eastern Australia and New Guinea, and *Araucaria hunsteinii* Schumann in New Guinea. In Brazil, since the last century the original extension of the *A. angustifolia* forest, estimated at approximately 200,000 km² has declined by more than 97% (Medina-Macedo *et al.*, 2014). The interest in its very high-quality timber and the expansion of agricultural frontiers in high fertility lands in the southern regions are the main causes of this severe reduction, Despite specific laws now in Brazil aiming at the conservation of the species, listed as 'Critically Endangered' according to IUCN Red List of Threatened Species (Thomas, 2013), some farmers still insist on clandestine thinning of trees.

Species and ecosystem conservation in Brazil has advanced by the creation and management of Conservation Units throughout the country (Montagna *et al.*, 2012). Additionally, ex situ conservation strategies have played a key role for the maintenance of biological and genetic resources (Ferreira *et al.*, 2012). In the case of highly valuable forest trees such as *A. angustifolia*, forest plantations could serve as supplementary important storehouses of genetic diversity, by ensuring genetic composition mirroring natural populations (Stefenon *et al.*, 2008). The investment in breeding programs of the species could therefore have a positive effect both on the conservation of genetic diversity in germplasm banks paired with the utilitarian purpose of producing highly valuable wood. Genetic improvement strategies that also include conservation efforts have been a reality for conifers in countries with temperate climate. In addition, the slow tree growth compels the forest growers to adopt sustainable exploitation plans (Farjon and Page, 1999). In Brazil, while exotic *Eucalyptus* and *Pinus* species, have received great attention from the germplasm conservation and breeding efforts given their silvicultural and economic importance (Feffer *et al.*, 2019), advances in conservation allied to breeding for native forest trees is still timid. *A. angustifolia*, has been somewhat of an exception given its valuable wood, with a some efforts throughout the years with the establishment and evaluation of provenance and progeny trials (Kageyama and Jacob, 1979; Shimizu *et al.*, 2000; Sebbenn *et al.*, 2003; da Silva *et al.*, 2018). Considerable variation has been reported for growth rate and stem form between different populations of *A. angustifolia*, and also between individual trees within populations (Sebbenn *et al.*, 2003). Opportunities, therefore, exist to improve the silvicultural value of the species by identifying the best wild seed sources and selecting individuals within them to develop varieties that are considerably better than the wild material. Understanding the geographical distribution of ecologically relevant genetic variation and the environmental factors driving adaptive divergence within species will help ensuring appropriate sourcing of material not only for the structuring of tree breeding programs but also for ecological restoration and conservation prioritization (Lu *et al.*, 2016).

Provenance trials combined with progeny tests provide a rich foundation to inform breeding and serve as valuable repositories to source material for conservation and restoration (O'Brien *et al.*, 2007; White *et al.*, 2007). These kinds of studies have been important drivers of breeding for a large number of economically important forest trees including species of *Pinus* (Dieters *et al.*, 1995; Haapanen, 2001; Hodge and Dvorak, 2001; Kroon *et al.*, 2011), *Cryptomeria* (Hiraoka *et al.*, 2019) and *Eucalyptus* (Stackpole *et al.*, 2010) no name a few. To date while much is known about patterns of geographic variation for temperate and subtropical forest trees, fewer are studies with tropical species. A noteworthy exception is the important effort of gene conservation and breeding through provenance/progeny trials carried out by CAMCORE for tropical pines (Hodge and Dvorak, 2001) and more recently with tropical *Eucalyptus* (Hodge and Dvorak, 2015). Following provenance and progeny trials, tree breeding involves sequential steps of mating, testing and selection to increase the frequency of useful alleles for several traits concurrently in a target population. Quantitative data are used for estimating genetic variances, types of genetic action, heritabilities and genetic correlations for the key traits, and results used to predict and estimate gain with successive selection cycles (Lynch and Walsh, 1998). In practice, however, there is strong economic pressure to reduce the time needed to complete a breeding cycle (White *et al.*, 2007) and to shorten the rotation cycle of a production forest (Haapanen *et al.*, 2016). Especially in slow growing species, early indirect selection is key and has been widely investigated for conifer species of *Pinus* (Lambeth, 1980; Foster, 1986; Carter *et al.*, 1990; Gwaze *et al.*, 2002; Weng *et al.*, 2007; Chauhan *et al.*, 2013), *Picea sp.* (Newton, 2003) and *Eucalyptus* (Leksono *et al.*, 2006).

Besides quantitative genetic parameters, knowledge of the growth behavior of a forest tree species is a key element, be it for early genetic selection, prediction of future biomass production, or to understand patterns of adaptive fitness (Bowman *et al.*, 2013). Hess and Schneider (2009), evaluating three sites in southern Brazil, described *A. angustifolia* height growth as a sigmoid form, with higher rates of increase between 15 and 20 years, and a trend of stagnation after 30 years. These same authors demonstrated that the Araucaria diameter presents a sigmoid growth form in the three environments evaluated, with higher rates of increase between 20 and 33 years, depending on the region (Hess *et al.*, 2009). Similarly, height growth of Scots Pine, between ages four and 18 years, was described as approximately linear by Haapanen (2001), but suggesting a moment of inflection after these earlier ages. In forestry, it is common to use nonlinear functions such as Weibull, Chapman-Richards and Logistics, to describe tree growth, and in particular, random-effect models are interesting for tree-to-tree growth estimates (Subedi and Sharma, 2011), allowing for greater flexibility of the model for growth projections taking into account the particular effects of the site on individual development or even its genetic features.

In this study, we investigated the genetic variation for growth traits in a *A. angustifolia* provenance and progeny trial involving 122 open pollinated families from 15 origins across four Brazilian states during a 35-year growth period. Our objectives were: i) to present an efficient methodology to estimate growth in *A. angustifolia* for unquantified ages; ii) to assess the variation in genetic parameters for growth traits across provenances; iii) to evaluate the efficacy of early genetic selection within and between progenies and provenances and iv) to present an efficient methodology to estimate missing measurement data. The underlying goal of the study was to provide an update on the long-term growth patterns of the most comprehensive genetic trial of *Araucaria angustifolia* currently running in Brazil to potentially foster private initiatives toward more intensive breeding and plantation of this iconic Brazilian conifer.

2. Material and Methods

2.1. Sampling and Experimental Design

The field experiment was originally described in Sebbenn *et al.* (2003). In brief, seeds from open-pollinated families were collected from trees sampled in 15 natural

A. angustifolia provenances in four Brazilian States - Minas Gerais (MG) São Paulo (SP), Paraná (PR) e Santa Catarina (SC) (Figure 1). A total of 122 families were sampled, with the number of families per provenance varying from four to 14 (Table 1). The experiment was set up in the Itapeva Experimental Station of the São Paulo State Forest Institute (24°17' S, 48°54' W and 930 m altitude). The trial was established in a compact-family design, with 15 provenances (plots), with four to 14 progenies per provenance (subplots), 10 individuals per subplot and three replicates in a 3 m x 2 m spacing and borders consisting of two rows. Seeds were collected in May 1979 and seedlings planted in March 1980. The trial was measured after four, seven, 24, 32 and 35 years for total height (HEI) and after seven, 24, 32, 33 and 35 years for diameter at breast height (DBH) (1.3 m). All measurements were used to adjust the individual growth curves for all 2,158 trees but for standardization and better quality of the adjusted values only data from measurements conducted between ages seven and 35 were used in the genetic models analyzed.

2.2. Extrapolation models for HEI and DBH to the entire 7-35yr growth range

Modeling was applied to provide a full range of year-to-year estimates of growth in DBH and HEI for individual trees using adequate growth functions (Burkhart and Tomé, 2012). The main motivation for using non-linear random models was to predict data outside the sampled range, especially for trees that have missing data for early or late ages. Modeling presents also the opportunity to correct errors that occur during data collection, such as non-sampling errors, or trees that oddly seem to have shrunk over time. We used the methodology described by Calegario *et al.* (2005) and Lindstrom and Bates (1990) to adjust the continuous growth of the trees based on nonlinear mixed-effect models.

Random nonlinear models were applied using the function 'nlme' in R (Pinheiro *et al.*, 2016). The variable considered x_{ij} represents the i-th tree on j-th measurement time, being $i = [1, 2, \dots, 2158]$ trees and $j = \{4, 7, 24, 32, 35\}$ for HEI and $j = \{7, 24, 32, 33, 35\}$ for DBH, in years. All ages contain approximately the same number of trees, except for missing data. The non-linear function $x_{ij} = \mathrm{F}(\theta_{ij}, v_{ij}) + \epsilon_{ij}$ could be applied to represent the relationship between the response variable and the covariates within the ith tree, where F is a general function of a group-specific parameter vector

θ_{ij} and a covariate vector v_{ij}, and ϵ_{ij} is a normally distributed within-group error term. The parameter vector θ_{ij} has the form:

$$\theta_{ij} = L_{ij}\beta_i + K_{ij}r_i \qquad [1]$$

$$r_i \sim N(0, \varphi),$$

where β is a vector of fixed effects; r_i is a vector of random effects associated with the i th tree; and L_{ij} and K_{ij} are incidence matrices of fixed and random effects, respectively. In the basic assumptions, the within group errors are independently distributed with mean zero and variance σ^2 and independent of the random effects.

The Chapman-Richards (Eq. [2]), function was chosen to estimate plant height over time following the instructions described earlier (Hess and Schneider, 2009). The function chosen to estimate DBH over time was the Logistic function (Eq. [3]), which has proved to be precise and flexible in this case based on the comparison to other non-linear functions (results not shown).

Chapman-Richards model for estimating HEI:

$$HEI_{ij} = \theta_{1_i}\left[1 - exp^{(-\theta_{2_i} t_{ij})}\right]^{\theta_{3_i}} + \varepsilon_{ij}, \qquad [2]$$

Logistic model for estimating DBH:

$$DAP_{ij} = \frac{\theta_{1_i}}{1 + exp^{(\theta_{2_i} - t_{ij})/\theta_{3_i}}} + \varepsilon_{ij}, \qquad [3]$$

where, HEI_{ij} and DAP_{ij} are respectively plant height and diameter at breast height for the i-th tree on the j-th age; t_{ij} is the age (time) in j years of the tree i; ε_{ij} is the random error.

$$\theta_i = \begin{bmatrix} \theta_{1_i} \\ \theta_{2_i} \\ \theta_{3_i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} r_{1_i} \\ r_{2_i} \\ r_{3_i} \end{bmatrix} = \beta + r_i, \qquad [4]$$

$$r_i \sim N(0, \varphi);\ \varepsilon_{ij} \sim N(0, \sigma^2)$$

Here, β is a vector of fixed effects and r_i represents the vector of random effects. ε_{ij} and r_i are independents.

The individual volume estimate of the tree (VOL) was obtained according to Sanquetta *et al.* (2016), using a Spurr Log model, as follows:

$$lnVOL = -9{,}6687 + 0{,}9650[ln(DBH^2HEI)], \quad [5]$$

2.3. Genetic modeling

The mixed model of Eq. [6] was used for the phenotypic growth values (HEI, DBH and VOL) using the function 'regress' (Clifford *et al.*, 2014) in the R environment. To speed up simultaneous adjustments for each of the ages, its configuration was inspired by model #5 of the free software SELEGEN REML-BLUP (Resende, 2016): complete blocks, several provenances, half-sib progenies tested in one location (Figure 1):

$$y = Xb + Za + Wp + Ts + e, \quad [6]$$

where y is the data vector (HEI, DBH or VOL); b is the blocking effect vector (presumed as fixed effects) summed to overall average; a is the vector of individual additive genetic effects (presumed as random effects); p is the plot effect vector (presumed as random effects); A is the Half-sib relationship matrix between all 2,158 individuals; s is the provenance effect vector (random) and e is the error or residuals vector (random). The capital letters X, Z, W, and T represent the incidence matrices for these effects. The variance structure of the model was as follows:

$a|A, \sigma_a^2 \sim N(0, A\sigma_a^2)$;
$p|\sigma_{pop}^2 \sim N(0, I\sigma_{pop}^2)$;
$s|\sigma_{parc}^2 \sim N(0, I\sigma_{parc}^2)$;
$e|\sigma_e^2 \sim N(0, I\sigma_e^2)$;

Narrow-sense heritability was obtained by: $h_a^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_{pop}^2 + \sigma_{parc}^2 + \sigma_e^2)$; the coefficient of population, i.e. provenance, determination was obtained by: $c_{pop}^2 = \sigma_{pop}^2 / (\sigma_a^2 + \sigma_{pop}^2 + \sigma_{parc}^2 + \sigma_e^2)$. It should be noted that $\sigma_a^2 + \sigma_{pop}^2 + \sigma_{parc}^2 + \sigma_e^2$ correspond to the component of the phenotypic variance of the model. To adjust the parameters for each provenance k, being $k = [1, 2, \ldots, 15]$, the reduced form of the model from Eq. [6] was adopted:

$$y_k = X_k b_k + Z_k a_k + T_k s_k + e_k, \quad [7]$$

where, y, the effects b, a, s, e, the incidence matrices X, Z and T and the variance structures are corresponding to those of Eq. [6]. Heritability in the narrow sense for each k provenance was obtained by: $h^2_{a_k} = \sigma^2_{a\,k}/\left(\sigma^2_{a\,k} + \sigma^2_{parc_k} + \sigma^2_{e\,k}\right)$. It is important to note that the models of equations [6] and [7] were repeated individually for each of the ages (7 to 35 years).

Genetic correlations (r_{gg}) within progenies were estimated based on Pearson correlations between predicted genetic values for each year ($\hat{a}$, from Eq. [6]). This strategy was adopted to generate inferences about early selection, since when measuring data at a given time j, i.e., data at $j+1$ is unkown. It was also for this same reason that the models were adjusted individually for each year, so that the data increment of other years did not provide an unrealistically better fit than expected.

The calculation of the genetic correlations between progenies was done in a manner analogous to the genetic correlation between individuals (within progenies), replacing the additive component of Eq. [6] with the family index vector making $y = Xb + Qf + Wp + Ts + e$, where f is the random effects of the 122 progenies and Q is their respective incidence matrix. The other components are described in Eq. [6]. Genetic correlations between progenies were finally obtained through Pearson's correlation between the random effects of progenies between each of the evaluated years. Models of equations [6] and [7] are reciprocal given that $\sigma^2_a = 4\sigma^2_f$.

3. Results

3.1. Random regression estimates

Annual growth was estimated from age seven to 35 years for 2,158 *A. angustifolia* trees (Figure 2). Parts ‘a’ and ‘b’ show the relationship between the observed and model estimated values. Points scattered below the dashed line (45°) correspond to trees that have “shrunk” over time. In other words, measurement or annotation errors that occurred during data collection, causing the data to differ from the true values, suggesting greater reliability in the estimated values than in the observed ones. Scattered points above this line are likely recording errors super estimating the actual measurements. Graphs in ‘c’ and ‘d’ show that the residues concentrate strongly around zero, indicating some rare observed values with deviation between 10% and 20%. Parts ‘e’ and ‘f’ show the behavior of the adjusted growth data over time for a randomly taken sample of 25 trees. It is noted that for HEI, growth stagnation generally starts around age 30 years, while for DBH the growth is still in full swing at that same age for most trees. The relationship between DBH and HEI for different time scales is shown in Supplementary Figure S1. Even at age 35 *A. angustifolia* does not show biomass growth stagnation, but a deceleration of height increment can already be noticed. The red curves corresponding to the relationship for the sequential ages show that the trees invest slightly more in height growth early on and as time passes this trend shifts to diameter growth. To validate the estimated data, the parameter estimates were compared for both the observed data and the estimated data (Table 2). For comparison, only ages seven, 24, 32 and 35 were used when coincident measurement for HEI and DBH were taken.

3.2. Genetic parameters from observed and modeled data

The modeling approach used to provide estimates of growth data for the entire time span showed good agreement between the observed and estimated data at the four ages for which coincidental measurements were taken for HEI and DBH. A coefficient of determination R^2 above 0.9 was already observed for both traits at age 7, increasing above 0.98 at age 35 with a concurrent reduction of Root Mean Square Error (RMSE) down to the 3-4% range (Table 2). This result is also illustrated by the overall small differences between the estimates of genetic parameters (heritability and variance components) obtained with observed and estimated data for essentially all genetic parameters. The good agreement observed between the observed and estimated data support the growth data modeling employed for the unsampled ages (Figure 2). This agreement might be explained by an improved capture of the genetic variance in the estimated data due to the absence of non-sampling errors. This is also shown by a reduction of the coefficient of residual variation CV_e in practically all the ages for both traits. The coefficients of determination of provenances (c_{pop}^2) were constant across the four measured ages and greater than the heritability in the narrow sense (h_a^2) only at age seven years. By performing a naïve Chi-square test of observed against estimated data among the four variances of the models, *p-values* obtained were 0.63 to 0.99 for HEI and 0.83 to 0.99 for DBH, indicating no significant difference between the parameters for the two traits.

The additive genetic variance showed an exponential increase from age 7 to age 24 for both traits but while for HEI it practically leveled off at age 24, for DBH it had a further increase from age 24 to 35 (Table 2). This increase in genetic variance is mirrored by an equivalent increase in heritabilities from age 7 to 24 and the same pattern of leveling off after that age (Table 2). Provenance specific estimates of individual narrow sense heritabilities (colored lines) were obtained across all ages showing considerable variation across provenances, although generally increasing with age, as also shown by the overall estimate (dashed black line) (Figure 3). Note that the overall heritability within provenances (black dashed line) is greater than the heritability between provenances (red dotted line) as expected (Vencovsky *et al.* (2012), while most individual-provenance heritabilities (colored lines) are substantially higher. Provenances 8, 2, 6, 10 (for HEI); 6, 10, 8, 13, 11, 9, 2 (for DBH); and 6, 10, 8, 2, 11, 13 (for VOL) are the ones with greater genetic variability and probably those that will allow greater gains from directional selection while provenances 1, 3, 12, and 15 conversely are the ones displaying the lowest genetic variance in the trial site.

Estimates of average growth traits for the 15 provenances show the same overall ranking across ages, with Northern region provenances 1 to 5 showing considerably higher growth rates for all traits when compared to the remaining provenances and provenance 4 from the municipality of Lambari, state of Minas Gerais, showing the most outstanding average performance (Figure 4). Southern provenances 12, 13 and 15 had the worst performances overall and provenances 8 and 9, located closest to the experimental site showed an average performance.

Age-age genetic correlations between all ages and age 35 for the three traits are shown in Figure 5a. Both correlations for individual trees within families and among families followed the same trend although slightly higher values were seen for among family's correlations. Correlations were high and increased over time and approached unity at age 35. Correlations ranged from 0.83 to 1.00 for HEI, 0.81 to 1.00 for DBH and from 0.75 to 1.00 for VOL. The strongest age-age correlations at both levels were seen for HEI. Expected genetic gains for growth at age 35 following early selection at different ages were estimated based on breeding values (Figure 5-b). Consistent with the age-age correlations graphs, expected gains from early selection increased when approaching the target age without showing any plateauing, indicating that the maximal gain will likely be made only at age 35. Three selection intensities were simulated on the 2,158 trees evaluated, namely, high: 1% (22 individuals), intermediate: 5% (108 individuals) and low: 10% (216 individuals). Percent gains were 5 to 10X higher when selecting for VOL when compared to HEI and DBH and, as expected, increased with increasing selection intensity (lower percent selected). Despite the large performance differences among the provenances, superior trees for growth were observed in all fifteen provenances, especially when performing selection with intermediate (5%) and low intensity (10%) (Table 3), consistent with the large amount of within-provenance genetic variation observed in the species.

4. Discussion

We have presented a comprehensive picture of time trends in genetic parameters and growth curves of the most genetically inclusive provenance and progeny trial of *A. araucaria* available to date. Using a modeling approach that allowed generating data for unsampled ages, a set of high-quality year-to-year growth and genetic parameter estimates were obtained for diameter at breast height (DBH), total tree height (HEI) and individual volume (VOL) for 2,158 individuals, corresponding to about 59% of the trees that survived the originally planted experiment. This unique long-term provenance, progeny and individual level growth dataset encompassing a countrywide representation of *A. araucaria* natural populations, represents a valuable asset, together with the actual trial, for genetic improvement, germplasm conservation and restoration efforts of this highly valuable and species, still largely unexploited from the point of view of sustainable forest plantation.

4.1. Patterns of heritability change with age

Genetic variation for growth traits in *A. angustifolia*, namely DBH, HEI and VOL, tends to increase over time as shown by an increase in narrow-sense heritability (h_a^2) (Figure 3). Such an increase is expected following the differentiation of biomass accumulation of some individual trees compared to others over time. While in early ages trees tend to have a smaller difference in biomass, over time this difference becomes increasingly pronounced. The individual narrow sense heritability for HEI and DBH in *A. angustifolia* reached values in the range of ~0.2 to 0.25 at age 35. These values are lower than some of the estimates reported in earlier studies where values between 0.03 and 0.6 were observed (Kageyama and Jacob, 1979; da Silva *et al.*, 2018). This difference could be explained by the significantly broader country-wide sampling in this trial in comparison to previous ones, including provenances from a much wider geographical range and correspondingly variable sampling of genetic variation and heritability behavior (Figure 3).

The increase in genetic variation for growth with age resulting in higher narrow sense heritability has been reported in a number of studies of different conifer species, particularly of genus *Pinus* (McKeand, 1988; Hodge and White, 1992; Balocchi *et al.*, 1993; Dieters *et al.*, 1995; Costa and Durel, 1996; Li *et al.*, 1996; Jansson *et al.*, 2003; Weng *et al.*, 2007), although other reports have shown either non-linear (Gwaze *et al.*, 2002) or constant (Haapanen, 2001) heritabilities with age. When individual *Araucaria angustifolia* provenances are considered, narrow-sense heritabilities for growth traits showed quite variable patterns across the 15 provenances of *A. angustifolia* and along the ages evaluated (Figure 3), in line with the fact that the expression of genetic and residual variances may depend on the provenance and also on the age of the individuals (Jansson *et al.*, 2003). A decrease in heritability for DBH was observed in provenances 9, 13 and 14 while in provenances 1, 14, 9 and 10 the decrease was in heritability for HEI (Figure 3).

Our results highlight the fact that using an average estimate of heritability for all provenances irrespectively of age, would lead to significant inaccuracies in the selection of the best individuals within families. For example, for provenance 6, using an average heritability would be problematic as low values of h_a^2 were estimated at advanced ages and high values at early ages, with the highest estimate of h_a^2 observed for VOL at age 36. On the other hand, for provenance 4 a more stable behavior is observed across ages, while provenance 13 has a variable behavior and provenance 1 results are practically null throughout the evaluation period. There are some possible explanations for this observed variability in heritability time-trend across provenances: i) a fluctuating growth interaction between individuals over time; ii) variable interaction of provenances and progenies with the environmental variation in the testing site with time; iii) differential expression of inbreeding depression across provenances and progenies within provenances as the magnitude of inbreeding depression can be expressed at different life stages (Tambarussi *et al.*, 2017); iv) patchy mortality rates over time across provenances, which reduces genetic variance and / or increases residual variance (Kroon *et al.*, 2011).

4.2. Age-age genetic correlation and early selection gains

While heritabilities are measures of the degree to which trait variances are governed by genetic rather than environmental factors, genetic correlations describe the extent to which breeding values (i.e., measures of the additive 'genetic worth' of individuals for a specific trait, age and environment) co-vary. Genetic correlations between all ages and age 35 were high both for individual trees within families and among families. For DBH and HEI, regardless of the growth period, genetic correlations were always greater than 0.85, and for volume, greater than 0.70. Selecting for HEI and DBH individually is slightly better than on VOL, as there is a tendency to achieve higher correlations with age 35 faster (Figure 5-a). When the seventh year of growth is reached, genetic correlations with age 35 years are already above 0.80, except for selection within progenies for variable VOL, which is approximately 0.75. To reach genetic correlations higher than 0.90, selection needs to be carried out beyond age 12 years for all three traits HEI, DBH and VOL. From age 30 onward the parametric estimates demonstrate a genetic correlation approximately equal to 1.0 with age 35 years. Our age-age correlations estimates are equivalent or higher than those reported for other conifers. While this could be an intrinsic biological property of *A. angustofolia*, the modeling approach used to improve data quality might have also contributed to this result. Reported results for other conifers vary. While age-age genetic correlations of early height with 8-year volume increased significantly in the first 3–4 years reaching values above 0.8 for loblolly pine (Xiang *et al.*, 2003), relatively modest genetic correlations of 0.468 for HEI and 0.531 for DBH between ages 5 and 30 years were reported for *Cryptomeria japonica* (Hiraoka *et al.*, 2019) and 0.51 for HEI between ages 7 and 24 years for *Pinus contorta* (Xie and Ying, 1996).

The high genetic correlations observed in *A. angustifolia* represent an exciting result, opening promising opportunities for carrying out efficient early selection. A commonly cited limitation to increase investment in *A. angustifolia* plantation forestry in Brazil is the slow growth and consequently the long investment timeframe necessary to capture returns and the long breeding cycles needed to improve populations. This becomes even more evident when comparisons with fast growing *Pinus taeda* or *Eucalyptus sp.* are made, although such comparisons are not legitimate given the significant differences in the final wood product and market price. *A. angustifolia* wood is fine in texture, uniform in color, with fiber length around 5 mm, much higher than *Eucalyptus* (1 mm), and pines (3 to 4 mm), superior mechanical strength and flexibility when compared to other commercial conifers (Santini *et al.*, 2000), basic density between 400 to 500 kg/m^3 (Trevisan *et al.*, 2016) and cellulose content up to 60%.

Currently there are no ongoing systematic efforts to attempt to reduce the breeding cycles of *A. angustifolia,* mostly due to the still limited interest in extensive plantation forestry of the species. Our study indicates that correlations between ages 7 and 35 would already allow efficient early selection (Figure 5a). Nevertheless recombination of selected *A. angustifolia* trees depends on emission of strobilus which generally takes place only around age 10-15 years in isolated trees and from 20 years of age onward in homogeneous plantations (Carvalho, 1994). Unless some early flowering techniques commonly used in conifers are optimized for *A. angustifolia,* such as induction of grafted scions with gibberellin (Greenwood, 1982) or top grafting on reproductively older trees (Perez *et al.*, 2007), a considerable time lag will be necessary to complete a breeding cycle despite early selection. To date no attempts have been made to induce early flowering in *A. angustifolia*, an area of research that should merit attention as an important tool to accelerate breeding, reminding that the species is dioecious therefore requiring flower induction in individuals of both sexes.

Our data show that the abundant genetic variation found both within and between families could drive efficient early selection at both levels. In theory family selection is expected to provide greater genetic gains at any selection intensity (Kageyama and Jacob, 1979; Diao *et al.*, 2016). However, our analyses indicate that early individual or family selection would provide similar efficiencies (Figure 5a), suggesting that, operationally, in an initial stage of improvement it may be more advantageous to select the best individuals within the best families. Additionally, it might also be recommended to consider selecting the best individuals within families in more geographically distinct provenances. Although the northern provenances displayed a considerably higher average growth performance at least in the trial site (Figure 4), fast growing trees can also be found in southern provenances. Recently, genome-wide single nucleotide polymorphism data have shown a considerably higher genetic divergence between northern and southern populations of *A. angustifolia* challenging previous microsatellite based estimates (Silva *et al.*, 2020). Selecting top trees in genetically divergent provenances might allow exploiting inter-provenance heterotic effects as demonstrated in other conifers such as *Picea* and *Pinus* (Kaya and Lindgren, 1992; Harfouche *et al.*, 2000; de la Mata *et al.*, 2014), a breeding strategy still generally underappreciated in conifer breeding. Finally, although scalable sustainable conifer cloning still represents a technical challenge, unless somatic embryogenesis is developed (Park, 2002), vegetative propagation of elite Araucaria trees could be considered by rooted cutting for the establishment of clonal plantations, clonal seed orchards or for conservation purposes (Wendling *et al.*, 2016).

Our results provide useful information to identify the growth patterns of individual provenances, families or individuals to allow selection at these different levels to shorten the final rotation age. The few existing commercial *A. angustifolia* plantations are grown in pure stands and harvested at a rotation age of approximately 30 years. Wood products include pulp for paper and cardboard, timber for construction, and veneer (Nutto *et al.*, 2005). In this study, the Mean Annual Increment (MAI) was estimated for each provenance, with the maximum volume observed around ages 26 to 28 years for all provenances (Figure 6), which indicates that this could be an ideal rotation age as far as volume growth. However, observing the individual trees growth curves, 239 trees were observed reaching the maximum growth increment before age 20 years, and 17 trees at age 15 years (Suppl. Figure S2). All provenances included individual trees with good potential to be selected for shorter rotations. Considerable variation was observed across provenances for the optimal age (Figure 6a) and optimal tree volume (Figure 6b) at which annual growth increment peaks, a tipping point that could be used as a predictor of the optimal rotation age and expected tree volume. For example, provenance 4 from Lambari, showed a remarkably higher growth rate when compared to all others, with a peak of average growth rate at 0.56 m^3 per tree. On the other hand, other provenances, such as provenance 13 from Caçador, peaks its average growth rate at just 0.2 m^3 per tree. It is important to mention that increasing growth rates may reduce the longevity of conifers. Rapid and large growth rates may mean reduced investment in defenses, lower wood density and mechanical resistance, greater hydraulic resistance, as well as problems with negative growth regulation during periods of stress (Bigler and Veblen, 2009). Data from individual growth curves therefore represent valuable information to be integrated into selection decision to potentially reduce the rotation cycle of *A. angustifolia*, although considerations regarding optimal age for wood properties traits need to be taken into account as well (Nutto *et al.*, 2005).

4.3. Araucaria as a viable tropical commercial conifer option

A. angustifolia is currently considered an endangered species at the international level and protected by a 20-year-old law in Brazil (BRASIL, 2001). Although the management of naturally forested areas of this species is mostly forbidden, recent initiatives have proposed that Mixed Ombrophilous Araucaria forests in Southern Brazil can be managed as sustainable sources of environmental, social and economic benefits (Longhi *et al.*, 2018; Arnoni Costa *et al.*, 2020). Clearly, however, much research is still needed to develop solid scientific data to support truly sustainable management strategies for such complex mixed stands subtropical forests, as models for low species diversity, temperate conifer stands, do not necessarily apply (Hess *et al.*, 2018).

A. angustifolia plantation, however, is fully legal and increasingly seen as a viable alternative that has recently attracted renewed interest, especially by small and medium size farmers that have to restore forested areas in their properties to abide to the new Brazilian forest code (E. Schaitza pers. comm.). The common question posed by both small- and large-scale forest enterprises is the economic viability of *A. angustifolia* versus exotic conifer species of *Pinus*. A recent economic analysis based on formal economic metrics such as NVP (Net Present Value), IRR (Internal Rate of Return) and ROI (Return on Investment), concluded that Araucaria plantations only become competitive with pines on average quality sites that would support an MAI of 23 m^3/ha/year for a 1.111 tree pure species stand at age 15 (Eisfeld *et al.*, 2018). Although our experimental trial data does not provide direct measures of MAI, estimates were calculated for the top 50 and 100 individual trees in the trial (Supplementary Figure S2) . Data show that such MAI could be potentially reached by a few individual trees and specific progenies in the top performing provenances.

Clearly, a systematic breeding effort based on the data and germplasm provided in this study, together with further improvements in silvicultural practices specifically tailored to the species, could represent an important move toward economic viability of extensive *A. angustifolia* commercial forest plantation. Advanced genomic-based breeding approaches exploiting the power of DNA marker data would be particularly useful to accelerate *A. angustifolia* breeding in the same way as it is currently happening with mainstream conifers and hardwoods (Grattapaglia *et al.*, 2018). The recently developed high-throughput genotyping chip for *A. angustifolia* with 3,000 SNPs (single nucleotide polymorphisms) markers (Silva *et al.*, 2020) opens the prospects of adopting genomic selection to accelerate breeding cycles, increase selection intensity, improve the accuracy of breeding values and innovate in genetic parameters' estimation and breeding approaches. Due to its country-wide distribution, studies involving multiple environmental variables could also be explored for site-specific recommendation of the best genotypes using enviromics approaches (Resende *et al.*, 2020). Finally, given the iconic relevance of the species, breeding programs should also value genetic diversity and the establishment of forests with high environmental adaptive value (Marcatti *et al.*, 2017).

Concluding remarks

Although the species has a long and unfortunate history of over-exploitation, high levels of genetic diversity are still found in the remnant natural populations in Brazil, both at the DNA sequence level (Stefenon *et al.*, 2007; Silva *et al.*, 2020) and at the phenotypic level for growth, as shown in our study. The growth data surveyed in this work match the DNA sequence data described previously as far as pointing to a major separation of the existing provenances into two groups: northern (1 - 7) and southern provenances (8 - 15) (Figure 4). As expected, when evaluated in a common garden trial, provenances coming from close geographical proximity showed similar growth patterns, revealing that the genetic variation within provenances is greater than the variation found between provenances. Our data also underscore the potential for early selection for growth with high prediction accuracy of later ages, and the possibility of shortening the harvest cycle by selecting the best individuals and families. Taken together all the data presented provide significant opportunities for directional selection toward systematic breeding of the species which could in turn foster greater interest and investment in sampling, characterizing and ultimately conserving a wider germplasm base of this valuable keystone Brazilian conifer.

Acknowledgements

This work was supported by (a) PRONEX-FAP-DF (Foundation for Scientific Research of the Federal District) grant NEXTREE 193.000.570/2009, and additional funding from EMBRAPA project 02.11.08.005.00.03 and a CNPq (Brazilian National Council for Scientific and Technological Development) fellowship productivity grant 306866/2018/8 to DG. We would like to thank Carlos Pedro B. Soares (Universidade Federal de Viçosa) for assistance with growth curve modeling, and Ananias de Almeida S. Pontinha, Miguel L. Menezes Freitas, Alexandre Sebbenn and the field staff of the Instituto Florestal de São Paulo (Itapeva experimental station) for technical and logistic support during the field work.

Table 1. Descriptive statistics of the *Araucaria angustifolia* provenance and progeny trial studied and evaluated trait means at age 35 years adopted as a benchmark.

Provenance	No. of Trees	No. of living Trees	Surviving rate (%)	Progenies count	Average number of individuals within family	Trait average (at age 35yr)		
						HEI	DBH	VOL
1	270	154	57.04	9	17.11	18.86 ± 0.31	21.92 ± 0.57	0.49 ± 0.03
2	430	238	55.35	14	17.00	18.37 ± 0.27	21.64 ± 0.51	0.50 ± 0.03
3	180	112	62.22	6	18.67	18.93 ± 0.40	21.46 ± 0.73	0.49 ± 0.04
4	150	84	56.00	5	16.80	19.58 ± 0.41	24.28 ± 0.97	0.64 ± 0.05
5	150	110	73.33	5	22.00	19.27 ± 0.32	20.65 ± 0.65	0.45 ± 0.03
6	210	113	53.81	7	16.14	17.70 ± 0.35	20.01 ± 0.75	0.42 ± 0.04
7	270	186	68.89	9	20.67	15.74 ± 0.30	18.25 ± 0.51	0.32 ± 0.02
8	270	134	49.63	9	14.89	17.46 ± 0.34	19.98 ± 0.62	0.40 ± 0.03
9	300	146	48.67	10	14.60	16.94 ± 0.34	19.04 ± 0.60	0.36 ± 0.02
10	210	117	55.71	7	16.71	17.91 ± 0.32	20.45 ± 0.62	0.41 ± 0.03
11	300	189	63.00	10	18.90	17.28 ± 0.26	18.37 ± 0.46	0.33 ± 0.02

12	270	180	66.67	9	20.00	15.14 ± 0.29	17.25 ± 0.48	0.27 ± 0.02
13	120	67	55.83	4	16.75	14.18 ± 0.46	16.07 ± 0.81	0.23 ± 0.03
14	270	159	58.89	9	17.67	17.31 ± 0.32	19.29 ± 0.53	0.37 ± 0.02
15	270	169	62.59	9	18.78	16.54 ± 0.27	17.45 ± 0.44	0.28 ± 0.02
All	3,670	2,158	59.18	122	17.78	17.37 ± 0.09	19.65 ± 0.16	0.39 ± 0.01

Table 2. Summary of the quantitative genetics parameters for height (HEI) and diameter at breast height (DBH) from observed data and estimated data with the non-linear random regression adjustments.

Parameter	HEI			
	07 yr	**24 yr**	**32 yr**	**35 yr**
R^2	0.914	0.962	0.982	0.984
RMSE (%)	9.984	5.819	3.159	3.502
$\bar{y}$	5.48 / 5.66	16.47 / 16.14	17.14 / 17.17	17.22 / 17.52
h_a^2	0.004 / 0.009	0.157 / 0.153	0.190 / 0.176	0.177 / 0.181
c_{pop}^2	0.114 / 0.110	0.076 / 0.118	0.121 / 0.121	0.133 / 0.122
σ_a^2	0.011 / 0.021	2.774 / 2.070	2.973 / 2.701	2.727 / 2.860
σ_{pop}^2	0.317 / 0.243	1.343 / 1.601	1.894 / 1.863	2.049 / 1.922
σ_{parc}^2	0.339 / 0.304	1.362 / 1.022	1.133 / 1.118	1.921 / 1.148
σ_e^2	2.107 / 1.641	12.242 / 8.834	9.620 / 9.664	9.369 / 9.848
CV_e (%)	13.597 / 11.975	10.330 / 8.841	8.821 / 8.746	8.993 / 8.759
	DBH			
	07 yr	**24 yr**	**32 yr**	**35 yr**
R^2	0.967	0.971	0.991	0.985
RMSE (%)	8.206	5.785	3.676	4.738
$\bar{y}$	8.13 / 8.31	17.28 / 17.25	19.04 / 19.21	20.03 / 20.08
h_a^2	0.005 / 0.007	0.218 / 0.206	0.250 / 0.256	0.233 / 0.266

c^2_{pop}	0.087 / 0.087	0.052 / 0.068	0.060 / 0.062	0.062 / 0.061
σ^2_a	0.052 / 0.052	7.284 / 6.537	11.884 / 11.931	12.464 / 13.638
σ^2_{pop}	0.824 / 0.611	1.740 / 2.161	2.835 / 2.895	3.294 / 3.112
σ^2_{parc}	0.834 / 0.753	1.197 / 1.265	1.059 / 1.321	1.830 / 1.341
σ^2_e	7.773 / 5.633	23.220 / 21.745	31.697 / 30.507	35.970 / 33.185
CV$_e$ (%)	15.926 / 13.705	11.991 / 11.404	11.785 / 11.851	12.474 / 12.033

OBS / EST: Estimation of parameters using the Observed data (on the left of the '/' bar) and those Estimated by the random regression models (on the right of the '/'); h^2_a: heritability in the narrow sense (i.e. additive); c^2_{pop}: provenance coefficient of determination; σ^2_a : additive genetic variance; σ^2_{pop} : variance between populations (or provenances); σ^2_{parc}: variance between experimental plots; σ^2_e: variance within experimental plots.

Table 3. Number of individuals and families (in parentheses) selected for each trait (columns) and provenance (rows) in the trial with 2,158 trees. Three selection intensities were carried out (1%: 22 individuals; 5%: 108 individuals; and 10%: 216 individuals). Selection was carried out at each age (7 - 35 years) and only individual trees that were selected at all ages were kept.

Provenance	HEI			DBH			VOL		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
1	0 (0)	3 (3)	15 (6)	0 (0)	12 (5)	26 (7)	5 (4)	16 (5)	27 (7)
2	6 (2)	21 (5)	49 (9)	10 (2)	25 (5)	49 (8)	8 (3)	31 (8)	58 (9)
3	8 (1)	15 (2)	20 (4)	1 (1)	10 (4)	16 (5)	3 (3)	10 (4)	18 (5)
4	1 (1)	10 (3)	18 (3)	5 (2)	17 (3)	27 (4)	5 (3)	17 (4)	26 (4)
5	0 (0)	6 (2)	13 (3)	0 (0)	9 (3)	17 (4)	1 (1)	12 (4)	16 (4)
6	3 (1)	12 (2)	22 (4)	7 (2)	13 (4)	24 (5)	5 (2)	15 (5)	25 (5)
7	0 (0)	5 (1)	13 (4)	1 (1)	7 (3)	16 (5)	1 (1)	7 (4)	14 (7)
8	5 (3)	26 (4)	39 (4)	7 (3)	17 (3)	30 (6)	2 (2)	17 (6)	26 (7)
9	3 (3)	15 (5)	29 (6)	0 (0)	8 (4)	27 (6)	2 (2)	5 (4)	14 (6)
10	11 (1)	18 (1)	20 (1)	7 (1)	19 (1)	25 (2)	3 (1)	14 (2)	23 (4)
11	2 (2)	15 (3)	26 (5)	1 (1)	12 (3)	25 (5)	0 (0)	11 (5)	23 (7)
12	1 (1)	4 (3)	9 (4)	0 (0)	2 (2)	8 (4)	0 (0)	5 (4)	8 (5)
13	0 (0)	5 (1)	8 (1)	0 (0)	6 (1)	7 (1)	0 (0)	3 (1)	7 (1)
14	5 (1)	18 (5)	29 (6)	3 (1)	14 (5)	26 (5)	5 (3)	16 (6)	25 (7)
15	1 (1)	3 (2)	17 (7)	0 (0)	2 (2)	10 (4)	1 (1)	3 (3)	12 (7)
nID	46	176	327	42	173	333	41	182	322
nFam	17	42	67	14	48	71	26	65	85
nProv	11	15	15	9	15	15	12	15	15

nID: total number of individual trees selected (max=2,158); **nFam**: number of families (max=122); **nProv**: number of provenances (max=15).

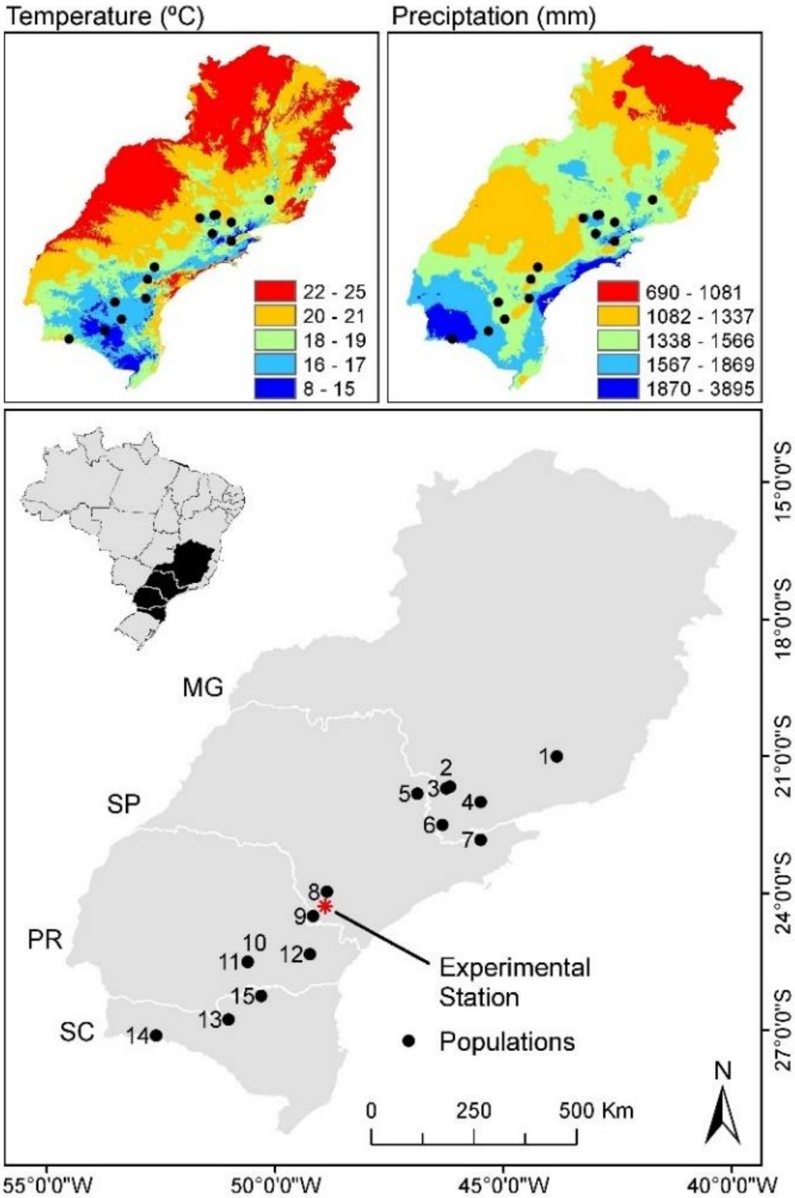

Figure 1. Geographic distribution of the fifteen Araucaria angustifolia provenances studied sampled across the entire natural range of the species in southeastern and southern Brazil

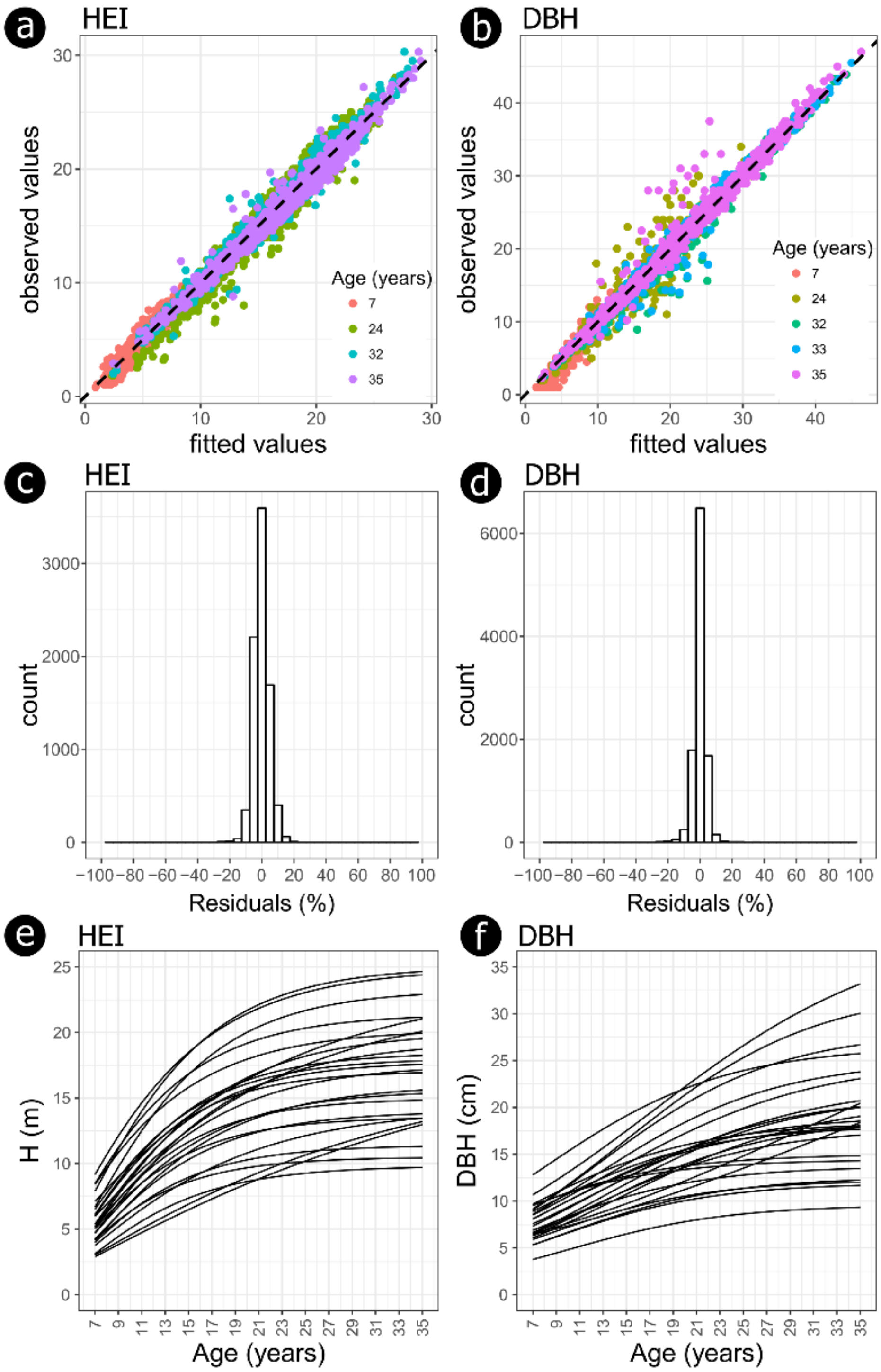

Figure 2. Year-to-year estimate of Tree Height (HEI) and Diameter at Breast Height (DBH) of 2,158 *Araucaria angustifolia* trees in the trial. Panels “a” and “b”: observed *versus* fitted values dispersion of traits values. Panels “c” and “d”: residuals histogram in percentage. Panels “e and “f”: growth curves of 25-tree samples across ages 7 to 35-years.

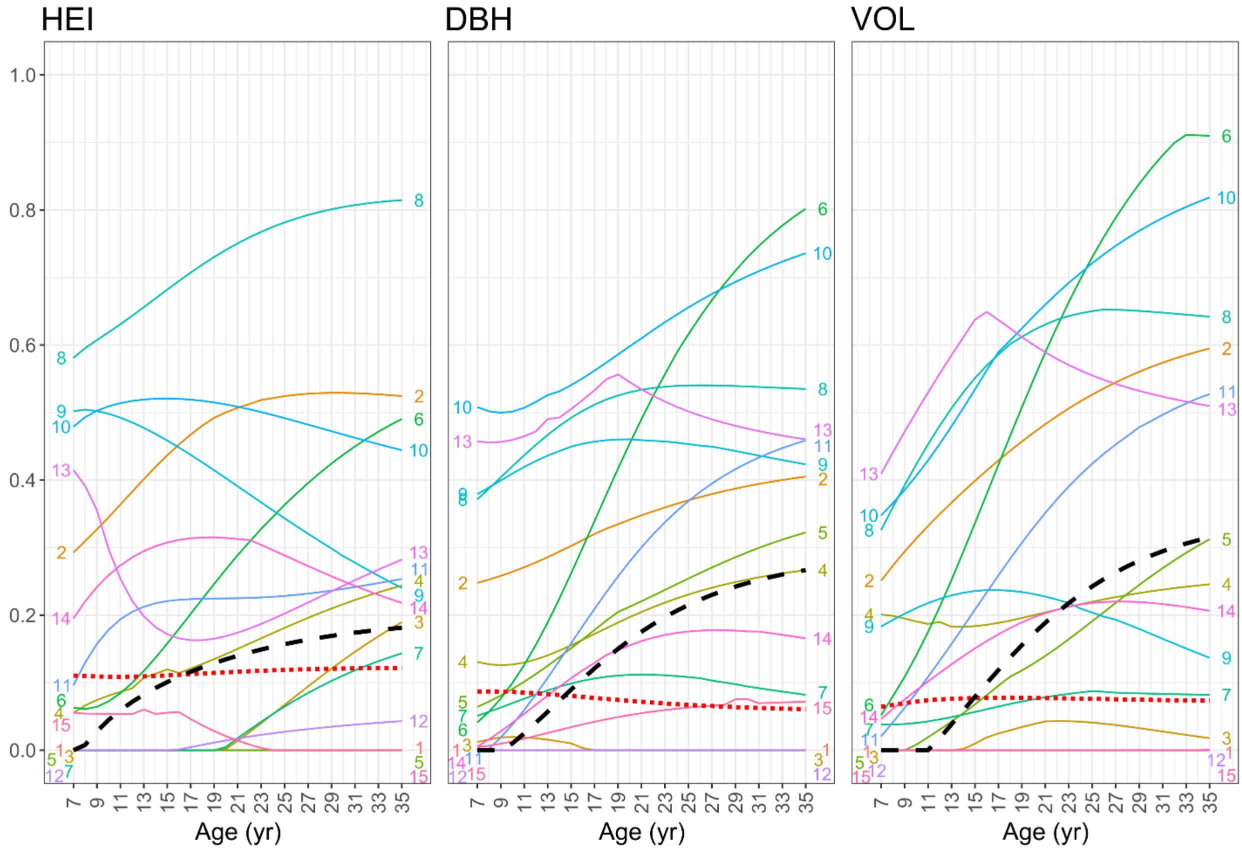

Figure 3. Narrow-sense heritabilities (h_a^2) (Y axis) across ages for traits HEI (tree height), DBH (diameter at breast height) and VOL (individual tree volume). The charts present h_a^2 estimates for a model with all 15 provenances (dashed black lines) and for each individual provenance (colored lines). The red dotted lines are coefficient of population, i.e. provenance, determination (c_{pop}^2) for the complete model with all 15 provenances.

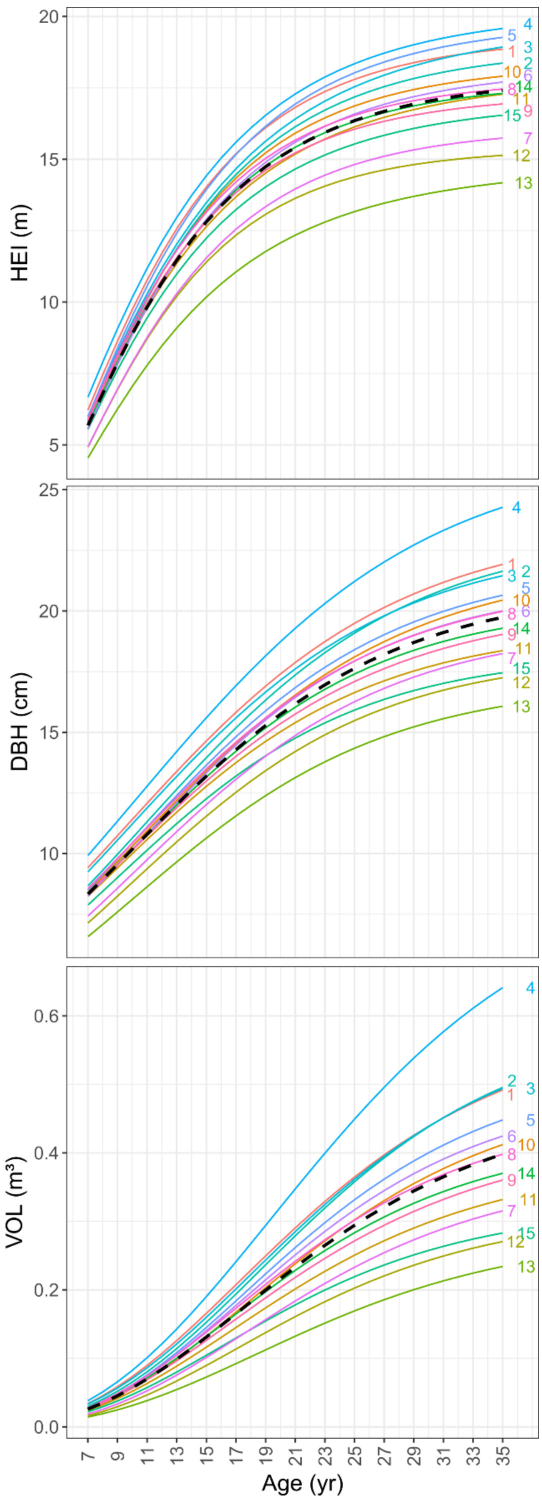

Figure 4. Growth Means of HEI (tree height), DBH (diameter at breast height) and VOL (individual tree volume) fr the 15 *Araucaria angustifolia* provenances. The dashed line is the average accounting for all fifteen provenances.

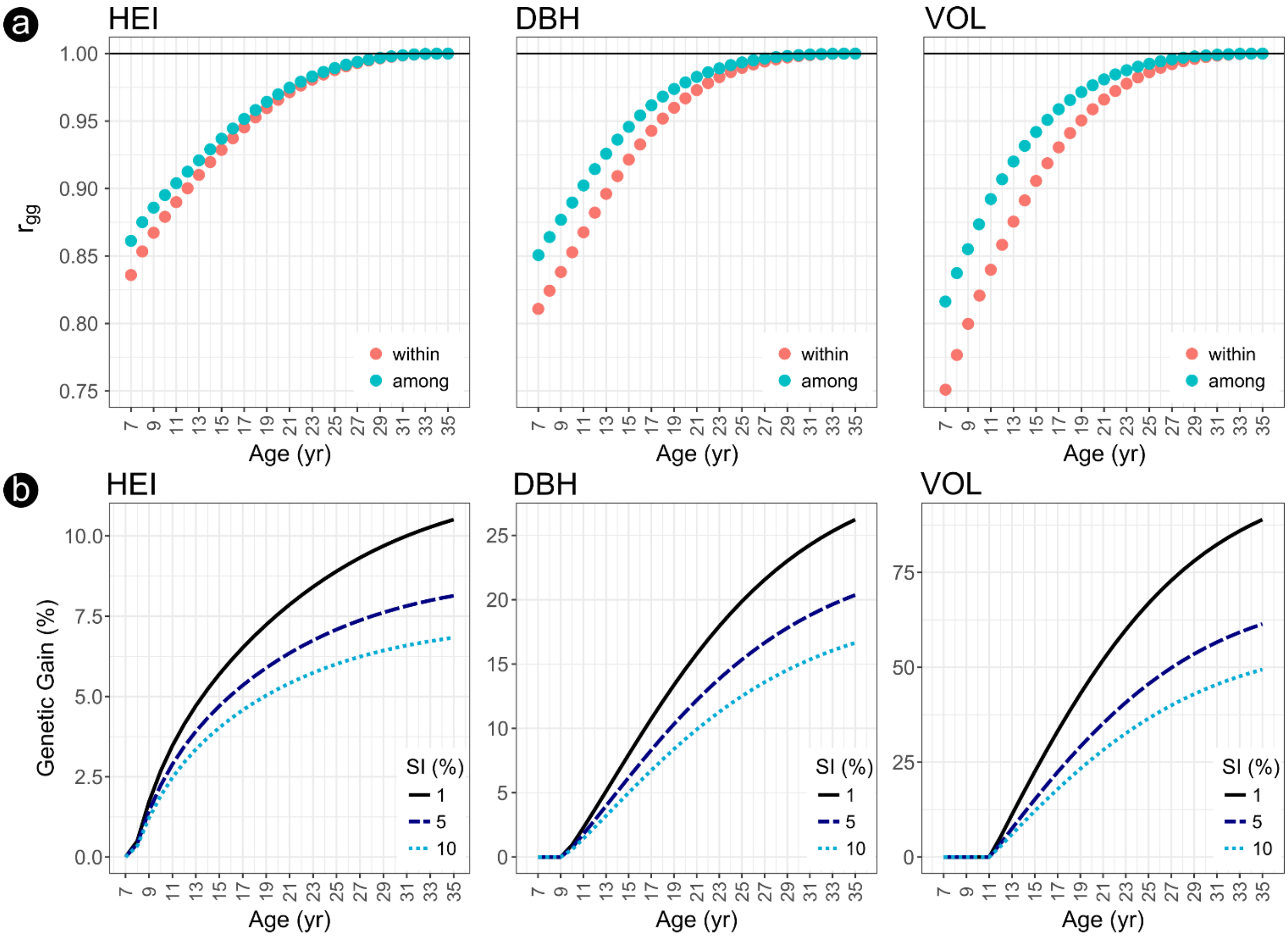

Figure 5. a) Genetic correlation (r_{gg}) between ages and the final age 35 year are presented both as genetic correlations of within (individual) and among progenies to illustrate early selection expectations. **b**) Expected genetic gain following individual tree selection for different selection intensities (SI) (1% (22 individuals), 5% (108 individuals) and 10% (216 individuals).

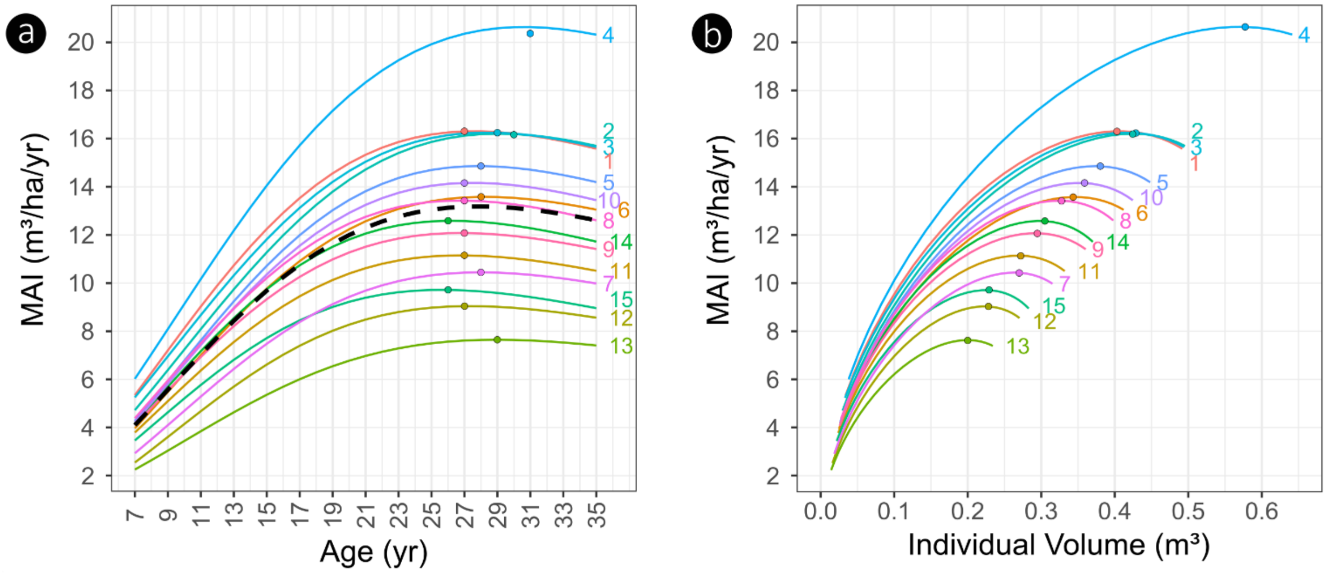

Figure 6. **a)** Relationship of the average tree mean annual increment (MAI) in volume with increasing age ; **b)** Relationship of the average tree mean annual increment (MAI) in volume with the average volume per tree suggesting appropriate moments for rotation age termination (indicated by dots) for the 15 provenances (colored lines).

References

Arnoni Costa, E., Liesenberg, V., Felipe Hess, A., Guimarães Finger, C.A., Renato Schneider, P., Villanova Longhi, R., Schons, C.T., Adriano Borsoi, G., 2020. Simulating Araucaria angustifolia (Bertol.) Kuntze Timber Stocks With Liocourt's Law in a Natural Forest in Southern Brazil. Forests 11, 339.

Balocchi, C.E., Bridgwater, F.E., Zobel, B.J., Jahromi, S., 1993. Age Trends in Genetic Parameters for Tree Height in a Nonselected Population of Loblolly Pine. Forest Science 39, 231-251.

Bigler, C., Veblen, T.T., 2009. Increased early growth rates decrease longevities of conifers in subalpine forests. Oikos 118, 1130-1138.

Bowman, D.M.J.S., Brienen, R.J.W., Gloor, E., Phillips, O.L., Prior, L.D., 2013. Detecting trends in tree growth: not so simple. Trends in Plant Science 18, 11-17.

BRASIL, 2001. Resolução CONAMA nº 278, de 24 de maio de 2001. Dispõe Contra o Corte e Exploração de Espécies Ameaçadas de Extinção da Flora da Mata Atlântica. In: CONAMA (Ed.), 278, Diário Oficial da República Federativa do Brasil - MMA - Brazilian Ministry of the Environment, pp. 157-158.

Burkhart, H.E., Tomé, M., 2012. Modeling forest trees and stands. Springer Science & Business Media.

Calegario, N., Daniels, R.F., Maestri, R., Neiva, R., 2005. Modeling dominant height growth based on nonlinear mixed-effects model: a clonal Eucalyptus plantation case study. Forest Ecology and Management 204, 11-21.

Carter, K.K., Adams, G.W., Greenwood, M.S., Nitschke, P., 1990. Early family selection in jack pine. Canadian Journal of Forest Research 20, 285-291.

Carvalho, P.E.R., 1994. Espécies florestais brasileiras: recomendações silvicultuarais, potencialidades e uso da madeira. EMBRAPA-SPI, Colombo.

Chauhan, S.S., Sharma, M., Thomas, J., Apiolaza, L.A., Collings, D.A., Walker, J.C.F., 2013. Methods for the very early selection of Pinus radiata D. Don. for solid wood products. Annals of Forest Science 70, 439-449.

Clifford, D., McCullagh, P., Clifford, M.D., 2014. The regress package. R package version, 1.3-14.

Costa, P., Durel, C., 1996. Time trends in genetic control over height and diameter in maritime pine. Canadian Journal of Forest Research 26, 1209-1217.

da Silva, J.R., dos Santos, W., de Moraes, M.L.T., Shimizu, J.Y., de Sousa, V.A., de Aguiar, A.V., 2018. Selection of provenances and progenies of Araucaria angustifolia (Bert.) O. Kuntze for wood and seed production. Scientia Forestalis 46, 519-531.

de la Mata, R., Merlo, E., Zas, R., 2014. Among-population variation and plasticity to drought of Atlantic, Mediterranean, and interprovenance hybrid populations of maritime pine. Tree Genetics & Genomes 10, 1191-1203.

Diao, S., Hou, Y., Xie, Y., Sun, X., 2016. Age trends of genetic parameters, early selection and family by site interactions for growth traits in Larix kaempferi open-pollinated families. BMC genetics 17, 1-12.

Dieters, M., White, T., Hodge, G., 1995. Genetic parameter estimates for volume from fuil-sib tests of slash pine (Pinus elliottii). Canadian Journal of Forest Research 25, 1397-1408.

Eisfeld, R.L., Arce, J.E., Sanquetta, C.R., Braz, E.M., 2018. É economicamente viável o plantio de araucária? Uma análise entre a espécie e seu principal substituto, o pinus. Scientia Forestalis 48, e3408.

Farjon, A., Page, C., 1999. Status survey and conservation action plan: conifers.

Feffer, D., Piva, H., Hartung, P., 2019. Brazilian Tree Industry Report / Indústria Brasileira de Árvores (IBÁ). Year-base:2018. In.

Ferreira, D.K., Nazareno, A.G., Mantovani, A., Bittencourt, R., Sebbenn, A.M., Dos Reis, M.S., 2012. Genetic analysis of 50-year old Brazilian pine (Araucaria angustifolia) plantations: implications for conservation planning. Conservation genetics 13, 435-442.

Foster, G.S., 1986. Trends in Genetic Parameters with Stand Development and Their Influence on Early Selection for Volume Growth in Loblolly Pine. Forest Science 32, 944-959.

Grattapaglia, D., Silva-Junior, O.B., Resende, R.T., Cappa, E.P., Muller, B.S.F., Tan, B.Y., Isik, F., Ratcliffe, B., El-Kassaby, Y.A., 2018. Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. Frontiers in Plant Science 9.

Greenwood, M.S., 1982. RATE, TIMING, AND MODE OF GIBBERELLIN APPLICATION FOR FEMALE STROBILUS PRODUCTION BY GRAFTED LOBLOLLY-PINE. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 12, 998-1002.

Gwaze, D., Bridgwater, F., Williams, C., 2002. Genetic analysis of growth curves for a woody perennial species, Pinus taeda L. Theoretical and Applied Genetics 105, 526-531.

Haapanen, M., 2001. Time trends in gentic parameter estimates and selection efficiency for Scots pine in relation to field testing method. Forest genetics 8, 129-144.

Haapanen, M., Hynynen, J., Ruotsalainen, S., Siipilehto, J., Kilpeläinen, M.-L., 2016. Realised and projected gains in growth, quality and simulated yield of genetically improved Scots pine in southern Finland. European journal of forest research 135, 997-1009.

Harfouche, A., Bahrman, N., Baradat, P., Guyon, J.P., Petit, R.J., Kremer, A., 2000. Provenance hybridization in a diallel mating scheme of maritime pine (Pinus pinaster). II. Heterosis. Canadian Journal of Forest Research 30, 10-16.

Hess, A.F., Schneider, P.R., 2009. Crescimento em altura de Araucaria angustifolia (Bertol.) Kuntze em três locais do Rio Grande do Sul. Ambiência 5, 213-232.

Hess, A.F., Schneider, P.R., Finger, C.A.G., 2009. Diameter growth in function of the age of Araucaria angustifolia (Bertol.) Kuntze in three regions of Rio Grande do Sul. Ciência Florestal 19, 7-22.

Hess, A.F.F., Loiola, T., Souza, I.A.d., Minatti, M., Ricken, P., Borsoi, G.A., 2018. FOREST MANAGEMENT FOR THE CONSERVATION OF Araucaria angustifolia IN SOUTHERN BRAZIL. FLORESTA; v. 48, n. 3 (2018).

Hiraoka, Y., Miura, M., Fukatsu, E., Iki, T., Yamanobe, T., Kurita, M., Isoda, K., Kubota, M., Takahashi, M., 2019. Time trends of genetic parameters and genetic gains and optimum selection age for growth traits in sugi (Cryptomeria japonica) based on progeny tests conducted throughout Japan. Journal of Forest Research 24, 303-312.

Hodge, G., Dvorak, W., 2001. Genetic parameters and provenenance variation of Pinus caribaea var. hondurensis in 48 international trials. Canadian Journal of Forest Research 31, 496-511.

Hodge, G.R., Dvorak, W.S., 2015. Provenance variation and within-provenance genetic parameters in Eucalyptus urophylla across 125 test sites in Brazil, Colombia, Mexico, South Africa and Venezuela. Tree Genetics & Genomes 11, 57.

Hodge, G.R., White, T.L., 1992. GENETIC PARAMETER ESTIMATES FOR GROWTH TRAITS AT DIFFERENT AGES IN SLASH PINE AND SOME IMPLICATIONS FOR BREEDING. Silvae Genetica 41, 252-262.

Jansson, G., Li, B., Hannrup, B., 2003. Time trends in genetic parameters for height and optimal age for parental selection in Scots pine. Forest Science 49, 696-705.

Kageyama, P.Y., Jacob, W.S., 1979. Variação genética entre e dentro de populações de Araucaria angustifolia (Bert.) O. Ktze. In, IUFRO (International Union of Forest Research Organizations) Meeting on Forestry Problems of the Genus Araucaria, pp. 83-86.

Kaya, Z., Lindgren, D., 1992. The genetic variation of inter-provenance hybrids of picea abies and possible breeding consequences. Scandinavian Journal of Forest Research 7, 15-26.

Kroon, J., Ericsson, T., Jansson, G., Andersson, B., 2011. Patterns of genetic parameters for height in field genetic tests of Picea abies and Pinus sylvestris in Sweden. Tree genetics & genomes 7, 1099-1111.

Lambeth, C.C., 1980. Juvenile-Mature Correlations in Pinaceae and Implications for Early Selection. Forest Science 26, 571-580.

Leksono, B., Kurinobu, S., Ide, Y., 2006. Optimum age for selection based on a time trend of genetic parameters related to diameter growth in seedling seed orchards of Eucalyptus pellita in Indonesia. Journal of forest research 11, 359-364.

Li, B., Mckeand, S.E., Weir, R., 1996. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the south-eastern US. In, Proceedings of QFRI–IUFRO Conference on Tree Improvement for Sustainable Tropical Forestry, pp. 164-168.

Lindstrom, M.J., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. Biometrics, 673-687.

Longhi, R.V., Schneider, P.R., Longhi, S.J., Marangon, G.P., Costa, E.A., 2018. Growth Dynamics of Araucaria after Management Interventions in Natural Forest. Floresta e Ambiente 25.

Lu, P., Parker, W.C., Colombo, S.J., Man, R., 2016. Restructuring tree provenance test data to conform to reciprocal transplant experiments for detecting local adaptation. Journal of Applied Ecology 53, 1088-1097.

Lynch, M., Walsh, B., 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

Marcatti, G.E., Resende, R.T., Resende, M.D.V., Ribeiro, C.A.A., dos Santos, A.R., da Cruz, J.P., Leite, H.G., 2017. GIS-based approach applied to optimizing recommendations of Eucalyptus genotypes. Forest Ecology and Management 392, 144-153.

McKeand, S.E., 1988. Optimum Age For Family Selection for Growth in Genetic Tests of Loblolly Pine. Forest Science 34, 400-411.

Medina-Macedo, L., Sebbenn, A.M., Lacerda, A.E.B., Ribeiro, J.Z., Soccol, C.R., Bittencourt, J.V.M., 2014. High levels of genetic diversity through pollen flow of the coniferous Araucaria angustifolia: a landscape level study in Southern Brazil. Tree Genetics & Genomes 11, 814.

Montagna, T., Ferreira, D.K., Steiner, F., da Silva, F.A.L.S., Bittencourt, R., da Silva, J.Z., Mantovani, A., dos Reis, M.S., 2012. A importância das unidades de conservação na manutenção da diversidade genética de araucária (Araucaria angustifolia) no Estado de Santa Catarina. Biodiversidade Brasileira 2, 18-25.

Newton, P.F., 2003. Systematic review of yield responses of four North American conifers to forest tree improvement practices. Forest Ecology and Management 172, 29-51.

Nutto, L., Spathelf, P., Rogers, R., 2005. Managing diameter growth and natural pruning of Parana pine, Araucaria angustifolia (Bert.) O Ktze., to produce high value timber. Annals of Forest Science 62, 163-173.

O'Brien, E.K., Mazanec, R.A., Krauss, S.L., 2007. Provenance variation of ecologically important traits of forest trees: implications for restoration. Journal of Applied Ecology 44, 583-593.

Park, Y.S., 2002. Implementation of conifer somatic embryogenesis in clonal forestry: technical requirements and deployment considerations. Annals of Forest Science 59, 651-656.

Perez, A.M.M., White, T.L., Huber, D.A., Martin, T.A., 2007. Graft survival and promotion of female and male strobili by topgrafting in a third-cycle slash pine (Pinus elliottii var. elliottii) breeding program. Canadian Journal of Forest Research 37, 1244-1252.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., 2016. R Core Team (2016) nlme: Linear and nonlinear mixed effects models. R package version 3.1-127. Vienna, Austria: R Foundation for Statistical Computing.

Reis, M.S., Ladio, A., Peroni, N., 2014. Landscapes with Araucaria in South America: evidence for a cultural dimension. Ecology and Society 19.

Resende, M.D.V.d., 2016. Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breeding and Applied Biotechnology 16, 330-339.

Resende, R., Piepho, H., Rosa, G., Silva-Junior, O., FF, E.S., de Resende, M., Grattapaglia, D., 2020. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. TAG. Theoretical and Applied genetics. Theoretische und Angewandte Genetik.

Sanquetta, C.R., Dolci, M., Dalla Corte, A.P., Niroh, M., Sanquetta, I., Pelissari, A.L., Meissner, A.L., Botânico, J., 2016. Estimação de volumes de Araucaria angustifolia (Bertol.) O. Kuntze por fatores de forma em classes diamétricas e modelos de regressão. Centro Científico Conhecer 13, 588-597.

Santini, E., Haselein, C., Gatto, D., 2000. Comparative analysis of physical and mechanical properties of wood from three softwood plantations. Ciência Florestal 10, 85-93.

Sebbenn, A., Pontinha, A., Giannotti, E., Kageyama, P., 2003. Genetic variation in provenance-progeny test of Araucaria angustifolia (Bert.) O. Ktze. in Sao Paulo, Brazil. Silvae genetica 52, 181-184.

Shimizu, J.Y., Jaeger, P., Sopchaki, S.A., 2000. Variabilidade genética em uma população remanescente de Araucária no Parque Nacional do Iguaçu, Brasil. Embrapa Florestas-Artigo em periódico indexado (ALICE).

Silva, P.I.T., Silva-Junior, O.B., Resende, L.V., Sousa, V.A., Aguiar, A.V., Grattapaglia, D., 2020. A 3K Axiom SNP array from a transcriptome-wide SNP resource sheds new light on the genetic diversity and structure of the iconic subtropical conifer tree Araucaria angustifolia (Bert.) Kuntze. PLOS ONE 15, e0230404.

Stackpole, D.J., Vaillancourt, R.E., de Aguigar, M., Potts, B.M., 2010. Age trends in genetic parameters for growth and wood density in Eucalyptus globulus. Tree Genetics & Genomes 6, 179-193.

Stefenon, V.M., Gailing, O., Finkeldey, R., 2007. Genetic structure of Araucaria angustifolia (Araucariaceae) populations in Brazil: implications for the in situ conservation of genetic resources. Plant Biol (Stuttg) 9, 516-525.

Stefenon, V.M., Gailing, O., Finkeldey, R., 2008. Genetic structure of plantations and the conservation of genetic resources of Brazilian pine (Araucaria angustifolia). Forest Ecology and Management 255, 2718-2725.

Subedi, N., Sharma, M., 2011. Individual-tree diameter growth models for black spruce and jack pine plantations in northern Ontario. Forest Ecology and Management 261, 2140-2148.

Tambarussi, E.V., Boshier, D., Vencovsky, R., Freitas, M.L.M., Sebbenn, A.M., 2017. Inbreeding depression from selfing and mating between relatives in the Neotropical tree Cariniana legalis Mart. Kuntze. Conservation Genetics 18, 225-234.

Thomas, P., 2013. Araucaria angustifolia. The IUCN red list of threatened species 2013.

Trevisan, R., Zanella, A., Silva, F.M.d., Rosa, M., Fioresi, T., Fortes, F.d.O., 2016. Axial variation of basic density of Araucaria angustifolia wood in different diameter classes. Ciência Rural 46, 1969-1972.

Vencovsky, R., Chaves, L.J., Crossa, J., 2012. Variance effective population size for Dioecious species. Crop science 52, 79-90.

Wendling, I., Stuepp, C.A., Zuffellato-Ribas, K.C., 2016. Araucaria clonal forestry: types of cuttings and mother tree sex in field survival and growth. Cerne 22, 19-26.

Weng, Y., Tosh, K., Park, Y., Fullarton, M., 2007. Age-related trends in genetic parameters for jack pine and their implications for early selection. Silvae Genetica 56, 242-252.

White, T.L., Adams, W.T., Neale, D.B., 2007. Forest Genetics. CABI Publishing, 682 pp., Cambridge, MA.

Xiang, B., Li, B., Isik, F., 2003. Time trend of genetic parameters in growth traits of Pinus taeda L. Silvae genetica 52, 114-120.

Xie, C., Ying, C., 1996. Heritabilities, age-age correlations, and early selection in lodgepole pine (Pinus contorta ssp. latifolia). Silvae Genetica 45, 101-106.

Capítulo 2

A 3K Axiom® SNP array from a transcriptome-wide SNP resource sheds new light on the genetic diversity and structure of the iconic subtropical conifer tree *Araucaria angustifolia* (Bert.) Kuntze

Pedro I.T. Silva [1,2,#], Orzenil Bonfim Silva-Junior [1], Lucileide V. Resende.[1], Valderes A. Sousa [3], Ananda V. Aguiar [3], and Dario Grattapaglia [1,4*]

[1]Plant Genetics Laboratory, EMBRAPA Genetic Resources and Biotechnology, CEP 70770-970, DF, Brasilia, Brazil.

[2]University of Brasília, Cell Biology Department, Campus Universitário, Asa Norte 70910-900, DF, Brasília, Brazil.

[3] Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Florestas. CEP: 83411-000. PR, Colombo, Brazil.

[4] Graduate Program in Genomic Sciences, Universidade Católica de Brasília, Brasília, DF, Brazil.

[#]Current address: Corteva Agriscience™ Guarapuava Research Station - Rodovia PR 540, Km 11, Colônia Vitória. CEP 85.139-400, PR, Guarapuava, Brazil

Keywords

SNP array, transcriptome, genetic diversity, genetic structure, SNPs, microsatellites, conifer

Os resultados descritos neste capítulo encontram-se também publicados na forma de artigo científico no periódico PLOS One (https://doi.org/10.1371/journal.pone.0230404)

Abstract

High-throughput SNP genotyping has become a precondition to move to higher precision and wider genome coverage genetic analysis of natural and breeding populations of non-model species. We developed a 44,318 annotated SNP catalog for *Araucaria angustifolia*, a grandiose subtropical conifer tree, one of the only two native Brazilian gymnosperms, critically endangered due to its valuable wood and seeds. Following transcriptome assembly and annotation, SNPs were discovered from RNA-seq and pooled RAD-seq data. From the SNP catalog, an Axiom® SNP array with 3,038 validated SNPs was developed and used to provide a comprehensive look at the genetic diversity and structure of 15 populations across the natural range of the species. RNA-seq was a far superior source of SNPs when compared to RAD-seq in terms of conversion rate to polymorphic markers on the array, likely due to the more efficient complexity reduction of the huge conifer genome. By matching microsatellite and SNP data on the same set of *A. angustifolia* individuals, we show that SNPs reflect more precisely the actual genome-wide patterns of genetic diversity and structure, challenging previous microsatellite-based assessments. Moreover, SNPs corroborated the known major north-south genetic cline, but allowed a more accurate attribution to regional versus among-population differentiation, indicating the potential to select ancestry-informative markers. The availability of a public, user-friendly 3K SNP array for *A. angustifolia* and a catalog of 44,318 SNPs predicted to provide ~29,000 informative SNPs across ~20,000 loci across the genome, will allow tackling still unsettled questions on its evolutionary history, toward a more comprehensive picture of the origin, past dynamics and future trend of the species' genetic resources. Additionally, but not less importantly, the SNP array described, unlocks the potential to adopt genomic prediction methods to accelerate the still very timid efforts of systematic tree breeding of *A. angustifolia*.

Introduction

The development of high throughput genotyping tools based on large numbers of SNP markers (single nucleotide polymorphisms) has become a prerequisite to move to a higher level of precision and genome coverage for the genetic analysis of natural and breeding populations of non-model organisms [1]. Genome-wide genotyping technologies provide exceptional opportunities to advance the understanding of the overall patterns of genetic diversity to drive conservation efforts [2] and inform genomic assisted breeding [3]. Next-generation sequencing (NGS) technologies have facilitated the task of SNP discovery in plant and animal genomes using different approaches that allow genome complexity reduction and more recently methods based on low to ultra-low sequencing of the whole genome. The most traditional and affordable complexity reduction method has been by RNA sequencing (RNA-seq), in which cDNA is made through reverse transcription from only a fraction of the transcribed genome, followed by sequencing and variant calling. Other methods have adopted different approaches of restriction enzyme digestion followed by high throughput sequencing such as Restriction site associated DNA (RAD-seq) [4] the various alternative protocols of genotyping by sequencing (GbS) [5], and targeted enrichment by sequence capture [6]. These methods have allowed not only the discovery of large numbers of SNPs, but also direct SNP genotyping at accessible costs for under resourced plant and animal species [7].

While targeted enrichment by sequence capture do provide generally reliable and portable SNP genotyping data in highly heterozygous species [8], several are the challenges of restriction enzyme based methods for robust SNP genotyping due to variable sequencing coverage, irregular sampling of loci and lack of a reference genome, causing frequent allele dropout and variable genotype reproducibility [9]. The final number of robust and portable SNPs across experiments is typically only a small fraction of the initial set, defeating the alleged cost advantage and possibly biasing genetic diversity measures [10,11]. For high reproducibility, high throughput genotyping fixed content SNP arrays are currently the gold standard and the only validated platform adopted in humans, major animal and crop species. With the substantial price reductions of competing technologies [12] and the possibility of designing multi-species SNP arrays [13], these platforms have become accessible at a fraction of what the cost used to be, translating to equivalent or lower price per informative data point when compared to GbS methods [11].

Araucaria angustifolia (Bertol.) Kuntze is an iconic long-lived subtropical conifer tree endemic to Southern and Southeastern Brazil and to minor areas in Argentina and Paraguay. It stands out as the keystone gymnosperm species native to Brazil and the leading species in the mixed Ombrophylous Forest (a.k.a. Araucaria Forest)[14]. Currently, with a strong reduction of its original old-growth forest area, the Araucaria Forest biome is one of the most threatened in Brazil, with *A. angustifolia* also called Paraná Pine included as critically endangered in the IUCN Red List of Threatened Species [15]. Besides its ecologically keystone role, *A. angustifolia* had a historically important social and economic role during the European colonization of Southern Brazil [16] and as a outstanding looking tree it is frequently planted for its ornamental appearance in gardens and homes for its aesthetic value. Isozymes allowed the first estimates of genetic diversity, structure and mating system [17-21]. Next, studies were carried out using dominant AFLP markers [22-24], or small sets of five to 15 microsatellites to compare the genetic diversity among natural and planted forest stands or estimate spatial genetic structure, mating system and gene flow [22,25-31]. The ultimate goal of these studies has been to provide evidence-based information for supporting conservation strategies. Generally, high levels of genetic diversity have been found, suggesting that *A. angustifolia* is resilient to forest fragmentation, maintaining adequate diversity for sustainable evolution [30,32]. Notwithstanding the existing information on the population genetics of *A. angustifolia*, marker resources and data gathered thereof are still restricted to very few microsatellite loci. Due to their small number, ascertainment bias and mutational behavior, microsatellites are limited as reliable predictors of genetic variation and historical demography of natural populations [33-35]. Clearly, the current molecular toolbox for the analysis of sequence variation in *A. angustifolia* is insufficient for answering important remaining questions, or else, corroborating or challenging the current view on the levels, distribution and dynamics of the current genetic diversity in this iconic subtropical conifer.

In conifers, due to relatively high past implementation costs of fixed content SNP arrays, moderate to high density chips with hundreds or up to several thousand markers have been developed exclusively for the mainstream commercially relevant genera for which funding is abundant. These have included *Picea* spp. [36-39], *Pinus* spp. [38,40-43], *Cunninghamia* spp.[44] and *Cryptomeria japonica* [45,46]. More recently, however, with advances in technology and cost reductions, SNP arrays have become very competitive with alternative sequence-based SNP genotyping methods.

In this study, we describe the development of a large annotated SNP catalog for *A. angustifolia*, together with an Axiom® SNP array with ~3,000 validated SNPs. The array was subsequently used to provide a comprehensive look at the genetic diversity and structure of population samples across the entire natural geographical range of the species in Brazil and compare the estimates with microsatellite markers genotyped on the same individuals. The Axiom SNP array described is fully public and accessible to anyone interested.

Materials and Methods

Plant material and RAD-seq data

RAD-seq libraries were prepared from two genomic DNA samples. Taking advantage of the haploid biology of conifers and the large *A. angustifolia* seeds, the first DNA sample was extracted from a single haploid megagametophyte. The second sample was an equimolar pool of DNA extracted from diploid needle tissue of twelve unrelated trees to serve as a representative sample of diversity for SNP detection. Genomic DNA was extracted with an optimized protocol for challenging samples [47]. The two genomic DNA samples were sent to Floragenex (Portland, OR, USA) for RAD-seq [4], with the difference that the restriction enzyme *PstI* was used for plant genome complexity reduction. The sheared, sequencer-ready fragments were size selected (~200-500 bp) and the RAD-seq libraries sequenced on a Genome Analyzer II (Illumina, San Diego, CA), paired-end 2 x 100 bp mode. RAD tags were used for SNP discovery and selection of SNPs. SNPs were evaluated by a population survey involving a set of 185 35-year old trees sampled in a provenance-progeny trial established in an experimental station in Itapeva, São Paulo State (23°58′56″S 48°52′32″W) with seeds collected in 15 different natural populations covering the geographical range of the species (Fig. 7 and S1 Table). Sample collection from natural populations for this genetic study (authorization number 02001.007609/2012-77) was issued by the Brazilian Institute of the Environment (IBAMA), the regulating body of the Brazilian Ministry of Environment. Genomic DNA for downstream SNP and microsatellite genotyping was extracted from needle or bark tissue with the same protocol as for the RAD-seq samples.

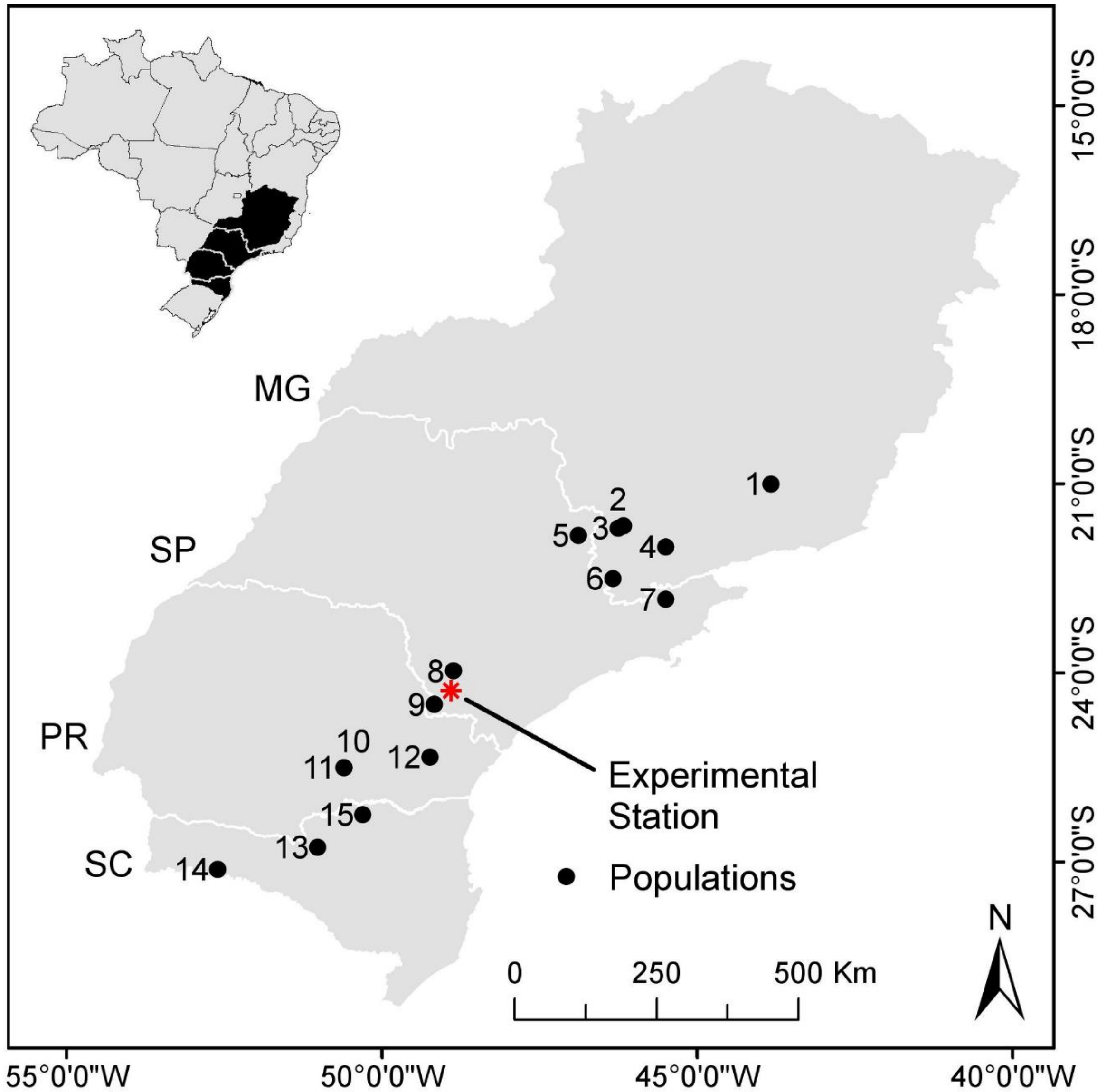

Figure 7 - Geographic distribution of the 15 Araucaria angustifolia populations studied. Population code used and respective locations are: 1.BAR: Barbacena – MG; 2.IPI: Ipiúna de Calda, MG; 3.CON: Congonhal, MG; 4. LAM: Lambarí, MG; 5.VAR: Vargem Grande do Sul, SP; 6.CAM: Camanducaia, MG; 7.CJO: Campos do Jordão, SP; 8.ITA: Itapeva, SP; 9.ITR: Itararé, SP; 10.IRA: Iratí, PR; 11. IRT: Iratí (Tardio), PR; 12.QBA: Quatro Barras, PR; 13.CAC: Caçador, SC; 14.CHA: Chapecó, SC; 15:TRB: Três Barras, SC. Indicated also the experimental station where the provenance/progeny field trial was established and actual samples collected for the population survey.

RNA sequence data

RNA-seq sequence data with insert size of approximately 200 bp was generated by Elbl et al. [48] and obtained from the NCBI SRA (short read archive) repository. The plant tissue material used for cDNA sequence data generation consisted of three pools of seeds from three different megastrobiles and three somatic embryogenic cultures of *A. angustifolia* as described. The raw dataset of 642 million 100 bp reads was downloaded, quality filtered down to 326 million reads, and used for *de novo* transcript assembly and SNPs discovery. Due to the large size (1C= 22 Gb) and highly repetitive nature of the *A. angustifolia* genome [49], a thorough analytical procedure for SNP discovery and ascertainment was adopted to maximize the likelihood of successful downstream SNPs genotyping.

Transcriptome assembly

As a first step we used the StringTie method to reconstruct cDNA fragments for both end of the reads [50]. StringTie makes use of super-reads module from MaSuRCA [51] to vastly reduce the original data set of reads contributing to a better performance of the subsequent transcriptome assembly using a *de novo* approach. A perl script (http://ccb.jhu.edu/software/stringtie/dl/superreads.pl) was used to inspect the super-reads to identify all pairs of reads that belonged to a unique string and to extract the sequence containing the pair plus the sequence between them. These steps converted many of the original paired-end reads into single 200 bp super-reads as though coming from a 200-bp fragment library. Afterwards, miraSearchESTSNPs from MIRA 4 assembler [52] was used to generate a transcript assembly from the super-reads. The assembly was inspected for any positions with conflicts that could not be resolved automatically by MIRA. These suspicious positions were collected and formatted as a single file containing coordinates along the contigs in BED format. The consensus sequences in the assembly were further evaluated with TransRate [53] to detect contigs that exhibited signs of chimerism, structural errors, incomplete assembly and transcripts that were likely paralogs.

RAD Sequences alignment and variants discovery

RAD-seq data was analyzed using a reference-free bioinformatics strategy and parameters described in Senn et al. [54] to end with sufficient flanking region around the SNPs to allow adequate probe design for the fixed content SNP array. To make use of all the sequence information when building the catalogue, the RAD sequenced samples were treated as two 'parent' individuals, named *single_haploid* and *pool_diploid*, respectively. Based on the catalog of formed loci, Read 2 tags were collated separately for each 'parent'. To avoid SNP calls at low coverage for Read 2 tags, for downstream analyses we only kept loci where 20 or more reads were available in total. Afterwards, we used CORTEX_VAR [55] to simultaneously form the entire sequence of loci and calling of variants using Read 2 tags that had representation across the two 'parent' individuals. Again, we followed standard steps described in Senn et al. [54] to simultaneously assemble contigs for each of the loci from collated Read 2 tags, to identify putative SNPs, to characterize the actual alleles of both 'parents' individuals and to genotype. We used the output of CORTEX_VAR to form a pseudo-reference genome containing one 'chromosome' per identified SNP, whose sequence provided the 5-prime flank, the SNP alternative alleles and the 3-prime flank for the variant in the 'parent' named single. We then processed the output of CORTEX_VAR to build a VCF-like file using this pseudo-reference to provide a coordinate system using the script process_calls.pl in CORTEX_VAR. This script collects the information on the variant calls, alleles and their read coverage and the most likely genotype at each SNP/InDEL for each 'parent' individual, along with a genotype confidence (GT_CONF) using a maximum likelihood approach against the confidence of the second most likely genotype.

RNA-seq alignment and sequence variants discovery

Contigs in the transcriptome assembly were used as reference to align back the original RNA-seq paired-end reads. We used only well represented contigs in the sequence reads in the Transrate analysis. We used the Novocraft Version 3.03 (Novocraft Technologies Sdn. Bhd., Malaysia) [56] suite of programs to perform all the alignments to the reference assembly, using standard paired-end processing with base quality calibration options disabled and setting SAM ouput format using "-o SAM". SAM formatted files were subsequently converted to BAM file of alignments for the reads from the eight cDNA libraries. To make use of all sequence information for polymorphism screening, samples were treated as coming from two 'parent' individuals, named *single_haploid* and single*_diploid* respectively. While all seeds were collected from a single individual tree, being the tree outcrossed and thus highly heterozygous, the haploid samples embodied within-tree variation and the diploid both the within and between tree variation. A single BAM file of alignments with read group information for the two contributing samples was processed for marking of duplicates and sorted using Picard [57] and Indel Realignment. The output BAM file was then processed for SNP calls and genotyping using GATK HaplotypeCaller analysis [58]. A VCF file of putative variants (SNPs/InDels) was ultimately obtained.

SNP filtering and prioritization for array design

VCF files from the sequence variant discovery carried out on RAD-seq and RNA-seq data were further refined to retain only positions in the references denoting putative high-quality SNPs. First we inspected each variant and set the genotype calls to NULL for samples where the genotype confidence (GT_CONF) was lower than three and genotype quality (GQ) lower than 30 for the RAD-seq and RNA-seq analyses, respectively. Variant sites without any genotype call across the two samples were removed from the VCF files. Then, several filter tags were applied to the remaining putative variant sites by inspecting their sequence vicinity and context. Each variant classified as InDel was tagged as *TypeIndel*. Each SNP variant located in the vicinity, 60-bp in each direction, of an InDel was tagged as *SnpGap*. Additionally, if more than one SNP variant was found within a 30 bp window on both sides of the target variant, all of them were deemed hotspots of clustered SNPs and tagged as *snpCluster*. Additionally, as low complexity regions in the DNA sequences contribute to highly variable variant calls between callers [59], we inspected the set of reference sequences targeting the identification of low complexity regions (LCRs) and a BED file of LCRs and repeats was generated and all the SNPs within their coordinates were tagged as *LowComplexityRegion*. For SNPs in the RNA-seq analysis, GATK's suggested hard filter criteria were applied so that variants that failed the filter were tagged as *FailureOnGatkHardFilter*. Taking advantage of the haploid biology of *Araucaria*, we set a tag named *FailHaploidTest* at each variant site where the haploid sample genotype was declared as heterozygous opposed to homozygous (when a genotype was emitted) or NULL (in the case of absence of a confident genotype). Worth to mention that the same SNP could end up having several of these tags since the filtering steps were simultaneous. Finally, we only used A/C, A/G, C/T and G/T polymorphisms as these require a single probe therefore optimizing the space available on the array. To adhere to the Axiom's platform recommendation for probe design, we extracted 35-bp sequences in each direction of any target SNP variant that passed all the filtering tags. All the 71-bp long sequences that likely encompassed a single putative SNP variant were inspected against the complete set of reference sequences obtained by the concatenation of contigs from the transcriptome assembly and pseudo-reference from the RAD-seq loci. Probes were then classified as not recommended when more than 100 matches of any 16-mers in its sequence were found in the reference.

Functional annotation and classification of the SNP catalog

The full collection of 22,983 sequences representing RAD loci and RNA transcripts from RAD-Seq and RNA-Seq assemblies, respectively, containing SNP variants with the status of "PASS" was subjected to further characterization. First, the sequences were corrected for possible frameshifts and missing signals using default parameters in FrameDP [60] and then characterized for protein similarity & classification, structural properties, gene ontology and annotation using Blast2GO 5 PRO [61]. We used Blast2GO to run BlastX against the Non-redundant protein sequences (nr), with an e-value cutoff of 1.0E-3, word size of three, a maximum number of hits of 20 and using the application low complexity filter. InterProScan was also ran to search for protein families, domains and sites against several databases – CDD, HAMAP, HMMPanther, HMMPfam, HMMPIR, FPrintScan, BlastProDom, ProfileScan, HMMTigr, Gene3D, SFLD, SuperFamily and MobiDBLite. A GO mapping and annotation of the protein Blast hits was carried out against curated Gene Ontology annotated proteins using default parameters.

SNP array design and validation

For the 44,318 assayable SNPs potential probes were designed for each one in both the forward and reverse direction. A set of 3,400 randomly selected SNPs among the 44,318 was submitted to ThermoFisher to populate the fixed content SNP array. This number of probes was defined based on the available space on a multispecies Axiom® myDesign ™ array that contained a total of 51,867 SNPs, shared among five different plant species, significantly reducing the individual sample genotyping cost while at the same time allowing access to a high quality SNP array for the underfunded species *A. angustifolia* [62]. The four additional species on the array besides *Araucaria angustifolia* (order Pinales) not only belong to different families but also to different phylogenetic orders to minimize the possibility of genome sequence homology. These were: *Anacardium occidentale* (cashew – order Sapindales), *Manihot esculenta* (cassava – order Malpighiales), *Coffea robusta* (coffee – order Gentianales) and *Eucalyptus* sp. (order Myrtales). Additionally, all probes designed for the five species were subject to a detailed sequence evaluation to avoid SNP probe cross talking.

SNP genotyping of a total of 192 samples, 185 unique and seven duplicated for reproducibility estimates, was carried out at ThermoFisher (Santa Clara, CA) and data analyzed using the Axiom Analysis Suite 3.1 [63]. Samples with a dish quality control (DQC) value >0.82 and call rate, CR >0.97 following the recommended "Best Practices Workflow" were considered to have passed the sample quality control assessment. Two criteria were used to classify successfully genotyped SNPs and converted polymorphic SNPs. The first stricter criteria set, suggested as default by the Axiom Analysis Suite used in human SNP evaluation, required a SNP CR > 97%. By this method, SNPs were classified into the following categories: (i) PHR (Polymorphic High Resolution) when the SNP passes the quality criteria and polymorphism measured by the presence of the minor allele in two or more samples, which, for N=185 translated to a minimum allele frequency MAF≥ 0.005; (ii) MHR (Monomorphic High Resolution) when the SNP passes the quality criteria except for polymorphism; (iii) CRBT (Call Rate Below Threshold) when the SNP CR was <97%; (iv) NMH (No Minor Homozygote) when the SNP passes all QC but only two genotype clusters are detected, (v) OTV (Off-Target Variant) where additional clusters arise from unaccounted sequence variants in the SNP flanking region, and (vi) OTH (Other) when the SNP can't be classified into any of the previous categories. Under this strict classification, only SNPs in categories PHR, MHR and NMH are retained as successful SNPs, while the SNPs in PHR class are those considered converted. The second criteria to declare a successful SNP adopted a more liberal CR >90% commonly used for plant and animal SNP genotyping. The same MAF≥ 0.005 as in the Axiom criteria was used to declare a polymorphic SNP.

Genetic diversity and structure analyses with microsatellites and SNPs

To assess the performance of the SNP array for population genetics analyses, the same 185 trees were also genotyped with a set of 8 previously published and widely used *A. angustifolia* microsatellites [64]. Microsatellite genotyping was carried out by fluorescence detection using previously described protocols [65]. Briefly, PCRs were carried out in tetraplex systems with primers labeled with fluorochromes (6-FAM, NED, VIC) and the PCR mixture with ROX-labeled size standard [66] electroinjected in an ABI 3100XL genetic analyzer and data collected using GeneMapper (Thermo Fisher Scientific). SNP and microsatellites estimates of allele frequencies, population diversity , i.e. observed (H_o) and expected (H_e) heterozygosity, coefficient of inbreeding (F_{is}) based on Weir and Cockerham inbreeding estimator (f) [67] and the fixation index as a measure of population differentiation (F_{st}) with their respective 95% confidence intervals were obtained using GDA (Genetic Data Analysis) [68]. AMOVA (Analysis of Molecular Variance) and Principal coordinate analyses (PCoA) plots were obtained using GENALEX 6.501 [69]. Genetic structure was further assessed using the approach implemented by STRUCTURE [70] under an admixture model with correlated allele frequencies. Jobs were run applying a burn-in length of 50,000 and 50,000 iterations for data collection with K ranging from 2 up to 15 potentially inferred clusters, performed with 10 independent runs each. For the microsatellite data and reduced 80 SNPs set the analysis was carried out using STRUCTURE v2.3.4 on a single computer while for the large SNP data set a multi-core computer was used with ParallelStructure[71]. Outputs of STRUCTURE were used to define the most probable number of K clusters by the 'ΔK' metric [72] using Structure Harvester [73] and to generate a consensus solution and plots of the 10 independent runs by a Markov clustering algorithm implemented by CLUMPAK [74].

Results and discussion

RAD-Seq vs RNA-seq for SNP discovery in a complex genome

Using RAD-seq, 44,332,020 and 15,920,336 reads were generated from the single sample haploid library and the multi-sample pooled library, respectively, totaling 60,252,356 reads. Reads from the single sample haploid library were used for the assembly of a haploid pseudoreference. From all 60,252,356 reads, a total of 176,629 contigs with size >120bp and average size of 325 bp was obtained and used for SNP discovery. For the RNA-seq resource, from the full set of 642 million RNA-seq reads, after applying the quality filters, a total of 326 million reads were ultimately used in transcriptome assembly, resulting in 43,608 unique contigs. The total number of raw sequence variants discovered following alignment and sequence variants discovery were 17,428 for RAD-seq and 309,509 for RNA-seq. The simultaneously applied quality filter steps adopted excluded the vast majority of SNPs due to failure in one or more criteria adopted for SNP selection (Table 4).

Table 4 - Summary of the number of single-nucleotide polymorphisms (SNPs) filtered out from each sequence source (RAD and RNA sequences) following the simultaneous filters applied for SNP selection toward the construction of the *Araucaria angustifolia* 3K SNP Axiom® Array.

	RAD-seq	**RNA-seq**
Number of high quality reads used for contig assembly	20,720,596	326,000,000
Number of high quality reads used SNP discovery	60,252,356	326,000,000
Number of contigs used for SNP detection	176,629	43,608
Total number of raw SNPs discovered	17,428	309,509
Simultaneously applied refining filter for SNP exclusion		
Number of SNPs located in contigs matching known transposon	-	40,538
Number of SNPs failing GATK best practice filter	-	45,279
Number of SNPs failing haploid test (heterozygous in haploid sample)	5,226	-
Number of SNPs located in homopolymer regions	763	-
Non assayable InDel variants	580	39,204
Number of SNPs with < 35 pb of available flanking sequence for probe design	1,818	984
Number of SNPs located in low complexity region	162	3,862
Number of SNPs failing genotype confidence score	3,217	0
Number of SNPs in sequences with similarity to other sequences	0	9,332
Number of SNPs with additional SNPs within a 30 bp window on both sides	3,916	236,723
Number of SNPs located up to 60 bp from an indel variant	49	75,374
Number of SNPs located in suspicious contigs according to TransRate evaluation	-	1,538
Total Number of retained SNPs	**4,508**	**39,810**

As expected, the genome complexity reduction generated by RAD-seq provided a much larger number of contigs on a per single quality read basis (176,629/60,252,356 = 0.29%) than the RNA-seq source (43,608/326,000,000 = 0.013%). On the other hand, RNA-seq was three times more efficient in providing raw sequence variants (309,509/326,000,000 = 0.095%) than RAD-seq (17,428/60,252,356 = 0.029%), despite the fact that a smaller number of individual trees were represented in the sequenced sample. This can partly be explained by the fact that the length and sequence coverage of each RAD sequence contig is shorter that RNA-seq such that the final SNPs count is lower, and only few eventually abide to the flanking sequence requirements of a SNP array. When all filters are considered, the final efficiency of RNA-seq in terms of assayable SNPs per quality raw sequence read was 1.6X higher than RAD-seq (0.012% versus 0.0075%). RNA-seq or exome-capture with RNA-derived probes has been the method of choice for SNP discovery and SNP array development in the very large and complex conifer genomes [38,39,42,43,75]. Our results confirm that RNA-seq is an efficient strategy for the identification of bona fide sequence variants. Additionally, our results also show that RNA-seq was far superior to RAD-seq in terms of coverting these putative variants to polymorphic SNPs, at least under the discovery and SNPs selection criteria used here.

Functional annotation of the SNP containing sequences

The full catalog of SNP variants retained following all quality filters contained 4,508 and 39,810 positions from RAD-Seq and RNA-Seq data respectively (Table 4). This total set of 44,318 assayable SNPs were located in 4,508 and 18,475 unique reference loci and contigs in the RAD-Seq and RNA-Seq *de novo* assemblies, respectively (S1 File). Therefore out of the 43,608 RNA-seq contig, only 18,475 (42%) could be actually sampled as far as identifying SNPs that passed all the quality and requirement filters. However, these 18,475 RNA-seq contigs provided more than one assayable SNP with an average of 2.15 SNPs and up to 19 SNPs in a single contig, thus providing a rich source of well curated SNPs for genotyping a reasonable portion of the *A. angustifolia* gene space. The average RNA-seq contig size was 1,729 bp, varying between 114 and 12,890 bp. This consolidated resource of 22,983 unique sequences was characterized for protein similarity, structural properties and gene ontology. Out of the 22,983 sequences, 15,144 (66%) had similarity at the used e-value threshold to database proteins. We also found that 9,316 sequences (41%) had

InterProScan hits, from which, 5,456 had corresponding GO terms. The gene set covers 1,945 identified domains across 3,231 InterPro protein families, the largest of which being the family P-loop containing nucleoside triphosphate hydrolase with 365 sequences (S1 Fig.). According to the similarity reported for the top hits in the BlastX searches against the nr database, we found that 10,889 sequences (47%) had significant similarity with average of 78% (e-value less than 1e-45) with 10,183 sequences with score greater than 60%. From a species perspective, we found that the highest proportion of top hits against the nr database matched to the gymnosperm *Picea sitchensis* (5,301; 34%), followed by matches to the basal angiosperm *Amborella trichopoda* (1,495; 10%). We assigned a total of 54,297 gene ontology terms to 13,351 (58%) of the putative transcripts of protein-coding genes. Most of the assignments (22,219; 41%) belonged to the 'Cellular Component' category, while the remaining belonged to the 'Biological Process' (19,472; 36%) and the 'Molecular Function' categories (12,606; 23%) (S2 Fig. and S1 File). In addition to the markers developed and validated, the newly established transcriptome for *A. araucaria* is similar to other published conifer transcriptomes in terms of size (54 Mbp), number of contigs (43,608 transcripts) and average transcript length (1,205 bp) [76]. Additionally, the score given by the TransRate to the assembly (score > 0.3) indicates further evidence of the assembly quality [53].

SNP genotyping performance on the Axiom array

From the full catalog of 44,318 putatively assayable SNPs, 3,400 SNPs (2,565 from RNA-seq and 835 from RAD-seq) were randomly selected at a rate of one SNP per unique contig to maximize transcriptome coverage, and in a few cases up to two SNPs for longer contigs (S1 File). Probes for these 3,400 SNPs were evaluated by the array manufacturer (Thermo Fisher) based on the company's *in silico* scoring system that uses a proprietary software that calculates a 'p-convert' value for each submitted SNP, i.e. the probability of a given SNP converting to a reliable SNP assay. Of the 3,400 SNP probes, 3,224 were classified as recommended, 170 as neutral and only six not recommended. In other words, 99.8% of the designed SNP probes successfully abided to the parameters of the Axiom array technology. The 3,400 probes were deliberately randomly selected from the entire 44,318 probes set such that the *in silico* evaluation would provide a bona fide estimate for the predicted success of all developed SNP probes. Assuming this same success rate of 99.8% for the entire SNP catalog, a larger SNP array could be designed to include at least ~40,000 SNPs covering all 22,983 transcriptome contigs.

Out of the 3,400 SNPs tested, using a global success based on a CR ≥ 90%, 3,038 SNPs (89.4%) were successfully genotyped (Table 5 and S2 File) with an average reproducibility rate of 99.95%, ultimately constituting the operational 3K SNP array. Using a global rate based on the Axiom criteria (the sum of PHR, MHR and NMH SNPs) 2,521 SNPs were successfully genotyped (74.1%). When considering a conversion rate (CR ≥ 90% and MAF≥ 0.005), 2,022 SNPs (59.5%) were converted and subsequently used in population genetic analyses. With the Axiom quality criteria, only the 1,650 PHR SNPs (48.5%) would be considered converted. RNA-seq was a significantly more efficient source of polymorphic SNPs (76.4%) compared to RAD-seq (7.4%). Due to the size, repetitive nature and lack of quality genome assembly with high continuity/contiguity, success rates of SNP genotyping for conifers in fixed content arrays have been generally slightly lower than those reported for other plant or animal species, varying between 60% and 85% [39,40,43]. Overall, our rates of 73.9% to 76.4% conversion for RNA-seq SNPs are in the similar range of those reported for other conifers SNP arrays derived from RNA-seq data. More importantly, if one applies the global and conversion rates obtained from RNA-seq data to all 39,810 RNA-seq derived SNPs in the catalog, we predict that this resource should provide ~29,000 polymorphic SNPs with CR ≥ 90% and MAF≥ 0.005, or ~25,000 PHR converted SNPs.

Table 5 - Summary of performance of SNPs derived from different sources (RAD-seq and RNA-seq) according to the two different performance criteria adopted (see Methods for details).

	RAD-seq	%	RNA-seq	%	Total	%
Total number of SNPs assayed	835		2,565		3,400	
Conventional performance criteria						
Global success rate - SNPs with Call Rate ≥ 90%	731	87.5	2,307	89.9	3,038	89.4
Conversion rate - SNPs with Call Rate ≥ 90% and MAF≥ 0.005	62	7.4	1,960	76.4	2,022	59.5
Axiom performance criteria						
Global success rate - Sum of PHR, MHR and NMH	626	75.0	1895	73.9	2521	74.1
PolyHighResolution (PHR)	4	0.5	1,646	64.2	1,650	48.5
Mono High Resolution (MHR)	607	72.7	207	8.0	814	23.9
No Minor Homozygous (NMH)	15	1.8	42	1.6	57	1.7
Call rate below threshold	13	1.6	165	6.4	178	5.2
OTV	28	3.4	5	0.2	33	1.0
Other	168	20.1	500	19.5	668	19.6

SNP polymorphism across populations

From the analysis of a sample of 185 individuals, the site frequency spectrum (SFS) of all successfully genotyped SNPs showed an enrichment toward higher frequency SNPs (S3 Fig.), possibly because of the criteria used to maximize probability of genotyping success. The rare SNP category, with MAF ≤ 0.005 shows that the array also contains a large number of rare SNPs, most of them coming from the RAD-seq data discovery. Although practically monomorphic in the sample analyzed, these variants could turn out to be polymorphic in larger sample sets. Just like the common practice of selecting the most polymorphic microsatellite markers leads to an ascertainment bias [77], SNP arrays also experience this trend caused by the SNP discovery and selection process. It has been shown, however, that multiple ways exist to correct ascertainment bias [78], and that typically it affects more profoundly analyses involving scans of selection signatures while population differentiation and diversity analysis are relatively robust to it [79]. Looking at the distribution of MAF for all 2,022 polymorphic SNPs in the samples analyzed (S3 File), a minimum of 1,415 in Population LAM and a maximum of 1,913 in population CHA are polymorphic, with an average of 1,703 per population. These results suggest that no relevant ascertainment bias toward any population exists in the array such that it should provide high-resolution power for detailed intra-population genetic analyses (e.g. mating system and kinship) across the natural range of the species.

Population genetic diversity: SNPs versus microsatellites

Allele frequencies, heterozygosities and inbreeding coefficient by marker in each population are provided as supplementary files for SNPs (S3 File) and microsatellites (S4 File). Individual population and overall heterozygosities, both observed and expected, were more than twice larger for microsatellites (0.644/0.708) when compared to SNPs (0.311/0.312), and the differences in value between observed and expected values for each population were much greater for microsatellites than SNPs, resulting in larger nominal estimates of inbreeding for microsatellites (Table 6). However, when the confidence intervals are taken into account, large nominal estimates of (*f*) obtained with microsatellite are not significantly different from zero for 10 out of 15 populations. On the other hand, estimates of inbreeding obtained with SNPs are considerably lower indicating very low nominal inbreeding in all populations, except CJO (f = 0.140) and CHA (f = 0.084), and not significantly different from zero in eight out of 15. Agreement between microsatellites and SNPs as far as the direction and significance of inbreeding was observed only for eight out of the 15 populations (BAR, IPI, CON, CAM, ITR, IRA, IRT and CHA). For the remaining, either disagreement as far as significance was observed (LAM, VAR, CJO, ITA and QBA), or the direction of the significant inbreeding was different (CAC and TRB). The much wider confidence intervals observed around the estimates of inbreeding with microsatellites obtained by bootstrapping over loci indicate the higher variability across loci and the much lower power of these markers to detect this occurrence. As expected, due to their larger number and the different intrinsic mutational behavior, SNPs, conversely, were able to detect significant levels of inbreeding both positive or negative at a much finer scale. Northern populations were on average 25% less genetically diverse (H_o = 0.262) than southern populations (H_o= 0.353) based on SNPs, but this pattern was not as clearly detected with the microsatellite dataset (Table 6).

Table 6 - Comparative summary of genetic diversity parameters (H_o observed heterozigosity; H_e expected heterozygosity) and inbreeding coefficient (F_{is}) with its respective 95% confidence interval (C.I) obtained with GDA for the two different data sets for the 15 *A. angustifolia* populations.

			Microsatellite data analysis (8 loci)						SNP data analysis (2,022 loci)					
#	**Region[a]**	**Population**	H_o	H_e	F_{is}[#]	**Lower 95% C.I.**	**Upper 95% C.I.**	**Sign.[#]**	H_o	H_e	F_{is}[#]	**Lower 95% C.I.**	**Upper 95% C.I.**	**Sign.[#]**
1	**N**	**BAR**	0.557	0.763	0.276	0.100	0.495	*	0.234	0.241	0.029	0.013	0.046	*
2	**N**	**IPI**	0.601	0.695	0.140	0.000	0.312	ns	0.272	0.275	0.011	-0.005	0.027	ns
3	**N**	**CON**	0.594	0.660	0.105	-0.128	0.307	ns	0.270	0.271	0.000	-0.018	0.020	ns
4	**N**	**LAM**	0.615	0.656	0.066	-0.026	0.162	ns	0.261	0.252	-0.037	-0.020	-0.056	*
5	**N**	**VAR**	0.509	0.585	0.135	-0.161	0.522	ns	0.276	0.261	-0.059	-0.044	-0.076	*
6	**N**	**CAM**	0.673	0.672	-0.001	-0.128	0.113	ns	0.257	0.254	-0.010	-0.029	0.007	ns
7	**N**	**CJO**	0.558	0.664	0.166	-0.001	0.399	ns	0.268	0.310	0.140	0.123	0.157	*
8	**S**	**ITA**	0.719	0.765	0.064	-0.068	0.181	ns	0.354	0.337	-0.054	-0.037	-0.071	*
9	**S**	**ITR**	0.767	0.752	-0.020	-0.122	0.082	ns	0.354	0.354	-0.001	-0.015	0.014	ns
10	**S**	**IRA**	0.700	0.742	0.060	-0.006	0.118	ns	0.355	0.353	-0.004	-0.019	0.011	ns
11	**S**	**IRT**	0.655	0.731	0.108	-0.090	0.296	ns	0.358	0.357	-0.004	-0.018	0.009	ns
12	**S**	**QBA**	0.756	0.653	-0.172	-0.103	-0.255	*	0.350	0.348	-0.006	-0.023	0.012	ns
13	**S**	**CAC**	0.695	0.765	0.095	0.016	0.181	*	0.361	0.345	-0.048	-0.063	-0.035	*
14	**S**	**CHA**	0.635	0.755	0.164	0.078	0.236	*	0.334	0.364	0.084	0.069	0.100	*

15	**S**	**TRB**	0.625	0.768	0.193	0.094	0.301	*	0.360	0.354	-0.017	-0.031	-0.002	*
		Overall	0.644	0.708	0.096	0.032	0.188	*	0.311	0.312	0.003	0.000	0.010	ns

[#] Estimates of within-population inbreeding based on Weir and Cockerham estimator (*f*); estimates contained in a 95% confidence interval containing zero are declared not significantly different (ns) from zero; otherwise inbreeding was declared significantly different from zero (*)

[a] Region: N = Northern population; S = Southern population

Overall populations, a relatively high and significant inbreeding (f = 0.096) was estimated with microsatellites while the SNPs estimate was very low and non-significant (f = 0.003). As *A. angustifolia* is outcrossed and dioecious [14] this high overall inbreeding estimated with microsatellites is definitely not expected suggesting a biologically misleading estimate. On the other hand, the trivial either positive or negative F_{is} values obtained with SNPs for all populations but CJO and CHA, or the overall non-significant estimate, much better fit the expectations. The relatively high inbreeding detected with SNPs in populations CJO and CHA, possibly resulting from mating among relatives is, most likely authentic and could be explained by their distinctive topographically isolated position from the larger northern and southern Araucaria formations therefore limiting potential gene flow.

Population differentiation: SNPs versus microsatellites

The AMOVA based partition of the genetic variation using the 2,022 SNPs indicated that ~32% of the variation is found between populations and 68% within populations, while the eight microsatellites indicated 11% between and 89% within populations (Table 7). In other words, estimates of F_{st} (F_{st} = 0.318) were almost three times higher than with eight microsatellites (F_{st} = 0.110). Additionally, SNPs data estimated that the majority of the differentiation is due to the between region differentiation (F_{stR} = 0.281) when compared to within region between populations ($F_{stP/R}$ = 0.052). Microsatellites on the other hand suggested the inverse with much closer estimates (F_{stR} = 0.049 and $F_{stP/R}$ = 0.065). Results indicate that microsatellites underestimated population differentiation and provided an incorrect view of the relative relevance of regional versus population differentiation. To discard the potential bias due to the much larger number of SNPs to detect population differentiation, we next generated estimates from 50 random sets of 80 SNPs to estimate population parameters. The number 80 came from a general indication that around four to 12 SNPs are expected to provide the equivalent power of a single microsatellite for population structure analyses [80]. Interestingly, essentially the same average estimates of heterozygosity, inbreeding (F_{is}) and partitions of population differentiation were obtained with the replicates of 80 SNPs as with all 2,022 SNPs (Table 7 and S5 File). Likewise, F_{is} was not significant while F_{st} was highly significant with different sets of SNPs. These results show that even small numbers of SNPs are able to correctly estimate heterozigosty, inbreeding and better detect and partition the genetic variation among populations.

Table 7 - Comparative summary of genetic variation parameters and F-statistics of population differentiation via AMOVA (Analysis of Molecular Variance) for *A. angustifolia* populations using different molecular marker sets together with previously published estimates for similarly regionally located populations.

Marker type	# Markers	# Populations	# Regions	H_o	H_e	F_{is}	F_{stR}	$F_{stP/R}$	F_{stT}	Ref.
Microsatellites	8	15	2	0.644	0.708	0.096	0.049	0.065	0.110	this study
SNP (50 random sets)*	80	15	2	0.316	0.317	0.003	0.283	0.050	0.319	this study
SNP (all)	2022	15	2	0.311	0.312	0.003	0.281	0.052	0.318	this study
Isozymes	15	9	1	0.073	0.084	0.148	-	-	0.044	[21]
Isozymes	7	13	3	0.132	0.128	-0.036	0.129	0.013	0.141	[19]
Microsatellites	5	6	2	0.580	0.710	0.110	-	-	0.112	[22]
AFLPs	166	6	2	-	0.300	-	-	-	0.144	[22]
Microsatellites	15	12	2	0.610	0.670	n.r.	-	-	$0.187^{\#}$	[31]
cpDNA	Sequence	39	3	-	-	-	0.521	0.183	$0.700^{\#}$	[81]

H_o = observed heterozygosity, H_e = expected heterozygosity, F_{is} = coefficient of inbreeding; F_{stR} = AMOVA based F_{st} between regions; $F_{stP/R}$ = AMOVA based F_{st} between populations within regions; F_{stT} = AMOVA based total F_{st} between populations and regions; *Average estimate of 50 random sets of 80 SNP markers; $^{\#}$ Estimate of Spatial AMOVA $F_{ct.}$; n.r. not reported

To illustrate further the power of the SNP array to detect population structure analysis we compared principal coordinate analysis (PCoA) plots based on matrix of genetic distances obtained with increasing numbers of SNPs or eight microsatellites. Consistent with the much higher estimates of F_{st} , using all 2,022 polymorphic SNPs the 15 populations split into three distinct genetic groups (Fig. 8). The axis reflecting the north-south latitudinal split is the major contributor explaining the majority of the variation (29.06%) while the additional separation of the northernmost population BAR contributes only an additional 2.07%. The microsatellite data, however, hardly detects the north-south signal explaining only 9.24% of the variation along the first PCoA axis. With just 80 randomly selected SNPs the north-south cline is clearly captured with 24.38% of the variation explained along the first axis, and population BAR already shows differentiation. With 800 SNPs, for example, the same result as using all 2,022 SNPs is already obtained (Fig. 8). In a further attempt to check the possibility of separating the populations within each region using all 2,022 SNPs, individuals in the northern regions showed a clustering pattern into their respective populations, with some overlap, while in the southern regions only individuals of population 8 more clearly split from all others (S4 Fig.). Our results are consistent with a review of a number of studies that compared SNPs with microsatellites showing that SNPs had greater accuracy for detecting and clustering groups of related individuals when three to several hundred fold larger numbers of SNPs are used than microsatellites[82].

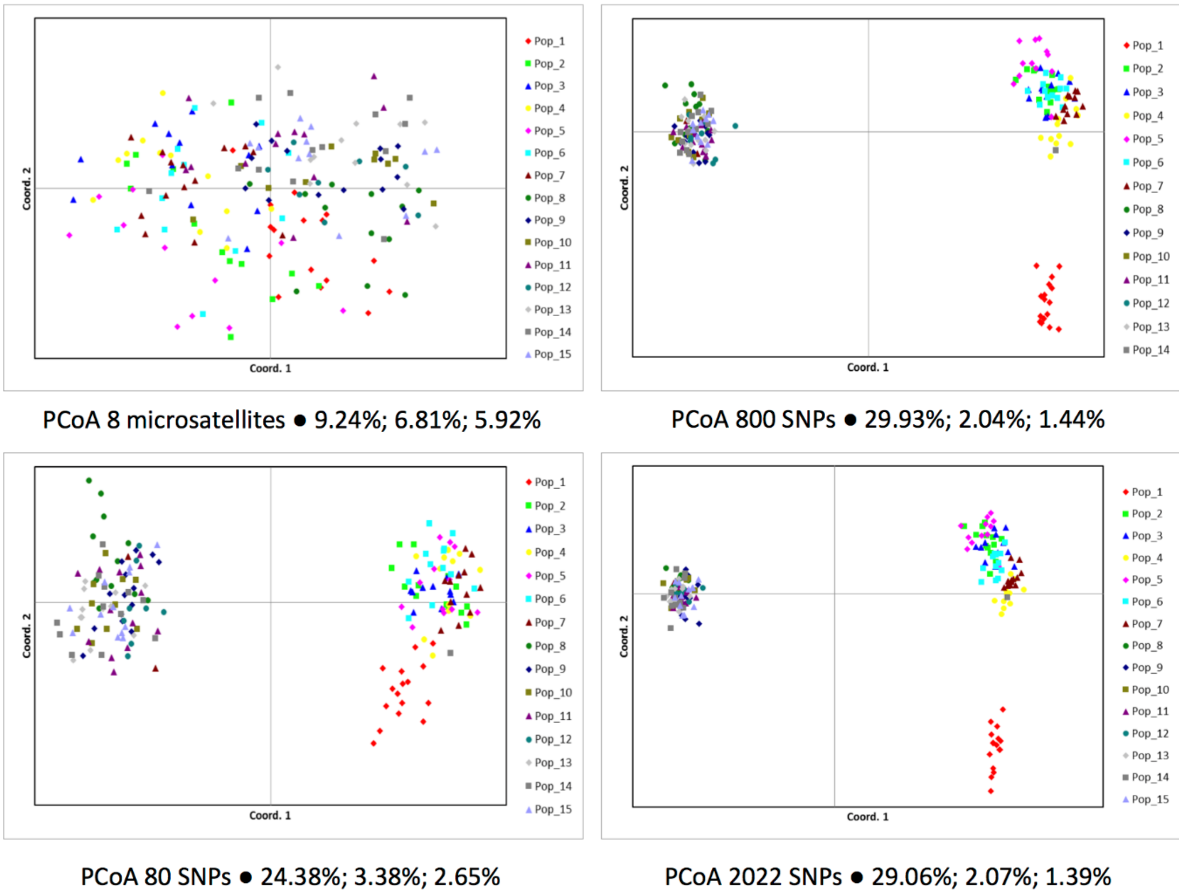

Figure 8 - Population structure analyses of the 15 *A. angustifolia* populations using a multidimensional Principal Coordinate analysis (PCoA) and different markers types (microsatellites or SNPs) and different numbers of SNPs. The proportions of variation explained by the first three PCoA axes are indicated from left to right respectively in each plot.

Finally, we carried out an exploratory comparison of our results with previously published *A. angustifolia* studies contemplating populations with a similarly wide sampling range, both across the northern and southern regions (Table 7). The geographic distribution of remnant Araucaria populations is relatively restricted and well determined in Brazil[32], therefore suggesting valid comparisons with previous reports. Regarding population heterozigosity and inbreeding, isozymes provided considerably lower estimates than SNPs and contrasting results regarding inbreeding [19,21]. Sets of five microsatellites resulted in equally high estimates of heterozygosity with high and significant inbreeding [22,25] equivalent to what we obtained with our eight microsatellites dataset. Dominant AFLP data, on the other hand resulted in estimates of expected heterozigosity similar to SNPs while inbreeding evidently could not be estimated. Regarding population differentiation, isozymes, microsatellites and AFLPs provided equally low estimates as our microsatellite dataset. However, a recent study with 15 microsatellites showed a slightly higher estimate based on spatial AMOVA (F_{ct} = 0.187) [31], suggesting that if a larger number of microsatellites is used the estimates might eventually converge to those obtained with SNPs. Chloroplast DNA, as expected, provided significantly higher estimates of population differentiation.

STRUCTURE analyses: SNPs versus microsatellites

From the STRUCTURE analyses and based on the Evanno's 'ΔK', both the SNP and the microsatellites datasets indicated k= 2 as the most likely number of clusters, corresponding to the northern and southern groups of populations (S5 Fig.). CLUMPAK generated plots of the consensus solution of the 10 independent runs are provided for all k´s tested for the different data sets (80 and 2,022 SNPs and 8 microsatellites) (S6 to S8 Files). The 2,022 SNP was only able to reveal the three separate clusters at k =3, while the microsatellites at k= 3 very clearly indicated population BAR corresponding to a third group, although at all k´s microsatellites indicated what seems an unexpected level of admixture (Fig. 9). At k> 2 the 80 SNPs set also showed an unexpected level of admixture but only in the southern populations and a clear separation of population BAR at k= 4. Both SNP datasets provided a unambiguous assignment of the individuals to their respective regions, highlighting, however, one tree in the southern population CHA that most likely was mislabeled at some point, while microsatellites did not clearly detect this individual and suggested an apparent level of admixture in both clusters (Fig. 9).

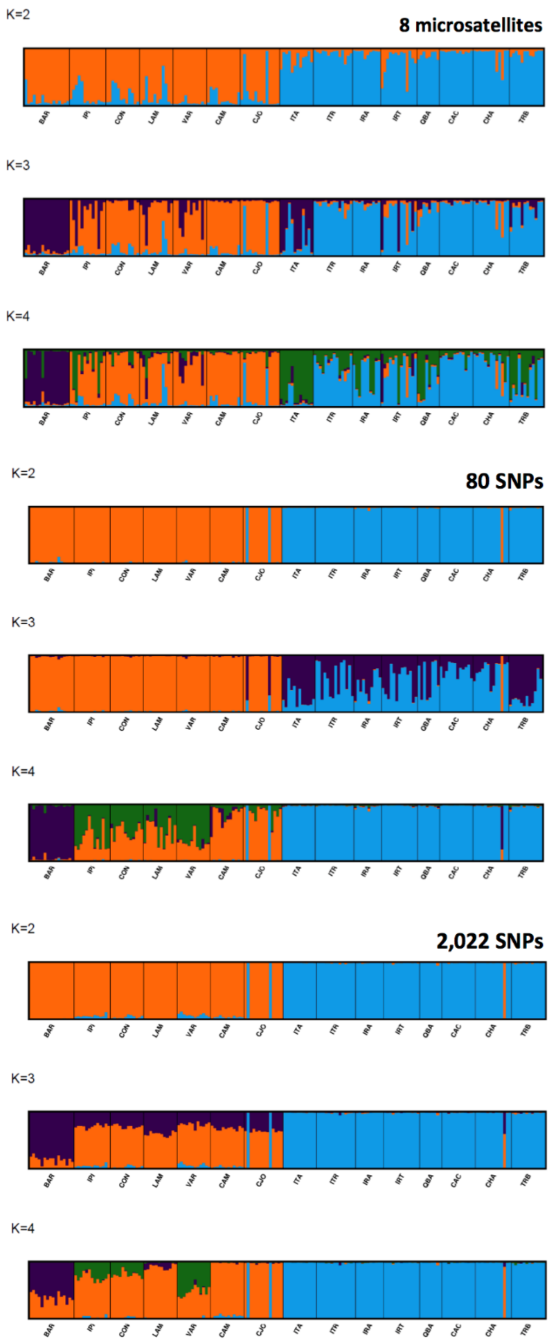

Figure 9 - Comparative population structure analyses for the 15 *A. angustifolia* populations indicated at the bottom of each panel obtained with the software STRUCTURE using a reduced (80) and full (2,022) set of SNPs and an eight-microsatellite set for different numbers of ‘k’ clusters (K=2 to 4).

A review of similar microsatellite vs. SNPs comparative studies reported a number of cases of discordancy between PCoA and the number of k clusters found using different marker types and numbers [82]. In the description of the ΔK method, Evanno et al.[72] emphasized that while the method identifies the correct number of clusters in many situations, it should not be used exclusively. Furthermore, the 'ΔK' metric was found by the same authors to be sensitive to the type and number of genetic markers used, and the number of populations and individuals typed in each sample. In our study, we have strong additional evidence based on F_{st} estimates and PcoA plots that three clusters exist such that this seems to be the most probable number of groups, also matching what a recent cpDNA phylogeographic study has shown [81]. Studies in humans have shown that random dinucleotide microsatellites are on average five to eight times more informative for structure and population assignment than random single-nucleotide polymorphisms (SNPs), but that a small proportion of carefully selected SNPs can be found with higher informativeness than the median for dinucleotides[80]. Considerable differences in MAF are seen between populations for several hundred SNPs indicating a good potential of selecting population specific SNP panels of ancestry informative markers (AIM) to track sample origin. Differences are seen between all pairs of populations within each region despite the overall low but significant population differentiation ($p<0.01$) as estimated by pair-wise F_{st} (S2 Table). While the overall average F_{st} between northern and southern populations is 0.198, between populations it is 0.057 and 0.037 within the Northern and Southern region respectively. It will be interesting in a follow up study to evaluate whether the selection of panels of AIM SNPs panels based on informative metrics of ancestry inference[80] will allow detecting additional structure and discriminating and assigning individuals to their respective populations. Panels of specific AIM SNPs on this same array could be useful, for example, to track illegal Araucaria logging by assigning timber samples to their geographic origin as shown for *Quercus* in Europe [83] or to check the origin of conserved seeds of unknown origin in germplasm banks.

SNP array data to advance Araucaria genetics, conservation and breeding

A potential criticism of our population survey could be the relatively small samples sizes used. However, we deliberately had the goal of assessing SNP performance in a country-wide sample of as many populations possible even if smaller samples sizes had to be used to demonstrate its utility irrespective of where future studies will be carried out. While larger sample sizes have been typically used in microsatellites surveys, with SNP genotyping this might not be strictly necessary, adding a further advantage to the use of large SNP sets. A number of studies have shown that large numbers of biallelic SNPs (> 1000) efficiently compensate for small sample sizes as small as n = 2 to 6 [84-87]. We are therefore confident that our results of population diversity and differentiation with 2,022 SNPs are robust and comparable with previous studies.

The SNP data gathered validate the ample applicability of the Axiom array and casts new light on the overall picture of diversity and structure of *A. angustifolia* along its extensive natural range. SNPs more accurately showed that southern populations (H_o=0.353) are significantly more genetically diverse that northern populations (H_o = 0.262) (Table 6), consistent with recent indications [31]. When compared to microsatellites, SNPs provided considerably lower estimates of heterozigosity and scant, if any, evidence of trivial inbreeding either positive or negative within-populations, and none at range wide level. While microsatellites have frequently delivered large nominal estimates of F_{is}, SNPs provide precise estimates that allow confidently detecting even minor levels of inbreeding if they in fact exist (Table 6). The major north-south genetic cline detected by SNPs (Figs. 2 and 3) is consistent with previous reports based on isozymes [19] and microsatellites [22]. Nevertheless, SNPs discriminated these two groups with a considerably higher magnitude (Table 7), together with the identification of a third PCoA cluster (Fig. 8) and further confirmed by a STRUCTURE analysis, robustly assigning individuals to their respective regions (Fig. 9), and showing a considerably higher population differentiation due to regional than population within-region difference (Table 7). This genetic cline has been explained by the combined effect of post-glacial migration from different refugia [22,81] and a north-south isolation most likely due to niche suitability [31]. The possibility of an additional phylogeographic separation of a third genetic group in the northern populations was also suggested earlier based on tenuous evidence from microsatellites [22] and recently confirmed based on sequence data of three intragenic regions in the chloroplast genome (cpDNA) [81]. Interestingly, the 2,022 autosomal SNP data set was able to provide equally strong evidence as the chloroplast sequence-based data for a third group, therefore matching the stronger phylogenetic signal typically obtained from the uniparental non-recombining inheritance of cpDNA [88].

The somewhat conflicting results between SNPs and microsatellite based population estimates are not surprising and should not be taken as criticisms to previous studies in *Araucaria angustifolia*. The four to six orders of magnitude higher mutation rate of microsatellites when compared to SNPs [35], and the consequential allele hypervariability will tend to result in higher heterozygosities. Moreover, the twice higher estimates of genetic diversity in *A. angustifolia* obtained with microsatellites when compared to SNPs (Table 7) may also be a consequence of a strong ascertainment bias when selecting the most polymorphic markers [77]. Combined with the small number of markers that sample a narrow fraction of the genome, biased estimates of genetic diversity may result [33]. A microsatellite-SNP comparative analysis in the outcrossed *Arabidopsis halleri* for example, showed the same pattern of twice to three times higher estimates of heterozigosity with microsatellites when compared to SNPs. More importantly, however, was the fact that while the microsatellite heterozigosity showed no correlation at all with the canonical Watterson theta (θ) genetic diversity metric, the SNP-based heterozigosity was highly correlated, properly reflecting the genome-wide genetic diversity [89]. Concerning the considerably higher estimates of F_{st} obtained with SNPs when compared to microsatellites, the recurrent mutation of the latter frequently leads to homoplasic alleles which are identical by state (size) but not identical by descent [90]. This fact tends to dampen the signal of population structure that SNPs correctly capture, even when only small random subsets of only 80 out of the 2,022 SNPs were used (Fig. 8 and S5 File). A number of previous studies in plant and animal species have in fact shown higher F_{st} estimates with SNPs when compared to microsatellites and a superior ability of SNPs to resolve fine-scale population structure and phylogeographic signals [87,91,92].

Conclusions

To summarize, we have developed the first comprehensive SNP resource for A*raucaria angustifolia*, a keystone subtropical conifer tree, critically endangered due to its valuable wood and seeds. From the transcriptome-wide catalog of 44,318 annotated SNP an Axiom® SNP array with ~3,000 validated SNPs was developed as part of a multi-species SNP array strategy significantly reducing the individual sample genotyping cost therefore allowing access to a high quality SNP array even for this generally underfunded species. Data obtained with this newly developed SNP genotyping platform provided a comprehensive look at the range-wide genetic diversity and structure of the species. By matching SNP with microsatellite data on the exact same individuals, our results indicate that microsatellite markers may have led to estimates of genetic diversity and differentiation that have not precisely reflected the actual genome-wide patterns of variation and structuring. The generally established microsatellite-based inference that *A. angustifolia* has been resilient against rapid losses of its genetic diversity due to forest fragmentation might not be fully warranted and should receive further attention [30,32]. Overestimated diversity, inaccurate inbreeding estimates and underestimated population differentiation can have relevant consequences on decisions on how to approach and manage *ex situ* and *in situ* conservation of the species' genetic resources. Our results do not doubt the usefulness of microsatellites in general as efficient tool for studying mating systems, kinship and relatedness at low taxonomic levels, but we caution on taking their diversity, inbreeding and differentiation estimates at face value given their known limitations. Because genetics applications rely on multilocus estimators of differentiation, panels of several hundred to thousand genome-wide SNPs will always be more powerful, representative and accurate than a dozen microsatellites [1]. It is relevant to note, however, that even SNPs when genotyped using methods based on sequencing reduced genomic representations, face significant challenges to provide consistent and interchangeable SNP genotypes, limiting, for example, valid across-study comparisons and meta-analyses due to the stochastic genome sampling and the sources of sequencing bias involved in these techniques [93].

The availability of a public, user-friendly 3K SNP array for *A. angustifolia* and the catalog of 44,318 SNPs predicted to provide ~29,000 informative SNPs across ~20,000 loci across the genome, will allow tackling still unsettled questions on its evolutionary history, toward a more comprehensive picture of the origin, past dynamics and future trend of the species' genetic resources. Additionally, but not less importantly, the SNP array described, unlocks the potential to consider adopting genomic prediction methods [94] to accelerate the still very timid efforts of systematic tree breeding of *A. angustifolia,* a species with enormous potential for its valuable wood and seed products but with very long generation times. In conclusion, this first fully public fixed content SNP array for *A. angustifolia* and the additional extensive SNP catalog provided in this work for the future manufacture of even denser arrays, raises this iconic species to a higher level for genetic research, opening opportunities to increase the breadth, precision, long-term portability and impact of the genetic data generated.

Acknowledgments

This work was mainly supported by a competitive grant PRONEX FAP-DF/CNPq "NEXTREE" 193.000.570/2009 to DG and additional funding from EMBRAPA project 02.11.08.005.00.03. A productivity grant awarded to DG by CNPq was also used to partially fund this work. We would like to thank Ananias de Almeida S. Pontinha, Miguel L. Menezes Freitas and the field staff of the Instituto Florestal de São Paulo (Itapeva experimental station) for technical and logistic support for sample collection.

References

1. Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, et al. (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources 11: 1-8.
2. Morin PA, Luikart G, Wayne RK, Grp SW (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19: 208-216.
3. Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.
4. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3: e3376.
5. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GbS) approach for high diversity species. Plos One 6: e19379.
6. Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. Molecular Ecology 25: 185-202.
7. Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology 22: 2841-2847.
8. Silva-Junior OB, Grattapaglia D, Novaes E, Collevatti RG (2018) Design and evaluation of a sequence capture system for genome-wide SNP genotyping in highly heterozygous plant genomes: a case study with a keystone Neotropical hardwood tree genome. DNA Research 25: 535-545.
9. Myles S (2013) Improving fruit and wine: what does genomics have to offer? Trends in Genetics 29: 190-196.
10. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, et al. (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 22: 3165-3178.
11. Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, et al. (2019) A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources. Frontiers in Plant Science 10.

12. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, et al. (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources 11: 123-136.
13. Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing 240 *Eucalyptus* tree genomes across 12 species. New Phytologist 206: 1527-1540.
14. Klein RM (1960) O aspecto dinâmico do pinheiro brasileiro. Sellowia 12: 17-44.
15. Thomas P (2013) *Araucaria angustifolia*. The IUCN Red List of Threatened Species 2013: e.T32975A2829141. doi: 102305/IUCNUK2013-1RLTST32975A2829141en Accessed 10 April 2019.
16. Guerra MP, Silveira V, Reis MS, Schneider L (2002) Exploração, manejo e conservação da araucária (*Araucaria angustifolia*). In: L.L. S, Lino CF, editors. Sustentável Mata Atlântica: a exploraçao de seus recursos florestais. São Paulo: Editora SENAC. pp. 85-102.
17. Shimizu JY, Jaeger P, Sopchaki SA (2000) Genetic variability in a remnant population of Araucaria in the Iguaçu National Park, Brazil. Boletim de Pesquisa Florestal 41: 18-36.
18. Medri C, Ruas PM, Higa AR, Murakami M, Ruas CD (2003) Effects of forest management on the genetic diversity in a population of Araucaria angustifolia (bert.) O. Kuntze. Silvae Genetica 52: 202-205.
19. Sousa VA, Robinson IP, Hattemer HH (2004) Variation and population structure at enzyme gene loci in Araucaria angustifolia (Bert.) O. Ktze. Silvae Genetica 53: 12-19.
20. Mantovani A, Morellato LPC, dos Reis MS (2006) Internal genetic structure and outcrossing rate in a natural population of Araucaria angustifolia (Bert.) O. Kuntze. Journal of Heredity 97: 466-472.
21. Auler NMF, Reis MSd, Guerra MP, Nodari RO (2002) The genetics and conservation of Araucaria angustifolia: I. Genetic structure and diversity of natural populations by means of non-adaptive variation in the state of Santa Catarina, Brazil. Genetics and Molecular Biology 25: 329-338.
22. Stefenon VM, Gailing O, Finkeldey R (2007) Genetic structure of Araucaria angustifolia (Araucariaceae) populations in Brazil: Implications for the in situ conservation of genetic resources. Plant Biology 9: 516-525.

23. de Souza MIF, Salgueiro F, Carnavale-Bottino M, Félix DB, Alves-Ferreira M, et al. (2009) Patterns of genetic diversity in southern and southeastern Araucaria angustifolia (Bert.) O. Kuntze relict populations. Genetics and Molecular Biology 32: 546-556.
24. Inza MV, Aguirre NC, Torales SL, Pahr NM, Fassola HE, et al. (2018) Genetic variability of Araucaria angustifolia in the Argentinean Parana Forest and implications for management and conservation. Trees-Structure and Function 32: 1135-1146.
25. Stefenon VM, Gailing O, Finkeldey R (2008) Genetic structure of plantations and the conservation of genetic resources of Brazilian pine (Araucaria angustifolia). Forest Ecology and Management 255: 2718-2725.
26. Sant'Anna CS, Sebbenn AM, Klabunde GHF, Bittencourt R, Nodari RO, et al. (2013) Realized pollen and seed dispersal within a continuous population of the dioecious coniferous Brazilian pine Araucaria angustifolia (Bertol.) Kuntze. Conservation Genetics 14: 601-613.
27. Bittencourt JVM, Sebbenn AM (2009) Genetic effects of forest fragmentation in high-density Araucaria angustifolia populations in Southern Brazil. Tree Genetics & Genomes 5: 573-582.
28. Bittencourt JVM, Sebbenn AM (2008) Pollen movement within a continuous forest of wind-pollinated Araucaria angustifolia, inferred from paternity and TwoGENER analysis. Conservation Genetics 9: 855-868.
29. Bittencourt JVM, Sebbenn AM (2007) Patterns of pollen and seed dispersal in a small, fragmented population of the wind-pollinated tree Araucaria angustifolia in southern Brazil. Heredity 99: 580-591.
30. Medina-Macedo L, Sebbenn AM, Lacerda AEB, Ribeiro JZ, Soccol CR, et al. (2015) High levels of genetic diversity through pollen flow of the coniferous Araucaria angustifolia: a landscape level study in Southern Brazil. Tree Genetics & Genomes 11.
31. de Sousa VA, Reeves PA, Reilley A, de Aguiar AV, Stefenon VM, et al. (2020) Genetic diversity and biogeographic determinants of population structure in Araucaria angustifolia (Bert.) O. Ktze. Conservation Genetics 10.1007/s10592-019-01242-9.

32. Stefenon VM, Steiner N, Guerra MP, Nodari RO (2009) Integrating approaches towards the conservation of forest genetic resources: a case study of Araucaria angustifolia. Biodiversity and Conservation 18: 2433-2448.
33. Väli Ü, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? Molecular Ecology 17: 3808-3817.
34. Hedrick PW (2001) Conservation genetics: where are we now? Trends in Ecology & Evolution 16: 629-636.
35. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5: 435-445.
36. Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9: 1537-1544.
37. Heer K, Ullrich KK, Liepelt S, Rensing SA, Zhou J, et al. (2016) Detection of SNPs based on transcriptome sequencing in Norway spruce (Picea abies (L.) Karst). Conservation Genetics Resources 8: 105-107.
38. Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, et al. (2016) Convergent local adaptation to climate in distantly related conifers. Science 353: 1431.
39. Pavy N, Gagnon F, Rigault P, Blais S, Deschenes A, et al. (2013) Development of high-density SNP genotyping arrays for white spruce (Picea glauca) and transferability to subtropical and nordic congeners. Molecular Ecology Resources 13: 324-336.
40. Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, et al. (2009) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (Pinus taeda L.). Tree Genetics & Genomes 5: 225-234.
41. Liu J-J, Sniezko RA, Sturrock RN, Chen H (2014) Western white pine SNP discovery and high-throughput genotyping for breeding and conservation applications. BMC Plant Biology 14: 380.
42. Plomion C, Bartholome J, Lesur I, Boury C, Rodriguez-Quilon I, et al. (2016) High-density SNP assay development for genetic analysis in maritime pine (Pinus pinaster). Molecular Ecology Resources 16: 574-587.

43. Pinosio S, Gonzalez-Martinez SC, Bagnoli F, Cattonaro F, Grivet D, et al. (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, Pinus halepensis Mill. Molecular Ecology Resources 14: 846-856.

44. Su Y, Hu DH, Zheng HQ (2016) Detection of SNPs based on DNA specific-locus amplified fragment sequencing in Chinese fir (*Cunninghamia lanceolata* (Lamb.) Hook). Dendrobiology 76: 73-79.

45. Moriguchi Y, Uchiyama K, Ueno S, Ujino-Ihara T, Matsumoto A, et al. (2016) A high-density linkage map with 2560 markers and its application for the localization of the male-sterile genes ms3 and ms4 in Cryptomeria japonica D. Don. Tree Genetics & Genomes 12: 57.

46. Mishima K, Hirao T, Tsubomura M, Tamura M, Kurita M, et al. (2018) Identification of novel putative causative genes and genetic marker for male sterility in Japanese cedar (Cryptomeria japonica D.Don). Bmc Genomics 19.

47. Inglis PW, Pappas MdCR, Resende LV, Grattapaglia D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLosOne 13: e0206085.

48. Elbl P, Lira BS, Andrade SCS, Jo L, dos Santos ALW, et al. (2015) Comparative transcriptome analysis of early somatic embryo formation and seed development in Brazilian pine, Araucaria angustifolia (Bertol.) Kuntze. Plant Cell, Tissue and Organ Culture (PCTOC) 120: 903-915.

49. Zonneveld BJM (2012) Genome sizes of all 19 Araucaria species are correlated with their geographical distribution. Plant Systematics and Evolution 298: 1249-1255.

50. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33: 290.

51. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, et al. (2013) The MaSuRCA genome assembler. Bioinformatics 29: 2669-2677.

52. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, et al. (2004) Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Research 14: 1147-1159.

53. Smith-Unna R, Boursnell C, Patro R, Hibberd J, Kelly S (2016) TransRate: reference free quality assessment of de novo transcriptome assemblies. Genome Research 26: 1134-1144.

54. Senn H, Ogden R, Cezard T, Gharbi K, Iqbal Z, et al. (2013) Reference-free SNP discovery for the Eurasian beaver from restriction site-associated DNA paired-end data. Mol Ecol 22: 3141-3150.

55. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature genetics 44: 226-232.

56. Hercus C (2009) Novocraft short read alignment package. Website http://www.novocraft.com.

57. BroadInstitute (2017) Picard Tools. http://broadinstitute.github.io/picard/: Broad Institute.

58. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303.

59. Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843-2851.

60. Gouzy J, Carrere S, Schiex T (2009) FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25: 670-671.

61. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

62. Grattapaglia D, Silva-Junior OB, Resende LV, Silva PIT (2017) A five-species 50K Axiom SNP microarray allows high quality genotyping of Coffee, Cashew, Cassava, Brazilian Pine and Eucalyptus. Plant & Animal Genome XXIV. San Diego: https://pag.confex.com/pag/xxv/meetingapp.cgi/Paper/26564. pp. 26564.

63. ThermoFisher (2017) Axiom Analysis Suite 3.1 - User Manual. Carlsbad, CA.

64. Schmidt AB, Ciampi AY, Guerra MP, Nodari RO (2007) Isolation and characterization of microsatellite markers for Araucaria angustifolia (Araucariaceae). Molecular Ecology Notes 7: 340-342.

65. Faria DA, Mamani EMC, Pappas MR, Pappas GJ, Grattapaglia D (2010) A Selected Set of EST-Derived Microsatellites, Polymorphic and Transferable across 6 Species of Eucalyptus. Journal of Heredity 101: 512-520.
66. Brondani RP, Grattapaglia D (2001) Cost-effective method to synthesize a fluorescent internal DNA standard for automated fragment sizing. Biotechniques 31: 793-795, 798, 800.
67. Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38: 1358-1370.
68. Lewis P, Zaykin. D (2001) Genetic data analysis: computer program for the analysis of allelic data (software). https://phylogenyuconnedu/software/.
69. Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28: 2537-2539.
70. Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945-959.
71. Besnier F, Glover KA (2013) ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers. PLOS ONE 8: e70651.
72. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14: 2611-2620.
73. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.
74. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular ecology resources 15: 1179-1191.
75. Telfer E, Graham N, Macdonald L, Li Y, Klápště J, et al. (2019) A high-density exome capture genotype-by-sequencing panel for forestry breeding in Pinus radiata. PLOS ONE 14: e0222640.
76. Baker EAG, Wegrzyn JL, Sezen UU, Falk T, Maloney PE, et al. (2018) Comparative Transcriptomics Among Four White Pine Species. G3: Genes|Genomes|Genetics 8: 1461.

77. Queirós J, Godinho R, Lopes S, Gortazar C, de la Fuente J, et al. (2015) Effect of microsatellite selection on individual and population genetic inferences: an empirical study using cross-specific and species-specific amplifications. Molecular Ecology Resources 15: 747-760.
78. Lachance J, Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays 35: 780-786.
79. Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity 106: 158-171.
80. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of Genetic Markers for Inference of Ancestry*. The American Journal of Human Genetics 73: 1402-1422.
81. Stefenon VM, Klabunde G, Lemos RPM, Rogalski M, Nodari RO (2019) Phylogeography of plastid DNA sequences suggests post-glacial southward demographic expansion and the existence of several glacial refugia for Araucaria angustifolia. Scientific Reports 9: 2752.
82. Puckett EE (2017) Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. Conservation Genetics Resources 9: 289-304.
83. Céline B-J, Liesebach M (2015) Tracing the origin and species identity of Quercus robur and Quercus petraea in Europe: a review. Silvae Genetica 64: 182-193.
84. Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. PLOS ONE 7: e42649.
85. Flesch EP, Rotella JJ, Thomson JM, Graves TA, Garrott RA (2018) Evaluating sample size to estimate genetic management metrics in the genomics era. Molecular Ecology Resources 18: 1077-1091.
86. Nazareno AG, Bemmels JB, Dick CW, Lohmann LG (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. Molecular Ecology Resources 17: 1136-1147.
87. Jeffries DL, Copp GH, Lawson Handley L, Olsén KH, Sayer CD, et al. (2016) Comparing RADseq and microsatellites to infer complex phylogeographic

patterns, an empirical perspective in the Crucian carp, Carassius carassius, L. Molecular Ecology 25: 2997-3018.

88. Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS (1997) Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. Plant Systematics and Evolution 206: 353-373.

89. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, et al. (2017) Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. BMC Genomics 18: 69.

90. Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research 22: 201-204.

91. Vendrami DLJ, Telesca L, Weigand H, Weiss M, Fawcett K, et al. (2017) RAD sequencing resolves fine-scale population structure in a benthic invertebrate: implications for understanding phenotypic plasticity. Royal Society Open Science 4: 16.

92. Hodel RGJ, Chen S, Payton AC, McDaniel SF, Soltis P, et al. (2017) Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. Scientific Reports 7: 17598.

93. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, et al. (2013) Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology 22: 3151-3164.

94. Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BSF, et al. (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. Frontiers in Plant Science 9: 1693.

Supporting information

S1 Table. Description of the genotyped samples from different populations of *Araucaria angustifolia* used to validate the 3K SNP Axiom® Array.

S2 Table. Heat map of pairwise F_{st} estimates based on 2,022 polymorphic SNPs among all 15 populations indicating the higher differentiation between the populations in the northern (1 to 7) and southern (8 to 15) regions, and lower differentiation between populations within regions. Populations 7 and 8 in the transition zone display slightly differentiated F_{st} estimates from their regionally associated populations as indicated by the heatmap. All F_{st} estimates were significant (p<0.001) based on a permutation test by bootstrapping over loci.

S1 Fig. Top 20 most abundant InterPro domains and families identified in the *A. angustifolia* non-redundant gene set**.** (A) InterPro domain annotation was performed on the non-redundant gene set obtained from the transcriptome assembly of *A. angustifolia*. (B) InterPro family annotation was performed on the non-redundant gene set obtained from the transcriptome assembly of *A. araucaria*.

S2 Fig. Distribution of most abundant gene ontology (GO) terms in the three GO categories assigned to the *Araucaria angustifolia* contigs. Only level 3 terms are represented.

S3 Fig. SNP frequency spectrum of all 3,038 successfully genotyped SNPs by the two performance evaluation criteria. Also plotted are only the SNPs derived from RNA-seq data to show that the majority of monomorphic SNPs came from RAD-seq data.

S4 Fig. Population structure analyses using a multidimensional Principal Coordinate analysis (PCoA) with the full set of 2,022 SNPs. Top panel: PCoA plot involving only the seven northern region populations; Bottom panel: PCoA plot involving only the eight southern region populations. The proportions of variation explained by the first three PCoA axes are indicated from left to right respectively in each plot.

S5 Fig. Results of Evanno´s Delta K analysis to define the most probable number of populations with different sets of markers as indicated in the figure.

S1 File. Catalog of all 44,318 high-quality SNPs detected from the RNA-seq and RAD-seq data with their corresponding probes designed and annotated contig.

S2 File. List of the 3,400 SNPs on the Axiom array with their corresponding SNP probe and performance data (Call Rate, MAF, Axiom performance criteria).

S3 File. Data of the 2,022 SNPs used in the analyses with estimates of allele frequencies, observed (Ho) and expected (He) heterozygosity and test for Hardy Weinberg Equilibrium.

S4 File. Data of the 8 microsatellites used in the analyses and estimates of allele frequencies, observed (Ho) and expected (He) heterozygosity and test for Hardy Weinberg Equilibrium.

S5 File. Estimates of heterozygosity (H_o e H_e) within-population inbreeding (Fi_s) , fixation index (F_{st}) and total reduction of heterozigosity (F_{it}) obtained with the 50 replicates of 80 randomly selected SNPs.

S6 File. CLUMPAK generated plots of the consensus solution of the 10 independent runs for all 15 k´s tested using 8 microsatellites.

S7 File. CLUMPAK generated plots of the consensus solution of the 10 independent runs for all 15 k´s tested using 80 SNPs.

S8 File. CLUMPAK generated plots of the consensus solution of the 10 independent runs for all 15 k´s tested using 2,022 SNPs.

Data archiving

1. All RAD-seq raw sequencing data have been deposited in the NCBI SRA (Short Read Archive) under BioProject, PRJNA602322 at https://www.ncbi.nlm.nih.gov/bioproject/602322
2. Araucaria_snps.RNA-seq.vcf: vcf file containing all SNPs discovered using RNA-seq sequences including those that did not pass the quality filters are available at 10.6084/m9.figshare.11861712
3. Araucaria_snps.RAD-seq.vcf: vcf file containing all SNPs discovered using RAD-seq sequences including those that did not pass the quality filters are available at 10.6084/m9.figshare.11861682
4. Araucaria_RNA-seq_contigs.fasta: fasta archive of contigs from RNA-seq are available at 10.6084/m9.figshare.11861754
5. Araucaria_RAD-seq_contigs.fasta: fasta archive of contigs from RAD-seq are available at 10.6084/m9.figshare.11861718

Capítulo 3

Genomic prediction using a 35-year old *Araucaria angustifolia* trial addressing additive, dominant and epistatic effects performing selection within and between provenances

Abstract

A. angustifolia is a South American conifer, predominantly found in the southern and southeastern states of Brazil and parts of Argentina and Paraguay. Given the characteristics of its wood, such as the long-length fiber for specific cellulosic uses and the vast food use of its chestnut, the species is promising for exploration, confronting exotic conifers such as *Pinus sp*., in subtropical regions. This work aimed to propose a methodology based on genomic selection (GS) to accelerate the genetic improvement of this native conifer. We used 15 provenances from four Brazilian states (zone of natural occurrence of the species), accounting for 2,158 trees and 122 open-pollinated families. A panel with 1,710 SNPs combined with 26 SSRs was used, 857 plants were genotyped with both marker platforms. In 35 years of measuring growth, components of additive, dominant and epistatic variance were observed, the first being the most prevalent in volumetric expression throughout the growth period of the trees. Using only additive effects, abandoning non-additives, is the best for GS. In addition, phenotypic data from all 15 provenances should be included, aiming at better GS adjustments. You can train models as early as 12 years of age, aiming at the indirect selection of individuals at 35 years of age. Given the very late growth of the plants and the good quality of the models, GS proved to be very competitive in comparison to the phenotypic selection, opening new paths for the improvement of this iconic Brazilian conifer.

Keywords: Mixed models; individual growth; SNP; SSR; response to selection; non-additive effects.

Introduction

Paraná pine [*Araucaria angustifolia* (Bertol.) Kuntze] is an iconic Brazilian conifer that mainly occurs in South of Brazil and parts of Argentina and Paraguay. The reasons for exploitation that goes from pulp for paper and cardboard manufacturing given its longer fibers than most of the conifers, high-quality timber for construction and veneer to the tasty nut-like seed that is an important source of calories for local fauna and also humans (OLIVEIRA; PÁDUA; ZUCCHI; VENCOVSKY *et al.*, 2006). Its growth for production purposes is still too slow when compared to other conifer species explored in Brazil. Most of the commercial Paraná pine commonly grows in pure stands and is harvested at a rotation age of approximately 30 years with a mean annual increment that varies from 10 to 25 m^3 ha^{-1} yr^{-1} (NUTTO; SPATHELF; ROGERS, 2005). This can be highly competitive when we compared with the productivity of exotic conifers such as *Pinus sp*. in Brazil which grows around 30 m^3 ha^{-1} yr^{-1} with a rotation at 20 years old given that investments in silvicultural technologies, has tripled productivity in the USA in the last 50 years (ZHAO; KANE; TESKEY; FOX *et al.*, 2016). In this context, it is possible to assume that *Araucaria* is still underestimated and underused due to its close to zero rate of improvement.

The breeding conifers follows very similar strategies to any other forest tree, except for its selection cycle that in general is four to six times longer and more complex cloning compared to hardwoods. The reduction of cycle length is vital for the successful implementation of an *A. angustifolia* breeding program and when traditional pedigree-based selection method is compared to GS, GS hastens the breeding cycle and boosts the genetic gain per unit time by shortening generation intervals and omitting a series of progeny tests. Several studies have produced promising results in regard to the acceleration of the breeding cycle using GS (CHEN; BAISON; PAN; KARLSSON *et al.*, 2018; RESENDE; RESENDE; SANSALONI; PETROLI *et al.*, 2012; RESENDE; MUÑOZ; ACOSTA; PETER *et al.*, 2012; RESENDE; MUÑOZ; RESENDE; GARRICK *et al.*, 2012; TAN; GRATTAPAGLIA; MARTINS; FERREIRA *et al.*, 2017; THISTLETHWAITE; RATCLIFFE; KLÁPŠTĚ; PORTH *et al.*, 2017).

GS combines phenotypic and genetic information of a training population to develop a prediction model that produce Genomic Estimated Breeding Values

(GEBV) for candidate selection, requiring only their genotypic information (MEUWISSEN; HAYES; GODDARD, 2001). Another advantage of using GS is regarding the use of the realized relationship matrix ($\boldsymbol{G}$) (VANRADEN, 2008) that offers increased accuracy when considering the genetic variance of the breeding population and can assist in the management of genetic diversity and within family selection (HEFFNER; SORRELLS; JANNINK, 2009). The genetic relationships described by $\boldsymbol{A}$ are based on expected values, however, relatedness within families can deviate from this expectation due to the Mendelian segregation of alleles, which is better captured using $\boldsymbol{G}$ (HEFFNER; SORRELLS; JANNINK, 2009).

GS prediction models are influenced by several factors including the extent of LD between genetic markers and QTL, heritability of the target trait, Ne, number of genetic markers, and training population size, being the last three factors the ones that can be efficiently regulated by the breeder (GRATTAPAGLIA, 2014). When you consider a breeding program for a species that has never been improved before, such as the Paraná pine, it is important to know patterns of the behavior of genetic variance (additive effects, due to dominance and epistatic), useful not only to calibrate models of GS throughout future cycles, but to direct phenotypic selection strategies throughout breeding cycles when thinking about selection gains in complex productivity traits. Phenotypically speaking, studies evaluating additive and non-additive effects have also been done in conifers (ISIK; KLEINSCHMIT; STEINER, 2010; LI; WU, 2005). Additionally, effects of dominance and epistasis via molecular markers have been tested in GS procedures in *Eucalyptus sp*. and *Pinus taeda*, however studies indicate that, more commonly, only the additive effects and eventually those of dominance stand out (DE ALMEIDA FILHO; GUIMARÃES; E SILVA; DE RESENDE *et al.*, 2016; TAN; GRATTAPAGLIA; WU; INGVARSSON, 2018). There are no reports in the literature of using genomics to assist *A. angustifolia* breeding efforts.

In conifer breeding, linkage disequilibrium (LD) is rarefied over the generation cycles (IWATA; JANNINK, 2011), so using a satisfactory amount of markers to better capture the population's LD is necessary for the success of GS. A recent study in maritime pine, using 4,436 SNP, indicates that the GS model adjusted in two early generations can successfully predict the performance in the third generation using additive effects (BARTHOLOMÉ; VAN HEERWAARDEN; ISIK; BOURY *et al.*, 2016).

This work aimed to evaluate the behavior of the effects of additivity, dominance and epistasis by using codominant markers - SNP and SSR - over 35 years of volumetric growth of *A. angustifolia* in provenances originated from highly representative environments of the species in Brazil. Based on the profile of gene expression, we adjusted the GS models to predict the future behavior of new families and thus be able to infer about early selection in the species aiming at genetic gains in an unprecedented program for the improvement of this iconic Brazilian conifer.

Materials and Methods

1. Plant Material

The field experiment used is fully described in SEBBENN; PONTINHA; GIANNOTTI e KAGEYAMA (2003). In brief, seeds from open-pollinated families were collected from trees distributed in 15 natural *A. angustifolia* provenances from four Brazilian States - Minas Gerais (MG) São Paulo (SP), Paraná (PR) e Santa Catarina (SC). All provenances were identified during collection and a total of 122 families were sampled, with the number of families per provenance varying from four to 14.

The provenance-progeny trial was established on Itapeva Experimental Station of São Paulo, State Forest Institute. The site's latitude, longitude and altitude are, respectively, 24°17' S, 48°54' W and 930 m. The trial was established in a compact-family design, with 15 provenances (plots), from four to 14 families per provenance (subplots), 10 individuals per subplot and three replicates, 3 m x 2 m spacing was used and borders consisted of two rows. A total of 2,158 trees were phenotyped. The phenotypic data used in this study is fully described in Tassinari (2020) unpublished. In brief we used the methodology described by CALEGARIO; DANIELS; MAESTRI e NEIVA (2005) and LINDSTROM e BATES (1990) to adjust the continuous growth of the trees based on the nonlinear mixed-effect models. On this study we evaluated GS in *A. araucaria* for Volume (VOL) only, which was obtained using an equation based on DBH and HEI, for every year starting at age 7 to 35 years.

2. SNP and SSR Genotyping

Cambium tissues from 857 trees comprising all provenances and all families were silica-dried and stored at room temperature until DNA extraction. Total DNA was isolated using the CTAB protocol DNA extraction described by INGLIS; MARILIA DE CASTRO; RESENDE e GRATTAPAGLIA (2018), using a sorbitol pre-wash to grind 20–30 mg of tissue. DNA isolation was scaled to a 96 x 1.2 ml polypropylene cluster tube format. DNA purity was estimated using a spectrophotometer (Nanodrop 2000; Thermo Fisher Scientific). DNA yield was estimated, using a fluorimeter and fluorescent DNA-binding dye (QubitTM dsDNA BR Assay Kit; Thermo Fisher Scientific), according to the manufacturer´s instructions.

SNP genotyping was carried out at ThermoFisher (Santa Clara, CA) following standard protocols for the Axiom Platform an using the multispecies Axiom® myDesign ™ array that contained a total of 3,400 SNPs of *A. angustifolia* (GRATTAPAGLIA; SILVA-JUNIOR; RESENDE; SILVA, 2017). The genotyping data was analyzed using the Axiom Analysis Suite 3.1.

The same 857 trees were also genotyped with a set of 26 microsatellites (unpublished results). PCRs were carried out in tri or tetraplex systems with different fluorochromes (6-FAM (blue), NED (yellow), VIC (green)) and the PCR mixture with ROX-labeled size standard 75 was electroinjected in an ABI 3100XL genetic analyzer and data collected using GeneMapper (Applied Biosystems).

3. Statistical/Genomic analysis

3.1. Phenotypic model

The R package sommer was used to adjust the model below, the phenotypic model in Eq. 1 was based on model # 5 (Complete Blocks, Various Populations, Half-Sibling Families in one Location) of the SELEGEN-REML/BLUP software (RESENDE, 2016)

$$y = Xb + Za + Tp + Qs + e, \qquad [1]$$

where, $\boldsymbol{b}$ is the fixed effects of general average (intercept) and experimental blocks; $\boldsymbol{g}$ the random effects of individuals, with the structure of variances and covariance of individuals given by $g \sim N(0, A\sigma_a^2)$, since $\boldsymbol{A}$ is the matrix of kinship coefficients derived from pedigree (also known as identity-by-descent matrix); $\boldsymbol{p}$ are the random effects of populations/origins; $\boldsymbol{s}$ the random effects of plots; and $\boldsymbol{e}$ the random effects of residuals. The other structures of variance and covariance are given by $p \sim N(0, I\sigma_{pop}^2)$, $s \sim N(0, I\sigma_{parc}^2)$, $e \sim N(0, I\sigma_e^2)$. $\boldsymbol{X}$ is the matrix of incidence on fixed effects; $\boldsymbol{Z}, \boldsymbol{W}$ and $\boldsymbol{Q}$ are the matrices of incidence on random effects. The estimated breeding values (EBV) were considered the predicted values of $\hat{a}$, and EGV (total genetic values) were assumed to be $\hat{a} + \hat{s} + \hat{e}$ (ie, with additive genetic values plus non-additive effects between and within plots).

3.2. Genomic matrices: additive, dominance and epistatic

Genomic matrices were obtained using data from both SNPs, SSRs and the combination of these two types of markers. Because of its multi-allelic nature, the SSR data matrix was adapted to represent numerically a matrix of biallelic markers, ie, the alleles of each locus were expanded in several sub-columns to compose the $\boldsymbol{M}_{ssr}$ matrix of marks (CAVALLI-SFORZA; BODMER, 1999). With that, the additive parameterization matrix (concatenation of the $\boldsymbol{M}_{snp}$ and $\boldsymbol{M}_{ssr}$ matrices) is given by $M \subset \{2 - p_i, 1 - p_i, 0 - p_i\}$, where $\boldsymbol{M}$ can also marginally assume both $\boldsymbol{M}_{snp}$ and $\boldsymbol{M}_{ssr}$; and the dominance parameterization matrix (concatenation of the $\boldsymbol{H}_{snp}$ and $\boldsymbol{H}_{ssr}$ matrices) given by $H \subset \{-2q_i^2, -2p_iq_i, -2p_i^2\}$, where $\boldsymbol{H}$ can also marginally assume both $\boldsymbol{H}_{snp}$ and $\boldsymbol{H}_{ssr}$. The derivations of these parameterizations can be found in (VITEZICA; VARONA; LEGARRA, 2013). The constructs of the genomic matrices to compose the genomic selection models (described in the next topic) are shown below ($\boldsymbol{m}_i$ being the *i*-th marker of matrices $\boldsymbol{M}$ and $\boldsymbol{H}$; and '#' the Hadamard product between matrices):

Additive genomic matrix: $\boldsymbol{G} = \frac{\boldsymbol{MM'}}{\sum_1^{m_i} \boldsymbol{2p_iq_i}}$

Dominance genomic matrix: $\boldsymbol{D} = \frac{\boldsymbol{HH'}}{\sum_{\boldsymbol{1}}^{\boldsymbol{m_i}} (\boldsymbol{2p_i q_i})^2}$

Additive by additive Epistasis genomic matrix: $\boldsymbol{E_1} = \boldsymbol{G}\#\boldsymbol{G}$

Dominance by Dominance Epistasis genomic matrix: $\boldsymbol{E_2} = \boldsymbol{D}\#\boldsymbol{D}$

Additive by Dominance Epistasis genomic matrix: $\boldsymbol{E_3} = \boldsymbol{G}\#\boldsymbol{D}$;

Although the components of genomic variance for SNP and SSR markers are shown marginally in the results, the genomic selection models have been adjusted by combining both types of markers, with gains in capturing lost genetic variances, as will be better addressed in the discussions in this paper, but also to adopt a more parsimonious genomic selection model without numerous factors of variation.

3.3. Genomic model

Complete genomic model, addressing both additive genetic values, dominance and three types of epistasis (as described in the previous session), these models were adjusted using the free R package [sommer] (COVARRUBIAS-PAZARAN, 2016).

$$y^* = W\beta + Z_1 g_G + Z_2 g_D + Z_3 g_{GG} + Z_4 g_{DD} + Z_4 g_{GD} + \varepsilon \quad , \quad [2]$$

where $\boldsymbol{y}^+$ is the phenotypic value corrected by the model shown in Eq. [1], where $\boldsymbol{y}^+ = \hat{\boldsymbol{a}} + \hat{\boldsymbol{p}} + \hat{\boldsymbol{s}} + \hat{\boldsymbol{e}}$; β is the fixed general average effect (intercept); g_G, g_D, g_{GG}, g_{DD}, g_{GD} are the additive, dominance and epistatic genomic effects additive, dominant epistatic by dominant and epistatic additive by dominant, respectively, with variances and covariance in the form $g_G \sim N(0, G\sigma_G^2)$, $g_D \sim N(0, D\sigma_D^2)$, $g_{GG} \sim N(0, E_1\sigma_{G\#G}^2)$, $g_{DD} \sim N(0, E_2\sigma_{D\#D}^2)$, $g_{GD} \sim N(0, E_3\sigma_{G\#D}^2)$; ε is the model residue with a vcov structure equal to $\varepsilon \sim N(0, D\sigma_\varepsilon^2)$. $\boldsymbol{W}$ is the matrix of incidence on the fixed effect; Z_1, Z_2, Z_3, Z_4 and Z_5 are the matrices of incidence on random effects.

Reduced model used in the practice of genomic selection, using only the epistatic effect of the additive type by dominant (the justification for choosing this single type of epistasis will be found in the discussion of this work):

$$y^{*} = W\beta + Z_1 g_G + Z_2 g_D + Z_3 g_{GD} + \varepsilon \quad , \quad [3]$$

where, $y^{*}$is the phenotypic value adjusted by the model shown in Eq. [1], assuming EBVs and assuming EGVs; the other components of the model are exactly as described for the model in Eq. 2.

The total genomic heritability h^2_{GT}, as well as its additivity, dominance and epistasis partitions (h^2_G, $h^2_{D\,k}$ e $h^2_{E\,k}$, respectively) were obtained by the following equations, using the components of variance estimated in the Eq. 3 model (each one was calculated for the *k*-th measurement over time, being $k = \{7,8, \ldots ,35\}$.

$$h^2_{GT\,k} = \frac{\sigma^2_{G\,k} + \sigma^2_{D\,k} + \sigma^2_{GD\,k}}{\sigma^2_{G\,k} + \sigma^2_{D\,k} + \sigma^2_{GD\,k} + \sigma^2_{\varepsilon\,k}},$$

$$h^2_{G\,k} = \frac{\sigma^2_{G\,k}}{\sigma^2_{G\,k} + \sigma^2_{D\,k} + \sigma^2_{GD\,k} + \sigma^2_{\varepsilon\,k}},$$

$$h^2_{D\,k} = \frac{\sigma^2_{D\,k}}{\sigma^2_{G\,k} + \sigma^2_{D\,k} + \sigma^2_{GD\,k} + \sigma^2_{\varepsilon\,k}},$$

$$h^2_{E\,k} = \frac{\sigma^2_{GD\,k}}{\sigma^2_{G\,k} + \sigma^2_{D\,k} + \sigma^2_{GD\,k} + \sigma^2_{\varepsilon\,k}}.$$

3.4. Cross validation

Two types of cross-validation were performed: i) training in different types of populations (see Figure 10), sampling 14 populations and validation in the individuals in the remaining population; ii) training in different groups with sampling of individuals in all populations, ie, 60 individuals were taken for the training group, with 4 individuals in each of the 15 populations with the restriction of 2 individuals from the same family never being sampled in this group of validation. To avoid being at the mercy of chance in this second type of random validation, this procedure was repeated 100 times. The prediction of individuals not observed by the GS model was performed based on the genetic covariance used in the model, whether G, G+D or G+D+E_3.

3.5. Response to selection

The response to selection was calculated by adjusting the genomic selection model as early as possible and aiming to select the best individuals with phenotypic performance at 35 years of age. An evaluation of both the selection of individuals, selection of families and selection of the entire population was carried out. The best units (individuals, families or populations) were subsequently taken, starting from the best scenario, which is the selection of only one unit (greatest gain or best response to selection, in%), until the selection in which all units were sampled (scenario with zero gain, ie, 0%). The response to selection was calculated according to equation [4]:

$$RGS(\%) = \left(\frac{\overline{EBV_s} - \overline{EBV_0}}{\overline{EBV_0}}\right)100, \qquad [4]$$

where, $\overline{EBV_s}$ is the average of the selected portion and EBV_0 is the general population average. Two situations were considered: i) direct phenotypic selection, based on the ranking and averages of the selected EBVs; ii) indirect selection via genomics, based on the ranking of GEBVs, but with a response on the average of EBVs. Therefore, in this way, we have the selection response via phenotypic selection and via genomic selection.

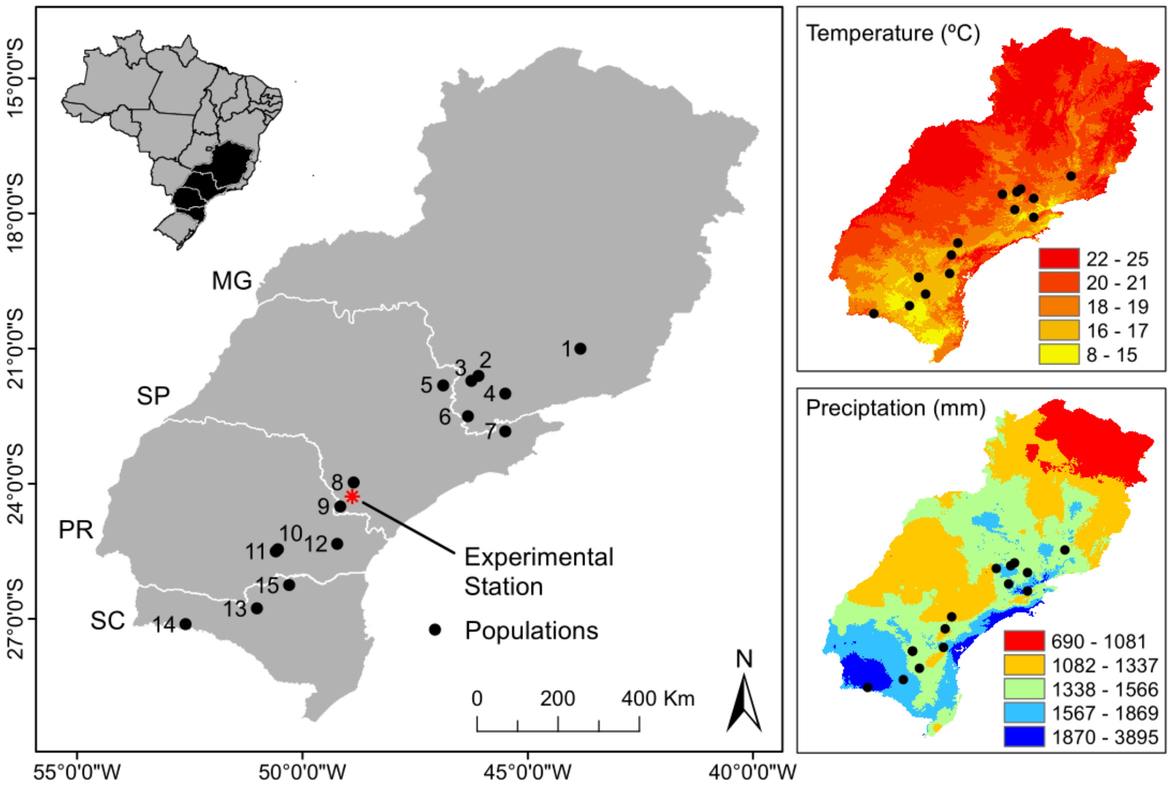

Figure 10 - Fifteen *A. angustifolia* populations into four Brazilian states (MG, SP, PR and SC) and the experimental station location. Also, the patterns of temperature and precipitation at the studied area are shown.

Results

Phenotypic and Genomic variance components

Among the phenotypic variation, the additive variance increased according to the increase in age, note that the phenotypic model was not able to capture additive variances at an earlier age. On the other hand, the environmental variance and the variation between plots decreased when increasing age. The variance between provenances or origins remained approximately stable over the ages, with a share of around 7% under the total phenotypic variation (Table 8).

In order to fully exploit the two types of available markers, we did some tests to justify the use of a single G matrix containing both SNP and SSR data. For this, we adopted molecular heritabilities (h_G^2) over time as a reference (Figure 11-a). When calculating h_G^2 in a single model incorporating both matrices (G_{SNP} e G_{SSR}), we observed a decrease in such heritabilities in comparison with the use of separate, individualized models, for each one. Note that by joining both matrices, the models over time are able to capture a greater amount of additive genetic variance, both by adding the two components and using a single matrix covering the two types of markers (see the two overlapping top lines of Figure 11-a). From these results, all development of this article was carried out with a unique G matrix containing the two types of available markers.

The molecular variances (additive, dominant and epistatic) were obtained with less than half of the data (857 individuals). In order to provide a better capture of their magnitudes, the phenotypes were corrected using all phenotypic data from all 2,158 individuals. Table 8 shows the behavior of additive, dominant and epistatic variances (both in their additive-by-additive, dominant-by-dominant and additive-by-dominant fractions), in five moments of the growth of *A. angustifolia* trees. With a clear objective of parsimony of the GS models, we chose to use a single Epistatic matrix. This choice was based on the amount of variance captured by each of the matrices, the reduction of the residual variance provided and the AIC index. The additive-by-additive epistatic matrix (E_1) captured the variance more than the ***G*** matrix, which seems somewhat unrealistic, and presented less favorable AICs than the others. The dominant-by-dominant matrix overestimated the residual variance in all scenarios and showed zero variances. The chosen one was the additive-by-dominant matrix (E_3), with the best AIC and lowest estimates of residual variances.

The gradual behavior of molecular variation can be seen in Figure 11-b. Note that additive effects have now been obtained starting at the earliest age (7 years), (remembering that it was not possible to capture additive variances using the phenotypic matrix) and gradually increased until it stabilized around 25 years. Dominance effects were observed as early as 19 years old, and it continued to grow until the last age of measurement at 35 years old, suggesting that it was in a moment before its stabilization point. Epistatic effects were observed throughout the growth period of the plants, however, always with low magnitude and with a subtle tendency to decrease over time.

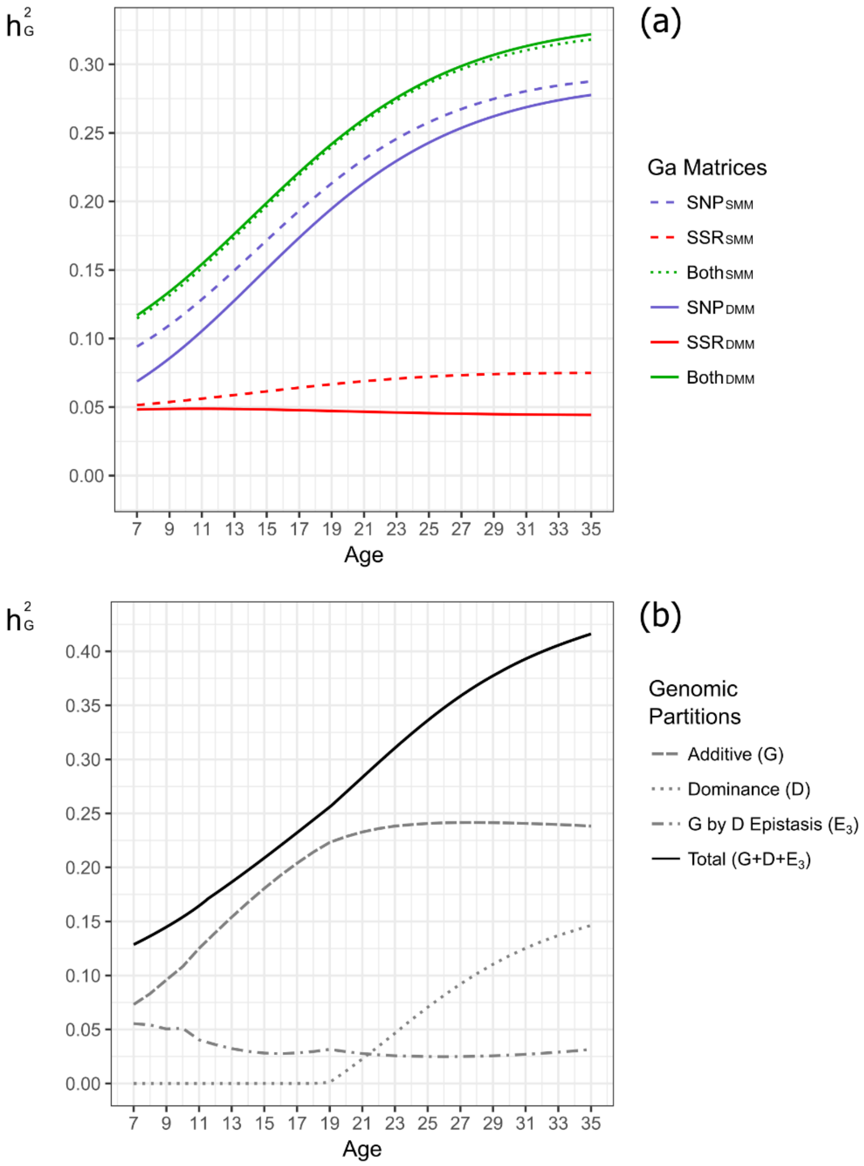

Figure 11 - Missing genomic heritability estimates. **a)** Genomic heritabilities with different marker types and with both. SMM: Single Matrix Model (SNP, SSR or both into a single matrix); DMM: Dual Matrices Model (SNP and SSR IBS matrices); **b)** Partitions of genomic effects into additive, dominance and epistasis fractions. The solid line is the sum of these three effects.

Table 8 - Variance components of phenotypic model using 2,158 phenotyped individuals.

	VOL07		VOL14		VOL21		VOL28		VOL35	
	value	%	value	%	value	%	value	%	value	%
σ^2_{parc}	7.02E-5	13.38%	9.35E-4	12.30%	2.84E-3	9.06%	4.64E-3	6.58%	5.63E-3	5.10%
σ^2_{a}	2.06E-9	0.00%	4.48E-4	5.88%	5.93E-3	18.88%	1.92E-2	27.25%	3.48E-2	31.52%
σ^2_{pop}	3.37E-5	6.43%	5.84E-4	7.68%	2.42E-3	7.72%	5.29E-3	7.51%	8.11E-3	7.34%
σ^2_{e}	4.20E-4	80.19%	5.64E-3	74.15%	2.02E-2	64.34%	4.13E-2	58.66%	6.19E-2	56.04%

Table 9 - Variance fractions using different GBLUP type models (including additivity, dominance and three shapes of epistasis matrices).

Trait	Factor	GBLUP_G		GBLUP_GD		$GBLUP\_GDE_1$		$GBLUP\_GDE_2$		$GBLUP\_GDE_3$		$GBLUP\_GDE_1E_2E_3$	
		value	%	value	%	value	%	value	%	value	%	value	%
VOL07	σ^2_G	7.60E-05	11.5	7.46E-05	11.3	6.18E-06	1.0	7.46E-05	11.3	4.78E-05	7.3	6.25E-06	1.0
	σ^2_D			1.62E-06	0.2	2.06E-09	0.0	1.64E-06	0.2	2.06E-09	0.0	1.87E-08	0.0
	$\sigma^2_{G\#G}$					6.49E-05	10.3					6.48E-05	10.3
	$\sigma^2_{D\#D}$							2.06E-09	0.0			2.06E-09	0.0
	$\sigma^2_{G\#D}$									3.62E-05	5.5	2.06E-09	0.0
	σ^2_ε	0.000586	88.5	0.000586	88.5	0.000561	88.8	0.000586	88.5	0.000569	87.1	0.000561	88.7
	AIC	-11.789		-9.789		-7.789		-7.789		-7.789		-3.789	
VOL14	σ^2_G	0.001544	18.5	0.001543	18.5	0.000364	4.7	0.001544	18.5	0.001389	16.8	0.000367	4.7
	σ^2_D			9.77E-07	0.0	5.50E-07	0.0	2.06E-09	0.0	1.10E-08	0.0	2.06E-09	0.0
	$\sigma^2_{G\#G}$					0.001201	15.4					0.001199	15.4
	$\sigma^2_{D\#D}$							1.82E-08	0.0			2.06E-09	0.0
	$\sigma^2_{G\#D}$									0.000246	3.0	2.06E-09	0.0
	σ^2_ε	0.006792	81.5	0.006792	81.5	0.006234	79.9	0.006792	81.5	0.006651	80.3	0.006234	79.9
	AIC	-10.789		-8.789		-6.791		-6.789		-6.789		-2.791	
VOL21	σ^2_G	0.008324	25.8	0.007853	24.2	0.0032	10.7	0.007853	24.2	0.007498	23.3	0.003199	10.7
	σ^2_D			0.000985	3.0	1.22E-08	0.0	0.000986	3.0	0.000733	2.3	6.08E-08	0.0
	$\sigma^2_{G\#G}$					0.00565	18.9					0.00565	18.9
	$\sigma^2_{D\#D}$							2.06E-09	0.0			2.06E-09	0.0
	$\sigma^2_{G\#D}$									0.000894	2.8	2.06E-09	0.0
	σ^2_ε	0.023908	74.2	0.023549	72.7	0.020973	70.3	0.023549	72.7	0.023084	71.7	0.020973	70.3
	AIC	-9.955		-7.955		-5.960		-5.955		-5.955		-1.960	
VOL28	σ^2_G	0.021928	30.1	0.018633	25.0	0.008639	12.7	0.018639	25.0	0.017927	24.2	0.008646	12.7
	σ^2_D			0.008044	10.8	0.003453	5.1	0.008026	10.8	0.007537	10.2	0.003429	5.0
	$\sigma^2_{G\#G}$					0.01377	20.3					0.01377	20.3
	$\sigma^2_{D\#D}$							8.77E-08	0.0			2.06E-09	0.0
	$\sigma^2_{G\#D}$									0.001871	2.5	2.06E-09	0.0
	σ^2_ε	0.051011	69.9	0.047867	64.2	0.04213	62.0	0.047874	64.2	0.046867	63.2	0.04214	62.0
	AIC	-9.207		-7.209		-5.218		-5.209		-5.209		-1.218	
VOL35	σ^2_G	0.03751	31.8	0.030336	24.9	0.014558	13.1	0.030325	24.9	0.028912	23.8	0.014556	13.1
	σ^2_D			0.018756	15.4	0.011029	9.9	0.018794	15.4	0.017755	14.6	0.011035	9.9
	$\sigma^2_{G\#G}$					0.022386	20.1					0.022386	20.1
	$\sigma^2_{D\#D}$							3.74E-09	0.0			2.06E-09	0.0

$\sigma^2_{G\#D}$									0.00383	3.2	2.06E-09	0.0
σ^2_{ε}	0.080422	68.2	0.072924	59.8	0.06346	56.9	0.072909	59.7	0.07085	58.4	0.063458	56.9
AIC	-8.581		-6.586		-4.598		-4.586		-4.586		-0.598	

Forms of epistasis matrices: E_1= GG; E_2= DD; E_3= GD. Genomic variance components: σ^2_G= traditional additive IBS matrix; σ^2_D= dominance matrix; σ^2_{GG}= additive by additive epistasis matrix; σ^2_{DD}= dominance by dominance epistasis matrix; σ^2_{GD}= additive by dominance epistasis matrix; σ^2_{ε}= residual component. AIC: Akaike Information Criteria.

Genomic selection within and between provenances

In this study we tested the opportunity to include dominant and epistatic effects in GS models in addition to the classic mandatory additive infinitesimal effects in every GS model. We aim to objectively make selections between and within the provenances. We used estimated breeding values (EBVs) as the basis for phenotypic selection, containing exclusively estimated additive effects; and the estimated genetic values (EGV) that contain additive and non-additive effects, which in turn culminate in clonal recommendations (usually in the final stages of a forest improvement program).

EBVs were used to adjust purely additive GS models (the classic GBLUP). EGVs were used in three GS adjustment scenarios: GBLUP, GBLUP plus genomic effects of dominance, and finally GBLUP plus genomic effects of dominance and epistasis. The GBLUP model adjusting for EBV data, not surprisingly, was the best, reaching predictive capacities of up to 46%, at more advanced ages. However, the selection made genomically between populations did not prove to be a good option. We observe at Table 10 that the best scenario is to perform GS always including on the models samples from all populations. Although the components of variance indicate that there are magnitudes of dominance and epistasis, especially expressing themselves over time, including such effects in genomic models didn't yield good results.

When dealing with conifers, selection at an early age is desired, targeting the desired trait at later ages. Figure 12 shows the behavior of predictive capacities correlating the predicted and observed both at the age of adjustment of the models and comparing the predicted at a given age with the values observed at other ages. For that, two scenarios are shown: in the upper part of this figure there is the cross validation between provenances and in the lower part, the validation within provenances. Note that by training the models by excluding entirely a provenance is not a good strategy, with low predictive capacities ranging from 6.5 to 9.5%. On the other hand, by including individuals from all provenances it was possible to achieve predictive capacities ranging from 39 to 41%. Considering the long selection cycle of the species, such values will allow for very significant selection gains. Carrying out indirect selection after 25 years provides almost the same result as when carrying out direct selection at 35 years of age (45% of predictive capacity). However, note that from the 12th year, the predictive ability for direct selection is approx. 42.5% and to perform indirect selection for the 35th year it is almost 41%.

When dealing with breeding values (EBVs), selecting individuals provided a greater response to selection than the selection of entire families, which showed a greater gain than selecting the entire provenance. In addition, we see at Figure 13 that genomic selection provided results as good as phenotypic selection (these results presented without cross-validation, ie., Consist of the direct application of the model already validated in this work).

Table 10 - Predictive abilities with standard deviations of GS models.

Trait	Cross validation between populations				Cross validation within populations (100× random)			
	EBV	EGV			EBV	EGV		
	GBLUP_G	GBLUP_G	GBLUP_GD	GBLUP_GDE$_3$	GBLUP_G	GBLUP_G	GBLUP_GD	GBLUP_GDE$_3$
V07	–	–	0.0047 (0.1457)	0.0023 (0.1675)	–	-0.1834 (0.0136)	0.0057 (0.1187)	-0.0183 (0.1163)
V14	0.0889 (0.1194)	–	-0.0386 (0.1589)	0.0503 (0.2046)	0.4398 (0.0827)	–	-0.0633 (0.0880)	-0.0386 (0.1106)
V21	0.0837 (0.1447)	0.0889 (0.1297)	0.0720 (0.1440)	0.0570 (0.1547)	0.4593 (0.0868)	0.1414 (0.1095)	0.1299 (0.1127)	0.1200 (0.1117)
V28	0.0729 (0.1655)	0.0714 (0.1310)	0.0916 (0.1308)	0.0861 (0.1297)	0.4673 (0.0915)	0.2030 (0.1173)	0.2083 (0.1186)	0.2020 (0.1161)
V35	0.0647 (0.1761)	0.0594 (0.1398)	0.1005 (0.1265)	0.0942 (0.1221)	0.4670 (0.0964)	0.2279 (0.1226)	0.2428 (0.1222)	0.2379 (0.1196)

OBS: Results between 7, 14, 21, 28 and 35 years were omitted for a leaner table, see Fig. 3 for complete information.

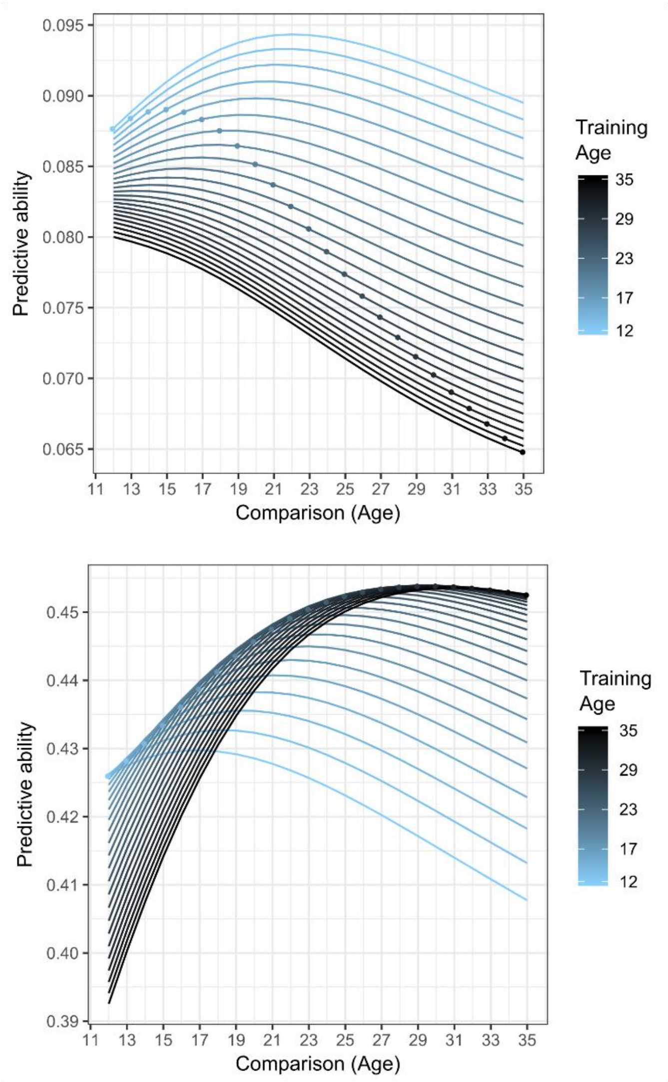

Figure 12 - Predictive abilities to indirect selection practice among growth ages, facing EBVs and GEBVs (both additive effects to pedigree and genomic values). The gradient colors (black-blue) represent the individuals age set used on GS model prediction. The lines represent the indirect selection at different growth ages. **Top chart)** Indirect selection comparison accounting cross-validation between populations. (**Bottom chart)** Indirect selection comparison accounting cross-validation within populations (random sampling).

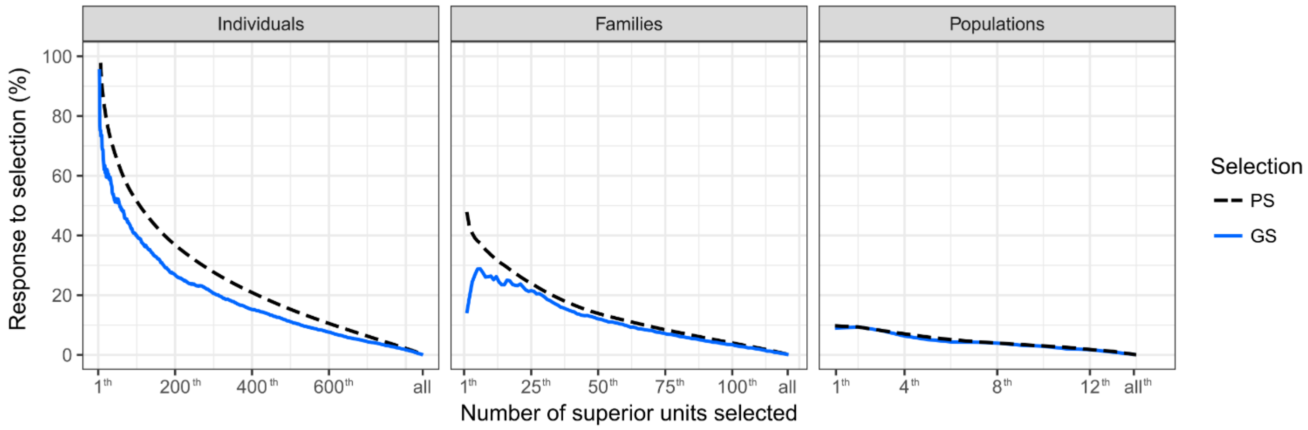

Figure 13 - Response to selection within families, i.e., between individuals (first box), between families (second box) and between populations (latest box). 'PS' means Phenotypic Selection and 'GS' means Genomic Selection. GS model was fitted at growth age 12 (i.e., GEBV at age 12) and compared with phenotypic ranking (i.e., EBV ranking) at growth age 35. These GS models' adjustments were done without any cross-validation procedure.

Discussion

Paraná pine is an iconic species in the southern and southeastern states of Brazil and due to its slow growth and also for legal reasons, the maintenance of individuals in preserved environments is a reality in the country (MONTAGNA; FERREIRA; STEINER; DA SILVA *et al.*, 2012). In this study, we used the genetic variability present in 15 provenances (Figure 10) in order to enable genomic selection to advance the improvement process and the domestication of the species. SG models proposed here comprise the initial stage of a feasible recurrent intrapopulation selection program - which in turn originated from random crossing within different populations - leading to a greater opportunity to obtain and select superior segregating individuals in terms of additive and heterotic values. The next step will be to cross the best individuals among populations, thus making a new selection using the adjusted GS models, and use vegetative propagation methods aiming at the fully preservation of superior genotypes, using somatic embryogenesis techniques (GRATTAPAGLIA; SILVA-JUNIOR; RESENDE; CAPPA *et al.*, 2018). The adoption of genomics as part of selecting individuals over multiple generations of conifers has shown good results as described by BARTHOLOMÉ; VAN HEERWAARDEN; ISIK; BOURY *et al.* (2016), evaluating three generations (G0, G1 and G2) of maritime pine, these authors reached accuracy of 0.70-0.80 in G2 selection.

Thinking about the effectiveness of the models proposed here, in addition to linkage disequilibrium (LD) and the effective population size, which are essential for the successful implementation of genomic selection (GRATTAPAGLIA; RESENDE, 2011), another component of equal importance is genetic variance of the trait to be improved in the population. Given the complexity of genomes in conifers (AMANDA; BIROL; BOUSQUET; INGVARSSON *et al.*, 2014) and in order to optimize the number of markers aiming at greater capture of additive variances, a blend of SNP and SSR were used (Figure 11-a). Although both markers share a part of the explanation of the additive variance, using both can aggregate more information about the inheritable nature of the trait Volume in Araucaria families. In *Phaseolus vulgaris*, the blend between Genotyping by Sequencing (GbS) and Diversity Arrays Technology (DArT) also demonstrated effectiveness in better capturing genetic variances of quantitative production traits (RESENDE; DE RESENDE; AZEVEDO; E SILVA *et al.*, 2018). It was observed that in Volume in Paraná pine trees, the additive phenotypic component disappears in the initial ages and reappears in later ages (Table 8). Similar results were obtained by LI e WU (2005) for DBH evaluating 26 years of growth in *Pinus radiata*. In the same direction, (XIANG; LI; ISIK, 2003) verified the absence of additive variances for volume in the early years in *Pinus taeda*, however these authors only evaluated eight years of growth. However, when using the genomic adjustment, it was possible to observe the magnitudes of distribution of the genetic variances - total and partial - in the first assessments at the age of seven (Figure 11-b). The families of *A. angustifolia* used in this study are open-pollinated, therefore, half-sib families, which are more difficult to capture additive effects and even more difficult to capture non-additives (LYNCH; WALSH, 1998). On the other hand, when we use DNA markers we benefit from the genomic relationship of individuals, making it possible to achieve the missing heritability lost in this case. In *Arabidopsis thaliana*, for example, it is shown that markers with wide coverage in the genome help to capture heritabilities that were previously unmeasurable (WAINSCHTEIN; JAIN; YENGO; ZHENG *et al.*, 2019).

In Table 9, we observe the components of variance achieved by the genomic models, addressing their additive and non-additive (dominant and epistatic) components. In this study we chose to only follow with the additive per dominant type epistatic matrix (G # D), in addition to the marginally additive and dominant effects, based on the quality of the models' fit (AIC). Going further, methods based on the infinitesimal additivity of the markers, such as GBLUP, are well-known (CAPPA; DE LIMA; DA SILVA-JUNIOR; GARCIA *et al.*, 2019), so investing in epistatic components that represent greater magnitudes than the additive values can be very risky. As is the case with the additive-by-additive epigenetic matrix, it captures much of the variation of the pure additive, something that did not seem reasonable to us. Genomic selection based on epistatic effects is generally discouraged, as shown by (TAN; GRATTAPAGLIA; WU; INGVARSSON, 2018). We can confirm this premise by observing what happened with the predictive capacity curves in Figure 14. Note these are very unstable and therefore little predictable over time, in contrary to what happens with genomic selection supported by additive or dominant effects. Therefore, although the E_1 matrix has a good ability to reduce residuals the AIC indicates a worse adjustment than that of the other models (Table 9).

EBVs, purely additive phenotypic values, are reference values throughout the improvement cycles of any crop. In forest breeding, when carrying out a forward type of selection (evaluation on the progeny and selection on the progeny itself) the effects of additivity and dominance are jointly important. Otherwise, backward selection (progeny assessment and selection of the best parents) is more appropriate to use only additive components. LENZ; NADEAU; AZAIEZ; GÉRARDI *et al.* (2020) carrying out forward selection failed to estimate dominance effects in a White spruce polycross test, however they still obtained good results for forward selection using only the additive effects. Although the components due to dominance were expressed, the predictive abilities using dominance effects were not satisfactory (Table 10). In contrast, using EGVs is interesting for cloning, without knowing male parents, they are merely corrected values by removing the effects of blocks and populations. The lack of knowledge of male parents is a factor for the dominant effects that were not well used in this work; the other factor is the fact that the populations were not crossed, which reduces the capture of the heterotic effects of individuals. Therefore, for selection in the middle of an improvement cycle the choice of the best individuals based on EBV (or genomically through GEBV) is recognized as the most suitable strategy for our data structure, providing predictive capabilities equivalent to other values reported in the literature, in the range of 0.45 (CHEN; BAISON; PAN; KARLSSON *et al.*, 2018; ISIK; BARTHOLOMÉ; FARJAT; CHANCEREL *et al.*, 2016; RATCLIFFE; EL-DIEN; KLÁPŠTĚ; PORTH *et al.*, 2015).

GS is dependent on linkage disequilibrium from relationship to function in its ideal. Our results show that when removing the entire validation population from the training model, predictive capacities drop significantly (upper part of Figure 14), and this behavior is observed for both additive and non-additive models. On the other hand, by performing cross-validation with a mix of populations and using additive models, there is the potential to achieve predictive capacities with 45% accuracy on average and it can reach around 65% accuracy for some validation groups. Although we assume that these results may be linked to the low amount of markers used, therefore low coverage throughout the genome, CHEN; BAISON; PAN; KARLSSON *et al.* (2018) evaluating Norway-spruce with ~117,000 SNPs observed the same pattern of low predictive capabilities when training and validating GS models in different families. In addition, in the case of LD, THISTLETHWAITE; RATCLIFFE; KLÁPŠTĚ; PORTH *et al.* (2017) had very optimistic predictive capacities ranging from 0.80-0.90 for full-sib families of another conifer - Douglas-fir, a level desired for *A. angustifolia* when increasing the coverage of the markers along the genome in the future.

In species with a very long rotation cycle like *Araucaria*, having to wait for 35 years of measurements to perform selection or even adjusting the GS models is not ideal. Dealing with phenotypic data, several studies have proposed ways to successfully reduce cycles of conifer selection (LI; WU, 2005; WU; DUAN; ZHANG, 2019). In our study we introduced the selection via DNA markers to the early selection process. The reduction of selection cycles via GS is a reality in long-cycle conifers (RATCLIFFE; EL-DIEN; KLÁPŠTĚ; PORTH *et al.*, 2015). As we have already discussed, the use of dominance effects or even epistasis did not provide good results of predictive capacity, in this sense we can also observe that the best strategy is to use all the data to compose the GS model, meaning, we believe that our model it is more efficient to additively select individuals in some selection cycles than to predict the behavior of individuals from different backgrounds (comparing the results of the upper and lower parts of Figure 12). Looking at the lower part of Figure 12, we can see that the later the moment of adjustment of the models, the better the indirect selection at different ages, however when training the model as early as 12 years old, the predictive ability (PA) in of the 35th year is approximately 0.41, an optimistic result, given that conducting training and selection simultaneously at 35 years of age (direct selection), PA has an insignificant increase, presenting a magnitude of approximately 0.45.

Although some studies indicate a low genetic variability of *A. angustifolia* (THOMAS, 2013), sparking a fear of practicing breeding with the species, our results indicate that this is a mistake, and that there is enough genetic variability to invest in breeding and commercial exploitation of the species. This variability can be explored both in the selection of individuals, families or provenances, as shown in Figure 13. Selection gains can be better utilized by pinning the best individuals within families, thus practicing a selection intensity and 10%, that is around 86 individuals compared to the 857 genotyped ones, it is possible to achieve a selection response equivalent to 52% dealing with phenotypic selection, and 42% dealing with genomic selection. From there, it is recommended that the composition of seed orchards, a well-recommended practice when breeding conifers, and from there the seeds are generated through the recombination of the best individuals (IWATA; HAYASHI; TSUMURA, 2011).The selection of the best (whole) families or even the best provenances did not provide selection responses as high as the individual selection, because no matter how much we carry superior individuals within these families, our results show that there is sufficient variability in families and provenances higher than the point of presenting both excellent and not so good individuals, pulling the selection average down, consequently compromising the selection response. In view of these results, the next steps indicate the selection and crossing of the best individuals forming a new generation of breeding, especially respecting the selection in the best families or provenances.

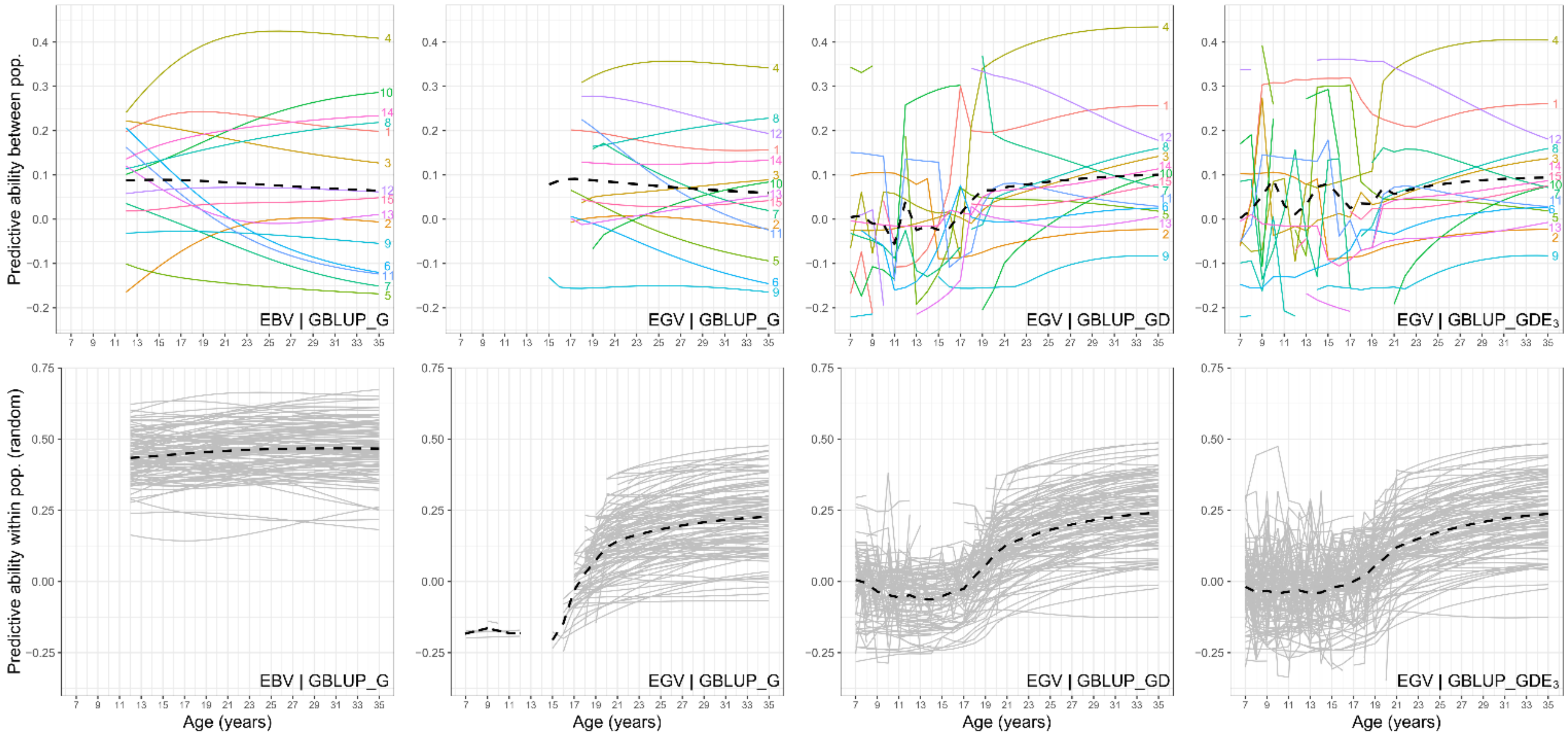

Figure 14 - Predictive abilities accounting the cross validation between (**top charts**) and within populations (**bottom charts**) along growth ages. EBV are estimated breeding values based on additive prediction effects ($\hat{a}$) and EGV are total genetic values based on the summation of additive values and within and between plot effects ($\hat{a} + \hat{s} + \hat{e}$) from the Eq. [1]. GBLUP_G, GBLUP_GD and GBLUP_GDE$_3$ are the reduced GS models from the complete model of Eq. [3].

Conclusion

Mixing SNP and SSR markers into a single genomic matrix proved to be a good strategy. It makes it possible to capture a greater amount of genetic variance, around +5%, compared to using only SNPs; and around +30% when compared to SSRs only. It is possible to observe components of additive, dominant and epistatic variances throughout the growth of *A. angustifolia* trees, the additive components are the most prevalent and tend to increase with age, those of dominance appear only from the 19th year, and the epistatic have low magnitude and are constant over time. Although there are additive and non-additive components of variance over time, using the purely additive GBLUP method is the best strategy to proceed with GS in *A. angustifolia*. Individuals of all populations or backgrounds must always be kept in the models to ensure the best predictive capabilities. Regarding early selection, a good alternative is to adjust the GS models at age of 12 years, which will guarantee a reliable indirect selection at 35 years. GS proved to be competitive in comparison to the merely phenotypic selection in *A. angustifolia*, both for individuals, families and provenances, opening horizons for breeding for this slow growing iconic tropical conifer.

References

AMANDA, R.; BIROL, I.; BOUSQUET, J.; INGVARSSON, P. K. *et al.* Insights into conifer giga-genomes. **Plant Physiology**, 166, n. 4, p. 1724-1732, 2014.

BARTHOLOMÉ, J.; VAN HEERWAARDEN, J.; ISIK, F.; BOURY, C. *et al.* Performance of genomic prediction within and across generations in maritime pine. **BMC genomics**, 17, n. 1, p. 604, 2016.

CALEGARIO, N.; DANIELS, R. F.; MAESTRI, R.; NEIVA, R. Modeling dominant height growth based on nonlinear mixed-effects model: a clonal Eucalyptus plantation case study. **Forest Ecology and Management**, 204, n. 1, p. 11-21, 2005/01/03/ 2005.

CAPPA, E. P.; DE LIMA, B. M.; DA SILVA-JUNIOR, O. B.; GARCIA, C. C. *et al.* Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. **Plant science**, 284, p. 9-15, 2019.

CAVALLI-SFORZA, L. L.; BODMER, W. F. **The genetics of human populations**. Courier Corporation, 1999. 0486406938.

CHEN, Z.-Q.; BAISON, J.; PAN, J.; KARLSSON, B. *et al.* Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. **BMC genomics**, 19, n. 1, p. 946, 2018.

COVARRUBIAS-PAZARAN, G. Genome-assisted prediction of quantitative traits using the R package sommer. **PloS one**, 11, n. 6, 2016.

DE ALMEIDA FILHO, J.; GUIMARÃES, J.; E SILVA, F. F.; DE RESENDE, M. *et al.* The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, 117, n. 1, p. 33-41, 2016.

GRATTAPAGLIA, D. Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. *In*: TUBEROSA, R.;GRANER, A.*, et al* (Ed.). **Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources**. Dordrecht: Springer Netherlands, 2014. p. 651-682.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genet Genomes**, 7, 2011.

GRATTAPAGLIA, D.; SILVA-JUNIOR, O. B.; RESENDE, L. V.; SILVA, P. I. T. A five-species 50K Axiom SNP microarray allows high quality genotyping of Coffee, Cashew, Cassava, Brazilian Pine and Eucalyptus. *In*: Plant & Animal Genome XXIV, 2017, Scherago, San Diego. p. pp. 26564.

GRATTAPAGLIA, D.; SILVA-JUNIOR, O. B.; RESENDE, R. T.; CAPPA, E. P. *et al.* Quantitative genetics and genomics converge to accelerate forest tree breeding. **Frontiers in Plant Science**, 9, p. 1693, 2018.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J.-L. Genomic Selection for Crop Improvement. **Crop Science**, 49, n. 1, p. 1-12, 2009.

INGLIS, P. W.; MARILIA DE CASTRO, R. P.; RESENDE, L. V.; GRATTAPAGLIA, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. **PloS one**, 13, n. 10, 2018.

ISIK, F.; BARTHOLOMÉ, J.; FARJAT, A.; CHANCEREL, E. *et al.* Genomic selection in maritime pine. **Plant Science**, 242, p. 108-119, 1// 2016.

ISIK, K.; KLEINSCHMIT, J.; STEINER, W. Age–age correlations and early selection for height in a clonal genetic test of Norway spruce. **Forest science**, 56, n. 2, p. 212-221, 2010.

IWATA, H.; HAYASHI, T.; TSUMURA, Y. Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. **Tree Genet Genomes**, 7, 2011.

IWATA, H.; JANNINK, J.-L. Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. **Crop Science**, 51, n. 5, p. 1915-1927, 2011.

LENZ, P. R.; NADEAU, S.; AZAIEZ, A.; GÉRARDI, S. *et al.* Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs: an example from white spruce. **Heredity**, p. 1-17, 2020.

LI, L.; WU, H. X. Efficiency of early selection for rotation-aged growth and wood density traits in Pinus radiata. **Canadian Journal of Forest Research**, 35, n. 8, p. 2019-2029, 2005.

LINDSTROM, M. J.; BATES, D. M. Nonlinear mixed effects models for repeated measures data. **Biometrics**, p. 673-687, 1990.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sinauer Sunderland, MA, 1998.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157, n. 4, p. 1819-1829, Apr 2001.

MONTAGNA, T.; FERREIRA, D. K.; STEINER, F.; DA SILVA, F. A. L. S. *et al.* A importância das Unidades de Conservação na manutenção da diversidade genética de araucária (Araucaria angustifolia) no Estado de Santa Catarina. **Biodiversidade Brasileira**, 2, n. 2, p. 18-25, 2012.

NUTTO, L.; SPATHELF, P.; ROGERS, R. Managing diameter growth and natural pruning of Parana pine, Araucaria angustifolia (Bert.) O Ktze., to produce high value timber. **Annals of forest science**, 62, n. 2, p. 163-173, 2005.

OLIVEIRA, E. J.; PÁDUA, J. G.; ZUCCHI, M. I.; VENCOVSKY, R. *et al.* Origin, evolution and genome distribution of microsatellites. **Genetics and Molecular Biology**, 29, p. 294-307, 2006.

RATCLIFFE, B.; EL-DIEN, O. G.; KLÁPŠTĚ, J.; PORTH, I. *et al.* A comparison of genomic selection models across time in interior spruce (Picea engelmannii× glauca) using unordered SNP imputation methods. **Heredity**, 115, n. 6, p. 547-555, 2015.

RESENDE, M. D. V.; RESENDE, M. F. R.; SANSALONI, C. P.; PETROLI, C. D. *et al.* Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194, n. 1, p. 116-128, 2012.

RESENDE, M. D. V. d. Software Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology**, 16, n. 4, p. 330-339, 2016.

RESENDE, M. F. R.; MUÑOZ, P.; ACOSTA, J. J.; PETER, G. F. *et al.* Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytol**, 193, 2012.

RESENDE, M. F. R.; MUÑOZ, P.; RESENDE, M. D. V.; GARRICK, D. J. *et al.* Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). **Genetics**, 190, 2012.

RESENDE, R. T.; DE RESENDE, M. D. V.; AZEVEDO, C. F.; E SILVA, F. F. *et al.* Genome-wide association and regional heritability mapping of plant architecture, lodging and productivity in Phaseolus vulgaris. **G3: Genes, Genomes, Genetics**, 8, n. 8, p. 2841-2854, 2018.

SEBBENN, A.; PONTINHA, A.; GIANNOTTI, E.; KAGEYAMA, P. Genetic variation in provenance-progeny test of Araucaria angustifolia (Bert.) O. Ktze. in São Paulo, Brazil. **Silvae genetica**, 52, n. 5-6, p. 181-184, 2003.

TAN, B.; GRATTAPAGLIA, D.; MARTINS, G. S.; FERREIRA, K. Z. *et al.* Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F 1 hybrids. **BMC plant biology**, 17, n. 1, p. 110, 2017.

TAN, B.; GRATTAPAGLIA, D.; WU, H. X.; INGVARSSON, P. K. Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. **Plant Science**, 267, p. 84-93, 2018.

THISTLETHWAITE, F. R.; RATCLIFFE, B.; KLÁPŠTĚ, J.; PORTH, I. *et al.* Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. **BMC genomics**, 18, n. 1, p. 930, 2017.

THOMAS, P. Araucaria angustifolia. **The IUCN red list of threatened species**, 2013, 2013.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **J Dairy Sci**, 91, n. 11, p. 4414-4423, Nov 2008.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, 195, n. 4, p. 1223-1230, 2013.

WAINSCHTEIN, P.; JAIN, D. P.; YENGO, L.; ZHENG, Z. *et al.* Recovery of trait heritability from whole genome sequence data. **BioRxiv**, p. 588020, 2019.

WU, H.; DUAN, A.; ZHANG, J. Long-term Growth Variation and Selection of Geographical Provenances of Cunninghamialanceolata (Lamb.) Hook. **Forests**, 10, n. 10, p. 876, 2019.

XIANG, B.; LI, B.; ISIK, F. Time trend of genetic parameters in growth traits of Pinus taeda L. **Silvae genetica**, 52, n. 3-4, p. 114-120, 2003.

ZHAO, D.; KANE, M.; TESKEY, R.; FOX, T. R. *et al.* Maximum response of loblolly pine plantations to silvicultural management in the southern United States. **Forest Ecology and Management**, 375, p. 105-111, 2016.

Conclusões

Dados fenotípicos além dos registros de pedigree têm sido a base dos programas de melhoramento genético na agricultura. O surgimento da seleção genômica oferece novas oportunidades para acelerar as taxas de ganho genético a um custo mais baixo, além de reduzir a dependência do registro de pedigree, permitindo que as matrizes de relacionamento estimado sejam substituídas por matrizes de relacionamento genômico derivada de genótipos. Sendo assim, o sucesso da seleção tradicional e genômica depende da coleta de fenótipos com qualidade para fornecer estimativas precisas do mérito genético, mas também depende da disponibilidade de marcadores moleculares de alta qualidade. Neste trabalho foi disponibilizado um conjunto de dados de alta qualidade para características de crescimento de 2.158 árvores de *Araucaria angustifolia* provenientes de 15 procedências, incluindo 122 famílias de meios-irmãos. Após um ajuste nos dados fenotípicos coletados usando funções de crescimento adequadas e modelagem individual de crescimento de árvore. Esses novos valores estimados tornaram possível capturar melhor a variação genética. Os dados estimados mostram que embora a espécie tenha uma longa história de exploração descontrolada os dados fenotípicos indicam altos níveis de diversidade genética para crescimento sendo possível agrupar as procedências em dois grupos: norte e sul. Também foi possível observar que as procedências mais próximas geograficamente mostraram padrões de crescimento semelhantes quando plantadas em um experimento de jardim comum. A metodologia de estimação de dados fenotípicos em idades não amostradas apresentada nesse estudo fornece uma ferramenta valiosa na implementação de um futuro programa de melhoramento de *A*. *angustifolia*. Neste trabalho também desenvolvemos o primeiro conjunto de SNPs para *Araucaria angustifólia*. A partir do catálogo de 44.318 SNPs anotados em todo o transcriptoma, foi desenvolvido um array Axiom® SNP com ~ 3.000 SNPs validados. Comparando-se os SNP com os dados de microssatélites, nossos resultados indicam que o uso de marcadores de microssatélites para cálculo de estimativas de diversidade e diferenciação genética não refletem precisamente os padrões reais de variação e estruturação em todo o genoma. A inferência geralmente baseada em microssatélites de que *A. angustifolia* tem sido resistente a perdas rápidas de diversidade genética devido à fragmentação da floresta pode não ser totalmente justificada e deve receber mais atenção. Diversidade genética superestimada, estimativas imprecisas de endogamia e diferenciação populacional subestimada podem ter consequências sérias nas decisões sobre como abordar e gerenciar a conservação dos recursos genéticos *ex situ* e *in situ* da espécie. Nossos resultados não questionam a utilidade dos microssatélites em geral como ferramenta eficiente para o estudo de sistemas de cruzamento e parentesco, mas alertamos sobre as estimativas de diversidade, endogamia e diferenciação dadas as limitações conhecidas. Como as aplicações genéticas dependem de estimadores de vários locus de diferenciação, painéis de várias centenas a milhares de SNPs distribuídos em todo o genoma sempre serão mais poderosos, representativos e precisos do que poucos microssatélites. Com dados fenotípicos ajustados e genótipos de qualidade em mãos foi possível também avaliar a capacidade preditiva de modelos de predição genômica para seleção de indivíduos e famílias para características de crescimento em diferentes idades e comparar o potencial da seleção genômica em relação ao melhoramento convencional. Nosso trabalho mostra que a utilização de marcadores SNPs e SSRs combinados em uma única matriz genômica provou ser uma boa estratégia em *A. angustifolia*. Permitindo capturar uma quantidade maior de variação genética quando comparado ao uso de SNPs ou SSRs separadamente. Foi possível também observar componentes de variações aditivas, dominantes e epistáticas ao longo do crescimento de árvores. Sendo os componentes aditivos os mais prevalentes e tendendo a aumentar com a idade. Os componentes dominantes apareceram apenas a partir do 19º ano de idade e os epistáticos se mostraram de baixa magnitude e constantes ao longo do tempo. Embora existam componentes de variação aditivos e não aditivos ao longo do tempo, o uso do método GBLUP puramente aditivo foi a melhor estratégia para prosseguir com a seleção genômica em *A. angustifolia*. Indivíduos de todas as populações ou procedências devem sempre ser mantidos nos modelos de predição para garantir as melhores capacidades preditivas. Em relação à seleção precoce a alternativa proposta é se ajustar os modelos aos 12 anos de idade, o que garantirá uma seleção indireta confiável aos 35 anos. A seleção genômica provou ser competitiva em comparação com a seleção meramente fenotípica tanto para indivíduos quanto para famílias e procedências abrindo horizontes para programas de melhoramento.

Os métodos descritos neste trabalho bem como os conjuntos de dados disponibilizados, ampliam enormemente as possibilidades de estudos com genéticos em *A. angustifolia*, mostrando o potencial de acelerar os esforços ainda muito tímidos do melhoramento dessa espécie brasileira icônica de enorme importância social e econômica.