University of São Paulo
"Luiz de Queiroz" College of Agriculture

SNP discovery, high-density genetic map construction, and identification of genes associated with climate adaptation, and lack of intermuscular bone in tambaqui (*Colossoma macropomum*)

José de Ribamar da Silva Nunes

Thesis presented to obtain the degree of Doctor in Science. Area: Animal Science and Pastures

Piracicaba
2017

José de Ribamar da Silva Nunes
Animal Scientist

SNP discovery, high-density genetic map construction, and identification of genes associated with climate adaptation, and lack of intermuscular bone in tambaqui (*Colossoma macropomum*)

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **LUIZ LEHMANN COUTINHO**

Thesis presented to obtain the degree of Doctor in Science. Area: Animal Science and Pastures

Piracicaba
2017

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

Nunes, José de Ribamar da Silva

SNP discovery, high-density genetic map construction, and identification of genes associated with climate adaptation, and lack of intermuscular bone in tambaqui (*Colossoma macropomum*) / José de Ribamar da Silva Nunes. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2017.

74 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Tambaqui 2. Descoberta de SNP 3. Mapa genético 4. Associação 5. Osso intermuscular I. Título

DEDICATION

I dedicate this work to my wife Margareth and my son Miguel for all patience, sacrifice, abdication and understanding during my doctoral studies and during this phase of our lives.

ACKNOWLEDGEMENTS

First of all, my special thanks to author and perfecter of my faith, who provided all resources for this moment. Who has sustained me and who placed special people in my way. To Him, I give all honor, all glory, all praise, and all adoration. Thank you, my God.

I am really grateful to my advisor, Dr. Luiz Coutinho for his support, guidance, professionalism, commitment and patience. During this time, I received more than guidance, I received respect and opportunities. I could choose my ways and I was respected and supported in what was possible and warned when was needed. I am grateful for the opportunity to work with Dr. Luiz Coutinho.

Thanks to University of São Paulo - "Luiz de Queiroz" College of Agriculture (USP/ESALQ) and to Federal University of Amazon (UFAM) for opportunity, knowledge, and support.

I would like to say thanks to Dr. Alexandre Wagner Silva Hilsdorf at University of Mogi das Cruzes for his work suggestions, help, and valuable partnership during the development of this project and Caio Perazza for help with sample collection to analysis. I also want to say thank you Prof. Dr. Vera Maria Almeida Val de Fonseca at Brazilian National Institute for Research of the Amazon who provide resources for the continuation and expansion of this work.

Thanks to all members of the Animal Biotechnology Laboratory, especially Fabio Pértille for the help in analysis and friendship, Priscilla Villela for the support in the libraries of GBS, Sonia Andrade and Horacio Montenegro for their help and patience in bioinformatic analysis. For technical staff Ricardo Brassaloti, Marcela Paduan, Gustavo Gasparin, Nirlei A. Silva, Pilar Mariani and Jorge Luís F. de Andrade. To friends Gabriel Costa, Thaís Godoy, Gabriella Borba, Karina Yu, Paula Soarez, Aline Cesar, Andrezza Felício, Mirele Poleti, Rosy Simas. Thanks for your friendship on the difficult and happy moments. I am grateful for the opportunity to work with y'all at Animal Biotechnology Laboratory.

To Dr. Zhanjiang (John) Liu at Auburn University that accepted me as a visiting researcher in Fish Molecular Genetics and Biotechnology Laboratory at Auburn University and to Shikai Liu for the valuable help during analysis and writing process.

Thanks for Amazonas State Research Support Foundation (FAPEAM) by the scholarship granted since the beginning of this project in Brazil and to Coordination for the Improvement of Higher Education Personnel (CAPES) by the scholarship granted in abroad by the Doctoral Sandwich Abroad Program (PDSE).

To my friends, brothers and sisters in Christ Ramona and Mark Davis, Pavel Kulakov, Brianna Shearer that supported me during all of my time in Alabama. Dale Simpson for support during the writing process of my thesis. To all of my friends from Brazil Rose Zanatta, Alexandre and Talita Ceschim for help and friendship.

Special thanks to my family. To my sisters and others relatives who believed and supported me. To my parents Domingos Reis and Maria das Graças who provided the base to all of my victories and who always are praying for me. Particularly I am so grateful for my son Miguel Nunes which often was penalized and sacrificed in this process, but always faced everything with optimism and confidence. And I am so grateful also to my wife Margareth Nunes, who always told me what I needed to hear instead of what I wanted to hear. You are my great reasons to go ahead.

EPIGRAPH

"The greatest want of the world is the want of men—men who will not be bought or sold, men who in their inmost souls are true and honest, men who do not fear to call sin by its right name, men whose conscience is as true to duty as the needle to the pole, men who will stand for the right though the heavens fall.

Ellen G. White

SUMMARY

RESUMO

Descoberta de SNP, construção de mapa genético de alta densidade e identificação de genes associados com adaptação climática e ausência da espinha intermuscular em tambaqui (*Colossoma macropomum*)

O tambaqui (*Colossoma macropomum*) é a maior espécie nativa de Characiforme da América do Sul e é encontrado nas bacias do rio Amazonas e Orinoco. O cultivo do tambaqui está crescendo rapidamente no Brasil, sua produção atingiu 139.209 toneladas em 2014, o que corresponde a 57,7% de aumento em relação a 2013. No entanto, poucos estudos genéticos realizados com o tambaqui estão disponíveis atualmente. Estudos genéticos em tambaqui, tanto em populações cultivadas quanto em populações selvagens, necessitam de uma abordagem holística para uma ação racional frente aos desafios ecológicos e mercadológicos na aquicultura. Abordagens baseadas em estudos genéticos têm fornecido ferramentas importantes para se entender a dinâmica populacional, adaptação local e função gênica visando melhorar as estratégias de seleção a serem aplicadas em programas de melhoramento genético. O sequenciamento de nova geração (NGS) permitiu um grande avanço nas abordagens genômicas e transcriptômicas, especialmente relacionadas a espécies não-modelo. A genotipagem por sequenciamento (GBS) é uma dessas abordagens que utilizam enzimas de restrição (REs) para reduzir a complexidade do genoma. Esta tese apresenta a aplicação desta abordagem objetivando proporcionar avanços significativos nos estudos genéticos de base para tambaqui. A técnica de GBS forneceu um painel de SNPs de alta densidade que nos permitiu desenvolver o primeiro mapa de ligação e estudos de associação com variáveis ambientais, adaptação local e ausência de ossos intermusculares no tambaqui. Este trabalho pode nos dar muitas referências teóricas a serem aplicadas em programas de melhoramento genético do tambaqui, permitindo uma melhor compreensão dos processos genéticos relacionados a traços de interesse na aquicultura.

Palavras-chave: Tambaqui; Descoberta de SNP; Mapa genético; Associação; Espinha intermuscular

ABSTRACT

SNP discovery, high-density genetic map construction, and identification of genes associated with climate adaptation, and lack of intermuscular bone in tambaqui (*Colossoma macropomum*)

Tambaqui (*Colossoma macropomum*) is the largest native Characiform species from the Amazon and Orinoco river basins of South America. Tambaqui farming is growing rapidly in Brazil, its production reached 139.209 tons in 2014, what corresponds to 57.7% of increase compared with 2013. However, few genetic studies of tambaqui are currently available. The tambaqui genetic studies for cultured and wild populations need a holistic approach for a rational action facing ecological and market challenges in aquaculture. Approaches based on genetic studies have provided important tools to understand population dynamics, local adaptation, and gene function to improve selection strategies to be applied in breeding programs. The next-generation sequencing (NGS) allowed a great advance in genomic and transcriptomic approaches, especially related to non-model species. The genotype-by-sequencing (GBS) is one of this approaches based on genome complexity reduction using restriction enzymes (REs). This thesis presents the application of these approaches to provide advances in the genetic background for tambaqui studies. The GBS approach provided a high-density SNPs panel that allowed us to develop the first linkage map, and association studies with environmental variables, local adaptation, and lack of intermuscular bones, both using tambaqui as a model. This work can give us many theoretical references to be applied in genetic breeding programs for tambaqui, allowing a better understanding of genetic processes related to traits of interest in aquaculture.

Keywords: Tambaqui; SNP discovery; Genetic map; Association; Intermuscular bone

1. INTRODUCTION

Tambaqui (*Colossoma macropomum*) is the largest native Characiform species from the Amazon and Orinoco river basins of South America [1]. It is a semi-migratory fish [2], with seasonal migrations restrict to floodplains and floodplain lakes for reproduction and feeding [3]. In aquaculture, tambaqui has been recognized as an economically valuable species because it has an omnivorous behavior, fast-growing, poor water quality tolerance, and has a well market acceptance by the consumer.[4]. Tambaqui farming is growing rapidly in Brazil, its production reached 139,209 tons in 2014, what corresponds to 57.7% of increase compared with 2013 [5]. This species is also farmed in different countries around the Amazon boundary and was introduced into other countries of Latin America, Asia, and Africa [6]. However, genetic studies of tambaqui are limited to genetic diversity surveys [7,8,9], transcriptome analysis under specific conditions [10], a couple of SNPs studies [11] and mitochondrial genome sequencing [12]. The tambaqui genetic studies of farmed and wild populations need a holistic approach for a rational action facing ecological and market challenges in aquaculture.

Genetic studies have provided important information to understand population dynamics, local adaptation, gene function and improve selection strategies based on genomic information in breeding programs for several species [13,14]. Generally, these studies use molecular markers to characterize a population or to evaluate traits of interest associated with them. Genetic markers are polymorphic DNA regions distributed along the chromosomes [15] as restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), amplified fragment length polymorphisms (AFLPs), single-nucleotide polymorphisms (SNPs) and short tandem repeats (STR) [16]. SNPs have been well-recognized as the marker of choice for genetic studies in various organisms [17]. SNPs are the most abundant molecular markers in any vertebrate genome, with a SNP presented in each 100–500 bp on average [18]. SNPs are mostly bi-allelic, making them amenable for high-throughput genotyping using SNP arrays [19,20]. The SNP markers development before next-generation sequencing (NGS) technology had been a costly, laborious and time-consuming work [21].

The NGS allowed huge advance in genomic and transcriptomic approaches [22,23,24] especially for non-model species [21]. The reduced complexity libraries from NGS approaches based on digestion by restriction enzyme are now highlighted because they allow efficient and cost-effective SNP discovery on a genome-scale in any organisms without prior genome information [25]. Specifically, the genotype-by-sequencing (GBS) is one of these NGS approaches [24] characterized as simple, quick, specific, reproducible [24], and has been extensively used to identify a large number of SNPs in various model and non-model species [26,27,28,29].

Genetic linkage maps are essential resources for genome and genetic research. They provide frameworks for understanding genome structure and function. Linkage maps are essential for quantitative trait locus (QTL) identification and mapping. QTLs mapping has been an important way to identify genes related to trait variations within and between populations or species, allowing implementation of genetic and breeding programs [30]. Genetic maps can be used to facilitate genome assembly corrections, anchoring scaffolds onto linkage groups to build chromosomal assembly [31].

The recent introduction of tambaqui in several regions of Brazil could create new patterns of adaptive variation. The identification of patterns of adaptive variation in non-model species has been challenging particularly for species with recently demographic histories[34]. Genomic datasets are now permitting the identification of outlier loci with unprecedented accuracy. Outliers may be an important tool to indicate regions under selection. allowing a quick response to the challenges posed by climate changes[35]. Also, allows the identification of putative loci associated

or integrated to genes with a known adaptive function. These genes can contribute to important particular processes that can be related to metabolic pathways, or be associated with important phenotypes in aquaculture [36].

The presence of intermuscular bones has been a problem in consumption of many farmed fish species [37,38,39]. This phenotype originates from incomplete membranous ossifications of connective tissue in the muscular septum [40] as a response to strains suffered by muscles[41]. It can have important functions in transmission of force, contraction, body firmness, and reduction of myomere deformation[40]. These characteristics can be of great importance to wild fishes [42,43], however, they can be secondary in farmed fishes that are not exposed to stresses characteristic of wild environments, such as food competition, predation, migration and seasonal variations. Tambaqui lacking intermuscular bones were found and reported by Perazza et al. (2016), and can be a model in studies of lack intermuscular bones for other species[44]. However, the genetic mechanisms involved in this phenotype have not yet been discovered. Advances in genomic research have significantly improved the tools available for the study of commercially important traits in aquaculture [45,46]. The use of association studies can help to unravel the biological mechanisms involved in this characteristic.

This thesis presents the effort in the development of studies and tools that can provide advances in the genetic improvement of tambaqui. Our approach based on GBS provided a high-density SNPs panel that allowed us to develop the first linkage map of tambaqui and association studies with environmental variables, local adaptation, and lack intermuscular bones in tambaqui. This thesis is a first effort to discover theoretical references in tambaqui breeding, allowing the understanding of genetic processes that are involved with important traits in aquaculture of this species.

References

1. Araujo-Lima, C. & Goulding, M. So Fruitful a Fish: Ecology, Conservation and Aquaculture of the Amazon's Tambaqui. (Columbia Univ Press, 1997).
2. Araújo-Lima, C. A. R. M. & Ruffino, M. L. Migratory fishes of the Brazilian Amazon. Migratory Fishes of South America: Biology, Fisheries and Conservation Status (2004).
3. Da Silva, J. A. M., Pereira Filho, M. & De Oliveira-Pereira, M. I. Frutos e sementes consumidos pelo tambaqui, Colossoma macrompum (Cuvier, 1818) incorporados em rações. Digestibilidade e velocidade de trânsito pelo trato gastrointestinal. Rev. Bras. Zootec. 32, 1815–1824 (2003).
4. Campos-baca, L. & Kohler, C. C. Aquaculture of Colossoma macropomum and Related Species in Latin America. Am. Fish. Soc. Symp. 46, 541–561 (2005).
5. Instituto Brasileiro de Geografia e Estatística - IBGE. Produção da Pecuária Municipal. IBGE 42, (2014).
6. Animals, A. Production of Aquatic Animals -Fishes, Anim. Sci. 194, 280–281 (1995).
7. García-Berthou, E. The characteristics of invasive fishes: What has been learned so far? J. Fish Biol. 71, 33–55 (2007).
8. Jacometo, C. B. et al. Variabilidade genética em tambaquis (Teleostei: Characidae) de diferentes regiões do Brasil. Pesqui. Agropecu. Bras. 45, 481–487 (2010).
9. Santos, C. H. A., Santana, G. X., Sá Leitão, C. S., Paula-Silva, M. N. & Almeida-Val, V. M. F. Loss of genetic diversity in farmed populations of Colossoma macropomum estimated by microsatellites. Anim. Genet. 47, 373–376 (2016).
10. Prado-Lima, M. et al. Transcriptomic Characterization of Tambaqui (Colossoma macropomum, Cuvier, 1818) Exposed to Three Climate Change Scenarios. PLoS One 11, e0152366 (2016).
11. Martínez, J. G. et al. SNPs markers for the heavily overfished tambaqui Colossoma macropomum, a Neotropical fish, using next-generation sequencing-based de novo genotyping. Conserv. Genet. Resour. online, 1–5 (2016).
12. Wu, Y.-P., Xie, J.-F., He, Q.-S. & Xie, J.-L. The complete mitochondrial genome sequence of Colossoma macropomum (Characiformes: Serrasalmidae). Mitochondrial DNA 1–2 (2015). doi:10.3109/19401736.2014.1003853
13. Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S. & Bull, J. K. Molecular markers in a commercial breeding program. in Crop Science 47, S–154 (Crop Science Society of America, 2007).
14. Heffner, E. L., Jannink, J.-L. & Sorrells, M. E. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome 4, 65–75 (2011).

15. Sunnucks, P. Efficient genetic markers for population biology. Trends in Ecology and Evolution 15, 199–203 (2000).
16. Chong, Z., Ruan, J. & Wu, C.-I. Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics 28, 2732–7 (2012).
17. 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, 275 (2002).
18. 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, 275 (2002).
19. Liu, S. et al. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12, 53 (2011).
20. Liu, S. et al. Development of the catfish 250K SNP array for genome-wide association studies. BMC Res. Notes 7, 135 (2014).
21. Davey, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev Genet 12, 499–510 (2011).
22. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248 (2007).
23. Van Tassell, C. P. et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 247–252 (2008).
24. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379 (2011).
25. Starks, H. A., Clemento, A. J. & Garza, J. C. Discovery and characterization of single nucleotide polymorphisms in coho salmon, Oncorhynchus kisutch. Mol. Ecol. Resour. 16, 277–287 (2016).
26. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7, e32253 (2012).
27. Annicchiarico, P. et al. Assessment of Cultivar Distinctness in Alfalfa: A Comparison of Genotyping-by-Sequencing, Simple-Sequence Repeat Marker, and Morphophysiological Observations. Plant Genome 9, 0 (2016).
28. Pértille, F. et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Sci. Rep. 6, 26929 (2016).
29. Boutet, G. et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics 17, 121 (2016).
30. Zeng, Z.-B. Precision Mapping of Quantitative Trait Loci. 1468, 1457–1468 (1994).
31. Carlson, B. M., Onusko, S. W. & Gross, J. B. A high-density linkage map for Astyanax mexicanus using genotyping-by-sequencing technology. G3 (Bethesda). 5, 241–51 (2015).
32. Bodénès, C. High-density linkage mapping and distribution of segregation distortion regions in the oak genome. DNA Res. In press, 1–10 (2015).
33. Liu, S. et al. High-density interspecific genetic linkage mapping provides insights into genomic incompatibility between channel catfish and blue catfish. Anim. Genet. 47, 81–90 (2016).
34. Akagi, T., Hanada, T., Yaegaki, H., Gradziel, T. M. & Tao, R. Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. DNA Res. 23, 271–282 (2016).
35. Schunter, C. et al. Molecular signatures of transgenerational response to ocean acidification in a species of reef fish. Nat. Clim. Chang. 1–5 (2016). doi:10.1038/nclimate3087
36. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: Practical considerations. Nat. Rev. Genet. 3, 391–397 (2002).
37. Moav, R., Finkel, A. & Wohlfarth, G. Variability of intermuscular bones, vertebrae, ribs, dorsal fin rays and skeletal disorders in the common carp. Theor. Appl. Genet. 46, 33–43 (1975).
38. Sahu, B. B. et al. Record of Skeletal System and Pin Bones in Table Size Hilsa Tenualosa ilisha (Hamilton, 1822). World J. Fish Mar. Sci. 6, 241–244 (2014).
39. Li, L. et al. Comparative analysis of intermuscular bones in fish of different ploidies. Sci. China Life Sci. 56, 341–350 (2013).
40. Danos, N. & Ward, A. B. The homology and origins of intermuscular bones in fishes: Phylogenetic or biomechanical determinants? Biol. J. Linn. Soc. 106, 607–622 (2012).
41. NURSALL, J. R. The lateral musculature and the swimming of fish. Proc. Zool. Soc. London 126, 127–144 (1956).
42. Sfakiotakis, M., Lane, D. M. & Davies, J. B. C. Review of fish swimming modes for aquatic locomotion. IEEE J. Ocean. Eng. 24, 237–252 (1999).
43. STACHOWICZ, J. J. Mutualism, Facilitation, and the Structure of Ecological Communities. Bioscience 51, 235 (2001).
44. Perazza, C. A., de Menezes, J. T. B., Ferraz, J. B. S. & Hilsdorf, A. W. S. Lack of intermuscular bones in specimens of Colossoma macropomum: An unusual phenotype to be incorporated into genetic improvement programs. Aquaculture (2016). doi:10.1016/j.aquaculture.2016.05.014

45. Wringe, B. F. et al. Growth-related quantitative trait loci in domestic and wild rainbow trout (Oncorhynchus mykiss). BMC Genet. 11, 63 (2010).
46. Gutierrez, A. P. et al. Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). PLoS One 10, e0119730 (2015).

2. LARGE-SCALE SNP DISCOVERY AND CONSTRUCTION OF A HIGH-DENSITY GENETIC MAP OF C*olossoma macropomum* THROUGH GENOTYPING-BY-SEQUENCING

ABSTRACT

Colossoma macropomum, or tambaqui, is the largest native Characiform species found in the Amazon and Orinoco river basins, yet few resources for genetic studies and the genetic improvement of tambaqui exist. In this study, we identified a large number of single-nucleotide polymorphisms (SNPs) for tambaqui and constructed a high-resolution genetic linkage map from a full-sib family of 124 individuals and their parents using the genotyping by sequencing method. In all, 68,584 SNPs were initially identified using minimum minor allele frequency (MAF) of 5%. Filtering parameters were used to select high-quality markers for linkage analysis. We selected 7,734 SNPs for linkage mapping, resulting in 27 linkage groups with a minimum logarithm of odds (LOD) of 8 and maximum recombination fraction of 0.35. The final genetic map contains 7,192 successfully mapped markers that span a total of 2,811 cM, with an average marker interval of 0.39 cM. Comparative genomic analysis between tambaqui and zebrafish revealed variable levels of genomic conservation across the 27 linkage groups which allowed for functional SNP annotations. The large-scale SNP discovery obtained here, allowed us to build a high-density linkage map in tambaqui, which will be useful to enhance genetic studies that can be applied in breeding programs.

Keywords: SNP; Linkage map; GBS; Tambaqui; *PstI*; Restriction enzyme

2.1. Introduction

Tambaqui (*Colossoma macropomum*) is the most important native aquaculture species in Brazil. In 2014, production reached 139,209 tons [1]. This species has been broadly farmed in different Amazon boundary countries and has been introduced to other Latin American countries [2], as it is well-suited to aquaculture farming, accepts artificial feed, grows rapidly, and is accepted by consumer markets [3]. However, genetic studies of tambaqui are limited to genetic diversity surveys [4,5,6], transcriptome analysis under specific conditions [7], some SNPs markers [8], and mitochondrial genome sequencing [9].

Genetic linkage maps are essential resources for genomic and genetic research. They provide frameworks for understanding genome structure and function. Linkage maps are essential for quantitative trait locus (QTL) identification and mapping. QTL mapping is an important way to identify genes that are related to trait variations within and between populations or species, allowing for implementation of genetic and breeding programs [10]. Genetic maps can be used to facilitate genome assembly corrections, anchoring scaffolds onto linkage groups to build chromosomal assembly [11]. With interspecific crosses, high-density genetic mapping can provide a genome-scan of segregation distortion within the genome [12] and can investigate genomic incompatibilities between species at the genome level [13]. Genetic studies such as genetic linkage mapping [14], QTL mapping [15], populational genetic analysis [16], and genome-wide association studies [17] require a large number of reliable molecular markers across the genome.

Single nucleotide polymorphisms (SNPs) are the marker of choice for genetic studies in various organisms. SNPs are the most abundant molecular markers in any vertebrate genome, with a SNP present in 100–500 bp on average [18]. SNPs are mostly bi-allelic, making them amenable for high-throughput genotyping using SNP arrays [19,20].

Over the last decade, with the development of next-generation sequencing technologies [21,22,23], genome-scale SNP markers can now be efficiently and cost-effectively identified in any organism for which prior genomic information does not yet exist [24]. Genotype-by-sequencing (GBS) is one next-generation sequencing technique, and is based on the reduction of genome complexity using restriction enzymes [23]. GBS is characterized as a simple, quick, specific, reproducible technique [23], and has been extensively used to identify a large number of SNPs in various model and non-model species [25,26,27,28].

In the present study, we applied GBS to identify large-scale SNPs and construct a high-density genetic linkage map for tambaqui, using the Illumina HiSeq 2500 platform. In addition, we conducted syntenic relationship and functional annotation analyses by aligning tambaqui against the zebrafish (*Danio rerio*) genome. This study provides large-scale SNP markers and high-density linkage maps in tambaqui, which can be a useful resource for facilitating the tambaqui physical map construction, genome assembly, and QTL mapping to enhance genetic studies and breeding programs.

2.2. Materials and Methods

2.2.1. Ethical statement

All experimental protocols employed in the present study that relate to animal experimentation were performed in accordance with Brazilian Directive for the Care and Use of Animals in Teaching or Scientific Research Activities, resolution number 30/2016 approved by the National Animal Experimentation Control Council to ensure compliance with international guidelines for animal welfare.

The samples were acquired from a commercial hatchery and were not subjected to any experimental manipulation or euthanasia. No specific permits were required for the work described here.

2.2.2. Fish materials and DNA extraction

The *Colossoma macropomum* subjects used in this study were collected at a fisheries farm located in the Rondônia state, in the northwestern region of Brazil. A full-sib F1 family of 124 offspring was created by crossing a wild female with a wild male. When the subject's mean body weight reached ~6 g (mean body length of ~5.5 cM), tail fin clips were collected from each progeny and the two parents, and preserved in 90% ethanol. DNA extraction followed these steps: proteinase K digestion (Promega), DNA precipitation in absolute ethanol, washing in 70% ethanol, and resuspension in ultrapure water. DNA concentration was quantified using a Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA, USA) and Nanodrop®2000c spectrophotometer. DNA integrity was checked in 1% agarose gel. All DNA samples were stored at −20°C prior to sequencing library preparation.

2.2.3. Construction of GBS sequencing libraries

To select an enzyme that uniformly distributed cutting sites across the tambaqui genome, we performed an *in-silico* DNA cleavage on the zebrafish genome with *PstI* and *SbfI* in R using the subsequent Bioconductor [49] packages: Biostrings, BSgenome.Drerio.UCSC.danRer7, plyr, ggplot2, reshape2 and scales (https://github.com/). Additionally, we performed in vitro genomic cleavage of tambaqui DNA samples with the *PstI* and *SbfI* enzymes, according to the manufacturer's protocol (New England BioLabs®).

GBS library construction and sequencing were conducted at the Animal Biotechnology Laboratory at the University of São Paulo (Piracicaba, Brazil), using the protocol described by Elshire *et al.* (2011) [23] with modifications. In brief, 100 ng of high-quality DNA were cleaved with 0.2 μL *PstI* (10U/uL) at 37 °C for 2 hours. After digestion, the restriction enzyme was deactivated at 85 °C for 20 seconds, and the samples were dehydrated. To perform the ligation reaction, all samples were rehydrated in 6 μL of adapter solutions and incubated at 22 °C for 2 hours in binding mix with T4 DNA ligase (New England BioLabs®). For post-ligation reactions, 10 μl from each of the 126 samples (124 progenies + 2 parents) were aliquoted and pooled with 32 samples per group, followed by polymerase chain reaction (PCR) cleanup using the QIAquick PCR Purification Kit® (Quiagen), resulting in four "pre-PCR" GBS libraries. Within each library, PCR amplification was conducted using specific primers for sequencing, using the Illumina platform. PCR purification was performed using the Agencourt AMPure XP PCR purification kit®. PCR products were quantified by quantitative PCR, using the KAPA Library Quantification Kit (KAPA Biosystems). For each library, 2 pools of 32 samples were mixed and diluted to 16 pM. For each library, 64 barcoded samples were pooled. The libraries were clustered using TruSeq SR Cluster Kit v3-cBot-HS on the cBOT (Illumina) equipment. The libraries were sequenced using Illumina TruSeq SBS Kit v3-HS on the Illumina HiSeq2500 sequencer (Illumina, San Diego) on two lanes of single-end reads with a read length of 100 bp.

2.2.4. Read processing and SNP discovery

For all samples, quality trimming was performed with SeqyClean tool v. 1.9.10 (https://bitbucket.org/izhbannikov/seqyclean/) using Phred quality score ≥24 and fragment size ≥50. A contaminant database was provided to the program to remove vector, adapter, and other sequence contaminations. The sequence processing was performed using the UNEAK[39] (Universal Network Enabled Analysis Kit) pipeline with default parameters. UNEAK separates all reads that have an exact match to a barcode plus the subsequent five nucleotides that are expected to remain from a PstI cut-site (i.e., 5´… CTGCA´G…3´) but no missing data in the first 64 bp subsequently the barcode. Identical reads were clustered into tags; rare or singleton tags represented by fewer than five reads were excluded to reduce possible sequencing errors. All tags were then aligned pair wisely, and 1-bp mismatches were detected as potential SNPs. To reduce false SNP calls, the UNEAK pipeline applied an error tolerance filter of 0.03.

2.2.5. Linkage map construction

T The UNEAK output files were imported to Tassel v.3.0[52] in order to apply post-UNEAK filtering. Genotype data were filtered for both parent and progeny samples. Only markers that had no missing genotypes for

both parents were retained. Were applied a sample call rate and marker call rate of >80% (i.e., at least 80% SNPs had genotypes called in each sample and a SNP was called in at least 80% of the samples). In addition, only markers that were heterozygous in at least one of the two parents (i.e., AA x AB, AB x AA, or AB x AB) were used for linkage mapping analysis. Next, markers that significantly deviated from Mendelian inheritance, as determined by a chi-squared test ($P < 0.001$), were excluded. The genetic maps were constructed using the pseudo-testcross strategy. The program R/OneMap was used to assign markers to linkage groups that had a minimum logarithm of odds (LOD) score of 8 and a maximum recombination fraction of 0.35. The program JoinMap4 was then used to construct the map for each linkage group, using the regression mapping algorithm and Kosambi mapping function.

2.2.6. Analysis of syntenic relationships

The SNPs in the tambaqui linkage maps were used to analyze the syntenic relationships with zebrafish (Danio rerio). The sequence reads that harbored the mapped SNPs were extracted and aligned with the zebrafish genome (Danio rerio, Zv9) using Bowtie2 v.2.2.5. To maximize the number of alignments, the parameters –D, –R, –N, –L, and –i were optimized manually (-D 100 -R 7 -N 1 -L 20 -i S,1,0.01). Sequences with multiple alignments were removed. The Circos [53] software was used to plot the relationship between the zebrafish chromosomes and tambaqui linkage groups.

2.2.1. Functional annotation

The SNPs that successfully aligned with the zebrafish genome (Danio rerio, Zv9) were annotated using the Variant Effect Predictor (VEP) tool v.71 [54]. The annotations included a range of variant types such as intronic, downstream gene, upstream gene, synonymous, missense, intergenic, splice region, and non-coding transcript.

2.3. Results

2.3.1. Enzyme selection

Based on *in silico* and *in vitro* genomic fragmentation, we tested the enzymes *PstI* and *SbfI* to generate expected number of reads required to obtain ~7X of sequencing coverage[33]. Each enzyme generated a different distribution of fragment lengths across the entire genome (APPENDIX A). The enzyme *PstI* yielded a larger number of fragments ranging between 200 bp and 500 bp (see APPENDIX A), providing suitable sizes for GBS to be clustered on cBOT (Illumina) in bridge amplification.

2.3.2. Sequencing

As summarized in Table 2.1, we generated over 352 million single-end reads with length of 100 bp from the 126 samples. After read trimming, ~285 million quality reads (81%) were obtained, with over 27 Giga bases,

equivalent to >8X genome coverage. An average 3.2 million quality reads were obtained from parents and over 2.2 million quality reads were obtained from each progeny (Table 2.1). The vast majority of samples had genome sequencing depth > 8X, with 15 progenies having fewer than one million reads (~4.5X) (APPENDIX B).

Table 2.1. Summary of the GBS reads before and after of trimmed.

	Reads	**Total bases** (bp)	**Trimmed Reads**	**Trimmed bases**	**Sequencing depth**
Total number	352,425,578	35,242,557,800	285,194,978	27,663,912,870	8.04X
Average per parent	3,870,142	387,014,200	3,264,068	316,614,596	8.60X
Average per offspring	2,779,720	277,972,000	2,247,313	217,989,361	8.03X

2.3.3. SNP discovery

A total of 81,222 pairwise alignments were obtained with UNEAK pipeline. After filtering with default parameters, 68,584 putative SNPs were identified with minor allele frequency (MAF) higher than 0.05 (Table 2.2). SNPs were classified into transitions (Ti) and transversions (Tv) based on nucleotide substitution. The number of A/G transitions was about two times greater than C/T transitions; the numbers of A/C and A/T transversions were relatively higher than C/G and G/T transversions. The transitions are the most common type of nucleotide substitutions, 64% of the base changes were transitions and 36% were transversions, with an observed transition to transversion ratio (Ti/Tv) of 1.8:1.

Table 2.2. Identification of SNPs from the tambaqui (*Colossoma macropomum*).

SNP	**Nucleotide substitution**	**SNP number**	**Proportion**
Transitions	A-G	28,744	0.42
	C-T	15,288	0.22
	-	**44,032**	**0.64**
Transversions	A-C	9,767	0.14
	A-T	7,963	0.12
	C-G	3,483	0.05
	G-T	3,339	0.05
	-	**24,552**	**0.36**
Total		**68,584**	**1.00**

2.3.4. SNP filtering

A set of 10,288 high quality SNP markers with sample call rate >80% and heterozygous for at least one parent were retained for further analysis. After removal of markers with significant segregation distortion, a total of 7,734 SNPs were retained for genetic map construction. Of these SNPs, 3,641 were heterozygous only in female, 2,565 were heterozygous only in male, and the remaining 1,528 were heterozygous in both parents. A total of 118 samples with SNP call rate >80% were retained for genetic map construction.

2.3.5. Construction of high-density linkage map

A total of 7,192 SNPs were mapped to 27 linkage groups (LGs), consistent with the haploid chromosome number (n=27) of tambaqui [29]. The genetic map spanned a total of 2,810.9 cM, with an average marker-interval of 0.39 cM (Figure 2.1).

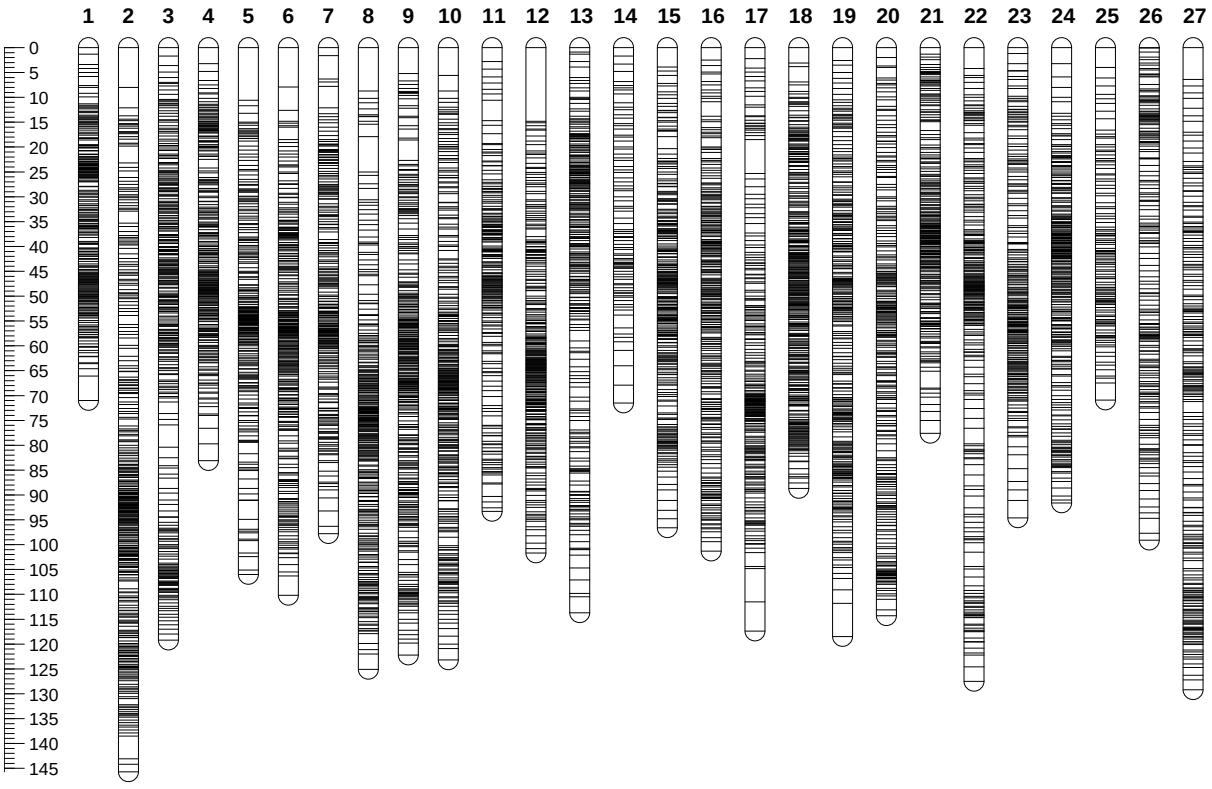


Figure 2.1. Graphical presentation of genetic linkage map of tambaqui. The 27 linkage groups (LGs) are represented by vertical bars while horizontal lines represent markers. Genetic distances between adjacent markers are shown on side ruler in centiMorgan (cM).

On average, each LG contained 266 markers that spanned an average length of 104 cM, with an average marker interval of 0.41 cM. The number of mapped markers per LG varied from 86 markers on LG14 to 362 markers on LG2 and LG9. The smallest LG was LG25, which contained 120 markers spanning a length of 70.94 cM. The largest LG was LG2 with 362 markers and a length of 145.66 cM. The maximum gap size in each LG ranged from 2.08 cM on LG20 to 14.79 cM on LG12, with an average of 5.27 cM (Table 2.3).

Table 2.3. Summary of genetic linkage map of tambaqui.

Linkage group	No. of mapped markers	Genetic size (cM)	Average marker interval	Max Gap (cM)
LG1	269	71.0	0.26	4.85
LG2	362	145.7	0.40	8.04
LG3	316	119.2	0.38	4.54
LG4	287	83.1	0.29	3.39
LG5	243	106.0	0.44	10.62
LG6	316	110.2	0.35	7.89
LG7	242	97.8	0.40	4.72
LG8	328	125.1	0.38	8.67
LG9	362	122.2	0.34	5.19
LG10	331	123.2	0.37	5.59
LG11	210	93.3	0.44	4.11
LG12	315	101.7	0.32	14.79
LG13	261	113.7	0.44	3.12
LG14	86	71.5	0.83	3.85
LG15	274	96.6	0.35	3.86
LG16	279	101.3	0.36	2.85
LG17	238	117.4	0.49	6.81
LG18	361	88.8	0.25	3.12
LG19	305	118.5	0.39	6.67
LG20	258	114.4	0.44	2.08
LG21	233	77.6	0.33	3.31
LG22	271	127.5	0.47	4.22
LG23	238	94.6	0.40	3.48
LG24	269	91.6	0.34	3.18
LG25	120	70.9	0.59	4.00
LG26	193	99.1	0.51	3.08
LG27	225	129.2	0.57	6.42
Total	**7192**	**2810.9**	**0.39**	**5.27**

The distribution of markers across each linkage group was assessed using the sliding window approach. The number of markers within a window was counted by a sliding window of 10 cM with a step size of 1 cM. The density value for each window was calculated by dividing the total of markers within a window by the window length. As shown in Figure 2.2, the SNP markers are evenly distributed across the 27 LGs. The linkage groups LG2, LG4, LG6, LG8, LG9, LG12 and LG22 had windows with high density markers (>8 markers per cM), most of which were clustered at the same genetic positions. The LG14 had more windows with low density of markers. The LG12 had the window with the highest density (10.4 markers per cM) and also the window with the lowest density (0 marker per cM). In general, the regions with the high density of markers were located near the centromeres (Figure 2.2).

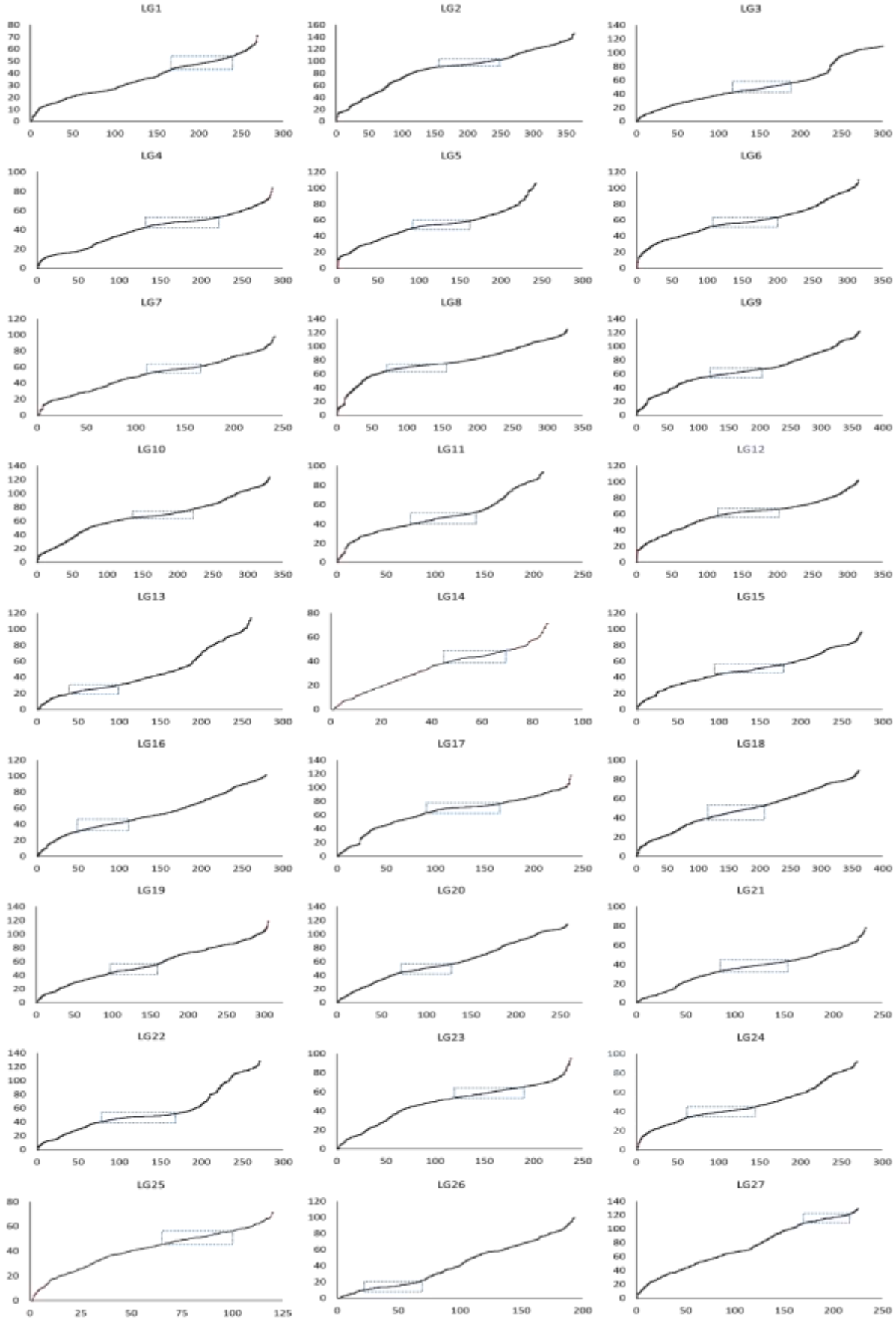


Figure 2.2. Patterns of marker distribution along the tambaqui linkage groups. Dotted box indicates region with higher marker density. The red lines indicate gaps without markers.

2.3.6. Comparative genomic analysis

Out of the 7,192 SNP markers mapped to the linkage map, 1,237 marker containing sequences were successfully aligned against zebrafish genome (Figure 2.3). The synteny analysis between tambaqui and zebrafish showed variable levels of genomic conservation across the 27 linkage groups. Most linkage groups of tambaqui showed relationship with homologous chromosomes in zebrafish. Through comparative analysis, a high level of genomic conservation was found between tambaqui and zebrafish for most linkage groups. For instance, more than 30% of SNPs mapped in LGs 17, 3, 19, 14, 10, 18, 1, 11, 8, 25, 21, 6, 20 and 15 of tambaqui were aligned on zebrafish chromosomes 21, 15, 23, 7, 8, 3, 9, 5, 13, 1, 10, 6, 17 and 18, respectively, suggesting orthologous chromosomal relationships. The zebrafish chromosome 7 corresponded to tambaqui linkage groups 14 and 16, while zebrafish chromosome 5 corresponded to linkage groups 11 and 23 in tambaqui. The linkage group 27 had a similar level of consensus between homologous chromosomes 11 and 22 in zebrafish. However, the linkages groups 2, 4, 5, 13, 22, 24 and 26 appear scattered among several chromosomes, suggesting the large-scale chromosomal rearrangements between species.

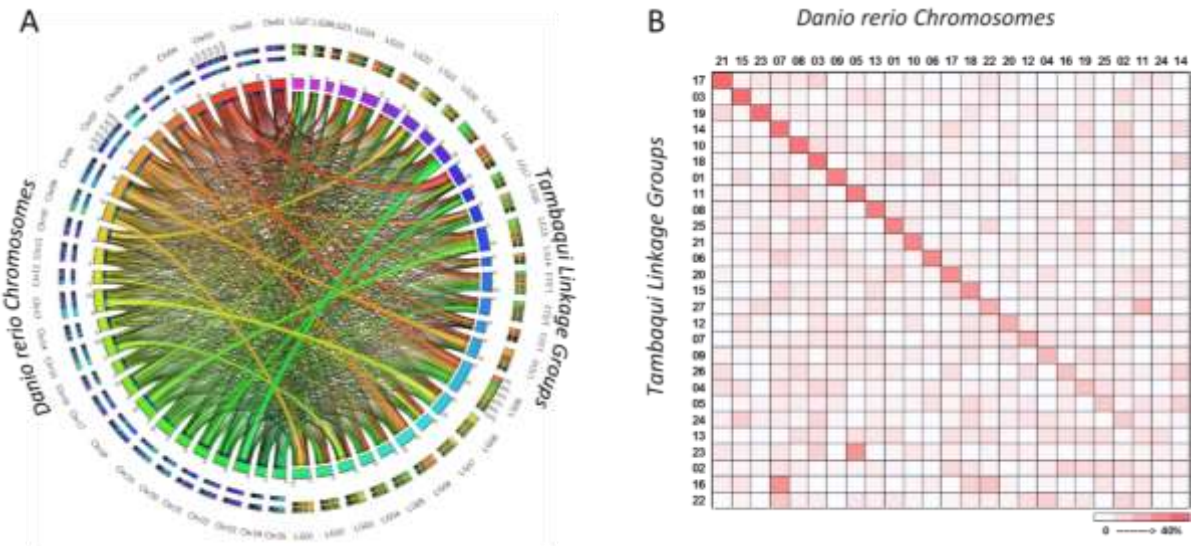


Figure 2.3. Comparisons between tambaqui linkage map and zebrafish genome. Syntenic links between tambaqui linkage map and zebrafish genome using Circos (A). Plotting the percentage of SNPs shared between tambaqui linkage groups and zebrafish chromosomes (B). In synteny comparison, the connecting links are color coded according to the zebrafish chromosomes arc to their tambaqui linkage-group corresponding.

2.3.7. Functional annotation

The SNPs for each linkage group were annotated against the genes from zebrafish (ENSEMBL release 84). This allowed us to evaluate the potential use of our genetic map with respect to possible important traits in breeding programs. Approximately 60% of the SNPs evaluated were annotated to be intronic, downstream or upstream of gene regions. These SNPs could, thus have direct effect or be associated with potential performance traits of interest (Table 2.4). From the 64% of exonic SNP detected, 58% were classified as synonymous variants with no effects on protein function and 39% were classified as missense mutations resulting in codons that coded for a different amino

acid. 2% of the SNPs resulted in stop lost codon, and 1% results in stop gain codon, which can lead to a protein truncation.

Table 2.4. Annotation of SNPs for each linkage group of tambaqui against the genes from zebrafish genome (ENSEMBL release 84).

		Variants Consequences (%)								
LG	**VP**	**IN**	**DW**	**UP**	**SY**	**MI**	**IT**	**SP**	**NC**	**OT**
1	52	25	15	15	12	12	8	4	3	6
2	46	34	7	10	3	7	14	11	12	2
3	46	34	8	10	24	10	7	2	1	4
4	41	37	10	7	5	6	16	-	15	4
5	36	34	14	7	12	2	12	6	7	6
6	56	49	6	11	7	7	13	1	5	1
7	46	30	16	13	4	10	16	9	-	2
8	61	40	14	5	13	8	7	4	4	5
9	79	34	17	8	10	14	8	3	4	2
10	55	29	6	15	8	10	14	5	9	4
11	33	39	14	13	19	3	8	-	2	2
12	49	44	11	5	6	3	19	-	10	2
13	40	49	22	7	3	7	7	-	3	2
14	24	39	12	2	22	10	6	-	8	1
15	69	26	13	10	18	8	16	-	3	6
16	46	26	13	13	18	5	6	-	5	14
17	33	38	16	7	7	8	7	5	10	2
18	58	35	22	15	9	9	5	2	2	1
19	53	29	13	14	17	8	8	3	2	6
20	38	45	6	15	11	5	10	4	1	3
21	47	30	14	10	18	8	8	5	3	4
22	50	47	13	14	7	3	7	1	3	5
23	51	47	7	10	10	5	6	8	5	2
24	49	39	18	12	7	6	14	-	3	1
25	20	34	16	16	3	11	8	-	12	-
26	25	52	12	12	-	-	19	-	-	5
27	34	32	12	14	20	2	12	3	2	3
All	1237	36	13	11	11	7	10	3	4	5

LG: linkage group; VP: variants processed; IN: intron variant; DW: downstream gene variant; UP: upstream gene variant; SY: synonymous variant; MI: missense variant IT: intergenic variant; SP: splice region variant; NC: non coding transcript variant; OT: others variants.

2.3.8. Allele frequency and coverage evaluation

To evaluate the GBS approach capability, as it is related to accurate determination of allelic frequencies, we compared the allelic frequency obtained in our genotyping with the expected for a diploid cross. We also analyzed allele frequency calls into population using different SNP coverage thresholds (Figure 2.4). The SNPs were placed into six groups depending on the coverage (5-10X, 10-20X, 20-40X, 40-80X, 80-160X, ≥160X). With the coverage ≥ 160X we observed three peaks of the allele frequency distribution for the SNPs. The first and third peak represents the crosses of AAxAa, with allele frequencies of 0.25 and 0.75, respectively, with 1:1 segregation of AA and Aa genotypes. The second peak represents the crosses of AaxAa, with allele frequency of 0.5 and 1:2:1 segregation of AA, Aa and aa genotypes. Lower coverages resulted in allelic frequency deviations from the expected.

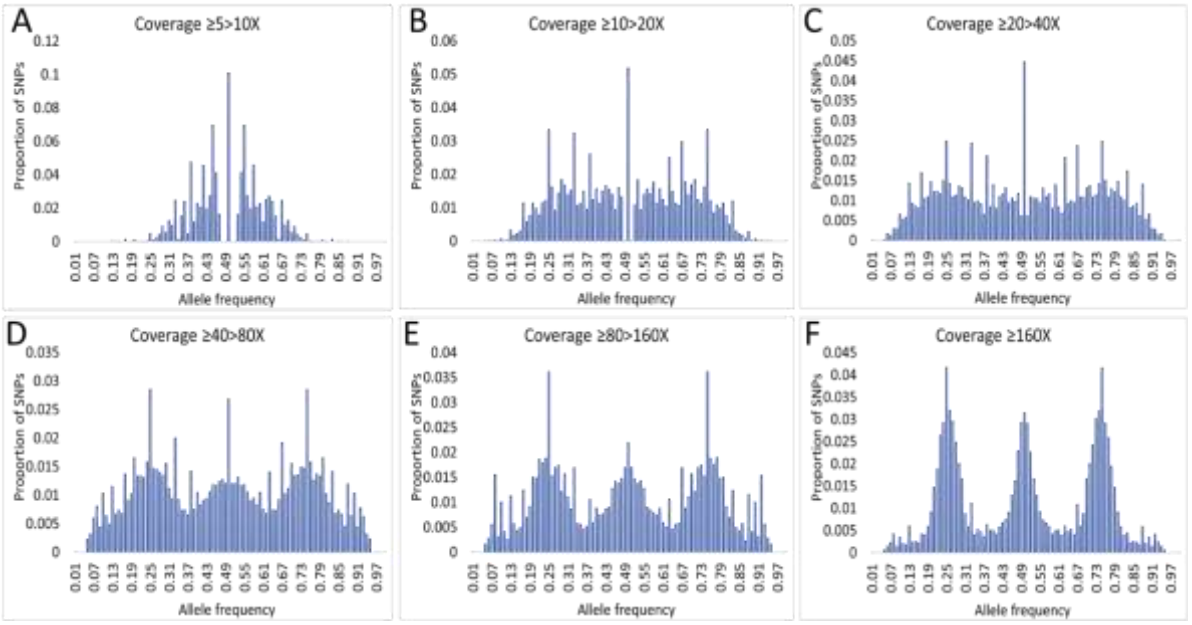


Figure 2.4. Effect of coverage depth on allele frequency of the tambaqui (*Colossoma macropomum*) genotyped with GBS.

2.4. Discussion

Development of molecular markers panels and genetic linkage maps is essential for genetic improvement programs in aquaculture. As already mentioned, despite of economic and ecological importance of tambaqui, limited genomic resources are available. In this study, we report the first genome-scale SNP discovery and high-density genetic map construction in tambaqui. Sequencing a F1 family of 124 offspring and the two parents allowed the identification and genotyping of 68,584 SNP markers in an efficient and cost-effective way. The linkage mapping analysis resulted in 27 linkage groups corresponding to the haploid chromosome number (n = 27) of tambaqui [29]. The genetic map contains a total of 7,192 markers that spanned a total of 2,810.9 cM, with an average marker interval of 0.39 cM. SNP discovery and genetic map construction using GBS approach have been conducted in a number of aquaculture species including asian seabass (*Lates calcarifer*) [30], common pandora (*Pagellus erythrinus*) [31], sablefish (*Anoplopoma fimbria*) [32], scallop (*Chlamys farreri*) [33], oysters (*Crassostrea gigas* × *Crassostrea angulata*) [34] and small abalone

(*Haliotis diversicolor*) [14,35,34,30,31,13,24]. This tambaqui genetic map will be valuable for the study of genome research and applications in genetic enhancement such as identification of interspecific hybrids [36] and implementation of genome-wide association studies [37].

The successful application of GBS approach depends on the choice of an effective restriction enzyme. This determines the number of genomic fragments produced by the complexity reduction for the species under study to identify genome-wide sequence polymorphisms [38]. The *in silico* fragmentation showed the large difference in numbers of restriction sites between *PstI* and *SbfI*. The *PstI* enzyme was more suitable for fragmentation of the genome tested. This enzyme has also been successfully used in prawn kuruma (*Marsupenaeus japonicas)* for a high-resolution genetic linkage map construction and quantitative trait locus (QTL) mapping [35].

The per-base quality of sequencing reads has great impacts on the accuracy of marker detection and genotype calling[39]. In our analysis, we trimmed off 19% of all reads with low-quality scores. This resulted in 28% larger numbers of identified SNPs than that from non-trimed reads (unpublished data). Low quality reads result in more complicated networks. It reduces read counts below a specified error tolerance rate. UNEAK Pipeline considers this networks as a result of sequencing errors and exclude them from genotyping [40], reducing SNP coverage. In our analysis, the SNP coverage had direct effect on the allelic frequency (Figure 2.4). The results showed that higher SNP coverage results in an allele frequency distribution that is consistent with a diploid expected crossing. For instance, the coverage of ≥160X provided smaller levels of missing data and a higher percentage of called genotypes (Figure 2.4).

The GBS method was efficient for the discovery of tens of thousands of SNPs in the genome without a reference sequence. However, genetic maps require high-quality genotypes of markers from a certain numbers of mapping samples. In this study, we used a high criteria by setting 80% SNP call rate and 80% sample call rate, i.e., at least 80 % SNPs had genotypes called in each sample and a SNP was called in at least 80% of the samples. Although the filtering steps, dropped about 85% of the 68,584 SNPs for linkage mapping, the remaining markers and samples with high quality scores ensured the construction of genetic map with a high level of accuracy. Similar numbers of makers were also reported in other studies using GBS strategies [41,42,34,43], which generally show high levels of missing genotypes due to the relatively low sequencing depth.

The GBS approach generated marker-containing sequences, allowing for analysis with other closely-related model species based on sequence homology (Figure 2.3). A total of 1,237 (17%) marker containing sequences of tambaqui were successfully aligned with the zebrafish genome. This percentage is higher than that reported by Calrson et al. (2015). [11] (14.2%) where GBS marker-containing sequences from mexican tetra (*Astyanax mexicanus*) were aligned with zebrafish genome. However, this could be a consequence of genetic distance of the species used or a consequence of manually maximizing in Bowtie2 alignment parameters, since this alignment, using the default parameter, falls to 9%. A variable degree of synteny is observed between tambaqui linkage groups and zebrafish chromosomes with some linkages groups showing homology with several chromosomes of zebrafish. The majority of the linkage groups in tambaqui have one-to-one orthologous relationships with zebrafish chromosomes. Notably, zebrafish chromosomes 5 and 7 have two-to-one syntenic relationships with tambaqui linkage groups. The use of zebrafish to perform synteny and collinearity analyses may provide framework for mapping candidate genes responsible for the traits related to phenotypic divergence in tambaqui. This also enables transferring of genomic information between zebrafish and others species, such as mexican tetra [11], gudgeons [44], common carp [45], rainbow trout [46], and channel catfish [47].

The syntenic relationships and functional annotation of zebrafish allowed annotation of SNPs from the tambaqui linkage map. This allowed identification of 1,237 variants, from which 36% were annotated in genomic regions. Candidate genes for important traits were identified in tambaqui such as *IGFBP*, a hypoxia-inducible gene that acts in regulating embryonic growth and development under hypoxic stress [48]. We also found variants in genes of medical interest such as *IGF1RA* (insulin-like growth factor 1a receptor), a gene that has been studied in animals with relatively slow progression for insulin resistance in, which would better help to understand genes or metabolic situations that may accelerate the progress of diabetes and offer new therapeutic targets [49]. In addition, variants of upstream, downstream and intronic regions of *IGHV* gene were annotated. Interesting, this gene have been reported as the most important indicator on chronic lymphocytic leukaemia prognostic [50]. Although the tambaqui genome sequences are still not available, through comparative genome analysis, the identification of large number of markers in genic regions enabled the potential use of our linkage map in breeding programs in tambaqui.

In this study, we showed that GBS is an effective approach for SNP discovery and development of high-density genetic linkage maps in tambaqui. We, for the first time, reported the identification of genome-scale SNP markers and construction of the first high-density genetic map. Comparative genomic analysis with zebrafish revealed variable levels of homologous relationships with zebrafish chromosomes for all tambaqui linkage groups. In addition, large number of markers in genic regions were annotated, and genes with potential functions for performance traits were identified. The SNPs and genetic map reported in this work should be valuable tools for genetic studies and aquaculture improvement in tambaqui.

References

1. Instituto Brasileiro de Geografia e Estatística - IBGE. De Geografia E Estatística - Ibge. Produção da Pecuária Munic. 42, 1 – 36 (2014).
2. Kapetsky, J. M. & Nath, S. S. A strategic assessment of the potential for freshwater fish farming in Latin America: Annex 2, Water Temperature Model. COPESCAL Tech. Pap. 10, 128p (1997).
3. Cuvier, C. M., Lima, C. D. E. S., Antonio, M., Bomfim, D. & De, J. C. Crude protein levels in the diets of tambaqui, Colossoma macropomum (Cuvier, 1818), fingerlings. Rev. Caatinga 2125, 183–190 (2016).
4. García-Berthou, E. The characteristics of invasive fishes: What has been learned so far? J. Fish Biol. 71, 33–55 (2007).
5. Jacometo, C. B. et al. Variabilidade genética em tambaquis (Teleostei: Characidae) de diferentes regiões do Brasil. Pesqui. Agropecu. Bras. 45, 481–487 (2010).
6. Santos, C. H. A., Santana, G. X., Sá Leitão, C. S., Paula-Silva, M. N. & Almeida-Val, V. M. F. Loss of genetic diversity in farmed populations of Colossoma macropomum estimated by microsatellites. Anim. Genet. 47, 373–376 (2016).
7. Prado-Lima, M. et al. Transcriptomic Characterization of Tambaqui (Colossoma macropomum, Cuvier, 1818) Exposed to Three Climate Change Scenarios. PLoS One 11, e0152366 (2016).
8. Martínez, J. G. et al. SNPs markers for the heavily overfished tambaqui Colossoma macropomum, a Neotropical fish, using next-generation sequencing-based de novo genotyping. Conserv. Genet. Resour. online, 1–5 (2016).
9. Wu, Y.-P., Xie, J.-F., He, Q.-S. & Xie, J.-L. The complete mitochondrial genome sequence of Colossoma macropomum (Characiformes: Serrasalmidae). Mitochondrial DNA 1–2 (2015). doi:10.3109/19401736.2014.1003853

10. Zeng, Z.-B. Precision Mapping of Quantitative Trait Loci. 1468, 1457–1468 (1994).
11. Carlson, B. M., Onusko, S. W. & Gross, J. B. A high-density linkage map for Astyanax mexicanus using genotyping-by-sequencing technology. G3 (Bethesda). 5, 241–51 (2015).
12. Bodénès, C. High-density linkage mapping and distribution of segregation distortion regions in the oak genome. DNA Res. In press, 1–10 (2015).
13. Liu, S. et al. High-density interspecific genetic linkage mapping provides insights into genomic incompatibility between channel catfish and blue catfish. Anim. Genet. 47, 81–90 (2016).
14. Ren, P. et al. Genetic mapping and quantitative trait loci analysis of growth-related traits in the small abalone Haliotis diversicolor using restriction-site-associated DNA sequencing. Aquaculture 454, 163–170 (2016).
15. Yu, H. et al. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 6, e17595 (2011).
16. Yáñez, J. M. et al. Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. Mol. Ecol. Resour. n/a–n/a (2016). doi:10.1111/1755-0998.12503
17. Correa, K. et al. Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (Salmo salar L.) using a 50K SNP genotyping array. Aquaculture (2016). doi:10.1016/j.aquaculture.2016.04.008
18. 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, 275 (2002).
19. Liu, S. et al. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12, 53 (2011).
20. Liu, S. et al. Development of the catfish 250K SNP array for genome-wide association studies. BMC Res. Notes 7, 135 (2014).
21. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248 (2007).
22. Van Tassell, C. P. et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 247–252 (2008).
23. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379 (2011).
24. Starks, H. A., Clemento, A. J. & Garza, J. C. Discovery and characterization of single nucleotide polymorphisms in coho salmon, Oncorhynchus kisutch. Mol. Ecol. Resour. 16, 277–287 (2016).
25. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7, e32253 (2012).
26. Annicchiarico, P. et al. Assessment of Cultivar Distinctness in Alfalfa: A Comparison of Genotyping-by-Sequencing, Simple-Sequence Repeat Marker, and Morphophysiological Observations. Plant Genome 9, 0 (2016).
27. Pértille, F. et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Sci. Rep. 6, 26929 (2016).
28. Boutet, G. et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics 17, 121 (2016).

29. Hashimoto, D. T., Senhorini, J. A., Foresti, F., Mart??nez, P. & Porto-Foresti, F. Genetic identification of F1 and post-F1 serrasalmid juvenile hybrids in Brazilian aquaculture. PLoS One 9, e89902 (2014).
30. Wang, L. et al. Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. Sci. Rep. 5, 16358 (2015).
31. Manousaki, T. et al. Exploring a Non-model Teleost Genome Through RAD Sequencing - Linkage Mapping in Common Pandora, Pagellus erythrinus and Comparative Genomic Analysis. G3 Genes Genomes Genet. 6, g3.115.023432– (2015).
32. Rondeau, E. B. et al. Genomics of sablefish (Anoplopoma fimbria): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. BMC Genomics 14, 452 (2013).
33. Jiao, W. et al. High-resolution linkage and quantitative trait locus mapping aided by genome survey sequencing: Building up an integrative genomic framework for a bivalve mollusc. DNA Res. 21, 85–101 (2014).
34. Wang, J., Li, L. & Zhang, G. A High-Density SNP Genetic Linkage Map and QTL Analysis of Growth-Related Traits in a Hybrid Family of Oysters (Crassostrea gigas × C. angulate) Using Genotyping-by-Sequencing. G3 (Bethesda). 6, g3.116.026971– (2016).
35. Lu, X. et al. High-resolution genetic linkage mapping, high-temperature tolerance and growth-related quantitative trait locus (QTL) identification in Marsupenaeus japonicus. Molecular Genetics and Genomics 291, 1–15 (2016).
36. Hashimoto, D. T., Senhorini, J. A., Foresti, F., Mart??nez, P. & Porto-Foresti, F. Genetic identification of F1 and post-F1 serrasalmid juvenile hybrids in Brazilian aquaculture. PLoS One 9, e89902 (2014).
37. Perazza, C. A., de Menezes, J. T. B., Ferraz, J. B. S. & Hilsdorf, A. W. S. Lack of intermuscular bones in specimens of Colossoma macropomum: An unusual phenotype to be incorporated into genetic improvement programs. Aquaculture (2016). doi:10.1016/j.aquaculture.2016.05.014
38. Poland, J. A. & Rife, T. W. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J. 5, 92–102 (2012).
39. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–51 (2011).
40. Lu, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9, e1003215 (2013).
41. Wang, L. et al. Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. Sci. Rep. 5, 16358 (2015).
42. Ward, J. A. et al. Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. BMC Genomics 14, 2 (2013).
43. Li, C. et al. SNP discovery in wild and domesticated populations of blue catfish, Ictalurus furcatus, using genotyping-by-sequencing and subsequent SNP validation. Mol. Ecol. Resour. 14, 1261–1270 (2014).
44. Kakioka, R. et al. A RAD-based linkage map and comparative genomics in the gudgeons (genus Gnathopogon, Cyprinidae). BMC Genomics 14, 32 (2013).
45. Zhang, X. et al. A Consensus Linkage Map Provides Insights on Genome Character and Evolution in Common Carp (Cyprinus carpio L.). Mar. Biotechnol. 15, 275–312 (2013).
46. Guyomard, R., Boussaha, M., Krieg, F., Hervet, C. & Quillet, E. A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. BMC Genet. 13, 15 (2012).

47. Zhang, Y. et al. Comparative genomic analysis of catfish linkage group 8 reveals two homologous chromosomes in zebrafish and other teleosts with extensive inter-chromosomal rearrangements. BMC Genomics 14, 387 (2013).

48. Kajimura, S., Aida, K. & Duan, C. Understanding hypoxia-induced gene expression in early development: in vitro and in vivo analysis of hypoxia-inducible factor 1-regulated zebra fish insulin-like growth factor binding protein 1 gene expression. Mol. Cell. Biol. 26, 1142–55 (2006).

49. Maddison, L. a, Joest, K. E., Kammeyer, R. M. & Chen, W. Skeletal muscle insulin resistance in zebrafish induces alterations in β-cell number and glucose tolerance in an age- and diet-dependent manner. Am. J. Physiol. Endocrinol. Metab. 308, E662–9 (2015).

50. Ghia, P. et al. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K 21, 1–3 (2007).

51. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12, 115–21 (2015).

52. Glaubitz, J. C. et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9, e90346 (2014).

53. Krzywinski, M. et al. Circos: An information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009).

54. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070 (2010).

3. IDENTIFICATION OF OUTLIER LOCI ASSOCIATED WITH REGIONAL ADAPTATION AND CLIMATE VARIABLES IN AN AMAZON FRESHWATER FISH

ABSTRACT

The intensity of climate changes impact on aquaculture activity will depend on velocity of the response and adaptation to new production conditions. The identification of candidate loci of response to climate variables can provide a valuable tool for selection of individuals more adapted to the new climatic reality. We evaluated tambaquis (*Colossoma macropomum*) populations from three Brazilian climatic regions using three different outlier detection methods. Out of 888 loci identified, 81 were supported by up to 3 statistical methods, while 449 loci were associated with differences in temperature, atmospheric pressure, sunshine or cloudiness. Functional annotation of these 888 loci, allowed identify important pathways and candidate genes with functions of the adaptive process. This represents the first report to identify outliers loci to climate adaptation in this species and it will be a valuable source of information for selection of individuals adapted for different geographic regions.

Keywords: *Colossoma macropomum*; Outliers; GBS; Climate changes; Restriction enzyme

3.1. Introduction

Although the theory of animal and plant breeding were started during the 1930s [1], fish aquaculture breeding programs have rarely been used, and only between 1% and 2% of production is based on genetically improved stocks [2]. However, before the planning and start a breeding program, the genetic resources available must be accurately identified and characterized [3]. A correct characterization of available genetic resources should consider the place where which these resources were developed and future changes to be faced by aquaculture.

The challenges posed by climate change will have direct impacts on the way of life and production around the world where large impacts are predicted on the global economy [4]. Aquaculture is an area that can be directly and indirectly affected due to the organisms sensitivity to environmental factors [5]. These impacts can be predicted through analysis of variability in annual yield in response to climate anomalies year-to-year. However, variations in the pattern response to annual climate changes does not reflect the organism adaptation to lasting change [6]. On another hand, the study of the adaptive differences between animals of different climatic regimes can provide valuable information about definite changes in the patterns of climate change responses [7].

The autoregulation inability of body temperature by aquatic organisms makes them highly sensitive to environmental fluctuations [6]. Changes in environmental variables such as temperature, atmospheric pressure, sunshine and cloudiness, can alter the environment in which these organisms are inserted, causing changes in acidification patterns of water, oxygen dilution and alterations in energy balance [8,9,10,4]. Furthermore, it can collaborate to toxic organism growth that reduce water quality and decrease the aquaculture production [11].

Tambaqui (*Colossoma macropomum*) is the largest native *Characiform* species from the Amazon and Orinoco river basins of South America[12]. This species has been recognized as an economically valuable species owing to fast-growing, omnivorous behavior, which tolerates poor water quality, and by the consumer market acceptance [13]. Tambaqui has been also farmed in different countries around the Amazon boundary and has been introduced into other countries of Latin America, Asia, and Africa[14]. Tambaqui farming is growing rapidly in Brazil, its production

reached 139.209 tons in 2014, what corresponds to 57.7% of increase compared with 2013 [15]. Most of this growth occurred in northern of Brazil. This region gathers the ideal conditions for its production [16], however, since the 60's, it has been also produced in other regions with less favorable weather conditions [17].

In the present study, we evaluated eight populations of tambaquis from the north, northeastern and southeastern Brazil using three different outlier detection methods. We found many loci associated with regional adaptation and climate variables in tambaqui. We also identify important pathways and candidate genes to the adaptive process.

3.2. Materials and Methods

3.2.1. Ethical statement

All experimental protocols employed in the present study that relate to animal experimentation were performed in accordance with Brazilian director for the care and use of animals in teaching or scientific research activities – DBCA, resolution number 30/2016 approved by the National Animal Experimentation Control Council to ensure compliance with international guidelines for animal welfare. The individuals were not subjected to any experimental manipulation or the euthanasia procedure.

3.2.2. Sample processing, discovery and filtering of SNPs

A total of 229 farmed and wild individuals were collected in eight different places of three Brazilian regions (Figure 3.1). Tail fin clips were collected from each individual, and preserved in 90% ethanol. DNA extraction was conducted using proteinase K protocol. DNA integrity was checked in 1% agarose gel and all DNA samples were stored at −20°C prior to sequencing library preparation.

Genotype-by-sequencing (GBS) library construction and sequencing were conducted at the Animal Biotechnology Laboratory at University of São Paulo (Piracicaba, Brazil) using the protocol described by Elshire et al. (2011)[23] with modifications described by Nunes et al. (manuscript under review). The libraries were sequenced using Illumina TruSeq SBS Kit v3-HS on the Illumina HiSeq2500 sequencer (Illumina, San Diego). Quality trimming was performed with SeqyClean tool v. 1.9.10 (https://bitbucket.org/izhbannikov/seqyclean/) using a Phred quality score ≥24, a fragment size ≥50 and a contaminant database to remove vector, adapter and other sequence contaminations. The SNP discovery was performed using UNEAK[60] (Universal Network Enabled Analysis Kit) pipeline with default parameters.

The genotype data were filtered for all samples, using marker call rate >90% (i.e., a SNP was called in at least 90% of the samples). The filtering was performed using Tassel v.5.0 [61]. We also used this program to estimate linkage disequilibrium between each pair of loci. When loci pair had a r^2 value >0.8, the locus with highest missing data was removed. Additionally, we used the GENEPOP *v.*4 (http://genepop.curtin.edu.au/) to perform the Hardy–Weinberg equilibrium (HWE) deviations tests. Putative SNPs showing significant HWE deviations ($P<0.05$) were removed. Finally, the genotype data was filtered using sample call rate >80% (i.e., at least 80% SNPs had genotypes called considering all samples).

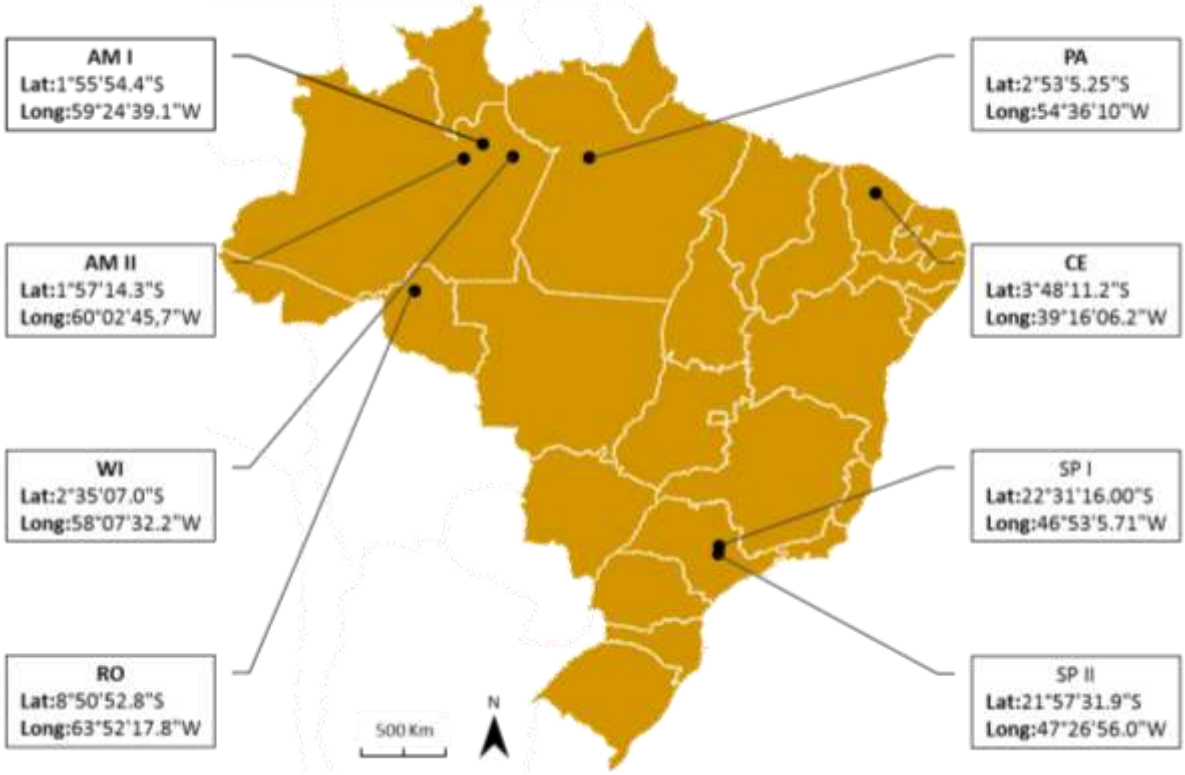


Figure 3.1: Tambaqui sample localities map with sample code and geographical coordinates. Details for each site were provided in Table 3.1.

Table 3.1: Characteristics of tambaqui populations used in this study. Population code, sampling site, annual temperature mean (Tmean), annual sunshine (Sun), cloudiness (Cloud), atmospheric pressure (ATM) and sample size (N) of eight tambaqui populations. The populations AM-I, AM-II, PA, CE, RO, SP-I and SP-II were from fish farms and the population WI came from nature.

Population code	Sampling site	T_{mean}(°C)	Sun(h)	Cloud(tenths)	ATM(hPa)	N
WI	Rio Uatumã-AM	26.4	1532.2	0.7	1008.2	30
AM-I	Balbina-AM	26.6	1828.5	0.6	1003.7	38
AM-II	Pres. Figueiredo-AM	27.4	1828.5	0.6	1003.7	30
PA	Mojuí de Campos-PA	26.0	2031.6	0.5	1008.8	30
CE	Pentecoste-CE	26.6	2843.4	0.5	1008.7	31
RO	Porto Velho-RO	25.6	1988.4	0.7	999.7	30
SP-I	Mogi Mirim-SP	20.6	2384.7	0.5	918.4	28
SP-II	Pirassununga-SP	20.1	2384.7	0.5	918.4	12

3.2.3. Population structure

The first genotype attribution analysis was conducted using STRUCTURE 2.3.4 [62]. This program was used to estimate the genetic population number and assign individuals to populations. We assumed a population number (K) ranging from 1 to 10, with 10 interactions each. Each run was conducted under the condition of 1,000,000

Markov chain Monte Carlo (MCMC) replications followed by 500,000 burn-in periods. Structure Harvester [63] was used to define the most probable K number.

We used principal component analysis (PCA) as an alternative method to the population structure analysis. This method allows the identification of highlight clusters of individuals, using the level of minor allele frequency to group these individuals through successive orthogonal PCs accounted for the maximum variance. The PCA were performed using Tassel v.5.2.26 [61] considering five principal components.

3.2.4. Identification of outlier loci

The first method used the distributions of heterozygosity and F_{ST} outlier approach based on simulation methods to test neutrality desviations[64]. This method was implemented in LOSITAN by Antão et al. (2008) [65]. It is expected that loci under positive selection shows high F_{ST} values, whereas loci under balancing selection shows low F_{ST} values. We perform 100,000 coalescent simulations with forced F_{ST} mean calculated for each pairwise comparison. The F_{ST} mean value found in the first analysis was added and the program was run again. A third analysis was performed with the F_{ST} value generated based on our dataset and tested for outliers at the 95% of confidence interval level and false discovery rate (FDR) of 0.05.

The second method used was suggested by Foll and Gaggiotti (2008) and implemented in BayeScan[66] program. This method uses a logistic regression model which explains the observed pattern of diversity by dividing it in a locus- and a population-specific component. This statistical methodology decomposes locus population F_{ST} values in two classes of components (locus-specific components (α) and population-specific components (β)) and then estimates the probability of a locus being under selection comparing their α and β allele frequencies [66]. We performed BayeScan following 20 pilot runs of 5,000 iterations with a 100,000 burn-in on a sample size of 5,000 and a thinning interval of 20. We also used the prior odds of 10 for a neutral mode running.

Considering that loci under selection can be correlated with environmental variables, we tested this hypothesis using Samβada [67] software. The methodology implemented in Samβada use genome–environment associations to identify which variables are influencing the selection of a specific molecular marker. It allows the study of local adaptation of a certain species based on a set of polymorphic markers. Therefore, this method does not make predictions for the genotype of individuals based on its sampling site environmental characteristics, but the focus is detecting which loci are potentially locally adapted. We must consider that Samβada was used to identify outlier loci based on environmental variables not in population structure. The individuals were coded with the presence (1) or absence (0) of an allele; it generated 3 possible genotypes (00, 01, 11) for each SNP tested. The Bonferroni-corrected threshold corresponding to $\alpha = 0.01$ was applied to significant tests.

3.2.5. Environmental data

The climate data were obtained from a public website (http://www.inmet.gov.br/) using the geographical coordinates where tambaquis were sampled. Among the available variables, we chose those that may be directly or indirectly influencing the tambaqui adaptation in the study regions. Associations between markers and environmental variables were tested for four environmental variables (temperature, atmospheric pressure, sunshine and cloudiness)

as described in Table 3.1. The mean for each environmental variable used in this study was calculated considering data collected in the last 30 years.

3.2.6. Distribution of outlier loci in tambaqui linkage map

To investigate the distribution of outlier loci among tambaqui linkage groups we aligned the sequences from de outliers SNPs with the tambaqui linkage map developed by Nunes et al. (manuscript under review) using Bowtie2 v.2.2.5 [68]. Sequences with multiple alignments were removed.

3.2.7. Annotation and pathways analysis

We used two methods for annotation the outlier loci. The first annotation was implemented using the Variant Effect Predictor (VEP) tool .7.1[69]. We use VEP to annotate the outlier loci using the zebrafish genome (*Danio rerio*, Zv9) as reference genome. We also annotated the outlier loci using Basic Local Alignment Search Tool (BLAST) [70]. The BLAST was used to search by sequence similarity against biological sequence databases for all organism on NCBI databases.

We used the internet-based tool Reactome [71], to perform an analysis of biological pathways in *Danio rerio* genome enriched by genes identified by outlier loci. The Reactome data model allows the representation of many diverse processes in biological system, including the pathways of intermediary metabolism, regulatory pathways, and signal transduction, and high-level processes, such as the cell cycle. Reactome accept the input of zebrafish genes with the ENSEMBL identifier and provides a curated peer-reviewed resource of biological processes.

3.3. Results

3.3.1. Sequencing, discovery and SNPs filtering

A total of 229 individuals were sequenced from eight populations using GBS approach to discover SNPs in tambaqui. Over 652 million single-end reads with 100bp were generated. The number of single-end reads obtained for each individual ranged from 1.5 to 7.3 million, with an average of 2.8 million. A total of 183,196 pairwise alignments were obtained with UNEAK pipeline. Default filtering parameters were applied to reveal 134,430 putative SNPs genotyped with minor allele frequency (MAF) higher than 0.05. Out of these, 119,688 were removed because they had marker call rate <90%; 3,881 were removed due the linkage disequilibrium ($r^2 > 0.8$); and finally, 2,078 were removed due deviations from HWE ($P<0.05$). Finally, 8,783 SNPs were retained to be used in further analyses. A total of 18 individuals showed more that 20% of missing data and were removed: 13 from AM-II and 5 from WI. Our final dataset consisted of 211 individuals genotyped at 8,783 SNPs.

3.3.2. Population clustering

Structure Harvester indicated K = 4 (the highest peak of DeltaK) as the best partition, suggesting four genotypic clusters (Figure 3.2). Two populations were composed mostly by individuals from de North, one population from the Northeast and one from the Southeast of Brazil. The North I cluster (n = 70) was genetic composed by individuals from AM I (n = 37), RO (n = 20), AM II (n = 7), WI (n = 3), SPI (n = 2) and CE (n = 1). The North II population (n = 69) was composed by individuals from PA (n = 30), WI (n = 23), RO (n = 10) and AM II (n = 6). The Northeast population (n = 29) was genetic composed only by individuals from CE (n = 29). The Southeast population (n = 43) was genetic composed by individuals from SPI (n = 26), SPII (n =12), AM II (n = 4) and AM I (n = 1).

In North I, North II, Northeast and Southeast populations, the number of individuals with probability of assignment to their respective genetic populations greater than 75% were of 37(53%), 35(51%), 25(86%) and 33(77%) respectively. These four clusters were selected to be used in subsequent analyses.

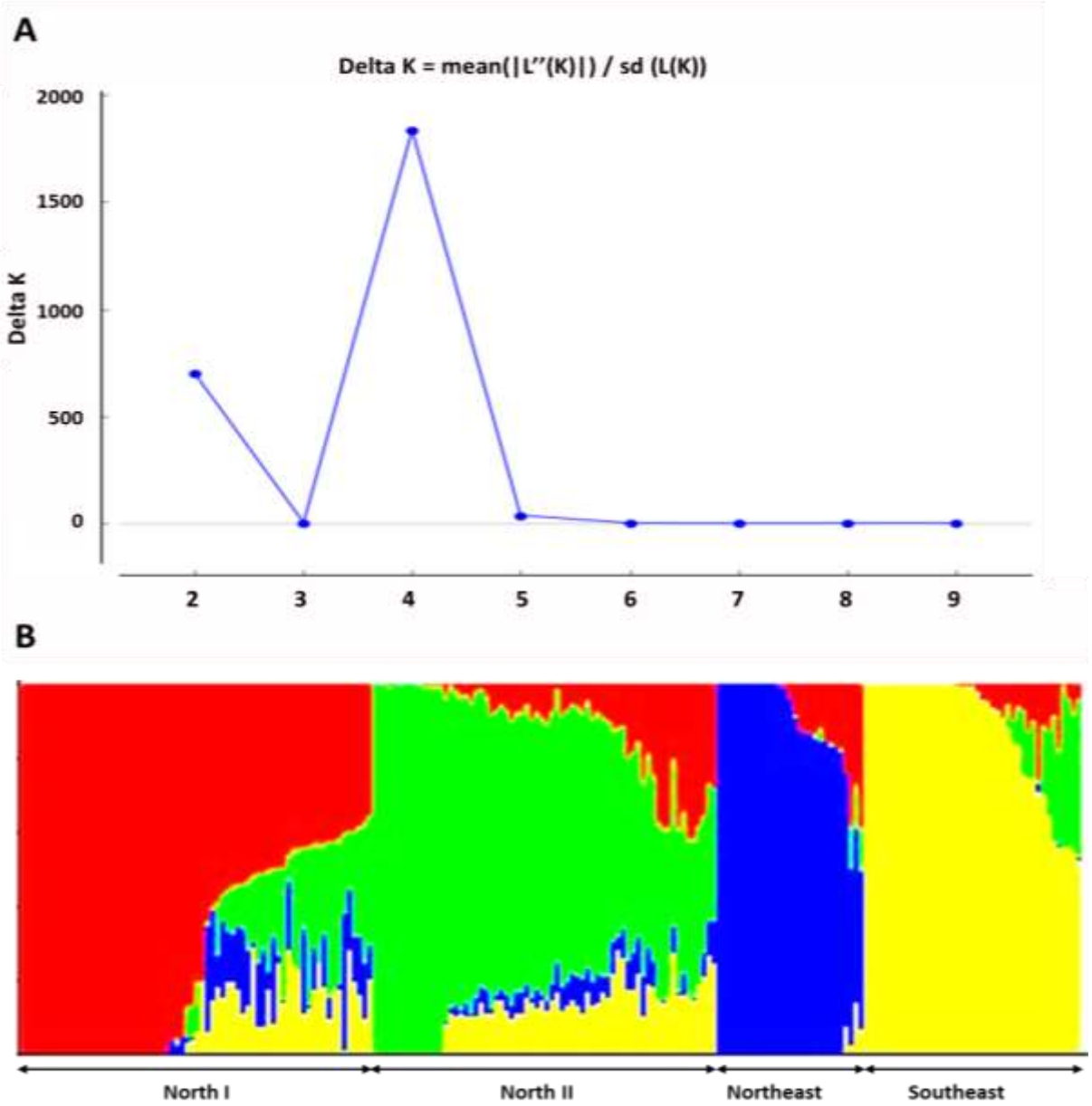


Figure 3.2: (A) Delta K values for K = 2 - 9 indicating highest likelihood at K = 4. (B) STRUCTURE barplot (k = 4). Each individual is shown as a vertical bar. The groups include: North I (red); North II (green); Northeast (blue); Southeast (yellow).

PCA analysis showed similar results to those found with STRUCTURE. Although a few individuals were allocated in different groups, four distinct clusters from tambaqui were shown in Figure 3.3. The highest variance axis (PC1) mainly separating North I and Southeast populations from North II and Northeast, while the second highest variance axis (PC2) splitting North I from Southeast individuals and North II from Northeast individuals.

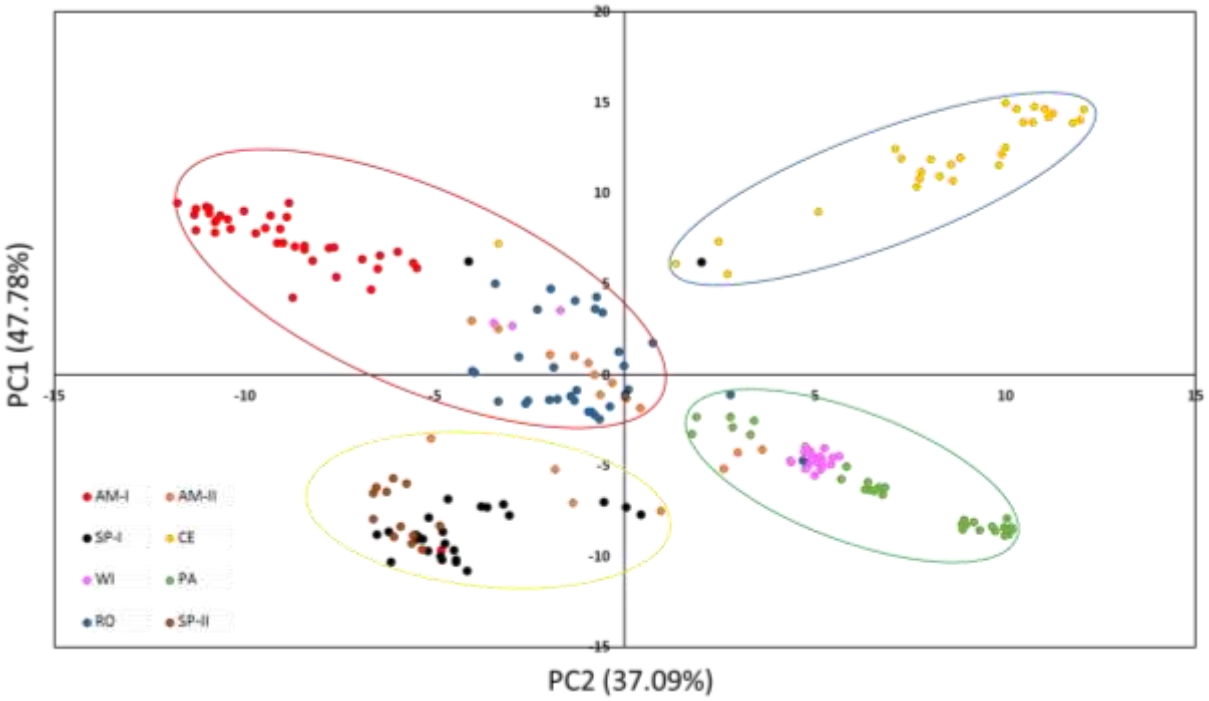


Figure 3.3: Principal component analysis (PCA) for eight populations of tambaqui based in 8,783 SNPs. The PCA reveals four clusters: North I (red circle); North II (green circle); Northeast (blue circle) and Southeast (yellow circle). The proportion of variance explained by the PCs is showed in parentheses along each axis.

3.3.3. Outlier Detection

Using Lositan, we identified 430 outliers across the four populations (0.05 confidence interval after FDR correction, Figure 3.4). Of these 153 SNPs presented higher F_{ST} values than expected under neutrality and were identified as being under positive selection. 277 presented a lower F_{ST} values than expected under neutrality and were identified as being under balancing selection.

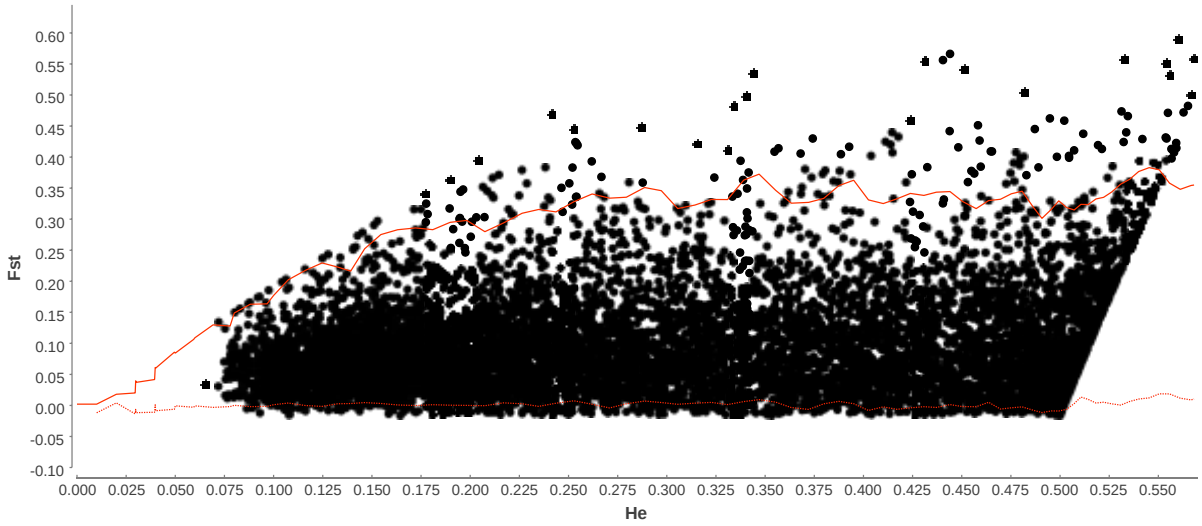


Figure 3.4: Detection of outlier SNP loci in tambaqui using the Lositan workbench at the 0.05 threshold. Each point corresponds to an SNP locus. Above the solid red line, is the confidence area for candidate loci under positive selection; below the dotted red line is the confidence area for candidate loci potentially under balancing selection; between the lines is the confidence area for neutral loci.

The BayeScan revealed 22 outlier loci after correction using a false discovery rate of 0.05. The decisive threshold value (log10 PO) used for identifying loci under selection was 0.5. The 22 outlier loci had F_{ST} ranging from 0.259 to 0.372 and log10 of the posterior odds (PO) ranging from 0.541 to 2.467 (Figure 3.5). Of these outliers, 8 had log10 Bayes factor values above 1.5.

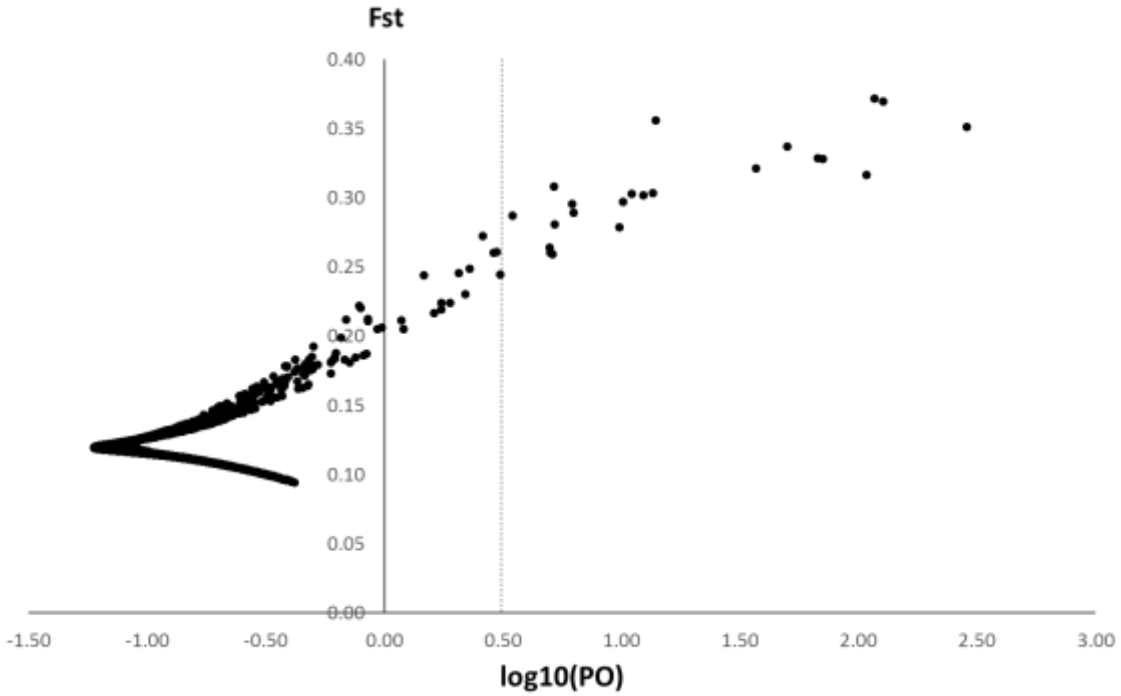


Figure 3.5: Detection of outlier SNP loci in tambaqui using the BayeScan. Each point corresponds to an SNP locus. F_{ST} shows evidence whether the locus is under selection or not. F_{ST} were plotted against the log10 of the posterior odds (PO). The vertical dashed line provides the decisive threshold value (log10 PO = 0.5) used for identifying loci under selection.

Using Samβada, we identified 669 significant associations with environmental variables. The outlier loci with significant associations were related to temperature (222 loci), atmospheric pressure (228 loci), sunshine (185 loci) and cloudiness (34 loci). 215 loci were associated with more than one environmental variable.

A total of 821 outlier loci were identified in the three tests for outlier. Out of these, 74 loci were identified as outliers in at least two of the methods; 12 were identified by Lositan and BayeScan; 56 by Lositan and Samβada; and 6 for all methods (Figure 3.6).

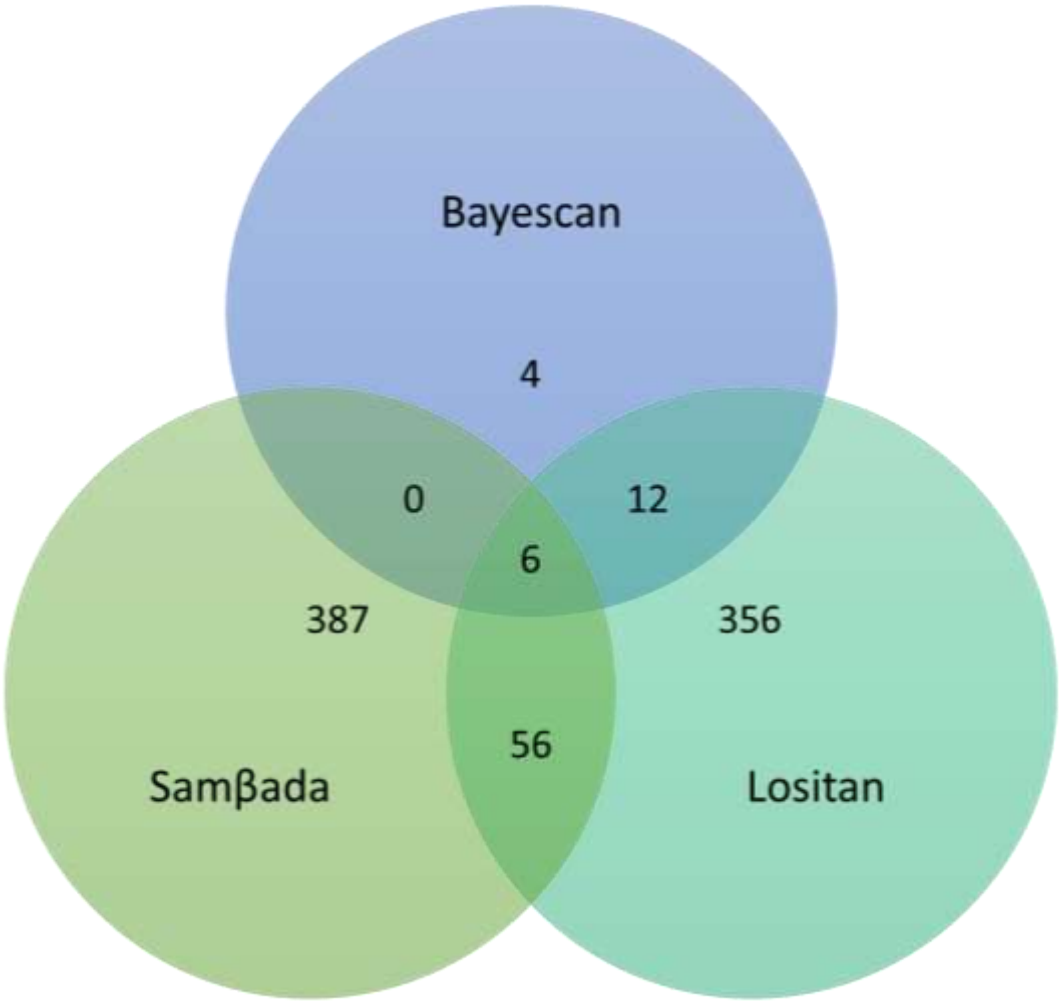


Figure 3.6: Venn diagrams illustrating the comparisons of outliers detected for BayeScan, Lositan and Samβada methods.

3.3.4. Distribution of outlier loci in tambaqui linkage map

Of the 821 outlier loci discovered, 339 (41.3%) were successfully aligned against tambaqui map among all linkage groups (Figure 3.7). However, the SNPs were not evenly distributed across the 27 linkage groups. We observed that many loci identified by one method are close from others identified by other. For example, on linkage group 12, between 63 and 65 cM, we can see 3 loci identified as outlier by the three different programs.

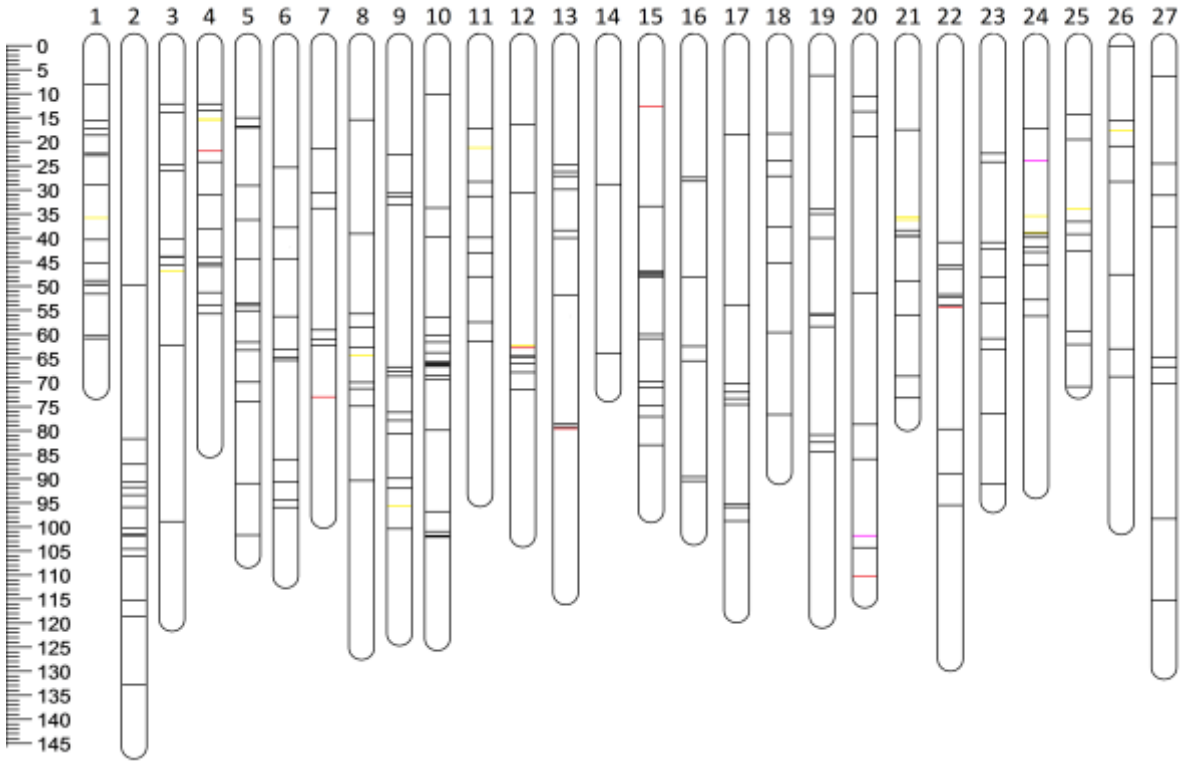


Figure 3.7: Distribution of 339 outlier loci in tambaqui linkage map. Genetic distances between adjacent markers are shown on side ruler in centiMorgan (cM). Horizontal black lines represent markers aligned with outlier loci discovered by just one outlier detection method while horizontal colored lines represent markers aligned with outlier loci discovered by two or three different outlier detection methods (Red line - BayeScan and Lositan; Yellow line - Lositan and Samβada; Pink line - All methods).

3.3.5. Annotation and pathways analysis

The outlier loci were annotated against the zebrafish genes (ENSEMBL release 84). This allowed us to evaluate the potential use of identified outlier loci with respect to possible important traits in breeding programs. A total of 286 loci were successfully aligned against zebrafish genome. The annotations of the sequences included a range of variant types such as intronic, downstream gene, upstream gene, synonymous, missense, intergenic, splice region, and non-coding transcript (Figure 3.8). The functional annotation of the 286 loci aligned against zebrafish genome were classified into of all categories. Some loci were classified into more than one category, which resulted in the sum of the loci ratio in each category exceeding 100%. In VEP annotation approximately 66% of the SNPs evaluated were annotated to be intronic, downstream or upstream of gene regions, suggesting that many putative outlier loci detected in tambaqui were located in protein coding sequences or close to them. (Figure 3.8).

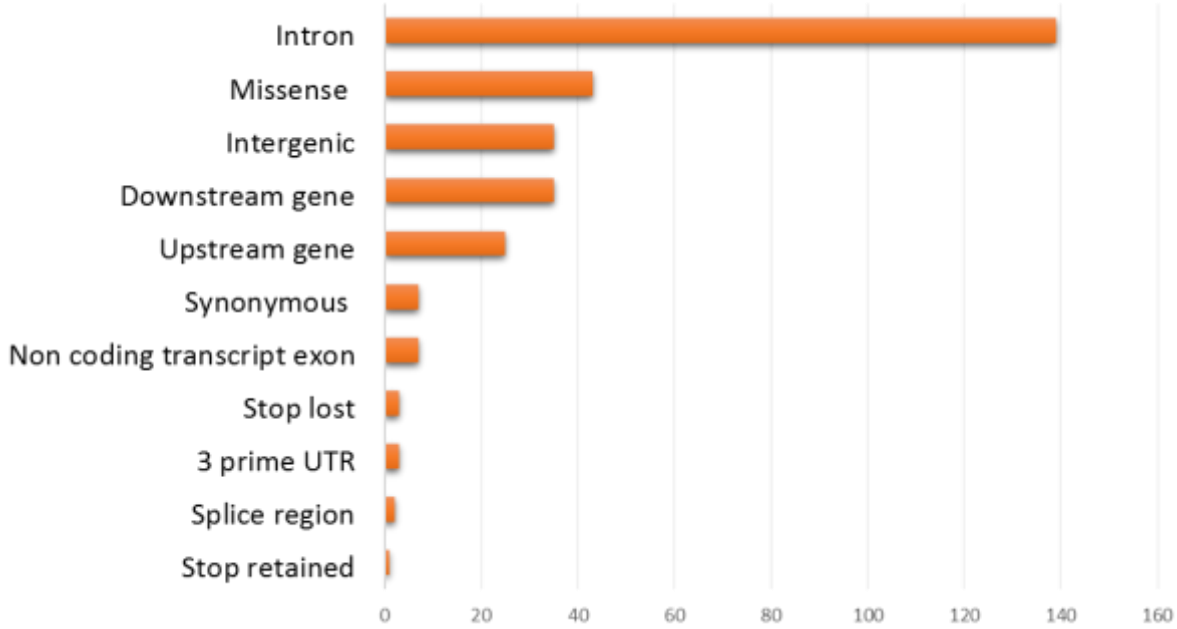


Figure 3.8: Annotated outlier loci aligned against genes from zebrafish genome.

The BLAST tool was used to annotate eighteen loci putatively under selection, selected by BayeScan and Lositan (APPENDIX C). The annotation showed hits against genes with immune, regulatory and structural functions. Of these eighteen loci, six loci had also been identified by Samβada and had significant associations with one or more environmental variables. The environmental variables with significant associations were related to temperature, annual sunshine, and atmospheric pressure. In our annotation using BLAST, others species also were identified (APPENDIX C). However, the majority of outlier loci were annotated in fish species (10 loci in 7 species), with three species accounting for over 75% of results: *Pygocentrus nattereri* (31.25%), *Cyprinus carpio* (25%), and *Danio rerio* (18.75 %).

The pathway analysis performed with Reactome in the 18 outlier loci common by BayeScan and Lositan, shows that all genes in our list were FGF (Fibroblast growth factors) family member allocated in signal transduction pathway. The analysis performed with the outlier loci selected by Samβada shows that the genes encode by these outlier plays a role in the pathways of the immune system, signal transduction, hemostasis, metabolism and metabolism of proteins (Figure 3.9).

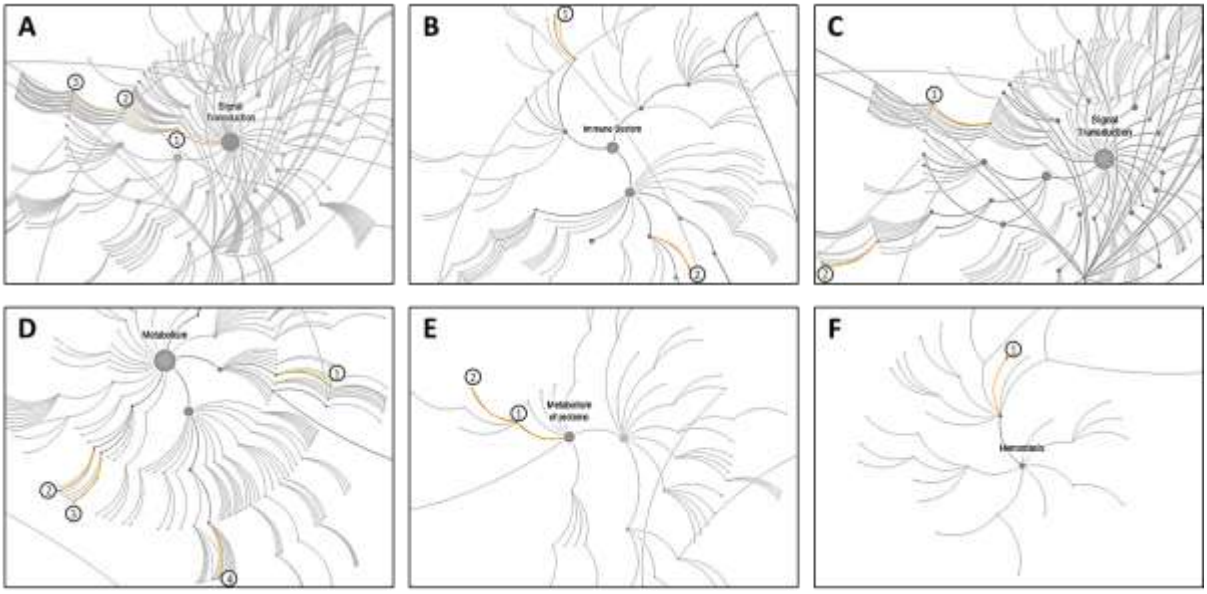


Figure 3.9: Pathway analysis performed with Reactome showing the association between genes related with outlier loci and biological pathways. Significantly enriched pathways to the 18 outlier loci common by BayeScan and Lositan (A) and the 669 selected by Samβada (B, C, D, E, F). The numbers showed in pathways indicate: A - Signaling by FGFR (1), Signaling by FGFR1 (2), FGFRL1 modulation of FGFR1 signaling (3); B - Generation of second messenger molecules (1), FCERI mediated Ca+2 mobilization (2); C - Signaling by FGFR1 (1), Vasopressin-like receptors (2); D - Glycolysis (1), Pyrimidine catabolism (2), Purine catabolism (3), Synthesis of PA (4); E - Unfolded Protein Response (UPR) (1), ATF6 (ATF6-alpha) activates chaperones (2); F- GPVI-mediated activation cascade (1).

3.4. Discussion

The identification of patterns of adaptive variation in non-model species has been a challenging, particularly for species with recent demographic history [18]. Although the tambaqui is considered a semi-migratory fish [19], the increasing demand of this species has favored its introduction in several regions of Brazil, creating new patterns of adaptive variation. A good example that illustrates this condition is the temperature average. The annual temperature average can vary up to 14º C between northern cities and southern cities in Brazil. Variations of this magnitude are now classified by Intergovernmental Panel on Climate Change (IPCC) as changes that may occur in an extreme CO^2 emissions scenario [4,20]. Our study of genetic variations between tambaquis populations selected in different Brazilian regions provided important data that will subsidize selection programs for this species.

The increased accessibility of high-throughput sequencing and genotyping methods allowed the discovery and genotyping of genome-scale SNPs in model and non-model organisms [21,22,23]. Genomic datasets are now allowing the identification of outlier loci with unprecedented accuracy. We presented here the first study that identified outlier loci to tambaqui using high density of SNPs discovered by genotype-by-sequencing (GBS) approach. The large number of SNPs discovered and genotyped in this study showed the efficiency of GBS approach [23]. Indeed, our analysis produced high proportions of putative SNPs. Although the filtering steps dropped about ~93% of the 134,430 SNPs, the number of remaining markers are high (8,783) for a non-model organism. Just as in tambaqui, GBS approach have allowed the SNPs discovery of SNPs for many non-model organisms [24,25,26,27].

The concordance between population structure observed in both Structure and PCA enabled us to define four major clustering groups with clear and consistent separation among them. The number of populations detected by structure analysis shown few variation among the eight populations. Our results are consistent with Jacometo et al. (2010) that evaluated the genetic diversity of four broodstocks of tambaqui from different regions of Brazil using RAPD markers, and also, the same study showed that most of the variation is within each broodstock and not among them[28]. The genetic proximity among the populations from the northern region of Brazil showed by PCA analysis, may have been due to the introduction of new individuals from nature or from others commercial hatcheries. The capture of new animals to compose the stock of reproductive matrices is a common practice in the northern region, as well as the purchase or exchange of these individuals. The individuals from the northeast and southeastern Brazil remained in their original clusters, showing a higher level of structure than the populations from northern Brazil. However, individuals from SP-I and SP-II populations were grouped in the same cluster indicating possible common origin. The knowledge of genetic structure between populations are important to detect loci under selection in genome scans [29]. The incorrect population assignments to groups can results in errors and false positive detection [30]. A study evaluated the effect of incorrect population assignments to groups [29]. The genetic diversity was simulated in 10 populations arranged in various numbers of groups (k={1, 2, 5, or 8}) and then was investigated the data set by assigning populations to the similar or a different number of groups (in this case g={1, 2, 5, or 8}). Their results showed that the proportion of false positives were overestimated when k≠g. In our study, we have assumed that individuals were correctly assigned on their groups based on Structure and PCA analysis.

Outliers may be an important tool to indicate genomic regions under selection, allowing a quick response to the challenges posed by climate changes [31]. The number of outlier loci detected with Bayesian method (BayeScan), shown the conservative nature of this method. The proportion of outlier found by BayeScan was 94.9% lower than Lositan and 95.1% lower than Samβada. The higher proportion of outliers reported by Lositan and Samβada is likely attributable to their application of relaxed statistics approaches. Despite of low concordance between outliers detected by the three approaches, the alignment with tambaqui linkage map showed that many outlier loci were close to each other. It can indicate that these methods could be used as complementary approaches in outlier loci identification.

The environmental contrasts of Brazilian regions allowed the identification of outlier loci. Our results from Samβada analysis revealed possible adaptive forces along the Brazilian regions. Large differences in environmental variables may contribute to local adaptation [32]. It can accelerate the signature of selection at loci associated or integrated to genes with an adaptive function [31]. In addition, the large number of individuals from fish farms results in a reduced gene flow, potentiating the effect of selection. Here the largest number of associations occur with the variable atmospheric pressure. This variable has direct effect on floatability of fishes. These same effects were already evaluated in chum salmon (*Oncorhynchus keta*) during an atmospheric depression [9]. Atmospheric pressure, also can influence other important variables in aquatic systems as temperature, CO_2 concentration and water movement [9,33]. We found a high number of outlier loci associated with annual temperature mean. Temperature is a very important physical regulatory factor in fish metabolism and this effect is expressed in the control of many biological processes[6]. The autoregulation inability of body temperature make fishes dependent of environmental variations[34]. However, some species shows capacity for acclimation and adaptation to large temperature gradients, it is difficult to assess the temperature effects on these species [6]. The sensitivity of fishes to temperature variations is often evaluated in experimental conditions, but there is a potentially important distinction between thermal tolerances and behavioral preferences [6]. The use of outlier loci together with temperature response information could help uncover differences

between this effects. The large number of loci associated with annual sunlight confirms the importance of this variable in fish production. The sunlight is a primary energy source indispensable in ecosystems [35]. There is a strong association between sunlight and others important variables in fish production. The sunlight is very important in oxygen rates, food production, temperature, health and development of aquatic organisms [36]. However, increases in sunlight results in a high ultraviolet radiation. In fish, the effects of ultraviolet radiation exposure include DNA damage [37], damage to tissues of the skin [38] and the brain [39], behavioral changes [10] and can lead to increased mortality [40,41].

The use of outlier detection methods allows the identification of putative loci associated or integrated to genes with a known adaptive function. These genes also known as candidate genes, have a known function related in previous studies [42]. They can contribute with a particular process, a metabolic pathway, or are associated to a phenotype [42].

The 18 loci putatively under selection, selected by BayeScan and Lositan are involved in immunity, metabolism, biorhythm, and growth. The genes trim54, bod1l1, nol6, sec16a and pde6d encode important proteins in metabolism and cellular function [43, 44, 45, 46, 47]. The genes *mcc1* and *tnfsf13b* encode important proteins in immunologic system and they are candidate genes to tumor-suppressor gene (*mcc1*) [48] and tumor necrosis factor (*tnfsf13b*) [49]. We found an outlier loci associated with *rorb* gene, which encodes a protein that can bind as a monomer or as a homodimer to hormone response elements upstream of several genes to enhance their expression [50]. These genes code, e.g., the NM23-2 protein, a nucleoside diphosphate kinase involved in organogenesis and differentiation, which regulates the expression of some genes involved in circadian rhythm [51]. This locus also was associated with variations in temperature and atmospheric pressure. The gene *avpr2ab* found in signal transduction pathway encode arginine vasopressin receptor 2a that is associated with significant decreases in cardiac output in animals under hypoxia effect [52]. The gene *atf6* found in metabolism of proteins pathway also may be associated with hypoxia [53]. The hypoxia has important effects of on fishes with significant changes in genetic expression of skeletal muscle, liver and cell growth [54]. The environmental variables can increase the hypoxia effects in tropical fishes. A study evaluated the tambaqui adaptations to hypoxic conditions and conclude that the seasonal exposure to hypoxia caused increases in hemoglobin concentration and erythrocyte counts especially in the summer months, when high temperature reduce dissolved oxygen [55]. The FGF family members together with heparin sulfate proteoglycan or heparin promotes receptor dimerization and autophosphorylation on tyrosine residues [56]. The action of tyrosine residues is related with the production of catecholamines [56]. The catecholamine is released by the sympathetic nervous system as a primary neurohormonal response in organisms under stress[57]. The exposure of animals to cold, results in an increase of the gene expression of tyrosine hydroxylase and enhances the synthesis and release of catecholamine in the adrenal medulla, improving the response to cold stress [58]. This effects was already demonstrated in several animal model species [57,59,58]. Our results show that the gene *fgfr1* is a strong candidate gene to cold adaptation in tambaqui.

We presented here the first study that identified outlier loci of tambaqui genome using high density of SNPs discovered by GBS approach. Our genetic study of tambaqui population's structure in different regions can give us many theoretical references in breeding, allowing the understanding of foundation of breeding resources in these regions and operating as a guideline for genetic breeding. The identification of outliers may provide an important tool in identifying genomic regions under selection allowing a quick response to the challenges posed by climate changes.

References

1. Gjedrem, T. Genetic improvement of cold-water fish species. Aquac. Res. 31, 25–33 (2000).
2. Gjedrem, T., Salte, R. & Gjøen, H. M. Genetic variation in susceptibility of Atlantic salmon to furunculosis. Aquaculture 97, 1–6 (1991).
3. Ryman, N. Conservation of genetic resources: experiences from the brown trout ('Salmo trutta'). Fish Gene Pools Ecol. Bull. 17, 61–7461 (1981).
4. Ipcc. Climate Change 2007: impacts, adaptation and vulnerability: contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel. Genebra, Suíça (2007). doi:10.1256/004316502320517344
5. Rijnsdorp, A. D. et al. Resolving the effect of climate change on fish populations. ICES J. Mar. Sci. 66, 1570–1583 (2009).
6. Pankhurst, N. W. & Munday, P. L. Effects of climate change on fish reproduction and early life history stages. Mar. Freshw. Res. 62, 1015–1026 (2011).
7. Song, Z. et al. Genome scans for divergent selection in natural populations of the widespread hardwood species Eucalyptus grandis (Myrtaceae) using microsatellites. Sci. Rep. 6, 34941 (2016).
8. Buentello, J. A., Gatlin, D. M. & Neill, W. H. Effects of water temperature and dissolved oxygen on daily feed consumption, feed utilization and growth of channel catfish (Ictalurus punctatus). Aquaculture 182, 339–352 (2000).
9. Kitagawa, T., Hyodo, S. & Sato, K. Atmospheric depression-mediated water temperature changes affect the vertical movement of chum salmon Oncorhynchus keta. Mar. Environ. Res. 119, 72–78 (2016).
10. Kelly, D. J. & Bothwell, M. L. Avoidance of solar ultraviolet radiation by juvenile coho salmon (Oncorhynchus kisutch). Can. J. Fish. Aquat. Sci. 482, 474–482 (2002).
11. Van Dolah, F. Marine algal toxins: Origins, health, effects, and their increased occurrence. Environ. Health Perspect. 108, 133–141 (2000).
12. Araujo-Lima, C. & Goulding, M. So. Fruitful a Fish: Ecology, Conservation and Aquaculture of of the Amazon's Tambaqui. (Columbia Univ Press, 1997).
13. Campos-baca, L. & Kohler, C. C. Aquaculture of Colossoma macropomum and Related Species in Latin America. Am. Fish. Soc. Symp. 46, 541–561 (2005).
14. Animals, A. Production of Aquatic Animals -Fishes,. Anim. Sci. 194, 280–281 (1995).
15. Instituto Brasileiro de Geografia e Estatística - IBGE. Produção da Pecuária Municipal. IBGE 42, (2014).
16. Da Silva, J. A. M., Pereira Filho, M. & De Oliveira-Pereira, M. I. Frutos e sementes consumidos pelo tambaqui, Colossoma macrompum (Cuvier, 1818) incorporados em rações. Digestibilidade e velocidade de trânsito pelo trato gastrointestinal. Rev. Bras. Zootec. 32, 1815–1824 (2003).
17. William, J., Silva, B. E. & Inês, M. Resultados de um experimento sobre o cultivo do tambaqui, colossoma macropomum cuvier, 1818, na densidade de estocagem de 10.000 peixes/ha. Ciên. Agron. 20, (1989).
18. Akagi, T., Hanada, T., Yaegaki, H., Gradziel, T. M. & Tao, R. Genome-wide view of genetic diversity reveals paths 0of selection and cultivar differentiation in peach domestication. DNA Res. 23, 271–282 (2016).
19. Araújo-Lima, C. A. R. M. & Ruffino, M. L. Migratory fishes of the Brazilian Amazon. Migratory Fishes of South America: Biology, Fisheries and Conservation Status (2004).

20. Marengo, J. A., Jones, R., Alves, L. M. & Valverde, M. C. Future change of temperature and precipitation extremes in South America as derived from the PRECIS regional climate modeling system. Int. J. Climatol. 29, 2241–2255 (2009).
21. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248 (2007).
22. Van Tassell, C. P. et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 247–252 (2008).
23. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379 (2011).
24. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7, e32253 (2012).
25. Annicchiarico, P. et al. Assessment of Cultivar Distinctness in Alfalfa: A Comparison of Genotyping-by-Sequencing, Simple-Sequence Repeat Marker, and Morphophysiological Observations. Plant Genome 9, 0 (2016).
26. Pértille, F. et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Sci. Rep. 6, 26929 (2016).
27. Boutet, G. et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics 17, 121 (2016).
28. Jacometo, C. B. et al. Variabilidade genética em tambaquis (Teleostei: Characidae) de diferentes regiões do Brasil. Pesqui. Agropecu. Bras. 45, 481–487 (2010).
29. Excoffier, L., Hofer, T. & Foll, M. Detecting loci under selection in a hierarchically structured population. Heredity (Edinb). 103, 285–298 (2009).
30. Robertson, A. Letters to the editors: Remarks on the Lewontin-Krakauer test. Genetics 80, 396 (1975).
31. Schunter, C. et al. Molecular signatures of transgenerational response to ocean acidification in a species of reef fish. Nat. Clim. Chang. 1–5 (2016). doi:10.1038/nclimate3087
32. NUNES, V. L., BEAUMONT, M. A., BUTLIN, R. K. & PAULO, O. S. Multiple approaches to detect outliers in a genome scan for selection in ocellated lizards (Lacerta lepida) along an environmental gradient. Mol. Ecol. 20, 193–205 (2011).
33. Ishimatsu, A., Kikkawa, T., Hayashi, M., Lee, K.-S. & Kita, J. Effects of CO2 on Marine Fish: Larvae and Adults. J. Oceanogr. 60, 731–741 (2004).
34. Kurt Gamperl, A., Swafford, B. L. & Rodnick, K. J. Elevated temperature, per se, does not limit the ability of rainbow trout to increase stroke volume. J. Therm. Biol. 36, 7–14 (2011).
35. Lewis, N. S. Solar Energy Use. Science (80-.). 315, 798–802 (2007).
36. Olson, T. A. Some Observations on the Interrelationships of Sunlight, Aquatic Plant Life and Fishes. Trans. Am. Fish. Soc. 62, 278–289 (1932).
37. Lesser, M. P., Farrel, J. H. & Walker, C. W. Oxidative stress, DNA damage and p53 expression in the larvae of atlantic cod (Gadus morhua) explosed to ultraviolet (290 - 400nm) radiation. J. Exp. Biol. 204, 157–164 (2001).
38. Armstrong, T. N., Reimschuessel, R. & Bradley, B. P. DNA damage, histologial changes and DNA repair in larval Japanese medaka (Oryzias latipes) exposed to ultraviolet-B radiation. Aquat. Toxicol. 58, 1–14 (2002).

39. Vehniäinen, E. R., Vähäkangas, K. & Oikari, A. UV-B exposure causes DNA damage and changes in protein expression in northern pike (Esox lucius) posthatched embryos. Photochem. Photobiol. 88, 363–370 (2012).
40. Dong, Q., Svoboda, K., Tiersch, T. R. & Todd Monroe, W. Photobiological effects of UVA and UVB light in zebrafish embryos: Evidence for a competent photorepair system. J. Photochem. Photobiol. B Biol. 88, 137–146 (2007).
41. Braun, C., Reef, R. & Siebeck, U. E. Ultraviolet absorbing compounds provide a rapid response mechanism for UV protection in some reef fish. J. Photochem. Photobiol. B Biol. 160, 400–407 (2016).
42. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: Practical considerations. Nat. Rev. Genet. 3, 391–397 (2002).
43. Spencer, J. A., Eliazer, S., Ilaria, R. L., Richardson, J. A. & Olson, E. N. Regulation of microtubule dynamics and myogenic differentiation by MURF, a striated muscle RING-finger protein. J. Cell Biol. 150, 771–784 (2000).
44. Higgs, M. R. et al. BOD1L Is Required to Suppress Deleterious Resection of Stressed Replication Forks. Mol. Cell 59, 462–477 (2015).
45. Utama, B., Kennedy, D., Ru, K. & Mattick, J. S. Isolation and characterization of a new nucleolar protein, Nrap, that is conserved from yeast to humans. Genes Cells 7, 115–32 (2002).
46. O'Rielly, D. D. et al. Private rare deletions in SEC16A and MAMDC4 may represent novel pathogenic variants in familial axial spondyloarthritis. Ann. Rheum. Dis. 75, 1–8 (2015).
47. Ershova, G. et al. cDNA sequence, genomic organization and mapping of PDE6D, the human gene encoding the delta subunit of the cGMP phosphodiesterase of retinal rod cells to chromosome 2q36. Cytogenet. Cell Genet. 79, 139–141 (1997).
48. Poursoltan, P. et al. Loss of heterozygosity of the Mutated in Colorectal Cancer gene is not associated with promoter methylation in non-small cell lung cancer. Lung Cancer 77, 272–276 (2012).
49. Roth, A., Glaesener, S., Schütz, K. & Meyer-Bahlburg, A. Reduced Number of Transitional and Naive B Cells in Addition to Decreased BAFF Levels in Response to the T Cell Independent Immunogen Pneumovax®23. PLoS One 11, e0152215 (2016).
50. Sumi, Y. et al. Rhythmic expression of ROR?? mRNA in the mice suprachiasmatic nucleus. Neurosci. Lett. 320, 13–16 (2002).
51. Feng, S., Xu, S., Wen, Z. & Zhu, Y. Retinoic acid-related orphan receptor RORβ, circadian rhythm abnormalities and tumorigenesis (Review). International Journal of Molecular Medicine 35, 1493–1500 (2015).
52. Jin, H. K. et al. Hemodynamic effects of arginine vasopressin in rats adapted to chronic hypoxia. J. Appl. Physiol. 66, 151–60 (1989).
53. Liu, L., Liu, C., Lu, Y., Liu, L. & Jiang, Y. ER stress related factor ATF6 and caspase-12 trigger apoptosis in neonatal hypoxic-ischemic encephalopathy. Int. J. Clin. Exp. Pathol. 8, 6960–6 (2015).
54. Gracey, A. Y., Troll, J. V & Somero, G. N. Hypoxia-induced gene expression profiling in the euryoxic fish Gillichthys mirabilis. Proc. Natl. Acad. Sci. U. S. A. 98, 1993–8 (2001).
55. Saint-Paul, U. Physiological adaptation to hypoxia of a neotropical characoid fish Colossoma macropomum, Serrasalmidae. Environ. Biol. Fishes 11, 53–62 (1984).
56. Mamalaki, E., Kvetnansky, R., Brady, L. S., Gold, P. W. & Herkenham, M. Repeated immobilization stress alters tyrosine hydroxylase, corticotropin-releasing hormone and corticosteroid receptor messenger ribonucleic acid levels in rat brain. J. Neuroendocrinol. 4, 689–699 (1992).

57. Zhang, L. et al. Functional allelic heterogeneity and pleiotropy of a repeat polymorphism in tyrosine hydroxylase: prediction of catecholamines and response to stress in twins. Physiol. Genomics 19, 277–291 (2004).

58. Sabban, E. L. & Kvetnanský, R. Stress-triggered activation of gene expression in catecholaminergic systems: dynamics of transcriptional events. Trends Neurosci. 24, 91–8 (2001).

59. Richard, F. et al. Modulation of tyrosine hydroxylase gene expression in rat brain and adrenals by exposure to cold. J. Neurosci. Res. 20, 32–37 (1988).

60. Lu, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9, e1003215 (2013).

61. Glaubitz, J. C. et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9, e90346 (2014).

62. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).

63. Earl, D. A. & VonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4, 359–361 (2012).

64. Beaumont, M. A., Nichols, R. A. & Beaumont, M.A.; Nichols, R. A. Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proc. R. Soc. Lond. B 263, 1619–1626 (1996).

65. Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A. & Luikart, G. LOSITAN: A workbench to detect molecular adaptation based on a F st -outlier method. BMC Bioinformatics 9, 1–5 (2008).

66. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics 180, 977–993 (2008).

67. JOOST, S. et al. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol. Ecol. 16, 3955–3969 (2007).

68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).

69. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070 (2010).

70. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–10 (1990).

71. Croft, D. et al. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–D697 (2011).

4. GENOME-WIDE ASSOCIATION STUDY (GWAS) REVEALS GENES RELATED WITH LACK OF INTERMUSCULAR BONES IN TAMBAQUI (*Colossoma macropomum*)

ABSTRACT

Intermuscular bones offer a limitation in the consumption and consequently in the commercialization of many fish species, including the tambaqui (Colossoma macropomum). These bones can cause medical emergencies, losses in fish processing and financial losses, hindering expansion of farming of some species. Recent discovery of specimens lacking intermuscular bones can be useful in genetic studies to understand cartilage ossification and fish breeding programs. In this study, we carried out a genome-wide association study (GWAS) among tambaqui populations to identify markers associated with lack of intermuscular bone. After analyzing 11,423 SNPs in 360 individuals (12 cases and 348 controls), we reported 675 significant (Padj < 0.03) associations for this trait. Out of these, 13 associations were located around genes related to the reduction of bone mass, promotion of bone formation, inhibition of bone resorption, central control of bone remodeling, bone mineralization, and correlated functions. To the best of our knowledge, this is the first study to successfully identify genes related to lack of intermuscular bone using GWAS in a non-model species.

Keywords: *Colossoma macropomum*; GWAS; GBS; Intermuscular bone; Bone development

4.1. Introduction

The intermuscular bones have been a problem in the consumption of many fish species [1,2,3].including the tambaqui (*Colossoma Macropomum*) [4]. These bones are an common cause of medical emergencies in patients presenting sharp pain in their throat after eating fish [5]. The consequences of these clinical disorders may include esophageal perforation, mediastinitis, mediastinal abscess or introduction of a foreign body to the subcutaneous tissue of the esophagus.[6]. Mediastinitis is a surgical emergency with a high mortality rate [7]. Intermuscular bones also cause losses in the industrial fish processing resulting in financial losses and is an obstacle to the expansion of farming for some species [8].

Intermuscular bones originates from incomplete membranous ossifications of connective tissue in the muscular septum [9] as a response to strains suffered by muscles [10]. These small bones can have important role in the force transmission, contraction, body firmness, and reduction of myomere deformation [9] which are traits of great importance to wild fishes [11,12]. However, these traits are of secondary importance in farmed fishes, which are not exposed to wild environmental stress, such as food competition, predation, migration and seasonal variations [12].

Tambaqui (*Colossoma macropomum*) is the largest native Characiform species from the Amazon and Orinoco river basins of South America [13]. Tambaqui farming is growing rapidly in Brazil. Its production reached 139,209 tons in 2014, what corresponds to 57.7% of increase compared with 2013 [14], but, the presence of intramuscular Y-shaped bones in the fillet reduces substantially their economic potential [4]. Tambaqui lacking intermuscular bones already were found and reported before [15], however, the genetic mechanism involved in this phenotype is not known. [16,17].The discovery of specimens with the lack of intermuscular bone in tambaqui can help us understand the genetic process in bone formation and may allow selection for this phenotype in this and other species [15].

The emergence of Next Generation Sequencing (NGS) approaches allows development of cost-effective methods for large-scale identification of single nucleotide polymorphisms (SNPs) and genome-wide association study (GWAS) [18]. Advances in genomic research have significantly improved the tools available for the study of commercially important traits in aquaculture [19]. The use of association studies can help to unravel the biological mechanisms involved in economically important characteristics.

The challenge in the study of the lack of intramuscular bones is to identify the genetic process involved in bone formation. The understanding of these process can give us information to select specific markers to be applied in breeding program, and to plan effective strategies for genetic improvement. In this study, we did GWAS to identify loci associated with lack of intermuscular bones in specimens of tambaqui. Using these associated loci, we also performed a functional annotation to identify candidate genes related with enriched biological pathways using *Danio rerio* model as reference genome.

4.2. Materials and Methods

4.2.1. Ethical statement

All experimental protocols employed in the present study that relate to animal experimentation were performed in accordance with Brazilian director for the care and use of animals in teaching or scientific research activities – DBCA, resolution number 30/2016 approved by the National Animal Experimentation Control Council to ensure compliance with international guidelines for animal welfare.

4.2.2. Animals and Phenotypes

This study was conducted using 400 tambaquis from eight different places from Brazil, including a farm located in a northwestern region of Brazil in the state of Rondônia, where individuals lacking intermuscular bones were identified. The approach used to detect tambaqui lacking intermuscular bones was previously reported [15]. The procedures include a screening among broodstock using X-ray and ultrasound imaging. The phenotypes were classified into three categories: presence, total or partial lack of intermuscular bone. Tail fin clips were collected from each individual, and preserved in 90% ethanol. DNA extraction was conducted using proteinase K protocol. DNA integrity was checked in 1% agarose gel and all DNA samples were stored at −20°C prior to sequencing library preparation.

4.2.3. SNP discovery and filtering

Genotype-by-sequencing (GBS) library construction and sequencing were conducted using the protocol described by Elshire et al. (2011) [46] with modifications described by Nunes et al. (manuscript under review). We used SeqyClean tool v. 1.9.10 (https://bitbucket.org/izhbannikov/seqyclean/) to perform the quality trimming. The program was runed using a Phred quality score ≥24, a fragment size ≥50 and a contaminant database to remove vector, adapter, and other sequence contaminations. The UNEAK [47] (Universal Network Enabled Analysis Kit)

pipeline was used to perform the SNP discovery using default parameters. The SNPs discovered were filtered for all samples using Tassel v.5.0 [48] with marker call rate >80% and sample call rate >80%.

4.2.4. Genetic relationships

Genetic relationships among individuals were investigated using a technique of classification/ordination based on neighbor-joining and plotted in a cladogram. We also used principal component analysis (PCA) as a complementary method. Principal component analysis (PCA) reduces the dimensionality of the data while retaining most of the variation in the data set making possible to visually assess similarities and differences between samples and determine whether samples can be grouped [49]. Finally, we use a kinship analysis to determine if individuals with lack of intermuscular bone are from the same family. We use the Centered_IBS to produce a kinship matrix with a reasonable estimate of additive genetic variance. The cladogram, PCA and kinship were performed using Tassel v.5.0 [48] and cladogram was plotted using Archaeopteryx v.0.98 (http://www.phylosoft.org/archaeopteryx).

4.2.5. Association analysis

We did a case control analysis considering individuals with partial and total lack of intramuscular bone as belonging to the affected group. The association analysis was performed using standard methodology for case/control analysis implemented in PLINK [50] package. The basic association test is based on comparing allele frequencies between cases and controls. We follow a standard genetic approach using a model in which the traits of interest are minimally adjusted: we correct for sex and adjusting for a kinship structure. We also made a genome-wide multiple-testing correction with a false-discovery rate (FDR) control performed by computing q-values. SNPs with FDR q-values less than 0.03 were declared significant.

4.2.6. Distribution of markers in tambaqui linkage map

The loci associated with lack of intermuscular bone were aligned against the zebrafish genome (ENSEMBL release 84) and tambaqui linkage map developed by Nunes et al. (manuscript under review) using Bowtie2 v.2.2.5 [51]. Sequences with multiple alignments were removed.

4.2.7. Linkage disequilibrium (LD) analysis

The linkage disequilibrium (LD) analysis was performed and plotted using Haploview 4.2 (http://www.broadinstitute.org/haploview). The pairwise comparison of our SNPs was performed considering as linked the SNPs that were at a distance of up to 1cM. We compared the 149 genome-wide associated SNPs with the lack of intermuscular bone in tambaqui. We have defined the haplotype blocks by the solid spine of LD of Haploview 4.2.

4.2.8. Annotation and pathways analysis

The annotation was implemented using the Variant Effect Predictor (VEP) tool .7. 1 [52]. We used VEP to the loci annotations aligned against the zebrafish genome (*Danio rerio*, Zv9). The gene list from Ensembl VEP was used for biological pathways enrichment in *Danio rerio* reference genome using the Reactome [53] internet-based tool. Reactome accepts the input of zebrafish genes with the ENSEMBL identifier and provides a curated peer-reviewed resource of biological processes.

4.3. Results

4.3.1. Identification of tambaqui without intermuscular bones

The screening using X-ray allowed the identification of 28 individuals that lacked intramuscular bones. The use of ultrasound pictures confirmed complete lack of intramuscular bones in six individuals and six others showed partial remnant of bones in one or both of body sides.

4.3.2. Sequencing, discovery and filtering of SNPs

We used 400 individuals for sequencing and generated over ~921 million single-end reads with 100bp, with the number of reads obtained for each individual raning from 1.5 to 6.5 million, with an average of 2.3 million of reads. After default filtering parameters were applied, 182,416 putative SNPs with minor allele frequency (MAF) higher than 0.05 were obtained. From these 170,993 were removed because they had marker call rate <80%. From 400 samples, 40 showing more that 20% of missing data were removed. Our final dataset consisted of 11,423 SNPs genotyped for 360 individuals.

4.3.3. Population structure

Genetic relationships between tambaquis were calculated and the neighbor-joining tree is presented in Figure 4.1. The tree shows the formation of many intermediate subgroups with a short distance among them. Although the individuals with total or partial lack of intermuscular bones were spread in three subgroups, a relatively small genetic dissimilarity was observed among them. Many individuals with intermuscular bone showed small genetic dissimilarity when compared with individuals with total or partial lack of intermuscular bone. The branch-length varied between the individuals with an average of 6.24E-02 (stdev: 5.77E-02) and ranging from 1.59E-01 to 7.6E-05.

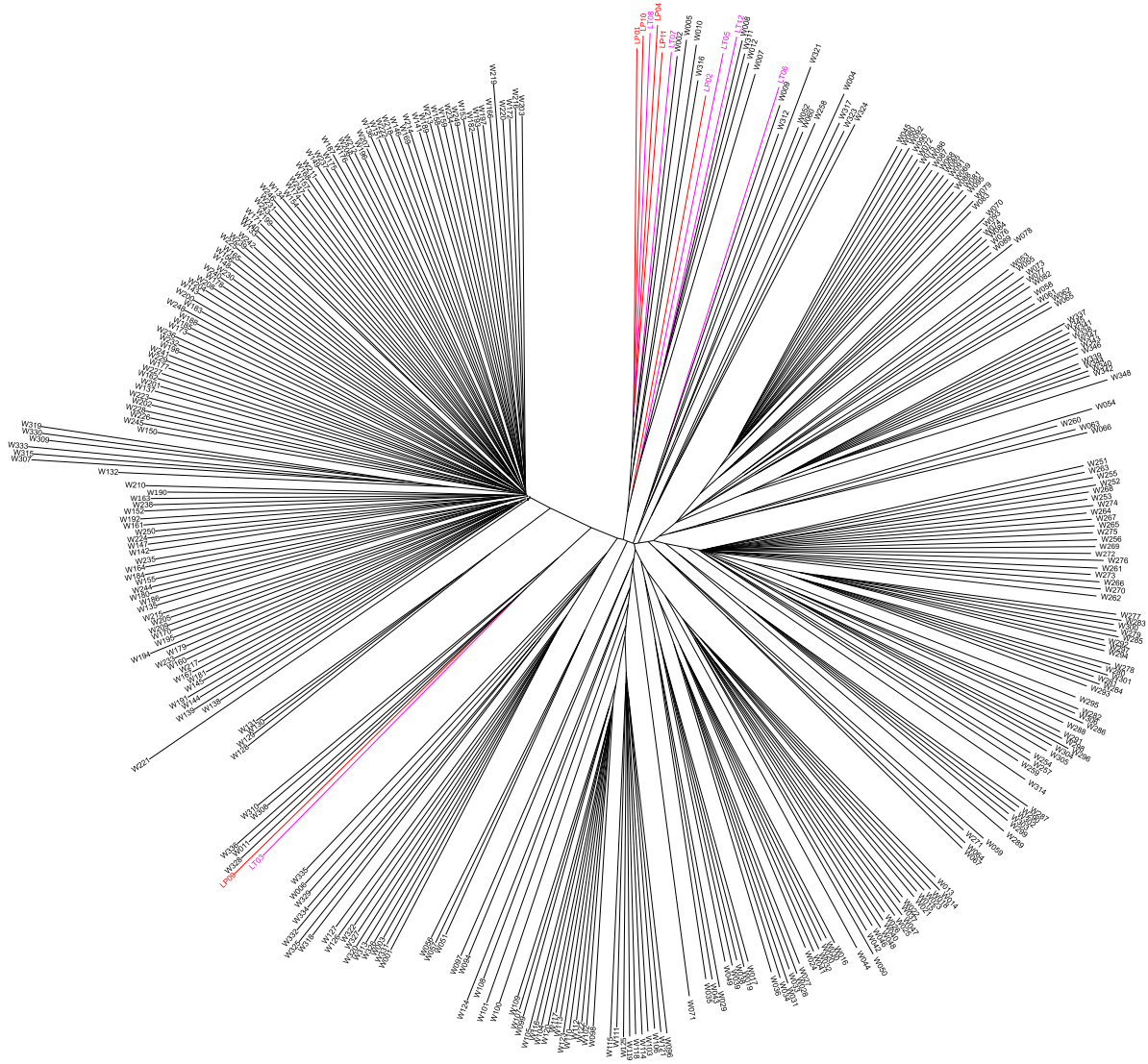


Figure 4.1: Neighbor-joining tree representing tambaquis relationships based on 11,423 SNPs. The genotypes of tambaquis with a total lack of intermuscular bones are represented in pink; genotypes of tambaquis with a partial lack of intermuscular bones are represented in red; genotypes of tambaquis without lack of intermuscular bones are represented in black.

PCA analysis showed similar genetic relationships of those found with neighbor-joining tree (Figure 4.2). The individuals with total or partial lack of intermuscular bones were mixed with many individuals with intermuscular bone. The individuals with total or partial lack of intermuscular bones were in a relatively centralized position among the individuals with of intermuscular bone.

The kinship analysis revealed that most of the full-sib and half-sib relationships within and among tambaquis with total, partial or without lack of intermuscular were evident in the neighbor-joining tree and PCA.

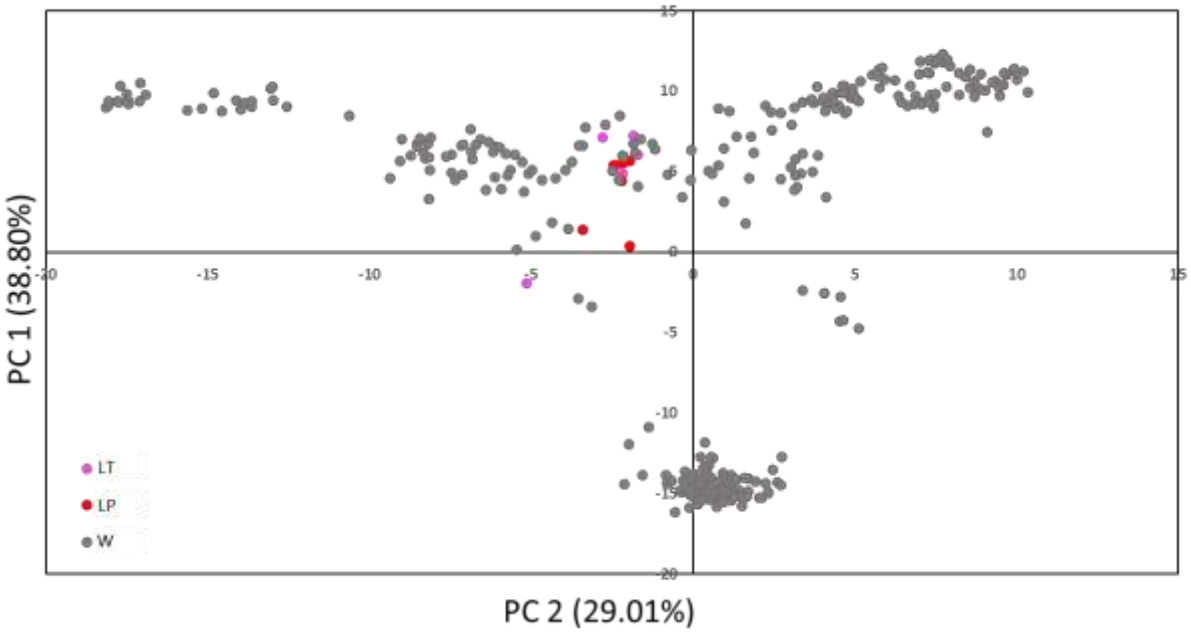


Figure 4.2: Principal component analysis (PCA) of tambaquis based in 11,423 SNPs. The pink dots represent individuals with the total lack (LT) of intramuscular bones; red dots represent individuals with the partial lack (LP) with vestiges of bone in one or both sides; black dots represent individuals without lack (W) of intramuscular bones.

4.3.4. Association and linkage disequilibrium analysis

We identified 675 significant SNPs ($P < 0.03$) associated with lack of intermuscular bone (Figure 4.3). Of these, 109 (16.15%) were successfully aligned against zebrafish genome and 149 (22.07%) were successfully aligned among 23 linkage groups.

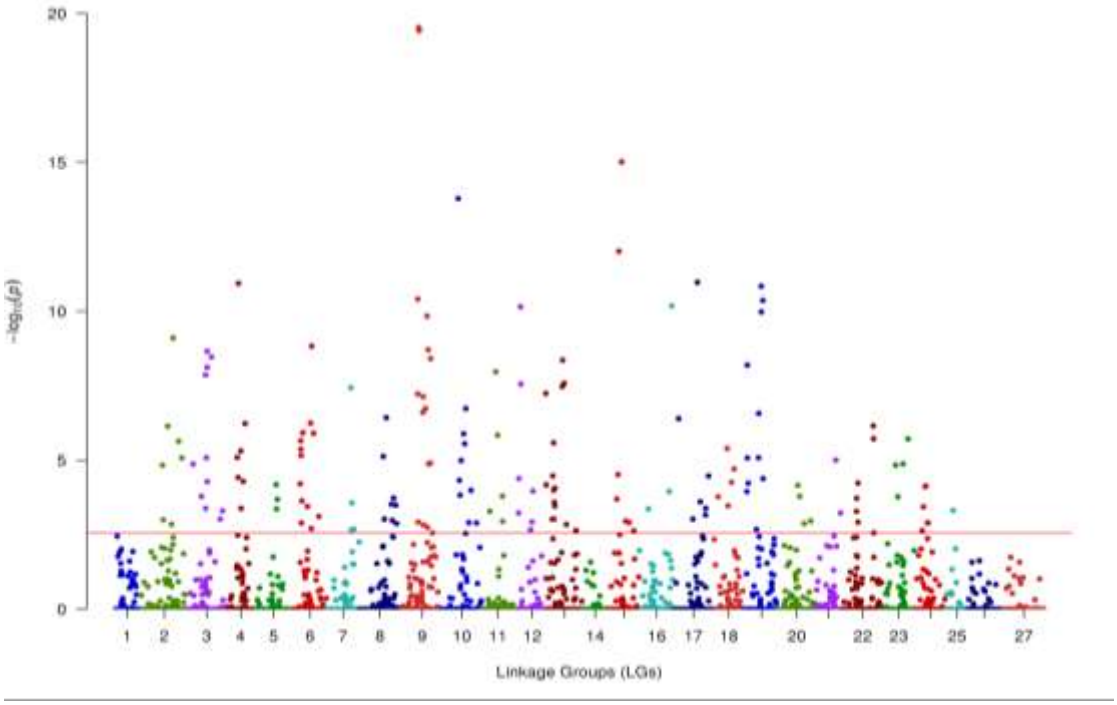


Figure 4.3: Manhattan plot of associated SNPs with lack of intermuscular bone in linkage groups of tambaqui. The y-axis is shown as –log10 (p-value) being that P-value was set to false discovery rate (FDR) correction. The red line indicates genome-wide association (Padj<0.03).

Twenty-nine haplotype blocks, distributed in 15 tambaqui linkage groups were obtained from the SNPs associated with lack of intermuscular bone. The linkage groups 9 and 13 have four haplotype blocks each. The linkage group 9 also showed a higher association with phenotype, their SNPs had padj-value ranging from 2.99E-20 to 1.00E-03.

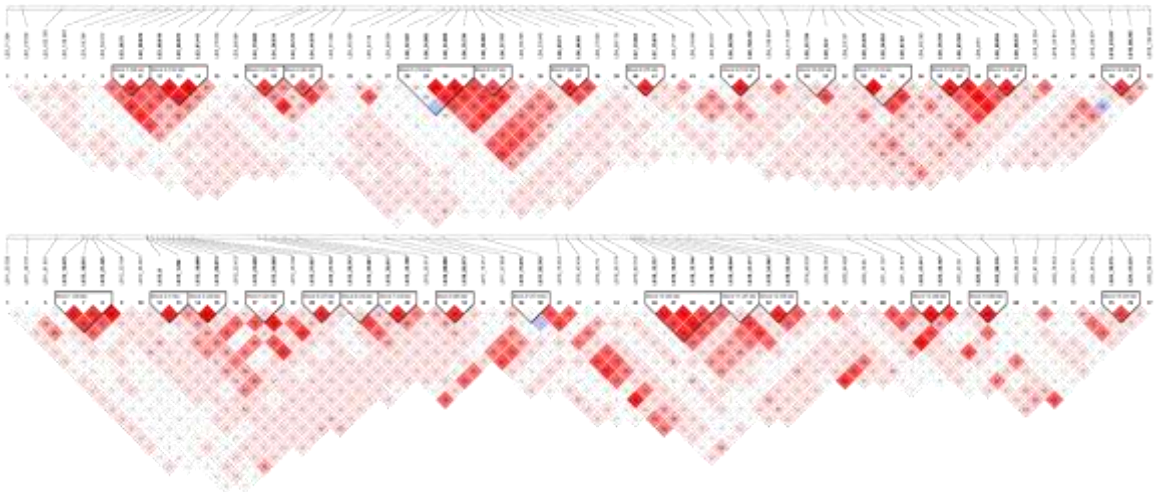


Figure 4.4: Haplotype blocks obtained by the solid spine of LD to lack of intermuscular bone in linkage groups of tambaqui using Haploview 4.2

4.3.5. Allelic Frequencies

We also observed a difference between allelic frequencies and the proportion of heterozygotes among individuals with and without lack of intermuscular bone (APPENDIX D). The frequency of heterozygotes was higher in individuals with lack of intermuscular bone (0.52).

4.3.6. Annotation and pathways analysis

Although we found twenty-nine haplotype blocks in the tambaqui linkage map, VEP annotations found only ten candidate genes within those regions (APPENDIX E). However, in VEP annotations for SNPs aligned with the *Danio rerio* genome we found a total of 109 genes. Considering all annotation categories, the number of annotations were higher (183) than the number of variants. In exon regions, 65% of the variations were classified as missense variant, 33% as a synonymous variant, and 2% as stop gained (Figure 4.5). We investigated in literature the reported function for each gene from the gene list obtained by VEP. We found thirteen genes related with reduction of bone mass, promotion of bone formation, inhibition of bone resorption, central control of bone remodeling, bone mineralization, and others functions (Table 4.1). However, these markers do not align with the tambaqui linkage map, which prevents us from knowing if they are within the twenty-nine haplotype regions.

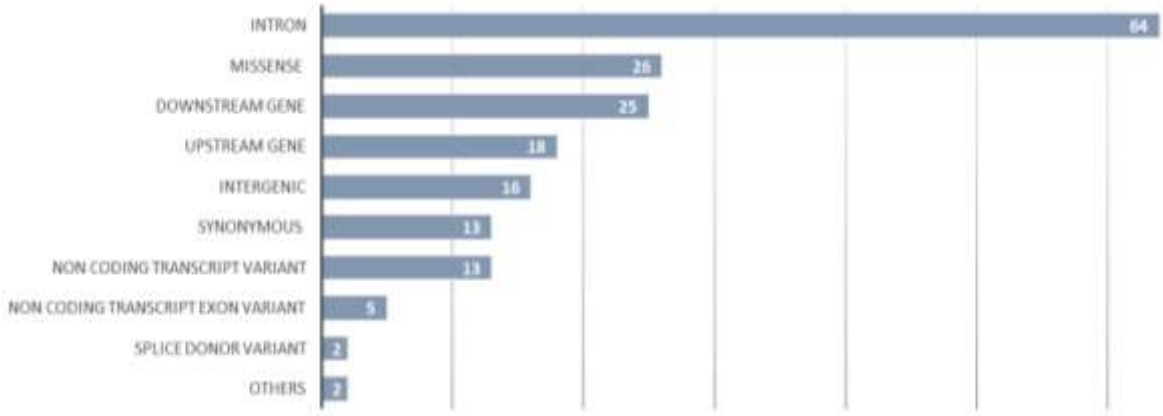


Figure 4.5: Annotation of associated SNPs aligned against the genes using zebrafish as reference genome (ENSEMBL release 84).

Table 4.1: Markers and genes associated with lack of intramuscular bone with a function related to the bone development.

Marker	P-value	Gene	Association in literature	Pathway
TP25478	0.02824	actn3b	Reduced bone mass in human and mouse [20]	Muscle contraction
TP94182	0.01829	adamtsl2	Rare bone dysplasia [21]	-
TP76585	0.00106	atp6v0a1a	Potential inhibitor of bone resorption [22]	Signal transduction; Immune System; Transmembrane transport of small molecules.
TP159881	0.00014	atp6v0ca	Potential inhibitor of bone resorption [22]	Signal transduction; Immune System; Transmembrane transport of small molecules.
TP126617	0.02814	dchs1b	The abnormal ossification of the sternum bone [23]	-
TP13280	1.60E-14	ebf3a	May exhibit a unique functional role during somite development and axial skeletogenesis [24]	-
TP183824	0.00135	efnb1	Promotion of bone formation [25]	Developmental biology
TP121179	0.00018	nmu	Central control of bone remodeling [26]	Signal transduction
TP238461	0.01584	ntn4	Inhibition of osteoclast differentiation in vitro and prevention of bone loss in vivo [27]	Developmental biology
TP20763	2.62E-07	pde4d	Genetic contribution to bone mineral density variation in humans [28]	Signal transduction
TP202945	0.003431	plek	Related with bone loss process in periodontitis [29]	Hemostasis
TP170872	0.009769	wisp1b	Regulation of bone turnover and signaling[30]	-
TP73476	0.00609	xpr1b	Regulation of phosphate balance in cells involved in bone resorption [31]	-

The gene list obtained by VEP was used in the enrichment of biological pathways performed using Reactome. The analysis shows that all genes associated with lack of intermuscular bone were related with 8 pathways (Figure 4.6). The pathways with more interactions with the phenotype were immune system, followed by metabolism and signal transduction pathways. However, these pathways were not inter-connected.

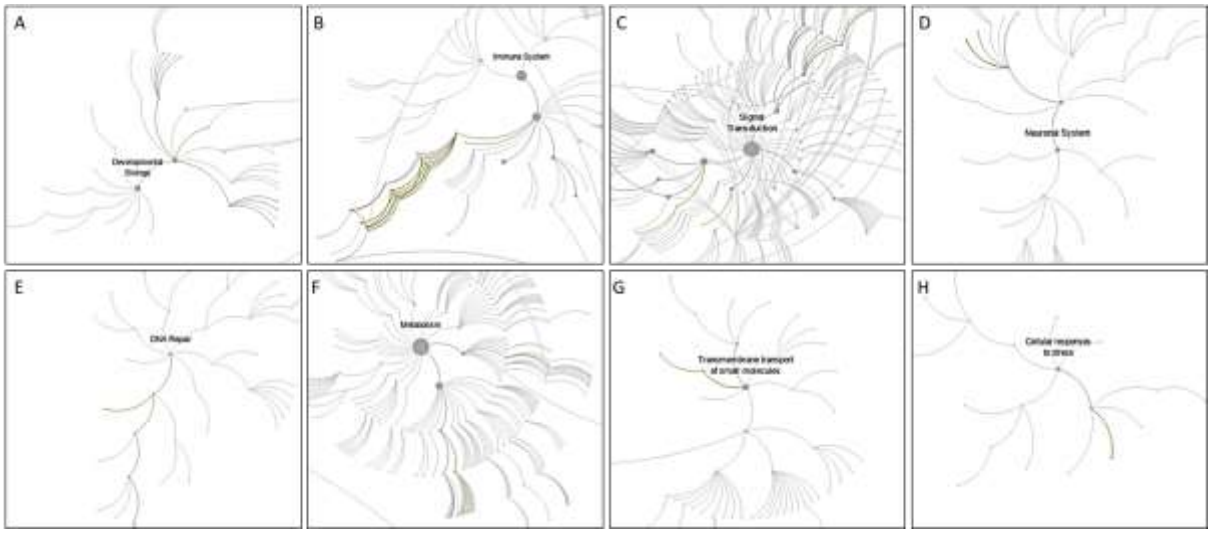


Figure 4.6: Pathways analysis performed with Reactome showing the association between genes related with lack of intermuscular bone and biological pathways.

4.4. Discussion

The presence of intermuscular bone is an exclusive characteristic of teleosts fishes [32]. These bones can be classified into of three groups according to the location of the bone ligament [33]. The epineural bones are attached to the neural arches, while that epicentral bones are attached to the central arch; and epipleural bones are attached to the hemal arches [32]. They are originated from incomplete membranous ossifications of connective tissue in the muscular septum [9] as a response to strains suffered by muscles [10]. Some studies about intermuscular bone already were reported in many fish species [32] [34] [35]. These studies mainly describe the homology of intermuscular bones among these species. An effort was made in attempt to understand genetic mechanisms involved in the development of this trait [36]. In this study, it was shown that the expression of muscle segment homeobox C (MsxC) gene was primarily expressed in the myosepta and brain. MsxC was expressed in the myosepta from 26 to 41 days after hatching, coinciding with the onset of intermuscular bone ossification. However, the lack of a model organism presenting individuals with and without lack of intermuscular bone was limited in this study. In our study, the screening using X-ray and the use of ultrasound pictures as described before[15] allow the identification of individuals with complete or partial lack of intramuscular bones for our study. The discovery of individuals with lack of intermuscular bones makes possible associative studies aiming the understanding of the genetic mechanisms involved in intermuscular bone development.

Genome-wide association studies (GWAS) have contributed significantly in the identification of genomic regions associated with diseases or other phenotypic traits of interest [37,38]. The progress in SNP discovery technology for the model and non-model species made GWAS possible for aquaculture species [39]. Generally, GWAS dataset has a large number of genotyped markers in a large number of individuals [38]. The discovery of only a few individuals with total or partial lack of intermuscular bone in our dataset is a drawback of our analysis. However, the genetic relationships calculated between tambaquis shown in neighbor-joining tree, PCA and kinship revealed a reduced number of clusters with a short distance between these clusters. The presence of strong population structure can result in an elevated false-positive rate [40]. The estimates of population structure generally are used to reduce effects

due to grouping, avoiding spurious associations. In our analysis, the individuals with lack of intermuscular bone do not be grouped into of an exclusive cluster. The formation of an exclusive cluster of individuals lacking intermuscular bone would increase the number of spurious associations, making it difficult to identify really associated markers.

As a direct or indirect consequence of the imbalance in our dataset, many SNPs were associated with lack of intramuscular bone. Despite careful selection of the most appropriate statistical analysis, a large number of spurious associations were expected. In our approach, we used FDR and Bonferroni correction as statistical methods to correct for multiple comparisons. However, we maintained the significance threshold padj<0.03, and we adopted complementary approaches to select genes with functions associated with the phenotype.

Alignment with the zebrafish genome and with the tambaqui linkage map greatly reduced the number of SNPs in 83 .85%, and 77.93% respectively, in order to be investigated in the complementary analysis. The poor performance in alignment may be explained by genetic differences between tambaqui and zebrafish genomes, and by the fact that the tambaqui linkage map has only 7,192 markers while the number of SNPs we tried to align was 11,423. However, the number of SNPs aligned with the tambaqui linkage map was sufficient to detect 29 haplotype blocks with an average length of less than 0.2 cM. The formation blocks grouping may be associated with population structural proximity. Populations with similar structural proximity will likely have similar block structures [41]. In addition to structural proximity, haplotype blocks may originate from recombination hot spots [42] [43] or by stochastic variation in models assuming that recombination is randomly distributed [44].

Among the annotated genes identified by VEP, we identified thirteen genes that were previously reported to affect bone development. Some of these genes also were represented in biological pathways. VEP annotation report an intron variant in the *actn3b* gene. This gene participates in muscle contraction pathway and was associated with low bone mass in humans and mice, with clear disruptions in mineralization and resorption in the mouse model [20]. Our annotation also identified splice donor variant in *adamtsl2* gene. A recently study revealed that mutations in this gene are affecting conserved amino acids residues and causing a rare bone dysplasia in humans [21]. Other genes found in our annotations were *atp6v0a1a* and *atp6v0ca* that encodes V-ATPase subunits. Genetic studies made in mice and humans revealed a critical role for V-ATPase subunits in osteoclast-related diseases including osteopetrosis and osteoporosis [22]. The *xpr1b* is also an important gene found in our study. This gene participates in bone resorption by regulation of phosphate balance in cells involved in bone resorption [31]. The overexpression of *efnb1* gene, also found in this study, can enhance levels of full-length ephrin B1 protein. Modifications in ephrin B1 actions in bone may provide a means to sensitize the skeleton to mechanical strain to stimulate new bone formation [25]. Also we found the *pde4d* gene that accounts for some of the genetic contributions to bone mineral density variation in humans [28]. We also found the NMU gene. It acts in the central nervous system, rather than directly on bone cells, to regulate bone remodeling [45]. A study showed that Nmu-deficient mice have high bone mass owing to an increase in bone formation [26]. In the same study, a treatment of wild-type mice with a natural agonist for the NMU receptor decreased bone mass.

An additional insight into the biological pathways and molecular functions that are affected by these variants was added by Reactome. Some biological pathways were significantly over-represented. For example, the most enriched pathway corresponding to biological processes were related to the regulation of the immune system. It suggests that the immune system can be affected by the genetic modifications that led to the appearance of a lack of intermuscular bone in tambaqui.

To the best of our knowledge, we performed for the first time a GWAS for the lack of intermuscular bone. We have identified many genes that perform important functions in bone development that can be elected as candidate genes associated with this trait to potentially be included in breeding programs in aquaculture area.

References

1. Moav, R., Finkel, A. & Wohlfarth, G. Variability of intermuscular bones, vertebrae, ribs, dorsal fin rays and skeletal disorders in the common carp. Theor. Appl. Genet. 46, 33–43 (1975).
2. Sahu, B. B. et al. Record of Skeletal System and Pin Bones in Table Size Hilsa Tenualosa ilisha (Hamilton, 1822). World J. Fish Mar. Sci. 6, 241–244 (2014).
3. Li, L. et al. Comparative analysis of intermuscular bones in fish of different ploidies. Sci. China Life Sci. 56, 341–350 (2013).
4. Guimarães, I. G. & Martins, G. P. Nutritional requirement of two Amazonian aquacultured fish species, Colossoma macropomum (Cuvier, 1816) and Piaractus brachypomus (Cuvier, 1818): A mini review. J. Appl. Ichthyol. 31, 57–66 (2015).
5. Knight, L. C. & Lesser, T. H. Fish bones in the throat. Arch. Emerg. Med. 6, 13–6 (1989).
6. Lin, M.-P., Chen, Y.-L. & Tzeng, W.-S. Incarceration of a Y-shaped fish bone in the upper thoracic oesophagus. Bmj 2014, 4742 (2014).
7. Shihada, R., Goldsher, M., Sbait, S. & Luntz, M. Three-dimensional computed tomography for detection and management of ingested foreign bodies. Ear. Nose. Throat J. 88, 910–1 (2009).
8. Reidel, A. et al. Evaluation of yield and morphometric characteristics of curimbatá Prochilodus lineatus (Valenciennes, 1836), and piavuçu Leporinus macrocephalus (Garavello & Britski, 1988) males and females. Rev. Varia Sci. 4, 71–78 (1988).
9. Danos, N. & Ward, A. B. The homology and origins of intermuscular bones in fishes: Phylogenetic or biomechanical determinants? Biol. J. Linn. Soc. 106, 607–622 (2012).
10. Nursall, J. R. The lateral musculature and the swimming of fish. Proc. Zool. Soc. London 126, 127–144 (1956).
11. Sfakiotakis, M., Lane, D. M. & Davies, J. B. C. Review of fish swimming modes for aquatic locomotion. IEEE J. Ocean. Eng. 24, 237–252 (1999).
12. Stachowicz, J. J. Mutualism, Facilitation, and the Structure of Ecological Communities. Bioscience 51, 235 (2001).
13. Araujo-Lima, C. & Goulding, M. So Fruitful a Fish: Ecology, Conservation and Aquaculture of of the Amazon's Tambaqui. (Columbia Univ Press, 1997).
14. Instituto Brasileiro de Geografia e Estatística - IBGE. Produção da Pecuária Municipal. IBGE 42, (2014).
15. Perazza, C. A., de Menezes, J. T. B., Ferraz, J. B. S. & Hilsdorf, A. W. S. Lack of intermuscular bones in specimens of Colossoma macropomum: An unusual phenotype to be incorporated into genetic improvement programs. Aquaculture (2016). doi:10.1016/j.aquaculture.2016.05.014
16. Wringe, B. F. et al. Growth-related quantitative trait loci in domestic and wild rainbow trout (Oncorhynchus mykiss). BMC Genet. 11, 63 (2010).
17. Gutierrez, A. P. et al. Genome-Wide Association Study (GWAS) for Growth Rate and Age at Sexual Maturation in Atlantic Salmon (Salmo salar). PLoS One 10, e0119730 (2015).
18. Andersson, L. Genome-wide association analysis in domestic animals: A powerful approach for genetic dissection of trait loci. Genetica 136, 341–349 (2009).

19. Lind, C. E., Ponzoni, R. W., Nguyen, N. H. & Khaw, H. L. Selective breeding in fish and conservation of genetic resources for aquaculture. Reprod. Domest. Anim. 47, 255–263 (2012).
20. Yang, N. et al. α-Actinin-3 deficiency is associated with reduced bone mass in human and mouse. Bone 49, 790–798 (2011).
21. Ben-Salem, S., Hertecant, J., Al-Shamsi, A. M., Ali, B. R. & Al-Gazali, L. Novel mutations in ADAMTSL2 gene underlying geleophysic dysplasia in families from United Arab Emirates. Birth Defects Res. Part A - Clin. Mol. Teratol. 97, 764–769 (2013).
22. Qin, A. et al. V-ATPases in osteoclasts: Structure, function and potential inhibitors of bone resorption. International Journal of Biochemistry and Cell Biology 44, 1422–1435 (2012).
23. Mao, Y. et al. Characterization of a Dchs1 mutant mouse reveals requirements for Dchs1-Fat4 signaling during mammalian development. Development 138, 947–57 (2011).
24. El-Magd, M. A., Allen, S., McGonnell, I., Otto, A. & Patel, K. Bmp4 regulates chick Ebf2 and Ebf3 gene expression in somite development. Dev. Growth Differ. 55, n/a-n/a (2013).
25. Cheng, S. et al. Transgenic Overexpression of Ephrin B1 in Bone Cells Promotes Bone Formation and an Anabolic Response to Mechanical Loading in Mice. PLoS One 8, e69051 (2013).
26. Sato, S. et al. Central control of bone remodeling by neuromedin U. Nat. Med. 13, 1234–1240 (2007).
27. Enoki, Y. et al. Netrin-4 derived from murine vascular endothelial cells inhibits osteoclast differentiation in vitro and prevents bone loss in vivo. FEBS Lett. 588, 2262–2269 (2014).
28. Reneland, R. H. et al. Association between a variation in the phosphodiesterase 4D gene and bone mineral density. BMC Med. Genet. 6, 9 (2005).
29. Song, L., Yao, J., He, Z. & Xu, B. Genes related to inflammation and bone loss process in periodontitis suggested by bioinformatics methods. BMC Oral Health 15, 105 (2015).
30. Maeda, A. et al. WNT1-Induced Secreted Protein-1 (WISP1), a novel regulator of bone turnover and Wnt signaling. J. Biol. Chem. 290, 14004–14018 (2015).
31. Meireles, A. M. et al. The phosphate exporter xpr1b is required for differentiation of tissue-resident macrophages. Cell Rep. 8, 1659–1667 (2014).
32. Patterson, C. & Johnson, G. D. The intermuscular bones and ligaments of teleostean fishes. Smithson. Contrib. to Zool. 1–83 (1995). doi:10.5479/si.00810282.559
33. Owen, R. & Owen, R. Lectures on the comparative anatomy and physiology of the vertebrate animals, delivered at the Royal College of Surgeons of England, in 1844 and 1846. By Richard Owen ... Part I.--Fishes ... (Printed for Longman, Brown, Green, and Longmans, 1846). doi:10.5962/bhl.title.13539
34. Gemballa, S. Homology of intermuscular bones in acanthomorph fishes. AmMusNov (1998).
35. Gemballa, S. et al. Evolutionary transformations of myoseptal tendons in gnathostomes. Proc. Biol. Sci. 270, 1229–1235 (2003).
36. Lv, Y. P. P., Yao, W. J. J., Chen, J. & Bao, B. L. L. Newly identified gene muscle segment homeobox C may play a role in intermuscular bone development of Hemibarbus labeo. 14, 11324–11334 (2015).
37. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6, 95–108 (2005).
38. Yue, G. H. Recent advances of genome mapping and marker-assisted selection in aquaculture. Fish Fish. 15, 1–21 (2013).

39. Huete-Pérez, J. A. & Quezada, F. Genomic approaches in marine biodiversity and aquaculture. Biological Research 46, 353–361 (2013).

40. Brzyski, D. et al. Controlling the rate of GWAS false discoveries. bioRxiv (2016). doi:10.1101/058230

41. Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. The American Journal of Human Genetics 71, (2002).

42. Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. The American Journal of Human Genetics 71, (2002).

43. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. Nat. Genet. 29, 229–232 (2001).

44. Subrahmanyan, L., Eberle, M. A., Clark, A. G., Kruglyak, L. & Nickerson, D. A. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am. J. Hum. Genet. 69, 381–95 (2001).

45. Teitelbaum, S. L. & Ross, F. P. Genetic regulation of osteoclast development and function. Nat. Rev. Genet. 4, 638–49 (2003).

46. Elshire, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379 (2011).

47. Lu, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9, e1003215 (2013).

48. Glaubitz, J. C. et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9, e90346 (2014).

49. Ringnér, M. What is principal component analysis? Nat. Biotechnol. 26, 303–304 (2008).

50. Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, (2007).

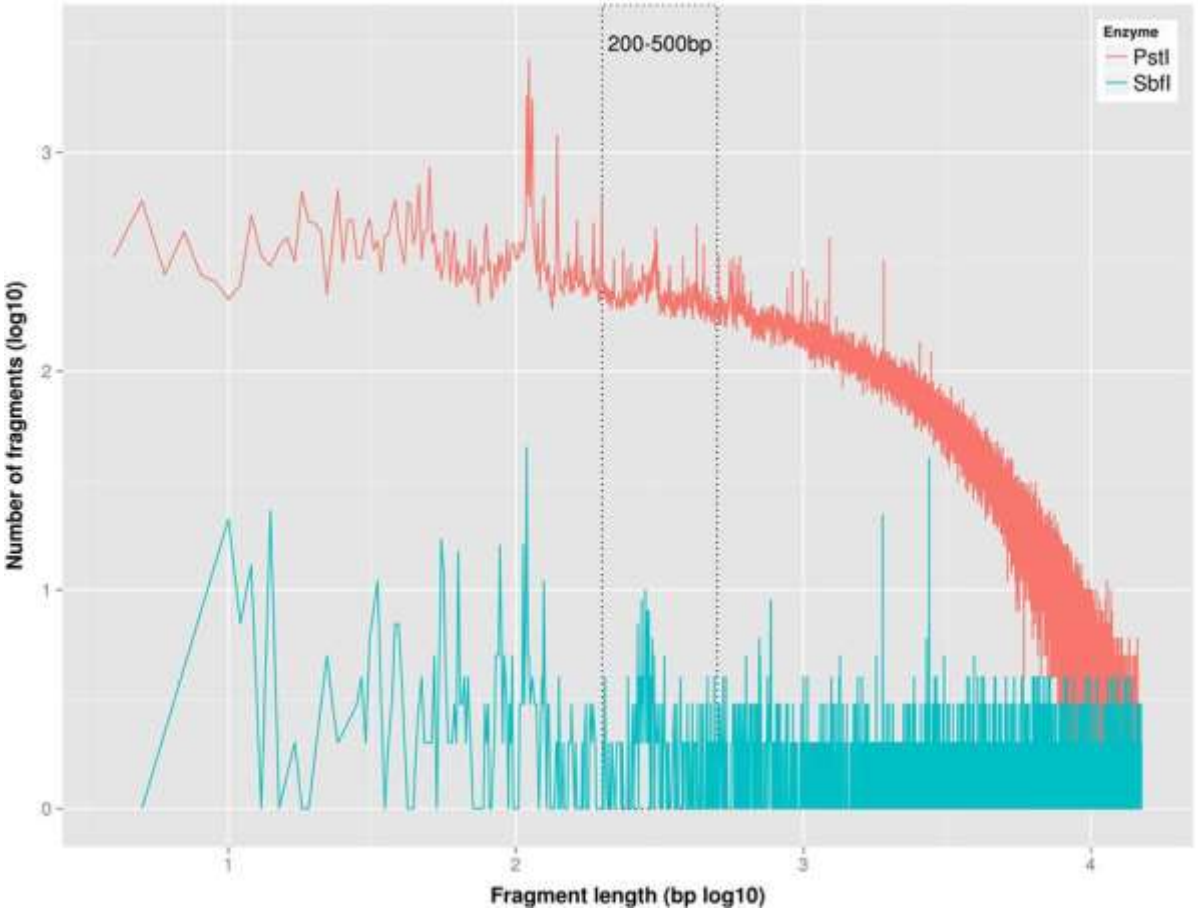
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).

52. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070 (2010).

53. Croft, D. et al. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–D697 (2011).

APPENDIX

APPENDIX A. Fragment size distribution generated by *PstI* and *SbfI* in silico digestion of zebrafish genome. Fragment length ranging between 200-500 bps are showed within the dotted box.



APPENDIX B. Summary of single end sequenced reads from 126 individuals before trimmed. Number of bases (N.Bases), unique reads (U.Reads), total reads (T. Reads) and depth of sequencing (Depth).

Fish ID	N. Bases	U. Reads	T. Reads	Depth
F	27797879	458015	4665280	10.18586728
M	16351514	265852	1862856	7.007116742
1	13229406	213522	1455593	6.817063347
2	20474889	332424	2894090	8.706020023
3	12249077	196366	1106784	5.636332155
4	16389859	264879	1988642	7.507737495
5	15995120	258112	1997680	7.739585916
6	8984984	143577	629860	4.386914339
7	14485001	233322	1551299	6.648747225
8	14241901	229270	1393661	6.078688882
9	19273438	310826	2347341	7.551945461
10	24942431	405649	3729130	9.192996901
11	19408108	313707	2533389	8.075653396
12	23545428	383232	3529934	9.210958375
13	25493318	413713	4145924	10.02125628
14	27506128	446127	4257847	9.544024459
15	24763897	401318	3812517	9.499990033
16	20030082	324044	2633015	8.125486045
17	21488239	347514	3008678	8.657717387
18	17842027	288797	2283856	7.908170791
19	19104311	309998	2564728	8.273369506
20	18184331	294498	2310584	7.845839361
21	11851401	191486	1204598	6.290788883
22	14533994	233846	1400543	5.989168085
23	17036389	275826	2067956	7.497320775
24	14988231	241908	1643447	6.793686029
25	18124365	293958	2012091	6.844824771
26	19471481	316430	2562465	8.098046961
27	19198645	311229	2489857	8.000080327
28	21600223	351282	3080954	8.770600258
29	23550956	384676	3468717	9.017243082
30	25557880	417110	3664585	8.785656062
31	12986927	209181	980080	4.685320368
32	18310843	297655	2238331	7.519883758
33	21693787	355553	4509692	12.68359991
34	13805501	222824	1502830	6.744470973
35	17732221	290035	2735463	9.431492751
36	9514094	153041	824393	5.386746035
37	7476527	119761	453048	3.782934344
38	20339966	334640	3523292	10.52860387
39	20035411	329201	3465505	10.52701845
40	16529747	270169	2365574	8.755904637
41	13378519	217786	1711813	7.860069059

42	11352644	184246	1250151	6.785227359
43	12893703	209759	1542471	7.353539061
44	15961447	261187	2420440	9.267076845
45	11800190	191438	1336028	6.978907009
46	12089584	196240	1438820	7.331940481
47	12415007	201783	1582919	7.844659857
48	14424796	235810	2035885	8.633582121
49	14844871	241992	2142064	8.851796754
50	16514501	270570	2582490	9.544628008
51	9695288	156135	831935	5.328305633
52	17785910	291796	2702059	9.260096095
53	10763408	175196	1202502	6.863752597
54	18969618	311928	2970657	9.523534277
55	16091074	264293	2361658	8.935756906
56	13354776	218504	1793516	8.208160949
57	10701308	173994	1186342	6.818292585
58	9658982	157413	827452	5.25656712
59	15289125	251226	2318091	9.227114232
60	15727871	258032	2337567	9.059213586
61	12683115	206785	1322561	6.395826583
62	22037332	366575	3478448	9.489048626
63	17889741	293897	2813865	9.574323658
64	15826224	258813	2145831	8.291047977
65	11266380	181817	1124257	6.183453692
66	12050496	195146	1244028	6.374857799
67	9908306	159406	859110	5.389445818
68	11209629	180765	1060320	5.865737283
69	9101450	145980	721678	4.943677216
70	6922488	110602	410020	3.707166236
71	8690721	139357	639759	4.590791995
72	15073741	244742	1831572	7.48368486
73	13475140	217961	1428803	6.555314942
74	12730277	205518	1252806	6.095845619
75	13083592	211639	1434877	6.77983264
76	14759250	240708	1533404	6.370390681
77	12979181	209677	1370131	6.534483992
78	18320958	297941	2248144	7.54560131
79	11661936	187760	1045516	5.568363869
80	12178777	196525	1220756	6.211708434
81	14522411	234718	1540696	6.564030028
82	16987622	277271	2214721	7.987568119
83	14005294	227431	1524520	6.703219878
84	15858796	258211	1977191	7.657268668
85	11865570	191521	1108232	5.786477723
86	7031380	118552	2792042	23.55120116
87	12775187	211110	3253091	15.40945952

88	7292929	123511	2348076	19.01106784
89	23233128	383872	9820085	25.58166524
90	23168497	380309	3681836	9.681169786
91	12616946	207978	3414352	16.41689025
92	29843059	499987	11168262	22.33710476
93	28975725	480783	6583681	13.69366429
94	36526115	602878	6018790	9.983429483
95	24346665	400231	3626870	9.061941729
96	25095740	413292	5572160	13.4823805
97	21385709	349037	2665469	7.636637377
98	13549940	218736	1078062	4.928598859
99	23418280	384036	3326273	8.661357269
100	19280385	313724	2203686	7.02428249
101	25151809	413379	3734147	9.033228587
102	20919794	342645	2765961	8.072381036
103	14513534	234943	1406911	5.988307802
104	12665279	205453	1180426	5.745479501
105	12598758	203900	1090542	5.34841589
106	24607616	406336	3592680	8.841648291
107	22737027	373982	3174985	8.4896733
108	10748840	173604	744757	4.289976037
109	14613125	237048	1267608	5.347473929
110	19520931	321481	2652775	8.251731829
111	10501244	170006	811585	4.773860923
112	18232856	300017	2188085	7.293203385
113	16897376	277134	1989924	7.18036762
114	16344347	267253	1871675	7.003382563
115	2141348	34230	53714	1.569208297
116	13451556	218812	1096974	5.013317368
117	10836684	175319	830239	4.735590552
118	10078175	162791	780475	4.794337525
119	11712430	189113	1162054	6.144760011
120	15310186	248858	1860223	7.475037973
121	13567861	219463	1427019	6.502321576
122	17414135	283157	2262221	7.989281565
123	14101210	228204	1599462	7.008913078
124	11813015	191750	1075586	5.609314211

APPENDIX C. BLAST annotation of the loci identified by Lositan and BayeScan as potentially under selection for tambaqui.

Selected loci	Environmental correlation Samβada	Gene	Function	Organism associated
TP101941	Annual sunshine	-	-	*Homo sapiens, Apteryx australis mantelli*
TP10330	-	cd81	Antigen (target of antiproliferative antibody 1)	*Pygocentrus nattereri, Cyprinus carpio, Mus musculus*
TP116573	Temperature Atmosferic pressure	rorb	Nuclear receptor	*Heligmosomoides polygyrus, Cyprinus carpio, Danio rerio, Bubalus bubalis*
TP128622	-	mcc1	Candidate tumor suppressor gene	*Homo sapiens, Danio rerio, Pan troglodytes*
TP131461	-	trim54	Muscle-specific RING finger protein	*Pygocentrus nattereri, Homo sapiens, Soboliphyme baturini, Ovis canadensis canadensis, Mus musculus*
TP151405	-	bod1l1	Biorientation of chromosomes in cell division	*Cyprinus carpio, Danio rerio, Sus scrofa, Ovis canadensis canadensis*
TP160108		-	-	*Drosophila melanogaster*
TP171581	Annual sunshine	nol6	Nucleolar RNA-Associated Protein	*Pygocentrus nattereri, Scleropages formosus, Ovis canadensis canadensis, Drosophila melanogaster, Otolemur garnettii*
TP171655	-	BAC clone RP11	BAC clone	*Homo sapiens, Ovis canadensis canadensis, Mus musculus, Macaca fascicularis, Pan troglodytes, Macaca mulata*
TP175199	-	LOC109118167	Fanconi anemia group A protein-like	*Mus musculus, Fukomys damarensis, Microcebus murinus, Oncideres albomarginata, Chrysochloris asiatica*
TP20640	Temperature Atmosferic pressure	LOC107809337	ncRNA	*Nicotiana tabacum*
TP21835	Temperature Atmosferic pressure	tnfsf13b	Coding of cytokine that belongs to the tumor necrosis factor	*Cyprinus carpio, Pygocentrus nattereri, Ovis canadensis canadensis,*
TP31591	-	AYM39_00100	Transporter ATP-binding protein	*Methylomonas sp., Trichobilharzia regenti*
TP31890	-	sec16a	Endoplasmic reticulum export factor	*Ictalurus punctatus, Brachypodium distachyon*
TP48275	-	-	Promoter (TATA_box)	*Oryzias latipes, Takifugu rubripes, Angiostrongylus cantonensis*
TP81620	-	pde6d	Encodes the delta subunit of rod-specific photoreceptor phosphodiesterase	*Oryctolagus cuniculus, Gemmatimonadetes bacterium, Conexibacter woesei*
TP82108	Temperature	-	-	*Dicrocoelium dendriticum*
TP9499	-	rnf10	Ring finger protein	*Pygocentrus nattereri, Lepidothrix coronata, Mus musculus*

APPENDIX D. Allelic frequencies and the proportion of heterozygotes among individuals with and without lack of intermuscular bone.

Site Name	Major Allele		Major Allele Frequency		Minor Allele		Minor Allele Frequency		Number Heterozygous		Proportion Heterozygous	
	BONE	BONELESS	BONE	BONELESS	BONE	BONELESS	BONE	BONELESS	BONE	BONELESS	BONE	BONELESS
TP3511	G	A	0.81	0.67	A	G	0.19	0.33	67.00	4.00	0.20	0.33
TP5898	G	G	0.98	0.54	A	A	0.02	0.46	16.00	11.00	0.05	0.92
TP6999	G	G	0.76	0.63	A	A	0.24	0.38	159.00	9.00	0.46	0.75
TP8811	C	G	0.71	0.92	G	C	0.29	0.08	81.00	2.00	0.24	0.17
TP11574	C	C	0.85	0.58	T	T	0.15	0.42	103.00	8.00	0.30	0.67
TP13280	C	C	0.98	0.67	T	T	0.02	0.33	9.00	8.00	0.03	0.67
TP20763	A	A	0.87	0.67	G	G	0.13	0.33	62.00	8.00	0.19	0.67
TP24660	C	A	0.85	0.50	A	C	0.15	0.50	97.00	12.00	0.30	1.00
TP25478	C	C	0.96	0.75	T	T	0.04	0.25	19.00	6.00	0.06	0.50
TP27854	T	T	0.97	0.67	A	A	0.03	0.33	10.00	8.00	0.03	0.67
TP34042	A	G	0.69	0.58	G	A	0.31	0.42	75.00	8.00	0.26	0.67
TP34875	G	G	0.85	0.57	A	A	0.15	0.43	101.00	2.00	0.29	0.29
TP36329	G	A	0.80	0.58	A	G	0.20	0.42	108.00	4.00	0.32	0.33
TP36378	A	A	0.90	0.63	T	T	0.10	0.38	71.00	9.00	0.20	0.75
TP38495	C	T	0.77	0.63	T	C	0.23	0.38	135.00	5.00	0.39	0.42
TP39241	G	G	0.89	0.75	A	A	0.11	0.25	51.00	6.00	0.16	0.50
TP40101	A	A	0.50	0.58	G	G	0.50	0.42	72.00	6.00	0.21	0.50
TP40466	C	C	0.89	0.79	T	T	0.11	0.21	69.00	5.00	0.20	0.42
TP42850	T	T	0.97	0.75	C	C	0.03	0.25	14.00	4.00	0.04	0.33
TP52423	C	G	0.82	0.54	G	C	0.18	0.46	83.00	9.00	0.26	0.75
TP59706	G	A	0.73	0.50	A	G	0.27	0.50	168.00	8.00	0.48	0.67
TP67282	C	T	0.79	0.58	T	C	0.21	0.42	128.00	6.00	0.37	0.50
TP67781	A	A	0.58	0.83	G	G	0.42	0.17	61.00	4.00	0.21	0.33
TP68840	G	G	0.86	0.58	A	A	0.14	0.42	93.00	10.00	0.27	0.83
TP70266	A	G	0.85	0.58	G	A	0.15	0.42	51.00	4.00	0.15	0.33
TP72351	G	G	0.86	0.71	A	A	0.14	0.29	94.00	5.00	0.27	0.42
TP73476	T	T	0.87	0.54	A	A	0.13	0.46	87.00	11.00	0.26	0.92
TP76585	C	T	0.78	0.58	T	C	0.22	0.42	124.00	6.00	0.36	0.50
TP77927	T	T	0.90	0.79	C	C	0.10	0.21	62.00	5.00	0.18	0.42
TP78633	C	C	0.89	0.83	A	A	0.11	0.17	74.00	4.00	0.22	0.33
TP78842	A	A	0.89	0.71	G	G	0.11	0.29	67.00	7.00	0.20	0.58
TP80260	C	T	0.80	0.54	T	C	0.20	0.46	105.00	5.00	0.31	0.42
TP83289	G	A	0.79	0.54	A	G	0.21	0.46	57.00	7.00	0.18	0.58
TP85032	C	C	0.81	0.54	T	T	0.19	0.46	110.00	9.00	0.32	0.75
TP88991	G	G	0.85	0.67	A	A	0.15	0.33	85.00	6.00	0.25	0.50
TP91172	C	G	0.81	0.54	G	C	0.19	0.46	118.00	5.00	0.36	0.42
TP94182	C	T	0.64	0.83	T	C	0.36	0.17	126.00	2.00	0.37	0.17
TP94651	T	T	0.94	0.67	C	C	0.06	0.33	30.00	8.00	0.10	0.67

TP100065	T	T	0.91	0.58	C	C	0.09	0.42	44.00	10.00	0.13	0.83
TP100773	G	A	0.86	0.58	A	G	0.14	0.42	90.00	6.00	0.26	0.50
TP100800	C	A	0.66	0.63	A	C	0.34	0.38	83.00	3.00	0.26	0.25
TP109672	C	T	0.60	0.54	T	C	0.40	0.46	124.00	5.00	0.36	0.42
TP111413	T	A	0.83	0.58	A	T	0.17	0.42	82.00	8.00	0.24	0.67
TP115690	A	A	0.95	0.67	G	G	0.05	0.33	31.00	6.00	0.09	0.50
TP116678	G	G	0.96	0.67	A	A	0.04	0.33	29.00	8.00	0.08	0.67
TP121179	C	C	0.84	0.63	T	T	0.16	0.38	110.00	7.00	0.32	0.58
TP122043	G	G	0.79	0.54	A	A	0.21	0.46	134.00	5.00	0.39	0.42
TP125121	C	T	0.72	0.54	T	C	0.28	0.46	179.00	7.00	0.51	0.58
TP126617	C	C	0.83	0.58	T	T	0.17	0.42	83.00	4.00	0.24	0.33
TP126764	G	G	0.80	0.75	C	C	0.20	0.25	117.00	2.00	0.35	0.17
TP130504	A	A	0.96	0.71	G	G	0.04	0.29	27.00	7.00	0.08	0.58
TP131758	C	C	0.83	0.63	T	T	0.17	0.38	101.00	5.00	0.30	0.42
TP136005	T	A	0.80	0.50	A	T	0.20	0.50	109.00	10.00	0.32	0.83
TP136223	A	A	0.88	0.54	G	G	0.12	0.46	78.00	9.00	0.22	0.75
TP136258	T	T	0.81	0.54	C	C	0.19	0.46	116.00	9.00	0.34	0.75
TP141369	C	C	0.85	0.54	T	T	0.15	0.46	98.00	11.00	0.28	0.92
TP147166	C	C	0.89	0.50	T	T	0.11	0.50	55.00	6.00	0.16	0.50
TP150629	T	T	0.83	0.63	G	G	0.17	0.38	103.00	5.00	0.30	0.42
TP152016	A	A	0.96	0.67	G	G	0.04	0.33	23.00	4.00	0.07	0.33
TP159881	A	A	0.89	0.50	T	T	0.11	0.50	48.00	10.00	0.14	0.83
TP162772	A	C	0.57	0.83	C	A	0.43	0.17	192.00	4.00	0.57	0.33
TP162910	T	T	0.95	0.71	C	C	0.05	0.29	20.00	7.00	0.06	0.58
TP163079	G	A	0.63	0.71	A	G	0.37	0.29	196.00	3.00	0.57	0.25
TP164396	C	C	0.94	0.64	T	T	0.06	0.36	27.00	6.00	0.08	0.55
TP165619	G	G	0.88	0.79	A	A	0.12	0.21	82.00	5.00	0.24	0.42
TP166069	G	G	0.97	0.71	A	A	0.03	0.29	23.00	7.00	0.07	0.58
TP166329	G	C	0.81	0.63	C	G	0.19	0.38	88.00	7.00	0.28	0.58
TP166772	G	G	0.84	0.50	T	T	0.16	0.50	87.00	4.00	0.26	0.33
TP170737	C	C	0.97	0.71	T	T	0.03	0.29	19.00	7.00	0.05	0.58
TP170872	C	C	0.80	0.50	T	T	0.20	0.50	121.00	6.00	0.35	0.50
TP171494	A	A	0.89	0.54	G	G	0.11	0.46	78.00	11.00	0.23	0.92
TP174453	G	G	0.94	0.67	A	A	0.06	0.33	41.00	8.00	0.12	0.67
TP175005	T	C	0.58	0.92	C	T	0.42	0.08	123.00	2.00	0.37	0.17
TP183541	G	G	0.87	0.67	A	A	0.13	0.33	75.00	8.00	0.24	0.67
TP183824	G	A	0.72	0.58	A	G	0.28	0.42	133.00	6.00	0.38	0.50
TP187105	A	A	0.87	0.71	G	G	0.13	0.29	91.00	7.00	0.26	0.58
TP18767	C	C	0.88	0.88	T	T	0.12	0.13	76.00	3.00	0.22	0.25

0												
TP192617	G	G	0.76	0.54	C	C	0.24	0.46	89.00	9.00	0.27	0.75
TP194131	G	G	0.85	0.58	A	A	0.15	0.42	88.00	6.00	0.26	0.50
TP197438	C	T	0.57	0.88	T	C	0.43	0.13	147.00	3.00	0.43	0.25
TP199795	T	G	0.70	0.54	G	T	0.30	0.46	79.00	7.00	0.25	0.58
TP200097	C	C	0.86	0.54	T	T	0.14	0.46	86.00	9.00	0.25	0.75
TP201331	G	G	0.87	0.55	C	C	0.13	0.45	73.00	2.00	0.21	0.18
TP202177	T	T	0.94	0.67	C	C	0.06	0.33	39.00	8.00	0.11	0.67
TP202569	C	C	0.86	0.58	T	T	0.14	0.42	83.00	6.00	0.24	0.50
TP202729	G	G	0.94	0.71	A	A	0.06	0.29	20.00	5.00	0.07	0.42
TP202945	A	C	0.73	0.75	C	A	0.27	0.25	103.00	6.00	0.31	0.50
TP203236	T	C	0.88	0.54	C	T	0.12	0.46	81.00	5.00	0.23	0.42
TP205341	C	A	0.68	0.67	A	C	0.32	0.33	111.00	4.00	0.33	0.33
TP206429	A	A	0.92	0.54	T	T	0.08	0.46	32.00	7.00	0.09	0.58
TP206557	T	T	0.93	0.58	A	A	0.07	0.42	32.00	6.00	0.09	0.50
TP206668	C	A	0.67	0.58	A	C	0.33	0.42	203.00	6.00	0.58	0.50
TP207209	G	G	0.84	0.67	C	C	0.16	0.33	79.00	6.00	0.25	0.50
TP208178	G	G	0.95	0.63	A	A	0.05	0.38	25.00	7.00	0.08	0.58
TP210342	G	G	0.92	0.58	A	A	0.08	0.42	41.00	4.00	0.12	0.33
TP211530	G	G	0.86	0.71	A	A	0.14	0.29	71.00	7.00	0.21	0.58
TP217012	A	A	0.72	0.50	G	G	0.28	0.50	194.00	12.00	0.56	1.00
TP217295	G	A	0.68	0.67	A	G	0.32	0.33	133.00	6.00	0.39	0.50
TP218246	A	G	0.58	0.79	G	A	0.42	0.21	178.00	5.00	0.51	0.42
TP220450	C	C	0.98	0.67	T	T	0.02	0.33	13.00	8.00	0.04	0.67
TP228711	T	T	0.97	0.75	C	C	0.03	0.25	15.00	6.00	0.04	0.50
TP229075	A	G	0.73	0.67	G	A	0.27	0.33	91.00	8.00	0.29	0.67
TP230826	A	C	0.91	0.63	C	A	0.09	0.38	26.00	1.00	0.08	0.08
TP230937	C	T	0.82	0.71	T	C	0.18	0.29	87.00	5.00	0.26	0.42
TP232881	G	G	0.88	0.63	A	A	0.12	0.38	74.00	7.00	0.21	0.58
TP234348	G	G	0.95	0.71	A	A	0.05	0.29	23.00	7.00	0.07	0.58
TP234852	A	A	0.84	0.58	G	G	0.16	0.42	104.00	10.00	0.30	0.83
TP237156	G	G	0.56	0.54	A	A	0.44	0.46	43.00	5.00	0.13	0.42
TP238461	G	G	0.88	0.75	A	A	0.12	0.25	77.00	4.00	0.22	0.33
Average	-	-	0.83	0.64	-	-	0.17	0.36	81.82	6.34	0.24	0.53

APPENDIX E. Haplotype blocks obtained by solid spine of LD in tambaqui using Haploview 4.2, and genes associated with lack of intramuscular bone using VEP.

Haplotype Block Nº	Block Size (cM)	Multiallelic Dprime	Haplotype	Haplotype Frequency	Markers	Genes
1	0.058	0.58	TA	0.766	TP85981	-
			CG	0.193	TP213117	
			CA	0.021		
			TG	0.02		
2	0.122	0.21	GCA	0.753	TP88391	-
			ATG	0.121	TP112285	
			GCG	0.112	TP10536	
3	0.036	52	AC	0.781	TP160634	-
			GA	0.133	TP48305	
			AA	0.069		
			GC	0.019		
4	0.044	0.36	CG	0.686	TP52139	-
			CA	0.191	TP75527	
			TA	0.105		
			TG	0.019		
5	0.104	0.59	TCGT	0.708	TP61469	ngfb
			CCGT	0.115	TP141369	
			TTAC	0.135	TP68840	
			TTGT	0.015	TP43384	
			TCGC	0.014		
6	0.037	0.64	GC	0.765	TP13271	-
			GT	0.082	TP46954	
			CT	0.13		
			CC	0.023		
7	0.066	0.3	GT	0.841	TP41945	-
			AT	0.021	TP201714	
			AC	0.125		
			GC	0.013		
8	0.075	0.25	AC	0.843	TP78842	CABZ01078427.1
			AT	0.045	TP31971	ICA1
			GT	0.102		
9	0.102	0.21	CG	0.763	TP76585	atp6v0a1a
			CA	0.016	TP51853	
			TA	0.221		
			TG	0.011		
10	0.052	0.55	GC	0.763	TP206826	
			AT	0.073	TP100773	
			AC	0.085	TP200097	pi4kaa
			GT	0.078		
11	0.123	0.66	ATA	0.77	TP22240	tspan9a
			GCG	0.066	TP203236	
			GCA	0.059	TP234852	
			ATG	0.091		
12	0.083	0.79	TG	0.839	TP26465	-
			CA	0.121	TP183559	
			CG	0.032		
13	0.09	0.49	CC	0.836	TP68809	-
			TA	0.133	TP194881	
			TC	0.021		
			CA	0.01		
14	0.088	0.37	CT	0.877	TP78633	-

			AG	0.094	TP82766	
			CG	0.012		
			AT	0.018		
15	0.038	0.26	CGT	0.783	TP52423	-
			GCA	0.094	TP166329	
			GCT	0.086	TP215482	
16	0.001	0.51	CG	0.83	TP162773	-
			AA	0.131	TP171096	
			AG	0.03		
17	0.02	0.46	GC	0.704	TP238595	-
			AT	0.262	TP35873	
			AC	0.03		
18	0.024	0.68	GG	0.763	TP14165	c13h10orf11
			GA	0.088	TP194131	
			AG	0.077		
			AA	0.072		
19	0.027	0.27	CG	0.836	TP202569	gpat2
			TG	0.037	TP206826	
			CA	0.011		
			TA	0.117		
20	0.033	0.74	CG	0.76	TP137678	-
			TG	0.053	TP190055	
			TA	0.068		
			CA	0.119		
21	0.043	0.41	TC	0.864	TP76866	-
			GC	0.01	TP103486	
			GT	0.111		
			TT	0.016		
22	0.095	0.76	GG	0.865	TP228646	-
			AG	0.027	TP218889	
			AA	0.099		
23	0.174	0.31	CC	0.556	TP147502	-
			TC	0.027	TP26042	
			CT	0.394		
			TT	0.023		
24	0.043	0.41	GGGC	0.749	TP165619	camta1a
			GGAC	0.015	TP10473	mxra8b
			AAAT	0.102	TP173049	
			GGGT	0.1	TP67282	
25	0.047	0.82	CT	0.795	TP69006	-
			TC	0.041	TP107256	
			CC	0.112		
			TT	0.052		
26	0.055	0.56	GC	0.886	TP228745	-
			AT	0.096	TP17032	
27	0.048	0.13	AC	0.45	TP111445	-
			CC	0.436	TP125838	
			CT	0.111		
28	0.098	0.17	GC	0.817	TP40634	-
			AT	0.167	TP153241	
29	0.041	-	CC	0.719	TP192899	-
			AA	0.2	TP110025	
			AC	0.053		
			CA	0.028		