1 Assessing the sex-related genomic composition difference using a K-mer-based
2 approach: a case of study in *Arapaima gigas* (Pirarucu).

3 Cavalcante, R.L.D¹, Ortega, J.M. ^{1,2}, Souza, J.E.S¹ and Sakamoto, T.^{1*}

4

5 ¹Bioinformatics Multidisciplinary Environment, IMD, Federal University of Rio Grande
6 do Norte - UFRN, Brazil.

7 ²Lab Biodados, Department of Biochemistry and Immunology, Federal University
8 of Minas Gerais - UFMG, Brazil.

9 *To whom correspondence should be addressed: tetsu@imd.ufrn.br

10

11 Keywords: Osteoglossiformes; Sexual differentiation; Amazon fish; Repetitive se-
12 quence; Molecular marker; Fish-farming;

13

14

15

16

17

18

19

20

21

22

23

24

25

ABSTRACT

Arapaima gigas is one of the largest freshwater bony fish in the world, in which adults could weigh 200 kilograms and measure 3 meters in length. Due to its large size and its low-fat meat, *Arapaima gigas* has quickly become a species of special interest in fish-farming. One challenge faced during their production is the lack of an efficient sexing methodology, since their sexual maturation occurs late (around the third to the fifth year) and the genetic mechanisms linked to their sex determination system are not known yet. For a more sustainable management, it is of paramount importance to seek an effective and non-invasive method to differentiate sexually juvenile individuals of *Arapaima gigas*. For this, the establishment of genetic markers associated with sexual differentiation would be an advantageous tool. In this study, we proposed a k-mer based approach to identify genome features with sex-determining properties. For this purpose, we used genomic data from four adult representatives of *Arapaima gigas*, two males and two females, and counted the k-mers comprising them. As result, we found k-mers from repetitive regions with high difference and disproportion in the count among individuals of the opposite sex. These differences in the k-mer-based genomic composition indicate the existence of genetic factors involved in the sexing of individuals in *Arapaima gigas*.

52 BACKGROUND

53 *Arapaima gigas*, commonly known as “Pirarucu” or “Paiche”, is the largest bony
54 freshwater fish in the world. It belongs to the bonytongues (Order Osteoglossi-
55 formes) Arapaimidae family, and has a natural habitat in the Amazon Basin [1].
56 Adult specimens may weigh around 200 kilograms and measure about 3 meters
57 [2, 3]. Due to its large size, its flesh containing low-fat and low fishbone, along with
58 its physiological characteristic of emerging to the surface at intervals of 15 minutes
59 to assimilate oxygen, *Arapaima gigas* became a vulnerable species to overfish-
60 ing in the Amazon region [4] leading to greater surveillance of the marketing of
61 Pirarucu by the Brazilian government in the early 2000s [5, 6].

62 Some studies show that the use of Pirarucu in intensive fish farming is facili-
63 tated, in part, by the physiological characteristics of the animal that guarantee the
64 rusticity of the species [7]. For example, the obligate air-breathing causes this
65 species can tolerate environments with low concentrations of dissolved oxygen in
66 the water [8]. In addition to the facility for captive management, attributes such as
67 the low content of fat, combined with the rapid growth of the species, which weighs
68 an average of 10 kg in its first year of life, add value for an intensification in the
69 commercial exploitation of the *Arapaima gigas* [5, 9, 10].

70 One of the problems related to its fishing exploitation and fish-farming is that
71 we do not know for certain the genetic mechanisms linked to sex-differentiation in
72 *Arapaima gigas* [11], its sexual maturation occurs around the third to the fifth year
73 of life, and sexual dimorphism is not a strong feature of the species [10]. In recent
74 decades, the creation of *Arapaima gigas* in captivity has been increasingly stimu-
75 lated, either to develop research to better know the particularities of the species or
76 to exploit its economic potential [12, 13]. For more sustainable management, it is
77 of paramount importance to seek an effective and non-invasive method to differen-

78 tiate sexually juvenile individuals.

79 For this, the establishment of a molecular genetic marker related to sexual dif-
80 ferentiation would be an advantageous tool. Previous analysis of the *Arapaima*
81 *gigas* genome found no genes associated with the identification of the sex deter-
82 mination system of these individuals [14, 15]. And chromosomal characterization
83 studies could not distinguishable cytologically a sex chromosome in *Arapaima gi-*
84 *gas* [16, 17]. In this study, we proposed to asses the genomic composition of *A.*
85 *gigas* using a k-mer-based approach to identify regions in excess or missing in one
86 of the sexes.

87

88 MATERIALS AND METHODS

89 *A. gigas* sequencing data and data processing

90 For the following analyses, genomic data were used from four adult *Arapaima*
91 *gigas* representatives, two males (M1 and M2) and two females (F1 and F2). All
92 samples were collected from Bioproject PRJEB22808 available in National Cen-
93 ter for Biotechnology Information (NCBI) database [14]. The quality of the reads
94 was verified with the help of FastQC (v. 0.11.4) [18], and low-quality reads were
95 trimmed with the help of the Sickle paired-end (v. 1.33) [19].

96

97 K-mer Analysis

98 The K-mer analysis was performed in three steps: (1) k-mer count, (2) k-mer
99 count normalization and (3) k-mer count comparison. The k-mer count was per-
100 formed with the help of the tool Alignment and Assembly Free (AAF, v. 20171001)
101 [20], a free software written in Python (v. 3.7.0) whose premise is to assemble a
102 phylogenetic distribution based on the k-mers count shared by the samples. The
103 trimmed fastq files of the 4 representatives *Arapaima gigas* were submitted as in-

put data to the script `aaf_phylokmer.py` of AAF using the parameter `-k` (k-mer size) as 23. The output of this script is a table with $n \times m$ elements, in which n is the total number of distinct k-mers ($i = 1, 2, \dots, n$), m is the number of samples ($j = 1, 2, \dots, m$) and C_{ij} is the count of k-mer i in sample j . For further analysis, we only considered k-mers in which the sum of the count in all samples were greater than 5.

Since the genomic data submitted to this analysis have different coverage values, in order to normalize the data and to decrease the number of false positives, the method Quantile Normalization (QN), a global adjustment method, which consists of a non-parametric methods that makes two or more distributions identical on statistical properties [21].

Finally, to verify the difference in the k-mer abundance among the samples, we compared the samples in pairwise manner (six comparisons in the total) and calculated, for each k-mer, (1) the difference ($D_{i,j \times j'} = C_{ij} - C_{ij'}$) and (2) the logarithm of the ratio ($L_{i,j \times j'} = \log_2(C_{ij} / C_{ij'})$) of their count values. To select those k-mers that showed high $D_{i,j \times j'}$ and extreme values of $L_{i,j \times j'}$, we calculated the product of both values ($P_{i,j \times j} = D_{i,j \times j} \times L_{i,j \times j}$) and extracted the 500 k-mers with highest product value. In the end, 3000 k-mers were extracted from all six comparisons and submitted to CD-HIT tool (v. 4.7) [22] for cluster analysis. Clusters with k-mers that were extracted from a comparison between samples of the same sex was discarded for further analysis. Then, in the remaining clusters, those which contain k-mers extracted from all comparison between samples of opposite sex were considered as candidate genetic markers for sex determination. All graphic plots were generated using Matplotlib (v. 3.0.1) [23] and Seaborn (v. 0.9.0) [24], which are Python 2D plotting libraries.

RESULTS

The genome sequencing data of 4 samples of *A. gigas*, two female (F1 and F2) and two male (M1 and M2) samples, was obtained and processed. After trimming for low-quality sequences, 8.8 million, 37.7 million, 37.1 million, 13.5 million reads remained for F1, F2, M1 and M2, respectively. The remained reads was submitted to the software AAF [20] to count the k-mers comprising each genome samples. The total k-mers counted was 1.03 billion, 2.15 billion, 2.17 billion and 1.18 billion k-mers for F1, F2, M1 and M2 samples, respectively. Because of the difference on the sequencing depth of each sample (Figure 1A), the Quantile Normalization method (QN) was applied to make the counts comparable between samples (Figure 1B).

The comparison of k-mers count between individuals of the opposite sex and the same sex, either by the count difference or logarithm of the ratio, showed some k-mers which were in excess in one of the samples (Figure 2). In each comparison between a pair of samples (6 comparisons in total), we selected 500 k-mers that showed the highest value of the product of the count difference and the log ratio. All selected k-mers (6 x 500 k-mers) had their sequences clustered with CD-HIT [22]. This analysis resulted in the formation of 1129 clusters, of which 104 were comprised of k-mers that was selected only in comparisons between samples of opposite sex. Furthermore, of these 104 clusters, four of them had at least one representative of each comparison made between samples of opposite sex, which were comprised of k-mers of a repetitive regions (Figure 3).

DISCUSSION

Understand the mechanisms of sexual determination in fish are essential for a sustainable management of ichthyofauna, either for commercial or conservation

156 purposes [25]. Introspection of these mechanisms in *A. gigas* has an appeal be-
 157 cause this species plays an important role in the economy of the north region of
 158 Brazil [12, 13]. But elucidating these mechanisms in fish is challenging, since most
 159 of them, including *A. gigas*, do not bear a typical sex chromosome in their genome
 160 [16, 17] and the difference of the genome sequences between samples of opposite
 161 sex seems to be minimal. The first genome sequencing of *A. gigas* [14] did not find
 162 significant differences between the genomic content of male and female samples.
 163 The second one [15] suggested the XY system for *A. gigas*, but their results conflict
 164 with the genome data of the first one.

165 The approach of finding specific regions in one sex using genomic data has
 166 demonstrated to be not so effective in finding those regions associated with the
 167 sex-determination. In this context, we explored other genomic features in *A. gigas*
 168 to find some clues about the genetic factors involved in the sex determining system.
 169 For this, we analysed the genomic composition of Pirarucu using k-mer based
 170 approach.

171 In this study, we have noticed the existence of k-mers over or underrepresented
 172 in one of the sexes, indicating potential differences in the genetic composition be-
 173 tween males and females of *A. gigas*. The difference is not so expressive, which
 174 corroborates with the reports that estimate 0.01% [14] to 0.1% [15] of the genome
 175 of this species as linked to the sexual determination. The four sequences reported
 176 in this work (Figure 3) are all part of repetitive sequences. Despite of their low
 177 complexity, repetitive regions have been reported to have important role on sex
 178 determination [26]. In medaka, which has the XY system, there is a large stretch
 179 of repetitive regions on the Y-specific regions [27]. The chromosome Y of Pacific
 180 salmon bears a specific repetitive regions (OtY1) that is used as genetic marker to
 181 differentiate sex [28].

182 In this context, the repetitive sequences found in this study could be a compo-
183 nent that could be used to determine sex in individuals of *Arapaima gigas*. We
184 recognize, however, the necessity to perform analyses with a greater number of
185 samples to obtain a statistical support for our results. Despite of that, the kmer-
186 based method applied on this work has demonstrated to be an interesting strategy
187 to help us discover the sex-determination system in Pirarucu specimens and can
188 be extended to other species.

189

190 CONCLUSIONS

191 With this study, we were able to suggest some repetitive genome sequences
192 that may be differentiated in quantity in male and female of *A. gigas*. We consider
193 these regions as strong candidates for a molecular marker that could be used for
194 sexing individuals of *A. gigas*. However, because of the small sample size, new
195 analyses with a larger number of individuals are necessary to allow statistical sup-
196 port for our conclusion, as well as to suggest bench trials for the validation of the
197 in-silico analyses. Furthermore, kmer-based methods demonstrated to be an inter-
198 esting strategy to assist us in unraveling the sex-determination system in Pirarucu.

199

200 ACKNOWLEDGMENTS

201 This work was supported by CAPES (Coordenação de Aperfeiçoamento de
202 Pessoal do Ensino Superior). To the Instituto Metrópole Digital - IMD, and specifi-
203 cally to Bioinformatics Multidisciplinary Environment - BioME for the computational
204 structure for the accomplishment of this work. To the Federal University of Rio
205 Grande do Norte - UFRN for structural support and resources have been ceded.
206 And to Dr. Sidney Emanuel Batista dos Santos from Federal University of Pará -
207 UFPA for contributing in discussions in this project.

AUTHOR CONTRIBUTIONS

RLDC designed the pipeline and evaluated the data. JMO assisted with computational structure. JESS assisted with project development and discussions. TS designed the project, evaluated the data and assisted with project development and discussions.

References

- [1] Roberto E Reis, Sven O Kullander, and Carl J Ferraris. *Check list of the freshwater fishes of South and Central America*. Edipucrs, 2003.
- [2] RF Bezerra, MCF Soares, EVM Maciel Carvalho, and LCBB Coelho. Pirarucu, arapaima gigas, the amazonian giant fish is briefly reviewed. *Fish, fishing and fisheries series. Nova Science, New York*, 41, 2013.
- [3] Joseph S Nelson, Terry C Grande, and Mark VH Wilson. *Fishes of the World*. John Wiley & Sons, 2016.
- [4] Michael Goulding, Nigel JH Smith, and Dennis J Mahar. *Floods of fortune: ecology and economy along the Amazon*. Columbia University Press, 1996.
- [5] Peter B Bayley and Miguel Petrere Jr. Amazon fisheries: Assessment methods, current status and management options. *Canadian special publication of fisheries and aquatic sciences/Publication speciale canadienne des sciences halieutiques et aquatiques*, 1989.
- [6] Don E Wilson and David Burnie. Animal: the definitive visual guide to the world's wildlife. *Dorling Kindersley, London, UK*, 2001.

- 229 [7] Elenise Gonçalves de Oliveira, Alberto Banhos Pinheiro, Valdemir Queiroz
230 de Oliveira, Antônio Ronaldo Melo da Silva Júnior, Manuella Gazzineo
231 de Moraes, Ítalo Régis Castelo Branco Rocha, Rommel Rocha de Sousa,
232 and Francisco Hiran Farias Costa. Effects of stocking density on the perfor-
233 mance of juvenile pirarucu (*arapaima gigas*) in cages. *Aquaculture*, 370:96–
234 101, 2012.
- 235 [8] Jacques Bard and Emir Palmeira IMBIRIBA. Piscicultura do pirarucu, ara-
236 paima gigas. *Embrapa Amazônia Oriental-Circular Técnica (INFOTECA-E)*,
237 1986.
- 238 [9] LOD de M Carvalho and CNB do NASCIMENTO. Engorda de pirarucus
239 (*arapaima gigas*) em associação com búfalos e suínos. *Embrapa Amazônia*
240 *Oriental-Circular Técnica (INFOTECA-E)*, 1992.
- 241 [10] IG Almeida, P Ianella, MT Faria, SR Paiva, and AR Caetano. Bulkéd segre-
242 gant analysis of the pirarucu (*arapaima gigas*) genome for identification of sex-
243 specific molecular markers. *Embrapa Amazônia Oriental-Artigo em periódico*
244 *indexado (ALICE)*, 2013.
- 245 [11] Tomas Hrbek and Izeni Pires Farias. The complete mitochondrial genome of
246 the pirarucu (*arapaima gigas*, *arapaimidae*, *osteoglossiformes*). *Genetics and*
247 *Molecular Biology*, 31(1):293–302, 2008.
- 248 [12] O Fontenele. Contribuição para o conhecimento da biologia do pirarucu
249 *arapaima gigas* (cuvier), em cativeiro (*actinopterygii*, *osteoglossidae*)[ceara,
250 brasil]. 1955.
- 251 [13] Emir Palmeira IMBIRIBA. Potencial de criação de pirarucu, *arapaima gigas*,
252 em cativeiro. *Acta Amazonica*, 31(2):299–299, 2001.

- 253 [14] Ricardo Assunção Vialle, Jorge Estefano Santana de Souza, Katia de Paiva
254 Lopes, Diego Gomes Teixeira, Pitágoras de Azevedo Alves Sobrinho,
255 André M Ribeiro-dos Santos, Carolina Furtado, Tetsu Sakamoto, Fábio Au-
256 gusto Oliveira Silva, Edivaldo Herculano Corrêa de Oliveira, et al. Whole
257 genome sequencing of the pirarucu (*arapaima gigas*) supports indepen-
258 dent emergence of major teleost clades. *Genome biology and evolution*,
259 10(9):2366–2379, 2018.
- 260 [15] Kang Du, Sven Wuertz, Mateus Adolphi, Susanne Kneitz, Matthias Stöck, Mar-
261 cos Oliveira, Rafael Nóbrega, Jenny Ormanns, Werner Kloas, Romain Feron,
262 et al. The genome of the arapaima (*arapaima gigas*) provides insights into
263 gigantism, fast growth and chromosomal sex determination system. *Scientific*
264 *reports*, 9(1):5293, 2019.
- 265 [16] Debora Karla Marques, Paulo Cesar Venere, and Pedro Manoel Galetti Junior.
266 Chromosomal characterization of the bonytongue arapaima gigas (*osteoglos-*
267 *siformes*: *Arapaimidae*). *Neotropical Ichthyology*, 4(2):215–218, 2006.
- 268 [17] Ezequiel Aguiar de Oliveira, Luiz Antonio Carlos Bertollo, Petr Rab, Tariq
269 Ezaz, Cassia Fernanda Yano, Terumi Hatanaka, Oladele Ilesanmi Jegede,
270 Alongklod Tanomtong, Thomas Liehr, Alexandr Sember, et al. Cytogenetics,
271 genomics and biodiversity of the south american and african arapaimidae fish
272 family (teleostei, osteoglossiformes). *PloS one*, 14(3), 2019.
- 273 [18] Steven W Wingett and Simon Andrews. Fastq screen: A tool for multi-genome
274 mapping and quality control. *F1000Research*, 7, 2018.
- 275 [19] NA Joshi and JN Fass. Sickel: A sliding-window, adaptive, quality-based
276 trimming tool for fastq files (version 1.33)[software], 2011.

- 277 [20] Huan Fan, Anthony R Ives, Yann Surget-Groba, and Charles H Cannon. An
278 assembly and alignment-free method of phylogeny reconstruction from next-
279 generation sequencing data. *BMC genomics*, 16(1):522, 2015.
- 280 [21] Stephanie C Hicks and Rafael A Irizarry. Quantro: a data-driven approach to
281 guide the choice of an appropriate normalization method. *Genome biology*,
282 16(1):117, 2015.
- 283 [22] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: ac-
284 celerated for clustering the next-generation sequencing data. *Bioinformatics*,
285 28(23):3150–3152, 2012.
- 286 [23] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science*
287 *& engineering*, 9(3):90–95, 2007.
- 288 [24] Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Saulius
289 Lukauskas, David C Gemperline, Tom Augspurger, Y Halchenko, JB Cole,
290 J Warmenhoven, et al. mwaskom/seaborn: v0. 8.1 (september 2017). *Zen-*
291 *odo*, doi: <https://doi.org/10.5281/zenodo.883859>, 2017.
- 292 [25] Paulino Martínez, Ana M Viñas, Laura Sánchez, Noelia Díaz, Laia Ribas, and
293 Francesc Piferrer. Genetic architecture of sex determination in fish: applica-
294 tions to sex ratio control in aquaculture. *Frontiers in genetics*, 5:340, 2014.
- 295 [26] Tariq Ezaz and Janine E Deakin. Repetitive sequence and sex chromosome
296 evolution in vertebrates. *Advances in Evolutionary Biology*, 2014, 2014.
- 297 [27] Mariko Kondo, Ute Hornung, Indrajit Nanda, Shuichiro Imai, Takashi Sasaki,
298 Atsushi Shimizu, Shuichi Asakawa, Hiroshi Hori, Michael Schmid, Nobuyoshi
299 Shimizu, et al. Genomic organization of the sex-determining and adjacent

regions of the sex chromosomes of medaka. *Genome Research*, 16(7):815–826, 2006.

[28] RH Devlin, CA Biagi, and DE Smailus. Genetic mapping of y-chromosomal dna markers in pacific salmon. *Genetica*, 111(1-3):43–58, 2001.

FIGURE LEGENDS

Figure 1 - K-mers count size distribution. The distributions were generated using raw (A) and QN normalized (B) k-mer count values. X-axis represents the k-mer count size and Y-axis represents the number of k-mer with that size in log scale.

Figure 2 - Scatter plot of the k-mer count difference and log ratio between two sample. The k-mer count comparisons between samples F1xM1 (A), F1xM2 (B), F2xM1 (C), F2xM2 (D), F1xF2 (E) and M1xM2 (F) are shown. For all plots, each dot represents a k-mer, X-axis represents the k-mer count difference and Y-axis represents the log ratio of the k-mer count dots far from the origin represent k-mers in excess in one of the sample.

Figure 3 - The 104 clusters comprised of k-mers extracted only from the comparison between individuals of opposite sex. A) Venn Diagram showing the k-mers shared among those extracted from the comparison F1xM1, F1xM2, F2xM1 and F2xM2. Four clusters showed are in the intersection of all comparison. B) Table detailing the four clusters in the intersection. F: female; M: male.

A

B

C

D

E

F

A

B

Cluster	Sequence	Excess in
305	CTCTCTCTCTCTCTCTCTC	F
308	ACTGACTGACTGACTGACTGAC	F
315	CAGTCAGTCAGTCAGTCAGTCA	F
326	AACCCTAACCCTAACCCTAAC	M