Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.2626

Let them eat cake: when the small aims at being LARGE or the empowering effects of bioinformatics in NGS wonderland

Gabriel M. Yazbeck [**DEZOO, Universidade Federal de São João Del Rei** | ***dna@ufsj.edu.br***]
Raíssa C. D. Graciano [**Programa de Pós Graduação em Biotecnologia, Universidade Federal de São João Del Rei** | ***raissadgraciano@gmail.com***]
Rosiane P. Santos [**Programa de Pós Graduação em Bioinformática, Universidade Federal de Minas Gerais** | ***rosianeps2007@yahoo.com.br***]
Rafael Sachetto Oliveira [**Departamento de Ciência da Computação, Universidade Federal de São João Del Rei** | ***sachetto@ufsj.edu.br***]

Laboratório de Recursos Genéticos (LARGE), CTAN, DEZOO, UFSJ, Av. Visconde do Rio Preto s/n, Fábricas, São João del-Rei, MG, 36301-360, Brazil

Received: 4 April 2022 • **Published:** 16 February 2024

Abstract This report summarizes the path (and pitfalls) in the way of the Genetic Resources Laboratory (LARGE-UFSJ), trailed with the aid of bioinformatics, in the field of massive DNA data analyses and its application in the field of conservation of biodiversity, particularly of Neotropical migratory fish. We use the metaphor of DNA sequencing as the cake, both as a prized delicacy formerly inaccessible to the masses, as in the infamous "let them eat cake", scornfully exclaimed by Marie-Antoinette during bread shortage in the French Revolution, but also as a means to achieve rapid growth for small research groups, as the plot device in Lewis Carroll' Alice in Wonderland. Next-Generation Sequencing (NGS) methods have been known to promote a true revolution in the Life Sciences, empowering groups with limited resources to explore the relatively new, still unknown and often surprising world of genetic sequences. Indeed, we argue for the inertia breaking potential of NGS and give our group's trajectory as a testimony. It all begun with the fortuitous union of providential fish DNA big-data gathered by Genetics professor, Dr. Yazbeck, and Computer Science professor, Dr. Sachetto's curiosity onto biological research, along the wit of some young researchers. Our initial NGS challenge was to provide the assembly and annotation of the first mitochondrial genome for the Anostomidae fish family. The LARGE's NGS research program was able to promote the characterization of what was then arguably the highest number of microsatellite DNA markers for the flagship species, *Salminus brasiliensis* (dourado) and *Brycon orbignyanus* (piracanjuba), useful in environmental applications for conservation (green biotechnology). We also have provided this large raw datasets, as well as elaborated massive results, freely available to the scientific community in data repositories such as GenBank, SRA and FigShare, such as genomic assemblies and gene annotation in these fish. Technological spin offs with application in the environmental protection and food production fields have also been devised as direct consequence of the availability of such rich and diverse data.

Keywords: Next-Generation Sequencing, High-Throughput Sequencing, Genomics, DNA Markers

1 Introduction - Down the rabbit hole

When faced with rising animosity by the starving mob, due to lack of bread, Marie-Antoinette, the queen of France during the French Revolution, alleged (most likely false) and infamously said "Let them eat cake, then!" [Barker, 1993]. The delicacy also features in another well-known tale, Alice's Adventures in Wonderland, by Lewis Carroll (1865), as a plot device and a further suitable allegory here. Spoiler ahead! There, the iconic protagonist is confronted with a glass bottle neatly marked "Drink Me!", which she cautiously did. That led to her shrinking in size, just so she could find a glass box with a small cake inside marked to be eaten, which made her grow instantly tall, or open out "like the largest telescope there ever was".

Next-Generation Sequencing (NGS) is for Genetics, the same as a telescope is to Astronomy. The umbrella term for a series of massively parallelized nucleic acid sequencing methods and platforms [Metzker, 2010; Ekblom and Galindo, 2011], embodies tools that have provided a previously unseen technological advancement in any area, ever, mostly so if judged by the "bang-for-the-buck". Traditional sequencing methodologies can cost around U$2,400.00 per million DNA bases sequenced, while NGS approaches can yield the same amount of DNA data for around U$0.15 [Park and Kim, 2016]. Its effects are metaphorically akin to a fantastic cake capable of suddenly and quickly inflating the technical capabilities of small research groups. That cake, which used to be more expensive than bread, only readily available to "royalty" (*i.e.* well established, large, structured, adequately equipped and staffed) groups, is now available to the common folk. One should not, however, underrate the hidden costs and challenges for data storage, treatment, and analysis costs, besides the skills needed to handle such massive datasets [Sboner *et al*., 2011]. This report intends to present the main results (as well as some obstacles) reached by our relatively young and small research group in the countryside of Minas Gerais, Brazil, from the 2010s to the present day, as

a testimony of the empowering capacities of bioinformatics and its application to NGS data.

2 False start

Our first incursion into bioinformatics came from a rather serendipitous simple non-NGS related project involving then students Rosiane P. Santos and Érico M. Polo, along with expert chemist Dr. Leonardo Marmo. It concerned theoretical reverse translation of amino acid chains from the external giant hemoglobin from *Lumbricus paulistus*, a large earthworm species, aiming oligonucleotide (*i.e.* primers) design for Polymerase Chain Reaction (PCR). Polo and Santos used a C++ in-house code capable of providing a sliding codon frame, to identify regions in the polypeptide least affected by genetic code degeneration, in order to minimize the number of given candidate subjacent nucleotide sequences, and using information from *Lumbricus terrestris* (the regular earthworm) to infer codon usage frequency and similarity. This allowed filtering the astronomical number of possibilities into a couple of potential PCR primer pairs. Awkwardly, these results were inadvertently published in a journal associated with a publisher deemed as predatory by Beall's List [Krawczyk and Kulczycki, 2021; Moed *et al*., 2021], which only later we found out about. It happened when R. Santos registered for attending a scientific conference in Australia, during her stay in the notable but now defunct *Ciências Sem Fronteiras* student's exchange program. The conference announced the publication of articles for the works accepted by it [Polo *et al*., 2014]. Regardless, we stand and ratify the work's integrity.

3 Stairway to heaven

When granted the funds to generate the first massive NGS data on important migratory Neotropical fish species of socio-environmental concern, such as *Salminus brasiliensis* (dourado), *Brycon orbignyanus* (piracanjuba), *Prochilodus lineatus* (curimba) and *Megaleporinus obtusidens* (piau), and following a fortuitous visit by Professor Rafael Sachetto, from the Computer Science Department to the Genetic Resources Laboratory (or LARGE, our small but ironically named research group) in 2014, the group took a one way ticket for the bioinformatics train departing into the NGS wonder-landscape. These fish are aimed because of their importance in conservation programs and initiatives associated with hydroelectric power generation, which relies upon dam implementation which, in turn, interrupts migratory routes and leads to environmental fragmentation and degradation. As such, the quick generation of genomic data allows for the rapid and abundant development of molecular markers capable of being applied in stock delimitation and other conservation genetics issues [Ekblom and Galindo, 2011]. Before NGS, the typical DNA marker development took long, convoluted, technically challenging, and rather expensive methods [Zane *et al*., 2002]. This would later prove the inertia-breaking capacity of NGS, in the face of small groups' lack of adequate infrastructure, such as equipment and even support personnel, and due to the fact that the most inexpensive way of working with NGS is arguably delegating the DNA sequencing itself to a third party specialized service provider.

We used a service provider (SP) to generate NGS data from DNA extracted at LARGE. The results came in the form of raw filtered paired-end short reads data (Table 1), as FASTQ files, which aggregate 90 bases long reads with associated quality (Q) for each base call. Some preliminary elaborated results (massive tables of microsatellite candidate-loci, a class of molecular DNA markers) from SP bioinformatics analyses were provided, although without the detailed middle-steps results (*e.g.* genomic assembly from the short reads). This challenged us to recreate the step-by-step approach, from the filtered data, independently performing assembly of genomic contigs and scaffolds, from where DNA markers could be searched and characterized *de novo*. It took us approximately two years, only working through the theory of genome assembly from short reads, De Bruijn Graphs, Eulerian walks, NGS chemistry, etc. This was achieved through an organized, regular study group, named GATTACA. At this stage, a core team from the study group was able to fully assemble and annotated the first mitochondrial genome for the Anostomidae fish family, available online from 2014, as an early communication [Yazbeck *et al*., 2016], from this NGS data. This was achieved using a simple but intensive bioinformatics pipeline, with the aid of an almost complete reference mitogenome from GenBank. This was considered by us a true feat for our small group and revealed, right away, the power of bioinformatics mining of NGS data. Meanwhile, Prof. Sachetto spent time harnessing computational infrastructure and devising in-house Bash, Python, and other scripts, in response to demands and feedback from Yazbeck, in order to execute our parallel approach, according to the general steps outlined by the SP in marker search. Although we succeed, this sort of computational endeavor is likely to generate slightly different results [Sandve *et al*., 2013], even if using the same version, of the same software, with the same input data, same parameters (*e.g.* k-mer) and, indeed, it generated a marginally divergent genome assembly, as judged by the back mapping of SP determined microsatellite DNA markers.

4 Nirvana

While we evaluated different alternatives for maximizing mapped microsatellite hits in the bioinformatics front, new developments were underway for results concerning potential DNA microsatellite markers. The wet-bench, molecular biology arm of LARGE was putting the SP's elaborated results (the DNA maker panels) to the test. The empirical PCR-based results could not have been better, yielding an amount of validated molecular markers unimaginable only 10 years before, for the same amount of money spent. Our group validated the first 29 microsatellite DNA markers for *B. orbignyanus*, a species virtually extinct in the wild in Minas Gerais [Arias *et al*., 2016] and 47 new molecular markers for *S. brasiliensis* [Cao *et al*., 2016], one of the most emblematic fish in Brazil and South America. This was safely a new record. The accession numbers for these markers, deposited

Table 1. Next-generation sequencing DNA short reads data generated for four different migratory fish. Each line represents one different genomic library and *P. lineatus* and *M. obtusidens* have multiple libraries from the same specimen. Q ≥ 20 = Quality value of 1% average error, maximum. Mb=Megabases and Gb=Gigabases.

Species	GC Content (%)	Q≥ 20 rate (%)	Reads (Mb)	Bases (Gb)
B. orbignyanus	41.11	97.73	178.21	16.04
S. brasiliensis	41.05	97.44	178.08	16.03
P. lineatus	36.92	97.87	77.3	6.95
	36.8	97.62	83.3	7.6
	36.81	97.62	81.2	7.1
	35.5	98.57	47.2	4.25
M. obtusidens	40.33	98.69	169.30	15.24
	34.56	97.66	62.8	5.65
	34.55	96.62	64.8	5.83

Table 2. List of microsatellite DNA molecular markers developed from the NGS data for *B. orbignyanus* deposited at NCBI's GenBank.

Locus	Accession numbers
Borb04	KT827796.1
Borb06	KT827797.1
Borb07	KT827798.1
Borb08	KT827799.1
Borb09	KT827800.1
Borb11	KT827801.1
Borb12	KT827802.1
Borb13	KT827803.1
Borb14	KT827804.1
Borb15	KT827805.1
Borb17	KT827807.1
Borb18	KT827808.1
Borb21	KT827809.1
Borb25	KT827811.1
Borb28	KT827812.1
Borb29	KT827813.1
Borb30	KT827814.1
Borb33	KT827815.1
Borb34	KT827816.1
Borb35	KT827817.1
Borb36	KT827818.1
Borb38	KT827819.1
Borb39	KT827820.1
Borb43	KT827821.1
Borb44	KT827822.1
Borb46	KT827823.1

in GenBank can be found in Tables 2 and 3, respectively.

It is also worth mentioning the punctual but useful, parallel bioinformatics solution developed in-house, using Python for use in the wet-bench front. The program, Helplex is able to generate automated picks for alternative sets of previously validated molecular markers, in order to combine them as PCR multiplexes, applying graph theory, written by then student José Mauro Ribeiro, as his undergrad supervised project in Computer Science. It had its application validated in the wet-bench and even resulted in two patent requests in the area of green biotechnology, in Brazil, regarding the use of multiplex (*i.e.* combined) microsatellite systems in commercial operations (*e.g.* [Carvalho *et al*., 2018]).

5 Wonderland

When it came to characterizing our own DNA marker panel, we could independently find more than 95% of the markers found by the SP. Some non-recovered markers (*i.e.* putative markers found in the unavailable SP's genomic assembly, which could not be traced back to our in-house *de novo* assembly), nevertheless, had already been empirically confirmed in actual PCR bench assays. This led us to retain the SP's original table since it was slightly larger than the new alternative, and we took a hard turn, mapping loci characterized in the unknown assembly, onto our new genomic assembly from short reads data. This was achieved through BLAST and SWIPE, in collaboration with French scientist Dr. Dominique Lavenier. This approach was published along with the short read data, the assembly (approx. 1.1 billion organized DNA bases in total), the mapping of short reads into the assembly (a Binary Alignment Mapping file, or BAM), extracted from a human-readable version, a Sequence Alignment Mapping or SAM file, as well as the potential microsatellite DNA marker table for this species [Yazbeck *et al*., 2018].

R. Santos used a pipeline to mine the *B. orbignyanus* NGS raw data in search of potential molecular markers (single nucleotide polymorphisms or SNP type, as well as insertion/deletions or anonymous nuclear loci, ANL, from non-coding regions of the genome) suitable as useful tools for molecular taxonomy and conservation genetics. This was achieved through the application programming interface GATB - Genome Assembly and Analysis Tool Box, using the methods implemented in DiscoSNP. It yielded an extensive list containing candidate SNPs and INDELs. Subsequently, a bioinformatics script was created for the selection of the best candidates. Pairs of primers for PCR were proposed for *in vitro* tests. These results (in preparation) provide a new and broad avenue for molecular marker development in a cost-effective way, with a pipeline applicable to many other different organisms. Parallel to this project, Santos, now a Ph.D. candidate in Bioinformatics, was able to take out the *P. lineatus* mitochondrial genome assembly from LARGE to Professor Daniel Carvalho's research group and conduct another research making evolutionary inference on the eastern Brazilian *Prochilodus* species group, along other four species, using full mitochondrial DNA sequence data [Santos *et al*., 2021].

Table 3. List of microsatellite DNA molecular markers developed from the NGS data *S. brasilensis* deposited at NCBI's GenBank.

Locus	Accession numbers
Sbra01	KX421479.1
Sbra03	KX421480.1
Sbra04	KX421481.1
Sbra05	KX421482.1
Sbra06	KX421483.1
Sbra07	KX421484.1
Sbra08	KX421485.1
Sbra09	KX421486.1
Sbra10	KX421487.1
Sbra11	KX421488.1
Sbra12	KX421489.1
Sbra14	KX421490.1
Sbra15	KX421491.1
Sbra16	KX421492.1
Sbra18	KX421493.1
Sbra19	KX421494.1
Sbra20	KX421495.1
Sbra21	KX421496.1
Sbra22	KX421497.1
Sbra23	KX421498.1
Sbra24	KX421499.1
Sbra25	KX421500.1
Sbra26	KX421501.1
Sbra27	KX421502.1
Sbra28	KX421503.1
Sbra29	KX421504.1
Sbra30	KX421505.1
Sbra31	KX421506.1
Sbra32	KX421507.1
Sbra34	KX421508.1
Sbra35	KX421509.1
Sbra36	KX421510.1
Sbra37	KX421511.1
Sbra38	KX421512.1
Sbra39	KX421513.1
Sbra40	KX421514.1
Sbra41	KX421515.1
Sbra42	KX421516.1
Sbra43	KX421517.1
Sbra44	KX421518.1
Sbra45	KX421519.1
Sbra47	KX421520.1
Sbra48	KX421521.1
Sbra49	KX421522.1
Sbra50	KX421523.1
Sbra51	KX421524.1
Sbra52	KX421525.1

In 2020-2021 Raíssa Graciano performed the functional gene annotation of the genomic assembly presented for *B. orbignyanus* [Graciano *et al*., 2021], using the intensive MAKER pipeline. This means that every stretch of DNA that could be assigned a biological function, such as a coding sequence, or gene family, was cataloged and mapped in the genomic assembly. Although heavily fragmented and incomplete in nature, this assembly allowed for the description of around 13 thousand new genes from this threatened fish (Table 3), whereas only 13 mitochondrial genes were previously characterized. These results represent a step forward in the search for biotechnological novelty and in the basic molecular genetics information for this migratory species.

The same approach was applied to *S. brasiliensis* and published as novel genomic resources for this fish species. NGS short reads, our own genomic assembly (totaling around 1 billion DNA bases), BAM file, mitogenome, and the potential microsatellite loci massive panel were all made publicly available in the Short Sequence Archives from NCBI or the FigShare data repository. We also provided the functional annotation for this fish, a potential flagship species with charisma for leading the conservation of aquatic ecosystems [Graciano *et al*., 2022]. All these genomic resources data and results are summarized in Table 4 and Table 5. These results also represent major advances in molecular knowledge about this fish and are expected to generate biotechnological spin-offs soon, such as hormones for aquaculture and the management of captive fish.

There are still NGS data consisting of multiple libraries to be analyzed, which is being hindered by a lack of access to computer resources with high-end random access memory availability (*i.e.* 128 GB or more), which has been limiting us to explore one library at a time. Our more immediate goals involve gaining access to such computation power, so similar contributions (and beyond) can be promptly made available to *P. lineatus* and the elusive *M. obtusidens*, a species at the center of a complex taxonomic issue.

It follows from our exposition that NGS data was indeed capable of catapulting this relatively young research group, LARGE, with limited equipment and facilities, to new heights, as judged by the amount of data, results, and publications made available to the scientific community worldwide in the last five years. This corroborates our metaphor for this new research horizon, NGS, as a cake, once a privilege of a few, now within the reach of many. One can go further and view bioinformatics as the knife or tool, allowing anyone to cut a slice from the cake for themselves, since the amount of data accumulating in repositories isn't possibly treatable in the short-term with current human and computer resources available, making it a *bona fide* democratic field.

6 Conclusion

This report presented a summary of the results achieved, as well as obstacles overcome, by a rather small research group, thanks to the application of bioinformatic tools and resources to next-generation DNA sequencing data and other biological problems. We hope it may serve as a tale of encouragement to young research groups aiming for relative independence and a way forward with their investigation goals. Bioinformatic exploration of NGS data permitted us to produce important practical contributions in the field of molecular marker development with an efficiency not attainable just a decade ago. We also increased the number of described genes for two important fish species, *S. brasiliensis* and *B. orbignyanus*, approximately 1,000-fold. The results and data produced have been publicly made available. Two more datasets are being processed for the same end, with a lack of

Table 4. ***Brycon orbignyanus*** NGS data sets and elaborated results publicly available.

Data set	Access link
Assembly	https://doi.org/10.6084/m9.figshare.5661802.v1
Genomic annotation (GFF3)	https://doi.org/10.6084/m9.figshare.11793627.v1
Table of genes characterized	https://doi.org/10.6084/m9.figshare.16734751.v1
Genome-wide microsatellite panel	https://doi.org/10.6084/m9.figshare.5661988.v1
Short reads	https://www.ncbi.nlm.nih.gov/sra/SRX3350440
BAM alignments	https://www.ncbi.nlm.nih.gov/sra/SRX3427716
Mitochondrial genome	https://www.ncbi.nlm.nih.gov/nuccore/KY825192.1

Table 5. ***Salminus brasiliensis*** NGS data sets and elaborated results publicly available.

Data set	Access link
Assembly	https://doi.org/10.6084/m9.figshare.18131495.v1
Genomic annotation (GFF3)	https://doi.org/10.6084/m9.figshare.11796468.v1
Table of genes characterized	https://doi.org/10.6084/m9.figshare.19169756.v1
Genome-wide microsatellite panel	https://doi.org/10.6084/m9.figshare.17754188.v1
Short reads	https://www.ncbi.nlm.nih.gov/sra/SRX13579730
BAM alignments	https://www.ncbi.nlm.nih.gov/biosample/SRS11543029
Mitochondrial genome	https://www.ncbi.nlm.nih.gov/nuccore/KY825190.1

prompt access to adequate scientific computing power being a main bridge yet to be crossed. Even practical technological developments in the area of molecular biology applied to the aquaculture field could be achieved, thanks to this new wonderland of possibilities NGS opens for those taking a bite off the cake. It is an honest testimony to the empowering effects of new technology, which was only possible because of the application of bioinformatic techniques. The expensive cake once only available to the high-brass is now within the grasp of the people. The NGS data and associated results generated by the LARGE will certainly be further mined by our own and other research groups and will aid in the future and inevitable telomere-to-telomere genome projects for these fish species.

Acknowledgements

The authors would like to acknowledge the larger team of students and collaborators, past and present, at LARGE-UFSJ. We also had support from PROPE-UFSJ and the Peixe Vivo team, Companhia Energética de Minas Gerais (CEMIG).

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). The author has been granted a DT-2 scholarship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Funding was also provided by Companhia Energética de Minas Gerais (CEMIG).

Authors' Contributions

GMY secured funds, managed and contributed to the conception of this study, performed bioinformatic analysis and wrote the manuscript. RDG performed bioinformatic assembly and annotation pipelines, analyzed data and edited the manuscript. RPS performed bioinformatic analysis. RSO supervised bioinformatic implementations, provided computer infrastructure, wrote scripts, analyzed data and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the links provided in Tables 4 and 5. Other links are provided along the text.

References

Arias, M. C., Aulagnier, S., Baerwald, E. F., Barclay, R. M., Batista, J. S., Beasley, R. R., Bezerra, R. A., Blanc, F., Bridge, E. S., Cabria, M. T., *et al*. (2016). Microsatellite records for volume 8, issue 1. *Conservation Genetics Resources*, 8(1):43–81.

Barker, N. N. (1993). "let them eat cake": The mythical marie antoinette and the french revolution. *The Historian*, 55(4):709–724.

Cao, Y.-L., Caputo, L. I., Cheng, H., da Silva Carmo, F. M., de Carvalho, L. C., de Menezes Yazbeck, G., de Oliveira Teixeira, Z., Fu, J., Guerrero, J. A., Hu, G., *et al*. (2016). Microsatellite records for volume 8, issue 3. *Conservation Genetics Resources*, 8(3):359–370.

Carvalho, L., Ribeiro, J., and Yazbeck, G. (2018). Microsatellite multiplex systems for brycon orbignyanus (characiformes: Bryconidae). *Brazilian Journal of Biology*, 79:751–754.

Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.

Graciano, R., Oliveira, R. S., dos Santos, I. M., and de Menezes Yazbeck, G. (2021). Functional dna annotation from a preliminary de novo genome assembly of brycon orbignyanus, an endangered neotropical migratory fish. *Latin American Data in Science*, 1(2):42–48.

Graciano, R. C. D., Oliveira, R. S., Santos, I. M., and Yazbeck, G. M. (2022). Genomic resources for salminus brasiliensis. *Frontiers in Genetics*, 13. DOI: 10.3389/fgene.2022.855718.

Krawczyk, F. and Kulczycki, E. (2021). How is open access accused of being predatory? the impact of beall's lists of predatory journals on academic publishing. *The Journal of Academic Librarianship*, 47(2):102271.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46.

Moed, H. F., de Moya-Anegon, F., Guerrero-Bote, V., Lopez-Illescas, C., and Hladchenko, M. (2021). Bibliometric assessment of national scientific journals. *Scientometrics*, 126(4):3641–3666.

Park, S. T. and Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International neurourology journal*, 20(Suppl 2):S76.

Polo, É. M., Santos, R. P., Moreira, L. M., and Yazbeck, G. M. (2014). Bioinformatical development of oligonucleotides for the d chain gene of the giant extracellular hemoglobin of glossoscolex paulistus. *International Journal of Bioscience, Biochemistry and Bioinformatics*, 4(2):67.

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285.

Santos, R. P., Melo, B. F., Yazbeck, G. M., Oliveira, R. S., Hilário, H. O., Prosdocimi, F., and Carvalho, D. C. (2021). Diversification of prochilodus in the eastern brazilian shield: Evidence from complete mitochondrial genomes (teleostei, prochilodontidae). *Journal of Zoological Systematics and Evolutionary Research*, 59(5):1053–1063.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology*, 12(8):1–10.

Yazbeck, G. d. M., Oliveira, R. S., Ribeiro, J. M., Carmo, F. M. d. S., and Carvalho, M. B. (2016). First complete mitochondrial genome for any anostomid fish: Leporinus piavussu, a recently described piracema species. *Mitochondrial DNA Part A*, 27(3):2293–2294.

Yazbeck, G. M., Oliveira, R. S., Ribeiro, J. M., Graciano, R. D., Santos, R. P., Carmo, F., and Lavenier, D. (2018). A broad genomic panel of microsatellite loci from brycon orbignyanus (characiformes: Bryconidae) an endangered migratory neotropical fish. *Scientific Reports*, 8(1):1–5.

Zane, L., Bargelloni, L., and Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Molecular ecology*, 11(1):1–16.