Full chloroplast genome assembly and phylogeny of “red and “yellow” Bixa orellana, “achiote”; popular source of food coloring and traditional medicine.

Jorge Villacres Vallejo
Instituto de Medicina Tradicional
José Aranda Ventura
Instituto de Medicina Tradicional
Anna Wallis
Cornell University
Robin Cagle
University of Washington
Sara M. Handy
US Food and Drug Administration
Jeffery Davis
University of Maryland
Monica Pava-Ripoll
US Food and Drug Administration
David Erickson
Joint Institute for Food Safety and Applied Nutrition
Padmini Ramachandran
US Food and Drug Administration
Andrea Ottesen (andrea.ottesen@fda.hhs.gov)
US Food and Drug Administration https://orcid.org/0000-0002-6425-7943

Research article

Keywords: Bixa orellana, achiote, annatto, achote, red and yellow achiote, bixin, norbixin, food coloring, Bixaceae, Malvales

Posted Date: April 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20035/v1

Version of Record: A version of this preprint was published at BMC Genomics on August 6th, 2020. See the published version at https://doi.org/10.1186/s12864-020-06916-0.

Title Page

Title

Full chloroplast genome assembly and phylogeny of “red and “yellow” *Bixa orellana*, “achiote”; popular source of food coloring and traditional medicine.

Authors

Jorge Villacres Vallejo[1,2], José Alberto Aranda Ventura[1, 2], Anna Wallis[4], Robin Cagle[5], Sara M. Handy[3], Jeffery Davis[6], Monica Pava-Ripoll[8], David Erickson[7], Padmini Ramachandran[3], Andrea Ottesen*[3, 9]

1. Seguro Social de Salud, Instituto de Medicina Tradicional, Iquitos, Perú
2. Universidad Nacional de la Amazonía Peruana, Facultad de Agronomía, Iquitos, Perú
3. Office of Regulatory Science, Center for Food Safety and Applied Nutrition (CFSAN) College Park, MD, USA
4. Department of Plant Pathology, Cornell University, Ithaca, NY, USA
5. Department of Microbiology, University of Washington, Seattle, WA, USA
6. Department of Chemistry and Biochemistry, University of Maryland, College Park, USA
7. Joint Institute for Food Safety and Applied Nutrition (JIFSAN), College Park, Maryland USA
8. Office of Food Safety, Center for Food Safety and Applied Nutrition (CFSAN) College Park, MD, USA
9. Department of Plant Sciences and Landscape Architecture, University of Maryland, College Park, MD, USA

*To whom correspondence should be addressed: Andrea.Ottesen@fda.hhs.gov

Abstract

Background

Seeds from *Bixa orellana*, commonly known as “achiote” and “annatto” produce bixin and norbixin apocarotenoids which impart bright red and orange colors that have been used for thousands of years for food, medicine and body painting by indigenous Americans, and by Europeans for ~ 500 years as food coloring, especially for cheeses. Use of *Bixa* colorants continues to grow as synthetic dyes come under increased scrutiny for toxicity to human and environmental systems. There is a wide range of color variation in pods of *Bixa orellana* for which genetic loci that delineate phenotypes have not yet been identified. Whole chloroplast genomes and raw genome skims provide a wide variety of genetic markers that can be used for identification purposes as well as phylogenetic inference of broad scale evolutionary relationships. Here we apply whole chloroplast genome sequencing of “red” and “yellow” individuals of *Bixa orellana* for phylogenetic analyses to explore the position of Bixaceae relative to other families within the Malvales as well as to underpin future work that may delineate diverse color phenotypes.

Results

Fully assembled chloroplast genomes were produced for both red and yellow *Bixa orellana* accessions (158,918 and 158,823 bp respectively). Synteny and gene content was identical to the only other previously reported full chloroplast genome of *Bixa orellana* (NC_041550). We observed a 17 base pair deletion at position 58399-58415 in both of our accessions, relative to NC_041550 and a 6 base pair deletion at position 75531-75526 in the accession of “red” *Bixa*. A phylogeny based on alignment free kmer distance metrics was used to confirm monophylly of *Bixa* accessions, and to place Bixaceae relative to other families within the Malvales.

Conclusions

Our data support Bixaceae as sister to Malvaceae and identified several potentially diagnostic insertion-deletion mutations that may with future work, reliably distinguish between red and yellow phenotypes. In addition to utility for phylogenic questions and development of identity markers, we demonstrate that chloroplast genomes can be used in conjunction with modern bioinformatic search tools (kmer based) to provide rapid and precise identification of *Bixa orellana* for Next Generation Sequencing approaches to natural product authentication.

Keywords *Bixa orellana*, achiote, annatto, achote, red and yellow achiote, bixin, norbixin, food coloring, Bixaceae, Malvales

Background

The neotropical plant *Bixa orellana* (also known as “annatto”, “achiote”, “achiote amarrillo”, and “achote”) is widely used in commercial food production as a food colorant[1]. The red, orange, yellow and magenta colors of bixin and norbixin apocarotenoids occur in arils that surround the seeds. *Bixa* has been used in food, medicine, and body decoration in the Americas for millennia and by Europeans since the 16th century, especially for cheddar cheeses. In poultry production, the colorant is added to chicken feed to create richer shades of yellow and orange in egg yolks. *Bixa orellana* also has a long history of use in traditional medicine throughout the Americas[2-4].

Use of natural color from *Bixa orellana* in the food and textile industries continues to grow in popularity as the use of synthetic dyes (especially azo dyes) comes under increased scrutiny due to their potential toxicity and detrimental impact on the environment[5-8]. While there have been sparse reporting of allergens potentially associated with annatto[9], it is general considered safe (GRAS) by the United States Food and Drug Administration [10, 11].

Effective means of identifying *Bixa* for authentication purposes in food and other commercial products is important. Recently, use of whole genomes as references (as opposed to DNA barcodes[12-16] for species identification programs has provided important reference databases and a greater variety of genetic markers that can be applied to identification/authentication protocols[17, 18]. Development of species specific markers that confirm the identify of natural products, relies on well curated databases of multiple high quality accessions[16].

An interesting phenomenon in *Bixa orellana* is the occurrence of multiple variations of seed pod color (from green to yellow to magenta). Even arils possess a great range of diverse reds, oranges and yellows that significantly impact color and use[19]. To identify genetic loci that distinguish color variation, it may be necessary to interrogate nuclear DNA, however, chloroplast genomes have previously demonstrated utility for disambiguation of closely related species. Thus, in the hope that full chloroplast genomes might differentiate “red” and “yellow” *Bixa* varieties, one individual of each of color, (accessions from the Instituto de Medicina Tradicional de ES SALUD (IMET) Iquitos, Peru), were used for full chloroplast sequencing.

While many more individuals will be needed to fully understand the utility of chloroplast loci to differentiate color variants, this work represents a valuable starting point to address this data gap. Whole chloroplast genomes are a powerful tool for phylogenetic reconstruction and inference of broad scale evolutionary relationships. Here, we compare the assembled chloroplast genomes of “red” and “yellow” *Bixa orellana* individuals with existing published *Bixa orellana* accessions, comparing gene count, organization and synteny among representative Malvales. A phylogeny based on alignment free kmer distance metrics is used to confirm monophylly of the *Bixa* accessions, and to explore the position of Bixaceae relative to other families within the Malvales. We observed a 100% accuracy in assigning *Bixa* sequence data to *Bixa* reference chloroplast genomes (e.g. Figure 3a). When *Bixa* sequence data was

combined with other data, we again observed that *Bixa* could be found even when combined as only 5% of all sequence data in a query (Figure 3b).

In addition to utility for phylogenic questions and development of identity markers, we demonstrate that chloroplast genomes can be used in conjunction with modern bioinformatic search tools (kmer based) to provide rapid and precise identification of *Bixa orellana* for modernized Next Generation Sequencing (NGS) approaches to natural product authentication and identification.

Results

Sequence Assembly

An assembled chloroplast genome was produced for both yellow and red *Bixa orellana* accessions (Figure 1). The assembled size of the chloroplast was 158,823 and 158,918 for the yellow and red accessions respectively (Table 1). The synteny and gene content of the two accessions reported here was identical to that of NC_041550. The three *Bixa* accessions contained 129 total genes, with 85 protein coding genes, 36 tRNA and 8 rRNA (Table 1 & Table S2). The 129 genes in *Bixa* represented a gain of four genes relative to *Theobroma*, due to rpl2 being captured (and therefor duplicated) by the Inverted Repeat in Bixa, as well as gain of rpl22, trnG-GCC in the LSC, and ycf15 in the Inverted Repeat. Bixa did not have the transfer-rna trnG-UCC, which was present in *Theobroma*, resulting in a net gain of 4 genes in *Bixa* relative to *Theobroma cacao*. The overall synteny of *Theobroma* and *Bixa* was continuous, and contrasted to *Heritiera fomes*, exhibited a much smaller inverted repeat (25,000 for *Bixa* and *Theobroma* contrasted to 34,494 for *Heritiera fomes*) with an accompanying decrease in gene number (129 and 125 for *Bixa* and *Theobroma* versus 130 for *Heritiera*) due to duplication of genes in single copy portions of the genome. The representative Diperocarpaceae genome was more similar in gene structure and content to *Bixa* and *Theobroma* than was the *Heritiera*, with a similar size of the Inverted Repeat (23,911) gene number (130) and relative number of CDS, tRNA and rRNA (86, 36 and 8) (Table 1).

Sequence variation and genetic markers

MAFFT sequence alignment of the three chloroplast genomes of *Bixa* allow us to screen the samples for genetic variants. We observed 17 base pair deletion at position 58399-58415 in in yellow *Bixa* FDAWG01 relative to NC_041550 that was present in both of our accessions, and a 6 base pair deletion at position 75531-75526 in “red” *Bixa* FDARB01.

Phylogenetic relationships

The use of alignment free kmer distance metrics allows for the comparison of all DNA sequence data among distantly related chloroplast genomes, not just among genes which can be easily compared through sequence alignment. As such, we were able to make direct comparisons among the full genomes of relatively divergent species (4 families of Malvales and one Brassicales out group) (Figure 2). The D2 metric distance matrix (Table S3) produced a NJ tree that places Bixaceae sister to Malvaceae within the Malvales (similar to that observed by Pacheco et al (2019) [11]. The three *Bixa* accessions grouped together in a discrete

monophyletic group, reflecting the accuracy of the assembly, and the putative value of the entire chloroplast as a diagnostic marker for the group. The Thymelaeaceae were sister to the Bixaceae-Malvaceae clade, with the Dipterocarpaceae basal within the order. This differs to the APG IV assessment (www.mobot.org/MOBOT/research/APweb/) which places Thymelaeaceae basal to both Dipterocarpaceae and Bixaceae.

NGS Applications for Identification

We combined our two assembled *Bixa* chloroplast genomes with all the chloroplast genomes for Malvales found within the NCBI Refseq repository (Table S1). This combined reference sequence data set was then used to format a searchable kmer based reference database with the software Genome2-ID. There were three reference *Bixa* chloroplast genomes available (the two presented here and NC_041550). We evaluated the accuracy of using whole chloroplast genomes by preparing three separate databases where all combinations of two *Bixa* chloroplasts genomes were included, and then used the whole genome shotgun/skim data from our two samples (SRR10320715 and SRR10320716) as well as data from available SRA accessions (SRR7941588 - SRR7941591) as input to test if the metagenomics software would correctly and unambiguously identify the raw data as that of *Bixa orellana*. The use of three databases allows us to avoid the circularity of testing the same data used in chloroplast assembly against its own assembled chloroplast genome (referred to as a 'take-one-out' strategy). We observed a 100% accuracy in assigning *Bixa* sequence data to *Bixa* reference chloroplast genomes (e.g. Figure 3a). When *Bixa* sequence data was combined with other data, we again observed that *Bixa* could be found even when combined as only 5% of all sequence data in a query (Figure 3b).

Discussion

Recent reports have suggested that for many organisms, particularly plants and microbes, the traditional DNA barcodes (rbcL+matK and 16S respectively) are not sufficient for the accurate discrimination among related species (CBOL et al., 2009; Coissac et al., 2016; Edgar, 2018). This observation motivates use of entire genomes, plastomes and mitochondria, to enable the unambiguous discrimination among species. Work on closely related *Echinacea* species by Zhang et al. demonstrated that complete plastomes provide unambiguous resolution among closely related species when considering the entire set of species in the genus, whereas use of the official DNA barcodes fail to discriminate among closely related species with statistical support, even collapsing the closely related genus, *Parthenium*, into a single clade with *Echinacea[15]*.

The use of multiple plastid genes in identifying and resolving species is widely used[20]. Because the use of the entire plastome as the core reference marker gene is now routinely possible many challenges may be sufficiently resolved with this approach. Use of DNA barcodes aimed to standardize the set of genes used for species identification, thereby making comparisons orthogonal, but the use of plastomes as the *defacto* DNA barcode will greatly improve identification rates while maintaining the value of orthogonal contrasts among studies and the standardization of reference sequence databases.

We observed that in all comparisons where shotgun/skim WGS data from *Bixa orellana* were queried against kmer formatted reference genome databases that contained *Bixa* chloroplast reference genomes, we consistently and uniquely identified *Bixa orellana* (Figure 3a). Likewise, when data from *Bixa orellana* was mixed with other species, e.g. *Oryza sativa*, both species could be recovered and identified even when the proportion of *Bixa* declined, with the lowest limit of 5% exhibiting clearly diagnostic identification (Figure 3b). As such, use of the entire chloroplast genome functioned as a genome scale DNA barcode and readily distinguished *Bixa*, even in admixture.

Conclusions

Use of complete chloroplast genomes may provide insights into the phylogenetic relationships among groups, allowing us to take advantage of all data within these genomes, not solely those elements which can be easily analyzed through multiple sequence alignment. An increasing number of published studies have used a combination of coding genes extracted from diverse chloroplast genomes for use in phylogenetic reconstruction[21, 22]. We employed kmer based alignment free distance metrics, D2 and D2*, to estimate genetic distances among a set of representative Malvales chloroplast genomes (Table S1). We observed that Bixaceae was sister to the Malvaceae, which contrasts with APG IV, but which agrees with Pacheco et al (2019) which similarly placed Bixaceae sister to Malvaceae.

The structure of the two chloroplast genomes presented here was effectively identical to that of the published chloroplast genome of *Bixa orellana* (NC_041550.1). Complete synteny was maintained, and the size and gene content of the inverted repeats was the same in all three accessions (Table 1). We found several potentially diagnostic insertion-deletion mutations that may distinguish among *Bixa* red and yellow phenotypes which will be the subject of future work. Accession NC_041550.1 exhibited a 17 base gap at 58,299 in a intergenic spacer between *atp*B and *rbc*L relative to the other accessions, our Red *Bixa* contained a 6 base gap at 75,537 in the intronic region of *clp*P, in addition to 5 SNP. The number of genes total was consistent with that observed in other Malvales (Table 1) with the exception of *Heritiera fomes* which had an expanded inverted repeat which resulted in the duplication of several genes that are single copy in other Malvales.

Future work with increased representation of individuals from diverse color phenotypes will be used to see if the potentially diagnostic insertion-deletion mutations reliably differentiate color variants. Additionally, we demonstrated that chloroplast genomes, used in conjunction with modern bioinformatic search tools (kmer based) can provide rapid and precise identification of *Bixa orellana* for use with modernized Next Generation Sequencing (NGS) approaches to natural product authentication.

Methods

Sampling, Library Preparation and Sequencing

Individuals were collected from *Bixa orellana* accessions in the display/research garden of the Instituto de Medicina Tradicional de ES SALUD (IMET), (Iquitos, Peru). DNA extraction from leaves was performed using DNeasy Plant Mini Kit from Qiagen® according to the manufacturers specifications. DNA was prepared for sequencing using the Nextera™ DNA Flex Library Prep Kit according to the manufacturer's specifications. Libraries were loaded onto an Illumina MiSeq using a V2 cartridge with read lengths set for (2 x 250).

Sequence Assembly

Following DNA sequencing, data were trimmed to remove adapters and indexes using BCL2fastq. Sequences were then paired and quality trimmed using Usearch11[17]. Quality trimming parameters trimmed sequences to a median Q score of 25, with minimum length of 100bp, and sequences below Q25 or 100 bp length discarded. Following read preparation, genome assembly followed the workflow described in [23] where reads are initially mapped to the nearest relative to enrich for chloroplast derived reads, and then denovo assembled to establish initial contigs, followed by subsequent subdivision and reassembly of contigs in an iterative fashion until the set of contigs can be combined to single complete chromosome. Completion of assembly was inferred by identifying inverted repeats and overlap between beginning and ending sequence reflecting assembly of a complete circle.

Paired, filtered reads were mapped to *Bixa orellana* (NC_041550) to enrich reads for chloroplast specific sequences using Geneious R11 (https://www.geneious.com) Mapped reads were then denovo assembled using SPAdes[24], with kmer sizes 21, 35 and 55. In parallel, all paired, filtered reads were denovo assembled with SPAdes to generate contigs independently of the reference genome to ensure all changes in structure and synteny could be identified. The denovo contigs of both assemblies were merged and assembled in Geneious R11. Assembled contigs were compared to the reference genome and gaps were filled by extracting sequences adjacent to the gaps and mapping the paired filtered reads to those sequences using Bowtie assembler [25] in a reference guided assembly. Contigs extended by Bowtie were then subsequently combined with previous contig sets and again compared to the reference genome. The assembled plastome was tested for the large inverted repeats, and the presence of both repeats was indicative of complete chromosome assembly. Raw sequence data were deposited in NCBI Sequence Read Archive (SRR10320715 and SRR10320716) as part of the FDA GenomeTrakrCP: chloroplast DNA for botanical product identification ([16], BioProject PRJNA325670).

Sequence Annotation and Comparative gene content

Sequences were annotated with the Verdant online chloroplast assembly tool [26] (see: http://verdant.iplantcollaborative.org/plastidDB/). The existing reference for *Bixa*, NC_041550, was added to the Verdant database prior to upload and analysis of our sequences. The resulting gff files were downloaded and used to annotate the assembled *Bixa* chloroplast genomes (Figure 1). Each CDS was extracted and translated to ensure correct open reading frames, and tRNA were extracted and assembled with homologous tRNA from the reference *Bixa* to ensure annotation captured the entire gene. Following evaluation and validation of all

annotated genes, annotations were exported with the position, identity and phase of each gene. We then compared the number, synteny and organization of the set of genes relative to other chloroplast genomes within the Malvales (*Theobroma* and *Heritiera* (Malvaceae), *Vatica* (Dipterocarpaceae) and *Daphne* (Thymelaeceae)). Genes were classified based on the type of gene (CDS, tRNA, rRNA) as well as their organization within the chromosome (Large Single Copy (LSC), Inverted Repeat (IR), Small Single Copy (SSC)). The order of all genes was established and changes in the structural order of genes along the chromosome for all 5 chloroplasts was compared. Gain or loss of genes was denoted, and movement of genes among subunits of the genome (LSC, IR, SSC) based on expansion of the Inverted Repeat was recorded.

Phylogenetic Inference

Representative Chloroplast genome sequences for Malvales were downloaded from the GenBank RefSeq library for use in evaluation of the monophylly of the *Bixa orellana* genomes, as well as for inference of the relationship of Bixaceae to Malvales, and in particular with respect to its status as sister group to either Malvaceae or Thymelaeceae. The two Bixa assembled in this study in addition to the existing Bixa reference genome were included in a group of 37 additional Malvales chloroplast genomes (see table S1) as well as *Arabidopsis thaliana* from the sister order Brassicales as an outgroup. The 41 chloroplast genomes were analyzed with the Café (accelerated alignment free sequence analysis [27] kmer distance program to infer genetic distance (using k=8 in conjunction with the D2 genetic distance metric[28]. The resulting phylip formatted distance matrix was imported in PAUP [29] where *Arabidopsis thaliana* was set as the outgroup, and a NJ distance tree was computed. The resulting phylogram was exported and annotated with family groupings.

List of abbreviations

NGS: Next Generation Sequencing
LSC: Large Single Copy
IR: Inverted Repeat
SSC: Small Single Copy
CDS: CoDing Sequence
DNA: DeoxyriboNucleic Acid
NCBI: National Center for Biotechnology Information
Q score: Quality score

References

1. Raddatz-Mota, D., et al., *Achiote (Bixa orellana L.): a natural source of pigment and vitamin E.* Journal of food science and technology, 2017. **54**(6): p. 1729-1741.
2. Duke, J.A., M.J. BogenSchutz-Godwin, and A.R. Ottesen, *Duke's handbook of medicinal plants of Latin America*. 2008, CRC Press.
3. Khare, C.P., *Indian Medicinal Plants*. 2007: Springer Science and Business Media.
4. Vasquez, R., J.A. Duke, and A. Ottesen, *Amazonian Ethnobotanical Dictionary*. 2020: Vertvolta.
5. Kusic, H., et al., *Photooxidation processes for an azo dye in aqueous media: Modeling of degradation kinetic and ecological parameters evaluation.* Journal of Hazardous Materials, 2011. **185**(2): p. 1558-1568.
6. Khalid, A., M. Arshad, and D.E. Crowley, *Accelerated decolorization of structurally different azo dyes by newly isolated bacterial strains.* Applied Microbiology and Biotechnology, 2008. **78**(2): p. 361-369.
7. Bae, J.-S. and H.S. Freeman, *Aquatic toxicity evaluation of new direct dyes to the Daphnia magna.* Dyes and Pigments, 2007. **73**(1): p. 81-85.
8. Öztürk, A. and M.I. Abdullah, *Toxicological effect of indole and its azo dye derivatives on some microorganisms under aerobic conditions.* Science of The Total Environment, 2006. **358**(1): p. 137-142.
9. Ramsey, N.B., et al., *Annatto seed hypersensitivity in a pediatric patient.* Annals of Allergy, Asthma & Immunology, 2016. **117**(3): p. 331-333.
10. FDA, *GRAS Notices.* https://www.accessdata.fda.gov/scripts/fdcc/index.cfm?set=grasnotices&id=471. **GRN No. 471**.
11. FDA, *CFR (Code of Federal Regulations) Title 21.* LISTING OF COLOR ADDITIVES EXEMPT FROM CERTIFICATION, 2019. **Subpart A Foods**.
12. Coissac, E., et al., *From barcodes to genomes: extending the concept of DNA barcoding.* Molecular Ecology, 2016. **25**(7): p. 1423-1428.
13. Hollingsworth, P.M., et al., *A DNA barcode for land plants.* Proceedings of the National Academy of Sciences, 2009. **106**(31): p. 12794.
14. Ottesen, A. and G. Ziobro, *Use of DNA Biocoding to Detect Adulteration of Star Anise (Illicium verum) by Other Illicium Species.* Economic Botany: Applied Plant Biology, 2006.
15. Zhang, N., et al., *An analysis of Echinacea chloroplast genomes: Implications for future botanical identification.* Scientific reports, 2017. **7**(1): p. 216.
16. Zhang, N., et al., *Development of a reference standard library of chloroplast genome sequences, GenomeTrakrCP.* Planta medica, 2017. **83**(18): p. 1420-1430.
17. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-2461.
18. Hebert, P.D.N., et al., *Biological identifications through DNA barcodes.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 2003. **270**(1512): p. 313-321.
19. Barrera Ruiz, T.U.L., Jessy Luis, *EVALUACIÓN DEL CONTENIDO DE BIXINA EN Bixa orellana L. (ACHIOTE) DEL BANCO DE GERMOPLASMA DE ZUNGAROCOCHA*, in

FACULTAD DE AGRONOMIA 2019, Universidad Nacional de la Amazonia Peruana: *Iquitos, Peru*.
20. Shaw, J., et al., *The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis.* American Journal of Botany, 2005. **92**(1): p. 142-166.
21. Lloyd Evans, D., S.V. Joshi, and J. Wang, *Whole chloroplast genome and gene locus phylogenies reveal the taxonomic placement and relationship of Tripidium (Panicoideae: Andropogoneae) to sugarcane.* BMC Evolutionary Biology, 2019. **19**(1): p. 33.
22. Jansen, R.K., et al., *Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids.* BMC Evolutionary Biology, 2006. **6**(1): p. 32.
23. Lischer, H.E.L. and K.K. Shimizu, *Reference-guided de novo assembly approach improves genome reconstruction for related species.* BMC Bioinformatics, 2017. **18**(1): p. 474.
24. Nurk, S., et al. *Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads*. in *Research in Computational Molecular Biology*. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.
25. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
26. McKain, M.R., et al., *Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes.* Bioinformatics, 2016. **33**(1): p. 130-132.
27. Lu, Y.Y., et al., *CAFE: aCcelerated Alignment-FrEe sequence analysis.* Nucleic Acids Research, 2017. **45**(W1): p. W554-W559.
28. Song, K., et al., *New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing.* Briefings in Bioinformatics, 2013. **15**(3): p. 343-353.
29. Swofford, D.L., *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0.b5*
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.458.6867, 2001.

Ethics Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Raw sequence data were deposited in NCBI Sequence Read Archive (SRR10320715 and SRR10320716) as part of the FDA GenomeTrakrCP: chloroplast DNA for botanical product identification, BioProject PRJNA325670).

Competing interests

The authors declare that they have no competing interests.

Funding

We would like to acknowledge the lab of Dr. Kerik Cox at Cornell University, the lab of Dr. Villacrés Vallejo, and the Division of Microbiology, Office of Regulatory Science, Center for Food Safety and Applied Nutrition (CFSAN), FDA for provision of materials and reagents.

Authors' Contributions

AO, JV, PR, DE, SH designed the study. JV, RC, AW, AO, PR participated in field and laboratory work. DE, PR provided conducted bioinformatic analyses. DE, MPR, JVV, JAAV, AO, SH, JD wrote and edited the manuscript. JD provided chemistry editorial guidance.

Acknowledgements

We would like to thank the staff of the Instituto de Medicina Traditional in Iquitos, Peru for their beautiful up-keep of the Medicinal Plants Garden. Additionally, we would like to thank Basilio Sahuarico for guidance at the garden.

Figures

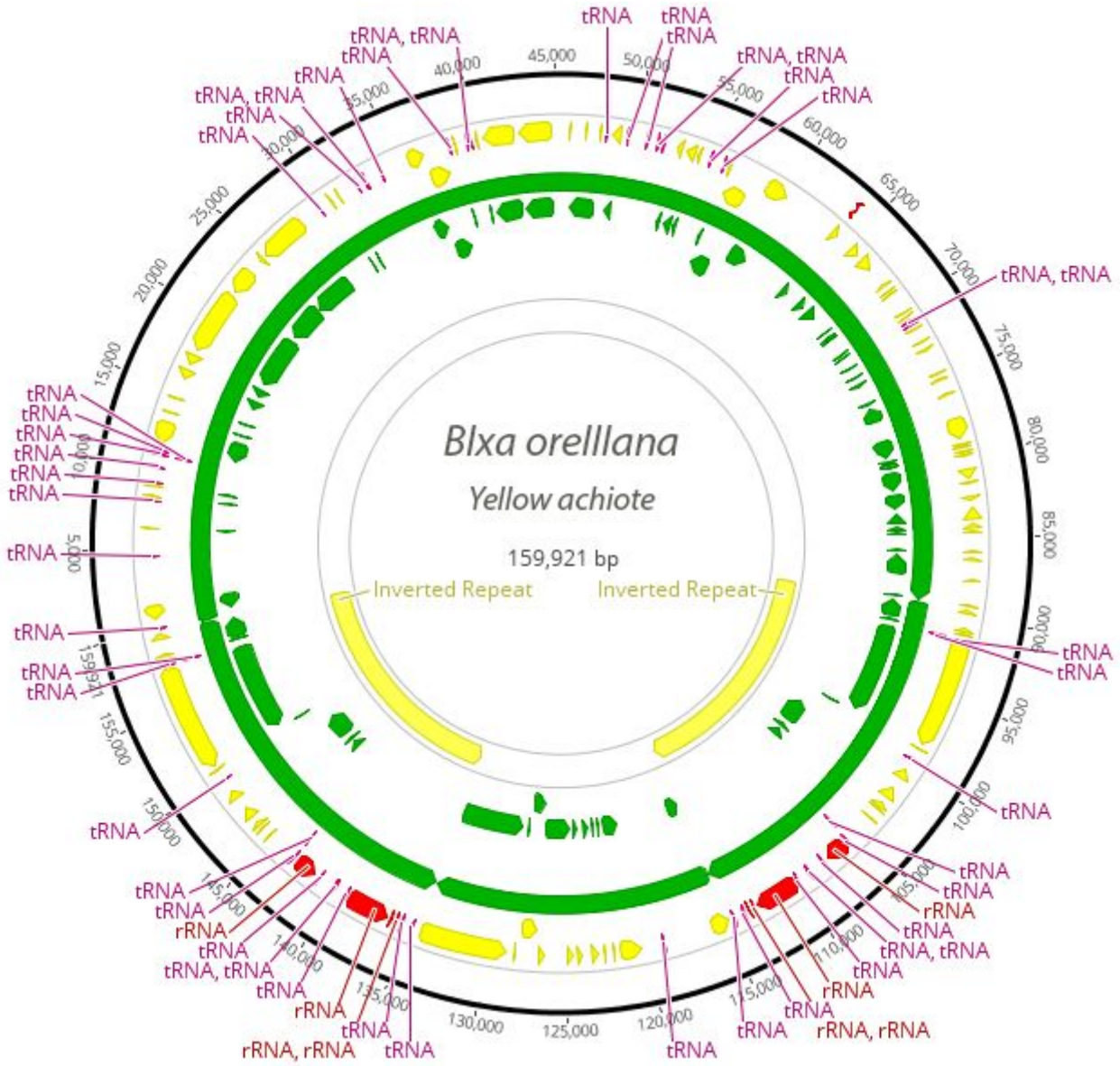

Figure 1

An assembled chloroplast genome was produced for both yellow and red Bixa orellana accessions

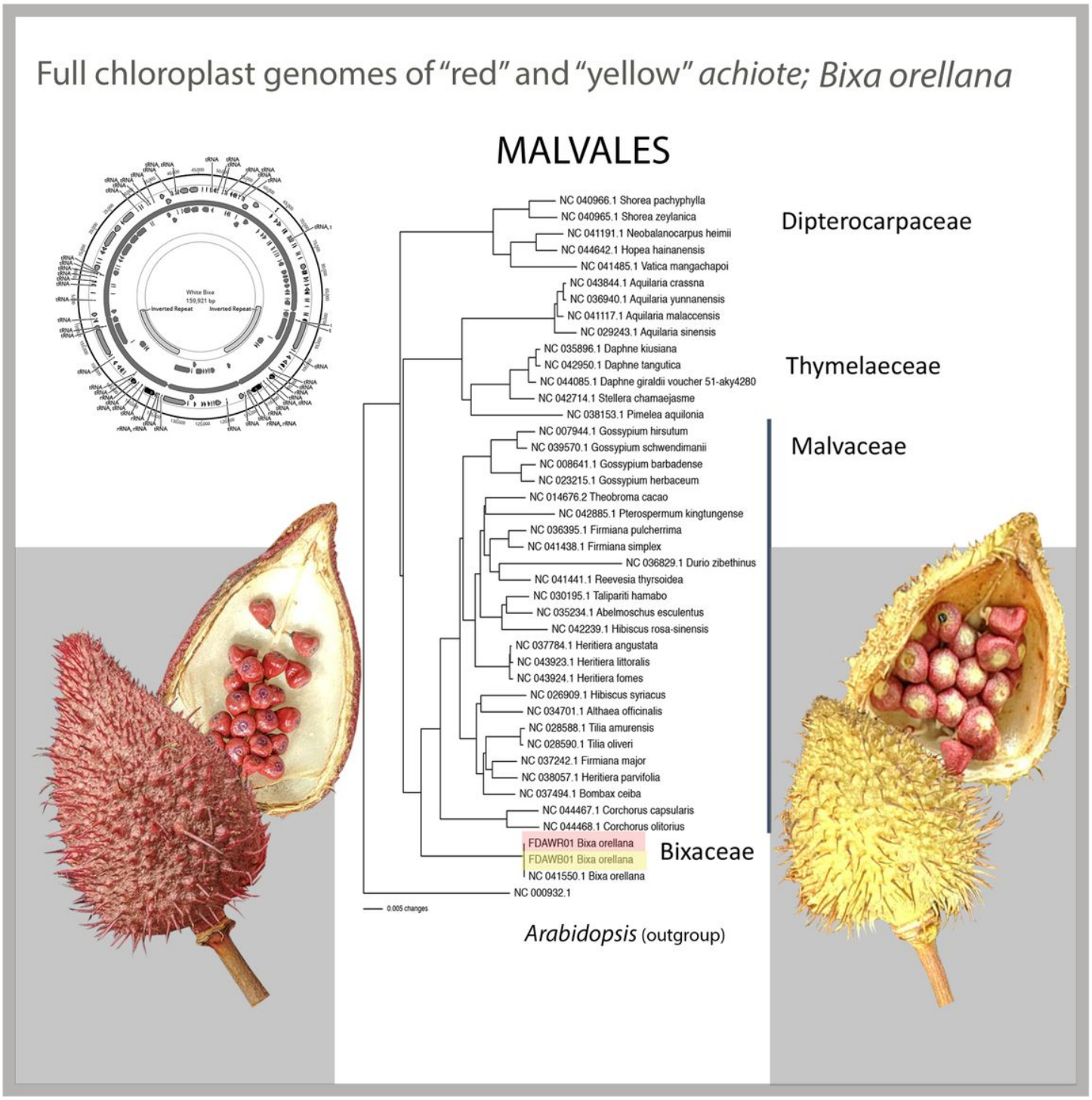

Figure 2

We were able to make direct comparisons among the full genomes of relatively divergent species

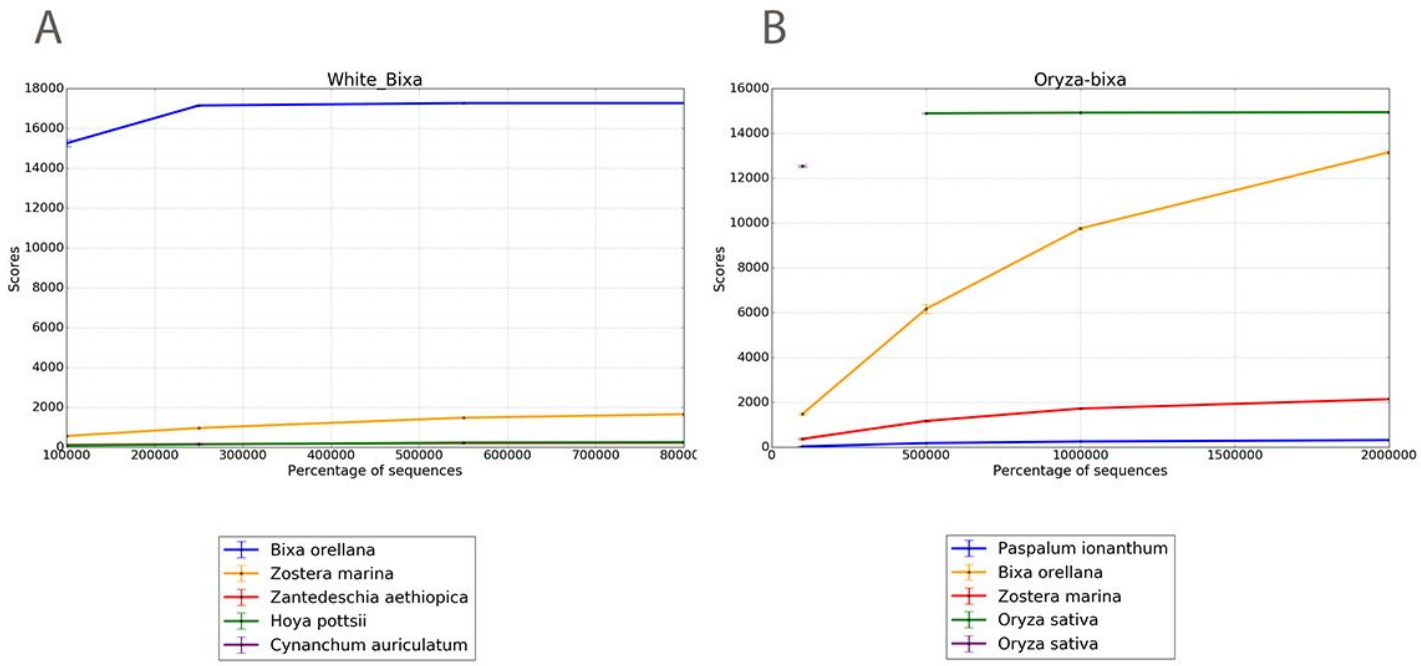

Figure 3

a. We observed a 100% accuracy in assigning Bixa sequence data to Bixa reference chloroplast genomes. b. When Bixa sequence data was combined with other data, we again observed that Bixa could be found even when combined as only 5% of all sequence data in a query

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2.xlsx
- Table1.docx
- TableS3result.D2V1.phylip
- TableS1.xlsx