

Comparison of *de novo* assemblies from Illumina and Oxford nanopore sequences in *Dynastes tityus*

Silvia García Juan
Jaime Níguez Baeza

Contents Index

1. Background	1
2. Methods	1
2.1. Data collection	1
2.2. Quality control	2
2.2.1. FASTQC	2
2.2.2. Trimmomatic	3
2.2.3. Cutadapt	4
2.3. Assembly of reads into contigs and scaffolds	4
2.4. Genome annotation	5
2.5. Annotation analysis	5
3. Results	5
3.1. Quality control of raw sequences	5
3.2. Annotation	6
4. Discussion	8
5. Conclusions	9
6. References	11

Figures Index

1. Figure 1	3
2. Figure 2	4
3. Figure 3	7
4. Figure 4	7
5. Figure 5	8
6. Figure 6	8

1. Background

The development during the last decades of new sequencing technologies has transformed the field of genomics, allowing the obtaining of high-quality genomes for a wide variety of organisms at an increasingly reduced price (Lu and Huang, 2018) (Mardis, 2017). In particular, Illumina’s short read sequencing technology has been widely used for the generation of reference genomes due to its high read coverage, precision (lower error rate), and low cost per base. (Kolmogorov *et al.*, 2020) (Goodwin *et al.*, 2016). However, this technology also presents a series of limitations when it comes to sequencing complex and repetitive genomic regions, in which assembly errors can appear. These drawbacks of Illumina sequencing when obtaining genomes with certain characteristics may make it more recommendable to alternatively apply long read sequencing offered by Oxford Nanopore Technologies (Wenger *et al.*, 2019) (Trofimova *et al.*, 2023). . The disadvantage of this technology is that its cost per base is higher, it has less reading coverage and a higher error rate, which ultimately results in lower quality readings (Hu *et al.*, 2021).

Taking into account these differences between both sequencing technologies, in this work two de novo assemblies will be carried out from Illumina and Oxford Nanopore Technologies sequences for *Dynastes tityus*, a species of beetle, with the aim of comparing the assembly results of both sequencing techniques independently, evaluating the strengths and limitations of each one for the specific organism.

Once the different assemblies have been obtained, an attempt will be made to make an annotation of their content.

Finally, the convenience of using short-read or long-read sequencing technology, will be evaluated.

The results of this project aim to contribute to the selection of the most appropriate sequencing technology to obtain precise and complete reference genomes for Coleoptera species, one of the groups with the greatest diversity and interest due to its multiple interactions with humans.

2. Methods

A workflow has been carried out in Galaxy in combination with the use of other programs outside the platform with the following steps.

2.1. Data collection

The sequences for this study have been obtained from the NCBI SRA database in FASTQ (using the Wrappers Galaxy tool). The downloaded sequences correspond to results of the sequencing of the Hercules beetle species *Dynastes tityus*.

It is important to highlight that those sequences obtained using the Illumina technique have followed a paired-end layout, unlike those received through Oxford Nanopore, which have been sequenced following a single-end process.

This difference will be taken into account when performing the analysis of the results and the comparison between both sequencing technologies.

The sequencing data to be used are:

- Those obtained by Illumina sequencing with accession number SAMN26643022.
- Those obtained by Nanopore sequencing with accession number SAMN26643021.

Sequences in FASTQ format will be uploaded to the Galaxy platform using the “Faster Download and Extract Reads in FASTQ format from NCBI SRA” tool via “Get Data”.

2.2. Quality control

Quality control will be carried out through FASTQC and Trimmomatic/Cutadapt. In the case of Illumina, this process will be double, since the quality of the forward and reverse sequences will be analyzed as they are paired-end.

2.2.1. FASTQC

In the "FastQC" tool (Andrews, 2010), we will select the previously loaded fastq file and run the program to start the analysis. The files used for each of the processes are those named in the Galaxy history as:

Dynastes tityus Illumina paired-end:

- SAMN26643022_reverse.gz
- SAMN26643022_forward.gz

Dynastes tityus Nanopore single-end:

- SAMN26643021.gz

The parameters used will be the same for all the analyses and are the ones that FASTQC assigns by default.

Once the quality analysis is completed, we will obtain the results as output in the form of a FASTQC report, which contains graphs and statistics that we will analyze below, the sections of this report include:

1. **Basic Statistics:** This section provides general information about the number of sequences, the average length, the average quality, and the quality of the GC content of the sequences.

2. **Per base sequence quality:** This section displays a quality graph for each position in the sequence. Quality is represented by colors, where green represents high quality, orange is medium, and red is low quality.
3. **Per sequence quality scores:** This section displays a graph of the distribution of quality scores for all sequences.
4. **Per base sequence content:** This section shows the percentage of each base in each position of the sequence.
5. **Per sequence GC content:** This section displays a graph showing the distribution of GC content across sequences.
6. **Per base N content:** This section shows the percentage of "N"bases, nucleotides that could not be identified, at each position in the sequence.
7. **Sequence Length Distribution:** In this section, the distribution of the length of the sequences is shown.
8. **Overrepresented sequences:** This section shows the sequences that appear with a higher frequency than expected in the samples.

2.2.2. Trimmomatic

Starting from the original Illumina fastq files, a preprocessing of the sequences will be carried out using Trimmomatic (Bolger *et al.*, 2014), seeking to increase their quality by eliminating those readings below the established threshold or that correspond to adapter sequences typical of the sequencing process of illuminates. After this, a second quality analysis will be carried out using FASTQC.

The aim of this step is facing the subsequent steps of assembling these reads in contigs and ordering them in scaffolds with the maximum guarantee.

Specific parameters that will be set for this preprocessing of data (Figure 1).

Tool Parameters







Input Parameter	Value
Single-end or paired-end reads?	pair_of_files
Input FASTQ file (R1/first of pair)	5 : SAMN26643022:forward   
Input FASTQ file (R2/second of pair)	6 : SAMN26643022:reverse   
Perform initial ILLUMINACLIP step?	no
Select Trimmomatic operation to perform	SLIDINGWINDOW
Number of bases to average across	4
Average quality required	20
Output trimlog file?	False
Output trimmomatic log messages?	False
Job Resource Parameters	no

Figure 1: Input parameters for Trimmomatic.

In this case we are going to obtain 4 files as output, of which we will use those called "paired", one corresponds to the forward sequence and the other to the reverse, to perform the second FASTQC, in the same way as we did in the previous step. In this way, we will check if the quality of the sequences has been improved.

2.2.3. Cutadapt

In the same way as with Trimmomatic, we will process the sequences obtained with Nanopore using the Galaxy Cutadapt 4.0 tool (Kechin *et al.*, 2017), in order to increase the average quality of the reads.

Subsequently, the quality results will be compared with those obtained previously using the FASTQC tool, in Annex 3.

For this purpose some inputs and parameters will be set (Figure 2).

Input Parameter	Value
Single-end or Paired-end reads?	single
FASTQ/A file	6 : SAMN26643921
r1	
Source	user
Enter custom 3' adapter name (Optional if Multiple output is 'No')	Empty.
Enter custom 3' adapter sequence	AGACGTGTGCTCTTCCGATCT
Disallow indels for this adapter	False
Cut bases from reads before adapter trimming	0
adapter_options	
What to do if a match is found	Trim: trim adapter and upstream or downstream sequence
Disallow internal adaptor occurrences	Disabled
Maximum error rate	0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).	False
Match times	1
Minimum overlap length	3
Match wildcards	In the adapters but not in the reads
Look for adapters in the reverse complement	False
filter_options	
Discard Trimmed Reads	False

Figure 2: Input parameters for Cutadapt.

2.3. Assembly of reads into contigs and scaffolds

In the case of Illumina, it will be performed using SPAdes (Bankevich *et al.*, 2012), an assembler that can handle different types of sequences, including short-reads, long-reads, and mate-pairs. In addition, it is also optimized for the assembly of low-quality and fragmented genomes.

This program assembles reads into contigs (sequences longer than a single read but still do not represent the whole genome) and eventually into scaffolds (sequences larger than contigs and that

can be associated with a position in the genome).

The inputs that will be used are the ones that have reached the best quality results in the previous steps.

In the case of Nanopore assembly, the program used is wtdbg v1.2.8 ([github.com/fantasticair/wtdbg-1.2.8](https://github.com/fantasticair/wtdbg)) (Ruan and Li, 2019), which has allowed for better performance when assembling long-reads.

2.4. Genome annotation

After assembly, genome annotation will be performed, in order to predict and characterize genes and other functional regions in the genome.

For this purpose, the genomic annotation tool Augustus (Keller *et al.*, 2011), available in Galaxy, will be used. It uses sequence databases to find similarities and predict gene locations, nucleotide and amino acid sequences.

For this process, the files obtained in the scaffolds assembly in the previous step will be used, which will be compared against the model organisms *Drosophila melanogaster* and *Homo sapiens* in order to compare annotated genes with these two species as reference.

After the annotation process, the genomic characteristics found will be obtained in a GFF file, which will be analyzed in the following section.

2.5. Annotation analysis

For a better manipulation and visualization of the data contained in the GFF file, the R package rtracklayer v1.58.0 (Lawrence *et al.*, 2009) will be used to display the number of annotated elements and a statistical summary of their size distribution in a clearer way.

3. Results

3.1. Quality control of raw sequences

In the 2 reports for Illumina (forward and reverse) available in Annex 1.1 (forward) and 1.2 (reverse) from github repository, the graph "Per base sequence quality" can be observed, a uniform quality is observed throughout the sequence positions, forming a practically horizontal line, except both extremes, which is to be expected in readings generated by Illumina. In addition, the quality is high at all times, except in certain positions near the end of the reverse sequence, in which, although the average is still high, values between 22 and 28 (medium quality) begin to appear.

In the case of Nanopore, as we might expect, the average quality is lower than in Illumina sequencing. However, it is also constant throughout the entire sequence with the exception of the extremes,

being around 25, medium quality.

Somewhat redundantly, we can observe a very high quality in the "Per sequence quality scores"graph, with most sequences presenting an average quality above 37 for Illumina sequences and 26 for Oxford Nanopore sequences.

As a curiosity, it is worth noting the presence of a lateral peak in the "Per sequence GC contentrepresentation with a%GC around 55 %, compatible with the expected values of the adapters necessary in Illumina sequencing. Another much less pronounced lateral peak appears in the case of Oxford Nanopore, which may be due to some noise.

Despite the fact that the quality obtained is quite high, in Illumina Trimmomatic was used next, a tool which by filtering contaminating adapter sequences and those that present low quality.

The quality results for the sequences after applying this preprocessing are found in Annex 2.1 (forward) and 2.2 (reverse).

When comparing these results with those obtained from the raw sequences, very few differences are observed, the most notable being the increase in quality at the ends, as we can see in the "Per base sequence quality"graph, which we can infer. which is due to the elimination of those sequences with lower quality.

The rest of the results are equivalent to those obtained previously, so it was considered that the slight increase in quality after Trimmomatic did not compensate for the loss of information involved in eliminating part of the readings, which also had sufficient quality as described.

3.2. Annotation

In the assembly obtained from Illumina reads, 508 candidate genes were annotated in 1060 CDS and 552 introns using *Drosophila melanogaster* as a model organism. Subsequently, using the human genome as a model, this number increased to 705 candidate genes in 1055 CDS, and 350 introns were identified. The assembly annotation files have been attached to the specified github repository. The size of the annotated genes has the following characteristics (Figure 3). We can also observe the frequency of genes size in each case (Figure 4).

Size (pb)	Annotation Illumina <i>D. melanogaster</i>	Annotation Illumina <i>H. sapiens</i>
Min	200	200
First quantile	679.2	566
Median	1170.5	926
Mean	1687.2	1160
Third quantile	2046.5	1499
Max	16431.0	8119

Figure 3: Statistical summary of genetic sizes for prediction from short-reads.

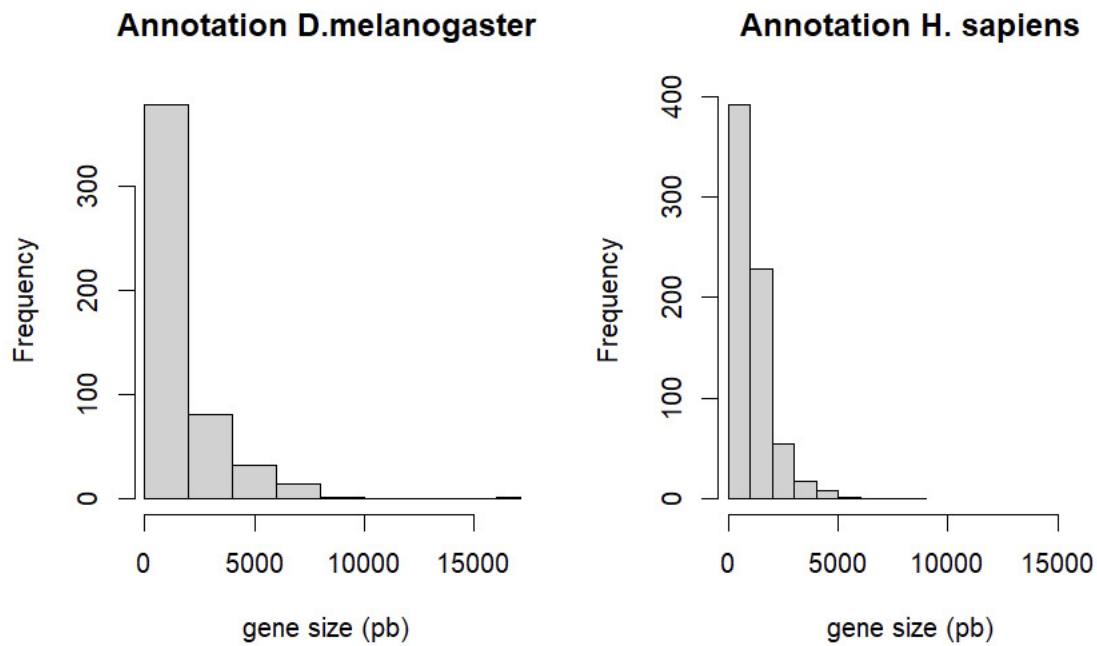


Figure 4: Frequency of predicted genes according to their length in base pairs.

On the other hand, the results for the assembly generated from long reads (Oxford Nanopore) allowed us to predict a total of 12056 genes in 40243 CDS, as well as 28187 introns.

The statistical analysis of these genes can be observe as a result (Figure 5) and again the frequency of genes size can be observed (Figure 6).

Size (pb)	Annotation Nanopore
Min	200
First quantil	485
Median	1712
Mean	4301
Third quantil	4473
Max	297971

Figure 5: Statistical summary of gene sizes for prediction from long-reads.

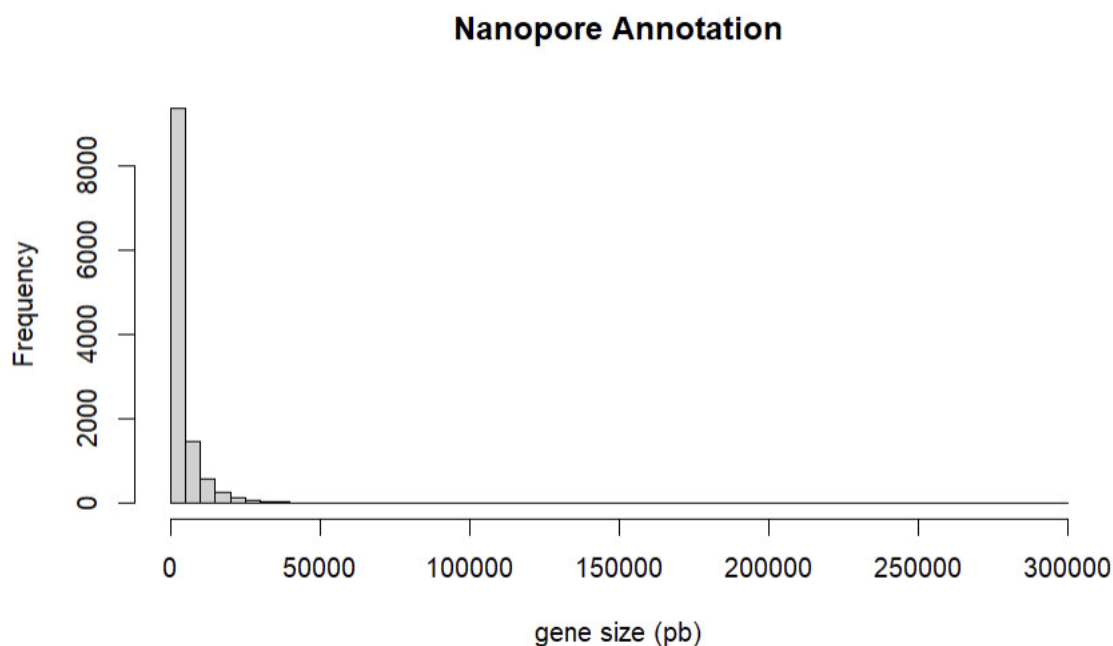


Figure 6: Frequency of genes predicted according to their length in base pairs.

4. Discussion

Based on the several differences between the results of both analyses, a comparison can be established between the results of both assemblies, using short-read sequencing with Illumina and long-read sequencing with Oxford Nanopore.

Since the first quality control, it was clear that the quality was much higher in the short-reads of Illumina, however, sequences obtained with long-reads Oxford Nanopore technology presented an acceptable and consistent quality. This difference in quality, nevertheless, did not result in better

performance of the annotation from the assembly produced with short-reads, as a much larger number of genes could be predicted in the case of the Nanopore assembly.

In future studies, it would be interesting to perform an analysis of the quality of the assemblies with programs such as QUAST. In the context of this project, it was not computationally possible to obtain this result, but it is considered as a feasible idea to continue analyzing these results. More exhaustive analyses that are beyond the scope of this work could also be carried out, such it was done in other studies consulted (Gan *et al.*, 2019).

Thanks to the de novo assemblies of long-reads and short-reads of the *Dynastes tityus* beetle, valuable and complementary information about its genome has been obtained. On the one hand, the use of long-reads allows the identification of highly repetitive genomic regions that probably could not be assembled with short-reads, which allows for a more complete and less fragmented assembly. On the other hand, short-reads allowed for better coverage and sequence quality.

With a more thorough analysis of the information, the function of the genes could be characterized and related to the biology of the *Dynastes tityus* species, and how these genes and genomic regions are related to important biological characteristics in this species could be analyzed. In this way, new information could also be obtained about important factors in its adaptation to the environment.

To obtain a more complete set of genes and a better understanding of the biology of this species, a hybrid assembly combining the advantages of both methods could be performed, allowing for a more complete and precise assembly of the beetle's genome. In this work, this was not possible due to computational and storage limitations, since the genome of beetle species is very extensive. RNA (mRNA) and protein read mapping techniques could also be used to validate and improve genome assembly and identify functional genes.

The completion and results of this project highlight the complexity of de novo assembly, especially when it comes to eukaryotic genomes. Furthermore, the differences between the results of both assemblies may be due to the fact that they are not the same sample, so the optimal option may be to perform a hybrid of both assemblies to combine the information provided by both techniques.

5. Conclusions

Fundamental differences have been observed in the results of the different assemblies for the genome of *Dynastes tityus*, which reinforces the need to use both sequencing technologies to improve the knowledge of the genome of this beetle, which could be extrapolated to future studies in Coleoptera.

The fact that genome annotation using long-reads sequences improved the performance in comparison with the short-reads could indicate that this specie could present a complex genome or repetitive regions, as this is one of the main advantages of this technology over short-reads.

This study sets important implications for future research in the field of insect genomics. The use of next-generation sequencing technologies and the combination of assembly approaches can provide a more complete understanding of the biology and genetic diversity of Coleoptera, which could have

applications in species conservation, agriculture, and public health. Additionally, the results of this study provide a basis for future research on the biology and evolution of Coleoptera and other insects species.

6. References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16. PMID: 22506599; PMCID: PMC3342519.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

Gan HM, Tan MH, Austin CM, Sherman CDH, Wong YT, Strugnell J, Gervis M, McPherson L, Miller AD. Best Foot Forward: Nanopore Long Reads, Hybrid Meta-Assembly, and Haplotig Purging Optimizes the First Genome Assembly for the Southern Hemisphere Blacklip Abalone (*Haliotis rubra*). *Front Genet.* 2019 Sep 25;10:889. doi: 10.3389/fgene.2019.00889. PMID: 31608118; PMCID: PMC6774278.

Goodwin, S., McPherson, J. and McCombie, W. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17, 333–351 (2016). <https://doi.org/10.1038/nrg.2016.49>

Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol.* 2021 Nov;82(11):801-811. doi: 10.1016/j.humimm.2021.02.012. Epub 2021 Mar 19. PMID: 33745759.

Keller O, Kollmar M, Stanke M, Waack S, A novel hybrid gene prediction method employing protein multiple sequence alignments, *Bioinformatics*, Volume 27, Issue 6, March 2011, Pages 757–763, <https://doi.org/10.1093/bioinformatics/btr010>

Kechin A, Boyarskikh U, Kel A, Filipenko M. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol.* 2017 Nov;24(11):1138-1143. doi: 10.1089/cmb.2017.0096. Epub 2017 Jul 17. PMID: 28715235.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Pevnikov E, Smith TPL, Pevzner PA. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods.* 2020 Nov;17(11):1103-1110. doi: 10.1038/s41592-020-00971-x. Epub 2020 Oct 5. PMID: 33020656.

Lawrence M, Gentleman R, Carey V (2009). “rtracklayer: an R package for interfacing with genome browsers.” *Bioinformatics*, 25, 1841-1842. doi: 10.1093/bioinformatics/btp328, rtracklayer: an R package for interfacing with genome browsers | *Bioinformatics* | Oxford Academic.

Lu, H., Huang, X. (2018). Current methods for genomic sequencing. *Genomics, Proteomics*

and Bioinformatics, 16(6), 1-5. <https://doi.org/10.1016/j.csbj.2019.11.002>

Mardis ER. DNA sequencing technologies: 2006-2016. Nat Protoc. 2017 Feb;12(2):213-218. doi: 10.1038/nprot.2016.182. Epub 2017 Jan 5. PMID: 28055035.

Ruan, J., and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. bioRxiv 530972. doi: 10.1101/530972

Trofimova E, Asgharzadeh Kangachar S, Weynberg KD, Willows RD, Jaschke PR. A bacterial genome assembly and annotation laboratory using a virtual machine. Biochem Mol Biol Educ. 2023 Mar 3. doi: 10.1002/bmb.21720. Epub ahead of print. PMID: 36866633.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct;37(10):1155-1162. doi: 10.1038/s41587-019-0217-9. Epub 2019 Aug 12. PMID: 31406327; PMCID: PMC6776680.