# Review of "Cryptocurrencies and Zero Mode Wave guides: An unclouded path to a more contiguous Cannabis sativa L. genome assembly."

Here (doi:10.31219/osf.io/7d968; preprint version 4.0), McKernan et al. report new genome assemblies for the flowering plant *Cannabis sativa* L., cultivar Jamaican Lion, commonly known as Cannabis or Marijuana. This sequencing and assembly project was funded by a distributed voting and funding project using the Dash cryptocurrency and a Dash "Digital Autonomous Organization" voting system.

This review evaluates the genomic and bioinformatic methods and results that are presented in the preprint, and does not intend to comment informatively on the blockchain/cryptocurrency aspects. For those most interested in the review of the biological / genomic aspects of the work, see the **"Comments on genomics aspects"** section.

Nonetheless, this preprint is an exceptional case of use of such technologies, so although the reviewer is not an expert in these areas, the reviewer cannot help but comment on it, despite their relative lack of expertise in this area. For those who are interested, see the **"Comments on blockchain/cryptocurrency aspects"**

Finally, the reviewer's involvement and review was incentivized by a monetary payment of $1000 USD total. The reviewer believes that monetary compensation and pricing signals will be a valuable component in future systems to enhance the quality and accountability of peer review, and considers this current monetarily incentivized review an experiment in this area. Naturally, the reviewer recognizes that in a preliminary & uncontrolled monetarily-incentivized peer-review system such as the one presented here, there is a conflict of interest that the monetary payment could have affected the quality of the review. The current receipt of payment and expectation of further payment upon review completion did not affect the tone, results, or quality of the review. See the "**Comments on incentivized peer review",** for more detail.

## Comments on genomics aspects:

First, assemblies reported in this manuscript (doi:10.31219/osf.io/7d968; preprint version 4.0) were downloaded from provided sources (e.g. NCBI, Mega.nz, Amazon Web Services), and reanalyzed for their assembly statistics using Seqkit (Shen et al., 2016), QUAST (Gurevich et al., 2013), and BUSCO (Simão et al., 2015). A recent preprint (Grassa et al., 2018), and peer-reviewed publication (Laverty et al., 2018) reporting genome assemblies for other Cannabis cultivars were also analyzed. At the time of writing, final genome assemblies for (Laverty et al., 2018) are still being processed on NCBI, and are not available. However, the authors shared equivalent assemblies with the reviewer (personal communication: Dr. Harm van Bakel). (Grassa et al., 2018) do report a NCBI BioProject identifier containing their assembly, but at the time of writing the assembly is not available through NCBI, so author-provided statistics are used. Summarized statistics from these analyses can be found in (Review Table 1)

| Assembly name | Source | Size (Mbp) | Contig or scaffold # | Hash[4] | Contig N50 (Mbp)[5] | Scaffold N50 (Mbp)[5] | BUSCO[6] |
|---|---|---|---|---|---|---|---|
| jlion_newasm_filt[1] | AWS | 669.7 | 35 | 494a5cd234156afe6cedebf076817790 | 4.6 | 75 | C:88.7%[S:74.6%,D:14.1%],F:1.0%,M:10.3%,n:2121 |
| Jamaican_Lion_polished_primary_181018[1] | Mega.nz | 1,074 | 558 | 6448fffd63bcee35ffdd46d917d24f43 | 3.8 | N/A | C:97.1%[S:55.2%,D:41.9%],F:0.8%,M:2.1%,n:2121 |
| GCA_003660325.1_MGC_Can.sativa_JamaicanLion_DASH_genomic[1] | NCBI | 1,032 | 2603 | 9d993e7421e32d27610fae3bebd4dd36 | 0.665 | N/A | C:95.7%[S:59.7%,D:36.0%],F:1.5%,M:2.8%,n:2121 |
| FN[2] | Authors | 1,009 | 5,304 | 04d4ac0f2e66678c150372b7918c376c | 0.370 | N/A[7] | C:94.2%[S:81.0%,D:13.2%],F:2.3%,M:3.5%,n:2121 |
| PK[2] | Authors | 891 | 12,884 | b8ff12d1897e233a80c884835818d661 | 0.133 | N/A[7] | C:92.5%[S:66.5%,D:26.0%],F:2.5%,M:5.0%,n:2121 |
| CBDRx[3] | N/A | 747.5 | 1,986 | N/A | 0.742 | N/A | C: 90.8 |
| F1 (Carmen x Skunk #1)[3] | N/A | 1,389.2 | 12,202 | N/A | 0.172 | N/A | Not available |

**Review Table 1:** Assembly statistics for recent *Cannabis* de novo genome assemblies

[1]=Calculated from provided assemblies in the reviewed manuscript by McKernan et al.

[2]=Calculated from (Laverty et al., 2018) author provided assemblies (personal communication).

[3]=Assemblies not yet available. Author-provided statistics from (Grassa et al., 2018) used.

[4]=Calculated with (seqkit seq -l $GENOME_PATH | seqkit sort -s | grep -v ">" | openssl md5)(Shen et al., 2016)
[5]=Calculated with QUAST (v5.0.1)(Gurevich et al., 2013), using parameters (--split_scaffolds)
[6]=Calculated with BUSCO (v3.0.2) (Simão et al., 2015), using the eudicotyledons_odb10 profile and default parameters
[7]=Linkage group level scaffolding provided with assembly, but not analyzed here

The assemblies reported by McKernan et al., are the most complete *Cannabis* genomes reported to date in terms of their contig N50, scaffold N50, and BUSCO conserved gene content presence statistics (%C)(Review Table 1). However, these assemblies have worrisome levels of haplotype-redundancy as measured by duplicate BUSCOs (~30+ %D), while the PK and FN genomes assemblies (Laverty et al., 2018), which have quite similar BUSCO completeness statistics with reduced BUSCO duplicate levels, also have further chromosomal linkage group level scaffolding, which was not analyzed here. It is unclear how well the scaffolding produced by McKernan et al. relates to the true chromosomes.

The most contiguous assembly from McKernan et al., both at the contig and scaffold level, is "jlion_newasm_filt", however a striking decrease in BUSCO content is observed with this assembly relative to the other Jamaican Lion cultivar assemblies presented in this manuscript. Notably, McKernan et al. report an identical BUSCO completeness statistic (94.1%) for both "jlion_newasm_filt" and "`Jamaican_Lion_polished_primary_181018`" in Table 1. The reviewer was not able to confirm this BUSCO completeness statistic for "lion_newasm_filt"; all BUSCO analyses showed a BUSCO completeness statistic less than 90% for "jlion_newasm_filt", so the "94.1%" statistic for "jlion_newasm_filt" is presumed to be either a typo or an error.

The key assemblies, and raw data (e.g. PacBio, Hi-C), were confirmed to be available by the reviewer.

## Major concerns:

**1)** Assemblies are quite contiguous and complete in terms of gene-content, however there are remaining concerns with large amounts of redundant haplotypes (e.g. Jamaican_Lion_polished_primary_181018), and/or the presumed loss of both haplotypes in the haplotype purging process (e.g. jlion_newasm_filt). Continued assembly work should be performed to optimize these assemblies.

**2)** With respect to non-bioinformatics methods, there are major methodological details missing. A link to a "daily lab journal of both successful and failed experiments" (manuscript pg. 8) is not a sufficient substitute, as (i) there is no guarantee this daily lab journal hosted on a 3rd-party site will be accessible in the future, and (ii), part of the process of manuscript writing is "compressing" such information into a form which accurately represents the key aspects while reducing the burden to readers. The transparency provided by such a "lab journal", is certainly a positive, and in a revised manuscript the reviewer would have no problem with its inclusion, but presenting it instead of well-written methods are not acceptable. Furthermore, there are several key methods missing, listed below:

**2a)** Description of growth of plants, including minimal description of permits (permit issuer, permit #) if permits for growth were necessary.

**2b)** Better description of the methods/rationale used to select the sequenced plant

**2c)** Statement of whether/how the sequenced plant/seed is available to scientists. If it is available, ideally it should be through a 3rd party mechanism. Clearly federal seed banks are out of the question, but perhaps alternative avenues could be used. If the seed/plant is a proprietary cultivar, and it is not available to scientists, this should be stated.

**2d)** Description of the "modified CTAB, Chloroform and SPRI technique" DNA extraction technique, including relevant citations.

**2e)** Description of who the sequencing provider was, and to the best degree available, the methods used for library preparation (e.g. size selection etc.)

**2f)** The instrument used to sequence the PacBio data (e.g. RSII, Sequel) was not stated. The number of SMRT cells sequenced was not stated.

**2g)** How many rounds of Arrow polishing were performed?

**2h)** **Absence of Hi-C scaffolding methods is unacceptable** SALSA is cited in Methods, but Proximo is cited in assemblies and figure legends

**3)** There are introduction and/or results information in the methods sections. Depending on the ultimate journal format, consider moving these statements to other sections.

**4)** Structure of the article is very strange. E.g. the "Data Access and Notarization" section is poorly formatted, and falls in the middle of the results disrupting the reading flow. Consider using major section headings, minor section headings, tables, etc., to better organize.

**5)** The reviewer found it very difficult to follow the assembly process (e.g. the various PacBio libraries that are at play), and key information, such as the versions of software, and the parameters used, were not included.

The authors should rewrite to ensure such information is included. Consider presenting the sequencing and assembly process as an annotated flow chart figure, or explicitly document the assembly process in a software Workflow system (e.g. NextFlow, Snakemake, https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems)

**6)** The authors should include informative summary statistics (e.g. subread length distributions, number of reads) for their input data. If there are multiple libraries from the same DNA extractions but different parameters (e.g. different size selection), or different

DNA extractions / individuals, it should be clearly stated. Tables are preferred over the current inclusion of such statistics in the methods maintext.

**7)** There is not an explicit and obvious link between the assembly naming schemes used in the paper (e.g. Table 1), and the assemblies available in the "Data Access and Notarization" section. By reading between the lines, and recalculating identifying statistics, the reviewer was able to make the connections, but this is not ideal. The filenames used for the FASTA assemblies should obviously match to the assembly IDs in the table. Also, consider providing a unique identifier to the contigs themselves. For example, if a given assembly were Jamaican Lion version 6, the assemblies could be called "JL_v6", and that string could be appended to the beginning of the contig identifier (e.g., using seqkit: seqkit replace -n -p "^" -r "JL_v6_" JL_v6.fa)

**8)** The full BUSCO string, including duplicates, and the profile which was used for evaluation, must be included. Presenting only the complete statistic (%C), without presenting the duplicates statistic (%D) and other statistics, gives the impression of cherry-picking of the results. The reviewer recommends using the eudicotyledons_odb10 profile for most reliable assessment of the assembly.

**9)** To the understanding of the reviewer, there are two main tracks in the paper, speaking to two largely separate audiences: (i) New blockchain based technologies and strategies, and how they apply to the scientific publication process, really to an in-depth level of detail, almost methods paper level of detail (e.g. manuscript Figure 1), and then (ii) genomics of *Cannabis sativa*. While indeed, the blockchain project produced the genomic assembly, and the intersection of these two topics is interesting and relevant, it is the reviewer's opinion that the two narrative tracks dilute rather than enhance the paper. Perhaps it would be better to publish a genomics oriented work, and also a work which describes the (quite interesting) voting, funding, and incentivized review parts, which could then reciprocally cite one another at relevant sections. This would also ensure review by scientific communities which are well-qualified to review.

**10)** The conflict of interest statement must explicitly list which authors have financial conflicts, not give them as "Many authors"

**11)** Although author KJM indicated to the reviewer over Twitter direct message that the co-authors on the manuscript reviewed the manuscript before submission, this should be stated explicitly somewhere in the manuscript.

**12)** What is the status of the mitochondrial genome in this assembly? Has it been accounted for / filtered out?

**13)** What is the status of the chloroplast genome in this assembly? Has it been accounted for / filtered out?

**14)** Assemblies should be screened and filtered for taxonomically unrelated (e.g. for bacterial and fungal sequences, using the blobtools pipeline or a suitable alternative).

This includes the assembly which is already uploaded to NCBI, as the authors have a responsibility to ensure assemblies which are submitted to public databases have been screened of contaminants.

**15)** No mention or direct non-bioinformatic measurement (e.g. by flow cytometry) of the expected genome size of Jamaican Lion

**16)** No mention or presented measurement of the heterozygosity rate of Jamaican lion.

**17)** No mention or direct measurement of the expected chromosome count and/or ploidy for Jamaican lion

**18)** Mega.nz or other third-party file sharing services are fine as a preliminary data storage mechanism, but raw data should ultimately be uploaded to the standard NCBI/EMBL-EBI/DDBJ SRA database, as this enforces that appropriate metadata is included. Assemblies are best served through the NCBI/EMBL-EBI/DDBJ databases as well. The reviewer did not test files stored on IPFS, but generally believes the NCBI sources are preferable given the experimental nature of IPFS and the technical barrier to retrieval of files.

**19)** Discussion of (Grassa et al) should likely fall into the discussion/conclusion section, rather than the results.

**20)** Future manuscript versions should also consider (Laverty et al., 2018) for comparisons.

**21)** Ideally future assemblies would also have repeat annotation, and gene-structure annotation.

**22)** Given the contentious status of research on Cannabis in the United States, consider an ethics compliance statement which states that research was conducted legally at all stages.

## Minor concerns:

**1)** Table 1,2 are embedded raster images, rather than a proper tables. It is essential for accessibility and machine readability to use actual tables.
**2)** Formatting error with “New England Biolabs” in affiliations
**3)** Preprint should include line-numbers, to allow for easy citation in the review process
**4) “**93.6” BUSCO completeness still cited in maintext.
**5)** Field-specific abbreviations are used without definition at the first usage (e.g. CTAB, SPRI, LTR)
**6)** Numbering of figures has typos. E.g. there are 2 Figure 4s.

## Review summary:

McKernan et al. have clearly generated valuable datasets using the cutting edge of sequencing technology to solve the problem of assembly of the Jamaican Lion cultivar, and their assemblies reach a level of contiguity that other Cannabis assemblies have not achieved. The tandem repetitive nature of the THCAS and CBDAS gene clusters is an interesting result that their data is well suited to address. That being said, the analysis of the assemblies is still at an early stage, and with the absence of most methodological details, and the unclear writing of the manuscript, it is not acceptable to this reviewer.

## Review conclusion:

Recommend revision.

## Review completed on:

2018-11-09 through 2018-11-14
Cambridge, Massachusetts, USA
Submitted to KJM in PDF format via Email.

## Review completed by:

Timothy R. Fallon, BA
PhD Candidate
Department of Biology
Massachusetts Institute of Technology
CV: https://photocyte.github.io/about/

## Comments on blockchain/cryptocurrency aspects.

The introduction to the preprint covers many interesting aspects of cryptocurrencies, including details on the Dash Distributed Autonomous Organization (DAO), used to fund this project via a voting mechanism. In the introduction, various rationales are given for the advantages of a distributed blockchain/cryptocurrency process over current approaches. However, it is the reviewer's concern that blockchains/cryptocurrencies have significant environmental impacts due to their extreme energy use relative to the computational benefit they provide (e.g. https://www.nytimes.com/2018/01/21/technology/bitcoin-mining-energy-consumption.html).

It is reassuring to read in this preprint that DASH has implemented some technological solutions to reduce energy consumption of the system, but ultimately the reviewers understanding is that (intentionally energetically expensive) cryptographic proofs are required to secure blockchain technologies, and blockchains/cryptocurrency to date have generated most of their interest as an "distributed and censorship free" currency and investment opportunity, rather than a technological solution per se. The reviewer's expectation is that uses of a blockchain which are used at any level as a representation of currency would naturally produce "developer communities" or "steering committees"

that are averse to any technically avoidable loss of value or security of their digital "property" as it were, due to weakened cryptographic security, necessitating that expensive cryptographic proofs will continue to be used, and that blockchains will be expensive relative to digital data stores to those that do not have such stringent security requirements, or will be expensive relative to existing secure stores of value that use presumably less energy intensive methods to ensure security (e.g. physical security, organizational or hierarchical security, law enforcement of rule breakers &/or seizure of stolen assets).

Traceability and immutability are certainly valuable goals in the publication of scientific works, but they can be currently implemented with existing non-blockchain technologies. The "democratic", "censorship free" or "censorship resistant" nature of the blockchain, and how censorship is presented in the manuscript introduction as being an impediment to truth, I think it an interesting point, given the nature of the scientific method. Scientific peer review, is perhaps the most pure example of undemocratic censorship. How else would one describe a rejection after peer review? But clearly, censorship, in the context of peer review, is not a negative, and in fact is the basis upon which hundreds of years of scientific progress have been laid. To use a biological analogy, freedom or chaos (e.g. mutation) is essential to explore the space of possibilities. But only evaluation, and if needed censorship (e.g. selection), can derive order and knowledge from this freedom. A scientific system designed to perfectly "democratic" or "free", either using blockchains or other technologies, would also be perfectly uninformative, as there would be no selection to determine reliable scientific works from unreliable works. In short, the reviewer is certainly intrigued by exploring alternative technologies to make the scientific method more robust, more efficient, more reliable, to enhance the goal of the search for truth, which is my goal as a scientist. In the current zeitgeist, blockchain technologies are certainly the area where new solutions are being explored, and I support that exploration, including my support through this review process. But, in the end blockchain solutions may be one of many, and selection must play a role. To supplant the current scientific system, a new system must be objectively superior in many aspects, and widely accepted across different fields of science, as well as nations. It is an impressive challenge, but experiments such as this one here will be necessary to discover alternative approaches.

## Comments on incentivized peer review.

In addition to the more cryptocurrency related explorations given above, this project also experiments in utilizing monetary payments to incentivize reviewers. This reviewer was paid $500 US dollars (equivalent based on Dash market value), at the start of the reviewing process, and is expected to receive $500 dollars at the end of the process, with the understanding that these payments will not affect the quality or tone of the review, nor are the payments contingent on the content of the review.

This understanding can be found at this link:
https://www.medicinalgenomics.com/cryptocurrency-incentivized-blockchain-recorded-peer-review-cibr/

The reviewer did not see this webpage before reviewing, and rejects the claim that they will "Verify the Cannabis genome presented is the most complete Cannabis genome that exists as of its notarized publication time stamp on (August 17th, 2018). " Given the continued revision of the project since that point, the reviewer is reviewing the manuscript as presented, rather than at the time of the time stamp.

The reviewer did not know this project nor the authors before seeing their request via Twitter through a "retweet" of their request for a monetarily incentivized scientific review by a mutually unrelated party. Further contact between the reviewer and the author KJM were through Twitter direct messages, and were related to logistical issues of the review process, including arranging payment using the DASH cryptocurrency, and repairing one(1) URL link that was not accessible to the reviewer in the review process, but were otherwise not related to the content of the review.

Even with this understanding, and transparency in the review process, naturally, there remains a conflict of interest. The reviewer recognizes that Academic / industrial partnerships with monetary conflicts, have led to a corruption of the research process in several well-publicized cases. The reviewer states that: An honest and uncorrupted scientific review process is essential to science, and is amongst their most closest held beliefs. They have not allowed the monetary compensation to affect the quality or tone of their review in any way. The reviewer believes the amount of time spent on this review is higher than a typical review they would perform.

## References:


Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, Schwartz CJ, Weiblen GD. 2018. A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. *bioRxiv* 458083. doi:10.1101/458083 (Preprint)

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075. doi:10.1093/bioinformatics/btt086

Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, Deikus G, Sebra R, Hughes TR, Page JE, Bakel H van. 2018. A physical and genetic map of Cannabis sativa identifies extensive rearrangement at the THC/CBD acid synthase locus. *Genome Res* gr.242594.118. doi:10.1101/gr.242594.118

McKernan K, Helbert Y, Kane LT, Ebling H, Zhang L, Liu B, Eaton Z, Sun L, Dimalanta ET, Kingan S, Baybayan P, Press M, Barbazuk W, Harkins T. 2018. Cryptocurrencies and Zero Mode Wave guides: An unclouded path to a more contiguous Cannabis sativa L. genome assembly. doi:10.31219/osf.io/7d968 (Preprint)

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**:e0163962. doi:10.1371/journal.pone.0163962

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212. doi:10.1093/bioinformatics/btv351