We thank the reviewers for their valuable suggestions. Their input has made the paper much stronger and for this we are grateful. We apologize for somewhat non-standard formatting of references; this is due to the technical complexities of MSWord and will be fixed in the final version of the document.

Herein we have reproduced the reviewer's comments with our responses printed **in bold**.

Sincerely,

The Assemblathon authors.

---Reviewer 1----

Assemblathon 1 Review

The authors describe an ambitious and overdue effort to compare the ever-growing list of genome assemblers available to the sequencing community. The paper is a good general introduction and survey of the various assembly strategies, is generally well-written and simultaneously tackles the challenge of analyzing genome assembly methods and developing novel methods for assessing assembly accuracy.
The latter is something that has been done poorly in many different individual genome assembly papers and the authors make novel contributions to the process of measuring accuracy.

The only significant flaw is the simulation of the error profiles of the reads which is reversed, leading to uncharacteristic and unnaturally high error rates at the start of the simulated data. I think I believe their assertion that their mistake does not impact the different assemblers, but it would have been far more convincing had they stated explicitly that all of the assembler authors concurred.

Thank you for the suggestion. We polled the participants regarding the read error mistake issue and report the result in the text as follows:

'*…we surveyed participants on this matter and only one group, L'IRISA, indicated that their methodology was possibly harmed more than other methods due to the mistake*'

Furthermore, it would greatly improve the manuscript if this detail was mentioned more prominently and not in the fourth paragraph of section 5.2.5.

We agree, we have made this more prominent, moving the discussion of this error to the start of the results section.

Technical comments

The authors make mostly pragmatic decisions regarding genomic evaluations which is hard work. I was a little disappointed to see the test set scaled back to a rather small total assembly size. One of the challenges with assembly is the scale and given that gigabase sized genomes are a common challenge, it would be helpful to get a sense at least of which assemblers are capable of dealing with realistic data sets. No matter how accurate the assembler, it is not very useful if it can't assemble.

Thanks, this is a limitation of the current Assemblathon, which was made primarily to allow as many smaller groups compete as possible, regardless of wether they had access to the appropriate hardware to handle a mammalian genome. We have modified Section 3.1 to include the sentence (when discussing the simulation):

'...we were able to precisely tailor the proportions of the simulated genome to those desired for this experimental analysis, i.e. to limit the size of the genome to less than that of a full mammalian genome and thus allow the maximum number of participants, while still maintaining a size that posed a reasonable challenge..'

We have also modified the discussion of this limitation to suggest that in the second competition (Assemblathon 2, now under way), a larger, gigabase genome should be included to test just this:

'..., it should feature at least one mammalian genome scale data set, to test the scaling of the assembly pipelines...'

It was unclear why the E. coli contamination at 5% was included. It

could benefit from some explanation. That is a relatively high rate for any genome assembly and given the premise of a completely next generation data set, there should be no bacterial termination from traditional cloning sources. Furthermore, it obviously wasn't explicit to the participants what to expect judging from the heterogeneous way in which they treated the E. coli. The manuscript would benefit from stating explicitly what the contestants were told about the data set.

Thanks, we added the following into the simulated reads section:

'Bacterial sequence was included as an attempt model the sort of contamination occasionally present in data from sequencing centres, though the 5% level was fairly arbitrary. Participants in the contest were notified that some bacterial contamination was present in the data, though they were not told about its precise nature or explicitly told to remove it.'

I was struck in section 3.4.8 by the very low success rate capturing the entire transcript and the authors attribute this to intron polymorphism. This suggests that the assemblers are actually doing very poorly in repetitive regions, at least over gene scales. There isn't much treatment of repeat accuracy other than copy number and this deserves more attention in the manuscript.

Thanks, we have completely reworked the the section of gene analysis, expanding it to include a simple analysis of repeat assembly within our MSAs as well as also looking at conserved non-coding elements. Please see Section 3.4.8 for more details.

(Notably, in doing this we refined our predictions on the accuracy of genic reconstruction (around 10% of genes are perfectly assembled in the best case, not 2.5% as reported from earlier work with blast)).

The authors do mention a number of limitations of the work and overall the manuscript is balanced and reflective. I would have enjoyed seeing a brief recommendation at the end for the next assemblathon. The authors beg this question with the provocatively numbered title and I feel the reader should at least get a preview of the next installment.

We appreciate this suggestion and have added/modified the final part of

the discussion as follows:

'There are a number of important limitations with the current work. Many of these limitations relate to the fact that we assessed only one, simulated dataset, which was about 30x smaller than the human genome. To address these issues a second competitive evaluation, Assemblathon 2, is now underway….

..., we conclude by making three distinguishing suggestions for Assemblathon 2 that would sufficiently expand its scope from this initial competition. Firstly, it should feature at least one mammalian genome scale data set, to test the scaling of the assembly pipelines. Secondly, it should feature real data, to compare with the simulation results presented in this competition; this may necessitate the use of a different set of evaluation metrics, where the "correct" answer is unknown. Thirdly, it should be expanded to include other sequencing technologies, so that a better comparative, unbiased understanding of available sequencing technologies can be made.'

The error assembly section was somewhat weak. The authors did not do a thorough review of the field and there are many different read simulators, which provide a number of the features that they complained did not exist. Given that the overall premise of the paper was the benefit of using simulation (with which this reviewer strongly agrees), investment in the accuracy of the simulation is paramount. The authors went to considerable trouble to simulate issues associated with circularized mate pair generation but then decided not to model chimeric reads which are a considerable headache for many assemblers. They note that their strategy produces effectively models of uncorrelated errors and overlooked a common strategy for simulating more realistic error profiles, namely using the quality profiles of individual real reads. This area could be considerably improved should this exercise be repeated. Given the downstream investment in genome generation and analysis, this is an obvious area to invest in.

We have included a more thorough review of existing Illumina read simulation software, along with details of an additional program that we considered using but originally omitted from the text. Additionally, we

apologize that that section of the text left Reviewer 1 with the impression that our read simulator did not model chimeric reads when, in fact, it does. We have changed the text to try to prevent such misimpressions.

Minor remarks

I'm not 100% sure but I thought that phrap predated the E myers 1995 et al. and was inspired more by the Staden package rather than being derivative.

Thank you.

“chiefly because it allowed us design evaluations “ -> “chiefly because it allowed us to design evaluations”

Thank you.

this is a very minor criticism, but the last two paragraphs of
section 2 were very awkwardly constructed and stood out relative in an overall well-written manuscript. They could benefit from a slight massage.

Thank you.

“types of errors simulated and the simulators limitations” - > “types of errors simulated and the simulator’s limitations”

Thank you.

The manuscript refers to supplementary figures 7 and 8, but figure 7 is actually included in the manuscript

The supplement's figure and table numbers begin at 1, even though the supplement's section number starts at 7.

The acknowledgments section has a [please insert!] comment, which I am sure your respective funding agencies would be more than happy to fill.

;)

---Reviewer 2---

The submitted paper by Earl et al. describes the process and the results of the Assemblathon competition, where various groups submitted assemblies of simulated reads from an artificial, unknown, genome. These submissions were then evaluated using various metrics. The paper mainly introduces some novel evaluation metrics and applies them to the submitted assemblies. I think the competition is itself is a valuable activity for the community, and that its results are of great interest to the community. However, I think the presentation of the paper should be improved.

The introduction is too long for what it brings, and it does not clearly state the contribution of the paper. For example, it ends (the last three paragraphs) by justifying various aspects of the competition, which are not at all discussed in the results section (use of simulation and Illumina data). On the other hand, the novel idea of using MSA during evaluation to allow for mixing of haplotypes is only briefly mentioned.

Thanks, the original introduction has been completely rewritten to make it both more didactic for the purposes of the paper's exposition. In particular:
(1) we have strived to improve its clarity by reducing the use of jargon and defining and/or referencing all the ideas, concepts and programs introduced.
(2) We have also reduced the amount of detail regarding individual assembly programs, which others felt was excessive.
(3) Strived to give the reader a clearer picture of assembly assessment and relevant information about sequencing technology.
(4) Rewritten the summary paragraph describing the work in the paper, it now better describes the achievements of the paper.

The results in Tables 3-5 and other figures can be used to evaluate the various assembly strategies described in Table 2 and in the supplement. However, I felt like this analysis was not included in the results or discussion. Which strategies worked the best? Did some strategies dominate over the others? Were there trade-offs involved, e.g. some strategy was better in one metric and worse in the other? From the perspective of a bio-informatician who wants to assemble a genome for some particular purpose, which strategy works best? For example, if only a “bag of genes” is desired, then the long range contiguity metric or the scaffold metrics would be less important then the Genic analysis, while if the goal is to finish the genome, the scaffold metrics would be much more important.

We present in Table 2 a summary of the programs employed by the participants and in Supplemental Section 8.2 self-reported details about the computational details involved in each strategy. However, as you note, we make no effort to synthesize this information with the results shown in Table 3. We debated this issue internally both within the core writing group and the extended participants list and our feeling is that any simple attempt to synthesize the results with the strategies employed would be potentially misleading, and is otherwise beyond the scope of this work. We added the following to the introduction to better explain our position:

'In general, most sequence assembly programs are multi stage pipelines, dealing with correcting measurement errors within the reads, constructing contigs, repeat resolution (i.e. disambiguating false positive alignments between reads) and scaffolding in separate phases. Since a number of solutions are available for each task, several projects have been initiated to explore the parameter space of the assembly problem, in particular in the context of short read sequencing ((Zhang, Chen, Yang, Tang, Shang, & Shen, 2011) (Phillippy, Schatz, & Pop, 2008) (Hubisz, Lin, Kellis, & Siepel, 2011) (Alkan, Sajjadian, & Eichler, 2011)). In this work we are concerned with evaluating assembly programs as whole, with the aim of comprehensively evaluating different aspects of assemblies.'

The reason for this is that the strategies employed by the participants are quite complex and can include not just a particular assembly software but different data pre-processing steps and filters; in general there is an enormous amount of variation present in the approaches taken.

Additionally, not all assembly software could be unanimously and unambiguously classified into algorithmic or data structure groups. Even were such a classification available, we don't think there is enough data to reliably pick out the signal of a particular assembler class (greedy, seed and extend, branch and bound, overlap layout consensus, sequencing by hybridization graph, deBruijn graph, string graph, etc etc) out of the surrounding noise of preprocessing, parameter, implementation and post-processing variation.

We agree that this is a large limitation of our study but we assert that this

illustrative of the scope and complexity of the problem of assessing assembly techniques.

I realize that some remarks with regards to the above questions are in fact in the paper but interspersed between different sections. It may help to separate the definition of the various metrics from the results of evaluating the assemblies with them.

The authors discuss various metrics: N50, NG50, etc. I think it would be helpful to include a table summarizing the various metrics and what quality of the assembly they reflect.

Thank you for this helpful suggestion. We have added a table (Table 7) to centralize and describe the various metrics used in the paper. We hope this will help the casual reader better navigate the paper.

Minor comments:

Intro, second paragraph, first sentence. The citation for string graphs should be Myers05, not Myers95.

Thanks, this is fixed.

Also, from what I recall, Arachne did not use overlap or string graphs.

We checked this and, though you are correct, they don't explicitly describe use overlap graphs, they implicitly use the framework (from the abstract):

'ARACHNE starts by detecting and aligning pairs of apparently overlapping reads, referred to here simply as overlaps. Some of these are false overlaps resulting from repeated sequences in the genome and will be eliminated in subsequent steps.'

We thus modified the paragraph as follows:

'As the field of sequencing has changed so has the field of sequence assembly, for a recent review see Miller, Koren & Sutton (2010). In brief, using Sanger sequencing, contigs were initially built using overlap or string graphs (Myers E. , 2005) (or data structures closely related to them), in tools such as Phrap (http://

www.phrap.org/), GigAssembler (Kent & Haussler, 2001), Celera (Myers, et al., 2000) (Venter, et al., 2001), ARACHNE (Batzoglou, et al., 2002), and Phusion (Mullikin & Ning, 2003), which were used for numerous high quality assemblies such as human (Lander, et al., 2001) and mouse (Mouse Genome Sequencing Consortium, et al., 2002). However, these programs were not generally efficient enough to handle the volume of sequences produced by the next generation sequencing technologies, spurring the development of a new generation of assembly software.'

Also, what do you mean by “tighter programming?” This is not clear.

Thank you, we have removed this statement and revised the paragraph as it was unclear.

Some of the citations give two authors followed by et al (e.g. Butler, Maccallum, et al), while it should be just one.

Thank you.

Intro, 5th paragraph: I think SBH was a much older technologies used in the 90s, not the modern technology used by SOLID.

Thank you, in revising the document we have removed this.

intro, 6th paragraph, “The predominate method...” should be “The predominant method”

Thank you.

section 3.4.4, second paragraph, “bidirected bi-graph” should be “bidirected graph.”

Oops, thank you.

Also, the term “adjacency graph” was previously used in this context in Bergeron, Mixtacki, and Stoye in “A Unifying View of Genome Rearrangements,” and the authors should cite this paper. I think that their graph is the same (modulo the handling of telomeres) as the adjacency graph here if the block

edges are collapsed and only two genomes are present.

Thanks, we have added this reference to the paper.

section 6, there is a phrase that says “[please insert]”

Thank you.

bibliography, the citations need to be fixed to follow the necessary format.

Thank you.

--Reviewer 3---

In this work, the authors describe a large-scale project to assess the performance of a variety of genome assemblers. While I think this is a worthy goal and the authors are to be commended for taking on this task. It is clear that the area of assembly assessment needs more rigor and the authors have tried to address this. However, I think the authors fell short in some areas of assessment and are a bit overly enthusiastic about the performance of the current field of genome assemblers. Additionally, the group failed to address issues related to how sequencing technology affects assembly performance. It is clear that there are algorithmic limitations to assembling short reads and that robust assemblies will only be achieved with significant technology improvements as well.

Major criticisms of the experimental approach:

- the use of simulated data: while I understand the allure of knowing the 'right' answer, it is incredibly difficult to robustly simulate biological data (both at the genome level as well as the read level). While it is fair to argue that the current human and mouse assemblies do not necessarily represent the 'right' answer, these assemblies are the highest quality mammalian genome assemblies that are available. Using these assemblies to represent 'truth' might produce a more realistic assessment of the assemblies (Gnerre et al., 2010: http://www.ncbi.nlm.nih.gov/pubmed/21187386). It is unclear to me how performance in these sets of tests will translate to real data.

Thanks. We modified section 3.1 (moving some text previously from the

introduction) to better assess our motivations for using a simulated genome:

'Rather than use an existing reference genome for assessment, we opted to simulate a novel genome. We did this primarily for three reasons. Firstly, it gave us a genome that had no reasonable homology to anything other than out-group genomes that we generated and subsequently provided. This allowed for a fair, blind test in which none of the assembly contributors had access to the underlying genomes during the competition. Secondly, we were able to precisely tailor the proportions of the simulated genome to those desired for this experimental analysis, i.e. to limit the size of the genome to less than that of a full mammalian genome and thus allow the maximum number of participants, while still maintaining a size that posed a reasonable challenge. Thirdly, as mentioned, we could simulate a diploid genome; we know of no existing diploid dataset in which the contributions of the two haplotypes are precisely and fully known.'

We also mention the limitations of using simulation in the discussion as follows:

'There are a number of important limitations with the current work. Many of these limitations relate to the fact that we assessed only one, simulated dataset, which was about 30x smaller than the human genome. To address these issues a second competitive evaluation, Assemblathon 2, is now underway. …. we conclude by making three distinguishing suggestions for Assemblathon 2 that would sufficiently expand its scope from this initial competition. Firstly, it should feature at least one mammalian genome scale data set, to test the scaling of the assembly pipelines. Secondly, it should feature real data, to compare with the simulation results presented in this competition; this may necessitate the use of a different set of evaluation metrics, where the "correct" answer is unknown….'

- While the authors spent a great deal of time discussing contig and scaffold metrics, there was no mention of the assemblers ability to reconstruct

chromosome sequences. If an assembly is to be a representation of an organism's genome, then understanding chromosome structure is key to this. Generating a bucket of unplaced scaffolds will not be as useful for biologists as being able to have a set of chromosome representations.

Thanks, this is a difficult issue. We assessed the output, in the form of contigs and scaffolds, provided by the assemblers; this is currently the standard.

Theoretically, if the assemblers had been able to scaffold the contigs completely then we would have expected to receive files containing only three scaffolds, one for each chromosome, though clearly this did not happen. On one level, therefore, we can conclude that the assemblers failed to properly construct the complete chromosomal structure.

However, we do manage to show, by going to extended lengths to assess the accuracy of the scaffolds, and provide coverage plots (see Figure 4 and numerous figures in the supplement) that the scaffolds correctly construct large regions of the chromosomes. It is unclear beyond this and the work on large scale contiguity (see Figure 5), given the information provided to the assemblers and the format of the output, what more we could do.

It is also important to note that Evolver does not model centromeres or telomeres and we worry that addressing the question of regional (peri-centromeric, telomeric, etc.) chromosome reconstruction directly might overstep the capabilities of our underlying simulation.

Additional issues:
- I would like to see a more detailed description of the differences between the a1 and a2 haplotypes. Perhaps a VCF file with the differences would be a good start, but understanding just how different the simulated chromosomes are would be useful. The authors allude to the fact that their simulated data likely under-represents recent duplications, which is unfortunate as we know these are the problem areas for all assemblers. Are the difference simple indels (even if they are big)? Are there varying regions of SNP diversity? Are there inversions?

The number of inversions present between the a1 genome and its MRCA with a2 (and vice versa) is shown in Table1. We have added to this table, describing in further detail the differences between the a1 and a2 haplotypes. In the Supplement we have included a dot-plot showing the

pairwise alignment of the a1 and a2 haplotypes. We have posted on the paper's website the pairwise multiple alignment format (MAF) file between the a1 and a2 haplotypes, as extracted by evolver. (http:// compbio.soe.ucsc.edu/assemblathon1/) and noted this in the methods

- The authors use an awful lot of unreferenced sources. In some cases (Phred, LastZ) the authors could at least point to a web site from which a description or source code could be obtained. Where this is not available the authors should provide more details about how the software works.

The websites were previously listed in the bibliography, which is not Genome Research format. We have corrected this and placed the citations to websites in-line.

- The flow of the paper was a bit odd as the sequencing technology bit came after the assembly bit. It has typically been the change in sequencing technology that drives the changes in assembly technology and things might have flowed a bit better with some re-arrangement.

We have completely rewritten the introduction to try to achieve a more coherent flow and didactic presentation of relevant background, including placing the discussion of sequencing technology before the discussion of assembly methods.

- Page 3: You should probably also provide the reference for the Celera human assembly, which was pretty important with respect to assembly development.

Thanks, we have included this.

- Page 4: You comment that short read length is compensated for by long distance mate pairs. This is a bit of an over statement. The 40Kb di-tags have only just come into more common practice and were not even part of your simulated data. 10Kb is relatively short for mammalian genomes and typically insufficient to get across complex regions. Additionally, even the 40Kb links fall short of the 150 - 200Kb links afforded by BACs.

Thanks, we have removed this statement.

- End of page 4 and beginning of page 5: The authors omit some methods that have been used to assess new assemblies- specifically the alignment of the

assembly to sequences not used to generate the assembly in the first place- these can include genomic sequences (clones or paired ends) or transcripts. The draft human and mouse assembly assessments used these methods as did the 'finished' mouse paper. Both the Church et al 2009 and Lindblad-Toh,et al 2005 paper provided many additional metrics other than just statistics to assess the validity of the assemblies.

Thank you. We have expanded this section to be a more comprehensive review of assembly assessment, while being more didactic and less jargon filled.

- Genic assessments: Just looking at coverage is not very sophisticated, nor is it sufficient for assessing the quality of an assembly with respect to gene annotation. One of the primary uses of most genome assemblies is to produce gene annotation (folks even tried to do this with the 2x genomes!). Coverage is only part of this- the authors should have checked both for continuity (that is, are the exons in the right order and how many are split by scaffold gaps) and ability to produce a protein product. One of the biggest difficulties in annotating a WGS assembly are base level errors that produce a nonsense or frameshift and make a protein-coding gene look inactive.

Thanks, we have completely redone this section of analysis, as we agree with you that it was the weak point of the paper. We have incorporated your suggestion and also included simple analysis of repeats and other annotations. Please see Section 3.4.8.

- The authors are a bit optimistic that assemblies using NGS data are approaching the same level of quality as capillary sequences for complex mammalian genomes. Gnerre at al made this assertion in their paper, and while the methods (both algorithmic and technological) described in that paper are significant improvements the assembly representation (both N50s and chromosome structure) are still significantly worse than the HuRef (Venter) assembly. Additionally, even if NGS assemblies really do get to the level of capillary WGS assemblies, there are still significant shortcomings with respect to the data. If the goal is to model an organism's genome, then phased haplotypes and representation of repetitive and structural divergent regions have got to be addressed.

Thank you. We discussed the Gnerre 2011 and MacCallum 2009 papers in the discussion. On the issue of phasing, supplementary section 7.2

investigates the question of phasing in the assemblies. Whilst the assemblies are not phased we stand by the concordance between their results and ours.