A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding

Salla Vartia, José L. Villanueva-Cañas, John Finarelli, Edward D. Farrell, Patrick C. Collins, Graham M. Hughes, Jeanette E. L. Carlsson, David T. Gauthier, Philip McGinnity, Thomas F. Cross, Richard D. FitzGerald, Luca Mirimin, Fiona Crispie, Paul D. Cotter and Jens Carlsson

Article citation details
R. Soc. open sci. 2: 150565.
http://dx.doi.org/10.1098/rsos.150565

Review timeline
Original submission: 2 July 2015
1st revised submission: 21 October 2015
2nd revised submission: 9 December 2015
Final acceptance: 10 December 2015

Note: This manuscript was transferred from another Royal Society journal without peer review.
Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept as is

Comments to the Author(s)
None

Review form: Reviewer 2

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
No

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Supplementary material are adequate and clear. However, raw sequencing data does not seem to have been made available by the authors: you the resulting genotypic data are provided as a supplementary material. Raw sequencing data have to be submitted in a publicly accessible database such as NCBI Short Read Archive or its European counterpart.

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
Yes

Recommendation?
Major revision is needed (please make suggestions in comments)

Comments to the Author(s)
The manuscript submitted by Salla Vartia and collaborators for consideration of publication in Royal Society Open Science reports on a new method to genotype microsatellite markers, a widely used kind of genetic marker in various fields including applied biology, ecology and evolutionary biology, using high throughput sequencing. The authors compared their new approach with traditionally used capillary sequencer based genotyping method to analyze more than 50 microsatellite markers in Atlantic cod (Gadus morhua L.). They also tested the relevance of using already developed set of multiplexed markers that they compared to a way to handle markers that have not previously be optimized for multiplexing. Finally, they provided detailed protocol allowing to multiplex 96 individuals in a single sequencer lane using a dual barcode approach which seem to be a cost effective method to make the most of the high throughput capacity of the Roche 454 sequencing platform they used.

The manuscript is generally well written and I greatly appreciated the introduction because the authors presented clearly the challenge associated with electrophoresis-based microsatellite genotyping and the need for more universal, reliable and transferable genotyping methods. The usual technical reply to improve genotyping workflow is to switch to SNP markers, yet SNP are they own limitations that the authors developed convincingly. Therefore, combining microsatellite strengths to absolute and high throughput genotyping allowed by modern
sequencing platform would be a major innovative step forward and this makes the manuscript very exciting to read.

However, I found that the manuscript incompletely dealt with some of the challenges associated with microsatellite genotyping by sequencing. From my point of view, microsatellite sequencing with high throughput sequencing platform has two main difficulties that had hampered its development: (1) the need to combine enough markers and individuals in a single sequencing lane to take advantage of the high sequencing output, and (2) the accurate genotyping of very unstable short tandem repeats structure of microsatellites which make the signal treatment from the reads to the number of repeat very challenging due to slippage mutation especially when amplification based library is used to generate the library. Although the authors partially dealt with the first point (which still would need some more discussion), and second point is barely detailed or explained which I believe make the developed method quite difficult to reproduce. I am going to develop more on these two main points below.

(1) Markers and individuals multiplexing level
The authors used a somewhat outdated sequencing platform (454 GS-FLX Titanium) which is not an issue by itself because their results are still totally relevant by themselves. The only downside is for the reproducibility of the method they developed. As they emphasized page 12 (line 264-277) modern sequencing platform produced shorter reads (which may not be a too big issue) at a much higher throughput (10 to 30 times more reads). I think that the authors should spend more time to discuss how their method could be leverage on more modern platform to show readers how it can be relevant to nowadays technologies. For instance, why sequencing 10 to 30 times more microsatellites for 96 individuals, what could be the application behind such analysis? I think a more useful way would be to open up new perspectives by multiplexing 10 to 30 more individuals (a few thousands) at a moderate number of microsatellite loci (20 - 50) because this kind of data is very common in population genetics applied to evolutionary biology or ecology. This high level of individual multiplexing would require a slight modification of the proposed dual indexing protocol, for instance using the approach of Campbell et al. (2015) consisting of using one barcode to index each plate well and the second barcode to index each plate. This would use 96 different barcodes for well indexing plus one barcode for each plate. The slightly highest cost for the barcodes would quickly be recovered given the thousands of individuals that could be multiplexed in a single sequencing run. I think that providing more perspective about how your method could be adapted to suite more recent sequencing platform on the market today and an updated comparison of the advantages of your method with alternative microsatellite sequencing method already developed (e.g., Carlson et al. 2015) would increase the interest and generality of your manuscript.

(2) Data analysis reproducibility and genotypic data limitations
While the main aim of a methodological paper is to be fully reproducible, I was very surprised to only found 9 lines (page 8, lies 156-164) describing the bioinformatics data analysis. With so few details about how the authors used the few hundred of thousand of reads output by the sequencer to generate the genotypic data, I cannot figure out how to redo the analysis myself which is a shame because all other aspects of the method are otherwise fully described. The authors have to provided additional details and justification about the bioinformatics analyses they used. How was the number of errors allowed for the barcode and primer sequence determined and how does this affect the results? More importantly how do they handled allele sequence instability (due for instance to slippage and partial adenylation) and could identify alleles from noise especially when two alleles had similar number of repeats? Frankly, just saying that a commercial software with the default parameters was used to verify read alignment is far from enough for a standard reporting of a bioinformatics analysis especially when the method had to be reproducible. In addition, it is a very interesting point that sequencing microsatellite allow for the detection of homoplasy, but again more analytical details are needed for the reader to understand how the analysis was performed, in particular, how was it possible to differentiate real mutation from sequencing errors given the relatively low coverage of 10X chosen for the analysis (page 8, line 162-164)?

Finally, a good genotyping method should match the requirement need for the application that will be made with the genotypic dataset. However, high rate of missing genotypic data is a notorious issue when using high throughput sequencing platforms, which with SNP is usually
compensated by the very high number of markers studied. However, the described microsatellite sequencing method seems to generate dataset with a high rate of missing genotype. With data for 3325 genotypes out of 5088 (page 10, line 200) amounting to 45% of missing data and 10 loci out of 53 with less than 50% of the individual genotyped (lines 201-203), I was wondering what analytical power could be expected from such a dataset. Would linkage mapping, parentage analysis or genetic assignment still be possible despite such a high missing data rate? I think you should discuss this explicitly and propose improvement of the protocol that could help to lower missing data in the final dataset.

Minor comments
Page 6, line 85: I think that the field of medical research are also interested in sequencing microsatellite because short tandem repeat are found to be associated with diseases, so that not only forensic science is involved in sequencing microsatellites. In addition, with the realization that short repeats main may have important functions and be linked to phenotypic variability, there would be a increasing interest in sequencing microsatellites in near future which could make you study even more interesting to readers (for instance, see the very good review of Press et al. 2014).
Page 11, lines 226: I am absolutely not convince by the argument that the proposed new method is "cost-effective" or "cheaper than current approaches while offering better and more data" (page 13, line 293). I will not come back on data quality here (although beyond homoplasy, missing data rate is very low when using capillary sequencer), but I do not think that the described method is cheaper to genotype 96 individuals at 50 microsatellite than a traditionally used capillary-based genotyping method when it is well optimized with high multiplexing level (Guichoux et al. 2011). While it can be difficult to estimate laboratory procedure cost because price became quickly outdated as technology improved, I still think that you should give some justification about why you think your method is cheaper (besides labor costs that is not really relevant in a publicly founded research institute context).
Page 12, line 276: I did not understood here if "short amplicons (less than 300bp)" included primers and adaptors or just the targeted microsatellite loci (repeats and flanking regions). Please specify.

References:

Decision letter (RSOS-150312)

24-Sep-2015

Dear Mr Vartia:

Manuscript ID RSOS-150312 entitled "A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding" which you submitted to Royal Society Open Science, has been reviewed. The comments from reviewers are included at the bottom of this letter.
In view of the criticisms of the reviewers, the manuscript has been rejected in its current form. However, a new manuscript may be submitted which takes into consideration these comments.

Please note that resubmitting your manuscript does not guarantee eventual acceptance, and that your resubmission will be subject to peer review before a decision is made.

You will be unable to make your revisions on the originally submitted version of your manuscript. Instead, revise your manuscript and upload the files via your author centre.

Once you have revised your manuscript, go to https://mc.manuscriptcentral.com/rsos and login to your Author Center. Click on "Manuscripts with Decisions," and then click on "Create a Resubmission" located next to the manuscript number. Then, follow the steps for resubmitting your manuscript.

Your resubmitted manuscript should be submitted by 23-Mar-2016. If you are unable to submit by this date please contact the Editorial Office.

We look forward to receiving your resubmission.

Sincerely,
Emilie Aime
Senior Publishing Editor, Royal Society Open Science

Associate Editor Comments to Author (Dr Ivo Gut):
Associate Editor: 1
Comments to the Author:
Dear Authors,

two reviewers have read your manuscript. One of the reviewers made a number of interesting suggestions that would help improve the manuscript. We would appreciate if you could respond to issues raised by this reviewer by modifying the manuscript and with a point-by-point response.

With best regards,
Ivo Gut

Reviewers' Comments to Author:
Reviewer: 1

Comments to the Author(s)
The manuscript submitted by Salla Vartia and collaborators for consideration of publication in Royal Society Open Science reports on a new method to genotype microsatellite markers, a widely used kind of genetic marker in various fields including applied biology, ecology and evolutionary biology, using high throughput sequencing. The authors compared their new approach with traditionally used capillary sequencer based genotyping method to analyze more than 50 microsatellite markers in Atlantic cod (Gadus morhua L.). They also tested the relevance of using already developed set of multiplexed markers that they compared to a way to handle markers that have not previously be optimized for multiplexing. Finally, they provided detailed protocol allowing to multiplex 96 individuals in a single sequencer lane using a dual barcode approach which seem to be a cost effective method to make the most of the high throughput capacity of the Roche 454 sequencing platform they used.
The manuscript is generally well written and I greatly appreciated the introduction because the authors presented clearly the challenge associated with electrophoresis-based microsatellite genotyping and the need for more universal, reliable and transferable genotyping methods. The
usual technical reply to improve genotyping workflow is to switch to SNP markers, yet SNP are they own limitations that the authors developed convincingly. Therefore, combining microsatellite strengths to absolute and high throughput genotyping allowed by modern sequencing platform would be a major innovative step forward and this makes the manuscript very exciting to read.

However, I found that the manuscript incompletely dealt with some of the challenges associated with microsatellite genotyping by sequencing. From my point of view, microsatellite sequencing with high throughput sequencing platform has two mains difficulties that had hampered its development: (1) the need to combine enough markers and individuals in a single sequencing lane to take advantage of the high sequencing output, and (2) the accurate genotyping of very unstable short tandem repeats structure of microsatellites which make the signal treatment from the reads to the number of repeat very challenging due to slippage mutation especially when amplification based library is used to generate the library. Although the authors partially dealt with the first point (which still would need some more discussion), and second point is barely detailed or explained which I believe make the developed method quite difficult to reproduce. I am going to develop more on these two main points below.

(1) Markers and individuals multiplexing level
The authors used a somewhat outdated sequencing platform (454 GS-FLX Titanium) which is not an issue by itself because their results are still totally relevant by themselves. The only downside is for the reproducibility of the method they developed. As they emphasized page 12 (line 264-277) modern sequencing platform produced shorter reads (which may not be a too big issue) at a much higher throughput (10 to 30 times more reads). I think that the authors should spend more time to discuss how their method could be leverage on more modern platform to show readers how it can be relevant to nowadays technologies. For instance, why sequencing 10 to 30 times more microsatellites for 96 individuals, what could be the application behind such analysis? I think a more useful way would be to open up new perspectives by multiplexing 10 to 30 more individuals (a few thousands) at a moderate number of microsatellite loci (20 - 50) because this kind of data is very common in population genetics applied to evolutionary biology or ecology. This high level of individual multiplexing would require a slight modification of the proposed dual indexing protocol, for instance using the approach of Campbell et al. (2015) consisting of using one barcode to index each plate well and the second barcode to index each plate. This would use 96 different barcodes for well indexing plus one barcode for each plate. The slightly highest cost for the barcodes would quickly be recovered given the thousands of individuals that could be multiplexed in a single sequencing run. I think that providing more perspective about how your method could be adapted to suite more recent sequencing platform on the market today and an updated comparison of the advantages of your method with alternative microsatellite sequencing method already developed (e.g., Carlson et al. 2015) would increase the interest and generality of your manuscript.

(2) Data analysis reproducibility and genotypic data limitations
While the main aim of a methodological paper is to be fully reproducible, I was very surprised to only found 9 lines (page 8, lies 156-164) describing the bioinformatics data analysis. With so few details about how the authors used the few hundred of thousand of reads output by the sequencer to generate the genotypic data, I cannot figure out how to redo the analysis myself which is a shame because all other aspects of the method are otherwise fully described. The authors have to provided additional details and justification about the bioinformatics analyses they used. How was the number of errors allowed for the barcode and primer sequence determined and how does this affect the results? More importantly how do they handled allele sequence instability (due for instance to slippage and partial adenylation) and could identify alleles from noise especially when two alleles had similar number of repeats? Frankly, just saying that a commercial software with the default parameters was used to verify read alignment is far from enough for a standard reporting of a bioinformatics analysis especially when the method had to be reproducible. In addition, it is a very interesting point that sequencing microsatellite allow for the detection of homoplasy, but again more analytical details are needed for the reader to understand how the analysis was performed, in particular, how was it possible to differentiate real mutation from sequencing errors given the relatively low coverage of 10X chosen for the analysis (page 8, line 162-164)?
Finally, a good genotyping method should match the requirement need for the application that will be made with the genotypic dataset. However, high rate of missing genotypic data is a notorious issue when using high throughput sequencing platforms, which with SNP is usually compensated by the very high number of markers studied. However, the described microsatellite sequencing method seems to generate dataset with a high rate of missing genotype. With data for 3325 genotypes out of 5088 (page 10, line 200) amounting to 45% of missing data and 10 loci out of 53 with less than 50% of the individual genotyped (lines 201-203), I was wondering what analytical power could be expected from such a dataset. Would linkage mapping, parentage analysis or genetic assignment still be possible despite such a high missing data rate? I think you should discuss this explicitly and propose improvement of the protocol that could help to lower missing data in the final dataset.

Minor comments

Page 6, line 85: I think that the field of medical research are also interested in sequencing microsatellite because short tandem repeat are found to be associated with diseases, so that not only forensic science is involved in sequencing microsatellites. In addition, with the realization that short repeats main may have important functions and be linked to phenotypic variability, there would be a increasing interest in sequencing microsatellites in near future which could make you study even more interesting to readers (for instance, see the very good review of Press et al. 2014).

Page 11, lines 226: I am absolutely not convince by the argument that the proposed new method is "cost-effective" or "cheaper than current approaches while offering better and more data" (page 13, line 293). I will not come back on data quality here (although beyond homoplasy, missing data rate is very low when using capillary sequencer), but I do not think that the described method is cheaper to genotype 96 individuals at 50 microsatellite than a traditionally used capillary-based genotyping method when it is well optimized with high multiplexing level (Guichoux et al. 2011). While it can be difficult to estimate laboratory procedure cost because price became quickly outdated as technology improved, I still think that you should give some justification about why you think your method is cheaper (besides labor costs that is not really relevant in a publicly founded research institute context).

Page 12, line 276: I did not understood here if "short amplicons (less than 300bp)" included primers and adaptors or just the targeted microsatellite loci (repeats and flanking regions). Please specify.

References:
Reviewer: 2

Comments to the Author(s)
None

***************************************************************

Journal Name: Royal Society Open Science
Journal Code: RSOS
Online ISSN: 2054-5703
Journal Admin Email: openscience@royalsociety.org
Journal Editor: Emilie Aime
Journal Editor Email: emilie.aime@royalsociety.org
MS Reference Number: RSOS-150312
Article Status: REJECTED
MS Dryad ID: RSOS-150312
MS Title: A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding
MS Authors: Vartia, Salla; Villanueva-Cañas, José; Collins, Patrick; Farrell, Edward; Finarelli, John; Hughes, Graham; Carlsson, Jeanette; Gauthier, David; McGinnity, Philip; Cross, Tom; FitzGerald, Richard; Mirimin, Luca; Crispie, Fiona; Cotter, Paul; Carlsson, Jens
Contact Author: Salla Vartia
Contact Author Email: salla.vartia@gmail.com, sallavar@hotmail.com
Contact Author Address 1: Area 52 Research Group, School of Biology & Environment Science and Earth Institute, University College Dublin, Belfield, Dublin, Ireland
Contact Author Address 2:
Contact Author Address 3:
Contact Author City:
Contact Author State:
Contact Author Country: Ireland
Contact Author ZIP/Postal Code:
Keywords: Amplicon sequencing, GBS, SSR, next generation sequencing, universal primer, Gadus morhua
Abstract: This study validates next generation sequencing based ‘Genotyping By Sequencing’ (GBS) of microsatellite loci for rapid and cost-effective genotyping in large-scale population genetic studies. The recovery of individual genotypes from large sequence pools was achieved by PCR-incorporated combinatorial barcoding using universal primers. Three experimental conditions were employed to explore the possibility of using this approach with existing and new multiplex marker panels and weighted amplicon mixture. The GBS approach was validated against microsatellite data generated by capillary electrophoresis. GBS allows access to the underlying nucleotide sequences that can reveal homoplasy, even in large data sets. GBS of microsatellites, using individual combinatorial barcoding, is faster and cheaper than current microsatellite approaches and offers better and more data.
EndDryadContent

Author's Response to Decision Letter for (RSOS-150565)

See Appendix A.
RSOS-150565.R1 (Revision)

Review form: Reviewer 2 (Olivier Lepais)

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Some supporting results are available but all supporting data are not included in the supporting material: contrary to what is stated in Data accessibility section, Table S10 is missing from the supporting material at this stage.

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept with minor revision (please list in comments)

Comments to the Author(s)

The revised version of the manuscript by Salla Vartia and collaborators has been much improved and the authors responded convincingly to my different comments on the previous version of the manuscript. I think that the relevance of their work appears more clearly and the discussion provides several useful advises to implement their approach by using updated technologies and adapting the protocol to scale genotyping up to typical needs of population genetics applications. The discussion section, especially page 14, make me think about two aspects that I felt was missing and that I would like, at the discretion of the authors, the authors to shortly address in the discussion.

I agree with the authors that homoplasy leads to underestimation of diversity (page 14, lines 319-322). However, beyond homoplasy, determining microsatellite repeat number also has the advantages of improving genetic diversity comparison (e.g. Petit et al. 2005) and more generally population genetic inferences (such as demography reconstruction) because longer microsatellite loci are usually more prone to instability and thus have a higher mutation rate. In short, accessing the molecular structure of the microsatellite markers will increase our understanding of the mutation model of the studied loci and thus the quality of the information we retrieve from the data.

At lines 302-303, page 14, the authors explain that accessing to the microsatellite sequence will allow for the development of automated genotyping. However, I think that bioinformatics tools allowing automated genotyping from high-throughput microsatellite sequencing are already available. A tool called MicNeSs has been released and very recently published (Suez et al. 2015). It may be interesting to cite this reference because it is appropriate at this stage of the discussion and also because it will create a link between the two studies helping to consolidate the research field and benefiting future readers interested in sequencing microsatellites.
Olivier Lepais. INRA, UMR 1224 Ecobiop.

References:

Decision letter (RSOS-150565)

30-Nov-2015

Dear Mr Vartia

On behalf of the Editor, I am pleased to inform you that your Manuscript RSOS-150565 entitled "A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding" has been accepted for publication in Royal Society Open Science subject to minor revision in accordance with the referee suggestions. Please find the referees' comments at the end of this email.

The reviewers and Subject Editor have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the comments and revise your manuscript.

• Ethics statement
If your study uses humans or animals please include details of the ethical approval received, including the name of the committee that granted approval. For human studies please also detail whether informed consent was obtained. For field studies on animals please include details of all permissions, licences and/or approvals granted to carry out the fieldwork.

• Data accessibility
It is a condition of publication that all supporting data are made available either as supplementary information or preferably in a suitable permanent repository. The data accessibility section should state where the article's supporting data can be accessed. This section should also include details, where possible of where to access other relevant research materials such as statistical tools, protocols, software etc can be accessed. If the data has been deposited in an external repository this section should list the database, accession number and link to the DOI for all data from the article that has been made publicly available. Data sets that have been deposited in an external repository and have a DOI should also be appropriately cited in the manuscript and included in the reference list.

If you wish to submit your supporting data or code to Dryad (http://datadryad.org/), or modify your current submission to dryad, please use the following link: http://datadryad.org/submit?journalID=RSOS&manu=RSOS-150565

• Competing interests
Please declare any financial or non-financial competing interests, or state that you have no competing interests.
• Authors’ contributions
All submissions, other than those with a single author, must include an Authors’ Contributions section which individually lists the specific contribution of each author. The list of Authors should meet all of the following criteria: 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published.

All contributors who do not meet all of these criteria should be included in the acknowledgements.

We suggest the following format:
AB carried out the molecular lab work, participated in data analysis, carried out sequence alignments, participated in the design of the study and drafted the manuscript; CD carried out the statistical analyses; EF collected field data; GH conceived of the study, designed the study, coordinated the study and helped draft the manuscript. All authors gave final approval for publication.

• Acknowledgements
Please acknowledge anyone who contributed to the study but did not meet the authorship criteria.

• Funding statement
Please list the source of funding for each author.

Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days (i.e. by the 09-Dec-2015). If you do not think you will be able to meet this date please let me know immediately.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions". Under "Actions," click on "Create a Revision." You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referees and upload a file "Response to Referees" in "Section 6 - File Upload". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referees.

When uploading your revised files please make sure that you have:

1) A text file of the manuscript (tex, txt, rtf, docx or doc), references, tables (including captions) and figure captions. Do not upload a PDF as your "Main Document".
2) A separate electronic file of each figure (EPS or print-quality PDF preferred (either format should be produced directly from original creation package), or original software format)
3) Included a 100 word media summary of your paper when requested at submission. Please ensure you have entered correct contact details (email, institution and telephone) in your user account
4) Included the raw data to support the claims made in your paper. You can either include your data as electronic supplementary material or upload to a repository and include the relevant doi within your manuscript
5) Included your supplementary files in a format you are happy with (no line numbers, vancouver referencing, track changes removed etc) as these files will NOT be edited in production
Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes

Emilie Aime
Senior Publishing Editor, Royal Society Open Science

Author’s Response to Decision Letter for (RSOS-150565)

Dear Emilie Aime,

On behalf of all co-authors, we are delighted with your decision to publish our manuscript in Royal Society Open Science. We agree with the reviewer's comments and have made the suggested changes to the discussion on page 14. The two suggested references have been added and are discussed in the text.

Best wishes,
Salla Vartia
Appendix A

21st October, 2015

Dear Prof. Gut,

Thank you very much for the thorough review of our manuscript (ID RSOS-150312) entitled "A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding". The comments made by the reviewer were very valuable and have helped us to significantly improve the focus and clarity of the manuscript.

We have considered all of the comments of Reviewer 1 and have provided detailed answers below. We have also made a number of changes to the manuscript and have used track changes to mark these. Reviewer 2 did not provide any comments.

We hope you will find this revision satisfactory and look forward to hearing from you in due course. Please feel free to contact me if you have any questions with regards to our resubmission at my email (salla.vartia@gmail.com).

Yours truly and on behalf of all co-authors,

Salla Vartia
Corresponding Author

Reviewer 1 Comments.

(1) Markers and individuals multiplexing level

Comment 1. The authors used a somewhat outdated sequencing platform (454 GS-FLX Titanium) which is not an issue by itself because their results are still totally relevant by themselves. The only downside is for the reproducibility of the method they developed.

Response 1. We agree with this comment, however, the method presented in the ms is not specifically for the 454 NGS platform as it is a proof of a generic concept study and we realise that much more data could be generated using more modern NGS platforms. We were in this case limited to the 454 NGS platform as we used already developed loci that had amplicon sizes larger than attainable by other NGS platforms. However, de-novo development of loci with shorter amplicon length than used here would be much more
efficient using the newer NGS platforms. These issues have now been clarified in the revised ms (see attached ms with track changes).

**Comment 2.** As they emphasized page 12 (line 264-277) modern sequencing platform produced shorter reads (which may not be a too big issue) at a much higher throughput (10 to 30 times more reads). I think that the authors should spend more time to discuss how their method could be leverage on more modern platform to show readers how it can be relevant to nowadays technologies. For instance, why sequencing 10 to 30 times more microsatellites for 96 individuals, what could be the application behind such analysis? I think a more useful way would be to open up new perspectives by multiplexing 10 to 30 more individuals (a few thousands) at a moderate number of microsatellite loci (20 - 50) because this kind of data is very common in population genetics applied to evolutionary biology or ecology.

**Response 2.** We agree with this comment that it would, from a population genetics standpoint, be much more efficient to run fewer loci but more individuals as this would maximize the cost effectiveness of generating population genetic data. In the current study we used 12 forward and 8 reverse barcodes yielding 96 unique individual combinations. This number of individuals and loci was sufficient for examining the potential for genotyping by sequencing of microsatellite loci as was the aim of the study. However, it is in fact possible to use an unlimited forward and reverse barcodes. For example using 48 forward and 48 reverse barcodes would allow for analysing 2304 individuals in a single sequencing run. Hence, the number of individuals versus loci is a trade-off depending on the number of sequences produced by each NGS platform. In comparison to Campbell et al. (2015) which use a plate specific barcode and 96 well specific reverse barcodes, our approach also achieves higher sequence diversity in the start of the sequence which is an important consideration when using Illumina based NGS as cluster identification (in Illumina based NGS) and hence sequence yield can be affected by low diversity libraries. This means that our approach is applicable for multiple NGS platforms. These issues have now been clarified in the revised ms (see attached ms with track changes).

Furthermore, our ms is only an examination of the potential for using genotyping by sequencing and deployment of the approach could easily use less loci but more individuals as suggested by the reviewer. However, one aim of the study was to simulate how loci could be developed de-novo when the yield of microsatellite containing sequences to working microsatellites is unknown and thus a larger initial number of loci would ensure that more loci would be available for downstream deployment. Another advantage of using a larger number of loci is that it will allow for high-grading informative markers which in turn allows for more efficient future data generation. These issues have now been clarified in the revised ms (see attached ms with track changes).

(2) Data analysis reproducibility and genotypic data limitations

**Comment 3.** While the main aim of a methodological paper is to be fully reproducible, I was very surprised to only found 9 lines (page 8, lies 156-164) describing the bioinformatics data analysis. With so few details about how the authors used the few hundred of thousand of reads output by the sequencer to generate the genotypic data, I cannot figure out how to redo the analysis myself which is a shame because all other aspects of the method are
otherwise fully described. The authors have to provided additional details and justification about the bioinformatics analyses they used. How was the number of errors allowed for the barcode and primer sequence determined and how does this affect the results?

**Response 3.** We agree that more information is needed and we have now included a more detailed description of the bioinformatics. In short, a simple python script was developed to identify raw sequence reads which contained both the forward and reverse barcodes and classified them by their locus specific primers. Once sorted and grouped into individuals and loci the sequences can be genotyped manually based on fragment length. The python script will also be made available on GitHub. We have clarified this in the revised ms (see attached ms with track changes). From a draft analysis allowing for more errors also lead to more reads allocated (data not shown) as expected.

**Comment 4.** More importantly how do they handle allele sequence instability (due for instance to slippage and partial adenylation) and could identify alleles from noise especially when two alleles had similar number of repeats? Frankly, just saying that a commercial software with the default parameters was used to verify read alignment is far from enough for a standard reporting of a bioinformatics analysis especially when the method had to be reproducible.

**Response 4.** Similar to capillary sequencer based genotyping our approach creates stutter and other PCR artifacts. We see no apparent solution to this unless the amplification stage based on PCR can be omitted. However, having access to the underlying sequence reads allows for closer inspection, unattainable by fragment length based genotyping, and dubious allele calls can be further examined. When alleles had similar number of repeats, we called the alleles based on sequence length frequency histograms (much like ABI based methods). The method described in this ms is also a "proof of concept" study and unless there is a need to use larger microsatellite loci (in excess of 300 bp) we would argue that using more modern NGS platforms (Ion Torrent or MiSeq) would allow for increasing the read depth significantly and hence improve the chances of detecting PCR artifacts and sequencing errors.

**Comment 5.** In addition, it is a very interesting point that sequencing microsatellite allow for the detection of homoplasy, but again more analytical details are needed for the reader to understand how the analysis was performed, in particular, how was it possible to differentiate real mutation from sequencing errors given the relatively low coverage of 10X chosen for the analysis (page 8, line 162-164)?

**Response 5.** This is a good point. The corresponding section has been modified to provide more information of the analysis. The point of this was not an in depth homoplasy analysis, but rather a rough analysis to estimate how prevalent the issue of homoplasy may be. At the low coverages it may not be possible to differentiate real mutation from sequencing errors, this has now been clarified in the text.

**Comment 6.** Finally, a good genotyping method should match the requirement need for the application that will be made with the genotypic dataset. However, high rate of missing genotypic data is a notorious issue when using high throughput sequencing platforms,
which with SNP is usually compensated by the very high number of markers studied. However, the described microsatellite sequencing method seems to generate dataset with a high rate of missing genotype. With data for 3325 genotypes out of 5088 (page 10, line 200) amounting to 45% of missing data and 10 loci out of 53 with less than 50% of the individual genotyped (lines 201-203), I was wondering what analytical power could be expected from such a dataset. Would linkage mapping, parentage analysis or genetic assignment still be possible despite such a high missing data rate? I think you should discuss this explicitly and propose improvement of the protocol that could help to lower missing data in the final dataset.

**Response 6.** This is a very good point and we appreciate the reviewer highlighting it. As previously noted the 454 technology has been superseded by Ion Torrent and Illumina technologies which deliver significantly more sequences/reads per sequencing run. For instance using the MiSeq would yield approximately 25m sequences. This would enable significantly more reads to be generated for each locus for each individual. Whilst there will still be an error rate it is expected that greater read depth will enable differentiation between errors and true alleles. Our method can still be implemented on many existing microsatellite panels with fragment lengths in excess of 400 bp. However, we would not suggest to use as many loci as we deployed in the current study. This would allow for both increasing read depth and the number of individuals analysed. As NGS continues to improve it is not unlikely that other NGS platforms than 454 would be able to accommodate larger amplicons. However, that would not change the approach we present in this ms as it is a generic approach that is capable of using any NGS platform. In fact, we recommend that future implementations of the approach described here should take advantage of more modern NGS platforms. Similarly, increased numbers of reads would reduce the amount of missing data significantly as more loci would have sufficient read depth to be genotyped. These statements have now been included in the discussion to present the method more clearly.

**Minor comments**

**Comment 7.** Page 6, line 85: I think that the field of medical research are also interested in sequencing microsatellite because short tandem repeat are found to be associated with diseases, so that not only forensic science is involved in sequencing microsatellites. In addition, with the realization that short repeats main may have important functions and be linked to phenotypic variability, there would be a increasing interest in sequencing microsatellites in near future which could make you study even more interesting to readers (for instance, see the very good review of Press et al. 2014).

**Response 7.** Thanks for this helpful comment. It is certainly something of interest. For the moment though we are focussed on potential application for population genetics and do not want to make our discussion too broad. We have revised the ms so that it does not read that only forensic sciences are interested in this research.

**Comment 8.** Page 11, lines 226: I am absolutely not convinced by the argument that the proposed new method is "cost-effective" or "cheaper than current approaches while offering better and more data" (page 13, line 293). I will not come back on data quality here (although beyond homoplasy, missing data rate is very low when using capillary sequencer), but I do not think that the described method is cheaper to genotype 96 individuals at 50
microsatellite than a traditionally used capillary-based genotyping method when it is well optimized with high multiplexing level (Guichoux et al. 2011). While it can be difficult to estimate laboratory procedure cost because price became quickly outdated as technology improved, I still think that you should give some justification about why you think your method is cheaper (besides labour costs that is not really relevant in a publicly founded research institute context).

**Response 8.** We agree with the reviewer. It would not be cheaper to run 96 individuals for 50 loci using the 454 approach outlined in this ms. Hence, to clarify we have changed the text as pointed out by the reviewer and elsewhere in the discussion to highlight the potential of the technique to be more cost-effective than ABI based approach by using less loci and more individuals or by changing to an NGS platform that has higher sequence yield for less cost. Regarding the data quality issue it is important to note that inter-laboratory reproducibility of capillary based genotyping is inherently difficult and often unsuccessful. GBS removes many of the issues associated with this as the actual sequence data is available and may be scrutinised for comparison.

Further, we disagree with the comment that labour costs are not relevant in a publically funded research institute context. Labour when considered both in time and financial costs are the largest component of any grant application and are often included at the expense of equipment and consumables budgets. For example, the large amount of time and cost spent performing multiple small genotyping runs with small multiplexes on an ABI could be replaced with a single NGS run with thousands of individuals and a large number of loci. Thus saving time and budget for other research.

**Comment 9.** Page 12, line 276: I did not understood here if "short amplicons (less than 300bp)" included primers and adaptors or just the targeted microsatellite loci (repeats and flanking regions). Please specify.

**Response 9.** We have clarified this in the text