We appreciate the work of the reviewers and editor in providing feedback on our manuscript. We are happy to make the changes requested and have included our responses in this document in blue font.

Editorial comments:

Changes to be made by the Author(s):

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.

We have done so.

2. Please add more details to your protocol steps. Please ensure you answer the “how” question, i.e., how is the step performed? Alternatively, add references to published material specifying how to perform the protocol action.

We have significantly revised the protocol section with additional clarifying details and steps, including the ones suggested by reviewers. For Reptile error correction we cite a highly detailed step-by-step protocol published by the Eisen lab.

3. Please provide all user input commands, either the terminal line commands or the click by click instructions if there is a GUI. If using terminal line commands, the entire line command for all steps must be provided.

Commands and click-by-click instructions have been added where needed.

4. 1.1: How is PEAR used?

A more detailed description has been added, including commands, for how to run PEAR.

5. 1.2.4: How is Reptile error correction done?

We have added a citation to a publication giving a step-by-step protocol for Reptile error correction of RNA-seq reads.

6. 1.3: How is this done?

Details of running Trinity have been added.

7. 5.1: What is the reaction mixture? Please provide all volumes and concentrations throughout.

We have added in the volumes and concentrations needed for the reaction mixture.

8. 5.2: How is qPCR signals measured? What device is used? How much is used for measurement?

We have added in a statement about the device that is used to run and analyze qPCR.

9. 5.3: What is actually done here?

We have clarified how to perform this step.

Reviewers' comments:

Reviewer #1:

Manuscript Summary:

This manuscript provides methodological detail behind a recently published gene discovery. The text here includes step-by-step bioinformatics that was presumably not published previously. The method details might be obvious to someone practiced in the art of bioinformatics but they would be nevertheless critical for anyone attempting to reproduce this study. The manuscript also includes helpful python scripts that presumably appear for the first time.

Major Concerns:

The writing requires significant attention. The title is off-topic. The abstract fails to summarize the method described. The introduction fails to provide relevant background. The figures are generic and unexplained. A (hopefully helpful) section-by-section critique follows.

We thank the reviewer for the feedback. We feel that the reviewer was looking for an extended methods section to our publication in Current Biology. However, our goals for this manuscript were to describe the subtractive method in a way that is broadly generalizable to other studies, and to provide general guidelines for implementation of these methods in other work. This helps explain the concerns of the reviewer regarding 'generic' figures etc. We have however enhanced the focus and detail about the specific study we published in Current Biology.

Title ========

The portion of the title following the colon hardly describes the current work. No fog is described here, not even metaphorically. The title should indicate that a gene was discovered by examination of the small portion of RNA sequence that could not be characterized after sequencing RNA from a model organism. Unless the paper demonstrates that the method generalizes to other species, the title should probably mention the species.

We deleted the reference to a metaphorical fog, as suggested by the reviewer, but trying to make sense of a vague or indistinct object in a fog is a suitable picture for certain kinds of genomic analysis, in which vast amounts of data are generated and then filtered and refined over time to get a clearer picture. We describe the implementation of the method in viral, bacterial, and other studies, in addition to our own, and also describe cases where the method could be helpful but was not used. Thus, the text demonstrates that the method is generalizable to a broad range of other studies in a variety of other species, using either DNA or RNA, and is not just limited to our zebra finch work.

Abstract ======

The Abstract does not adequately summarize the paper. It gives lots of background and lots of jargon but hardly any mention of the novel work presented in this paper.

We intended the Abstract to succinctly describe the method, the challenges, and our implementation of it. The novelty of our approach comes from linking the various steps of subtractive genomics into a cycle, which can be implemented in a variety of scientific settings and can be iterated multiple times as required by a specific study. We have edited the abstract to make this clearer and would be happy to address specific additional comments from the reviewer.

Germline restricted chromosomes are unusual in biology and the current work may represent the first study of them by subtractive genomics. The abstract would be helpful if it contained a sentence explaining this phenomenon.

The germline-restricted chromosome is not the focus of this manuscript. More details on the germline restricted chromosome are included in the introduction where we find it is better suited for this manuscript. Furthermore, the abstract is space constrained, so we do not have room to include the description in the abstract.

All of these sentences in the Abstract could be unclear to someone who has not yet read the article:

\*particularly useful when T cannot be easily separated from R

We moved this sentence to later in the abstract where it is more clearly understandable.

\*missing the sequence of interest, but does not need to be perfectly known for the method to work, because the sequences remaining after subtraction are tested through quantitative Polymerase Chain Reaction (qPCR)

We have clarified the sentence.

\*Multiple references can be used to improve filtering efficiency either in the first subtraction step or during cycles of filtering and qPCR testing.

We have changed the wording in order to make the sentence less confusing.

\*A drawback of the method is finding a suitable reference for subtraction and obtaining T-positive and negative samples for qPCR testing.

We were told in the JoVE guidelines to list pitfalls and drawbacks of the method in the abstract. We feel this sentence is clear as written but would be happy to entertain specific comments from the reviewer that s/he feels would improve it.

Introduction ======

The Introduction provides too much ancillary information and not enough crucial background. Three paragraphs about other investigations of other organisms in which subtractive genomics was or was not used is wide of the mark. Only the last paragraph of the Introduction provides the essential facts but in compressed form. The Introduction should review the biology: the unusual germline restricted chromosome in finch, what was known about it prior, why it has been uncharacterized for so long, whether it would be contained in all ovary and testis cells used here, whether it was expected to contain novel genes, and how the new gene that was discovered and characterized. The Introduction should also prep the bioinformatics: prior work with subtraction (briefly), why the subtraction method was selected rather than physical purification, how subtraction led to the new gene discovery, whether the existence of a paralog in the somatic genome presented a challenge. Finally, since the Introduction notes that gonad RNA will be mapped to muscle DNA and to raw Sanger genomic reads (why?) and to brain transcripts, the Introduction should discuss the motivation for, and challenges presented by, this RNA/DNA disconnect.

We appreciate the feedback. We have added some specifics of the zebra finch work, but we feel many of the background details requested by the reviewer would unnecessarily clog up the manuscript, making it instead a review of the field, which has been done elsewhere by others and by ourselves in Current Biology. We have, however, clarified that the use of three different reference sequences (two DNA, one RNA) was driven entirely by the qPCR results obtained during the iterations of the cycle. This helps more strongly link the motivation of what we did in our study with the cyclical method we describe in this manuscript.

Protocol =======

The protocol description is too general in places and too specific in others. At one extreme, readers are told to map with BWA or bowtie without mention of parameters. At the other extreme, the instructions refer to some unnamed specific file format (quality scores "in the third column") and to a file (the ".lsf" file) that would only be present on one kind of compute environment. For accuracy, the authors should describe their compute environment and give the specific commands and parameters, but for generality, they should also describe each step in general terms.

We agree with the reviewer our protocol was too vague in general. We have clarified the Protocol and added the missing detail.

Step 5.3 includes a Note that says "abundance … is key". This is not a procedure nor is it helpful as written. The description of step 6.1 is totally unclear.

We have removed the problematic Step 5.3 text entirely and have re-written step 6.1 to be clearer.

Figures =======

Figure 1. The box sizes are probably not proportional to real data. If a Venn diagram is to be shown, it should be proportional.

We respectfully disagree with the reviewer here. While some Venn diagrams are proportional, some are not. As this one is a conceptual Venn, it is appropriately non-proportional.

The figure should include specific numbers (# sequences and # bp) from each loop of the experiment. Since the process involves a loop, the figure really should have a time dimension to show how the sequence set evolves over 2 or more iterations. As is, the figure is unhelpful.

We have addressed this by revising Figure 3 into a more sequential loop-type figure 2B. This also allows us to show the numbers at each subtractive iteration as requested by the reviewer alongside each loop. We have therefore removed Figure 3.

Figure 2, or at least its caption, should clarify what the last blue arrow represents (the arrow heading back into BLAST). It is not clear what sequences are emitted from the prior step ("validation of sequence") or what are the subject and query sequences input to BLAST. More detail would help.

We agree and have made changes to this figure, also integrating a modified Figure 3 into a Figure 2B. We have added an explicit decision node, guided by qPCR, to Figure 2A along with an 'end' outcome, that helps to clarify precisely how our validation by qPCR fed into the overall cycle.

Figure 3 is also probably not drawn to scale. Its caption provides no information at all. It is unclear whether labels such as "78 proteins" refer to the number that were characterized or that remained uncharacterized. It is unclear why 338 genes exit the pipeline part way through. The figure does not demonstrate the looping behavior of the process that was illustrated in Figure 2.

Figure 3 has been re-drawn as Figure 2B, with looping and clarified numbers as requested. The figure legend has been edited and clarified.

Figure 4 is valuable but the caption only serves to label the charts. Especially for readers looking at the figures before reading all the text, the caption should explain the significance of this figure i.e. repeatable detection of gonad expression of the novel genes.

We thank the reviewer for the feedback and have added the clarifications as requested.

Representative results =======

This section should start by stating which subset of results are being presented, why they were selected, and to what extent they are representative. As is, the text gets around to this in its fourth paragraph (line 234) when it says, "For simplicity we will describe the processing of the ovary sequence…". What makes this subset simple?

We meant to imply that the Testis subtraction was omitted from our description to simplify it. We have clarified the sentence.

This section goes beyond representative results. It seems to be partly a FAQ including things that can go wrong and what to check when they do. A FAQ is problematic here since it does not report scientific results and is impossible to review for accuracy. Perhaps a FAQ document should be provided as a supplement. Even as a FAQ, this section is inconsistent e.g. it provides suggestions for the case of no BLAST hits but none for the case of no qPCR amplification.

A FAQ ('frequently asked questions') would have questions and answers. Our section has no questions so it is not a FAQ and it does not adhere to FAQ format. We respectfully disagree with the reviewer's claim that only novel scientific results can be reviewed for accuracy, since accuracy of a method can also be assessed. JoVE papers are visualized methods so the main concern when reviewing the text portion is whether the method, pitfalls, and controls necessary are adequately described. If the goal of the manuscript were to report overall new scientific results, we would have submitted to a different journal. As it is, this section adheres to JoVE guidelines specifying that authors include a discussion of what can go wrong with the method and what to do when this happens. As no other reviewer raised an objection we suggest it remain as written, but we have added a portion about what to do if qPCR fails as suggested by the reviewer.

Reviewer #2:

Manuscript Summary:

I like the manuscript and its very useful approach. Sequencing and in-silico work is the future and its importance will even increase. I found just some minor points to correct for you.

The zebra finch example, with which you showed the methodology was very interesting. The methodology is described in detail except for 2 points I remarked in the minor concerns. The results and applicability of the method are discussed in various directions.

Minor Concerns:

Line 111+113: As it is a methodology paper, I would suggest to give some more details how to run PEAR and Reptile (as you gave them also for other tools) or if only default parameters are used, state that the default parameters were used.

We thank the reviewer for the useful comments and are gratified that it was found interesting. We are happy to address the two main concerns, and have thoroughly re-worked the protocol section to provide all necessary missing detail including for PEAR and Reptile.

Line 136: You refer to section 2b, but I guess 2.1 is meant.

We thank the reviewer for catching the typo. We meant 2.2 and have updated the text.

Reviewer #3:

Novel sequence discovery by subtractive genomics: Peering into the fog of the unknown JoVE58877

General comments:

This manuscript describes a useful method to find novel sequences that have little or no similarity to sequences in the databases.

Reference to figure 4 is done before figure 3 in the text (Lines 224 and 227). The order of the figure numbering should change to correct this.

We thank the reviewer for catching this error, based off of other comments figure 3 has been changed and is now figure 2B. This also resolves the numbering order.

The text is somewhat chatty and lacking in precision at some points, but is also very detailed in its description of the procedure.

Some improvements of the text are listed below.

Specific comments:

Line 27 The meaning of "separated" is unclear to me in this context. Please rephrase and specify if physical separation (i.e purification) is the desired meaning.

We thank the reviewer for the suggested clarification, that statement has been rephrased to “physically isolated.”

Line 31 No sequence is "perfectly known", please rephrase or omit

We have addressed this by changing the phrase to "complete,” to communicate that sequences cannot be perfectly known.

Line 48 "experimentally isolated" should be "physically separated"? Experiments do encompass both wet lab and

We have clarified by changing the wording to “physically separated” to signify wet lab experiments and we acknowledge the point that both bench and computational methods can be used in separation.

Lines 64 - 73 In the beginning of this section it is claimed that "Readhead et al. used subtractive genomics". This seems to contradict Line 67 where

it is stated that "Subtractive genomics could have been used" referring to the same study. Please rephrase this section to clarify.

This paragraph has been edited for clarity.

Line 80 "BLAST was able to positively identify the target sequences". BLAST is a tool by which the target sequences can be identified but BLAST

can not identify anything by itself.

We have restructured the sentence so that we are not implying that BLAST identified a sequence. We thank the reviewer for catching this error.

Line 109 "De Novo Assemble Sequence of Interest", please add the letter T for coherence, for example "De Novo Assemble Sequence of Interest (T)"

We agree this helps the sentence and have added the suggested '(T)'.

Line 167 and 172 The provided Python scripts work only with Python 2 in their present form. This is not stated anywhere in the manuscript. Using the link

provided as a source for python (in Table of Materials) will provide the reader with Python 3. Further, Python 2 will not be maintained past 1 Jan 2020.

A recommendation for the Python version has to be added in this section. The best would be to also provide Python 3 versions of the scripts.

We have added a statement about the scripts requiring Python 2.7, which is a standard Python release. We have corrected the link in the Table of Materials as well.

Line 176 "extract a fasta with these" --> "extract a fasta file with these".

This problem has been resolved and we thank the reviewer for catching this error.

Line 202 "isolated without a host cell's genome" --> "isolated without a host cell's genomic DNA"

We have fixed this sentence and thank the reviewer for noting it.

Reviewer #4:

In this manuscript the authors present a method called „subtractive genomics" and outline a step-by-step instruction to perform the analysis. The manuscript is clearly structured; however, a few issues remain as outlined below:

MAJOR

\*) The whole protocol is very much designed with respect to the previously published manuscript. It would be great if the authors could demonstrate the suitability of their protocol with another (maybe publicly available) dataset and rerun the whole protocol.

While we agree analyzing another dataset would be valuable, we demonstrate in the text that subtractive genomics have already been used in other contexts, such as identification of non-cross-reacting bacterial vaccine targets and novel virus identification. Given that we did not invent the method ourselves, but that others have already implemented it in myriad other contexts, we feel rerunning the whole protocol would do little to further improve the manuscript, while considerably delaying publication of our paper which provides valuable methodological details that are absent from the literature to date. We also innovate in making the whole process cyclical, with qPCR to determine whether more iterations are necessary, a structure we feel may benefit many other studies.

\*) Major parts (e.g., "Positive results") are repeats of the previously published manuscript. Please correct.

Our JoVE manuscript provides details not provided in the previously published manuscript. The text in the "Positive results" section has not appeared elsewhere. Given that neither the text nor the details have been published previously, we do not see this as an error in need of correction. We do note, however, that Figure 3 has been replaced with a new Figure 2B which more conceptually captures the pipeline we employed and is more distinct from our published Figure in Current Biology, which we hope alleviates this Reviewer's concerns. Furthermore, the data in Figure 4A-D (now 3A-D) are all previously unpublished, showing our raw initial data results for both negative and positive results. Again, none of this is a repeat of previous work so no correction is required.

\*) The title of the manuscript infers that this protocol can be used for "arbitrary" sequence discovery, while the manuscript only talks about "gene identification". Please adjust.

The method can be used for any sequence discovery: coding sequence or noncoding, using RNA or DNA. It just so happens that some examples we discuss in greatest detail are based on gene discovery but we also discuss mitochondrial genome reconstruction, where it might apply to noncoding DNA, in plants and ciliates. We do not want to arbitrarily limit the use of the method to gene identification so we suggest the title remain broadly applicable.

\*) The introduction into the field of "subtractive genomics" is not detailed enough and should be rewritten.

We introduce the concept and describe two specific examples in which the method is currently used, describe some cases where parts of the method are used and then discuss the specific example from our own studies. We are happy to edit or add detail as the Reviewer suggests, but it is unclear from his/her statement which portions are objectionable. If the Reviewer would provide more specific guidance we could address it.

\*) The authors should mention what data can be analyzed with this protocol: NGS, RNASeq, … currently, this should be explicitly stated.

We agree and have indicated in Step 1 of the protocol that any NGS data, and its corresponding assembly, may be used to start the process (we also note that NGS is an umbrella term incorporating RNA-Seq so separating the concepts as the reviewer does is technically imprecise). We make it explicit in the text that the method works on any NGS data assembly, derived from either DNA or RNA.

\*) Please indicated what quality of the draft sequence is needed. Max number of contigs. N50 size, …

This will be highly project-dependent and impossible to specify beforehand. Given that the input data may come from a genome or transcriptome project and any number of different sequencing methods, we cannot give specifics here without compromising the applicability of our method to various projects. For example a good N50 for a transcriptome project (many short contigs corresponding to transcripts) would be quite poor for a genomic DNA assembly. Fortunately the standards are well established and we provide references describing both assembly quality control and assessment metrics for the user to examine for more information, rather than trying to reinvent the wheel in our manuscript.

\*) Please report why no adapter trimming of the RNASeq data is needed?

We appreciate the Reviewer noting this accidental omission. Adaptor and quality trimming were performed and accidentally omitted; we have included this information in the Protocol as Step 1.1.

\*) Please mention that your protocol requires RNASeq data. Also mention what coverage is needed, what RNASeq protocol (I guess you require PE sequencing), what sequencer, …

Our protocol does not require RNA-seq data, as it would work equally well on DNA-seq data or DNA-based data from PacBio or other technologies like Oxford Nanopore. A specific coverage is not needed for RNA-seq data, but we used 10 million paired-end reads. This has been added to the manuscript.

\*) Please comment why no filtering of the RNASeq data is required.

The RNA-seq data were quality filtered with trimmomatic as described above.

\*) Please mention in your protocol what prerequisites are required. What programs need to be installed, …

Adding this information would simply bloat the length of our text without appreciably improving the usability of it, given that dependencies are described in the README files and user manuals of each software package. Indeed we would obscure the flow of the protocol with so many comments and digressions that we feel the manuscript would become extremely hard to use.

\*) How do the authors ensure specificity of the designed primers? Is this required? Please comment.

We tested each primer set by PCR off genomic DNA from testis. Testis DNA encodes both the target (T) and the reference (R) so it is a suitable positive control template for primer testing. We describe this in the text under "What could go wrong" and also have added details to Protocol step 4.3 to make it clearer.

\*) Step 6 - what if no "new reference" is available.

Then the subtraction process is finished, as no more iterations of the method are possible. As in all research, a successful outcome (high-confidence identification of T) cannot be guaranteed. The method is still a valuable addition to the bioinformatic toolbox.

MINOR

\*) Please correct: "Convert the FASTQ file into FASTA and FASTQ files using Reptile readme file". I guess you mean the commands in the readme file; also convert FASTQ into FASTQ is misleading

We have removed this portion in simplifying the Reptile implementation section, citing a 2013 paper describing in step-by-step detail the method we used.

\*) It would be great if the authors could provide a Docker image to remove the burden of installing tools.

Because individual applications of subtractive genomics might use different software packages, this is not feasible. Furthermore, many tools require a cluster for parallel computing, making them incompatible with a Docker implementation. The two most important software packages are BLAST and BWA, and are quite easy to install and run locally; this can be demonstrated in the video portion of this JoVE publication. One valuable aspect of our manuscript is an opportunity to educate the scientific community about installing and running bioinformatic tools.

\*) Please mention the used versions of the tools.

We have added this information.

\*) Is your protocol also suitable for long-read sequencing methods (PacBio, Nanopore, …)? Please comment.

Yes, as long as an assembly is generated from the data, anything can be used as a starting point. We have added this to the text.

\*) Sentences should be polished and need to be more precise.

We have done so whenever possible.

\*) Line 217: What does "the total sequence" mean?

We have added the notation of R+T after the statement total sequence to remain consistent with earlier explanation of the flow of protocol.