Journal of Visualized Experiments

Novel sequence discovery by subtractive genomics

--Manuscript Draft--

Article Type:	Methods Article - JoVE Produced Video
Manuscript Number:	JoVE58877R1
Full Title:	Novel sequence discovery by subtractive genomics
Keywords:	Genomic subtraction, qPCR, BLAST, Python, Read mapping, De novo assembly, Primer design
Corresponding Author:	John R Bracht American University Washington, DC UNITED STATES
Corresponding Author's Institution:	American University
Corresponding Author E-Mail:	jbracht@american.edu
Order of Authors:	Kathryn Celestia Asalone
	Megan M. Nelson
	John R Bracht
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Washington, District of Columbia, United States of America

JOURNAL OF VISUALIZED EXPERIMENTS
1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

TITLE:

Novel Sequence Discovery by Subtractive Genomics

AUTHORS AND AFFILIATIONS:

Kathryn C. Asalone[1], Megan M. Nelson[1], John R Bracht[1] *

[1] Biology Department, American University, Washington, DC

ka5144a@student.american.edu
mn1298b@american.edu
jbracht@american.edu

Corresponding Author:
John R. Bracht

KEYWORDS:

Genomic subtraction, qPCR, BLAST, Python, Read mapping, *De novo* assembly, Primer design

SUMMARY:

The purpose of this protocol is to use a combination of computational and bench research to find novel sequences that cannot be easily separated from a co-purifying sequence, which may be only partially known.

ABSTRACT:

Subtractive genomics can be used in any research where the goal is to identify the sequence of a gene, protein, or general region that is embedded in a larger genomic context. Subtractive genomics enables a researcher to isolate a target sequence of interest (T) by comprehensive sequencing and subtracting out known genetic elements (reference, R). The method can be used to identify novel sequences such as mitochondria, chloroplasts, viruses, or germline restricted chromosomes, and is particularly useful when T cannot be easily isolated from R. Beginning with the comprehensive genomic data (R + T), the method uses Basic Local Alignment Search Tool (BLAST) against a reference sequence, or sequences, to remove the matching known sequences (R), leaving behind the target (T). For subtraction to work best, R should be a relatively complete draft that is missing T. Since sequences remaining after subtraction are tested through quantitative Polymerase Chain Reaction (qPCR), R does not need to be complete for the method to work. Here we link computational steps with experimental steps into a cycle that can be iterated as needed, sequentially removing multiple reference sequences and refining the search for T. The advantage of subtractive genomics is that a completely novel target sequence can be identified even in cases in which physical purification is difficult, impossible, or expensive. A drawback of the method is finding a suitable reference for subtraction and obtaining T-positive and negative samples for qPCR testing. We describe our implementation of the method in the identification of the first gene from the germline-restricted chromosome of zebra finch. In that case computational filtering involved three

references (R), sequentially removed over three cycles: an incomplete genomic assembly, raw genomic data, and transcriptomic data.

INTRODUCTION:

The purpose of this method is to identify a novel target (T) genomic sequence, either DNA or RNA, from a genomic context, or reference (R) (**Figure 1**). The method is most useful if the target cannot be physically separated, or it would be expensive to do so. Only a few organisms have perfectly finished genomes for subtraction, so a key innovation of our method is the combination of computational and bench methods into a cycle enabling researchers to isolate target sequences when the reference is imperfect, or a draft genome from a non-model organism. At the end of a cycle, qPCR testing is used to determine whether more subtraction is needed. A validated candidate T sequence will show statistically greater detection in known T-positive samples by qPCR.

Incarnations of the method have been implemented in discovery of new bacterial drug targets that do not have host homologs[1-4] and identification of novel viruses from infected hosts[5,6]. In addition to identification of T, the method can improve R: we recently used the method to identify 936 missing genes from the zebra finch reference genome and a new gene from a germline-only chromosome (T)[7]. Subtractive genomics is particularly valuable when T is likely to be extremely divergent from known sequences, or when the identity of T is broadly undefined, as in the zebra finch germline-restricted chromosome[7].

By not requiring positive identification of T beforehand, a key advantage of subtractive genomics is that it is unbiased. In a recent study, Readhead *et al.* examined the relationship between Alzheimer’s disease and viral abundance in four brain regions. For viral identification, Readhead *et al.* created a database of 515 viruses[8], severely limiting the viral agents that their study could identify. Subtractive genomics could have been used to compare the healthy and Alzheimer’s genomes in order to isolate possible novel viruses associated with the disease, regardless of their similarity to known infectious agents. While there are 263 known human-targeting viruses, it has been estimated that approximately 1.67 million undiscovered viral species exist, with 631,000-827,000 of them having a potential to infect humans[9].

Isolation of novel viruses is an area in which subtractive genomics is particularly effective, but some studies may not need such a stringent method. For example, studies identifying novel viruses have used unbiased high-throughput sequencing followed by reverse transcription and BLASTx for viral sequences[5] or enriching of viral nucleic acids to extract and reverse transcribe viral sequences[6]. While these studies employed *de novo* sequencing and assembly, subtraction was not used because the target sequences were positively identified through BLAST. If the viruses were completely novel and not related (or distantly related) to other viruses, subtractive genomics would have been a useful technique. The benefit of subtractive genomics is that sequences that are completely new can be obtained. If the organism’s genome is known, it can be subtracted out to leave any viral sequences. For example, in our published study we isolated a novel viral sequence from zebra finch through subtractive genomics, though it was not our original intent[7].

Subtractive genomics has also proved useful in the identification of bacterial vaccine targets, motivated by the dramatic rise in antibiotic resistance[1-4]. To minimize the risk of autoimmune reaction, researchers narrowed down the potential vaccine targets by subtracting any proteins that have homologs in the human host. One particular study, looking at *Corynebacterium pseudotuberculosis*, performed subtraction of vertebrate host genomes from several bacterial genomes to ensure that possible drug targets would not affect proteins in the hosts leading to side effects[1]. The basic work flow of these studies is to download the bacterial proteome, determine vital proteins, remove redundant proteins, use BLASTp to isolate the essential proteins, and BLASTp against host proteome to remove any proteins with host homologs[1-4]. In this case, subtractive genomics ensure that the vaccines developed will not have any off-target effects in the host[1-4].

We used subtractive genomics to identify the first protein-coding gene on a germline-restricted chromosome (GRC) (in this case, T), which is found in germlines but not somatic tissue of both sexes[10]. Before this study, the only genomic information that was known about the GRC was a repetitive region[11]. *De novo* assembly was performed on RNA sequenced from ovary and teste tissues (R+T) from adult zebra finches. The computational elimination of sequences was performed using published somatic (muscle) genome sequence (R_1)[12], its raw (Sanger) read data (R_2), and a somatic (brain) transcriptome (R_3)[13]. The sequential use of three references was driven by the qPCR testing at step 5 of each cycle (**Figure 2A**), showing that additional filtering was required. The discovered α-SNAP gene was confirmed through qPCR from DNA and RNA, and cloning and sequencing. We show in our example that this method is flexible: it is not dependent on matching nucleic acids (DNA vs RNA) and that subtraction can be performed with references (R) that are comprised of assemblies or raw reads.

PROTOCOL:

1. *De novo* Assemble Starting Sequence

NOTE: Any Next-Generation Sequence (NGS) data can be used, as long as an assembly can be produced from those data. Suitable input data includes Illumina, PacBio, or Oxford Nanopore reads assembled into a fasta file. For concreteness, this section describes an Illumina-based transcriptomic assembly specific to the zebra finch study we performed[7] ;however be aware that the specifics will vary by project. For our example project, raw data were derived from a MiSeq and approximately 10 million paired reads were obtained from each sample.

1.1. Use Trimmomatic 0.32[14] to remove Illumina adaptors and low-quality bases. On the command line, enter:

```
java -jar trimmomatic-0.32.jar PE -phred33 forward.fq.gz reverse.fq.gz -baseout quality_and_adaptor_trimmed ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:40
```

1.2. Use PEAR[15] v. 0.9.6 to create high-quality merged reads from trimmomatic output paired reads, using default parameters. On the command line, enter:

```
pear -f <quality_and_adaptor_trimmed_forward_paired_reads.fq> -r <quality_and_adaptor_trimmed_reverse_paired_reads.fq>
```

1.3. Use Reptile v. 1.1[16] to error-correct the reads produced through PEAR. Follow the step-by-step protocol described in[17].

1.4. Use Trinity v. 2.4.0[18] in default mode to assemble the corrected sequences. For strand-specific libraries, use the -SS_lib_type parameter. The output is a fasta file (your_assembly.fasta). On the command line, enter:

```
Trinity --seqType fq --SS_lib --max_memory 10G --output Trinity_output --left
quality_and_adaptor_trimmed_forward_paired_reads.fq --right
quality_and_adaptor_trimmed_reverse_paired_reads.fq --CPU 10
```

NOTE: The output will be placed in a new directory, Trinity_output, and the assembly will be named 'Trinity.fasta' which can be renamed as Your_assembly.fasta if desired. See the Trinity website for more details: https://github.com/trinityrnaseq/trinityrnaseq/wiki/Running-Trinity.

2. BLAST the Assembly against the Reference Sequence

NOTE: Use this step when the reference is an assembly or long reads like Sanger; if it is composed of raw Illumina reads, see step 3 below for mapping reads to the query. All BLAST steps were completed with version 2.2.29+ though the commands should work on any recent BLAST version.

2.1. Make a BLAST database of the reference sequence (nucleotide_reference.fasta) at the command line. Enter into the command line the following:

```
makeblastdb -dbtype nucl -in nucleotide_reference.fasta -out nucleotide_reference.db
```

2.2. BLAST-match the query assembly (generated in step 1) to the reference database. To obtain an output file, use [**-out BLAST_results.txt**] and to generate tabular output (required for subsequent processing steps with Python scripts), use [**-outfmt 6**]. These options can be combined in any order, so an example complete command is [**blastn -query your_assembly.fasta -db nucleotide_reference.db -out BLAST_results.txt -outfmt 6**]. If an e-value setting is desired, use the -evalue option with an appropriate number, for example [**-evalue 1e-6**]. Be aware however that the subtractive cycle effectively inverts the evalue setting in as described in the discussion.

2.3. For increased stringency, use protein sequences from the assembly as the BLAST query with translated nucleotide BLAST (tBLASTn), which performs 6-way translation of the (nucleotide) database. This method is recommended for most non-model systems, avoiding the problem of incomplete protein annotations.

2.3.1. Ensure the correct genetic code is selected for the organism being studied, using the -

db_gencode option. To obtain protein sequences for the query, run the TransDecoder.LongOrfs command (from TransDecoder package v. 3.0.1) to identify the longest open reading frames from assembled query sequences. The command is [**TransDecoder.LongOrfs -t your_assembly.fasta**]; the output will be placed in directory called 'transcripts.transdecoder_dir' and will contain a file called longest_orfs.pep containing the longest predicted protein sequences from each sequence in your_assembly.fasta.

2.3.2. To use tBLASTn, run the command [**tblastn -query longest_orfs.pep -db nucleotide_reference.db -out BLAST_results.txt -outfmt 6**]. If a high-quality protein reference is available, use protein-protein matching with BLASTp rather than tBLASTn.

2.3.3. Make a BLAST database of the protein reference [**makeblastdb -dbtype prot -in protein_reference.fasta -out protein_reference.db**] and then [**blastp -query longest_orfs.pep -db protein_reference.db -out BLAST_results.txt -outfmt 6**]. Make sure to save the results as a file for downstream processing, and use tabular (outfmt 6) to ensure the Python scripts can parse them correctly.

3. Map Reads onto the Assembly

NOTE: This method can be used if the reference dataset consists of raw genomic reads, rather than assembled sequences or Sanger sequences, in which case use BLAST (step 2.1).

3.1. Using BWA –MEM v. 0.7.12[19] or bowtie2[20], map the downloaded raw reads (raw_reads.fastq) onto the query assembly. The output will be .sam format. Commands are as follows: first index the assembly: [**bwa index your_assembly.fasta**], and then map the reads [**bwa mem your_assembly.fasta raw_reads.fastq >mapped.sam**]. (Note the '>' symbol here is not a greater-than sign; instead it instructs the output to go into the file mapped.sam).

4. Use Python Script to Remove any Matching Sequences

NOTE: Provided scripts work with Python 2.7.

4.1. Following Step 2, use subtractive Python script by using the command [**./Non-matching_sequences.py your_assembly.fasta BLAST_results.txt**]. Before running the script, ensure that the BLAST output file is in format 6 (tabular). The script will output a file with non-matching sequences in fasta format named your_assembly.fasta_non-matching_sequences_BLAST_results.txt.fasta and also the matching sequences for records, as your_assembly.fasta_matching_sequences_BLAST_results.txt.fasta. The non-matching file will be the most important, as a source of potential T sequences for testing and further cycles of subtractive genomics.

4.2. Following Step 3, run the Python script removeUnmapped.py to take as input the .sam from step 3.1, and identifies the names of query sequences without any matching reads and saves them to a new text file. Use the command [**./removeUnmapped.py mapped.sam**] and the output will be mapped.sam_contigs_with_no_reads.txt. (The program will generate a slimmed-down sam file with all unmapped reads removed; this file can be ignored for purposes of this protocol but may be useful for other analyses).

4.3. As the output of the previous step is a list of sequence names in a text file called mapped.sam_contigs_with_no_reads.txt, extract a fasta file with these sequences: [**./getContig.py your_assembly.fasta mapped.sam_contigs_with_no_reads.txt**]. The output will be a file called mapped.sam_contigs_with_no_reads.txt.fasta.

5. Design Primers for the Sequence that Remains

NOTE: At this point there is a fasta file containing candidate T sequences. This section describes qPCR to experimentally test whether they come from T or from previously unknown regions of R. If the subtraction in step 4 removed all sequences, then either the initial assembly failed to include T, or the subtraction may have been too stringent.

5.1. Use Geneious[21] to determine optimal primer sequences manually.

5.1.1. Highlight a candidate sequence of 21-28 bp for the Forward primer. Avoid runs of 4 or more of any base. Try to target a region with a fairly uniform combination of all basepairs. A single G or C at the 3' end is beneficial, helping to anchor the primer.

5.1.2. Click on the **Statistics** tab on the right-hand side of the screen to view that sequence's estimated melting temperature (Tm) as the candidate region is highlighted. Look to obtain a melting temperature between 55-60 °C, while avoiding repeats and long runs of G/C.

5.1.3. Follow steps 5.1.1. and 5.1.2 to choose a reverse primer, situated 150-250 base pairs 3' of the forward primer. While the primer lengths do not need to match, the predicted Tm should be as close as possible to the Tm of the forward primer. Be sure to reverse complement the sequence (if right-clicking in Geneious while the sequence is highlighted it is a menu option).

5.2. Use the **Primer Design** function, which is found in the top tool bar in the sequence window.

5.2.1. Click on the **Primer Design** button. Insert the region to amplify under **Target Region**.

5.2.2. Under the **Characteristics** tab, insert desired size, melting temperature (Tm), and %GC (see step 5.1.1.).

5.2.3. Click **OK** to have primers generated. Order the primers through a custom oligo service.

5.3. Validate primers with control DNA (encoding both T and R) to optimize Tm and extension time. Use regular Taq and gel electrophoresis to see the band size, but optimization can also be performed with qPCR following the methods in step 6.

5.3.1. Make 10X dilutions of both forward and reverse primers so that the primers have a concentration of 10 μM.

5.3.2. Use a PCR mix of 0.5 μL of dNTP, 0.5 μL of forward primer, 0.5 μL of reverse primer, 0.1 μL of Taq polymerase, 2 μL of template, 0.75 μL of magnesium, 2.5 μL of buffer, and 18.15 μL of water so that there is 25 μL per template with a concentration of 5 ng/μL.

5.3.3. Test the primers at different melting temperatures in the PCR program. Usually optimal performance is observed melt temperatures slightly below the predicted Tm of the primers, but not usually above 60 °C. Also test for optimal extension times using this guide: 1 min per 1000 bp (thus, usually 10-30 seconds depending on amplicon length).

5.3.4. Perform end-point gel electrophoresis to confirm that the primers amplify the expected sequence. Run 25 μL of the qPCR product mixed with 5 μL of 6X glycerol dye on a 2% TAE agarose gel at 200 V for 20 min.

6. *qPCR* Validation of the Remaining Sequence

NOTE: This step requires primers validated and PCR conditions established in step 5.

6.1. Run each template in triplicate with the following mix; 12.5 μL of PowerSYBR Green master mix, 0.5 μL of forward primer with a concentration of 10 μM, 0.5 μL of reverse primer with a concentration of 10 μM, 10.5 μL of water, and 1 μL of template DNA (at a concentration of 2 ng/μL), so that each well contains 25 μL of total volume.

6.2. Run a qPCR program informed by the validated temperature and extension time from step 4. We designed and validated all primers to be compatible with a two-stage cycle, 95 °C for 10 min initial melt, then 40 cycles of 95 °C for 30 s and 60 °C for 1 min. However, a three-stage (melt-anneal-extend) program may be more optimal for the primers and should be adapted if necessary. We recommend that final denaturing curves be generated at least the first time the primers are employed in qPCR to validate the amplification of a single DNA product.

6.3. Measure qPCR/SYBR Green signals relative to actin (or any other suitable 'R' control) by ΔCt. For all cases calculate the average and standard deviation of $2^{-(\text{gene Ct - }\beta\text{-actin Ct})}$.

6.4. (Optional) Perform end-point gel electrophoresis to confirm correct product size detection by qPCR. Here, run 25 μL of the qPCR product mixed with 5 μL of 6x glycerol dye on a 2% TAE agarose gel at 200 V for 20 min.

7. *Repeat* with a New Reference to Pare Down the Data.

NOTE: If step 6 validated the identified sequences from T, end the cycle here (**Figure 2A**).

However, a variety of considerations may motivate a continuation of the cycle, for example if many R sequences remain in the file or if none of the candidate T sequences were validated by qPCR in step 6.

7.1. Obtain a new reference. This step enables a new iteration of the cycle and may include raw genomic data, raw RNA-seq data, or other assembled datasets. Valuable resources for reference data include the Genome database at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome) which stores assembled genomes accessible through FTP (ftp://ftp.ncbi.nlm.nih.gov/genomes/), and the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) where raw next-generation sequence reads are stored. Genome projects may provide their raw sequence data through other project-associated websites and databases.

REPRESENTATIVE RESULTS:

After running BLAST, the output file will have a list of sequences from the query that match the database. After Python subtraction, a number of nonmatching sequences will be obtained, and tested by qPCR. The results of this, and next steps, are discussed below.

Negative result. There are two possible negative results that can be seen after BLAST to the reference sequence. There may be no BLAST results, meaning that the total sequence does not have any similar sequences to the reference. This may be an error in selecting the right reference sequence for the sample sequenced. Another possibility is that there are no unique sequences in the starting assembly (everything is subtracted away), therefore no genes are found for the sequence of interest. Check where the reference came from and ensure that it is not the same tissue as the query assembly.

After computational filtering, qPCR may yield a negative result, for examples see **Figure 3A, 3B, C** in which there was no difference in detection across bird tissues. Panels A through C are representative genes from different subtraction cycles, which motivated additional subtractive cycle iterations and the development of the method (**Figure 2A**, **2B**).

Positive result. A positive result--the identification of a true target sequence--is confirmed when genomic DNA qPCR shows statistically greater detection in the tissue / sample of interest relative to the reference (**Figure 3D**). The subtractive project in this case started with sequencing the RNA from germline tissue of male and female adult zebra finch, obtaining 10 million read pairs from each sex. For brevity, we will describe the processing of the ovary sequence only, in which 167,929 transcripts were obtained by *de novo* assembly. The subtractive genomics method (BLASTn) was used to eliminate any sequences that matched the published somatic genome[12], which left 5,060 transcripts corresponding to 598 unique proteins, indicating that many of the transcripts were noncoding. The Sanger raw reads used to generate the assembly were then used for the next level of subtraction by tBLASTn, yielding 78 proteins. One final subtraction was performed using RNA-seq raw reads from the auditory lobule[13], which left eight proteins. When these proteins were run through NCBI nr BLAST, six of the proteins were viral, one was a repetitive region in birds, and the last was an α-SNAP that is germline restricted[7] (**Figure 2B**). During this process, 935 somatic genes that were not previously included in the whole genome annotation were identified; several showed uniform qPCR amplification across tissues (**Figure 3A**, **3B, 3C**). The α-SNAP gene was validated to be germline restricted using qPCR, because it was depleted in somatic tissue relative to testis DNA where it was present at levels equivalent to actin (**Figure 3D**).

What could go wrong. The main problem that must be overcome when using this method is ensuring that the proper reference sequence is used. The best reference sequence encapsulates, in the broadest sense, the genomic complexity in which the sequence of interest (T) is embedded. This may mean that sequences in different forms; transcriptome, assembly, raw data, or data from multiple studies need to be used as references (**Figure 1**). In the zebra finch study, we developed primers from RNA sequencing data; however, the primers did not always work due to the presence of introns between or within primer binding sites in DNA. We tested each primer set by PCR off genomic DNA from testis DNA, which encodes both the target (T) and the reference (R), making it a suitable positive control. Primer failure at this stage necessitates the design and testing of new primers until a suitable set is identified. Standard pitfalls of PCR-based methods apply: amplification conditions must be optimized, amplification specificity confirmed by testing and/or cloning, and no-template controls must be included in all experiments. For more information on qPCR assays, see[22].

FIGURE AND TABLE LEGENDS:

Figure 1. The subtractive approach can iteratively remove multiple references (R) to recover only the target sequence of interest (T) from total genomic data. The reference sequences of individual projects may not overlap in precisely this way and may include datasets not indicated on the figure.

Figure 2. Visual methods. (A) Subtractive cycle schematic. The cycle can be iterated as many times as needed, each time utilizing distinct reference sequences, to obtain the best results. (B) Specific example of the subtractive cycle of steps carried out in Biederman *et al.*[7], with steps numbered as in A, and with the number of sequences remaining at each stage shown.

Figure 3. Example data of qPCR results including negative and positive outcomes. (A) Genomic DNA qPCR of CHD8, a negative outcome. (B) Genomic DNA qPCR of DNMT1, a negative outcome. (C) Genomic DNA qPCR of CHD7, a negative outcome. (D) Genomic DNA qPCR of NAPAG, confirming presence specifically in testis samples and depletion from liver and ovary relative to actin, a positive outcome. All panels indicate average +/- standard deviation of three measurements.

DISCUSSION:

While subtractive genomics is powerful, it is not a cookie-cutter approach, requiring customization at several key steps, and careful selection of reference sequences and test samples. If the query assembly is of poor quality, filtering steps might only isolate assembly artifacts. Therefore, it is important to thoroughly validate the *de novo* assembly using an appropriate validation protocol to the specific project. For RNA-seq, guidelines are provided on the Trinity website[18] and for DNA, a tool like REAPR[23] can be used. Another critical step when using BLAST is selection of appropriate e-value, which will determine whether the subtraction will be relaxed or stringent. However, an inversion occurs in the method: a more stringent match to reference is actually a less-stringent subtraction, as non-matching sequences are not subtracted. Therefore, a larger (less stringent) e-value should be used in BLAST for a more stringent subtraction. The final essential step of the protocol is reference selection. For greatest efficiency the reference should be as complete as possible; however, it does not need to be perfect because qPCR testing confirms whether remaining sequences are from T or R, and whether more filtering is necessary. During the implementation of the protocol, new references may be used to further narrow down the genes to be validated. We note that sometimes the matching method may change: for the last subtractive step we used the algorithm BWA to map raw reads onto the query sequences, and used custom python scripts to identify query sequences with no matching reads (**Figure 2B**).

Limitations of this method include availability of a reference sequence. For example, Meyer *et al.* evaluated the mitochondrial genome of a new hominin; they used human and Denisovan probes to capture mitochondrial DNA, which was sequenced and mapped to a human reference[24]. In this case, there were no existing nuclear genome reference data that the researchers could have subtracted against to obtain the mitochondrial genome, necessitating the read-mapping alternative strategy[24]. Any extensively diverged regions of the novel mitochondrion relative to the human mitochondrial reference would be lost by read-mapping. Subtractive genomics offers a less-biased approach than read-mapping but is not always applicable depending on the research question, and in this case the low levels of ancient DNA precluded the kind of sequence coverage required for *de novo* assembly (step 1 of subtractive genomics).

Physical purification provides another alternative method to subtractive genomics. Purification of DNA or RNA is often used in sequencing whole chloroplast and mitochondrial genomes because these organellar genomes are much smaller than nuclear genomes[25-28]. Human and other smaller mitochondrial genomes can be isolated for sequencing through amplification using two primer sets followed by purification[25]. However, subtractive genomics may be helpful for cases in which mitochondrial genomes are unusually large, the primer binding sites are divergent or will not result in the full genome. An example of this is in ciliates, which have large, divergent, linear mitochondrial genomes[29]. Mapping to a reference genome is not a viable option for ciliates due to high divergence across species and lack of homologs even across genuses[30]. By using subtractive genomics, the ciliate mitochondrial genome can be isolated and analyzed while minimizing the potential of missing segments of the genome. Similarly, while a *de novo* assembly approach was used in the Sitka spruce chloroplast genome assembly, gap-closing involved comparative read mapping against the white spruce, potentially introducing bias at these sites[31].

Depending on the project, subtractive genomics may offer time and cost advantages relative to purification or mapping approaches, while offering less bias in the discovery process. In some situations, the target sequence cannot be easily isolated because it is completely unknown, is vital to cell survival (mitochondria), or too large to separate by standard gel electrophoresis. Size-based electrophoretic purification is slow and requires significant starting material (which may be expensive) while optimizing conditions over multiple attempts. Pulse-field gel electrophoresis (PFGE) enables separation of DNA fragments up to 10^7 bp (10 Mb) but takes 2-3 days, large amounts of material, and sometimes specialized equipment that is not commercially available[32]. In Biederman *et al.*, the only sequence that was known from the germline-restricted chromosome was a noncoding repeat[7]**.** As this chromosome is the largest in the bird, over 100 Mb in length[10], purification would have been impossible; therefore, subtractive genomics was able to do what other methods could not. In the genomic era it is often cheaper and faster to sequence now, and filter by computer later. Enabling the discovery of completely novel sequences, subtractive genomics utilizes a combination of approaches to isolate novel sequences even without a perfect reference sequence.

ACKNOWLEDGMENTS:

The authors acknowledge Michelle Biederman, Alyssa Pedersen, and Colin J. Saldanha for their assistance with the zebra finch genomics project at various stages. We also acknowledge Evgeny Bisk for computing cluster system administration and NIH grant 1K22CA184297 (to J.R.B.) and NIH NS 042767 (to C.J.S).

DISCLOSURES:

The authors have nothing to disclose.

REFERENCES:

1. Barh, D., *et al.* A Novel Comparative Genomics Analysis for Common Drug and Vaccine Targets in Corynebacterium pseudotuberculosis and other CMN Group of Human Pathogens. *Chemical Biology & Drug Design*. **78** (1), 73-84 (2011).

2. Sarangi, A. N., Aggarwal, R., Rahman, Q., & Trivedi, N. Subtractive Genomics Approach for in Silico Identification and Characterization of Novel Drug Targets in Neisseria Meningitides Serogroup B. *Journal of Computer Science & Systems Biology*. **2** (5), doi:10.4172/jcsb.1000038 (2009).

3. Kaur, N., *et al.* Identification of Druggable Targets for Acinetobacter baumannii *Via* Subtractive Genomics and Plausible Inhibitors for MurA and MurB. *Applied Biochemistry and Biotechnology*. **171** (2), 417-436 (2013).

4. Rathi, B., Sarangi, A. N., & Trivedi, N. Genome subtraction for novel target definition in Salmonella typhi. *Bioinformation*. **4** (4), 143-150 (2009).

5.Epstein, J. H., *et al.* Identification of GBV-D, a Novel GB-like Flavivirus from Old World Frugivorous Bats (*Pteropus giganteus*) in Bangladesh. *PLoS Pathogens,* **6** (7). doi:10.1371/journal.ppat.1000972 (2010).

6. Kapoor, A., *et al.* Identification of Rodent Homologs of Hepatitis C Virus and Pegiviruses. *MBio, **4*** (2). doi:10.1128/mbio.00216-13 (2013).

7. Biederman, M. K., *et al.* R. Discovery of the First Germline-Restricted Gene by Subtractive Transcriptomic Analysis in the Zebra Finch, Taeniopygia guttata. *Current Biology*. **28** (10), 1620-1627 (2018).

8. Readhead, B., *et al.* Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus. *Neuron.* **99**, 1-19 (2018).

9. Carroll, D., *et al.* The global virome project. *Science*, **359** (6378), 872-874 (2016).

10. Pigozzi, M.I., Solari, A.J. Germ cell restriction and regular transmission of an accessory chromosome that mimics a sex body in the zebra finch. Taeniopygia guttata. *Chromosome Research*. **6**, 105–113 (1998).

11. Itoh, Y., Kampf, K., Pigozzi, M.I., and Arnold, A.P. Molecular cloning and characterization of the germline-restricted chromosome sequence in the zebra finch. *Chromosoma,* **118**, 527-536 (2009).

12. Warren, W.C., *et al.* The genome of a songbird. *Nature*. **464**, 757–762 (2010).

13. Balakrishnan, C.N., Lin, Y.C., London, S.E., and Clayton, D.F. RNAseq transcriptome analysis of male and female zebra finch cell lines. *Genomics*. **100**, 363–369 (2012).

14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30** (15), 2114-20 (2014).

15. Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. **30**, 614–620 (2014).

16. Yang, X., Dorman, K.S., and Aluru, S. Reptile: representative tiling for short read error correction. *Bioinformatic*s. **26**, 2526–2533 (2010).

17. MacManes M.D., Eisen M.B. Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ.,* **1** (113). doi:10.7717/peerj.113 (2013).

18. Grabherr, M.G., *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*. **29**, 644–652 (2011).

19. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]. (2013).

20. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9**. 357-359. (2012).

21. Kearse, M., *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28** (12), 1647-1649 (2012).

22. Peirson SN, Butler JN. Quantitative polymerase chain reaction. *Methods in Molecular Biology,* **362**, 349-62 (2007).

23. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR citation: REAPR: a universal tool for genome assembly evaluation. *Genome Biology*. **14** (5). doi:10.1186/gb-2013-14-5-r47 (2013).

24. Meyer, M., *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*. **505** (7483), 403-406 (2013).

25. Gunnarsdóttir, E. D., Li, M., Bauchet, M., Finstermeier, K., & Stoneking, M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Research. **21***(1), 1-11 (2010).

26. King, J. L., *et al.* High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Science International: Genetics. **12***, 128-135 (2014).

27. Yao, X., *et al.* The First Complete Chloroplast Genome Sequences in Actinidiaceae: Genome Structure and Comparative Analysis. *Plos One.* **10** (6), doi:10.1371/journal.pone.0129347 (2015).

28. Zhang, Y., *et al.* The Complete Chloroplast Genome Sequences of Five Epimedium Species: Lights into Phylogenetic and Taxonomic Analyses. *Frontiers in Plant Science.* ***7***, doi:10.3389/fpls.2016.00306 (2016).

29. Swart, E. C., *et al.* The Oxytricha trifallax Mitochondrial Genome. *Genome Biologyyogy and Evolution.* **4** (2), 136-154. doi:10.1093/gbe/evr136 (2011).

30. Barth, D., & Berendonk, T. U. The mitochondrial genome sequence of the ciliate Paramecium caudatum reveals a shift in nucleotide composition and codon usage within the genus Paramecium. *BMC Genomics.* **12** (1). doi:10.1186/1471-2164-12-272 (2011).

31. Coombe, L. *et al.* Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data. *Plos One.* **11** (9), doi:10.1371/journal.pone.0163059 (2016).

32. Herschleb J, Ananiev G, Schwartz DC. Pulsed-field gel electrophoresis. *Nature Protocols*., **2** (3), 677-84 (2007).

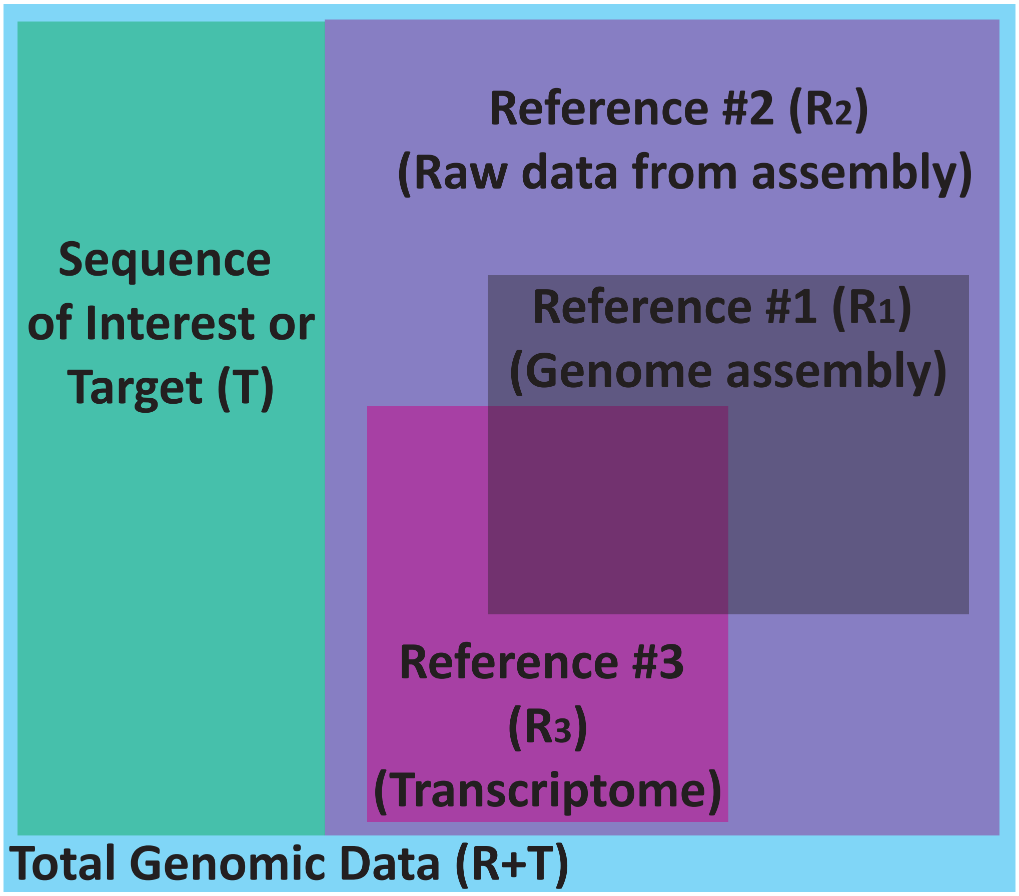

Sequence of Interest or Target (T)
Reference #2 (R_2) (Raw data from assembly)
Reference #1 (R_1) (Genome assembly)
Reference #3 (R_3) (Transcriptome)
Total Genomic Data (R+T)

Figure 2

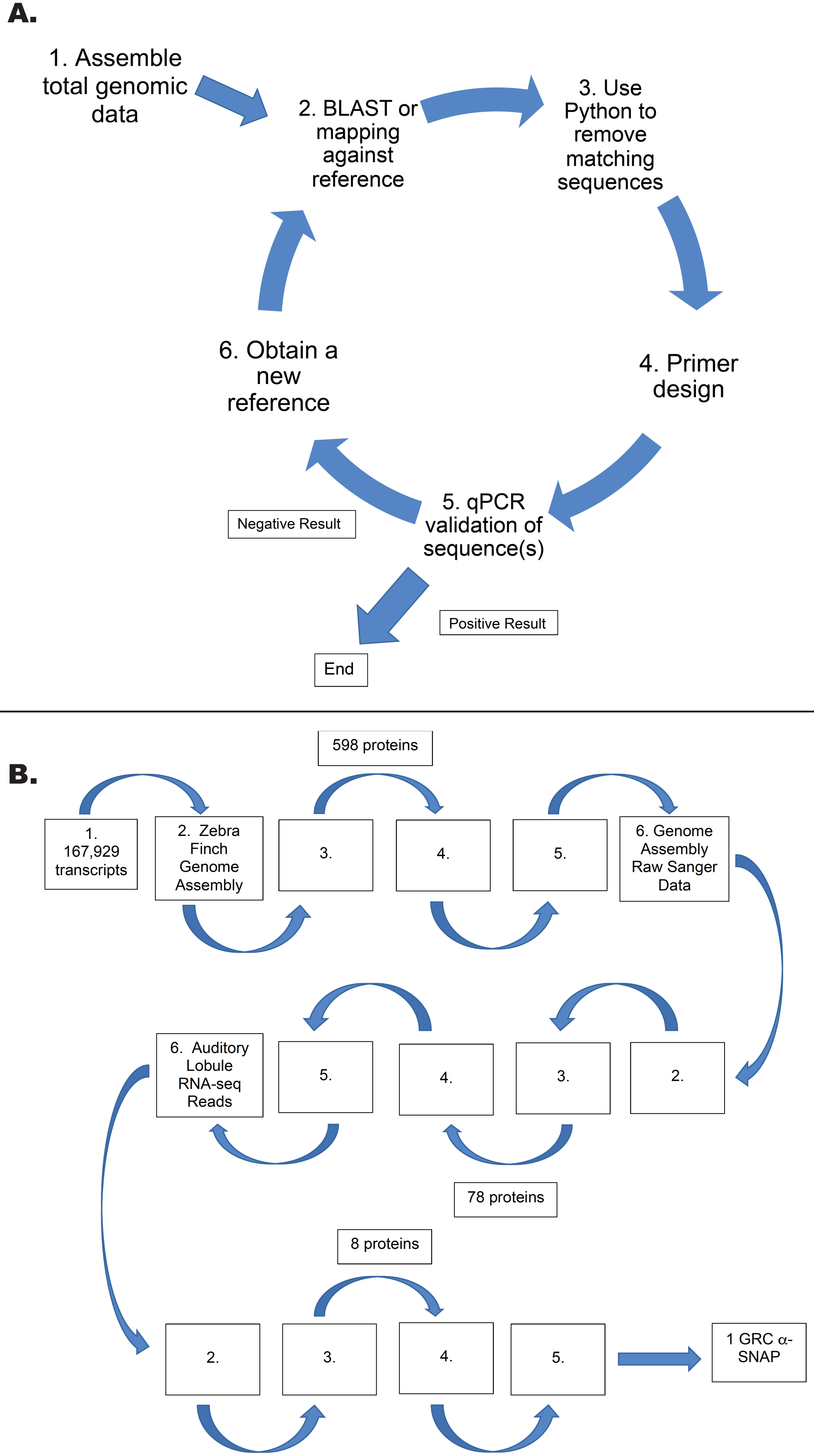
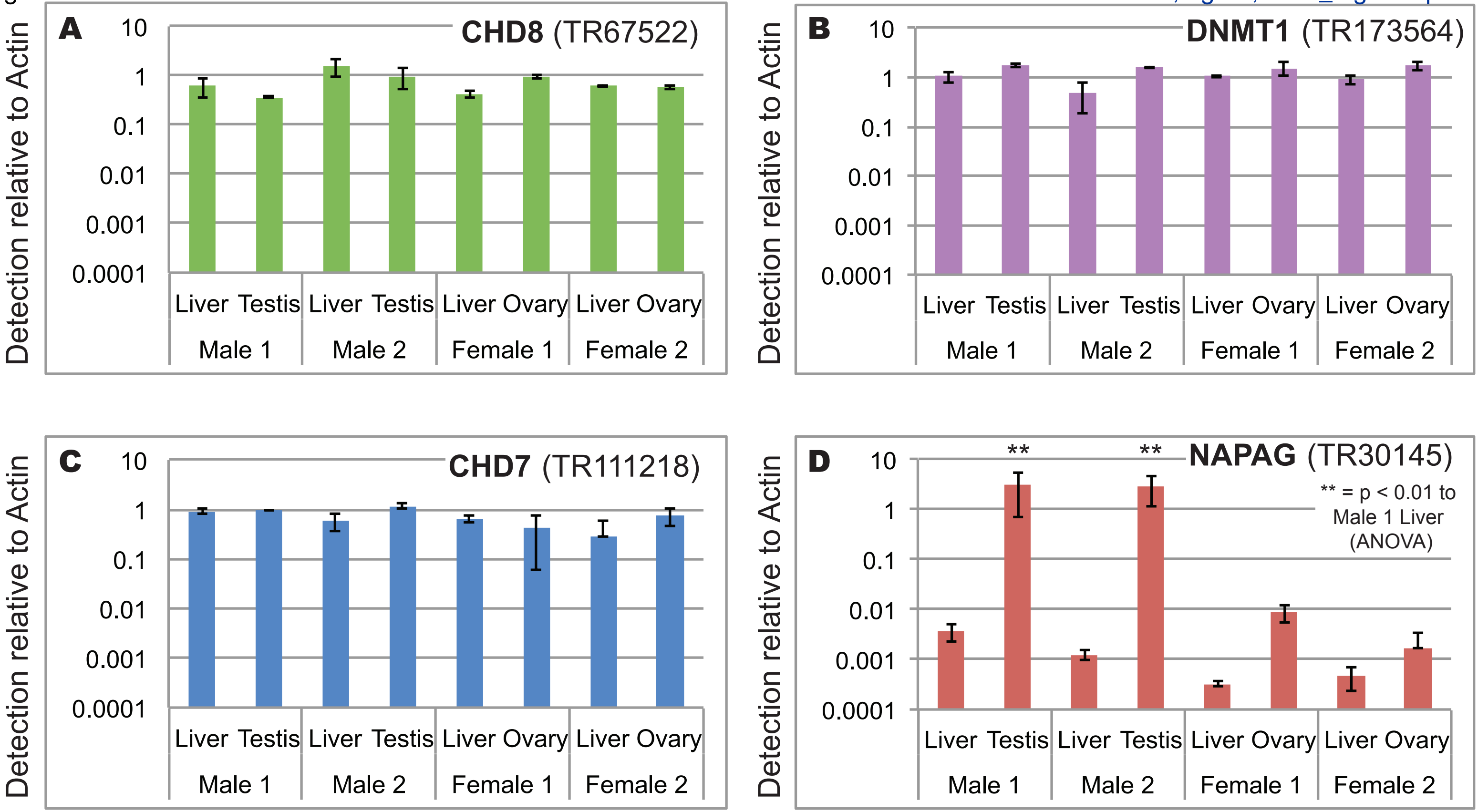

A.
1. Assemble total genomic data
2. BLAST or mapping against reference
3. Use Python to remove matching sequences
4. Primer design
5. qPCR validation of sequence(s)
6. Obtain a new reference
Negative Result
Positive Result
End
B.
598 proteins
1. 167,929 transcripts
2. Zebra Finch Genome Assembly
3.
4.
5.
6. Genome Assembly Raw Sanger Data
2.
3.
4.
5.
6. Auditory Lobule RNA-seq Reads
78 proteins
8 proteins
2.
3.
4.
5.
1 GRC α-SNAP

Figure 3


A
CHD8 (TR67522)
B
DNMT1 (TR173564)
C
CHD7 (TR111218)
D
NAPAG (TR30145)
**
**
** = p < 0.01 to Male 1 Liver (ANOVA)
Detection relative to Actin
10
1
0.1
0.01
0.001
0.0001
Liver
Testis
Liver
Testis
Liver
Ovary
Liver
Ovary
Male 1
Male 2
Female 1
Female 2

Name of Material/ Equipment	Company	Catalog Number	Comments/Description
Accustart II Taq DNA Polymerase	Quanta Bio	95141	
Blasic Local Alignment Search Tool (BLAST)			https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Assembly-Quality-Assessment
Bowtie 2			https://www.python.org/download/releases/2.7/
BWA-MEM v. 0.7.12			https://github.com/BenLangmead/bowtie2
Geneious			https://blast.ncbi.nlm.nih.gov/Blast.cgi
PEAR v. 0.9.6			http://www.mybiosoftware.com/reptile-1-1-short-read-error-correction.html
Personal Computer	Biomatters		http://www.geneious.com/
PowerSYBR qPCR mix	ThermoFisher	4367659	
Python v. 2.7			https://sco.h-its.org/exelixis/web/software/pear/
Reptile v.1.1			https://alurulab.cc.gatech.edu/reptile
Stratagene Mx3005P	Agilent Technologies	401456	
TransDecoder v. 3.0.1			https://sourceforge.net/projects/bio-bwa/files/
Trinity v. 2.4.0			https://github.com/TransDecoder/TransDecoder/wiki

jove
JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:	Novel sequence discovery by subtractive genomics
Author(s):	Kathryn Asalone, Megan Nelson, John Bracht

Item 1: The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via:

☒ Standard Access ☐ Open Access

Item 2: Please select one of the following items:

☒ The Author is **NOT** a United States government employee.

☐ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and(c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

CORRESPONDING AUTHOR

Name: John Bracht

Department: Biology

Institution: American University

Title: Assistant Professor

Signature: John Bracht Date: 7/31/18

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140

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.

Supplemental Coding Files