MetaPlex: An Ion Torrent COI metabarcoding workflow and toolkit to increase experimental efficacy and efficiency

Nick Gabry[1], Jeff Kinne[2], and Rusty Gonser[1]

[1]Indiana State University Department of Biology
[2]Indiana State University

September 15, 2022

Abstract

DNA barcoding has become a dependable way to assign taxonomic identifications to otherwise unknown DNA samples. Metabarcoding using next-generation sequencing (NGS) maintains this same caliber of taxonomic identification and can occur on a mass scale in pooled, or multiplexed, samples from a wide variety of sources when properly multiplexed. However, this methodology has several pitfalls, including inaccurate source sample assignment, multi-step workflows that leave room for human error, and high costs. To combat these issues, library preparation protocols have been established for the popular Illumina sequencing platform which reduce sequence assignment errors to levels equivalent to background noise, while also greatly reducing cost. An equivalent protocol has yet to be established for the Ion Torrent sequencing ecosystem. Here, we present MetaPlex, a library preparation workflow and post-processing toolkit for efficient and accurate COI metabarcoding on Ion Torrent sequencers. These methods significantly decrease the costs of multiplexed sample sequencing by nearly 9x when compared to commercially available kits through the reduction of reagents needed for library preparation. In addition, our workflow provides increased reliability through elimination of laboratory processes which are known to negatively bias NGS outputs and source-sample tracing, limiting the common pitfall known as index jumps to 0.17%. Finally, our accompanying bioinformatic toolkit increases user accessibility by providing an easy-to-use command line tool for processing and correcting for errors in our, and others', dual-indexed reads.

Title

MetaPlex: An Ion Torrent COI metabarcoding workflow and toolkit to increase experimental efficacy and efficiency

Running Title

MetaPlex: A COI metabarcoding workflow and toolkit

Authors

Nick Gabry[1,2,3], Jeff Kinne[2,3] & Rusty A. Gonser[1,2]

Affiliations

[1] Department of Biology, Indiana State University

[2] The Center for Genomic Advocacy, Indiana State University

[3] Department of Mathematics and Computer Science, Indiana State University

Corresponding authors

Rusty Gonser, rusty.gonser@indstate.edu

Nick Gabry, ngabry@sycamores.indstate.edu

Abstract

DNA barcoding has become a dependable way to assign taxonomic identifications to otherwise unknown DNA samples. Metabarcoding using next-generation sequencing (NGS) maintains this same caliber of taxonomic identification and can occur on a mass scale in pooled, or multiplexed, samples from a wide variety of sources when properly multiplexed. However, this methodology has several pitfalls, including inaccurate source sample assignment, multi-step workflows that leave room for human error, and high costs. To combat these issues, library preparation protocols have been established for the popular Illumina sequencing platform which reduce sequence assignment errors to levels equivalent to background noise, while also greatly reducing cost. An equivalent protocol has yet to be established for the Ion Torrent sequencing ecosystem. Here, we present MetaPlex, a library preparation workflow and post-processing toolkit for efficient and accurate COI metabarcoding on Ion Torrent sequencers. These methods significantly decrease the costs of multiplexed sample sequencing by nearly 9x when compared to commercially available kits through the reduction of reagents needed for library preparation. In addition, our workflow provides increased reliability through elimination of laboratory processes which are known to negatively bias NGS outputs and source-sample tracing, limiting the common pitfall known as index jumps to 0.17%. Finally, our accompanying bioinformatic toolkit increases user accessibility by providing an easy-to-use command line tool for processing and correcting for errors in our, and others', dual-indexed reads.

Keywords

Ion Torrent, Metabarcoding, Amplicon sequencing, Multiplex, Index jumps

Introduction

Advancements in sequencing and computational technology have allowed DNA barcoding to become an efficient way to assign taxonomic identifications to unknown DNA samples . The process of metabarcoding using high-throughput sequencers allows for identification to occur on a mass scale in multiplexed samples from a wide variety of sources . This is typically achieved by following the standardized workflow of indiscriminate DNA extraction, targeted PCR amplification of a known barcode region to create amplicons, library preparation of these amplicons, high-throughput sequencing of these amplicons, and computational taxonomic identification of the resulting sequences. With the high costs of a single sequencing run, adding unique DNA index tags is a way to reduce costs by pooling multiple samples per sequencing run and still allowing for sample-source tracing . However, despite the promises of this method, imperfect and overcomplicated experimental designs are still often used which introduce errors into sequencing outputs and add extra costs to project budgets.

One prominent error that occurs when uniquely indexing samples is tag switching, or index jumping. Index jumping is the process of individual index tags cleaving and annealing to alternate sequences which leads to the improper assignment of sequences to source samples. This has been found to commonly occur in a range of 2-49% of sequenced reads when using standard multiplexed library preparation protocols . Two steps in the standard library creation process have been identified as the cause of most index jumping: blunt-end repair, and post-ligation PCR of amplicons . One method which increases chance of detection of index switches is the use of dual-indexes, as opposed to a single index, which entails placing the same identifying index on both the forward and reverse end of amplicons (i.e. F1-R1, F2-R2, etc.). If this method is used, then any amplicon which contains non-matching indices is known to have experienced an index jump and can be filtered out of the data. When using expensive commercial kits, however, this method still requires blunt-end repair and post-ligation PCR, will fail to identify the occurrence of dual index jumps, and can be costly. This raises the apparent need for cost effective and error reducing third-party library preparation methodologies.

Thus far, the most efficient method has proven to be avoiding the blunt-end repair and post-ligation PCR steps altogether using a single-PCR library creation method . While an efficient protocol combining dual-indexing and single-PCR library creation (Tagsteady) has been developed for sequencing in the Illumina ecosystem , and a single-index single-PCR library creation exists for sequencing on Ion Torrent systems , a uniquely dual-indexed (i.e. F1-R2, F2-R3, etc.) single-PCR library protocol does not yet exist for sequencing on Ion Torrent systems. Here we present MetaPlex, a novel single-step PCR sequencing workflow designed for cytochrome oxidase 1 (COI) metabarcoding specifically for the Ion Torrent sequencing ecosystem. Using DNA from bulk insect collections, we show that our novel primers and sequencing workflow can be used to simplify the library preparation process in both complexity and cost, as well as increase overall read retainment through reduction of index jumps.

In addition to simplifying the library preparation workflow, we also address the need for simplifying the bioinformatic processing of dual-indexed reads with the MetaPlex toolkit. We identified QIIME2 as common microbiome analysis platform used widely across COI metabarcoding projects (. However, with QIIME2 there is currently no easy way to analyze dual-indexed reads without a greater knowledge of programming. Without a way of processing these reads in an accurate and streamlined fashion, the dual-indexed workflows which provide increased experimental accuracy may be seen as a hinderance to researchers. To greater increase the utility of the MetaPlex workflow, and other dual-indexing workflows, we also provide a free and open-source bioinformatic toolkit that integrates seamlessly with QIIME2. With the goal of increasing accessibility and user friendliness of COI metabarcoding, the MetaPlex workflow and toolkit provide all the necessary tools for a cost effective and efficient research pipeline complete with simplified library creation, read-processing, error detection, and sample filtering.

Materials and Methods

MetaPlex primer design

MetaPlex primers are uniquely indexed fusion-primers designed using the ANML primer pair as a base for the purpose of cheap and accurate multiplexed COI metabarcoding applications . The ANML primer pair (LCO1490 5'-GGTCAACAAATCATAAAGATATTGG-3' / CO1-CFMRa 5'-GGWACTAATCAATTTCCAAATCC-3') has been proven to have increased detection of arthropod taxa when applied in COI barcoding applications compared to other popular primers . The MetaPlex forward fusion-primer is 68bp and consists of an Ion Torrent 'A' sequencing adapter, the required TCAG 'key signal', a 10bp swappable index sequence, a 'GAT' spacer sequence, and the LCO1490 forward primer. The MetaPlex reverse fusion-primer is 59 bp and consists of an Ion Torrent 'trP1' sequencing adapter, a second 10bp swappable index sequence, a 'GAT' spacer sequence, and the CO1-CFMRa reverse primer. A full MetaPlex read is visualized in Figure 1. On both the forward and reverse primer, the 10bp swappable indexes are the same 10bp sequences used in the Ion Xpress Barcode Adapters 1-96 Kit. Primers were ordered from idtdna.com as 4 nmole Ultramer® DNA Oligos, were diluted to 10 µM in DNA-free H2O, and stored at -4°C upon arrival until use. A full list of MetaPlex primers used here is available in Supplementary File 1.

Library Preparation and Sequencing

To test our primers, DNA was first extracted from bulk arthropod communities using NucleoSpin DNA Insect kits (Macherey-Nagel). DNA extracts were amplified in triplicate on 96-well plates using unmodified reagent volume and concentrations per 15uL reactions from Jusino et al. (2019): 7.88 µl DNA-free H2O, 3 µl Green GoTaq 5x buffer (Promega), 0.12 µl of 20 mg/ml BSA (New England BioLabs), 0.3 µl of 10 mM dNTPs (Promega), 0.3 µl of each 10 µM primer, and 0.1 µl of 5u/µl GoTaq polymerase (Promega). Thermal Cycler conditions for amplification were as follows:

94°C for 60s;

5 cycles of: 94°C for 60 s, 45°C for 90 s, 72°C for 90s;

35 cycles of: 94°C for 60 s, 50°C for 90 s, 72°C for 60s;

72°C for 7 min

10°C hold

Once amplified, PCR products were verified on a 1% agarose gel ran at 90V for 30 minutes. Positive PCR products were then purified and size-selected for 300bp using AMPure magnetic bead-based clean-up (Beckman Coulter). Purified products were quantified using a Qubit 4 1xds DNA Assay (Fisher), after which were pooled equimolarly prior to being templated onto an Ion 530 Chip (Fisher) using the Ion Chef (Fisher). After templating, the amplicon libraries were sequenced using the Ion S5 Prime system. In total, 3 chips were sequenced on 3 separate runs with 79, 71, and 68 multiplexed samples.

Bioinformatic Toolkit

To easily process MetaPlex reads using the popular microbiome analysis platform QIIME2, we provide the MetaPlex toolkit compatible with UNIX-based machines. While the MetaPlex toolkit is designed to work seamlessly with the QIIME2 analysis platform, it also contains flexibility for managing any data in fastq formats for use in other pipelines which require tools outside of QIIME2. We have made this open-source toolkit accessible on GitHub, and it can be easily installed using the Anaconda package manager through the BioConda channel or using the Pip package manager. This toolkit is usable from the command line or as importable python modules and contains the following functionalities: 'remultiplexing', index jump calculating, per-sample frequency-based filtering, and all sample length-based filtering.

Remultiplexing

Ion Torrent sequencers produce single-end reads, as opposed to paired-end reads produced by Illumina sequencers. For proper analysis of multiplexed and indexed reads produced from single-end sequencers QIIME2 requires that the index appear at the immediate 5' end of the read. To achieve this format we have designed a process we call 'remultiplexing', which takes MetaPlex (i.e. dual-indexed) reads, trims the 5' and 3' ends of the reads past the specified indexes, and moves the 3' index to immediately follow the 5' index (Figure 2.). The remultiplexing process can start from either a raw unmapped bam file such as what is given by an Ion Torrent sequencer, or a fastq/fastq.gz file, and requires a sample map containing all the index tag sequences used in the sequencing pool (Supplementary File 2). Remultiplexing produces a single gzipped fastq containing all sequences where both a forward and reverse index were found, allowing for immediate import into QIIME2. While designed for single end sequencers, this method is also applicable to paired-end reads if they have been merged as is standard practice for metabarcoding workflows.

Index jump calculating

One major purpose of MetaPlex is to reduce index jumps and allow for correction when they do occur. For this, we provide a method of calculating the rate at which index jumps occur per a given sequencing run and estimate the total number of false reads within the total pool, as well as each individual sample within the pool. Though this calculation process is usable across all dual-indexed workflows, there are a few key requirements we feel must be met for it to be accurate, all of which are achieved through the MetaPlex library preparation workflow. First is the use of dual-indexed reads, or reads containing unique indexes on both the forward and reverse end. Second is the use of two or more dual-index pairs, i.e. F01-F02 and F11-F12. Thirdly, we highly suggest the use of a sample spike containing what we have termed 'calibrator tags', or the use of at least one forward and one reverse index pair exclusively with each other (Figure 3). While only a single pair of calibrator tags is recommended for increased index-jump calculation accuracy, we also provide support for specifying multiple calibrator tag pairs, or no calibrator tags, though this last option is not recommended.

Using a user defined sample map (Supplementary File 3a) the index jump rate of each forward and reverse calibrator tag is calculated by taking the number of false reads with the specified tag and dividing by the total number of that tag present in the pool. In the instance where F01 and R11 are used as calibrator tags (Figure 3), a false read is any read which has an F01 tag and not an R11 tag, or any read with an R11 tag and not an F01 tag. The jump rate for the F01 tag and R11 tag are then averaged as a best estimate for the overall rate at which any index jumps occur. Compared to calculating jump rate based off just the total false reads within a pool, with the use of calibrator tags we can detect every instance in which a tag jump occurred in regard to a single index tag. With this rate, we then calculate an estimate for the total number of false reads in the data set by multiplying the total read count by the jump rate. This is used to give an estimate of the overall percentage of true and false reads in the data set. Additionally, for more precise per-sample metrics, we take into account individual abundances of index tags within the pool to generate an estimate of false reads per sample (Supplementary File 3b). Ultimately, this tool generates a table with false read estimates both per-pool and per-sample for future filtering.

Sample filtering

Using the per-pool and per-sample metrics generated from our index jump calculator, we provide tools which can then be used to assist in quality control through per-pool or per-sample frequency filtering, as well as a length-based filtering method. The per-sample frequency-based filtering removes false reads from each sample based off the calculated expectancies either provided by the index jump calculator, or at user specified levels. The all-sample length-based filtering removes sequences below a length threshold from the pool.

Results

An average of 21,621,596 (SD = 895,568) usable reads were produced among our sequencing runs, and all runs produced reads with median read lengths of 308bp (68bp forward primer, 181bp COI barcode region, 59bp reverse primer). Using our calibrator tag approach, we calculated an average of 0.17% of reads as experiencing an index jump (SD = 0.17%). This percentage of reads which experienced index jumps using our MetaPlex workflow is on par with the average of 0.25% of index jumps which occurred using the dual-indexed 'Tagsteady' workflow devised for the Illumina platform (Figure 4.). Aside from the viability of this workflow, the cost of this workflow is of equal note. The total cost of non-overlapping library preparation reagents for creating 96 uniquely indexed samples along with a calibrator is approximately $1,292. This is a nearly 9x reduction in cost when compared to the Ion Xpress^{TM} Barcode Adapters 1-96 Kit (ThermoFisher Scientific, Catalog number: 4474517).

Discussion

In this study we set out to address two main issues in COI metabarcoding workflows on Ion Torrent sequencers - index jumps and high costs. To do so, we present the MetaPlex workflow and accompanying post-processing command line toolkit (Figure 5). In our use of this laboratory protocol and index jump detection algorithm, an average of 0.17% (SD: 0.17) of reads from our sequencing outputs contained index jumps. Previous studies done on other sequencing platforms with a similar dual-indexed approach have detected a similar average precent of index jumps . These studies concluded that these rates are likely attributed to unavoidable sequencer errors and background contamination. The low rate of index jumps that we achieved were through our use of specially designed fusion-primers with swappable 10bp index tags on both the forward and reverse ends. In designing the dual-indexed primers with Ion Torrent sequencing adapters, we create a single-PCR library preparation process that eliminates the two steps which are most responsible for index jumps: blunt-end repair and post-ligation PCR of amplicons. Importantly, this method also greatly decreases the total cost of experimentation. For comparison, the current cost of an Ion Xpress^{TM}Barcode Adapters 1-96 Kit (ThermoFisher Scientific, Catalog number: 4474517), which is used for creating 96 single-indexed libraries, is $11,270, while creating 100 unique dual-indexed libraries with MetaPlex primers costs an estimated $1,184. Other methods have used a similar approach in creating Ion Torrent compatible fusion primers but have only included indexes on the forward primer . Comparatively, 96 uniquely indexed libraries with this method would cost an estimated $5,610, while also failing to reduce index jumps (Table 1.). For these reasons, we feel the MetaPlex workflow should be adopted for most COI metabarcoding applications to reduce both costs and source-sample tracing errors.

Apart from MetaPlex being designed for a sequencing platform where an equivalent workflow did not yet exist, our accompanying command-line toolkit increases the utility of our workflow.. All functionalities of the toolkit – remultiplexing, index jump calculating, frequency-based filtering, and length-based filtering – can be applied if proper formatting is followed as explained on the MetaPlex manual page (https://github.com/NGabry/MetaPlex). Of note is our novel approach to detecting and estimating index jumps. This calculation method is still a conservative measure, as it doesn't take into account indexes which jump but don't actually change the sample assignment, such as an R11 index jumping from an F01-R11 sample to another F0-1R11 sample, nor does it account for the potential for indexes to jump at differential rates. This fact is of greater importance if our index jump calculation tool is used on non-MetaPlex reads, as the workflow is inherently designed to eliminate workflow process which introduce the possibility of index jumps. Ultimately though, our toolkit can be used to process not only MetaPlex reads, but other dual-indexed reads as long as they're merged into a single read.

In summation, this novel workflow significantly increases sequencing efficacy through elimination of workflow processes which are known to bias NGS outputs and source-sample tracing, as well as increases sequencing efficiency through elimination of need for costly commercial library preparation kits. With this research, we provide an easily accessible and replicable way of producing reliable multiplexed sample sequencing libraries and provide a seamless way to post-process these reads in popular microbiome analysis platform QIIME2.

Acknowledgements

This work was supported by grants from the Northern New York Audubon, American Ornithological Society, and Indiana State University to N. Gabry, funding provided through The Center for Genomic Advocacy at Indiana State University to R.Gonser, and NIH grant 1R25MD011712-01 to R. Gonser and J. Kinne (Big Data 4 ISU).

References

Data Availability Statement

Sequencing data sample set and library preparation information can be found in the open-source project repository, https://github.com/NGabry/MetaPlex . Full sequencing outputs will be made publicly available on Dryad.

ORCID

Nick Gabry https://orcid.org/0000-0002-2015-4483

Author Contributions

N.G. conceived project idea, acquired funding, performed experiments, designed, and programmed toolkit, analyzed data, and wrote the manuscript. R.G. helped with experimental design and project funding. J.K. provided input on toolkit design and revision.

Conflict of Interest

None.

Tables and Figures

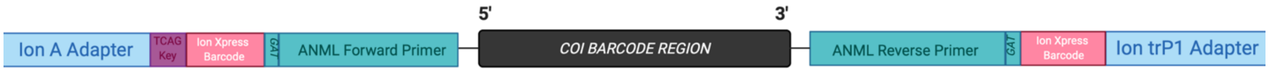


Figure 1. Visualization of a full MetaPlex read forward fusion primer on the 5' end, the target COI Barcode region in black, and the reverse fusion-primer on the 3' end.

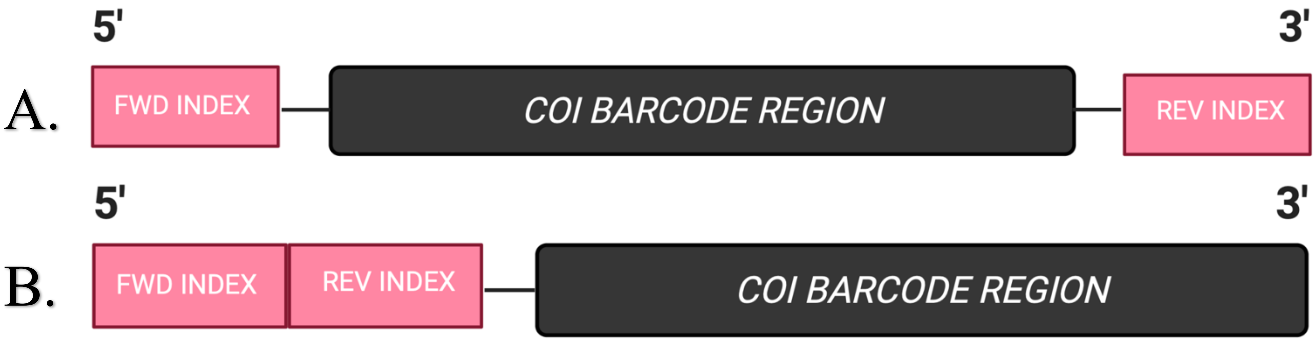


Figure 2. Visualization of MetaPlex read before (A) and after (B) the remultiplexing process, which turns a dual-indexed read into a single combined index on the 5' end for single-end processing in QIIME2.

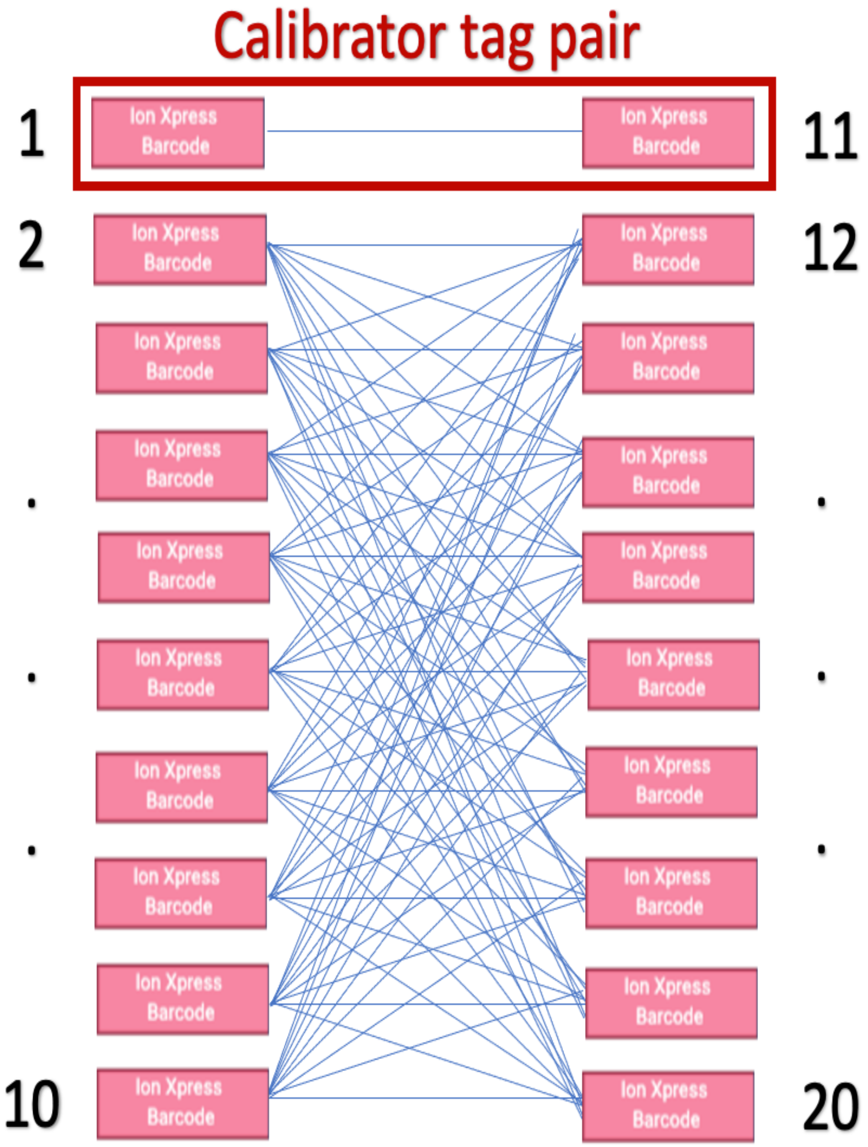


Figure 3. Visualization of a set of 20 MetaPlex primers with 2 reserved as calibrator tags, and the remaining 18 being used to create 81 unique combinations of dual-indexed pairs.

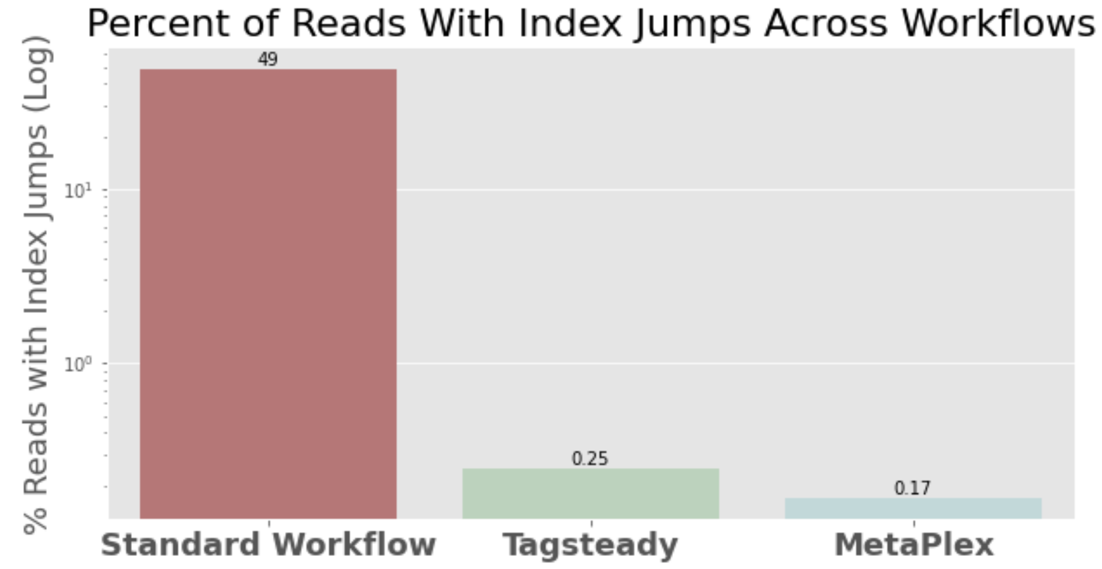
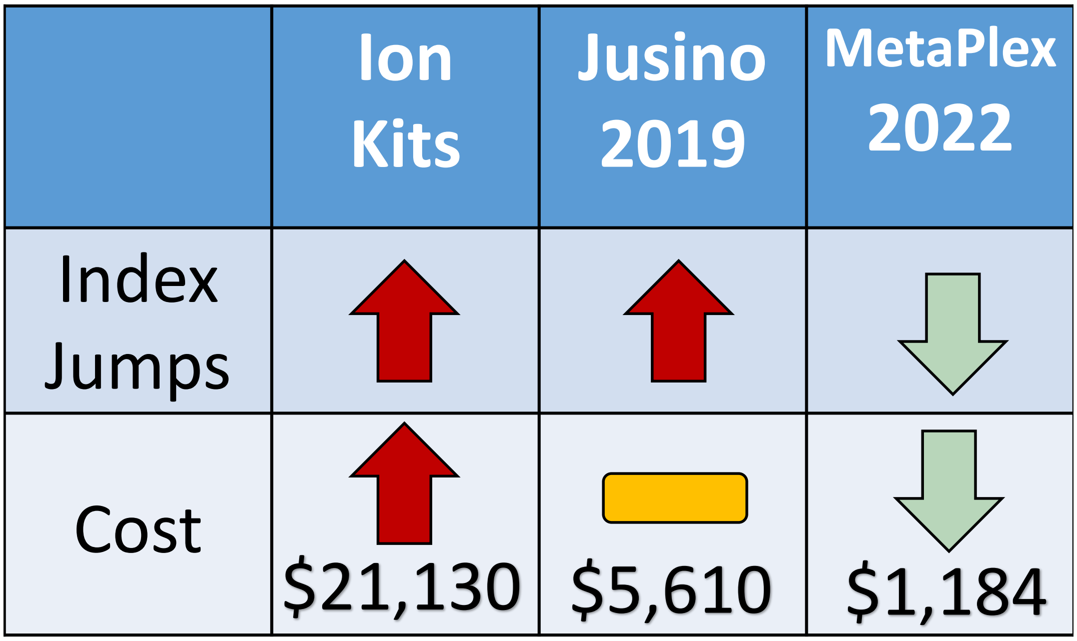


Figure 4. Bar plot showing how the MetaPlex protocol compares to the Illumina Tagsteady protocol, and the high-end of a standard Ion Torrent/Illumina library kit in terms of percentage of reads with index jumps.



	Ion Kits	Jusino 2019	MetaPlex 2022
Index Jumps	⬆	⬆	⬇
Cost	⬆ $21,130	▬ $5,610	⬇ $1,184

Table 1. An overview of the efficacy (gauged by amount of index jumps), and efficiency (gauged by cost) of three Ion Torrent indexed metabarcoding workflows.

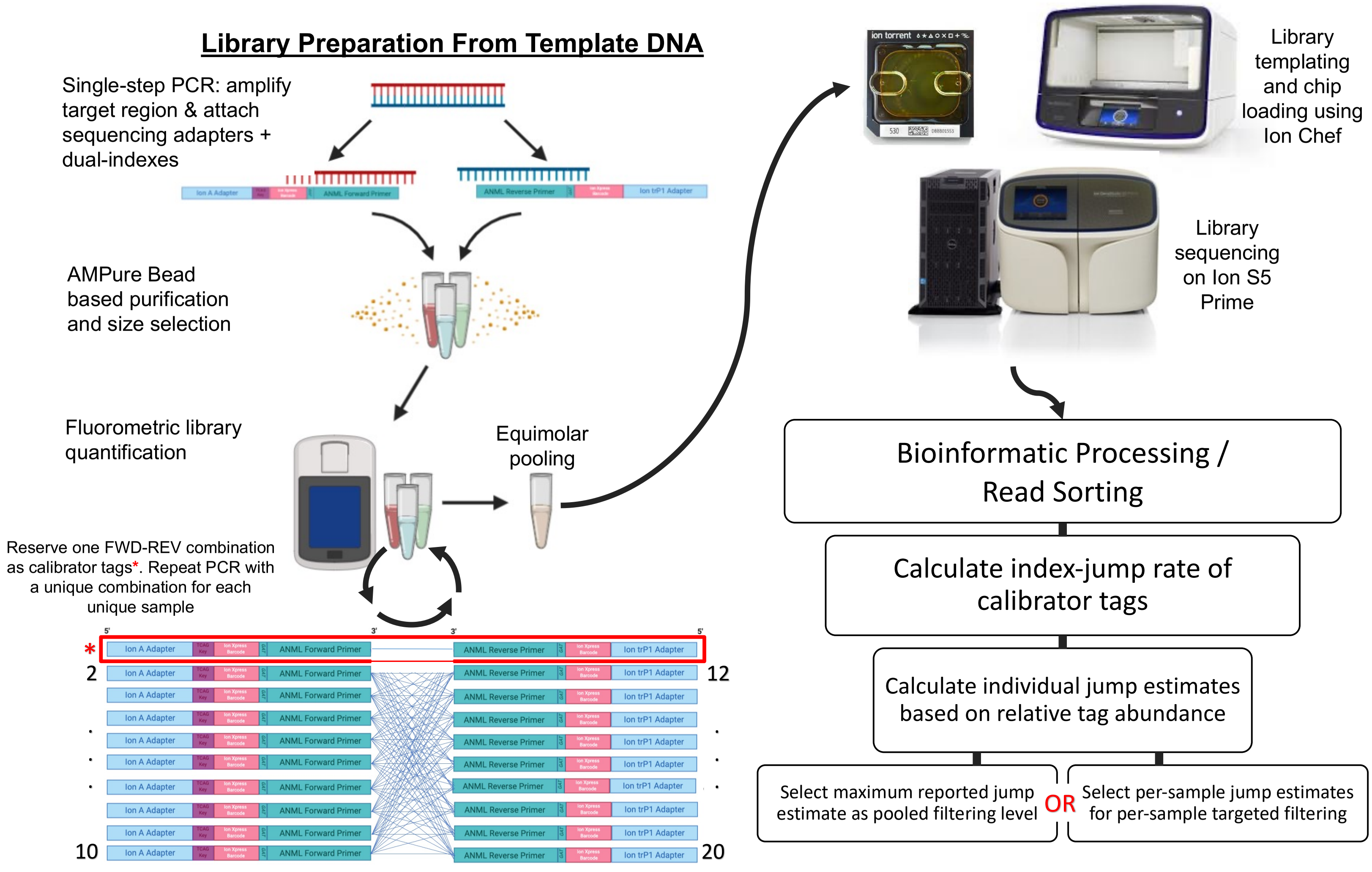

Library Preparation From Template DNA
Single-step PCR: amplify target region & attach sequencing adapters + dual-indexes
Ion A Adapter
Ion Xpress Barcode
ANML Forward Primer
ANML Reverse Primer
Ion trP1 Adapter
AMPure Bead based purification and size selection
Fluorometric library quantification
Equimolar pooling
Reserve one FWD-REV combination as calibrator tags*. Repeat PCR with a unique combination for each unique sample
*
5'
3'
3'
5'
TCAG Key
GAT
2
10
12
20
ion torrent
530
Library templating and chip loading using Ion Chef
Library sequencing on Ion S5 Prime
Bioinformatic Processing / Read Sorting
Calculate index-jump rate of calibrator tags
Calculate individual jump estimates based on relative tag abundance
Select maximum reported jump estimate as pooled filtering level
OR
Select per-sample jump estimates for per-sample targeted filtering

Figure 5. Flowchart of the MetaPlex workflow, from library creation all the way through bioinformatic processing utilizing the open-source MetaPlex toolkit.