Optimising ddRAD sequencing for population genomic studies with ddgRADer

Aparna Lajmi*, Felix Glinka, Eyal Privman*

Department of Evolutionary and Environmental Biology, Institute of Evolution,
* Corresponding authors

Keywords: genomic sequencing, reduced representation genome sequencing, population genomics, RAD sequencing, size selection, user friendly, webtool

Running title: Optimising ddRADseq with ddgRADer

Abstract

Double-digest Restriction-site Associated DNA sequencing (ddRADseq) is widely used to generate genomic data for non-model organisms in evolutionary and ecological studies. Along with affordable paired-end sequencing, this method makes population genomic analysis more accessible. However, when designing a ddRADseq experiment multiple factors should be considered, which can be challenging for new users. The generated data often suffer from substantial read overlaps and adaptor contamination, severely reducing sequencing efficiency and at times affecting data quality. Here, we analyse diverse datasets from the literature and carry out controlled experiments to understand the effects of enzyme choice and size selection on sequencing efficiency. The empirical data reveal that size selection is imprecise and has limited efficacy. In certain scenarios, a substantial proportion of short fragments pass below the lower size-selection cutoff resulting in low sequencing efficiency. However, enzyme choice can considerably mitigate inadvertent inclusion of these shorter fragments. A simple model based on these experiments is implemented to predict the number of genomic fragments generated after digestion and size selection, number of SNPs

genotyped, number of samples that can be multiplexed, and the expected sequencing efficiency. We developed ddgRADer, a user-friendly webtool, to aid in designing ddRADseq experiments while optimising sequencing efficiency. Given user-defined study goals, ddgRADer recommends enzyme pairs and allows users to compare and choose enzymes and size-selection criteria. ddgRADer improves the accessibility and ease of designing ddRADseq experiments and increases the probability of success of the first population genomic study conducted in labs with no prior experience in genomics.
http://ddgrader.haifa.ac.il/

1. Introduction

Restriction-site Associated DNA sequencing (RADseq; Baird et al., 2008) is a widely used method to generate reduced representation genomic libraries. This approach dramatically reduces the genotyping cost per sample compared to whole genome sequencing, accommodating a larger sample size in a given budget. RADseq has been particularly beneficial in SNP discovery and genotyping of non-model organisms for population, ecological, evolutionary, and conservation genomic studies (Clugston et al., 2019; Davey et al., 2011; Eaton & Ree, 2013; Luikart et al., 2003; Wagner et al., 2013). The key advantage of this method is the flexibility it offers in the number of cut-site that can be targeted across the genome, through the choice of appropriate restriction enzymes. This allows for a wide range of applications, from demographic surveys and population structure analyses for guiding conservation efforts (for example Natesh et al., 2017; Zecherle et al., 2021), to high-resolution scans looking for the genomic basis of local adaptation (for example Magalhaes et al., 2021).

Double-digest RAD sequencing (ddRADseq) is one of the most commonly used variations of RADseq where genomic DNA is digested with two different restriction enzymes, generating

genomic fragments in a reproducible manner (Peterson et al., 2012). Typically, a combination of a rare- and a frequent-cutting enzyme is used. The rare cutter determines the number of fragments generated and the frequent cutter determines the average length of these fragments. While the general rule of thumb is that enzymes that recognise shorter target sequences cut more often, the GC content in the recognition sequence also affects the number of cut sites. When choosing enzymes for a ddRADseq experiment, the size and GC content of the genome, and the polymorphism expected in the study population should be taken into consideration as they determine the number of polymorphic loci that will be genotyped, the depth per locus, and hence the number of samples that can be multiplexed in a single lane of sequencing. Taking these numerous factors into account can make designing a ddRADseq study challenging for beginners and easily result in a suboptimal study design even for experienced researchers. This can lead to substantial waste of sequencing effort. Moreover, it can generate sequencing data that is ill-fit for the downstream analyses and study goals (Puritz et al., 2014). Given the wide applicability of this method, it is crucial to make designing ddRADseq experiments more accessible and aid researchers in optimal experimental design.

Two of the biggest challenges for using ddRADseq are lack of appropriate training in designing experiments and the overall high cost of the project compared to traditional approaches (such as Sanger sequencing or microsatellites genotyping), despite the relatively low per-sample cost (compared to whole genome sequencing). Given the large capacity of a single lane of Illumina sequencing, most projects tend to conduct only one sequencing run on all samples together, multiplexed in one or a few sequencing lanes. Thus, unlike Sanger sequencing experiments, most researchers cannot afford sequential improvements to the experimental design. These challenges are particularly consequential for researchers in developing regions, including the tropics where most of the biodiversity lies and conservation effort is needed. Several tools are available for processing of ddRADseq data and SNP

calling (for example Catchen etal 2011; Catchen etal 2013; Puritz etal 2014; Sovic etal 2015; Ravidran etal 2018; Eaten and Overcast, 2020). However, these applications are meaningful only when the sequence data are generated with an appropriate experimental design. Relatively fewer tools can help in designing and optimising the experimental design (but see Lepais & Weir, 2014; Mora-Márquez et al., 2017; Rivera-Colón et al., 2021). While these tools may help in enzyme choice and size-selection, they are based only on predictions generated in silico that may be unrealistic.

With the advent of Illumina paired-end sequencing, where reads are sequenced from both ends of a DNA fragment, twice the amount of data can be generated at a relatively low additional cost compared to single-end sequencing. Thus, paired-end sequencing is nowadays the common practice. When carrying out paired-end sequencing of 150bp reads, if the DNA fragment (insert size) is less than 300bp (2 read lengths), the ends of the reads overlap resulting in some bases being sequenced twice. If the insert size is less than 150bp (1 read length), the adaptors are inadvertently sequenced leading to adaptor contamination at the end of each read, in addition to complete read overlap. Any sequenced fragments shorter than two read lengths lead to considerable loss of sequencing effort apart from a dataset contaminated with adaptors . For instance, ddRADseq libraries for 117 ant samples were constructed in a previous study in our lab (Inbar et al., in prep) using the commonly used enzyme pair MseI/EcoRI and size selection by magnetic beads. Paired-end sequencing of these libraries resulted in 74% of the sequenced fragments that were shorter than two read lengths. As a result, 69% of sequenced bases were wasted, that is, sequencing efficiency of only 31%. While the data was adequate to address the question in hand, this massive waste of sequencing effort motivated us to conduct a methodological study to optimise our sequencing efforts.

In this study, we set out to optimise ddRADseq experimental design by minimising adaptor contamination and overlapping reads, and designed a user-friendly webtool to aid in designing ddRADseq experiments. To this end, a literature survey was first carried out to look at the common choices of enzyme combinations, the size-selection methods and cutoffs used in ddRADseq studies. Some of these datasets were re-analysed to quantify adaptor contamination and overlaps, and these empirical results were compared with the theoretical expectations from in silico digestion. We conducted a controlled experiment to test switching the frequent cutting enzyme to an enzyme with less cut sites, to increase the fragment sizes and hence reduce the waste and improve sequencing efficiency. This led to insights into the efficiency of size selection, allowing more realistic predictions for the implications of alternative enzyme choices and size selection cutoffs. Lastly, we developed a user-friendly webtool to aid in designing ddRADseq experiments that allows researchers to examine alternative enzyme combinations, predict adaptor contamination and read overlaps while incorporating observations from empirical studies, and aid in deciding size-selection cutoffs and the number of samples to be multiplexed given the requirements of the study goals. This tool will help researchers to maximise their ddRAD-sequencing efficiency and optimise their experimental design.

2. Materials and methods

2.1 Literature survey

A brief literature survey was first carried out to estimate the proportion of short fragments given the study organism, enzyme pair and size-selection cutoffs used. To keep the number of papers manageable, we limited our search to papers published in Molecular Ecology and Molecular Ecology Resources as these journals routinely publish studies involving ddRADseq. The keywords “ddRAD” and “paired-end” were used for the search using the

software Publish or Perish (PoP; Harzing, 2007), which allowed us to get the data in a tabular format. The enzymes used, size-selection method and cutoffs, the stage in which size selection was performed (before PCR or after PCR) and read length were noted. We then downloaded the raw sequencing data from NCBI and re-analysed them for these studies whenever possible. Most of the published datasets could not be used as submitted datasets were already trimmed, and therefore were not suitable for examining adaptor contamination. We analysed five datasets (Combs et al., 2018; Fritz et al., 2018; Ryan et al., 2017; Termignoni-García et al., 2017; Trense et al., 2021) where size-selection was carried out using Pippin Prep or Bluepippin before the PCR step. We used FastQC (Andrews, 2010) to estimate the level of adaptor contamination in each of the datasets. Adaptors were then trimmed using Trimmomatic (Bolger et al., 2014) with the default settings, which are to identify a minimum of 8bp adaptor sequence. Sequences were not filtered or trimmed for low quality. FLASH (Magoč & Salzberg, 2011) was used to combine the trimmed overlapping paired-reads and get a fragment size distribution of the combined reads. We obtained the theoretically predicted fragment size distribution from the in silico digestion of the closest available reference genome using a custom python script. To investigate the relationship between in silico fragment size distribution and the empirical one, we conducted a linear regression on the five datasets as well as the data from our controlled experiment (described below). We looked at the relationship between the in silico predicted proportion of fragments with adaptor contamination (between 1bp – and one read length) and fragments with overlapping reads (between 1bp and two read lengths or the lower –size selection cutoff) versus the empirically observed proportions of fragments in these ranges (based on the FLASH results).

2.2 Controlled experiment to study the effect of enzyme choice

In silico digestion and size-selection cutoffs

We conducted a controlled experiment to understand the effect of switching enzymes on the proportion of short fragments and consequently on sequencing efficiency. The *Camponotus floridanus* reference genome (NCBI accession QANI01000000) was in silico digested to compare the fragment size distributions generated by either pair of enzymes using a custom python script as described above.

To estimate the number of genomic sites that were expected to be sequenced at different size-selection cutoffs, taking into account the waste due to adaptor contamination and read overlaps, an excel sheet calculator was created (Supplementary file 1). Additionally, we estimated the number of samples that could be multiplexed with a given yield of an Illumina sequencing lane and depth requirement of the study. These calculations were subsequently implemented in the webtool ddgRADer described below.

Sampling and library construction

Samples of 125 male *Camponotus fellah* were collected from Rehovot, Israel (31°54'32.1"N 34°48'45.7"E) from a single nest and stored at -80C. DNA was extracted from the abdomen of the ants using the Qiagen DNeasy blood and tissue kit followed by ethanol precipitation of the DNA extracts. Genomic libraries were constructed following Brelsford et al., (2016), a modified double-digest Restriction-site Associated DNA sequencing (ddRADseq) protocol based on Parchman et al. (2012) and Peterson et al. (2012) (Supplementary file 2). To understand the effect of enzyme choice on the resulting fragment size distribution and the proportion of waste due to short fragments, we compared libraries built using different restriction enzyme pairs. One way to switch enzymes where the same Illumina sequencing

adaptors can be used, is to switch to an enzyme that generates the same 3' overhang. One set of libraries was constructed using the commonly used combination of EcoRI and MseI. EcoRI is a 6-base rare cutter (recognition sequence 5′-G|AATTC-3′) while MseI is a 4-base frequent cutter (recognition sequence 5′-T|TAA-3′). Given that the recognition site for MseI is GC poor, for a genome with <50% GC content, MseI cuts frequently, generating a large proportion of small fragments. BfaI is a 4-base frequent cutter generating an overhang identical to MseI, yet its recognition site contains more GC than MseI (recognition sequence 5′-C|TAG-3′). Therefore, the GC-poor genome of *C. fellah* contains less BfaI than MseI sites, which should increase the fragment sizes and improve sequencing efficiency (in general, ant genomes contain between 30-50% GC (Simola et al., 2013), and the genome of *C. floridanus* has only 33% GC).

Briefly, genomic DNA was digested using two restriction enzymes and adaptors containing unique barcodes were ligated to each sample for multiplexing (using CutSmart buffer® NEB). The ligated products were PCR amplified in replicates of four using Q5 Hot Start Polymerase® NEB for 20 cycles with a starting DNA volume increased to 7ul per reaction. Additionally, primers and dNTPs were added to a final thermal cycle step to reduce single-stranded or heteroduplex PCR products. All the samples for each pair of enzymes were pooled and the two sets of libraries were size selected using Bluepippin (Sage Science, Beverly, MA, USA). 125 samples were used for MseI/EcoRI and 117 of them were also used for building BfaI/EcoRI libraries. The BfaI/EcoRI libraries were selected for a range of 330 to 700bp and MseI/EcoRI for a range of 300bp to 700 bp (this size includes 128bp of the adaptors and primers). The rationale behind choosing different size-selection ranges are discussed in detail in the Results and discussion below. The libraries produced with the two sets of enzymes were sequenced with paired-end 150bp reads on a single lane of an illumina Hiseq X sequencer to generate 340 million paired end reads (Accession: xxxxxx).

2.3 ddRADseq data analyses

The quality of the raw sequence data was examined using FastQC. The mean phred quality score was 37.9 and more than 90% of the reads had a mean score >= 30. The total number of paired-reads in the BfaI/EcoRI raw data were 143,893,548 (117 samples) while in MseI/EcoRI were 143,444,471 (125 samples). We removed the excess samples from the MseI/EcoRI dataset to have a comparable set of samples in the two datasets. The BfaI raw data was then subsampled (randomly excluded reads) to have a comparable number of reads in both the datasets. These two datasets were then analysed similar to literature survey datasets using trimmomatic and FLASH. The command process_radtags.pl from the program stacks2 (Rochette et al., 2019) was used to demultiplex the data. The resulting reads were mapped to the *Camponotus floridanus* genome (NCBI accession number: GCA_003227725.1 (Shields et al., 2018)) using *Bowtie2* (Langmead B & Salzberg SL, 2012). Using Ref_map.pl from stacks2 the mapped sequences were further processed to call SNPs. The details on the number of reads at each step are given in supplementary file 3, Table 1.

3. Results and discussion

Adaptor contamination and read overlap are major factors reducing the efficiency of sequencing. For example, Inbar et al. (in prep) constructed ddRADseq libraries using the ant *Cataglyphis niger* samples, with the enzymes MseI/EcoRI. The results contained overlaps between 74% of read pairs (fragments shorter than two read lengths) and 16% of the data had adaptor contamination (fragments shorter than one read length; see Fig. 1a). The fact that we only see adaptor contamination in 16% was due to the bead cleanup step in the library construction protocol, which excludes a significant proportion of short fragments. And yet, 69%of the sequencing output in this study was wasted on adaptor contamination and read overlaps. Many researchers try to deal with this issue using a more precise size

selection method, such as manually cutting fragments of a certain size from the gel or using instruments like Pippin Prep, to exclude short fragments from the library.

A literature survey was carried out to understand the range of size selection criteria and commonly used enzymes in ddRADseq protocols. We found a total of 66 studies from the year 2015 - 2021 where ddRADseq data were generated using paired-end sequencing in the journals Molecular Ecology and Molecular Ecology Resources (Supplementary File 4). Of these, 10 studies used only magnetic beads for size selection, four studies used manual size selection, 38 used some form of automated gel electrophoresis-based size selection such as Pippin Prep, and 14 studies did not report the size selection method used. EcoRI/MspI was the most used enzyme pair (10), followed by EcoRI/SphI (8), MluCI/NlaIII (6), SbfI/MseI (6), SbfI/MspI (6), and SphI/MluCI (5) (Supplementary file 3, table 2).

We then conducted meta-analyses, reanalysing and comparing five ddRADseq datasets from diverse organisms (Fig 1b-f) that used Bluepippin size selection, to determine the extent of resulting adaptor contamination and read overlaps. These highly diverse datasets demonstrate a range of possible outcomes from ddRADseq, which depend on a combination of factors in the experimental design, including the genome of the focal organism, the choice of restriction enzymes, size selection method and cutoffs, and other technical variations in the protocol. Therefore, we conducted a controlled comparative study to examine the effect of alternative enzyme choices on the resulting adaptor contamination, overlapping reads, and consequently, efficiency of sequencing. We generated two sets of libraries for the same set of samples with two different frequently cutting enzymes. Both sets of libraries were built using the same rare cutter (EcoRI), but the frequent cutter (MseI) was switched to an enzyme with a more GC rich recognition site (BfaI) to increase the average fragment size (Fig. 1g-h).

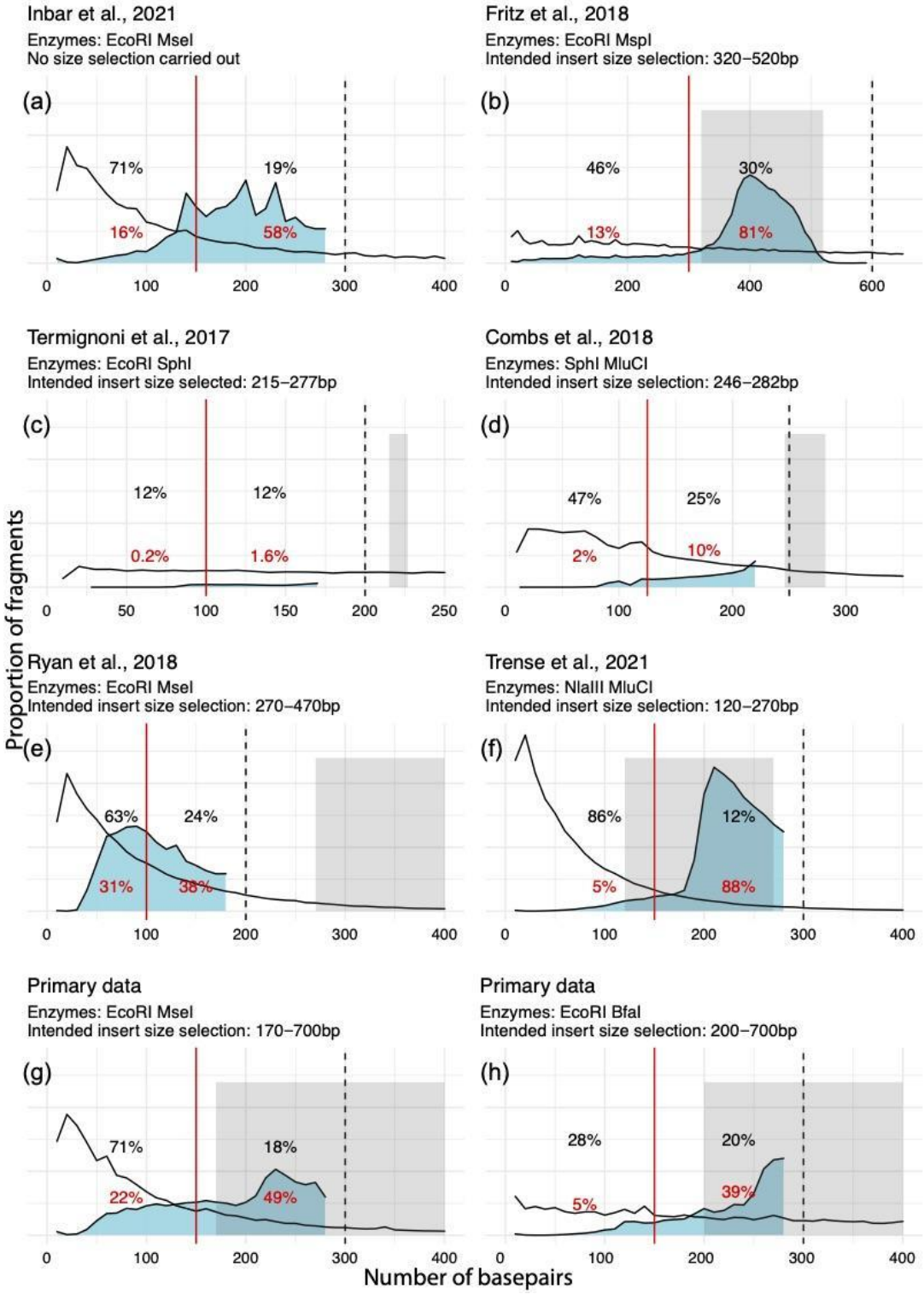

Figure 1: *In silico predictions (black trend line) compared with empirical results (blue histogram) obtained by merging overlapping reads. Black dashed line indicates two read lengths below which we expect overlapping reads. Red line indicates one read length below which we expect adaptor contamination. Grey box represents the intended size selection*

criteria chosen in the study. ***(a)*** *A study that did not use Bluepippin for size selection, only magnetic beads cleanup,* ***(b-f)*** *studies that used Bluepippin,* ***(g-h)*** *controlled experiments carried out on the same set of samples while switching the frequently cutting enzyme.*

The first major factor differentiating these studies is the in silico predicted distribution of fragment sizes. The results of the meta-analyses and our controlled experiment reveal high variation in the shape of the in silico distributions, dramatically differing in their skewness. Some studies might have most of the expected distribution in the size range of one read length, which would result in high level of adaptor contamination. For example, Trense et al. and our MseI/EcoRI experiment have 86% and 71% of the in silico distribution below one read length, respectively. In contrast, Termignoni et al. and our BfaI/EcoRI experiment have 12% and 28%, respectively. These results indicate that the initial fragment size distribution, before any size selection, is highly dependent on the combination of enzyme choice and studied genome.

The second important distinction between these studies was the choice of size selection range and the effectiveness of size selection. In theory, one might expect no adaptor contamination and read overlaps if the size selection cutoff is larger than two read lengths. However, in practice we observe large proportions of sequencing reads with overlaps and adaptor contamination in some of the studies. For example, Combs et al. chose a size selection cutoff that is approximately equal to two read lengths, and yet 10% of the sequencing output had read overlaps and 2.1% had adaptor contamination. Ryan et al. chose a similar size selection criterion, but their results show much larger proportions of adaptor contamination (31%) and read overlaps (38%). The observed empirical distribution is similar to the in silico distribution, possibly indicating that the size selection procedure was not effective. Other studies that intentionally allowed read overlaps but had a size selection cutoff just bigger than one read length were not expected to have adaptor contamination, and

yet large proportions of adaptors are observed in their result (e.g., 13.1% in Fritz et al.). These results demonstrate that a large proportion of short fragments might pass through the size selection procedure, even when using the most precise instruments available. This issue is worse when the in silico distribution is steeper, i.e., having a larger proportion of short fragments before size selection. This effect could explain why Termignoni et al. only had 0.2% adaptor contamination, while Combs et al. had 2.1% and our MseI/EcoRI experiment had 22%, corresponding to the difference in the in silico proportions of 11.7%, 46.5%, and 71%, respectively.

Size selection not only leads to variable outcomes, but it also appears to be imprecise. In our controlled experiment (Fig 1g and h), we see a shift of nearly 50bp between the intended size selection range and the outcome in the empirical data. A similar shift of nearly 60bp can be observed in Trense et al.. Despite these shortcomings of size selection, we see a considerable effect of size selection in enriching fragments within the desired range. In Fritz et al. (Fig 1b), Trense et al. (Fig 1f), and our experiments (Fig 1g and h), size selection cutoffs were below two read lengths, which allows us to observe the proportion of the empirical distribution within the size selection range. Size selection in these studies resulted in a much larger proportion of fragments in the selected range, higher than that expected in the in silico distribution, indicating that size selection was successful in enriching target fragment sizes.

To understand the relationship between the in silico predictions and empirical outcomes, we carried out a linear regression of expected vs. observed proportion of adaptor contamination (Fig 2a) and read overlap (Fig 2b) from the seven studies that used instrument based size selection. When we looked at the proportion of fragments with adaptor contamination, although we found a trend of increasing proportion of adaptor contamination in empirical data with increasing predicted adaptor contamination, this relationship was not statistically

significant (slope: 0.213; intercept: 0.003; *p*-value: 0.284). Similarly, the predictions for read overlaps below the size selection cutoff showed a positive trend (we used the observed sharp drop in the distribution, not the intended cutoff), but this was not significant either (slope: 0.40941; intercept: -0.08044; *p*-value: 0.3069). We also noted a trend of increase in variance with increase in the in silico predictions. This suggests that we observe the effect of the in silico prediction compounded with other effects due to technical variations and various sources of noise, including the choice of size selection cutoffs, variability in the bead clean-up and/or Bluepippin steps, and other variations in the details of the ddRADseq protocol.

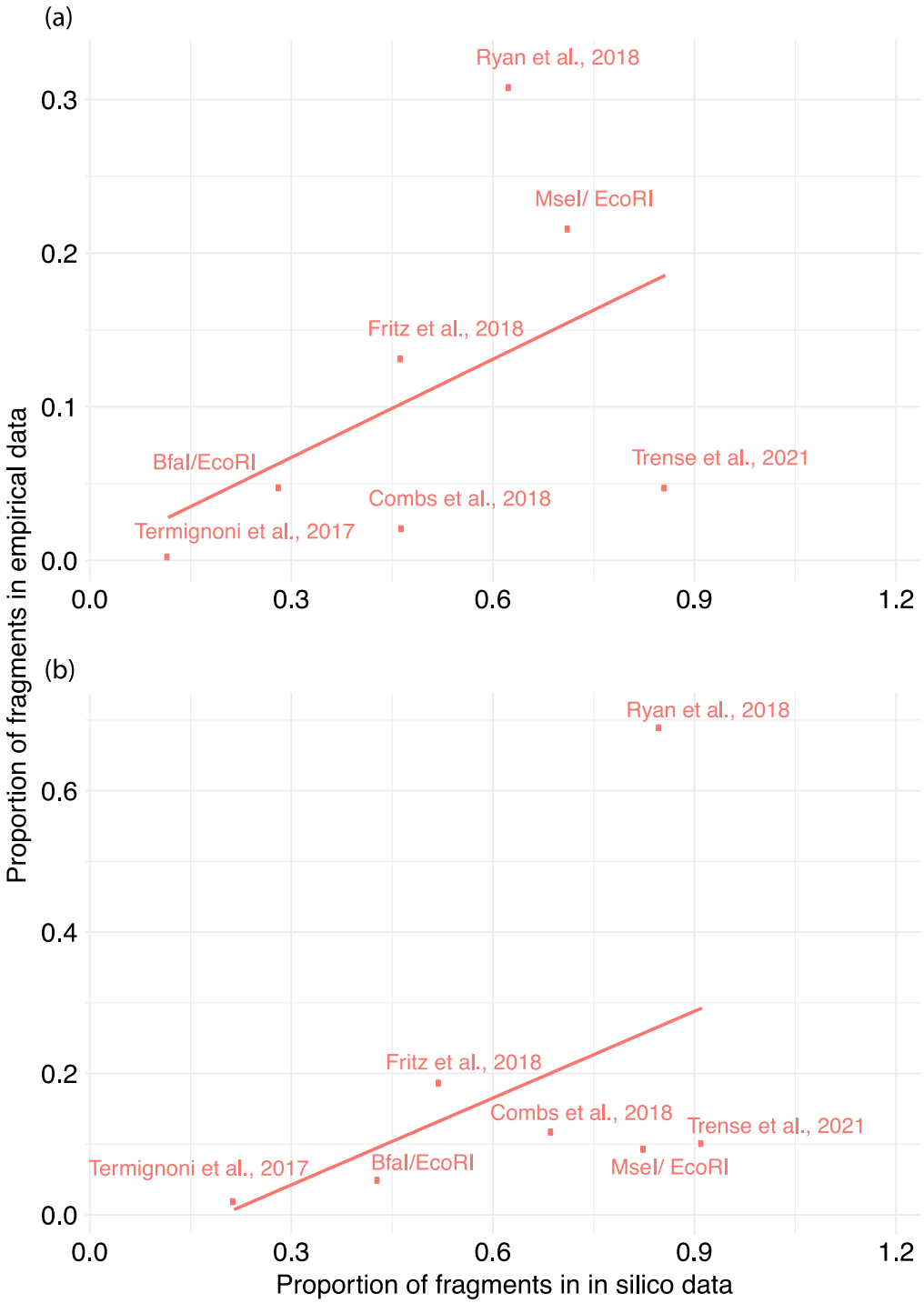

Figure 2: *Regression plot of empirical vs. in silico proportions of adaptor contamination* ***(a)*** *and read overlaps* ***(b)*** *for the various studies (i.e., fragments shorter than one read length or shorter than two read lengths, respectively).*

We also detect an effect of the distance between the size selection cutoff and the read length on the efficacy of size selection. For example, both Fritz et al. (Fig. 1b) and Combs et al. (Fig. 1d) had a similar proportion of predicted fragments below one read length (46% vs 47%). Combs et al. chose a size selection range beginning at two read lengths while Fritz chose a cutoff just larger than one read length. This resulted in only 2.1% adaptor contamination for Combs et al. compared to 13.1% for Fritz et al., suggesting that as the lower cutoff of size selection decreases, adaptor contamination increases. This might be due to the tendency of some short fragments to migrate with slightly longer fragments in gel electrophoresis. However, choosing a high selection cutoff might not necessarily be the optimal choice, because it might dramatically reduce the number of loci in the sequencing output.

One of the differences between our two controlled experiments, other than the frequently cutting enzyme, was the size selection cutoff. In the MseI/EcoRI experiment, the in-silico distribution predicted close to 90% of the fragments to be below 300bp compared to only 48% in BfaI/EcoRI. While selecting a cutoff above 300bp would result in efficient sequencing with low wastage, this would also exclude ~70% of the genomic sites from being sequenced (only 5 million out of a potential 18 million genomic sites would be sequenced for MseI/EcoRI; Fig. 3). Therefore, we considered the trade-off between genomic sites sequenced and sequencing efficiency, defined as the proportion of sequenced bases not wasted due to adaptor contamination or read overlaps (see Methods for details). As the lower cutoff moves below 300bp, we allow some read overlaps and sequencing efficiency drops. This drop is more pronounced for MseI than BfaI (dashed lines in Fig. 3) in line with their different in silico distributions. Hence, different cutoffs were chosen for the two experiments to optimise this trade-off: 170bp for MseI/EcoRI and 200bp for BfaI/EcoRI.

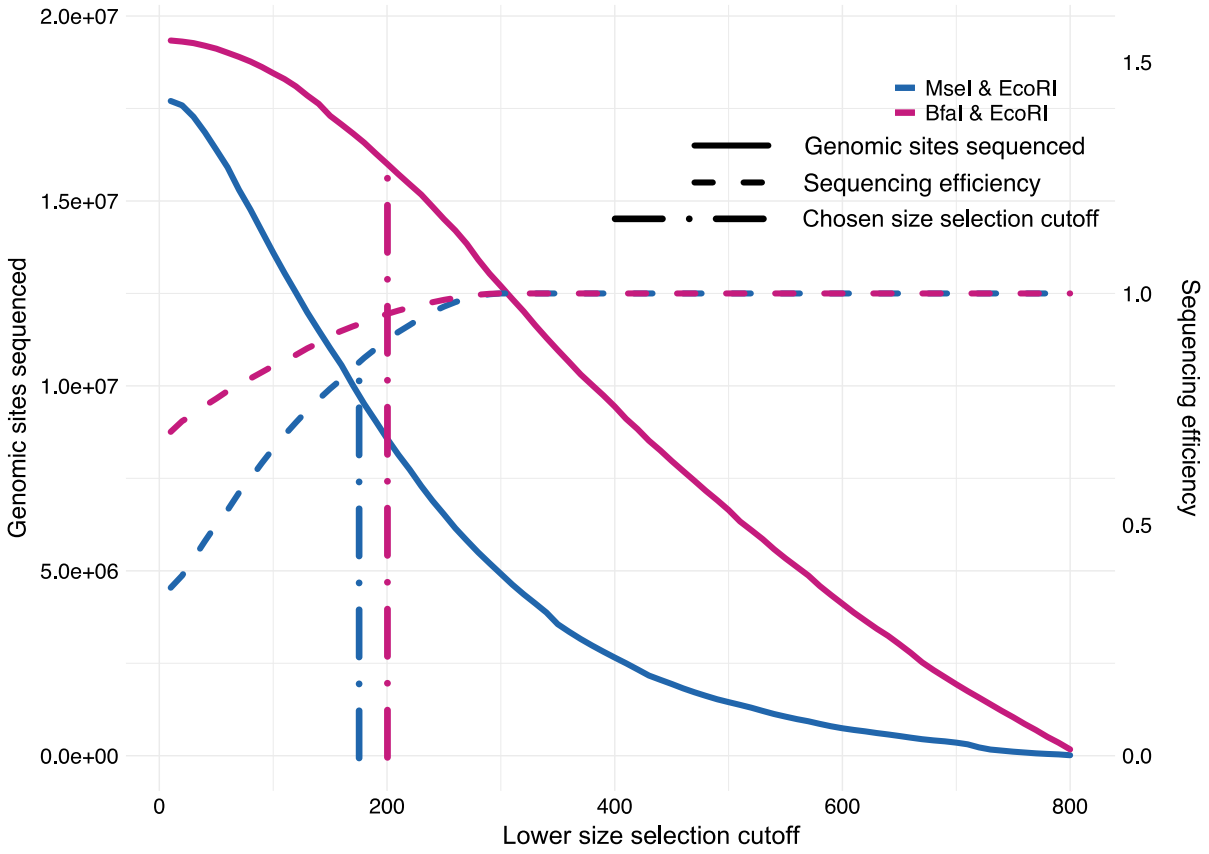

Figure 3: *Trade-off between sequencing efficiency and the number of genomic sites sequenced plotted against the size selection cutoff for the controlled experiment.*

Theoretically, we expected an efficiency of 92% in MseI/EcoRI (for the observed sharp drop in the empirical distribution at 210bp in Fig. 1g) and 98% in BfaI/EcoRI (at 240bp in Fig. 1h) based on the excel sheet calculator (Supplementary file 3). However, due to poor efficiency of size selection that allows unintended shorter fragments to go through, we observed lower sequencing efficiency. Moreover, the considerably lower sequencing efficiency of MseI/EcoRI relative to BfaI/EcoRI is a reflection of the large number of short fragments produced by MseI/EcoRI compared to BfaI/EcoRI as predicted from the in silico distribution and observed in the empirical distribution. The empirical data from the controlled experiment revealed a sequencing waste of 27% of the genomic fragments (sequencing efficiency of

73%) in MseI/EcoRI compared to 2.7% in BfaI/EcoRI (sequencing efficiency of 97%) as a result of adaptor contamination and read overlaps.

Interestingly, we observe higher variance of sequencing depth for MseI/EcoRI compared to BfaI/EcoRI (Fig 4a). Accordingly, if we do not set any threshold on the minimum depth, we have 8,003,694 unique genomic sites sequenced with MseI/EcoRI compared to 5,404,131 with BfaI/EcoRI. However, as in common practice, when we do require a minimum depth of 10X (--minDP parameter), less genomic sites are usable in MseI/EcoRI (513,891) compared to BfaI/EcoRI (737,766) (Fig. 4b). Therefore, the enzyme choices that give a more uniform fragment size distribution ultimately result in a larger number of genomic sites that can be genotyped (for a given sequencing budget, and with a commonly used depth cutoff).

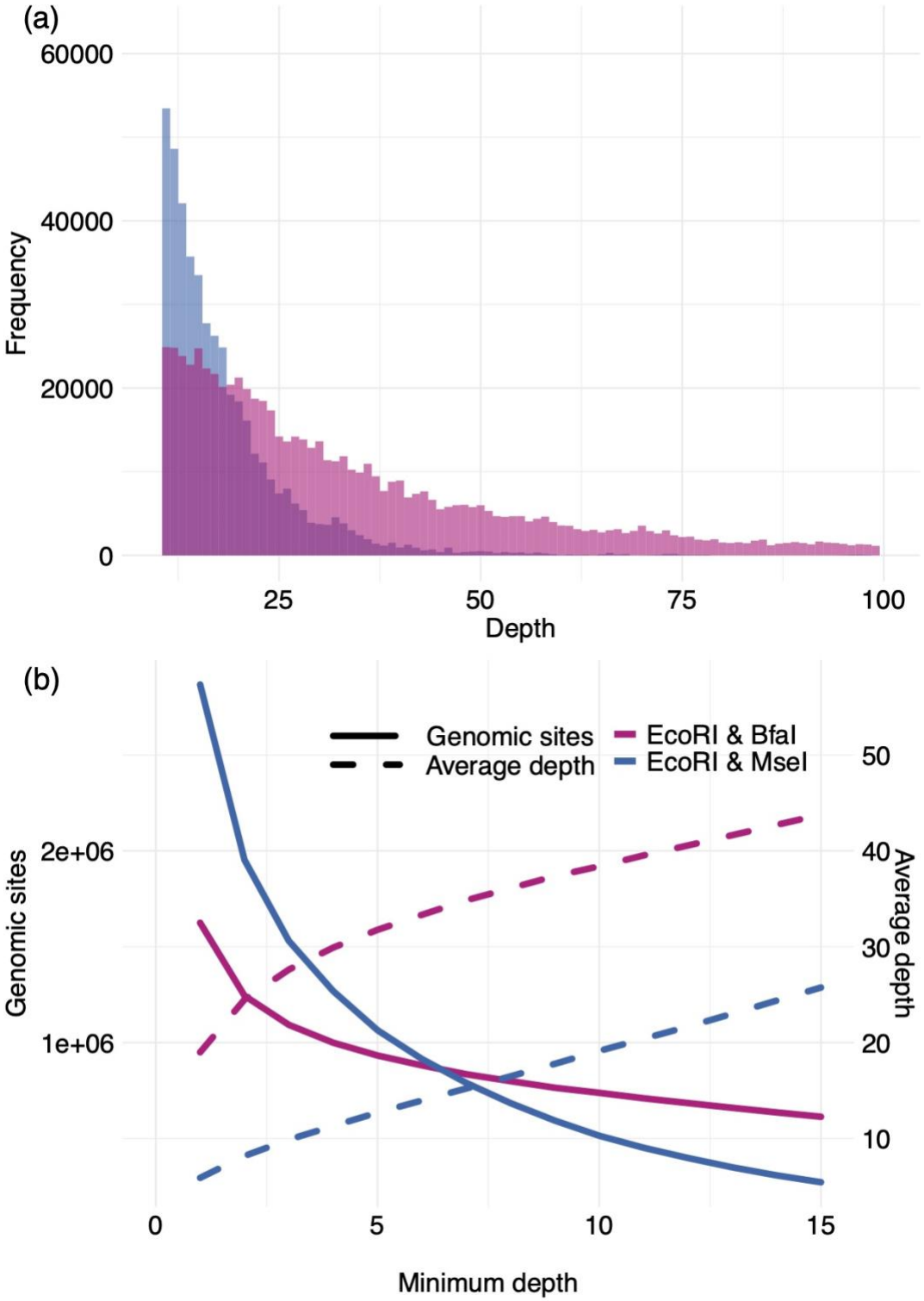

Figure 4: (*a*) *Distribution of sequencing depth for the two enzyme choices.* ***(b)*** *Number of unique genomic sites sequenced plotted against the minimum depth cutoff.*

Webtool: ddgRADer

ddgRADer (http://ddgrader.haifa.ac.il/) was built to optimise sequencing efficiency by aiding enzyme and size-selection choices. It was created as a webtool to make designing ddRADseq experiments more accessible. The webtool works on the user-provided reference genome along with some additional basic information. The genome is in silico digested with either a predetermined list of 14 commonly used enzyme pairs or any enzyme combinations chosen by the user from a comprehensive list (see the two usage modes “I have no idea!” and “Lemme try out!” in Fig. 5a). Based on the nature of the question being addressed, the desired number of SNPs to be genotyped (for example to carry out demographic analysis) or the desired SNP density (for a genomic scan analysis) are specified by the user. The user also enters an estimate of the expected SNP density in the study population, which is needed to predict the number of SNPs that will be genotyped by ddRADseq. Additional input information includes read length, single- or paired-end sequencing, expected yield from one sequencing lane, and the desired average sequencing depth for each sample. Based on the results of the in silico digestion, ddgRADer gives out the expected number of fragments in each fragment size bin that have cut sites of both the enzymes on either end. The predicted fragment size distributions from digestion by various enzyme pairs are plotted. Further outputs are given as an interactive table where the user can try out different size selection cutoffs for each of the enzyme combinations to compare the number of SNPs in the digested sample, the number of samples that can be multiplexed, and the expected sequencing efficiency (Fig. 5c). In addition to these theoretical values, the table also lists predicted values based on the meta-analysis and our controlled experiment, incorporating the effect of incomplete size selection. This is done using the two linear models based on the linear regression analysis above, one for read overlaps and one for adaptor contamination (Fig. 2). These estimates cannot be taken as precise predictions because of the large variation we observed across the datasets included in the meta-analyses. They are meant to be

suggestive of the scale of the effect of the combination of incomplete effectiveness of size selection and the enzyme choice on sequencing efficiency, and as a cautionary indication of how unrealistic the theoretical predictions might be. The workflow of ddgRADer is shown in Fig. 6 and details are given in Supplementary File 3.

A few tools are already available to help in designing ddRADseq experiments and optimise the sequencing outcome. SimRAD (Lepais & Weir, 2014) is designed to choose between different reduced representation library protocols by predicting the number of loci that can be generated. It performs in silico digestion and fragment size-selection allowing the user to choose the enzyme and size-selection criteria. Another software that carries out a similar analysis is DDRADSEQTOOLS (Mora-Márquez et al., 2017). In addition to in silico digestion, this tool also simulates the effect of allele dropout and PCR duplicates on the coverage taking into account potential sources of error. Both simRAD and DDRADSEQTOOLS are R packages. RADinitio (Rivera-Colón et al., 2021) is a python software that simulates population-level data based on a demographic model input by the user. The user can implement various reduced representation protocols, in silico digestion, size selection, library amplification, and sequencing depth, while taking into account errors such as allelic dropout. Additionally, RADinitio carries out retrospective simulation, where an in silico library is generated based on an empirical RADseq dataset. Thus, RADinitio takes into account a lot more variables, particularly inherent variability in natural populations, that contribute to the quality of the data. While some of these tools help optimise experiments by taking into account allele dropouts and PCR duplication, none of them help in making the experiment cost effective by optimising the sequencing efficiency. Moreover, they assume perfect size-selection and do not take into account its incomplete effectiveness, which results in additional adaptor contamination and read overlaps. This is an important factor to consider when selecting the enzyme pair because in certain cases it has a dramatic effect on sequencing efficiency, the expected number of SNPs, the sequencing depth, and the number

of samples that can be multiplexed per lane. Additionally, all these softwares are command-line tools and might require some skill in R, which would involve a steep learning curve particularly for researchers without prior experience with command-based programs.

The easy-to-use graphical user interface of ddgRADer makes it accessible to even a beginner, considerably reducing the time taken to design ddRADseq experiments. Further, multiple popovers and an explanation page are included to support the user and provide explanations about the process and guidance on decision making. Thereby, this tool aids researchers that do not yet have experience or in-depth understanding of RADseq protocols and the associated issues. ddgRADer explicitly incorporates the error associated with size-selection by accounting for additional shorter fragments that result in adaptor contamination and read overlaps. This helps the user make several informed enzyme choices taking into account sequencing efficiency. Although enzymes aren't very expensive, the multiplexing barcodes require a large initial investment. These barcodes attach to the overhang generated by the restriction enzyme making enzyme choice a consequential decision. Additionally, ddgRADer allows the user to compare several enzyme pairs and the size-selection cutoffs that best suit each of these pairs. The interactive size-selection slide bar helps the user understand the trade-off between the number of genomic sites in the digested sample and sequencing efficiency, as a function of different size-selection cutoffs, further aiding enzyme choice.

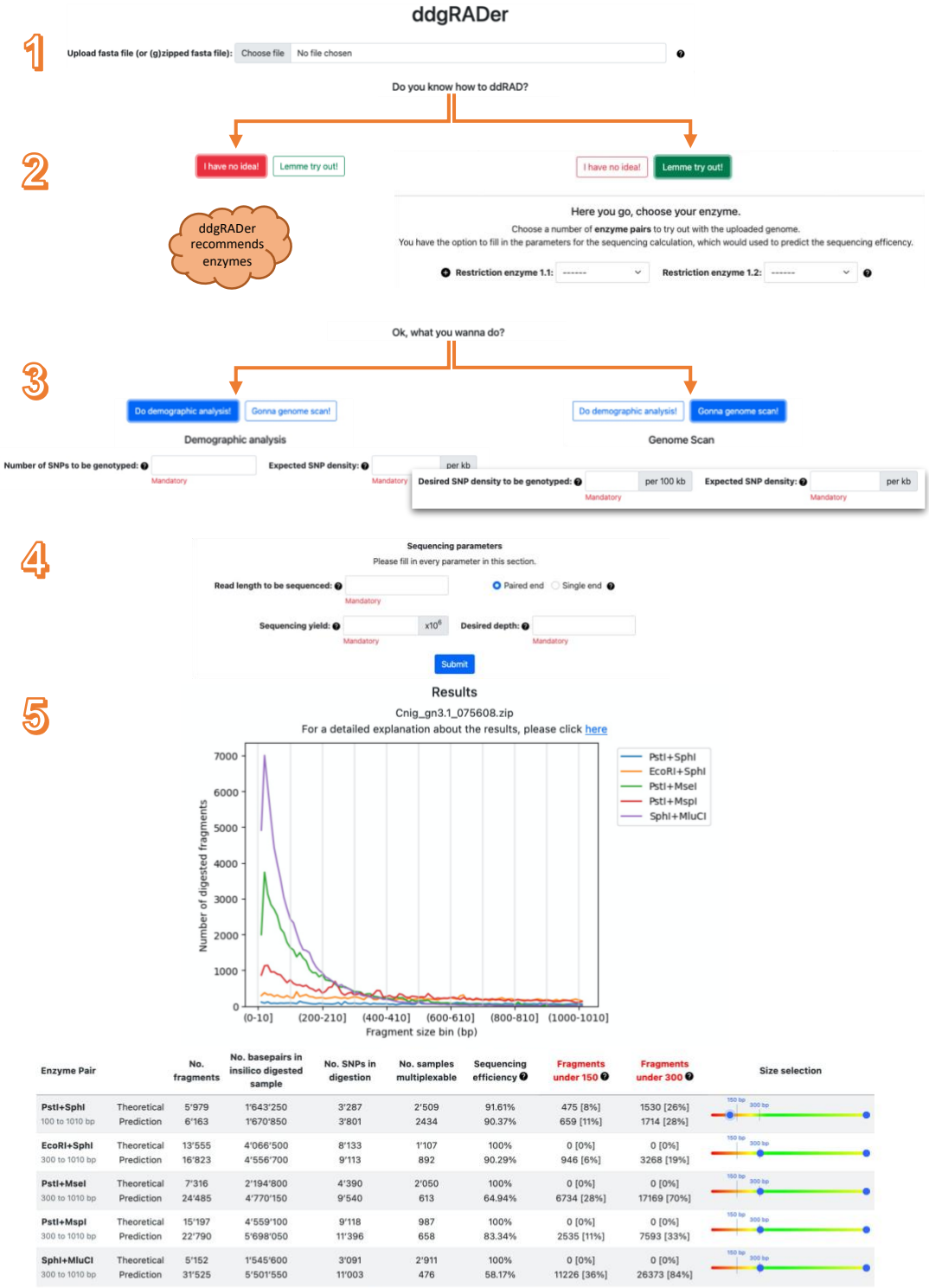

Figure 5: *Overview of the workflow of the ddgRADer webtool and output of ddgRADer for the genome of Cataglyphis niger with five different enzyme combinations (with 150bp paired-end sequencing, 300M expected yield, and 20X desired depth). The user uploads the*

reference genome sequence and enters numerical parameters including the desired number of SNPs and sequencing parameters. The output consists of the in silico predicted fragment size distribution and a table that lists theoretical (based on in silico digestion) and predicted (based on the empirical studies) values from the webtool.

Conclusion

Here, we demonstrate using empirical data and controlled experiments that enzyme choice greatly affects inadvertent inclusion of adaptor contamination and read overlaps, in spite of size-selection. We observed that efficacy of size-selection is variable across studies and imprecise. Therefore, enzyme choice is crucial in determining sequencing efficiency and optimising ddRADseq experimental design. We provide a user-friendly webtool that guides beginners through considering alternative enzyme choices and optimising size-selection. The tool provides realistic predictions about the expected outcomes from ddRADseq, based on our empirical characterization of size-selection efficacy in a meta-analysis across diverse studies from the literature as well as the controlled experiments. This will make ddRADseq far more accessible and facilitate the smooth transition of more researchers into the field of population genomics.

Acknowledgements

We thank Jessica Purcell and Alan Brelsford for advice on ddRADseq methodology, and Pnina Cohen for help in sample collection. This study was funded by US-Israel Binational Science Foundation (Grant no. 2017319).

References

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Clugston, J. A. R., Kenicer, G. J., Milne, R., Overcast, I., Wilson, T. C., & Nagalingum, N. S. (2019). RADseq as a valuable tool for plants with large genomes—A case study in cycads. *Molecular Ecology Resources*, *19*(6), 1610–1622. https://doi.org/10.1111/1755-0998.13085

Combs, M., Puckett, E. E., Richardson, J., Mims, D., & Munshi-South, J. (2018). Spatial population genomics of the brown rat (Rattus norvegicus) in New York City. *Molecular Ecology*, *27*(1), 83–98.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Eaton, D. A. R., & Ree, R. H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology*, *62*(5), 689–706. https://doi.org/10.1093/sysbio/syt032

Fritz, M. L., DeYonke, A. M., Papanicolaou, A., Micinski, S., Westbrook, J., & Gould, F. (2018). Contemporary evolution of a Lepidopteran species, Heliothis virescens, in response to modern agricultural practices. *Molecular Ecology*, *27*(1), 167–181. https://doi.org/10.1111/mec.14430

Harzing, A. W. (2007). *Publish or perish 6*.

Inbar, S., Cohen, P., Yahav, T., & Privman, E. (2020). Comparative study of population genomic approaches for mapping colony-level traits. *PLoS Computational Biology*, *16*(3), e1007653.

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

Lepais, O., & Weir, J. T. (2014). SimRAD: An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, *14*(6), 1314–1321. https://doi.org/10.1111/1755-0998.12273

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Magalhaes, I. S., Whiting, J. R., D'Agostino, D., Hohenlohe, P. A., Mahmud, M., Bell, M. A., Skúlason, S., & MacColl, A. D. C. (2021). Intercontinental genomic parallelism in multiple three-spined stickleback adaptive radiations. *Nature Ecology & Evolution*, *5*(2), 251–261. https://doi.org/10.1038/s41559-020-01341-8

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Mora-Márquez, F., García-Olivares, V., Emerson, B. C., & López de Heredia, U. (2017). ddRADseqTools: A software package for in silico simulation and testing of double-digest RADseq experiments. *Molecular Ecology Resources*, *17*(2), 230–246. https://doi.org/10.1111/1755-0998.12550

Natesh, M., Atla, G., Nigam, P., Jhala, Y. V., Zachariah, A., Borthakur, U., & Ramakrishnan, U. (2017). Conservation priorities for endangered Indian tigers through a genomic lens. *Scientific Reports*, *7*(1), 9614. https://doi.org/10.1038/s41598-017-09748-3

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, *21*(12), 2991–3005.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). *dDocent*: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431. https://doi.org/10.7717/peerj.431

Rivera-Colón, A. G., Rochette, N. C., & Catchen, J. M. (2021). Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Molecular Ecology Resources*, *21*(2), 363–378. https://doi.org/10.1111/1755-0998.13163

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Ryan, S. F., Fontaine, M. C., Scriber, J. M., Pfrender, M. E., O'Neil, S. T., & Hellmann, J. J. (2017). Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climatic gradient. *Molecular Ecology*, *26*(18), 4725–4742.

Shields, E. J., Sheng, L., Weiner, A. K., Garcia, B. A., & Bonasio, R. (2018). High-Quality Genome Assemblies Reveal Long Non-coding RNAs Expressed in Ant Brains. *Cell Reports*, *23*(10), 3078–3090. https://doi.org/10.1016/j.celrep.2018.05.014

Simola, D. F., Wissler, L., Donahue, G., Waterhouse, R. M., Helmkampf, M., Roux, J., ... & Gadau, J. (2013). Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome research*, *23*(8), 1235-1247.

Termignoni-García, F., Jaramillo-Correa, J. P., Chablé-Santos, J., Liu, M., Shultz, A. J., Edwards, S. V., & Escalante-Pliego, P. (2017). Genomic footprints of adaptation in a cooperatively breeding tropical bird across a vegetation gradient. *Molecular Ecology*, *26*(17), 4483–4496. https://doi.org/10.1111/mec.14224

Trense, D., Schmidt, T. L., Yang, Q., Chung, J., Hoffmann, A. A., & Fischer, K. (2021). Anthropogenic and natural barriers affect genetic connectivity in an Alpine butterfly. *Molecular Ecology*, *30*(1), 114–130.

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., Sivasundar, A., & Seehausen, O. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, *22*(3), 787–798. https://doi.org/10.1111/mec.12023

Zecherle, L. J., Nichols, H. J., Bar-David, S., Brown, R. P., Hipperson, H., Horsburgh, G. J., & Templeton, A. R. (2021). Subspecies hybridization as a potential conservation tool in species reintroductions. *Evolutionary Applications*, *14*(5), 1216–1224. https://doi.org/10.1111/eva.13191

[dataset] Lajmi, Glinka and Privman 2022; ddRADseq data for *Camponotus fellah*; National Centre for Biotechnology Information: Sequence Read Archive (NCBI SRA);