Restriction Site Associated DNA Python-Digested Simulation (RApyDS)

Kristianne Arielle Gabriel
Mark Lenczner Mendoza
Daniella Jean Pamulaklakin

August 2018

Project for the 2018 Internship Program in Bioinformatics

Contents

1 Introduction

Genetic markers are variations in the genes or sequences in a specific location that can be used to identify individuals. Today, with next-generation sequencing, we are able to simultaneously discover and score large numbers of markers. As a result, we can now easily sequence and study whole genomes. However, along with this is a high sequencing cost and a huge amount of data which is often unnecessary (Hamblin & Rabbi, 2014). As an alternative to whole genome sequencing, restriction site-associated DNA sequencing (RADseq) is a technology which combines molecular biology techniques with next-generation sequencing to sample only a subset of the target genome. It utilizes restriction enzymes in order to sample specific sequences in the genome that is flanked by the restriction sites. The short DNA fragments produced can then be screened for genetic variation. RAD sequencing techniques is advantageous over other techniques because it can produce a lot of information at a much lower cost (Shirasawa, Hirakawa, & Isobe, 2016; Rowe, Renaut, & Guggisberg, 2011; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016) and a greater depth of coverage per locus which increases the confidence in the genotype calls (Andrews et al., 2016). RADseq can be used in studies involving single-nucleotide polymorphism (SNP) discovery (Baird et al., 2008; Davey & Blaxter, 2010), genome-wide genotyping (Davey et al., 2011), de novo assembly of genomes without reference genome (Baird et al., 2008; Shirasawa et al., 2016), linkage analysis, genus- or family-level phylogenetics (Yang et al., 2016), and phylogeographic studies (Davey & Blaxter, 2010).

1.1 RAD Sequencing

In designing a successful RADseq study, there are several factors researchers have to consider. First is the choice of the restriction enzyme/s to be used for the target organism (Herrera, Reyes-Herrera, & Shank, 2015; Davey et al., 2011; Bonatelli, Carstens, & Moraes, 2015). This decision is affected by the number of RAD markers desired for a specific study.

In order to help researchers with the problem of choosing the best restriction enzyme to use for a RADseq study, we developed a tool which aims to provide the user with an estimation of the number of restriction sites, number of fragments and number of fragments that is within a gene region that will be produced after digestion with a specific restriction enzyme. The goal is to evaluate restriction enzymes and rank which is the best one to use for a specific RADseq study. In the development of the tool, there are smaller objectives:

1. Given the genome sequence, the restriction enzyme should be able to cut the genome into fragments that are desirable for the sequencing run;

2. Given the gene annotation, when compared, the location of the fragments produced should be a part of a gene region; and

3. Determine repetitive fragments produced.

This is an exploratory study which means that the scope can be broadened and that the program can be further developed, adding more functions.

1.2 RADseq Techniques

There are several RADseq techniques developed to become more suitable for ecological and evolutionary genomic studies of different species. These includes complexity reduction of polymorphic sequences (CRoPS), genotyping-by-sequencing and reduced representation libraries (Andrews et al., 2016; Davey et al., 2011). In this study, we only focused on two specific protocols namely, the original RADseq and ddRADseq. RADseq techniques are advantageous than other techniques as prior genomic knowledge about the target organism is not required.

All RADseq protocols require a high-molecular-weight DNA sample (Andrews et al., 2016) in order to produce a good quality sequencing library. The original RADseq protocol involves the digestion of the genomic DNA using one specific restriction enzymes producing fragments with sticky-ends (Shirasawa et al., 2016; Bonatelli et al., 2015). The fragments are ligated with P1 adapters that will allow binding to the Illumina flow cell. These adapters are also composed of molecular identifiers (Davey & Blaxter, 2010) that will allow the reads to be associated with particular individuals. The tagged fragments from different individuals are pooled and then randomly sheared. P2 adapters are ligated to the sheared fragments and the fragments are PCR amplified. The P2 adapter has a specialized structure such that it will not bind to the P2 primer unless the P1 adapter has been completely amplified. This is a mechanisms that guarantees that all the amplified fragments have P1 and P2 adapters, MID and the sequence flanking the restriction site iteradseq and that only the target region will be sequenced (Andrews et al., 2016). This approach increases the confidence in the base calls as it increases the coverage for a given site (Rowe et al., 2011). The fragments that will be sequenced has different characteristic size due to the random shearing process that produced fragments that have a restriction site on one end and a randomly sheared on the other.

On the other hand, ddRADseq protocol is more simplified. It does not involve a random shearing process instead, it involves a second digestion with the same or a different restriction enzyme which offers customization for a specific study (Puritz, Matz, Toonen, Weber, & Bird, 2014). This protocol builds a sequencing library consisting only of fragments that are flanked by two different restriction sites (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012).

1.3 Choice of Restriction Enzyme

The study of genetic markers helps us in better understanding how genes function. Baird et al (2008), combined RAD marker isolation with Illumina sequencing in developing a genotyping platform that can discovery novel SNP markers and simultaneously genotype individuals. Many studies based on these method have been conducted since then.

The discovery of restriction enzymes had a great impact on the field of molecular biology (Moore & Moore, 1982). It allowed us to analyze and manipulate the DNA. These enzymes became essential tools for sampling loci and generation of information on population-level variation (Andrews et al., 2016).The choice of restriction enzyme is a critical step when designing a RADseq study. has a huge effect in the success of a RAD sequencing study. The choice can be affected by the nature of the study, the number of markers required (Baird et al., 2008; Davey et al., 2011; Herrera et al., 2015; Bonatelli et al., 2015; Shirasawa et al., 2016) and the characteristics of the genome to be studied (Shirasawa et al., 2016), thus the success of the study. An estimation of the number of restriction sites can be obtained given the genome size and the GC content of the genome.

Several studies were reviewed in order to determine the possible applications of RADseq and the factors that affect the choice of restriction enzyme for RADseq studies. In choosing the best restriction enzyme, Bontelli et al. (2015) recommended to use a combination of a rare and common cutters when working with a non-model organism. When working with a species with high genetic diversity, that is likely to show mutations in the restriction sites, it is recommended to use a rare cutter. On the other hand, species with low genetic diversity and those with high numbers of repeat sequences require a common cutter to produce sufficient polymorphic markers (Andrews et al., 2016).

The choice of restriction enzyme is also affected by the goals of the study. If conducting a study regarding phylogenetic relationships, geographic population structure, gene flow and individual inbreeding, only a several hundred to a few thousand of RADseq loci can be used to sample the whole genome. Those studies requiring examination of the functionally important regions will require more markers (Andrews et al., 2016).

In the study conducted by Yang et al. (2016) which aims to develop a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants, several combinations of restriction enzymes for ddRAD seq were used in 23 plant species most of which are angiosperms. They found out that AvaII+MspI (a 4.5-base cutter and a four-base cutter) enzyme pair produced a consistently high number of fragments (approximately 13, 958 segments between 400-700bp in length) and was predicted to produce sufficient tags in a broad range of plant species. Among the combinations, EcoRI+MspI and PstI+MspI were also recommend to be taken into consid-

eration as they may also produce hundreds to thousands of markers. Of the three recommended combinations, AvaII+MspI was the highly recommended, followed by EcoRI+MspI and then PstI+MspI.

2 RApyDS Program

To address the problems mentioned earlier, RApyDS was created to help in choosing the best restriction enzyme to be used for RAD sequencing.

RApyDS is a python script that performs in silico digestion of a given genome or known GC content. It can simulate single-enzyme or double-enzyme digestion on a single file or multiple FASTA files. After digestion, the program can match fragments to a given annotation file and run alignment analysis to check for *unique* and *repeat regions*. It provides a tabular report and visualization in forms of electrophoresis simulation and cut site mapping of the simulated digestion.

2.1 Workflow

The program requires a genome or a DNA sequence in `FASTA` format as its input. For cases that the genome is known, the assembled `FASTA` file can be immediately used. However, for cases that the genome is unknown but the GC frequency and length are known, a sequence can be generated.

For the original RAD protocol, the genome is first digested by the enzyme at restriction sites. This produces fragments that are sheared then size selected. In this protocol, it results to 3 types of fragments: (1) those that are too small for shearing and size selection, (2) after shearing, those that does not contain one sticky end and one sheared-end, and (3) those fragments that where successfully sheared and size selected which proceeds to the next part of the process. Fragments of Type 1 are filtered out by the program during size selection of the in silico-digested fragments. Those that do not meet the minimum size are discarded. During pair-end sequencing, only the start and end parts of a fragments that contains a sticky (cut by restriction enzyme) and a *sheared end* will be sequenced. Which means that for long fragments, the ends and some middle parts can still be sequenced after random shearing. To accomodate this, the program cuts the fragments into *reads* that are by default 100bp from the ends. This number can be changed depending on the type of sequencer that will be used. Type 2 fragments are the one in the middle of these *reads*. This process already accounts for the shearing and size selection. And these *reads* becomes the Type 3 fragments.

On the other hand, for the ddRAD protocol, it starts with the similar process of the genome being digested by the restriction enzymes. Since ddRAD cuts on two restriction site (let's call them *site A* and *site B*), it produces different

fragments than the original RAD. These types of fragments are: (1) those that ends with both *site A*, (2) those with ends of both *site B*, and (3) those with one end is *site A* and the other is *site B*. For ddRAD, only fragments of Type 3 is filtered and furthered used. Unlike in original RAD, these fragments undergo size selection only and no random shearing occurs. *Reads* are also taken from the Type 3 fragments.

In prepartion for the next part, the qualified fragments which we will now call *reads*, will be written on a `FASTQ` file. To take into account for repetition in the genome, the program uses an external tool Burrows-Wheeler Aligner (`BWA`) to align the *reads* into the genome. `BWA` then outputs a `SAM` file. Since we are concerned with the repeat regions, we are only interested in the `XA` and `XU` tags of the `SAM` file. This file is parsed by the program to look for: *unique reads* or those that `BWA` did not find a repeat in the genome, *repeat reads* or those that `BWA` did find a repeat (whether with another *read* or in another part of the genome), and *repeat regions* or regions in the genome that were found as repeat because it is similar to any of the *reads*. The sizes of the *reads* and the *repeat regions* are then summed to count the breadth of coverage.

In case the user attaches a `GFF` file and/or a desired feature (default is set to `gene`), the program counts the number of *reads* that is within the desired feature region and the percentage of feature regions that contains at least one *read*.

3 Output

After running RApyDS, it will produce a `.zip` file containing the generated report. The zipped file contains the following structure:

```
report/
├── output/
│   └── ...data files generated
├── src/
│   └── ...source files
├── cutsite.html
├── gel.html
└── index.html
```

There will be 3 html files:

- `index.html` contains a tabular information about the fragments from in silico digestion

- `gel.html` is the electrophoresis simulation
- `cutsite.html` contains the cut site location simulation

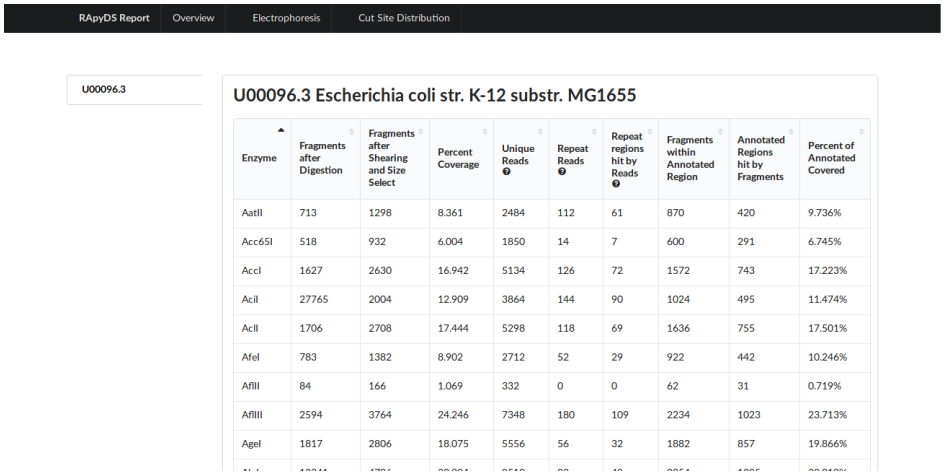
3.1 Overview (index.html)

This html file shows information about the fragments after in silico digestion. Column headers can be clicked to sort the table according to the column's value in increasing or decreasing order. And on the left is a side bar with list of sequence identifier in the input FASTA file.

The first column is the enzyme's name (for ddRAD the two enzymes' name) followed by the number of fragments it produced after digestion. The third column is the number of *reads* in both original RAD and ddRAD. Then followed by the coverage which is simply the sum of the length of *reads* divided by the total length of the genome.

The next three columns are the ones that account for repeat regions. The number of *unique reads*, *repead reads*, and *repeat regions*.

Lastly, the last 3 columns are for the targeted feature. *Fragments within annotated region* is the number of fragments whose start falls within the desired region, let's say for example a `gene`. *Annotated regions hit by fragments* and *Percent of Annotation Covered* represents the number of desired feature regions where there is at least one fragment covering it and that number divided by the total number of desired feature region respectively.



Enzyme	Fragments after Digestion	Fragments after Shearing and Size Select	Percent Coverage	Unique Reads	Repeat Reads	Repeat regions hit by Reads	Fragments within Annotated Region	Annotated Regions hit by Fragments	Percent of Annotated Covered
AatII	713	1298	8.361	2484	112	61	870	420	9.736%
Acc65I	518	932	6.004	1850	14	7	600	291	6.745%
AccI	1627	2630	16.942	5134	126	72	1572	743	17.223%
AciI	27765	2004	12.909	3864	144	90	1024	495	11.474%
AclI	1706	2708	17.444	5298	118	69	1636	755	17.501%
AfeI	783	1382	8.902	2712	52	29	922	442	10.246%
AflII	84	166	1.069	332	0	0	62	31	0.719%
AflIII	2594	3764	24.246	7348	180	109	2234	1023	23.713%
AgeI	1817	2806	18.075	5556	56	32	1882	857	19.866%

Figure 1: Overview table of the in silico digestion (original RAD) for E.Coli

3.2 Electrophoresis (gel.html)

One of the two visualisation the program presents is the electrophoresis simulation of the fragments after digestion. This makes use of the `d3.js` and `d3-electrophoresis` javascript plug-ins.

To begin the simulation, the user will first select the genome from the list, input the desired markers, and then selects up to five restriction enzymes. The webpage will read from the `json` output of RApyDS program then displays an electrophoresis simulation after. The markers can be customised as well as the fragments because they can be furthered size-selected for simulation purposes.

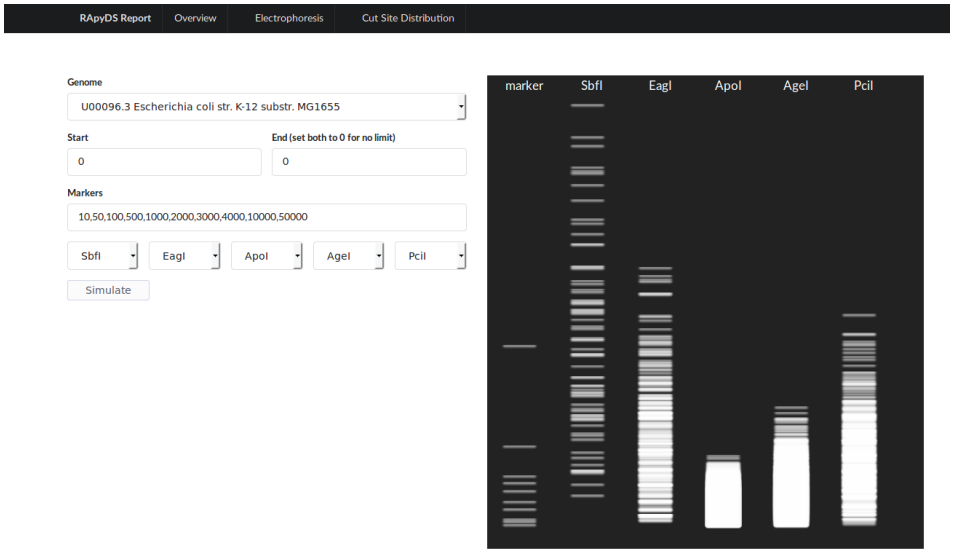


Figure 2: Electrophoresis simulation for E.Coli with enzymes AgeI, ApoI, EagI, PciI, SbfI

3.3 Cut Sites (cutsite.html)

The second and last visualisation is the cut site distribution. This also uses `d3.js` javascript plugin.

For this visualisation, the data is from another `json` file also generated by the program which contains the location of the cut sites in the genome. It is then rendered as a black line in perpendicular to the genome location as the x-axis. The image can be zoomed in to see the more exact location.

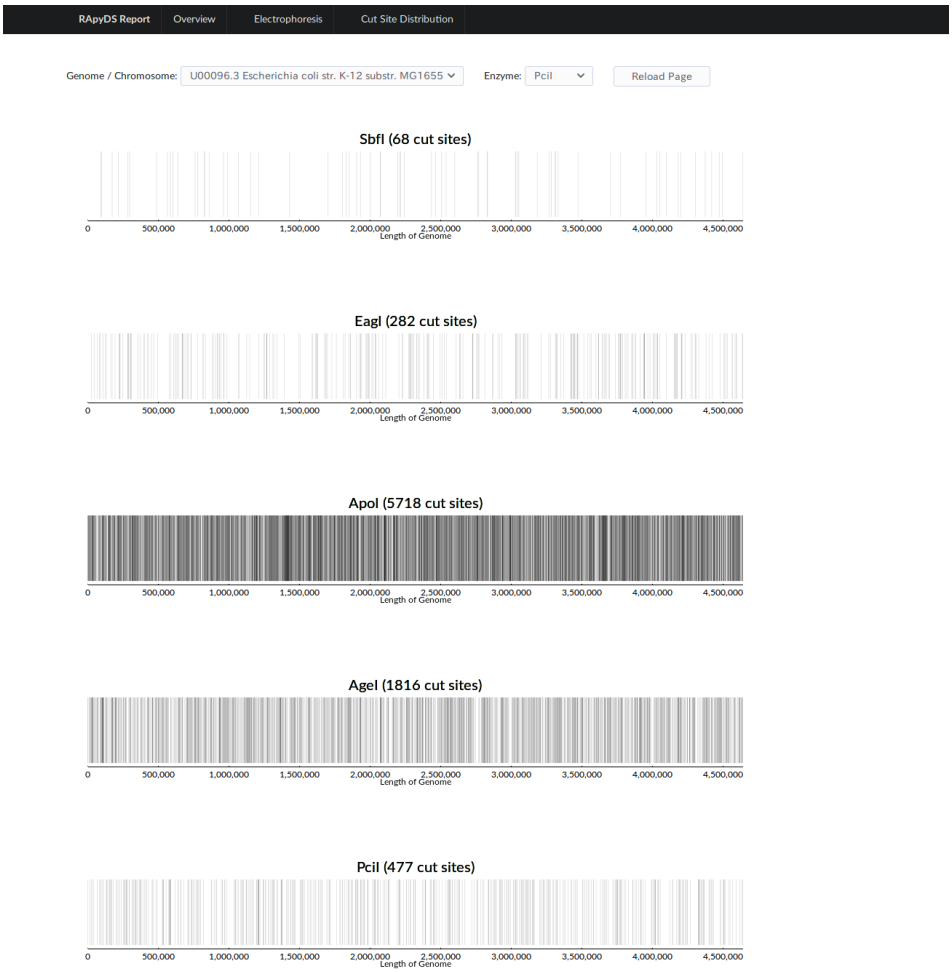


Figure 3: Cut site markers for E.Coli with enzymes AgeI, ApoI, EagI, PciI, SbfI

4 Appedix

4.1 Technical Specifications

4.1.1 Software Requirements

- Linux OS
- Python 2.7 or greater
- numpy (`pip install numpy`)
- BWA 0.7.12 (http://bio-bwa.sourceforge.net/)

- Firefox (for viewing html files)

4.2 Running RApyDS

4.2.1 Arguments

`-h, --help` - show this help message and exit
`-i [I]` - input genome sequence file (FASTA)
`-db [DB]` - resteriction enzyme dabatase file.
Format per line: SbfI,CCTGCA|GG
`-re [RE]` - file of list of restriction enzyme to be tested
`-a [A]` - annotation file for genome (GFF)
`-at [AT]` - what to look for in gene annotation file (ex. gene region, exon, intron, etc) (default: gene)
`-min [MIN]` - minimum fragment size (default: 200)
`-max [MAX]` - maximum fragment size (default: 300)
`-bp [BP]` - base pair read length for FASTQ generation (default: 100)
`-p [P]` - radseq protocol: use ddrad for double digestion (default: orig)
`-gc [GC]` - input gc frequency. Value must be between 0 and 1
`-dna [DNA]` - input dna estimated length
`-o [O]` - output file name (default: report)
`-t [T]` - number of processes (default: 4)

4.2.2 Common Usage

Original RADSeq with:

- known or given genome file (FASTA format)
  `./rapyds.py -i <genome_file.fasta> [other arguments]`
- known or given genome file (FASTA format) with custom restriction enzyme list
  `./rapyds.py -i <genome_file.fasta> -re <restriction_enzymes> [other arguments]`
- unknown genome but with GC content/frequency
  `./rapyds.py -gc <GC frequency> -dna <sequence length> [other arguments]`

DDRad with known genome:

`./rapyds.py -i <genome_file.fasta> -p ddrad -re <re_file.txt> [other arguments]`

References

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of radseq for ecological and evolutionary genomics. *Nature Review Genetics*, *17(2)*.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid snp discovery and genetic mapping using sequenced rad markers. *PLoS ONE*, *3(10)*.

Bonatelli, I. A. S., Carstens, B. C., & Moraes, E. M. (2015). Using next generation rad sequencing to isolate multispecies microsatellites for pilosocereus (cactaceae). *PLoS ONE*, *10*(11).

Davey, J. W., & Blaxter, M. L. (2010). Radseq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5-6), 416-423.

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 Review Genetics*, *12(7)*, 499-510.

Hamblin, M. T., & Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotype-by-sequencing libraries: a study in cassava (manihot esculenta). *Crop Science*, *54*.

Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting rad-seq marker numbers across the eukaryotic tree of life. *Genome Biology and Evolution*, *7*(12), 3207-3225.

Moore, G. P., & Moore, A. R. (1982). The average spacing of restriction enzyme recognition sites in dna. *Journal of Theoretical Biology*, *98*, 165-169.

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).

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., & Bird, D. I. B. C. E. (2014). Demystifying the rad fad. *Molecular Ecology*, *23*, 5937-5942.

Rowe, H. C., Renaut, S., & Guggisberg, A. M. (2011). Rad in the realm of next-generation sequencing technologies. *Molecular Ecology*, *20*.

Shirasawa, K., Hirakawa, H., & Isobe, S. (2016). Analytical workflow of double-digest restriction site-associated dna sequencing based on empirical and in silico optimization. *DNA Research*, 145-153.

Yang, G.-Q., Chen, Y.-M., Wang, J.-P., Guo, C., Zhao, L., Wang, X.-Y., ... Guo, Z.-H. (2016). Development of a universal and simplified ddrad library preparation approach for snp discovery and genotyping in angiosperm plants. *Plant Methods*, *12*(39).