Received Date : 27-Jan-2014

Revised Date : 30-Apr-2014

Accepted Date : 30-Apr-2014

Article type : Resource Article

SimRAD: a R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches.

[1]Olivier Lepais, [2]Jason T Weir

[1] INRA, UMR 1224, Ecologie Comportementale et Biologie des Populations de Poissons, Saint Pée sur Nivelle, France

[1] Univ Pau & Pays Adour, UMR 1224, Ecologie Comportementale et Biologie des Populations de Poissons, UFR Sciences et Techniques de la Côte Basque, Anglet, France

[2] Department of Biological Sciences, and Department of Ecology and Evolutionary Biology, University of Toronto Scarborough, Toronto, ON, M1C 1A4, Canada

Keywords: restriction site associated DNA polymorphism, GBS, next generation sequencing, genome complexity reduction, *in silico* digestion, single nucleotide polymorphism.

Corresponding author: Dr Olivier Lepais, INRA, UMR 1224 ECOBIOP, Aquapole, 64310 Saint Pée sur Nivelle, France; fax: +33 (0)5 59 54 51 52; email: olepais@st-pee.inra.fr.

Running title: SimRAD: predicting RAD and GBS loci number

Abstract

Application of high throughput sequencing platforms in the field of ecology and evolutionary biology is developing quickly since the introduction of efficient methods to reduce genome complexity. Numerous approaches for genome complexity reduction have been developed using different combinations of restriction enzymes, library construction strategies and fragment size selection. As a result, the choice of which techniques to use may become cumbersome, because it is difficult to anticipate the number of loci resulting from each method. We develop SimRAD, an R package that performs *in silico* restriction enzyme digests and fragment size selection as implemented in most restriction associated DNA polymorphism and genotyping by sequencing methods. *In silico* digestion is performed on a reference genome or on a randomly generated DNA sequence when no reference genome sequence is available. SimRAD accurately predicts the number of loci under alternative protocols when a reference genome sequence is available for the targeted species (or a close relative) but may be unreliable when no reference genome is available. SimRAD is also useful for fine-tuning a given protocol to adjust the number of targeted loci. Here, we outline the functionality of SimRAD and provide an illustrative example of the use of the package (available on the CRAN at http://cran.r-project.org/web/packages/SimRAD).

Introduction

Application of post-Sanger sequencing technologies in the field of ecology and evolutionary biology is developing at an unprecedented pace since the introduction of Restriction site Associated DNA (RAD) sequencing (Baird *et al.* 2008). RAD sequencing reduces genome complexity by targeting regions of specific restriction sites, thereby genotyping thousands of single nucleotide polymorphisms (SNPs) (Davey *et al.* 2011; Rowe *et al.* 2011). Such reduction of genome complexity has enabled multiplexing of tens to hundreds of individuals in a single run on high throughput platforms such as Illumina GAIIx and HiSeq (Glenn 2011). Since the advent of the first RAD protocol, several alternative approaches (collectively referred as genotyping by sequencing in the following, specific methods being named using acronyms) have been developed (Table 1), aiming to decrease cost and laboratory workload, and to increase flexibility in genome complexity reduction. Reducing cost and time needed for library construction has been the subject of continuous development (Elshire *et al.* 2011; Peterson *et al.* 2012; Poland *et al.* 2012; Toonen *et al.* 2013). In addition, the increasing availability of user-friendly analytical tools to handle genotyping by sequencing data (Catchen *et al.* 2011, 2013; Chong *et al.* 2012; Lu *et al.* 2013) will contribute to make genotyping by sequencing approaches more accessible. Alternative genotyping by sequencing approaches allow the user to adjust genome complexity reduction level, which has proven to be particularly useful for studying complex genomes (Elshire *et al.* 2011; Poland *et al.* 2012; Lu *et al.* 2013). Optimization of genome complexity reduction can be achieved by using more than one restriction enzyme (Peterson *et al.* 2012; Poland *et al.* 2012; Stolle & Moritz 2013), selecting fragments within a particular size range (Peterson *et al.* 2012; Toonen *et al.* 2013) and removing fragments containing complex AT or GC repeated regions (Stolle & Moritz 2013). Recent adaptation of genotyping by sequencing to new sequencing platforms (such as Life Technologies Ion Torrent) has also contributed to the development of new protocols (Stolle & Moritz 2013; Mascher *et al.* 2013).

As a result, confusion may arise as to which of the available protocols may be most appropriate for a given experimental design. In particular, selecting a cost effective method to sufficiently reduce the DNA library complexity of a given species so that a large number of individuals can be multiplexed in a single sequencing run is not straightforward. The first step is to decide how many loci and individuals will be needed to achieve sufficient statistical power for the question at hand, and how this translates into the number of libraries to prepare and sequencing runs to perform. While theoretical predictions of the number of loci expected using one restriction enzyme in the RAD protocol can be performed using a probabilistic approach (e.g. the RadCounter tool from The GenePool, Edinburgh, UK; Davey J.W., 2009, unpublished), the amount of genome complexity reduction by more than one restriction enzyme is far more difficult to estimate probabilistically, especially when combined with a fragment size selection step.

Here we present a new R package, SimRAD, a simulation-based tool that may be used to estimate the number of loci expected from the most common genotyping by sequencing approaches. SimRAD is useful for both broad scale comparison of alternative protocols as applied to a newly studied species and for fine-scale optimization of a given genotyping by sequencing protocol for that species. We present the general principles used in SimRAD to model genotyping by sequencing approaches, assess prediction quality and provide examples illustrating potential usages. Our example simulations highlight the wide-scale differences in the amount of complexity reduction allowed by the alternative genotyping by sequencing methods. In addition, we discuss the limit of a simulation-based predictive approach, especially the limited accuracy of predictive results obtained from randomly generated DNA sequence.

SimRAD workflow and functions

A subsample or the full reference genome sequence of a species (*ref.DNAseq,* SimRAD functions in italic) or a randomly generated DNA sequence (*sim.DNAseq*) can be used to simulate restriction enzyme digestion (*insilico.digest*), library construction (*adapt.select*) and fragment selection (*size.select*, *exclude.seqsite*) to predict the number of loci expected from different genotyping by sequencing approaches (Figure 1).

Data input

When reference sequences for a species are available (full or draft reference genome, Figure 1), the function *ref.DNAseq* can be used to load the sequences contained in a FASTA file. When a full reference genome is available, the entire genome can be easily analyzed by digesting each chromosome independently allowing prediction of the exact number of restriction sites in the genome and the genetic sequence of each loci. When a draft genome sequence is available, contigs are randomly concatenated to form a continuous DNA sequence as input data because they do not generally represent real entities in the genome (like chromosomes in a full reference genome), but rather DNA segments left apart due to technical reasons. This choice was made to avoid confounding DNA fragments originating from separate contigs and DNA fragments originating from restriction digestion. A randomly sampled fraction of the contigs can be loaded using specific parameter of *ref.DNAseq* to lower computational cost. In such a case, contig sub-sampling can be performed several times to estimate the accuracy of the prediction (Table 2 and script in Supporting Material File 1).

When no reference genome is available (Figure 1), the function *sim.DNAseq* can be used to randomly generate DNA sequence of a given length and percentage of GC content representative of the studied species genome. However, results from simulated versus reference genomes often do not agree closely (Table 2), probably due to the complexity of genome structure (e.g. repeat regions…), which is difficult to replicate in simulation. When using simulated DNA, we also suggest that predictions should be compared with results obtained from genome sequence of a related species expected to have similar genome structure as the target species (Table 2).

In silico digestion

Up to four restriction enzymes (current methods use only one or two) can be used to virtually digest the DNA sequence (*insilico.digest* function), with both the number of restriction sites and the DNA fragments flanked by each recognition site returned. Note that GBS (Elshire *et al.* 2011) which uses restriction enzymes that recognize a degenerated site, such as ApeKI (GCWGC), can be modeled by specifying the two alternative recognition sites as if two different enzymes were used (GCAGC and GCTGC). Methods that use digestion as the sole complexity reduction approach, such as RAD and GBS, are completed at this step. For methods that use additional complexity reduction steps, the resulting digested fragments can be kept in memory for subsequent analyses (Figure 1).

Library construction process

When more than one restriction enzyme is used to digest a genome, different types of fragments are generated that differ in flanking sequence motifs. This offers an additional opportunity to further reduce the number of targeted fragments during library construction. Depending of the type of adapter primers selected and their complementarities to the sticky ends left after restriction digestion, it is possible to select fragments flanked by two identical restriction sites (Stolle & Moritz 2013), or fragments flanked only by two different restriction sites (Peterson *et al.* 2012; Poland *et al.* 2012). These two different library construction strategies may result in a wide range of numbers of targeted loci and can be simulated using the *adapt.select* function (Figure 1). In addition, more stringent reduction of the number of targeted fragments can be achieve by selecting fragments flanked by only one of the two restriction enzymes used for digestion. This method is equivalent to removing fragments containing particular restriction sites (RESTseq, Stolle & Moritz 2013) and can be modeled using the *excluding.seqsite* function (Figure 1).

Fragment size selection

For ddRAD (Peterson *et al.* 2012), RESTseq (Stolle & Moritz 2013) and ezRAD (Toonen *et al.* 2013) methods, digested fragments within a specified size range are selected to further reduce the number of targeted loci (Table 1). The complexity reduction efficiency at this stage will strongly depend on the adequacy of the selected fragment size range and the distribution of the fragment lengths after restriction digestion. In such protocols, it is quite difficult to anticipate the distribution of the obtained digested fragments and hence, the effects of size selection parameters, such as minimum size, maximum size and range, are difficult to intuit. The function *size.select* (Figure 1) will help the user by plotting the distribution of the digested fragment lengths and indicate the number of loci within a specified size range (Figure S3), which will help optimizing the number of targeted loci.

Prediction quality

In this section we evaluate SimRAD efficiency by comparing simulation to real data and assessing the effect of different types of input data on prediction accuracy.

Comparing SimRAD in silico digestions to real data

We analyzed previously published genotyping by sequencing data using SimRAD (Table 2). In general, we found good concordance between observed and predicted number of loci using draft or full genome sequence with an overall median deviation of 16.2% (Table 2). Most notable differences may be explained by several non-mutually exclusive factors. First, an incomplete draft genome sequence or a reference genome from a related species may not be representative of the whole genome sequence of a targeted species (Table 2, for instance *Brassica napus* genome that may differ from *Brassia rapa* genome characteristics). Second, DNA methylation that is not accounted for in SimRAD may decrease the number of accessible restriction sites and thus the observed loci compared to predictions (*Apis mellifera* example in Table 2). Third, errors in laboratory procedure or data analysis may impact the number of observed loci (e.g. fragment size selection step performed using manual gel excision for *Patiria miniata* may be imprecise).

Effect of genome sequence knowledge on SimRAD predictions

Predictions made from randomly generated DNA sequence can be particularly unreliable with an overall median deviation from number of observed loci of 38% (Table 2). The deviation between observation and prediction decreases when additional genome sequence knowledge are used for predictions: from 26% with the use of a draft genome or a full genome of a related species to 14% when a full genome sequence is available for a species (Table 2). When contigs are not assembled into chromosomes in a draft genome sequence, sub-sampling of a portion of the contigs (10% or more; see Supporting Material Figure S1) leads to unbiased and precise estimation as showed by small standard deviation around the mean when sub-sampling is performed repeatedly (Table 2, Table 3). The effect of sub-sampling contigs on the estimation will, however, depend on the contig size distribution within the draft genome sequence and should be evaluated on a case by case basis by repeated analyses to insure congruent results (or using a repeated sub-sampling procedure as illustrated in the companion script Supporting Material File 1).

Illustrative application examples

Eels – *Anguilla anguilla* in Europe and *Anguilla rostrata* in America – are catadromous fishes that are unique biological models from a population genetics perspective, because they are nearly panmictic despite having extensive geographic distributions (Avise 2011). This combination of large range size and panmixia renders these species particularly interesting to the study of spatially varying selection (Gagnaire *et al.* 2012) and marker-based demographical reconstruction (Côté *et al.* 2013). Both species have experienced a global population collapse and are considered critically endangered. Genomic resources have recently increased for *A. anguilla* with a published draft genome representing about 84% of the estimated genome (Henkel *et al.* 2012). In addition, RAD sequencing using the EcoRI restriction enzyme produced a total of 422,634 loci, 82,425 of which were polymorphic. This number of loci is appropriate for development of SNP based tools, but may be too high for cost-effectively sequencing multiplexed libraries comprising a high number of individuals due to capacity limits of available sequencing platforms (Glenn 2011).

Here, we illustrate the use of SimRAD in *Anguilla anguilla* by predicting the number of loci expected under alternative genotyping by sequencing approaches and protocol parameters. The scenario consists of finding protocols to target approximately 50,000 loci to get a cost effective genotyping method that would necessitate approximately 1.5 million reads per individual to sequence each loci to an average depth of 30X, thus allowing for multiplexing around 48 individuals in a Life Technologies Proton P1 or Illumina GAiiX run or 192 individuals in an Illumina HiSeq 2500 run. The script used to run SimRAD for this example is available in full as a Supporting Material File 1 and can be used to reproduce the results and have a precise indication and contextual example on how to run SimRAD functions.

Broad scale example: comparing methods

The *A. auguilla* RAD experiment using digestion with the EcoRI restriction enzyme on 30 individuals, produced a total of 422,634 loci (Pujolar *et al.* 2013) which is very close to the number of loci predicted (mean: 441,113 and standard error: 4,253; Table 3) over 30 simulations using sub-samplings of 10% of the draft genome sequences. The range of the expected number of loci varies greatly across RAD and genotyping by sequencing methods (Table 3) from as few as a few thousand (RESTseq using TaqI and excluding fragments containing one of 5 restriction sites, followed by size selection of fragments with lengths between 240 and 290 bp) to millions in the case of GBS approach using ApeKI restriction enzyme (Table 3). Several alternative approaches could be selected to target approximately 50,000 loci in *A. anguilla*: (1) the classical RAD using SbfI, (2) ddRAD with PstI and MseI selecting fragments between 210 and 260 bp (the later approach could be further adjusted, see below), (3) RESTseq using TaqI and excluding fragment containing one of five restriction sites, followed by fragment size selection between 70 and 105 bp, or (4) RESTseq using TaqI and excluding fragments containing MseI restriction site and selecting fragment size between 240 and 290 bp (Table 2). The final choice should then take into account other constraints such as availability of sequencing platform and consumables, commercial offers and local technical expertise.

Fine scale example: fine tuning ddRAD protocols

Once an appropriate approach has been selected, it is then possible to adjust protocol parameters to achieve the targeted number of loci. For instance, among all examples above, the ddRAD approach using PstI and MspI followed by fragment size selection between 210 and 260 bp should yield approximately 38,000 loci. Modifying the fragment selection size range using the function *size.select* allowed finding alternative parameterizations that should result in the targeted number of loci. The first solution corresponds to selecting fragments between 120 and 170 bp (narrow size selection range setting using a Pippin Prep system) resulting in 50,214 fragments (Table 3; Supporting Information Figure S2A). The second solution consists in selecting fragments between 230 and 300 bp (wide size selection range setting using a Pippin Prep system) yielding 49,423 loci (Table 3; Supporting Information Figure S2B).

Other applications and limitations

Digestion in repetitive genomic regions

Sequencing repetitive fraction of the genome is particularly undesirable because highly repeated fragments can capture a high percentage of the total number of generated reads (Beissinger *et al.* 2013) and compromise SNP detection and genotyping. *In silico* digestion in SimRAD followed by fragment size distribution inspection (Supporting Material Figure S3) could be used to screen alternative enzymes for species with full reference genome sequences. Such screening could be perform as a first step to detect potential digestion within the repetitive fraction of the genome.

Taking advantage of R flexibility

Due to the fact that SimRAD is embedded within the R environment, it greatly benefits from R flexibility. For instance, several SimRAD functions can be joined within a single line of code to simulate the different steps needed for a given genotyping by sequencing method, saving computation time and memory usage. Alternatively, output from any individual function can be saved as a FASTA file for subsequent analysis (e.g. fragment mapping outside the R environment). In the spirit of R user-interface principle, the user can choose which information to retain for further analysis: the whole fragment sequences for use in downstream external analysis or just output of a fragment size distribution plot to help in restriction enzyme choice. Such flexibility should encourage creative uses of SimRAD that may result in unanticipated additional applications.

Limitations

While SimRAD can be used on any species, it is clear that accurate estimation can only be reached if some representative sequence of a genome is used as an input. Randomly generated sequence, even following a given GC content, is generally far from representing the realistic genome sequence of actual organisms. Therefore, predictions made from simulated DNA sequence should be taken with caution, as showed by the illustrative examples (Table 2 and Table 3). We strongly advise the user to only consider predictions from a randomly generated DNA sequence as a first preliminary step prior to additional analyses of reference genomes from related species or real digestion experiments to confirm results obtained *in silico*. In any case, laboratory experimental validation are the only reliable means to detect additional factors that may impact the number of obtained loci and that are not accounted for in SimRAD such as DNA methylation and presence of unreferenced repetitive genomic regions.

Conclusion

SimRAD should contribute to the spread of alternative genotyping by sequencing techniques in the field of ecology and evolutionary biology by providing a user-friendly predictive tool to help with the experimental design planning, protocol optimization and data analysis. The SimRAD package is available for R on the Comprehensive R Archive Network at http://cran.r-project.org/web/packages/SimRAD.

Acknowledgments

We would like to thank Franck Salin, Christophe Boury, Erwan Guichoux and Yec'han Laizet at the Genomic and Sequencing Facility of Bordeaux for constructive discussions that led to the development of SimRAD. Paul Hohenlohe and three anonymous reviewers provided valuable comments that improved previous versions of the manuscript. OL was supported by a Marie Curie Action Career Integration Grant of the European Union FP7-PEOPLE-2011-CIG grant agreement number 303526. JW was supported by an NSERC Discovery Grant

References

Avise JC (2011) Catadromous eels continue to be slippery research subjects. *Molecular Ecology*, **20**, 1317–1319.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Beissinger TM, Hirsch CN, Sekhon RS *et al.* (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, **193**, 1073–81.

Catchen J, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Chen X, Li X, Zhang B *et al.* (2013) Detection and genotyping of restriction fragment associated polymorphisms in polyploid crops with a pseudo-reference sequence: a case study in allotetraploid Brassica napus. *BMC Genomics*, **14**, 346.

Chong Z, Ruan J, Wu C-I (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics*, **28**, 2732–2737.

Côté CL, Gagnaire P-A, Bourret V *et al.* (2013) Population genetics of the American eel (Anguilla rostrata): FST = 0 and North Atlantic Oscillation effects on demographic fluctuations of a panmictic species. *Molecular Ecology*, **22**, 1763–1776.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davidson WS, Koop BF, Jones SJM *et al.* (2010) Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biology*, **11**, 403.

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One*, **8**, e62137.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Gagnaire P-A, Normandeau E, Côté C, Møller Hansen M, Bernatchez L (2012) The genetic consequences of spatially varying selection in the panmictic American eel (Anguilla rostrata). *Genetics*, **190**, 725–736.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Gonen S, Lowe NR, Cezard T *et al.* (2014) Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics*, **15**, 166.

Henkel C V, Burgerhout E, de Wijze DL *et al.* (2012) Primitive duplicate Hox clusters in the European eel's genome. *PLoS One*, **7**, e32231.

Hohenlohe P a., Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Houston RD, Davey JW, Bishop SC *et al.* (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics*, **13**, 244.

Lu F, Lipka AE, Glaubitz J *et al.* (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, **9**, e1003215.

Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. *PLoS One*, **8**, e76925.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, **7**, e32253.

Pujolar JM, Jacobsen MW, Frydenberg J *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, **13**, 706–714.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

Schmutz J, Cannon SB, Schlueter J *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Sonah H, Bastien M, Iquira E *et al.* (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One*, **8**, e54603.

Stolle E, Moritz R (2013) RESTseq – Efficient benchtop population genomics with RESTriction fragment SEQuencing. *PLoS One*, **8**, e63960.

The Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **443**, 931–949.

Toonen R, Puritz J, Forsman Z *et al.* (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Wang X, Wang H, Wang J *et al.* (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics*, **43**, 1035–1039.

Zimin A V, Delcher AL, Florea L *et al.* (2009) A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology*, **10**, R42.

Data Accessibility

SimRAD package, including user manual, is available for R on the Comprehensive R Archive Network at http://cran.r-project.org/web/packages/SimRAD. *A. anguilla* draft genome version 1 used in this manuscript is available online at http://www.zfgenomics.org/sub/eel.

Supporting Material

File 1: R script used to reproduce all results presented in this manuscript.

Figure S1: Effect of the proportion of draft genome sequence sub-sampling (from 1% to full genome analysis) on the precision and accuracy of the number of loci predicted for a ddRAD protocol using PstI and MspI and 210-260 bp fragment size selection on *Anguilla anguilla*.

Figure S2: Alternative ddRAD protocol parameterizations to target 50,000 loci at the *A. anguilla* genome scale using PstI and MspI enzyme combination with a narrow size selection range (A, 50 pb) and a wide size selection range (B, 70 pb), typically setup using a Pippin Prep ™ system. Light grey histogram represent the fragment size distribution following double digestion and adapter-based fragment selection; the red portion illustrates the size selected fragments. Each sub-figure illustrates the graphical output of SimRAD function *size.select*.

Figure S3: GBS fragment size distribution obtained by in silico digestion of *Bos taurus* whole reference genome using different restriction enzymes (ApeKI, PstI or EcoT22I) or combination (EcoT22I & PstI) illustrating digestion in repetitive genomic regions for ApeKI and the double digestion EcoT22I and PstI, as already obtained from real digestion experiment followed by Agilent BioAnalyzer 2100 analysis (see Supplemental Figure S1 in De Donato *et al.* 2013).

Author Contributions

OL and JTW programmed and documented the functions, OL performed the analyses, OL and JTW wrote the paper.

Figure Legends

Figure 1: SimRAD workflow. The succession of functions to apply (in rectangles) are indicated according to the availability of reference genome sequence and the genotyping by sequencing protocol to simulate (rounded rectangle in the right side). The functions *ref.DNAseq* and *sim.DNAseq* are used to input sequence data, *insilico.digest* performs *in silico* restriction enzyme digestion, *adapt.select* simulates library construction, *size.select* and *exclude.seqsite* perform fragment selection steps.

Tables

Table 1: Comparison of approaches used to reduce the number of loci targeted in common genotyping by sequencing methods.

Method	Restriction enzyme	Restriction exclusion	Size selection	Reference
RAD	1	n	n	(Baird *et al.* 2008)
GBS	1	n	n	(Elshire *et al.* 2011)
teGBS	2	n	n	(Poland *et al.* 2012)
ddRAD	2	n	y	(Peterson *et al.* 2012)
ezRAD	1 or more	n	y	(Toonen *et al.* 2013)
RESTseq	2 or more	y	y	(Stolle & Moritz 2013)

n: no; y: yes; RAD: Restriction site Associated DNA, GBS: Genotyping-By-Sequencing, teGBS: two-enzymes Genotyping-By-Sequencing, ddRAD: double digest Restriction site Associated DNA, ezRAD: easy Restriction site Associated DNA, RESTseq: Restriction fragment sequencing.

Table 2: Comparison of the number of loci predicted using SimRAD and reported in the literature. We reported mean and standard deviation (in parenthesis) over ten replicated sub-samplings of draft genomes, or alternatively exact figures form full genome in silico digestions.

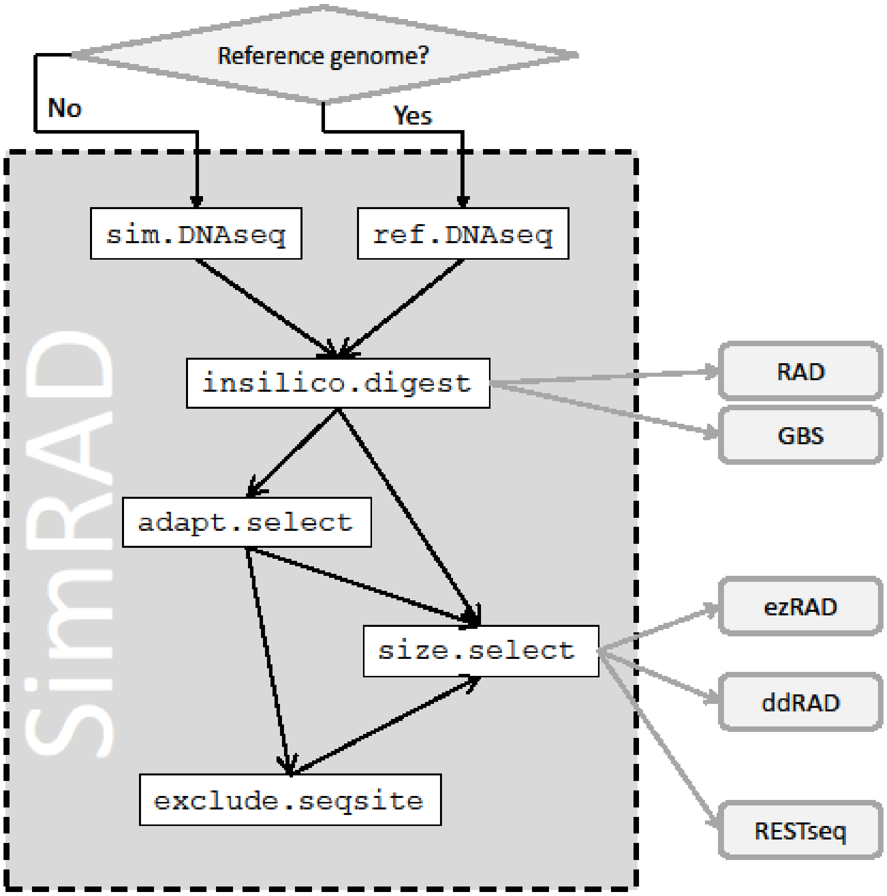
Species	Reference genome	Genome size (Mb)	%GC	Method	Reference	Restriction enzyme	Size selection	Restriction exclusion	N_{rep}	$N_{sim-0.1}$	$N_{ref-0.1}$	$N_{ref-full}$
Anguilla anguilla	draft[1]	1101.18	43.5	RAD	(Pujolar *et al.* 2013)	EcoRI			422,634	676,659 (3,711)	438,589 (4,935)	
Oncorhynchus spp.; *Salmo salar*	draft (*Salmo salar*)[2]	2435.04	42.6	RAD	(Hohenlohe *et al.* 2011; Houston *et al.* 2012; Gonen *et al.* 2014)	SbfI			98,190; 75,000	37,346 (1,172)	108,098 (3,258)	
Bos taurus	full[3]	2670.42	41.9	GBS	(De Donato *et al.* 2013)	PstI			1,400,000	433,235 (2,903)	1,700,448 (132,968)	1,584,146
Glycine max	full[4]	973.78	35.0	GBS	(Sonah *et al.* 2013)	ApeKI			800,000[a]	582,747 (6,305)	738,973 (63,545)	755,623
						MseI			9,500,000[a]	10,864,021 (9,870)	9,610,908 (342,364)	9,511,761
						PstI			100,000[a]	98,819 (4,007)	103,898 (27,291)	115,170
Patiria miniata	draft[5]	811.03	40.2	ezRAD	(Toonen *et al.* 2013)	MboI & Sau3A	280-380		635,376	323,259 (2,485)	186,043 (1,108)	
Brassica napus	full (*Brassica rapa*)[6]	1200[b]	35.4	ddRAD	(Chen *et al.* 2013)	SacI & MseI	141-420		180,991; 147,000[a]	179,723 (1,822)	119,283 (5,996)	
Apis mellifera	full[7]	250.29	34.1	RESTseq	(Stolle & Moritz 2013)	TaqI	70-105	MseI	131,732	114,465 (640)	204,716 (9,385)	176,883
						TaqI	155-195	MseI, MluCI, BstUI, MspI & HimP1I	2,250	3,946 (231)	3,009 (126)	2,637

N_{rep}: number of loci report in the literature; $N_{sim-0.1}$: predicted number of loci using randomly generated DNA sequence of length corresponding to a subsample of 10% of the draft genome sequences; $N_{ref-0.1}$: predicted number of loci using a 10% subsampled of the genome sequences; $N_{ref-full}$: predicted number of loci using the whole reference genome sequence. [a] number of loci expected from the reference genome; [b] reference genome available for *Brassica rapa* with a genome size of 283.98 MB, estimation are based on a total genome size of 1200 MB for *Brassica napus* (Chen *et al.* 2013). [1]: draft genome version 1 (Henkel *et al.* 2012, available at http://www.zfgenomics.org/sub/eel); [2]: ASM23337v1 AGKD00000000.1 (Davidson *et al.* 2010); [3]: UMD3.1

DAAA00000000.2 (Zimin *et al.* 2009); [4]: V1.1 ACUP00000000.1 (Schmutz *et al.* 2010); [5]: Pmin_1.0 AKZP00000000.1 (Liu et al., unpublished); [6]: Brapa_1.0 AENI00000000.1 (Wang *et al.* 2011); [7]: Amel_4.5 AADG00000000.6 (The Honeybee Genome Sequencing Consortium 2006).

Table 3: Illustrative application of SimRAD. We reported the number of predicted loci (mean and standard deviation over 30 simulations using sub-samples of 10% of the draft genome sequences) for various genotyping by sequencing methods in *Anguilla anguilla*. Methods yielding a number of loci close to the targeted 50,000 loci are highlight in bold. The two last lines in italic show predictions from the full draft genome sequence demonstrating a fine tuning of ddRAD protocols to reach the targeted number of loci.

Method	Restriction enzyme	Size selection	Restriction exclusion	Mean number of predicted loci (sd)
RAD	EcoRI			441,113 (4,253)
RAD	**SbfI**			**79,074 (1,319)**
GBS	ApeKI			2,340,917 (11,602)
teGBS	PstI and MspI			756,655 (3,911)
ddRAD	**PstI and MspI**	**210-260**		**37,549 (696)**
teGBS	EcoRI and MspI			343,410 (2,654)
ddRAD	EcoRI and MspI	210-260		15,374 (400)
ddRAD	EcoRI and MspI	200-270		21,595 (479)
ddRAD	**PstI and MseI**	**210-260**		**71,047 (690)**
ddRAD	PstI and MseI	200-270		100,516 (947)
RESTseq	TaqI	70-105	MseI	124,927 (1,460)
RESTseq	**TaqI**	**70-105**	**MseI, MliCI, HaeIII, MspI, HinP1I**	**47,140 (702)**
RESTseq	**TaqI**	**240-290**	**MseI**	**51,839 (723)**
RESTseq	TaqI	240-290	MseI, MliCI, HaeIII, MspI, HinP1I	3,932 (151)
ezRAD	MboI and Sau3A	280-380		188,184 (1,595)
ddRAD	***PstI and MspI***	***120-170***		***50,214***
ddRAD	***PstI and MspI***	***230-300***		***49,423***


Reference genome?
No
Yes
SimRAD
sim.DNAseq
ref.DNAseq
insilico.digest
adapt.select
size.select
exclude.seqsite
RAD
GBS
ezRAD
ddRAD
RESTseq