Variant calling and genotyping accuracy of ddRAD-seq: Comparison with 20X WGS in layers

Mathilde Doublet, Fabien Degalez, Sandrine Lagarrigue, Laetitia Lagoutte, Elise Gueret, Sophie Allais, Frédéric Lecerf

RESEARCH ARTICLE

Variant calling and genotyping accuracy of ddRAD-seq: Comparison with 20X WGS in layers

Mathilde Doublet[1], **Fabien Degalez**[1], **Sandrine Lagarrigue**[1], **Laetitia Lagoutte**[1], **Elise Gueret**[2], **Sophie Allais**[1☯], **Frédéric Lecerf**[1☯*]

1 PEGASE, INRAE, Institut Agro, Saint Gilles, France, **2** MGX-Montpellier GenomiX, Univ. Montpellier, CNRS, INSERM, Montpellier, France

☯ These authors contributed equally to this work.
* frederic.lecerf@institut-agro.com

Abstract

Whole Genome Sequencing (WGS) remains a costly or unsuitable method for routine genotyping of laying hens. Until now, breeding companies have been using or developing SNP chips. Nevertheless, alternatives methods based on sequencing have been developed. Among these, reduced representation sequencing approaches can offer sequencing quality and cost-effectiveness by reducing the genomic regions covered by sequencing. The aim of this study was to evaluate the ability of *double digested Restriction site Associated DNA sequencing* (ddRAD-seq) to identify and genotype SNPs in laying hens, by comparison with a presumed reliable WGS approach. Firstly, the sensitivity and precision of variant calling and the genotyping reliability of ddRADseq were determined. Next, the SNP Call Rate (CR_{SNP}) and mean depth of sequencing per SNP (DP_{SNP}) were compared between both methods. Finally, the effect of multiple combinations of thresholds for these parameters on genotyping reliability and amount of remaining SNPs in ddRAD-seq was studied. In raw form, the ddRAD-seq identified 349,497 SNPs evenly distributed on the genome with a CR_{SNP} of 0.55, a DP_{SNP} of 11X and a mean genotyping reliability rate per SNP of 80%. Considering genomic regions covered by expected enzymatic fragments (EFs), the sensitivity of the ddRAD-seq was estimated at 32.4% and its precision at 96.4%. The low CR_{SNP} and DP_{SNP} values were explained by the detection of SNPs outside the EFs theoretically generated by the ddRAD-seq protocol. Indeed, SNPs outside the EFs had significantly lower CR_{SNP} (0.25) and DP_{SNP} (1X) values than SNPs within the EFs (0.7 and 17X, resp.). The study demonstrated the relationship between CR_{SNP}, DP_{SNP}, genotyping reliability and the number of SNPs retained, to provide a decision-support tool for defining filtration thresholds. Severe quality control over ddRAD-seq data allowed to retain a minimum of 40% of the SNPs with a CcR of 98%. Then, ddRAD-seq was defined as a suitable method for variant calling and genotyping in layers.

OPEN ACCESS

Citation: Doublet M, Degalez F, Lagarrigue S, Lagoutte L, Gueret E, Allais S, et al. (2024) Variant calling and genotyping accuracy of ddRAD-seq: Comparison with 20X WGS in layers. PLoS ONE 19(7): e0298565. https://doi.org/10.1371/journal.pone.0298565

Editor: Elingarami Sauli, Nelson Mandela African Institute of Science and Technology, UNITED REPUBLIC OF TANZANIA

Received: January 26, 2024

Accepted: May 23, 2024

Published: July 26, 2024

Data Availability Statement: Data is deposited in the European Nucleotide Archive (ENA) with the accession number PRJEB58821 and PRJEB71464.

Funding: This research project was supported by the French National Research Agency ANR, within the framework of project ANR-10-GENOM_BTV-015 UtOpIGe, and by the French Institut national de la recherche agronomique et de l'environnement (INRAE), within the framework of the SelGen metaprogram. Mathilde Doublet is a PhD fellow supported by the French Institut national de la recherche agronomique et de l'environnement (INRAE) and L'institut Agro Rennes-Angers.

Competing interests: The authors declare that they have no competing interests.

Introduction

The development of Next-Generation Sequencing (NGS) approaches has revolutionized genetic marker discovery and genotyping. Depending on the chosen approach, the balance between the marker density, genotype accuracy, the degree of multiplexing of individuals and the experimental costs may vary. Among these approaches, whole genome sequencing (WGS) allows the simultaneous detection and genotyping of the majority of individual polymorphisms at a chosen sequencing depth [1–3]. However, it is still challenging mostly because of the sequencing costs per sample. Then, cheaper alternative methods such as low depth WGS have been studied [4, 5]. Compared to deeper WGS, low depth WGS offers a large panel of single nucleotide polymorphisms (SNPs) per sample while increasing inter-individual variability and lowering genotyping accuracy [2, 6]. Moreover, sequencing the whole genome of every individual in a population is often unnecessary, as many biological questions requirering genomic markers (population genetics, genomic selection, genetic diversity studies. . .) can be answered using only a subset of genomic regions [6]. Alternative approaches as SNP chips can be used to genotype only a subset of SNP [7]. But, compared to WGS, the selection of a subset of SNPs distributed equidistantly on the genome with maximal MAF values leads to ascertainment bias, causing issues in the interpretation of genetic diversity in the population [8–10].

Restriction site associated DNA sequencing (RAD-seq) is a great alternative to WGS. RAD-seq are *de novo* approaches targeting a subset of the genome, thus reducing its complexity and providing a reliable set of markers. Practically, these sequencing methods begin by an enzymatic digestion followed by a filtration step based on the size of the enzymatic fragments (EFs). Then, remaining EFs are amplified by PCR to create a library. The fragments of this library are then sequenced from each end. Depending on sequencing capabilities and fragment size, the central part may not be sequenced (The size filtration step helps to limit this gap). The diversity of restriction enzymes (REs) available and ways to combine them make RAD-seq methods versatile assay tools [6]. With a good reference genome, the reads can be mapped, which improves the proportion of markers shared between individuals [6, 11, 12].

The low proportion of shared markers between individuals is a common drawback of RAD-seq approaches. During DNA digestion, REs recognize a specific motif called a restriction site (RS) to cut the DNA. With ddRAD-seq, two different REs will recognize two different restriction sites. When a restriction site is methylated, an RE sensitive to methylation will not be able to access it, and the fragment will not be created. When a mutation occurs in a restriction site, this may be modified and the fragment will not be created either. A mutation can also be responsible for the creation of a restriction site and therefore the creation of a new fragment. These phenomena usually occur on only one of the 2 chromosomes. So, when the enzymatic fragment is sequenced, only one chromosome will be sequenced. The result will be incomplete genotyping, leading to an interpretation of a homozygous genotype.

Thus, mutations in a RS tends to misestimate genetic diversity within the population [13]. PCR bias are also a common bias of RAD-seq approaches leading to a miss estimation of genetic diversity [14]. Numerous studies have demonstrated methods to mitigate variant calling and genotyping errors from library preparation to bioinformatics processing of sequencing data [15].

To ensure the lowest possible error rate, all studies systematically apply quality control to variant calling and genotyping data [16]. The most common quality control filters are the call rate SNP (CR_{SNP}) [17, 18] and average sequencing depth per SNP (DP_{SNP}) [13, 19, 20]. These filters increase the chances for a SNP to be detected and that allelic frequencies will be well represented in the population. They also ensure to limit the impact of mutations in the RS and methylations on the final data set [13, 21]. Other filters, such as the probability that the allele frequency distribution respects the Hardy-Weinberg equilibrium (HWE) or the minor allele frequency (MAF) [11] are commonly used to identify and remove SNPs with a high genotyping error rate. The thresholds chosen for each of these filters are rarely justified in the bibliography [22] and depend greatly on the application of the study [6]. Some quality control filter thresholds are almost standardized, while others vary from study to study. However, in the case of sequencing methods based on genome reduction, losing genetic information, or wrongly assuming that all SNPs are correct after quality control, can have significant consequences for the conclusions of a study, depending on the application [15, 22–26]. But, because of the infinite number of events that cause genotyping errors, it is impossible to hope to get rid of them entirely [27]. If not, the studies recommend at least to quantify them in order to give a confidence interval to their results [26–28].

Variant calling and genotyping errors can be quantified through the introduction of replicate sample, by comparing the results of the two sequencings with each other. Variants that are not common to both replicates will be considered erroneous [12, 26, 28]. However, this method cannot identify genotyping errors due to mutations in RS [29] which is a major biases for RAD-seq approaches. Then, the best way to quantify the variant calling and genotyping quality of a sequencing method is to compare it to another more reliable reference method [27, 29]. The availability of sequencing data reliable enough to be considered as a reference representing the "truth", on the same genomic regions as the RAD-seq method, offers the possibility of calculating the sensitivity and accuracy of the RAD-seq method. Sensitivity corresponds to the ability of the RAD-seq method to detect all the SNPs detected by the reference method, in the genomic regions it covers. Precision, on the other hand, reflects the rate of loci wrongly considered as SNPs by RAD-seq. Sensitivity and accuracy are two indicators of the amount of variant calling and genotyping errors that are rarely found in the literature, and even less so for RAD-seq approaches [3, 20] although they have been reported in other studies [30, 31]. The reliability of sensitivity and precision measurements depends on the quality of the sequencing data used as a reference. In literature, genotypes obtained by a RAD-seq approaches have already been compared to considered “more reliable approaches” such as Sanger sequencing [32], SNP chips [19], or even WGS [29] but never for layers.

Among RAD-seq approaches, double digest Restriction site Associated DNA sequencing (ddRAD-seq) is a method which, thanks to the use of two different RE to digest DNA and a size filtration step for EFs, reduces the inter-individual variability of EFs generated and SNPs detected compared with other RAD-seq methods. It drastically reduces the rate of variant calling and genotyping errors compared with other RAD-seq approaches. The use of two REs also facilitates adapter design and reduces sequencing costs per individual and per base, thus offering the best multiplexing capability compared with other RAD-seq approaches [14, 24]. Moreover, ddRAD-seq is a sequencing method that can be customized (RE choice, filtering method) to suit the needs of each study [6, 24, 33]. This makes ddRAD-seq a reliable method for variant discovery and genotyping of plants [33–38] and animals [39–43]. But, as with other RAD-seq methods, there is no consensus in the literature on the quality control and its filters, according to application, species or protocol features. Furthermore, the thresholds chosen for quality control filters are often not justified or based on other studies. Their real impact on the quality of genotyping data is rarely studied.

But, despite its bias, ddRAD-seq represents a major opportunity for the poultry industry as a small number of markers is sufficient to perform for various applications such as genome-wide association studies [44], linkage disequilibrium calculation [45], CNV detection [46], genomic selection [47, 48] or new variant discovery. ddRAD-seq is cheaper than deep WGS or HD chips and most of the time more accurate than low depth WGS in animal species. Compared with LD chips, RAD-seq sequencing enables the integration of the whole genome including micro-chromosomes and new markers unavailable on the HD chip. These micro-chromosomes are not well represented on commercial chips. The markers present on the LD chips come from the Affimetrix HD commercial chip [45], which was designed at the time of the galGal4 version of the reference genome (Nov, 2011). At that time, many chromosomes were not fully sequenced. As a result, several micro-chromosomes do not show any SNPs on the HD chip.

Then, the aim of this study is to assess ddRAD-seq quality of sequencing in terms of variant calling and genotyping, for bi-allelic markers, according to a set of population scale filtering options by comparison with 20X WGS. The 20X WGS was taken as a reference, as an effective coverage of 15X is considered as sufficient to achieve high-quality genotyping for WGS [2], The three parts of this work are (i) to describe and compare the SNP calling and genotyping data between WGS and ddRAD-seq, more precisely, (ii) to estimate sensitivity and precision of ddRADseq SNP calling and finally, (iii) to study the genotype concordance of the common SNP between deep WGS and ddRAD-seq.

Results

Genome scale parameters

With the 20X WGS, 9,219,123 bi-allelic SNPs were detected on chromosomes 1 to 39 and Z and 51,050 on contigs. With ddRAD-seq, 349,497 bi-allelic SNPs were detected on chromosomes 1 to 39 and Z and 712 on contigs. There were 327,364 SNP common to both methods.

In both cases, at least half of the SNPs (60.9% in 20X WGS and 50.8% in ddRAD-seq) were located on the macro-chromosomes (1-5). Both methods obtained similar results regarding the percentage of SNPs detected on the intermediate chromosomes (6-10) with 15.4% and 16.3% for the 20X and ddRAD-seq respectively. On the contrary, 29.7% of the SNPs identified in ddRAD-seq were located on the micro chromosomes (11-39) against only 19.5% of those found in 20X WGS. Finally, 4.2% of the SNPs from 20X WGS and 3.2% of the SNPs from ddRAD-seq were found on the Z sexual chromosome (Table 1, Fig 1).

As shown in the Fig 1, the number of SNPs were similarly distributed on chromosomes in 20X WGS or in ddRAD-seq. The number of SNP identified on each chromosome was significantly correlated with the length of each chromosome in 20X WGS (ρ = 0.99) and in ddRAD-seq (ρ = 0.98). The mean distance between two adjacent SNPs called was 3,221 bp and 130 bp for ddRAD-seq and 20X WGS respectively.

Sensitivity and precision

In silico estimation of the genomic regions that should be covered by enzymatic fragments (EFs) theoretically generated by the ddRAD-seq protocol was performed as described in the *Materials and Methods*. So, 860,138 Pst1 restriction sites (RS) and 447,997 Taq1 RS were identified on the reference genome GRCg7b. Moreover, 33,600 Pst1 and 63,564 Taq1 RS were

Table 1. Number of SNPs called in 20X WGS and in ddRAD-seq and their percentage on the total number of SNPs by chromosome type.

	WGS 20X		*ddRAD-seq*	
	#SNPs	**%SNPs**	**#SNPs**	**%SNPs**
Macro (1–5)	5 611 994	60.9	177 658	50.8
Intermediate (6–10)	1 420 343	15.4	56 989	16.3
Micro (11–39)	1 803 092	19.5	103 639	29.7
Z	383 694	4.2	11 211	3.2

https://doi.org/10.1371/journal.pone.0298565.t001

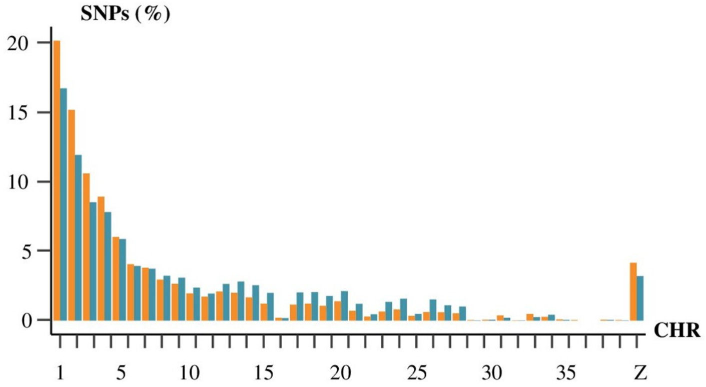

Fig 1. Percentage of total SNP detected by 20X WGS (*in orange*) and ddRAD-seq (*in blue*) on chromosomes 1 to 39 and Z.

https://doi.org/10.1371/journal.pone.0298565.g001

created by mutations in our population when considering every individual. A total of 285,004 EFs between 200 and 500 base pair (bp) should have been theoretically generated and pair-end sequenced on an average of 150 bp according to *in silico* prediction.

In 20X WGS, considered here as the reference sequencing approach, 638,135 of the 9,219,123 SNPs were identified inside the expected EFs. Therefore, we expected a sensitivity of 6.9% for ddRAD-seq at a genome-wide scale and a precision surrounding 100%. But, the real sensitivity of ddRAD-seq at a genomic scale was 3.6% and its precision was 93.7%. It represents a loss of sensitivity of 57.9% compared to what was expected.

Then, inside the expected EFs, 214,495 SNPs were identified by ddRAD-seq and 206,872 SNPs were commonly identified by ddRAD-seq and 20X WGS. Using 20X WGS as the reference approach, the effective sensitivity of ddRAD-seq inside the expected EFs was 32.4% and its precision was 96.2%. It represents an even greater loss of sensitivity (67,6%) compared to what was expected then at a genomic scale. This suggested that, in ddRAD-seq, some SNPs were detected outside the expected EFs. Considering the total number of 349,497 SNPs detected in ddRAD-seq, only 61.5% of them were found inside the expected EFs.

Locations of SNPs outside the EFs

The location of SNPs outside of the expected EFs (*i.e*., 200 to 500 bp framed by the two enzymatic restriction sites) was investigated as described in Materials and methods. Out of the 134,552 SNPs identified outside the EFs, 18,816 SNPs were found in the 10 bases on each end of these EFs. A total of 47,407 SNPs were located in regions covered by EFs less than 200 bp long but generated by the combination of Taq1 and Pst1. Also, 12 342 SNPs were located in genomic regions covered by EFs cut by Taq1 on both ends and 59 125 by Pst1 on both ends. Additionally, 119,116 SNPs were located in regions corresponding to EFs resulting from the failure of both sizing and selection on RE steps in the ddRAD-seq protocol. Genomic regions concerned by each scenario can overlap with one another. All the effectives of SNPs that could be identified by multiple scenarios overlapping was described in the Fig 2. Finally, considering each scenario and their overlaps, 4,265 SNPs called out of the EFs were not explained by any of these hypotheses.

For ddRAD-seq, the ratio between SNPs inside and outside the theoretical EFs was not different between the chromosomes except for the sexual chromosome Z, where there was as many SNPs inside and SNPs outside the theoretical EFs (Fig 3). There was no difference of location in specific chromosomic regions between the two sets of SNPs (S1 Fig).

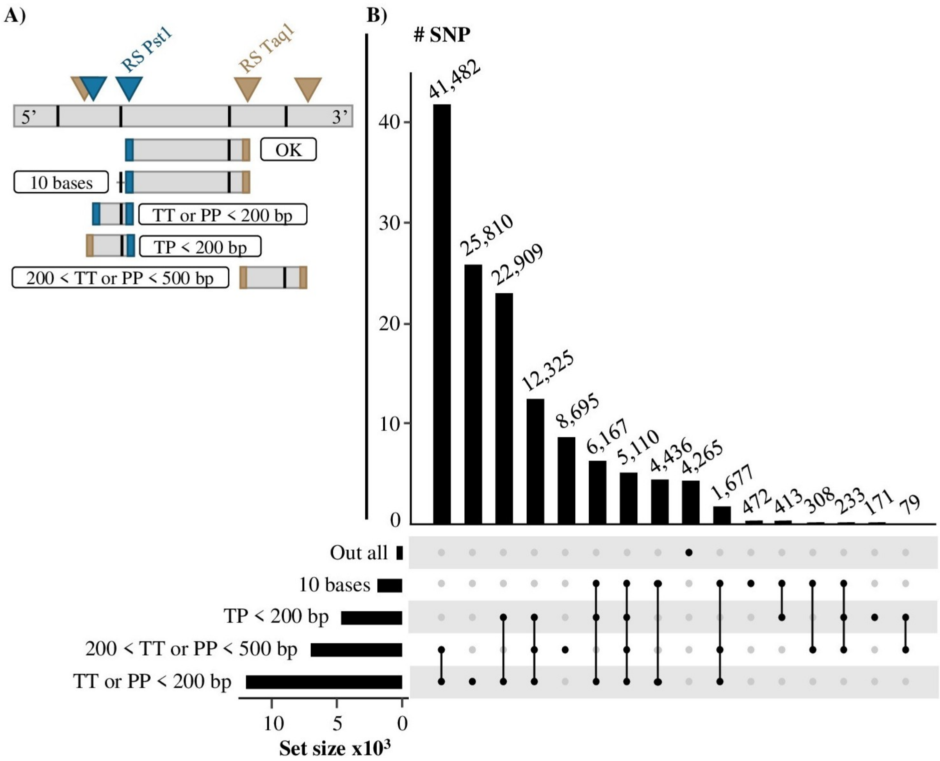

Fig 2. Possible reason for SNPs to be called outside of the theoretical enzymatic fragments theoretically generated by the ddRAD-seq protocol. 10 bases: SNPs called in the 10 bases on each side out of an enzymatic fragment (EF) between 200 and 500 bp and generated by the combination of Taq1 and Pst1. TP < 200 bp: SNPs located inside the EFs generated by the combination of Taq1 and Pst1 under 200 bp. 200 < TT or PP < 500 bp: SNPs located inside the EFs between 200 and 500 bp, generated by the same restriction enzyme at both ends. TT or PP < 200 bp: SNPs located inside the EFs under 200 bp and generated by the same RS on both sides. Out all: SNPs that don't fit in any of our scenarios.

https://doi.org/10.1371/journal.pone.0298565.g002

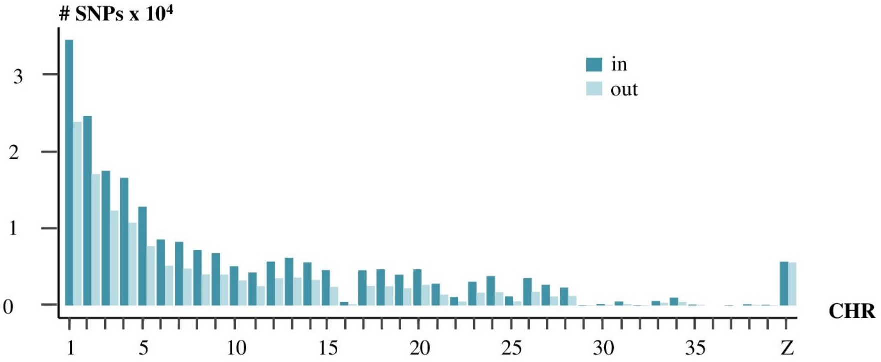

Fig 3. Localization of SNPs in (*in blue*) and out (*in light blue*) genomic regions theoretically sequenced in ddRAD-seq.

https://doi.org/10.1371/journal.pone.0298565.g003

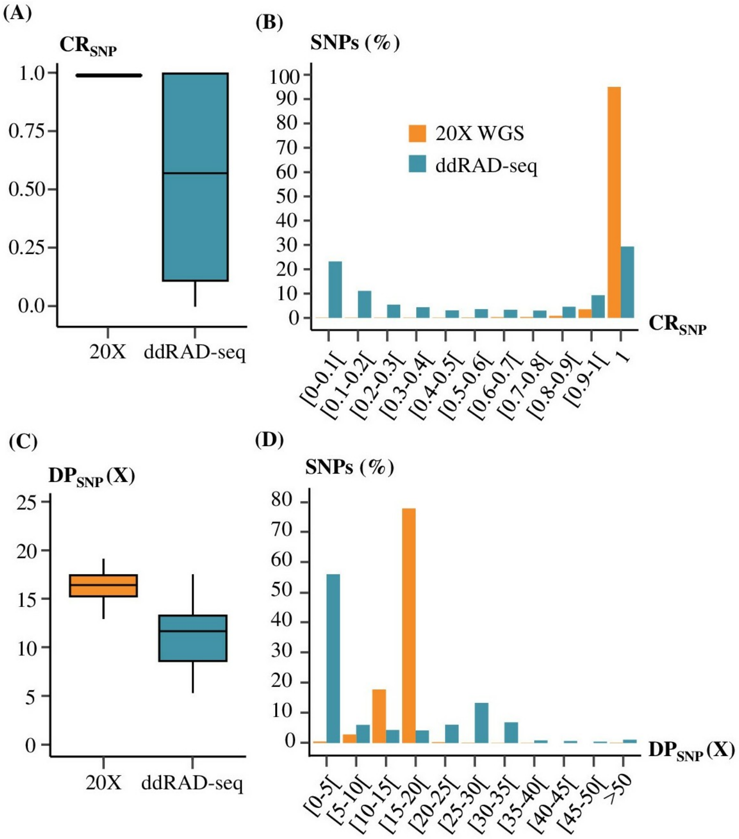

Fig 4. Comparison of variant calling results between 20X WGS (*in orange*) and ddRAD-seq (*in blue*). (A) Distribution of CR_{SNP} values for 20X WGS and ddRAD-seq. (B) Percentage of SNPs per Call Rate SNPs (CR_{SNP}) categories. (C) Distribution of DP_{SNP} values for 20X WGS and ddRAD-seq. (D) Percentage of SNPs per SNPs mean depth of sequencing (DP_{SNP}) categories.

https://doi.org/10.1371/journal.pone.0298565.g004

Distribution of CR_{SNP}, DP_{SNP} and MAF values

The CR_{SNP}, on average, was significantly higher in 20X WGS (0.99) than in ddRAD-seq (0.55) (t-test p-value < 0,05, Fig 4A). 95.2% of the SNPs obtained with the 20X WGS were genotyped for all individuals (CR_{SNP} = 1) while only 29.3% were in ddRAD-seq (Fig 4B). 56.8% of the ddRAD-seq SNPs had a CR_{SNP} below 0.8 against only 0.9% of the 20X WGS SNPs. Furthermore, the parabolic distribution of CR_{SNP} values in ddRAD-seq showed two higher points, with a large proportion of SNPs having low CR_{SNP} values (0–0.1) and an equivalent proportion having higher CR_{SNP} values (Fig 4B).

Similarly to CR_{SNP}, the average DP_{SNP} in ddRAD-seq was lower (11X) than the average DP_{SNP} observed in 20X WGS (16X) and even lower than the average DP_{SNP} expected (~45X, Fig 4C). Theoretically in ddRAD-seq, 285,004 EFs between 200 and 500 bp should be generated and pair-end sequenced on an average of 150 bp. Therefore, we estimated that 85.5 Mb should be covered by ddRAD-seq. In laying hens, the genome size is 1.26 Gb [49] which means that we expect a mean DP_{SNP} close to 45X in ddRAD-seq (450 Gb per Novaseq 6000 flowcell with 120 individuals per flowcell).

The distribution of SNPs in DP_{SNP} categories of ddRAD-seq showed two peaks of density: one at low DP_{SNP} values (0-5X) and one at higher values (25-30, Fig 4D). It was observed that

25% of the deepest ddRAD-seq DP_{SNP} values were superior to 24X compare with 17X in 20X WGS (Fig 4D). 56.0% of the ddRAD-seq SNPs were genotyped with less than 5X on average, against 0.5% in 20X WGS (Fig 4D).

The mean MAF was similar between 20X WGS (0.18) and ddRAD-seq (0.19). The distribution of MAF values was not different between ddRAD-seq and 20X WGS (S2 Fig). Individuals mean depth of sequencing (DP_{ind}) were 16X for 20X WGS and 11X for ddRAD-seq. Individual call rate (CR_{ind}) were respectively 99.4% for 20X WGS and 55.4% for ddRAD-seq.

Variant calling parameters at EFs scale

For the SNPs inside and outside of the expected EFs, the CR_{SNP} and the DP_{SNP} were calculated. These two parameters were lower for the set of SNPs outside the theoretical EFs then for the SNPs inside the expected EFs (Fig 5). Mean CR_{SNP} was 0.74 for SNPs inside the expected EFs and 0.25 for SNPs outside. The standard deviation of CR_{SNP} was 0.35 for ddRAD-seq SNPs inside and 0.27 for SNPs outside the expected EFs. These values appeared consistent with the peaks observed at a genomic scale (Fig 4B).

Mean DP_{SNP} values were respectively 17X and 1X for SNPs inside and SNPs outside the expected EFs (Fig 5). The standard deviation of DP_{SNP} for ddRAD-seq SNPs inside the expected EFs was 15X while it was 3X for SNPs outside. These DP_{SNP} values also corresponded to those observed for each peak at a genomic scale (Fig 4D).

The MAF was similar between 20X WGS (0.18) and ddRAD-seq (0.19). The distribution of MAF values was not different between ddRAD-seq and 20X WGS (S2 Fig). Individuals mean depth of sequencing (DP_{ind}) were 16X for 20X WGS and 11X for ddRAD-seq. Individual call rate (CR_{ind}) were respectively 99.4% for 20X WGS and 55.4% for ddRAD-seq.

ddRAD-seq genotyping reliability

The concordance of ddRAD-seq genotypes with 20X WGS was assessed by comparing, when it was possible, ddRAD-seq genotypes to 20X WGS ones. Genotypes were said *comparable*

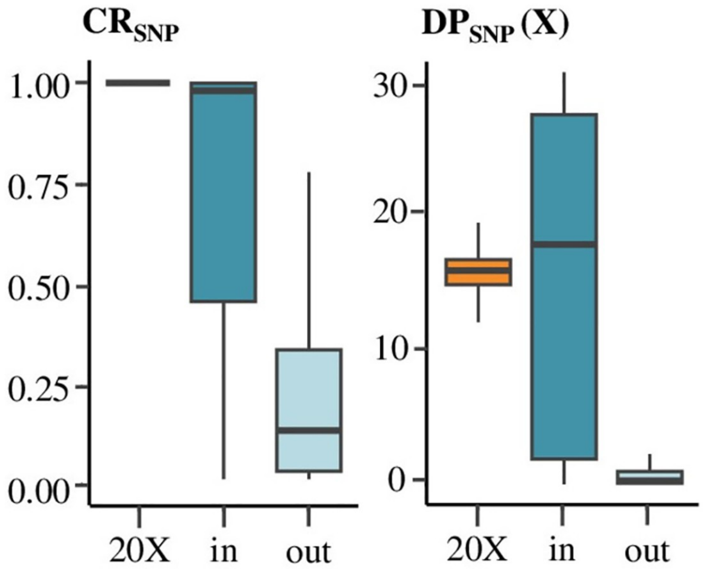

Fig 5. Distribution of CR_{SNP}, DP_{SNP} between SNPs in (*in blue*) and out (*light blue*) of enzymatic fragments for ddRAD-seq compared to 20X WGS (*in orange*).

https://doi.org/10.1371/journal.pone.0298565.g005

when, for an individual, the SNP was genotyped by both ddRAD-seq and 20X WGS. Then, 9,398,316 genotypes were identified as comparable.

When comparable genotypes were identical between ddRAD-seq and 20X WGS, ddRAD-seq genotypes were said *concordant* with 20X WGS ones. Globally, 90.0% of the comparable genotypes from ddRAD-seq and 20X WGS were concordant. Most of these concordant genotypes (76.7%) were those of SNPs located inside the expected EFs. The other 13.3% of the concordant genotypes were those of SNPs located outside the expected EFs.

Discordant genotypes between ddRAD-seq and 20X WGS represented 10.0% of the total number of comparable genotypes. Among these discordant genotypes, 6.1% were genotypes from SNPs located inside the expected EFs. The other 3.8% of the discordant genotypes were from SNPs located outside the expected EFs.

The majority (7,784,818) of all comparable genotypes were those of SNPs located inside the EFs expected to be sequenced by the ddRAD-seq protocol (Table 2). Genotypes were more concordant between ddRAD-seq and 20X WGS for SNPs inside the expected EFs (92.6%) then for SNPs outside the expected EFs (77.7%).

We also observed that the concordance between genotypes in ddRAD-seq and in 20X WGS, and the DP_{GT} of those genotypes were linked. Whether inside or outside expected EFs, concordant genotypes between ddRAD-seq and 20X WGS had greater DP_{GT} (15X) than discordant genotypes (7X, Fig 6A).

Then, for a SNP, the number of individuals with identical genotypes between ddRAD-seq and 20X WGS were quantified to calculate a concordance rates per SNP (CcR). The mean CcR for all the SNPs was 80.0%. The mean CcR was 87.3% for SNPs inside the expected EFs and 72.7% for SNPs outside the expected EFs. The correlation of CcR with the CR_{SNP} and the DP_{SNP} was investigated. High CcR values were associated with SNPs with high CR_{SNP} and DP_{SNP} values (Fig 6B). Most of the SNPs with high CcR, CR_{SNP} and DP_{SNP} values were located inside the expected EFs. Inside the expected EFs, 27.8% of the SNPs were correctly genotyped for all individuals (CR = 1 and CcR = 1) with a mean DP_{SNP} of 29X. Outside the expected EFs, only 0.1% of the SNPs were genotyped with a DP_{SNP} of 26X (Fig 6B).

Finally, ddRAD-seq SNPs were filtered according to multiple combinations of DP_{SNP} and CR_{SNP} threshold. Afterwards, the mean CcR for remaining SNPs and the percentage of ddRAD-seq SNPs kept after applying these filters were calculated (Fig 6C). This multi-criteria filtering approach (DP_{SNP}, CR_{SNP}) makes it possible to assess the number of SNPs retained depending on the objectives of genotyping reliability. For example, by filtering ddRAD-seq SNPs according to a threshold of 5X for the DP_{SNP} and a CR_{SNP} of 0.8 in ddRAD-seq, approximately 40–50% of ddRAD-seq SNPs will be retained, and 95.1% of these retained SNP will have a concordant genotype with 20X WGS. With the application of QC filters on ddRAD-seq data (CR_{SNP} and DP_{SNP}) with the highest thresholds possible, poor-quality SNPs can be eliminated by retaining a minimum of 27% genotyped SNPs for all individuals with 98% of reliable genotypes.

Table 2. Effectives of comparable genotypes (GTs) according to the location of the associated SNPs, inside or outside the expected enzymatic fragments (EFs) theoretically generated by the ddRAD-seq protocol. The effectives of ddRAD-seq GTs matching (concordant) or not matching (discordant) with 20X WGS genotypes, when compared one by one for an individual, for a SNP.

	Inside the expected EFs		*Outside the expected EFs*	
	Concordant	*Discordant*	*Concordant*	*Discordant*
Number of comparable GTs	7,209,100	575,718	1,253,604	359,894
TOTAL	7,784,818		1,613,498	
	9,398,316			

https://doi.org/10.1371/journal.pone.0298565.t002

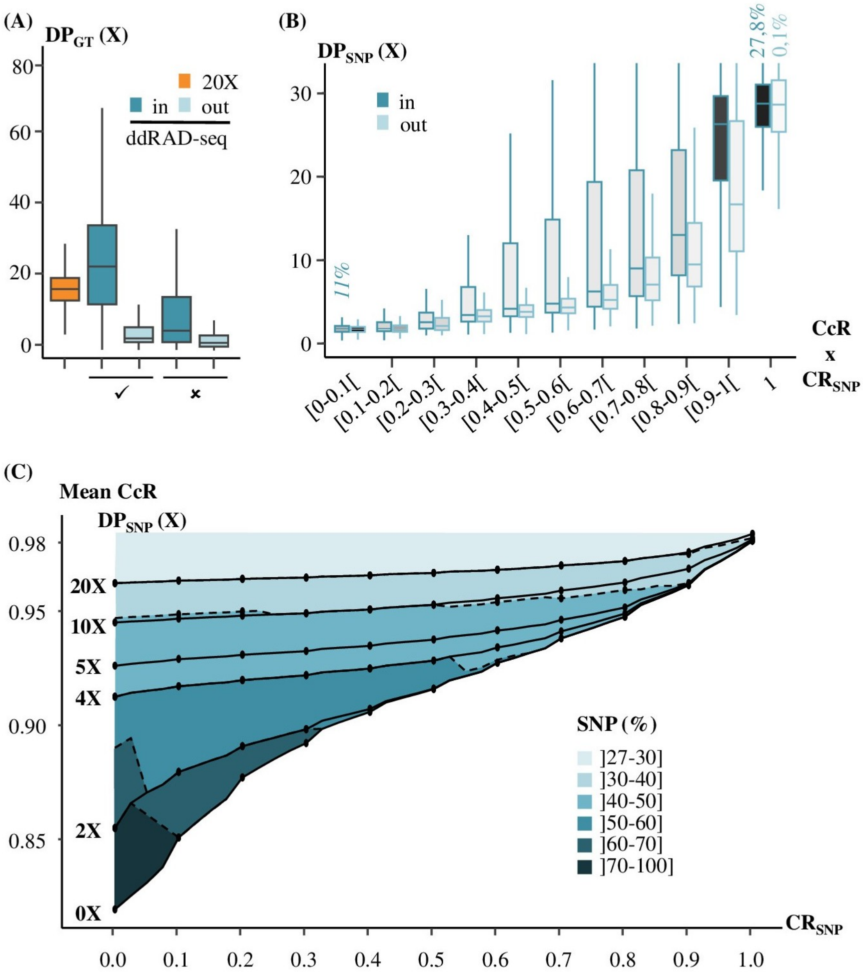

Fig 6. Genotyping quality of ddRAD-seq is correlated with the CR$_{SNP}$ and the DP$_{SNP}$. (A) Distribution of genotypes sequencing depth (DP$_{GT}$) for 20X WGS, the reference, (in orange) and concordant (✓) or discordant (✗) genotypes in ddRAD-seq for SNPs in the enzymatic fragments theoretically generated by the ddRAD-seq protocol (in blue) or out the EFs theoretically generated by the ddRAD-seq protocol (in light blue). Percentages of each category of ddRAD-seq genotype on the whole amount of genotype was displayed (grey shades). (B) Distribution of SNPs sequencing depths (DP$_{SNP}$) according to their category of genotype concordance rate (CcR) and SNP call rate (CR$_{SNP}$). (C) Mean CcR of ddRAD-seq according to the CR$_{SNP}$ and the mean sequencing depth threshold. The blue gradient represents the proportion of ddRAD-seq SNPs kept according to each filter combination.

https://doi.org/10.1371/journal.pone.0298565.g006

Discussion

The aim of the study was to assess the variant calling and genotyping quality of ddRAD-seq in laying hens, as an alternative to low-density sequencing methods (LD chip or low-depth WGS). Due to the diverse nature of ddRAD-seq protocols (enzyme pairs, size filtration method, and bioinformatics processing), estimating the quality of our variant calling and genotyping data by comparing them to the literature was challenging. Therefore, the results from ddRAD-seq were compared to results from 20X WGS obtained for the same individuals. Prior research has demonstrated that comparing a test sequencing method to a reference method allows for estimating the quality of variant calling and genotyping [29]. Additionally, for WGS, an effective coverage of 15X is considered as sufficient to achieve high-quality genotyping [2].

With an observed average DP_{SNP} of 16X for the 20X WGS, it was deemed a reliable basis for comparison with ddRAD-seq and representative of reality.

Initially, the study revealed that ddRAD-seq exhibited high precision (93.7%). Literature reports precision rates for RAD-seq studies ranging around 90 to 94% [29]. The SNPs were well-distributed across all chromosomes, including micro-chromosomes. This represented an advantage for ddRAD-seq compared to commercial chips that do not provide information on all micro-chromosomes. Also, the MAF distribution showed that unlike the SNP chips, ddRAD-seq isn't affected by an ascertainment bias. The genome-scale sensitivity of ddRAD-seq at 3.6% was comparable to similar methods found in the literature (0.5 to 5.5% depending on enzyme pairs) [20].

However, at the scale of the theoretically expected enzymatic EFs by the protocol in our population, a significantly lower sensitivity of ddRAD-seq than expected was observed (32.4%). Half of the SNPs detected by ddRAD-seq were located outside the expected EFs, indicating that the sequenced EFs did not correspond to reality. This discrepancy is due to in silico estimations based on the known reference genome in the literature [6, 50]. To approach the actual digestion results and leveraging data from the 20X WGS, restriction sites created and destroyed by mutations were integrated from the outset. These mutations, documented as significant sources of variability in generated EFs among individuals, were described in the literature. It is also acknowledged that the quality of the reference genome significantly impacts in silico simulation of EFs [37]. To mitigate this known bias, the latest version of the reference genome was used [51]. Despite these precautions, a substantial difference between the expected REFs, based on protocol descriptions, and reality was observed.

Some SNPs were found in genomic regions covered by EFs smaller than 200 bp, despite the protocol's filtration step. Literature describes that for ddRAD-seq, depending on the enzyme pair and size filtration method chosen, a significant difference in size distribution between the expected and sequenced EFs. This holds particularly true when using a 4-base cutter and a 6-base cutter as the enzyme pair, as in our case [50]. Therefore, it is not surprising to find some SNPs in genomic regions covered by EFs smaller than 200 bp with ddRAD-seq.

We also noted a small proportion of SNPs (14.0%) located outside the anticipated EFs, within a 10-base proximity of the expected EFs. Drawing from existing literature, we hypothesized that these might have originated from degraded DNA fragments. Specifically, ddRAD-seq is highly sensitive to DNA quality among RAD-seq methods [12], known to significantly influence enzymatic digestion efficiency and susceptibility to UV light exposure [15]. Some EFs may have been cleaved by a restriction site on one end and, despite the use of sticky end sites during adaptor ligation, improperly bound to these adaptors on the other end. Usually, after adaptor linking, EFs proceed through the rest of the protocol for amplification and sequencing. Based on previous observations in the literature, it's plausible that despite the protocol's trimming step, certain EFs smaller than twice the size of a read (~300 bp) might have been sequenced in pair-end, potentially causing partial adaptor contamination [50].

Among the remaining SNPs that didn't align with expected EFs or the previously described scenarios, some were genotyped by ddRAD-seq in genomic regions corresponding to EFs generated by the same restriction enzyme on both ends. This suggests that some EFs weren't appropriately filtered to have two distinct adaptors at each end. According to the protocol, for an EF to be sequenced, it must be produced by the combination of Taq1 and Pst1. The majority of SNPs outside the EFs were situated in regions that corresponded to EFs cut by the same restriction site at both ends and were less than 200 bp long. We thus inferred that, within our ddRAD-seq protocol, the adaptor filtering step might not have been completely efficient. Several studies have also observed this phenomenon [52, 53]. These studies describe the presence of EFs generated exclusively by the first introduced RE in the protocol but not by the second one. In our case, we observed SNPs located in both Taq1-Taq1 and Pst1-Pst1 EFs with more Pst1-Pst1. Regarding the frequency of Taq1 and Pst1 RS on the reference genome, we hypothesized that the difference between the number of SNPs located in Taq1-Taq1 EFs or Pst1-Pst1 EFs was more due to the larger number of Pst1 RS. It is possible to assess the efficacy of the adaptor filtration step for EFs using methods like qPCR, yet very few studies do so [37, 53], and they didn't report the results. Similar to most studies employing ddRAD-seq, no quality control for this step was performed in our study. Therefore, it's conceivable that the retention of a portion of EFs generated by a single restriction enzyme despite the filtration step is a generally acknowledged characteristic of ddRAD-seq.

Nevertheless, the random nature of previously cited bias leading events, among the pool of samples, should lead to average CR_{SNP} and DP_{SNP} values lower than those of the SNPs genotyped in the regions covered by the expected EFs. Globally, this would result in a loss of CR_{SNP} and DP_{SNP} at a genome-wide scale [15].

SNPs outside the expected EFs exhibited lower CR_{SNP} and DP_{SNP} compared to SNPs within the EFs, impacting the overall average values of CR_{SNP} and DP_{SNP} more negatively than anticipated. The sequencing depth intended for ddRAD-seq, originally allocated to specific regions, was spread across a larger area than expected. The theoretical calculation of wrongly assumed that only the theoretical EFs had been sequenced by the protocol. As a result, the average DP_{SNP} value decreased from the expected 45X to an observed 11X. Given that a minimum of 30X is recommended for ddRAD-seq, based on a reference genome, to ensure comprehensive genotyping in all individuals, the decrease in CR_{SNP} for SNPs called by ddRAD-seq was expected [22]. As DP_{SNP} and sensitivity are correlated [1, 54], this decline in DP_{SNP} is responsible for the low sensitivity observed at the expected EFs. Considering that low CR_{SNP} and DP_{SNP} values can impact genotyping reliability [55], it was hypothesized that genotyping SNPs outside the expected EFs might be less reliable than those within.

Upon comparison, the genotype concordance between ddRAD-seq and 20X WGS at 90% was highly satisfying. SNPs with high rates of erroneous genotypes (CcR) were indeed associated with lower CR_{SNP} and DP_{SNP} values than reliable genotypes.

Having genotype data for both ddRAD-seq and 20X WGS was a significant asset for our study, allowing for a detailed individual-level analysis of ddRAD-seq genotyping by pairwise genotype comparison. Quantifying genotyping errors enabled a thorough examination of the most common quality control filters' impact on genotyping reliability. Many studies apply consensus threshold filters from the literature without quantifying their impact on genotyping reliability, which, depending on the applications, can significantly affect study conclusions [15, 26].

Our study described the relationship between genotyping reliability and commonly used quality control filters for ddRAD-seq data (CR_{SNP} and DP_{SNP}), while measuring their impact on the retained SNP quantity. Fig 6C allows for comparison among different quality control scenarios' impact on genotyping reliability in terms of genotype reliability and SNP quantity. It offers the opportunity to establish decision rules on quality control thresholds tailored to each study's needs.

Applying the most stringent quality control filters on CR_{SNP} and DP_{SNP} still allow for the detection a reasonable number of markers. According to the literature, this marker count is largely sufficient for various applications in laying hens [39] such as genome-wide association studies [44], linkage disequilibrium calculation [45], CNV detection [46] or genomic selection [47]. Hence, ddRAD-seq proves to be a reliable tool for laying hen genotyping, offering a superior number of SNPs with reliable genotypes, evenly distributed across the entire genome, making it a compelling alternative to LD chips and LD WGS.

Materials and methods

Ethics approval

All blood samples were carried out as part of the commercial and selection activities of Novogen. These animals studied and the scientific investigations described herein are therefore not to be considered as experimental animals per se, as defined in EU directive 2010/63 and subsequent national application texts. Consequently, we did not seek ethical review and approval of this study as regarding the use of experimental animals. All animals were reared in compliance with national regulations pertaining to livestock production and according to procedures approved by the French Veterinary Services.

Animals

All animals consisted in a commercial pure line of laying hen of Rhode Island. This line was created and selected by *Novogen* (Plédran, France). The population studied was constituted of 50 roosters from the same generation, bred in individual cages.

Whole genome sequencing

All 50 individuals were sequenced by the Genomics and Transcriptomics platform GeT-PlaGe (Toulouse, France) with the Illumina HiSeq2000 technology expecting a global coverage of 20X Firstly, 38 individuals were sequenced as part of UtOpIGe project. Secondly, 12 individuals from the project OptiSeq were sequenced. These individuals were chosen because they have been selected as breeders for further generation. Data were aligned to the GRCg7b chicken reference genome [51] with BurrowsWheeler Aligner V0.7.15 [56] with default parameters for paired-end alignment. SNP calling was performed with GATK V3.7 [57]. Bi-allelic SNP have been extracted with the SelectVariant function and the "—restrictAllelesTo BIALLELIC" option. Remaining SNP have been filtered using the "VariantFiltration" option and hard filters for DNA-sequencing "FS > 60.0", "QD < 2.0", "MQ < 40.0", "MQRankSum < -12.5", "ReadPosRankSum < -8.0" and "SOR > 3.0".

ddRAD-sequencing

The same 50 individuals have been also sequenced with the ddRAD-seq technology as described in [14] by the **Montpellier GenomiX facility (MGX, France)**. Enzymatic digestion was performed using enzymes Taq1-v2 and Pst1-HF (New Englands Biolabs, 1assachusetts, USA), in agreement with the simulation results of Herry et al (2023). Only fragments ranging from 200 to 500 bp were selected as it is the appropriate length for sequencing fragments with **Illumina's** sequencing systems **and more precisely the Novasesq 6000** [58]. Mapping and variant calling were carried out in the same way as for the WGS sequences with the exception of the HaplotypeCaller module of GATK V3.7: -drf DuplicateRead argument, which was added to keep duplicated reads, as it is one of the principles of ddRAD-seq method.

Identification of genomic regions theoretically covered by ddRAD-seq

Genomic regions defined by the ddRAD-seq protocol were estimated on the reference genome GRCg7b [51]. First, all Taq1 and Pst1 RS were identified on the reference genome thanks to R package *Biostrings* [59]. Then, RS created by mutations were identified using SNPs detected in 20X WGS as the list of possible mutations within our population. Then, EFs between 200 and 500 bp were generated and only the first and last 150 bp were kept. SNPs called in these regions by ddRAD-seq or 20X WGS were identified using bedtools v2.30.0.

SNPs called by ddRAD-seq outside of these regions were identified and their locations were studied. The location of theses SNPs regarding the position of RS on the genome was empirically observed using IGV 2.7.2. Hypothesis about the possible reasons of the identification of SNPs outside expected EFs were made based on these observations and comparisons with the literature. SNPs were then classified according to these hypotheses and counted using bedtools v2.30.0.

SNP calling summary

For both ddRAD-seq and 20X WGS, the number and the distribution of genotyped SNPs along the chromosomes were estimated and compared using R V4.0.4. The correlation between the length of the chromosomes and the number of SNPs found on them was performed with the method of spearman for both methods using R V4.0.4. The average distance between two adjacent SNPs in bp was calculated on the whole genome and for each chromosome.

Then, for both ddRAD-seq and 20X WGS, three filtering parameters at the population scale were computed: (*i*) the ratio between the number of individuals with a non-missing genotype for a SNP and the total number of individuals, called the SNP call rate (CR_{SNP}), (*ii*) the minimum allele frequency (MAF) in the population and (*iii*) the SNP sequencing depth (DP_{SNP}) at the population scale. The genotype sequencing depth (DP_{GT}) corresponds to the number of reads supporting a genotype. DP_{SNP} is the sum of each DP_{GT} divided by the total number of genotyped individuals for this SNP. CR_{SNP} and MAF calculations were performed using Plink V1.9 [60] and DP_{SNP} using VCFtools V0.1.16 [61] and R V4.0.4 [62].

ddRAD-seq variant calling sensitivity and precision

The variant calling sensitivity of ddRAD-seq was calculated as the number of SNP commonly called by ddRAD-seq and 20X WGS divided by the total number of SNPs called by 20X WGS. The variant calling precision of ddRAD-seq was calculated as the number of SNP commonly called by 20X WGS and ddRAD-seq divided by the total number of SNPs called by ddRAD-seq. Sensitivity and precision were also calculated considering only regions covered by expected EFs.

20X WGS and ddRAD-seq genotype concordance

First, the SNPs that were called with both 20X WGS and ddRAD-seq approaches were kept. The number and the repartition of these common SNPs on the chromosomes were analyzed. Then, individually, each genotype was compared between 20X WGS and ddRAD-seq. If one or both methods didn't allow to genotype the individual for a SNP, the genotypes were considered **incomparable,** and the SNP was excluded for this individual. Concerning the **comparable** genotypes, they were considered **concordant** when both alleles were the same between ddRAD-seq and 20X WGS. On the contrary, if one or two alleles were different for a genotype between ddRAD-seq and 20X WGS sequencing methods, they were considered **discordant**. For the concordant and discordant genotypes, DP_{GT} were compared between ddRAD-seq and 20X WGS. They were obtained using VCFtools V0.1.16. Moreover, the Concordance rate (CcR) was computed as the ratio between the number of concordant genotypes and the number of comparable genotypes for each SNP. The mean CcR was calculated on all the common SNPs. Finally, ddRAD-seq data were filtered according to multiple combinations of Mean DP_{SNP} and SNP CR_{SNP} threshold. The evolution of the mean CcR and the number of retained SNPs were studied under these conditions.

Supporting information

S1 Fig. SNP distribution across the chicken chromosome. For each type of chromosome category of the chicken genome (macro-chromosome, intermediate chromosome, and micro-chromosome), the SNP distribution of the ddRAD-Seq data (in blue) and 20X experiment (in orange) is displayed. The black bar represents the theorical restriction fragment location.
(PDF)

S2 Fig. MAF distribution between 20X (in orange) and ddRAD-Seq (in blue).
(TIF)

Author Contributions

Conceptualization: Sophie Allais, Frédéric Lecerf.

Data curation: Mathilde Doublet.

Formal analysis: Mathilde Doublet.

Investigation: Mathilde Doublet, Sophie Allais, Frédéric Lecerf.

Methodology: Mathilde Doublet, Fabien Degalez, Laetitia Lagoutte, Elise Gueret, Frédéric Lecerf.

Project administration: Sophie Allais, Frédéric Lecerf.

Resources: Laetitia Lagoutte.

Supervision: Fabien Degalez, Sandrine Lagarrigue, Sophie Allais, Frédéric Lecerf.

Validation: Sophie Allais, Frédéric Lecerf.

Writing – original draft: Mathilde Doublet.

Writing – review & editing: Fabien Degalez, Sandrine Lagarrigue, Sophie Allais, Frédéric Lecerf.

References

1. Wong LP, Ong RTH, Poh WT, Liu X, Chen P, Li R, et al. Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. The American Journal of Human Genetics. 2013; 92(1):52–66. https://doi.org/10.1016/j.ajhg.2012.12.005 PMID: 23290073
2. Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. Sci Rep. 2019; 9(1):1784. https://doi.org/10.1038/s41598-018-38346-0 PMID: 30741997
3. Szarmach SJ, Brelsford A, Witt CC, Toews DPL. Comparing divergence landscapes from reduced-representation and whole genome resequencing in the yellow-rumped warbler (Setophaga coronata) species complex. Molecular Ecology. 2021; 30(23):5994–6005. https://doi.org/10.1111/mec.15940 PMID: 33934424
4. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low-depth whole-genome sequencing in complex trait association studies. Bioinformatics. 1 août 2019; 35(15):2555–61. https://doi.org/10.1093/bioinformatics/bty1032 PMID: 30576415
5. Olofsson JK, Cantera I, Van de Paer C, Hong-Wa C, Zedane L, Dunning LT, et al. Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. Molecular Ecology Resources. 2019; 19(4):877–92. https://doi.org/10.1111/1755-0998.13016 PMID: 30934146
6. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. juill 2011; 12(7):499–510. https://doi.org/10.1038/nrg3012 PMID: 21681211
7. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 28 janv 2013; 14(1):59. https://doi.org/10.1186/1471-2164-14-59 PMID: 23356797

8. Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How array design creates SNP ascertainment bias. PLOS ONE. 30 mars 2021; 16(3):e0245178. https://doi.org/10.1371/journal.pone.0245178 PMID: 33784304

9. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution. 1 nov 2010; 27(11):2534–47. https://doi.org/10.1093/molbev/msq148 PMID: 20558595

10. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays. 2013; 35(9):780–6. https://doi.org/10.1002/bies.201300014 PMID: 23836388

11. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. févr 2016; 17(2):81–92. https://doi.org/10.1038/nrg.2015.28 PMID: 26729255

12. Maroso F, Hillen JEJ, Pardo BG, Gkagkavouzis K, Coscia I, Hermida M, et al. Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species. Marine Genomics. 2018; 39:64–72. https://doi.org/10.1016/j.margen.2018.02.002 PMID: 29496460

13. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, et al. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology. 2013; 22 (11):3165–78. https://doi.org/10.1111/mec.12089 PMID: 23110526

14. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLOS ONE. 2012; 7(5):e37135. https://doi.org/10.1371/journal.pone.0037135 PMID: 22675423

15. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Molecular Ecology Resources. 2015; 15(1):28–41. https://doi.org/10.1111/1755-0998.12291 PMID: 24916682

16. Pool JE, Hellmann I, Jensen JD, Nielsen R. Population genetic inference from genomic sequence variation. Genome Res. 2010; 20(3):291–300. https://doi.org/10.1101/gr.079509.108 PMID: 20067940

17. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010; 5(9):1564–73. https://doi.org/10.1038/nprot.2010.116 PMID: 21085122

18. Ulaszewski B, Meger J, Burczyk J. Comparative Analysis of SNP Discovery and Genotyping in Fagus sylvatica L. and Quercus robur L. Using RADseq, GBS, and ddRAD Methods. Forests. 2021; 12 (2):222.

19. Brouard JS, Boyle B, Ibeagha-Awemu EM, Bissonnette N. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genet. 2017; 18(1):32. https://doi.org/10.1186/s12863-017-0501-y PMID: 28381212

20. Magbanua ZV, Hsu CY, Pechanova O, Arick M, Grover CE, Peterson DG. Innovations in double digest restriction-site associated DNA sequencing (ddRAD-Seq) method for more efficient SNP identification [Internet]. Genomics; 2022 sept [cité 18 nov 2022]. Disponible sur: http://biorxiv.org/lookup/doi/10.1101/2022.09.06.506835 PMID: 36481242

21. Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, et al. GBStools: A Statistical Method for Estimating Allelic Dropout in Reduced Representation Sequencing Data. PLOS Genetics. 2016; 12(2): e1005631. https://doi.org/10.1371/journal.pgen.1005631 PMID: 26828719

22. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology. 2013; 22(11):3151–64. https://doi.org/10.1111/mec.12084 PMID: 23110438

23. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 15 oct 2014; 30(20):2843–51. https://doi.org/10.1093/bioinformatics/btu356 PMID: 24974202

24. Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE. Demystifying the RAD fad. Molecular Ecology. 2014; 23(24):5937–42. https://doi.org/10.1111/mec.12965 PMID: 25319241

25. Meirmans PG. Seven common mistakes in population genetics and how to avoid them. Molecular Ecology. 2015; 24(13):3223–31. https://doi.org/10.1111/mec.13243 PMID: 25974103

26. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you've looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology. 2018; 27(16):3193–206.

27. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12(6):443–51. https://doi.org/10.1038/nrg2986 PMID: 21587300

28. Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. How to track and assess genotyping errors in population genetics studies. Mol Ecol. 2004; 13(11):3261–73. https://doi.org/10.1111/j.1365-294X.2004.02346.x PMID: 15487987

29. Bresadola L, Link V, Buerkle CA, Lexer C, Wegmann D. Estimating and accounting for genotyping errors in RAD-seq experiments. Molecular Ecology Resources. 2020; 20(4):856–70. https://doi.org/10.1111/1755-0998.13153 PMID: 32142201

30. Roux PF, Marthey S, Djari A, Moroldo M, Esquerré D, Estellé J, et al. Comparison of whole-genome (13X) and capture (87X) resequencing methods for SNP and genotype callings. Anim Genet. févr 2015; 46(1):82–6. https://doi.org/10.1111/age.12248 PMID: 25515399

31. Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, et al. RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Frontiers in Genetics [Internet]. 2021 [cité 21 nov 2023]; 12. Disponible sur: https://www.frontiersin.org/articles/10.3389/fgene.2021.655707 PMID: 34262593

32. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLOS ONE. 2013; 8(1):e54603. https://doi.org/10.1371/journal.pone.0054603 PMID: 23372741

33. Cumer T, Pouchon C, Boyer F, Yannic G, Rioux D, Bonin A, et al. Double-digest RAD-sequencing: do pre- and post-sequencing protocol parameters impact biological results? Mol Genet Genomics. 2021; 296(2):457–71. https://doi.org/10.1007/s00438-020-01756-9 PMID: 33469716

34. Hossain MR, Natarajan S, Kim HT, Jesse DMI, Lee CG, Park JI, et al. High density linkage map construction and QTL mapping for runner production in allo-octoploid strawberry Fragaria × ananassa based on ddRAD-seq derived SNPs. Sci Rep. 2019; 9(1):3275.

35. Jaiswal V, Gupta S, Gahlaut V, Muthamilarasan M, Bandyopadhyay T, Ramchiary N, et al. Genome-Wide Association Study of Major Agronomic Traits in Foxtail Millet (Setaria italica L.) Using ddRAD Sequencing. Sci Rep. 2019; 9(1):5020. https://doi.org/10.1038/s41598-019-41602-6 PMID: 30903013

36. Nugroho YA, Tanjung ZA, Yono D, Mulyana AS, Simbolon HM, Ardi AS, et al. Genome-wide SNP-discovery and analysis of genetic diversity in oil palm using double digest restriction site associated DNA sequencing. IOP Conf Ser: Earth Environ Sci. juin 2019; 293(1):012041.

37. Aballay MM, Aguirre NC, Filippi CV, Valentini GH, Sánchez G. Fine-tuning the performance of ddRAD-seq in the peach genome. Sci Rep. 2021; 11(1):6298. https://doi.org/10.1038/s41598-021-85815-0 PMID: 33737671

38. Gargiulo R, Kull T, Fay MF. Effective double-digest RAD sequencing and genotyping despite large genome size. Molecular Ecology Resources. mai 2021; 21(4):1037–55. https://doi.org/10.1111/1755-0998.13314 PMID: 33351289

39. Zhai Z, Zhao W, He C, Yang K, Tang L, Liu S, et al. SNP discovery and genotyping using restriction-site-associated DNA sequencing in chickens. Animal Genetics. 2015; 46(2):216–9. https://doi.org/10.1111/age.12250 PMID: 25591076

40. Ba H, Jia B, Wang G, Yang Y, Kedem G, Li C. Genome-Wide SNP Discovery and Analysis of Genetic Diversity in Farmed Sika Deer (Cervus nippon) in Northeast China Using Double-Digest Restriction Site-Associated DNA Sequencing. G3 Genes|Genomes|Genetics. 2017; 7(9):3169–76. https://doi.org/10.1534/g3.117.300082 PMID: 28751500

41. Janjua S, Peters JL, Weckworth B, Abbas FI, Bahn V, Johansson O, et al. Improving our conservation genetic toolkit: ddRAD-seq for SNPs in snow leopards. Conservation Genet Resour. juin 2020; 12(2):257–61.

42. Shepherd L, Bulgarella M, Haddrath O, Miskelly C. Genetic analyses reveal an unexpected refugial population of subantarctic snipe (Coenocorypha aucklandica). Notornis. 2020; 67.

43. Magris G, Marroni F, D'Agaro E, Vischi M, Chiabà C, Scaglione D, et al. ddRAD-seq reveals the genetic structure and detects signals of selection in Italian brown trout. Genetics Selection Evolution. 31 janv 2022; 54(1):8. https://doi.org/10.1186/s12711-022-00698-7 PMID: 35100964

44. Doekes HP, Bovenhuis H, Berghof TVL, Peeters K, Visscher J, Mulder HA. Research Note: Genome-wide association study for natural antibodies and resilience in a purebred layer chicken line. Poultry Science. 1 janv 2023; 102(1):102312. https://doi.org/10.1016/j.psj.2022.102312 PMID: 36473374

45. Fu W, Dekkers JC, Lee WR, Abasht B. Linkage disequilibrium in crossbred and pure line chickens. Genet Sel Evol. 26 févr 2015; 47(1):11. https://doi.org/10.1186/s12711-015-0098-4 PMID: 25887184

46. Drobik-Czwarno W, Wolc A, Fulton JE, Dekkers JCM. Detection of copy number variations in brown and white layers based on genotyping panels with different densities. Genet Sel Evol. 6 nov 2018; 50(1):54. https://doi.org/10.1186/s12711-018-0428-4 PMID: 30400769

47. Herry F. Stratégies de génotypage pour la sélection génomique chez la poule pondeuse [Internet] [Theses]. Agrocampus Ouest; 2019 [cité 15 févr 2022]. Disponible sur: https://hal.inrae.fr/tel-02789314

48. Herry F, Hérault F, Lecerf F, Lagoutte L, Doublet M, Picard-Druet D, et al. Restriction site-associated DNA sequencing technologies as an alternative to low-density SNP chips for genomic selection: a simulation study in layer chickens. BMC Genomics. 2023; 24(1):271. https://doi.org/10.1186/s12864-023-09321-5 PMID: 37208589

49. Tixier-Boichard M, Lecerf F, Herault F, Bardou P, Klopp C. Le projet «Mille Génomes Gallus»: partager les données de séquences pour mieux les utiliser. INRAE Prod Anim. 2020; 33(3):189–202.

50. Lajmi A, Glinka F, Privman E. Optimizing ddRAD sequencing for population genomic studies with ddgRADer. Molecular Ecology Resources [Internet]. [cité 3 oct 2023]; n/a(n/a). Disponible sur: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13870 PMID: 37732396

51. Warren W, Fedrigo O, Tracey A, Mason A, Formenti. Multiple chicken (Gallus gallus) genome references to advance genetic variation studies.

52. DaCosta JM, Sorenson MD. Amplification Biases and Consistent Recovery of Loci in a Double-Digest RAD-seq Protocol. PLOS ONE. 4 sept 2014; 9(9):e106713. https://doi.org/10.1371/journal.pone.0106713 PMID: 25188270

53. Kess T, Gross J, Harper F, Boulding EG. Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle Littorina saxatilis. Journal of Molluscan Studies. 2016; 82(1):104–9.

54. Shirasawa K, Hirakawa H, Isobe S. Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. DNA Research. 1 avr 2016; 23 (2):145–53. https://doi.org/10.1093/dnares/dsw004 PMID: 26932983

55. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet. janv 2011; Chapter 1:Unit1.19. https://doi.org/10.1002/0471142905.hg0119s68 PMID: 21234875

56. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

57. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

58. Quail MA, Gu Y, Swerdlow H, Mayho M. Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. ELECTROPHORESIS. 2012; 33(23):3521–8. https://doi.org/10.1002/elps.201200128 PMID: 23147856

59. Pagès H, Aboyoun P, Gentleman R, DebRoy S, Carey V, Delhomme N, et al. Biostrings: Efficient manipulation of biological strings [Internet]. Bioconductor version: Release (3.17); 2023 [cité 20 oct 2023]. Disponible sur: https://bioconductor.org/packages/Biostrings/

60. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4(1):s13742-015-0047-8.

61. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 1 août 2011; 27(15):2156– https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

62. R Core Team. R: A language and environment for statistical [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2022. Disponible sur: https://www.R-project.org/.