Received: 18 November 2022 | Revised: 3 March 2023 | Accepted: 5 April 2023

DOI: 10.1111/1755-0998.13800

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity

Nicolas C. Rochette[1,2] | Angel G. Rivera-Colón[2] | Jessica Walsh[3] | Thomas J. Sanger[4] | Shane C. Campbell-Staton[1] | Julian M. Catchen[2]

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA

[2]Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[3]Floragenex, Inc., Portland, Oregon, USA

[4]Department of Biology, Loyola University Chicago, Chicago, Illinois, USA

Correspondence
Julian M. Catchen, Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
Email: jcatchen@illinois.edu

Handling Editor: Paul A. Hohenlohe

Abstract

Library preparation protocols for most sequencing technologies involve PCR amplification of the template DNA, which open the possibility that a given template DNA molecule is sequenced multiple times. Reads arising from this phenomenon, known as PCR duplicates, inflate the cost of sequencing and can jeopardize the reliability of affected experiments. Despite the pervasiveness of this artefact, our understanding of its causes and of its impact on downstream statistical analyses remains essentially empirical. Here, we develop a general quantitative model of amplification distortions in sequencing data sets, which we leverage to investigate the factors controlling the occurrence of PCR duplicates. We show that the PCR duplicate rate is determined primarily by the ratio between library complexity and sequencing depth, and that amplification noise (including in its dependence on the number of PCR cycles) only plays a secondary role for this artefact. We confirm our predictions using new and published RAD-seq libraries and provide a method to estimate library complexity and amplification noise in any data set containing PCR duplicates. We discuss how amplification-related artefacts impact downstream analyses, and in particular genotyping accuracy. The proposed framework unites the numerous observations made on PCR duplicates and will be useful to experimenters of all sequencing technologies where DNA availability is a concern.

KEYWORDS
allele dropout, bioinformatics, PCR duplicates, RAD-seq, sequencing library

1 | INTRODUCTION

The occurrence of polymerase chain reaction (PCR) duplicates is an artefact present in most current sequencing technologies, from whole genome resequencing, to single-cell RNA (Marx, 2017), and to reduced-representation methods such as Restriction site-Associated DNA (RAD) sequencing (Table 1). Sequencing library preparation protocols often include a PCR step to improve yield or create molecular species of interest, but this also introduces artefacts (Aird et al., 2011; Kebschull & Zador, 2015). In particular, since these amplified libraries comprise multiple copies of each original template molecule, it becomes possible to independently sequence several reads that correspond to the same template; such reads are known as PCR duplicates (Casbon et al., 2011).

TABLE 1 PCR duplicate rates across sequencing technologies.

Method	PCR duplicate rate (%)	Reference
Whole genome sequencing	1–5	Ebbert et al. (2016)
Exome sequencing	3–15	Shigemizu et al. (2015)
Exome sequencing	2–13	Bonfiglio et al. (2016)
Exome sequencing	15–30	Borgström et al. (2017)
Linked-read DNA sequencing	11–59	Meier et al. (2021)
Deep targeted DNA sequencing	3–63	Smith et al. (2014)
Ancient DNA	2–80	Gamba et al. (2016)
Ancient DNA	2–42	Kapp et al. (2021)
Single-cell DNA sequencing	15–25	Gonzalez-Pena et al. (2021)
Single-cell bisulfite sequencing	2–6	Farlik et al. (2015)
RNA-seq	1–14	Fu et al. (2018)
RNA-seq	1–68	Parekh et al. (2016)
Single-cell RNA-seq	85–95	Ziegenhain et al. (2017)
Hi-C	14–64	Niu et al. (2019)
Micro-Capture-C	80–96	Hua et al. (2021)

PCR duplicates create a distorted view of the abundancies of molecules in the original sample, which may bias or degrade downstream statistical analyses (Casbon et al., 2011; DePristo et al., 2011; Fu et al., 2018; Niu et al., 2019). These concerns have motivated the development of amplification-free methods (Kozarewa et al., 2009; Niu et al., 2019), but such protocols come with substantial technical constraints, particularly with regard to input biological materials. Most methods have instead accepted the artefact as an inherent feature of sequencing data and focus on tracking and removing PCR duplicates bioinformatically (Marx, 2017; Sims et al., 2014), which can be done either based on read (or read pair) mapping coordinates (DePristo et al., 2011; Li et al., 2009), by tagging template molecules before amplification using unique molecular identifiers (UMIs) (Casbon et al., 2011; Kivioja et al., 2011), or by combining both approaches (Islam et al., 2014; Smith et al., 2017).

Although a posteriori removal of PCR duplicates has proven effective to mitigate bias, it is not without limitations. This approach can increase the cost of sequencing noticeably—up to severalfold when duplicates represent most of the initial reads (Table 1). In the worst case, after deduplication, coverage may become insufficient for the intended purpose causing experiments to fail. The bioinformatic identification of duplicates is also imperfect. Coordinate-based tracking is unsuitable for experiments in which coverage in some genomic regions is high enough that they become saturated, such as bulk RNA-seq (Fu et al., 2018; Parekh et al., 2016), ChIP-seq (Tian et al., 2019) or Pool-seq (Kofler et al., 2016). UMI-based tracking is confounded by sequencing errors, which must be rigorously accounted for to avoid incomplete deduplication (Islam et al., 2014; Marx, 2017; Smith et al., 2017). Lastly, and more fundamentally, removing PCR duplicates is a pragmatic approach that targets the visible consequences of amplification, but distortions can be expected to remain present even after deduplication.

Therefore, developing a better understanding of amplification-related artefacts and gaining control of the rate of PCR duplicates a priori by optimizing library preparation procedures remains highly desirable. Despite the prevalence of these artefacts, we still lack a realistic quantitative model for library amplification and the generation of PCR duplicates. As a result, our comprehension of the phenomenon remains highly empirical (Marx, 2017), and there remains some uncertainty and confusion regarding the precise experimental factors that control their occurrence, how these factors interact with one another, and the consequences of PCR duplicates on downstream statistical analyses.

Among the factors believed to have an effect on the PCR duplicate rate, the most important one is library complexity, which has been alternatively defined as the complement of the duplicate rate (Chen et al., 2012), as the information content of a library (Zhang et al., 2015), or as the number of distinct molecular species represented in a sequencing library (Daley & Smith, 2013; following Lander & Waterman, 1988). Insufficient library complexity is frequently given as the probable cause of high duplicate rates (Chen et al., 2012; Marx, 2017; Parekh et al., 2016; Smith et al., 2014; Tin et al., 2015), and several studies have demonstrated that the amount of starting biological material used, which presumably correlates with library complexity, had a marked effect on PCR duplicates (Casbon et al., 2011; Fu et al., 2018; Kapp et al., 2021; Smith et al., 2014). A few authors have pointed out that sequencing depth should also be considered (Daley & Smith, 2013; Fu et al., 2018; Marx, 2017; Smith et al., 2014). Lastly, it is often claimed that PCR duplicate rates depend on the number of PCR amplification cycles that the library was subjected to (Andrews et al., 2016; Ebbert et al., 2016; Flanagan & Jones, 2018; Marx, 2017; Orlando et al., 2021; Smith et al., 2017; Stuart et al., 2018; Vargas-Landin et al., 2018). Some studies have indeed found this to be the case (Lu et al., 2017; Niu et al., 2019; Parekh et al., 2016), but others have concluded that no such relationship existed (Fu et al., 2018; Tin et al., 2015). Thus, several factors relevant to the phenomenon have been identified, but their precise roles remain unclear. A more quantitative understanding of the PCR duplicate artefact would help clarify which methodological alterations are likely to suppress it.

Another question that has been particularly debated in the context of RAD-seq —a restriction enzyme-based reduced-representation sequencing approach that is widely used for population genomics studies of nonmodel organisms (Andrews et al., 2016; Catchen et al., 2017; Daley & Smith, 2013)—is the extent to which PCR duplicates affect the reliability of genotyping. While Tin et al. (2015) and Flanagan and Jones (2018) found it to be an important source of error, Euclide et al. (2019) did not. In any case, substantial (>60%) PCR duplicate rates have been reported in some data sets (Andrews et al., 2016; Davey et al., 2013; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Hoffberg et al., 2016; Schweyen et al., 2014), which

is especially concerning as some popular protocols do not allow the monitoring of this artefact (e.g., double-digest RAD-seq without UMIs). In addition, as sample availability is often a limiting factor in nonmodel organisms (Andrews et al., 2016; Peterson et al., 2012; Tin et al., 2015), better understanding the extent to which observed PCR duplicate patterns result from the use of reduced amounts of DNA would be of great experimental interest. Finally, missing data and allelic dropout in RAD-seq has been attributed to restriction site polymorphism (a RAD-seq-specific artefact in which alleles carrying mutations in the restriction enzyme recognition sequence are absent from the data set; Andrews et al., 2016), but this may also reflect an incomplete understanding of the consequences of PCR duplicates.

Here, we develop a general quantitative framework able to realistically model amplification-related artefacts in sequencing experiments based on library complexity, sequencing depth, and amplification noise. We provide a method, *Decoratio*, to estimate these factors for any sequencing data set based on PCR duplicate patterns. We apply this method to new and previously published RAD-seq data sets to demonstrate that amplification artefacts are often an important feature of this sequencing approach and that the model recapitulates the main properties of these experiments. We show that our model reconciles the numerous earlier observations made on PCR duplicate rates, and we discuss how amplification-related artefacts increase variance in downstream analyses, with particular application to genotyping accuracy. Overall, this work furthers our understanding of the properties and consequences of amplification artefacts and will facilitate the optimization and deployment of novel sequencing technologies.

2 | MATERIALS AND METHODS

2.1 | PCR duplicate model

We identify three stages of library preparation and sequencing as being critical to the modelling of PCR duplicates: (1) the pre-amplification pool of template molecules; (2) the pool of amplified molecules; and (3) the pool of molecules that are sampled for sequencing into digital reads (Figure 1). Accordingly, our model for the occurrence of PCR duplicates comprises two disjoint steps that respectively connect the first and second, and the second and third of these stages.

In the first step, we model the amplification of template molecules. Namely, given a pool of template molecules, we determine a probability distribution for their individual amplification factors. As the precise method used to determine the distribution does not matter, this approach is highly flexible; we propose two amplification models. First, a distribution of amplification factors can be obtained using forward simulations. For instance, we implemented the PCR model developed empirically by Best et al. (2015): starting with one molecule, we apply the amplification model, then record the final number of molecules in the resulting clone (i.e. the amplification factor), and repeat the amplification process independently, one molecule at a time, to obtain an estimate of the distribution of clone sizes. Alternatively, the distribution of amplification factors can be set to some relevant parametric distribution, such as the log-normal distribution.

In the second step, we model the sequencing of reads from the amplified pool of molecules. Crucially, we treat the pool of amplified molecules as infinite, and use the distribution of amplification factors (i.e. of amplification clone sizes) as a statistical description of this pool. This assumption is always reasonable because in practice the amplified pool has to be much larger than the number of reads derived from it—for instance, in the case of Illumina sequencing, only a small fraction of the PCR product is eventually loaded onto a flow cell and bridge-amplified. In addition, for simplicity, we assume that there is no sequencing bias, that is all clones and all molecules within a clone have an equal probability to be sequenced (we note, however, that under the proposed framework it is also reasonable to let the amplification factor distribution aggregate the variance introduced during both amplification and sequencing). Under these assumptions, the occurrence of PCR duplicates can be modelled using a Poisson mixture model, as follows.

We refer to the set of duplicate molecules that were amplified from a particular template molecule as an 'amplification clone,' and to the set of reads that derive from a particular template molecule (i.e. duplicate reads) as a 'sequencing clone.' Importantly, for a given read data set, the PCR duplicate rate r is a function of the distribution of sequencing clone sizes. For each clone, we count one read as unique and the rest as duplicates, which gives the formula

$$r = 1 - \sum_{k=1}^{\infty} \frac{1}{k} P(S = k)$$

where $P(S = k)$ is the distribution of sequencing clone sizes. The distribution of sequenced clone sizes can be modelled as

$$p(S = k) = \sum_{a=1}^{\infty} p(A = a) p(S = k \mid AC = a)$$

where $P(A = a)$ is the distribution of amplification clone sizes (i.e. the distribution of amplification factors), and

$$P(S = k|A = a) \sim \text{Binomial}\left(N_r, \frac{a}{N_m \times \bar{a}}\right)$$

where N_r is the number of read pairs that were sequenced, N_m is the number of template molecules in the pre-amplification pool of DNA, and $\bar{a}$ is the mean amplification factor, so that $N_m \times \bar{a}$ is the number of molecules in the amplified pool of DNA. Assuming N_r is large and all individual species frequencies are low, this can alternatively be written as

$$P(S = k|A = a) \sim \text{Poisson}\left(\frac{N_r}{N_m} \times \frac{a}{\bar{a}}\right).$$

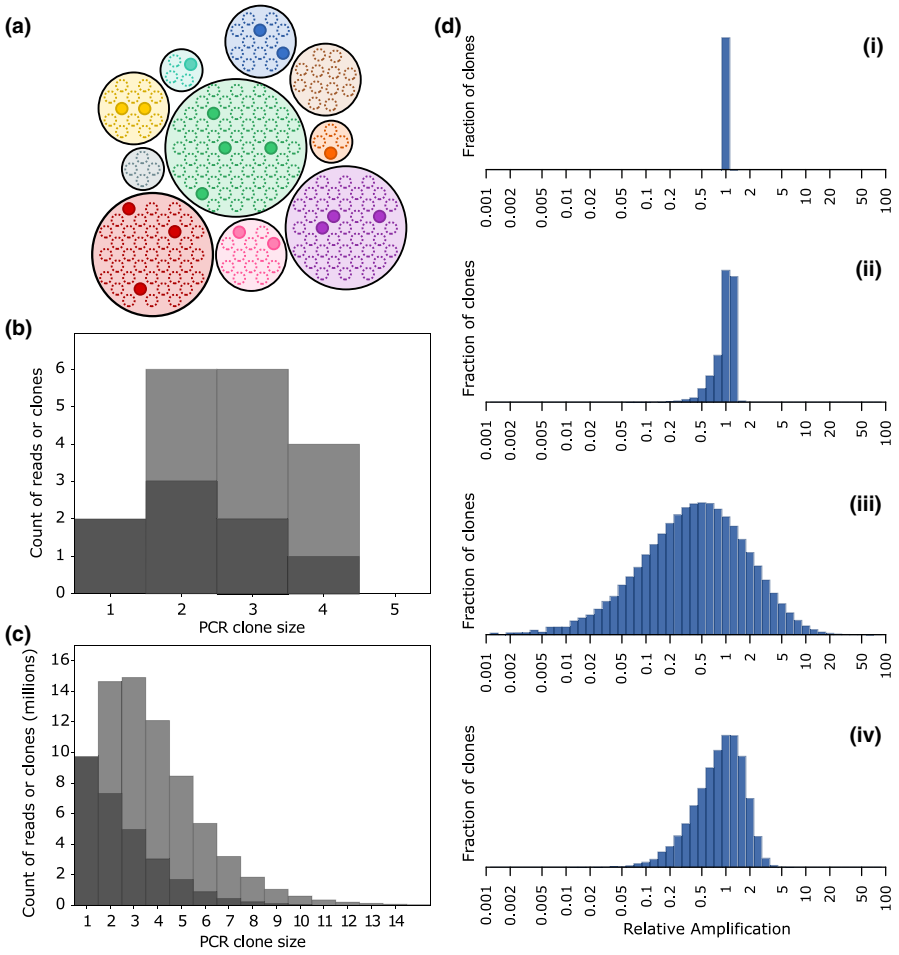

FIGURE 1 Overview of the PCR duplicate rate model. (a) Schematic representation of the processes involved in the occurrence of PCR duplicates. During library preparation, template DNA molecules are amplified, so that for each initial molecule a clone (coloured circles) of amplified molecules (dots) is generated. Since PCR is a stochastic and biased process, amplification efficiency is heterogenous across templates and amplification clones vary in size. Subsequently, a small fraction of amplified molecules are randomly sampled for sequencing and become reads (solid dots). Reads that correspond to the same template, that is, that belong to the same clone, are known as PCR duplicates. Duplicate reads are more likely to be sampled from large clones (i.e. templates whose amplification was most successful) than from small clones. (b, c) In practice, what can be observed in sequencing data is the size of clones at the read level (i.e. the number of singletons, of pairs of two duplicate reads, etc.). This serves as the main empirical input for the model. The histograms show the distribution of read clone sizes respectively for the schematic in the first panel (b) and for the brown anole data set (c), either as the number of clones of each size (dark grey) or as the number of reads that belong to clones of a particular size (dark and light grey combined). These views are equivalent by definition, as for instance the number of reads in clones of size two (duplicate pairs) must be twice the number of such clones. Furthermore, as the light grey bars represent redundant reads, the PCR duplicate rate can be calculated from such histograms by taking the ratio of the light grey area over the total area; in this case respectively 56% (18 reads in 8 distinct clones) and 61%. (d) Amplification noise is modelled as the distribution of relative amplification factors among templates. Histograms show the amplification factor distributions for (i) a perfect, noiseless amplification; (ii, iii) for the empirically developed amplification model of Best et al. with mean amplification efficiency and bias parameters set to 'low noise' and 'high noise' values, respectively $m = 0.7$, $s = 0.01$, 12 cycles, and $m = 0.45$, $s = 0.1$, 18 cycles; and (iv) for the log-skew normal distribution fitted to the brown anole data set. These distributions are used to model sequencing data as a mixture of Poisson distributions (see Section 2).

The PCR duplicate rate then depends on two separate parameters: (i) the ratio between the number of sequenced read pairs and the number of pre-amplification template molecules (i.e. the number of unique species in the library, its absolute complexity), which we hereafter refer to as the 'depth-complexity ratio,' and (ii) the distribution of amplification factors relative to the mean amplification factor, *that is* the noisiness of the amplification (Figure 1d). It can be noted that in the model the absolute value of the average amplification factor has no effect; this results from our assumption that the pool of amplified molecules is infinite, which, as argued above, is always realistic, and from the independent parametrization of the amplification variance.

Importantly, it is possible to re-write the depth-complexity ratio in terms of coverage. Specifically, if we consider a particular read

(or read pair) length l_r and a set of target genomic regions (e.g. the whole genome) of total length L_T from which the reads derive, then nucleotide-wise coverage can be written as

$$C = \frac{N_r \times l_r}{L_T}.$$

Similarly, we can define molecular coverage—hereafter 'molecular density' to avoid confusion around the term coverage—as

$$M = \frac{N_m \times l_r}{L_T}.$$

This allows us to re-write the depth-complexity ratio as

$$\frac{N_r}{N_m} = \frac{C}{M}.$$

Molecular density defined above is fundamentally a locus-wise measure of library complexity and corresponds to the number of unamplified template molecules that cover an average position of the sequencing target. We note that when coverage is tallied at the nucleotide level, the complexity of a library depends on the read length that is used to sequence it and is maximized when the full length of the molecules is sequenced. This is expected, because the read length influences the number of useful nucleotides within each molecule, and therefore the information content of the library.

2.2 | Implementation

In the model described above, the PCR duplicate rate depends on two parameters: the depth-complexity ratio, and the distribution of amplification factors. We provide a method, *Decoratio* (for 'depth-complexity ratio'), to jointly estimate these two parameters based on the distribution of PCR clone sizes observed in a sequencing data set. Given that the experimenter already knows the sequencing depth, the depth-complexity ratio also corresponds to an estimate of library complexity.

The program requires two inputs, a distribution of PCR clone sizes and a class of PCR models, and outputs the optimized depth-complexity ratio and PCR model parameters, as well as a plot of the input distribution and fitted model. An example of the expected input, command line call, and outputs of the program is shown in Figure S1. The distribution of PCR clone sizes should be formatted as a TSV table giving the number of clones of each size, and can be obtained using programs such as *SAMtools-Markdup* (Li et al., 2009), *Picard-MarkDuplicates* (McKenna et al., 2010), *UMItools* (Smith et al., 2017), or *Stacks-gstacks* (Rochette et al., 2019), as described in *Decoratio*'s online manual. If an experiment comprises multiple libraries, we stress that clone size distributions should be derived on a per-library basis, rather than on an aggregated data set, as the properties of the data are likely to vary across libraries.

For the PCR model, the program currently implements log-normal and log-skew-normal distributions of amplification factors, as well as the empirical 'inherited efficiency' model class of Best et al. (2015) which we described above. The model may be fully specified or only partly so, in which case the program will optimize the model parameters. For most users, using the default log-skew-normal model should work well and the program will fit the standard deviation and skew parameters (Figure 1). For the inherited efficiency model, the distribution of amplification factors is obtained by forward simulation. Each clone is assigned a duplication probability drawn from a normal distribution, and then amplified by a succession of binomial samplings. In practice, the parameter space is binned to increase computational efficiency, and by default 1 million simulations are performed. We added a slightly modified variant of this model that uses a Beta distribution instead of the normal distribution (Figure S2) so as to avoid parametrization issues related to the need to truncate the normal distribution between 0 and 1. The program jointly optimizes the depth-complexity ratio and amplification model by minimizing the sum of squared residuals to the observed distribution of the fraction of reads in each clone size class, using simplicial homology global optimization (SHGO) as implemented in SciPy.

2.3 | RAD-seq data generation and analysis

We tested the PCR model in *Decoratio* on five empirical RAD-seq data sets. Three of these data sets are from previously published studies: (1) stickleback (*Gasterosteus aculeatus*) (Nelson & Cresko, 2018), (2) yellow warbler (*Setophaga petechia*) (Bay et al., 2018), and (3) Emperor penguin (*Aptenodytes forsteri*) (Cristofari et al., 2016). These are hereafter referred to as the stickleback, warbler, and penguin data sets, respectively. For these data sets, we split and separately analysed according to their respective libraries, for example the warbler data set was constructed as three separate RAD-seq libraries, each comprised of multiple individuals, and was thus analysed as *Warbler-1*, *Warbler-2*, and *Warbler-3* (see Supplementary Methods).

In addition, we used two newly generated RAD-seq data sets to guarantee complete control of the library preparation process, allowing for a more detailed validation of the PCR model. First, the robin data set, comprised of 150 American robins (*Turdus migratorius*) collected from central Illinois, USA (A. B. Luro, A. G. Rivera-Colón, J. M. Catchen, M. E. Hauber, unpublished). Briefly, DNA was extracted from all individuals, prepared into a single-digest RAD-seq (sdRAD) library digested with *SbfI* (Baird et al., 2008; Etter et al., 2011) and sequenced on an Illumina NovaSeq-6000 SP 2 × 150 bp lane. Second, the anole data set, which was generated from 39 *Anolis sagrei* embryos. All 39 individuals were sequenced in two separate single-digest *SbfI* RAD-seq libraries (Baird et al., 2008; Etter et al., 2011): a high template (Anole-600) library in which a pool of 600 ng of DNA was used as template for the PCR, and a low template (Anole-30) library which instead used 30 ng of DNA as the PCR template. After

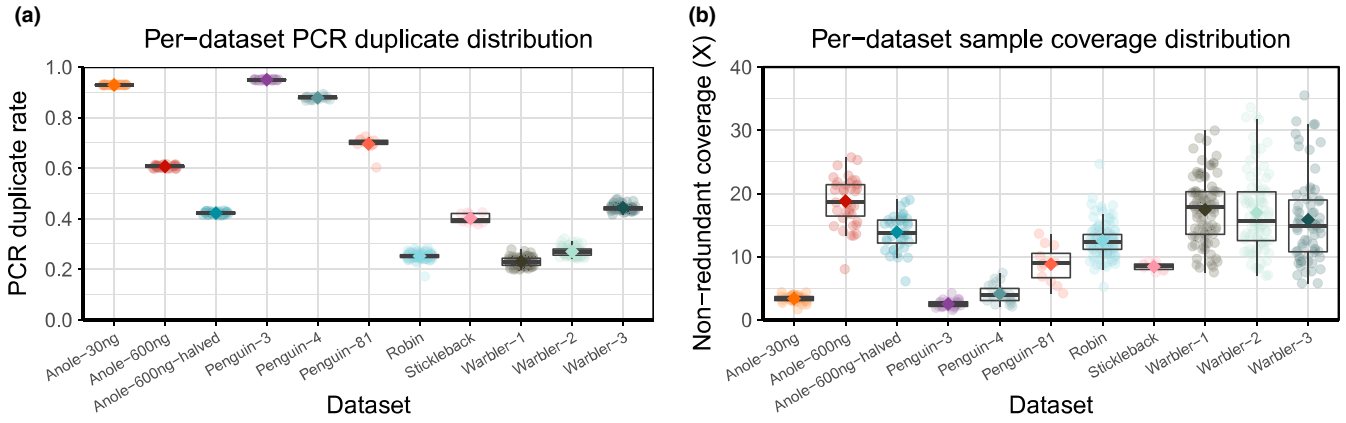

FIGURE 2 PCR duplicate rates in reanalyzed sdRAD data sets. (a) Distribution of PCR duplicates for all samples in each data set. Coloured diamonds show the mean per-library duplicate rate. The PCR duplicate rate is consistent across all samples within a molecular library. For data sets comprised of multiple libraries, the duplicate rate can vary among the different libraries. (b) Non-redundant coverage distribution for all samples in each data set. Coloured diamonds show the mean non-redundant coverage. In contrast to PCR duplicate rate, coverage is highly variable within these libraries.

amplification, each library was sequenced 2 × 100 bp on an Illumina HighSeq-4000 (see Supplementary Methods for detail).

Sequencing data for each library were analysed separately using *Stacks* v2.5 (Rochette et al., 2019). Raw reads were processed and demultiplexed using the PROCESS_RADTAGS program. RAD loci were assembled de novo using the `denovo_map.pl` script. The mismatch parameters `M` and `n` were kept equal and set separately for each data set to account for differences in read length. We identified and removed PCR duplicates (`--rm-pcr-duplicates`) and selected only loci present in over half the samples in the library (using the flag `-X 'gstacks: --dbg-min-loc-spls'`). The final loci and variant sites were exported in VCF format using the POPULATIONS program, keeping variants genotyped in at least 50% of samples (`-r 0.5`) (see Supplementary Methods for more detail).

2.4 | qPCR quantification of library complexity

To obtain an empirical measure of the complexity of the robin library, the amount of template DNA was quantified using qPCR. During RAD-seq library preparation, this template is composed of the fraction of the DNA molecules in the pool that have been successfully ligated with both P1 and P2 adapter sequences. This template DNA, which we obtained in the library preparation process after P2 ligation, but prior to PCR, was amplified alongside a control made of the final RAD-seq library, which is cleaned and quantified prior to sequencing. The known concentration of the final library was used to then calculate an absolute concentration of template in the pre-amplified library. To create a standard curve, five 1:10 sequential dilutions of the 0.1 nM control library were prepared and amplified in triplicate, alongside a negative control. Similarly, the library template was diluted in two sequential 1:10 dilutions and amplified in triplicate. qPCR reactions were prepared using the KAPA Library Quantification Kit (KAPA Biosystems). While the KAPA kit by default uses primers compatible with Illumina's P5 and P7 oligo sequences, the complimentary sequence for these primers is not present in single-digest RAD-seq libraries until it is reconstructed during PCR amplification (Baird et al., 2008; Etter et al., 2011). Instead, we performed qPCR amplification using the primers designed for sdRAD library enrichment PCR (see Supplementary Methods for primer sequences). For the reaction, the forward and reverse primers were combined into a single 10 μM primer mix. Each 20 μL reaction consisted of 10.4 μL of KAPA SYBR FAST mix with ROX, 2 μL of standard Primer Mix, 3.6 μL of PCR-grade water, and 4 μL of DNA. The qPCR reactions were run on a QuantStudioTM 3 Real-Time PCR System (Applied Biosystems) using the default $\Delta\Delta C_T$ protocol. This protocol consists of an initial denaturation step of 1 min at 95°C, 35 cycles of a 95°C 30 s denaturation followed by a 60°C annealing/extension/data acquisition for 45 s, and a 65-95°C melting curve analysis.

The C_T of a given sample was obtained by calculating the average across its replicates. The concentration of the library template in nM was obtained by regressing its average CT value against the standard curve obtained for all the known control.

3 | RESULTS

3.1 | PCR duplicate rates in RAD-seq data sets

To assess the general occurrence of PCR duplicates in RAD-seq studies, we reanalyzed a series of new and published data sets. Specifically, we generated two sdRAD (Baird et al., 2008) paired-end data sets, respectively, comprising 39 brown anole (*A. sagrei*) individuals and 150 American robin (*T. migratorius*) individuals. We also considered published paired-end data sets that use sdRAD in the Emperor Penguin (Cristofari et al., 2016) (3 libraries; Penguin-3, Penguin-4 and Penguin-81), sdRAD in the Threespine Stickleback (Nelson & Cresko, 2018) (1 library), and bestRAD (Ali et al., 2016) in the Yellow Warbler (Bay et al., 2018) (3 libraries; Warbler-1 to Warbler-3). We did not include any libraries based on double-digest

TABLE 2 PCR duplicate rates and library complexities in newand reanalyzed RAD-seq data sets.

Library	Anole 600ng	Anole 30ng	Robin	Stickle-back	Penguin 3	Penguin 4	Penguin 81	Warbler 1	Warbler 2	Warbler 3
Raw library coverage (total)	1899×	1895×	2452×	144×	1532×	820×	351×	2218×	2133×	2006×
Samples in library	39	39	150	10	30	24	12	95	89	68
PCR duplicate rate	61 ± 1%	93 ± 1%	21 ± 1%	40 ± 2%	95 ± 0.2%	88 ± 1%	70 ± 3%	23 ± 2%	27 ± 2%	44 ± 1%
Depth-complexity ratio	1.93	14.2	0.29	0.87	19.9	8.16	1.85	0.38	0.46	0.85
Library density (total)	984×	133×	8455×	166×	77×	101×	190×	5837×	4637×	2360×
Library density (per-sample mean)	25×	3.4×	56×	17×	2.6×	4.2×	16×	61×	52×	35×
Mean non-redundant coverage	19.1 ± 3.8×	3.4 ± 0.6×	12.8 ± 2.4×	8.6 ± 0.6×	2.6 ± 0.6×	4.2 ± 1.5×	8.9 ± 2.8×	17.9 ± 5.3×	17.3 ± 6.4×	16.3 ± 7.2×

RAD (ddRAD; Peterson et al., 2012) as this protocol does not allow the tracking of PCR duplicates (but see Section 4). PCR duplicates were identified based on read mapping coordinates (Rochette et al., 2019). Given that in sdRAD the first read of a pair always maps to the restriction site (a fixed position), this approach may incorrectly flag independent reads as duplicates if they by chance map to the same coordinates. However, given the nonredundant depths (3–19×) and insert size distribution widths (100-300 bp) in the data sets considered, we expect that these coordinate collisions should be rare and amount to only 1%–5% of the estimated duplicate rates.

Overall, we found substantial PCR duplicate levels in all analysed data sets. Mean per-library duplicate rates ranged from 21% for the Robin library to 95% for Penguin-3 (Figure 2; Table 2). Unsurprisingly, because a 95% duplicate rate corresponds to a sequencing efficiency of just 5%, nonredundant coverage (i.e. coverage after removing PCR duplicates) was very low for the samples of the Penguin-3 library, ranging from 1.6 to 4.3×. However, even the better libraries were subject to an appreciable sequencing efficiency loss. The bestRAD protocol, yielding three libraries with low duplicate rates, tended to perform better than the sdRAD protocol, which yielded libraries with both high and low duplicate rates, although our sample of data sets may be too small for generalization.

Perhaps most importantly for the purposes of this work, it appeared very clearly that the relevant level at which to look at PCR duplicates was the library, rather than the individual sample or the study. Indeed, PCR duplicate rates varied greatly across libraries within each data set (for data sets comprising several libraries) but were highly consistent across samples within each molecular library (Figure 2; Table 2). Remarkably, quality differences between individual DNA samples are likely present in at least some of the tested libraries but did not appear to impact PCR duplicate rates.

These observations prompted us to investigate the causes underlying the occurrence of PCR duplicates, the potential detrimental effects of their presence, and the steps that should be taken to reduce these rates.

3.2 | A realistic model for PCR duplicates

We present a quantitative model that captures the steps of library preparation and sequencing that are critical with regard to PCR duplicates. Briefly, this model comprises two steps. First, we model how the relative molecule abundancies within a DNA library are distorted by PCR amplification. Specifically, for a chosen PCR model, we derive a distribution of the relative amplification factors across molecules within a library (Figure 1c). Second, we derive expected patterns of PCR duplicates (i.e. the distribution of PCR clone sizes in the sequencing data; Figure 1b) by modelling the stochastic sequencing process—accounting for sequencing depth, the complexity of the library, and amplification distortions—using a compound Poisson model. In practice, the model uses two parameters: a PCR model, and the ratio between the number of sequenced reads and the number of unique molecules in the library, which we refer to

as the depth-complexity ratio. These parameters are fitted to individual data sets by matching the predicted distribution of PCR clone sizes to the observed one.

Using this method, we were able to reproduce the clone size distributions that were observed experimentally in the data sets introduced above (Figure 3a,b, Figure S3). Accounting for amplification led to considerably better fits than using a noiseless, simple Poisson model, especially in data sets that have more pronounced amounts of PCR duplicates (e.g. Figure 3a). Such data sets provide more information on the underlying true distribution of amplification factors, whereas in data sets with low duplicate rates, most reads are singletons and are uninformative in this regard. In addition, even if the noiseless model may appear to be a reasonable fit for data sets with fewer duplicates (e.g. Figure 3b), ignoring amplification noise still led to skewed library complexity estimates (Table S1). For instance, for the 22%-duplicate Robin data set, the true library complexity was likely more than 60% larger than the estimate based on a noiseless model.

Remarkably, our model also captures one striking property of empirical RAD-seq libraries: that all samples within a library had almost identical PCR duplicate rates, regardless of any differences in coverage or DNA quality that may exist among them (Figure 2). The model suggests this happens because the primary determinant of the rate of PCR duplicates is the depth-complexity ratio. This ratio is expected to be identical for all samples within a library, because for each sample both the number of reads and the number of (active) template molecules are measures of the relative abundancy of that sample in the library. For instance, a sample representing one percent of a library's template molecules can be expected to later receive one percent of the total sequencing coverage for this library, so that the depth-complexity ratio of that sample will be equal to the global depth-complexity ratio of the library.

Nevertheless, we note that there remained some within-library variation of the PCR duplicate rate (Figure 2), which attests to the presence of sample-specific effects. Some, but not all of this residual variation could be explained by differences in library representation (41% and 71% of the variance for the Anolis and Robin libraries respectively; Figure S4). Conceivably, differences in DNA quality among samples may lead to different responses during PCR amplification or sequencing. For instance, the fragment size distribution

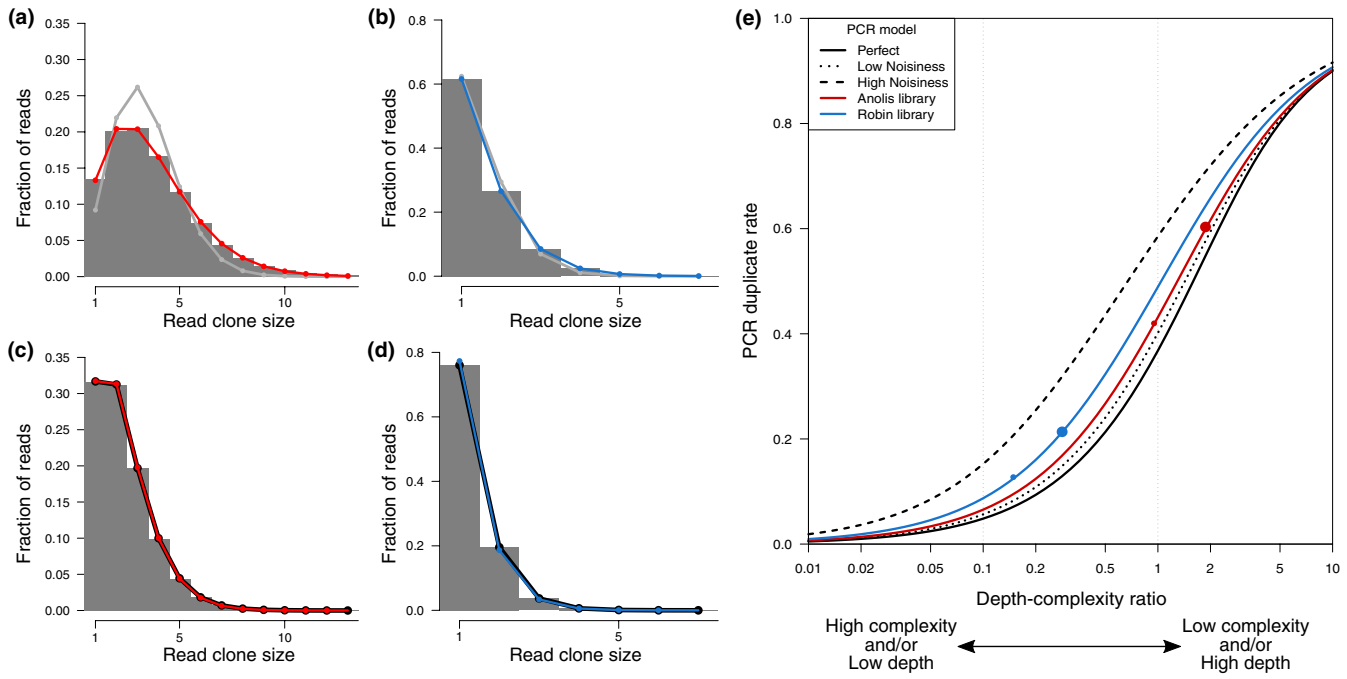

FIGURE 3 Expected relationship between the depth-complexity ratio and the PCR duplicate rate. (a, b) Histogram of the read clone sizes observed in the Anole-600ng and Robin data sets, respectively, overlaid with the fitted depth-complexity ratio and amplification noise model (solid red and blue lines respectively) or a noiseless amplification model (Poisson distribution; grey lines; Figure S7). (c, d) Same as (a) and (b) respectively, but for re-analyses of each data set after randomly discarding half the reads (twofold downsampling). The red (respectively blue) line shows the distribution predicted when halving the depth-complexity ratio in the model fitted on the full data set (i.e. the coloured lines in panels A and C respectively). In both panels, the thick black line shows the model fitted to the downsampled data set. (e) Expected relationship between the PCR duplicate rate, the sequencing depth, the library complexity, and the PCR itself. The solid, dotted, dashed and red amplification noise models are the same as in (Figure 1d), and the blue model is the best fit to the American robin data set. Large coloured dots mark the observed PCR duplicate rates and estimated depth-complexity ratios for the brown anole (red) and American robin (red) RAD-seq data sets, respectively, and small dots the values observed when considering only half the reads in those data sets, as shown in panels (A–D). PCR duplicate rates are determined primarily by the ratio between the sequencing depth and the complexity of the library, and modulated by the noisiness of the model used for the amplification. The curves shown in this panel may also be considered from the perspective of sequencing saturation (Figure S8).

may vary across samples, and this could result in differences in mean amplification factor and/or sequencing efficiency.

3.3 | PCR duplicate rates vary predictably with sequencing depth

While differences in coverage among the samples in a library due to unequal representation do not prevent them from exhibiting the same PCR duplicate rate, the dependence of PCR duplicate rates on the depth-complexity ratio nevertheless implies a more general dependence between coverage and PCR duplicates. Specifically, for a given library, the ratio's denominator (i.e. complexity) is fixed once the library has been prepared, whereas the numerator depends on the sequencing effort that is subsequently applied. Consequently, for a given library, increasing the sequencing effort should increase the PCR duplicate rate by a predictable amount.

We tested this prediction using the Anolis data set by down-sampling the original reads to reduce the sequencing coverage by half, with the expectation that the PCR duplicate rate would decrease accordingly. The original data set had a raw coverage of 49× and an observed PCR duplicate rate of 61% (Table 2). The fitted model predicted that the depth-complexity ratio in this experiment was 1.93, and that halving the coverage should decrease the PCR duplicate rate to 42%. After processing the down-sampled data set, a PCR duplicate rate of 42% was observed, exactly matching the prediction. At a finer level, the change in the shape of the distribution of PCR clone sizes was also predicted precisely (Figure 3a–d). We conclude that our model captures essential properties of the PCR duplicate phenomenon and has predictive value.

3.4 | PCR duplicate rates depend primarily on the complexity of a library

Next, we summarize the predictions of the model regarding the behaviour of PCR duplicate rates under a range of scenarios. As previously described, the model relies on two parameters: the depth-complexity ratio and a PCR model. The results obtained by varying these parameters are presented in (Figure 3e).

Importantly, while the PCR duplicate rate depends on both depth and complexity, it is only sensitive to the value of the ratio between the two, regardless of their respective absolute values. In the model, this property derives from assumptions on the sequencing sampling process, but it also holds for real data (see above results regarding the uniformity of duplicate rates within libraries). We thus only investigate variations of the depth-complexity ratio and did not assess the effects of coverage and complexity individually. Similarly, with regard to the PCR model parameter, the duplicate rate only depends on the distribution of relative amplification factors (Figure 1d), regardless of the mean absolute amplification factor or of the precise mechanism generating the spread of amplification factors. What matters especially is the overall 'noisiness' of the PCR—whether all molecules are amplified equally or, on the contrary, whether some molecular species become much more abundant than others. For this reason, the results presented here, derived using the class of PCR models developed empirically by Best et al. (2015), should be robust to the choice PCR model class, and specifically should generally hold for any PCR model producing an approximately log-normal distribution of amplification factors. We note that the amplification noisiness range assessed here is relevant for typical library amplification reactions; some specialized amplification techniques, such as multiple displacement amplification, are much noisier (Gawad et al., 2016) and may thus fall outside of the considered range.

Our central observation is that depth-complexity ratios much less than one always yield low duplicate rates, and that depth-complexity ratios greater than one always yield high duplicate rates. The duplicate rate is significantly influenced by the PCR model only when the depth-complexity ratio is between one tenth and one (Figure 3e). Nevertheless, the duplicate rate is always in the 5%–15% range if the depth-complexity ratio is 0.1, or in the 35%–60% range if that ratio is 1, regardless of the noisiness of the PCR.

Thus, we find that the PCR duplicate rate depends primarily on the depth-complexity ratio, and only marginally on the PCR model. From an experimental perspective, it should be noted that the sequencing coverage is set according to experimental needs (e.g. to 30×), whereas complexity is an intrinsic and difficult to control property of the prepared library. Consequently, a high depth-complexity ratio typically occurs because the library has a molecular density that is too low with regards to experimental needs. Thus, the above result can most simply be understood as: high PCR duplicate rates occur when pre-amplification library complexity is insufficient, largely independently of the PCR protocol being used.

3.5 | qPCR measurement of the molecular density of RAD-seq libraries suggests in silico estimates are realistic

It is notable that the library complexities measured above (Table 2) do not match the values expected from the physical mass of DNA used to prepare the libraries. For instance, the PCR of the 150-sample Robin library used a total of 400 ng of DNA (Table 3). Given that the weight of one haploid American robin genome (i.e. its C-value) is 1.39×10^{-3} ng (Andrews et al., 2009), the DNA mass used is equivalent to 288,000 genome copies, so that the library could be expected to have a density of 288,000x (or 1918× per sample, on average). In contrast, our modelling of PCR duplicate patterns in the resulting sequencing data suggests a total density of 8500× (57× per sample). Thus, the diversity of molecules represented in sequence reads was 34 times less than the mass-based expectation.

To explain this discrepancy, we hypothesized that only a small fraction of the molecules present could actually be amplified and subsequently sequenced, while most molecules would be degraded or otherwise inert for the purpose of amplification and sequencing (Meyer et al., 2008). To test this experimentally, we used qPCR to

TABLE 3 Library and bioinformatic statistics needed to contextualize an experiment's PCR duplicate rate. For sequencing approaches other than RAD-seq, the number of RAD loci should be substituted with a relevant measure of the size of the genomic target, and most libraries will only include one sample. The American Robin C-value was sourced from Andrews et al. (2009) and the Anole C-value (1.97) was calculated assuming a genome size of 1.93 Gbp (Geneva et al., 2022) and a DNA weight of 1.023 pg/Gbp (Doležel et al., 2003).

Library	Anole-600	Robin
DNA amplified (ng)	600	400
DNA amplified (C-value equivalents)	305,000	288,000
Total aligned read pairs	72,942,759	189,422,046
PCR duplicate rate	61%	21%
Number of filtered RAD loci	43,906	89,161
Samples in library	39	150
Mean non-redundant coverage	19.1×	12.8×

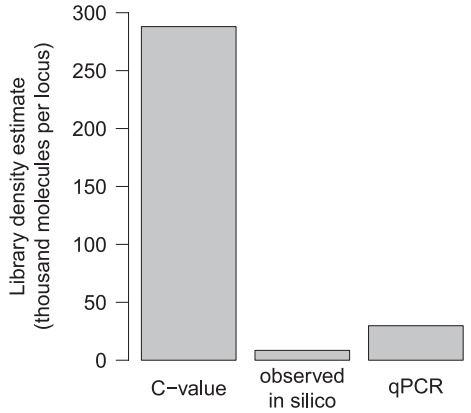

FIGURE 4 Library complexity is much smaller than suggested by DNA mass. Library complexity for the American robin RAD-seq library, measured as the total molecular density for 150 pooled samples, obtained using proxies available at different stages of the experiment. Calculating library complexity by dividing the mass of DNA used for amplification by the C-value of the organism vastly over-estimates the value actually observed in downstream in silico analyses. Quantifying this same DNA using qPCR highlights that most molecules are indeed not amplifiable and therefore do not contribute to library complexity, owing to sample preservation and partial yield at earlier experimental steps.

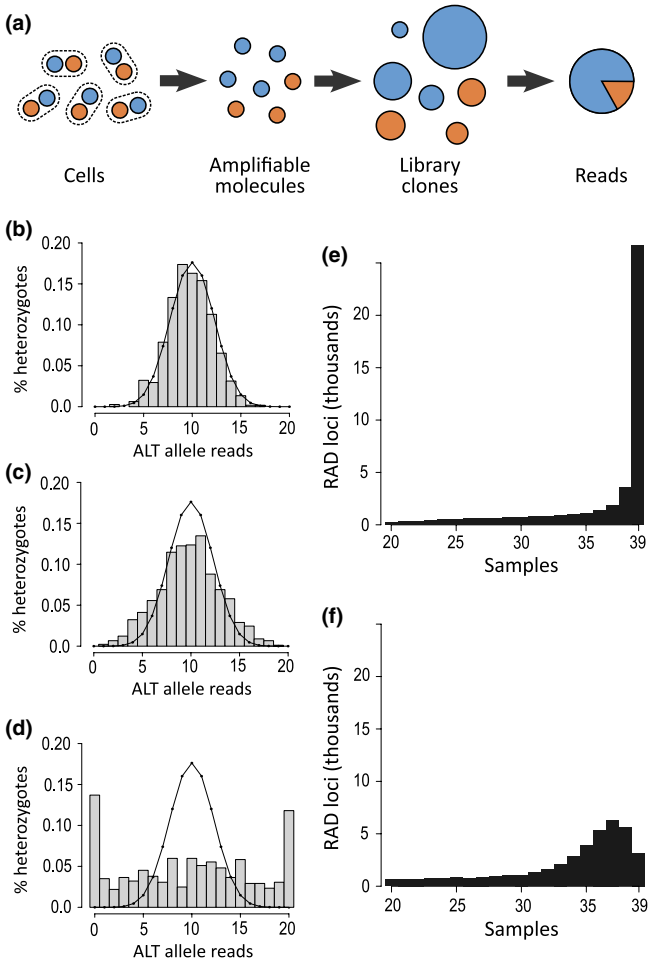

FIGURE 5 Consequences of low library complexity on genotyping. (a) Due to molecule sampling and noisy amplification, the alleles at a heterozygote locus may already be unequally represented in the library by the time it is sequenced, causing sequence data to be over-dispersed compared to the binomial distribution assumed by genotyping models. This artefact is expected to be particularly serious at low library complexities and can be mitigated by removing PCR duplicates (see Section 4). (b–d) Observed allelic counts at heterozygous sites of anole individual E35 that have a total depth of 20 reads, respectively in the Anole-600ng data set with (b) or without (c) PCR duplicate removal, and in the low-complexity Anole-30ng data set without PCR duplicate removal (d). The solid line shows the binomial distribution. Peaks at x = 0 and x = 20 in (d) are due to allelic dropout, where one allele is entirely absent from the library. Heterozygous sites were annotated based on the Anole-600ng data set with PCR duplicate removal. No figure is presented for the Anole-30ng data set with PCR duplicate removal as the low non-redundant coverage leaves no heterozygous sites with a depth of 20. (e, f) Histograms showing apparent locus sharing among the 39 individuals of the Anole data set, respectively for the Anole-600ng and Anole-30ng libraries. The latter, Anole-30ng library, exhibits 'locus dropout' because most individuals have such a low molecular density that at any locus several individuals typically fail to sample even a single amplifiable molecule.

quantify amplifiable DNA in the pre-PCR molecular library. We measured a template single-stranded DNA concentration of 0.192 nmol/L (Figure S5), which implies that the 40 µL used during library preparation corresponded to 7.68×10^{-6} nmol, *that is* 4.62 billion, single-stranded template molecules. Owing to the design of adapters in the RAD-seq protocol used here, these molecules each corresponded to independent double-stranded molecules, *that is* contributed fully to library complexity. Additionally, the density of the library was bioinformatically measured over the 89,162 loci found in more than half the samples, and these loci collectively represented 57.5% of the reads in the library (with the rest of read pairs corresponding to repetitive RAD loci, to RAD loci that are only found in one or a few individuals, or to genomic DNA not flanked by a restriction site). As this proportion must also hold in the library at the molecule level, we distributed 57.5% of 4.62×10^9 molecules (i.e. 2.66×10^9 molecules)

over 89,162 loci, so that we obtained a qPCR-based library density estimate of 29,800×.

Thus, the experimental, qPCR-based library density estimate was much lower (9.7 times less) than the estimate based on the C-value (288,000×) (Figure 4; Table 3). This confirmed that most of the DNA used for this amplification reaction was indeed inert, and that estimates of library complexity based on DNA mass alone may be inflated by more than an order of magnitude. Nevertheless, we note that the qPCR-based estimate remained 3.5 times higher than the in silico one; we cannot presently resolve this residual discrepancy.

3.6 | Increasing library complexity reduces PCR duplicate rates experimentally

Finally, the comparison between our two *A. sagrei* RAD-seq libraries, Anole-600ng and Anole-30ng, illustrates the experimental importance of library complexity. These two libraries were prepared from the same 39 DNA samples using protocols that were identical except for the volume used at the amplification step, respectively 6000 and 300 µL, which corresponded to 600 and 30 ng of template DNA. The yield of the smaller reaction was not technically limiting so that subsequent steps could be performed identically, and both libraries were then sequenced at the same depth of 49×.

As predicted, the observed library density was much higher for the Anole-600ng library than for the Anole-30ng one, respectively at 984× and 133×. Accordingly, the PCR duplicate rates were respectively 61% and 93% (for a sequencing depth of 49×), resulting nonredundant coverages of 19.1× and 3.4× per sample, on average (Table 2). This adds to previously published evidence (see Section 4) to demonstrate the critical role of the total amount of DNA used for PCR amplification, independently of all other experimental factors, in determining library complexity and PCR duplicate rates.

In addition, further inspection of these two data sets highlighted that library density was an intrinsic and separate property of these data sets, rather than merely a parameter fitted so that the model replicates PCR duplicate patterns. Since a low per-sample library density indicates that only a few molecules are stochastically sampled and amplified at each locus, we can expect that the makeup of the library itself will have a pronounced stochastic component, so that the resulting read data will be more variable than expected based on the randomness of the sequencing process alone (Figure 5a; and see Section 4).

This overdispersion is first apparent for coverage patterns at heterozygous sites. Considering the sequencing process alone, the number of reads observed for either allele should follow a binomial distribution. The Anole-600ng data approximately conforms to this expectation, especially when the effects of amplification stochasticity are mitigated by removing PCR duplicates Figure 5b,c. In the Anole-30ng data set, however, the number of observations for the two alleles deviate much more from equal representation, and in many cases one allele simply is not observed (Figure 5d). The latter is explained by allelic dropout: given that the per-sample density of the library is very low, at only 3.4 molecules per locus or 1.7 molecules per allele, on average, frequently no molecules will be sampled for a particular allele.

Second, at such low densities, there is also a chance that neither of the two alleles of a locus will be sampled, so that an entire locus may be stochastically dropped. While in RAD-seq analyses, it is normal that loci are not perfectly shared across individuals due to polymorphisms in restriction sites, comparing the extent of locus sharing in the Anole-600ng and Anole-30ng data sets (Figure 5e,f) makes it clear that most of the variation in locus composition across individuals in the Anole-30ng data is caused by stochastic locus dropout due to low library density rather than genetic polymorphism.

4 | DISCUSSION

4.1 | Determinants of PCR duplicate occurrence

PCR duplicates are a pervasive sequencing artefact and a wealth of empirical observations on their occurrence have been reported. However, this knowledge accumulation has often happened as a by-product of the development and validation of new molecular protocols, and the lack of a theoretical framework within which to connect individual results has not allowed a complete understanding of the artefact's causes and of its seriousness in specific applications (Marx, 2017). We fill this gap by introducing a mechanistic and quantitative model for the occurrence of PCR duplicates. This allows us to unify earlier results, to further clarify the effects of various experimental factors, and to draw expectations about the statistical properties of duplicate-containing data sets.

Our first observation is that the PCR duplicate rate fundamentally depends on the ratio between sequencing depth and library complexity (Figure 3). In particular, we stress the symmetry between depth and complexity, and that it is imperative to consider both when comparing PCR duplicate rates across experiments. The idea that PCR duplicates occur when a library is sequenced in excess relative to its complexity has been discussed in earlier works (Daley & Smith, 2013; Fu et al., 2018; Rao et al., 2014; Smith et al., 2017), and in principle could be extrapolated following Lander and Waterman (1988). However, these studies did not consider amplification artefacts or focused on a specific problem or application. For instance, the method of Daley and Smith (2013) proposed to estimate library complexity based on the species-saturation approach of Efron and Thisted (1976), which ultimately amounts to modelling amplification noise in the same way as presented here. However, as the authors' focus is solely on library complexity, this factor is then eliminated through nonparametric approximations and its effects are not discussed further.

The central role of library complexity and sequencing depth is experimentally supported by the RAD-seq-based results presented here, as well as by observations from earlier studies. The amount of material used as input for library preparation has been shown to strongly impact duplicate rates in RNA-seq (Fu et al., 2018) and

ancient DNA (Kapp et al., 2021). Very high duplicate rates are also apparent in whole genome resequencing libraries prepared from minute amounts of DNA (e.g. Bruinsma et al., 2018), and there is evidence that this parameter is also important in ATAC-seq and Hi-C (see below). As for the effect of sequencing depth, it is implicit in experiments where sequencing is performed to saturation (Daley & Smith, 2013; Farlik et al., 2015; Niu et al., 2019; Ziegenhain et al., 2017) and should generally be immediately apparent in any data set containing substantial levels of duplicates if the read data is down-sampled (Figure 3).

Second, since our model explicitly accounts for amplification, we were able to explore its behaviour across a range of amplification parameters. We find that PCR itself only plays a secondary role in the rate of occurrence of PCR duplicates. Although a greater variance in relative amplification factors across templates (i.e. a 'noisier' amplification) does elevate PCR duplicate rates, it only does so within a narrow range determined primarily by the depth-complexity ratio (Figure 3c).

This conclusion is in contrast with the frequent assertion that PCR duplicates are due to an excessive number of PCR cycles (Ebbert et al., 2016; Marx, 2017; Orlando et al., 2021; Smith et al., 2017; Stuart et al., 2018; Tin et al., 2015; Vargas-Landin et al., 2018). The absence of a major effect of PCR is nevertheless in agreement with published experimental results. In particular, Fu et al. (2018) specifically tested this interaction in RNA-seq and found none. Similarly, Tin et al. (2015) observed no change in RAD-seq duplicate rates when increasing the number of cycles. And while several studies have reported a strong effect of the number of cycles, in RNA-seq (Parekh et al., 2016), ATAC-seq (Lu et al., 2017), RAD-seq (Díaz-Arce & Rodríguez-Ezpeleta, 2019), and Hi-C (Niu et al., 2019), in each case the amount of starting material for the protocol was made to vary concurrently, so that either factor could be responsible for the observed differences. In fact, Fu et al. (2018) also made this observation, but subsequently found that amplification was not the causal factor. Thus, given theoretical expectations and other empirical results in this direction, these studies can more parsimoniously be interpreted as further evidence that library complexity plays a pivotal role in a wide range of sequencing approaches.

Although from an experimental perspective it makes sense to think jointly of the amount of material available and the number of amplification cycles to perform, confounding their effects hinders the formulation of clear and effective recommendations to improve experiments compromised by PCR duplicates. In particular, numerous authors have suggested that limiting the number of cycles was critical to minimize duplicates (Marx, 2017; Orlando et al., 2021; Stuart et al., 2018). Using fewer cycles should indeed lead to fewer PCR duplicates because it forces experimenters to use more starting material to achieve necessary yields. But the reverse is not true: if enough starting material is used, excess amplification should not appreciably impact the resulting duplicate rate. Thus, we find that with regard to PCR duplicates, in both RAD-seq and other experimental applications, experimenters should not be excessively concerned with the amplification protocol and should instead focus primarily on increasing the amount of (active) starting material for the amplification, even if this is the hardest thing to do as it may call for substantial alterations to the protocol (e.g. Kapp et al., 2021). Nevertheless, general optimization of the amplification is still advisable because a higher efficiency is directly associated to a lower amplification noise (see 'Amplification overdispersion' below).

4.2 | Measuring library complexity

Although library complexity and sequencing depth have exactly opposite roles in determining the PCR duplicate rate, this symmetry is only mathematical as in practice these two factors come with very different constraints. Sequencing depth is an experimental choice and is set by the experimenter to a level called for by the intended application (e.g. upwards of 30× for single-individual germline genotyping (Sims et al., 2014), 1× for low-coverage population genotyping (Lou et al., 2021), or 10–30 M reads for RNA-seq in mammals (Stark et al., 2019)) and can be adjusted a posteriori by sequencing again if required. Library complexity, in contrast, is established permanently during library preparation so that the only way around insufficient library complexity is to prepare new libraries (Rao et al., 2014). In this context, it is critical for experimenters to achieve an appropriate complexity at the time of library preparation. A complexity at least one order of magnitude higher than the required sequencing depth is ideal for most approaches, although an interesting exception is linked-read approaches (Meier et al., 2021; Zheng et al., 2016) where a high sequencing saturation is necessary to sample multiple fragments associated with each tag, so that library complexity should instead be matched to the intended sequencing depth (Weisenfeld et al., 2017).

The experimental number which is most relevant to library complexity is the number of initial molecules that effectively undergo amplification. Once a library has been amplified, each unique molecular species is present in many copies and is therefore unlikely to be lost, so that the complexity then remains essentially constant at later steps of the library preparation protocol. In other words, library complexity is fundamentally determined by the amount and quality of DNA that is input into the PCR. Consequently, reducing PCR duplicate levels primarily amounts to amplifying more template DNA. However, it is important to stress that only a fraction of a pool of DNA is usually amplifiable, so that mass is a poor proxy for the amount of useful DNA (Gansauge & Meyer, 2013; Kapp et al., 2021; Meyer et al., 2008) and the complexity of libraries prepared using the same input mass but from samples differing in quality or using distinct protocols may greatly differ. A more direct estimate of library complexity can be obtained by qPCR quantification of the library before it is amplified, as this directly measures the concentration of active molecules. This assay also informs on the percentage yield of the protocol that was used (Gansauge & Meyer, 2013; Kapp et al., 2021), which ultimately is what determines how complex a library derived from a given finite sample can be.

4.3 | Measuring library complexity as a molecular density

As noted above, the minimum acceptable library complexity—measured as an absolute number of distinct molecules—may vary by several orders of magnitude depending on the intended application. This definition of library complexity is thus only meaningful to compare closely related experiments, rather than in a general sense, making it difficult to establish guidelines. Here, we argue that measuring library complexity in terms of the number of molecules per locus will often be more informative than the above 'absolute' library complexity. In the context of sequencing approaches aiming for a homogenous coverage (e.g. whole-genome, exome, bisulfide, or RAD-seq), scaling absolute library complexity by the size of the genomic regions considered leads to a measure that can be directly compared to coverage, and which we refer to as the molecular density of a library. As this measure considers the sequencing target, it becomes possible to provide quantitative recommendations. For instance, for single-individual genotyping at 30× coverage, library molecular density should ideally be greater than 300×, assuming a one-order-of-magnitude margin. This target value applies regardless of whether the sequencing target is a whole genome or an exome, or if the organism considered has a genome size substantially different from that of mammals, as is the case for many species of medical, agricultural, or ecological interest (e.g. *Drosophila melanogaster*, *Danio rerio*, and *Arabidopsis thaliana*).

In addition, the molecular density of a library is more relevant than absolute complexity both when considering amplification-related statistical error patterns in downstream computational analyses (see below) and when evaluating experimental yield. The latter is because it is directly related to units that are used for experimental inputs, such as number of cells (Brind'Amour et al., 2015; Butler, 2015; Lu et al., 2017), genome equivalents (Lander & Waterman, 1988; Zheng et al., 2016) (i.e. mass expressed in units of the C-value), or template DNA units (Taberlet et al., 1996) (equal to two genome equivalents). The relationship with yield is most obvious when the input is measured in number of cells: if one starts an experiment with 1000 diploid cells, comprising 2000 genomes copies, the maximum number of unique molecules that may be sequenced for any genomic region is also 2000. Molecular density measures how many molecules per locus (on average across the genome) are present in the final library, and therefore represents the overall yield. Taking this logic to the limit highlights that molecular density is also closely related to the 'genome coverage' yield statistic used in single-cell whole-genome sequencing experiments (Daley & Smith, 2014; Zhang et al., 2015): especially, for haploid cells sequenced to saturation, the two become identical.

For the same reason, comparable library densities may be achieved from a given amount of input sample for whole-genome, exome, or RAD-seq, even though the absolute library complexities will vary by orders of magnitude (with no effect on the success of the experiment, as proportionately less molecules are needed to describe an exome than a whole genome). And remarkably, this rationale also applies to heterogenous-coverage approaches. For instance, when performing ChIP-seq on a histone mark for which few peaks exist, for a given number of input cells the resulting yield and absolute library complexity can be expected to be lower than for a more frequent mark, but this does not necessarily imply that the local molecular density (which determines the statistical properties of the data, see below) will be lower at the rare-mark peaks, because the molecules in the library are spread out over fewer loci.

Thus, when examining the yield of a library preparation protocol, we recommend expressing input sample amounts in number of cells or in genome equivalents (i.e. multiples of the C-value mass), and to measure library complexity as a molecular density rather than as an absolute number of unique molecular species.

4.4 | Modelling amplification-related overdispersion

Amplification-related artefacts cause increased technical variance (Casbon et al., 2011), which has been demonstrated for instance in RNA-seq (Castel et al., 2015; Fu et al., 2018; Parekh et al., 2016), single-cell RNA-seq (Grün et al., 2014; Islam et al., 2014; Ziegenhain et al., 2017), iCLIP (Smith et al., 2017), or Hi-C (Niu et al., 2019). These artefacts have also been reported to cause genotyping errors (Andrews & Luikart, 2014; Bresadola et al., 2020; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Taberlet et al., 1996; Tin et al., 2015). The conceptual framework presented here provides insights on the mechanisms that give rise to these error patterns.

Essentially, underlying our model is the view that technical variance in sequencing data is the result of three successive sampling steps: (i) the acquisition of actually amplifiable molecules from the biological sample, (ii) the noisy amplification of these molecules and (iii) the selection of amplified molecules for sequencing to generate reads (Figures 1a and 5a). In many methods of sequence data analysis, variance is modelled primarily after read depth, thus only the third step above is accounted for. This may lead to inaccurate variance estimations, particularly if the magnitude of amplification-related artefacts varies among libraries in the considered data set, or worse, this may cause methods that do not estimate residual variance (e.g. genotyping, see below) to produce unreliable results by failing to acknowledge the overdispersion of the data. How misspecified a model is for a particular data set depends on the relative scales of the variances introduced at each of the three steps, and especially whether the third step—sequencing—is actually the predominant source of variance in the data set.

Both the first and the third steps are Poisson processes, with means respectively equal to the molecular density of the library and to the sequencing coverage. The second step, amplification, is a complex process, but generally its variance (i.e. the unequal amplification of templates) can be partitioned in two components: systematic bias and stochasticity (Best et al., 2015; Kebschull & Zador, 2015). Systematic bias is caused by differences in amplification efficiency among templates (e.g. due to differences in molecule

length) and simply increases geometrically with the number of cycles. In contrast, stochasticity, which can represent most of the variance (Kebschull & Zador, 2015), is due to the partial efficiency of amplification at each cycle (some molecules are amplified and some are not) which is magnified by the exponential nature of PCR. This stochasticity is strongest during the first few cycles when clone sizes are still small, then subsides as clones grow and clone-wise efficiency becomes predictable (Kebschull & Zador, 2015). Figure S6 shows how these components combine to yield the overall amplification variance distribution. Importantly for experimental purposes, mechanisms underlying both variance components depend on amplification efficiency, so that amplification noisiness is minimal when per-cycle efficiency is high.

The PCR duplicate phenomenon intersects this three-step variance model in several ways. First, as discussed above, an absence of PCR duplicates is an indication that the molecular density of the library is much greater than the sequencing coverage. This directly implies that the variance of the first step's Poisson sampling (of amplifiable DNA molecules) is negligible in comparison with that of the third step's Poisson sampling (of reads). In addition, the presence of a large number of molecules for each locus causes stochastic amplification effects to average out, so that the relative variance of the second step is also reduced. Thus, if the PCR duplicate rate is negligible, the first and second steps can be neglected. On the contrary, a high PCR duplicate rate implies that sequencing coverage is comparable to or greater than the molecular density of the library. In this case, the variances associated with molecule sampling and noisy amplification can be comparable to or exceed the variance associated with sequencing itself, leading to increased technical variance and model mis-specification.

In addition, while PCR duplicates can indicate serious underlying amplification artefacts, at the same time they offer the opportunity to monitor and reduce these artefacts. Tracking PCR duplicates allows one to unwind the three-step process a posteriori. In the extreme case, sequencing a library to saturation while removing duplicates should lead to an apparent sequencing coverage equal to the library's molecular density and a complete reduction of amplification noise, essentially simplifying the entire sampling process to only its first step—the sampling of amplifiable molecules. This saturation effect has been observed in single-cell RNAseq (Islam et al., 2014).

In data sets with intermediate duplication levels, the effects of removing PCR duplicates are less straightforward. However, if we assume no systematic differences in amplification efficiency between templates, and considering that removing PCR duplicates is equivalent to keeping track of which amplification clones have received at least one read, then it becomes possible to merge all three steps into a single Poisson process with mean equal to the nonredundant coverage. This re-emergence of the Poisson distribution after PCR duplicate removal suggests that the common practice of using bioinformatic methods designed around the sequencing process alone to process de-duplicated data sets is reasonable.

4.5 | Effects of PCR duplicates on genotyping

We will now focus on genotyping, but we note that the mechanistic framework described here should also be useful to better understand technical noise in other sequencing technologies. For instance, in RNA-seq, it represents a possible explanation for the non-linear relationships between technical noise, input starting material, and gene abundancy (Brennecke et al., 2013)—technical variance at a particular gene should be determined by the local molecular density, which is proportional to both the absolute complexity of the library and to the gene's expression level. In single-cell RNA-seq, it could inform how the occurrence of a major technical noise feature, gene dropout (Kiselev et al., 2019), is influenced by the complexities of the sub-libraries for each cell, the noisiness of the amplification, and the relative sequencing effort.

Amplification artefacts have long been known to be a source of error for genotyping (Pompanon et al., 2005; Taberlet et al., 1996), and it is standard practice to remove PCR duplicates before genotyping. As discussed above, this approach should suppress model mis-specification and yield reliable genotypes, even in cases where initial PCR duplicates rates are high—provided that duplication clones can be identified accurately and that the resulting coverage is sufficient.

However, several recent studies have interrogated the degree to which retaining PCR duplicates impacted the accuracy of genotype calls (Andrews & Luikart, 2014; Bresadola et al., 2020; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Euclide et al., 2019; Flanagan & Jones, 2018; Tin et al., 2015) as certain RAD-seq approaches do not permit the identification and removal of duplicates (reviewed in Andrews et al., 2016). Surprisingly, while some studies have confirmed the intuitive expectation that PCR duplicates negatively impact genotyping, others have recently reported convincing evidence that genotyping could be reliable even in data sets containing substantial duplicate rates (Bresadola et al., 2020; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Euclide et al., 2019; Tin et al., 2015).

Our three-step model allows us to make specific predictions on the effects of amplification-related artefacts on genotyping and to reconcile these apparently contradictory results. The model suggests that amplification-related artefacts impact genotype calls by causing overdispersion of the allelic ratios observed at heterozygous sites. Genotyping models are based on the binomial distribution, with modifications to account for sequencing errors (DePristo et al., 2011; Li, 2011; Maruki & Lynch, 2017; Rochette et al., 2019). This, however, assumes that the allelic ratio is balanced in the final library being sequenced, which can be incorrect in amplified libraries (Figure 5a). Genotyping models thus underestimate the likelihood of imbalanced allelic ratios when the underlying genotype is a heterozygote. This leads to bias against heterozygotes, which manifests itself either as a loss of power to call heterozygotes, or in the worst case as calling a heterozygote site as homozygote.

Remarkably, how incorrect genotyping becomes if PCR duplicates are not removed does not just depend on the duplicate rate (itself determined by the depth-complexity ratio), but also on the

absolute values of sequencing coverage and library density. This is because calling discrete genotypes involves a strong threshold effect. Indeed, a heterozygote site will be called correctly as long as the heterozygote likelihood is significantly larger than the homozygote one. This of course depends on the model, but also on the data itself: if coverage and density are large enough that several reads are consistently observed for both alleles, homozygote likelihoods will be consistently small, so that correct calls can be made regardless of model mis-specification.

In practice, the key statistic is the nonredundant coverage—the apparent coverage after duplicates have been removed, which depends on both sequencing coverage and library density. If the nonredundant coverage is high enough for reliable genotype calls to be made, for instance >20×, it can be expected that calls made without removing duplicates should also be reasonable. This seemingly was the case in the data of Euclide et al. (2019), which may explain why they found, in apparent contradiction with other studies, that removing PCR duplicates had little influence on genotyping. The argument also applies to the results of Ebbert et al. (2016), whose data featured a PCR duplicate rate of just 2%.

In contrast, if the nonredundant coverage is too low for reliable genotype calling, attempting to genotype samples without removing PCR duplicates will lead to important biases. The strong overdispersion of the allelic ratios observed at heterozygous sites implies that genotyping models underestimate the chance that all the sampled reads come from the same allele and may thus confidently (but wrongly) call homozygous genotypes at sites that are actually heterozygous. This is essentially a theory of the occurrence of allelic dropout (Broquet & Petit, 2004; Taberlet et al., 1996) in high-throughput sequence data and explains the main error patterns reported in the literature, namely miscalled heterozygotes and deflated heterozygosity (Bresadola et al., 2020; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Flanagan & Jones, 2018; Tin et al., 2015).

This allelic dropout phenomenon is most obvious and most pronounced in low complexity libraries, that have a density smaller than 10×. For instance, if a heterozygote locus is represented by just three molecules, chances are high that all three will represent the same allele. If 20 reads are then sequenced and presented to a genotyping model, this model will confidently infer a homozygote, because it sees a single allele in a relatively large sample size. Assuming the three molecules have each been sequenced at least once, removing PCR duplicates would leave three reads, and the model would correctly conclude that such a small sample size comprises too little information to make a reliable genotype call.

We conclude that genotypes reported without information about the PCR duplicate rate may be unreliable, and therefore that it is crucial to monitor PCR duplicate rates in genotyping experiments. We confirm that removing PCR duplicates is a simple and efficient strategy to mitigate the detrimental effects of amplification-related artefacts on genotyping, and that it comes at a very small cost in power since the information that is discarded is mostly redundant. Although data sets with relatively high PCR duplicate rates can sometimes produce acceptable genotypes even without removing duplicates (Euclide et al., 2019), this happens, perhaps frustratingly, precisely when the nonredundant coverage is high and avoiding this filter presents little appeal. Nevertheless, removing PCR duplicates is preferable even in these cases, as full-coverage genotypes should still be expected to suffer from lesser artefacts such as biased genotyping confidence scores.

4.6 | Implications for RAD-seq studies

Our results and those of others (Schweyen et al., 2014; Tin et al., 2015) demonstrate that PCR duplicates are present at substantial levels in both single-digest and double-digest RAD-seq experiments (Table 1). Given that genotyping quality is primarily determined by non-redundant coverage (see above), and consequently that genotype calling is unreliable in experiments where PCR duplicates are not monitored, controlling for PCR duplicates in RAD-seq experiments is critical. We thus advocate the systematic use of protocols that allow their identification, *that is* either single-digest protocols (Ali et al., 2016; Baird et al., 2008) combined with paired-end sequencing, or double-digest protocols combined with UMIs (Hoffberg et al., 2016; Schweyen et al., 2014; Tin et al., 2015). This is essential for demographic analyses, as they tend to be more sensitive to genotyping errors than population structure analyses, and are sensitive in particular to biased estimates of heterozygosity and/or of the number of singleton alleles.

In addition, these results highlight that PCR duplicate rates vary extensively across experiments. This variation must be partly due to differences in library complexity, but we stress that sequencing depth also plays an important role (Figure 3). In comparing RAD-seq libraries, the measure of choice for library complexity should be library density, which determines both the maximum achievable nonredundant coverage and the cost-efficiency of sequencing (i.e. the number of reads required to achieve a given nonredundant coverage). Importantly, most studies do not report enough information to estimate the complexities of their libraries. We therefore encourage researchers to report all the technical values relevant to understanding the occurrence of PCR duplicates in RAD-seq, namely: the number of samples that were pooled in each library, the mass of DNA that was amplified, the number of RAD loci kept in the analysis, the per-library number of (possibly paired) reads aligning to these loci, the per-library average PCR duplicate rate, and the per-library average number of non-redundant reads per locus.

Nevertheless, the widespread occurrence of PCR duplicates suggests that RAD-seq experiments would typically benefit from libraries with higher densities. As discussed above, this involves amplifying greater quantities of template DNA whenever possible. How much DNA should ideally be amplified depends on the mass of the haploid genome of the study organism (i.e. its C-value), on the quality of the input DNA and percentage yield of the protocol, as well as—crucially for pooled libraries such as those typically seen in RAD-seq—on the number of individuals in the library. Importantly, pooling does not change the per-sample

library complexity requirements for genotyping. Pooled libraries therefore require handling considerable amounts of DNA (at least before amplification) to achieve similar per-sample complexities. For instance, amplifying 1 microgram of genomic DNA should not be out of the ordinary for a 100-sample library, as this represents 10 nanograms per sample. This logic is not directly applicable to bestRAD (Ali et al., 2016) because amplification is performed on purified (restriction-site digested, ligated and bead-captured) DNA rather than on genomic DNA; however, all numbers above can instead be applied to the genomic DNA input to the bead-capture step.

Importantly, insufficient library densities can explain allelic dropout, which is another error pattern that has been extensively discussed in the context of RAD-seq, often in combination with concern about restriction site polymorphism (Andrews et al., 2016; Davey et al., 2011; Gautier et al., 2013). However, as underlined by the heterozygosity deficit caused by amplification-related artefacts, allelic dropout can also result from the purely stochastic non-sampling of one of the two alleles of a heterozygote, which is particularly likely to happen in libraries that have a low molecular density. This is in fact the mechanism that the original definition of allelic dropout refers to (Broquet & Petit, 2004; Taberlet et al., 1996).

Moreover, population genetics simulations suggest that restriction site polymorphism should have little consequences unless genetic diversity is very high (Gautier et al., 2013; Rivera-Colón et al., 2021). In contrast, insufficient library density can create dramatic allelic dropout, and can even result in locus dropout for densities so low that it becomes likely that neither allele is sampled (Figure 5). Thus, we suggest that in the absence of PCR duplicate tracking, evidence for allelic dropout in a data set should be considered indicative of a low library density (and of a high duplicate rate) rather than of restriction site polymorphism. Finally, we note that stochastic allelic dropout is a non-issue in data sets where duplicates have been removed, as using nonredundant reads prompts the genotyping models to account for the phenomenon.

5 | CONCLUDING REMARKS

We propose a realistic and predictive model for the distortions that amplification introduces in sequencing libraries. This establishes a conceptual framework within which the numerous observations made on PCR duplicates can be united, the causes of the artefact can be quantified, and their consequences on downstream analyses understood. In particular, we find that:

- PCR duplicate rates are determined mainly by the ratio between sequencing depth and library complexity. Nevertheless, as most experiments target a particular sequencing depth, amplifying a large enough DNA pool is critical to maintain a low depth-complexity ratio and limit PCR duplicates; thus, the amount of active DNA input into the amplification reaction during library preparation is key.
- In most cases, it is advantageous to measure library complexity not as an absolute number of molecules, but as a molecular coverage—molecular density—as this statistic is more transferable between experiments, more relatable to experimental yield, and more directly interpretable into expected error patterns in statistical analyses.
- Because PCR duplicate rates depend on both library complexity and sequencing depth, the reporting and interpretation of PCR duplicate rates should always be accompanied by depth considerations.
- PCR amplification is stochastically uneven, and this has an effect on the statistical properties of libraries, including a secondary role in determining duplicate rates and a lowering of the apparent library complexity. However, the number of cycles can be expected to only have a marginal effect on duplicate rates, which is in contrast with a widespread conception, but in agreement with the experimental results that exist in the literature. It is important to maximize per-cycle amplification efficiency, as this reduces stochasticity and bias.
- Our framework can be leveraged to better understand the error patterns that amplification noisiness creates in downstream statistical analyses. We demonstrate this for the case of genotyping, which makes us able to propose recommendations to improve the effectiveness of RAD-seq experiments.
- For RAD-seq and other molecular applications, researchers should focus on maximizing the quantity of input DNA that is used for library preparation, particularly the amounts used as template for PCR. They should be aware of how larger genome sizes, a high number of individuals per library, and/or the presence of low-quality samples reduce the effective library complexity and modify molecular protocols accordingly.

AUTHOR CONTRIBUTIONS

NCR, AGR-C, SCC-S, and JMC designed experiments. NCR and AGR-C performed experiments. NCR drafted the manuscript. AGR-C and JMC edited the manuscript. JW prepared the *Anolis* sequencing libraries. TJS provided and prepared the *Anolis* samples. SCC-S and JMC provided funding. All authors approved the final manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Mark E. Hauber and Alec B. Luro for access to the robin RAD-seq data set, Sarai H. Stuart for assistance with the qPCR, and Giovanni Madrigal and Kira M. Long for their feedback in the writing of the manuscript. AGR-C and NCR were supported by NSF grant 1645087.

CONFLICT OF INTEREST STATEMENT

JW is an employee of Floragenex, Inc. Other authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Decoratio is written in Python using NumPy and SciPy and is available from PyPI or from Bitbucket (bitbucket.org/rochette/decoratio) under the GNU GPL v3 licence. Raw sequencing data for the Anole and Robin data sets is available in NCBI-SRA Bioproject PRJNA925469 and PRJNA925213, respectively.

BENEFIT SHARING STATEMENT

Benefits from this research accrue from the sharing of our experimental findings with the wider research community to improve future experimental analyses.

ORCID

Angel G. Rivera-Colón https://orcid.org/0000-0001-9097-3241

Julian M. Catchen https://orcid.org/0000-0002-4798-660X

REFERENCES

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Andrews, C. B., Mackenzie, S. A., & Gregory, T. R. (2009). Genome size and wing parameters in passerine birds. *Proceedings. Biological Sciences*, *276*(1654), 55–61. https://doi.org/10.1098/rspb.2008.1012

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis. *Molecular Ecology*, *23*(7), 1661–1667. https://doi.org/10.1111/mec.12686

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Bay, R. A., Harrigan, R. J., Underwood, V. L., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science*, *359*(6371), 83–86. https://doi.org/10.1126/science.aan4380

Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J., & Chain, B. (2015). Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, *5*, 14629. https://doi.org/10.1038/srep14629

Bonfiglio, S., Vanni, I., Rossella, V., Truini, A., Lazarevic, D., Dal Bello, M. G., Alama, A., Mora, M., Rijavec, E., Genova, C., Cittaro, D., Grossi, F., & Coco, S. (2016). Performance comparison of two commercial human whole-exome capture systems on formalin-fixed paraffin-embedded lung adenocarcinoma samples. *BMC Cancer*, *16*(1), 692. https://doi.org/10.1186/s12885-016-2720-4

Borgström, E., Paterlini, M., Mold, J. E., Frisen, J., & Lundeberg, J. (2017). Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS One*, *12*(2), e0171566. https://doi.org/10.1371/journal.pone.0171566

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., & Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–1095. https://doi.org/10.1038/nmeth.2645

Bresadola, L., Link, V., Buerkle, C. A., Lexer, C., & Wegmann, D. (2020). Estimating and accounting for genotyping errors in RAD-seq experiments. *Molecular Ecology Resources*, *20*(4), 856–870. https://doi.org/10.1111/1755-0998.13153

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M. M., & Lorincz, M. C. (2015). An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nature Communications*, *6*, 6033. https://doi.org/10.1038/ncomms7033

Broquet, T., & Petit, E. (2004). Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology*, *13*(11), 3601–3608. https://doi.org/10.1111/j.1365-294X.2004.02352.x

Bruinsma, S., Burgess, J., Schlingman, D., Czyz, A., Morrell, N., Ballenger, C., Meinholz, H., Brady, L., Khanna, A., Freeberg, L., Jackson, R. G., Mathonet, P., Verity, S. C., Slatter, A. F., Golshani, R., Grunenwald, H., Schroth, G. P., & Gormley, N. A. (2018). Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics*, *19*(1), 722. https://doi.org/10.1186/s12864-018-5096-9

Butler, J. M. (2015). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1674), 20140252. https://doi.org/10.1098/rstb.2014.0252

Casbon, J. A., Osborne, R. J., Brenner, S., & Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, *39*(12), e81. https://doi.org/10.1093/nar/gkr217

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, *16*, 195. https://doi.org/10.1186/s13059-015-0762-6

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*, 362–365. https://doi.org/10.1111/1755-0998.12669

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., Ruan, Y., Bickel, P. J., Myers, R. M., Wold, B. J., White, K. P., Lieb, J. D., & Liu, X. S. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, *9*(6), 609–614. https://doi.org/10.1038/nmeth.1985

Cristofari, R., Bertorelle, G., Ancel, A., Benazzo, A., Le Maho, Y., Ponganis, P. J., Stenseth, N. C., Trathan, P. N., Whittington, J. D., Zanetti, E., Zitterbart, D. P., Le Bohec, C., & Trucchi, E. (2016). Full circumpolar migration ensures evolutionary unity in the emperor penguin. *Nature Communications*, *7*, 11842. https://doi.org/10.1038/ncomms11842

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Daley, T., & Smith, A. D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics (Oxford, England)*, *30*(22), 3159–3165. https://doi.org/10.1093/bioinformatics/btu540

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD sequencing data: Implications for genotyping. *Molecular Ecology*, *22*(11), 3151–3164. https://doi.org/10.1111/mec.12084

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

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011).

A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-seq data analysis parameters for population genetics: The more the better? *Frontiers in Genetics*, *10*, 533. https://doi.org/10.3389/fgene.2019.00533

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Letter to the editor. *Cytometry Part A*, *51A*(2), 127–128. https://doi.org/10.1002/cyto.a.10013

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Alzheimer's Disease Neuroimaging Initiative, Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(Suppl 7), 239. https://doi.org/10.1186/s12859-016-1097-3

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*(3), 435–447. https://doi.org/10.1093/biomet/63.3.435

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., & Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, *6*(4), e18561. https://doi.org/10.1371/journal.pone.0018561

Euclide, P. T., McKinney, G. J., Bootsma, M., Tarsa, C., Meek, M. H., & Larson, W. A. (2019). Attack of the PCR clones: Rates of clonality have little effect on RAD-seq genotype calls. *Molecular Ecology Resources*., *20*, 66–78. https://doi.org/10.1111/1755-0998.13087

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., & Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, *10*(8), 1386–1397. https://doi.org/10.1016/j.celrep.2015.02.001

Flanagan, S. P., & Jones, A. G. (2018). Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. *Molecular Ecology Resources*, *18*(2), 264–280. https://doi.org/10.1111/1755-0998.12734

Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, *19*(1), 531. https://doi.org/10.1186/s12864-018-4933-1

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Bradley, D. G., & Orlando, L. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, *16*(2), 459–469. https://doi.org/10.1111/1755-0998.12470

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*(11), 3165–3178. https://doi.org/10.1111/mec.12089

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews. Genetics*, *17*(3), 175–188. https://doi.org/10.1038/nrg.2015.16

Geneva, A. J., Park, S., Bock, D. G., de Mello, P. L. H., Sarigol, F., Tollis, M., Donihue, C. M., Reynolds, R. G., Feiner, N., Rasys, A. M., Lauderdale, J. D., Minchey, S. G., Alcala, A. J., Infante, C. R., Kolbe, J. J., Schluter, D., Menke, D. B., & Losos, J. B. (2022). Chromosome-scale genome assembly of the brown anole (*Anolis sagrei*), an emerging model species. *Communications Biology*, *5*, 1126. https://doi.org/10.1038/s42003-022-04074-5

Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., Connelly, J., Pruett-Miller, S., Chen, X., Easton, J., & Gawad, C. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(24), e2024176118. https://doi.org/10.1073/pnas.2024176118

Grün, D., Kester, L., & van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, *11*(6), 637–640. https://doi.org/10.1038/nmeth.2930

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: Sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, *16*, 1264–1278. https://doi.org/10.1111/1755-0998.12566

Hua, P., Badat, M., Hanssen, L. L. P., Hentges, L. D., Crump, N., Downes, D. J., Jeziorska, D. M., Oudelaar, A. M., Schwessinger, R., Taylor, S., Milne, T. A., Hughes, J. R., Higgs, D. R., & Davies, J. O. J. (2021). Defining genome architecture at base-pair resolution. *Nature*, *595*(7865), 125–129. https://doi.org/10.1038/s41586-021-03639-4

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. https://doi.org/10.1038/nmeth.2772

Kapp, J. D., Green, R. E., & Shapiro, B. (2021). A Fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *The Journal of Heredity*, *112*(3), 241–249. https://doi.org/10.1093/jhered/esab012

Kebschull, J. M., & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, *43*(21), e143. https://doi.org/10.1093/nar/gkv717

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews. Genetics*, *20*(5), 273–282. https://doi.org/10.1038/s41576-018-0088-9

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. https://doi.org/10.1038/nmeth.1778

Kofler, R., Nolte, V., & Schlötterer, C. (2016). The impact of library preparation protocols on the consistency of allele frequency estimates in Pool-seq data. *Molecular Ecology Resources*, *16*(1), 118–122. https://doi.org/10.1111/1755-0998.12432

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295. https://doi.org/10.1038/nmeth.1311

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, *2*(3), 231–239.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*, 5966–5993. https://doi.org/10.1111/mec.16077

Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M., & Schmitz, R. J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Research*, *45*(6), e41. https://doi.org/10.1093/nar/gkw1179

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3 (Bethesda, Md.)*, *7*, 1393–1404. https://doi.org/10.1534/g3.117.039008

Marx, V. (2017). How to deduplicate PCR. *Nature Methods*, *14*(5), 473–476. https://doi.org/10.1038/nmeth.4268

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N. J., Bridle, J. R., Rolian, C., Barton, N. H., McMillan, W. O., Jiggins, C. D., & Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), e2015005118. https://doi.org/10.1073/pnas.2015005118

Meyer, M., Briggs, A. W., Maricic, T., Höber, B., Höffner, B., Krause, J., Weihmann, A., Pääbo, S., & Hofreiter, M. (2008). From micrograms to picograms: Quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research*, *36*(1), e5. https://doi.org/10.1093/nar/gkm1095

Nelson, T. C., & Cresko, W. A. (2018). Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evolution Letters*, *2*(1), 9–21. https://doi.org/10.1002/evl3.37

Niu, L., Shen, W., Huang, Y., He, N., Zhang, Y., Sun, J., Wan, J., Jiang, D., Yang, M., Tse, Y. C., Li, L., & Hou, C. (2019). Amplification-free library preparation with SAFE hi-C uses ligation products for deep sequencing to improve traditional hi-C analysis. *Communications Biology*, *2*, 267. https://doi.org/10.1038/s42003-019-0519-y

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 14. https://doi.org/10.1038/s43586-020-00011-0

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, *6*, 25533. https://doi.org/10.1038/srep25533

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews. Genetics*, *6*(11), 847–859. https://doi.org/10.1038/nrg1707

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

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

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

Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin*, *227*(2), 146–160. https://doi.org/10.1086/BBLv227n2p146

Shigemizu, D., Aiba, T., Nakagawa, H., Ozaki, K., Miya, F., Satake, W., Toda, T., Miyamoto, Y., Fujimoto, A., Suzuki, Y., Kubo, M., Tsunoda, T., Shimizu, W., & Tanaka, T. (2015). Exome analyses of long QT syndrome reveal candidate pathogenic mutations in calmodulin-interacting genes. *PLoS One*, *10*(7), e0130329. https://doi.org/10.1371/journal.pone.0130329

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Smith, E. N., Jepsen, K., Khosroheidari, M., Rassenti, L. Z., D'Antonio, M., Ghia, E. M., Carson, D. A., Jamieson, C. H., Kipps, T. J., & Frazer, K. A. (2014). Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biology*, *15*(8), 420. https://doi.org/10.1186/s13059-014-0420-4

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499. https://doi.org/10.1101/gr.209601.116

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stuart, T., Buckberry, S., & Lister, R. (2018). Approaches for the analysis and interpretation of whole genome bisulfite sequencing data. *Methods in Molecular Biology (Clifton, N.J.)*, *1767*, 299–310. https://doi.org/10.1007/978-1-4939-7774-1_17

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L. P., & Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, *24*(16), 3189–3194. https://doi.org/10.1093/nar/24.16.3189

Tian, S., Peng, S., Kalmbach, M., Gaonkar, K. S., Bhagwate, A., Ding, W., Eckel-Passow, J., Yan, H., & Slager, S. L. (2019). Identification of factors associated with duplicate rate in ChIP-seq data. *PLoS One*, *14*(4), e0214723. https://doi.org/10.1371/journal.pone.0214723

Tin, M. M. Y., Rheindt, F. E., Cros, E., & Mikheyev, A. S. (2015). Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, *15*(2), 329–336. https://doi.org/10.1111/1755-0998.12314

Vargas-Landin, D. B., Pflüger, J., & Lister, R. (2018). Generation of whole genome bisulfite sequencing libraries for comprehensive DNA methylome analysis. *Methods in Molecular Biology (Clifton, N.J.)*, *1767*, 291–298. https://doi.org/10.1007/978-1-4939-7774-1_16

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767. https://doi.org/10.1101/gr.214874.116

Zhang, C.-Z., Adalsteinsson, V. A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K. L., Meyerson, M., & Love, J. C. (2015). Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nature Communications*, *6*, 6822. https://doi.org/10.1038/ncomms7822

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., … Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, *34*(3), 303–311. https://doi.org/10.1038/nbt.3432

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, *65*(4), 631–643.e4. https://doi.org/10.1016/j.molcel.2017.01.023

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rochette, N. C., Rivera-Colón, A. G., Walsh, J., Sanger, T. J., Campbell-Staton, S. C., & Catchen, J. M. (2023). On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity. *Molecular Ecology Resources*, *23*, 1299–1318. https://doi.org/10.1111/1755-0998.13800