MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 1764–1779 doi: 10.1111/mec.12693

Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics

B. NEVADO,*†[1] S. E. RAMOS-ONSINS* and M. PEREZ-ENCISO*† ‡

Centre for Research in Agricultural Genomics, Campus UAB, 08193 Bellaterra, Spain, †Universitat Autònoma de Barcelona, Bellaterra, Spain, ‡Institut Català de Recerca I Estudis Avancats (ICREA), Barcelona, Spain

Abstract

Decreasing costs of next-generation sequencing (NGS) experiments have made a wide range of genomic questions open for study with nonmodel organisms. However, experimental designs and analysis of NGS data from less well-known species are challenging because of the lack of genomic resources. In this work, we investigate the performance of alternative experimental designs and bioinformatics approaches in estimating variability and neutrality tests based on the site-frequency-spectrum (SFS) from individual resequencing data. We pay particular attention to challenges faced in the study of nonmodel organisms, in particular the absence of a species-specific reference genome, although phylogenetically close genomes are assumed to be available. We compare the performance of three alternative bioinformatics approaches – genotype calling, genotype–haplotype calling and direct estimation without calling genotypes. We find that relying on genotype calls provides biased estimates of population genetic statistics at low to moderate read depth (2–8×). Genotype–haplotype calling returns more accurate estimates irrespective of the divergence to the reference genome, but requires moderate depth (8–20×). Direct estimation without calling genotypes returns the most accurate estimates of variability and of most SFS tests investigated, including at low read depth (2–4×). Studies without species-specific reference genome should thus aim for low read depth and avoid genotype calling whenever individual genotypes are not essential. Otherwise, aiming for moderate to high depth at the expense of number of individuals, and using genotype–haplotype calling, is recommended.

Keywords: neutrality tests, population genomics, population variability, resequencing studies, simulation study

Received 26 August 2013; revision received 5 February 2014; accepted 5 February 2014

Introduction

The emergence of next-generation sequencing (NGS) technologies has dramatically affected population genetic studies. Originally driven by theoretical insights, population genetics has now become a mostly data-driven discipline (Pool *et al.* 2010). Compared with Sanger sequencing instruments, NGS technologies deliver an amount of data that are several orders of magnitude higher, as well as significantly cheaper. This has made it possible for the sequencing of complete genomes from humans (http://www.1000genomes.org/) and several animal and plant model organisms (e.g. http://www.1001genomes.org/). At the same time, complete genomes for a variety of nonmodel organisms are currently being sequenced, a trend that will only accelerate in the future.

Nevertheless, NGS studies with nonmodel organisms are likely to be significantly different from projects

Correspondence: Bruno Nevado,
E-mail: brunonevado@gmail.com
[1]Present address: Campus UAB—Edifici CRAG, Bellaterra, 08193 Barcelona, Spain

© 2014 John Wiley & Sons Ltd

dealing with model organisms, both in experimental design and in the questions addressed (Helyar *et al.* 2011). The reasons for this are twofold: first, even if the cost of sequencing has decreased dramatically, projects dealing with nonmodel organisms will likely have significantly stronger budgetary constraints than those dealing with model organisms, resulting in data sets with less individuals and/or less coverage and depth; and second, the amount of information available *a priori* and the variety of additional resources that can be used, for example annotation or variation data, for nonmodel species is not comparable to the one already available for the latter, thus conditioning the range of biological questions that can be addressed. This also means that researchers dealing with NGS data from nonmodel organisms will be faced with specific constraints, which are not necessarily the same, nor will have the same solutions, as in studies of model organisms.

One of the main challenges in applying NGS techniques to nonmodel organisms stems from the absence of a species-specific reference genome (Ekblom & Galindo 2010). While resequencing projects can produce an enormous amount of NGS data from any organism for a relatively low price, analysis of these data presents particularly difficult challenges in the absence of a (high quality) reference genome to which to align the short reads.

An approach to overcome the lack of a species-specific reference genome in NGS studies is to use restriction-site-associated DNA sequencing (RAD-seq, Baird *et al.* 2008; Rowe *et al.* 2011). With RAD-seq, a small proportion of a species' genome can be sequenced at relatively high depth (>20×) from several tens of individuals for a fraction of the price involved in whole-genome sequencing, and the data generated can be analysed without alignment to a reference genome (Davey *et al.* 2011). The technique can be used both within and across species and has been applied in a number of studies with nonmodel organisms (e.g., Begun *et al.* 2010; Hohenlohe *et al.* 2011; Bruneaux *et al.* 2013; Wagner *et al.* 2013). However, the approach has drawbacks, and while phylogenetic inferences using thousands of independent SNPs recovered from RAD-seq seems a promising approach (Rubin *et al.* 2012; Cariou *et al.* 2013; McCormack *et al.* 2013), population genetic inferences based on this kind of data are more problematic and can be biased by polymorphisms within restriction sites (Arnold *et al.* 2013; Gautier *et al.* 2013b). Furthermore, analysis of RAD-seq data without using a reference genome is hampered by the confounding effect of repetitive regions (Hohenlohe *et al.* 2012) and by varying levels of variability and recombinations across the genome. Other authors (Gayral *et al.* 2013) have proposed instead to carry out *de novo* transcriptome assembly. However, the difficulty of obtaining RNA from many wild-sampled individuals, the particularity that variability in coding regions may be highly constrained, and the potential unbalanced expression of alleles, means this approach might be ill-suited for population genetic studies.

Another promising approach for NGS studies with nonmodel organisms relies on pooled sequencing. In this approach, DNA samples from many individuals are sequenced together, significantly decreasing sequencing costs while allowing analysis of a large number of individuals. This approach can be used to obtain accurate estimates of population allele frequencies as well as to detect a large number of SNPs from otherwise uncharacterized populations (Futschik & Schlotterer 2010; Gautier *et al.* 2013a). However, data collected in this way does not capture information regarding linkage disequilibrium, thus precluding a number of powerful tests to detect selection and demographic changes (Nielsen *et al.* 2005; Tang *et al.* 2007; Li & Durbin 2011), issues which are of central interest to many population genetic studies.

It thus seems timely to reanalyse the performance of individual resequencing studies in population genomics of nonmodel organisms, taking into account both alternative experimental designs and different bioinformatics approaches. A number of studies have addressed optimal experimental designs for resequencing projects, however primarily from the perspective of genomewide association studies aiming to find genetic markers associated with complex human diseases (e.g. Spencer *et al.* 2009; Sampson *et al.* 2011; Shi & Rao 2011; Goldstein *et al.* 2013). These studies have generally found that the optimal sampling design should include many individuals sequenced at low depth. However, dealing with the high error rate in SNP calling associated with low read depth is not trivial, and solutions to this problem have so far included using prior information on variable sites (Li *et al.* 2009b) or using low-depth data from many (usually hundreds) of individuals and genotype imputation (Le & Durbin 2010; Li *et al.* 2011; Pasaniuc *et al.* 2012). These solutions will typically not be available to researchers dealing with NGS data from nonmodel organisms, where *a priori* genomic resources will be scarce or inexistent. Crawford and Lazzaro (2012) reported on the biases introduced in population genetic inferences when using low-depth NGS data from organisms without *a priori* genomic resources. However, the authors did not consider the effect of genetic divergence to the reference genome, of alternative bioinformatics approaches or of the trade-off between number of individuals sampled and depth of data collected for each individual.

© 2014 John Wiley & Sons Ltd

In this work, we assess the suitability of alternative experimental designs and bioinformatics approaches in resequencing studies aiming to infer population genetic statistics. We assume the absence of a species-specific reference genome but the existence of a reference genome from a related species, and study the effect of increasing genetic divergence to the available reference genome. We consider alternative experimental designs (number of individuals sampled and read depth per individual) and alternative bioinformatics pipelines. We focus on estimating population and nucleotide variability, as well as neutrality tests based on the site-frequency-spectrum (SFS). We first characterize the biases introduced in population genetic estimates by the alternative experimental designs and bioinformatics approaches at different levels of genomic divergence, proceed to identify optimal strategies to infer these statistics, and contrast the results obtained from these simulations to those obtained in the analysis of an empirical data set from the western lowland gorilla.

Material and methods

We consider a single population from which we are interested in estimating population genome wide statistics using NGS data. To assess the effect of the reference genome divergence on the accuracy of population genetic estimates, we investigate four scenarios that differ in the genetic divergence between the reference genome and the population under study, ranging from 0 to about 2% sequence divergence. For each scenario considered, we sample a varying number of diploid individuals (2–20 individuals) to be sequenced at a varying average expected depth (2–20× per diploid individual), while keeping the total sequencing cost fixed at 40×. Table 1 details the different variables and the ranges considered and Appendix S1 (Supporting information) the relevant bioinformatics pipelines used in the analysis.

Obtaining the original SNP data

To simulate genetic data, we used the software SFS_CODE, a forward population genetic simulator allowing for selection, recombination and an arbitrary number of populations and migration patterns (Hernandez 2008). We set the per site scaled mutation and recombination rates to 0.001 and simulated a genomic region 100 kb long under the neutral model. Two populations, each consisting of 1000 individuals, were derived from a single ancestral population and allowed to evolve in isolation from each other (i.e. without migration). Population size was kept constant for each population. At different time points throughout the simulation, the two populations were sampled: a single individual from the first population was used as the reference genome in subsequent analyses, and a varying number of individuals (Table 1) from the second population was used as the population of interest, that is, the one from which we are interested in obtaining population genetic estimates. These settings allowed us to simulate different amounts of genetic divergence between the reference genome and the population of interest: *c.* 0.15%, 1% and 2% sequence divergence. To investigate the case where a reference genome is available from the population of interest, we simply sampled one additional haplotype in each population to be used as the reference genome in the 'no divergence' scenario. We obtained 1000 replicates from 10 independent runs with default burn-in values between sampling times in SFS_CODE: initial burn-in period of 10 000 generations, followed by shorter burn-in periods of 4000 generations between replicates. This approach ensures that each of the 1000 replicates is started from a random draw of a population at equilibrium, while significantly shortening the computational time (SFS_CODE manual, page 7). The data obtained were converted into FASTA format using custom software (Appendix S1, Supporting information) and are hereafter referred to as the presequencing data sets.

Table 1 Main variables considered in this study

Variable	Range	Details
Divergence to reference genome	*No divergence* 0.15% 1% 2%	Average% of fixed differences between *population of interest* and the reference genome. In the *no divergence* scenario, a single population was sampled for both *population of interest* and reference genome
Experimental design	20 individuals, 2× read depth 10 individuals, 4× read depth 5 individuals, 8× read depth 2 individuals, 20× read depth	Total sequencing cost kept constant (40×). Exp design with two individuals used only for variability estimates
Bioinformatics approach	*iSNPcall, mSNPcall* *GHcall* *Without Genotype calls*	*iSNPcall* and *mSNPcall* with SAMTOOLS and GATK *GHcall* with custom software *Without Genotype calls* with ANGSD

© 2014 John Wiley & Sons Ltd

Obtaining and analysing the NGS short reads

Next-generation sequencing reads were simulated, for each individual of each replicate, with the software ART v. 1.5.0 (Huang *et al.* 2012) using the presequencing data sets as input. We used the built-in profile for Illumina sequencing runs, with 75-bp reads, paired-ends (average fragment length of 500 bp, standard deviation of 10), and varying average expected read depth per individual (Table 1). ART simulates sequencing reads by mimicking the real sequencing process with empirical error models and quality profiles parameterized empirically in large sequencing data sets. Substitution errors are simulated according to the empirical, position-dependent distribution of base quality scores, and two different profiles are used for forward and reverse reads, both empirically determined (Huang *et al.* 2012).

The software BWA v 0.6.2 (Li & Durbin 2009) was used to align the resulting reads from each individual to the reference genome and to convert the aligned reads into bam format. Preliminary analysis revealed that using between 4 and 8 mismatches between each read and the reference genome resulted in highest power while retaining low false discovery rate (FDR) (data not shown), with the exact value depending on the genetic distance between reference genome and individuals analysed: with higher genetic distance, allowing for more mismatches produced better results (higher power, comparable FDR). We thus allowed 8 mismatches per read in the highest divergence scenario considered, and 6 mismatches in all other cases. We implemented a mapping quality threshold of 20 and removed duplicated reads from bam files using SAMTOOLS v 0.1.18 (Li *et al.* 2009a) *rmdup* command. The software GATK v 2.7.2 (DePristo *et al.* 2011) was used to realign reads near indels. Note that using a simple genome in our simulations (i.e. without any repetitions or duplication events) means the alignment of short reads to the reference genome proceeds for the most part without errors, which might not be the case in empirical studies.

Estimating population genetic statistics from NGS data

A number of approaches and software packages are available to perform SNP and genotype calling from the aligned bam files (Nielsen *et al.* 2011). In this work, we considered three approaches: genotype calling (specifically with SAMTOOLS and GATK), genotype–haplotype calling (Roesti *et al.* 2012) and inference of population genetic statistics directly from genotype likelihoods without calling genotypes (ANGSD v 0.553, available from http://popgen.dk/angsd/; Korneliussen *et al.* 2013). Details on these approaches are given in the next sections.

The first two approaches returned FASTA-formatted files that were used in MSTATSPOP (available from http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html) to obtain estimates of population variability (Theta), nucleotide diversity (Pi) and neutrality tests: Tajima's *D* (Tajima 1989), Fu and Li's *D* (Fu & Li 1993) and Fay and Wu's *H* (Fay & Wu 2000). This software implements the algorithms described in Ferretti *et al.* (2012) and provides unbiased estimates even in the presence of high levels of missing data between individuals. For the latter two neutrality tests, we used the reference genome as outgroup. For the third approach, population variability and neutrality tests' estimates are reported directly by ANGSD.

We obtained estimates using both presequencing and post-sequencing data sets, the former representing the situation if the true sequences were known, and the latter mimicking the data sets that one would obtain when performing resequencing studies. We then compared the different experimental designs and SNP calling algorithms in terms of bias introduced in relation to estimates obtained with the presequencing data sets. For neutrality tests, we consider only experimental designs involving 5, 10 and 20 individuals, as SFS tests with only two diploid individuals will normally be of little practical interest.

Population genetic statistics from genotype calls

The first method considered performs genotype calls using the standard packages SAMTOOLS and GATK. The approach used by these packages is known to work well for low-depth resequencing studies when a species-specific reference genome is available. For each package, we consider the effect of performing individual (iSNPcall) and multiple-sample (mSNPcall) SNP calling. In the former case, each individual's data is analysed separately, while in the latter, SNP calling is performing jointly for all individuals analysed.

For the analysis using SAMTOOLS, we employed the *mpileup* command, with default options except for the minimum base quality threshold (we used a threshold of 20), and disabling the probabilistic realignment for the computation of base alignment quality (−B). Preliminary analyses showed that disabling this option resulted in higher power and same FDR. Variant calls were filtered with the VCFUTILS.PL *varFilter* command (part of SAMTOOLS package) with default values except for depth (we used minimum depth of half and maximum depth of double the expected read depth per individual).

© 2014 John Wiley & Sons Ltd

For analysis with GATK, we used the *UnifiedGenotyper* command to obtain a list of potential variants, which were subsequently filtered with the *VariantFiltration* tool using the recommended filtering parameters by GATK best practice guide (www.broadinstitute.org/gatk/; see also Appendix S1, Supporting information). We note that GATK implements a variant quality score recalibration (VQSR) tool, which is the recommended approach to filter genotype calls obtained with the *UnifiedGenotyper*. In this work, we used a hard-filtering approach because the VQSR tool requires a set of high-quality known variants, which are typically absent from nonmodel organisms.

An additional issue with these two SNP calling approaches is that, typically, only the set of confident SNPs are reported. Genetic variability values need to be expressed on a per site basis, so that they can be compared across genomic regions as well as between populations. Thus, it is crucial to distinguish between sites that are confidently identified as invariant, from those where not enough information is available to perform a variant call. The latter class of sites should be coded as missing data, whereas the former can be confidently identified as homozygous for the reference allele.

For analysis with SAMTOOLS, for each individual, we used the command *depth* to obtain a list of sites passing the depth and mapping quality thresholds and intersected this list with all genotyped homozygous-reference calls (i.e. sites that bcftools reports with a Phred-scaled genotype likelihood of zero for homozygous reference, Appendix S1, Supporting information). To this list, we then added all positions that contained confident variants, that is, that passed the VCFUTILS.PL *varFilter* step, separately for iSNPcall and mSNPcall. With GATK, we used the option EMIT_ALL_CONFIDENT_SITES in the *UnifiedGenotyper* to obtain an initial list of confident sites for each individual. For homozygous-reference calls, we further filtered by depth (half to double expected read depth in each case), while for variant calls we used the filters described above. In both cases, the sites that contained potential variants, but were not included in the set of confident variant calls, that is, were present in the raw but not in the filtered vcf files, were not included in the list of confident sites.

For each of these SNP calling methods, the set of confident variants obtained was used, together with the reference sequence and the list of confident sites, to reconstruct FASTA format files with the aligned sequences for each replicate, which were then used to obtain estimates of population genetic statistics for the population of interest. In these data sets, ambiguous variant calls – those where none, or more than one genotype, had Phred-scaled likelihood of zero – and sites not present in the list of confident sites were coded as missing, remaining sites were coded according to genotype call in vcf file, or homozygous for reference allele when present in the list of confident sites but not in the vcf file. SAMTOOLS does not properly handle multiallelic SNPs (http://samtools.sourceforge.net/mpileup.shtml); thus, for analyses including this package, we only considered genotype calls including the reference and/or the first alternative allele – genotype calls including any other alternative allele were coded missing.

Population genetic statistics from genotype–haplotype calls

As an alternative approach to standard genotype calling methods, we used a genotype–haplotype calling approach (hereafter denoted GHcall) based on the work of Lynch (2009) and Roesti *et al.* (2012). Briefly, at each site and for each individual, the likelihood of each possible genotype is calculated using a multinomial approach, assuming a fixed raw sequencing error (Lynch 2009). The likelihood of the two most likely genotypes is then compared using a likelihood ratio test (LRT), and, if significant (compared with a chi-square distribution with 1 d.f. at the 0.05 threshold), the most likely genotype is called. If the LRT is not significant, or if the number of reads covering the site is below 6, a haplotype call is attempted: likelihoods of the two most likely homozygous genotypes are compared using an LRT, and, if significant, a haplotype call is performed, consisting of one known (A, C, G or T) and one unknown (N) base. For this method, we only considered bases with a base quality above 20 and set the raw sequencing error to 0.01. Sites were also filtered based on depth, with thresholds of half and double the expected depth per individual. The interest of this method is that it does not depend on which reference allele is present; and it allows for the possibility of calling one single allele, which can still be used for estimates of variability without introducing bias by assuming fully known genotypes at low read depth. Given the low read depth values analysed in this work, we used a threshold of 6 reads for genotype calls, but different thresholds (5, 6 or 7 reads) returned equivalent results (data not shown). Software implementing this method, and used to reconstruct FASTA-formatted files for each individual, is available from www.bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html.

Population genetic statistics without genotype calls

The third method used in this study does not perform genotype calls, but rather bases inferences on genotype likelihoods, thus directly dealing with missing data and

© 2014 John Wiley & Sons Ltd

the sequencing and mapping errors inherent in NGS data (Nielsen *et al.* 2012). We used the empirical Bayes approach detailed in Korneliussen *et al.* (2013) and implemented in the software ANGSD v 0.553, although other similar approaches are available (Buerkle & Gompert 2013). Briefly, the method used involves estimating and optimizing the SFS over a large genomic region, and using this estimate as an empirical prior to jointly estimate the SFS over smaller regions (windows). In this work, for each case, we first calculated an SFS prior using the first 500 replicates totalling 50 million base pairs – the default length used by ANGSD to calculate the SFS prior – and then used this prior to estimate population genetic statistics for each replicate separately. This pooling of replicates to obtain a large region to estimate the SFS prior is adequate, because all replicates used were obtained with the same evolutionary scenario. We used minimum base and mapping quality thresholds of 20, and the genotype likelihood calculation method of GATK. Variability values were divided by the effective number of sites reported for each replicate, and neutrality tests reported directly following the output of ANGSD.

Real data example

To validate the simulation results, we also performed an analysis of an empirical data set of the western lowland gorilla *Gorilla gorilla gorilla* (Prado-Martinez *et al.* 2013). We downloaded the unaligned Fastq files for 20 individuals from the short read archive database (www.ncbi.nlm.nih.gov/sra, accession numbers SRR747935-SRR747946, SRR747967-SRR747977, SRR747984-SRR747994, SRR748076-SRR748080, SRR748089-SRR748108, SRR748113-SRR748116, SRR748165-SRR748176). To analyse the effect of divergence to the reference genome, we analysed the data using either the gorilla genome (genome assembly gorGor3.1), or the human genome (genome assembly GRCh37), both obtained from www.ensembl.org.

Fastq files for each individual were aligned to each of the genomes using the options detailed in Prado-Martinez *et al.* (2013): dynamic quality trimming (-q 15), default read alignment identity thresholds when using the gorilla genome (-n 0.04) and increased edit distance parameter when using the human genome (-n 0.01). We note that the edit mapping used here is more conservative than the settings used in the simulations, which was chosen given the high complexity of the real genomes compared with the simple genomes used in the simulations. Aligned short reads were converted to bam format using the *sampe* option of BWA, and duplicated reads removed using SAMTOOLS *rmdup* command. We then merged all bam files for each individual into a single bam file using PICARD TOOLS' *MergeSamFiles* command (v 1.104, available from http://picard.sourceforge.net).

For computational constraints, we focused subsequent analysis on chromosome 12, for which all 20 individuals exhibited average read depth above 10, and similar number of reads mapped to both human and gorilla genomes (data not shown). Aligned bam files were subsampled using PICARD TOOLS' *DownsampleSam*, to obtain similar data sets to the ones used in the simulations: 20 individuals at *c.* 2×, 10 individuals at *c.* 4×, 5 individuals to 8× and 2 individuals at *c.* 20×. In each case, down-sampled bam files were realigned around putative indels using GATK, and population genomic analysis performed with the same methods described above with the same quality and depth filters. Population genetic statistics were calculated over windows of 100 kb for the entire chromosome (*c.* 130 Mb). For analysis with ANGSD, SFS prior was obtained over the entire chromosome and used to obtain estimates for each 100-kb window.

Results

Estimating population variability

The main challenge in estimating population variability from low-depth NGS data is to distinguish true variants from NGS errors. In this study, we used empirically derived read error models and base quality profiles (implemented in ART) and applied a value of population variability in the simulated data sets (0.001) similar to those observed in a variety of animals and plant species.

The distribution of read depths obtained with the different experimental designs, before and after base and mapping quality filtering, is shown in Fig. S1 (Supporting information). Actual read depth was reduced by about 20–25% after all quality control filters and was similar across divergence scenarios. The number of segregating sites identified and the FDR (percentage of wrongly identified segregating sites), at each divergence level and experimental design, is shown in Table 2, for both genotype and genotype–haplotype calling approaches.

We found a significant difference in the sensitivity with which heterozygous and homozygous SNPs are detected with genotype calling methods (Fig. S2, Supporting information). We calculated sensitivity as the percentage of genotypes in presequencing data sets that are correctly identified in post-sequencing data sets. We report it as average across individuals, and separately for heterozygous and homozygous SNPs – heterozygous SNPs being those where an individual carries one reference and one alternative allele (genotype RA) and homozygous SNPs those where the individual is homozygous for the alternative allele (genotype AA). At the lowest

© 2014 John Wiley & Sons Ltd

Table 2 Number of segregating sites identified and false discovery rate (FDR) (percentage of wrongly identified segregating sites) of different bioinformatics approaches, for different experimental designs and divergence levels. Shown is also the original number of segregating sites in the presequencing data sets

	No divergence				0.15% Divergence			
	20I2X	10I4X	5I8X	2I20X	20I2X	10I4X	5I8X	2I20X
Original	425	355	283	184	423	353	281	182
iSAM (mean)	212	249	245	176	211	248	245	174
iSAM (FDR)	0.86	1.53	1.56	0.14	1.49	2.2	2.31	0.25
mSAM (mean)	260	259	230	157	260	258	230	154
mSAM (FDR)	0.9	0.75	0.51	0.02	1.78	1.6	1.29	0.04
iGATK (mean)	202	214	218	147	201	213	218	146
iGATK (FDR)	0.25	0.18	0.16	0.05	1.15	1.3	1.5	0.22
mGATK (mean)	241	239	231	159	242	240	232	158
mGATK (FDR)	0.12	0.13	0.16	0.04	1.64	1.69	1.53	0.22
GHcall (mean)	206	231	213	153	204	228	211	151
GHcall (FDR)	0.85	1.15	1.73	0.01	0.84	1.18	1.71	0.01
	1% Divergence				2% Divergence			
	20I2X	10I4X	5I8X	2I20X	20I2X	10I4X	5I8X	2I20X
Original	426	356	284	184	425	355	283	184
iSAM (mean)	218	257	258	175	234	276	278	176
iSAM (FDR)	6.99	7.64	8.75	1.83	12.25	13.05	14.54	3.75
mSAM (mean)	277	274	240	148	303	298	258	148
mSAM (FDR)	9.19	8.7	7.75	0.44	15.66	15.08	13.61	1.16
iGATK (mean)	202	219	234	155	188	200	223	155
iGATK (FDR)	8.43	9.14	10.98	2.51	15.71	14.72	18.07	7.38
mGATK (mean)	271	269	257	168	281	279	266	171
mGATK (FDR)	12.58	12.57	11.57	2.5	20.09	20.13	19.21	7.38
GHcall (mean)	196	222	202	144	177	205	176	123
GHcall (FDR)	0.84	1.12	1.73	0.06	0.91	1.19	1.8	0.2

read depth analysed (2×), about 80% of homozygous AA genotypes were correctly identified by multiple-sample SNP calling approaches (mSNPcall), but only about 20% of heterozygous genotypes. Individual SNP calling (iSNPcall) at this low level of depth correctly recovered <5% of heterozygous genotypes, although it also resulted in less homozygous-alternative calls compared with mSNPcall. Individual and multiple-sample SNP calling approaches with SAMTOOLS and GATK returned comparable results, particularly at low to moderate read depths. The sensitivity values we obtained in Fig. S2 (Supporting information) are in line with those reported by Meynert *et al.* (2013) who in addition used GATK's base recalibration tool – a step expected to increase SNP calling accuracy, but not usually available for nonmodel organisms. Depending on genotype calling approach, up to 40% heterozygous genotypes are wrongly identified as homozygous for either the reference or the alternative alleles at low read depth (2–4×, Fig. S2, Supporting information). These errors do not increase the number of segregating sites identified (Table 2), as the positions are variable across the population, but they have a strong impact on the SFS (discussed in next section). For heterozygous genotypes, the GHcall method almost exclusively resulted in haplotype calls at 2–4×, with a similar number of haplotype calls carrying the reference and the alternative alleles (RN and AN, Fig. S2, Supporting information). With read depth of 8×, close to 40% of heterozygous genotypes were correctly identified by this method, and a further 30% called either RN or AN. Finally, as also found by Meynert *et al.* (2013), we detected a strong bias in genotype calls obtained with mSNPcall towards the

© 2014 John Wiley & Sons Ltd

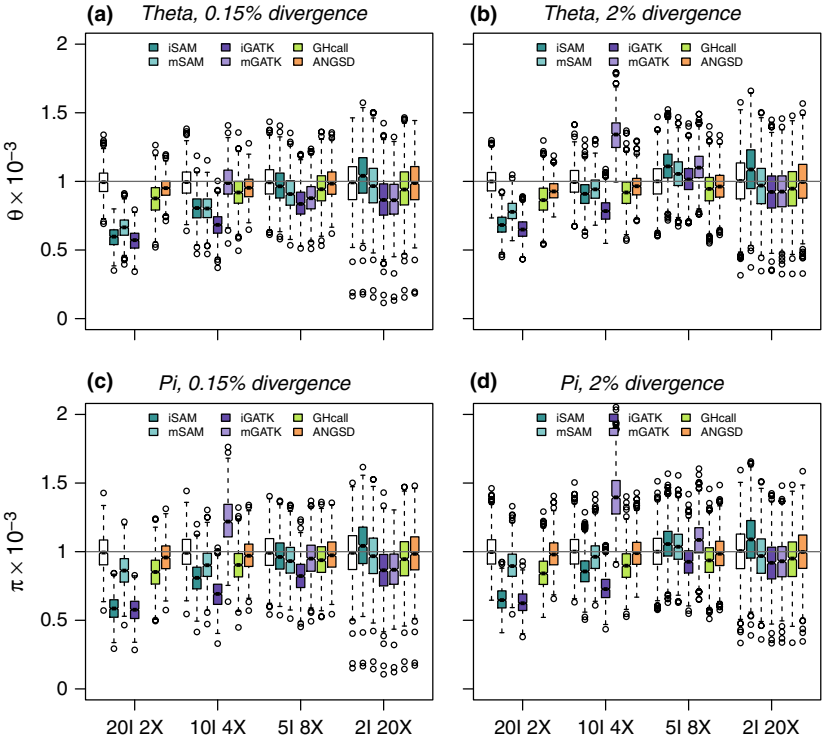

Fig. 1 Distributions of variability estimates (θ a,b; π c,d) obtained across 1000 replicates using the different experimental designs (columns within each panel) and bioinformatics approaches considered in this study (marked with different colours). Shown are the results for two levels of divergence between population of interest and reference genome (0.15% on the left; 2% on the right) obtained under the neutral scenario. Results for other levels of divergence are shown in Fig. S3 (Supporting information). The grey lines denote the 'real' value of variability used in the simulations, and the white boxplots the distributions obtained with the presequencing data sets (i.e. as if the sequence data were known without error nor missing data). For each distribution, the box represents the interquartile distance, and the notches the 95% confidence interval. Note that estimates obtained with mSNPcall in GATK at low read depth (2×) are not shown as they returned highly inflated values.

reference allele, which particularly affects nonreference singletons (data not shown).

As shown in Figs 1 and S3 (Supporting information), genetic divergence between population of interest and reference genome had a marked impact on estimates of population variability, as did the bioinformatics approach and the experimental design used. Read depths of 2–4× when a reference genome is available from the same or a closely related population resulted in underestimates of variability with all SNP calling methods except mSNPcall with GATK (Figs 1a,c and S3a,c, Supporting information). Estimates obtained without genotype calls returned less biased results, particularly for nucleotide variability (Fig. 1 and S3, Supporting information). Increased depth invariably resulted in less biased results, with an average read depth of 20× usually returning distributions whose 95% confidence interval included, or was very close to their true value with the exception of GATK (Fig. 1). In general, correlation values between values obtained with pre- and postsequencing data sets were high ($R > 0.8$) for all experimental designs, divergence scenarios and bioinformatics approaches, with the exception of mSNPcall with GATK at low read depth (Table S1, Supporting information).

The distinct behaviour of variability estimates using GATK (Fig. 1) stems from the inference of confident sites using the EMIT_ALL_CONFIDENT_SITES option. Almost only variable sites are reported as confident by GATK at very low depth (2×) when using mSNPcall, thus unrealistically inflating estimates of population variability per site. Conversely, at high read depths, this option seems to be biased towards invariable positions: at 2I 20× and 'no divergence', mSNPcall with GATK and SAMTOOLS recover an average of 159 and 157 segregating sites (out of 184, Table 2). However, GATK reported as confident about 99% of the sites, while the approach with SAMTOOLS reported <90% (Table 3), thus accounting for the difference in variability values *per site*.

Estimates from low depth (2–4×) were affected by divergence level irrespective of genotype calling method (Figs 1 and S3, Supporting information), with

© 2014 John Wiley & Sons Ltd

higher divergence resulting in higher estimates of variability due to an increase in FDR (Table 2). For instance, population variability estimates with 2I 20× in the 'no divergence' case were $(5.9 \pm 0.8) \times 10^{-4}$, $(6.6 \pm 0.8) \times 10^{-4}$ and $(5.7 \pm 0.8) \times 10^{-4}$, with iSAM, mSAM and iGATK (respectively). At 2% divergence, the same methods resulted in estimates of $(6.8 \pm 0.8) \times 10^{-4}$, $(7.8 \pm 0.8) \times 10^{-4}$ and $(6.5 \pm 0.7) \times 10^{-4}$. While FDR was typically below 1% at low depth without divergence, it approached 10–20% for all genotype calling methods as divergence increased. Variability estimates become more stable across divergence scenarios for all genotype calling methods with moderate to high read depths (8–20×, Fig. 1).

Analysis of vcf and bam files showed that the increase in FDR is due to wrong genotype calls in homozygous SNP sites, that is, those where the individual is homozygous for the alternative allele. In these sites, sequencing errors that incorporate a reference base result in erroneous genotype calls with both SAMTOOLS and GATK. At moderate or high read depths, these sites are inferred as heterozygous (reference-alternative), whereas at low depth, both heterozygous and homozygous-reference calls occur. At higher levels of divergence, these sites are more common and often fixed across the entire population; as a result, homozygous-reference or heterozygous calls increase the total number of segregating sites in the post-sequencing data sets. At 2% sequence divergence, each individual carries on average 2000 homozygous SNP sites over each 100 kb window. At 2× depth, and assuming a raw sequencing error of 1%, an average of 40 wrong bases are incorporated in homozygous SNP sites. Assuming for simplicity that each base can be incorporated with equal probability, we then expect an average of 13 homozygous SNP sites to carry a reference allele due to sequencing errors – which results in an erroneous genotype call. Compounding the issue, this problem occurs for each individual independently, thus adding to the total number of SNPs identified on each data set. We note that the FDR estimates depend on the amount of variability present in the population: with higher variability, FDR decreases due to an increase in the number of correctly identified SNPs. In this regards, it is relevant that many species, such as invertebrates with large population sizes, exhibit variability values an order of magnitude higher than the one simulated herein. In these cases, we would expect the direction of the bias in population genetic statistics to be the same, but its magnitude smaller, compared with our results.

Importantly, and in contrast to genotype calling methods, the other two methods used in this work (GHcall and ANGSD) were not significantly affected by increasing divergence (Figs 1 and S3, Supporting information). The GHcall approach returned unbiased estimates with depth of 8× and above and underestimates at lower depths. Increasing divergence did not affect FDR and only slightly reduced the number of segregating sites recovered with GHcall (Table 2). However, the method resulted in a significant increase in the proportion of missing data compared with genotype calling methods, particularly at low to moderate read depths, due to the threshold of 6 reads for genotype calling: at 2× the method results in about 80% missing data, and *c*. 60% at 4× (Table 3). Estimates obtained with ANGSD usually returned unbiased results irrespective of read depth, particularly for nucleotide variability. At very low depth (2×), ANGSD tended to slightly underestimate variability values (Fig. 1). However, this method returned the less

Table 3 Proportion of missing data in post-sequencing data sets, for different scenarios and SNP calling algorithms

	No divergence				0.15% Divergence			
	20I2X	10I4X	5I8X	2I20X	20I2X	10I4X	5I8X	2I20X
iSAM	0.273	0.253	0.218	0.133	0.274	0.255	0.219	0.135
mSAM	0.272	0.253	0.218	0.133	0.273	0.254	0.219	0.135
iGATK	0.283	0.141	0.068	0.013	0.284	0.141	0.068	0.014
mGATK	0.998	0.384	0.116	0.015	0.997	0.386	0.116	0.015
GHcall	0.81	0.596	0.374	0.133	0.81	0.597	0.376	0.135
	1% Divergence				2% Divergence			
	20I2X	10I4X	5I8X	2I20X	20I2X	10I4X	5I8X	2I20X
iSAM	0.289	0.273	0.242	0.165	0.291	0.27	0.236	0.159
mSAM	0.285	0.273	0.244	0.168	0.283	0.27	0.24	0.163
iGATK	0.295	0.153	0.078	0.019	0.298	0.154	0.077	0.019
mGATK	0.991	0.426	0.124	0.021	0.984	0.415	0.121	0.02
GHcall	0.818	0.61	0.406	0.167	0.817	0.608	0.401	0.162

© 2014 John Wiley & Sons Ltd

biased estimates across divergence and experimental designs considered.

Site-frequency-spectrum based tests

Our second aim was to assess the effect of alternative experimental designs and bioinformatics approaches on well-known neutrality tests. We found that all genotype calling approaches introduced important biases on the distribution of the neutrality tests analysed (Figs 2 and S4, Supporting information). While the magnitude of the bias increased with increasing divergence and decreasing depth, distributions of several neutrality tests became noticeably skewed even at low divergence (0.15%) and moderate depth (8×). In contrast to variability estimates, mSNPcall often performed worse than iSNPcall, particularly at the lowest levels of divergence considered (Fig. 2a,c,e). Both GHcall and ANGSD approaches returned much less biased distributions for all neutrality tests, particularly with moderate read depths (Figs 2 and S4, Supporting information). Correlations between pre- and post-sequencing SFS tests were weaker for neutrality tests than for variability estimates and were more strongly

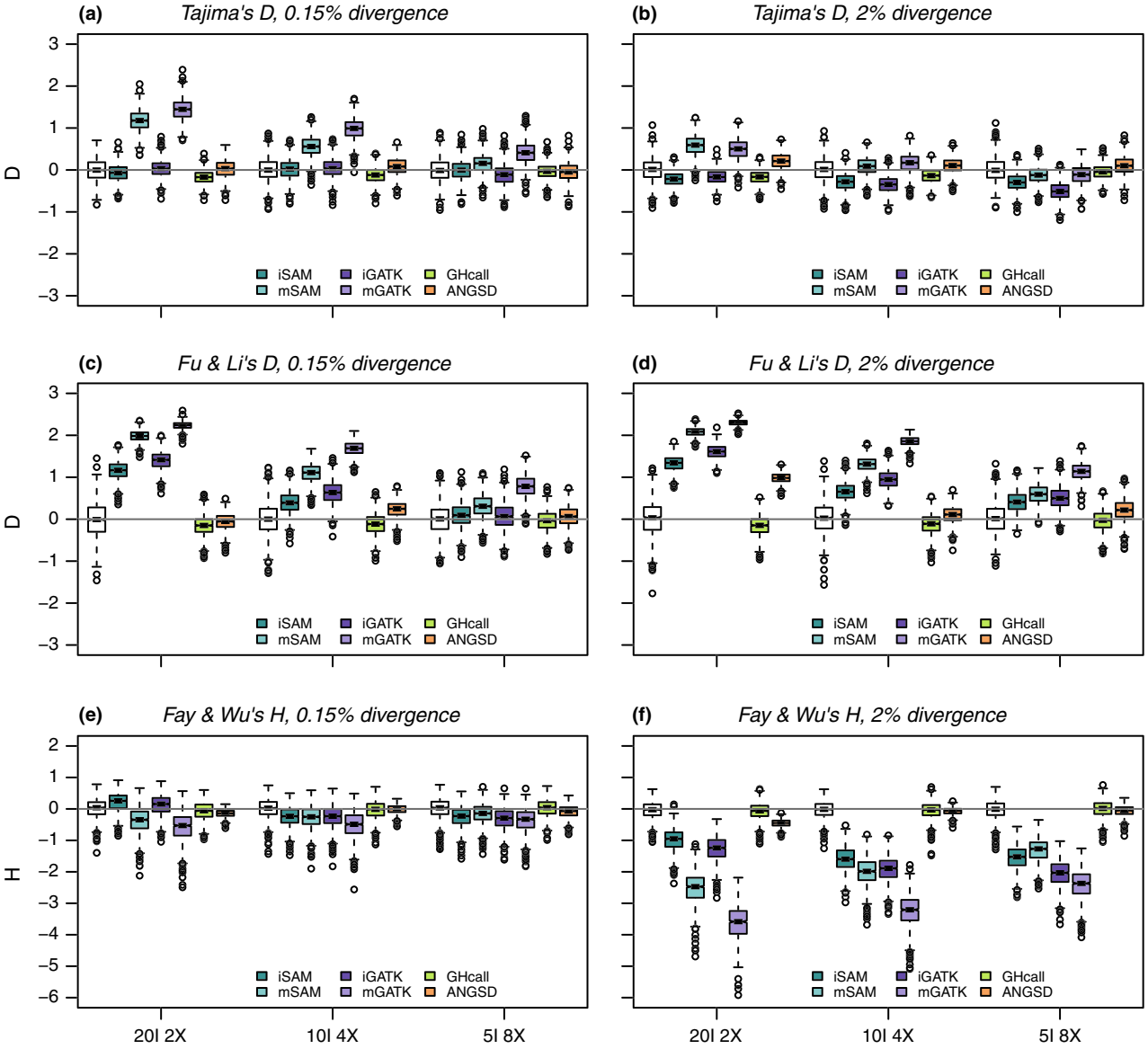

Fig. 2 Distributions of neutrality tests (Tajima's *D* a,b; Fu and Li's *D* c,d; Fay and Wu's *H* e,f) obtained across 1000 replicates under the neutral scenario. Data presented as in Fig. 1. Results for other levels of divergence shown in Fig. S4 (Supporting information). The horizontal dashed lines denote the expectation of each test under neutrality.

© 2014 John Wiley & Sons Ltd

affected by the level of divergence (Table S1, Supporting information). Nevertheless, values were moderate to high for most cases considered with the exception of Fu and Li's *D*.

For Tajima's *D*, estimates obtained with iSNPcall were in general unbiased at all read depths in the 'no divergence' scenario (Fig. S4, Supporting information), but became downward-biased with increasing divergence when depth was low (Fig. 2b). Conversely, mSNPcall returned more biased distributions at lower divergence levels (Figs 2 and S4, Supporting information). Distributions of Fu and Li's *D* were very strongly upward-biased when using any of the genotype calling approaches – more so when using mSNPcall – even at moderate read depths of 8×. Only at low divergence levels, and using the iSNPcall methods, did the distributions approach their true value when using read depth of 8× (Figs 2c and S4c, Supporting information). For Fay and Wu's *H*, all genotype calling methods returned strong downward-biased distributions with increasing divergence, even at moderate depths of 8× (Figs 2f and S4f, Supporting information).

In contrast to the standard genotype calling algorithms considered, GHcall method returned mostly unbiased distributions for all neutrality tests. The distributions did not significantly shift with increasing divergence nor decreasing read depth, although depths of 2–4× returned slightly negative values for Tajima's *D* and Fu and Li's *D* (Fig. 2). The ANGSD method returned distributions of neutrality tests similarly unbiased. However, at low depth and high divergence, some bias was evident, particularly in estimates of Fu and Li's *D* (Fig. 2d,f).

The skew in distributions of the different neutrality tests when using genotype calling approaches can be explained in the light of the results discussed previously. With individual SNP calling, Tajima's *D* becomes downward-biased at high divergence due to the increase in the number of low-frequency variants that results from errors in homozygous SNP sites. Conversely, the bias of mSNPcall algorithms towards medium-frequency variants results in overestimates of Tajima's *D*, even at 8× depth and low divergence (Fig. 2a). Fu and Li's *D* estimates are strongly upward-biased at low read depths, as singletons are particularly difficult to detect at these depths, irrespective of genotype calling algorithm. Multiple-sample SNP calling exacerbates this effect by rejecting even those singletons for which individual data contain enough support (Fig. 2c,d). The negative skew in Fay and Wu's *H* distributions with increased divergence is due to the effect discussed above concerning sequencing errors in homozygous SNPs. These positions are often fixed across the entire population, and a genotyping error including the reference allele will increase the number of inferred high-frequency variants, thus shifting Fay and Wu's *H* estimates downwards.

Real data example

Summary of the results of the analysis of chromosome 12 from 20 gorilla specimens is shown in Figs 3 and 4. Sequence divergence between the human and gorilla reference genomes, calculated across chromosome 12 in 100-kb windows using the sequence alignment available from Ensembl (http://www.ensembl.org), was 1.5–2.5%. The two reference genomes varied with respect to the amount of missing data: almost all windows in the human genome exhibited <1% missing data (excluding the terminal windows and centromere), while in the gorilla genome, most windows contained 5–15% sites coded as missing (N). This difference in amount of missing data in the reference is unlikely to affect our results because (1) sites with unknown reference were not used in the calculation of population genetic statistics, and (2) the variability values calculated were divided by the effective length of the alignment, thus accounting for varying length of each window.

Figure 3 shows the distribution of variability estimates (Theta), obtained across chromosome 12 of gorilla in 100 kb windows, with the four experimental designs considered. Figure 4 shows the estimates obtained for the three neutrality tests at moderate read depth (8× and five individuals) when using either the gorilla or the human genome as reference. In both figures, we omit windows with more than 80% missing data and outliers. Figure S5 (Supporting information) shows the estimates obtained across the chromosome, for the first 20 Mb, with alternative experimental designs and methods, and using either the gorilla or the human genome as reference.

For the empirical data, we do not know the real values for either variability or neutrality tests, so we can only compare the estimates obtained with different methods and with differing read depth and divergence level. Importantly, the real data analysis confirms the simulation results regarding the effect of read depth and divergence on estimates of variability and SFS tests.

Estimates of population variability obtained with genotype calling approaches strongly depended on read depth (Fig. 3), with low depth returning lower values with all methods except mSNPcall with GATK. Variability estimates obtained with GHcall and ANGSD were more stable across experimental designs considered (Fig. 3).

In what concerns neutrality tests, and similarly to the results obtained by simulation, we find that even at moderate read depth (8×), estimates strongly depended

© 2014 John Wiley & Sons Ltd

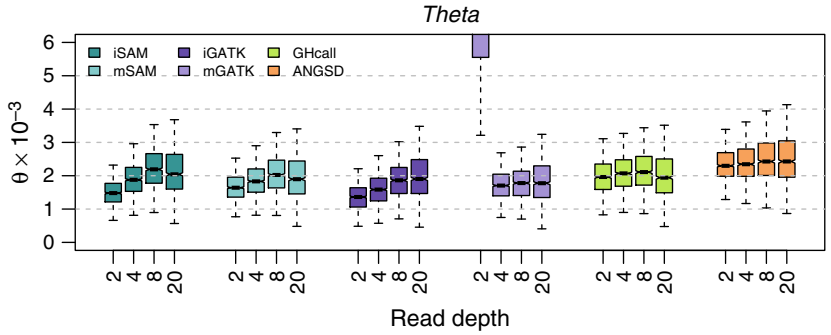

Fig. 3 Effect of read depth on estimates of population variability, in the analysis of chromosome 12 from gorilla. Distributions shown include all 100-kb windows with <80% missing data, after removing outliers. For each bioinformatics approach – coloured according to previous figures – we show the results obtained at different read depths (2× with 20 individuals, 4× with 10, 8× with 5 and 20× with 2 individuals).

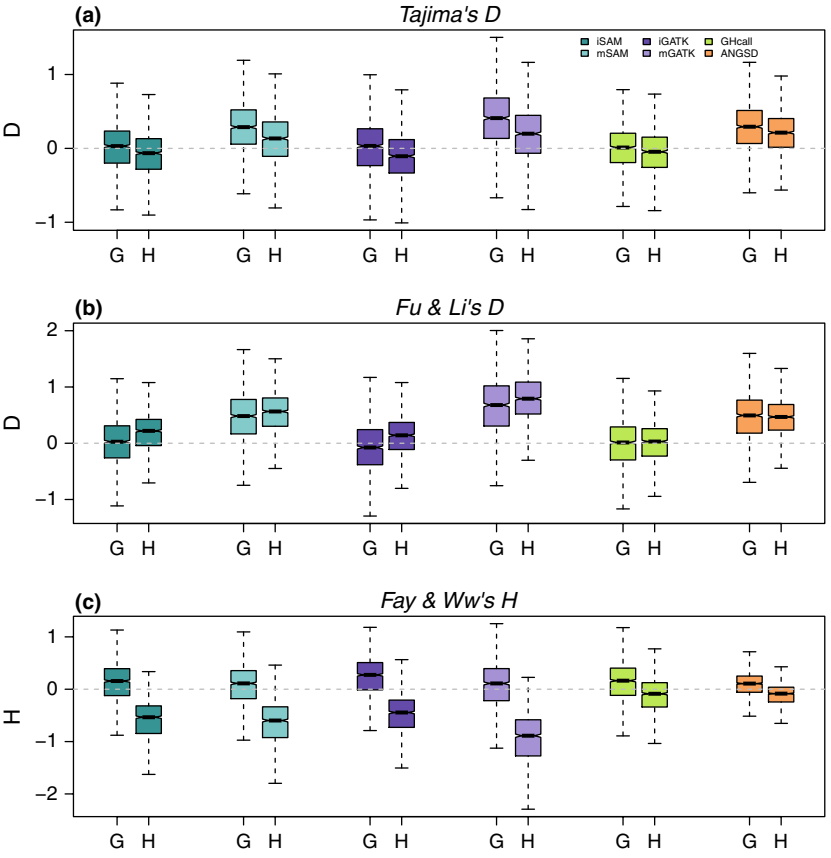

Fig. 4 Effect of genome divergence on site-frequency-spectrum tests (Tajima's *D* a; Fu and Li's *D* b; Fay and Wu's *H* c) in the analysis of chromosome 12 from gorilla. Shown are results obtained with moderate depth (five individuals at 8×) across chromosome 12, when using either the gorilla (G) or the human (H) genomes as reference. Outliers and windows with more than 80% missing data are excluded from this figure. Grey dashed line represents expectation under neutrality.

on both bioinformatics approach and divergence level (Fig. 4). The bias of mSNPcall towards medium-frequency variants is evident in the distributions of Tajima's *D* and Fu and Li's *D*, which are higher compared with iSNPcall. Even at this moderate read depth, all genotype calling approaches returned estimates that varied with increasing divergence: small downward bias in Tajima's *D* and upward in Fu and Li's *D*, and strong downward bias in Fay and Wu's *H* (Fig. 4). The shift in Fu and Li's *D* upwards results from a decrease

© 2014 John Wiley & Sons Ltd

in singletons for the reference allele: with increasing divergence, more singletons in the population carry the alternative allele and are increasingly difficult to detect. The changes in both Tajima's *D* and Fay and Wu's *H* reflect the increase in the number of high-frequency-derived variants with increasing divergence. These observations agree with our simulation results and support our interpretation of reference-biased genotyping errors in homozygous SNPs.

With GHcall and ANGSD methods, some effect of divergence was evident, particularly for Fay and Wu's *H* (Fig. 4c). Further, estimates obtained with the two methods were different, particularly for Tajima's *D* and Fu and Li's *D*. These results reflect that the simulation approach we used does not capture all the problems facing empirical studies. Given the sensitivity of ANGSD to prior misspecification (Korneliussen *et al.* 2013), we hypothesize that our approach of estimating a SFS prior over the entire chromosome might have affected estimates with this software. Regardless, and in agreement with the simulation results, ANGSD and GHcall approaches returned more consistent estimates of variability and SFS tests, across divergence levels and read depths, when compared to genotype calling approaches.

Discussion

Our work quantifies the biases inherent to different experimental designs and bioinformatics approaches, as well as the effect of genome divergence on population genetic estimates. An important consideration is that the optimal experimental design depends on the aim of the study.

For studies aiming to discover variable positions along the genome, an experimental design favouring number of individuals at the cost of read depth can be an optimal strategy, which agrees with previous work (Shi & Rao 2011; Mechanic *et al.* 2012; Pasaniuc *et al.* 2012). Using multiple-sample SNP calling in GATK or SAMTOOLS, we recovered 50–60% more SNPs at 2× and 20 individuals than at 20× and 2 individuals, with similar FDR (Table 2). However, if the reference genome available is distantly related to the population of interest, these methods result in a non-negligible amount of false SNPs. Under these conditions, it is advisable to use SNP calling algorithms that do not use information from the reference sequence, such as the GHcall method used here. The number of SNPs identified by GHcall is in fact comparable to individual SNP calling at low depth, without incurring in a high FDR with increasing divergence (Table 2). However, the proportion of missing data in the resulting data sets is very high, as the method only performs haplotype calls at this level of depth.

Conversely, an approach involving low read depth and many individuals will introduce strong biases in variability and SFS estimates when coupled with genotype calling approaches. Worryingly, both the direction and magnitude of these biases depend critically on read depth, divergence and SNP calling algorithm, making it difficult to account for in empirical studies and to give a practical recommendation. The amount of divergence to the reference genome might be unknown, the depth needs to be decided *a priori*, and both divergence and read depth might significantly vary across the genome. Underestimation of population variability values under these conditions has been reported previously (e.g. Crawford & Lazzaro 2012), as have the biases introduced in neutrality tests under similar circumstances (Korneliussen *et al.* 2013). In addition, here we find that divergence to the reference genome – which is more difficult to characterize than read depth – can have as well a strong impact in estimates of variability and in SFS tests. If these estimates are of central importance, our results suggest two alternative approaches: aiming for low read depth and using methods that do not perform SNP calling, or aiming for higher read depth coupled with an individual SNP calling approach or a genotype–haplotype method. The former approach has the advantage of performing better at low read depth; thus, resources can be allocated towards sampling more individuals: increasing the number of individuals will reduce the variance of the estimates obtained with all methods (see e.g., Fig. 1). Further, this approach can be used to obtain other statistics of interest for population genetics (Fumagalli *et al.* 2013). However, it can be affected by prior misspecification (Korneliussen *et al.* 2013) and returns less accurate estimates for some neutrality tests (Fig. 2d,f). In favour of the latter approach, aiming for moderate depth (8×) and using a GHcall approach returned unbiased estimates of both variability and neutrality tests, while confidently identifying SNPs irrespective of divergence level.

Another consideration when deciding on experimental designs for resequencing studies without a species-specific reference genome is the complexity of real genomes, which we have not modelled in our simulations. Real genomes contain features such as repetitive regions, copy number variants or indels, which can affect both the mapping and the SNP calling steps. Studying the effect of all these features in the SNP calling procedure is beyond the scope of the present study. We focused on the simplest case to highlight that even in the best of circumstances, genetic divergence can affect population genetic estimates. Given that mutation rate of repetitive and copy number variants is much higher than for SNPs, the effect of absence of reference genome can be a more serious one than that reported

© 2014 John Wiley & Sons Ltd

here. It is also likely that these genome features have a larger influence on the mapping step than in the SNP calling procedure itself. Alternative mapping approaches to handle these issues have received extensive attention, for example Zhao *et al.* (2013) and references therein. In general, these features result in lower actual read depth in empirical studies and poor mapping quality, further supporting the idea that studies without species-specific reference genome should aim for higher read depths.

In this work, we focused on the analysis of NGS data from species where a specific reference genome is not available, but our results also bear significance for a wide range of resequencing studies. Particularly, variability and divergence levels, as well as actual read depth resulting from resequencing studies, have been shown to vary across the genome of species in conspicuous ways (e.g. Begun *et al.* 2010; Roesti *et al.* 2012; Beissinger *et al.* 2013). Thus, even within a single species from which a reference genome is available, biased inference can occur in regions with higher divergence or lower depth. Sequence divergence as low as 0.15% may result in marked biases in several tests. Finally, our study suggests to proceed with caution in comparative analysis of genomic data between closely related species, where NGS reads are mapped and SNP calls performed against a common reference genome (e.g. Prado-Martinez *et al.* 2013). In those studies, at least moderate coverage levels are warranted and SNP calling algorithms that do not depend on reference alleles are recommended.

Acknowledgements

Work funded by Consolider CSD2007-00036 'Centre for Research in Agrigenomics' and AGL2010-14822 grants (Spain) to MPE and CGL2009-09346 grant (Spain) to SRO. We thank J. Davey and four anonymous referees for valuable comments on an earlier version of this manuscript.

References

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

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

Begun DJ, Hohenlohe PA, Bassham S *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Beissinger TM, Hirsch CN, Sekhon RS *et al.* (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, **193**, 1073–1081.

Bruneaux M, Johnston SE, Herczeg G *et al.* (2013) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology*, **22**, 565–582.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution*, **3**, 846–852.

Crawford JE, Lazzaro BP (2012) Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in Genetics*, **3**, 66.

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

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Ekblom R, Galindo J (2010) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Ferretti L, Raineri E, Ramos-Onsins S (2012) Neutrality tests for sequences with missing data. *Genetics*, **191**, 1397–1401.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Fumagalli M, Vieira FG, Korneliussen TS *et al.* (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Futschik A, Schlotterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Gautier M, Foucaud J, Gharbi K *et al.* (2013a) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.

Gautier M, Gharbi K, Cezard T *et al.* (2013b) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Gayral P, Melo-Ferreira J, Glemin S *et al.* (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics*, **9**, e1003457.

Goldstein DB, Allen A, Keebler J *et al.* (2013) Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, **14**, 460–470.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

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

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. *Methods in Molecular Biology*, **888**, 235–260.

© 2014 John Wiley & Sons Ltd

Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, **14**, 289.

Le SQ, Durbin R (2010) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, **21**, 952–960.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li H, Handsaker B, Wysoker A *et al.* (2009a) The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li R, Li Y, Fang X *et al.* (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Research*, **19**, 1124–1132.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, **21**, 940–951.

Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, **182**, 295–301.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Mechanic LE, Chen HS, Amos CI *et al.* (2012) Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genetic Epidemiology*, **36**, 22–35.

Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics*, **14**, 195.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, e37558.

Pasaniuc B, Rohland N, McLaren PJ *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, **44**, 631–635.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Prado-Martinez J, Sudmant PH, Kidd JM *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

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

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N (2011) Efficient study design for next generation sequencing. *Genetic Epidemiology*, **35**, 269–277.

Shi G, Rao DC (2011) Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genetic Epidemiology*, **35**, 572–579.

Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, **5**, e1000477.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, **5**, e171.

Wagner CE, Keller I, Wittwer S *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14** (Suppl 11), S1 .

B.N., S.E.R.O. and M.P.E. designed the research and wrote the manuscript. B.N. performed the research.

Data accessibility

Pipelines used in this work are provided in Appendix S1 (Supporting information). Bioinformatics tools to perform the simulations, including transforming data between different formats, and calculating genotyping sensitivity, are available from www.github.com/brunonevado/Pipeliner. Software for genotype–haplotype method available from http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html. DNA sequences used in the real data example available from ncbi: SRR747935-SRR747946, SRR747967-SRR747977, SRR747984-SRR747994, SRR748076-SRR748080, SRR748089-SRR748108, SRR748113-SRR748116, SRR748165-SRR748176.

© 2014 John Wiley & Sons Ltd

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Observed depth per site attained in the different experimental designs, before and after base and mapping quality filtering.

Fig. S2 Genotyping sensitivity obtained with SAMTOOLS, GATK and GHcall approach under the different experimental designs investigated.

Fig. S3 Distribution of genetic variability estimates (Theta, Pi) obtained across replicates for the 'no divergence' and the 1% divergence scenarios.

Fig. S4 Distribution of neutrality tests (Tajima's *D*, Fu and Li's *D*, Fay and Wu's *H*) obtained across replicates for the 'no divergence' and the 1% divergence scenarios.

Fig. S5 Estimates of variability and SFS tests obtained in 100 kb windows across the first 20 Mb of chromosome 12, for all experimental designs and methods considered, and when using either the gorilla or the human genome as reference.

Table S1 Pearson's correlation coefficient calculated between pre- and post-sequencing data sets for all statistics and scenarios considered.

Appendix S1 Bioinformatics pipelines used in this study.

© 2014 John Wiley & Sons Ltd