An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity

Badr Benjelloun, Frédéric Boyer, Ian Streeter, Wahid Zamani, Stefan Engelen, Adriana Alberti, Florian J Alberto, Mohamed Benbati, Mustapha Ibnelbachyr, Mouad Chentouf, et al.

An evaluation of whole genome genotyping strategies to assess neutral and adaptive diversity

Short running title: Genotyping strategies to assess genome diversity

Badr Benjelloun[1,2*], Frédéric Boyer[1], Ian Streeter[3], Wahid Zamani[1,4], Stefan Engelen[5], Adriana Alberti[5], Florian J. Alberto[1], Mohamed BenBati[2], Mustapha Ibnelbachyr[6], Mouad Chentouf[7], Abdelmajid Bechchari[8], Hamid R. Rezaei[9], Saeid Naderi[10], Alessandra Stella[11], Abdelkader Chikhi[6], Laura Clarke[3], James Kijas[12], Paul Flicek[3], Pierre Taberlet[1], François Pompanon[1*]

[1] Univ. Grenoble-Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, F-38000 Grenoble, France

[2] National Institute of Agronomic Research (INRA Maroc), Regional Centre of Agronomic Research, 23000 Beni-Mellal, Morocco

[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD UK

[4] Department of Environmental Sciences, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, 46417-76489 Noor, Mazandaran, Iran

[5] CEA - Institut de biologie François-Jacob, Genoscope, 2 Rue Gaston Cremieux 91057 Evry Cedex, France.

[6] National Institute of Agronomic Research (INRA Maroc), CRRA Errachidia, 52000 Errachidia, Morocco

[7] National Institute of Agronomic Research (INRA Maroc), CRRA Tangier, 90010 Tangier, Morocco

[8] National Institute of Agronomic Research (INRA Maroc), CRRA Oujda, 60000 Oujda, Morocco

[9] Department of Environmental Sci, Gorgan University of Agricultural Sciences & Natural Resources, 41996-13776 Gorgan, Iran

[10] Environmental Sciences Department, Natural Resources Faculty, University of Guilan, 49138-15749 Guilan, Iran

[11] PTP Science Park, Bioinformatics Unit, Via Einstein-Loc. Cascina Codazza, 26900 Lodi, Italy

[12] Commonwealth Scientific and Industrial Research Organisation Animal Food and Health Sciences, St Lucia, QLD 4067, Australia.

* Corresponding authors

Badr Benjelloun, François Pompanon

Emails badr.benjelloun@univ-grenoble-alpes.fr francois.pompanon@univ-grenoble-alpes.fr

Abstract

Whole genome sequences (WGS) greatly increase our ability to precisely infer population genetic parameters, demographic processes, and selection signatures. However WGS can still be not affordable for a representative number of individuals/populations. In this context, our goal was to assess the efficiency of several SNP genotyping strategies by testing their ability to accurately estimate parameters describing neutral diversity and to detect signatures of selection. We analysed 110 WGS at 12X coverage for four different species, i.e. sheep goats and their wild counterparts. From these data we generated 946 datasets corresponding to random panels of 1K to 5M variants, commercial 50K and HD (600K) SNP chips and exome capture, for sample sizes of 5 to 48 individuals. We also simulated low-coverage genome re-sequencing of 1X, 2X and 5X by randomly sub-sampling reads from the 12X re-sequencing data. Globally, 5K to 10K random variants (representing one variant each 500 Kb to 260 Kb) were enough for an accurate estimation of genome diversity. Conversely, commercial panels (i.e., 50K – 600K SNP chips) and exome capture displayed strong ascertainment biases and even sometimes modified the ranking of populations based on diversity estimates. Besides the characterization of the neutral diversity, the detection of the signature of selection and the accurate estimation of linkage disequilibrium required panels of at least 1 M variants. Finally, whole genome re-sequencing coverage of at least 5X appeared to be necessary for accurate assessment of genomic variations. These results have implications for studies seeking to deploy low-density SNP collections or genome scans across genetically diverse populations/species showing similar genetic characteristics and patterns of LD decay for a wide variety of purposes.

Key words

Whole Genome Sequencing, Mammals, depth of coverage, SNP chip, population genomics

Introduction

Demographic and adaptive processes such as migration, genetic bottlenecks and selection are evolutionary forces that have influenced patterns of variation in genomes. Combined with genetic processes such as recombination they result in a non-uniform distribution of genetic variation across the genome. Since the middle of the last century (Wright, 1931; Fisher, 1958), population genetics has been providing theoretical models to infer how these processes have shaped evolution by studying genetic variations among individuals, populations or species. This set up a conceptual framework for inferring the role of these processes from the study of current genetic variations. Thus, a mandatory prerequisite of evolutionary studies has been the design of panels of molecular markers representative of genome variations. This step has always been challenging. Until the last decade, the efficiency of molecular markers was mainly limited by technical issues. Co-dominant markers such as microsatellites give access to allelic frequencies and are informative for inferring demographic processes (e.g. Di Rienzo et al., 1998; Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999) but a maximum of a few dozen markers was usually genotyped. Other markers such as Amplified Fragment Length Polymorphism (AFLP) are more representative of whole genome variations, as a few hundred can be genotyped simultaneously, but they are dominant and do not give access to allelic frequencies. Even co-dominant markers that can be genotyped across the whole genome such as Single Nucleotide Polymorphisms (SNPs) were first revealed in limited numbers, e.g. (Holloway et al., 1999). Due to these limitations, there was a strong risk for misestimating whole genome variations and linkage disequilibrium (Jones et al., 2013), and to detect inaccurately both past demographic and selection signatures. Recent technological developments made possible the typing of very large numbers of co-dominant markers, mostly SNPs, which considerably increased the representativeness of the genomes and lead to the development of population

genomics approaches (Black, Baer, Antolin, & DuTeau, 2001; Goldstein & Weale, 2001; Jorde, Watkins, & Bamshad, 2001; Luikart, England, Tallmon, Jordan, & Taberlet, 2003). So far, such an approach has essentially been limited to model organisms because of the need for whole genome reference data. Today, methods that overcome this challenge have been developed (Davey et al., 2011; Everett, Grau, & Seeb, 2011), and it is possible to manage whole genome data on nearly any studied species. Moreover, references genomes now have been produced for more and more species (Howe et al., 2013; Dong et al., 2013; Bennetzen et al., 2012; Carneiro et al., 2014). Nevertheless, producing reference genomes and setting up population genomic studies by re-sequencing whole genomes of several individuals at sufficient coverage remains both costly and computationally time consuming. Thus, several strategies have been developed in order to reduce these costs while aiming to keep reliable and representative information of whole genome variations. One strategy is to reduce the depth of coverage of the WGS data to obtain information on the whole genome. A few studies promoted the use of low to medium coverage shotgun WGS (Bizon et al., 2014; Dastjerdi, Robert, & Watson, 2014; Jansen et al., 2013; Nina Overgaard Therkildsen). Nevertheless there is a strong risk of losing accuracy in variant calling and individual genotyping. These problems can be overcome by sequencing key individuals or increasing the number sequenced individuals (e.g. Y. Li, Sidore, Kang, Boehnke, & Abecasis, 2011; Pasaniuc et al., 2012; Alex Buerkle & Gompert, 2013; Han, Sinsheimer, & Novembre, 2014), or by imputing genotypes using genotype probabilities when available (Therkildsen & Palumbi). In addition, the reliability of low to medium coverage in WGS to infer individual genotypes has not yet been empirically evaluated. A second strategy to reduce the costs is to avoid whole genome sequencing and genotype a panel of a limited number of variants. For instance, commercial DNA chips or arrays for SNP typing are already available for several species (e.g. human, cattle, sheep, chicken) and can be designed for the purpose of any

species. The Restriction-site Associated DNA sequencing (RAD-seq) method reduces genome complexity by re-sequencing stretches of genomic DNA adjacent to restriction endonuclease sites (M. R. Miller, Dunham, Amores, Cresko, & Johnson, 2007; Baird et al., 2008). The RNA-seq method gives access to the transcriptome by sequencing the complementary DNA (cDNA) (Wilhelm et al., 2008; Mudge et al., 2011). Genome enrichment methods allow the extraction of targeted regions of the genome, and one main application is the exome capture used for sequencing protein-coding regions (Ng et al., 2009; Choi et al., 2009; Teer & Mullikin, 2010; Cosart et al., 2011), which can be used in population genomics studies (Mascher et al., 2013; Campbell et al., 2013). However, when using these approaches we face the key question of their ability to produce accurate genotypes. The panel of genotyped variants should reliably represent genome variations for all studied individuals to avoid the ascertainment bias that results in the misestimating of genetic parameters. Only very few studies evaluated the accuracy of such genotyping approaches, and showed the impact of ascertainment bias on measures of population divergence (Albrechtsen, Nielsen, & Nielsen, 2010). Moreover, until now, no study evaluated the impact of subsampling panels of variants compared to WGS data, when studying genome diversity, population genetic structure and genes under selection.

In this context, our study aimed at assessing the accuracy of different variant subsampling methods for describing whole genome diversity. We produced 110 WGS at 12X coverage for four mammal species: sheep (*Ovis aries*), goat (*Capra hircus*) and their closely related wild species, the Asiatic mouflon (*Ovis orientalis*) and the Bezoar ibex (*Capra aegagrus*). From these WGS data, we extracted panels of genomic variants corresponding to different genome sampling strategies (i.e., exome capture, commercial SNP chips or random panels) in order to evaluate the impact of variants subsampling on the estimation of genome diversity and on the detection of a selection signature. This allowed defining appropriate variant densities for

population genomic studies. We also simulated lower re-sequencing coverage to evaluate the impact of sequencing depth on the assessment of whole genome diversity.

Material and Methods

Sampled individuals

Tissue samples were collected for 48 sheep (*Ovis aries*) and 30 goats (*Capra hircus*) widely spread across the Northern half of Morocco (North of latitude 28°) between January 2008 and March 2012 (Table S1). Tissues from the distal part of the ear were collected and placed in alcohol for one day, before transfer into silica-gel tubes until DNA extraction. Tissues from 15 Asiatic mouflon (*Ovis orientalis*) and 20 Bezoar ibex (*Capra aegagrus*) were collected in Iran, either from captive or recently hunted animals and conserved in silica-gel after one day in alcohol, or from frozen corpses or tissues archived in alcohol by the Iranian local Department of Environment and transferred in silica-gel until extraction.

DNA extraction and re-sequencing

DNA extraction was done at *Parco Tecnologico Padano* (Lodi, Italy) using the Puregene Tissue Kit from Qiagen® following the manufacturer's instructions. Then, 500ng of genomic DNA were sheared to a 150-700 bp range using the Covaris® E210 instrument. Sheared DNA was used for Illumina® library preparation by a semi-automatized protocol. Briefly, end repair, A tailing and Illumina® compatible adaptors (BiooScientific) ligation were performed using the SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter), according to the manufacturer protocol. A 300-600 bp size selection was applied in order to recover most of the fragments. DNA fragments were amplified by 12

cycles PCR using Platinum Pfx Taq Polymerase Kit (Life Technologies®) and Illumina® adapter-specific primers. Libraries were purified with 0.8x AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent Technologies®) and qPCR quantification, the libraries were sequenced using 100 bp length read chemistry in paired-end flow cell on the Illumina® HiSeq2000.

Read mapping, SNP calling and filtering

Illumina paired-end reads of *Ovis* were mapped on the sheep reference genome (OAR v3.1, GenBank assembly GCA_000317765.1, Jiang et al.) and those of *Capra* on the goat reference genome (CHIR v1.0, GenBank assembly GCA_000317765.1, Dong et al., 2013) using BWA mem (H. Li & Durbin, 2009). 99.4% (± 0.1%), 99.3% (± 0.2%), 98.9% (± 0.1%) and 98.8% (± 0.4%) of the reads were mapped on the reference assembly for sheep, mouflon, goats and bezoar, respectively. The BAM files produced were then sorted using Picard SortSam and improved using Picard MarkDuplicates (http://picard.sourceforge.net), GATK RealignerTargetCreator, GATK IndelRealigner (DePristo et al., 2011) and Samtools calmd (H. Li et al., 2009).

Variant sites were initially called using three different algorithms: Samtools mpileup (H. Li et al., 2009), GATK UnifiedGenotyper (McKenna et al., 2010) and Freebayes (Garrison & Marth, 2012). Variants were called for each group independently: sheep, mouflon, goat, and Bezoar ibex. Note that a larger dataset than that used in this study was used for variant discovery in domestic groups (160 sheep and 161 goats from Morocco; for European Nucleotide Archive ID, see Table S2). Then we ran two successive rounds of filtering variant sites. Filtering stage 1 merged together calls from the three algorithms, whilst filtering out the lowest-confidence calls. A variant site passed if it was called by at least two different calling

algorithms with variant phred-scaled quality > 30. An alternate allele at a site passed if it was called by any one of the calling algorithms, and the genotype count > 0. Filtering stage 2 used Variant Quality Score Recalibration by GATK. First, we generated a training set of the highest-confidence variant sites where (i) the site is called by all three variant callers with variant phred-scaled quality > 100; (ii) the site is biallelic; (iii) the minor allele count is at least 3, counting only samples with genotype phred-scaled quality > 30. The training set was used to build a Gaussian model using the tool GATK VariantRecalibrator using the following variant annotations from UnifiedGenotyper: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, DP, Inbreeding Coefficient. The Gaussian model was applied to the full data set, generating a VQSLOD (log odds ratio of being a true variant). Sites were filtered out if VQSLOD < cutoff value. The cutoff value was set for each population by the following: Minimum VQSLOD = {the median value of VQSLOD for training set variants} - 3 * {the median absolute deviation VQSLOD of training set variants}. Measures of the transition / transversion ratio of SNPs suggest that this chosen cut-off criterion gives the best balance between selectivity and sensitivity.

Genotypes were improved and phased by Beagle 4 (Browning & Browning, 2013), and then filtered out where the genotype probability calculated by Beagle is less than 0.95. The genotype call sets generated at this stage constituted the WGS datasets used for within-population analyses. For cross-populations comparisons and validation of the identified WGS surrogates that were performed in each genus (i.e. *Capra* and *Ovis*), we generated a set of filtered variant sites per genus by merging the positions of filtered bi-allelic SNPs called in the different groups. For each sample, genotypes were called at each SNP position using GATK UnifiedGenotyper using the option GENOTYPE_GIVEN_ALLELES. Genotypes were improved and phased by Beagle 4 (Browning & Browning, 2013), and then filtered out where the genotype probability calculated by Beagle is less than 0.95.

Quality control of WGS data

To further assess the quality of the filtered WGS datasets, a subset of the sequenced individuals were genotyped using commercial SNP Chips by *Laboratorio Genetica e Servizi* (Cremona, Italy). 29 sheep and 8 Asiatic mouflon were genotyped with the Illumina® ovine 50K SNPs BeadChip, and 27 goats and 8 Bezoar ibex with the Illumina® caprine 50K SNPs BeadChip. In order to establish the concordance between WGS and chip data the coordinates of the SNPs on the chips were obtained by mapping the probes used for chip design onto the corresponding reference genome (OAR v3.1 or CHIR v1.0) using BWA aln and sample (H. Li & Durbin, 2009). The raw data in Plink format (Purcell et al., 2007) were updated for SNP coordinates and were filtered for each group by applying the following inclusion criteria: SNPs in a known chromosome (from our mapping); minor allele frequency (MAF) > 0.02, genotype call rate (SNPs) > 0.95, genotype call rate (Animals) > 0.95 and identity-by-state (Animals) < 0.95. The filtered datasets were converted to harmonize the reference alleles with the reference genomes using a script based on the programs PlinkSeq v 0.08 (http://atgu.mgh.harvard.edu/plinkseq/index.shtml) and Plink v 1.07 (Purcell et al., 2007) which was necessary for the quality control of the re-sequencing data. After removing the positions corresponding to short indels and tri-allelic variants, which are incorrectly genotyped by the BeadChips, the number of SNPs both genotyped with the Chip and by whole genome sequencing was 47,122 for sheep 49,467 for goats, 37,779 for Asiatic mouflon and 41,751 SNPs for Bezoar ibex. The comparison of the *ovine* and *caprine* 50K BeadChips genotyping data with the WGS data was performed. The average (± s.d.) genotype concordance between the *ovine/caprine* 50K BeadChips and the WGS was 99.9% (± 0.1%) in sheep, 99.7% (± 0.0%) in goats, 99.7% (± 0.1%) in Asiatic mouflon and 98.5% (± 0.3%) in Bezoar ibex.

Setting up datasets of variants

From the individuals sequenced, we defined different groups depending on the question addressed. First, to evaluate the impact of sampling panels of variants and reducing the WGS coverage on the estimation of genetic parameters we designed four groups corresponding to 30 sheep, 30 goats, 14 Asiatic mouflon and 18 Bezoar ibex. In order to assess the effect of individual sample size, each analysis was performed for the whole groups and for two random subsets corresponding approximately to one third and two thirds of the total (i.e. respectively 10 and 20 sheep and goats, 5 and 10 Asiatic mouflon and 8 and 13 Bezoar ibex). Second, for detecting a signal of selection associated to the *RXFP2* locus (J. W. Kijas et al., 2012) and related to the presence/absence of horns, we had to consider additional sheep to constitute 2 contrasted groups of 15 horned and 15 polled individuals (Figure S1; Table S1).

For each group of individuals, a 12X WGS dataset was composed of all the SNPs called (see 'Read mapping, SNP calling and filtering' section) and used for within population analyses. Note that for cross-populations comparisons we only kept the variants found polymorphic in both groups considered, in order to prevent from any possible biased calling or filtering error. Then, variant panels were extracted from the 12X WGS datasets. Random panels were extracted using GATK SelectVariants (McKenna et al., 2010) consisting in 5 independent replicates for each of the 8 following numbers of variants: 1K, 5K, 10K, 50K, 100K, 500K, 1M and 5M. We also created non-random panels simulating the data obtained with commercial BeadChips or through exome capture. BeadChip data were obtained by calling variants at the Illumina® 50K Ovine or Caprine BeadChip SNP coordinates. We successfully extracted 42,117 variants for sheep, 47,245 variants for goats, 26,141 variants for Asiatic mouflon and 33,951 variants for Bezoar ibex. The combined datasets used for cross-population analyses included 30,870 variants for *Ovis* and 38,641 variants for *Capra*. In

sheep, the High Density BeadChip genotyping was simulated by calling WGS variants at the coordinates of the Illumina® ovine HD BeadChip. This gave 601,456 variants for sheep and of 444,169 variants for Asiatic mouflon. The combined dataset had 419,041 variants.

We simulated an exome capture only for *Ovis* because of the annotation of the goat genome was insufficiently advanced. The exome annotation was obtained from the sheep genome annotation that was available in ENSEMBL database by the time of analysis (25th September 2013) (ftp://ftp.ensembl.org/pub/pre/) and corresponded to 224,871 exons in 45,972 genes. The number of variants from these regions extracted from 12X WGS was 278,568 for sheep and 155,236 for Asiatic mouflon. The 93,409 variants polymorphic in both groups constituted the combined dataset. Thus, for the identification of potential surrogates for the WGS, the genotypes produced for the different variant panels and the different groups of individuals constituted a total of 946 datasets of which 516 were used for estimating within-group genetic diversity and 430 for cross-populations comparisons.

Simulating low-coverage re-sequencing data

The 12X WGS data were subsampled to simulate the output of a sequencing experiment with fewer reads were generated. For each of the 30 sheep and the 30 goats groups, three sub-sampled WGS datasets were generated comprising (i) 15 million, (ii) 30 million, (iii) 75 million paired reads, corresponding approximately to a 1X, 2X, and 5X sequencing coverage of the genome, respectively. Paired reads were randomly chosen from the full sequencing data using Picard Downsample, in such a way that all reads had an equal probability of being chosen, including duplicate or unaligned ones. Next, Picard MarkDuplicates was used to tag reads that appeared as duplicates in the sub-sampled datasets.

For each variant of the list generated for the 12X WGS, genotypes were called using GATK UnifiedGenotyper with the option GENOTYPE_GIVEN_ALLELES. Genotypes were improved and phased with Beagle4, and filtered at the individual level when the genotype probability was less than 0.95. Variants were kept when the genotype was called for more than 95% of the individuals. For each of the simulated coverages, the genotype at each variant position was compared to that obtained for the 12X coverage and classified as matching, un-matching or missing, for homozygotes and heterozygotes separately. Additionally, the individual observed heterozygosity was inferred for each coverage and used to estimate (i) Pearson correlation, (ii) Spearman correlation with 12X inferences. Slope values (*b)* were estimated for each depth by setting the intercept to 0.

Description of genome diversity

Genetic diversity within groups. Using Vcftools (Danecek et al., 2011) we estimated the observed heterozygosity (*Ho*) and inbreeding coefficient (*F*) with polymorphic autosomal bi-allelic SNPs, and the nucleotide diversity (π) by taking the averaged nucleotide diversity over all autosomal variants. Pairwise SNPs linkage disequilibrium (r^2) was also estimated with Vcftools on all bi-allelic non-rare variants (SNPs and indels with MAF>=0.05) for 5 segments of 2Mbp selected on 5 chromosomes (physical positions between 1 Mbp and 3 Mbp on chromosomes 5, 10, 15, 20 and 25). The extent of the linkage disequilibrium was assessed by the physical distance corresponding to $r^2 = 0.15$ ($r^2$0.15).

Genetic differentiation and structure. The genetic structure and differentiation was measured between domestics and wilds, as representative of population having diverged about 10,000 years ago (i.e., at the time of domestication). The averaged Fst (Weir & Cockerham, 1984) was estimated for bi-allelic variants with Vcftools. Additionally, genetic

structure was investigated through the Bayesian clustering approach sNMF (Frichot, Mathieu, Trouillon, Bouchard, & Francois, 2014) using bi-allelic variants. This method was specifically developed to estimate individual admixture coefficients on large genomic datasets.

Detection of a selection signature. We targeted the genomic region surrounding the *Relaxin/insulin-like family peptide receptor 2* gene (*RXFP2*; Chr 10: 29,454,677 – 29,502,617bp), which already showed a signature of selection related to polledness in sheep (Kijas et al., 2012; Dominik, Henshall, & Hayes, 2012). We extracted variants between positions 20 Mb and 40 Mb on chromosome 10 for 15 horned and 15 polled sheep and searched for selected sweeps in this region using two methods: A standard Fst test (Weir & Cockerham, 1984) and XP-CLR (Chen, Patterson, & Reich, 2010). For the later method, we estimated a constant recombination rate for this region based on the random 1M variants dataset, using the PAIRWISE program of LDhat v2.2 (Auton & McVean, 2007) with recommended parameters. XP-CLR scores were calculated for each grid point placed along the segment considered with a spacing of 5Kb. A maximum of 300 bi-allelic variants was considered in a sliding window of 0.5cM around the grid point and we down-weighted contributions of highly correlated SNPs (r^2>0.99).

Results

Variant calling was done using 12X coverage whole genome sequencing data for each species. It allowed the discovery of 29.96, 29.04, 21.71 and 17.32 million polymorphic variants for sheep, Asiatic mouflon, goats and Bezoar ibex, respectively (see Table 1), which correspond mostly to SNPs but also to small indels (i.e., 6% to 10% of the variants). We created both non-random (i.e., exome and SNP BeadChips) and random variant panels across a wide range of densities by subsampling SNPs from this WGS data, and assessed the potential of each panel to accurately represent genome diversity in domestic and wild animals. Lastly, low-coverage re-sequencing data were generated by randomly sampling various fixed percentages of reads from the raw 12X re-sequencing data (see 'Material and Methods' section) and genotypes were compared to the 12X WGS variants.

Description of genome diversity

We assessed the effect of variant subsampling on estimating genetic diversity by comparing the observed heterozygosity (*Ho*), inbreeding coefficient (*F*), nucleotide diversity (π) and Linkage disequilibrium ($r^2_{0.15}$) for both the WGS dataset and variant panels (see Table 1). Even at low-densities, random panels returned diversity metrics similar to those derived from WGS. Accurate estimates were obtained with all random panels of 5K or more markers for inbreeding (*F*) and nucleotide diversity (π), (Figures 1, S2, S3, S4) and with random panels of 10K or more markers for observed heterozygosity (*Ho*) (Figure S5). On the contrary, non-random panels of variants generated strongly biased estimations. The *ovine* 50K SNP and HD BeadChips and the *caprine* 50K SNP BeadChip from Illumina® showed considerable ascertainment biases by overestimating the diversity in all groups and datasets (Figures 1-3, S6, S7). For example, in the 30 sheep there was an overestimation of 129%, 108% and 194%

for π and 61%, 47% and 102% for *Ho* for these three panels, respectively. This ascertainment bias did not affect the estimation of the inbreeding coefficient (F) (Figures 3, S8). Whatever the bias, the ranking of individual *Ho* and *F* were not affected by the panel of variants used (Figure 3), but for π, the wilds even appeared less diverse than the domestics while WGS data showed the opposite (Figure 2). The dataset simulating exome capture underestimated π and *Ho* (e.g., underestimation of 20% and 6% for π and 8% and 5% for *Ho* in sheep and mouflon, respectively). The underestimation of the inbreeding coefficient (*F*) was higher in domestic but not in wild animals. Otherwise, we did not detect any sample size effect on π inference.

At least 1M random markers in sheep and 500K random markers in the other groups were necessary to have an estimation of *LD* ($r^2_{0.15}$) similar to that obtained with the WGS dataset. Smaller random panels and non-random panels biased this estimation (Figures S9, S10, S11). Exome capture especially biased the *LD* estimation in sheep (but not in Asiatic mouflon with 10 individuals and more). Moreover, in all groups, decreasing the number of individuals ~~highly~~ increased $r^2_{0.15}$. In particular, Asiatic mouflon had an $r^2_{0.15}$ of 4.52Kb for 14 individuals and of 79.4Kb for 5 individuals (Figure S10).

We also assessed the influence of the variant panels on two methods describing the genetic differentiation of wild versus domestic populations. First, we estimated the Weir & Cockerham (Weir & Cockerham, 1984) differentiation index (*Fst*), which was rather high between wild and domestic animals (*Fst* = 0.105 in *Ovis* and *Fst* = 0.087 in *Capra* from WGS data; Figures 4, S11). Independently of the number of variants used, there was a strong sampling effect due to the individuals selected for estimating *Fst*. For a given set of individuals the number of random variants did not influence greatly the mean *Fst* values

compared to that obtained with WGS data. The smallest random panels (from 1K to 50K) increased the variance in *Fst* estimates among marker-set replicates for a given set of individuals (Figures 4, S11). The caprine 50K SNP BeadChip Illumina® overestimated *Fst* values by 28% on average (Figure 5) and the ovine 50K and HD SNP BeadChips, and the exome capture slightly underestimated the *Fst* (2 to 13%). However, all non-random panels kept the ranking found with the WGS datasets for *Fst* estimated with different sets of individuals (r always > 0.98). Except for the caprine 50K BeadChip, the effect of the subsampling strategy on the *Fst* estimation was lower than that of the sample size. Second, we used the clustering method implemented in sNMF (Frichot et al., 2014) to estimate individual ancestry coefficients. The estimations depended neither on the number of markers used nor on the number of individuals in the sample. For the most likely number of clusters (K=2 for *Ovis* and *Capra* from the sNMF cross-validation values (Frichot et al., 2014)), all variant panels led to similar results (Figure S12).

Finally, we assessed the effect of the panel of SNPs used on the ability to detect a signature of selection. By contrasting 15 horned and 15 polled sheep, a single Fst test and the XP-CLR method (Chen et al., 2010) applied on the WGS dataset allowed detection of the signal of selection previously reported on the Relaxin/Insulin-Like Family Peptide Receptor 2 gene *RXFP2* (Kijas et al., 2012). This signal could be also detected with random panels of 100K markers and more, with the *ovine* 50K and HD BeadChips and with the exome capture (Figure 5, Figure S13). However, the intensity of the signal decreased progressively with the density of markers. Therewith, another non-previously reported sweep was detected with the XP-CLR method only for the WGS dataset with random panels of 5M and 1M variants. This signal was located in the region of the Neurobeachin *NBEA* and Mab21-like 1 *MAB21L1*

genes on chromosome 10 (positions 26,007,917-26,592,574 and 26,231,353-26,232,432 on OAR v3.1, respectively).

Difference between random and non-random panels

One major difference in the design of random panels of variants and the BeadChips relies on the distribution of variants across the genome. Figure 6 illustrates this in showing the distributions of the physical distances between adjacent variants in various panels for Moroccan sheep and goats. The random 50K variants as well as the random 500K variants and the HD ovine BeadChip showed a similar L-shaped curve indicating that variants were evenly distributed across the genomes. On the other hand, as it might be predicted, the caprine 50K BeadChip displayed an almost complete lack of SNPs separated by less than around 30Kb, while for the ovine 50K BeadChip the lack of SNPs in these categories is less drastic, at most around a half of the expected distribution for the shorter distances. The exome capture simulation displayed a very high occurrence of distances lower than 200 bp and a quasi absence of distance larger than 10kb, which might be expected (Figures 6, S14).

Reliability of low-coverage re-sequencing

1X, 2X and 5X whole genome sequencing coverage were simulated by randomly sampling reads in the 12X WGS data, and used for calling genotypes in 30 goats and 30 sheep. The 12X WGS allowed genotyping at 31,775,474 variant sites (31,735,229 at which more than 95% of individuals had genotypes called) for goats and 43,478,084 for sheep (43,105,056 at which more than 95% of individuals had genotypes called), and decreasing the coverage strongly reduced the number of variants that could be genotyped (missing genotypes, Table 2), while the number of variants wrongly genotyped remained rather low (mis-matching

genotypes, Table 2). Heterozygous genotypes were more affected than homozygous ones. Moreover, the decreasing coverage resulted in an increasing underestimation of *Ho* (around 1.2, 3 and 6 times for 5X, 2X and 1X, respectively), and in a decreasing preservation of the relative ranking of *Ho* values among individuals (Table 2). This ranking was better preserved in sheep than in goats.

Discussion

A wide range of methods are used for assessing the diversity of genomes, from whole sequencing of individual genomes, e.g. (Kidd et al., 2012; Altshuler et al., 2012) to the genotyping of a panel of variants randomly chosen or specifically designed, e.g. (Kijas et al., 2012). Because the choice of the methods (e.g., commercial DNA chip, low/high coverage whole genome sequencing, random panel of SNPs) might not be straightforward depending on the goal of the study, we set up this study to test the ability and robustness of a wide range of genome sampling strategies to (i) assess genome variability, (ii) infer population genetic structure and (iii) detect genome regions under selection. We applied this benchmark analyses on four different wild and domesticated groups representing different levels of diversity and linkage disequilibrium (Table 1).

Effect of sequencing coverage on the assessment of whole genome variations

Overall, the genotypes inferred from the 12X WGS were highly reliable according to the high concordance between 12X re-sequencing data and the 50K SNP BeadChips genotyping.

The simulation of 1X, 2X and 5X WGS datasets from the 12X WGS confirmed the sensitivity of population genetics inferences to the sequencing coverage previously found (e.g. Jansen et al., 2013; Alex Buerkle & Gompert, 2013), and helped to depict the effect of

reducing the coverage. As might be expected, we found that homozygote genotypes were more correctly called than heterozygote ones whatever the coverage. This is due to the fact that more reads should be mapped at a position for calling the two alleles of an heterozygote than for calling the unique allele of an homozygote. Additionally, the filtering process for variant calling induced a higher percentage of missing data for heterozygotes because it discarded any heterozygous genotype for which one allele was under or over-represented.

Thus, the decrease in WGS coverage first resulted in a decrease in variant density (increasing proportion of missing data). The density of reliable variants obtained when decreasing the coverage (> 250k for 1X and > 3M from 2X, see Table 2) would still have been sufficient to allow accurate estimation of population genetics parameters and detection of selection signatures (see below 'effect of the density of variants'). However, the trend is combined to a bias that strongly affected the estimations. This bias concerned both missing and erroneous genotyping (Figure 7), which affected mostly heterozygotes (even more when the coverage decreases) where the erroneous genotyping mostly produces homozygotes. This resulted in an underestimation of heterozygosity (*Ho*). However, the values obtained for the 5X coverage appeared to be just as accurate as those inferred from the 12X WGS (highly correlated values of *Ho*, and thus of *F*), for the studied species. This result is coherent with the findings of Li et al. (2011) who showed that in association studies, genotyping 3,000 individuals at 4X depth provided similar power to 30X sequencing of about 2,000 individuals. A way to overcome the concerns due to low-coverage sequencing is to analyse the data with adapted methods that use account genotype probabilities (e.g. Therkildsen & Palumbi). However, such a reliable *a priori* information is not always available in the study group.

Effect of the density of variants

When assessing the effect of the density of variants for various sample sizes, we generally observed a sample size effect on the estimation of summary statistics. This was observed whatever the species and the panel of variants, and the effect was especially strong when measuring population differentiation and linkage disequilibrium, even greater than the effect of variant density. For any of the chosen sample sizes, the density of variants was determinant to get a representative view of genome variations.

Many population genetics studies that infer demographic processes still rely on just a few dozens to a few hundreds of genetic markers aiming to be representative of the whole genome variations (Alhaddad et al., 2013; Olson, Whittaker, & Rhodes, 2013; Garza et al., 2014; Huang, Wang, Li, Wu, & Chen, 2014). We could, in fact, get a representative view of the whole genome variations by using a relatively small set of variants, provided they are randomly distributed across the genome (Figure 7). Low-density random panels of variants (i.e. 5K or 10K corresponding to 1 variant every ~300 or ~600Kb) gave estimates of summary statistics similar to those calculated from 12X WGS data whatever the species and its demographic history. The assessment of population structure through calculation of coefficients of ancestry was reliable whatever the panel density, while the estimations of *Fst* required at least 100K variants in the different populations/species. Furthermore, the estimation of *LD* and the detection of signatures of selection required higher variant densities: around one variant every 3 to 6Kb, which gave similar estimates to 12X WGS data with roughly one variant every 100 to 200bp.

The adequate densities of variants required for a reliable description of genomic variations depend on the pattern of *LD* decay across the genome. In the four studied species, those patterns represent a wide range of variation, with r^2 dropping below 0.15 within 4.5Kb in Asiatic mouflon and within more than 10Kb in sheep while excluding rare variants (Table 1). Consequently, we needed 500k to 1M variants to accurately estimate *LD* decay. All panels of

fewer than 100K variants (~1 variant per 30kb) produced incorrect estimations of r^2 for small distances (until 50Kb depending on the panel). ~~We could expect that~~The same orders of magnitude of variant densities would be required in species characterized by similar patterns of *LD* decay such as true ungulates (Meadows, Chan, & Kijas; Wade et al.; Villa-Angulo et al.; Ai, Huang, & Ren; Veroneze et al.; McCue et al.) or even other mammals with similar genetic characteristics (e.g., Laurie)~~e.g. *Anopheles arabiensis* with r^2 dropping bellow 0.2 within 200 bp (Marsden et al., 2014)~~. However, genomic patterns of *LD* decay depend on the demographic histories of populations, and reflect the changes in effective population sizes~~. It is likely that populations with smaller effective population sizes, which have experienced for example strong bottlenecks such as industrial breeds, could require smaller variant densities.~~

Selective sweeps, when they occur, increase *LD* in regions of several Kb surrounding the selected allele. This signature is then reduced by recombination, and the older the selective sweep the smaller will be the region still influenced around the selected allele (Stephens et al., 1998; Kim & Nielsen, 2004). In the case of the selective sweep that has occurred in the *RXFP2* gene, the signal is still extending ~350Kb and required at least a random panel of 100K variants in order to be detected. Therefore, higher density random panels would be needed to detect any weaker selective sweep (i.e. associated to lower *LD*).

Ascertainment bias in non-random panels

As expected, the estimation of almost all population genetics parameters was biased when using variants from commercial SNP BeadChips or exome (Figure 7). Measurements both of genome diversity and of population differentiation were affected. SNPs included in the design of the commercial panels were intentionally chosen according to their high level of polymorphism in several breeds (mainly European industrials, Alhaddad et al., 2013). This is

because these panels were designed to deploy breeding programs in connection with genomic selection and genome wide association studies, for which an accurate estimate of true population genetic diversity is irrelevant. The resulting ascertainment bias lead to an overestimation of the genomic diversity. The *ovine* HD BeadChip suffered less from this bias compared to the 50K *ovine* BeadChip due to the inclusion of high, medium and low frequency SNPs (James W. Kijas et al., 2014).

The exome capture data, while representing highly conserved regions, logically underestimated genetic diversity. If scientists are widely aware of the possible consequences of such ascertainment biases, as already reported for microsatellites markers (Wan, Wu, Fujihara, & Fang, 2004; Vowles & Amos, 2006; Curtis, Vine, & Knight, 2008; J. M. Miller et al., 2014), only one study has addressed this question for SNP chips until now to our knowledge (Albrechtsen et al., 2010).

The biased estimation of genetic diversity and genetic differentiation would be less problematic as long as the ranking of estimated values is preserved (e.g., the most variable individuals are actually those with the highest measured diversity). For example, when estimating animal genetic resources, this will allow finding the more diverse populations/breeds. However, we showed that this ranking was inverted when comparing the diversity of wilds and domestics with the ovine and caprine SNP Beadchips, which should be used with caution when comparing well-differentiated populations.

Besides the biases they induce, non-random panels of variants such as SNP chips also impact the estimation of genomic diversity according to the density of variants and their distribution across the genome.

Distribution of variants across the genome

Besides the effects of variant density and ascertainment bias, the distribution of variants across the genome also impacts the reliability of the characterization of the genome. For similar numbers of variants, the ovine and caprine 50K BeadChips were less accurate than random panels for estimating the *LD* decay over short distances. This is not surprising given the underrepresentation of close adjacent SNPs (<6Kb in ovine and <30Kb in caprine BeadChips, Figure 6) in these Beadchips. Moreover, the local density of Beadchip SNPs varied across the genome with some regions being far well covered than others. This explains why, like (J. W. Kijas et al., 2012), we were able to detect the signal of selection associated to the *RXFP2* gene with the *ovine* 50K SNP BeadChip but not with 50K variants random panels. The commercial BeadChip has four SNPs in a 148 Kb window centred on the *RXFP2* gene, which appeared to be enough for detecting selection, while the random 50K panel used had no variant in that window. Similarly, the *NBEA* signal was detected by XP-CLR with 1M variants or more. This illustrates our expectation that medium and low SNPs densities are limiting for the detecting less intense selective events .

The distribution of variants across the genome obviously determines the ability to detect selection signatures**,** and high-density variant panels are required to detect selected regions. One needs variants from regions under selection to find the associated signature, which is not necessarily assumed by low and medium-density panels of variants. This is more limiting when studying populations characterised by low overall linkage disequilibrium and old or low-intensity selection signatures.

Consequences for population genomics analyses

Outcomes of this study (Figure 7) might help setting up genotyping strategies to accurately infer population genetics statistics and test hypotheses on the structure and evolution of study

populations. These informations might be useful to study populations with similar genetic characteristics and LD decays (i.e., ungulate species and even other mammals). When measuring population genetic diversity, we show that commercial SNP panels could invert the ranking of populations. The bias induced by especially medium-density chips was also substantial when assessing inter-populations differentiation. Fst could be either overestimated (e.g. in caprine) or underestimated (e.g., in ovine). Similarly, random panels of less than 50K SNPs could lead to inaccurate estimates but the bias due to low sample size clearly exceed that due to low SNP density. Furthermore, if a commercial SNP panel was able to detect a strong selective sweep, this was related to its design (i.e. number of SNPs in the region of interest) and, as shown by random panels, medium and low SNP densities are clearly inadequate for a genome-wide estimation of the number of selective sweeps and/or the proportion of hard/soft sweeps. Medium-coverage resequencing would be recommended for such goals.

Our study also demonstrate that low coverage re-sequencing (1x and 2x) is not effective to get reliable genomic information at the individual level, as it is required e.g., in several breeding programs or landscape genomics analyses. This is especially true when no a priori information such as genotype probability is available from the study population, otherwise methods are available to impute individual information (Therkildsen & Palumbi). However, low-coverage may be sufficient to get information at the population level provided that a large number of individuals is studied (Alex Buerkle & Gompert).

Conclusion

The accuracy of panels of variants to describe genome variations depends on the distribution of these variants across the genome, according to the level of *LD* and its proper variability.

While high to medium coverage genome sequencing produces reliable genotyping, it remains costly both in terms of money and in data management, and thus surrogates of WGS data are still needed.

For model species, commercial standardized panels are generally already available and one should know their potential biases and use them cautiously. This is particularly true if the studied populations or breeds are genetically divergent from the individuals used for designing the set of variants. Our results showed that a few thousands of markers randomly chosen across the genome provide unbiased information. Therefore, it could be valuable to include such sets of variants when designing new high density SNP chips. In non-model species, the genotyping of individuals by SNP chips could be replaced by genotyping by sequencing approaches (RAD-seq), which would theoretically approximate a random distribution of markers across the genome, and could thus provide convenient surrogates for WGS data. As shown by our results, a suitable variant density should be targeted according to the aim of the study and the resources allocated. Finally, when considering Whole Genome Sequencing approaches, low-coverage (< 5X) sequencing might not be appropriate for setting up population genomics studies due to the important underestimation of heterozygote genotypes, unless using analysis methods provided for such purpose (e.g., considering genotype probabilities), what is not always possible.

Acknowledgments

This work was funded by the UE FP7 project *NEXTGEN* 'Next generation methods to preserve farm animal biodiversity by optimizing present and future breeding options'; grant agreement no. *244356.* We thank Eric Coissac who helped in setting-up the overall approach

and Bertrand Servin for the useful discussions. We are grateful to R. Hadria, M. Laghmir, L. Haounou, E. Hafiani, E. Sekkour, M. ElOuatiq, A Dadouch, A. Lberji, C. Errouidi and M. Bouali for helping in sampling in Morocco.

References

Ai, H., Huang, L., & Ren, J. (2013). Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. *Plos One, 8*(2). doi: 10.1371/journal.pone.0056001

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution, 27*(11), 2534-2547. doi: 10.1093/molbev/msq148

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology, 22*(11), 3028-3035. doi: 10.1111/mec.12105

Alhaddad, H., Khan, R., Grahn, R. A., Gandolfi, B., Mullikin, J. C., Cole, S. A., . . . Lyons, L. A. (2013). Extent of Linkage Disequilibrium in the Domestic Cat, Felis silvestris catus, and Its Breeds. *Plos One, 8*(1). doi: 10.1371/journal.pone.0053537

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., . . . Genomes Project, C. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi: 10.1038/nature11632

Auton, A., & McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Research, 17*(8), 1219-1227. doi: 10.1101/gr.6386707

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One, 3*(10). doi: 10.1371/journal.pone.0003376

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., . . . Devos, K. M. (2012). Reference genome sequence of the model plant Setaria. *Nature Biotechnology, 30*(6), 555-+. doi: 10.1038/nbt.2196

Bizon, C., Spiegel, M., Chasse, S. A., Gizer, I. R., Li, Y., Malc, E. P., . . . Wilhelmsen, K. C. (2014). Variant calling in low-coverage whole genome sequencing of a Native American population sample. *Bmc Genomics, 15*. doi: 10.1186/1471-2164-15-85

Black, W. C., Baer, C. F., Antolin, M. F., & DuTeau, N. M. (2001). Population genomics: Genome-wide sampling of insect populations. *Annual Review of Entomology, 46*, 441-469. doi: 10.1146/annurev.ento.46.1.441

Browning, B. L., & Browning, S. R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics, 194*(2), 459-+. doi: 10.1534/genetics.113.150029

Campbell, N., Sinagra, G., Jones, K. L., Slavov, D., Gowan, K., Merlo, M., . . . Taylor, M. R. G. (2013). Whole Exome Sequencing Identifies a Troponin T Mutation Hot Spot in Familial Dilated Cardiomyopathy. *Plos One, 8*(10). doi: 10.1371/journal.pone.0078104

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F. W., Alfoeldi, J., Barrio, A. M., . . . Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science, 345*(6200), 1074-1079. doi: 10.1126/science.1253714

Cathy C Laurie, D. A. N., Amy D Anderson, Bruce S Weir, Robert J Livingston, Matthew D Dean, Kimberly L Smith, Eric E Schadt, Michael W Nachman. (2007). Linkage Disequilibrium in Wild Mice. *Plos Genetics*.

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research, 20*(3), 393-402. doi: 10.1101/gr.100545.109

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., . . . Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America, 106*(45), 19096-19101. doi: 10.1073/pnas.0910672106

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., & Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *Bmc Genomics, 12*. doi: 10.1186/1471-2164-12-347

Curtis, D., Vine, A. E., & Knight, J. (2008). Investigation into the ability of SNP chipsets and microsatellites to detect association with a disease locus. *Annals of Human Genetics, 72*, 547-556. doi: 10.1111/j.1469-1809.2008.00434.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Anal, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi: 10.1093/bioinformatics/btr330

Dastjerdi, A., Robert, C., & Watson, M. (2014). Low coverage sequencing of two Asian elephant (Elephas maximus) genomes. *GigaScience, 3*, 12-12. doi: 10.1186/2047-217x-3-12

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. doi: 10.1038/nrg3012

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-+. doi: 10.1038/ng.806

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M. L., . . . Barch, D. H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics, 148*(3), 1269-1284.

Dominik, S., Henshall, J. M., & Hayes, B. J. (2012). A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. *Animal Genetics, 43*(4), 468-470. doi: 10.1111/j.1365-2052.2011.02271.x

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., . . . Wang, W. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature Biotechnology, 31*(2), 135-141. doi: 10.1038/nbt.2478

Everett, M. V., Grau, E. D., & Seeb, J. E. (2011). Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources, 11*, 93-108. doi: 10.1111/j.1755-0998.2010.02969.x

Fisher, R. (1958). POLYMORPHISM AND NATURAL-SELECTION. *Bulletin of the International Statistical Institute, 36*(3), 284-289.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & Francois, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics, 196*(4), 973-+. doi: 10.1534/genetics.113.160572

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing (Vol. 1207.3907v2): arXiv.

Garza, J. C., Gilbert-Horvath, E. A., Spence, B. C., Williams, T. H., Fish, H., Gough, S. A., . . . Anderson, E. C. (2014). Population Structure of Steelhead in Coastal California. *Transactions of the American Fisheries Society, 143*(1), 134-152. doi: 10.1080/00028487.2013.822420

Goldstein, D. B., & Weale, M. E. (2001). Population genomics: Linkage disequilibrium holds the key. *Current Biology, 11*(14), R576-R579. doi: 10.1016/s0960-9822(01)00348-7

Han, E., Sinsheimer, J. S., & Novembre, J. (2014). Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. *Molecular Biology and Evolution, 31*(3), 723-735. doi: 10.1093/molbev/mst229

Holloway, J. W., Beghe, B., Turner, S., Hinks, L. J., Day, I. N. M., & Howell, W. M. (1999). Comparison of three methods for single nucleotide polymorphism typing for DNA bank studies: Sequence-specific oligonucleotide probe hybridisation, TaqMan liquid phase hybridisation, and microplate array diagonal gel electrophoresis (MADGE). *Human Mutation, 14*(4), 340-347. doi: 10.1002/(sici)1098-1004(199910)14:4<340::aid-humu10>3.0.co;2-z

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., . . . Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature, 496*(7446), 498-503. doi: 10.1038/nature12111

Huang, H., Wang, H., Li, L., Wu, Z., & Chen, J. (2014). Genetic Diversity and Population Demography of the Chinese Crocodile Lizard (Shinisaurus crocodilurus) in China. *Plos One, 9*(3). doi: 10.1371/journal.pone.0091570

Jansen, S., Aigner, B., Pausch, H., Wysocki, M., Eck, S., Benet-Pages, A., . . . Fries, R. (2013). Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *Bmc Genomics, 14*. doi: 10.1186/1471-2164-14-446

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., . . . Dalrymple, B. P. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science, 344*(6188), 1168-1173. doi: 10.1126/science.1252806

Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., . . . Manel, S. (2013). INTEGRATING LANDSCAPE GENOMICS AND SPATIALLY EXPLICIT APPROACHES TO DETECT LOCI UNDER SELECTION IN CLINAL POPULATIONS. *Evolution, 67*(12), 3455-3468. doi: 10.1111/evo.12237

Jorde, L. B., Watkins, W. S., & Bamshad, M. J. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics, 10*(20), 2199-2207. doi: 10.1093/hmg/10.20.2199

Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., . . . Bustamante, C. D. (2012). Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *American Journal of Human Genetics, 91*(4), 660-671. doi: 10.1016/j.ajhg.2012.08.025

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Neto, L. R. P., San Cristobal, M., . . . Int Sheep Genomics, C. (2012). Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *Plos Biology, 10*(2). doi: 10.1371/journal.pbio.1001258

Kijas, J. W., Porto-Neto, L., Dominik, S., Reverter, A., Bunch, R., McCulloch, R., . . . Int Sheep Genomics, C. (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Animal Genetics, 45*(5), 754-757. doi: 10.1111/age.12197

Kim, Y., & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics, 167*(3), 1513-1524. doi: 10.1534/genetics.103.025387

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data, P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi: 10.1093/bioinformatics/btp352

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research, 21*(6), 940-951. doi: 10.1101/gr.117259.110

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics, 4*(12), 981-994. doi: 10.1038/nrg1226

Marsden, C. D., Lee, Y., Kreppel, K., Weakley, A., Cornel, A., Ferguson, H. M., . . . Lanzaro, G. C. (2014). Diversity, Differentiation, and Linkage Disequilibrium: Prospects for Association Mapping in the Malaria Vector Anopheles arabiensis. *G3-Genes Genomes Genetics, 4*(1), 121-131. doi: 10.1534/g3.113.008326

Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A., Clissold, L., Sampath, D., . . . Stein, N. (2013). Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *Plant Journal, 76*(3), 494-505. doi: 10.1111/tpj.12294

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., . . . Mickelson, J. R. (2012). A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. *Plos Genetics, 8*(1). doi: 10.1371/journal.pgen.1002451

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for

analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi: 10.1101/gr.107524.110

Meadows, J. R. S., Chan, E. K. F., & Kijas, J. W. (2008). Linkage disequilibrium compared between five populations of domestic sheep. *Bmc Genetics, 9*. doi: 10.1186/1471-2156-9-61

Miller, J. M., Malenfant, R. M., David, P., Davis, C. S., Poissant, J., Hogg, J. T., . . . Coltman, D. W. (2014). Estimating genome-wide heterozygosity: effects of demographic history and marker type. *Heredity, 112*(3), 240-247. doi: 10.1038/hdy.2013.99

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research, 17*(2), 240-248. doi: 10.1101/gr.5681207

Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., . . . Harrow, J. (2011). The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. *Molecular Biology and Evolution, 28*(10), 2949-2959. doi: 10.1093/molbev/msr127

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., . . . Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461*(7261), 272-U153. doi: 10.1038/nature08250

Nina Overgaard Therkildsen, S. R. P. Practical low-coverage genome-wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in non-model species. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12593

Olson, Z. H., Whittaker, D. G., & Rhodes, O. E. (2013). Translocation History and Genetic Diversity in Reintroduced Bighorn Sheep. *Journal of Wildlife Management, 77*(8), 1553-1563. doi: 10.1002/jwmg.624

Palti, Y., Gao, G., Liu, S., Kent, M. P., Lien, S., Miller, M. R., . . . Moen, T. (2015). The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources, 15*(3), 662-672. doi: 10.1111/1755-0998.12337

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., . . . Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics, 44*(6), 631-U641. doi: 10.1038/ng.2283

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution, 16*(12), 1791-1798.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559-575. doi: 10.1086/519795

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., . . . Dean, M. (1998). Dating the origin of the CCR5-Delta 32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics, 62*(6), 1507-1515. doi: 10.1086/301867

Teer, J. K., & Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics, 19*, R145-R151. doi: 10.1093/hmg/ddq333XTherkildsen, N. O., & Palumbi, S. (2017). Practical low-coverage genome-wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in non-model species. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12593

Veroneze, R., Lopes, P. S., Guimaraes, S. E. F., Silva, F. F., Lopes, M. S., Harlizius, B., & Knol, E. F. (2013). Linkage disequilibrium and haplotype block structure in six commercial pig lines. *Journal of Animal Science, 91*(8), 3493-3501. doi: 10.2527/jas.2012-6052

Villa-Angulo, R., Matukumalli, L. K., Gill, C. A., Choi, J., Van Tassell, C. P., & Grefenstette, J. J. (2009). High-resolution haplotype block structure in the cattle genome. *Bmc Genetics, 10*. doi: 10.1186/1471-2156-10-19

Vowles, E. J., & Amos, W. (2006). Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Molecular Biology and Evolution, 23*(3), 598-607. doi: 10.1093/molbev/msj065

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., . . . Broad Inst Whole Genome Assembly, T. (2009). Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. *Science, 326*(5954), 865-867. doi: 10.1126/science.1178158

Wan, Q. H., Wu, H., Fujihara, T., & Fang, S. G. (2004). Which genetic marker for which conservation genetics issue? *Electrophoresis, 25*(14), 2165-2176. doi: 10.1002/elps.200305922

Weir, B. S., & Cockerham, C. C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION-STRUCTURE. *Evolution, 38*(6), 1358-1370. doi: 10.2307/2408641

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., . . . Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature, 453*(7199), 1239-U1239. doi: 10.1038/nature07002

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*(2), 0097-0159.

Data accessibility

The variant call sets are archived in the European Nucleotide Archive with accession numbers provided in Table S2. The accession of the sample in the Biosamples archive, and of the corresponding aligned bam file in the ENA archive are listed in Table S1.

Author contributions

The study was done within the NEXTGEN project (coordinated by P.T.). F.P. and P.T. designed and supervised the study. B.B., M.I, M.B, M.C., A.B, A.C, W.Z., H.R.R. and S.N. collected the samples. A.S. supervised the work of her research group. A.A. and S.E. produced whole-genome sequences. I.S., L.C., B.B., S.E., and F.B. contributed to bioinformatic analyses. B.B., F.B., I.S. and W.Z. did the analyses. B.B. and F.P. produced the figures and drafted the paper. I.S., F.B., F.J.A., P.T., J.K. and P.F. reviewed and amended the paper.

Tables

Table 1. Population genomics statistics from WGS data for wild and Moroccan domestic small ruminants.

Species/Populations	*Ovis*		*Capra*	
	sheep	Asiatic mouflon	goats	Bezoar ibex
Number of individuals (*n*)	30	14	30	18
Number of variants	43,478,084	29,274,713	31,775,474	17,449,771
Number of polymorphic variants	29,958,788	29,039,121	21,709,831	17,321,976
Short indels	2,805,416	2,713,334	2,139,714	1,344,653
Variants with > 2 alleles	817,859	265,998	219,706	109,520
Heterozygosity (*Ho*)	0.222 ± 0.026	0.223 ± 0.032	0.189 ± 0.018	0.194 ± 0.025
Inbreeding coefficient (*F*)	0.061 ± 0.108	0.186 ± 0.118	0.056 ± 0.092	0.182 ± 0.106
Linkage disequilibrium $r^2_{0.15}$ (Kb)	10.22	4.52	8.70	6.69
Nucleotide diversity (π)	0.165	0.273	0.137	0.237

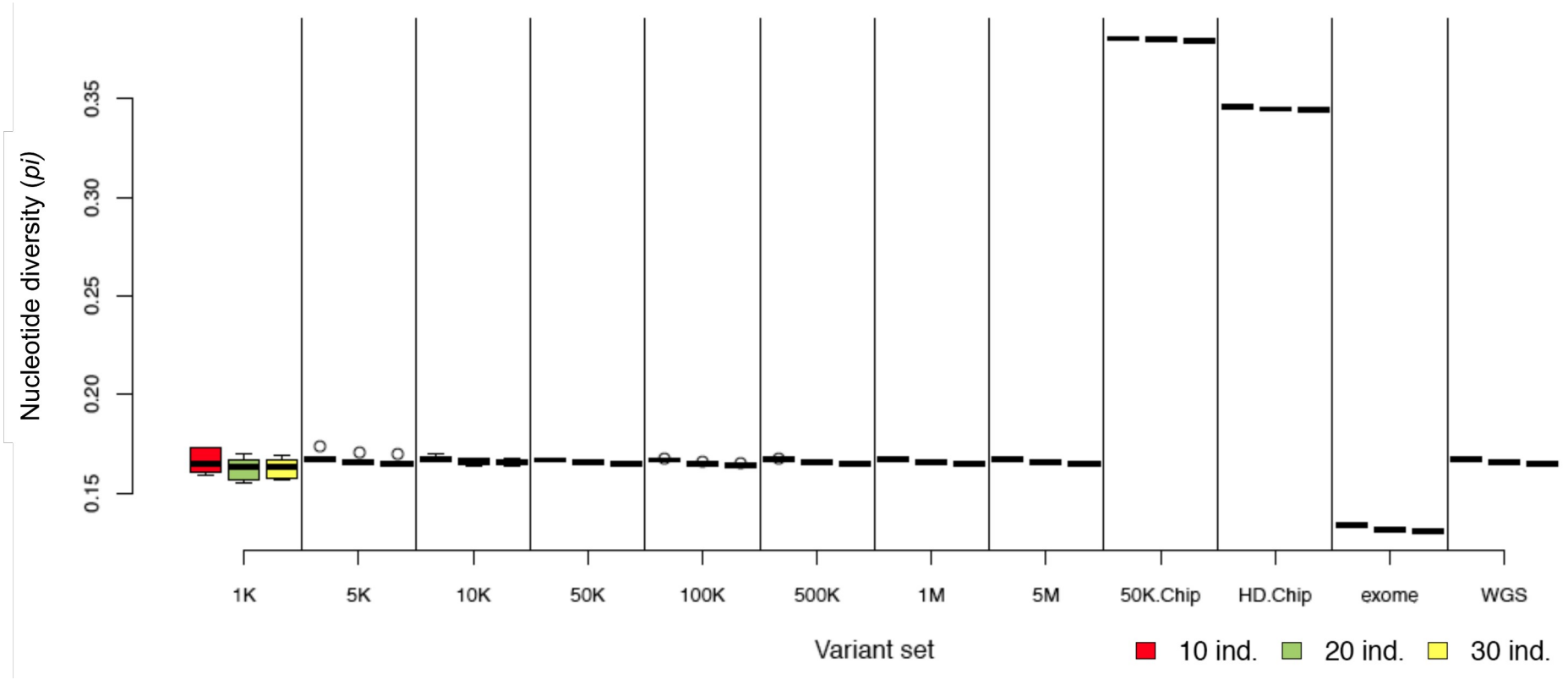
Table 2. Concordance between low-coverage re-sequencing and 12X coverage for homozygous and heterozygous genotypes.

Coverage		1x		2x		5x	
Species		Sheep	Goats	Sheep	Goats	Sheep	Goats
Genotypes for > 95% of individuals	# of sites	12,603,362	10,701,885	18,615,123	14,906,837	37,473,708	27,472,390
	# of polymorphic variants	268,491	254,313	4,327,012	3,324,740	24,074,435	17,108,812
Genotypes for 100% of individuals	# of sites	12,550,038	10,662,633	15,617,291	12,930,734	28,419,192	20,974,409
	# of polymorphic variants	259,177	249,056	1,783,255	1,737,328	15,847,527	11,296,131
Heterozygote genotypes	Matching 12x (%)	0.12 ± 0.02	0.19 ± 0.03	6.20 ± 0.79	5.68 ± 0.72	74.1 ± 8.6	71.5 ± 7.1
	Mis-matching 12x (%)	3.14 ± 0.33	4.38 ± 0.65	2.97 ± 0.30	3.66 ± 0.53	2.27 ± 0.24	2.16 ± 0.23
	Missing (%)	96.7 ± 11.1	95.4 ± 9.4	90.8 ± 10.3	90.7 ± 9.0	23.6 ± 2.6	26.4 ± 2.6
Homozygote genotypes	Matching 12x (%)	34.1 ± 0.1	38.1 ± 0.1	49.2 ± 0.2	52.3 ± 0.2	87.8 ± 1.7	87.5 ± 1.1
	Mis-matching 12x (%)	0.02 ± 0.02	0.02 ± 0.02	0.03 ± 0.01	0.02 ± 0.01	0.09 ± 0.10	0.07 ± 0.02
	Missing (%)	65.9 ± 2.1	61.9 ± 1.4	50.8 ± 1.9	47.7 ± 1.4	12.2 ± 0.4	12.5 ± 0.4
Correlations of *Ho* with 12X estimates	r (Pearson)	0.642	0.173	0.989	0.507	0.999	0.989
	Slope (Pearson)	0.149	0.176	0.329	0.281	0.864	0.818
	r (Spearman)	0.586	0.203	0.802	0.522	0.942	0.900

Number of sites and polymorphic variants were defined using two different percentages of genotyped individuals thresholds: > 95% and 100%. Other estimates were inferred from 95% filtering. *Ho* correlations were estimated according to Pearson and Spearman to compare rankings of individuals. Slopes were estimated by forcing the intercept of the linear regression to be 0.

Figures

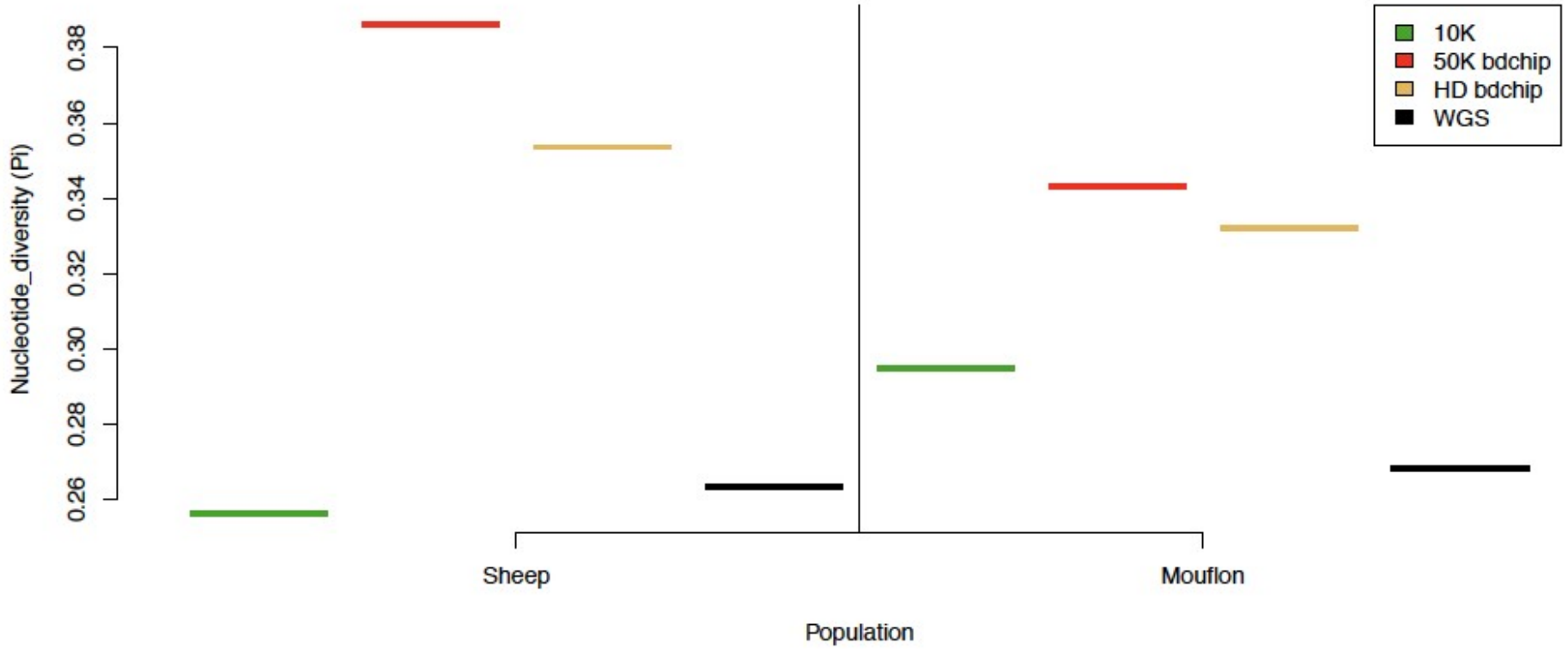
Figure 1. Nucleotide diversity (π) in sheep calculated from WGS data and from random and non-random panels of variants.



Nucleotide diversity (π) was estimated for each replicate of the different numbers of variants of the random panels and for each non-random panel. Sample sizes varied for each estimate from 10 to 30 individuals.

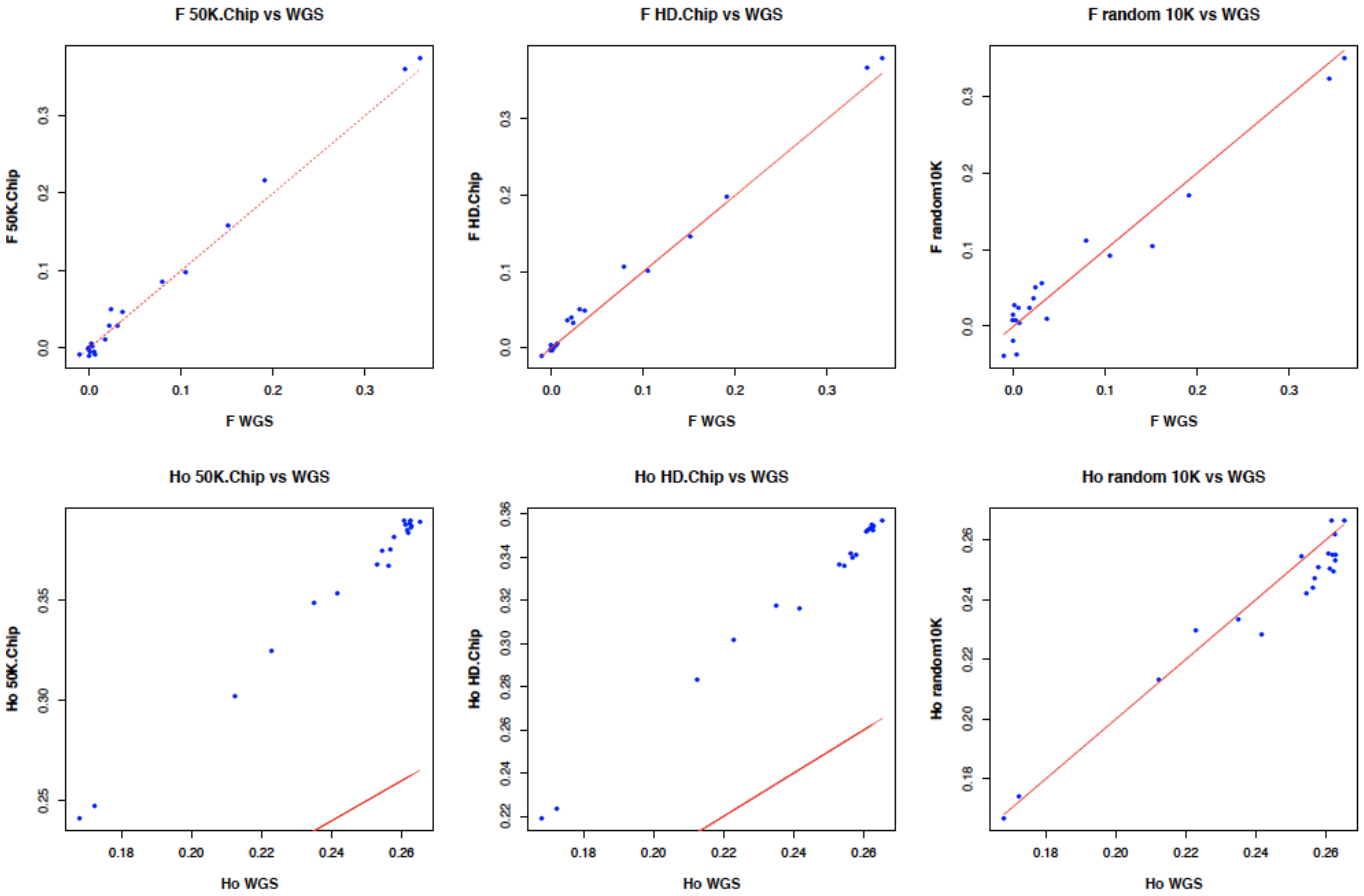
Random panels are denoted by their number of variants (from 1K to 5M) and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (all variants extracted from whole genome sequences). For each panel of variants the sample sizes are from left to right: 10 (red), 20 (green) and 30 (yellow) individuals.

Figure 2. Nucleotide diversity (π) estimated in two *Ovis* groups with random and commonly used panels of variants.



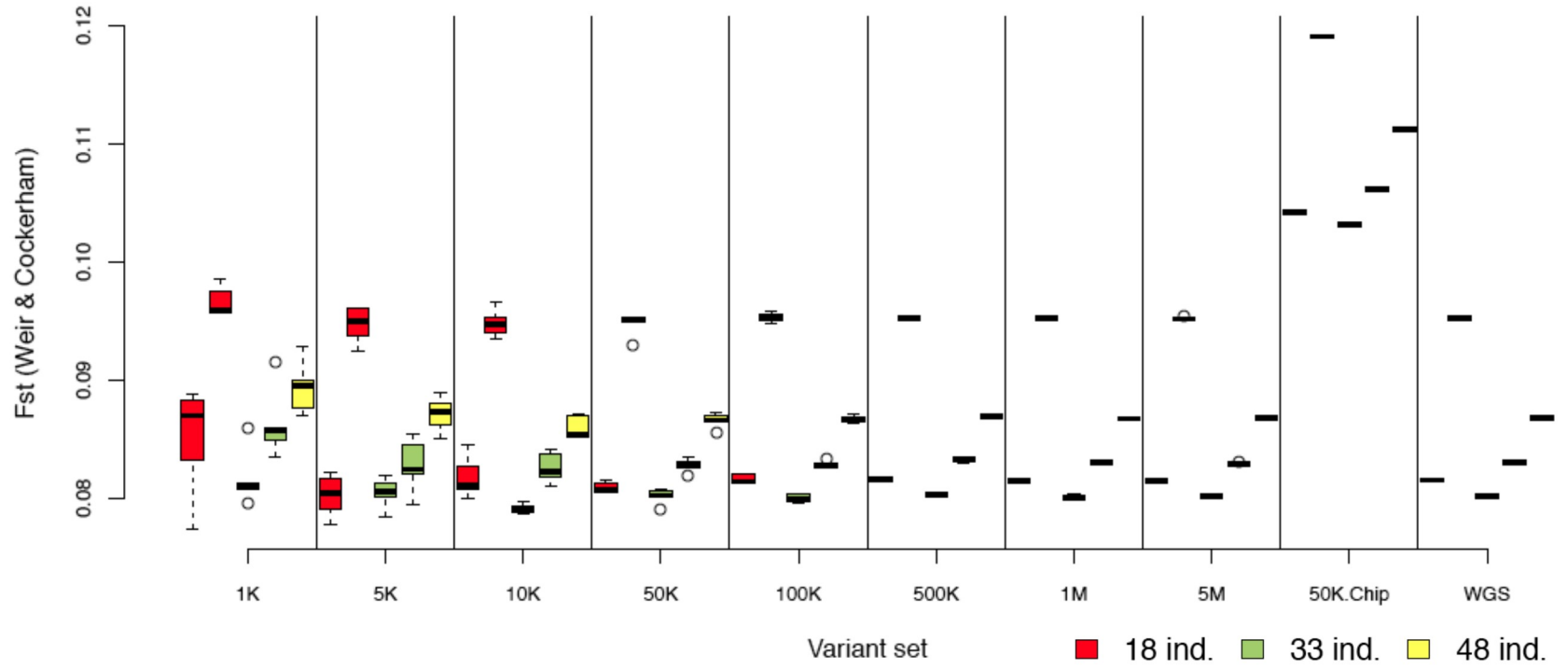
Plot of Nucleotide diversity (π) estimated with a random set of 10K variants sampled in sheep data (10K), and with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and variants extracted from whole genome sequences (WGS).

Figure 3. Estimates of individual inbreeding coefficient (*F*) and observed heterozygosity (*Ho*) from different panels of variants compared to WGS data estimates in sheep.



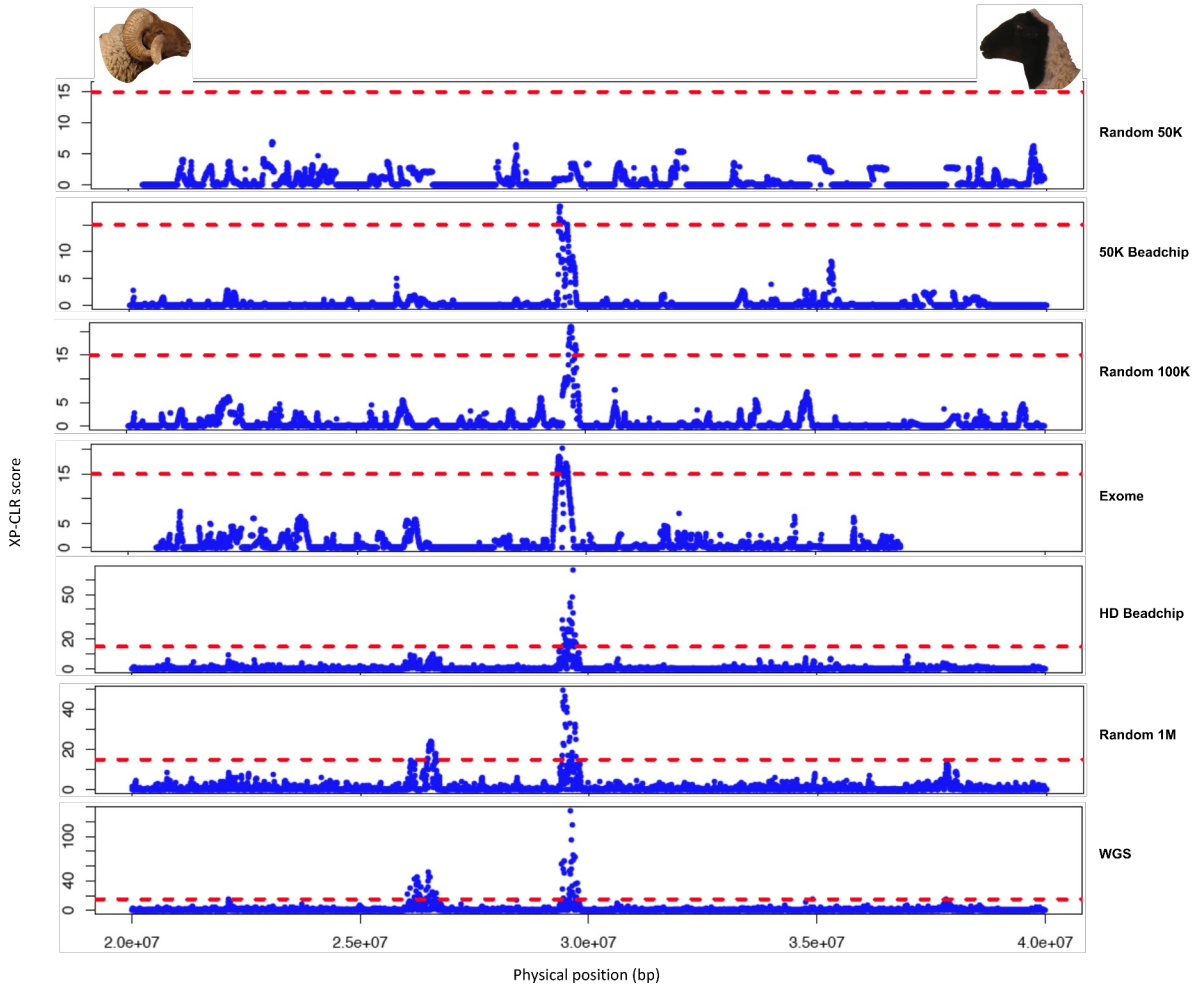
Plot of individual inbreeding coefficient (*F*; top) and observed Heterozygosity (*Ho*; bottom) estimated with variants extracted from whole genome sequences (WGS) versus inferences with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and 1 set of 10K variants defined in Moroccan sheep (random 10K). The red lines represent the relationship for which the estimates of the different panels are identical to the ones of WGS inferences.

Figure 4. Fixation index (*Fst*) between Moroccan goats and Bezoar ibex for different panels of variants and different samples of individuals.



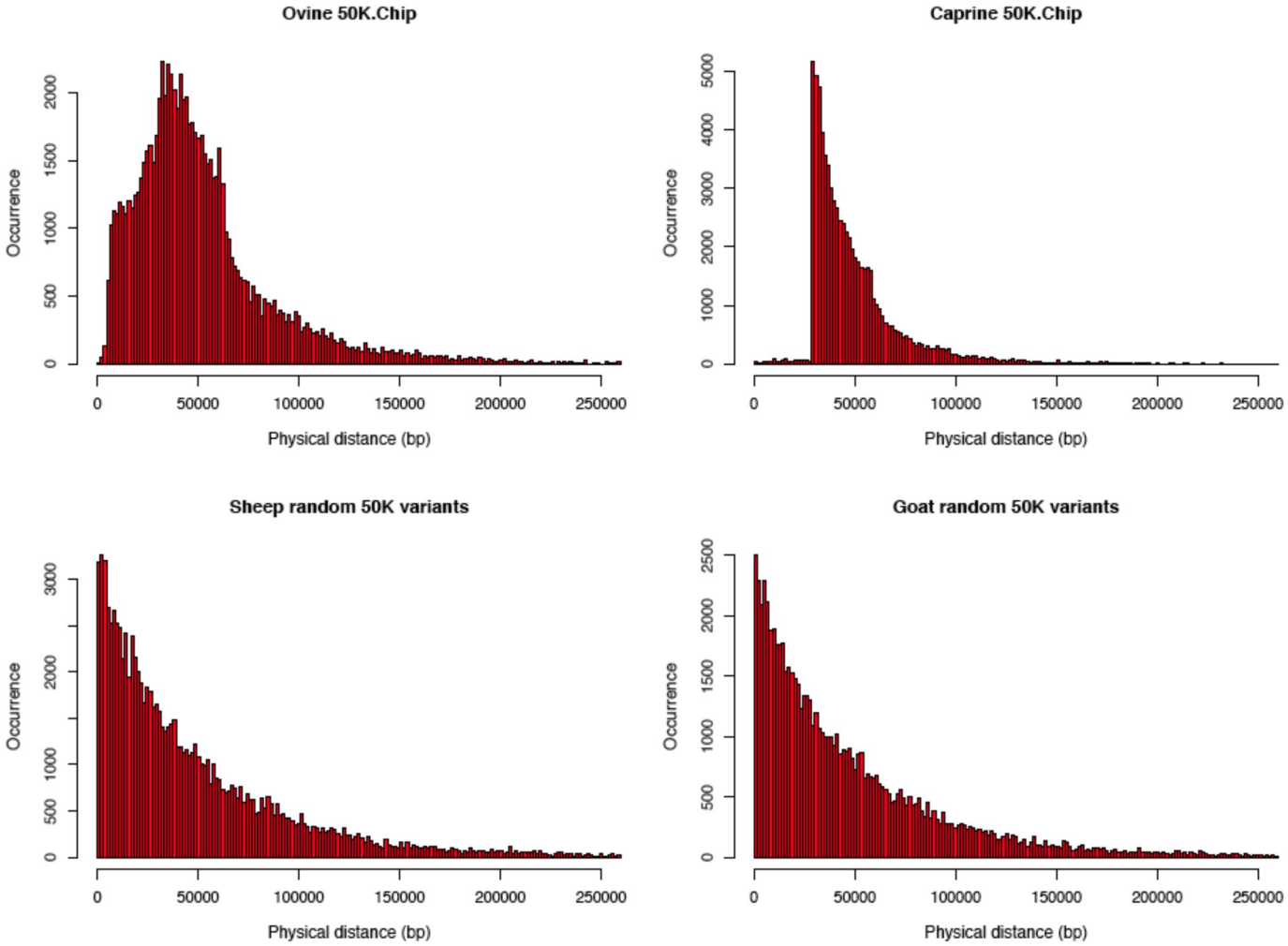
The fixation index *Fst* (Weir & Cockerham, 1984) was estimated for each random panel for the 5 independent replicates , and for each non-random dataset for each sample size. Random and non-random panels are denoted according to Figure 1.

Figure 5. XP-CLR scores calculated along the 20M-40M bp segment on chromosome 10 in a horned-polled Moroccan sheep comparison for different sets of variants.



The two peaks of XP-CLR scores showed in the WGS data plot are located respectively in the two genes *NBEA* (chr 10: 26,007,917 - 26,592,574) and *MAB21L1* (chr 10: 26,231,353 - 26,232,432) and in the *RXFP2* gene (chr 10: 29,454,677 - 29,502,617 bp). The horizontal dashed line represents a XP-CLR score of 15 to represent a scale among the different plots.

Figure 6. Distribution of physical distances between adjacent variants in 50K BeadChips and random panels of 50K variants.



Ovine 50K.Chip: Illumina® ovine 50K SNP Beadchip; Caprine 50K.Chip: Illumina® caprine 50K SNP Beadchip; Sheep random 50K variants: random panel of 50K variants defined in Moroccan sheep; Goat random 50K variants: random panel of 50K variants defined in Moroccan goats.

Figure 7. Efficiency and accuracy of different genome scan strategies

For each purpose, the different strategies are rated according to the accuracy of the estimates taking as a reference the WGS (12x depth) inferences. Grey dots indicate that the genotyping approach allow detecting some selection signatures but could miss some further signals detected by WGS (12x depth) data. MD chip = 50K SNP BeadChip (caprine and ovine); HD chip = 600K SNP Ovine BeadChip.