Batch effects in a multiyear sequencing study: false biological trends due to changes in read lengths

Leigh, D M ; Lischer, H E L ; Grossen, C ; Keller, L F

Abstract: High-throughput sequencing is a powerful tool, but suffers biases and errors that must be accounted for to prevent false biological conclusions. Such errors include batch effects; technical errors only present in subsets of data due to procedural changes within a study. If overlooked and multiple batches of data are combined, spurious biological signals can arise, particularly if batches of data are correlated with biological variables. Batch effects can be minimized through randomization of sample groups across batches. However, in long-term or multiyear studies where data are added incrementally, full randomization is impossible, and batch effects may be a common feature. Here, we present a case study where false signals of selection were detected due to a batch effect in a multiyear study of Alpine ibex (Capra ibex). The batch effect arose because sequencing read length changed over the course of the project and populations were added incrementally to the study, resulting in nonrandom distributions of populations across read lengths. The differences in read length caused small misalignments in a subset of the data, leading to false variant alleles and thus false SNPs. Pronounced allele frequency differences between populations arose at these SNPs because of the correlation between read length and population. This created highly statistically significant, but biologically spurious, signals of selection and false associations between allele frequencies and the environment. We highlight the risk of batch effects and discuss strategies to reduce the impacts of batch effects in multiyear high-throughput sequencing studies.

DOI: https://doi.org/10.1111/1755-0998.12779

Received: 9 October 2017 | Revised: 27 February 2018 | Accepted: 1 March 2018

DOI: 10.1111/1755-0998.12779

WILEY MOLECULAR ECOLOGY RESOURCES

SPECIAL ISSUE: ASSOCIATION MAPPING IN NATURAL POPULATIONS

Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths

D. M. Leigh[1,2,3] | H. E. L. Lischer[1,2] | C. Grossen[1] | L. F. Keller[1,4]

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[2]Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Genopode, Lausanne, Switzerland

[3]Department of Biology, Queen's University, Kingston, ON, Canada

[4]Zoological Museum, University of Zurich, Zurich, Switzerland

Correspondence
Deborah M. Leigh, Department of Biology, Biosciences Complex, Queen's University, Kingston, ON, Canada.
Email: deborahmleigh.research@gmail.com

Funding information
URPP Evolution in Action

Abstract

High-throughput sequencing is a powerful tool, but suffers biases and errors that must be accounted for to prevent false biological conclusions. Such errors include batch effects; technical errors only present in subsets of data due to procedural changes within a study. If overlooked and multiple batches of data are combined, spurious biological signals can arise, particularly if batches of data are correlated with biological variables. Batch effects can be minimized through randomization of sample groups across batches. However, in long-term or multiyear studies where data are added incrementally, full randomization is impossible, and batch effects may be a common feature. Here, we present a case study where false signals of selection were detected due to a batch effect in a multiyear study of Alpine ibex (*Capra ibex*). The batch effect arose because sequencing read length changed over the course of the project and populations were added incrementally to the study, resulting in nonrandom distributions of populations across read lengths. The differences in read length caused small misalignments in a subset of the data, leading to false variant alleles and thus false SNPs. Pronounced allele frequency differences between populations arose at these SNPs because of the correlation between read length and population. This created highly statistically significant, but biologically spurious, signals of selection and false associations between allele frequencies and the environment. We highlight the risk of batch effects and discuss strategies to reduce the impacts of batch effects in multiyear high-throughput sequencing studies.

KEYWORDS
genotyping error, GWAS, long-term data, outlier, RADseq, sequencing error

1 | INTRODUCTION

High-throughput sequencing (HTS) technologies have enabled marker acquisition on a genome-wide scale without the need for access to a reference genome (Birzele et al., 2010; Ratan, Zhang, Hayes, Schuster, & Miller, 2010). Consequently, genome-level studies are now possible in most nonmodel species, and HTS has been used widely to obtain genome-wide single nucleotide polymorphism (SNP) data in a variety of organisms. While invaluable to advancing genetic studies of nonmodel organisms, HTS studies are not error-free. The simplest and most widely known form of errors is false SNP calls arising from sequencing error (Sims, Sudbery, Ilott, Heger, & Ponting, 2014). HTS platforms have rates of sequencing error that range from 0.3% to 20% total nucleotide calling errors (Ross et al., 2013). In addition, biases can arise during sample storage or library preparation. For example, the type of sonicator used to shear DNA fragments to a size suitable for sequencing can introduce technical differences between libraries (Davey et al., 2013; Poptsova et al.,

2014). Furthermore, variation among sequencing lanes may create differences in coverage across specific sequence motifs (Ross et al., 2013). These types of biases are known as library biases, lane biases, and more broadly as batch effects. Batch effects are thus technical sources of variation that differ among subsets of the data (Leek et al., 2010). If batches of data are combined without accounting for their presence, batch effects can create systematic differences within the data set. When batch effects are correlated with biological variables, the systematic differences among batches may lead to invalid biological conclusions. For example, false SNP calls arising from batch effects caused by the combination of several sequencing technologies led to biologically invalid associations in a genome-wide association study (GWAS) of the genomic determinants of human longevity (Sebastiani et al., 2011).

One way to address batch effects in HTS studies is to randomly divide samples from a population or experimental group across libraries and sequencing lanes (Buhule et al., 2014; Taub, Corrada Bravo, & Irizarry, 2010). This reduces the risk of a correlation between technical artefacts and biological variables, thereby decreasing the risk that batch effects will lead to spurious biological trends. However, as HTS sequencing develops, an increasing number of multiyear or long-term studies will add sequencing data over time (Goodwin, McPherson, & McCombie, 2016). Due to the time limitations and monetary costs of HTS libraries, resequencing old and new libraries simultaneously to create a fully randomized data set will rarely be possible (discussed for microarrays in Müller et al., 2016). If populations or generations are added incrementally to a multiyear data set, false biological signals may arise.

The potential consequences of batch effects in such data will be study- and method-dependent. For instance, if batch effects are small and a study utilizes methods that average over a large genomic window of polymorphic markers (e.g., Xu et al., 2015 for an example of a selection study), or over thousands of markers across a genome (e.g., for heterozygosity estimations, Fumagalli, 2013), the technical artefacts may have little impact on the estimated quantities. Such studies may thus be less susceptible to biases arising from batch effects. However, batch effects may still add variation that reduces statistical power (Leek & Storey, 2007). In contrast, when scientific questions focus on specific SNPs, batch effects may be more problematic. For example, if a reduced representation sequencing method, such as restriction site-associated DNA sequencing (RADseq), is used in a GWAS (Yu et al., 2015), each relevant section of the genome will often only be represented by a single SNP (Lowry et al., 2017; Mckinney, Larson, Seeb, & Seeb, 2017). In such cases, batch effects can easily cause bias. It should be noted, however, that in studies with a higher marker density (e.g., from whole genome sequencing), associations would be confirmed by a cluster of markers rather than a single marker, making false associations due to batch effects less likely.

Here, we present a case study of a RADseq-based analysis of selection in Alpine ibex (*Capra ibex*), where a batch effect due to the addition of sequencing data over several years created statistically highly significant, but incorrect, biological signals. We discuss the origin of this batch effect, how it was identified, and its impact on the biological conclusions. We end with a brief discussion of ways in which the consequences of batch effects can be reduced.

2 | METHODS

2.1 | Study system

The Alpine ibex (*Capra ibex*), a species of wild goat once widespread across the Alpine arc of Europe, underwent a severe global bottleneck about 200 years ago due to overhunting, followed by a reintroduction programme that started in Switzerland (Stüwe and Nievergelt, 1991; Biebach & Keller, 2009). The reintroduction programme has enabled Alpine ibex to recolonize much of their former range, and population estimates suggest that over 40,000 Alpine ibex now live in the Alps (Stüwe & Grodinsky, 1987; Stüwe and Nievergelt, 1991; Shackleton and Group ISCI, 1997). This study focused on populations of Alpine ibex within Switzerland, where estimates indicate over 17,000 Alpine ibex are now spread across more than 40 populations, and for which the reintroduction history is often meticulously documented (Shackleton and Group, 1997; BAFU, 2015; Biebach & Keller, 2009).

2.2 | Data collection

To investigate potential signals of selection RADseq libraries were generated for 87 individuals in 2012 (Grossen, Biebach, Angelone-Alasaad, Keller, & Croll, 2017) and for 219 individuals in 2015, representing a total of 23 populations (Table 1). Samples were collected from legally hunted animals or using tissue biopsy darts (Biebach & Keller, 2009). DNA was extracted from all samples using the QIAGEN DNeasy Blood Tissue Kit (QIAGEN). DNA quantity was assessed using a PicoGreen assay (QuantIT)(Ahn, Costa, & Emanuel, 1996) and a 1% agarose gel to check for degradation of samples. RAD libraries were constructed following the Etter protocol (Etter, Bassham, Hohenlohe, Johnson, & Cresko, 2011) with minor modifications (see Grossen, Keller, Biebach, & Croll, 2014) using a SBF1 digest. Due to the staggered data collection, RADseq data were obtained in three batches that were sequenced to different read lengths because of sequencing technology changes during the project. Specifically, 87 individuals were sequenced paired-end to a read length of 100 base pairs (2012 data, 10 libraries and two sequencing lanes), 30 individuals were sequenced paired-end to 125 base pairs (2015a data, three libraries and one sequencing lane), and 189 individuals were sequenced paired-end to 140 base pairs (2015b data, 24 libraries and eight sequencing lanes; see Table 1). We aimed for a 15× coverage per individual. The 2012 RADseq libraries were sequenced at the Functional Genomics Centre of the University of Zurich and at the Genomics Facility Basel of the ETH Zurich. The 2015a and 2015b samples were sequenced only at the Genomics Facility Basel of the ETH Zurich. All samples were sequenced on an Illumina HiSeq. In an attempt to avoid correlations among technical and biological variables, samples from individual populations were

split across libraries and sequencing lanes where possible. However, due to the addition of populations to the study over time, only samples from five populations were sequenced in more than one batch.

2.3 | Data processing

Sequence data were demultiplexed into samples using the unique inline barcodes and the FASTX-TOOLKIT (version 0.0.14; Pearson, Wood, Zhang, & Miller, 1997) allowing one partial overlap and two mismatches (including deletions) in the barcode sequence. Due to a single base insertion found before the inline barcode of a number of unassigned reads, the first base pair of all forward unassigned reads was trimmed, and these reads were redemultiplexed allowing only one mismatch and partial overlap. Demultiplexed reads were then cleaned with TRIMMOMATIC (version 0.32: Bolger, Lohse, & Usadel, 2014) using the palindrome mode and a 16 base pair seed with two mismatches to remove any sequences of adaptors that may be present at the end of reads. TRIMMOMATIC also trimmed low quality regions of a read, that is bases at the start and end of reads with a PHRED score of less than three. In addition, reads were trimmed if a sliding window of four bases with a PHRED score below 15 arose. Reads shorter than 36 nucleotides long after trimming were discarded. All remaining reads were then mapped to the goat genome (*Capra hircus*, 01.Genome, scaffold file; Dong et al., 2013) with BOWTIE2 (version 2.2.5: Langmead & Salzberg, 2012) using the sensitive mode and an expected insert size of 50–800 base pair between reads. PCR-duplicate reads were removed after mapping with MarkDuplicates in Picard Tools (Broad, 2016).

Variant nucleotides were called using FREEBAYES (version v0.9.21-19-gc003c1e: Garrison & Marth, 2012). No prior population definitions were assumed, and to help reduce computational time, only a maximum of three alleles were evaluated for each variant site, and a maximum complex gap of half the minimum read length was allowed (18 base pairs). FREEBAYES identified over seven million SNPs before filtering. Fixed variants in the Alpine ibex were removed before quality filtering using custom scripts (available on dryad). SNPs were then filtered with vcfilter in FREEBAYES and VCFTOOLS 0.1.14 (Danecek et al., 2011; Garrison, 2016). SNPs were removed if the site quality-to-depth ratio was <2, if site quality was <1, or if mean mapping quality was <30. Similarly, SNPs with a genotyping quality of <20, a PHRED score <40 and a read depth below eight were also excluded. SNPs with a mapping quality ratio of the two alleles below 0.9 or above 1.05 were also removed to avoid strand bias, as recommended by DDOCENT (Puritz, Hollenbeck, & Gold, 2014). Furthermore, SNPs with

TABLE 1 Details of the populations used in this study

Population	Sample size after filtering	Years sequenced	Latitude	Longitude	Mean snow depth (cm)
Albris	11	2013, 2015a, 2015b	46.51	9.99	120.9
Alpstein	10	2015a	47.24	9.35	266.3
Arolla	10	2015b	46.04	7.53	20.2
Bire-Oeschinen	8	2015b	46.51	7.70	27.4
Brienzer Rothorn	10	2013, 2015b	46.78	8.04	*NA*
Churfirsten	10	2015b	47.15	9.27	278.0
Crap da Flem	10	2015b	46.88	9.28	27.3
Creux du Van	8	2015b	46.93	6.72	*NA*
Falknis	9	2015b	47.03	9.71	145.1
Fluebrig	10	2015b	47.01	8.85	27.3
Flueela	9	2015b	46.69	9.98	144.8
Graue Hoerner	6	2013, 2015b	46.95	9.39	27.3
Gross Lohner	10	2015a	46.42	7.60	27.4
Justistal	8	2015b	46.74	7.81	*NA*
Oberbauenstock	10	2015b	46.92	8.51	18.2
Pilatus	8	2015b	46.97	8.25	*NA*
Mont Pleureur	9	2013, 2015b	46.02	7.32	5.5
Schwarzmönch	9	2013, 2015b	46.55	7.89	27.4
Tanay	8	2015b	46.34	6.78	0.6
Val Bever	6	2015b	46.55	9.79	30.4
Weisshorn	7	2013	46.19	7.79	34.1
Wetterhorn	9	2015a	46.62	8.10	184.2
Wittenberg	10	2015b	46.40	7.21	*NA*

The latitude and longitude are the average point from which samples were obtained. Mean snow depth is the mean depth of snow on the ground (cm) from November to April across all years the population existed. The years sequenced indicate the library sequenced to 100 base pairs (2013), 125 base pairs (2015a) and 140 base pairs (2015b).

no observation on the other DNA strand were removed. In addition, SNPs with a mean maximum depth across all individuals that was above twice the mean depth were removed (Li, 2014), and SNPs with a heterozygosity of >0.6 were also excluded to ensure that paralogs and duplicate regions were not present. Finally, all singletons, non-biallelic SNPs and individuals with >50% of the final SNP set missing were removed. This resulted in a sample of 210 individuals, of which five were additionally removed, because they had been sampled in duplicate or displayed biologically impossible heterozygosity where nearly all SNPs were heterozygous or entirely homozygous. For the selection analysis, SNPs on the X-chromosome and SNPs that could not be mapped to the annotated goat genome (CHIR_1.0, Dong et al., 2013) with MUMMER (version 3.23: Delcher, Phillippy, Carlton, & Salzberg, 2002) were removed. After these SNP filtering and quality control steps, a final data set of 205 individuals and 6,677 SNPs remained and was used for selection detection. An average of 23% (±20%) of SNPs was missing in each individual. SNPEFF (version 4.1 l: Cingolani et al., 2012) was used to annotate the putative effect of a SNP and to determine the nearest gene to each SNP. A custom database for this analysis was created using the goat genome (version CHIR 1.0; *Capra hircus*, Dong et al., 2013).

2.4 | Selection detection

To identify putative signals of selection, we used both outlier analyses (F_{ST}-like approaches) and genetic-environment association (GEA) analyses. F_{ST}-like approaches detect selection by identifying loci with larger differences in allele frequencies between populations than expected based on the population average. GEA approaches identify selection by detecting correlations between allele frequencies and environmental variables (Hoban et al., 2016). Three selection detection programmes that can employ F_{ST}-like or GEA approaches were used as follows: BAYENV 2.0 (F_{ST}-like and GEA approaches, Günther & Coop, 2013), BAYPASS 2.1 (F_{ST}-like and GEA approaches, Gautier, 2015a) and OUTFLANK (F_{ST}-like approach only, Whitlock and Lotterhos, 2015a). These three programmes were chosen as they have previously been shown to perform well in species with complex patterns of relatedness among populations that arise from demographies that do not follow a simple island model (Günther & Coop, 2013; Lotterhos & Whitlock, 2014; Gautier, 2015a; Whitlock and Lotterhos, 2015a). Only SNPs identified by either all three softwares using the F_{ST}-like approaches, or by the GEA approaches in BAYENV 2.0 and BAYPASS 2.1 and the F_{ST}-like approach of OUTFLANK, were considered true positives. These are hereafter referred to as "triple positives."

The environmental variables used for the GEA approach were means of local snow conditions (new snow accumulation, snow depth and number of days above average snow depth), temperature (minimum, maximum and mean air temperature) and precipitation, variables that have previously been shown to affect Alpine ibex population dynamics (Jacobson, Provenzale, Von Hardenberg, Bassano, & Festa- Bianchet, 2004; Grøtan, Saether, Filli, & Engen, 2007; MeteoSwiss, 2016). All environmental variables were averages from across all years since the year after population founding (or the beginning of the weather records, if these started after a population was founded) and were measured at the closest meteorological station. The year after the first translocation was used to ensure the environmental conditions captured the first possible reproductive event onwards. Each environmental variable included was divided into winter (November–April) or summer (May–October).

BAYENV 2.0 was executed three times with 2×10^5 Markov chain Monte Carlo (MCMC) iterations both for the estimation of covariance in allele frequencies between populations, as well as for the tests for outliers and environmental correlations in allele frequencies (Blair, Granka, & Feldman, 2014). All polymorphic SNPs were used for the estimation of the covariance matrix of allele frequencies. The final covariance matrix estimated in each of the three runs was used in the selection detection analysis. Putatively selected SNPs were determined differently for the GEA approach and the F_{ST}-like approach. For the GEA approach, SNPs were considered to be under selection if (i) their Bayes factor (BF) was above three for all replicate runs, which indicates substantial support for a SNP being under selection (Nadeau, Meirmans, Aitken, Ritland, & Isabel, 2016), and (ii) their average Spearman's rho correlation coefficient between allele frequencies and environmental variable was in the top or bottom 2.5% of all SNPs across the three runs (Günther & Coop, 2013). For the F_{ST}-like approach, SNPs were considered outliers if their X^TX value was among the top 100 ranking SNPs across all three runs (Günther & Coop, 2013). X^TX is a differentiation statistic similar to F_{ST}, and high X^TX values signal excess differentiation and, hence, putative directional selection.

BAYPASS 2.1 was run under default conditions (20 pilot runs of 1,000 MCMC iterations, 5000 MCMC iterations for "burn-in") using the auxiliary model and an estimate of the covariance matrix produced by the core model (Gautier, 2015a). The covariance matrix was also used to correct for shared population history in the selection analysis. The thresholds used to identify loci putatively under selection were taken from the best-practice tutorial accompanying BAYPASS 2.1 (Gautier, 2015b) and included two criteria: (i) For the GEA approach, SNPs were considered to be under selection when they showed a $10\times$ $\log_{10}$ Bayes factor (db) >20 (Gautier, 2015a). (ii) For the F_{ST}-like approach, X^TX outliers were determined by creating three BAYPASS -simulated data sets of 1,000 SNPs to determine a 99% threshold for X^TX values. SNPs in the top 1% of X^TX values were considered outliers (Gautier, 2015b). Allele frequency distributions of SNPs used in the BAYPASS -simulated data were drawn from the input data set. To ensure that in the BAYPASS -simulated data sets, no singleton loci (loci where a minor allele is only seen once) were present, and a minimum minor allele frequency of 0.005 was applied. Furthermore in the BAYPASS -simulated data sets, correlations of SNPs with the environmental variables were set to zero, to ensure all SNPs behaved neutrally (Gautier, 2015b).

In OUTFLANK, outlier SNPs were identified using the default settings (Whitlock & Lotterhos, 2015b). Under these settings, OUTFLANK assesses if a locus is under selection by creating a chi-square distribution based on the distribution of F'_{ST} values. This chi-square distribution is used as a null distribution to calculate *p*-values for each

locus's likelihood of being an outlier. *Q*-values are then based on the right-tailed *p*-values (*q*-values are similar to *p*-values but are corrected for the false discovery rate). A SNP is defined as an outlier if it has a *q*-value of less than 0.05 (Whitlock and Lotterhos, 2015a). A SNP had to show a heterozygosity of greater than 10% to confirm its outlier status (Whitlock & Lotterhos, 2015b).

2.5 | Detection of false SNPs

Eight SNPs were triple positives and were considered putatively under selection in the Alpine ibex. To confirm these associations, we examined the alignments around the SNPs by eye. This demonstrated that seven of the SNPs were false variants caused by a misalignment of an insertion or deletion only in the libraries with a read length of 125 base pairs (SNP 1 to SNP 5) or 100 base pairs (SNP 8). This leads to a systematic difference in alignments across data batches and resulted in false SNPs.

2.6 | Preventing batch effects in selection analyses

To identify if including sequencing length as an environmental variable in BAYENV 2.0 and BAYPASS 2.1 can help prevent the observed batch effect, both programmes were rerun on a subset of the data using the same conditions, matrices and thresholds outlined above but with sequencing length as an environmental variable. Populations sequenced to 125 or 140 base pairs were given an environmental value of 125 or 140, respectively. Those sequenced to multiple lengths were given a missing environmental value (set to *NA*) due to the difficulty in determining the correct "environment" for these populations.

2.7 | Removing the false SNP calls

To identify what sequence data processing steps may help prevent the batch effect from arising, the use of different SNP filters, SNP callers, read trimming and alignment methods was examined. The SNP filter and SNP caller analyses involved the entire data set, while the read trimming and alignment method analyses used a subset of 21 individuals from three libraries (one from each batch).

To identify suitable SNP filters to remove the false SNPs, filters used in other studies were explored (Puritz et al., 2014). The quality metrics used included the following: (i) the number of reads placed to the left or right of the SNP call (denoted as RPR and RPL in FREEBAYES, a threshold of >1 read was applied; Garrison & Marth, 2012); (ii) the allelic balance which indicates the ratio of reads of the two alleles in heterozygote individuals (denoted as AB in FREEBAYES, a threshold of 0.25≤ and ≥0.75 and ≤0.01 was applied; Puritz et al., 2014); and (iii) the quality-to-depth ratio, which is the site quality score divided by the number of alternate allele observations (denoted as QUAL/AO in FREEBAYES, a threshold of >10 was applied; Garrison & Marth, 2012).

A different SNP caller was examined because local realignment as employed by GATK's IndelRealigner, or HaplotypeCaller can remove small misalignments like those causing the false SNPs in these data (McKenna et al., 2010). To explore the effect of SNP caller, SNPs were also identified using GATK's HaplotypeCaller (version 3.4-46, McKenna et al., 2010) instead of FREEBAYES. The same initial filtered read set that was used to call SNPs as in FREEBAYES. The minimum PHRED value to call a SNP was 30, and the minimum PHRED value to report a SNP was 10.

Upstream processing steps were also explored. Trimming reads to a uniform length before alignment would remove the largest cross-batch difference and thus may remove the batch effect. To explore this, reads were cropped to a maximum length of either 93 base pairs or 118 base pairs by TRIMMOMATIC (version 0.36) during the adaptor removal step. This corresponds to the read lengths in the 2013 and 2015a libraries when the inline barcodes are removed. Trimmed libraries were then realigned to the goat genome with BOWTIE2 (version 2.3.3.1); PCR duplicates were removed with MarkDuplicates (version 2.15.0), and SNPs were recalled for the regions housing the false SNPs with FREEBAYES (version v1.1.0-54-g49413aa) with parameters as described previously.

We aligned the Alpine ibex reads not to an ibex reference genome but to the genome of the closely related domestic goat (Dong et al., 2013). It is unclear whether this batch effect would have been less pronounced if an Alpine ibex reference genome had been available. Alignment to a reference genome of another species is known to introduce a number of biases, including bias towards regions that are slowly evolving (Schubert et al., 2012) and towards haplotypes that are most similar to the reference genome (Dilthey, Cox, Iqbal, Nelson, & McVean, 2015). Hence, it is conceivable that aligning to a divergent reference genome increased the risk of alignment errors and the batch effect identified. However, short insertions and deletions are common even within populations (Gudbjartsson et al., 2015); hence, the batch effect might have been equally pronounced even when using an Alpine ibex reference genome. Though no Alpine ibex genome is available, the use of a de novo contig assembly approach would circumvent the use of a divergent reference genome. To explore this, a de novo alignment was generated with the DDOCENT interactive pipeline (version 2.2.16, Puritz et al., 2014). We used eight individuals that were from the 2015b library, the library with the longest read length. The reference genome was generated considering unique sequences found with a coverage >4 (*k* value) and considering unique sequences found in more than two individuals (*c* value). For read trimming and alignment, the default conditions were used. To examine if the false SNPs were still present when aligning to the de novo genome, the contigs from the de novo genome were mapped against the goat genome using MUMMER (Delcher et al., 2002). This step was necessary to identify the contigs to which the reads containing the false SNPs were assembled and aligned to. SNPs were then called on the contigs that corresponded to the regions where the false SNPs were present on the goat genome using FREEBAYES.

3 | RESULTS

3.1 | Screening for signals of selection

Eight SNPs were triple positives and were considered putatively under selection in the Alpine ibex. Allele frequencies at five of the eight putatively selected SNPs were significantly associated with winter snow conditions (Figure 1, Tables 2 and 3). The nearest genes to these SNPs as found by SNPEFF included SENP1, NEDD1, MAD2L1, SH3TC1, CKMT2, RYR3, CADM1 and PRIMPOL. The SNPs identified in this way were not only statistically significant but also biologically plausible, because previous studies of cold and altitude adaptation in other species had highlighted the importance of some of the genes surrounding the eight SNPs putatively under selection (Cardona et al., 2014; Song et al., 2016; Wang et al., 2015; Zhang et al., 2014). However, in six of the putatively selected SNPs, the alternative allele was only detected in three populations, and one SNP was only identified in two populations. Examination of the alignments by eye suggested these seven SNPs were not true variants. Only the SNP that was present in nine populations (SNP 6, Table 2) was a true variant. The SNP calls at the seven false SNPs had resulted from the fact that several populations were only present in a single batch of sequencing and that the batches differed in read lengths. In individuals that were sequenced to longer read lengths (140 base pairs), an insertion or deletion relative to the reference genome was correctly identified a few base pairs away from the SNPs in question, leading to an absence of the alternative allele at the SNP site (Figure 2). In individuals with a shorter read length (125 base pairs or 100 base pairs) however, these insertions or deletions were misaligned, leading to the false SNP calls. Because read length differed systematically among populations, large apparent differences in SNP frequency between populations were identified (Table 2), leading to erroneous signals of selection.

The batch effect created strong differences in allele frequency among populations, but the false SNPs were not fixed within populations or individuals (Figure 1). Both alternative and reference alleles were present in each sequencing batch. This observation can be explained by the random shearing step combined with the use of "paired-end" sequencing. In this method of RADseq the forward reads, all have the same start point and length within a library; therefore, all forward reads were called for the false SNPs because they were all misaligned around the insertion or deletion. In contrast, the reverse reads have staggered starting points because of the random shearing. Therefore, subsets of the reverse reads in the short-read library (125 base pairs or 100 base pairs) aligned correctly across the insertion or deletion leading to the insertion or deletion being correctly called and the reference allele at the false SNP site. Overall, this leads to heterozygote SNP calls in some individuals.

3.2 | Preventing batch effects in selection analyses

To examine if including batch as a variable in a selection scan would identify the false SNPs, BAYENV 2.0 and BAYPASS 2.1 were rerun with sequencing length as an environmental variable. None of the false SNPs were significantly associated with the sequencing length in BAYPASS 2.0, despite the batch effect in the data. In BAYENV 2.1, 15 SNPs were associated with sequencing length, of which six were the false SNPs previously identified. This included SNP 1, SNP 2, SNP 3, SNP 4, SNP 5 and SNP 7 (Table 3).

3.3 | Removing false SNP calls

False SNPs such as the ones reported above can be removed with appropriate SNP filtering. In our data set, we found that certain filters based on quality metrics in FREEBAYES successfully removed the false variants. Though these filters removed all known false SNPs from the data set, true variants were also likely lost due to the increase in filter stringency. A filtering approach is advantageous over simple exclusion of the false SNPs because it is unlikely that the seven SNPs identified represent all the false calls in the data.

To examine the effect of the SNP caller on the batch effect, SNPs were recalled across all individuals using GATK's HaplotypeCaller. HaplotypeCaller did not call the false variants, except for SNP 2. SNP 2 was identified in the raw unfiltered SNP calls in six individuals and was always homozygous. The StrandsOddsRatio value, a GATK metric that identifies strand bias in SNP calls, was 6.9, well above the recommended threshold value of 3 (GATK Development Team, 2016). Thus, the default GATK filters removed this false SNP.

An alternative to additional SNP filtering would have been to trim all reads to the shortest read length (93 or 118 base pairs) among all batches. We tested this with a subset of our data. Trimming did help remove false variants from arising in the data that were trimmed to 93 base pairs because the reads were now too short to cover the false variant site. However, some false variants were found in reads that were trimmed to 118 base pairs due to our staggered read end and one base pair shifts at a small number of read starts. This amounted to three SNPs (SNPs 2, 3 and 5) that were present in 1–3 reads in up to three individuals.

Aligning to an ibex-specific de novo contig assembly also helped reduce false SNP calls due to batch effects. Only two regions surrounding the false SNPs were represented by the de novo contigs, and the false variants were not called on these two contigs because the sequences were correctly aligned across the indels.

4 | DISCUSSION

We identified a batch effect in RADseq data driven by differences in read length across libraries. Combining data from libraries sequenced to different lengths lead to misalignments of insertions/deletions and consequently to false SNPs and false signals of selection. This could easily have led to false biological conclusions, as biologically plausible genes were located in the neighbourhood of the false SNPs. That biologically plausible genes were identified near the false SNPs were likely due to chance, as biologically relevant associations surrounding

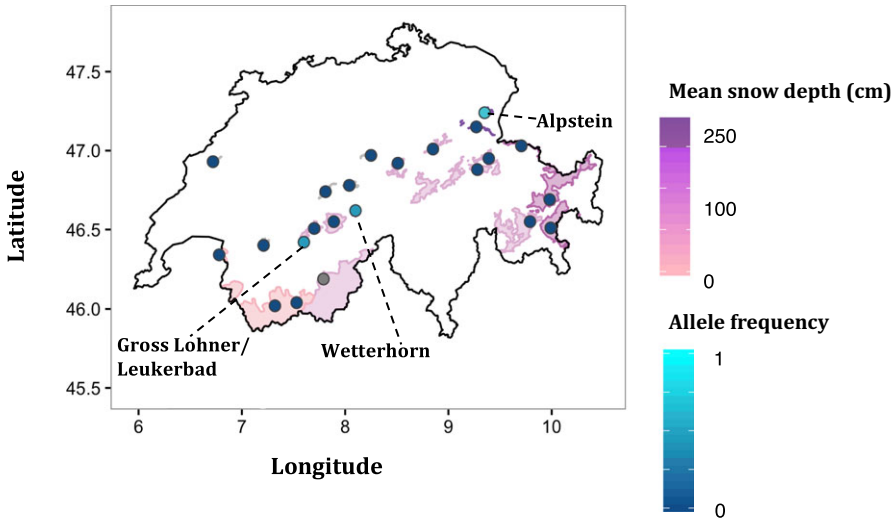

FIGURE 1 Allele frequencies across Switzerland of the false SNP 1, a significant outlier locus associated with winter snow depth. Circles represent the mean location of individuals sampled in each population. Grey indicates a single population where the SNP was not sampled because the corresponding DNA region was not sequenced. The irregular coloured areas represent the population range (shape) and snow depth (shading) as estimated from weather stations close to each population. Mean snow depth is the mean depth of snow on the ground (cm) from November to April. The SNP appeared to be polymorphic only in populations at Alpstein, Gross Lohner/Leukerbad and Wetterhorn, populations sequenced with a read length of 125 base pairs. Source of map data: Federal Office of Topography swisstopo. Weather data: MeteoSwiss. Ibex colony distribution: BAFU (2016)

focal SNPs have previously been shown to arise through chance alone (Pavlidis, Jensen, Stephan, & Stamatakis, 2012). Hence, as our case study illustrates, biological plausibility offers no protection from batch effects.

4.1 | Identifying batch effects

We initially identified the batch effect by examining the alignments surrounding our outlier SNPs by eye. To see if a more systematic approach might help pinpoint batch effects in tests for selection, we examined if including read length of a population as an environmental variable in the GEA analysis could identify the batch effect. No SNPs were significantly associated with read length in BAYPASS 2.1, despite the presence of the batch effect in the data. However, of the seven false SNPS, BAYENV 2.1 correctly identified six as being significantly associated with read length, but did not detect a batch effect on the remaining false SNP, SNP 8. This may have been the result of not being able to assign environmental values to populations with mixed read lengths (Table 2). Thus, including batch-related variables in downstream analysis can help to identify batch effects in the data, but such an approach is not sufficient, as it may miss batch effects, and it is limited to GEA-like analysis. Overall, several approaches (discussed below) must be used to identify if batch-related errors are present in a data set.

Studies that gather sequencing data over several years should look for alleles associated with specific batches of data, for instance, by testing for a significant association between allele frequencies and batch identity (The UK10K Consortium, 2015), and filter the data accordingly or change upstream bioinformatics steps like the SNP caller. In addition, batch effect detection and correction tools that have been designed for microarray data (Johnson, Li, & Rabinovic, 2007; Leek, Johnson, Parker, Jaffe, & Storey, 2012; Manimaran et al., 2016), and SNP data (e.g., batchTest in GWASTOOLS; Gogarten et al., 2012), can be employed. Together with other SNP quality checks, for example examining heterozygosity or *F*-statistics (Shafer et al., 2016), this will allow the detection of most obvious batch effects. Furthermore, long-term studies may benefit from rerunning a subset of samples when new data are added to a study, to help identify and quantify batch effects or errors. Alternatively, in situations where samples are limited or DNA may deteriorate over time, positive controls containing synthetic sequences with known variants or standard genomes similar to the positive controls used in metagenomics could also be employed (Miller et al., 2016). Finally, reporting of the most common sources of batch effects, that is differences in read length, sequencing technology, or sequencing centre should become standard procedure so that this information is available to data analysts (Leek et al., 2010).

4.2 | Removing false SNP calls

Previous studies of batch effects have recommended the randomization of samples across technical variables to prevent spurious observations (Buhule et al., 2014). If the Alpine ibex data had been fully randomized across the libraries of different read lengths, SNP calls would still have been wrong, but the false signals of selection would likely not have arisen. However, fully randomized study designs will often not be possible in HTS studies that extend over several years due to changes in sequencing technologies over time. Thus,

TABLE 2 Allele frequencies of outlier SNPs

SNP	Population	SNP position	Outlier type	Allele frequency	Library length (bp)
SNP 1	Alpstein	Chr 5: 31085709	F_{ST}-like and GEA	0.79	125
SNP 1	Gross Lohner/Leukerbad	Chr 5: 31085709	F_{ST}-like and GEA	0.5	125
SNP 1	Wetterhorn	Chr 5: 31085709	F_{ST}-like and GEA	0.5	125
SNP 2	Alpstein	Chr 5: 59336872	F_{ST}-like and GEA	0.75	125
SNP 2	Gross Lohner/Leukerbad	Chr 5: 59336872	F_{ST}-like and GEA	0.71	125
SNP 2	Wetterhorn	Chr 5: 59336872	F_{ST}-like and GEA	0.71	125
SNP 3	Alpstein	Chr 27: 13975771	F_{ST}-like and GEA	0.69	125
SNP 3	Gross Lohner/Leukerbad	Chr 27: 13975771	F_{ST}-like and GEA	0.67	125
SNP 3	Wetterhorn	Chr 27: 13975771	F_{ST}-like and GEA	0.5	125
SNP 4	Alpstein	Chr 7: 77330268	F_{ST}-like and GEA	0.75	125
SNP 4	Gross Lohner/Leukerbad	Chr 7: 77330268	F_{ST}-like and GEA	0.81	125
SNP 4	Wetterhorn	Chr 7: 77330268	F_{ST}-like and GEA	0.75	125
SNP 5	Alpstein	Chr 10: 25296960	F_{ST}-like and GEA	0.55	125
SNP 5	Gross Lohner/Leukerbad	Chr 10: 25296960	F_{ST}-like and GEA	0.67	125
SNP 5	Wetterhorn	Chr 10: 25296960	F_{ST}-like and GEA	0.63	125
SNP 6	**Arolla**	**Chr 15: 25267019**	**F_{ST}-like**	**0.06**	**140**
SNP 6	**Bire-Oeschinen**	**Chr 15: 25267019**	**F_{ST}-like**	**0.86**	**140**
SNP 6	**Brienzer-Rothorn**	**Chr 15: 25267019**	**F_{ST}-like**	**0.13**	**100** and **140**
SNP 6	**Creux du Van**	**Chr 15: 25267019**	**F_{ST}-like**	**0.1**	**140**
SNP 6	**Fluebrig**	**Chr 15: 25267019**	**F_{ST}-like**	**0.05**	**140**
SNP 6	**Justistal**	**Chr 15: 25267019**	**F_{ST}-like**	**0.3**	**140**
SNP 6	**Mont Pleureur**	**Chr 15: 25267019**	**F_{ST}-like**	**0.25**	**100** and **140**
SNP 6	**Schwarzmönch**	**Chr 15: 25267019**	**F_{ST}-like**	**0.38**	**100** and **140**
SNP 6	**Wittenberg**	**Chr 15: 25267019**	**F_{ST}-like**	**0.71**	**140**
SNP 7	Alpstein	Chr 6: 4799392	F_{ST}-like	0.56	125
SNP 7	Gross Lohner/Leukerbad	Chr 6: 4799392	F_{ST}-like	0.63	125
SNP 7	Wetterhorn	Chr 6: 4799392	F_{ST}-like	0.64	125
SNP 8	Albris	Chr 6: 111314676	F_{ST}-like	0.22	100 and 140
SNP 8	Weisshorn	Chr 6: 111314676	F_{ST}-like	1	100

Populations not shown are fixed for the reference allele. All loci were identified as putatively under selection by BAYENV 2.0, BAYPASS 2.1, and OUTFLANK. F_{ST}-like indicates loci identified using F_{ST}-based approaches, GEA indicates loci identified with genetic-environment associations. Bold indicates the one true SNP, all other SNPs are false and arose due to batch effects.

TABLE 3 Significant associations between the triple positive SNPs and environmental variables as found by two different selection detection softwares, BAYENV 2.0 and BAYPASS 2.1

Environmental variable (cm)	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 7
New snow accumulation (winter)		Be				
Snow depth (winter)	Be	BeBp	BeBp	BeBp	BeBp	
Snow depth (summer)			Be	Be		
Sequencing length	Be	Be	Be	Be	Be	Be

completely controlling for batch effects before sequencing may often be impossible in long-term studies and a bioinformatic control of batch effects will be necessary.

Various bioinformatics steps helped to remove the batch effect reported here. Firstly, using more stringent SNP filters removed the false SNP calls, but presumably at the cost of removing many true SNPs. A second, alternative approach was to employ a different SNP caller. Calling variants with GATK's HaplotypeCaller did lead to correct SNP calls. This indicates that is important to not only evaluate SNP filters but also SNP callers in projects that combine data from multiple sequencing runs. Further approaches to remove the batch effects, including trimming reads to a uniform length and aligning to an ibex-specific de novo contig assembly, were less efficient. Trimming did not remove all false variants and must be balanced with the loss of information incurred through reducing read length. Aligning to a de novo contig assembly also introduced new limitations because of missing genomic regions. Consequently, trimming and

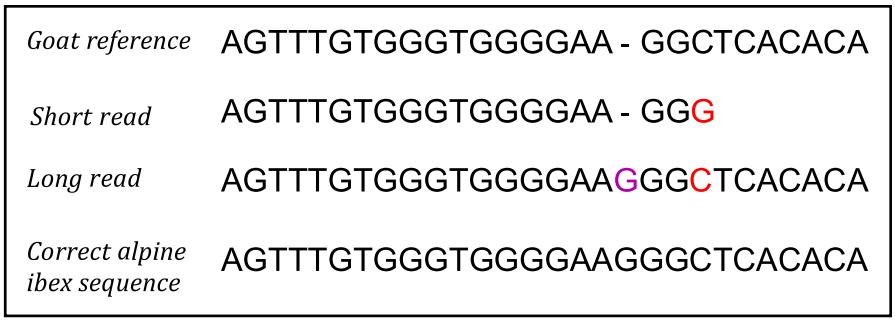

FIGURE 2 An example of a technical artefact that can lead to false SNP calls. The red G/C nucleotides represent a false SNP caused by an alignment error that occurred in the shorter read length libraries. In this example, the false SNP falls at the end of a short read. In the short read, a G insertion relative to the reference genome has been misaligned and was not identified (dashed line), which caused this spurious SNP. The longer read is correctly aligned, and an insertion relative to the goat genome sequence (purple G) is present in the Alpine ibex. Due to the mixing of libraries of different read lengths, the alignment error on short reads created a false polymorphic SNP in the Alpine ibex. The dash in the goat reference sequence represents the location of the G insertion that is found in the correct Alpine ibex sequence

alignment to a de novo contig assembly alone may be insufficient to ensure complete false SNP removal.

This study reports the occurrence of a batch effect in RADseq data that led to biologically plausible yet erroneous signals of selection. The study highlights the need for careful control of study design, bioinformatics pipelines and data analysis to prevent batch effects. Only a combination of approaches will ensure that biologically spurious conclusions from batch effects in HTS data are kept to a minimum.

ACKNOWLEDGEMENTS

We thank G. Camenisch for his help and support and A. Wagner, U. Tobler and three anonymous reviewers for comments that greatly improved this manuscript. We are also grateful to J. Pemberton for recommending our submission to this special issue. We are grateful for the laboratory and sequencing support from the Genetic Diversity Center, Fucntional Genomics Center, University of Zurich and the Genomics Facility Basel, ETH Zurich and the funding from the University of Zürich's Research Prioirity Program "Evolution in Action". The RADseq technologies used in this study are covered by patents owned by Keygene N.V., which protect its Sequence Based Genotyping technologies.

DATA ACCESSIBILITY

The eight outlier SNPs, including the seven false variants, can be viewed in the vcf file on Dryad. The script by HEL Lischer on removing variants that were fixed in the Alpine ibex but were polymorphic compared to the *Capra hircus* reference genome, representing fixed species differences, can be accessed on Dryad. The individuals, corresponding populations, and the length they were sequenced to can also be viewed on Dryad: http://datadryad.org/review?doi=doi:10.5061/dryad.8vm8d. Read data can be viewed on the short-read archive, ncbi project number PRJNA422727: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA422727.

AUTHOR CONTRIBUTIONS

DML created the RADseq libraries from 2015, analysed the data and co-wrote the paper. HELL offered bioinformatics supervision and support for the NGS data analysis. CG created the RADseq libraries in 2012 and identified and helped confirm the batch effect. LFK supervised the work and co-wrote the manuscript. All authors commented on the manuscript.

ORCID

D. M. Leigh http://orcid.org/0000-0003-3902-2568

REFERENCES

Ahn, S. J., Costa, J., & Emanuel, J. R. (1996). PicoGreen quantitation of DNA: Effective evaluation of samples pre- or post-PCR. *Nucleic Acids Research*, *24*, 2623–2625. https://doi.org/10.1093/nar/24.13.2623

BAFU (2015) Abschuss Steinbock, ganze Schweiz: 2000–2015. Retrieved from: www.wild.uzh.ch/jagdst/index.php

BAFU (2016) Ibex Colonies. Retrieved from: ch.bafu.fauna-steinbockkolonien

Biebach, I., & Keller, L. F. (2009). A strong genetic footprint of the reintroduction history of Alpine ibex (*Capra ibex ibex*). *Molecular Ecology*, *18*, 5046–5058. https://doi.org/10.1111/j.1365-294X.2009.04420.x

Birzele, F., Schaub, J., Rust, W., Clemens, C., Baum, P., Kaufmann, H., … Hildebrandt, T. (2010). Into the unknown: Expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Research*, *38*(12), 3999–4010. https://doi.org/10.1093/nar/gkq116

Blair, L. M., Granka, J. M., & Feldman, M. W. (2014). On the stability of the BAYENV method in assessing human SNP-environment associations. *Human Genomics*, *8*, 1–13. https://doi.org/10.1186/1479-7364-8-1

Bolger, A. M., Lohse, M., & Usadel, B. (2014). TRIMMOMATIC: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Broad (2016) Picard. Retrieved from: http://broadinstitute.github.io/picard.

Buhule, O. D., Minster, R. L., Hawley, N. L., Medvedovic, M., Sun, G., Viali, S., ... Weeks, D. E. (2014). Stratified randomization controls better for batch effects in 450K methylation analysis: A cautionary tale. *Frontiers in Genetics*, *5*, 1–11. https://doi.org/10.3389/fgene.2014.00354

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. A., Yngvadottir, B., ... Kivisild, T. (2014). Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS ONE*, *9*(5), e98076. https://doi.org/10.1371/journal.pone.0098076

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SNPEFF: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*, 80–92. https://doi.org/10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... McVean, G. (2011). The variant call format and VCFTOOLS. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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

Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, *30*, 2478–2483. https://doi.org/10.1093/nar/30.11.2478

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688. https://doi.org/10.1038/ng.3257

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*, 135–141. https://doi.org/10.1038/nbt.2478

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., & Cresko, W. A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In V. Orgogozo & M. V. Rockman (Eds.), *Methods in molecular biology: Molecular methods for evolutionary genetics*, Vol. *772* (pp. 157–178). New York, NY: Humana Press.

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, *8*(11), 14–17. https://doi.org/10.1371/journal.pone.0079667

Garrison, E. (2016). VCFLIB: A C++ library for parsing and manipulating VCF files. Retrieved from: https://github.com/vcflib/vcflib

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1–9

GATK Development Team (2016) Understanding and adapting the generic hard-filtering recommendations Doc #6925. https://software.broadinstitute.org/gatk/documentation/article?id=6925 Accessed: 17/12/17

Gautier, M. (2015a). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*, 1555–1579. https://doi.org/10.1534/genetics.115.181453

Gautier, M. (2015b). BAYPASS version 2.1 User Manual. Retrieved from: http://www1.montpellier.inra.fr/CBGP/software/baypass/index.html

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., ... Laurie, C. C. (2012). "GWASTOOLS: An R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, *28*(24), 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F., & Croll, D. (2017). Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evolutionary Applications*, *11*(2), 123–139. https://doi.org/10.1111/eva.12490

Grossen, C., Keller, L., Biebach, I., & Croll, D. (2014). Introgression from domestic goat generated variation at the major histocompatibility complex of Alpine ibex. *PLoS Geneteics*, *10*, e1004438. https://doi.org/10.1371/journal.pgen.1004438

Grøtan, V., Saether, B. E., Filli, F., & Engen, S. (2007). Effects of climate on population fluctuations of ibex. *Global Change Biology*, *14*, 218–228. https://doi.org/10.1111/j.1365-2486.2007.01484

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., ... Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, *188*(4), 379–397. https://doi.org/10.1086/688018

Jacobson, A. R., Provenzale, A., Von Hardenberg, A., Bassano, B., & Festa-Bianchet, M. (2004). Climate forcing and density dependence in a mountain ungulate population. *Ecology*, *85*, 1598–1610.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with BOWTIE2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.192

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883. https://doi.org/10.1093/bioinformatics/bts034

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739. https://doi.org/10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, *3*(9), 1724–1735. https://doi.org/10.1371/journal.pgen.0030161

Li, H. (2014). Toward better understanding of artefacts in variant calling from high-coverage samples. *Bioinformatics*, *30*, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, *23*, 2178–2192. https://doi.org/10.1111/mec.12725

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. https://doi.org/10.1111/1755-0998.12635

Manimaran, S., Selby, H. M., Okrah, K., Ruberman, C., Leek, J. T., Quackenbush, J., ... Johnson, W. E. (2016). BATCHQC: Interactive software for evaluating sample and batch effects in genomic data. *Bioinformatics*, *32*(24), 3836–3838. https://doi.org/10.1093/bioinformatics/btw538

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*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Mckinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, *17*, 356–361. https://doi.org/10.1111/1755-0998.12649

Miller, R. R., Uyaguari-Diaz, M., McCabe, M. N., Montoya, V., Gardy, J. L., Parker, S., … Patrick, D. (2016). Metagenomic investigation of plasma in individuals with ME/CFS highlights the importance of technical controls to elucidate contamination and batch effects. *PLoS ONE*, *11* (11), e0165691. https://doi.org/10.1371/journal.pone.0165691

Müller, C., Schillert, A., Röthemeier, C., Trégouët, D. A., Proust, C., Binder, H., … Ziegler, A. (2016). Removing batch effects from longitudinal gene expression - Quantile normalization plus comBat as best approach for microarray transcriptome data. *PLoS ONE*, *11*(6), 1–23. https://doi.org/10.1371/journal.pone.0156594

Nadeau, S., Meirmans, P. G., Aitken, S. N., Ritland, K., & Isabel, N. (2016). The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. *Ecology and Evolution*, *6*(24), 8649–8664. https://doi.org/10.1002/ece3.2550

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29* (10), 3237–3248. https://doi.org/10.1093/molbev/mss13

Pearson, W. R., Wood, T., Zhang, Z., & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics*, *36*, 24–36. https://doi.org/10.1006/geno.1997.4995

Poptsova, M. S., Il'icheva, I. A., Nechipurenko, D. Y., Panchenko, L. A., Khodikov, M. V., Oparina, N. Y., … Grokhovsky, S. L. (2014). Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports*, *4*(1), 4532. https://doi.org/10.1038/srep04532

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). DDOCENT: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431. https://doi.org/10.7717/peerj.431

Ratan, A., Zhang, Y., Hayes, V. M., Schuster, S. C., & Miller, W. (2010). Calling SNPs without a reference sequence. *BMC Bioinformatics*, *11* (1), 130. https://doi.org/10.1186/1471-2105-11-130

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Rasheid, K. A., Willerslev, E., … Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*(1), 178. https://doi.org/10.1186/1471-2164-13-178

Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Dworkis, D. A., … Perls, T. T. (2011). Retraction. *Science*, *333*, 404. https://doi.org/10.1126/science.1190532

Shackleton, D., & Group, ISCS. (1997). *Wild sheep and goats and their relatives. Status survey and conservation action plan for Caprinae*. Gland, Switzerland: IUCN.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917. https://doi.org/10.1111/2041-210X.12700

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

Song, S., Yao, N., Yang, M., Liu, X., Dong, K., Zhao, Q., …Ma, Y. (2016). Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the Tibetan cashmere goat (*Capra hircus*). *BMC Genomics*, *17*, 122. https://doi.org/10.1186/s12864-016-2449-0

Stüwe, M., & Grodinsky, C. (1987). Reproductive biology of captive Alpine ibex (*Capra i. ibex*). *Zoo Biology*, *6*, 331–339. https://doi.org/10.1002/zoo.1430060407

Stüwe, M., & Nievergelt, B. (1991). Recovery of Alpine ibex from near extinction – The result of effective protection, captive breeding, and reintroductions. *Applied Animal Behaviour Science*, *29*, 379–387. https://doi.org/10.1016/0168-1591(91)90262-V

Taub, M. A., Corrada Bravo, H., & Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, *2*(12), 87. https://doi.org/10.1186/gm208

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*, 82. https://doi.org/10.1038/nature14962

Wang, K., Yang, Y., Wang, L., Ma, T., Shang, H., Ding, L., … Qiu, Q. (2015). Different gene expressions between cattle and yak provide insights into high-altitude adaptation. *Animal Genetics*, *47*, 28–35. https://doi.org/10.1111/age.12377

Whitlock, M. C., & Lotterhos, K. E. (2015a). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F_{ST}. *The American Naturalist*, *186*(S1), S24–S36. https://doi.org/10.1086/682949

Whitlock, M. C., & Lotterhos, K. E. (2015b). OUTFLANK finding F_{ST} outliers with an inferred neutral distribution. https://github.com/whitlock/OutFLANK

Xu, L., Bickhart, D. M., Cole, J. B., Schroeder, S. G., Song, J., Van Tassell, C. P., … Liu, G. E. (2015). Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Molecular Biology and Evolution*, *32*(3), 711–725. https://doi.org/10.1093/molbev/msu333

Yu, S. G., Chu, W. W., Zhang, L. F., Han, H. M., Zhao, R. X., Wu, W., … Chen, J. (2015). Identification of laying-related SNP markers in geese using RAD sequencing. *PLoS ONE*, *10*(7), 1–19. https://doi.org/10.1371/journal.pone.0131572

Zhang, W., Fan, Z., Han, E., Hou, R., Zhang, L., Galaverni, M., … Zhang, Z. (2014). Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLoS Genetics*, *10*(7), e1004466. https://doi.org/10.1371/journal.pgen.1004466

How to cite this article: Leigh DM, Lischer HEL, Grossen C, Keller LF. Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Mol Ecol Resour*. 2018;18:778–788. https://doi.org/10.1111/1755-0998.12779