

This is a repository copy of *Identification and characterisation of endogenous Avian Leukosis Virus subgroup E (ALVE) insertions in chicken whole genome sequencing data.*

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/163045/>

Version: Published Version

Article:

Mason, Andrew Stephen orcid.org/0000-0002-8222-3974, Lund, Ashlee R, Hocking, Paul M et al. (2 more authors) (2020) Identification and characterisation of endogenous Avian Leukosis Virus subgroup E (ALVE) insertions in chicken whole genome sequencing data. *Mobile DNA*. 22. ISSN 1759-8753

<https://doi.org/10.1186/s13100-020-00216-w>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown


If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

RESEARCH

Open Access



Identification and characterisation of endogenous Avian Leukosis Virus subgroup E (ALVE) insertions in chicken whole genome sequencing data

Andrew S. Mason^{1,2*} , Ashlee R. Lund³, Paul M. Hocking^{1^}, Janet E. Fulton^{3†} and David W. Burt^{1,4†}

Abstract

Background: Endogenous retroviruses (ERVs) are the remnants of retroviral infections which can elicit prolonged genomic and immunological stress on their host organism. In chickens, endogenous Avian Leukosis Virus subgroup E (ALVE) expression has been associated with reductions in muscle growth rate and egg production, as well as providing the potential for novel recombinant viruses. However, ALVEs can remain in commercial stock due to their incomplete identification and association with desirable traits, such as ALVE21 and slow feathering. The availability of whole genome sequencing (WGS) data facilitates high-throughput identification and characterisation of these retroviral remnants.

Results: We have developed obsERVer, a new bioinformatic ERV identification pipeline which can identify ALVEs in WGS data without further sequencing. With this pipeline, 20 ALVEs were identified across eight elite layer lines from Hy-Line International, including four novel integrations and characterisation of a fast feathered phenotypic revertant that still contained ALVE21. These bioinformatically detected sites were subsequently validated using new high-throughput KASP assays, which showed that obsERVer was highly precise and exhibited a 0% false discovery rate. A further fifty-seven diverse chicken WGS datasets were analysed for their ALVE content, identifying a total of 322 integration sites, over 80% of which were novel. Like exogenous ALV, ALVEs show site preference for proximity to protein-coding genes, but also exhibit signs of selection against deleterious integrations within genes.

Conclusions: obsERVer is a highly precise and broadly applicable pipeline for identifying retroviral integrations in WGS data. ALVE identification in commercial layers has aided development of high-throughput diagnostic assays which will aid ALVE management, with the aim to eventually eradicate ALVEs from high performance lines. Analysis of non-commercial chicken datasets with obsERVer has revealed broad ALVE diversity and facilitates the study of the biological effects of these ERVs in wild and domesticated populations.

Keywords: ALVE, Ev gene, Avian Leukosis virus, Endogenous retrovirus, obsERVer, Chicken

* Correspondence: andrew.mason@york.ac.uk

[^]Paul M. Hocking is deceased.

[†]Janet E Fulton and David W Burt Equal senior authorship

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

²York Biomedical Research Institute, The Department of Biology, The University of York, York YO10 5DD, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Retroviruses present persistent and unique challenges to the vertebrate species they infect. Exogenous retroviruses typically evolve at rates up to six orders of magnitude faster than their hosts and transfer novel accessory genes horizontally from other vertebrates, their parasites or other viruses [1–4]. Retroviruses also impact long term genome evolution as endogenous retroviruses (ERVs) following integration into the germline. ERVs represent a ‘fossil record’ of previous retroviral infections, and constitute approximately 3% of the avian genome [5, 6]. Identifiable avian ERVs include lineage-specific evolutionary ‘recent’ loci, integrations found across birds, and ancient sites also seen in mammalian genomes [5, 6].

ERVs decay, or are epigenetically silenced, over long evolutionary timescales [2, 3, 7, 8]. However, recurrent infection and intracellular retrotransposition can generate new, structurally intact ERVs with the potential to impact gene expression, facilitate chromosomal rearrangements and modulate host response to retroviral infections [4, 9–14]. ERVs may also continue to express their own retroviral genes (*gag*, *pol*, and *env*), driven by promoters with the potential for bidirectional effects in the flanking long terminal repeats (LTRs). Furthermore, they can re-emerge from the genome by recombination to pose novel exogenous threats, such as Avian Leukosis Virus (ALV) subgroup J in chickens [1, 4, 11, 13, 15, 16].

ALV is an alpharetrovirus which infects galliform birds, and is the only known chicken (*Gallus gallus*) retrovirus with both exogenous and endogenous activity [13, 17]. Exogenous ALVs are generally slow-transforming viruses which induce lymphoid (subgroups A–D) or myeloid (subgroup J) tumour formation over weeks or months via insertional mutagenesis, and can spread through flocks via viral shedding [18, 19]. Endogenous ALVs can also infect horizontally but have a species-specific range. Subgroup E (the ALVEs; historically known as *ev* genes) are found in the domestic chicken and its wild progenitor, the red jungle fowl (RJF), but in no other *Gallus* species [15, 20].

As evolutionarily recent integrations, ALVEs are typically found at low copy number in the genome, but often retain high structural integrity [21–23]. The presence of replication-competent ALVEs, and ALVEs which express *gag* proteins, has been associated with impacts on traits including reductions in muscle growth rate, egg number, size and shell thickness, and an increased incidence of viral shedding [24–28]. Conversely, *env* expression can mediate ALV infection through receptor interference [29–31]. As genomic ALVE content increases, their cumulative influence becomes increasingly complex, particularly when lines are interbred. Furthermore, recent work has shown interactions between ALVEs and the increasingly virulent Marek’s Disease Virus (MDV), a chicken-specific alphaherpesvirus. MDV can induce

ALVE expression, even of normally silenced elements such as ALVE1 [32, 33], and MDV vaccines can induce higher incidence of spontaneous lymphoid tumours in lines containing ALVE21 [34–36].

Despite these effects, and the creation of ‘ALVE-free’ lines [37, 38], the commercial poultry breeding community has been unable to completely remove all ALVEs from breeding stock. This has been a combination of the inability to detect all ALVE insertions, the association between some ALVEs and commercially desirable traits (such as ALVE21 with slow feathering [39–43], and ALVE-TYR with white plumage [44, 45]), and managing selection programmes for multiple performance traits. Crucially, commercial breeding stock must be negative for exogenous ALV before shipment. This necessitates continual testing for the ALV-specific antigen p27, which generates an effective false positive when endogenous expression is detected. Gel-based PCR assays were developed to detect ALVEs common in layers [21], but such assays are uneconomical at commercial scales and were limited to known sites. Traditional ALVE identification methods utilised characteristic restriction fragment length polymorphisms (RFLPs), but these patterns were poorly conserved between breeds, and can be difficult to interpret when there are high ALVE numbers per bird [46–48]. Comprehensive ALVE identification and characterisation within commercial lines is therefore essential for improvements in both productivity and health monitoring, but any assays must be cost effective at commercial scales.

Beyond commercial chickens, little is known about ALVE diversity in other domesticated or wild chicken populations. The reference RJF assembly contains two ALVEs, one of which is commonly shared with commercial layers and broilers [21, 49, 50]. Short read, whole genome sequencing (WGS) datasets have been generated for many different chicken lines, breeds and populations (including wild-caught RJF) over the last decade, and present an opportunity for the identification of ALVEs. However, identification of such structural variants has been hindered by complex read mapping, incomplete reference genome assemblies and limited sequence coverage at insertion junction sites. Recent work has used target capture sequencing to enrich for repetitive DNA, including ALVEs [23, 51], to promote identification of novel integrations. However, these datasets have limited future applications, and most WGS data has not been mined fully for structural variants, including ERVs.

Here, we describe obsERVer, a bioinformatic pipeline developed for the detection of specific, user-determined ERV integrations in WGS datasets. We validated the bioinformatic predictions in eight elite layer lines with new high-throughput diagnostic assays for each

identified ALVE, which we then used to genotype the originally sequenced birds and over 9000 archived DNA samples from the same elite layer lines. We also utilised obsERVer to identify ALVEs in fifty-seven diverse chicken datasets encompassing commercial, experimental and heritage layers and broilers, native breeds, and wild populations, including RJF. This work has enabled a better understanding of ALVEs in both commercial and more diverse chicken lines including the identification of over 260 novel ALVE loci. In addition, these methods provide new opportunities to examine the biological effects of ALVEs in diverse domesticated and wild chicken populations.

Methods

Whole genome sequencing datasets

A total of sixty-five Illumina paired-end 101 base pair (bp) chicken WGS datasets were surveyed for ALVE integrations. Datasets were available either from public repositories or kindly shared by collaborators, derived from individual birds or multiple individual pools as indicated in Table S1. These datasets included: experimental White Leghorn (WL), Brown Leghorn (BL) and Rhode Island White (RIW) lines; commercial layers (WL, White Plymouth Rock (WPR) and Rhode Island Red (RIR)); heritage broilers; native breeds and village populations from Asia and Africa; and wild-caught RJF from China, Java and Sumatra.

All sequencing reads were quality checked by FastQC v0.11.2 [52]. TrimGalore v0.4.0 [53] and Cutadapt v1.4 [54] were used to remove sequencing adapters and trim reads where base quality dropped below 20 in a 4 bp sliding window, removing reads trimmed by more than half their length. Each dataset was aligned to the *Gallus gallus* 5.0 reference genome (Galgal5; GenBank: GCF_000002315.4) using BWA-mem v0.7.10 [55], and coverage calculated with samtools v0.1.19 mpileup [56].

Bioinformatic detection of ALVE integration sites

For detection of ALVE integrations, sequencing reads from each chicken WGS dataset were remapped to an “ALVE pseudochromosome” constructed of eleven publicly available ALV sequences (Table S2), each separated by 1 kilobase pair (kbp) of Ns (ambiguous bases). Reads which mapped to this ALVE pseudochromosome (and their read mates) were subtracted from the original FASTQ files for each dataset and then re-mapped against the Galgal5 reference genome to identify putative integration sites. Mapping quality [56] greater than 20 was required in both cases. This approach facilitated targeted analysis of user-defined integrations and reduced computational burden by first aligning to a limited reference sequence set rather than the entire genome, as in other recent analogous approaches [57, 58].

Assembled endogenous alpharetroviral sequences were detected by BLASTn [59] using publicly available ALV, EAV (endogenous avian virus) and ART-CH (avian retrotransposon of chicken) sequences (Table S2) and used to filter the list of putative integration sites. Remaining sites were filtered on the presence of split reads, where part of the read aligned to the reference genome and the remainder to an unassembled ALVE, precisely defining the integration site.

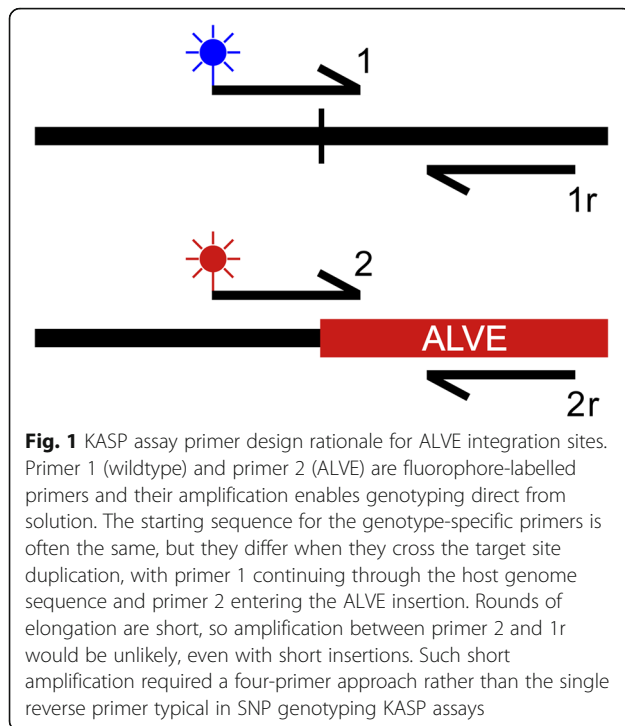
These scripts are contained within a single bioinformatics pipeline, obsERVer (Figure S1), available on GitHub (<https://github.com/andrewstephenmason/obsERVer>).

Putative ALVE integration sites were inspected in IGV Desktop for Windows v2.3.60 [60] and named using existing ALVE nomenclature where possible (Table S3). New names were given to novel insertions following the format “ALVE_ros001”, and to previously identified ALVEs with ambiguous names. Presence of the reference-genome-assembled ALVE-JFevA (ALVE6) and ALVE-JFevB integrations was checked manually in the full BAM files. ALVE orientation, terminal sequence integrity, and target site duplication (TSD) was identified using the split reads at each site. ALVEs which had integrated within other repeat elements were identified using RepeatMasker v4.0.3 [61].

Detection of ALVEs by obsERVer was developed using WGS datasets from eight Hy-Line (HL) elite layer lines (five independent WLs, sister WPRs, and one RIR), each derived from a pool of ten birds [62]. Following validation of the bioinformatically detected ALVE integrations (see below), obsERVer was implemented on the remaining WGS datasets.

Validation of ALVEs detected by obsERVer

Due to the complexities of structural variant detection from short read sequencing data, we assessed the number of false positives detected by obsERVer by validating the ALVEs detected in the HL WGS datasets. These datasets were the focus of obsERVer development and validation due to the availability of same-line DNA, including those from the original sequencing pools. Diagnostic assays were developed to validate each bioinformatically identified ALVE, as well as the *K* locus duplication site associated with ALVE21. Sequence from the ALVE-homologous split reads and the flanking genome at each putative integration site was used to develop KASP™ (Kompetitive Allele-Specific PCR; LGC, UK) assays using a four-primer approach (Fig. 1). Briefly, KASP assays use allele specific primers with fluorescent tags to directly genotype alternative alleles; enabling high-throughput detection of variants in a highly automated process. Initially designed for single nucleotide polymorphism (SNP)-genotyping, we have adapted the system to detect ALVE integrations.



Primers were designed with Kraken™ Primer Picker software (LGC, UK) with lengths of 20–25 bases and equal primer GC content in the 40–50% GC range (Table S4). Reactions used dehydrated DNA samples, primers and the LGC KASP™ 2x Mastermix V4.0 1536 formulation, following the original KBiosciences KASP™ protocol: a 61 °C to 55 °C touchdown for ten cycles and then 55 °C for twenty further cycles in a total reaction volume of 1 µl. Allele-specific fluorescence was detected using the PHERAstar Plus SNP plate reader and genotypes determined using Kraken™. For those ALVEs which had a previously reported PCR based test, we unambiguously validated genotype results between the KASP detection and gel-based detection method, thus confirming that we were detecting the same ALVE. All developed assays were used to genotype the original eighty individuals included in the sequencing pools (Table S1), as well as over 9000 banked samples from multiple generations of the eight sequenced lines.

In addition, traditional gel-based PCR assays were developed for those HL ALVEs that did not have previously published gel-based assays (Table S5). Primers were designed using Primer3 v2.3.7 [63].

Modelling obsERVer detection sensitivity

The likelihood of missing ALVEs by chance from each HL dataset was modelled for all genotype frequencies given the pool size of ten individuals used for sequencing, total line population size, and the average observed

genome coverage (11–18X). Mapping error rates were included based on the proportion of unmapped reads in each dataset, but the ‘sequenceability’ of the genome was not included due to limited literature on the estimation of this value [64–66]. For a given allele frequency, the model was run as follows: 1) a flock of given size was randomly assigned genotypes based on Hardy-Weinberg Equilibrium, 2) individuals and alleles were sampled according to a binomial distribution, and 3) then scaled for genome-average coverage and Poisson-varied coverage, both of which were further scaled by an underlying error rate. Models were run one million times. Probabilities were calculated for each sequenced HL line, for each possible insertion allele frequency. Probabilities were calculated for the HL ALVEs detected by KASP assay but not obsERVer, using the non-Poisson-scaled probabilities as site coverage was known.

Sanger sequencing of ALVE integrations

KASP assays were used to identify individuals homozygous for each HL ALVE. ALVEs were amplified using the Takara PrimeSTAR® GXL DNA polymerase kit with the flanking genomic primers developed for the gel-based PCRs (Table S5). Amplification followed the standard Takara protocol with eight-minute extensions in each cycle.

DNA was extracted from excised gel bands using the Invitrogen PureLink™ Quick Gel Extraction kit. DNA from ALVE inserts larger than 1kbp was cloned into the Invitrogen ZeroBlunt TOPO pCR®4 Blunt-TOPO® vector and used to transform One Shot® Mach1™-T1^R Competent *E. coli* cells. Plasmids were extracted using the Invitrogen PureLink™ Quick Plasmid Miniprep kit following the manufacturer’s instructions. Purified DNA from ALVE inserts shorter than 1kbp was not cloned, but cleaned after PCR with the ExoSAP protocol.

Purified ALVE DNA was amplified for sequencing using the Applied Biosystems BigDye Terminator v3.1 Cycle Sequencing kit and Sanger sequenced at Edinburgh Genomics (University of Edinburgh, UK). Full-length ALVE insertions required eighteen sequencing reactions, using primers designed from the ALVE1 reference sequence (GenBank: AY013303.1) spaced every 500 bp (Table S6; Figure S2). Consensus sequences were built using the Geneious v7.0.4 [67] ‘Map to Reference’ tool, ALVE domains and SNPs were annotated, and LTR pairs were aligned by MUSCLE v3.8.31 [68]. ORFs were identified in each sequence by GLIMMER3 [69], transcription factor binding sites annotated by the EMBOSS v6.6.0 [70] tfscan tool, and the EMBOSS fuzznuc tool was used to identify the miR-155 AGCATT A target sequence in the ALVE *envelope* domain [71].

High resolution optical mapping of the K locus

Whole blood samples were collected and stabilised in agarose plugs for one male individual from each of the HL WPR lines as well as one from slow feathering and fast feathering WLs. High molecular weight DNA was extracted at The Earlham Institute (UK) and analysed using the BioNano Irys platform with the Nt.BspQ1 restriction enzyme. Molecule object files were assembled and aligned to an in silico digest of the Galgal5 Z chromosome in IrysView v2.5.1 using the associated BioNano Knickers, Refaligner and Assembler software (release v5122; available: <https://bionanogenomics.com/support/software-downloads/>).

Analysis of ALVE integration site distribution

ALVE integration sites were overlapped with the Ensembl v87 Galgal5 feature GTF and the locations of all RepBase [72] repeat classes (20181026 release). GC bias was inspected across each target site duplication and windows of 100 bp, 1kbp, 10kbp and 100kbp centred on the integration. Values were compared to the genome and chromosomal average, as well as simulations of an equal number of random integrations repeated one million times.

A matrix was constructed for the presence/absence of each ALVE within each analysed dataset, and used to build a hierarchical binary cluster tree based on Jaccard distances. A general linear model (GLM) was fitted to identify whether differences in sequencing library type (individual vs pool) or average genome coverage influenced ALVE identification by obsERVer when lines were grouped into five broad categories: white-egg layers, brown-egg layers, broilers, native breeds, and 'wild', including the RJF samples. GLM category groups are shown in Table S1.

Results

Detection of ALVE integrations in HL elite layer lines by obsERVer

We have developed obsERVer (<https://github.com/andrewstephenmason/obsERVer>), a focused bioinformatic pipeline for the detection of ALVE integrations in WGS datasets, which utilises popular, freely available tools for processing next generation sequencing data. obsERVer was initially used to identify twenty different ALVEs across eight elite layer lines from Hy-Line International, of which four ALVEs were novel to this study (Table 1). All detected ALVEs were present in the full genome alignment maps, and no other known ALVE sites (Table S3) were detected, including the two ALVEs of the reference genome (ALVE6/ALVE-JFevA and ALVE-JFevB) [50].

In nineteen cases, manual inspection of the ALVE alignments in IGV was simple, with elevated coverage of

the TSD and split reads supporting both the 5' and 3' ends of the ALVE integration. The novel ALVE_ros007, however, was initially identified by obsERVer as two separate sites 1939 bp apart due to a post-integration deletion of over 8 kbp, excising the absent genomic sequence (intergenic with no predicted regulatory or conserved regions) and over 80% of the ALVE integration (Figure S3). Manual inspection of obsERVer-identified ALVEs therefore remains crucial, particularly when detected sites are only supported by 5' or 3' split reads alone.

We validated the bioinformatically-detected ALVEs by developing specific high-throughput KASP™ genotyping assays for each identified integration (Fig. 2; Figure S4; Table S4), and genotyped the original eighty males used for sequencing. Strikingly, obsERVer exhibited a 0% false discovery rate (FDR) as all bioinformatically identified ALVEs were subsequently detected by KASP assay in their appropriate lines. However, by KASP assay alone, ALVE3 was detected in a single RIR bird and ALVE5 in a single WPR1 bird; neither of which were identified by obsERVer (Table 1). The full genome alignment files for each dataset showed no supportive read evidence for either integration, supporting the loss of both sites due to allelic dropout in the sequencing pools. Further genotyping of over 9000 males from multiple generations across the eight elite layer lines identified a further four ALVE occurrences within the WPRs (at frequencies < 0.1) which were not found in the smaller subset of sequenced birds (Table 1).

With a 0% FDR, obsERVer is a highly specific detection tool limited only by the experimental design of the sequencing project. The HL WGS datasets, limited by technology at the time, were from 10-individual pools and had low average coverage of 11–18X. Based on our modelling of integration detection, only ALVEs with a flock frequency of 0.3–0.5 had a > 90% probability of detection. At frequencies of 0.1 there was only a 28–45% chance of detecting that integration. Whilst this may support the use of target enrichment sequencing (such as [23]), WGS using individual libraries with high coverage is now commonplace, resulting in > 90% probability of detection with ALVE frequencies < 0.1 at 30x coverage. Rare ALVEs may still be missed due to population sampling, but target enrichment sequencing would not improve this.

Characterisation of the obsERVer-identified ALVEs in the HL elite layer lines

Of the twenty identified ALVEs, eleven were shared between multiple lines (of which five were between multiple breeds; Table 1) giving a total of forty-two ALVE occurrences across the HL datasets. The white-egg layer WLs had fewest ALVEs with two to four loci per line,

Table 1 ALVEs of the Hy-Line elite layer lines

Name	Location	TSD	Gene	Length	WL1	WL2	WL3	WL4	WL5	WPR1	WPR2	RIR
ALVEB5	1:10,637,460	GGTGGT		7530 (F)						AD	✓	✓
ALVE1	1:65,993,542	ACGGTT	SOX5 int1	7530 (F)	✓	✓	✓	✓	✓			
ALVE_ros001 (COTW55)	1:101,668,931	GTTGTG		7531 (F)								✓
ALVE_ros002 (COTW69)	1:158,775,708	ATAAGT		–								✓
ALVE_ros003 (SGT-24)	1:163,248,553	CCTACT		7528 (F)								✓
ALVE-TYR	1:187,921,213	ACACTG	TYR int4	7534 (F)						✓	✓	
ALVE-NSAC1	2:120,868,843	CCTGTT		4838 (P-E-3 L)						✓	✓	✓
ALVE_ros004	2:124,432,997	CTTGAC		7530 (F)						NSI	NSI	✓
ALVE_ros005 (New11)	2:142,480,536	TTGATA		280 (SL)						NSI		✓
ALVE-NSAC3	3:53,639,776	ATAAAA		–						✓	✓	
ALVE_ros006 (N4)	3:57,337,987	GGACTC		–								✓
ALVE15	3:70,384,294	GTTTAT	GRIK2 int16	280 (SL)	✓	✓			✓			
ALVE_ros007	4:59,843,015	AATAGA		1400 (E-3 L)								✓
ALVE_ros008 (BK-59)	4:62,680,158	CTGTAG		7529 (F)				✓				
ALVE_ros009	4:71,095,932	GTCCAG		–							✓	
ALVE9	6:33,153,441	CTCAAA	DOCK1 int35	5077 (P-E-3 L)			✓					
ALVE-NSAC7	9:11,714,130	CTTCTC		7531 (F)						✓	✓	
ALVE_ros010	9:11,871,576	TCGGAT		–						NSI		✓
ALVE3	20:10,309,347	AACCAC	HCK int6	5848 (F, RT-)		✓	✓	✓				AD
ALVE21	Z:10,681,671	GGGTAG		7529 (F)				✓		✓	✓	

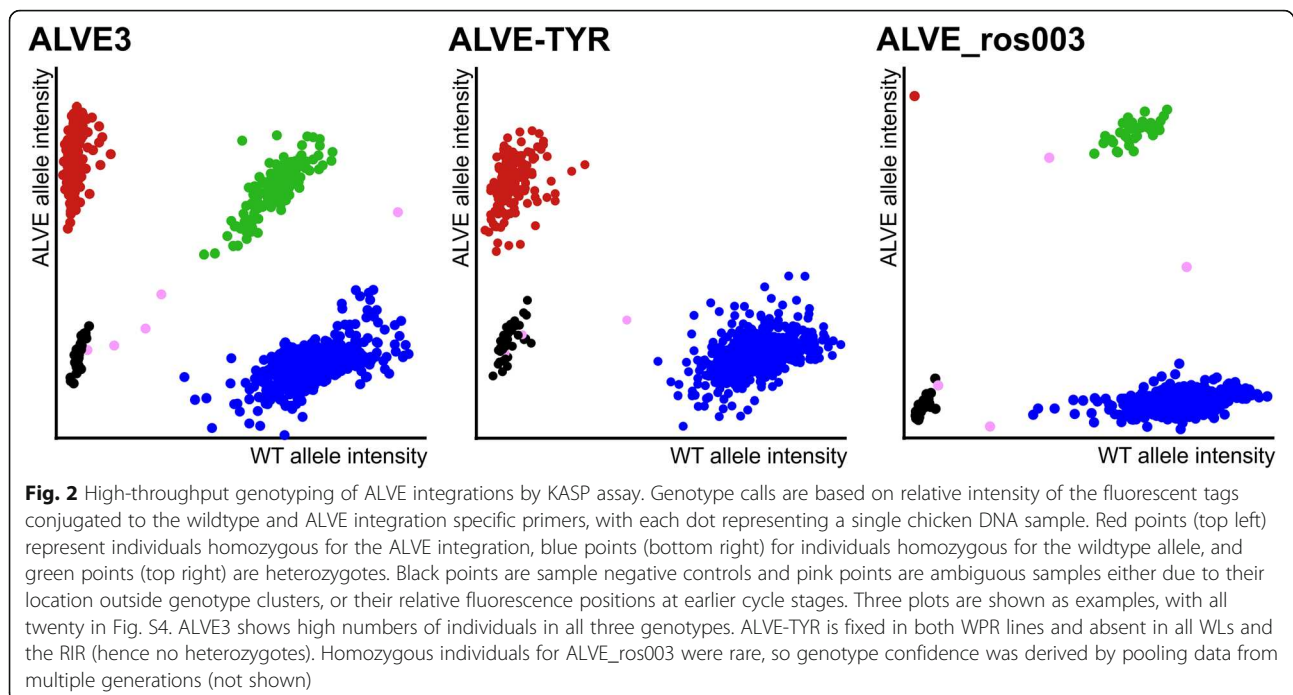
Identified ALVEs are shown with Galgal5 location, target site duplication (TSD), overlap with annotated gene, sequence integrity, and presence in each of the eight analysed lines. Ambiguous prior names are shown. ALVE integrity is shown under length where: F = full, P = polymerase, E = envelope, 3 L = 3'LTR, RT = missing reverse transcriptase, SL = solo LTR. Five ALVEs were not sequenced. ALVE detection: tick indicates detection by obsERVer; AD shows allelic dropout in the sequencing data, NSI shows that ALVE was present in the line, but not in the sequenced individuals.

but these were often fixed or at high frequency within each flock. The three brown-egg layer lines had a greater number of ALVE loci (eight and nine in the WPRs; 11 in the RIR), typically found at flock frequencies < 0.2, likely reflecting the broader genetic background of these breeds [73]. All four novel ALVEs identified in this study were intergenic and found in brown-egg layers.

Five ALVEs were within introns (Table 1). Except for ALVE-TYR, where the integration in the final *TYR* (Tyrosinase) intron causes transcript truncation [44, 45], none of these has any reported effect on their containing gene. Interestingly, ALVE15, a solo LTR widespread in

layers [21], is within the final intron of *GRIK2* (Glutamate Ionotropic Receptor Kainate Type Subunit 2) and Ensembl reports a *GRIK2* transcript which lacks the final exon, corresponding to the entire intracellular domain known to regulate channel dynamics in other glutamate receptor family members [74].

Fifteen of the twenty ALVEs were fully sequenced (Table 1; AF2). Of these, ALVE15 and ALVE_ros005 were solo LTRs, and ALVE9, ALVE-NSAC1 and ALVE_ros007 had varying 5' truncations. The remaining ten were 'intact' elements with both 5' and 3' LTRs, although ALVE3 had no *reverse transcriptase* domain,



matching the GenBank reference (AY013304.1). ALVE LTRs retained high identity (98.6% across all LTRs; 9/10 intact element LTR pairs had 100% identity; ALVEB5 5' LTR had a single SNP at G262T) and contained intact TATA boxes, transcription start sites, and two binding sites for serum response factor (SRF). The high sequence similarity and integrity seen for the LTRs was not observed throughout the internal coding domains. Six of the ten ALVEs with a gag domain contained one or more mutations in the p10 or p27 that truncated any potential transcripts (Figure S5). Intact p27 was detected from ALVE1, ALVE3, ALVE21 (the only ALVE with the potential for total expression) and ALVE-TYR. Whilst ALVE1 is not normally expressed [33, 75, 76], the others have well characterised expression in both commercial and experimental lines [27, 44]. Sequence integrity was better across the envelope domain (Figure S6), with ten of the thirteen represented ALVEs containing unbroken reading frames with four to six non-synonymous changes. However, envelope expression may be inhibited by the intact miR-155 target site found in all ALVEs (position 5634–5640 relative to AY013303.1) [71]. No sequence was obtained for the remaining five ALVEs, however the original split reads suggest these elements had intact LTR pairs.

ALVE21 and a K locus revertant

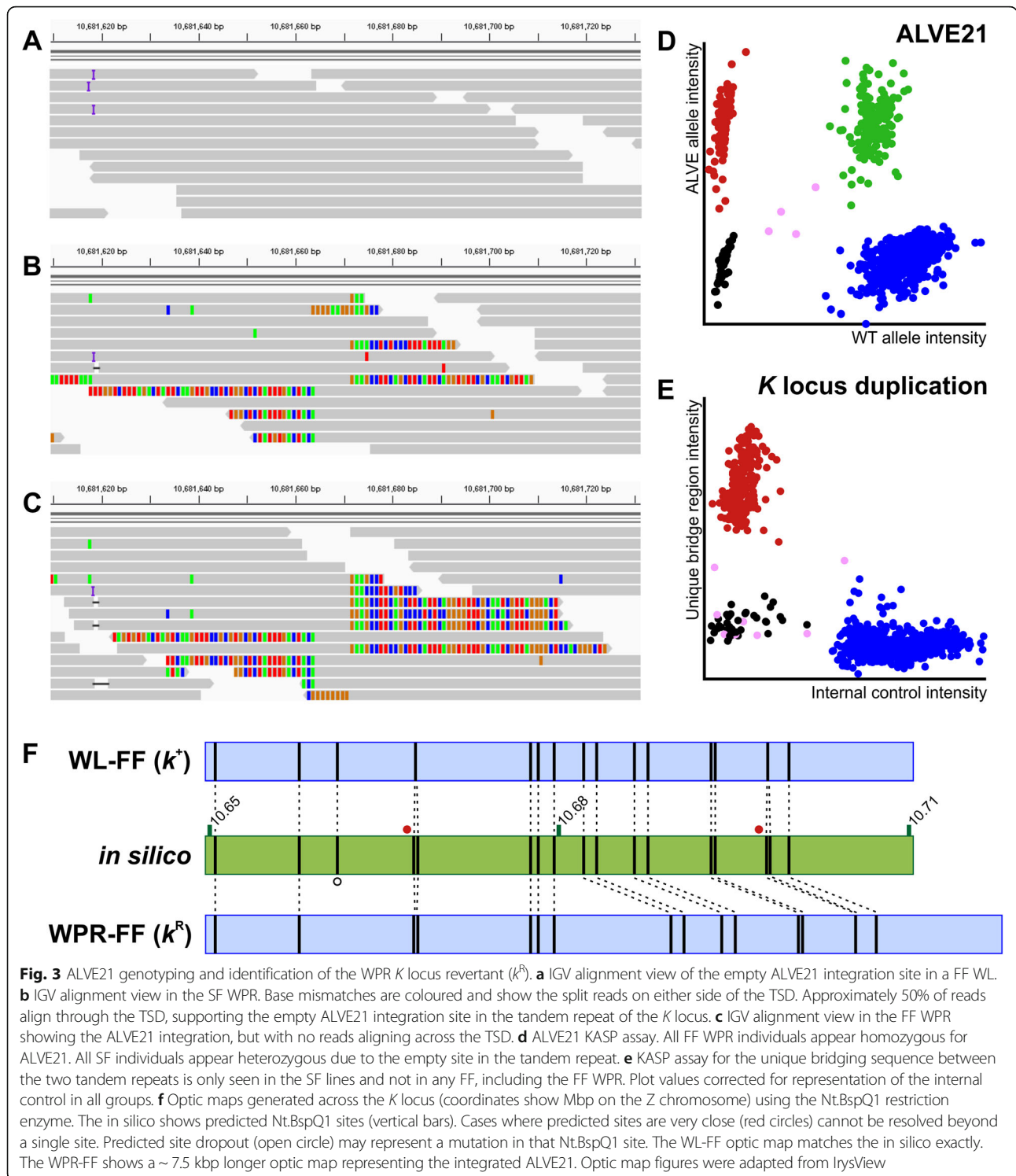
ALVE21 is an integration of great commercial interest as it is associated with the *K* locus mutation; a 180kbp tandem duplication on the Z chromosome which leads to gender-dependent, slowed feathering rate in day old-chicks [39, 42, 43]. The slow/fast feathering phenotype is

used extensively in the production industry as a rapid non-invasive means of determining gender at hatch. Consequently, ALVE21 is often present in commercial flocks, even though it is structurally intact (as shown above) and retains the potential to produce retroviral proteins [29].

ALVE21 was detected by obsERVer in both of the “slow feathering” (SF) HL lines (Fig. 3a-b), and by KASP assay (Fig. 3d) all SF birds appeared heterozygous, as ALVE21 is only found in one of the *K* locus tandem duplication sites (Figure S7). However, ALVE21 was also found in the wild-type, “fast feathering” (FF) WPR sister line, but exclusively in a homozygous state (Fig. 3c-d). This supports a phenotypic reversion (K^R) by recombination in the FF WPR (as has been previously reported [77]), retaining the copy containing the ALVE21 integration (Figure S7). We validated this by designing a KASP assay to the unique bridging sequence between the tandem duplications. Congruently, both SF lines were homozygous for this sequence, but it was absent in all FF lines, including the phenotypic revertant (Fig. 3e). The ALVE21 integration in the FF WPR was also detected using BioNano high resolution optic mapping relative to a truly wildtype FF WL (Fig. 3f). Unfortunately, there was insufficient molecule resolution to fully describe the *K* locus in any SF individual or combined dataset (Table S7).

Application of obsERVer to reveal broad ALVE diversity

Following validation of the twenty ALVE integration sites identified in the HL lines (with 0% FDR), obsERVer was used to identify ALVEs from fifty-seven diverse



chicken WGS datasets (Table S1). These included experimental and heritage layer and broiler lines, further commercial layer lines, African and Asian native breeds, and wild-caught RJF from China, Java and Sumatra. When combined with the HL ALVES, a total of 322 different ALVES were identified by obsERVer

in this study, of which 261 (81.1%) were novel (AF1). Neither of the two ALVES of the reference genome were identified in any dataset, even the wild-caught RJF samples. All identified datasets contained at least one ALVE, except the ADOL Line 0 which had been selected to be ALVE free [37].

The WL samples had the fewest ALVEs (up to six per dataset) and formed a distinct clade in the dendrogram constructed using ALVE content (Fig. 4). Twelve of the

nineteen ALVEs found in WLS had been identified in previous studies, and “typical” WL ALVEs were highly prevalent: ALVE1 (22/23), ALVE3 (14/23), ALVE9 (7/23) and

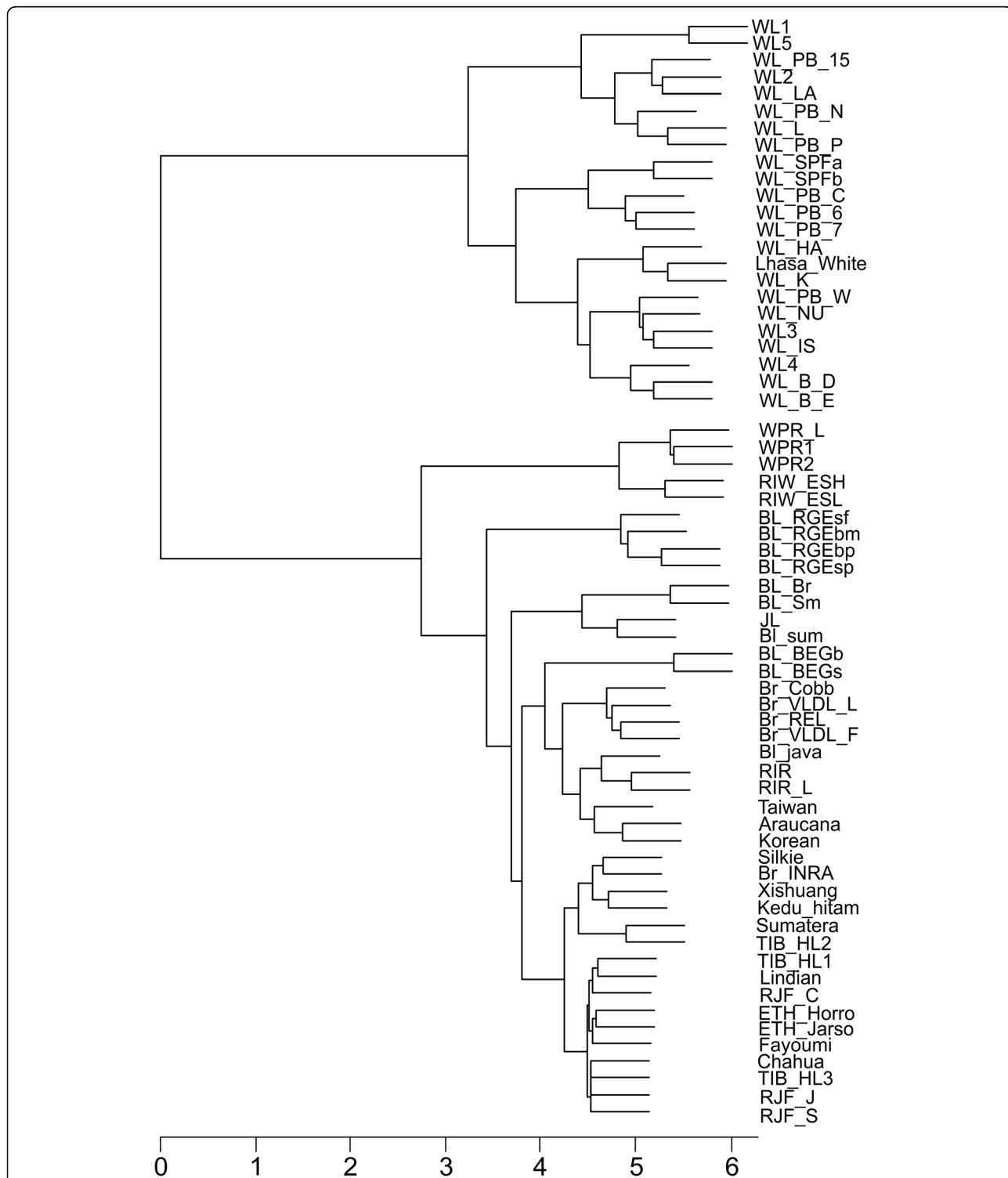


Fig. 4 ALVEs as genetic markers. Cladogram constructed based on ALVE presence/absence data for all sixty-five analysed datasets (Table S1; AF1). WLS dominate and cluster tightly together in the top clade. Brown Leghorns cluster with the brown egg layer WPRs, RIWs and RIRs, as well as the heritage broiler datasets, reflecting the broader genetic diversity of these breeds. The non-commercial datasets are highly diverse

ALVE15 (10/23). Brown egg layers, including WPR, RIR and RIW lines, had higher ALVE content (six to eleven) and the heritage broilers were higher again (thirteen to thirty), with many sites shared between these lines. Thirty-six of the eighty-three ALVEs found in commercially relevant breeds had been identified previously. Other, non-commercial WGS datasets were highly variable in their ALVE content, and also exhibited high lineage-specificity. Across the entire study, 260 ALVEs were specific to a single dataset (80.7%) with over 60% of these identified in the village chickens and wild-caught RJF.

It is unlikely these totals represent a comprehensive catalogue of ALVE content in these lines, particularly as some WGS datasets were derived from only a single bird, or a pool of 2–3 birds. Despite this, over 90.3% of the variation in ALVE content across the analysed datasets was due to line category (largely based on breed; Table S1): the only significant variable in the GLM ($P < 10^{-4}$). Genome coverage and derivation from an individual or pooled sequencing library were both non-significant terms.

ALVE integration site distribution

Previous studies of exogenous ALV integration sites suggested a preference for open chromatin, particularly near protein-coding genes [78–80]. Under a model of random integration 51.8% of sites would fall within coding regions, however only 26.7% of the 322 ALVEs identified in this study were within genes; a significant depletion ($P = 2.0 \times 10^{17}$). However, 32.9% of ALVEs were within 10kbp of a gene, compared to just 4.1% in the random integration model ($P = 3.1 \times 10^{-147}$). Furthermore, whilst there was no observable GC bias at any window size across the integration sites, the 6 bp TSDs were significantly more GC rich than expected (49.9% GC compared to genome mean of 42.4%; $t = 4.66$; $P = 3.8 \times 10^{-6}$). Taken together, these results support the previous ALV integration studies [78, 79], with depletion within genes likely due to post-integration selection on the host genome.

ALVE_ros012 (found only in the RJF-J dataset; Table S1) was the only ALVE found within an exon (exon 8/11 of carboxypeptidase A5 precursor; *CPA5*), and would likely cause a truncation of 170 amino acids (40.6%). The impact of this truncation on the host is unknown, but may be mediated by multiple *CPA5* paralogues, including the neighbouring *CPA1* and *CPA2*. Intronic integrations (25.8% of all identified sites) may also elicit effects on the host organism by causing truncations or exon skipping, as with ALVE-TYR [44].

A total of eighteen ALVEs (5.6%) were found to have integrated within assembled Chicken Repeat 1 (CR1) elements within the genome (AF1). These included thirteen

novel integrations (including ALVE_ros009 identified in the HL elite layer lines), as well as the previously described ALVE12, ALVE16, ALVEB10, ALVE_NSAC2 and ALVE_NSAC5. The observed number of ALVEs within assembled repetitive elements closely matched that of the random integration model (5.7%; $P = 1.00$).

Discussion

obsERVer enables detection of specific retroviral integrations from WGS data

High-throughput sequencing technologies have facilitated the description of many genomic features, although repetitive elements remain relatively understudied. Whilst repeat element-targeted sequencing technologies (such as [23]) seem appealing, the generated data are not typically applicable to other research questions, with data repurposing usually a major strength of sequencing projects. Here, we have described obsERVer, a pipeline developed for the identification of specific, user-determined retroviral integrations in existing WGS data, and then applied it to the identification of ALVE integrations in sixty-five chicken datasets, describing 322 ALVEs, of which 261 were novel (including 6 within commercial lines).

Development of diagnostic assays to 20 ALVEs identified across eight elite layer lines revealed a 0% FDR for obsERVer in detecting ALVE integrations, making it a highly precise method for identifying integration sites. It is unlikely that the ALVEs identified in this study represent a complete annotation of all integration sites within the WGS examined for this study. Integrations within difficult to sequence, or poorly assembled regions of the genome will not be detected, currently limiting any identification on many of the microchromosomes, the W chromosome, or near the centromeres and telomeres. For example, ALVE6 is common in commercial lines [21] but was not found in this study, likely due to its location near the chromosome 1 p arm telomere and its incomplete assembly in Galgal5. Further reference genome improvements will aid ALVE identification; recently shown specifically with ALVE6 [50].

Beyond the genome itself, the sequencing strategy also impacts obsERVer annotation completeness. As we saw in the HL data, rare ALVEs in a flock may be missed due to the specific individuals chosen for sequencing, and from allelic dropout from pooled sequencing libraries. With higher coverage and individual sequencing libraries now typical, allelic dropout is of less concern in future projects, however researchers should consider the minimum number of individuals needed to identify rare integrations. Target-enriched sequencing from multiple pooled-individual libraries (with high coverage) might be a more cost-effective way to ensure identification of all ALVEs in a population if the sole purpose of the

investigators is to design genotyping assays to those integrations.

The biological impact of ALVEs in chicken populations

A total of 322 different ALVE integrations were identified in this study. We confirmed previous work showing that commercial layers had fewer ALVEs than broilers [21, 81], and our novel assessment of non-commercial populations suggests that intensive poultry selection has successfully limited ALVE abundance within flocks. The greater number of highly lineage-specific ALVEs in non-commercial and RJF populations may suggest a high ancestral diversity of ALVEs before domestication [73, 82, 83], and would be consistent with recurrent infection and a role for ALVEs in ERV derived immunity (EDI) against exogenous ALV [84, 85]. Moreover, the RJF reference genome does not appear to be representative of observed ALVE diversity as it contains only two ALVEs [49, 50, 86]. A broader analysis of non-commercial and RJF datasets is needed to assess the role of ALVEs in wild populations, particularly as the structural integrity of each integration cannot be unambiguously determined from short read sequencing data alone.

The lower ALVE number in commercial lines is likely due to a combination of detrimental associations with productivity traits, relatively small effective population sizes, and the narrow genetic background of some breeds; factors which are less prominent in broilers compared with layers [73]. Flocks have also been subjected to decades of selection against the ALV-specific p27 antigen, and degradation of this region was seen in six of the ten HL ALVEs with an intact *gag* domain. However, with breeding programmes focused on multiple traits, and the close association of some ALVEs with desirable phenotypes, many ALVEs in commercial lines are found at very high frequencies, or have become fixed. Traditional selective breeding methods could gradually reduce ALVE allele frequencies, but fixed ALVEs, such as ALVE21 and ALVE-TYR in both HL WPRs, could only be removed by out crossing, which would likely create varied, undesirable production phenotypes. The CRISPR/Cas9 system was recently used to eradicate porcine ERVs from the pig genome [87], and could be an approach applied to commercial poultry to remove ALVEs such as ALVE21 whilst maintaining the associated slow feathering phenotype. Furthermore, accurate integration site identification by obsERVer facilitates highly specific genome editing for ALVE removal. Use of the high-throughput diagnostic KASP assays developed in this study have begun to identify phenotypic effects of segregating ALVEs in the HL flocks [88], and will identify priorities for future breeding programmes.

Conclusions

We have developed and utilised the obsERVer pipeline to identify 322 ALVE integration sites across sixty-five chicken WGS datasets without the need for additional targeted sequencing. Further work is needed to elucidate the biological impact, if any, of these ALVEs on exogenous ALV infection modulation, and on productivity traits. Development of high-throughput diagnostic assays will enable better management of ALVEs in commercial stock and may lead to their eventual eradication in these lines. Beyond ALVEs, obsERVer can be applied to the identification of any retroviral integration in any species.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13100-020-00216-w>.

Additional file 1: AF1. Presence/Absence matrix for all identified ALVEs. This file gives the location, names, previous names and any gene or repeat element overlaps for each of the identified ALVEs as well as their presence (1) or absence (0) within each analysed dataset. Dataset names are given as the codes listed in Table S1.

Additional file 2: AF2. Hy-Line ALVE sequences. This file contains all fifteen successfully sequenced ALVEs from the eight Hy-Line elite layer lines. FASTA headers include sequence genomic orientation, observed length and new GenBank accession numbers where applicable.

Additional file 3: AF3. Supplementary figures. This file includes seven additional figures which support the manuscript. These are referred to in the text as Fig. S1 etc. Full titles and legends are given for each figure.

Additional file 4: AF4. Supplementary tables. This file includes seven additional tables which support the manuscript, including lists of primers. These are referred to in the text as Table S1 etc. Full titles and legends are given for each table.

Abbreviations

ADOL: Avian Disease and Oncology Laboratory; ALV: Avian Leukosis Virus; ALVE: Avian Leukosis Virus subgroup E; ART-CH: Avian retrotransposon of chickens; BL: Brown Leghorn; bp: Base pairs; CR1: Chicken Repeat 1; EAV: Endogenous avian virus; EDI: ERV-derived immunity; ERV: Endogenous retrovirus; FDR: False discovery rate; FF: Fast feathering; GLM: General linear model; HL: Hy-Line; KASP: Kompetitive Allele-Specific PCR; kbp: Kilobase pairs; LTR: Long terminal repeat; MDV: Marek's Disease virus; NGS: Next generation sequencing; PCR: Polymerase chain reaction; RFLP: Restriction fragment length polymorphism; RIR: Rhode Island Red; RIW: Rhode Island White; RJF: Red junglefowl; SF: Slow feathering; SNP: Single nucleotide polymorphism; SRF: Serum response factor; T_m : Melting temperature; TSD: Target site duplication; UTR: Untranscribed region; WGS: Whole genome sequencing; WL: White Leghorn; WPR: White Plymouth Rock

Acknowledgements

The authors would like to take this opportunity to thank Professor Bernhard Benkel for kindly sharing data on previously identified ALVEs, Bob Paton for assisting with the ALVE sequencing, Dr. Gregor Gorjanc for his help in developing the obsERVer sensitivity model, Dr. Scott Tyack for helpful discussions regarding ALVE detection and genotyping, Dr. Jacqueline Smith and Dr. Samantha Lycett for support at the end of ASM's doctoral studentship, and Grant Liebe, Amy McCarron and Kara Pinegar for all their assistance with sample preparation and genotyping at Hy-Line International. We offer our sincere thanks to Professors Chris Ashwell, Marc Eloit, Olivier Hanotte, Susan Lamont, Rudolf Preisinger and Douglas Rhoads for so kindly sharing chicken WGS data for analysis. We would also like to thank our colleagues for critically reviewing this manuscript.

Authors' contributions

ASM, PMH, DWB and JEF participated in the design of the study. DWB and JEF supervised the work, design and initial concept of the study. ASM developed the obsERVer pipeline, performed the WGS data analysis, assisted in KASP assay design, sequenced the ALVE inserts and prepared the manuscript. ARL developed and tested the KASP assays, and genotyped the Hy-Line DNA sample bank. All authors critically reviewed and approved the final revision of the manuscript.

Funding

ASM was funded by BBSRC CASE studentship 1361596 in association with Hy-Line International. The Roslin Institute is supported by an Institute Core Strategic Grant from the Biotechnology and Biological Sciences Research Council (BB/J004235).

Availability of data and materials

The obsERVer bioinformatics pipeline is freely available on GitHub (<https://github.com/andrewstephenmason/obsERVer>). Accession values for publicly available WGS data have been indicated, or the relevant publication for non-public data (Table S1). Eleven previously unpublished ALVE sequences derived in this study have been uploaded to GenBank with accession numbers indicated in AF2.

Ethics approval and consent to participate

All samples were collected under the guidelines of the Hy-Line International Animal Use and Care Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no competing interests.

Author details

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK. ²York Biomedical Research Institute, The Department of Biology, The University of York, York YO10 5DD, UK. ³Hy-Line International, 2583 240th Street, Dallas Center, Iowa 50063, USA. ⁴The University of Queensland, Brisbane, Queensland 4072, Australia.

Received: 6 April 2020 Accepted: 17 June 2020

Published online: 30 June 2020

References

- Doolittle R, Feng D, Johnson MS, et al. Origins and evolutionary relationships of retroviruses. *Q Rev Biol*. 1989;64(1):1–30.
- Stoye JP. Endogenous retroviruses: still active after all these years? *Curr Biol*. 2001;11(22):R914–6.
- Reiss D, Mager DL. Stochastic epigenetic silencing of retrotransposons: does stability come with age? *Gene*. 2007;390(1–2):130–5.
- Kanda R, Tristem M, Coulson T. Exploring the effects of immunity and life history on the dynamics of an endogenous retrovirus. *Philos Trans R Soc Biol Sci*. 2013;368(1626):20120505.
- Mason AS, Fulton JE, Hocking PM, et al. A new look at the LTR retrotransposon content of the chicken genome. *BMC Genomics*. 2016;17(1):688.
- Kapusta A, Suh A. Evolution of bird genomes—a transposon's-eye view. *Ann N Y Acad Sci*. 2017;1389:164–85.
- Rigal M, Mathieu O. A “mille-feuille” of silencing: epigenetic control of transposable elements. *Biochim Biophys Acta*. 2011;1809(8):452–8.
- Magiorkinis G, Blanco-Melo D, Belshaw R. The decline of human endogenous retroviruses: extinction and survival. *Retrovirology*. 2015;12(1):8.
- Katz RA, Skalka AM. Generation of diversity in retroviruses. *Annu Rev Genet*. 1990;24:409–45.
- Bock M, Stoye JP. Endogenous retroviruses and the human germline. *Curr Opin Genet Dev*. 2000;10(6):651–5.
- Katzourakis A, Rambaut A, Pybus OG. The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol*. 2005;13(10):463–8.
- Isbel L, Whitelaw E. Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes. *Bioessays*. 2012;34(9):734–8.
- Payne LN, Nair V. The long view: 40 years of avian leukosis research. *Avian Pathol*. 2012;41(1):11–9.
- Magiorkinis G, Gifford RJ, Katzourakis A, et al. Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci U S A*. 2012;109(19):7385–90.
- Venugopal K. Avian leukosis virus subgroup J: a rapidly evolving group of oncogenic retroviruses. *Res Vet Sci*. 1999;67(2):113–9.
- Dunn CA, Romanish MT, Gutierrez LE, et al. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene*. 2006;366(2):335–42.
- Borysenko L, Stepanets V, Rynditch AV. Molecular characterization of full-length MLV-related endogenous retrovirus ChiRV1 from the chicken, *Gallus gallus*. *Virology*. 2008;376(1):199–204.
- Coffin JM, Tschlis PN, Conklin KF, Senior A, Robinson HL. Genomes of endogenous and exogenous avian retroviruses. *Virology*. 1983;126(1):51–72.
- Payne LN. Retrovirus-induced disease in poultry. *Poult Sci*. 1998;77(8):1204–12.
- Frisby DP, Weiss RA, Roussel M, et al. The distribution of endogenous chicken retrovirus sequences in the DNA of galliform birds does not coincide with avian phylogenetic relationships. *Cell*. 1979;17(3):623–34.
- Benkel BF. Locus-specific diagnostic tests for endogenous avian leukosis-type viral loci in chickens. *Poult Sci*. 1998;77(7):1027–35.
- Borisenko L. Avian endogenous retroviruses. *Folia Biol*. 2003;49(5):177–82.
- Rutherford K, Meehan CJ, Langille MGI, et al. Discovery of an expanded set of avian leukosis subgroup E proviruses in chickens using Vermillion, a novel sequence capture and analysis pipeline. *Poult Sci*. 2016;95:2250–8.
- Crittenden LB, Smith EJ, Fadly AM. Influence of endogenous viral (ev) gene expression and strain of exogenous avian leukosis virus (ALV) on mortality and ALV infection and shedding in chickens. *Avian Dis*. 1984;28(4):1037–56.
- Fox W, Smyth JRJ. The effects of recessive white and dominant white genotypes on early growth rate. *Poult Sci*. 1985;64(3):429–33.
- Kuhnlein U, Sabour M, Gavora JS, et al. Influence of selection for egg production and Marek's disease resistance on the incidence of endogenous viral genes in white leghorns. *Poult Sci*. 1989;68(9):1161–7.
- Gavora JS, Kuhnlein U, Crittenden LB, et al. Endogenous viral genes: association with reduced egg production rate and egg size in white leghorns. *Poult Sci*. 1991;70(3):618–23.
- Ka S, Kerje S, Bornold L, et al. Proviral integrations and expression of endogenous avian leukosis virus during long term selection for high and low body weight in two chicken lines. *Retrovirology*. 2009;6:68.
- Smith EJ, Fadly AM, Crittenden LB. Interactions between endogenous virus loci ev6 and ev21:1. Immune response to exogenous avian Leukosis virus infection. *Poult Sci*. 1990;69(8):1244–50.
- Smith EJ, Fadly AM, Crittenden LB. Interactions between endogenous virus loci ev6 and ev21:2. Congenital transmission of EV21 viral product to female progeny from slow-feathering dams. *Poult Sci*. 1990;69(8):1251–6.
- Smith EJ, Fadly AM, Levin I, et al. The influence of ev6 on the immune response to avian Leukosis virus infection in rapid-feathering progeny of slow- and rapid-feathering dams. *Poult Sci*. 1991;70(8):1673–8.
- Chang S, Xie Q, Wang C, et al. Genetic susceptibility to and presence of endogenous avian leukosis viruses impose no significant impact on survival days of chickens challenged with very virulent plus Marek's disease virus. *Ann Virol Res*. 2015;1(2):1007.
- Hu X, Zhu W, Chen S, et al. Expression patterns of endogenous avian retrovirus ALVE1 and its response to infection with exogenous avian tumour viruses. *Arch Virol*. 2017;162(1):89–101.
- Fadly A, Mays J, Zhang H, et al. Role of endogenous avian leukosis virus and serotype 2 Marek's disease virus in enhancement of spontaneous lymphoid-leukosis-like tumors in chickens. In: American Veterinary Medical Association Annual Convention, July 25–29, 2014, Denver, Colorado. 2014. Abstract No. 16756.
- Cao W, Mays J, Kulkarni G, et al. Further observations on serotype 2 Marek's disease virus-induced enhancement of spontaneous avian leukosis virus-like bursal lymphomas in ALVA6 transgenic chickens. *Avian Pathol*. 2015;44(1):23–7.
- Mays JK, Black-Pyrkosz A, Mansour T, et al. Endogenous avian leukosis virus in combination with serotype 2 Marek's disease virus significantly boosted

- the incidence of lymphoid leukosis-like bursal lymphomas in susceptible chickens. *J Virol.* 2019;93(23):1–18.
37. Crittenden LB, Fadly AM. Responses of chickens lacking or expressing endogenous avian leukosis virus genes to infection with exogenous virus. *Poult Sci.* 1985;64(3):454–63.
 38. Zhang H, Bacon LD, Fadly AM. Development of an endogenous virus-free line of chickens susceptible to all subgroups of avian leukosis virus. *Avian Dis.* 2008;52(3):412–8.
 39. Bacon LD, Smith E, Crittenden LB, et al. Association of the Slow Feathering (K) and an endogenous viral (ev21) gene on the Z chromosome of chickens. *Poult Sci.* 1988;67(2):191–7.
 40. Tixier-Boichard MH, Benkel BF, Chambers JR, et al. Screening chickens for endogenous virus ev21 viral element by the polymerase chain reaction. *Poult Sci.* 1994;73(10):1612–6.
 41. Tixier-Boichard M, Boulliou-Robic A. A deleted retroviral insertion at the ev21-K complex locus in Indonesian chickens. *Poult Sci.* 1997;76:733–42.
 42. Efferink MG, Vallée AAA, Jungerius AP, et al. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. *BMC Genomics.* 2008;9:391.
 43. Bu G, Huang G, Fu H, et al. Characterization of the novel duplicated PRLR gene at the late-feathering K locus in Lohmann chickens. *J Mol Endocrinol.* 2013;51(2):261–76.
 44. Chang CM, Coville JL, Coquerelle G, et al. Complete association between a retroviral insertion in the tyrosinase gene and the recessive white mutation in chickens. *BMC Genomics.* 2006;7:19.
 45. Chang CM, Furet JP, Coville JL, et al. Quantitative effects of an intronic retroviral insertion on the transcription of the tyrosinase gene in recessive white chickens. *Anim Genet.* 2007;38(2):162–7.
 46. Aarts HJM, van der Hulst-van Arkel MC, Beuving G, et al. Variations in endogenous viral gene patterns in white Leghorn, medium heavy, white Plymouth rock, and Cornish chickens. *Poult Sci.* 1991;70(6):1281–6.
 47. Tixier-Boichard M, Durand L, Morisson M, et al. Comparative analysis of avian leukosis virus-related endogenous viral genes in experimental strains of the domestic chicken. *Genet Sel Evol.* 1994;26:53–66.
 48. Grunder AA, Benkel BF, Chambers JR, et al. Characterization of eight endogenous viral (ev) genes of meat chickens in semi-congenic lines. *Poult Sci.* 1995;74(9):1506–14.
 49. Benkel B, Rutherford K. Endogenous avian leukosis viral loci in the red jungle fowl genome assembly. *Poult Sci.* 2014;93:2988–90.
 50. Mason AS, Fulton JE, Smith J. Endogenous avian Leukosis virus subgroup E elements of the chicken reference genome. *Poult Sci.* 2020;99(6):2911–5.
 51. Baillie JK, Barnett MW, Upton KR, et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature.* 2011;479(7374):534–7.
 52. Andrews S. FastQC. A quality control tool for high throughput sequence data. 2012. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
 53. Krueger F, Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Buisulfite-Seq) libraries. 2013. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
 54. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. Available from: <http://cutadapt.readthedocs.io/en/stable/index.html>.
 55. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* 2013;1303:3997v2.
 56. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.
 57. Chen X, Li D. ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. *Bioinformatics.* 2019;35(20):3913–22.
 58. Goubert C, Thomas J, Payr LM, et al. TypeTE: a tool to genotype mobile element insertions from whole genome resequencing data. *Nucleic Acids Res.* 2020;48(6):e36.
 59. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
 60. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 2013;14(2):178–92.
 61. Smit A, Hubley R, Green P. RepeatMasker Open-4.0.3. 2013. Available from <http://repeatmasker.org>.
 62. Kranis A, Gheyas AA, Boschiero C, et al. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics.* 2013;14(1):59.
 63. Rozen S, Skaletsky H. Primer3 on the WWW for general users and biologist programmers. *Methods Mol Biol.* 2000;132:365–86.
 64. Li H. Improving SNP discovery by base alignment quality. *Bioinformatics.* 2011;27(8):1157–8.
 65. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987–93.
 66. Li H. BFC: correcting Illumina sequencing errors. *Bioinformatics.* 2015;31(17):2885–7.
 67. Kearse M, Moir R, Wilson A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28(12):1647–9.
 68. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
 69. Delcher AL, Bratke KA, Powers EC, et al. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics.* 2007;23(6):673–9.
 70. Rice P, Longden I, Bleasby A. EMBOS: the European molecular biology open software suite. *Trends Genet.* 2000;16(6):276–7.
 71. Hu X, Zhu W, Chen S, et al. Expression of the env gene from the avian endogenous retrovirus ALVE and regulation by miR-155. *Arch Virol.* 2016;161(6):1623–32.
 72. Jurka J, Kapitonov WW, Pavlicek A, et al. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetics Genome Res.* 2005;110:462–7.
 73. Muir WM, Wong GK-S, Zhang Y, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proc Natl Acad Sci U S A.* 2008;105(45):17312–7.
 74. Maki BA, Aman TK, Amico-Ruvio SA, et al. C-terminal domains of N-methyl-D-aspartic acid receptor modulate unitary channel conductance and gating. *J Biol Chem.* 2012;287(43):36071–80.
 75. Conklin KF, Coffin JM, Robinson HL, et al. Role of methylation in the induced and spontaneous expression of the avian endogenous virus ev-1: DNA structure and gene products. *Mol Cell Biol.* 1982;2:638–52.
 76. Conklin K. Activation of an endogenous retrovirus enhancer by insertion into a heterologous context. *J Virol.* 1991;65(5):2525–32.
 77. Takenouchi A, Toshishige M, Ito N, et al. Endogenous viral gene ev21 is not responsible for the expression of late feathering in chickens. *Poult Sci.* 2018;97(2):403–11.
 78. Narezkina A, Taganov KD, Litwin S, et al. Genome-wide analyses of avian sarcoma virus integration sites. *J Virol.* 2004;78(21):11656–63.
 79. Serrao E, Ballandras-Colas A, Cherepanov P, et al. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology.* 2015;12:39.
 80. Grawenhoff J, Engelman AN. Retroviral integrase protein and intasome nucleoprotein complex structures. *World J Biol Chem.* 2017;26(81):32–44.
 81. Sabour MP, Chambers JR, Grunder AA, et al. Endogenous viral gene distribution in populations of meat-type chickens. *Poult Sci.* 1992;71(8):1259–70.
 82. Rubin C-J, Zody MC, Eriksson J, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature.* 2010;464(7288):587–91.
 83. Ellegren H. The avian genome uncovered. *Trends Ecol Evol.* 2005;20(4):180–6.
 84. Aswad A, Katzourakis A. Paleovirology and virally derived immunity. *Trends Ecol Evol.* 2012;27(11):627–36.
 85. Hurst T, Magiorkinis G. Activation of the innate immune response by endogenous retroviruses. *J Gen Virol.* 2015;96:1207–18.
 86. Ulfah M, Kawahara-Miki R, Farajallah A, et al. Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatera and Kedu Hitam. *BMC Genomics.* 2016;17(1):320.
 87. Niu D, Wei H-J, Lin L, et al. Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science.* 2017;357:1303–7.
 88. Mason AS, Wolc A, Arango J, et al. Effect of Endogenous Retroviral Elements (ALVE) on Egg Size in Commercial Egg production lines. In: Proceedings of the World Congress on Genetics Applied to Livestock Production, February 11–16, 2018, Auckland, New Zealand. Species - Avian. 2018;1:11.59. <http://www.wcgalp.org/proceedings/2018/effect-endogenous-retroviral-elements-alve-egg-size-commercial-egg-production-lines>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.