scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Pig and quail CpG methylation datasets from short and long read sequencing technologies

Paul Terzian[1,5], Céline Vandecasteele[2,5]✉, Joanna Lledo[2], Rémy-Félix Serre[2], Jules Sabban[2], Claire Kuchly[2], Frédérique Pitel[3], Sophie Leroux[3], Julie Demars[3], Nathalie Iannuccelli[3], Katia Fève[3], Michèle Bonnet[3], Christine Gaspin[1,4], Denis Milan[2,3], Carole Iampietro[2], Christophe Klopp[1,4] & Cécile Donnadieu[2]

CpG methylation, a key epigenetic mark involved in gene regulation, development, and other biological processes, is commonly analyzed using Whole-Genome Bisulfite Sequencing (WGBS). However, bisulfite treatment causes significant DNA degradation. Enzymatic Methyl-seq (EM-seq) offers a short-read alternative that preserves DNA integrity but requires conversion steps, limiting its compatibility with downstream analyses. Third-generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) and PacBio, enable direct detection of DNA modifications without altering the DNA, providing simultaneous genome and epigenome information. This work presents a comprehensive dataset combining long- and short-read sequencing data, including ONT, PacBio, Enzymatic Methyl-seq, and WGBS, for two agronomically relevant species: pig (*Sus scrofa*) and quail (*Coturnix japonica*). Data quality evaluation reveals high nucleotide quality scores for PacBio and short reads, robust alignment rates for long reads, and inter-method correlations in CpG methylation calling ranging from 0.76 to 0.99. This dataset is a valuable resource for training methylation callers and represents the first combined methylation dataset for these species, providing an essential benchmark for assessing emerging sequencing technologies.

Background & Summary

DNA modifications, such as CpG epigenetic marks, have long been studied to understand their role in gene expression regulation, genomic imprinting, X chromosome inactivation, transposable element repression, developmental processes, and disease associations. CpG sites are Cytosine-Guanine dinucleotides that can be methylated by the addition of a methyl group to the fifth carbon of the cytosine (5mC). Until now, most state-of-the-art detection methods have required chemical or enzymatic DNA treatment prior to sequencing. This is the case for Whole Genome Bisulfite Sequencing (WGBS) and the more recent method, Enzymatic Methyl-seq (EM-seq)[1]. Because these methods alter native DNA integrity and consequently modify bases, the resulting reads are more difficult to use for variant calling and genome assembly. Recently, third generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) and PacBio, have made it possible to detect DNA modifications without chemical or enzymatic alteration. Both methods rely on machine learning models trained to detect DNA methylation from raw or intermediate signal fluctuations during sequencing. The raw electrical signal from the nanopore is used to call modifications during DNA basecalling, while for PacBio reads, nucleotide incorporation kinetics during sequencing are used to detect epigenetic modifications.

Detecting CpG methylation using these technologies requires access to raw signals or additional data (such as fast5 or pod5 files for ONT and BAM files with kinetics for PacBio), along with extensive computational processing. For instance, a PromethION sequencer generates between hundreds and thousands of gigabytes of fast5 files per run, which are used exclusively for modification calling or rebasecalling. Consequently, most genomic

[1]Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France. [2]INRAE, US 1426, GeT-PlaGe, Genotoul, France Génomique, Université Fédérale de Toulouse, Castanet-Tolosan, France. [3]GenPhySE, Université Fédérale de Toulouse, INRAE, ENVT, 31326, Castanet-Tolosan, France. [4]Université Fédérale de Toulouse, INRAE, MIAT, 31326, Castanet-Tolosan, France. [5]These authors contributed equally: Paul Terzian, Céline Vandecasteele. ✉e-mail: celine.vandecasteele@inrae.fr

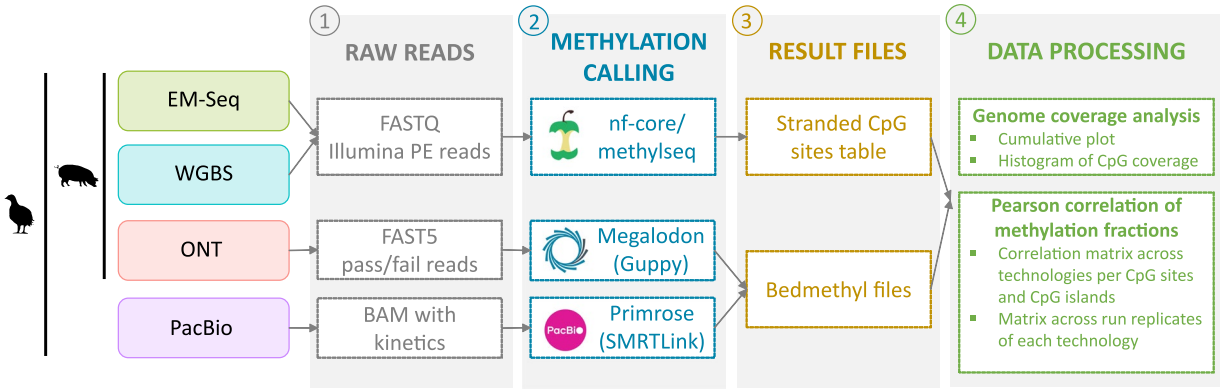

Fig. 1 Overview of study design and dataset produced. The quail and pig samples were sequenced using EM-Seq and WGBS methods on Illumina HiSeq 3000 or NovaSeq 6000, as well as Oxford Nanopore technology on PromethION. Additionally, the quail sample was sequenced using PacBio on Sequel II. The dataset was then processed by 1) analyzing the format of raw reads produced by each method, 2) applying specific CpG methylation calling pipelines, 3) compiling fractions of 5mC methylation and depth of coverage per CpG position and strand (one file per run), and 4) performing R analyses with the result data to evaluate and compare the technologies.

studies using ONT reads do not retain or publish raw signal files, resulting in limited availability of public datasets for benchmarking or model training. The most commonly used ONT dataset for CpG methylation calling comes from human B-Lymphocyte cells (Coriell NA12878). This dataset, produced with a MinION sequencer, was initially presented by Simpson *et al.*[2]. It was later used by Ni *et al.*[3] to train DeepSignal CpG calling models. Subsequently, Liu *et al.*[4] and Yuen *et al.*[5] incorporated it into a survey and benchmarking study, respectively. One of the reasons for its widespread use is that Simpson *et al.* also generated the corresponding WGBS data, which serves as ground truth for model training and method benchmarking. This approach was later adopted by Sigurpalsdottir *et al.*[6], who compared nanopore and PacBio data with corresponding oxBS data, but this time across a much larger number of samples.

Here, we present sequencing data and CpG methylation detection results for two species: the pig (*Sus scrofa*), a species genetically close to humans, and the quail (*Coturnix japonica*), a bird with microchromosomes exhibiting high GC content (Morris *et al.*[7]. Both DNA samples were sequenced using ONT, WGBS, and EM-seq. The quail sample was also sequenced with PacBio in HiFi mode. To assess data quality, we evaluated read quality, alignment rates, and compared methylation calling results (Fig. 1). For the first time, ONT and PacBio methylation detection reads are presented jointly with EM-seq and WGBS in two species, providing an essential dataset for benchmarking emerging sequencing methods for methylation calling.

Methods

Biological materials. The quail blood sample was harvested from a 6-month-old female of a high social reinstatement quail line, obtained through divergent selection on social motivation (Mills and Faure)[8]. The animal was bred at the INRAE Poultry Experimental Facility (PEAT) in Nouzilly, France (https://doi.org/10.15454/1.5572326250887292E12) in accordance with European Union guidelines for animal care, with approval from the local ethical committee for animal experimentation (Val de Loire) and the French Ministry of Higher Education and Scientific Research (authorization n 4609-2016032115196879). The pig blood sample was harvested from an adult Large White boar as part of the national selection program for the French LargeWhite breed, conducted by the three French breeding companies of the Alliance R&D association (composed of Axiom, Choice Genetics France, Nucleus, and IFIP). The procedure was performed in compliance with the guidelines of the French Ministry of Agriculture and Fisheries.

Genomic DNA extraction and quality control. High molecular weight genomic DNA (HMW DNA) was extracted from the quail blood sample using a high-salt extraction method (Roussot *et al.*)[9]. Pig HMW DNA was extracted from blood using the Genomic-tip 100/G kit (Qiagen, 10243) following the manufacturer's recommendations. Genomic DNA concentrations were measured using the Qubit fluorometry system with the Broad Range kit (Qiagen, 32853) for detection of double-stranded DNA (Thermo Fisher, Part #Q32854). Fragment size distributions were assessed using the Femto Pulse Genomic DNA 165 kb Kit (Agilent). Purity was evaluated using a Nanodrop system (Thermo Fisher). All samples were purified with beads to achieve expected purity ratios *i.e.* 260/280: 1.8–2 and 260/230: 2–2.2. For consistency purposes, the same DNA extractions were used across all three sequencing platforms (Illumina, ONT, PacBio). Prior to library preparation, the average gDNA molecular weight was estimated at 158 kb for quail and 174 kb for pig. Subsequently, DNA fragment size and quantity were carefully selected based on the specific requirements of each technology.

Library preparation and sequencing. Library preparation and sequencing were performed by the GeT-PlaGe core facility at INRAE Toulouse (https://doi.org/10.17180/NVXJ-5333).

Whole Genome Bisulfite Sequencing. WGBS libraries were prepared according to Bioo scientific's protocol using the NEXTflex™ Bisulfite Library Prep Kit for Illumina Sequencing. Briefly, 1500 ng of DNA were fragmented by sonication, size-selected using AMPure XP beads and ligated to adaptors prior to sequencing. Bisulfite treatment was then performed for 2.5 hours using the EZ DNA Methylation-Gold™ Kit from Zymo Research, and 12 PCR cycles were performed. Library quality was assessed using an Advanced Analytical Fragment Analyzer (Agilent Technologies) and libraries were quantified by qPCR using the Kapa Library Quantification Kit. WGBS libraries were sequenced in paired-end mode (2 × 150 pb) on an Illumina HiSeq 3000 with the Illumina HiSeq 3000 Reagent Kits and on a NovaSeq 6000 (Illumina, California, USA) with the Illumina NovaSeq Reagent Kits.

Enzymatic Methyl-seq. EM-seq libraries were prepared following NEB's protocol using the NEBNext Enzymatic Methyl-seq Kit. Briefly, 200 ng of DNA were fragmented by sonication using Pixul sonicator (Active Motif), followed by 5 PCR cycles. Library quality was assessed using an Advanced Analytical Fragment Analyzer (Agilent Technologies) and libraries were quantified by qPCR with the Kapa Library Quantification Kit (Roche). EM-seq libraries were sequenced on a NovaSeq 6000 (Illumina, California, USA) in paired-end mode (2 × 150 pb) with the Illumina NovaSeq Reagent Kits.

Oxford Nanopore Technology. ONT libraries were prepared following the manufacturer's instructions "1D gDNA selecting for long reads (SQK-LSK109)". At each step, DNA concentration was measured using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the Nanodrop spectrophotometer (Thermo Fisher Scientific). Size distribution and degradation were evaluated using the Fragment Analyzer (Agilent Technologies) DNF-464 HS Large Fragment Kit. Purification steps were performed using AMPure XP beads (Beckman Coulter). For a single flow cell, 5 µg of DNA were purified and sheared to 25 kb using the Megaruptor system (Diagenode), followed by a size selection step with the Short Read Eliminator M Kit (Circulomics). A single-step DNA damage repair, end-repair, and dA-tailing of double stranded DNA fragments was performed on 2 µg of DNA, followed by adapter ligation. The library was loaded onto a FLO-PRO002 R9.4.1 flow cell and sequenced on a PromethION instrument (Oxford Nanopore Technologies, UK) at 20 fmol for 72H. At 24 and 48 hours, a nuclease flush was performed on the same flow cell, followed by reloading of the library at 20 fmol.

PacBio. PacBio library preparation and sequencing was performed according to the manufacturer's instructions "Procedure & Checklist Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0". At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the Nanodrop (Thermo Fisher Scientific), and size distribution and degradation were evaluated using the Femto Pulse Genomic DNA 165 kb Kit (Agilent Technologies). Purification steps were performed using AMPure PB beads (PacBio). Fifteen micrograms (15 µg) of DNA were purified then sheared to 15 kb using the Megaruptor 3 system (Diagenode). Using SMRTbell Express Template prep kit 2.0, single-strand overhang removal, DNA damage repair, and END repair steps were performed on 10 µg of DNA. Blunt hairpin adapters were then ligated to the library. The library was treated with an exonuclease cocktail to digest unligated DNA fragments. A size selection step using a 12 kb cutoff was performed on the BluePippin Size Selection System (Sage Science) with "0,75% DF Marker S1 High Pass 15–20 kb" protocol. Using the Binding kit 2.0 kit and sequencing kit 2.0, the primer V2 annealed and polymerase 2.0 bounded library was sequenced by diffusion loading onto 2 SMRT cells on Sequel II instrument (PacBio, USA) at 30 to 70 pM with a 2 hours pre-extension and a 30 hours movie.

Methylation calling. Methylation calling for both species was performed on ONT, EM-seq and WGBS reads. Additionally, methylation was called on PacBio reads for *Coturnix japonica*. The reference assemblies used were Coturnix japonica version 2.0 and Sus scrofa version 11.1, with genome sizes of 0.9 Gb and 2.5 Gb, respectively.

Enzymatic Methyl-seq and Whole Genome Bisulfite Sequencing. Methylation extraction from EM-seq and WGBS reads was performed using the nfcore/methylseq pipeline v1.6.1[10] and Nextflow v21.04.1[11], following the Bismark[12] workflow. For WGBS reads, the *–zymo* option was applied due to the use of the Zymo kit for bisulfite conversion. The maximum insertion length was set to 800 using the *–maxins* parameter. For EM-seq reads, the *–em_seq* option was used. All downstream analyses were performed using the stranded CpG report files, which is an output specified by the *–cytosine_report* option.

Oxford Nanopore Technology. CpG methylation calling from ONT reads was performed using megalodon v2.5.0[13] along with guppy v5.0.17, using res_dna_r941_prom _modbases_5mC_CpG_v001.cfg as calling model with default parameters. Methylation calling was performed on fast5_pass classified reads (quality threshold greater than Q7). Fast5 files contain raw nanopore signal. Megalodon wraps Guppy, which calls methylation during basecalling. Basecalls in fastq format and read alignments can be saved on demand. The CpG methylation calling process for each sequencing run was performed with 3 GPUs (NVIDIA V100) and 40 CPUs. The time taken to complete methylation calling was approximately 12 hours for one PromethION run of *Coturnix japonica*, and approximately 40 hours for one PromethION run of *Sus scrofa*.

PacBio. Primrose software package in SMRT Link v11.0[14] was used to call CpG methylation from HiFi reads. HiFi reads were extracted from the subreads BAM file using the CCS command line tools with the *–hifi-kinetics* option. This generated a novel BAM used by Primrose to perform methylation calling. Modified base positions within the reads, along with their probability of being modified, were added as additional *MM* and *ML* SAM tags to the output BAM file. This BAM file was then mapped to the reference genome using pbmm2. Finally, the modbam2bed script from ONT was used to compute and extract methylation fractions at genomic positions in bedmethyl format.

Sample	Method	Instrument	Run	Raw reads ID	Result files
Coturnix japonica	WGBS	Illumina HiSeq 3000	1	ERR11591925	Seqoccin.Quail.WGBS_HiSeq 1.stranded_CpG_report.txt.gz
			2	ERR11591925	Seqoccin.Quail.WGBS_HiSeq 2_stranded.CpG_report.txt.gz
			3	ERR11591924	Seqoccin.Quail.WGBS_HiSeq 3.stranded_CpG_report.txt.gz
	EM-Seq	Illumina NovaSeq 6000 (1 SP lane)	1	ERR11591927	Seqoccin.Quail.EMSEQ_NovaSeq.stranded_CpG_report.txt.gz
	ONT	PromethION	1	ERR11591487	Seqoccin.Quail.ONT.PAE47187.modified_bases.5mC.bed
			2	ERR11591488	Seqoccin.Quail.ONT.PAE47965.modified_bases.5mC.bed
	PacBio	Sequel II	1	ERR11528759	Seqoccin.Quail.PacBio.203625.ccs.bedmethyl
			2	ERR11528760	Seqoccin.Quail.PacBio.030231.ccs.bedmethyl
Sus scrofa	WGBS	Illumina HiSeq 3000	1	ERR11635334	Seqoccin.Pig.WGBS_HiSeq 1.stranded_CpG_report.txt.gz
			2	ERR11635335	Seqoccin.Pig.WGBS_HiSeq 2.stranded_CpG_report.txt.gz
			3	ERR11635336	Seqoccin.Pig.WGBS_HiSeq 3.stranded_CpG_report.txt.gz
	WGBS	Illumina NovaSeq 6000 (1 SP lane)	1	ERR11635337	Seqoccin.Pig.WGBS_Novaseq.stranded_CpG_report.txt.gz
	EM-Seq	Illumina NovaSeq 6000 (1/4 S4 lane)	1	ERR11635338	Seqoccin.Pig.EMSEQ_Novaseq.stranded_CpG_report.txt.gz
	ONT	PromethION	1	ERR11635339	Seqoccin.Pig.ONT.PAD93746.modified_bases.5mC.bed
			2	ERR11635340	Seqoccin.Pig.ONT.PAE20806.modified_bases.5mC.bed

Table 1. Description of data files available in ENA and Recherche Data Gouv (see the Data Records section).

Results files. For WGBS and EM-seq, the nf-core/methylseq pipeline was used to generate stranded CpG report files from Bismark. These files are in tab-delimited text format and contain information about the methylation status of all cytosines in a given sample, including details such as chromosome name, genomic position, strand information, and counts of cytosines and thymines found after conversion at each genomic position. To calculate CpG position coverage, counts of cytosines and thymines were summed. The methylation fraction at each position was then determined by dividing the number of cytosines by the total number of thymines and cytosines. For ONT and PacBio reads, the output files are in standard bedmethyl format described in https://www.encodeproject.org/data-standards/wgbs/.

Data quality control processing. The resulting files were processed to assess the quality of each methylome dataset per technology. For this, methylation was called separately for each sequencing run (one output file per run) and then merged by technology.

Methylation and coverage calculation of CpG sites. The methylation fraction and CpG coverage were calculated at each position by summing the values from the forward and reverse strands. Positions with a coverage higher than 100X or lower than 5X were filtered out to mitigate potential biases in the calculation of the methylation fraction. This maximum threshold accounts for the fact that it is more than twice the average coverage,helping to address challenges associated with aligning reads to repetitive regions of the genome. The minimum coverage value of 5X was set to ensure a sufficient number of reads for methylation fraction calculation.

Annotation and methylation calculation of CpG Islands. CpG structural annotation is needed for computing the average methylation fraction in CpG islands. In this study, we used the unmasked CpG island region lists available in the UCSC database for both *Coturnix japonica* and *Sus scrofa* reference genomes. Bigbeds were converted to BED format using the UCSC's bigBedToBed tool and are available in our GitHub repository (see Code Availability section). To compare methylation fraction trends between different sequencing technologies at the CpG island level, we focused on regions where at least 40% of CpG sites had a coverage of at least 10X. By considering these specific regions, we calculated the average methylation fraction and compared it across technologies. For correlation analysis per CpG in CpG islands, we retained only the CpG positions that fell within UCSC CpG island regions.

Correlation of CpG methylation measurements across technologies. The evaluation of the correlation of CpG methylation fraction results between the four technologies using the Pearson method aims to provide more insight into the read quality and the reproducibility of the results. When computing Pearson correlation between technologies or technical replicates, only CpG sites or islands present in both datasets were retained through inner joins. The correlation values displayed in the heatmaps were rounded to two decimal places. Barplots above each correlation matrix indicate the proportion of covered reference CpG sites after coverage filtering and before the inner join (see the Code availability section for more information).

Data Records

All raw read files have been uploaded to ENA (European Nucleotide Archive)[15] at www.ebi.ac.uk/ena as part of the SeqOccIn project PRJEB60075[16] and can be accessed under project PRJEB55065[17] for quail and PRJEB55066[18] for pig. The result files have been uploaded to Recherche Data Gouv repository at https://entrepot.recherche.data.gouv.fr as part of the SeqOccIn Dataverse and can be accessed via https://doi.org/10.57745/PIHOIA[19] for the quail sample and https://doi.org/10.57745/EMMOSV[20] for the pig sample. The ENA run accession numbers and result file names are described in Table 1 based on species and sequencing methods.

Sample	Method	Instrument	Run	Raw data							Filtered and pre-processed data						
				#Reads (Millions)	Total bases (Gbp)	Raw coverage (X)	Mean length (bp)	Maximum length (bp)	N50 (bp)	Q20 (%)	% Mapped	% Dups	Total bases after filtering and mapping (Gbp)	% usable	Effective coverage (X)	CpG coverage (X)	
Coturnix japonica	WGBS	Illumina HiSeq 3000	1	572.18	85.83	72	150	150	150	93,68% (R1) 79,62% (R2)	43.8%	51.5%	14.44	16.8%	12	15	42
			2	576.96	86.54	72	150	150	150	94,36% (R1) 80,14% (R2)	44.1%	51.7%	14.56	16.8%	12	14	
			3	739.96	110.99	92	150	150	150	94,96% (R1) 76,78% (R2)	34.5%	50.2%	15.02	13.5%	13	14	
	EM-Seq	Illumina NovaSeq 6000 (1 SP lane)	1	601.76	90.26	75	150	150	150	95,33% (R1) 92,36% (R2)	59.1%	10.9%	42.17	46.7%	35	42	
	ONT	PromethION	1	3.02	42.22	35	13,995	313,987	18,566	30.71%	99.5%	0.0%	31.36	74.3%	26	24	39
			2	1.13	22.78	19	20,140	1,352,716	37,881	50.08%	98.9%	0.0%	20.49	89.9%	17	15	
	PacBio	Sequel II	1	0.50	8.06	7	16,221	39,247	16,371	98.70%	99.8%	0.0%	8.05	99.8%	7	8	28
			2	1.35	22.05	18	16,276	40,634	16,419	98.52%	99.8%	0.0%	22.00	99.8%	18	21	
Sus scrofa	WGBS	Illumina HiSeq 3000	1	523.72	78.56	26	150	150	150	96,81% (R1) 83,06% (R2)	70.5%	49.4%	22.67	28.9%	8	7	23
			2	579.00	86.85	29	150	150	150	96,78% (R1) 83,07% (R2)	69.0%	51.9%	23.44	27.0%	8	7	
			3	443.61	66.54	22	150	150	150	96,44% (R1) 87,7% (R2)	72.0%	55.5%	17.47	26.2%	6	6	
	WGBS	Illumina NovaSeq 6000 (1 SP lane)	1	944.76	141.71	47	150	150	150	96,57% (R1) 95% (R2)	63.7%	69.5%	23.27	16.4%	8	7	
	EM-Seq	Illumina NovaSeq 6000 (1/4 S4 lane)	1	815.46	122.32	41	150	150	150	94,8% (R1) 93,09% (R2)	75.0%	9.0%	74.07	60.6%	25	27	
	ONT	PromethION	1	5.17	81.70	27	15,804	611,639	21,325	50.49%	99.9%	0.0%	74.92	91.7%	25	25	41
			2	5.29	59.96	20	11,239	3,001,578	24,236	42.26%	99.8%	0.0%	48.39	80.7%	16	16	

Table 2. Dataset metrics across technologies. This table compares dataset metrics based on species and sequencing methods. The metrics provide insights into the quality and quantity of the dataset before and after raw data processing. Key metrics evaluated include: Q20(%): Percentage of the dataset with a Phred quality score above 20, indicative of a minimum base call accuracy of 99%. % Mapped: Percentage of reads uniquely aligned to the reference genome from the filtered dataset, reflecting mapping efficiency. % Dups: Percentage of reads identified as duplicates post-alignment, indicated dataset redundancy. % Usable: Percentage of data retained after quality filtering, mapping, and deduplication. Effective coverage (X): Mean depth of coverage across the dataset after quality filtering, mapping, and deduplication. CpG coverage (X): Mean depth of coverage at CpG sites, providing insight into sequencing uniformity and epigenetic analysis potential. Note: Coverage (X) represents the average number of times a nucleotide is sequenced, estimating dataset completeness and depth.

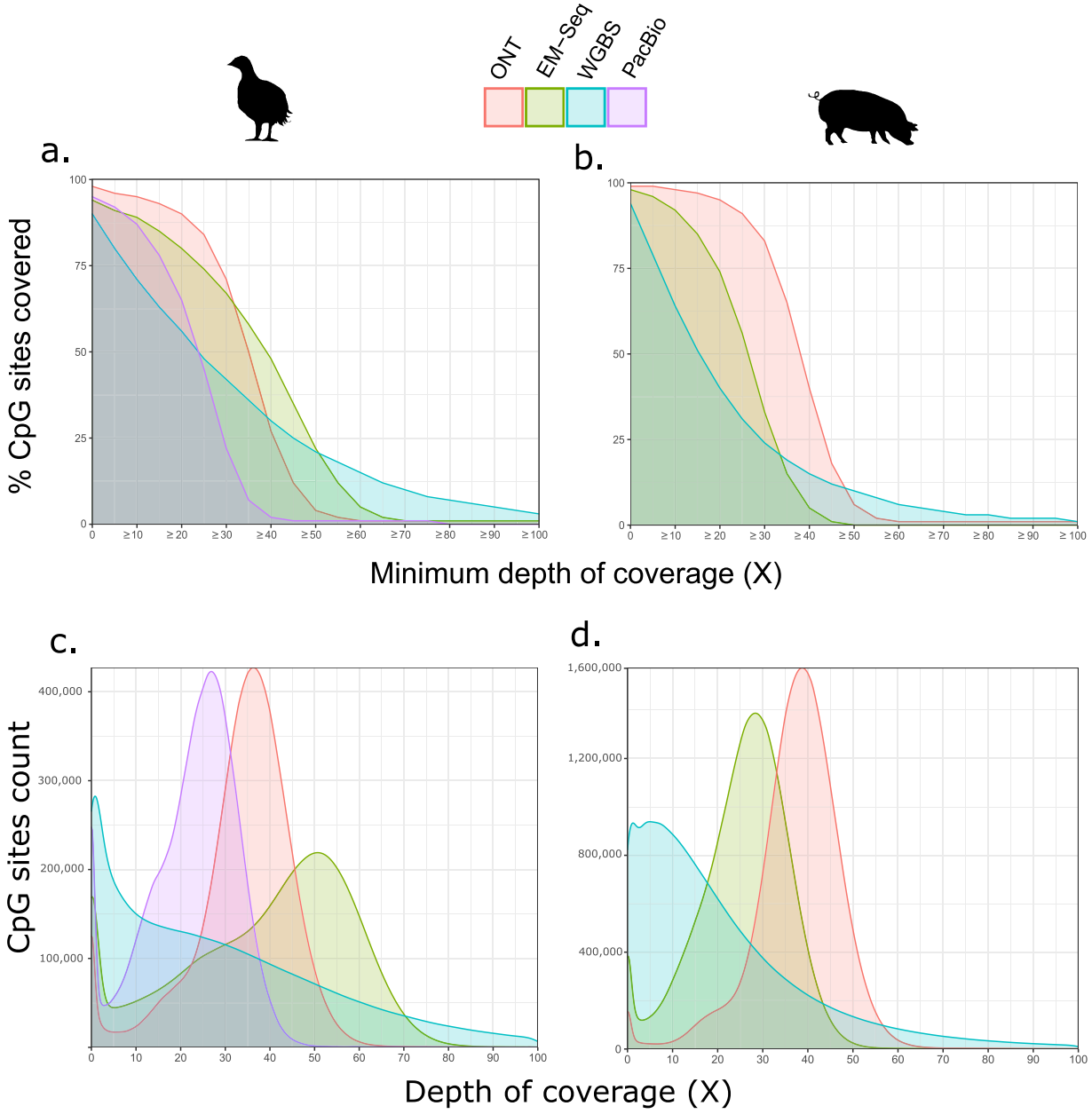

Fig. 2 Coverage of CpG sites across technologies. The percentage of CpG sites covered at a minimum depth of coverage (**a, b**) and the distribution of CpG site coverage (**c, d**) for ONT (red), EM-seq (green), WGBS (blue), and PacBio (purple) sequencing of *Coturnix japonica* (**a, c**) and *Sus scrofa* (**b, d**) samples.

Technical Validation

Data quality control. To assess raw data quality, FastQC[21] was used for Illumina short reads, MinIONQC[22] for Nanopore long reads, and SMRT Link v10.1[14] for PacBio long reads. Additionally, data metrics presented in Table 2 were obtained using seqkit[23]. The objective was to guarantee sufficient data quality for accurate methylome analysis and to enable comparability between the methods by achieving a similar CpG site coverage across sequencing technologies, with a minimum depth of coverage of 30X.

Enzymatic Methyl-seq and Whole Genome Bisulfite Sequencing. Achieving 30X CpG site coverage with bisulfite sequencing was challenging due to the substantial number of duplicates (filtered out during processing) and the low mapping rate. Table 2 shows that the usable amount of WGBS reads was significantly lower compared to other methods. For the quail, this value ranged from 13.5% to 16.8%, resulting in a total effective CpG coverage of 42X for the 3 pooled lanes (Table 2). For the pig, the number of usable reads was higher than for the quail, but still insufficient to compensate for the genome size difference. That's why an additional NovaSeq run was produced for the pig, in addition to the 3 HiSeq runs. Overall, the percentage of usable reads is lower for WGBS compared to EM-seq. Only one run of EM-seq was performed per species. On average, the quail sample achieved

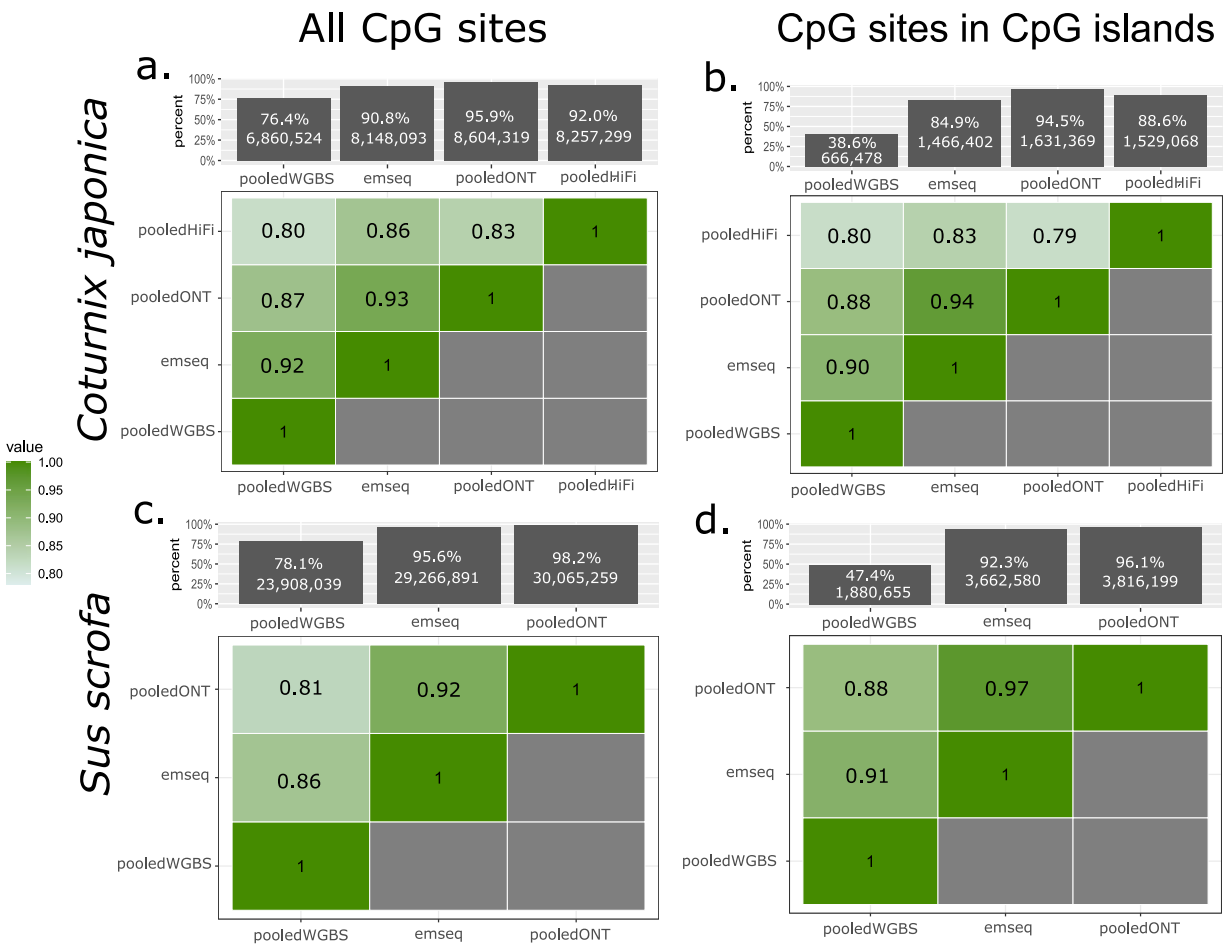

Fig. 3 Correlation matrix of CpG per-site methylation levels across technologies. (**a, c**) Correlation analysis using all shared CpG positions. (**b, d**) Correlation analysis focusing on shared CpG sites located within CpG islands. Barplots above each matrix show the number of CpG sites covered and their percentage compared to the reference genome for each technology. Panels (**a**) and (**b**) report results for quail DNA, whereas (**c**) and (**d**) detail findings for pig DNA.

42X CpG coverage, and the pig achieved 27X. EM-Seq duplicate rates were lower than WGBS (around 10%), and the mapping percentage was better with EM-seq, likely due to the enzymatic reaction preserving DNA integrity without damage or degradation. These observations are consistent with results obtained in a previous article[24] comparing bisulfite and EM-seq in *Arabidopsis thaliana*, as well as an article[1] from the EM-seq kit provider on human DNA and by Hubert *et al.*[25], showing that enzymatic methyl-seq consistently outperforms the bisulfite-based standard in capturing candidate regions for genomic imprinting in the pig.

Oxford Nanopore Technology. Quail and pig samples achieved CpG coverage of 39X and 41X, respectively, with only two PromethION runs (Table 2). The percentage of reads remaining after quality filtering (Q7) and mapping ranged from 74% to 92%, which was significantly higher compared to both EM-seq and WGBS. It is interesting to note that fewer reads were produced by the quail samples compared to the pig samples. This decrease in yield for bird genomes has already been observed with ONT sequencing, as described in this provider's presentation[26]. It is believed to be caused by the DNA conformation, which can obstruct the pores. Despite performing a nuclease wash to empty blocked pores, quail runs still had a notably lower read count, particularly the second run, with approximately 3-4 times fewer reads than pig runs.

PacBio. PacBio HiFi reads were only generated for the quail sample. The first run resulted in a limited number of reads (around 8 Gbp). However, with improved SMRTCell loading, the second run produced a higher number of reads (around 22 Gbp). Overall, pooling the data from these two runs yielded a CpG coverage of 28X.

CpG sites coverage for each method. Figure 2 illustrates the coverage of CpG sites across different technologies using two different approaches. The cumulative distribution (a, b) provides information on the percentage of CpG sites covered based on the minimum depth threshold. Typically, a minimum depth of 10X is used for accurate methylation fraction calculations. At this threshold, the graph shows that the WGBS approach covers less

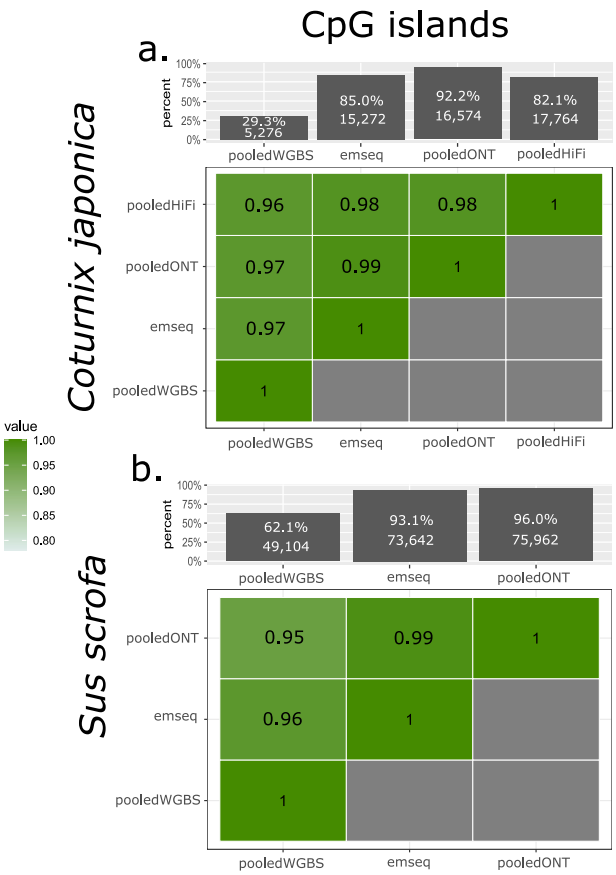

Fig. 4 Correlation matrix of CpG islands methylation levels across technologies. Only CpG islands with at least 40% of CpG sites covered by ten or more reads were included in the correlation analysis. Barplots above each matrix display the number of CpG islands meeting this criterion and the percentage of regions covered compared to the reference genome. (**a**) depicts results for quail DNA, while (**b**) presents results for pig DNA.

than 75% of CpG sites for both quail and pig samples, whereas EM-Seq and PacBio cover around 90%, and ONT achieves close to 100% coverage. The second part (c, d) demonstrates that the coverage distribution at CpG sites is narrower for long-read technologies (PacBio and ONT), as well as for the EM-Seq technology, compared to the WGBS method. Notably, WGBS exhibits significant heterogeneity in CpG site coverage, along with a high proportion of CpG sites uncovered, indicating a substantial alignment bias along the genome.

Methylation calling quality control. To assess the quality of ONT and PacBio datasets, methylation detection results were compared with those obtained from WGBS and EM-Seq assays at different levels (single CpG and CpG island regions). Figures 3 and 4 show methylation correlation matrices between technologies for both quail and pig samples. Correlations were calculated using three filters (columns): 1. All shared CpG positions in results (Fig. 3a,c "all CpG sites"); 2. All CpG sites falling into a CpG island (Fig. 3b,d "CpG sites in CpG islands"); 3. The average methylation fraction per CpG island (Fig. 4) (explained in the dedicated method subsection). Above the correlation matrices, barplots with the percentage of CpG covered in the reference genome for each technology and each filter can be found. Figure 5 shows correlation matrices between runs for all sequencing technologies, enabling the comparison of ONT and PacBio with WGBS and EM-Seq reproducibility. As in Figs. 3 and 4, barplots located above the matrices present CpG coverage fractions of the reference genome as well as the number of CpG sites covered per sequencing run.

Correlation of CpG methylation measurements between methods. A correlation of 0.93 was found between ONT and EM-Seq fractions for the quail sample, while the correlation between ONT and WGBS was 0.87, and that between PacBio and EM-Seq was 0.86, and PacBio and WGBS was 0.80 (Fig. 3a). Furthermore, the ONT method demonstrated the highest percentage of CpG sites covered, with 95.9%, in comparison to the WGBS method, which achieved a coverage of only 76.4% (Fig. 3a). Looking at the coverage of CpG sites in CpG islands (Fig. 3b,d), WGBS sequencing was found to poorly cover many of these genomic regions, especially for *C. japonica,* with only 38.6% of covered CpG sites in CpG islands, but the coverage was also suboptimal for the pig sample with 47.4%. It has indeed been demonstrated by Feng *et al.*[24] that high GC% regions were tedious to sequence with WGBS, which is the case of CpG islands. Other technologies did not show this bias.

Overall, the results for *Sus scrofa* were consistent with those of *Coturnix japonica*, showing correlations between ONT and EM-Seq ranging from 0.92 (Fig. 3c) to 0.97 (Fig. 3d). The exception was the pig ONT reads'

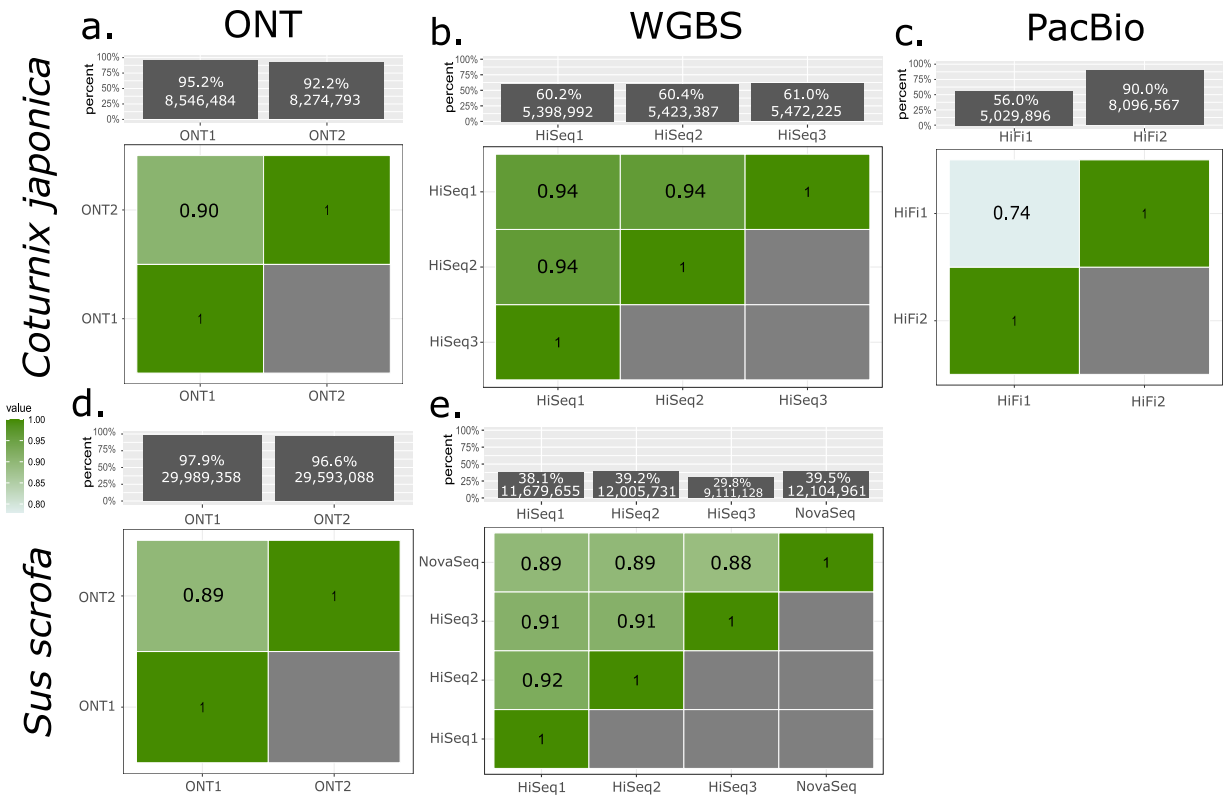

Fig. 5 Correlation matrix of CpG per-site methylation levels between technical replicates. Correlation was calculated using all shared CpG positions. Barplots above each matrix show the number of CpG positions covered by runs and the percentage of positions covered compared to the reference.

methylation fractions, which exhibited less correlation than quail ONT reads when compared to WGBS controls (0.81 versus 0.87) (Fig. 3a,c). However, we are confident that the methylation prediction for pig ONT reads is at least of the same quality as *Coturnix japonica*, considering the WGBS control was of poor quality, which was not the case for the EM-seq control. ONT and EM-Seq were the only combination where the correlation was consistently close to 1 in all approaches. Interestingly, it was also the only combination where the correlation of CpG sites in CpG islands (Fig. 3d) was better than the per CpG site approach (Fig. 3c) for the pig sample.

For both species, we can see that both ONT and PacBio (in the case of *C. japonica*) had systematically higher correlation with EM-seq than with WGBS, suggesting that EM-Seq is a more accurate and reliable detection method than WGBS, potentially serving as a better reference to represent the short-read sequencing approach for long-read method benchmarking.

Finally, when considering the average methylation fractions in CpG islands, all approaches showed a correlation close to 1 (0.95 to 0.99), indicating their ability to accurately assess methylation levels within CpG islands (Fig. 4). Specifically, PacBio shows a higher correlation with other sequencing technologies when comparing the mean methylation fractions of CpG islands to the methylation fractions at individual CpG sites, suggesting a potential precision shortfall in site-specific methylation detection.

It's important to note that the highest correlation is still obtained between ONT and EM-Seq.

Correlation of CpG methylation measurements between runs. Figure 5 shows correlations between runs of the same technology.

The correlation between ONT runs for both pig and quail is close to 0.90 (Fig. 5a,d), which is comparable to the observed correlation with WGBS (0.88 to 0.94) (Fig. 5b,e). Notably, the correlation between HiSeq runs and the NovaSeq run is slightly lower (0.88 to 0.89) compared to the correlation between HiSeq runs (0.91 to 0.94).

The correlation between PacBio runs for quail was 0.74 (Fig. 5c). One of these runs had exceptionally low CpG coverage compared to the other, which could explain the low correlation and highlights the importance of sufficient coverage when analysing vertebrate CpG methylation data.

EM-seq reproducibility results were not calculated because only one run was enough to obtain 40X average coverage.

We expected higher correlation values when comparing runs of the same technology. This is particularly true for WGBS, where they are very low for the pig. This can be attributed to the low coverage of CpG sites per run, resulting in less accurate methylation fraction values compared to those obtained by pooling all datasets per technology to achieve 30–40X of coverage per CpG site. For quail, the CpG coverage per run is

approximately 15x, whereas for pig it is only 7x. The data are filtered with a minimum of 5x to calculate fractions, but a greater depth would allow more accurate fractions to be calculated. The correlation values slightly increase with a minimum depth of 10X (+0.01 to +0.03 on average), but applying this filter would have drastically reduced the number of CpGs covered. In any case, this correlation between technical replicates allowed us to assess whether the data could be pooled according to the ENCODE recommendations (https://www.encodeproject.org/data-standards/wgbs/).

Code availability

Comparative analysis from methylation calling results were performed using R 3.6.3 with Biostring (2.54.0), data.table (1.13.6) and ggplot2 (3.3.3) packages. The egg (0.4.5) and scales (1.1.1) packages were used for heatmap panels. All scripts for preprocessing and analysis are available on the github page https://github.com/GeTPlaGe/SeqOccIn/tree/main/Data%20paper/Epigenetics%20data%20paper as markdown files.

Received: 2 July 2024; Accepted: 6 March 2025;
Published online: 01 April 2025

References

1. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects dna methylation at single-base resolution from picograms of dna. *Genome Res.* **31**, 1280–1289, https://doi.org/10.1101/gr.266551.120 (2021).
2. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410, https://doi.org/10.1038/nmeth.4184 (2017).
3. Ni, P. *et al.* DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* **35**, 4586–4595, https://doi.org/10.1093/bioinformatics/btz276 (2019).
4. Liu, Y. *et al.* DNA methylation-calling tools for oxford nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol.* **22**, https://doi.org/10.1186/s13059-021-02510-z (2021).
5. Yuen, Z. W.-S. *et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun.* **12**, https://doi.org/10.1038/s41467-021-23778-6 (2021).
6. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting dna methylation from long-read sequencing of human genomes. *Genome Biol.* **25**, https://doi.org/10.1186/s13059-024-03207-9 (2024).
7. Morris, K. M. *et al.* The quail genome: insights into social behaviour, seasonal biology and infectious disease response. *BMC Biol.* **18**, https://doi.org/10.1186/s12915-020-0743-4 (2020).
8. Mills, A. D. & Faure, J.-M. Divergent selection for duration of tonic immobility and social reinstatement behavior in japanese quail (*Coturnix coturnix japonica*) chicks. *J. Comp. Psychol.* **105**, 25–38, https://doi.org/10.1037/0735-7036.105.1.25 (1991).
9. Roussot, O. *et al.*Aflp linkage map of the japanese quail *Coturnix japonica*. *Genet. Sel. Evol.* **35**, https://doi.org/10.1186/1297-9686-35-6-559 (2003).
10. Ewels, P. *et al.* nf-core/methylseq: nf-core/methylseq version 1.6.1 [nauseous serpent], https://doi.org/10.5281/zenodo.4744708 (2021).
11. Ewels, P. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278, https://doi.org/10.1038/s41587-020-0439-x (2020).
12. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572, https://doi.org/10.1093/bioinformatics/btr167 (2011).
13. *Oxford nanopore technologies: Megalodon.* https://nanoporetech.github.io/megalodon/ (2019).
14. Pacbio: Smrt link. https://www.pacb.com/support/software-downloads/.
15. Leinonen, R. *et al.* The european nucleotide archive. *Nucleic acids research* **39**, D28–D31, https://doi.org/10.1093/nar/gkq967 (2010).
16. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB60075 (2023).
17. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB55065 (2023).
18. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB55066 (2023).
19. Terzian, P. & Vandecasteele, C. CpG methylation calling of the quail genome, https://doi.org/10.57745/PIHOIA (2024).
20. Terzian, P. & Vandecasteele, C. CpG methylation calling of the pig genome, https://doi.org/10.57745/EMMOSV (2024).
21. Andrews, S. *et al.* Fastqc: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2012).
22. Lanfear, R., Schalamun, M., Kainer, D., Wang, W. & Schwessinger, B. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics* **35**, 523–525, https://doi.org/10.1093/bioinformatics/bty654 (2018).
23. Shen, W., Le, S., Li, Y. & Hu, F. Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. *PloS one* **11**, e0163962, https://doi.org/10.1371/journal.pone.0163962 (2016).
24. Feng, S., Zhong, Z., Wang, M. & Jacobsen, S. E. Efficient and accurate determination of genome-wide dna methylation patterns in arabidopsis thaliana with enzymatic methyl sequencing. *Epigenetics & Chromatin* **13**, https://doi.org/10.1186/s13072-020-00361-9(2020).
25. Hubert, J.-N. *et al.* Detection of dna methylation signatures through the lens of genomic imprinting. *Sci. Reports* **14**, https://doi.org/10.1038/s41598-024-52114-3 (2024).
26. Harrington, E. From mixed bags to magic bullets: solving challenging problems in genomics. https://nanoporetech.com/resource-centre/eoghan-harrington-mixed-bags-magic-bullets-solving-challenging-problems-genomics (2018).

Acknowledgements

We thank members of the SeqOccIn Consortium for their feedback before and during this study. We thank *La Région Occitanie* and the European Union for financing project in the call for projects "Regional Platforms of Research and Innovation" of the Occitanie region, under the Operational Program FEDER-FSE MIDI-PYRENEES ET GARONNE 2014–2020 and SYSAAF, ADISSEO, and Alliance R&D for their contributions to the financing of these analyses. We thank the INRAE experimental units of GENESI (https://doi.org/10.15454/1.5572415481185847E12) and Nouzilly (PEAT, https://doi.org/10.15454/1.5572326250887292E12) for producing the samples and the SIGENAE team for carrying out part of the analyses.

Author contributions

C.D., D.M. and Ch.G. conceived and supervised the "SeqOccIn" project. N.I., S.L., K.F. and M.B. collected samples and carried out DNA extraction. J.L. and R-F.S. optimized and performed the sequencing. J.S. collected the data. P.T. performed the methylation calling for ONT and PacBio data and the post-processing of the result files. C.V. performed the methylation calling for WGBS and EM-Seq data. Ch.K. supervised with C.V. the bioinformatic analysis. C.I., C.D., C.V., Ch.K., J.D., F.P., conceived the experimental design and supervised the technical aspects of the study. C.V. and Cl.K. submitted the data. P.T. and C.V. wrote the article and revised the manuscript with Ch.K. All authors read and approved the final manuscript. P.T. and C.V. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.V.

Reprints and permissions information is available at www.nature.com/reprints.

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