



# OPEN Long-read sequencing disentangles isoform complexity at allele-specific loci

Lison Lemoine<sup>1,2</sup>, Sarah Hoelzl<sup>1,2</sup>, Tim P. Hasenbein<sup>1,2</sup>, Elisabeth Graf<sup>3</sup> & Daniel Andergassen<sup>1,2</sup>✉

In recent years, long-read sequencing technologies have detected transcript isoforms with unprecedented accuracy and resolution. However, it remains unclear whether long-read sequencing can effectively disentangle the isoform landscape of complex allele-specific loci that arise from genetic or epigenetic differences between alleles. Here, we combine the PacBio Iso-Seq workflow with the established phasing approach WhatsHap to assign long reads to the corresponding allele in polymorphic F1 mouse hybrids. Upon comparing the long-read sequencing results with matched short reads, we observed general consistency in the allele-specific information and were able to confirm the imprinting status of known imprinted genes. We then explored the complex imprinted *Gnas* locus known for allele-specific non-coding and coding isoforms and were able to benchmark historical observations. This approach also allowed us to detect isoforms from both the active and inactive X chromosomes of genes that escape X chromosome inactivation. The described workflow offers a promising framework and demonstrates the power of long-read transcriptomic data to provide mechanistic insight into complex allele-specific loci.

Mammals are diploid and, therefore, have two copies of each chromosome, one inherited from each parent. Given that both gene copies on the parental alleles are regulated independently, genes can be expressed from both alleles, known as biallelic expression, or from one allele, known as allele-specific expression. Allele-specific expression can arise due to genetic or epigenetic differences between the alleles, with a substantial proportion attributed to genetic differences<sup>1–3</sup>. These allelic genetic variations have been demonstrated to induce allele-specific expression by influencing transcription factor binding or impacting post-transcriptional processes in an allele-specific manner<sup>3–5</sup>. Genomic imprinting is a prominent example of allele-specific expression arising from epigenetic differences between alleles<sup>6</sup>. Mechanistically imprinted genes are typically clustered, with allele-specific silencing regulated by differential DNA methylation<sup>6,7</sup>. For several imprinted clusters, the unmethylated imprint control element acts as a promoter for a long non-coding RNA (lncRNA) that silences nearby genes by targeting histone-modifying complexes<sup>7,8</sup>. These genes play a crucial role in various biological processes, such as fetal development and homeostasis, and their dysregulation is associated with several diseases<sup>9,10</sup>. Another classic example of an epigenetic phenomenon leading to allele-specific expression is X chromosome inactivation (XCI), a process critical for dosage compensation between the sexes<sup>11</sup>. During XCI, one of the female X chromosomes undergoes epigenetic silencing facilitated by the lncRNA *Xist*, which is expressed exclusively from the future-inactive X chromosome<sup>12,13</sup>. Interestingly, some genes can escape this process and are consequently expressed from both X chromosomes<sup>14–16</sup>.

Current approaches for mapping allele-specific expression primarily rely on short-read sequencing data<sup>1,17–20</sup>. In mice, these approaches depend on single nucleotide polymorphisms (SNPs) between genetically distinct inbred strains to assign sequencing reads to the corresponding parental allele<sup>2</sup>. This strategy allowed for a nearly complete map of the imprinted landscape, spanning a comprehensive set of murine organs and developmental stages<sup>5,18,21,22</sup>. A limitation of short-read sequencing is its inability to accurately quantify entire allele-specific isoforms, a characteristic of imprinted loci, which often harbor multiple overlapping allele-specific isoforms. A prime example of such a complex locus is the imprinted *Gnas* cluster, where the same gene locus is transcribed into distinct biallelic, maternal, and paternal isoforms<sup>23,24</sup>. Dysregulation of the *Gnas* cluster can lead to growth defects and pseudohypoparathyroidism<sup>23</sup>. Hence, investigating the isoform landscape of disease-relevant loci displaying diverse isoform profiles can illuminate their underlying mechanisms and functions. Recent advances

<sup>1</sup>Institute of Pharmacology and Toxicology, Technical University of Munich, Munich, Germany. <sup>2</sup>DZHK (German Centre for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich, Germany. <sup>3</sup>Institute of Human Genetics, Klinikum Rechts Der Isar, Technical University of Munich, Munich, Germany. ✉email: daniel.andergassen@tum.de

in long-read sequencing technologies, such as the Oxford Nanopore and Pacific Biosciences (PacBio) platforms, have allowed for transcriptomic analysis with a high resolution, span, and accuracy<sup>25,26</sup>. Long-read sequencing can span challenging genomic regions and capture full-length transcripts, enabling a comprehensive gene expression analysis<sup>27,28</sup>. However, allele-specific analysis of long-read data poses unique challenges due to the need for specialized bioinformatic approaches that differ from those used for short-read sequencing analysis, as recently shown in a large allele-specific human study<sup>29</sup>. Here, we present a user-friendly framework for analyzing allele-specific expression using long-read sequencing. To accomplish this, we analyzed the whole brain of a highly polymorphic F1 mouse by integrating the PacBio Iso-Seq pipeline with the established phasing approach WhatsHap<sup>30,31</sup>. This strategy allowed us to capture and quantify the allele-specific isoform diversity of protein-coding genes and lncRNAs. Upon comparing the long-read sequencing results with organ- and background-matched short-read sequencing data, we observed a general consistency in the allele-specific expression data across most genes. We further highlight the effectiveness of the described workflow by benchmarking historical observations at the imprinted *Gnas* locus and known XCI escape genes. In summary, the described workflow offers a promising framework for future studies aiming to shed light on loci with allele-specific isoform diversity.

## Results

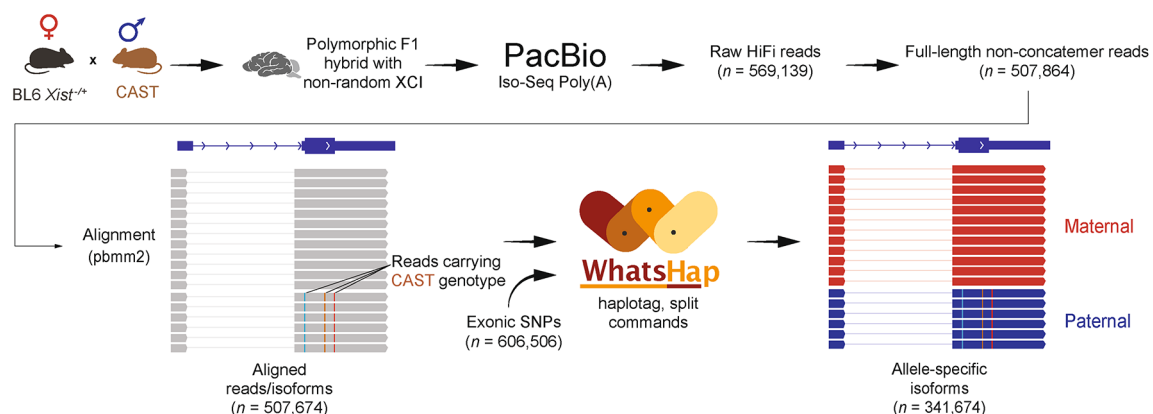
### Assigning long reads to the corresponding allele

To capture the allele-specific expression landscape through long-read sequencing, we bred C57BL/6J (BL6) females heterozygous for a deletion of the proximal A-repeat of *Xist* (*Xist*<sup>-/-</sup>) with CAST/EiJ (CAST) males (Fig. 1)<sup>32</sup>. Next, we collected a whole adult brain from a resultant F1 female hybrid (BL6<sup>ΔXist</sup> × CAST) and applied the PacBio technology-based Iso-Seq workflow. This poly-A selection-based method allows for a comprehensive exploration of isoform discovery at the gene level in the transcriptome (see "Methods"). The resulting dataset enables the detection of allele-specific expression for autosomal genes and X-linked genes due to the *Xist* deletion, preventing XCI of the maternal BL6 allele and resulting in a fully skewed XCI pattern. A total of 569,139 HiFi reads were recovered, with an average read length of 2836 base pairs. These reads were further processed and filtered, resulting in 507,864 full-length non-concatemer reads, defined as reads that contain all essential transcript features, including both 5' and 3' primers, a poly-A tail, and the full insert sequence with no truncations or internal priming artifacts. The long reads were aligned to the reference genome using pbmm2, a PacBio wrapper of Minimap2, resulting in 507,674 uniquely aligned reads<sup>33</sup>.

To investigate allele-specific expression using long-read sequencing, we applied the established phasing method, WhatsHap, using the haplotag and split subcommands<sup>30,31</sup> (see "Methods" and Supplementary Methods). Using available exonic SNP information between BL6 and CAST ( $n = 606,506$ )<sup>2</sup>, the WhatsHap tool was used to assign long reads to maternal or paternal alleles. A single SNP was sufficient to map long reads to the corresponding alleles, as shown for the paternally expressed imprinted gene *Ndn* (Supplementary Fig. 1). Of the 341,674 tagged long reads, 52.4% ( $n = 179,197$ ) were assigned to the maternal BL6 allele, and 47.5% ( $n = 162,477$ ) were assigned to the paternal CAST allele (Fig. 1), covering 13,732 genes. A minimum cutoff of 10 reads overlapping at least one SNP per gene was applied, resulting in 6203 informative genes (coding genes = 6123, non-coding genes = 80), including 6009 autosomal and 194 X-linked genes (Fig. 2a, Supplementary Table S1, sheet a). In summary, we combined the PacBio Iso-Seq workflow with established phasing methods, allowing us to explore the allele-specific expression landscape in highly polymorphic F1 hybrids using long-read sequencing.

### The allelic status shows general consistency between short- and long-read sequencing, with the highest confirmation for imprinted genes

After implementing our long-read sequencing approach for allele-specific expression analysis, we focused on validating the allelic long-read landscape. To accomplish this, we conducted a comparative analysis with



**Fig. 1.** Integrating PacBio Iso-Seq with WhatsHap for allele assignment in a polymorphic F1 mouse hybrid. The raw reads obtained from long-read sequencing are filtered using the Iso-Seq pipeline. These filtered reads are called full-length non-concatemer reads and are aligned using PacBio's minimap2 wrapper. Using known exonic SNPs between BL6 and CAST<sup>2</sup>, long reads are assigned to the corresponding allele (maternal or paternal).

a background-matched short-read dataset, utilizing an adult F1 brain sample in the BL6xCAST genetic background<sup>18</sup>. We conducted an in-depth analysis using the established Allelome.PRO approach, a pipeline for defining allele-specific expression in short-read sequencing<sup>1</sup>, with 24,716,728 uniquely aligned short reads as input (Fig. 2a, see "Methods"). We recovered allelic information for 13,476 autosomal genes (reads overlapping a SNP  $\geq 20$ ) and found that of the 6009 autosomal genes detected by long-read sequencing, 99% ( $n=5938$ ) overlapped (Fig. 2a, Supplementary Table S1, sheet a-c). Among the genes that overlapped, most autosomal genes exhibited biallelic expression with both long and short reads (Fig. 2b). We used the previously defined allelic ratio cutoff of  $\leq 0.3$  or  $\geq 0.7$  to classify expression as paternal or maternal, respectively, while all other cases were classified as biallelic, a threshold shown to effectively distinguish known imprinted genes from biallelically expressed genes<sup>1</sup>. Within the 5204 and 5771 biallelic genes identified in the long-read and short-read datasets, respectively, 98% ( $n=5102$ ) overlapped. (Fig. 2c and d). However, there was a notable difference in the proportion of allele-specific genes. For short reads, 2.8% ( $n=167$ ) of the genes showed allele-specific expression, while a higher proportion of genes, 12.4% ( $n=734$ ), was detected in the long-read dataset, with 65 genes present in both datasets (Fig. 2c and d, Supplementary Table S1, sheet d). Of these shared 65 allele-specific genes, the majority had the same allelic direction ( $n=58$ , 89.23%). Among these, 15 are known imprinted genes in the brain whose imprinted status was correctly determined<sup>5</sup>. The remaining seven genes showed different patterns due to the following observations. *Trappc9* suffers from the lack of strand specificity in short-read data due to its overlap with the paternally expressed imprinted gene *Peg13*. *Gm5424* is a highly repetitive retroposed gene, which may introduce sequencing and alignment bias. The remaining genes (*Sdccag8*, *Sox8*, *Ogdhl*, *Csnk2a1*, and *Churc1*) can be explained as genes expressed slightly below or above the allelic ratio cutoff. Overall, the correlation of the allelic ratio for all informative genes shared between short- and long-read sequencing was positive but relatively weak ( $r=0.212$ ,  $p=2.2 \times 10^{-16}$ , Fig. 2e). However, we observed a markedly high correlation ( $r=0.989$ ,  $p=2.989 \times 10^{-12}$ ,  $n=15$ ) for known imprinted genes in the brain (Fig. 2e). To illustrate the power of this approach, we visualized the biallelically expressed gene *Ube2g2* ( $n=29$  maternal and 28 paternal long reads) and the paternally expressed imprinted gene *Impact* ( $n=2$  maternal and 217 paternal long reads, Fig. 2f). This approach also enabled us to correctly assign the allele-specific status of lncRNAs, including the biallelically expressed lncRNA *Malat1* ( $n=598$  maternal and 705 paternal long reads) and the maternally expressed imprinted lncRNA *Meg3* ( $n=737$  maternal and 0 paternal long reads, Fig. 2g). In conclusion, we observed general agreement in allelic status between short- and long-read sequencing data, with the highest level of confirmation for biallelic and imprinted genes.

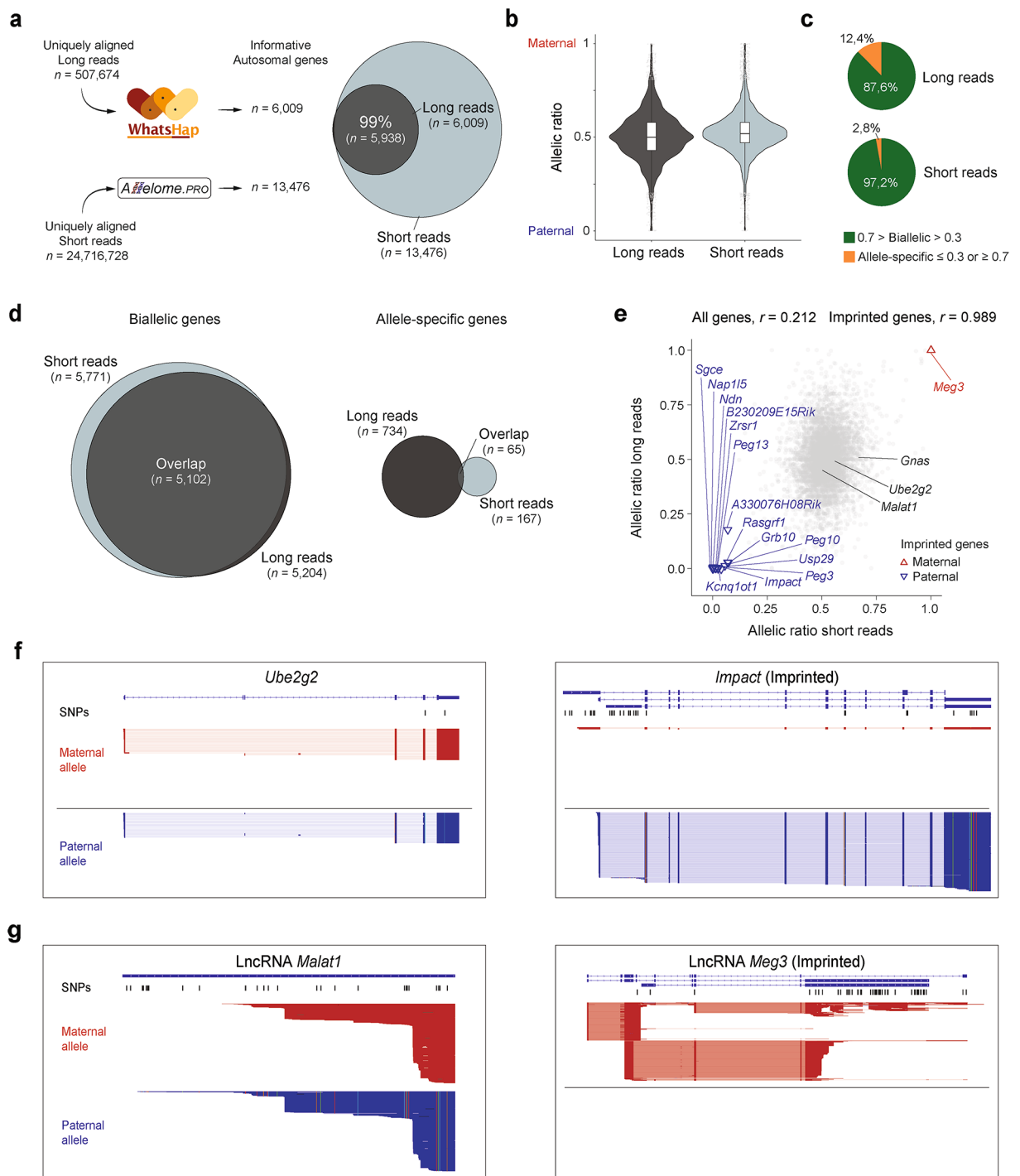
#### Allele-specific long-read analysis of the *Gnas* cluster benchmarks historical observations

To further highlight the strength of the allele-specific long-read analysis, we conducted a detailed inspection of the imprinted *Gnas* cluster. At least 50 different transcripts derived from alternative promoters have been identified in the *Gnas* cluster with multiple allele-specific isoforms<sup>23,24,34</sup>. The most notable *Gnas* gene products include the biallelically expressed *Gsa* isoform, the paternally expressed isoforms *Gnasxl* and *Ex1A*, and the maternally expressed *Nesp* isoform. In addition, the *Gnas* locus encodes the paternally expressed antisense lncRNA *Nespas* (Fig. 3a)<sup>23</sup>.

In our long-read data, we detected and recovered allele-specific information from three isoforms of the *Gnas* cluster, namely, *Gsa*, *Nesp*, and *Ex1A* (Fig. 3b). The *Nespas* and *Gnasxl* isoforms are not expressed in the adult brain, supported by short- and long-read sequencing data (Fig. 3b). We identified *Gsa* as biallelically expressed with 117 maternal and 98 paternal reads, consistent with previous observations (Fig. 3c)<sup>23</sup>. The imprinted *Nesp* isoform was correctly identified as maternally expressed, with all six reads coming from the maternal allele. Similarly, we confirmed paternal expression of *Ex1A* by detecting all 26 reads exclusively from the paternal allele. In summary, our allele-specific long-read analysis of the *Gnas* cluster validates previous findings and underscores the effectiveness of this approach in providing mechanistic insights within imprinted loci characterized by overlapping allele-specific isoforms.

#### Allele-specific long-read sequencing enables the detection of escape gene isoforms from both the active and inactive X chromosomes

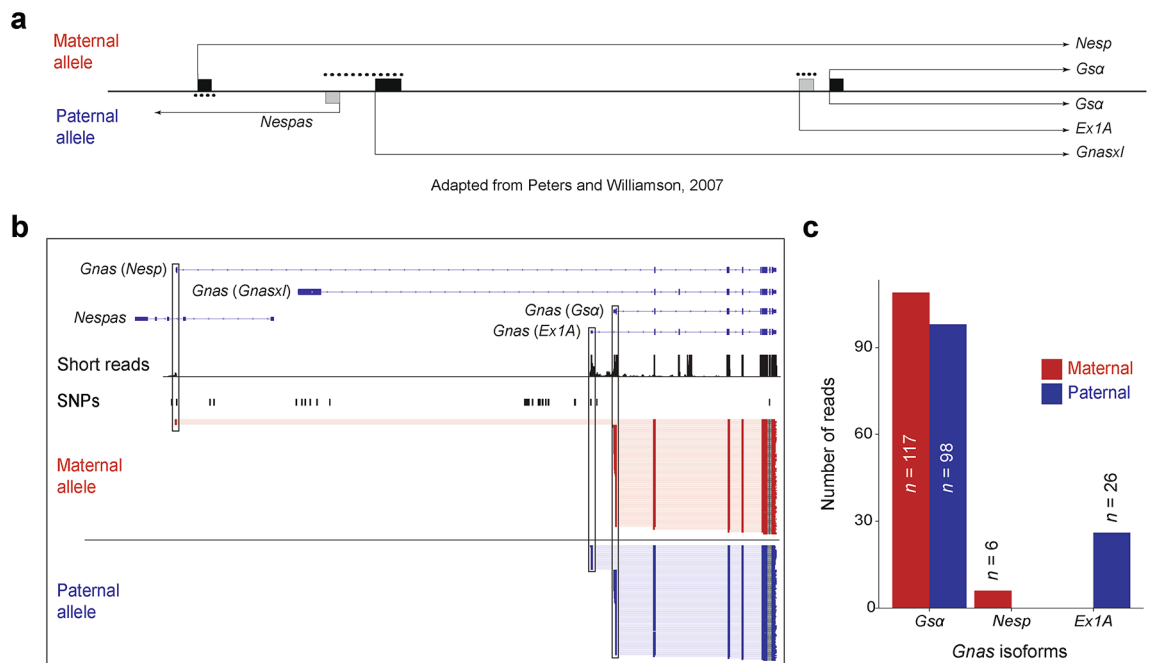
We then examined the allele-specific long-read landscape on the X chromosome by using our skewed XCI model caused by the *Xist* deletion (BL6 <sup>$\Delta$ Xist</sup>  $\times$  CAST), which results in the expression of the maternal BL6 X chromosome (see "Methods")<sup>32</sup>. For all informative X-linked genes, we observed 12,623 long reads aligned to the X chromosome. The vast majority originated from the maternal active X chromosome, with only 69 (0.54%) aligning to the paternal inactive X chromosome. Of these, 39 reads correspond to *Xist* and the two known escape genes, *Ddx3x* and *Kdm5c*. As anticipated by our model, *Xist* was found to be exclusively expressed from the inactive X chromosome ( $n=10$  paternal long reads, Fig. 4a), resulting in maternal expression of the majority of X-linked genes ( $n=191/194$ , allelic ratio  $> 0.9$ , Supplementary Table S1, sheet a). For instance, the *Ubqln2* gene is subject to XCI and displays exclusively maternal reads ( $n=72$  maternal and 0 paternal reads), which is in line with the findings in previous reports<sup>30,35</sup> (Fig. 4b). Furthermore, we examined the isoforms of well-established genes known to escape XCI, such as *Ddx3x* and *Kdm5c*<sup>36</sup> (Fig. 4c). As expected, we detected biallelic expression of these escape genes. Specifically, for *Ddx3x*, we identified 36 maternal reads and 18 paternal reads, while for *Kdm5c*, we observed 26 maternal reads and 11 paternal reads. This approach also sheds light on whether isoform differences can be detected depending on whether a gene is transcribed from the active or the epigenetically silenced X chromosome. For the two escape genes detected in our long-read data, no apparent isoform differences were observed between the active and inactive X chromosomes (Fig. 4c). In conclusion, the described workflow demonstrates the feasibility of detecting escape genes using long-read sequencing and enables the investigation of whether the inactive X chromosome influences isoform patterns.



## Discussion

In this study, we combined PacBio long-read sequencing with an established phasing approach, WhatsHap, to capture the allele-specific expression landscape of autosomal and X-linked genes<sup>26,30,31</sup>. To this end, we performed long-read sequencing on an adult brain isolated from a highly polymorphic F1 mouse, allowing us to use SNPs to assign long reads to the corresponding allele. To benchmark the allele-specific long-read results, we compared them to short-read sequencing analysis from a background-matched brain<sup>18</sup>. Overall, we observed a high degree of overlap for biallelic and imprinted genes, representing the largest proportion of informative genes. In contrast, we observed a notable difference in the proportion of allele-specific genes, likely driven by genetic differences between the alleles. Specifically, a higher proportion of allele-specific genes was found in the long-read dataset, which may be attributed to the ability of long-read sequencing to detect full-length isoforms. Several other factors could also contribute to these variations, including the lack of strand-specific information in the short-read comparison, leading to false positive allelic calls due to overlapping transcripts and differences in read coverage, and technical explanations such as alignment biases. Considering that disagreement was found only for allele-specific non-imprinted genes, which generally tend to have a lower allelic bias than imprinted

**Fig. 2.** Comparative validation of allele-specific long-read results using short-read sequencing results. (a) Number of uniquely aligned reads used as input for allele-specific analysis and the number of informative genes in short (reads overlapping a SNP  $\geq 20$ ) and long (reads overlapping a SNP  $\geq 10$ ) reads. High overlap of informative autosomal genes between the datasets is shown with a Venn diagram. Long reads are represented in dark gray, and short reads are represented in light gray. (b) Violin plots of the allelic ratios of autosomal genes from the long- and short-read analysis. In our setup, allelic ratios range from 0 to 1 (0 = 100% expression from the paternal allele, 1 = 100% expression from the maternal allele). (c) Pie charts showing the proportion of biallelic genes ( $0.3 < \text{allelic ratio} < 0.7$ ) and allele-specific genes for the overlapping autosomal genes between the long- and short-read datasets. (d) Venn diagrams showing the overlap of genes labeled as biallelic (left) and allelic (right) in both datasets. (e) Pearson correlation of short- and long-read allelic ratios for all genes (gray, opacity 20%) and for known imprinted genes ( $n = 15$ ). Known imprinted genes are indicated in blue (paternally expressed) or red (maternally expressed), while genes labeled in black are presented in other figure panels. (f) Genome browser visualization showing the long-read data for the biallelic gene *Ube2g2* and the known paternally expressed imprinted gene *Impact*. Maternal reads are indicated in red, and paternal reads are indicated in blue. SNPs used to assign long reads to the BL6 or CAST allele are indicated in black. To visualize the long reads in the IGV browser, the following options were used: the 'Link supplementary alignments', represented by a thin blue/red connecting line, and 'Hide small indels'. Black lines indicate regions of the primary alignment that span splice junctions. No reads have been collapsed, and all reads aligned to each locus are shown. (g) Genome browser visualization showing the long-read data for the biallelic lncRNA *Malat1* and the known maternally expressed imprinted lncRNA *Meg3*.

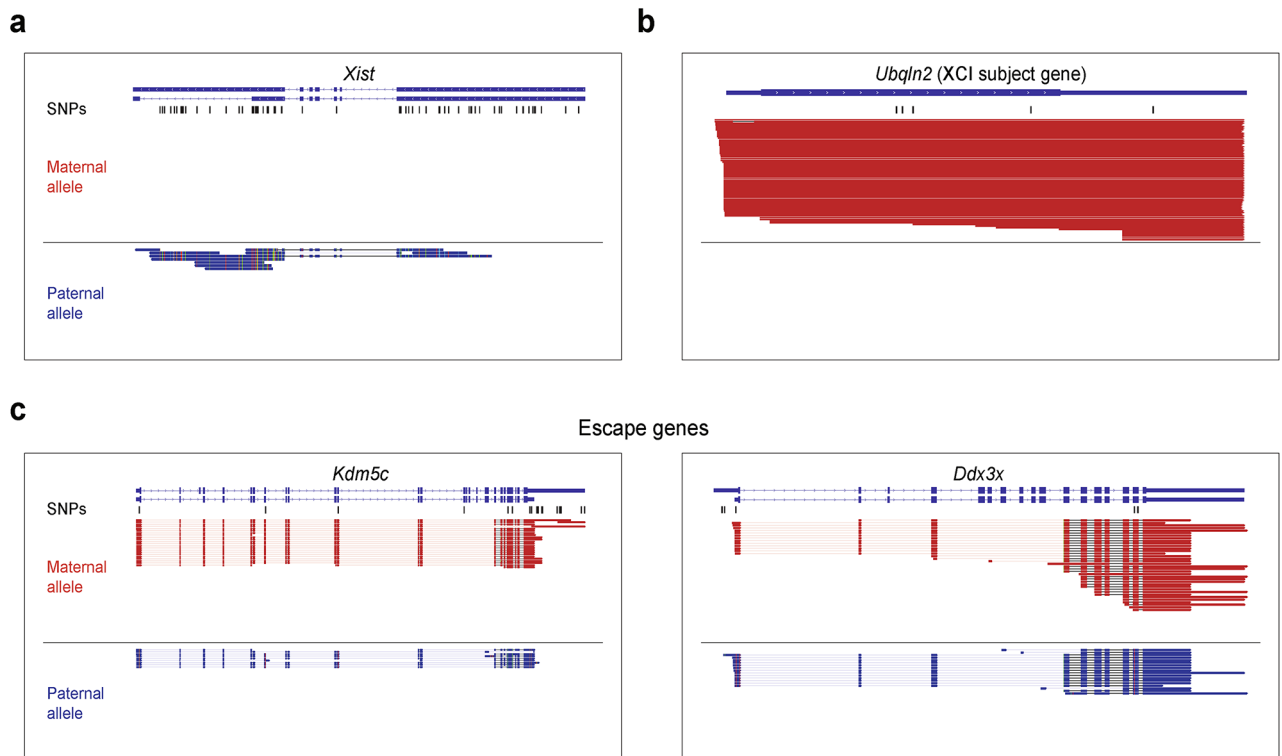


**Fig. 3.** Allele-specific long-read analysis of the *Gnas* cluster (a) Organization of the *Gnas* cluster, adapted from Peters and Williamson<sup>23</sup>. While many transcripts have been identified within this cluster, only the most notable isoforms are shown for simplicity. Dots indicate methylated promoters on either the maternal or paternal allele. First exons of protein-coding (black rectangles) and non-coding isoforms (gray rectangles) are shown. (b) Genome browser tracks at the *Gnas* locus showing long-read data for the maternal (red) and paternal (blue) alleles as well as short-read data. SNPs used to assign long reads to the BL6 or CAST allele are indicated in black. (c) Quantification of *Gnas* isoform counts for the maternal and paternal alleles.

genes, the variation may also be attributed to genes that fluctuate around the allelic ratio cutoff. Thus, for future studies, it is recommended to incorporate additional replicates to enhance the robustness and reproducibility of allele-specific expression detection, as previously suggested<sup>37,38</sup>.

Nonetheless, we were able to accurately assign the allele-specific status of known imprinted genes, including lncRNAs. We highlight the power of this approach by exploring allele-specific isoforms in the imprinted *Gnas* cluster. Our allele-specific long-read quantification confirmed the allelic status of the biallelically expressed *Gsa*, maternal *Nesp*, and paternal *Ex1A* isoforms, emphasizing the potential of this approach to elucidate the mechanism of complex allele-specific loci. To identify novel imprinted genes using long-read sequencing, a reverse cross would be required. This approach would help determine whether allele-specific expression arises from imprinting or caused by genetic differences between the alleles<sup>1,5,18</sup>. Furthermore, we correctly assigned the





**Fig. 4.** Long-read sequencing enables allele-specific assignment of known X chromosome inactivation escape genes. (a) Genome browser visualization of long-read data for *Xist*. (b) Genome browser visualization of long-read data for X-linked genes showing the non-escape gene *Ubqln2*. (c) Genome browser visualization of long-read data for the known escape genes *Kdm5c* and *Ddx3x*. SNPs used to assign long reads to the BL6 or CAST allele are indicated in black.

majority of X-linked genes as exclusively expressed from the maternal allele, as expected from our model. We also identified the lncRNA *Xist* as being expressed from the paternal inactive X chromosome, along with known escape genes showing the expected expression from both the active and inactive X chromosomes. Interestingly, no obvious isoform differences were detected for these escape genes, regardless of whether they were expressed from the active or inactive X chromosome, suggesting that the repressive epigenetic signature of the Barr body does not give rise to inactive X-specific isoforms.

Future studies should use long-read sequencing across multiple organs, cell types, and developmental timepoints to map the allele-specific landscape, as this approach provides critical insights into the regulation of gene expression at the isoform level. Using the described framework, these findings could be extended to humans when phased SNP information is already available. Otherwise, SNP calling and chromosome-wide phasing would be required for each human sample to accurately distinguish between alleles. For such challenging datasets, more complex pipelines like LORALS can enable detailed transcriptomic analyses, including classification of alternative transcript structure, rare variant effects, and isoform usage<sup>29</sup>. In summary, we provide a user-friendly framework to encourage future studies to adopt allele-specific long-read analysis, facilitating the disentangling of complex loci and offering mechanistic insights into gene regulation.

## Methods

### Mouse model

We obtained C57BL/6J (BL6) females carrying a heterozygous deletion of the proximal A-repeat (*Xist*<sup>-/+</sup>) from Riken BRC Japan (RBRC02655: B6;129-Xist<tm5Sado>) and crossed them with a wild-type CAST/EiJ (CAST) male (JAX: Strain #000928)<sup>32</sup>. From the resulting highly polymorphic F1 offspring (BL6<sup>ΔXist</sup> × CAST), we isolated the whole adult brain (9 weeks) of a female and promptly snap-froze it, storing it at -80 °C for subsequent use.

### PacBio Iso-Seq library preparation and sequencing

RNA was extracted from the whole brain according to the manufacturer's instructions for Invitrogen TRIzol reagent. RNA concentration and purity were measured using a NanoDrop™ One spectrophotometer (Thermo Fisher Scientific). The integrity of the total RNA was assessed on an Agilent 4200 TapeStation System (Agilent Technologies) using an RNA Screentape analysis kit. The RNA was of high quality, with a RIN greater than 9. cDNA libraries were prepared according to the Procedure & Checklist—Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems protocol. Briefly, 300 ng of total RNA was used to generate full-length cDNA using oligo-dT primers and the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New

England Biolabs). cDNA amplification was performed, followed by end repair and adapter ligation. Primer and polymerase binding was performed according to the SMRT Link Sample Setup using the Sequel II Binding Kit 2.1. The entire pool was sequenced on one SMRT cell on the Sequel II sequencing platform using the Sequel II Sequencing Kit 2.0 (Pacific Biosciences) with a 4-h pre-extension and 20-h run time.

### Long-read data processing and allele-specific analysis

Raw long reads were demultiplexed and processed using the PacBio standard pipeline (SMRT Link version 11.0). The demultiplexed brain sample yielded 569,139 long reads with an average length of 2836 bp and read quality of Q40. These reads were further processed and filtered, resulting in 507,864 full-length non-concatemer reads, which were uniquely aligned to the mm10 genome using pbmm2 (version 1.10.0), a minimap2 (version 2.15) wrapper for PacBio<sup>33</sup>.

To conduct allele-specific long-read analysis, each read was individually tagged using the 'haplotag' command from WhatsHap (version 1.7.0)<sup>30,31</sup>. This tagging process involved incorporating exonic SNPs ( $n = 606,506$ , exons annotated from RefSeq) obtained from the Sanger database, specifically those differentiating between CAST and BL6<sup>2</sup>. The resulting BAM file was then separated into two haplotype BAM files, one for CAST and one for BL6, using the 'split' command from WhatsHap. Genes were visualized in the BAM files of the two haplotypes using the Integrative Genomics Viewer (version 2.16.0) with the following options: 'hide indels,' and 'link supplementary alignments'<sup>39,40</sup>.

To compute the allelic ratio for every gene, the long-read counts were first quantified for both haplotype BAM files using htseq-count for every gene within the RefSeq annotation (downloaded February 2018) with the stranded option (HTSeq version 0.12.4)<sup>41</sup>. Subsequently, we computed allelic ratios for every informative gene with at least 10 reads overlapping at least one exonic SNP. For this computation, we divided maternal reads by the total (maternal and paternal reads) for each gene, resulting in a ratio of 1 for maternally expressed genes and 0 for paternally expressed genes. This analysis was performed using R (version 4.2.3).

### Short-read data processing and allele-specific analysis

Short-read data (generated from Illumina RNA-Seq TruSeq v3) was downloaded for a male F1 adult brain (BL6xCAST) (Sequence Read Archive [SRP020526](https://www.ncbi.nlm.nih.gov/sra/SRP020526))<sup>18</sup>. The short reads were aligned to the mm10 genome using STAR, excluding alignments with an intron size over 100,000, non-canonical junctions, or that were multimapped (version 2.6.0c)<sup>42</sup>. We performed allele-specific expression analysis for informative genes, defined as having at least 20 reads overlapping a SNP, using the Allelome.PRO pipeline with the exonic SNP file as input<sup>1,2</sup>.

### Data availability

The long-read Hi-Fi raw reads (brain\_finc.bam) for the whole brain, together with the processed maternal and paternal long-reads BAM files generated for this study, have been deposited in the Gene Expression Omnibus (GEO) and are accessible under [GSE246857](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE246857). The whole-brain short-read sequencing dataset was downloaded from the Sequence Read Archive (SRA) under the following accession number: [SRP020526](https://www.ncbi.nlm.nih.gov/sra/SRP020526). Access to the allele-specific long-read resource is provided via the Integrative Genomics Viewer at: [https://github.com/AndergassenLab/Allelic\\_Longreads](https://github.com/AndergassenLab/Allelic_Longreads).

Received: 15 March 2024; Accepted: 2 April 2025

Published online: 11 November 2025

### References

- Andergassen, D. et al. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Res.* **43**, e146. <https://doi.org/10.1093/nar/gkv727> (2015).
- Keane, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294. <https://doi.org/10.1038/nature10413> (2011).
- Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511. <https://doi.org/10.1038/nature12531> (2013).
- Leung, D. et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354. <https://doi.org/10.1038/nature14217> (2015).
- Andergassen, D. et al. Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. *Elife* <https://doi.org/10.7554/eLife.25125> (2017).
- Barlow, D. P. & Bartolomei, M. S. Genomic imprinting in mammals. *Cold Spring Harb. Perspect. Biol.* <https://doi.org/10.1101/cshperspect.a018382> (2014).
- Barlow, D. P. Genomic imprinting: A mammalian epigenetic discovery model. *Annu. Rev. Genet.* **45**, 379–403. <https://doi.org/10.1146/annurev-genet-110410-132459> (2011).
- Andergassen, D. & Rinn, J. L. From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo. *Nat. Rev. Genet.* **23**, 229–243. <https://doi.org/10.1038/s41576-021-00427-8> (2022).
- Reik, W. & Walter, J. Genomic imprinting: Parental influence on the genome. *Nat. Rev. Genet.* **2**, 21–32. <https://doi.org/10.1038/35047554> (2001).
- Tucci, V., Isles, A. R., Kelsey, G., Ferguson-Smith, A. C. & Erice Imprinting, G. Genomic imprinting and physiological processes in mammals. *Cell* **176**, 952–965. <https://doi.org/10.1016/j.cell.2019.01.043> (2019).
- Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373. <https://doi.org/10.1038/190372a0> (1961).
- Loda, A., Collombet, S. & Heard, E. Gene regulation in time and space during X-chromosome inactivation. *Nat. Rev. Mol. Cell Biol.* **23**, 231–249. <https://doi.org/10.1038/s41580-021-00438-7> (2022).
- Wutz, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.* **12**, 542–553. <https://doi.org/10.1038/nrg3035> (2011).
- Berlitch, J. B. et al. Escape from X inactivation varies in mouse tissues. *PLoS Genet.* **11**, e1005079. <https://doi.org/10.1371/journal.pgen.1005079> (2015).

15. Cotton, A. M. et al. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* **14**, R122. <https://doi.org/10.1186/gb-2013-14-11-r122> (2013).
16. Hoelzl, S., Hasenbein, T. P., Engelhardt, S. & Andergassen, D. Aging promotes reactivation of the Barr body at distal chromosome regions. *Nat. Aging* <https://doi.org/10.1038/s43587-025-00856-8> (2025).
17. Babak, T. et al. Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* **18**, 1735–1741. <https://doi.org/10.1016/j.cub.2008.09.044> (2008).
18. Babak, T. et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nat. Genet.* **47**, 544–549. <https://doi.org/10.1038/ng.3274> (2015).
19. Krueger, F. & Andrews, S. R. SNPsplite: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research* **5**, 1479. <https://doi.org/10.12688/f1000research.9037.2> (2016).
20. Hasenbein, T. P., Hoelzl, S., Engelhardt, S. & Andergassen, D. Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome. *Nucleic. Acids Res.* **53**(19), 1362–4962. <https://doi.org/10.1093/nar/gkaf912> (2025).
21. Hanna, C. W. et al. Endogenous retroviral insertions drive non-canonical imprinting in extra-embryonic tissues. *Genome Biol.* **20**, 225. <https://doi.org/10.1186/s13059-019-1833-x> (2019).
22. Andergassen, D., Smith, Z. D., Kretzmer, H., Rinn, J. L. & Meissner, A. Diverse epigenetic mechanisms maintain parental imprints within the embryonic and extraembryonic lineages. *Dev. Cell* **56**, 2995–3005.e4. <https://doi.org/10.1016/j.devcel.2021.10.010> (2021).
23. Peters, J. & Williamson, C. M. Control of imprinting at the Gnas cluster. *Epigenetics* **2**, 207–213. <https://doi.org/10.4161/epi.2.4.5380> (2007).
24. Williamson, C. M. et al. Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. *Nat. Genet.* **38**, 350–355. <https://doi.org/10.1038/ng1731> (2006).
25. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239. <https://doi.org/10.1186/s13059-016-1103-0> (2016).
26. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162. <https://doi.org/10.1038/s41587-019-0217-9> (2019).
27. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA* **111**, 9869–9874. <https://doi.org/10.1073/pnas.1400447111> (2014).
28. De Paoli-Iseppi, R., Gleeson, J. & Clark, M. B. Isoform age—splice isoform profiling using long-read technologies. *Front. Mol. Biosci.* <https://doi.org/10.3389/FMOLB.2021.711733/FULL> (2021).
29. Glinos, D. A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359. <https://doi.org/10.1038/s41586-022-05035-y> (2022).
30. Martin, M. et al. WhatsHap: Fast and accurate read-based phasing. *bioRxiv* <https://doi.org/10.1101/085050> (2016).
31. Patterson, M. et al. WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509. <https://doi.org/10.1089/cmb.2014.0157> (2015).
32. Hoki, Y. et al. A proximal conserved repeat in the Xist gene is essential as a genomic element for X-inactivation in mouse. *Development* **136**, 139–146. <https://doi.org/10.1242/dev.026427> (2009).
33. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> (2018).
34. Holmes, R. et al. A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Res.* **13**, 1410–1415. <https://doi.org/10.1101/gr.955503> (2003).
35. Peeters, S. B., Cotton, A. M. & Brown, C. J. Variable escape from X-chromosome inactivation: Identifying factors that tip the scales towards expression. *BioEssays* **36**, 746–756. <https://doi.org/10.1002/bies.201400032> (2014).
36. Yang, F., Babak, T., Shendure, J. & Distche, C. M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* **20**, 614–622. <https://doi.org/10.1101/gr.103200.109> (2010).
37. DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA-Seq: A new perspective. *PLoS Genet.* **8**, e1002600. <https://doi.org/10.1371/journal.pgen.1002600> (2012).
38. Wang, X. & Clark, A. G. Using next-generation RNA sequencing to identify imprinted genes. *Heredity* **113**, 156–166. <https://doi.org/10.1038/hdy.2014.18> (2014).
39. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. <https://doi.org/10.1038/nbt.1754> (2011).
40. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192. <https://doi.org/10.1093/bib/bbs017> (2013).
41. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169. <https://doi.org/10.1093/bioinformatics/btu638> (2015).
42. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. <https://doi.org/10.1093/bioinformatics/bts635> (2013).

## Acknowledgements

We acknowledge Thomas Meitinger, Thomas Carell, and Pascal Giehr for providing access to the PacBio sequencing platform. We would like to thank David Stucki for providing valuable feedback on the manuscript and on the bioinformatic long-read workflow. This work was partly supported by the DZHK (Junior Research Group), the Deutsche Forschungsgemeinschaft (Project ID: 403584255-TRR 267), and the BMBF in the framework of the Cluster4Future program CNATM (Project ID: 03ZU1201BA).

## Author contributions

SH and EG planned and conducted the laboratory experiments. LL carried out computational and statistical analysis and contributed to data curation. TPH provided essential analytical tools for the research. DA assisted in the interpretation of the results and provided critical feedback and supervision. LL and DA took the lead in writing the manuscript. All the authors participated in critical discussions and provided feedback. Funding acquisition for this research was carried out by DA.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Declarations

## Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial



relationships that could be construed as a potential conflict of interest.

### **Additional information**

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-97362-z>.

**Correspondence** and requests for materials should be addressed to D.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**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 Author(s) 2025