Genome and single-cell RNA-sequencing of the earthworm *Eisenia andrei* identifies cellular mechanisms underlying regeneration

Yong Shao\textsuperscript{1,13}, Xiao-Bo Wang\textsuperscript{2,13}, Jin-Jin Zhang\textsuperscript{1,3,13}, Ming-Li Li\textsuperscript{1,3,13}, Shou-Song Wu\textsuperscript{4}, Xi-Yao Ma\textsuperscript{1}, Xue Wang\textsuperscript{5,6}, Hui-Fang Zhao\textsuperscript{5,6}, Yuan Li\textsuperscript{7}, Helen He Zhu\textsuperscript{5,6}, David M. Irwin\textsuperscript{1,8,9}, De-Peng Wang\textsuperscript{7}, Guo-Jie Zhang\textsuperscript{1,10,11,12,✉}, Jue Ruan\textsuperscript{2✉} & Dong-Dong Wu\textsuperscript{1,12,✉}

The earthworm is particularly fascinating to biologists because of its strong regenerative capacity. However, many aspects of its regeneration in nature remain elusive. Here we report chromosome-level genome, large-scale transcriptome and single-cell RNA-sequencing data during earthworm (*Eisenia andrei*) regeneration. We observe expansion of LINE2 transposable elements and gene families functionally related to regeneration (for example, *EGFR*, epidermal growth factor receptor) particularly for genes exhibiting differential expression during earthworm regeneration. Temporal gene expression trajectories identify transcriptional regulatory factors that are potentially crucial for initiating cell proliferation and differentiation during regeneration. Furthermore, early growth response genes related to regeneration are transcriptionally activated in both the earthworm and planarian. Meanwhile, single-cell RNA-sequencing provides insight into the regenerative process at a cellular level and finds that the largest proportion of cells present during regeneration are stem cells.
Regeneration is one of the most complex and intriguing biological processes that can occur throughout the lifetime of some organisms. However, the regenerative capacity of many animals is extremely limited; in contrast to differentiated tissues and organs, only fetal tissues can be recreated without fibrosis. However, some organisms still retain strong regenerative abilities. For example, the zebrafish, salamander, axolotl, and gecko are regenerative vertebrates that can scarlessly heal wounds and regenerate lost organs and appendages such as fins, heart, jaws, limbs, tails, gills and lenses. In addition to vertebrates, several invertebrates such as planarian and hydrids possess even stronger regenerative ability and can regenerate almost an entirely new organism because of the abundance of somatic stem cells (neoblasts) in their bodies. Therefore, these animals are often viewed as important models in stem cell biology and regenerative medicine.

Charles Darwin performed a large amount of work emphasizing the importance of the earthworm on soil formation and ecosystem development. Earthworms influence the physical characteristics of the soil as they dig burrows, deposit casts on the soil surface and within it and overturn dead organic matter. In Annelida, only three whole genomes, a marine polychaete and two freshwater worms, have been sequenced. However, some organisms still retain strong regenerative capacity, such as *Macrostomum lignano* (contig N50 = 64 Kb) and *Apostichopus japonicus* (contig N50 = 192 Kb). The genome size of the final assembly was approximately 1.3 Gb, which was close to the estimated size of 1.28 Gb from k-mer estimation and ~1.3 Gb from flow cytometry. The assembly exhibited a much better continuity, with a contig N50 size of approximately 740 kb, than the genomes of several other invertebrates with strong regenerative capacity, such as *Macrostomum lignano* (contig N50 = 64 Kb) and *Apostichopus japonicus* (contig N50 = 192 Kb)

To assess the completeness of our genome assembly, we aligned the short reads and the transcriptome unigenes to the genome and found that over 98.2% of the short reads and ~94.5% of the de novo transcriptome unigenes could be mapped to the assembly, demonstrating the high completeness of the assembly.

Phenotypic and transcriptomic changes during regeneration. Some studies have documented transcriptomic and some phenotypic changes of posterior regeneration in the earthworms, particularly for *E. andrei*. Therefore, in the present study, we focused on the anterior regeneration in the earthworm *E. andrei*. Phenotypic observations on multiple time points after the anterior amputation (Fig. 2a–c and Supplementary Figs. 4 and 5), could help us to evaluate whole regenerative processes after anterior amputation (the first four body segments), especially for early regenerative events. Using Ki-67 fluorescent labeling, we found that cell proliferation initiated at 24 h post-amputation, and at 48 and 72 h post-amputation the proliferating cells increased rapidly and gradually migrated to the center of cross sections (Fig. 2d and Supplementary Fig. 6). At 5 days post-amputation, the wound healing was fully accomplished and a small blastema appeared in center of the amputation plane (Supplementary Fig. 4). At 6 and 7 days post-amputation, the blastema persistently experienced growth and elongation (Supplementary Fig. 4). Although the newly produced body segments were not observed at 14 days post-amputation, the base of outgrowth has accumulated pigments (Supplementary Fig. 4). At 18 days post-amputation, new body segments arise, and at 28 days post-amputation the obvious body segments take shape in regenerative appendages (Supplementary Fig. 4).
To understand the genetic regulatory mechanisms underlying the early regenerative process in earthworms, we further sequenced the transcriptome of head, the first four body segments containing the central nervous system, during the regeneration process at 6 time points (0, 6, 12, 24, 48 and 72 h after cutting, with 5 biological replicates for each time point) (Fig. 2a). Principal component analysis of the gene expression profiles clearly split the 0 hour point from the remaining regeneration stage transcriptomes, indicating a high level of gene activity remodeling initiated by the regeneration process (Supplementary Fig. 7). Differentially expressed genes (DEGs) were identified for each regeneration stage compared to the control stage (Fig. 2e; fold change >2 and false discovery rate (FDR) < 0.05). In total, 6,048 DEGs that changed their expression at one or more regeneration time points were identified, and these genes demonstrated a temporal order in their expression profiles (Supplementary Fig. 8).

Gene enrichment analysis found that many biological processes important for development were commonly upregulated across all regeneration stages, including gene transcription (GO:0006351), Wnt signaling pathway (GO:0016055), cell surface receptor signaling pathway (GO:0007166), multicellular organism development (GO:0007275), and anatomical structure development (GO:0048856) (Supplementary Data 1). These results indicate that regeneration, as a very complex process, involves multiple genes and pathways. Next, we integrated genomic and transcriptomic analyses to reveal the molecular mechanisms underlying regeneration.

**Expansion of LINE2 transposable elements.** Transposable elements (TEs) make up a large fraction of the genome and play important roles in genome function and evolution.

In the earthworm, TEs comprise ~56.72% of the genome, posing a challenge for genome assembly (Supplementary Table 8). Among them, DNA transposons and long interspersed nuclear elements (LINEs) comprise the majority of the repeats, spanning 349.6 Mb of the genome (Fig. 3a and Supplementary Table 8). Of particular note, LINE2 has undergone significant expansion (7.49%) in the earthworm compared to other representative metazoan species (2.52% in C. teleta, 3.90% in H. robusta, 0.00% in M. lignano, and 0.84% in A. japonicus), and the closely related species E. fetida also harbors a high LINE2 proportion (~4.10%) compared to other un-earthworm species, although a low genome assembly quality may underestimate this possibility (Fig. 3b, Supplementary Figs. 9 and 10, and Supplementary Data 2). The number of substitutions to replace consensus4, which is an estimate of the relative age of the LINE2, implied that the earthworm LINE2 has undergone a recent and apparent burst of expansion with a peak at 25~30 Mya (Fig. 3c), which is much more recent than its divergence time (309 Mya) from H. robusta (Supplementary Fig. 11).

Approximately 43.54% of the LINE2 elements in the earthworm genome are located in intron regions, and 6.66% are located within the 5-kb flanking regions of genes (Fig. 3d). This suggests that the function of LINE2 is potentially involved in regulatory roles. To test it, we performed further analyses by integrating transcriptomes described above. We discovered that the proportion of DEGs (described above) harboring LINE2 elements, was significantly higher than that of non-DEGs (background genes) harboring LINE2 elements (Fig. 3e, \( P = 7.641E-07, \chi^2 \) test). Further, 44 and 119 significantly differentially expressed LINE2 elements (DEL2s, FDR < 0.05), located in 5k 5′-flanking and 5k 3′-flanking of coding genes, respectively, were identified, which potentially were activated during regeneration.
process because of their increasing expression trends (Fig. 3f and Supplementary Fig. 12), especially for DEL2s in 5k 5'-flanking ($P < 0.05$, Mann-Whitney $U$ test). Among these DEL2s within the 5-kb flanking regions of coding genes, we found 19 DEL2s were transcriptionally activated with significantly increased expression during the regenerative process and their neighboring genes also demonstrated similar increasing expression trends (Fig. 3g, FDR < 0.05, Benjamini-Hochberg FDR). The neighboring genes of 19 DEL2s, such as $EGR1$, $FOSL$, $BMP10$, $HUNB$ and $MMP17$, are frequently reported to participate in regeneration$^{22-24}$. For example, $EGR1$ functions as a pioneer factor to directly regulate early wound-induced genes in acoels$^{22}$. Our analyses suggest
that partial LINE2 elements in earthworms might regulate the expression of neighboring genes by coopting them into regeneration-regulatory networks. However, we acknowledge that further experiments are needed to elucidate how LINE2 elements regulate gene expression during earthworm regeneration. Overall, our study suggests that LINE2 elements in earthworms may play important roles in early regenerative processes.

Evolution of gene families in the earthworm genome. Expansion or contraction of gene families is associated with the evolution of specific phenotypes and physiological functions. In the present study, we identified 26,926 gene families from 12 invertebrates (Supplementary Fig. 13). 4,877 gene families were shared by five species (E. andrei, H. robusta, C. teleta, Crassostrea gigas, and Lottia gigantea) (Supplementary Fig. 14), while 1165 gene families were unique to earthworms (Supplementary Fig. 15 and Supplementary Table 9). In line with a previous study, which identified extensive gene duplications functioning as regulating early development in the E. fetida genome, we also found abundant expanded gene families in the earthworms (i.e., +2776 in E. andrei and +3537 in E. fetida) (Fig. 4a and Supplementary Fig. 16). We further estimated the time of these duplication events by using Ks distributions, where Ks is the synonymous distance or defined number of synonymous substitutions per synonymous site. Ks distributions of duplication events in the E. andrei genome were obviously larger than the Ks distribution of one-to-one orthologs between E. andrei and E. fetida, which implied that these gene duplications occurred before the divergence of E. andrei and E. fetida (Fig. 4b). Furthermore, these expanded gene families were mainly enriched in GO terms including cell-cell signaling (GO:0007267, \( P = 2.38E-02 \)), Wnt signaling pathway (GO:0016055, \( P = 2.32E-02 \)), cell surface receptor signaling pathway (GO:0007166, \( P = 6.05E-03 \)), regulation of cell communication (GO:0010646, \( P = 3.91E-06 \)), development process (GO:0009613, \( P = 2.38E-02 \)), Wnt signaling pathway (GO:0000077, \( P = 5.47E-03 \)), organeller organization (GO:0006996, \( P = 3.56E-02 \)), and regulation of cellular biosynthetic process (GO:0031326, \( P = 7.98E-05 \)) (Supplementary Data 3). We speculated members of these expanded gene families in E. andrei may potentially participate in special phenotypic evolution of the earthworm, such as regeneration. Similarly, a previous study using expressed sequence tags also found that biological processes such as cell-cell communication and biosynthesis could occur during the regeneration stages in P. excavates, another earthworm. Of particular interest, the Wnt signaling pathway, a canonical regeneration pathway controlling anteroposterior polarity during planarian regeneration and regulating progenitor cell fate and proliferation during regeneration of zebrafish fins and deer antlers, have displayed a substantial expansion in the earthworm. For example, the genes APC and DVL3 showed expansions in the Wnt signaling pathway and exhibited increasing trends in expression during regeneration (Supplementary Fig. 17).

Among 186 significantly expanded gene families in the earthworm branch (Viterbi P-value ≤0.05), 35 gene families harbor over 10% of their family members displaying significant expression changes during regeneration (Fig. 4c and Supplementary Fig. 18). Furthermore, we performed a randomization test and found five gene families standing out as showing significantly higher proportion of differentially expressed genes (\( P < 0.05, \chi^2 \) test), including ZNFX1, EGFRI, NNP, HEL22 and SACS. For example, ZNFX1 is activated in both newt and axolotl incompe-
tent iris regeneration, and 9 of 11 copies of ZNFX1 exhibit significant expression changes in earthworm during regeneration processes (\( P = 0.0105, \chi^2 \) test, and Supplementary Fig. 19). Gene, EGFRI, encodes an epidermal growth factor receptor, which is a transmembrane receptor with tyrosine kinase activity that can regulate cell proliferation and differentiation. In planarians, silencing of EGFRI and EGFRI-3 can result in abnormal morphogenesis and disorganized developmental structures during regeneration. EGFRI experienced a significant expansion with a significantly increased copy number in the earthworm (12 copies) relative to other species, which have 0–2 copies (Fig. 4d, and \( P = 0.0114, \chi^2 \) test). Eight of the 12 members showed differential expression levels during regeneration and real-time Quantitative PCR further validated expression trends of these duplications in regenerative process of the earthworm (Fig. 4e, Supplementary Fig. 20, and Supplementary Table 10). Although these gene families have diverse roles during development across life cycle, the members of them were significantly differentially expressed during the earthworm regeneration processes and their duplications might potentially play a role in the evolution of regeneration in E. andrei.

Temporal gene regulation patterns in regeneration. To understand the large-scale gene interactions involved in regeneration, we conducted a weighted gene coexpression network analysis (WGCNA)\(^{34} \). This quantitative network-based approach has proven to be a powerful tool for elucidating cell type, anatomic and convergent gene networks across species. Here, we identified 19 gene coexpression modules in response to temporal changes during the regeneration process (Fig. 5a and Supplementary Figs. 21–24). These modules represent genes that share highly similar expression patterns during regeneration (Fig. 5a).

Of these 19 modules, five modules (tan, brown, lightcyan, grey60, and cyan) were dominated by genes showing upregulation at the early stage of the regeneration process (6 h) (Supplementary Figs. 23, 24). Among them, the expression of the brown module was most significantly correlated with the regeneration stage (6 h) (\( r = 0.53, P = 0.003 \)) (Fig. 5b and Supplementary Fig. 23). Genes enriched in this module participate in signal transduction, transcription and translation, implying an increasing level of cell communication and biochemical processes via the synthesis of mRNAs and proteins in response to regeneration (Supplementary Data 4). The list of driver genes in the brown
module triggered by the regeneration process includes several genes involved in cellular proliferation, differentiation and programmed cell death, such as FOSS (intramodule membership = 0.9587) and HUNB (intramodule membership = 0.934) (Fig. 5c, and Supplementary Table 11). Previous studies reveal that FOSS participates in neoblast maintenance and the wound response program in planarians36,37 and is a key factor in the cell signaling system activated immediately after cell damage38.

Two other modules, red and blue, containing genes with increased expression until 12 h of regeneration (Fig. 5d, e, and Supplementary Fig. 24), were also enriched in genes involved in biosynthetic processes and the regulation of cell growth (Supplementary Data 5 and 6). Additionally, the blue module was enriched in genes involved in energy metabolism that are necessary for cell proliferation and growth. However, intriguingly, both of these networks lacked core driver genes.
Fig. 4 Evolution of gene families in the earthworm genome. a Expansion/construction of gene families for 12 invertebrates. Expanded gene families were shown in green and contracted gene families in red at the whole genome levels. b E. andrei and E. fetida paranome $K_s$ distribution and $K_s$ distribution of one-to-one orthologs of E. andrei and E. fetida. We constructed and visualized the $K_s$ distribution of paralogs and orthologs using 'ksd' with default parameters and 'viz' command in 'wdg' tools, respectively. c Gene families possessing higher proportions of time-dependently DEGs in regeneration. The numbers in the grids were the copy numbers of the gene families for each species. The % column showed the proportion of DEGs in the specific gene family. And the gene families with higher copy numbers were indicated in dark blue, and the gene families with lower copy numbers were indicated in light yellow. For each regenerate stage of each copy included 5 biological replicates ($n=5$). The error bars were showed by using standard error of the mean (s.e.m.). And copies were named by contig orders.
functioning as regeneration regulators (Fig. 5f, g). Therefore, we proposed that the two modules might be vital for regulating the preparation process of the cell proliferation in the early phase of earthworm regeneration.

The black module contains genes that exhibit upregulation within 6 hours after amputation and then gradually increase in expression until 72 h (Fig. 5h, $r = 0.49, P = 0.0006$, and Supplementary Figs. 23, 24). This module presumably has an important...
functional role, especially at 48 and 72 h of the early phase of regeneration, because of its sustained and increasing activity. Gene enrichment analysis found that this module was significantly enriched in genes with functions in phosphorylation, cell surface receptor, enzyme activity and ATP binding, all of which are vital for signal transduction (Supplementary Table 12). We uncovered driver genes in the black module, such as AGRIN, which had a higher network connectivity (intramodule membership = 0.9276) and is a component of the extracellular matrix, affecting regenerative capacity and development processes in mammals36,37 (Fig. 5i and Supplementary Table 13). Thus, we propose that the black module genes, with their increasing consistent temporal regulation patterns, may play an important functional role in earthworm regeneration.

Transcriptional activation of immediate early response genes. We next sought to discover genetic toolkits that participate in the wound-induced regeneration processes of earthworms and planarians by comparing the temporal transcriptome data from these two species37. We found the early growth response genes were transcriptionally induced as a rapid response to injury healing in both species. In earthworms, the expression of the early growth response protein 1 gene (EGR1) and the immediate early response gene (IERSL) was significantly up-regulated at all regeneration stages, while the expression level of the early growth response protein 1-B gene (EGR1B) was significantly elevated at 6, 12 and 24 hours (Fig. 2f). Similarly, in planarians, we noticed that genes involved in early growth responses, i.e., EGR1, EGR2, EGR1l and EGR3, were also transcriptionally activated (Fig. 2f). Importantly, EGR1 was a shared gene in both of two species during regeneration. EGR1, as a member of the immediate early response gene transcription factor family, is implicated in the regulation of multiple cellular processes, such as cell growth, development and stress responses in many tissues, and can control the proliferation and localization of stem cells41. Additionally, several important transcription factors, RUNT, JUN and FOS, regulating regeneration processes, were involved in early regeneration in both species (Fig. 2f). For example, the RUNT gene, encoding the Runx transcription factor, whose function specifies different cell types during regeneration and promotes heterogeneity in neoblasts near wounds in planarians37, was significantly upregulated in earthworms throughout the regeneration process and was also upregulated in planarians at 3, 6, and 12 h. Thus, our results suggest the earthworm and planarian potentially utilize a set of similar transcriptional activated immediate early response genes to regulate early regeneration process.

Single-cell RNA-sequencing reveals cytological mechanisms. To provide an in-depth understanding of the complex interplay among the molecular and cellular processes underlying earthworm regeneration, we performed single-cell RNA-sequencing using 10X Genomics Chromium platform to examine regenerating heads (the first four segments) at 72 h after cutting. In brief, once the head was amputated, we obtained regenerating segments of cell dissociation and cell sorting. We captured a total of 2080 cells with an average of 493 genes and 1904 transcripts per cell (Supplementary Fig. 25 and Supplementary Table 14). After cell dissociation and cell sorting, we captured a total of 2080 cells with an average of 493 genes and 1904 transcripts per cell. We successfully identified 20 clusters, where the expression levels of stem cell markers THY1 and SOX2 were transcriptionally induced as a rapid response to injury healing in both species. In earthworms, the expression of the early growth response protein 1 gene (GROWTH RESPONSE PROTEIN 1 gene) was significantly elevated at 6, 12 and 24 hours (Fig. 2f). Similarly, in planarians, we noticed that genes involved in early growth responses, i.e., EGR1, EGR2, EGR1l and EGR3, were also transcriptionally activated (Fig. 2f). Importantly, EGR1 was a shared gene in both of two species during regeneration. EGR1, as a member of the immediate early response gene transcription factor family, is implicated in the regulation of multiple cellular processes, such as cell growth, development and stress responses in many tissues, and can control the proliferation and localization of stem cells41. Additionally, several important transcription factors, RUNT, JUN and FOS, regulating regeneration processes, were involved in early regeneration in both species (Fig. 2f). For example, the RUNT gene, encoding the Runx transcription factor, whose function specifies different cell types during regeneration and promotes heterogeneity in neoblasts near wounds in planarians37, was significantly upregulated in earthworms throughout the regeneration process and was also upregulated in planarians at 3, 6, and 12 h. Thus, our results suggest the earthworm and planarian potentially utilize a set of similar transcriptional activated immediate early response genes to regulate early regeneration process.

**Discussion**

A mounting number of studies suggest the importance of earthworms in terms of understanding many aspects of biology. In particular, earthworms are of great interest from the perspective of regenerative biology. To date, apart from *C. teleta* and *H. robusta*, which are annelida, the genome of only one other species, *E. fetida*, from oligochaeta (also known as earthworms) has been sequenced using the next generation genome sequencing strategy, but provided poor assembly quality (contig N50 = 1,852 bp and contig N50 = 967 bp, respectively). Having no high quality genome severely hinders the development of earthworm regeneration biology. In this study, we present a chromosome level genome assembly of the earthworm *E. andrei* with a scaffold N50 = 111 Mbp using a single molecule sequencing (PacBio) integrating Hi-C assembly technology, up to now representing an optimal genome assembly in the phylum annelida. The earthworm *E. andrei* exhibits a high level of regenerative ability at both its anterior and posterior and is easy to culture and handle in laboratory12,19. Therefore, it can be potentially regarded as a valuable model to investigate the mechanisms underlying regeneration. We believe that this high-quality genome will supply a useful genetic resource for future research especially in regeneration biology.

Increasing genomes from diverse species indicate that nearly half of genome sequences are derived from TE{s}, which have played important functional roles in many biological processes. In this study, we propose a potential regulatory role of LINE2 in...
the evolution of the earthworm, possibly in earthworm regeneration. We discover that several LINE2 elements are inserted in the loci of DEGs during early stages of earthworm regeneration. Some specific differentially expressed LINE2 elements in the 5k-flanking sequences of coding genes and their neighboring genes harbored similar increased expression trends during earthworm regeneration. For example, EGR1, a core regulator of wound inducing process in diverse regenerative organisms\textsuperscript{48–51}, such as acel and planarian\textsuperscript{22,37}, displayed significant differential expression, and harbored differentially expressed LINE2 elements in the earthworm. However, future experiments are required to relate expanded LINE2 with regeneration of the earthworm.
Consistent with a previous study, a mount of gene duplication events (i.e., many potential expanded gene families) have occurred in the genome of earthworm. These expanded gene families in earthworms were significantly enriched in terms/paths representing development biology, which potentially reflect partly their roles in regeneration. Particularly, some expanded gene families, e.g., ZNFX1 and EGFR, show a higher proportion of their members undergoing significant differential expression during early phases of regeneration in the earthworm. Previous studies indicated that ZNFX1 (which encodes a NFX1-type zinc finger-containing protein 1) is up-regulated in both newt and axolotl lens differentiation in regeneration of E. anderi. The histological observations of blastema formation in anterior regeneration still remains largely unclear in the earthworm. Considering that the wound healing process is accomplished at 3–5 days post-amputation in E. anderi, we performed transcriptomic analyses at early phases of wound healing in this earthworm. Immediate early response genes (e.g., EGR1) were transcriptionally co-activated in the earthworms and planarians, implying a set of parallel activated mechanisms in early phases of anterior regeneration still remains largely unclear in the earthworm. Considering that the wound healing process is accomplished at 3–5 days post-amputation in E. anderi, we performed transcriptomic analyses at early phases of wound healing in this earthworm. Immediate early response genes (e.g., EGR1) were transcriptionally co-activated in the earthworms and planarians, implying a set of parallel activated mechanisms in early phases of regeneration. Four vital gene co-expression network modules (i.e., brown, blue, red and black) were identified and these show substantial transcriptional activation during early phases in earthworm regeneration. Functional enrichment of some of the genes expressed in these networks identified signal transduction, biosynthetic processes and the regulation of cell growth, suggesting that these genes may regulate wound healing process in the early phase of the earthworm regeneration.

The epimorphic process of earthworms is thought to occur mainly via dedifferentiation and subsequent re-differentiation of cells, without any contribution from totipotent stem cells (or neoblasts), and this process commonly involves blastema formation (dedifferentiated cells), which contributes to re-differentiation in regeneration of Enchytraeus japonicus and E. anderi. The histological observations of blastema formation at 1–3 days post-amputation during E. anderi tail regeneration showed that at 3 days post-amputation, the blastema cells, which are likely to be pluripotent cells, rapidly proliferated and migrated to coelom. Here, we performed single-cell RNA-sequencing data at 3 days (72h) anterior post-amputation in the earthworm, and found that the pluripotent stem cells, potentially representing blastema cells, were the largest proportion of cells at this time. Further ISH experiments supported large proportion of PSCs and found that highly enriched PSCs surrounding the EP (central area) of the cross section, which was consistent with formation of blastema at this time (Fig. 7 and Supplementary Fig. 30). However, single cell RNA-sequencing data from more different times will undoubtedly help to understand cellular process of regeneration.

Our study identifies some candidate genetic mechanisms underlying regeneration and highlights the earthworm as a promising model for future studies of regenerative biology. In the future, multiple OMICS strategies, interdisciplinary and functional experiments will provide further insight into the regenerative biology of the earthworm.

Methods
DNA isolation, PacBio library preparation and sequencing. One live earthworm (E. andrei, originating from Guangxi province in China) was prepared, and its intestinal tract was removed. After washing with saline solution, the earthworm genomic DNA was collected using a Qiagen kit. After assessing the quality of the DNA, we constructed a PacBio library with an insert size of 20 kb and utilized a single molecular RS sequencer to perform long-read sequencing. Hi-C was performed using the following protocol: the adult earthworm tissues were fixed in 1% formaldehyde solution. The nuclear chromatin was obtained from the fixed tissue and digested using HindIII (New England Biolabs). The overlaps resulting from HindIII digestion were blunted by bio-14-DCTP (Invitrogen) and the Klenow enzyme (NEB). After dilution and religation using T4 DNA ligase (NEB), the earthworm genomic DNA was extracted and sheared to a size of 350–500 bp with a Bioruptor (Diagenode). The biotin-labeled DNA fragments were enriched by utilizing streptavidin beads (Invitrogen) to further finish library preparation.

Estimation of earthworm genome size. The k-mer algorithm was applied to evaluate the earthworm genome size. The 17 k-mer and 34.7G next-generation sequencing reads were utilized in these analyses. Flow cytometry analysis further was used to evaluate the genome size of the earthworm. In brief, after cell suspensions were prepared, we added 500ul PI (C0080, Solarbio) dye working solution [0.85% PBS 9.4 ml, PI (1 mg/ml) 500 μl, DNA free Rnase (10 mg/ml) 50 μl, Triton X-100 10 μl, Sodium Citrate 10 mg, keep away from light] into the prepared earthworm cell suspension, chicken blood cell solution and mixture of earthworm cell and chicken blood cell, and then they were mixed and moved 400ul to flow tubes covered with fresh-keeping films to be tested. The estimation of genome size was performed using BD LSR Fortessa flow cytometer (BD Biosciences, USA). The genome size of chicken (Gallus gallus GRCg6a) (1.04 pg) was utilized as a reference control. Flow cytometry analysis was carried out using the laser excitation at 488 nm with minimum 10,000 events (cells) per sample. The mean fluorescence intensity was obtained using Flowjo (v7.1). The DNA content was estimated using the standard formula for genome size (pg) = (Sample fluorescence channel number FL/Chicken fluorescence channel number FL) x 1.04 pg.

Long-read de novo assembly of the genome. We used ~80X PacBio subreads to perform de novo genome assembly by using Wdubg (v1.2.7) ([https://github.com/ruanjun/wdubg](https://github.com/ruanjun/wdubg)), FALCON26 (v052016) and Canu26 (v1.7). Then, the assembled genome was corrected by aligning subreads using the Arrow program (v2.3.2) with the default parameters. Finally, Pilon (v1.22) was used to polish the resulting assembly with ~24X PE310 reads from the Illumina platform. The base accuracy of the assembly was estimated by Illumina reads alignment. The completeness of the assembly was evaluated by BUSCO genes ([http://busco.ezlab.org/](http://busco.ezlab.org/)). Furthermore, the completeness of the assembly was validated by six de novo transcriptomes using Trinity26 (v2.1.1).
Fig. 7 In situ hybridization of gene SOX2 in cross sections at 6 different time points post-amputation in the earthworm. The slice size was 10 μm. The 6 time points post-amputation included 0 (a), 6 (b), 12 (c), 24 (d), 48 (e) and 72 (f) hours. The red fluorescence represented positive signals and DAPI (blue fluorescence) was used to stain cell nucleus. Similar results in a-f could be ensured by three independently biological experiments.
**Genome annotation.** De novo and homology approaches were combined to identify repetitive sequences in the earthworm genome. For the de novo approach, we combined transcript and repeat library using RepeatMasker (www.repeatmasker.org/RepeatModeler/) with the default settings. Then, RepeatMasker (v4.0.7) (http://www.repeatmasker.org/) was run on the earthworm genome using the de novo library. RepeatMasker was also run against the RepBase (v20105807) (https://www.girinst.org/repbase/) for homologous repeat identification. The result of repeat annotation using these two approaches were integrated. To annotate the protein-coding genes of the earthworm genome, de novo, homology-based and transcriptome-based prediction methods were combined. Two de novo programs, Augustus67 (v3.0.3) and SNAPP68 (v2006-07-28), were performed to predict genes in the repeat-masked genome sequences. Long predicted genes were filtered using PASA69 (v20104417) to use the train gene model parameters for the two de novo programs. For the homology-based predictions, protein sequences from C. teleta and H. robusta (downloaded from the Ensembl database) were aligned to the earthworm genome using blat (e-value < 10^-5). We used genBlatA70 (v1.0.138) to cluster the adjacent HSPs (high-scoring pairs) from the same protein alignments, and GeneWise71 (v2.2.3) was used to identify accurate gene structures. After QC and filtering, reads from all RNA libraries were mapped to the earthworm genome using TopHat2 (v2.0.13) (http://ccbb.jhu.edu/software/tophat/) and Cufflinks (v2.1.1) (http://cole-trapnell-lab.github.io/cufflinks/) was subsequently used to predict gene models. All predicted genes from the three approaches were integrated with EVidenceModeler (EVM)72 (v2012-06-25) to generate high-confidence gene sets. To obtain gene function annotations, KEgg (https://www.genome.jp/kegg/), SwissProt and TrEMBL protein databases (https://www.uniprot.org/) were searched with BLASTP (ncbi-blast-2.2.28+) (e-value < 10^-5). The best hits were used to assign homology-based gene functions. Functional classification based on GO categories and InterPro entries was achieved using the InterProScan program (v5.21-60.0) (http://www.ebi.ac.uk/interpro/download/).

**Gene family clusters.** Comparisons among 12 species, including Caenorhabditis elegans, C. gigas, C. teleta, Drosophila melanogaster, H. robusta, L. gigantea, Schistosoma mansoni, Strongylocentrotus purpuratus, Saccofocus kowalevskii, A. japonicus, Ascaris suum and earthworms, were conducted to clans of homologous gene families. We selected the longest transcript for each gene and eliminated those with premature stop codons, non-potent length or fewer than 30 amino acids encoded. Subsequently, OrthoMCL69 (v2.0.9) was used to construct gene families via all-versus-all BLAST alignments. Changes in gene family size (expansion/contraction) were calculated by the CAFE program (v2.2.3) (https://hahlab-lab.isthec.tsinghua.edu.cn/software.html). To perform phylogenetic analyses, single-copy families were identified, and peptide alignments for each family using MUSCLE (v3.8.31) (http://drive5.com/muscle/downloads.html) and concatenated to form a supergene for each species. RAxML (v8.2.9) (https://cme.h-its.org/elixis/conf/software/raxml/index.html) with the PROTEINAAUTO model and 100 bootstraps were used to build a phylogenetic tree. The peptide alignment were converted to coding sequences, which were subjected to analysis with MCMMCre in the PAML package (v4.8a) (http://abacus.gene.ucl.ac.uk/software/paml.html) to estimate divergence times. Fossil calibration points were obtained from a web-based database—http://fossilweb.tamu.edu/silic Tongue number (gene) 12, 24, 48, and 72 h after post-amputation to the reference genome according to the two annotations based on the bowtie2 program in TopHat2 (v2.0.13). The expression abundance of each LINE2 was quantified by the cuffquant program in Cufflinks (v2.1.1), and the cuffdiff program in Cufflinks (v2.1.1) was utilized to determine DEGs (FDR < 0.05) at each regeneration time points (6, 12, 24, 48 and 72 h) after post-amputation to the reference genome according to the two annotations based on the bowtie2 program in TopHat2 (v2.0.13). The expression abundance of each LINE2 was quantified by the cuffquant program in Cufflinks (v2.1.1), and the cuffdiff program in Cufflinks (v2.1.1) was utilized to determine DEGs (FDR < 0.05) at each regeneration time points (6, 12, 24, 48 and 72 h) after post-amputation. Thus, we screened significantly differentially expressed LINE2 elements in 5′-flanking and 3′-flanking of coding genes, respectively.

**Identification of coexpression networks in early regenerative processes.** Analysis was carried out in R on a 64-bit LINUX platform with 65.7 GB memory. Modules/organ networks were constructed using WGCNA34 (v1.67). Modules were defined as branches of the hierarchical cluster tree using the dynamic tree cut method. For each module, the expression patterns were summarized by the module eigengene (ME), defined as the right singular vector of the standardized expression matrix. The MEs were also defined as the first principal component calculated using PCA, which can summarize module behavior. Pairs of modules with high ME correlations (R > 0.8) were merged. MEs for modules were plotted by using the ggplot2 library in R. These ME scores were tested for correlation with phenotypes (regeneration time points) adjusted by a linear regression model. In more detail, a weighted signed network was computed based on a fit to scale-free topology, with a threshold softPower of 10 chosen (as it was the smallest that resulted in a scale-free R² fit). A topological overlap dendrogram was used to define modules with a minimum module size of 80 genes and the deepSplit parameter set to 2. The connectivity of every gene in every module was assessed by correlation to the MEs, or kMEs. Module membership (MM) was regarded as intramodular connectivity. MM can be combined as a systematic biological method to obtain driver genes in networks, which are highly interconnected nodes within coexpression gene modules. The driver genes were defined by the WGCNA connectivity algorithm. Each module network was viewed by VisANT (v5.0) (http://www.visantbio.com/santnet.html), which allows users to input an edge file and a node file from a WGCNA module.

**Single-cell RNA-sequencing analysis.** The preparation of the earthworm single-cell samples was performed using the following protocol: (1) 15 earthworms were cleaned and soil was removed using PBS or ddH2O. (2) We used tweezers to drag the earthworms to make its head natural extended and then quickly amputated the first four body segments. The amputated cell samples was placed into soil with fertilizers and cultivated for 0, 6, 12, 24, 48 and 72 h, and then again amputated the injured segments to isolate total RNA by using TRIzol reagent (Invitrogen, 15596-026) and RNeasy® Mini Kit (50) (QIAGEN, 74104). The first-strand cDNA was synthesized with 1 μg total RNA using a HiScript® III RT SuperMix for qPCR (+g-dDNA) kit (Vazyme, R323-01). Quantitative real-time PCR was performed using ChamQ™ Universal SYBR qPCR Master Mix (Vazyme, Q711-03). 5 biological replicates for each time point were guided. The comparative cycle threshold (Ct) method was applied to quantify the expression levels by 2^-ΔΔCt method. The β-actin was served as a reference gene to normalize the relative mRNA expression levels.
6. Thummel, R., Ju, M., Sarras, M. P. Jr. & Godwin, A. R. Both Gekko japonicus.

Received: 18 May 2019; Accepted: 30 April 2020; The authors declare that all data supporting the findings of this study are available within the Article and its Supplementary Information files or from the corresponding authors upon reasonable request.

The authors declare that all data supporting the findings of this study are available within the Article and its Supplementary Information files or from the corresponding authors upon reasonable request.

The authors declare that all data supporting the findings of this study are available within the Article and its Supplementary Information files or from the corresponding authors upon reasonable request.

The authors declare that all data supporting the findings of this study are available within the Article and its Supplementary Information files or from the corresponding authors upon reasonable request.

Received: 18 May 2019; Accepted: 30 April 2020; Published online: 27 May 2020

References

1. Gurnett, G. C., Werner, S., Barrandon, Y. & Longaker, M. T. Wound repair and regeneration. Nature 453, 314–321 (2008).

2. Liu, Y. et al. Gekko japonicus genome reveals evolution of adhesive toe pads and tail regeneration. Nat. Commun. 6, 10033 (2015).

3. Murawala, P., Tanaka, E. M. & Currie, J. D. Regeneration: the ultimate example of wound healing. Semin Cell Dev. Biol. 23, 954–962 (2012).

4. Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. Nature 554, 50–55 (2018).

5. Poss, K. D., Wilson, L. G. & Keating, M. T. Heart regeneration in zebrafish. Science 298, 2188–2190 (2002).

6. Thummel, R., Ju, M., Sarras, M. P. Jr. & Godwin, A. R. Both HOXC13 orthologues are functionally important for zebrafish tail fin regeneration. Dev. Genes Evol. 217, 413–420 (2007).

7. Wasik, K. et al. Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano. Proc. Natl Acad. Sci. USA 112, 12462–12467 (2015).

8. Bosch, T. C. Why polyps regenerate and we don’t: towards a cellular and molecular framework for Hydra regeneration. Dev. Biol. 303, 421–433 (2007).

9. Sturzenbaum, S. R., Andre, J., Kille, P. & Morgan, A. J. Earthworm genomes, genes and proteins: the (re)discovery of Darwin’s worms. Proc. Biol. Sci. 276, 789–797 (2009).

10. Lavelle, P. Earthworm activities and the soil system. Biol. Fertil. Soils 6, 237–251 (1988).

11. Cho, S.-J., Koh, K. S., Lee, E. & Park, S. C. Differential expression of three labial genes during earthworm head regeneration. Biosci. Biotechnol. Biochem. 73, 2609–2614 (2009).

12. Cho, S. J. et al. Up-regulation of multiple serine proteinases during earthworm tail regeneration. Invertebr. Reprod. Dev. 40, 103–108 (2001).

13. Cho, S. J. et al. Gene expression profile in the anterior regeneration of the earthworm using expressed sequence tags. Biosci. Biotechnol. Biochem 73, 29–34 (2009).

14. Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. Nature 493, 526–531 (2013).

15. Bhamari, A. et al. Large scale changes in the transcriptome of Eisenia fetida during regeneration. PLoS ONE 13, e0204238 (2018).

16. Zwarycz, A. S., Nossa, C. W., Putnam, N. H. & Ryan, J. F. Timing and scope of genomic expansion within Annelida: evidence from homeoboxes in the genome of the Earthworm Eisenia fetida. Genome Biol. Evol. 8, 271–281 (2015).

17. Zhang, X. et al. The sea cucumber genome provides insights into morphological evolution and visceral regeneration. PLoS Biol. 15, e2003790 (2017).

18. Belton, J.-M. et al. Hi–C: A comprehensive technique to capture the comprehensive information of genomes. Methods 58, 268–276 (2012).

19. Park, S. K., Cho, S.-J. & Park, S. C. Histological observations of blastema formation during earthworm tail regeneration. Invertebr. Reprod. Dev. 57, 165–169 (2013).

20. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8, 272–285 (2007).

21. Yin, W. et al. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer vipers. Nat. Commun. 7, 13107 (2016).

22. Gehrke, A. R. et al. Acol genome reveals the regulatory landscape of whole-body regeneration. Science 363, eaau173 (2019).

23. Ruchke, K., Hiepen, C., Becker, J. & Knauß, P. BMPs are mediators in tissue crosstalk of the regenerating musculoskeletal system. Cell Tissue Res. 347, 521–544 (2012).

24. Vinarsky, V., Atkinson, D. L., Stevenson, T. J., Keating, M. T. & Edgell, S. J. Normal newt limb regeneration requires matrix metalloproteinase function. Dev. Biol. 279, 86–98 (2005).

25. Zwaanepoel, A. & Van de Peer, Y. wgd-simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35, 2153–2155 (2019).

26. Gurley, K. A., Rink, J. C. & Sanchez Alvarado, A. Beta-catenin defines head versus tail identity during planarian regeneration and homeostasis. Science 319, 323–327 (2008).

27. Iglesias, M., Gomez-Skarmeta, J. L., Salo, E. & Adell, T. Smelling of smed-betacatenin1 generates radial-like hypercellularized planarians. Development 135, 1215–1221 (2008).

28. Petersen, C. P. & Reddien, P. W. Smed-betacatenin-1 is required for anteroposterior blastema polarity in planarian regeneration. Science 319, 327–330 (2008).

29. Yazawa, S., Umezono, Y., Hayashi, T., Tarui, H. & Agata, K. Planarian hedgehog/patched establishes anterior-posterior polarity by regulating Wnt signaling. Proc. Natl Acad. Sci. USA 106, 22329–22334 (2009).

30. Stoick-Cooper, C. L. et al. Distinct Wnt signaling pathways have opposing roles in appendage regeneration. Dev. Dyn. 235, 1390–1399 (2006).

31. Soussanis, K., Athippozhoy, A. T., Voss, S. R. & Tsonis, P. A. Plasticity for axolotl lens regeneration is associated with age-related changes in gene expression. Regeneration 1, 47–57 (2014).

32. Fraguas, S., Barberan, S. & Cebria, F. Egfr signaling regulates cell proliferation, differentiation and morphogenesis during planarian regeneration and homeostasis. Dev. Biol. 354, 87–101 (2011).

33. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).

34. Johnson, M. R. et al. Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease. Nat. Neurosci. 19, 223–232 (2016).

35. Ringler, D. et al. Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. Cell Stem Cell 10, 299–311 (2012).
37. Wenemoser, D., Lapan, S. W., Wilkinson, A. W., Bell, G. W. & Reddien, P. W. A molecular wound response program associated with regeneration initiation in planarians. *Invest. Ophthalmol. Vis. Sci.* 43, 2424–2449 (2002).

38. Oshitari, T. et al. The role of c-fos in cell death and regeneration of retinal ganglion cells. *Invest. Ophthalmol. Vis. Sci.* 31, 109–116 (1990).

39. Hua, Y. et al. The extracellular matrix protein agrin promotes heart regeneration in mice. *Nature 547*, 179–184 (2017).

40. Dong, L., Payan, D. G., Magli-Sloc, C., Cowan, D. M. & Scheller, R. H. Structure and expression of a rat agrin. *Neuron* 8, 811–823 (1991).

41. Min, I. M. et al. The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell* 2, 380–391 (2008).

42. Yoda, K., Ohnuki, Y. & Kurouawa, H. Optimization of the treatment conditions with glycogen synthase-3 inhibitor towards enhancing the proliferation of human induced pluripotent stem cells while maintaining an unaffected state under feeder-free conditions. *J. Biosci. Bioeng.* 127, 381–387 (2018).

43. Huangfu, D. et al. Induction of pluripotent stem cells from primary human fibroblasts with only OCT4 and SOX2. *Nat. Biotechnol.* 26, 1269–1275 (2008).

44. Plass, M., Solana, J. & Wolf, F. A. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360, eaau723 (2018).

45. Qadota, H. et al. A novel protein phosphatase is a binding partner for the protein kinase domains of UNC-89 (Obscurin) in *Caenorhabditis elegans*. *Mol. Biol. Cell* 19, 2424–2432 (2008).

46. Bely, A. E. Early events in annelid regeneration: a cellular perspective. *Integr. Comp. Biol.* 54, 688–699 (2014).

47. van de Lagenmaat, L. N., Landry, J. R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19, 530–536 (2003).

48. Geurquin, M. J. et al. The transcription factor EGR1 directs tendon differentiation and promotes tendon repair. *J. Clin. Invest* 123, 3564–3576 (2013).

49. Mueller, L. et al. The induction of the immediate-early-genes Egr-1, PAI-1 and PRL-1 during liver regeneration in surgical models is related to increased portal flow. *J. Hepatol.* 37, 606–612 (2002).

50. Liu, H. et al. Crucial transcription factors in tendon development and differentiation: their potential for tendon regeneration. *Cell Tissue Res.* 356, 287–297 (2014).

51. Liu, B. et al. Suppression of liver regeneration and hepatocyte proliferation in hepatocyte-targeted glypican 3 transgenic mice. *Hepatology* 52, 1060–1067 (2010).

52. Huang, Y. et al. Igf signaling is required for cardiomyocyte proliferation during zebrafish heart development and regeneration. *PLoS ONE* 8, e67266 (2013).

53. Handberg-Thorsager, M. & Salo, E. The planarian nanos-like gene Smednos is expressed in germine and eye precursor cells during development and regeneration. *Dev. Genes Evol.* 217, 403–411 (2007).

54. Rink, J. C., Yu, H. T. & Sanchez Alvarado, A. The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development* 138, 3769–3780 (2011).

55. Collin de L’Hortet, A., Gilgenkrantz, H. & Guidotti, J. E. EGFR: a master piece in G1/S phase transition of liver regeneration. *Int. J. Hepatol.* 2012, 476910 (2012).

56. Gusterson, B. A. & Hunter, K. D. Should we be surprised at the paucity of genome data? *Nat. Biotechnol.* 29, 1050–1054 (2011).

57. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736 (2017).

58. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652 (2011).

59. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl 2), ii215–ii225 (2003).

60. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* 5, 59 (2004).

61. Haas, B. J. et al. Improving the A. arribidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666 (2003).

62. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling blast to identify homologous gene sequences. *Genome Res.* 19, 143–149 (2009).

63. Birney, E., Clamp, M. & Durbin, R. GeneWise and genowise. *Genome Res.* 14, 988–995 (2004).

64. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDencemodeler and the program to assemble spliced genes. *Genome Biol.* 9, R7 (2008).

65. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189 (2003).

66. Love, M. L., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with Deseq2. *Genome Biol.* 15, 550 (2014).

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No XDB13020600, National Natural Science Foundation of China (91731304), and the Animal Branch of the GermplasmBank of Wild Species of Chinese Academy of Science (the Large Research Infrastructure Funding).

Author contributions

D.D.W., J.R. and G.J.Z. designed and led the project. Y.S. analyzed genome and transcriptome, and drafted the paper. X.B.W. analyzed the assembly and annotation of genome data. M.L.L. analyzed single-cell RNA-sequencing data. H.H.Z., X.W. and H.F.Z. performed cell differentiation and cell sorting experiments. S.S.W. sampled and processed the experimental materials. I.J.Z. and X.Y.M. did experiments. Y.L. and D.P.W. finished genome sequencing. D.M.I. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16454-8.

Correspondence and requests for materials should be addressed to G.-J.Z., J.R. or D.-D.W.

Peer review information

*Nature Communications* thanks Michel Vervoort and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information

is available at http://www.nature.com/reprints

Publisher’s note

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