

Distinct gene expression dynamics in developing and regenerating crustacean limbs

Chiara Sinigaglia, Alba Almazán, Marie Lebel, Marie Sémon, Benjamin Gillet, Sandrine Hughes, Eric Edsinger, M. Averof, Mathilde Paris

▶ **To cite this version:**

Chiara Sinigaglia, Alba Almazán, Marie Lebel, Marie Sémon, Benjamin Gillet, et al.. Distinct gene expression dynamics in developing and regenerating crustacean limbs. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119 (27), pp.e2119297119. 10.1073/pnas.2119297119 . hal-03799485

HAL Id: hal-03799485
https://cnrs.hal.science/hal-03799485

Submitted on 8 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distinct gene expression dynamics in developing and regenerating crustacean limbs

Chiara Sinigaglia[1,*], Alba Almazán[1], Marie Lebel[1], Marie Sémon[2], Benjamin Gillet[1], Sandrine Hughes[1], Eric Edsinger[3], Michalis Averof[1,*] and Mathilde Paris[1,*]

[1] Institut de Génomique Fonctionnelle de Lyon (IGFL), Centre National de la Recherche Scientifique (CNRS), École Normale Supérieure de Lyon, and Université Claude Bernard Lyon 1, 32 avenue Tony Garnier, 69007 Lyon, France

[2] Laboratoire de Biologie et Modélisation de la Cellule (LBMC), École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon, France

[3] Salk Institute for Biological Studies, 10010 N Torrey Pines Road, La Jolla CA 92037, USA

*Corresponding authors (chi.sinigaglia@gmail.com, michalis.averof@ens-lyon.fr, mathilde.paris@ens-lyon.fr)

ABSTRACT

Regenerating animals have the ability to reproduce body parts that were originally made in the embryo and subsequently lost due to injury. Understanding whether regeneration mirrors development is an open question in most regenerative species. Here we take a transcriptomics approach to examine to what extent leg regeneration shows the same temporal patterns of gene expression as leg development in the embryo, in the crustacean *Parhyale hawaiensis*. We find that leg development in the embryo shows stereotypic temporal patterns of gene expression. In contrast, global patterns of gene expression during leg regeneration show a higher degree of variation, related to the physiology of individual animals. A major driver of this variation is the molting cycle. We dissect the transcriptional signals of individual physiology from regeneration, to obtain clearer temporal signals marking distinct phases of leg regeneration. Comparing the transcriptional dynamics of development and regeneration we find that, although the two processes use similar sets of genes, the temporal patterns in which these genes are deployed are different and cannot be systematically aligned.

SIGNIFICANCE STATEMENT

Some organisms have the fascinating capacity to regenerate lost body parts. To which extent regeneration entails the re-deployment of an embryonic developmental program is a long standing question of regenerative studies, with implications for development, evolution and regenerative medicine. In this study we address this question by comparing the global transcriptional dynamics of leg regeneration and leg development in the crustacean *Parhyale hawaiensis*. We show that despite extensive overlaps in gene usage, the development and regeneration of *Parhyale* legs show distinct temporal profiles of gene expression that cannot be aligned in a coherent fashion. These results suggest that regeneration does not simply mirror development, but deploys some of the same gene modules in a different overall framework.

INTRODUCTION

Many animals have the capacity to regenerate body parts that have been lost after a severe injury. In some cases regeneration produces faithful replicas of the lost organs, which are indistinguishable from those originally developed in the embryo. This similarity in the *outcome* of development and regeneration suggests that the *processes* generating these structures could also be similar, *i.e.* that regeneration could mirror embryonic development. The fact that both take place within the same organism, relying on the same genome, makes it easy to envisage that the same molecular mechanisms and gene regulatory networks could be used in both cases.

Besides this evident connection, however, there are important ways in which development and regeneration are likely to differ. Development is a stereotypic process, unfolding from a defined starting point, in the stable and well-provisioned environment of the fertilised egg. In contrast, regeneration starts with an injury whose extent and timing are unpredictable. Regeneration also unfolds in the context of adult physiology, e.g. influenced by the nutritional status of the animal, exposure to microbes, as well as circadian, seasonal and hormonal cycles. For instance, in arthropods, the molting cycle profoundly affects the physiology of the individual and imposes physical constraints on the growth of regenerating structures (1).

In the adult body, the cellular context in which regeneration takes place differs from the embryo: different pools of progenitors are available compared with development, and differentiated cell types such as immune cells and neurons (which are not yet formed at the onset of organ development) are known to play key roles in supporting regeneration (e.g. refs 2-6).

Significant differences can also be seen in the scales in which development and regeneration unfold. Embryonic organ primordia are usually hundreds of micrometres to millimetres in size, while

adult regenerating organs can be orders of magnitude larger (see ref. 7). Such differences in size are likely to have an impact on the mechanisms that coordinate cell behaviour and cell fate across developing tissues, such as diffusion-based morphogen gradients and long range cell-cell communication. Differences may also exist in the temporal scale over which developmental and regenerative processes unfold.

In spite of these differences, numerous studies indicate that development and regeneration could share significant similarities (reviewed in ref. 8). For example, a classic study used reciprocal tissue grafts between developing limb buds and regenerating blastemas in axolotls to reveal similar patterning activities in those tissues (9). Later studies contributing to this debate have compared the roles played by specific regulatory genes (e.g. refs 10, 11), the deployment of positional markers (e.g. refs 12, 13) and the transcriptional profiles of regenerating tissues (14-16) during development and regeneration, reaching different conclusions.

The crustacean *Parhyale hawaiensis* presents an excellent system for exploring the relationships between embryonic and regenerative processes, for several reasons. First, *Parhyale* are able to regenerate their legs with high fidelity; regenerated legs are indistinguishable from the original, unharmed adult legs (17). Second, *Parhyale* are direct developers (do not undergo metamorphosis), so the adult legs directly derive from the legs developing in the embryo (18). Third, although adult legs are larger than embryonic legs, the leg primordia in embryos and regenerating adults develop on similar spatio-temporal scales. The primordia are in the order of 100 micrometres in size and consist of a few hundred cells (19, 20). The temporal scales of leg development and regeneration are also similar, spanning 4-5 days at 26˚C from primordium/blastema formation to fully patterned leg (19-21). These shared features provide a common framework for comparing development and regeneration in *Parhyale* and testing to which extent the dynamics of regeneration mirror those of development.

To compare gene usage during development and regeneration on a genome-wide scale we performed RNAseq on single legs, covering the time course of each process, from early limb buds or freshly amputated legs to fully patterned legs. Our working hypothesis was that some phases of leg regeneration, such as wound closure, may be specific to regeneration, but others like patterning, morphogenesis, growth and cell differentiation could share significant similarities. Our goals were: 1) to compare expression dynamics on a global scale and determine whether specific phases of leg regeneration can be associated, on the basis of gene expression, to specific phases of leg development, 2) to identify sets of genes that are co-expressed in distinct phases of leg development and regeneration and determine whether similar clusters of co-expressed genes are involved in development and regeneration, and 3) to determine whether these sets of co-expressed genes are deployed in the same temporal order in the embryo and in the regenerating adult leg.

We expected to recover common temporal patterns of gene expression underpinning the embryonic and regenerative time courses, consistent with the idea that some aspects of leg regeneration re-deploy mechanisms used for leg embryonic development. A failure to detect a common temporal order of gene expression would suggest that development and regeneration follow distinct trajectories.

RESULTS

Transcriptional profiling of leg development reveals stereotypic developmental profiles

To investigate transcriptional dynamics and assess individual variation in developing embryonic legs we performed RNAseq on individual T4 legs during the time course of leg development, from young limb bud stages to fully patterned and differentiated legs (21) (**Figure 1A**). We collected entire T4 legs every 6 hours, from 96 to 192 hours post fertilisation (hpf) (E_f series). In order to account for the progressive regionalization of the primordium, we also collected, when possible, the distal portion of the T4 leg: the distal 1/3 of the leg from 120 to 138 hpf, and the carpus, propodus and dactylus from 144 to 192 hpf (E_d series). The full and the distal leg samples were collected in pairs, from contralateral T4 legs of the same embryos, yielding a total of 70 samples covering the time course of leg development (**Table S1**).

Principal component analysis (PCA) on the complete RNAseq dataset (**Datasets S3, S4**) showed that the principal axis of gene expression variation, explaining 14% of the variance, strongly correlates with developmental time in both the E_f (full) and the E_d (distal) leg samples (PC1, **Figure 1B**). A weaker source of variation was linked to specific samples (PC2, see Figure S1.1). To probe the strength of the temporal signal in those data, we applied RAPToR, a method that allows predicting the developmental stage of a sample from its gene expression profile, relative to a reference time series (22). We built a reference using the E_f leg samples and used this to estimate the stage of each sample. The predictions of the model match the real developmental age of each sample accurately, not only for the E_f samples (which were used to train the model) but also for the E_d leg samples (**Figure 1C**, **Table S2**). These results indicate that the temporal dynamics of gene expression in developing legs are highly stereotypic, and that the temporal dynamics captured in the E_f and E_d series are highly coherent with each other.

Further comparing the transcriptional dynamics in the E_f and E_d samples, we found 7,963 and 1,354 genes to be differentially expressed during embryogenesis in E_f and E_d samples respectively (DESeq2, padj < 0.001, in a total of 43,212 gene models; **Figure S1.2**), with an overlap of 1,121 genes (differentially expressed genes given in **Datasets S5, S6**). We attribute the higher number of differentially expressed genes in the E_f samples to the fact that this dataset spans a longer developmental period and includes additional tissues. The tissue dissections to collect the E_d samples

were also more challenging, possibly contributing to lower sample quality (**Figure S1.1**, **Table S1**). In spite of these differences, we noticed a high similarity in gene expression dynamics between these two datasets (**Figure 1D**). Overall, this analysis shows that the temporal signal of the distal part of the leg (E_d) is largely recapitulated in the full leg series (E_f).

Transcriptional profiling of leg regeneration reveals temporal dynamics with high inter-individual variation

We performed a similar series of RNAseq experiments to investigate the temporal dynamics of gene expression during the course of regeneration in adult T4 legs, amputated at the distal end of the carpus. Previous studies have shown that the cellular activity associated with regeneration occurs within 200 microns from the amputation site (19, 23), which in these experiments corresponds approximately to the distal half of the carpus. Samples were collected every 12 hours, from the moment of amputation until 120 hours post amputation (hpa), when the legs appear to be fully patterned (19). To ensure that we have sampled patterned and differentiated legs, we also collected samples at the onset of expression of a late distal leg marker (the *Distal*^{DsRed} exon trap, collected ~150 hpa; 23, 24) and after the first molt following regeneration.

From each of these legs we collected two fragments (**Figure 2A**): one consisting of the distal-most end of the leg stump (the carpus, including the blastema and newly regenerating structures; series R_d) and one from a more proximal podomere (the distal part of the merus; series R_p). The R_d samples capture the entire region that participates actively in regeneration (19), while the R_p samples serve as controls, intended to capture transcriptional variations associated with the physiological status of each individual (e.g. molting stage, nutritional state). Overall, we collected 120 samples from 37 individuals (paired samples were collected from the left and right T4 legs in 23 individuals), spanning 13 time points, yielding a total of 60 R_d and 60 R_p samples (**Figure 2A**, listed in **Table S1**).

Principal component analysis including both the R_d and R_p series reveals several distinct sources of variation in these data (count and tpm values in **Datasets S7, S8**, respectively). PC1 captures the difference between the R_d (regenerating) and R_p (control) samples, with the notable exception of the 0 hpa (pre-amputation) and the post-molt R_d samples, which group together with the R_p series (**Figure 2B**). This distribution shows that tissues undergoing active regeneration are transcriptionally distinct from the non-regenerating samples.

PC2 reveals marked differences between the groups of samples collected from each individual (particularly in individuals marked in bold, **Figure 2C**). This variation reflects real biological differences between individuals, as we find a much higher correlation of gene expression in samples collected from the same individual than among different individuals (**Figures 2C and S2.1A**). These samples were

processed in a randomised order, so these correlations could not arise from post-processing batch effects. We return to the source of this inter-individual variation in the next section.

PC3 and PC4 capture a temporal signal corresponding to the progress of regeneration in the R_d samples (**Figures 2D and S2.2A**). On these axes, the transcriptional profile of pre-amputation samples (0 hpa) matches the profile of samples collected at the end of the regenerative time course, consistent with our expectation that regeneration is largely completed by ~120 hpa and after the following molt.

Principal Component Analysis on the Rd samples alone captures the temporal signal in PC1 (Figure 2E) and both temporal and inter-individual variation in PC2 (Figure S2.2B). Overall, our analysis reveals that regeneration and inter-individual variation are the major sources of variation in the Rd samples.

Using a reference timeline based on the Rd samples, the stage of regenerating samples (R_d) can be partly predicted based on their transcriptome (linear regression r^2=0.33; **Figure 2F, Table S3**). In contrast, most of the control samples ($R_{p)}$ are assigned to the fully differentiated state, confirming that these samples do not carry a substantial regenerative signal (linear regression r^2~0; **Figure 2F**).

Impact of the molting cycle on the transcriptional profile of adult legs

We hypothesised that the observed 'individual signal' (PC2 in **Figure 2C**) is linked to the physiological state of each animal, as it is shared by all the samples collected from each individual. Since molting is a major physiological variable in adult crustaceans, we decided to test directly the impact that the molting cycle might have on the leg transcriptome.

Selecting animals of the same age/size as in the regeneration RNAseq experiments, we monitored the molting status of 66 animals over two successive molts; we observed that this cohort molted with a mean period of 27 days (SD 7.2 days). We then collected entire T4 legs from 20 of these animals at different stages of the molting cycle (**Figure 3A**, **Table S1**) and performed RNAseq on these samples (**Datasets S9, S10**). Principal component analysis on these 20 samples shows that different stages of the molt cycle are well separated on PC1 and PC2, representing almost half of total variation (large circles in **Figure 3B**). Major transcriptional changes can be observed in the 5 days that precede molting (orange, brown and yellow circles in **Figure 3B**), followed by more stable transcriptional profiles post-molting (blue and purple circles in **Figure 3B**).

On the principal components describing the molting cycle, we projected the expression data of our regeneration time series in order to assess the molting status of each sample and its potential impact on inter-individual variation. We observed that most samples are associated with the intermolt phase, except those highlighted in bold in **Figure 2C**, which are associated with near-molt stages (**Figure**

3B). Molt-associated genes are a major driver of the inter-individual variation (seen in PC2, Figure 2C) in the regenerating leg samples (Figure S3.1).

Applying a soft clustering approach (Mfuzz) on the molting cycle dataset, we defined eight distinct sets of co-expressed genes (see **Figure S3.2** for clustering parameters, **Figure S3.3** for cluster content, **Datasets S11-12** for cluster data). The samples collected shortly before molting show the largest changes in gene expression (orange, brown and yellow phases in **Figure 3C**). We identified 131 transcription factors whose expression changes during the molt cycle (**Figure S3.4**). These factors, which include the ecdysone receptor and other known mediators of molt responses in arthropods, are prime candidates for future studies on the interplay between molting and regeneration in *Parhyale*.

These analyses confirm our initial hypothesis that molting status has a strong transcriptional influence on the regenerating leg transcriptomes. In particular, the imminence of molting deeply modifies the transcriptional state of an adult leg (**Figures 2C-E**).

Disentangling the transcriptional signals of physiology and regeneration

To investigate the transcriptional dynamics driven by the regenerative process independently of the physiological/molting status of each animal, we developed a Bayesian modelling approach (using JAGS, 25; model outlined in **Figure 4A**) to dissect the contributions of regeneration (R, bold red in **Figure 4A**) and the individual's physiology on the the transcriptome of R_d samples (grey circle in **Figure 4A**). Based on the results presented earlier, we assumed that the variation due to an individual's physiology would be shared by all the samples collected from each individual (R_d and R_p from contralateral T4 legs). In contrast, the variation in gene expression driven by the regenerative process should be specific to each R_d sample. Previous observations suggested that individual limbs can regenerate at different speeds (19), we therefore modelled the regenerative signal separately in each R_d sample, even when we collected them from the left and right T4 legs of the same individual.

The regenerative signal R was modelled as an enrichment value, similar to a fold change between R_d and the same individual's control/physiological signal measured in the R_p samples, taking into account sampling errors/variation (see **Figure 4A**; R values given in **Dataset S13**). An R value of 1 conveys that there is no difference in the expression of a given gene between R_d and R_p samples, $R>1$ means that the gene is upregulated in the regenerating sample, and $R<1$ that the gene is downregulated. Comparing the temporal profiles of R and R_d shows that R values preserve the temporal signal of regeneration but largely reduce the inter-individual variation associated with molting (**Figures S4.1A,B**).

This modelling approach is successful in extracting the expression dynamics of regeneration from the overall transcriptional variation, without introducing unintended distortions or artefacts in the

data (Figure 4B,C). The principal axes of variation correlate better with regeneration time in the modelled R data compared with the Rd data (Figure S4.2), and predictions on the regenerative stage of each sample using RAPToR are also more accurate using R instead of Rd (Figure S4.3A,C, average distance 21 versus 30 hours respectively).

A more targeted approach for removing molt-related variation was to exclude from the Rd dataset the five samples collected close to molting and the genes whose expression is significantly affected by molting (>6000 genes). RAPToR predictions made using this alternative approach have a similar accuracy with the predictions made using Bayesian modelling on the entire dataset (Figure S4.3B,C). Given that the targeted approach excludes 5 samples and >6000 genes from the analysis, we decided to pursue our study using Bayesian modelling.

The regenerative stage of each sample is predicted more accurately in early phases of regeneration (0-36 hpa) than in later phases (48-120 hpa). This mirrors our observations in live imaging experiments, in which wound closure reliably takes place in the first 1-2 days after amputation, but the onset of later events varies (19). Given these variations, instead of sample collection time, we decided to use the predictions made by RAPToR to place the regenerating samples on a common temporal scale (pseudotime), reflecting each leg's progress in regeneration based on its transcriptional profile.

Distinct transcriptional dynamics in developing and regenerating legs

Having captured the transcriptional profiles of leg development and regeneration, we turned our attention to comparing these profiles, to determine whether the dynamics of leg regeneration could mirror to some extent the dynamics of leg development. To render the embryonic data (E_f and E_d) comparable with the modelled data from regenerating legs (R), raw counts in the embryonic datasets were transformed into enrichment values (fold changes) by applying a similar Bayesian modelling as for the R values (see Methods and **Figures S4.1D-G**, **Datasets S14-S15**); we refer to these transformed data as the E and D series (based on E_f and E_d, respectively). We find that a large proportion of genes showing temporal variation during regeneration also show dynamic expression during leg development (**Figure S4.4**, **Dataset S16**).

A combined principal component analysis on the E and the R datasets reveals that, overall, gene expression varies more during development than regeneration: the two major axes of variation (PC1 and PC2) capture well the transcriptional dynamics of leg development, but not the dynamics of leg regeneration (**Figure 5A**). There is no obvious alignment between the variation seen in the embryonic and the regenerating time series.

Next, we tried to temporally align these datasets using RAPToR. We built a reference time series based on the embryonic leg (E) data and tested whether the regenerating leg (R) samples can be aligned

to this reference. Pre-amputation and post-molt samples are consistently assigned to the latest stages of the leg development series, as expected of fully differentiated leg tissues (**Figure 5B**, **Table S4**). The other samples are inconsistently placed on the developmental series, some matching early and some later phases of development with no obvious pattern, suggesting that there is no straightforward relation between the phases of leg regeneration and leg development.

When using PCA and RAPToR, global expression profiles could be dominated by specific groups of genes that show strong differential expression (e.g. terminal differentiation genes), concealing relevant expression dynamics that occur on a smaller scale (e.g. in genes involved in patterning, the control of cell proliferation and morphogenesis). To dissect the temporal dynamics of genes associated with different phases of development and regeneration we turned to a clustering approach, which classified genes into 4 major co-expression clusters in the E series and 8 major co-expression clusters in the R series (**Figure S5.1** and **Datasets S17-S28**; cluster sizes in **Table S5**; the same analysis was also performed on the D series, see **Figures S5.1, S5.4, S5.5, S5.8-10**). All the E clusters and five out of the eight R clusters appear to be associated with specific phases of development or regeneration (**Figure 5C**). In the embryonic leg data, genes in cluster E2 are expressed predominantly in the early phases of leg development, genes in E4 and E1 in mid phases, and genes in E3 in the late phase (**Figure 5C**, left). In the regenerating leg data, genes in cluster R4 are expressed early, R1 in early-mid phases, R8 in mid-late phases, and R6 and R2 are associated with differentiated legs (both pre- and post-regeneration). The remaining three R clusters (R3, R5 and R7) do not show a consistent temporal signal (**Figure 5C**, right).

Having identified clusters of co-expressed genes in the embryonic and regenerative time series, we systematically compared the gene content of these clusters to determine whether similar sets of genes are co-expressed during development and regeneration. We measured their overlap in terms of enrichment (fold change) relative to random expectation (see Methods). The largest overlaps are observed between the early developmental cluster E2 and the mid regenerative clusters R1 and R8, and between the late developmental and regenerative clusters (E3, R2 and R6) which are associated with differentiated legs (**Figures 5E**, **S5.4D** and **S5.5**). In spite of this enrichment, we find that these clusters also show significant differences in gene content (**Figure 5F**).

Examining the expression profiles of R cluster genes during leg development and E cluster genes during regeneration (Figure S5.3) did not yield any additional insights.

A finer classification of co-expressed genes yielded 32 and 37 clusters for E and R, respectively. Similar overlaps in gene content were observed at this finer resolution (**Figures 5G**). In spite of these similarities, we did not detect a well-aligned temporal sequence of expression during development and regeneration (alternative strategies for temporally aligning the gene clusters were tested, see **Figures 5G, S5.9** and **S5.10**).

Comparing the deployment of specific functional categories of genes

To associate the identified clusters with biological functions we performed a GO enrichment analysis on each cluster (**Figures S5.7** and **S5.8**). We then grouped the enriched GO terms in categories that describe various processes that contribute to leg development and regeneration (see Methods; **Figures 5D** and **S5.7**). In the embryonic leg data, the most noticeable feature is the strong enrichment of the early gene cluster with GO terms associated with cell proliferation (cluster E2), followed by phases enriched in patterning and morphogenesis (cluster E4), and the cell differentiation (cluster E3) (**Figure 5D**, left). In the regenerating leg, we observe an initial phase associated with stress, wounding, immune responses and cell death (cluster R4), followed by a phase associated with cell dedifferentiation, cell proliferation, patterning and morphogenesis (cluster R8), and then a phase associated with cell differentiation (cluster R6); a gene cluster that is associated with TOR signalling and growth is expressed throughout the process (cluster R5) (**Figure 5D**, right). This temporal sequence is in perfect agreement with the regenerative phases identified by live imaging (19).

Overall, we observe that enriched GO terms have distinct temporal distributions in these two datasets. As expected, regeneration starts with a phase of wound healing (cluster R4) that is not represented in embryonic leg development. But notably, embryonic legs express genes associated with cell proliferation and patterning/morphogenesis in distinct phases (clusters E2 and E4), whereas in leg regeneration these processes occur simultaneously (cluster R8).

Taking a more targeted approach, we also examined the temporal profiles of specific genes that are likely to be involved in immunity, cell proliferation, leg patterning and cell differentiation (**Figures 6** and **S6**; gene list in **Table S6**). The genes associated with immunity, cell proliferation and patterning were selected based on published information (particularly on *Drosophila* orthologues, see Supplementary Methods) and on the GO term analysis; genes associated with differentiated neurons and muscles were identified from *Parhyale* single-cell transcriptomic data (17).

Immunity-associated genes are markedly upregulated during the early phases of regeneration, following wounding (**Figures 6A** and **S6.4B**). The same genes are expressed only in the later phases of embryonic development, possibly connected to the differentiation of circulating haemocytes (**Figures 6A, S6.1A** and **S6.3**).

The expression profiles of genes expressed in the G2/M phase of the cell cycle indicate that cell proliferation occurs mainly during the early phase of leg embryogenesis, and the mid-late phases of regeneration (**Figures 6B, S6.1B, S6.3, S6.4**). This is consistent with data from live imaging (19; **Figure S6.5**) and with the GO enrichment analysis (**Figure 5D**).

Leg patterning genes were predominantly expressed in the mid-late phases of leg development (from 120 hpf onward), after the downregulation of genes associated with cell proliferation (**Figures 6C, S6.1, S6.2, S6.3**). During leg regeneration, this set of genes is predominantly expressed during later phases, overlapping with the expression profiles of genes associated with cell proliferation (**Figures 6C, S6.2, S6.4**).

Finally, genes associated with differentiated muscles and nerve cells are strongly expressed during the late phases of leg development in the embryo, and in fully differentiated adult legs before and after regeneration (**Figures 6D,E**, **S6.1** and **S6.3**). They are downregulated during the course of regeneration, possibly reflecting neuron and muscle de-differentiation during these stages (**Figures 6D,E** and **S6.4**).

DISCUSSION

Comparing the transcriptional profiles of developing and regenerating legs has allowed us to probe whether the process of leg regeneration recapitulates parts of embryonic development, in terms of transcriptional dynamics, on a global, genome-wide scale.

We find that the transcriptional dynamics of leg development are stereotypic and highly reproducible across individuals (**Figure 1**). The developmental stage of a leg can be predicted from the transcriptome to within ~8 hours of developmental time (**Figure 1C**). In contrast, the transcriptional dynamics of leg regeneration are embedded within strong inter-individual variation (**Figure 2**) which is largely driven by the molting cycle (**Figure 3B**). This is consistent with the fact that, unlike development which occurs in the relatively stable environment of the egg, regeneration takes place in the context of the complex physiology of the adult. Even after correcting for inter-individual variation, regenerating legs show less stereotypic temporal profiles of expression than developing embryonic legs (compare **Figures 1C** and **4C**, expression profiles in **Figure 6**).

After filtering out inter-individual variation, through Bayesian modelling, we are able to recover more clearly the transcriptional dynamics of leg regeneration (**Figure 4**). Like the dynamics of leg development in the embryo, these reveal distinct phases of gene expression that unfold during regeneration, with most variation occurring during early-mid regenerative stages (**Figure 4C**); this is consistent with variation in cell dynamics observed by live imaging (19; **Figure S2.3**). Using GO term enrichment analysis we can assign putative gene functions to each of these phases. This analysis reveals distinct phases for wound healing, metabolic reprogramming (during a period previously described as a phase of quiescence; 19), cell proliferation and morphogenesis, and finally cell differentiation (**Figures 5C** and **6**).

We have tried to relate these phases of leg regeneration to the time course of leg development by comparing global transcriptional dynamics (**Figure 5A,B**), sets of co-expressed genes (**Figures 5C-G**, **S5.3** and **S5.10**), or the transcriptional dynamics of genes involved in cell proliferation, patterning and cell differentiation (**Figures 6B-E**, **S6.1-4**). While we observe that overlapping gene sets are implicated in both leg development and regeneration, we find that the temporal order in which they are deployed is not the same. This is true not only in phases and processes that are likely to be unique to regeneration – e.g. wound healing, immune/stress responses, metabolic reprogramming – but also in processes like cell proliferation, patterning and morphogenesis, which are shared between development and regeneration. We conclude that the time course of leg regeneration is not collinear with that of leg development.

A similar approach has been used recently to compare the transcriptional dynamics of development and regeneration in the sea anemone *Nematostella vectensis* (26) and in the zebrafish heart (27). Notwithstanding differences in experimental design (e.g. sample pooling masking inter-individual variation), the results of these studies echo some of the conclusions we present here. Similar to what we observe in crustacean legs, both in the body of *Nematostella* and in the zebrafish heart, the transcriptional dynamics of development are more pronounced than the dynamics of regeneration (e.g. compare **Figure 5A** with ref. 26 figure 1D) and the comparison of transcriptional dynamics revealed both shared and divergent patterns of gene deployment over time.

Our analysis does not exclude that a core set of regulatory genes could coordinate leg development and regeneration in similar ways. For example, patterning mechanisms in development and regeneration could share common regulators and regulatory interactions, as suggested by previous studies (see refs 8, 9). Our work highlights, however, that in spite of any putative common underlying regulators, the global transcriptional dynamics of development and regeneration are largely distinct. The similar results obtained in distant species – cnidarians, crustaceans and fish – start to build a coherent picture in which regeneration is not a straightforward replay of development, but deploys some common modules in a different overall framework.

METHODS

RNAseq design and sequencing

Embryonic dataset. Parhyale females of the Chicago-F inbred line (28) were collected after fertilization, and their embryos removed about 3 days post fertilization. Each brood was kept separately, in a temperature-controlled incubator set to 27˚C, in sterile 6-well plates (Costar, #3516), in filtered artificial seawater (FASW; salinity at 30 PSU) containing antibiotics (Gibco, #15240-062, at 1X final concentration). Embryos were staged 3-4 days post fertilization. Two series of single leg samples were collected from these embryos at 6-hour time intervals. The E_f series consisted of entire developing T4 legs, collected from 96 hpf to 192 hpf (40 samples in total). These samples included the primordia of the basis, the ischium, the merus, the carpus, the propodus, the dactylus, and of the coxal plates and the gills. The E_d series of samples included only the distal-most part of the leg: the prospective merus, carpus, propodus and dactylus from 120 to 138 hpf, when these podomeres cannot yet be morphologically distinguished (9 samples), and the developing carpus, propodus and dactylus from 144 to 192 hpf (21 samples). At least 2-3 samples were collected for each time point per series. E_f and E_d samples were collected from contralateral T4 legs of the same individuals.

Embryos were dissected in the lid of a 5 ml Protein LoBind Tube (Eppendorf, #0030108302), in 1% BSA in FASW (80 ul). The eggshell was removed with fine forceps (Fine Science Tools, #11254-20), and legs were dissected with borosilicate needles (pulled capillaries; Sutter, #B100-50-15). Samples were transferred in 100 ul of ice-cold lysis solution (Agilent Absolutely RNA Nanoprep kit, #400753), homogenized though brief pipetting, and flash-frozen in liquid nitrogen. RNA extraction was performed with the Agilent Absolutely RNA Nanoprep kit (#400753), following manufacturer's instructions, and eluted in 10 ul of elution buffer. RNA extraction was performed in randomized batches, to avoid shared batch effects in biological replicates. As the concentrations of RNA was too low to be directly quantified in these extracts, they were treated as follows: 9 ul from each sample were directly used for cDNA amplification over 15 cycles of LD-PCR, using the SMART-Seq v4 Ultra Low Input RNA kit for sequencing (Takara Bio, #634898) and the SeqAmp DNA Polymerase (Takara Bio, #638509); 1 ul of cDNA was then used for Qubit quantification (4.0 HS DNA), measuring in the range of 0.2- 0-7 ng/ul. Libraries were synthesized from 1 ng of cDNA, using the Nextera XT DNA Library Preparation Kit (Illumina, #FC-131-1096; with Dual indexing strategy, i7 and i5), and with a protocol that included an accelerated cooling step on ice after the 55°C step. Quantification and validation of libraries were done with both Qubit 4.0 (HS DNA kit, Thermofisher) and Tapestation D5000 equipment using the D5000 ScreenTape System (Agilent, #5067- 5588 and #5067- 5589). QC libraries were normalized and then loaded into an Illumina NextSeq 500 sequencing system using NextSeq 500 High Output Kit with 76 bp single-end sequencing,

according to the manufacturer's instructions (Illumina, San Diego, CA, USA). Further details about the sequenced samples are provided in **Table S1**. Raw sequencing data is deposited at GEO, accession number GSE196485.

Regeneration dataset. Adult males of the Chicago-F line (28) and the Distal^{DsRed} line (24), measuring ~1 cm in length and with no damaged appendages, were selected and kept individually in homogeneous conditions – photoperiod (12:12 hours light:dark cycle), temperature (25-26˚C) and medium (individual containers separated by mesh, sharing artificial seawater at 30 PSU) – for three months prior to experiment. The same conditions were kept during the course of sampling. Sampling was performed between 8:00 and 10:00 am. In order to test for the inter- and intra-individual variability of the regenerative process, both T4 legs of each animal were amputated simultaneously, proximally to the carpus/propodus joint. Samples from the Chicago-F line were harvested pre- and post regeneration, and every 12 hour, from 0 to 120 hpa; samples from the Distal^{DsRed} line were collected when the DsRed signal became visible (around 150 hpa). Samples from the same individual were collected at the same moment. Due to the observed variability in the regenerative sequence, samples were processed and sequenced individually, as follows: *i)* from each animal, either both or only one T4 leg was harvested, *ii)* from each leg, two fragments were isolated, one including the regenerating podomere(s) (R_d series, localized on the carpus) and one its proximal control podomere (R_p series, localized on the distal part of the merus from the same leg). Five paired samples were collected per time point, with the exception of Distal^{DsRed} line and post-molt samples (2 samples each). In total 60 regenerating and 60 control samples were collected; a scheme is provided in **Figure 2A**, and more details about the samples are provided in **Table S1**.

Leg fragments were immediately transferred in 1.5 ml LowBind tubes with 500 ul of ice-cold lysis buffer (Reliaprep RNA Tissue MiniPrep System, Promega, #Z6111), vortexed, then transferred to a sterile multiwell plate for manual disruption of cuticle with a clean surgical blade. The sample was then re-transferred to the tube, vortexed, and frozen in liquid nitrogen, for then being stored at -80°C until further processing. RNA extraction was randomized, avoiding to process at the same time related samples. RNA was extracted with the Reliaprep RNA Tissue MiniPrep System (Promega, #Z6111), and eluted in 15 ul of nuclease-free water (Invitrogen, #AM9937). RNA quality was assessed with TapeStation D5000 (Agilent RNA ScreenTape High Sensitivity system: #5067- 5579, #5067- 5580, #5067-5581), and 1 ng of RNA was used for cDNA synthesis (SMART-Seq v4 Ultra Low Input RNA kit, Takara Bio, #634898). Verification of library quality and sequencing were done as for the embryonic dataset (see above). Raw sequencing data is deposited at GEO, accession number GSE196485.

Molting dataset. 66 Chicago-F animals, selected as previously described, were individually kept and monitored over two successive molts (ca. 3 months), in order to determine their molting cycles: we observed that this cohort molted with a mean period of 27 days (SD 7.2). One entire T4 leg was harvested per animal, and RNA was extracted as described above. For the pre-molt samples, individual animals were monitored for molting every day after harvesting the leg; based on the time of molting, the harvested legs were assigned to one of the pre-molt categories (1-2, 3, 4 or 5 days before molting). The post-molt samples were collected 1, 2, 9 or 10 days post molting. A total of 20 legs deriving from animals at different stages of their molting cycle was sequenced. Library preparation and sequencing were done as described above. Details about the sequenced samples are provided in **Table S1**. Raw sequencing data is deposited at GEO, accession number GSE196485.

Adult entire leg dataset. In order to help build a new reference transcriptome (see below), two additional full T4 leg samples from non-regenerating Chicago-F males were collected and processed as described above. Details about the sequenced samples are provided in **Table S1**, sequencing results can be found in **Datasets S7** and **S8** (counts and tpm values, respectively). Raw sequencing data is deposited at GEO, accession number GSE196485.

Reference transcriptome assembly

Transcriptome assembly and annotation. Sequenced reads were mapped to a modified version of the available *Parhyale hawaiensis* genome assembly Phaw_5.0 (https://www.ncbi.nlm.nih.gov/assembly/GCA_001587735.2/, see Supplementary Methods), using hisat2 v2.1.0. Gene models were built from the RNAseq data generated in this study, combined with a previous gene annotation; overlapping genes were removed and split genes were identified based on sequence similarity with PacBio long reads (see Supplementary Methods). After a final manual curation step, the final gtf file contains 54,718 genes (**Dataset S1**) and the final list used for further analysis includes 52,759 genes (**Dataset S2**). Orthology annotation was performed using BLASTP (results given in **Dataset S29**). See the Supplementary Methods for more details.

Analyses of the RNAseq datasets

Read mapping and quantification. For all RNAseq datasets, reads were mapped to the 54,718 gene models (see above), using kallisto v. 0.42.5 (29). Count and tpm values are provided in **Datasets S3, S4** and **S7-S10**. See the Supplementary Methods for further details.

Genes for which more than 2 reads mapped on average on each sample of the embryonic dataset or the regeneration dataset, were kept for further analysis (43,212 and 43,968 genes for the embryogenesis and regeneration datasets, respectively).

Time series analysis. Rather than comparing gene expression among specific time points, our strategy is to use time as a continuous variable. Thus, rather than binning samples on discrete time points and considering these as replicates, we investigate temporal changes in gene expression by comparing individual samples over a continuous time course. To show that sampling density across the developing and the regenerating leg time course is sufficient for a robust analysis of transcriptional dynamics, we used a subsampling approach (see Supplementary Methods); the results are presented in Figure S5.9.

Normalization and visualization of transcriptional dynamics. Count and tpm matrices were first quantile-normalized (limma R package, v. 3.48.0; 30), then log transformed (log(x+1)). The JAGS-transformed values were log transformed. Details on Principal Component Analysis and heatmaps are provided in Supplementary Methods.

Differential expression analysis. We used the R package DESeq2 (31) on raw counts for identifying genes differentially expressed during embryogenesis - separately in the E_f and the E_d series - and using hpf as the explanatory variable. Genes with a p.adj < 0.001 were selected (**Datasets S5, S6**). In order to determine the list of molting-related genes (which we used at different steps to assess the efficiency of the removal of the physiological signal from the R values, **Figure S4.1B**), we also applied DESeq2: molting-related and unrelated genes were identified as genes significantly differentially expressed between time windows “1-2 days before molt” and “9-10 days after molt” (p.adj < 0.001) while genes unrelated to molting were not differentially expressed (p.adj > 0.5). The results of the DESeq2 analysis are provided at https://zenodo.org/record/6420682/files/DESeq2_IresO_molt.gz, as an R software object (rds format).

Identification of co-expressed gene sets. We applied a soft clustering approach, using the R package Mfuzz (v. 2.52.0; 32), setting sd = 0.2, and membership = 0.8 for calculating the eset object. The optimal number of clusters was estimated from the inflection points of the Dmin function (iterations = 100). In order to plot expression dynamics, we extracted the cluster centroids values, which were used to build an input matrix for the heatmaps. For the molting cycle dataset, we considered the entire transcriptome; we found that 15,646 genes were assigned to clusters, identifying 8 co-expressed gene sets (**Figures 3C and S3.1**, **Dataset S11**). For the rest of the analysis, we considered the union of the 20,000 most variable genes within each E, D and R dataset (var() function of base R), which resulted in 27,709 genes (**Figure S4.4**, **Dataset S16**). We identified 8 (19,731 genes were clustered) or 37 (15,665

519 genes) clusters for the R dataset, 4 (20,014 genes) or 32 (15,529 genes) for the E dataset, 4 (17,033
520 genes) or 12 (14,883 genes) for the D dataset. Data is available in **Figure S5.1**, **Tables S18-S29**.

521 *Comparison of clusters.* Fold enrichment scores were computed as follows: we took an approach
522 inspired from ref. 33, where a hypergeometric distribution on gene counts is used to estimate an
523 enrichment score between gene sets. The enrichment fold was calculated as the ratio of the observed
524 number of genes that overlapped between two clusters over the expected number. The overlap
525 between clusters was further assessed with chord diagrams (**Figure 5F**), for which the circlize package (v.
526 0.4.13; 34) was requested, and venn diagrams (**Figures S4.4, S5.5**), for which we used the RVenn
527 package (v. 1.1.0; 35).
528 The clustree package (v. 0.4.3; 36) was used to assess the correspondence between the clusters
529 (**Figures S5.6 and S5.9**).

530 *GO enrichment analysis.* Enriched GO terms were identified using the packages clusterProfiler (v. 4.0.0;
531 37), org.Dm.eg.db (v. 3.13.0; 38) and enrichplot (v. 1.12.0; 39), based on the GO terms assigned to the
532 best blastp hit of each *Parhyale* gene in *Drosophila* (for 14,741 *Parhyale* genes, e-value <0.001). The
533 parameters used for the GO term analysis were: ont = BP, pvalueCutoff = 0.05, and qvalueCutoff = 0.05,
534 minGSSize = 4. Results are presented in **Figures 5D, S3.2, S5.7** and **S5.8** (dotplots for Biological Process,
535 categories to display = 50). The list of significant GO terms (p-value < 0.005) was further trimmed for
536 display (**Figures 5D, S5.7**) using the Revigo algorithm (http://revigo.irb.hr/; 40). We set the following
537 parameters: allowed similarity = tiny, semantic similarity measure = SimRel.
538
539 **Bayesian modelling of regenerating and embryonic datasets**

540 Computations were performed using JAGS via the R package rjags (25). Details on modelling are
541 provided in the Supplementary Methods. The resulting R, E and D transformed values are provided in
542 **Datasets S13-S15** (values have been further log transformed).
543
544 **Predicting regeneration and developmental stages using RAPToR**

545 In order to infer the progression of regeneration or development based on transcriptomic data, we used
546 the R package RAPToR v1.1.4 (Real Age Prediction from Transcriptome staging on Reference), a recently
547 developed tool to accurately predict individual samples' developmental age from their gene expression
548 profiles (22).
549 For building RAPToR references we used the function ge_im with the formula formula = "X ~
550 s(hpa, bs = 'ts') and parameter dim_red="pca". For the RAPToR reference based on Ef values for **Figure**
551 **1C**, we used the normalized and log-transformed tpm values and selected the genes variable in the E_f

samples (intersection of the top 20,000 variable genes as calculated by DESeq2), excluding the genes with a low expression (75th percentile count above 10, with a final set of 16,199 genes) and 30 PCs (values in **Table S2**). To build the RAPToR reference based on Rd values for **Figure 2F**, we used the top 20,000 most variable genes in samples R_d as calculated by DESeq2, excluding the genes with a low expression (75th percentile count above 10, with a final set of 12,438 genes). As regeneration is less synchronous than embryonic development (19), we were concerned that the sampling timing would not faithfully reflect regeneration progression and that our samples were an imperfect reference. To avoid overfitting, we made two changes to the standard RAPToR protocol for building a reference: we used only 3 PCs, and for each sample we built a separate reference excluding the sample being tested. Pre-amputation samples were given a timing of 300 hpa that would place them at the end of the regenerative sequence, and post-molt samples were excluded from the reference (values are provided in **Table S3**). To build the RAPToR references based on R values for **Figures 4C, 5C, S5.2-4, S5.9, S5.10, 6, S6.1** and **S6.2**, we used the 20,000 top genes with the most variable R values, further excluding the genes with a low expression (75th percentile count above 10, for a final set of 12,434 genes). To avoid overfitting, we followed the same procedure described above (values in **Tables S7** for R samples, **S8** for E samples, and **S9** for D samples). For the RAPToR reference based on E values for **Figure 5B**, we used a gene list as exhaustive as possible, constituted of the union of the 20,000 top genes with the most variable R, E or D values, and excluding the genes with a low expression (75th percentile count above 10, for a final set of 16,759 genes) and 10 PCs (values in **Table S4**). For the RAPToR reference based on E values for Figures 5C, 5.9C, 6 and S6.2A, we used the 20,000 top genes with the most variable E values, excluding the genes with a low expression (75th percentile count above 10, yielding a final set of 14,632 genes), and 3 PCs. Correlation between predicted time of amputation and real time of amputation (**Figures 2F** and **4C**) was computed excluding post-molt and t0 samples, the real timing of the former being too uncertain and the later being considered as an end-point to regeneration.

Data and code availability

R code and input files are provided in https://zenodo.org/record/6420682/files/R_data_Sinigaglia_Embryo_reg_05042022-20220407T075816. Raw sequencing results are deposited at GEO, accession number GSE196485.

ACKNOWLEDGEMENTS

We thank Romain Bulteau for his help with RAPToR, Philippe Veber and Bastien Boussau for their help with rjargs, Nipam Patel for the *Parhyale* used for the Iso-seq experiment and the Salt Lake City

sequencing facility for sequencing the Iso-seq libraries, and Enrique Arboleda, Pascale Roux and Laetitia Lebre for technical support. We thank Margarida Cardoso, Nikos Konstantinides and Jordi Casanova for critical feedback, and an anonymous reviewer for detailed and thoughtful suggestions. This work was funded by the European Research Council, under the European Union Horizon 2020 programme (project ERC-2015-AdG #694918). Alba Almazán was supported by the Marie Curie ITN programme 'EvoCell', under the European Union Horizon 2020 programme (project H2020-MSCA-ITN-2017 #766053).

REFERENCES

1. Skinner, D. M. (1985). Molting and regeneration. In The Biology of Crustacea: Integument, Pigments, and Hormonal Processes (eds. Bliss, D. E. and Mantel, L. H., pp. 43–146. Academic Press.

2. Singer, M. (1952). The influence of the nerve in regeneration of the amphibian extremity. The Quarterly Review of Biology 27, 169–200.

3. Kumar, A., Godwin, J. W., Gates, P. B., Garza-Garcia, A. A. and Brockes, J. P. (2007). Molecular basis for the nerve dependence of limb regeneration in an adult vertebrate. Science 318, 772–777.

4. Kyritsis, N., Kizil, C., Zocher, S., Kroehne, V., Kaslin, J., Freudenreich, D., Iltzsche, A. and Brand, M. (2012). Acute inflammation initiates the regenerative response in the adult zebrafish brain. Science 338, 1353–1356.

5. Godwin, J. W., Pinto, A. R. and Rosenthal, N. A. (2013). Macrophages are required for adult salamander limb regeneration. Proc. Natl. Acad. Sci. U.S.A. 110, 9415–9420.

6. Sinigaglia, C. and Averof, M. (2019). The multifaceted role of nerves in animal regeneration. Curr. Opin. Genet. Dev. 57, 98–105.

7. Brockes, J. P. and Kumar, A. (2005). Appendage regeneration in adult vertebrates and implications for regenerative medicine. Science 310, 1919–1923.

8. Nacu, E. and Tanaka, E. M. (2011). Limb regeneration: a new development? Annu. Rev. Cell Dev. Biol. 27, 409–440.

9. Muneoka, K. and Bryant, S. V. (1982). Evidence that patterning mechanisms in developing and regenerating limbs are the same. Nature 298, 369–371.

10. Fan, C.-M., Li, L., Rozo, M. E. and Lepper, C. (2012). Making skeletal muscle from progenitor and stem cells: development versus regeneration. WIREs Dev Biol 1, 315–327.

11. Czarkwiani, A., Dylus, D. V., Carballo, L. and Oliveri, P. (2021). FGF signalling plays similar roles in development and regeneration of the skeleton in the brittle star *Amphiura filiformis*. Development 148, dev180760.

12. Bosch, M., Bishop, S.-A., Baguna, J. and Couso, J.-P. (2010). Leg regeneration in *Drosophila* abridges the normal developmental program. Int. J. Dev. Biol. 54, 1241–1250.

13. Roensch, K., Tazaki, A., Chara, O. and Tanaka, E. M. (2013). Progressive specification rather than intercalation of segments during limb regeneration. Science 342, 1375–1379.

14. Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M., Nowoshilow, S., Kageyama, J., Khattak, S., et al. (2018). Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science 362, 421.

15. Storer, M. A., et al. (2020). Acquisition of a Unique Mesenchymal Precursor-like Blastema State Underlies Successful Adult Mammalian Digit Tip Regeneration. Developmental Cell 52, 509–524.

16. Aztekin, C., Hiscock, T. W., Gurdon, J., Jullien, J., Marioni, J. and Simons, B. D. (2021). Secreted inhibitors drive the loss of regeneration competence in *Xenopus* limbs. Development 10.1242/dev.199158

17. Almazán, A., Çevrim, Ç., Musser, J. M., Averof, M. and Paris, M. (2021). Regenerated crustacean limbs are precise replicas. bioRxiv doi: https://doi.org/10.1101/2021.12.13.472338.

18. Paris, M., Wolff, C., Patel, N. and Averof, M. (2021). The crustacean model *Parhyale hawaiensis*. Preprints 10.20944/preprints202106.0018.v1

19. Alwes, F., Enjolras, C. and Averof, M. (2016). Live imaging reveals the progenitors and cell dynamics of limb regeneration. Elife 5, 73.

20. Wolff, C., Tinevez, J.-Y., Pietzsch, T., Stamataki, E., Harich, B., Guignard, L., Preibisch, S., Shorte, S., Keller, P. J., Tomancak, P., et al. (2018). Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. Elife 7, e34410.

21. Browne, W. E., Price, A. L., Gerberding, M. and Patel, N. H. (2005). Stages of embryonic development in the amphipod crustacean *Parhyale hawaiensis*. genesis 42, 124–149.

22. Bulteau R. and Francesconi M. (2021). Real age prediction from the transcriptome with RAPToR. bioRxiv 2021.09.07.459270

23. Konstantinides, N. and Averof, M. (2014). A common cellular basis for muscle regeneration in arthropods and vertebrates. Science 343, 788-791.

24. Kontarakis, Z., Pavlopoulos, A., Kiupakis, A., Konstantinides, N., Douris, V. and Averof M. (2011). A versatile strategy for gene trapping and trap conversion in emerging model organisms. Development 138, 2625-2630.

25. Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X.

26. Johnston, H., Warner, J. F., Amiel, A. R., Nedoncelle, K., Carvalho, J. E. and Röttinger, E. (2021). Whole body regeneration deploys a rewired embryonic gene regulatory network logic. bioRxiv doi.org/10.1101/658930.

27. Weinberger, M., Simões, F. C., Sauka-Spengler, T. and Riley, P. R. (2021). Distinct epicardial gene regulatory programmes drive development and regeneration of the zebrafish heart. bioRxiv 2021.06.29.450229.

28. Kao, D., et al. (2016). The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. Elife 5, e20062.

660 29. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq
661 quantification, Nature Biotechnology 34, 525-527.

662 30. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth GK (2015). limma powers
663 differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids
664 Research 43, e47.

665 31. Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and
666 dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550.

667 32. Kumar, L., and Futschik, M. (2007). Mfuzz: a software package for soft clustering of microarray
668 data. Bioinformation 2, 5-7.

669 33. Kowarsky, M., et al. (2021) Sexual and asexual development: two distinct programs producing
670 the same tunicate. Cell Reports 34, 108681.

671 34. Gu, Z., Gu, L., Eils, R., Schlesner, M. and Brors, B. (2014). circlize Implements and enhances
672 circular visualization in R. Bioinformatics 30: 2811-2.

673 35. Akyol, T. G. (2019). RVenn: Set Operations for Many Sets. R package version 1.1.0.
674 https://CRAN.R-project.org/package=RVenn

675 36. Zappia L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at
676 multiple resolutions. Gigascience 7, giy083.

677 37. Yu, G., Wang, L., Han, Y. and He, Q. (2012). clusterProfiler: an R package for comparing biological
678 themes among gene clusters. OMICS 16, 284-287.

679 38. Carlson, M. (2019). org.Dm.eg.db: Genome wide annotation for Fly. R package version 3.8.2.

680 39. Yu, G. (2021). enrichplot: Visualization of Functional Enrichment Result. R package version
681 1.12.0, https://yulab-smu.top/biomedical-knowledge-mining-book/

682 40. Supek, F., Bošnjak, M., Škunca, N. and Šmuc T. (2011). REVIGO summarizes and visualizes long
683 lists of Gene Ontology terms. PLoS ONE 6, e21800.
684

Figure 1: Transcriptional profiling of *Parhyale* leg development. (**A**) Morphology of *Parhyale* embryo and sampling of developing legs (E_f and E_d samples highlighted in grey and in blue, respectively). (**B**) Principal component analysis of the E_f and E_d series. PC1, representing 14% of the variance, correlates with developmental stage. (**C**) The developmental stage of the E_d samples is well predicted by RAPToR, using a reference built from the E_f samples (excluding the sample being tested, see Methods). (**D**) Heatmap representing the expression of 8196 genes that are differentially expressed in the E_f and E_d time series. The dashed rectangle marks the developmental period that is covered by both the E_f and the E_d series.

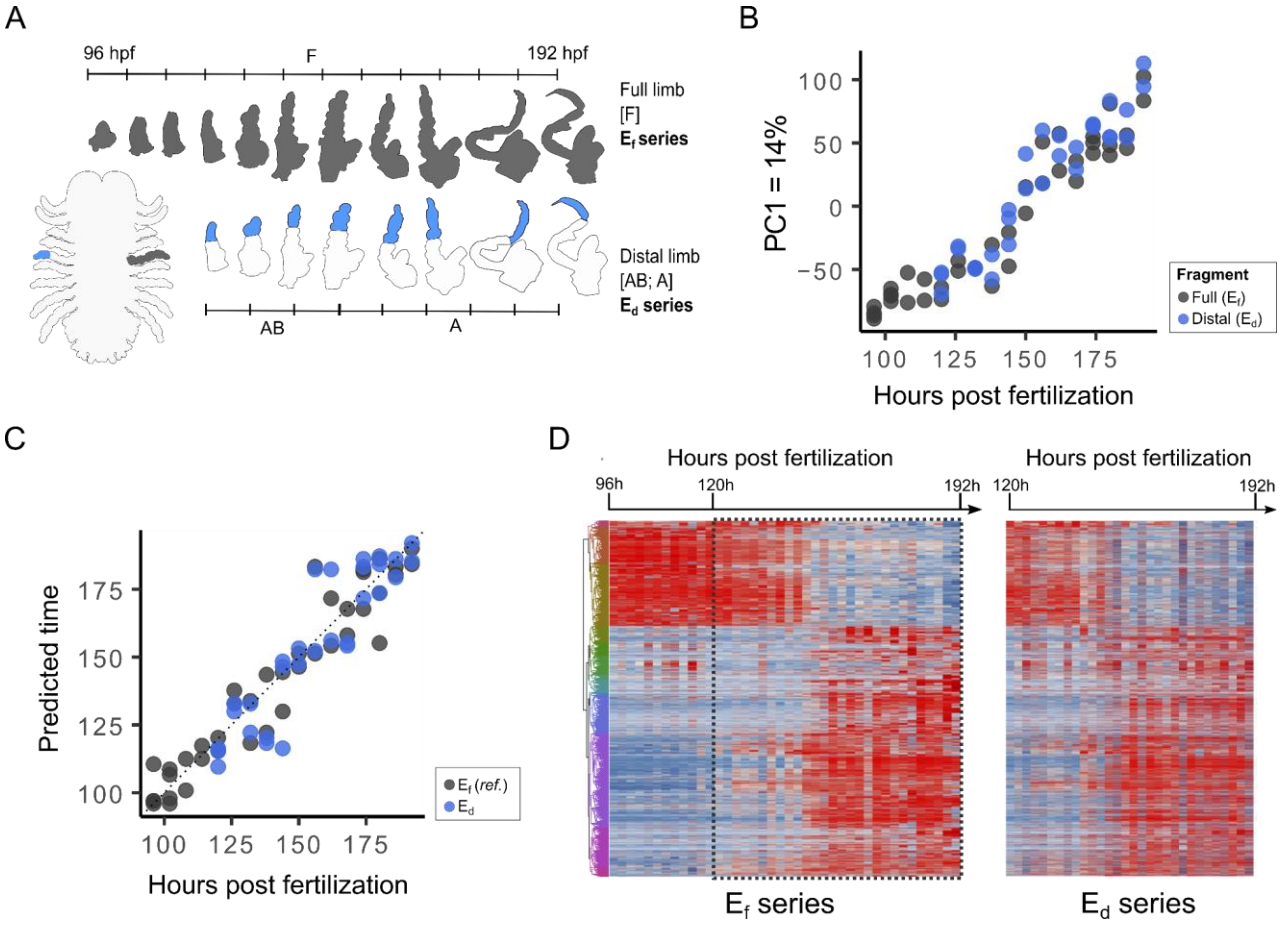
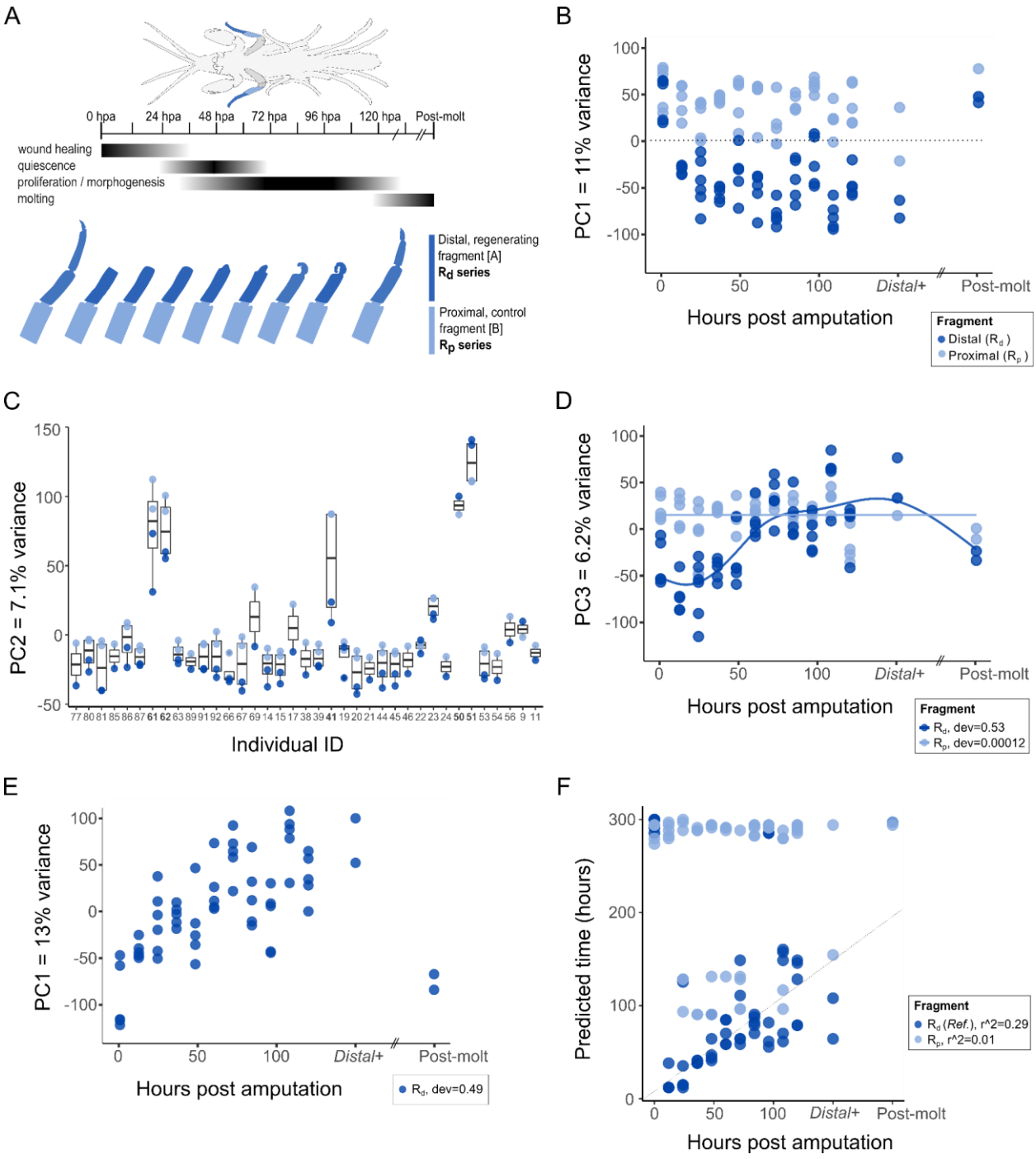


Figure 2: Transcriptional profiling of *Parhyale* leg regeneration. (**A**) Morphology of *Parhyale* adult and sampling of regenerating legs (regenerating R_d and control R_p samples, highlighted in dark and light blue, respectively). The events of the different phases of regeneration, as established by live imaging (19), are indicated. (**B-E**) Principal component analysis of the R_d and R_p series. **(B)** PC1 separates the regenerating R_d samples from the pre-amputation (0 hpa), post-molt and control (R_p) samples. (**C**) Variation in PC2 is associated with the individual from which each sample was collected; R_d and R_p samples from the same individual show similar values (x axis, individuals ordered by time after amputation). (**D**) PC3 captures temporal changes that occur during regeneration in R_d, but not R_p. (E) PC1 of principal components analysis applied to the Rd series only, capturing temporal changes during regeneration. (**F**) Prediction of regenerative stage by RAPToR, using a reference built on the R_d series. To build the reference, fully differentiated legs (pre-amputation) were assigned to 300 hpa and the sample being tested was excluded (see Methods). RAPToR makes reasonable predictions of the stage of most R_d samples, and matches most R_p samples with fully differentiated legs (300 hpa). The average absolute distances between real time of collection and predicted time for the Rd and Rp samples are 30 and 170 hours, respectively. dev: deviance explained (gam regression, excluding the post-molt samples); r^2: r squared (linear correlation).


A
0 hpa
24 hpa
48 hpa
72 hpa
96 hpa
120 hpa
Post-molt
wound healing
quiescence
proliferation / morphogenesis
molting
Distal, regenerating fragment [A]
Rd series
Proximal, control fragment [B]
Rp series
B
PC1 = 11% variance
Hours post amputation
Distal+
Post-molt
Fragment
Distal (Rd)
Proximal (Rp)
C
PC2 = 7.1% variance
Individual ID
D
PC3 = 6.2% variance
Hours post amputation
Fragment
Rd, dev=0.53
Rp, dev=0.00012
E
PC1 = 13% variance
Hours post amputation
Rd, dev=0.49
F
Predicted time (hours)
Hours post amputation
Fragment
Rd (Ref.), r^2=0.29
Rp, r^2=0.01

713

Figure 3: Impact of the molting cycle on the transcriptional profile of *Parhyale* legs. (**A**) Single T4 legs (dark grey) were sampled at different stages of the molting cycle: on the 5 days that precede molting (orange to yellow), 1-2 days post molt (blue) and 9-10 days post molt (purple). (**B**) Principal component analysis of these samples (large circles) captures molt-associated differences in PC1 and PC2. Projecting the R_d and R_p data on this PCA (in grey) reveals that the outliers of Figure 2C (identified by number) were in the process of molting, whereas most other samples were in post-molt/intermolt phases (also see Figure S3.5). (**C**) Fuzzy c-means clustering of genes, based on expression values from the molting dataset, reported as centroid values. Three main transcriptional phases are observable, corresponding to post-molt/intermolt (clusters 6, 7, 1, 5, 2), 5-3 days pre-molt (clusters 4 and 3) and 1-2 days pre-molt (cluster 8) periods.

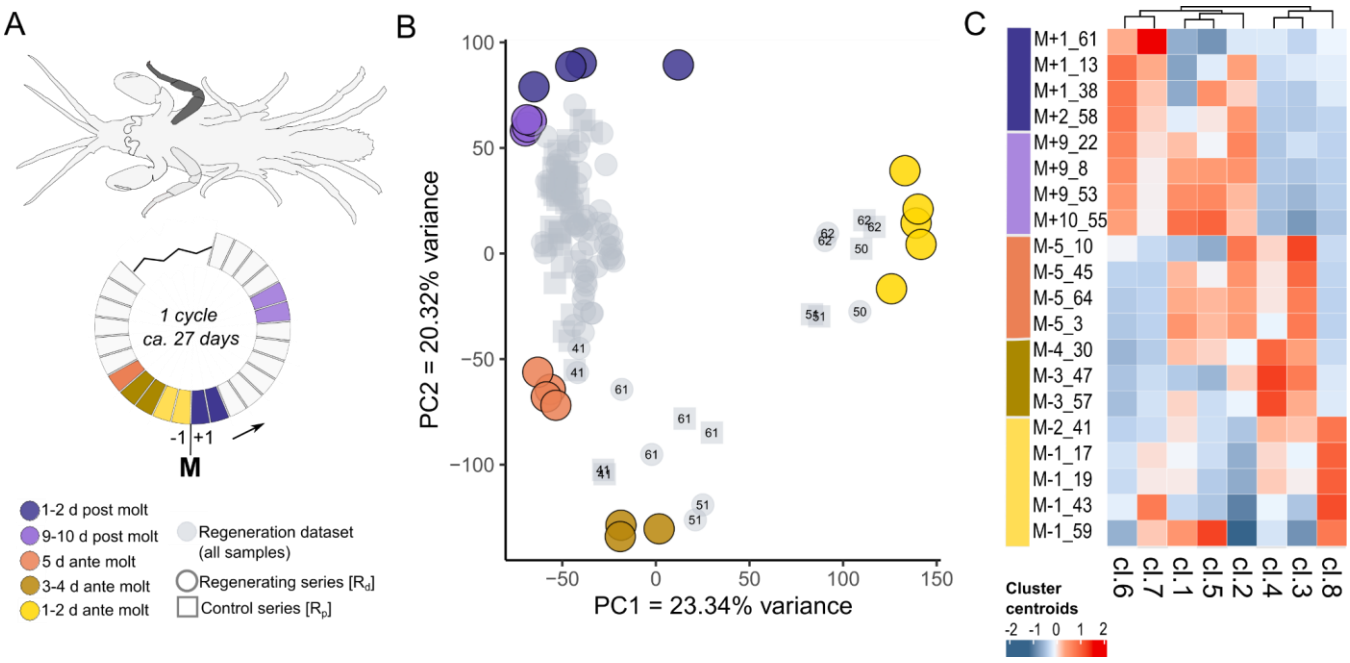
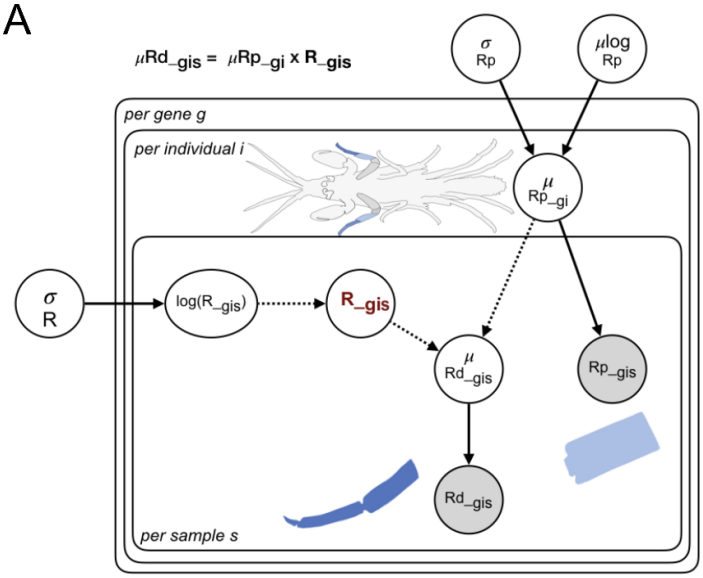
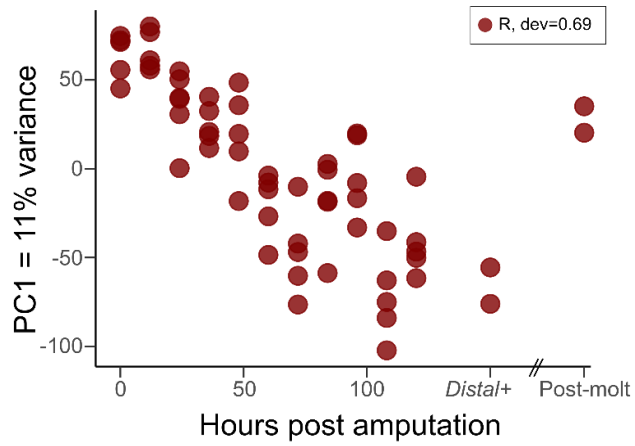


Figure 4: Modelling of the regenerative signal. (**A**) Directed acyclic graph illustrating the model used to extract R values (dark red) from the raw counts of R_d and R_p series (grey circles). Gene levels in R_d samples (dark blue) are modelled as the product of gene levels in the corresponding R_p samples (light blue) multiplied by an R value (sampling error taken into account). (**B**) Principal component analysis of the R values: PC1 is strongly associated with the stage of regeneration. (**C**) Prediction of regenerative stage by RAPToR, using a reference built on the R series. To build the reference, fully differentiated legs (pre-amputation) were assigned to 300 hpa and the reference excluded the sample being tested (see Methods). Predictions are robust particularly in the early stages and they are largely independent of the gene set used to build the reference (Figure S5.2A). The average absolute distance between real time of collection and predicted time for the R samples is 21 hours. dev: deviance explained (gam regression, excluding the post-molt samples); r^2: r squared (linear correlation).

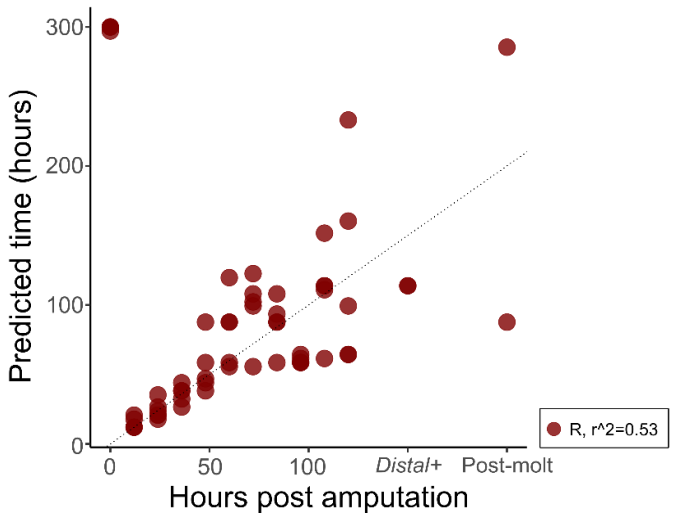
A


μRd_gis = μRp_gi x R_gis
σ Rp
μlog Rp
per gene g
per individual i
μ Rp_gi
σ R
log(R_gis)
R_gis
μ Rd_gis
Rp_gis
Rd_gis
per sample s

B


R, dev=0.69
PC1 = 11% variance
50
0
-50
-100
0
50
100
Distal+
Post-molt
Hours post amputation

C


300
200
100
0
Predicted time (hours)
R, r^2=0.53
0
50
100
Distal+
Post-molt
Hours post amputation

738

739

Figure 5: Comparing the transcriptional dynamics of leg embryonic development and regeneration. (**A**) Combined principal component analysis of development (E series) and regeneration (R series); samples color-coded according to RAPToR pseudotimes. Variation in PC1 and PC2 is largely driven by embryonic development. (**B**) RAPToR temporal predictions on the R samples using a reference based on the E series. Coherent predictions are only made on pre-amputation and late or post-regeneration samples. Other stages are poorly predicted, and different sets of genes make incoherent predictions (see Figure S5.2B). (**C**) Co-expression clusters defined by fuzzy c-means clustering of expression values in the developing (left) and regenerating (right) leg series. Four co-expressed gene clusters were identified in the E series (E2, E4, E1 and E3) and 8 clusters were identified in the R series (R4, R1, R8, R2, R6, R3, R5 and R7). Heatmaps represent the average profiles (centroids) of each cluster. Clusters are ordered according to their temporal profiles (except clusters R3, R5 and R7, which do not show clear temporal profiles); samples are ordered by pseudotime. Cluster sizes are given in Table S5. (**D**) Summary of the GO enrichment analysis for the E and R co-expression clusters; enriched GO terms were categorised as shown in Figure S5.5. (**E**) Number of genes shared between embryonic and regenerative co-expression clusters, expressed as a fold enrichment relative to equally sized random clusters. Clusters are ordered as in panels **C** and **D** (alternative ordering shown in Figure S5.4). Similar results were obtained using clusters defined on untransformed Ef data (Figure S5.6). (**F**) Chord diagram depicting the genes shared between regenerative (top) and embryonic (bottom) co-expression clusters (aligned temporally from left to right). Left: diagram highlighting the genes of the R8 cluster (purple), corresponding to the regenerative phase of cell proliferation and patterning. Right: matches between all the regenerative and embryonic clusters. A fraction of genes (> 5000) are not clustered in the embryonic dataset. (**G**) Overlap of co-expressed gene clusters applied on a finer gene clustering of the E and R datasets (as in panel E; see Methods). Alternative ordering of clusters presented in Figures S5.9, S5.10.

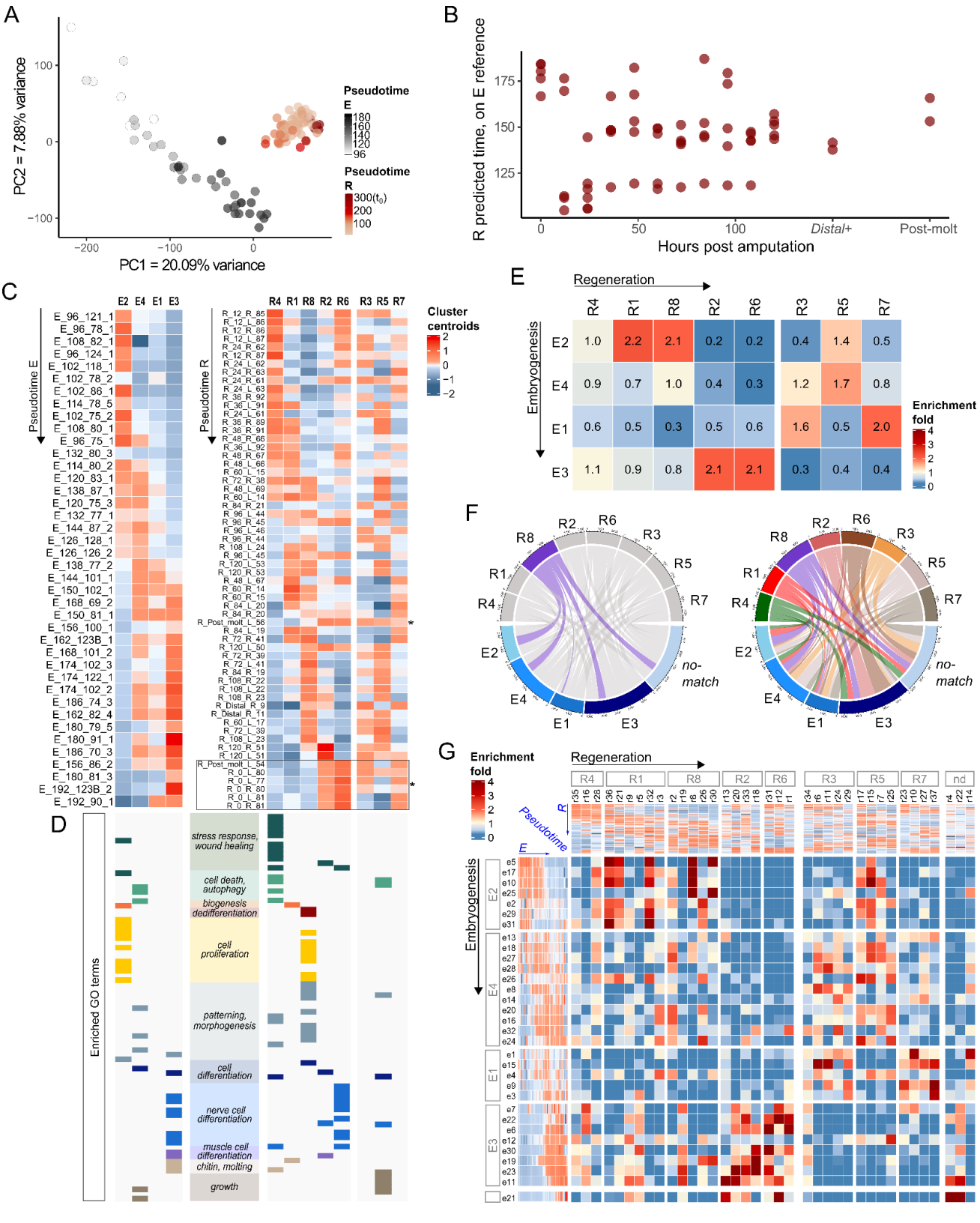

A
PC2 = 7.88% variance
PC1 = 20.09% variance
Pseudotime E
Pseudotime R
B
R predicted time, on E reference
Hours post amputation
Distal+
Post-molt
C
Pseudotime E
Pseudotime R
Cluster centroids
D
Enriched GO terms
stress response, wound healing
cell death, autophagy
biogenesis dedifferentiation
cell proliferation
patterning, morphogenesis
cell differentiation
nerve cell differentiation
muscle cell differentiation
chitin, molting
growth
E
Regeneration
Embryogenesis
Enrichment fold
F
no-match
G
Enrichment fold
Regeneration
Pseudotime
Embryogenesis

Figure 6. Temporal expression profiles of selected gene sets during leg development and regeneration. Expression in embryonic (E values, left) and regenerating (R values, right) legs, for genes associated with immune cells/responses (**A**), cell proliferation (**B**), patterning (**C**), differentiated nerves (**D**) and differentiated muscle (**E**). Samples ordered by pseudotime; t_0: pre-amputation; *pm*: post-molt.

