Estimer l'âge biologique à partir du transcriptome pour étudier l'impact intergénérationnel des phéromones sur les dynamiques d'expression génique chez l'embryon de C. elegans

Romain Bulteau

Université Claude Bernard Lyon 1

THESE de DOCTORAT DE
L'UNIVERSITE CLAUDE BERNARD LYON 1

Ecole doctorale N°340:
Biologie Moléculaire, Intégrative et Cellulaire (BMIC)

Discipline : Bioinformatique & Transcriptomique

Soutenue publiquement le 11/12/2023, par :

Romain BULTEAU

Inferring biological age from the transcriptome to uncover intergenerational effects of pheromone on embryonic gene expression dynamics in *C. elegans*

Devant le jury composé de :

DUPUY,	Denis	CR INSERM	IECB (Bordeaux)	Rapporteur
JARRIAULT,	Sophie	DR CNRS	IGBMC (Strasbourg)	Rapporteure
FÉLIX,	Marie-Anne	DR CRNS	IBENS (Paris)	Examinatrice
BESSEREAU,	Jean-Louis	Pr. Univ. UCBL	INMG (Lyon)	Président
FRANCESCONI,	Mirko	CR INSERM	LBMC (Lyon)	Directeur de thèse

PhD thesis summary

The environment influences not only the behavior and physiology of an organism, but can also impact its descendents. In the nematode model *C. elegans*, perception of social cues (pheromones) elicits such intergenerational effects, notably increasing generation time of the progeny. Here, I characterize the molecular changes in embryos caused by parental pheromone exposure by profiling gene expression in single individuals.

To achieve this, I first developed a robust computational method that infers age from the transcriptome in diverse organisms and sample types, makes it possible to detect and correct for developmental bias in gene expression data, and allows us to bypass synchronization and staging challenges for embryo collection. Then, I adapted experimental techniques used for sorting and profiling single cells to single embryos in order to improve throughput, revealing great potential for accessible and cost-efficient studies at large scale. Armed with these methods, I could then profile genome-wide gene expression across embryo development in the progeny of pheromone-exposed and control parents. I show that the developing nervous system and sensory organs are influenced by parental neuronal perception of the environment, likely changing how progeny will experience their own surroundings.

Résumé de la thèse

L'environnement n'influence pas seulement le comportement et la physiologie d'un organisme, mais peut également avoir un impact sur sa descendance. Dans le nématode modèle *C. elegans*, la perception de l'environnement social (phéromones) déclenche de tels effets intergénérationnels, augmentant notamment le temps de génération de la progéniture. Dans mes travaux, je caractérise les changements moléculaires dans les embryons causés par l'exposition des parents aux phéromones en profilant l'expression des gènes à l'échelle de l'individu.

Pour y parvenir, j'ai d'abord développé une méthode computationnelle robuste capable d'estimer l'âge à partir du transcriptome dans divers organismes et types d'échantillons, qui permet de détecter et de corriger les biais liés au développement dans les données d'expression génique, et nous permet de contourner les défis de synchronisation et de stadification pour la collecte des embryons. J'ai ensuite adapté des techniques expérimentales initialement utilisées pour trier et profiler les cellules uniques (*single-cell*) aux embryons individuels pour permettre un haut débit, révélant un important potentiel pour mener des études à grande échelle de manière accessible et à moindre coût. Armé de ces méthodes, j'ai ensuite pu profiler l'expression des gènes à l'échelle du génome tout au long du développement de l'embryon chez la progéniture de parents exposés aux phéromones et de témoins. Je montre que l'expression génique du système nerveux et des organes sensoriels est influencée au cours de leur développement par la perception neuronale de l'environnement des parents, ce qui change certainement la manière dont la progéniture percevra son propre environnement.

REMERCIEMENTS

A ceux qui de près comme de loin ont rendu possible ces travaux, y compris ceux que j'oublie ci-dessous, merci. Merci pour votre présence, votre écoute, votre soutien, et vos savoirs, qui m'ont permis d'arriver jusqu'ici.

To Mirko, my supervisor and mentor for these nearly five years as your master and PhD student. Thank you for your support on many levels, pushing me through the high and the lows, for the rich discussions, scientific or not, for your patience, for the excitement you showed in my work, and for the freedom you allowed it. I have learned a lot from you.

To Bianca Habermann and Marie Sémon, for following my thesis work as members of my CST/TMC, and for your valuable feedback and pointers. I went over my promised timeline, but I promise it was worth it!

To Sophie Jarriault and Denis Dupuy, for agreeing to read through this manuscript and evaluate my work. Thank you to Marie-Anne Félix and Jean-Louis Bessereau for also taking part in this evaluation as members of my defense jury.

A mes nombreux collègues et amis du LBMC, et en particulier,

A Merhnaz, Marie-Alice, et Noémie, membres successifs de notre équipe, et rares personnes pouvant rire aux vannes pointues autour de notre sujet d'étude (dont un certain nombre ornent toujours la porte de notre bureau, merci M-A). C'était un plaisir de travailler avec vous.

A Marie Delattre et son équipe, Caroline, Caroline, Eva, Carine, Brice, pour les nombreux échanges sur nos nématodes favoris (*C. elegans* est le plus mignon, c'est scientifiquement prouvé) autour de repas faisant eux-même souvent l'objet de nos discussions.

A l'équipe des levures, ceux de passage comme permanents, notamment Gaël, Alice, Gérard, Hélène, Arnaud, Fabien, pour ces lab meetings partagés. J'aurai beaucoup appris sur sur votre modèle, et vos retours sur les projets de notre équipe étaient précieux.

A Marie Sémon, pour m'avoir donné l'opportunité d'enseigner dans l'UE NGS, avec Carine, Corentin et Alice, moyennant un ratio enseignant-élève si confortable que j'ose à peine en parler à mes amis profs.

Au comité étudiant que Camille, Caroline, Léa et moi avons créé trois semaines avant le 1er confinement... et qui a malgré tout subsisté, et continue de faire vivre bon nombre d'activités au labo grâce à Timothée, Alice, Vinciane, et Markram.

A la direction et administration du laboratoire, notamment Isabelle, Corinne, Julie, et Didier, de rendre possible le bon fonctionnement de nos équipes, et les projets du comité étudiant.

A la Bioband, Benjamin, Sophie, et Gaël. On était pas vraiment assidus, mais on s'est quand même bien marré.

Merci aussi à Sébastien et Estelle de la plateforme de Cytométrie, d'avoir rendu possible toutes ces manips de FACS un peu bizarres avec les embryons de *C. elegans*.

Aux millions de nématodes morts pour la Science dans le contexte de mes travaux.

A tous mes amis hors du labo, qui ont suivi ma thèse autour d'une bière ou dans un salon Discord. Je pense en particulier à mes camarades de promo de l'IUT d'Aurillac, Alexandre et Clément et ceux de BIM à l'INSA Lyon, dont Amaury et Bastien.

Aux amis des chorales dont je fais partie, les Phonies Polies et les Chaussettes de Courtoisie, d'avoir rempli de musique, de bonne humeur, et de moments forts mon quotidien de ces dernières années.

A mon piano (et à celui chaché dans l'annexe de la salle des thèses).

Aux copains de Banyuls, qui me demandaient chaque été "Alors, c'est pour quand ton article ?" ou encore "C'est bientôt ta thèse, non ?". Oui, ça y est, elle est là !

A ma famille, et son précieux soutien. Merci pour qui vous êtes et pour ce que vous faites. Max et Isa, la pépette, et le petit pâté. Marie-Aure, à qui je conseille de feuilleter ce pavé avant de se lancer dans sa propre thèse. Papa, pour l'éducation que toi et Maman m'avez donné, cette curiosité, ce goût de la science, et bien plus encore, dont j'ai certainement hérité. Ushuaïa, pour les poils que je retrouve encore régulièrement dans mes vêtements.

A Cécile, mon coeur, pour son amour et son soutien pendant les moments les plus difficiles. Cette thèse, tu la mérite aussi !

A Maman, qui ne lira pas ces lignes mais qui, je le sais, aurait été fière de moi. Nul n'aurait douté de ta présence à ma soutenance, même après ton diagnostique. Tu t'es battue admirablement contre ton cancer. Sans broncher. A l'écoute, et attentionnée jusqu'au bout. Merci pour tout.

Résumé des travaux en français

Introduction

L'ensemble des caractères d'un individu, le phénotype, est parfois présenté comme la combinaison du génome et de l'environnement. Pourtant, même à environnement contrôlé et égal, des populations isogéniques peuvent présenter de fortes différences dans des traits importants. Par exemple, chez le nématode modèle *C. elegans*, la durée de vie d'individus génétiquement identiques dans le même environnement peut varier du simple au triple (KIRKWOOD et al., 2005 ; BANSAL et al., 2015).

De nombreuses études ont révélé que le vécu des générations précédentes est une source majeure de variation. Ces effets intergénérationnels peuvent impacter de nombreux phénotypes de la descendance comme son métabolisme (GRANDJEAN et al., 2015), l'hérédité de pathologies (CHEN et al., 2021), sa vitesse de développement (FRIDMANN-SIRKIS et al., 2014), voire sa durée de vie (RECHAVI et al., 2014), et leur existence a été démontrée dans de nombreux modèles, dont des mammifères (JABLONKA et al., 1992), et chez l'Homme (PEMBREY et al., 2006 ; PEMBREY et al., 2014).

Récemment, deux études ont démontré que la perception sensorielle des phéromones chez *C. elegans* impacte leur progéniture (PEREZ et al., 2021 ; WASSON et al., 2021), ce qui implique une transmission d'information du système nerveux vers la lignée germinale capable de changer des caractères cruciaux chez la descendance, tels que le temps de génération. L'étendue des effets de la perception parentale de phéromones sur la physiologie de la descendance n'est cependant pas connue. De plus, bien que plusieurs mécanismes de transmission inter- et transgénérationnels aient été découverts (PEREZ & LEHNER, 2019 ; FITZ-JAMES & CAVALLI, 2022), peu d'entre eux sont pleinement compris, et la manière dont l'information perçue par les neurones est transmise à la descendance reste inconnue.

L'objectif de ma thèse est donc de caractériser les changements moléculaires dans les embryons causés par l'exposition des parents aux phéromones afin de mieux comprendre ce phénomène. Pour cela, j'ai d'abord développé et adapté les outils expérimentaux et computationnels nécessaires pour profiler l'expression génique d'individus uniques.

Chapitre 1 : Prédire l'âge à partir du transcriptome

Mon travail a commencé par le développement de RAPToR (***R**eal **A**ge **P**rediction by **T**ranscriptome staging **o**n **R**eference*), un outil pouvant estimer l'âge à partir du transcriptome qui a fait l'objet d'une publication incluse dans le premier chapitre (BULTEAU & FRANCESCONI, 2022). RAPToR est robuste, capable de stadifier précisément le développement et le vieillissement à partir du profil d'expression génique de types d'échantillons biologiques variés (bulk, individus, tissus, cellules uniques), fonctionne avec les organismes modèles les plus utilisés (nématode, drosophile, zebrafish, souris) et chez l'Homme, peut estimer un âge tissu-spécifique à partir de données d'individu entier, et peut même stadifier des échantillons d'une espèce sur une référence d'une autre. L'estimation d'âge ainsi obtenue peut être utilisée pour détecter l'effet de perturbations d'intérêt ou variables expérimentales sur la vitesse de développement, et être utilisée comme covariable dans un modèle pour améliorer la puissance statistique de détection de gènes différentiellement

exprimés. De plus, il est possible de quantifier et de corriger l'effet d'une différence de développement entre deux groupes d'échantillons sur une analyse d'expression génique différentielle en intégrant des données de référence, y compris lorsqu'il n'y a aucun chevauchement entre les âges des deux groupes.

Mon premier chapitre rapporte également plusieurs améliorations et pistes de recherches explorées depuis la publication de l'outil. Je caractérise plus en profondeur l'impact de différences de développement sur les analyses d'expression différentielles et simplifie la méthode de correction initialement proposée. Ensuite, je propose une manière plus robuste de construire des références pour stadifier le vieillissement, démontre que RAPToR peut en principe stadifier entre plusieurs trajectoires de développement, et étends le champ d'applications possibles de l'outil concernant les échantillons provenant de tissus dissociés.

Enfin, stadifier des individus post-profilage nous permet de contourner les défis de synchronisation et de stadification qui entravent la collecte d'embryons en grand nombre.

Chapitre 2 : Vers le séquençage d'ARN d'individus uniques à grande échelle

L'engouement autour du *single-cell* (cellules uniques) a poussé les technologies de profilage à un point tel que les faibles apports d'ARN et la collecte d'échantillons à large échelle ne sont plus une barrière au *RNA-seq* (séquençage d'ARN à haut débit). Dans mon deuxième chapitre, je montre l'intérêt de ces techniques pour profiler l'expression génique d'individus (uniques) entiers, permettant ainsi l'étude des variations inter-individuelles à grande échelle et à faible coût avec un moindre ajustement des protocoles existants.

En effet, je montre premièrement que des embryons de *C. elegans* peuvent être triés avec les technologies de cytométrie de flux standard (*FACS*), permettant un haut débit d'échantillonnage. Ensuite, je démontre qu'avec ces mêmes instruments, il est également possible de sélectionner des embryons à des stades spécifiques à partir d'une population mixte en utilisant uniquement des paramètres physiques et l'autofluorescence mesurés, sans avoir besoin de marqueurs fluorescents. Enfin, j'ai adapté le protocole de RNA-seq Smart-seq3 (Hagemann-Jensen et al., 2020) pour établir le profil d'expression génique d'embryons uniques à haute complexité et faible coût, permettant dans l'ensemble une mise à l'échelle rentable du profilage d'un individu unique.

Chapitre 3 : La perception des phéromones modifie l'expression génique du système nerveux en développement dans la descendance

Grâce aux méthodes développées dans les chapitres précédents, j'ai pu profiler des embryons uniques de *C. elegans* et étudier l'impact des signaux sociaux perçus par les neurones sensoriels de la génération précédente sur l'expression des gènes au cours de l'embryogénèse. Ces données révèlent qu'en plus de retarder la lignée germinale de la progéniture, l'exposition aux phéromones parentales modifie probablement le développement de leur système nerveux, et en particulier de leurs organes sensoriels. Nous émettons l'hypothèse que ces changements pourraient altérer la perception et la réaction à l'environnement de la progéniture, influençant éventuellement les décisions importantes pour la survie du nématode, telles que l'entrée dans la voie de développement alternative *dauer*.

Conclusion

Dans le but de caractériser les changements moléculaires dans l'embryon en développement de *C. elegans* causés par l'exposition à la phéromone parentale, ma thèse s'intéresse à plusieurs méthodes, de la collecte d'échantillons et préparation de banques pour le séquençage, à l'analyse et l'intégration des données.

La collecte et l'étude d'embryons uniques par *FACS*, le profilage d'individus entiers avec Smart-Seq3 et l'inférence de l'âge à partir du transcriptome ont un potentiel allant au-delà de l'étude des effets intergénérationnels de la phéromone et de *C. elegans*, et fournissent des solutions accessibles pour étudier des individus uniques à haut débit. La grande partie de mes travaux dédiée au développement de méthodes expérimentales et computationnelles est donc certainement pertinente pour les recherches futures dans ce domaine.

Références

BANSAL, ANKITA et al. (2015). "Uncoupling lifespan and healthspan in Caenorhabditis elegans longevity mutants". In : *Proceedings of the National Academy of Sciences* 112.3, E277-E286.

BULTEAU, ROMAIN et MIRKO FRANCESCONI (août 2022). "Real age prediction from the transcriptome with RAPToR". en. In : *Nature Methods* 19.8, p. 969-975. ISSN : 1548-7105. DOI : 10.1038/s41592-022-01540-0.

CHEN, YAN-TING et al. (2021). "Imprinted lncRNA Dio3os preprograms intergenerational brown fat development and obesity resistance". In : *Nature communications* 12.1, p. 6845.

FITZ-JAMES, MAXIMILIAN H et GIACOMO CAVALLI (2022). "Molecular mechanisms of transgenerational epigenetic inheritance". In : *Nature Reviews Genetics* 23.6, p. 325-341.

FRIDMANN-SIRKIS, YAEL et al. (2014). "Delayed development induced by toxicity to the host can be inherited by a bacterial-dependent, transgenerational effect". In : *Frontiers in genetics* 5, p. 27.

GRANDJEAN, VALÉRIE et al. (2015). "RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders". In : *Scientific reports* 5.1, p. 18193.

HAGEMANN-JENSEN, MICHAEL et al. (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In : *Nature Biotechnology* 38.6, p. 708-714.

JABLONKA, EVA, MICHAEL LACHMANN et MARION J LAMB (1992). "Evidence, mechanisms and models for the inheritance of acquired characters". In : *Journal of theoretical biology* 158.2, p. 245-268.

KIRKWOOD, THOMAS BL et al. (2005). "What accounts for the wide variation in life span of genetically identical organisms reared in a constant environment?" In : *Mechanisms of ageing and development* 126.3, p. 439-443.

PEMBREY, MARCUS, RICHARD SAFFERY, LARS OLOV BYGREN et al. (2014). "Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research". In : *Journal of medical genetics* 51.9, p. 563-572.

PEMBREY, MARCUS E et al. (2006). "Sex-specific, male-line transgenerational responses in humans". In : *European journal of human genetics* 14.2, p. 159-166.

PEREZ, MARCOS FRANCISCO et BEN LEHNER (2019). "Intergenerational and transgenerational epigenetic inheritance in animals". In : *Nature cell biology* 21.2, p. 143-151.

PEREZ, MARCOS FRANCISCO et al. (2021). "Neuronal perception of the social environment generates an inherited memory that controls the development and generation time of C. elegans". In : *Current Biology* 31.19, p. 4256-4268.

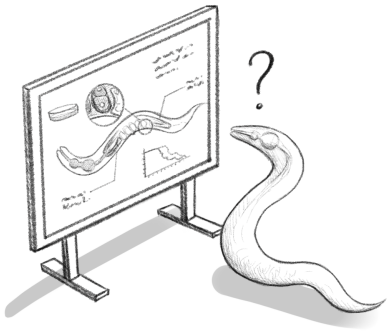
RECHAVI, ODED et al. (2014). "Starvation-induced transgenerational inheritance of small RNAs in C. elegans". In : *Cell* 158.2, p. 277-287.

WASSON, JADIEL A et al. (2021). "Neuronal control of maternal provisioning in response to social cues". In : *Science Advances* 7.34, eabf8782.

Table of contents

List of figures

List of tables



Introduction

Contents

In.1 Same genetics, same environment, differences persist

> " *Under the most carefully controlled conditions, biological material does whatever it damn well pleases.* "
>
> – Harvard Law of Biology

The apparent frustration expressed in this 'Harvard Law of Biology' (INSALL, 2001; HALLGRIMSSON & HALL, 2011) reflects the ubiquitous nature of variation in biology. Indeed, despite the best efforts of biologists, it persists even between individuals with identical genetics and environment, and is often ignored (and cursed) as a hindrance to experiments. Enst Mayr says "variation is an endless source of challenging questions" (HALLGRIMSSON & HALL, 2011), and we, as biologists, seek to understand its origins.

In.1.1 Stochasticity in isogenic populations

Some aspects of biology are inherently noisy. Gene expression, for example, is a stochastic process that is particularly variable when there are low copy numbers of a gene and of its regulatory elements (HARDO & BAKSHI, 2021), impacting crucial functions such as DNA repair in cells (UPHOFF et al., 2016). Dividing cells also impart variable fractions of their cytoplasmic content (RNAs, proteins, organelles) between daughters which causes heterogeneity between genetically identical single-celled organisms, or between cells of a tissue (HUH & PAULSSON, 2011).

In multicellular organisms, although cell-to-cell variability is to an extent buffered (SMITH & GRIMA, 2018), stochasticity can still drive differences between individuals. For example, random variations of histone acetylation at specific genomic loci explain behavior differences between isogenic zebrafish (ROMÁN et al., 2018).

Despite controlled environments, all kinds of traits (behavior, physiology, morphology, and life history) can substantially vary between genetically identical individuals. Within an isogenic population of *Caenorhabditis elegans* nematodes in a lab culture, the longest-lived individual will live 3 times longer than short-lived worms (KIRKWOOD et al., 2005; BANSAL et al., 2015), and the time worms spend in good health versus the declining "twilight" period during aging is also highly variable (ZHANG et al., 2016). Furthermore, faced with the same stress, genetically identical *C. elegans* worms respond to a different extent, and this variation comes with important consequences, as higher-stress resistance was shown to be a trade-off with reproductive fitness (CASANUEVA et al., 2012).

Stochasticity in isogenic populations has been proposed as a means to tackle fluctuating environments by diversifying individual responses (THATTAI & VAN OUDENAARDEN, 2004; ACAR et al., 2008; CASANUEVA et al., 2012). However, the origin of interindividual variation is rarely attributed to purely random effects (ROMÁN et al., 2018) and growing evidence implicates previous generations and their experiences as major sources of variation.

In.1.2 Influence of the previous generation and its environment

Daphnia water fleas are famous for developing defensive 'helmet' structures in the presence of predators (Fig. In.1a, AGRAWAL et al., 1999). This plasticity is not restrained to the perturbed generation, however, and the progeny of *Daphnia* exposed to predators are also born with larger defensive structures, and develop faster than controls (AGRAWAL et al., 1999; WALSH et al., 2015). Therefore, two clones living in identical environments can have very different morphologies and reach reproductive maturity at different times because of the environment experienced by their parents.

Evidence for such intergenerational effects – or transgenerational effects, when they persist for multiple generations – has been found with diverse triggers (e.g. environmental cues, stresses, age), and in many organisms (JABLONKA et al., 1992), including humans (PEMBREY et al., 2006;

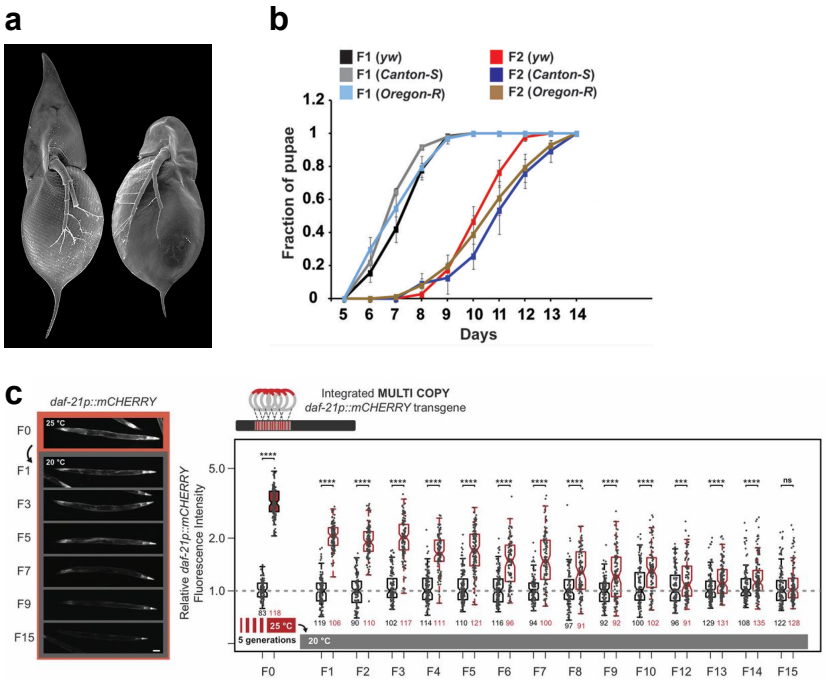


Figure In.1 – Examples of plasticity and transgenerational effects in several species
a, *Daphnia cucullata* water flea clones have distinct morphologies due to predator presence inducing the development of defensive 'helmet' structures (left). Electron micrograph from Fig. 2 of AGRAWAL et al., 1999.
b, Egg sterilization (bacterial depletion in P0) delays the development of F2 progeny into pupae in several strains of *Drosophila melanogaster* flies. Curves show the fraction of progeny that reached pupation in days post-hatching. Panel adapted from Fig. 1d of FRIDMANN-SIRKIS et al., 2014.
c, Heat stress alters gene expression in adult *Caenorhabditis elegans* nematodes for close to 15 generations, as evidenced by fluorescence of an integrated multi-copy mCherry transgene reporter. Scale bar (left images) is 0.1mm. Significance of the difference in fluorescence intensity (right) tested with a Wilcoxon test. False-discovery rates (FDR) q-values: ****: $q < 0.0001$; ***: $q < 0.001$; ns: $q > 0.05$. Panel from Fig. 1a of KLOSIN et al., 2017.

PEMBREY et al., 2014). Several inheritance mechanisms have been found for these effects, with some still poorly understood. I give a few examples below to illustrate this diversity, but see PEREZ & LEHNER, 2019; FITZ-JAMES & CAVALLI, 2022; BURTON & GREER, 2022 for excellent reviews on the subject.

In mice, parental diet has been shown to influence the physiology and metabolism of progeny. Specifically, paternal obesity and metabolic disorders like type II diabetes caused by high sugar and fat diets are inherited through small RNAs in sperm (GRANDJEAN et al., 2015), while preconception fasting of fathers decreases glucose levels in progeny by an unknown mechanism (ANDERSON et al., 2006). Furthermore, mice can also inherit obesity maternally through DNA methylation altering the expression of a long non-coding RNA in progeny (CHEN et al., 2021), thus showing that similar intergenerational effects can be mediated by distinct mechanisms within a species.

Antibiotic treatments in *Drosophila* flies can alter the composition of commensal gut bacteria, which leads to a developmental delay in the progeny that is transgenerationally inherited (Fig. In.1b, FRIDMANN-SIRKIS et al., 2014). Furthermore, body mass and reproductive success of progeny can be similarly altered transgenerationally through gut microbial transfer (MORIMOTO et al., 2017), altogether implicating the microbiome as a vector for cross-generation effects on important fitness traits.

Several studies document transgenerational inheritance of small RNAs in *C. elegans*, with effects such as learned avoidance of pathogenic bacteria (MOORE et al., 2019; KALETSKY et al., 2020), altered sexual attractiveness (TOKER et al., 2022), or lifespan extension (RECHAVI et al., 2014). Gene silencing induced by RNA interference (RNAi), can also be inherited for three or more generations through a similar mechanism (ALCAZAR et al., 2008; VASTENHOUW et al., 2006). However, other transgenerational inheritance mechanisms also operate in *C. elegans*. For instance, heat stress can alter gene expression transgenerationally for over 10 generations through altered histone methylation that is transmitted through both sperm and oocytes (Fig. In.1c, KLOSIN et al., 2017). Heavy metal exposure also confers increased resistance to progeny to similar stresses for several generations, likely through histone modifications, in an example of potentially adaptive transgenerational inheritance (KISHIMOTO et al., 2017).

Lastly, intergenerational effects are also plenty in the worm, implicating for example parental diet or age. TAUFFENBERGER & PARKER, 2014 show that a high-glucose diet impairs stress resistance and fecundity in progeny (likely through histone methylation, although the mechanism is unclear), while PEREZ et al., 2017 demonstrate that loading of vitellogenin (egg yolk) to embryos increases with maternal age, resulting in progeny born from older mothers being more resistant to starvation and developing faster.

Thanks to its extensive phenotypic plasticity in controlled isogenic populations and increasingly well-documented inheritance mechanisms, *C. elegans* has emerged as the star model amongst lab animals to study inter- and transgenerational effects (BAUGH & DAY, 2020). The present work too relies on the nematode, further described in the following section.

In.2 "The worm"

At first glance through the microscope, one could doubt what a small free-living nematode like *Caenorhabditis elegans* could bring to this discussion. Yet, since Sydney Brenner introduced it as a model for genetic studies some 50 years ago, the worm has rapidly spread to labs around the world, became the first multicellular organism to have its genome sequenced (SEQUENCING CONSORTIUM, 1998), and has been the vector for landmark discoveries in diverse fields including neurology, aging, or uncovering the role of small RNAs, notably RNAi.

In recent years, *C. elegans* has been instrumental to the study of inter- and trans-generational effects thanks to its short generation time, ease of genetic analysis and handling, and deep pool of accumulated knowledge. Here, I describe the general life cycle(s) and anatomy of the worm with a slight focus on the sensory organs.

In.2.1 Life cycles

In nature, *C. elegans* are typically found feeding on bacteria growing in rotting fruit and vegetal matter of compost heaps (FRÉZAL & FÉLIX, 2015), from which the original Bristol "N2" wild-type[1] strain used by labs around the world was isolated by BRENNER, 1974. They are self-fertilizing (selfing) hermaphrodites making both sperm and eggs required for their offspring, meaning populations tend to become homozygous and are essentially isogenic. A small fraction of males (<0.2%) also spontaneously occurs by losing one of the two X chromosomes through chromosomal nondisjunction, making it possible to cross strains by mating with hermaphrodites.

In around 65 hours at 20°C, a fertilized egg will develop, hatch and grow through four successive larval stages (L1-L4) to become a 1 mm long egg-laying adult (Fig. In.2) that is capable of birthing around 300 genetically identical offspring on its own over several days.

If a larva of the first stage (L1) encounters poor conditions (high temperatures, low food, overcrowding), it can enter an alternate developmental path leading to a diapause stage capable of enduring extremely harsh conditions. The *dauer* larva (German, "enduring" larva) is highly resistant to heat, desiccation, and starvation, capable of roaming for months without food (GOLDEN & RIDDLE, 1984). When a favorable environment is found, the worm resumes normal development by molting into the final (L4) larval stage (Fig. In.2).

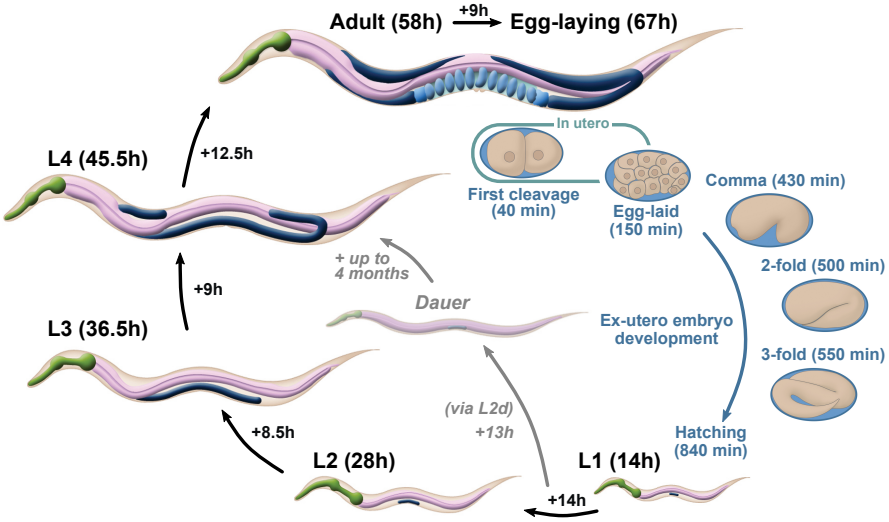


Figure In.2 – Overview of the *C. elegans* reproductive cycle
Adapted from (ALTUN et al., 2006), timings in hours/minutes post-fertilization at 20°C.

[1] though "wild-type" is debatable given the extent of lab domestication in *C. elegans*, see STERKEN et al., 2015; WEBER et al., 2010.

In.2.2 General anatomy

The anatomy of *C. elegans* is characterized by remarkable simplicity and determinism, with just 959 somatic cells in adult hermaphrodites, that arise from stereotyped cell divisions during development (SULSTON et al., 1983). Despite this limited cell count, the nematode has defined tissues and organs including epithelial, muscular, and digestive systems, as well as a complex nervous circuit, arranged in bilateral symmetry within its long tapered cylindrical body (Fig. In.3a).

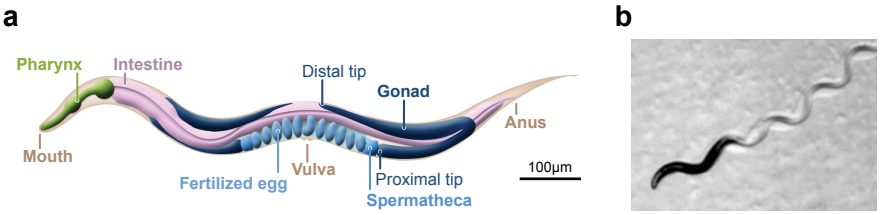


Figure In.3 – *C. elegans* anatomy and movement
a, General anatomy of *C. elegans*, adapted from (ALTUN et al., 2006).
b, Sinusoidal movement of the worm. Image from (LIKITLERSUANG et al., 2012)

In the lab, our nematode generally feeds on *Escerichia coli* (OP50), pumping and grinding the bacteria with its pharynx. Material then flows through its multi-functional intestine, that not only digests and absorbs nutrients but also synthesizes and stores lipids, before waste is discharged at the posterior end through the anus.

The skin (hypodermis) of *C. elegans* secretes a strong and flexible armor, the cuticle, mainly composed of collagens and a lipid-rich external membrane (CHISHOLM & XU, 2012), that is shed between each larval stage and before adulthood to allow growth. Beneath the hypodermis, the worm's body is lined with muscles enabling it to move through solid or liquid environments by sinusoidal undulations (Fig. In.3b).

The hermaphrodite reproductive system consists of two symmetric gonadal arms arranged in a U-shape that connect to a central uterus at their proximal end (Fig. In.3a). Meiosis occurs along the distal to proximal axis of the gonad, first producing a limited amount of sperm during the fourth larval stage (approx. 150 per gonad), then switching permanently to oogenesis after the final adult molt. Sperm is stored in the spermatheca at the proximal end of the gonad through which the oocytes pass and are fertilized before reaching the uterus. Fertilized eggs then remain in the uterus for a couple hours before being laid outside through the vulva, and finish embryogenesis *ex-utero*. The brood size of selfing hermaphrodites is limited by the initial amount of sperm produced, but mating with a male can increase the number of progeny up to 1400 (HIRSH et al., 1976).

In.2.3 Nervous system

In.2.3.1 Overview

The nervous system of an adult *C. elegans* hermaphrodite accounts for nearly a third of the somatic cell count of the animal, with 302 neurons. Most neuron cell bodies are grouped within the head ganglia (anterior, lateral, ventral, dorsal, and retrovesicular), with many nerve processes bundled in a loop around the pharynx, termed the "nerve ring", that has the highest synapse density within the body (Fig. In.4). Other notable groupings of neuron cell bodies and processes include the ventral nerve cord, and the tail ganglia (preanal, dorsorectal, and lumbar, Fig. In.4). The morphology of *C. elegans* neurons is nearly exclusively mono- or bi-polar and non-branching, with highly conserved shapes, positions, and to an extent connections, across individuals (WHITE et al., 1986; WITVLIET et al., 2021; BRITTIN et al., 2021; COOK et al., 2023).

As for most neuron-endowed creatures, connections are ensured by synapses that are electrical (gap junctions) or chemical, with the latter accounting for around 90% of the nervous system

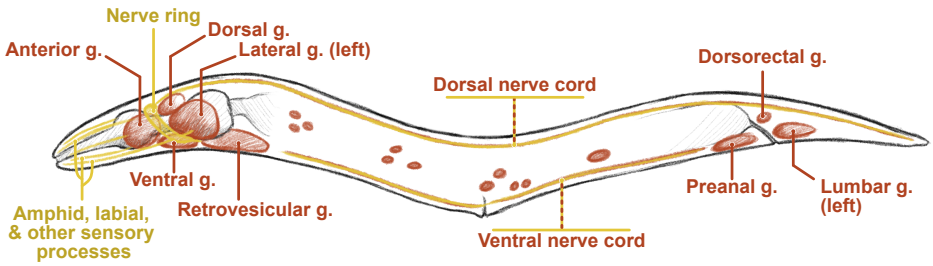


Figure In.4 – Main elements of the *C. elegans* nervous system
Cartoon depicting the main neuron ganglia (g.) and nerve processes of *C. elegans*. Lateral and lumbar ganglia have equivalents bilaterally symmetrical equivalents on the right side. **Ganglia** (and a few individual neurons) are depicted **in red**, while **nerve processes and bundles** are **in yellow**.

connectivity (comparatively over 99% in mammals GREENGARD, 2001) and being the best understood due to comparatively easier annotation.

Neuron classes are further grouped into 4 functional categories: sensory neurons, motor neurons, interneurons, or polymodal neurons (when performing more than one of the previous functions). Motor neurons are characterized by a synaptic output to muscle cells, while interneurons have both input and output synaptic connections to other neurons.

In.2.3.2 Sensory organs

C. elegans mainly senses its surroundings through 24 sensory neuron classes grouped in 7 sensilla : amphids (Am), phasmids (PH), inner and outer labials (IL, OL), anterior and posterior deirids (ADE, PDE), and cephalic sensilla (CEP). Each sensillum is composed of its respective sensory neurons, a sheath cell (sh) and one or more socket cells (so) that wrap around the sensory neuron processes and form a ring-like structure at their distal tip respectively. Aside from phasmids and posterior deirids, all sensilla are in the head of the worm.

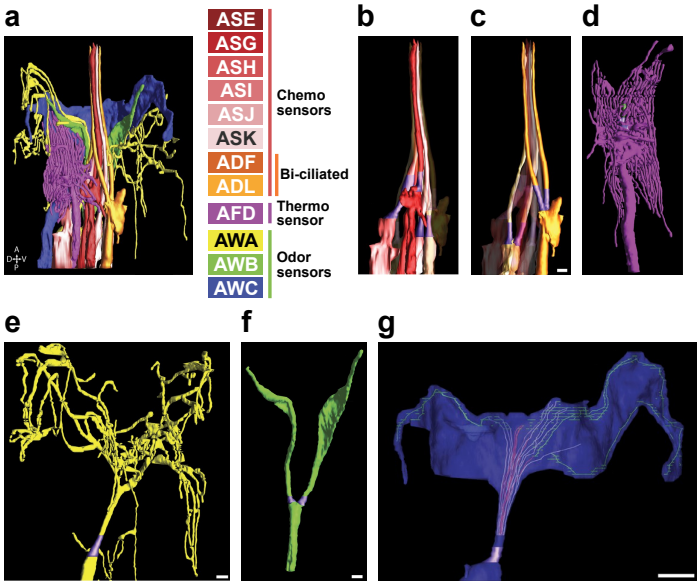


Figure In.5 – Structure of amphid sensory neurons
a, 3D reconstruction of amphid sensory neuron classes (neurons from the right amphid are show here).
b-g, Structure of individual neurons. Chemosensors are mono-ciliated (**b**) and bi-ciliated neurons (**c**), while thermosensor AFD (**d**) and volatile odor sensors AWA (**e**), AWB (**f**), and AWC (**g**) have more complex structures.
Scale bar in all images is 100nm, adapted from DOROQUEZ et al., 2014.

Amphids are the principal chemosensory organs of the worm, with 12 sensory neurons (Fig. In.5).

Of these, ASE, ASG, ASH, ASI, ASJ, and ASK have simple mono-ciliated and ADF, ADL, bi-ciliated endings (Fig. In.5b,c). These neurons mainly detect soluble ligands (BARGMANN, 2006), including ascaroside pheromones secreted by the worms ("dauer pheromones" and others, LUDEWIG & SCHROEDER, 2018, see below). ASI in particular is required for inhibition of dauer entry (BARGMANN & HORVITZ, 1991) and is the only source of DAF-7/TGF-β in normal conditions (REN et al., 1996). The "wing" cells AWA, AWB, and AWC have more complex branching cilia (Fig. In.5e-g), and primarily capture volatile ligands. AFD is thermosensory (GOODMAN & SENGUPTA, 2019), and also required for dauer-entry inhibition (BARGMANN & HORVITZ, 1991).

Phasmids are akin to smaller amphids, with just two chemosensory neuron classes (PHA, PHB) and a mechanosensor neuron (PQR) (ALTUN & HALL, 2003).

Labial (both inner and outer), deirid (both anterior and posterior) and cephalic neurons are mainly mechanosensory and involved in head withdrawal responses to touch, but some also likely function as interneurons, and IL2 is suggested to be chemosensory (PERKINS et al., 1986; ALTUN & HALL, 2003).

Beyond the sensilla, several neurons also perform sensory functions, notably perception of oxygen and touch.

To summarize, *C. elegans* has an exceptionally well-characterized nervous system in which behaviors, sensory and mechanical functions can be attributed to specific neurons, a knowledge made possible by the worm's suitability for the lab, including largely stereotyped cell positions and shapes, and powerful genetic tools.

In.3 Sensory perception of the environment changes progeny phenotypes

Several examples cited above show that environmental conditions experienced by parents can impact the progeny of *C. elegans* across several generations (SCHOTT et al., 2014; RECHAVI et al., 2014; KLOSIN et al., 2017; KALETSKY et al., 2020). However, this is often caused by a stress which directly affects the germline, such as a heat stress (KLOSIN et al., 2017), or by communication from the adjacent intestine through the canonical RNA interference pathway, for example after ingesting bacterial RNA (KALETSKY et al., 2020).

Two recent findings, however, show more provoking phenomena with sensory perception of social cues (pheromones) as the trigger, implicating a neuron to germline transfer of information that alters progeny phenotypes (PEREZ et al., 2021a; WASSON et al., 2021).

In.3.1 Neuron-germline communication

Communication from sensory inputs (neurons) to the germline is not a novel phenomenon when considered within the same generation. In *C. elegans*, it has been demonstrated that TGF-β produced in ASI neurons when sensing favorable environments (low population density, and abundant food) regulates germline proliferation through a pathway distinct from dauer entry (DALFÓ et al., 2012).

Olfactory imprinting, whereby worms behave differently upon sensing a previously-encountered chemical, has been shown to last several generations in proportion to exposure over multiple generations (REMY, 2010). The perception of these chemicals requires the AWC chemosensory neurons, and imprinting requires the downstream AIY interneurons (REMY & HOBERT, 2005). This suggests that an acquired sensory-dependent behavior can be inherited and implies neuron-germline transfer of information. The mechanism through which imprinting is inherited, and whether these neurons are actually required for the transgenerational effect is however not known.

DEVANAPALLY et al., 2015 have shown that exogenous double-stranded RNA made in the neurons can be transferred to the germline, and cause transgenerational gene silencing by RNAi. Furthermore, endogenous small-interfering RNAs (endo-siRNAs, a form of RNAi) have been implicated in neuron regulatory function within the same generation, in the context of behavioral adaptations to olfactory signals (JUANG et al., 2013). However, although transfer of endogenous small RNAs from neurons to the germline has been suggested (POSNER et al., 2019), to date, there is no direct evidence of this.

Of note, although this work focuses on *C. elegans*, neuron-germline transmission of information also occurs in mammals and can generate transgenerational effects. In mice, an odor-conditioned fear response can be paternally transmitted, modifying the behavior of grandsons and reducing DNA methylation at a gene encoding an olfactory receptor specific to the odor (DIAS & RESSLER, 2014). How the nervous system could induce such a change in sperm is not known.

To summarize, neuron to germline communication has been shown in *C. elegans*, with several sources suggesting neuronal information can also be transmitted to progeny. Mechanisms for information transfer to the germline and inheritance possibly implicate small RNAs, but are currently unclear.

In.3.2 Parental exposure to pheromone changes progeny phenotypes

In.3.2.1 Pheromone effect within the same generation

C. elegans releases hundreds of small chemicals into their environment, the best-studied and most abundant being ascarosides, first discovered by GOLDEN & RIDDLE, 1982 as the "dauer-inducing pheromone" conveying population density information. Ascarosides are widely-conserved signaling chemicals in nematodes (CHOE et al., 2012) that elicit diverse behavioral and

physiological response in *C. elegans* upon perception. Depending on the blend, pheromone can affect attraction, aggregation, and repulsion behaviors (SRINIVASAN et al., 2008; SRINIVASAN et al., 2012; MACOSKO et al., 2009; CHUTE et al., 2019), as well as developmental speed of larva (LUDEWIG et al., 2019), metabolism (HUSSEY et al., 2017), germline function (DALFÓ et al., 2012; MCKNIGHT et al., 2014), and even lifespan (LUDEWIG et al., 2013; MAURES et al., 2014).

Although very much unexplored, there is also evidence for the existence of non-ascaroside secreted molecules with physiological significance (ARTYUKHIN et al., 2018). For example, mutants defective in ascaroside production still communicate larval density (ARTYUKHIN et al., 2013).

Despite its numerous effects on worm behavior and physiology in the same generation, pheromone perception was only recently implicated in intergenerational effects (PEREZ et al., 2021a; WASSON et al., 2021).

In.3.2.2 Non-dauer pheromone perception alters maternal provisioning of translational machinery

WASSON et al., 2021 demonstrate that perception of pheromone by hermaphrodites promotes provisioning of translational machinery to the embryos and reduces stress resistance. They show that ascaroside pheromones distinct from the dauer-inducing cocktail cause the phenotype, and that ASI neurons and the FMRFamide-like peptide *flp-21* are required in parents for inheritance. Whether ASI is implicated in perception is unknown, but several genes implicated in its neuropeptide signaling pathways are required, namely *egl-3*, and *kpl-1*, *unc-13*, and *unc-31*.

Maternal mRNA contribution is proposed as the main inheritance mechanism for these effects, as transcripts found in early embryos (prior to the onset of zygotic transcription) substantially differed between *flp-21* mutants and wild-type. However, how *flp-21* mediates these changes is not known, and only chromatin modifications were (partially) ruled out as a means of intergenerational communication (WASSON et al., 2021).

WASSON et al., 2021 suggest this intergenerational effect could be adaptive, as although high concentrations of dauer pheromone generally signal poor environmental conditions, the total absence of different secretions could also indicate that the environment is severely detrimental to the worms (e.g. toxin, pathogen, zero-food). Therefore, preparing progeny by cutting down on energy-intensive processes (translation) and increasing stress responses could be beneficial.

In.3.2.3 Dauer pheromone exposure controls the generation time in the progeny

Pheromone perception can also regulate a crucial fitness trait in the progeny of *C. elegans*: generation time. PEREZ et al., 2021a show that the time between the final (L4 to adult) molt and the appearance of the first embryo (Δ soma-germline), can increase by up to 2 hours in worms whose parents sensed pheromone. This germline delay in the progeny is already visible at primordial germ cell division in early L1 stage, and can be induced in a dose-responsive manner with crude pheromone extract (filtered liquid culture medium) and a synthetic blend of two major dauer-inducing ascarosides (ascr#2, and ascr#3, see BUTCHER et al., 2007). Parents lacking chemosensory neuron ion channels TAX-2 and TAX-4 implicated in their downstream signal transduction do not induce the progeny phenotype, as well as genetic ablation strains for ASI, ASJ, ASK, or AWC, showing that sensory perception and/or processing by these neurons is required.

DAF-7/TGF-β is required in parents, further implicating ASI neurons. The principal downstream effector, DAF-3/co-SMAD, which antagonizes TGF-β, as well as another downstream nuclear hormone receptor DAF-12/NHR are however dispensable in parents but required in the progeny, thus implicating the TGF-β signaling pathway for both signal transmission in the parents and interpretation in the progeny. The fact that neither DAF-3/co-SMAD, DAF-12/NHR, or DAF-9 (a cytochrome P450 enzyme required to catalyze dafachronic acid, a DAF-12/NHR ligand) are required in the parents suggests DAF-7/TGF-β acts through a non-canonical axis to control progeny germline delay.

Of note, although insulin signaling has previously been implicated in intergenerational signaling (HIBSHMAN et al., 2016; JORDAN et al., 2019), PEREZ et al., 2021a exclude its role in the germline delay phenotype as null mutants of *daf-16*, transcription factor and downstream effector of the insulin pathway, are still responsive. Signaling from neuron to germline also appears distinct from the *flp-21*-dependent axis discussed above (WASSON et al., 2021), does not require inheritance machinery of RNAi which relies on HRDE-1 (BUCKLEY et al., 2012), nor dsRNA-selective importer SID-1 required for RNA transport from neuron to germline (DEVANAPALLY et al., 2015), as null mutants for these components still exhibit the intergenerational effect (Fig. In.6).

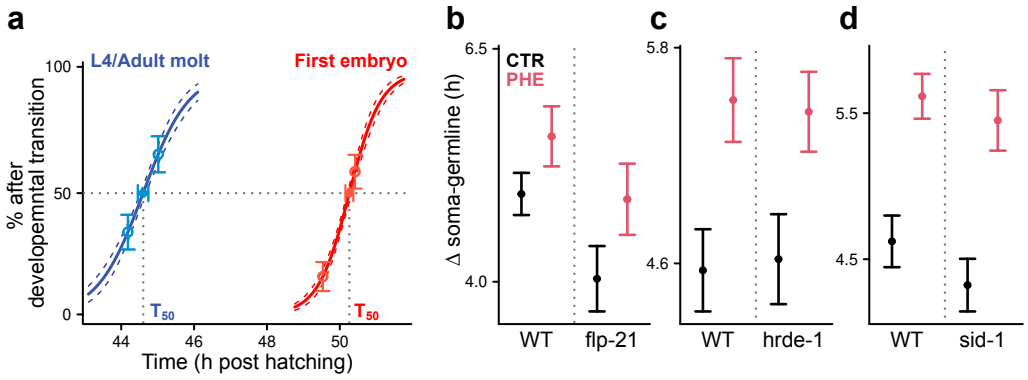


Figure In.6 – Parental control of soma-germline delay does not depend on *flp-21*, *hrde-1*, or *sid-1*
a, Scoring L4-adult molt (blue) and appearance of first embryo (red) for worms as soma and germline developmental transitions respectively. T_{50} (timing at which 50% of the population has passed the transition) is then used to compute Δ soma-germline.
b-d, Germline delay is induced by parental exposure to pheromone in *flp-21* (**b**), *hrde-1* (**c**), and *sid-1* (**d**) null mutants.
Data from (PEREZ et al., 2021b), experimental procedures as described in PEREZ et al., 2021a.

In.3.3 Open questions

What other changes occur in the progeny of pheromone-exposed individuals? Germline delay is not the only effect that parental perception of pheromone causes in the progeny. For instance, PEREZ et al., 2021a also report that the progeny of pheromone-exposed worms spend longer in the L1 stage, and it is likely that there are others. However, to what extent the physiology of the progeny is altered remains to be discovered and characterized. Given odor-learned fear response transgenerationally alters sensory organs specific to the odor in mice (DIAS & RESSLER, 2014), it would be interesting to find a similar feedback loop to amphids in this system.

What is(are) the transmitted signal(s)? Although several inheritance mechanisms have been ruled out (RNAi inheritance, direct RNA transport, chromatin modifiers, insulin/insulin-like pathway), the signal transmitted to progeny is currently unknown. Maternal provisioning of mRNA to the oocyte is a prime suspect, through a mechanism distinct from the *flp-21*-mediated response reported by WASSON et al., 2021. However, it is not excluded for different or complementary signals to exist, especially since crude pheromone extract contains many potentially physiologically relevant chemicals released by the worms that may act independently or in tandem with the ascarosides responsible for germline delay. For example, germline delay in progeny was initially imputed to maternal age because older mothers live in an environment with accumulated pheromone from their own secretions (PEREZ et al., 2017; PEREZ et al., 2021a).

The corollary question, namely how is this signal is interpreted by the progeny to generate the observed (and yet to be discovered) phenotypes is only partially resolved for the germline delay, implicating the TGF-β pathway with DAF-7/TGF-β in the parents, and DAF-3/co-SMAD, DAF-12/NHR in the progeny.

Is this intergenerational inheritance adaptive ? While effects such as pathogen avoidance, or stress resistance can be viewed as priming the next generation for their environment, delaying germline development is more puzzling. Organisms with rapid life cycles like *C. elegans* are critically dependent on their minimum generation time for fitness, more so than brood size, to colonize their environments (HODGKIN & BARNES, 1991). This would suggest germline delay is a maladaptive response. However, experiments up to date surveyed the progeny of pheromone-exposed parents in favorable environments, which may instead disadvantage worms primed for adverse conditions. This remains to be investigated.

In.4 Objectives of the thesis

As detailed in the introductory sections above, sensory perception of pheromone by *C. elegans* adults regulates crucial fitness traits of progeny, such as generation time. However, little is known about the changes that occur in the progeny because of parental exposure to pheromone, as well as what signals transmit parental sensory information to the next generation. The aim of my thesis is to characterize the molecular changes in the developing *C. elegans* embryo caused by parental pheromone exposure to better understand this phenomenon. To achieve this, I devised an efficient strategy to profile gene expression of single embryos across embryogenesis from pheromone-exposed or control parents, which involved adapting single-cell-specific experimental methods to single embryos and developing a novel computational method.

This method, which I named RAPToR, can precisely infer age from gene expression, an ability instrumental to this project for several reasons. Manual staging under a microscope can be imprecise and tedious, particularly when collecting large numbers of samples. Furthermore, imperfect synchronization of embryos and interindividual variation in developmental speed could hinder a sample collection strategy targeting specific time points. Estimating age post-profiling with RAPToR therefore lifted these constraints, allowing us to collect many embryos of all developmental stages at once from a single population, and exploit heterogeneity in developmental stages. We found RAPToR to be extremely robust and expanded its scope to include most common model organisms and stage aging, which are the subject of my first chapter.

Experimentally, I adapted single-cell technologies for large-scale collection and profiling of single embryos. I demonstrate in the second chapter that high-throughput RNA-sequencing of *C. elegans* single embryos can be achieved with minimal adjustments to existing protocols developed for single-cell. I show that single live wild-type embryos can be sorted with standard FACS, and that embryos of any stage can be staged and collected from physical parameters and autofluorescence measured by the FACS without requiring fluorescent markers. Furthermore, I adapt the Smart-seq3 RNA-seq library preparation protocol to profile the transcriptome of single embryos at high complexity for a fraction of the cost of previous protocols, altogether enabling cost-effective scaling up of single-individual profiling.

Lastly, I rely on these methods to collect and analyze preliminary data, and start investigating the changes induced by parental pheromone exposure in my third chapter. With gene expression profiling of single-individuals spanning all of embryogenesis, I reveal large-scale changes affecting the developing nervous system of embryos, particularly implicating sensory organs. In light of these expression changes, I discuss potential phenotypes and their implications for the worms, as well as further experiments to confirm these findings.

Each of the three chapters of my thesis can be approached independently, with separate discussions on their respective subjects, before a brief global discussion on my work concludes my manuscript.

References

ACAR, MURAT, JEROME T METTETAL, and ALEXANDER VAN OUDENAARDEN (2008). "Stochastic switching as a survival strategy in fluctuating environments". In: *Nature genetics* 40.4, pp. 471–475.

AGRAWAL, ANURAG A, CHRISTIAN LAFORSCH, and RALPH TOLLRIAN (1999). "Transgenerational induction of defences in animals and plants". In: *Nature* 401.6748, pp. 60–63.

ALCAZAR, ROSA M, RUEYLING LIN, and ANDREW Z FIRE (2008). "Transmission dynamics of heritable silencing induced by double-stranded RNA in Caenorhabditis elegans". In: *Genetics* 180.3, pp. 1275–1288.

ALTUN, ZF and DH HALL (2003). "WormAtlas Hermaphrodite Handbook-Nervous System-Neuronal Support Cells". In: *WormAtlas. doi* 10.

ALTUN, ZF, DH HALL, and LA HERNDON (2006). "WormAtlas Hermaphrodite Handbook-Introduction". In: *WormAtlas*.

ANDERSON, LUCY M et al. (2006). "Preconceptional fasting of fathers alters serum glucose in offspring of mice". In: *Nutrition* 22.3, pp. 327–331.

ARTYUKHIN, ALEXANDER B, FRANK C SCHROEDER, and LEON AVERY (2013). "Density dependence in Caenorhabditis larval starvation". In: *Scientific reports* 3.1, p. 2777.

ARTYUKHIN, ALEXANDER B et al. (2018). "Metabolomic "dark matter" dependent on peroxisomal β-oxidation in Caenorhabditis elegans". In: *Journal of the American Chemical Society* 140.8, pp. 2841–2852.

BANSAL, ANKITA et al. (2015). "Uncoupling lifespan and healthspan in Caenorhabditis elegans longevity mutants". In: *Proceedings of the National Academy of Sciences* 112.3, E277–E286.

BARGMANN, CORNELIA I (2006). "Chemosensation in C. elegans". In: *WormBook: The online review of C. elegans biology [Internet]*.

BARGMANN, CORNELIA I and H ROBERT HORVITZ (1991). "Control of larval development by chemosensory neurons in Caenorhabditis elegans". In: *Science* 251.4998, pp. 1243–1246.

BAUGH, L RYAN and TROY DAY (2020). "Nongenetic inheritance and multigenerational plasticity in the nematode C. elegans". In: *Elife* 9, e58498.

BRENNER, SYDNEY (1974). "The genetics of Caenorhabditis elegans". In: *Genetics* 77.1, pp. 71–94.

BRITTIN, CHRISTOPHER A et al. (2021). "A multi-scale brain map derived from whole-brain volumetric reconstructions". In: *Nature* 591.7848, pp. 105–110.

BUCKLEY, BETHANY A et al. (2012). "A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality". In: *Nature* 489.7416, pp. 447–451.

BURTON, NICHOLAS O and ERIC L GREER (2022). "Multigenerational epigenetic inheritance: Transmitting information across generations". In: *Seminars in cell & developmental biology*. Vol. 127. Elsevier, pp. 121–132.

BUTCHER, REBECCA A et al. (2007). "Small-molecule pheromones that control dauer development in Caenorhabditis elegans". In: *Nature chemical biology* 3.7, pp. 420–422.

CASANUEVA, M OLIVIA, ALEJANDRO BURGA, and BEN LEHNER (2012). "Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans". In: *Science* 335.6064, pp. 82–85.

CHEN, YAN-TING et al. (2021). "Imprinted lncRNA Dio3os preprograms intergenerational brown fat development and obesity resistance". In: *Nature communications* 12.1, p. 6845.

CHISHOLM, ANDREW D and SUHONG XU (2012). "The Caenorhabditis elegans epidermis as a model skin. II: differentiation and physiological roles". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 1.6, pp. 879–902.

CHOE, ANDREA et al. (2012). "Ascaroside signaling is widely conserved among nematodes". In: *Current Biology* 22.9, pp. 772–780.

CHUTE, CHRISTOPHER D et al. (2019). "Co-option of neurotransmitter signaling for inter-organismal communication in C. elegans". In: *Nature Communications* 10.1, p. 3186.

COOK, STEVEN J, CRISTINE A KALINSKI, and OLIVER HOBERT (2023). "Neuronal contact predicts connectivity in the C. elegans brain". In: *Current Biology* 33.11, pp. 2315–2320.

DALFÓ, DIANA, DAVID MICHAELSON, and E JANE ALBERT HUBBARD (2012). "Sensory regulation of the C. elegans germline through TGF-β-dependent signaling in the niche". In: *Current Biology* 22.8, pp. 712–719.

DEVANAPALLY, SINDHUJA, SNUSHA RAVIKUMAR, and ANTONY M JOSE (2015). "Double-stranded RNA made in C. elegans neurons can enter the germline and cause transgenerational gene silencing". In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2133–2138.

DIAS, BRIAN G and KERRY J RESSLER (2014). "Parental olfactory experience influences behavior and neural structure in subsequent generations". In: *Nature neuroscience* 17.1, pp. 89–96.

DOROQUEZ, DAVID B et al. (2014). "A high-resolution morphological and ultrastructural map of anterior sensory cilia and glia in Caenorhabditis elegans". In: *Elife* 3, e01948.

FITZ-JAMES, MAXIMILIAN H and GIACOMO CAVALLI (2022). "Molecular mechanisms of transgenerational epigenetic inheritance". In: *Nature Reviews Genetics* 23.6, pp. 325–341.

FRÉZAL, LISE and MARIE-ANNE FÉLIX (2015). "C. elegans outside the Petri dish". In: *elife* 4, e05849.

FRIDMANN-SIRKIS, YAEL et al. (2014). "Delayed development induced by toxicity to the host can be inherited by a bacterial-dependent, transgenerational effect". In: *Frontiers in genetics* 5, p. 27.

GOLDEN, JAMES W and DONALD L RIDDLE (1982). "A pheromone influences larval development in the nematode Caenorhabditis elegans". In: *Science* 218.4572, pp. 578–580.

GOLDEN, JAMES W and DONALD L RIDDLE (1984). "The Caenorhabditis elegans dauer larva: developmental effects of pheromone, food, and temperature". In: *Developmental biology* 102.2, pp. 368–378.

GOODMAN, MIRIAM B and PIALI SENGUPTA (2019). "How Caenorhabditis elegans senses mechanical stress, temperature, and other physical stimuli". In: *Genetics* 212.1, pp. 25–51.

GRANDJEAN, VALÉRIE et al. (2015). "RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders". In: *Scientific reports* 5.1, p. 18193.

GREENGARD, PAUL (2001). "The neurobiology of slow synaptic transmission". In: *Science* 294.5544, pp. 1024–1030.

HALLGRIMSSON, BENEDIKT and BRIAN K HALL (2011). *Variation: a central concept in biology*. Elsevier.

HARDO, GEORGEOS and SOMENATH BAKSHI (2021). "Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments". In: *Essays in Biochemistry* 65.1, pp. 67–79.

HIBSHMAN, JONATHAN D, ANTHONY HUNG, and L RYAN BAUGH (2016). "Maternal diet and insulin-like signaling control intergenerational plasticity of progeny size and starvation resistance". In: *PLoS genetics* 12.10, e1006396.

HIRSH, DAVID, DANIEL OPPENHEIM, and MICHAEL KLASS (1976). "Development of the reproductive system of Caenorhabditis elegans". In: *Developmental biology* 49.1, pp. 200–219.

HODGKIN, JONATHAN and THOMAS M BARNES (1991). "More is not better: brood size and population growth in a self-fertilizing nematode". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 246.1315, pp. 19–24.

HUH, DANN and JOHAN PAULSSON (2011). "Random partitioning of molecules at cell division". In: *Proceedings of the National Academy of Sciences* 108.36, pp. 15004–15009.

HUSSEY, ROSALIND et al. (2017). "Pheromone-sensing neurons regulate peripheral lipid metabolism in Caenorhabditis elegans". In: *PLoS genetics* 13.5, e1006806.

INSALL, ROBERT H (2001). "The Whole Organism, and nothing but the Organism". In: *Cell* 107.3, pp. 279–281.

JABLONKA, EVA, MICHAEL LACHMANN, and MARION J LAMB (1992). "Evidence, mechanisms and models for the inheritance of acquired characters". In: *Journal of theoretical biology* 158.2, pp. 245–268.

JORDAN, JAMES M et al. (2019). "Insulin/IGF signaling and vitellogenin provisioning mediate intergenerational adaptation to nutrient stress". In: *Current Biology* 29.14, pp. 2380–2388.

JUANG, BI-TZEN et al. (2013). "Endogenous nuclear RNAi mediates behavioral adaptation to odor". In: *Cell* 154.5, pp. 1010–1022.

KALETSKY, RACHEL et al. (2020). "C. elegans interprets bacterial non-coding RNAs to learn pathogenic avoidance". In: *Nature* 586.7829, pp. 445–451.

KIRKWOOD, THOMAS BL et al. (2005). "What accounts for the wide variation in life span of genetically identical organisms reared in a constant environment?" In: *Mechanisms of ageing and development* 126.3, pp. 439–443.

KISHIMOTO, SAYA et al. (2017). "Environmental stresses induce transgenerationally inheritable survival advantages via germline-to-soma communication in Caenorhabditis elegans". In: *Nature communications* 8.1, p. 14031.

KLOSIN, ADAM et al. (2017). "Transgenerational transmission of environmental information in C. elegans". In: *Science* 356.6335, pp. 320–323.

LIKITLERSUANG, JIRAPAT et al. (2012). "C. elegans tracking and behavioral measurement". In: *JoVE (Journal of Visualized Experiments)* 69, e4094.

LUDEWIG, ANDREAS H and FRANK C SCHROEDER (2018). "Ascaroside signaling in C. elegans". In: *WormBook: The Online Review of C. elegans Biology [Internet]*.

LUDEWIG, ANDREAS H et al. (2013). "Pheromone sensing regulates Caenorhabditis elegans lifespan and stress resistance via the deacetylase SIR-2.1". In: *Proceedings of the National Academy of Sciences* 110.14, pp. 5522–5527.

LUDEWIG, ANDREAS H et al. (2019). "An excreted small molecule promotes C. elegans reproductive development and aging". In: *Nature chemical biology* 15.8, pp. 838–845.

MACOSKO, EVAN Z et al. (2009). "A hub-and-spoke circuit drives pheromone attraction and social behaviour in C. elegans". In: *Nature* 458.7242, pp. 1171–1175.

MAURES, TRAVIS J et al. (2014). "Males shorten the life span of C. elegans hermaphrodites via secreted compounds". In: *Science* 343.6170, pp. 541–544.

MCKNIGHT, KATHERINE et al. (2014). "Neurosensory perception of environmental cues modulates sperm motility critical for fertilization". In: *Science* 344.6185, pp. 754–757.

MOORE, REBECCA S, RACHEL KALETSKY, and COLEEN T MURPHY (2019). "Piwi/PRG-1 argonaute and TGF-β mediate transgenerational learned pathogenic avoidance". In: *Cell* 177.7, pp. 1827–1841.

MORIMOTO, JULIANO, STEPHEN J SIMPSON, and FLEUR PONTON (2017). "Direct and trans-generational effects of male and female gut microbiota in Drosophila melanogaster". In: *Biology letters* 13.7, p. 20160966.

PEMBREY, MARCUS, RICHARD SAFFERY, LARS OLOV BYGREN, et al. (2014). "Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research". In: *Journal of medical genetics* 51.9, pp. 563–572.

PEMBREY, MARCUS E et al. (2006). "Sex-specific, male-line transgenerational responses in humans". In: *European journal of human genetics* 14.2, pp. 159–166.

PEREZ, MARCOS FRANCISCO and BEN LEHNER (2019). "Intergenerational and transgenerational epigenetic inheritance in animals". In: *Nature cell biology* 21.2, pp. 143–151.

PEREZ, MARCOS FRANCISCO et al. (2017). "Maternal age generates phenotypic variation in Caenorhabditis elegans". In: *Nature* 552.7683, pp. 106–109.

PEREZ, MARCOS FRANCISCO et al. (2021a). "Neuronal perception of the social environment generates an inherited memory that controls the development and generation time of C. elegans". In: *Current Biology* 31.19, pp. 4256–4268.

PEREZ, MARCOS FRANCISCO et al. (2021b). *Unpublished data related to Curr. biology article*. unpublished.

PERKINS, LIZABETH A et al. (1986). "Mutant sensory cilia in the nematode Caenorhabditis elegans". In: *Developmental biology* 117.2, pp. 456–487.

POSNER, RACHEL et al. (2019). "Neuronal small RNAs control behavior transgenerationally". In: *Cell* 177.7, pp. 1814–1826.

RECHAVI, ODED et al. (2014). "Starvation-induced transgenerational inheritance of small RNAs in C. elegans". In: *Cell* 158.2, pp. 277–287.

REMY, JEAN-JACQUES (2010). "Stable inheritance of an acquired behavior in Caenorhabditis elegans". In: *Current Biology* 20.20, R877–R878.

REMY, JEAN-JACQUES and OLIVER HOBERT (2005). "An interneuronal chemoreceptor required for olfactory imprinting in C. elegans". In: *Science* 309.5735, pp. 787–790.

REN, PEIFENG et al. (1996). "Control of C. elegans larval development by neuronal expression of a TGF-β homolog". In: *Science* 274.5291, pp. 1389–1391.

ROMÁN, ANGEL-CARLOS et al. (2018). "Histone H4 acetylation regulates behavioral inter-individual variability in zebrafish". In: *Genome biology* 19.1, pp. 1–21.

SCHOTT, DANIEL, ITAI YANAI, and CRAIG P HUNTER (2014). "Natural RNA interference directs a heritable response to the environment". In: *Scientific reports* 4.1, p. 7387.

SEQUENCING CONSORTIUM, C. ELEGANS (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology". In: *Science* 282.5396, pp. 2012–2018.

SMITH, STEPHEN and RAMON GRIMA (2018). "Single-cell variability in multicellular organisms". In: *Nature communications* 9.1, p. 345.

SRINIVASAN, JAGAN et al. (2008). "A blend of small molecules regulates both mating and development in Caenorhabditis elegans". In: *Nature* 454.7208, pp. 1115–1118.

SRINIVASAN, JAGAN et al. (2012). "A modular library of small molecule signals regulates social behaviors in Caenorhabditis elegans". In: *PLoS biology* 10.1, e1001237.

STERKEN, MARK G et al. (2015). "The laboratory domestication of Caenorhabditis elegans". In: *Trends in Genetics* 31.5, pp. 224–231.

SULSTON, JOHN E et al. (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans". In: *Developmental biology* 100.1, pp. 64–119.

TAUFFENBERGER, ARNAUD and J ALEX PARKER (2014). "Heritable transmission of stress resistance by high dietary glucose in Caenorhabditis elegans". In: *PLoS genetics* 10.5, e1004346.

THATTAI, MUKUND and ALEXANDER VAN OUDENAARDEN (2004). "Stochastic gene expression in fluctuating environments". In: *Genetics* 167.1, pp. 523–530.

TOKER, ITAI ANTOINE et al. (2022). "Transgenerational inheritance of sexual attractiveness via small RNAs enhances evolvability in C. elegans". In: *Developmental Cell* 57.3, pp. 298–309.

UPHOFF, STEPHAN et al. (2016). "Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation". In: *Science* 351.6277, pp. 1094–1097.

VASTENHOUW, NADINE L et al. (2006). "Long-term gene silencing by RNAi". In: *Nature* 442.7105, pp. 882–882.

WALSH, MATTHEW R et al. (2015). "Predator-induced phenotypic plasticity within-and across-generations: a challenge for theory?" In: *Proceedings of the Royal Society B: Biological Sciences* 282.1798, p. 20142205.

WASSON, JADIEL A et al. (2021). "Neuronal control of maternal provisioning in response to social cues". In: *Science Advances* 7.34, eabf8782.

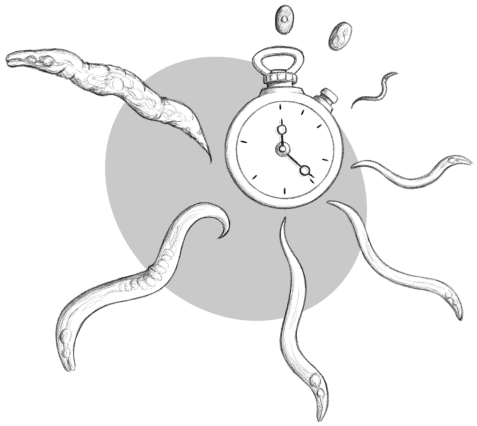
WEBER, KATHERINE P et al. (2010). "Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans". In: *PloS one* 5.11, e13922.

WHITE, JOHN G et al. (1986). "The structure of the nervous system of the nematode Caenorhabditis elegans". In: *Philos Trans R Soc Lond B Biol Sci* 314.1165, pp. 1–340.

WITVLIET, DANIEL et al. (2021). "Connectomes across development reveal principles of brain maturation". In: *Nature* 596.7871, pp. 257–261.

ZHANG, WILLIAM B et al. (2016). "Extended twilight among isogenic C. elegans causes a disproportionate scaling between lifespan and health". In: *Cell Systems* 3.4, pp. 333–345.

1. Predicting age from the transcriptome



Contents

Foreword

The first (and largest) part of my thesis was dedicated to developing RAPToR, a bioinformatic tool capable of **R**eal **A**ge **P**rediction by **T**ranscriptome staging **o**n **R**eference[1].

Initially envisioned specifically for *C. elegans* – where hidden developmental variation in expression studies is widespread – RAPToR proved much more flexible and robust than we expected (and was thus promptly renamed from 'wormAge'). This expansion to various other species (*e.g.*, flies, mice, humans) and sample types (*e.g.* dissected tissues, single cells) led to more substantial findings and a very broad scope of applications. As a result, my supervisor Mirko Francesconi and I were able to publish RAPToR in an excellent methods journal (BULTEAU & FRANCESCONI, 2022).

Alongside the publication of the main article, we contributed to a small briefing on our research, which I include below as a good introduction to this chapter and overview of our main findings.

Since publication, RAPToR has gone through updates, improvements, and further research described in section 1.2, followed by a general discussion.

1.1 Real age prediction from the transcriptome with RAPToR

1.1.1 Research briefing

> MIRKO FRANCESCONI and **ROMAIN BULTEAU** (2022). "Inferring biological age from the transcriptome with RAPToR". in: *Nature Methods* 19.8, pp. 936–937. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01542-y

[1] days of brainstorming.

Inferring biological age from the transcriptome with RAPToR

RAPToR (real age prediction from transcriptome staging on reference) is a new, broadly applicable method that can precisely estimate the age of a sample from a reference transcriptome time series.

This is a summary of:
Bulteau, R. & Francesconi, M. Real age prediction from the transcriptome with RAPToR. *Nat. Methods* https://doi.org/10.1038/s41592-022-01540-0 (2022).

Published online:
11 July 2022

The problem

Genome-wide gene expression profiling is a powerful tool that provides a rich and systematic characterization of biological systems. However, its potential has been limited by the presence of hidden and uncontrolled factors that can obscure and confound the effects of variables of interest. One such confounding factor, especially when profiling fast-developing systems (such as worm or fly embryos and larvae), is often unintended variation in age across samples and experimental conditions. For example, when comparing gene expression in mutant and wild-type animals, we are interested in understanding the specific effect of the mutation on gene expression. However, if the mutant animal develops more slowly than the wild-type and one does not account for this difference (Fig. 1a) — which can be difficult or impossible — the observed expression changes can be dominated by non-mutation-specific changes normally observed during development (Fig. 1b). It is surprising how many datasets are affected by this problem, leading to wrong conclusions[1].

The solution

The expression of genes changes during development and aging; thus, it should be possible to estimate age from gene expression. Many methods have been developed to place single-cell transcriptomes along developmental trajectories[2]. However, these methods require many samples to infer a sufficiently accurate trajectory and do not provide an absolute age estimate, but only rankings or arbitrary values called 'pseudo-time', which are not easily comparable across genetic backgrounds, environmental conditions or experiments. To solve this problem, we reasoned that we could leverage the wealth of time-series gene expression data already available for many organisms as a reference. First, we fill the gaps between the reference time points by interpolating the data, thereby overcoming the limitations of sparsely sampled datasets. Then, each expression profile of interest is independently compared to every time point of the interpolated reference, and the time point with the highest genome-wide correlation is determined as the age estimate. This method yields an absolute biological age estimate for samples across conditions and experiments, provided they are staged using the same interpolated reference.

This simple method, which we named RAPToR (real age prediction from transcriptome staging on reference) is precise (up to minutes for *Caenorhabditis elegans* larval development) and flexible: it works for pooled or single whole animals, dissected tissues and single-cell data; it works for the most common animal models and humans; and it even works for non-model organisms that lack reference data, by using closely related species as reference. When gene expression tissue specificity is known, RAPToR can provide tissue-specific age estimates from whole-animal data. Moreover, it works for both development and aging. When chronological age is known, its comparison with RAPToR age estimates can precisely quantify the effect of genetic or environmental perturbations on the speed of development or aging. Integrating RAPToR age estimates and reference data when performing a differential expression analysis enables the recovery of the specific effect of these perturbations on gene expression even when confounded by age. Thus, data suspected to be confounded by the effects of development or aging can be reanalyzed with RAPToR for validation (or rejection).

Future directions

RAPToR will be useful for large-scale gene expression profiling of single individuals, because it eliminates the need for accurate manual staging (which can be difficult and is usually limited to easily distinguishable stages) or synchronization (which is not always possible or effective with large inter-individual variations). RAPToR also has the potential to accelerate research into aging, as it can precisely quantify the effects of treatments on aging well before the onset of mortality (Fig. 1c).

RAPToR needs existing time-series gene expression data as a reference, which sometimes might be outdated and of poor quality compared with recently produced data. Although RAPToR is robust even when using sparse microarray data as a reference (Fig. 1d), this reliance on datasets could be a limitation. However, the amount and quality of expression data will only increase in the future.

RAPToR is theoretically applicable to any process with robust gene expression dynamics; thus, we plan to test it on disease progression in cancer or neurodegenerative diseases, although the heterogeneous nature of disease progression (compared with development, for example) might complicate the process.

Mirko Francesconi & Romain Bulteau
Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, Lyon, France.

EXPERT OPINION

"These results show how developmental variation discovered by RAPToR can be exploited to increase power, to detect differential expression and to untangle the signal of perturbations of interest even when it is completely confounded with development. This method is very useful for predicting the precise developmental stage from the transcriptome, and it will have a significant impact in the field." **An anonymous reviewer.**

FIGURE

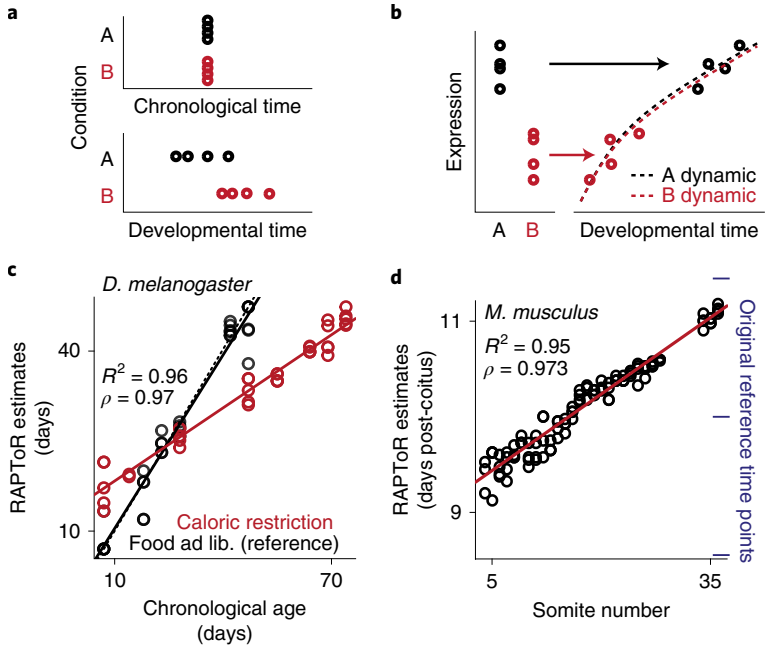


Fig. 1 | Inferring physiological age with RAPToR. a, Groups of individuals sampled at identical chronological times can differ in physiological age if a condition affects developmental or aging speed. **b**, Although A and B have the same gene expression dynamic along development, samples presumed to be at the same age have different gene expression, as condition B delays development. **c**, Chronological and estimated age in aging *Drosophila melanogaster*, showing the effect of caloric restriction compared with food available ad libitum (ad lib.) on lifespan. **d**, Morphological staging (from somite number) of *Mus musculus* embryos and RAPToR age estimates using a very sparse reference dataset. R^2, linear model fit; ρ, Spearman correlation.

BEHIND THE PAPER

We have been aware of how difficult it can be to control for development and how big of an effect this confounder can have since M.F. analyzed a large *C. elegans* expression dataset in which half the variance originated from unintended age differences[3]. The size of the dataset enabled the use of a strategy that unfortunately is not viable for experiments with a smaller sample size, such as the typical '3 wild-types and 3 mutants'. Thus, we aimed to solve this problem, initially by trying to combine reference and sample data using complex methods, with little success. In the end, we tried simple correlations with a reference dataset interpolated for missing data points. The results went far beyond our expectations, as RAPToR is robust and flexible in various organisms and profiling data types, showing that often 'the simpler, the better'. **R.B.**

REFERENCES

1. Snoek, L. B. et al. A rapid and massive gene expression shift marking adolescent transition in *C. elegans*. *Sci. Rep.* **4**, 3912 (2014).
This article shows a striking number of expression profiling experiments that are affected by unintended developmental variation.

2. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
This paper compares over 40 methods for trajectory inference from the transcriptome.

3. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505**, 208–211 (2014).
In this article, profiling data are reanalyzed taking into account unintended developmental spread.

FROM THE EDITOR

"Bulteau and Francesconi have mined the richness of transcriptomic data to develop a powerful strategy for determining the precise age of animals used in biological studies. This work will help to remove confounding age-related variables from comparative studies, and demonstrates through example why this is important. I envision the RAPToR technique becoming a staple in transcriptomics studies involving animals." **Rita Strack, Senior Editor, *Nature Methods*.**

1.1.2 Main article

ROMAIN BULTEAU and MIRKO FRANCESCONI (Aug. 2022). "Real age prediction from the transcriptome with RAPToR". en. In: *Nature Methods* 19.8, pp. 969–975. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01540-0

nature methods

ARTICLES

https://doi.org/10.1038/s41592-022-01540-0

Real age prediction from the transcriptome with RAPToR

Romain Bulteau and Mirko Francesconi ✉

Transcriptomic data is often affected by uncontrolled variation among samples that can obscure and confound the effects of interest. This variation is frequently due to unintended differences in developmental stages between samples. The transcriptome itself can be used to estimate developmental progression, but existing methods require many samples and do not estimate a specimen's real age. Here we present real-age prediction from transcriptome staging on reference (RAPToR), a computational method that precisely estimates the real age of a sample from its transcriptome, exploiting existing time-series data as reference. RAPToR works with whole animal, dissected tissue and single-cell data for the most common animal models, humans and even for non-model organisms lacking reference data. We show that RAPToR can be used to remove age as a confounding factor and allow recovery of a signal of interest in differential expression analysis. RAPToR will be especially useful in large-scale single-organism profiling because it eliminates the need for accurate staging or synchronisation before profiling.

Genome-wide profiling of gene expression is a powerful technique that provides a global and unbiased view of the transcriptional state of a biological system. However, the analysis of gene-expression data can be complicated by uncontrolled and unknown sources of variance—which may be technical or biological in nature[1]—that can mask or confound the effects of variables of interest.

To tackle this problem, several methods have been developed to learn and remove hidden covariates (or surrogate variables) from the data, such as remove unwanted variance[2], surrogate variable analysis[3], or probabilistic estimation of expression residuals[4]. However, a drawback of these methods is that the sources of variance usually remain obscure; therefore, potentially interesting biological variance might also be removed.

Unintended differences in developmental progression across biological replicates or experimental conditions are a major source of variance in gene-expression data of developing systems (Fig. 1a), which can confound (Fig. 1b) or mask (Fig. 1c) the effect of the variable of interest. This is especially true in organisms with rapid life cycles and highly variable growth speed such as worms, fruit fly, or zebrafish, where numerous factors like genetic background, temperature, diet, crowding[5–9], or even the physiological state of the previous generation[9] substantially impact developmental speed. Carefully controlling for all conditions influencing development is therefore particularly challenging, but failing to do so can strongly impact gene expression. For example, in *Caenorhabditis elegans* even a few hours of development may result in 10,000 differentially expressed genes[10]. Hence, it is not surprising that around 50% of variance in gene expression in the profiling of a large panel of *C. elegans* recombinant inbred lines[11] is due to unintended developmental variation[12] and that almost 38% of the datasets that did not intend to include development in a *C. elegans* gene-expression database[13] show substantial developmental variation in gene expression[10].

Estimating the real physiological age of the samples and identifying hidden developmental variation between them is important first to quantify the impact of the perturbation of interest on developmental speed; second, to distinguish perturbation-specific from unspecific changes in gene expression caused by development; third, to uncover time-specific effects of the perturbations under study[12] by including inferred age as a covariate in expression data analyses (such as differential expression analysis). In yeast, analogous ideas were successfully implemented to identify genetic and environmental perturbations impacting specific phases of the cell cycle[14] and direct and specific effects of 700 gene deletions on gene expression after removing the main source of variance (25%): a shared expression signature of cell cycle and growth rate[15].

Extracting developmental progression from transcriptomes has recently become a topic of intense research, especially after the advent of single-cell RNA sequencing. Many algorithms have been developed that learn developmental progression from large-scale bulk, single-cell, or whole-organism transcriptomic data and sort samples along those trajectories (for example Slingshot[16], DPT[17], Monocle[18], and BLIND[19]). However, a major drawback of these trajectory-learning algorithms is that they require large numbers of samples to learn the trajectory of developmental expression changes directly from the data. Moreover, they only output dataset-specific ranks or arbitrary values usually referred to as "pseudo-times" rather than real-age predictions, making it difficult to compare results across datasets or conditions.

To overcome these limitations, we developed a computational framework that exploits available time-series gene-expression data as reference to determine the absolute age of even a single sample from its transcriptome with high precision. We implemented RAPToR in R (available at https://github.com/LBMC/RAPToR), providing references to stage *C. elegans*, *Drosophila melanogaster*, *Danio rerio*, and *Mus musculus* development from gene expression.

We show that RAPToR successfully estimates age during development and ageing, estimates tissue-specific age from whole-organism data, works in dissected tissue and single-cell profiling, and can also estimate age of one species using another species as reference. Finally, we show how to use estimated ages to quantify a perturbation effect on developmental or ageing speed, and recover the specific effects of the variables of interest on gene expression even when completely confounded by age.

Laboratoire de Biologie et Modelisation de la Cellule, Ecole normale superieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Universite Claude Bernard Lyon 1, Lyon, France. ✉e-mail: mirko.francesconi@ens-lyon.fr

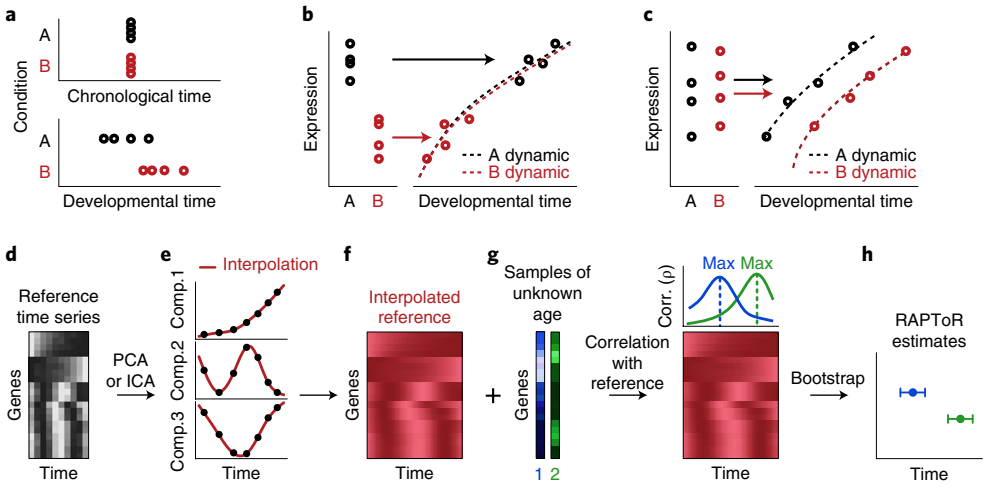


Fig. 1 | Estimating age from the transcriptome using RAPToR. a, Cartoon of individuals sampled at identical chronological times in two conditions, resulting in different developmental age between groups owing to the condition impacting developmental speed. **b**–**c**, Cartoons of differential expression analysis situations where hidden developmental variation is either misinterpreted as an effect of the condition due to development and condition being confounded (**b**), or masking an effect of the condition owing to developmental spread (**c**). **d**–**f**, RAPToR staging exploits existing reference time-series expression data (**d**). This data is first decomposed into principal/independent components, which are interpolated with respect to time (**e**). Comp., component. Interpolated reference is then reconstructed at gene level with interpolated components and gene loadings (**f**). **g**,**h**, For each sample, a correlation profile is built by computing genome-wide Spearman correlation with every time point of the reference (**g**). Corr., correlation. The reference time with maximal correlation becomes the estimate, and bootstrapping on random gene subsets defines a confidence interval (see Methods) (**h**).

Results

RAPToR design. We set out to develop a strategy to continuously stage development from gene expression that would be effective even for experiments with a limited number of samples, in which trajectory learning methods are not applicable. We reasoned that we could exploit existing developmental time-series data as reference to stage samples individually by taking the time point of the reference that has maximum correlation with a sample transcriptome as the age estimate. In this way, the age of each sample is inferred independently from others and outliers only influence their own staging (as opposed to trajectory-based approaches). Furthermore, age estimates acquired on the same reference should be comparable even across different experiments, conditions, and genetic backgrounds.

However, using the time of maximum correlation with the reference as the age estimate limits its precision to the temporal resolution of the reference. To overcome this limitation, we interpolate reference gene expression (Fig. 1d) with respect to time in a dimensionally-reduced space (Fig. 1e and Supplementary Note 1), generating interpolated expression profiles potentially at any time between the original reference time points (Fig. 1f).

The sample age estimate is simply the time point of maximum Spearman correlation between the interpolated reference and the sample gene expression (Fig. 1g). We then compute a confidence interval of the estimate by bootstrapping on genes (Fig. 1h; Methods).

We implemented this strategy in RAPToR, an R package where we provide functions to interpolate references and stage samples.

RAPToR accurately infers developmental age of model organisms. To test RAPToR in the most commonly used animal model organisms, we built interpolated references exploiting existing time-series data on *C. elegans* roundworm embryonic and larval development[20–22], zebrafish embryonic and larval development[23], mouse[24], and fruit fly[25] embryonic development (Supplementary Table 1) and then staged independent time-series experiments of *C. elegans* late-larval development[26] and zebrafish[27,28], mouse[29], and fruit fly[27] embryonic development.

We found RAPToR age estimates accurately match chronological age for both *C. elegans* and zebrafish ($R^2 > 0.99$; Figs. 2a,b), and morphological staging (somite number) for mice ($R^2 = 0.95$; Fig. 2c). Age estimates of fly single embryos[27] less accurately match chronological age, especially at later stages ($R^2 = 0.74$; Fig. 2d). However, this is likely due to the single-individual nature of data as any inter-individual variation in developmental speed would not be averaged out as in bulk data. Indeed, the authors used BLIND[19]—a trajectory-learning method—to re-rank their samples[27] similarly to RAPToR ($\rho > 0.99$; Extended Data Fig. 1). RAPToR estimates do in fact enhance expression dynamics captured by principal components (Figs. 2e,f and Extended Data Fig. 1) and for the majority of genes (Extended Data Fig. 1) in comparison to chronological age (Methods). Thus, RAPToR estimates the true physiological age of individuals and reveals the heterogeneity of their developmental speeds.

Reference interpolation greatly improves staging accuracy. Crucially, reference interpolation allows staging with an accuracy far beyond the original sampling resolution of reference time series. Indeed, RAPToR accurately stages a dense zebrafish time course[28] with over 40 times the temporal resolution of the reference before interpolation (Extended Data Fig. 2 and Supplementary Note 1; Methods). RAPToR estimates also stay remarkably accurate and precise even when staging samples with a few hundred genes or with noisy data (Supplementary Note 1 and Supplementary Figs. 1–4), and are robust to reference interpolation parameters (Supplementary Note 1, Supplementary Figs. 5 and 6, and Supplementary Table 2).

RAPToR correctly infers developmental speed scaling factors. RAPToR estimates are relative to the reference chronological age. Thus, one can use RAPToR to stage samples with known chronological age to estimate developmental speed differences or scaling factors with a reference. For example, staging a *C. elegans* developmental time series grown at 25 °C[26] on the reference grown at 20 °C[20] recapitulates the expected 1.5-fold increase in developmental speed owing to temperature increase[20] (Fig. 2a and Supplementary Note 1).

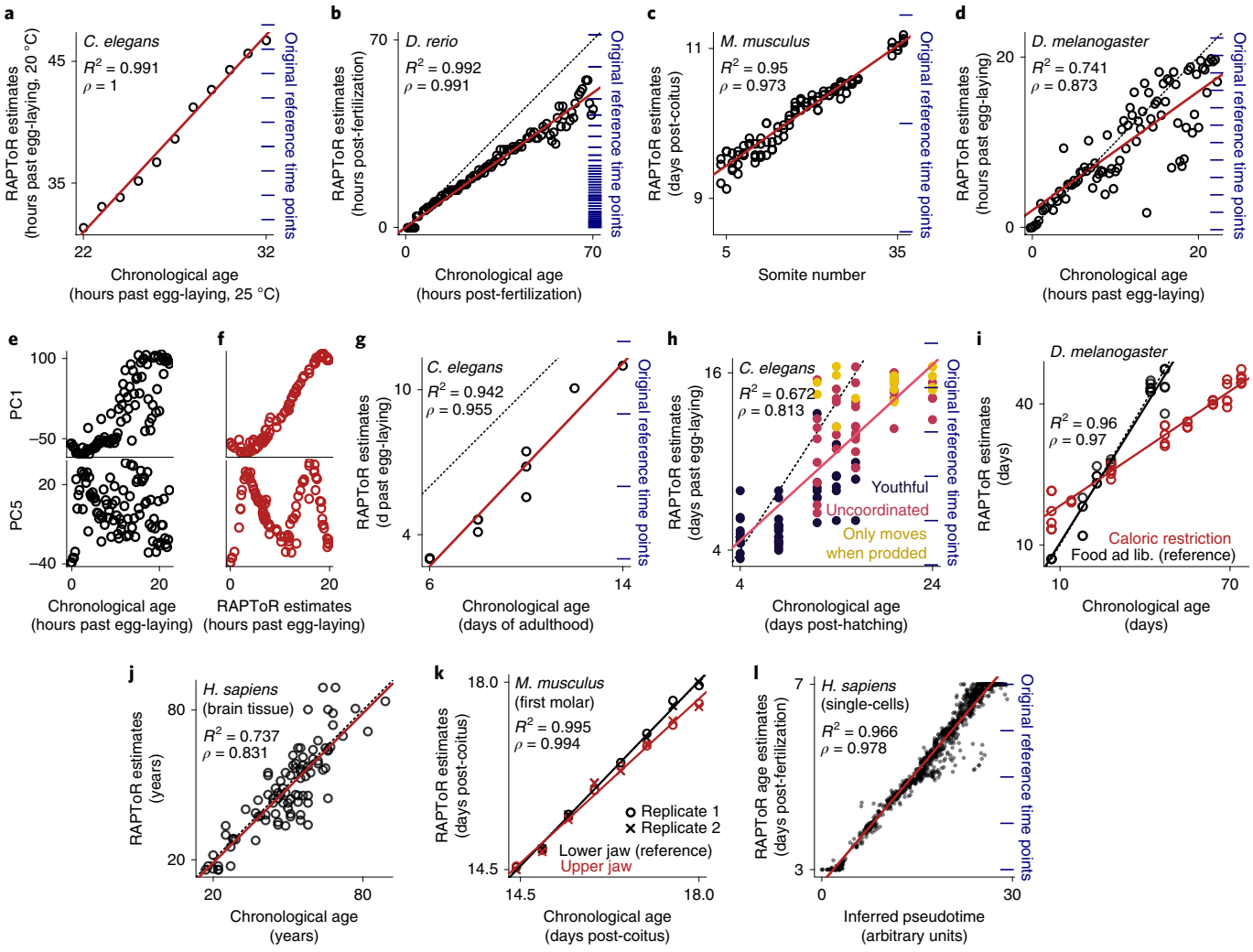


Fig. 2 | RAPToR precisely stages development and ageing, and works from whole-organism to single-cell data. a,**b**, Chronological age versus RAPToR estimates of *C. elegans* late-larval samples[26] (linear model is $y = -4.7 + 1.6x$; **a**) and *D. rerio* embryo samples[27] (linear model is $y = 0.7x$; **b**). **c**, Somite number versus RAPToR estimates of *M. musculus* embryo samples[29] (linear model is $y = 9.2 + 0.05x$). **d**, Chronological age versus RAPToR estimates of *D. melanogaster* embryo samples[27] (linear model is $y = 1.3 + 0.77x$). **e**,**f**, Selected principal components of the data staged in **d**, plotted in black along chronological age (**e**) and in red along RAPToR estimates (**f**). **g**–**j**, Chronological age versus RAPToR estimates of adult *C. elegans* bulk samples[33] (**g**) and single-worms[34] (**h**), of adult *D. melanogaster*[35] (**i**), and of human brain tissue[36] (**j**). **k**, Chronological age versus RAPToR estimates of dissected samples of upper jaw first molars from *M. musculus* embryos staged using the lower jaw samples as reference[37,38]. **l**, Inferred pseudo-time versus RAPToR estimates of *H. sapiens* embryo single cells[39]. **a**–**d**,**g**,**h**, Staged samples[26,27,29,33,34] and references[20,23–25] are from independent time-series experiments. Original time points of the references within the plot area are shown to the right (blue), but the references can span much longer coverage.

RAPToR performs well on ageing. While RAPToR works very well with robust expression changes during development, ageing and ageing-related changes in gene expression are widely known to be heterogeneous, stochastic, and strongly influenced by environmental factors[30–32]. This can potentially limit the applicability of RAPToR to ageing. In fact, RAPToR performs poorly between independent ageing time series (but works within experiments; Supplementary Note 1 and Supplementary Fig. 7) with references built using the whole transcriptome. We reasoned RAPToR performance could increase by strengthening the ageing signal in the reference. Indeed, by building RAPToR references restricted to genes with robust monotonous trends along ageing (Methods), we could successfully estimate ageing in *C. elegans* bulk[33] ($R^2 = 0.94$; Fig. 2g) and single-worm[34] samples ($R^2 = 0.67$, Fig. 2h), *Drosophila*[35] ($R^2 = 0.96$; Fig. 2i), and humans[36] ($R^2 = 0.74$; Fig.2j, Supplementary Note 1 and Supplementary Fig. 8). Importantly, single worms staged older than their chronological age behaved like older individuals[34] and vice versa (Fig. 2h), moreover age estimates of flies under caloric restriction are consistent with the expected lifespan extension[35] (Fig. 2i). This shows that RAPToR age estimates recapitulate true differences in biological age across individuals or environmental conditions. We conclude that RAPToR reliably infers ageing from transcriptomic data.

RAPToR accurately stages dissected tissue samples. We tested RAPToR on expression profiling from dissected tissues—where variation in cell-type composition and relative amount might potentially confound staging—using time series of *M. musculus* upper- and lower-jaw first molar development[37,38]. Since these two organs have very similar development[37], we built a lower-jaw reference to stage the upper-jaw samples (Methods). RAPToR not only accurately estimates age ($R^2 > 0.99$; Fig. 2k), but also correctly infers the known developmental delay of upper molars in comparison to lower molars[37,38]. Thus, despite potential confounders, RAPToR is effective and precise on dissected tissue samples.

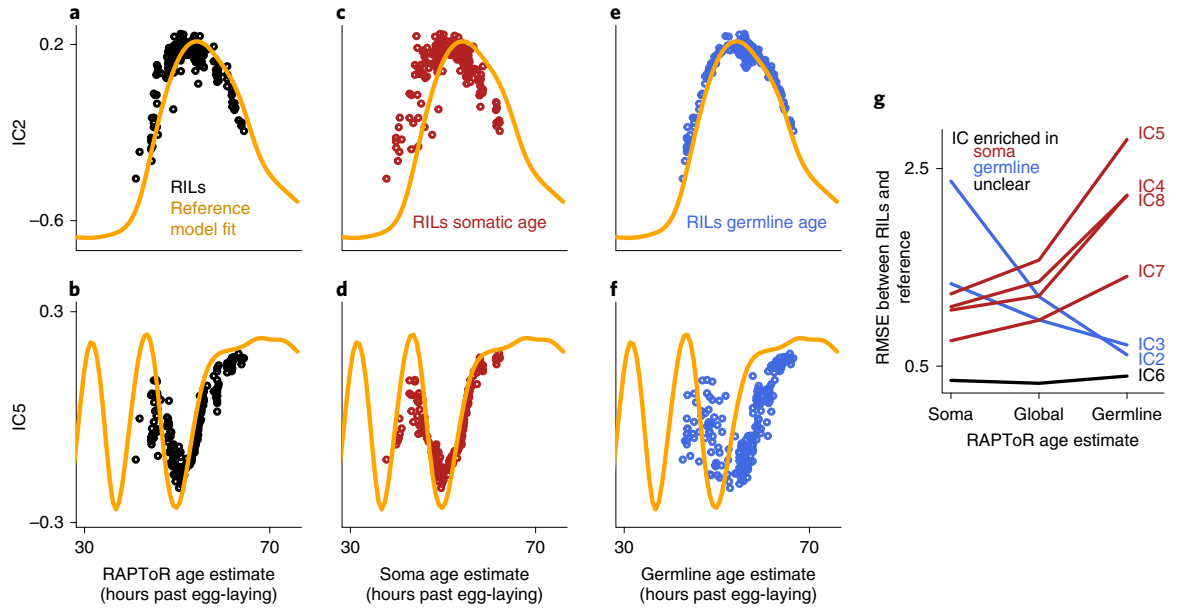


Fig. 3 | Tissue-specific staging. a,**b**, Selected independent components from ICA on joint *C. elegans* RILs[11] (dots) and reference data[21] they were staged on (orange line). **c**–**f**, As in (**a**,**b**), with RILs plotted in red along soma age (**c**,**d**), and in blue along germline age (**e**,**f**). **g**, Root mean square error between RILs and reference for independent components 2–8 when using soma, global or germline age estimates.

RAPToR accurately stages single-cell data. Single-cell expression data is usually much sparser and noisier than bulk data, which might potentially limit RAPToR performance. We therefore tested whether RAPToR could stage single cells from human early embryo development[39] by building a reference with a random subset of the data and staging the remaining cells (Methods). Age estimates not only match the chronological time ($R^2 = 0.87$; Supplementary Fig. 9), but also strongly correlate with the pseudo-times computed by the authors[39] ($R^2 = 0.97$; Fig. 2l). Therefore, despite the sparsity of the data, RAPToR ranks cells as well as pseudo-time methods specifically designed for single-cell data but at the same time provides a real biological time for each cell.

RAPToR is robust to genetic variation in gene expression. Variable genetic background is another potential confounder, so we tested RAPToR performance on expression data for over 200 *C. elegans* recombinant inbred lines (RILs) showing extensive genetic variation in gene expression[11]. RAPToR closely matches previous estimates from a trajectory-learning approach[13] ($R^2 = 0.94$; Extended Data Fig. 3), thus confirming the RILs span mid-larval to young-adult stage, a period with vast expression changes both in the soma (molting) and the germline (spermatogenesis and oogenesis) of worms.

We noticed that some gene-expression dynamics in the RILs are advanced and others delayed in comparison to the reference (Figs. 3a,b, Extended Data Fig. 4 and Supplementary Note 1). Shifts between soma and germline development (soma–germline heterochrony) are easily induced by environmental and physiological changes in *C. elegans*[9,40]. Indeed, a consistent enrichment of soma and germline genes in advanced and delayed dynamics respectively suggests soma–germline heterochrony between the reference and the RILs (Extended Data Fig. 4).

Quantifying heterochrony with tissue-specific staging. To confirm this we used germline- and soma-specific gene sets[22,26] to separately stage the germline and soma of the RILs (Methods; Extended Data Fig. 3). We find germline- and soma-specific dynamics align better on the reference when staged with the corresponding gene set (Figs. 3d,e,g) while they are otherwise shifted (Figs. 3c,f), confirming heterochrony between reference and RILs. Thus tissue-specific staging outperforms global staging in case of heterochrony between the reference and the samples to stage.

Beyond differences between the reference and RILs, we noticed that tissue-specific staging also decreases variance among the RILs. Indeed, germline genes are better fit by germline than soma age and vice versa, suggesting soma–germline heterochrony among the RILs (Extended Data Fig. 5). However, when searching for the genetic basis of this heterochrony with a multivariate quantitative trait loci (QTL) analysis, we found no significant genetic locus (even at a false discovery rate (FDR) of 0.5) and overall no significant amount of genetic variance in heterochrony (Supplementary Note 1), which is therefore likely due to unknown and uncontrolled environmental differences or to a very complex genetic architecture not captured by the model.

In summary, by using tissue-specific gene sets RAPToR provides accurate tissue-specific age estimates from whole-organism expression despite varying genetic background.

Staging on references of a different species. Developmental time-series data are often unavailable for non-model organisms. However, gene-expression dynamics during development are often well-conserved across related species, especially during the phylotypic stage[41]. Seeing the robustness of RAPToR to genetic variation within species, we decided to test how well RAPToR can stage one species on a related species.

Staging time series of embryo development across six *Drosophila* species[41] on a *D. melanogaster* reference using orthologs indeed results in accurate age estimates ($R^2 > 0.99$; Fig. 4a) despite decreasing overall correlation with increasing phylogenetic distance (Fig. 4b). Moreover, we infer between-species growth speed factors matching those found by the authors (Supplementary Table 3) and account for small age differences between replicates of each time point, which refines expression dynamics (Extended Data Fig. 6) and reduces unexplained variance in the data (Supplementary Fig. 10).

Encouraged by this, we probed RAPToR limits by staging samples on more distant reference species. We were able to stage

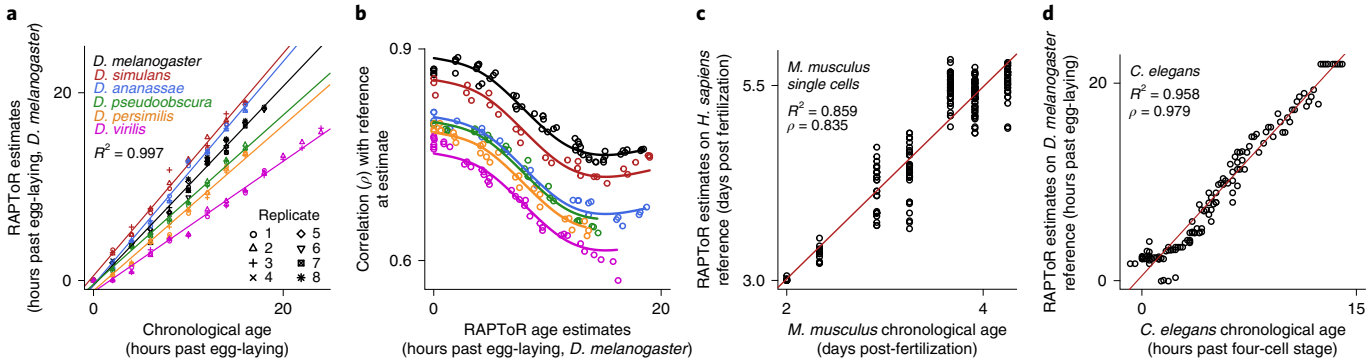


Fig. 4 | Staging samples cross-species. a, Chronological age versus RAPToR estimates for time series of embryo development of six *Drosophila* species[41] staged on a *D. melanogaster* reference[25] (Extended Data Fig. 6). **b**, Spearman correlation between samples from **a** and the reference at age estimate, along RAPToR estimates. **c**, Chronological age versus RAPToR estimates for single cells of *M. musculus* embryos[42], staged on a *H. sapiens* single-cell reference[39] using orthologs (Extended Data Fig. 7). **d**, Chronological age versus RAPToR estimates for *C. elegans* embryo samples[27], staged on a *D. melanogaster* reference[25] using orthologs.

early embryo mouse single cells[42] on a human reference[39], matching both chronological age ($R^2=0.86$; Fig. 4c) and pseudo-times ($\rho=0.95$; Extended Data Fig. 7), as well as human[43] on cow[44] whole embryos ($R^2=0.83$; Supplementary Fig. 11). To our surprise, we could even successfully stage *C. elegans* embryogenesis[27] on a *D. melanogaster* reference ($R^2=0.96$; Fig. 4d, Extended Data Fig. 8 and Supplementary Note 1), two species separated by 600 million years of evolution[45].

Which biological processes with highly conserved dynamics during embryogenesis could account for this accurate staging? We found that gene-expression signatures of decreasing cell proliferation shared across phyla[27] and signatures of muscle development or cell differentiation are necessary and almost sufficient for accurate staging (Supplementary Note 1, Extended Data Fig. 7,8 and Supplementary Tables 4–11; Methods).

Thus, RAPToR can stage non-model organisms using available close species data and perform well even in extremely distant species, when applied to developmental stages with highly conserved developmental dynamics.

To summarize, RAPToR performs well across the organisms, sample types, and diverging genetic backgrounds and species we tested, yielding estimates that are precise, accurate thanks to interpolation, and robust to gene-set size changes.

RAPToR finds hidden drug effects on germline development. RAPToR absolute age estimates are useful in many ways. First, rather than just getting a list of differentially expressed genes from profiling data, RAPToR precisely quantifies the effect of perturbations on developmental timing, including in a tissue-specific way. For example, tissue-specific staging of *C. elegans* exposed to three concentrations of mefloquine, dichlorvos, and fenamiphos[46] found that all three drugs induce a similar germline-specific and dose-dependent developmental delay (Fig. 5a, Supplementary Note 2 and Supplementary Fig. 12).

RAPToR improves differential expression analyses. Even when known chronological age is included as a model covariate in differential expression analyses, replacing it by RAPToR age estimates increases statistical power. For example, using RAPToR estimates instead of chronological age when analyzing expression changes in *C. elegans pash-1* versus wild type (WT)[47] (Fig. 5b), detects up to 60% more differentially expressed genes in *pash-1* and 10% more differentially expressed genes across development thanks to overall better model fits (Fig. 5c, Supplementary Fig. 13 and Supplementary Note 2).

Quantifying developmentally driven changes in gene expression. If an experimental condition strongly impacts developmental speed but perturbed and control samples are collected at the same chronological—and therefore different physiological—time (Fig. 1a), the variable of interest will be completely confounded with development. Thus, purely developmental expression changes are wrongly attributed to the perturbation of interest (Fig. 1b). As an example, we reanalyze a dataset comparing young-adult *C. elegans* that developed through dauer state (post-dauer) to controls that did not[48]. The authors found a downregulation of spermatogenesis-associated genes and an upregulation of oogenesis-associated genes from which they concluded that post-dauer animals have reduced spermatogenesis and increased oogenesis. However, as *C. elegans* switch from sperm to egg production during development, this pattern could simply be explained by post-dauer samples being physiologically older than controls. This is indeed what RAPToR found (Fig. 5d, Supplementary Fig. 14 and Supplementary Note 2). Furthermore, strong correlation ($r>0.8$) between the observed expression changes in germline genes and the expected developmental expression changes calculated from matching time points in the reference (Fig. 5e, Extended Data Fig. 9, Supplementary Fig. 14 and Supplementary Note 2) suggests that, despite synchronization efforts, most of the initially observed differential expression is due to uncontrolled differences in developmental progression.

Recovering confounded perturbation-specific effects. We reasoned that integrating RAPToR age estimates and developmental gene expression from the reference in the differential expression analysis should allow us to extract perturbation-specific expression changes even when the variable of interest is completely confounded with development (Supplementary Note 2 and Extended Data Fig. 10). We tested this using a *C. elegans* larval development time series of *xrn-2* mutant and relative WT control sampled every 1.5 h[49]. We defined a gold standard of truly differentially expressed genes in the mutant, which allowed us to vary the age difference between mutant and WT and quantify first the amount, intensity, and variance of expression changes owing to development (Fig. 5f,g, Extended Data Fig. 10 and Supplementary Note 2); second, the deleterious impact of these developmental expression changes on the performance of a standard analysis in detecting truly differentially expressed genes; and third, the improvement obtained by integrating RAPToR estimates and reference expression data in the model. As expected, with increasing age differences between mutant and WT, the performance of a standard test of differential expression sharply decreases (Fig. 5h). However, performance is greatly recovered by

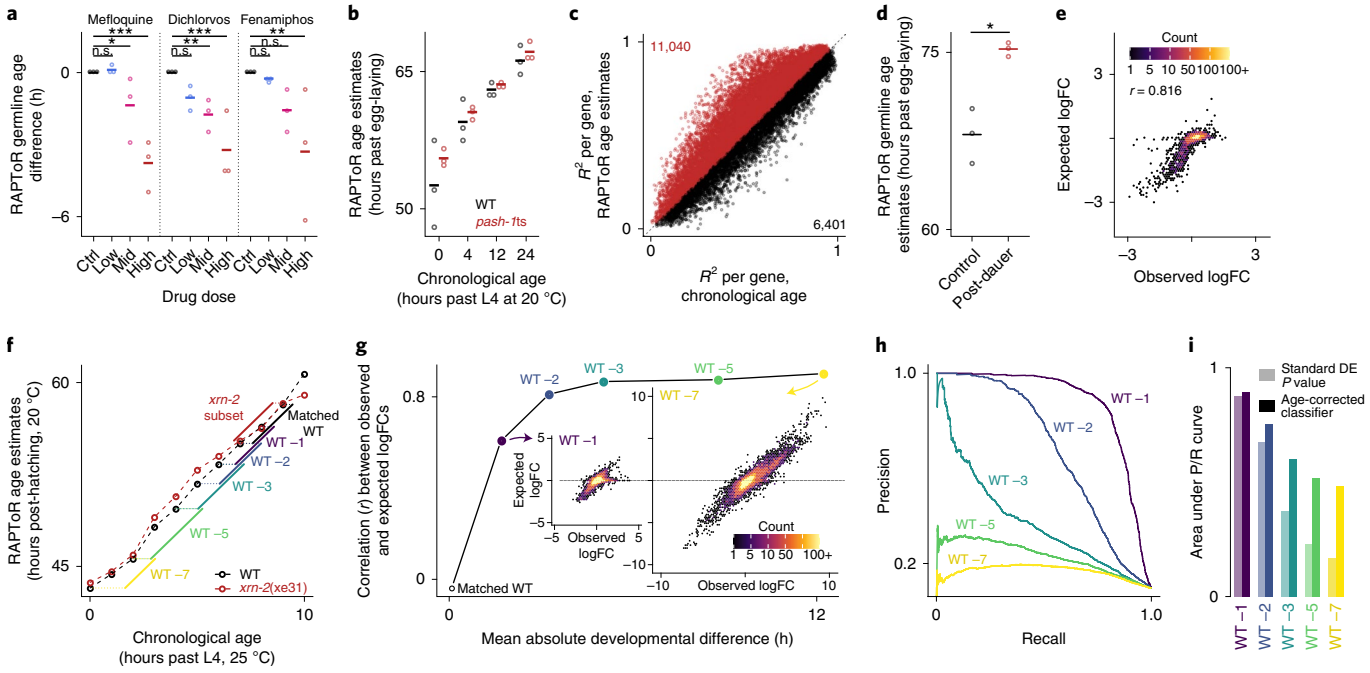


Fig. 5 | Quantifying and correcting for developmental effects using RAPToR age estimates. a, Effect of increasing drug dose exposure[46] on RAPToR estimates of *C. elegans* germline age (Methods; Supplementary Fig. 12). *P* values are derived from two-sided *t*-tests on linear model coefficients for each drug. From top to bottom, respectively: mefloquine, $P = 6.32 \times 10^{-5}$, $P = 0.0252$, $P = 0.8248$; dichlorvos, $P = 1.54 \times 10^{-4}$, $P = 0.0067$, $P = 0.0620$; fenamiphos, $P = 0.0040$, $P = 0.0946$, $P = 0.7702$. n.s., not significant. **b**, RAPToR estimates versus reported chronological age highlight large developmental spread within time points of WT *C. elegans* and *pash-1*ts time series[47] (Supplementary Note 2 and Supplementary Fig. 13). **c**, R^2 per gene of identical models with chronological age, or RAPToR estimates. Genes and gene counts above and below dashed line ($x = y$) are indicated in red and black, respectively. **d**, Germline age estimates of control and post-dauer (PD) *C. elegans* adults[48], $P = 0.025$, derived from a two-tailed *t*-test. **e**, Germline gene logFCs between control and PD from **d** in comparison to logFCs expected from developmental time difference only (Extended Data Fig. 10). **f**, Chronological age versus RAPToR estimates of a time series of WT *C. elegans* and *xrn-2* late larval development[49]. Sample subsets defining a gold standard of truly DE genes and shifted WT sets used in subsequent panels are color-coded. **g**, Correlation of observed logFCs and expected developmental logFCs computed from the interpolated reference between the *xrn-2* subset and increasingly shifted WT sets from **f** (Supplementary Note 2). **h**, PR curves showing the performance of a standard differential-expression model *P* value for each shifted WT subset in detecting gold-standard DE genes. **i**, AUPRC of standard differential expression model *P* value from **h**, or of the age-corrected classifier for each shifted WT subset in detecting gold-standard DE genes (Supplementary Note 2). DE, differentially expressed. In **a**,**b**,**d**, central bar denotes group mean.

integrating reference data in the model, especially for large age differences when the mutant effect would be fully confounded by development (Fig. 5i, Extended Data Fig. 10 and Supplementary Note 2).

In summary, we showed that using RAPToR and reference data it is possible to measure the impact of development in gene-expression analyses and recover the specific effect of a perturbation even when completely confounded with development.

Discussion

We present RAPToR, a computational strategy to accurately estimate the age of samples from their gene-expression profile. Unlike trajectory-based methods, RAPToR exploits existing reference time-series data to continuously stage each sample separately, providing several advantages: first, it eliminates the need for large datasets to infer developmental trajectories; second, it provides absolute developmental times that are comparable across datasets, conditions, genetic backgrounds, profiling technologies and other covariates; and third, outliers have no impact on the staging of other samples as each sample is staged independently.

While RAPToR is limited by the existence of reference time-series data, interpolation allows precise staging well beyond the resolution of the original reference data, enabling the use of sparse time series as references. RAPToR estimates age both during development or ageing in most common animal models and humans, from bulk, single-individual, dissected-tissue, or single-cell expression profiles, and can also infer tissue-specific age from whole-organism profiles. Importantly, RAPToR can stage one species using a close species as reference, which dramatically expands the scope of RAPToR, including to non-model organisms. We showed how RAPToR absolute estimates can be exploited in many ways: to detect the effect of a perturbation or treatment on developmental or ageing speed; as model covariates to increase statistical power to detect differential expression; finally, we showed that even in the extreme scenario when the perturbation of interest is completely confounded with development, it is still possible to recover genuine perturbation-specific expression changes by integrating reference data in differential expression analysis.

RAPToR can currently only stage on one developmental or ageing trajectory, so a future improvement will be to provide RAPToR with the ability to stage on multiple branching trajectories. Another avenue for improvement is to adapt this approach to data other than genome-wide gene expression, such as genome-wide binding data.

We anticipate our strategy of staging post-profiling with RAPToR will be especially useful in large-scale single-organism profiling experiments since it eliminates the need for synchronization or for tedious and potentially difficult steps of accurate staging before profiling.

To conclude, we remark that our approach is not restricted to development or ageing but can in principle be applied to any process with robust underlying reference gene-expression dynamics

(for example, cell differentiation, cell cycle, disease progression, and drug response) and its scope will only expand with the increasing availability of time-series profiling data.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgments, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01540-0.

Received: 8 September 2021; Accepted: 25 May 2022;
Published online: 11 July 2022

References

1. Francesconi, M. & Lehner, B. Reconstructing and analysing cellular states, space and time from gene expression profiles of many cells and single cells. *Mol. Biosyst.* **11**, 2690–2698 (2015).
2. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
3. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161 (2007).
4. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
5. Gómez-Orte, E. et al. Effect of the diet type and temperature on the *C. elegans* transcriptome. *Oncotarget* **9**, 9556–9571 (2018).
6. MacNeil, L. T., Watson, E., Arda, H. E., Zhu, L. J. & Walhout, A. J. Diet-induced developmental acceleration independent of TOR and insulin in *C. elegans*. *Cell* **153**, 240–252 (2013).
7. Ludewig, A. H. et al. Larval crowding accelerates *C. elegans* development and reduces lifespan. *PLoS Genet.* **13**, e1006717 (2017).
8. Kuntz, S. G. & Eisen, M. B. *Drosophila* embryogenesis scales uniformly across temperature in developmentally diverse species. *PLoS Genet.* **10**, e1004293 (2014).
9. Perez, M. F., Francesconi, M., Hidalgo-Carcedo, C. & Lehner, B. Maternal age generates phenotypic variation in *Caenorhabditis elegans*. *Nature* **552**, 106–109 (2017).
10. Snoek, L. B. et al. A rapid and massive gene expression shift marking adolescent transition in *C. elegans*. *Sci Rep.* **4**, 3912 (2014).
11. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372–376 (2010).
12. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505**, 208–211 (2014).
13. Hibbs, M. A. et al. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23**, 2692–2699 (2007).
14. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl Acad. Sci.* **100**, 10370–10375 (2003).
15. O'Duibhir, E. et al. Cell cycle population effects in perturbation studies. *Mol. Syst. Biol.* **10**, 732 (2014).
16. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
17. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
18. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
19. Anavy, L. et al. BLIND ordering of large-scale transcriptomic developmental timecourses. *Development* **141**, 1161–1166 (2014).
20. Kim, Dhyun, Grün, D. & van Oudenaarden, A. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. *Nat. Genet.* **45**, 1337–1344 (2013).
21. Meeuse, M. W. et al. Developmental function and state transitions of a gene expression oscillator in *Caenorhabditis elegans*. *Mol. Syst. Biol.* **16**, e9498 (2020).
22. Reinke, V., San Gil, I., Ward, S. & Kazmer, K. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131**, 311–323 (2004).
23. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
24. Xue, L. et al. Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis. *BMC Genomics* **14**, 568 (2013).
25. Graveley, B. R. et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**, 473–479 (2011).
26. Hendriks, G.-J., Gaidatzis, D., Aeschimann, F. & Großhans, H. Extensive oscillatory gene expression during *C. elegans* larval development. *Mol. Cell* **53**, 380–392 (2014).
27. Levin, M. et al. The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637–641 (2016).
28. Rauwerda, H. et al. Transcriptome dynamics in early zebrafish embryogenesis determined by high-resolution time course analysis of 180 successive, individual zebrafish embryos. *BMC Genomics* **18**, 287 (2017).
29. Collins, J. E. et al. Common and distinct transcriptional signatures of mammalian embryonic lethality. *Nat. Commun.* **10**, 2792 (2019).
30. Somel, M., Khaitovich, P., Bahn, S., Pääbo, S. & Lachmann, M. Gene expression becomes heterogeneous with age. *Curr. Biol.* **16**, R359–R360 (2006).
31. Kedlian, V. R., Donertas, H. M. & Thornton, J. M. The widespread increase in inter-individual variability of gene expression in the human brain with age. *Aging* **11**, 2253–2280 (2019).
32. Martinez-Jimenez, C. P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
33. Hou, L. et al. A systems approach to reverse engineer lifespan extension by dietary restriction. *Cell Metab.* **23**, 529–540 (2016).
34. Golden, T. R., Hubbard, A., Dando, C., Herren, M. A. & Melov, S. Age-related behaviors have distinct transcriptional profiles in *Caenorhabditis elegans*. *Aging Cell* **7**, 850–865 (2008).
35. Pletcher, S. D. et al. Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr. Biol.* **12**, 712–723 (2002).
36. Chen, C.-Y. et al. Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proc. Natl Acad. Sci.* **113**, 206–211 (2016).
37. Pantalacci, S. et al. Transcriptomic signatures shaped by cell proportions shed light on comparative developmental biology. *Genome Biol.* **18**, 29 (2017).
38. Sémon, M. *et al.* Comparison of developmental genome expression in rodent molars reveals extensive developmental system drift. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.22.043422 (2020).
39. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
40. Perez, M. F. Neuronal perception of the social environment generates an inherited memory that controls the development and generation time of *C. elegans*. *Curr. Biol.* **31**, 4256–4268 (2021).
41. Kalinka, A. T. et al. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).
42. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
43. Vassena, R. et al. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699–3709 (2011).
44. Cuthbert, J. M. et al. Comparing mRNA and sncRNA profiles during the maternal-to-embryonic transition in bovine IVF and scNT embryos. *Biol. Reprod.* **105**, 1401–1415 (2021).
45. Li, J. J., Huang, H., Bickel, P. J. & Brenner, S. E. Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.* **24**, 1086–1101 (2014).
46. Lewis, J. A., Szilagyi, M., Gehman, E., Dennis, W. E. & Jackson, D. A. Distinct patterns of gene and protein expression elicited by organophosphorus pesticides in *Caenorhabditis elegans*. *BMC Genomics* **10**, 202 (2009).
47. Lehrbach, N. J. et al. Post-developmental microRNA expression is required for normal physiology, and regulates aging in parallel to insulin/IGF-1 signaling in *C. elegans*. *RNA* **18**, 2220–2235 (2012).
48. Hall, S. E., Beverly, M., Russ, C., Nusbaum, C. & Sengupta, P. A cellular memory of developmental history generates phenotypic diversity in *C. elegans*. *Curr. Biol.* **20**, 149–155 (2010).
49. Miki, T. S., Carl, S. H. & Großhans, H. Two distinct transcription termination modes dictated by promoters. *Genes Dev.* **31**, 1870–1879 (2017).

Methods

Analyses were all performed using the R statistical software (v.4.1.2)

Data pre-processing. Probe or gene IDs of datasets were converted to standard IDs (WBGene IDs for *C. elegans*, FBgn IDs for *D. melanogaster*, Ensembl IDs for *D. rerio*, *M. musculus*, *H. sapiens* and *B. taurus*). When multiple probes or IDs matched a single standard ID, they were mean-aggregated for microarray, sum-aggregated for RNA-seq counts. IDs with no standard ID match were dropped.

For RNA-seq datasets, gene-level transcripts per kilobase million (TPM) data was used when available, or computed from raw counts using transcript lengths from the Ensembl biomart (v.99). No remapping of the transcriptomes was done, aside from the *M. musculus tooth* data (see below). No background correction was applied to microarray data.

Samples were considered of poor quality and discarded when the 99th percentile of the distribution of their Spearman correlation coefficients with other samples fell below a threshold defined below for each dataset.

Expression values for all datasets were quantile normalized using the normalizeBetweenArrays function from limma[50] (v.3.50.1) on $\log(\boldsymbol{X}+1)$ transformed values unless otherwise specified.

RAPToR implementation. Our method is implemented in an R package, RAPToR (v.1.1.6), which can be downloaded and installed from https://github.com/LBMC/RAPToR.

Functions for staging samples, plotting results, interpolation, and building references are included in the package. Detailed vignettes on general usage, reference building, and showcases are also provided with the package.

Auxiliary R data packages include references for *C. elegans* (embryonic, larval, and young-adult to adult development, https://github.com/LBMC/wormRef), *D. melanogaster* (embryonic development, https://github.com/LBMC/drosoRef), *D. rerio*, (embryonic and larval development, https://github.com/LBMC/zebraRef) and *M. musculus* (embryonic development, https://github.com/LBMC/mouseRef).

Reference interpolation. Let $\boldsymbol{X}$ $(m \times n)$ be the gene-expression matrix of m genes by n samples. The matrix is first gene-centered such that $\boldsymbol{X_0} = \boldsymbol{X} - \text{rowMeans}(\boldsymbol{X})$. We then use independent component analysis (ICA) ('ica' function, 'icafast' library v.1.0.2) or principal component analysis (PCA) ('prcomp' base R function) to decompose the data into a component space of dimension c such that $\boldsymbol{X_0} = \boldsymbol{G}\,\boldsymbol{S}^T$, with $\boldsymbol{G}$ $(m \times c)$ the gene loadings, and $\boldsymbol{S}$ $(n \times c)$ the sample scores. Columns of $\boldsymbol{S}$ are interpolated on with respect to time (and other potential variables of interest, for example, batch), forming a new matrix $\boldsymbol{T}$ $(l \times c)$ of l new time points in component space. The full interpolated expression matrix $\boldsymbol{Y}$ $(m \times l)$ is then reconstructed by multiplying the gene loadings matrix by the transposed $\boldsymbol{T}$ and by adding the gene centers $\boldsymbol{Y} = \boldsymbol{G}\,\boldsymbol{T}^T + \text{rowMeans}(\boldsymbol{X})$.

To interpolate the components, we fit generalized additive models to handle non-linear dynamics through splines with the 'gam' function in the 'mgcv' package (v.1.8.39) using a single model formula for all components selected by cross validation (CV) as follows. CV training sets are built with 80% of samples, with proportional representation of any covariate group (for example, batch). The model is evaluated using the average relative error, mean squared error (MSE), and average root MSE[51]. We compared generalized additive models fitted with different splines (cubic, thin plate, and duchon), and chose the model with minimal CV and prediction errors. Automatic spline parameter estimation from 'gam' function was used, unless the model was clearly performing poorly with automatic parameter estimation (overfitting, predictions not matching the component dynamics), in which case we performed further CV on reasonable spline parameter spaces to tweak the model (defining a number of knots). We further verified that RAPToR age estimates match chronological age of the original reference data and of independent time series when staged on the interpolated reference, using the R^2 of linear models (Supplementary Note 1).

The number of components to fit was selected by setting a cutoff on cumulative explained variance (for example, 99%). The cutoff was adjusted according to the number of components with intelligible dynamics with respect to time (Supplementary Note 1 and Supplementary Fig. 6). Interpolation (and subsequent staging) is robust to variation in the number of components used (Supplementary Note 1).

We implemented reference interpolation with the 'ge_im' function in the 'RAPToR' package. Model formulas and parameters for building all the references used in this study are displayed in Supplementary Table 1.

Age estimation. To perform age estimation, we implemented the 'ae' function that takes the gene-expression matrix to stage (genes as rows, samples as columns), the reference matrix (genes as rows and time points as columns), and the reference times (time values associated with the columns of the reference matrix) as inputs. The 'ae' function then finds common genes between sample and reference and computes the Spearman correlation between each sample and each reference time point. The age estimate for each sample is simply the reference time point with the highest correlation.

When an age estimate lands within 5% of the reference edges, RAPToR suggests to stage the samples on another appropriate reference if possible.

Confidence intervals on age estimates are computed by repeated staging on bootstrap gene sets of default size of one third of the total. Unless stated otherwise, the number of bootstraps is 30. A confidence interval is given by the median absolute deviation (MAD) of bootstrap estimates (est_{boot}) from the global estimate (est), and the resolution of the interpolation (res, time interval between two points of the interpolated reference): $[\text{est} - (\text{median}(|\text{est} - \text{est}_{\text{boot}}|) + \text{res}/2);\ \text{est} + (\text{median}(|\text{est} - \text{est}_{\text{boot}}|) + \text{res}/2)]$.

Staging using a prior probability. We implemented the possibility of providing a prior probability in the form of parameters for a Gaussian distribution per sample (mean, standard deviation) which must be given in the time scale of the reference. A Gaussian density function over the reference time is defined per sample from these parameters. During staging, all correlation peaks of the profile are determined and ranked by averaging their scaled correlation score (height of the peak in the correlation profile scaled to [0, 1]) and prior score (value of the Gaussian density function scaled to [0, 1], at the peak time point). The first peak of the ranking is then kept as the estimate. Since the ranking is determined by averaging normalized priors and correlation scores, changing the prior standard deviation parameter results in scaling the importance of the prior with respect to the correlation information.

No priors were used for staging unless explicitly stated.

Evaluating RAPToR performance. Staging *C. elegans* larval development. We built the reference from a time series of WT larval development at 20 °C sampled at 26 time points from L1 feeding to 48 h[20] (Supplementary Table 1), we set the number of interpolated time points to 500.

Staged samples are WT *C. elegans* collected during mid to late larval development at 25 °C from 22 to 37 h after L1 feeding[26]. Only samples aged below 32 h (corresponding to about 48 h at 20 °C) were staged, to stay within the reference boundaries.

Staging D. melanogaster embryonic development. We staged a *Drosophila* developmental time series[27] on an interpolated reference from another embryo developmental time series[25] (Dme_embryo reference of the drosoRef package; Supplementary Table 1). Samples were discarded when the 99th percentile of the distribution of their Spearman correlation coefficients with others samples fell below 0.6, leaving 90 samples to stage. The number of interpolated time points in the reference was set to 500.

We compared our rankings with the BLIND[19] rankings provided in the supplementary data[27] (restricting to 77 samples as the authors used a more stringent quality filter).

To test if our age estimates capture physiological development better than chronological age, we fit identical linear models using the 'lmFit' function of 'limma' with either chronological age or RAPToR estimates as the predictor. Age is modeled using a natural cubic spline with two to eight degrees of freedom (built with the ns function of the splines package). For each gene, we use R^2 to compare the goodness of fit of the models with chronological age or RAPToR age estimates.

Staging D. rerio embryonic development. We used the interpolated reference we built from embryo and larval development data[23] (Dre_emb_larv reference of the zebraRef package; Supplementary Table 1) to stage a zebrafish time series of embryonic development from fertilization to 72 h post-fertilization[27]. Samples were discarded when the 99th percentile of the distribution of their Spearman correlation coefficients with others samples fell below 0.6, leaving 93 samples. The number of interpolated time points in the reference was set to 1,000.

We then used the same reference, increasing the interpolation resolution between 0 and 15 h to 800 time points (resulting in a reference time density of around one time point per minute instead of the previous one time point per hour) to stage an additional dense embryonic time series of 180 zebrafish embryos around gastrula[28]. We compare RAPToR staging to rankings (Extended Data Fig. 2a) previously determined[28] as following: the ten youngest and oldest embryos (determined through the morphological criterion of epiboly coverage) are used to select the genes with the largest decrease in expression from start to end of the time series. The average expression of these genes then determines the ranking. To show the benefit of reference interpolation, we also staged the embryos on the non-interpolated reference time series (Extended Data Fig. 2c,d).

Staging M. musculus embryonic development. We used the interpolated reference we built from mouse embryonic development time series data[24] (Mmu_embryo reference of the mouseRef package; Supplementary Table 1) to stage an independent mouse somite-staged developmental time course[29]. The number of interpolated time points was set to 500. We compare RAPToR staging with the provided embryo somite number as no chronological age is given[29].

Staging M. musculus first-molar embryonic development. First and second data replicates for mouse first molar embryonic development are from Pantalacci et al.[37], and Sémon et al.[38], respectively. Reads from both replicates were processed together, trimmed with trimmomatic[52] (v.0.39) to remove adapters, and mapped using salmon[53] (v.0.14.1) and the Ensembl 98 version of the mouse transcriptome to obtain TPM values.

Genes with a median expression of log(TPM + 1) < 0.5 across all samples were filtered out, leaving 15,362 genes. A reference was built from both replicates of the lower jaw samples (Supplementary Table 1) and used to stage all 32 samples.

Estimating developmental speed factors and resolution increase factors. Developmental speed factors and R^2 between chronological and estimated age of samples are estimated with linear models.

We call 'resolution increase factor' the factor between sampling frequencies of a reference before interpolation and of a successfully staged independent time series.

C. elegans larval development is sampled every 2 h at 20 °C (0.5/h) in the reference[20] and every hour at 25 °C (1/h, 1.5 development speed factor) in the staged time series[26] resulting in a resolution increase factor rf = (1.5 × 1)/0.5 = 3.

Drosophila embryo development is sampled every 2 h (0.5/h) in the reference[25] and every 15 min (4/h) in the staged time series[27], resulting in a resolution increase factor rf = 4/0.5 = 8.

Mouse embryo development is sampled every 1.5 d (0.66/d) in the reference[24] and somite-staged in the target time series[29]. Since the first 30 somites of *M. musculus* grow in ~2.5 d[54], the somite-staged times series has a resolution of 12 time points per day (12/d) determining a resolution increase factor rf = 12/0.66 = 18.2.

Zebrafish embryo development is sampled every hour (1/h) in the reference[23] and at a rate equivalent to 47 per hour (47/h) in the staged samples[28] (180 samples are roughly evenly staged between 5.7 and 9.5 h post-fertilization: 180/(9.5 − 5.7) = 47/h), resulting in a resolution increase factor rf = 47.

Building ageing references. To build RAPToR references capable of staging adults across independent studies, we select genes with monotonous expression along chronological age. Monotonous genes are defined as those with Spearman correlation with chronological age above a threshold given for each dataset. A threshold of $\sqrt{0.33}$ selects genes where approximately a third of expression variance is explained by ageing progression. When less than 200 genes were kept by the filter, we used a more lenient threshold of $\sqrt{(0.25)}$. We then interpolate as described above, using only the first component (which is monotonous given the gene selection).

Staging *C. elegans* ageing. We built a *C. elegans ageing* reference with an unpublished time series (GSE93826) using monotonous genes, defined as those with spearman correlation with chronological age above $\sqrt{0.33}$ (see Supplementary Table 1). The number of interpolated time points was set to 500.

We then staged an RNA-seq bulk time series[33], and a single-worm microarray profiling[34]. Single-worm behavior was provided by the authors[34].

Staging D. melanogaster ageing. Pletcher et al.[35] profiled ageing *Drosophila* in food ad libitum and caloric restriction conditions. We built an ageing reference with the ad libitum food time series using monotonous genes, defined as those with Spearman correlation with chronological age above $\sqrt{0.33}$ (see Supplementary Table 1). The number of interpolated time points was set to 500. We then staged all flies (control and caloric restriction).

Staging human ageing from brain tissue. We removed outliers from Chen et al.[36] expression data when the 99th percentile of the distribution of their Spearman correlation coefficients with other samples fell below median + 2 s.d., and when the reported RNA integrity score was below seven. The remaining 360 samples were split per tissue (BA47 or BA11), and genes with Spearman correlation with chronological age above $\sqrt{0.25}$ were defined as monotonous. For each tissue, half of the samples were randomly selected to build a RAPToR reference with monotonous genes (Supplementary Table 1), on which all samples were then staged. Samples were also staged on the reference built from the other tissue (Supplementary Fig. 8).

Staging H. Sapiens single-cell embryo development. We filtered Petroupoulos et al.[39] single-cell counts to remove genes with a median expression of log(TPM + 1) < 0.5 across all samples, leaving 8,482 genes. Twenty percent of cells were randomly sampled from each of the five time points (305 cells) to build an interpolated reference (Supplementary Table 1), on which all 1,529 cells were then staged (only non-reference cells are shown in Fig. 2l). Cell pseudo-times are from the original study[39], provided as sample metadata in the ArrayExpress entry.

Probing robustness of reference interpolation. Robustness of reference interpolation to the choice of dimensionality reduction method and number of components was evaluated using either the *C. elegans* time series by Kim et al.[20] (as above), or the one by Meeuse et al.[21] as references.

Robustness was evaluated computing sum squared (SSQ) of gene-expression prediction error by reference models using PCA or ICA and 2–16 components with the Kim et al. time series, and 2–20 with the Meeuse et al. time series. The model formula was fixed to the one defined in Supplementary Table 1. The SSQ prediction error is defined as $\mathrm{SSQerror} = \frac{(\boldsymbol{X}_{n\times m} - \hat{\boldsymbol{X}}_{\mathrm{pred}})^2}{\mathbf{n}\times\mathbf{m}}$, with **n** samples, **m** genes.

For six conditions—ICA/PCA, each at three different numbers of components—we staged the reference samples as well as an independent *C. elegans* time series[26] on the interpolated reference (only samples within reference boundaries were staged on the Kim et al. reference). We evaluated models built from 4, 9, and 14 PCA or ICA components for the Kim et al. reference and models built from 10, 20 and 25 PCA or ICA components for the Meeuse et al. reference.

We then reported the R^2 value of a linear fit of RAPToR estimates by the chronological age of the samples in each condition (Supplementary Table 2), as well as correlation scores between samples and the interpolated reference at the estimate (Supplementary Fig. 5).

Estimating the impact of gene-set size on staging. The impact of gene-set size on staging was evaluated by staging the *C. elegans* larval time series by Hendriks et al.[26] on the reference built from the Kim et al.[20] samples, as above.

We staged the samples using 50 random gene sets of sizes 16,000, 12,000, 8,000, 4,000, 2,000, and 1,000. The resulting estimates were used to compute confidence intervals for varying bootstrap set sizes. We reported the median absolute deviation of estimates to the full gene set estimate plus interpolation resolution (that is, the size of half the confidence interval).

The same approach was repeated for smaller gene-set sizes of 2,000, 1,000, 500, and 250, this time staging the samples with and without priors (defined as 1.5 times the chronological age of the samples to account for the developmental speed difference with the reference; prior standard deviation was set to 10).

Tissue-specific staging and quantification of soma–germline heterochrony. Two hundred and eight RILs from a cross between N2 (Bristol) and CB4856 (Hawaii) strains of *C. elegans* were genotyped at 1,455 single-nucleotide polymorphism markers, and collected as young-adult hermaphrodites (originally intended as one time point) for profiling by microarray with one sample per RIL[11].

Microarray intensities were first normalized within arrays with LOESS using the 'normalizeWithinArrays' function of the 'limma' library. Arrays corresponding to pooled mixed stage controls were then discarded. Samples were discarded when the 99th percentile of the distribution of their Spearman correlation coefficients with other samples fell below 0.95, leaving 193 samples for analysis.

We staged the samples with the 'Cel_larv_YA' reference[21] of the wormRef package (Supplementary Table 1), using 1,000 interpolated time points.

Samples were first staged using the entire available gene set to obtain the global estimates, then with soma and germline-specific gene sets to obtain the corresponding tissue-specific estimates: the soma gene set corresponds to the oscillatory genes denoted 'osc' in Hendriks et al.[26]. The 'germline' gene set corresponds to the union of 'germline_intrinsic', 'spermatogenesis_enriched', and 'oogenesis_enriched' gene sets defined in Reinke et al.[26]. Estimating soma age required the use of the global estimate as prior (owing to oscillations in gene expression generating multiple correlation peaks), with the prior standard deviation set to ten for all samples. Germline age estimates required no prior.

To compare expression dynamics between reference and RILs, we kept the overlapping genes between the non-interpolated reference and the samples, quantile normalized both datasets together, and performed an ICA ('ica' function of 'icafast') extracting 46 components, explaining 95% of the variance in the joined data. For components 2–8, capturing developmental signal (IC1 captured batch effect, IC9 captured genetic variation), we defined contributing genes as those with an absolute loading above 1.96. We then tested for enrichment of soma, oogenesis, and spermatogenesis categories in these genes with a two-sided hypergeometric test, the *P* values of which were adjusted across all tests with the Benjamini–Hochberg method.

To test heterochrony between RILs and the reference, we fit splines on the reference samples in IC2–IC8 (using the same model as the RAPToR reference; Supplementary Table 1) and computed root mean square error between the fit and RILs for each component. This was done for global, soma and germline age estimates of RILs (Fig. 3g), and for shifted values of global age estimates (−5 h to +5 h; Extended Data Fig. 4c).

To test the existence of heterochrony among the RILs, we fit identical models on the RIL expression data using the 'lmFit' function in limma with global, soma, or germline age values as predictors. We used natural cubic splines ('ns' function in the 'splines' library) on the age with four, six, or eight degrees of freedom. Choice between models (at equal spline degrees of freedom) was done per gene on the basis of highest R^2 value.

QTL analysis on soma–germline heterochrony. The multivariate QTL analysis on soma–germline heterochrony among RILs defined as (soma age) − (germline age) was performed by random forest regression[55] with or without batch as a covariate. Each RIL was genotyped at 1,455 single-nucleotide polymorphism markers[11]. Redundant markers were filtered out from the selected 193 RILs, missing values for the remaining 1,105 markers are imputed with the 'rfImpute' function and random forest regression was fit with 5,000 trees using the 'randomForest' function; both functions are from the 'randomForest' package (v.4.6.14). The random forest selection frequency was used as importance measure, adjusted for selection bias[55], which was estimated by fitting 500 forests of 10 trees to Gaussian noise.

We estimated the null probability distribution of random forest selection frequency through 100 trait permutations, calculated empirical *P* values and adjusted them for FDR.

Cross-species staging. Staging non-model *Drosophila* on *D. melanogaster*. We used the interpolated reference we built from *D. melanogaster* embryo development data[25] (Dme_embryo reference of the drosoRef package; Supplementary Table 1) to stage developmental time series of six *Drosophila* species[41]: *D. melanogaster*, *D. simulans*, *D. ananassae*, *D. pseudoobscura*, *D. permisilis* and *D. virilis* profiled by microarrays. We used orthologs provided by the authors[41]. The number of interpolated time points in the reference was set to 500.

Developmental speed difference from *D. melanogaster* was determined with a linear model without intercept predicting RAPToR estimates with the chronological age of samples, with species as covariate and including interaction. A comparison with the original scaling factors[41] is shown in Supplementary Table 3.

To determine whether RAPToR estimates or the linearly scaled age from the study[41] is the better development indicator, we fit identical linear models on gene expression (lmFit function of limma) with either value as the predictor, and species as covariate. Age is modeled using a natural cubic spline with two to eight degrees of freedom (ns function of splines). For each gene, we use R^2 to compare the goodness of fit of either model (Supplementary Fig. 10). No interaction between age and species coefficients was considered as temporal scaling of development between species is already applied.

We evaluated the effect of species distance on staging through the maximal correlation score between the samples and the reference (that is, at their age estimate).

Staging C. elegans on *Drosophila*. We staged a *C. elegans* embryo time series[27] on the interpolated reference we built from the *D. melanogaster* embryo development time series[25] ('Dme_embryo' reference, drosoRef package). First, poor-quality *C. elegans* samples were discarded when the 99th percentile of the distribution of their Spearman correlation coefficients with other samples fell below 0.67. Additionally, a sample (GSM1487346, or 'sample_0029') was also excluded as it clearly appeared as an outlier on multiple ICA components (Extended Data Fig. 8). Four samples (GSM1487318, GSM1487319, GSM1487320, and GSM1487321, or 'sample_0001' to '_0004') were further removed owing to erroneous chronological age (Extended Data Fig. 8), leaving 127 samples.

We then performed the staging using a restricted worm–fly ortholog set[45]. We also did staging on a second reference interpolated as above but using the first two instead of eight components. For both references, the number of interpolated time points was set to 500.

Further analysis is restricted to the overlapping set of orthologs between worm and fly datasets (3,194 genes). We ranked genes by Spearman correlation between the *C. elegans* embryo time series and their matching timepoints in the second *D. melanogaster* reference. We then selected the 10% of genes with highest correlation (319 genes) and staged the *C. elegans* samples once more on the second *D. melanogaster* reference, evaluating staging performance with Spearman correlation and the R^2 of a linear model between chronological age and estimated age.

Hierarchical clustering of the top 10% genes in the original *D. melanogaster* reference data[25] ('hclust' function on the euclidean distance matrix of gene-centered log(TPM + 1)), resulted in 3 clusters with over 20 genes. We then evaluated Gene Ontology enrichment in each cluster with gProfiler[56] using the 3,194 overlapping set of worm–fly orthologs as background (Supplementary Table 4–6).

Staging M. musculus single cells on *H. sapiens* single cells. We filtered Deng et al.[42] mouse early embryo single-cell counts to remove genes with a median expression of log(TPM + 1) < 0.5 across all samples, leaving 6,506 genes. We then staged cells using all available mouse–human orthologs from ensembl (v.99), on the interpolated reference we built from *H. sapiens* embryo development single cells[39] (see above; Supplementary Table 1). We compared age estimates with chronological age, as well as with pseudo-time rankings similar to Petropoulos et al.[39] (Extended Data Fig. 7). We fit a principal curve (principal_curve function of princurve library) on the first three components of a PCA on the 1,000 most variable genes.

As done above for worm–fly staging, we then restricted further analysis to the overlapping set of orthologs between mouse and human datasets (6,057 genes), ranked genes to select the 10% of genes with highest correlation between both species (509 genes), and staged the *M. musculus* single cells once more on the human reference, evaluating staging performance with Spearman correlation and the R^2 of a linear model between chronological and estimated age.

Hierarchical clustering of the top 10% of genes in the original *H. sapiens* reference data (expression matrix aggregated per sampled time point, gene-centered log(TPM + 1)), resulted in 5 clusters with over 20 genes. We then evaluated Gene Ontology enrichment in each cluster with gProfiler[56] using the 6,057 overlapping set of mouse–human orthologs as background (Supplementary Tables 7–11).

Staging H. sapiens on *B. taurus*. We filtered an RNA-seq cow early embryo time series[44] to keep genes with median log(TPM + 1) expression >0, and built a reference with half of the samples. We similarly kept genes of a microarray human embryo time series[43] with log(X + 1) expression >2 and built a reference with half the samples (Supplementary Table 1). As only morphological stages (for example, two-cell, four-cell) and not timings were given in both datasets, we used timings from the literature[57]. The number of interpolated time points for both references was set to 100. We then staged all cow samples on the human reference and vice versa, using all available human–cow orthologs from ensembl (v.99) (Supplementary Fig. 11).

Exploiting RAPToR age estimates. *Drug dose response on developmental delay in C. elegans.* Expression profiles of young *C. elegans* adults exposed to drugs[46] were staged on the "Cel_larv_YA" reference[21] from the wormRef package (Supplementary Table 1), with 500 interpolated time points in the reference. We estimated global, soma-, and germline-specific ages (see *Tissue-specific staging*). For each age type, we then subtracted the age of the control sample within each replicate of each drug assay to compute the developmental difference by treatment group. We fit a linear model with drug, dose, and interaction on the age differences to assess the significance of the effects.

Increasing statistical power in differential expression analyses. WT and *pash-1*ts *C. elegans* samples[47] were staged on the 'Cel_YA_2' reference[22] from the wormRef package (Supplementary Table 1), with 500 interpolated time points in the reference. The second replicate of the first WT time point (wt_h0.2) was omitted from further analysis owing to its extreme developmental displacement and lack of comparable mutant sample.

We fit identical linear models with the 'lmFit' function in the 'limma' library to test for differential expression, including either chronological or estimated age modeled with a natural cubic spline ('ns' function in 'splines', degree of freedom = 2), strain, and their interaction.

Effect of strain and development was then assessed by considering the significance of appropriate model coefficients (interaction and strain coefficients for strain effect, spline and interaction coefficients for development effect), with the 'topTable' function in the 'limma' library. Differential expression was considered significant at 0.05 Benjamini–Hochberg FDR.

To test the effect of similar random age differences from chronological age, we generated 100 'random age' sets by sampling age differences from the distribution of (chronological age) − (estimated age) values, estimated with the 'density' function in R. Sampled age differences were then added to the chronological age, and the same model and analysis as above was applied. The goodness-of-fit per gene is assessed using R^2.

Quantifying developmentally driven changes in gene expression. Given any two groups of expression profiling samples 'A' and 'B', we first stage them, then fit a linear model per gene on $\log_2$(TPM + 1) (or $\log_2$(intensity + 1) for microarray data) to compute the observed $\log_2$(fold changes) (logFCs) of 'A' versus 'B' samples. Then we fit the same model on reference profiles at matching time points to compute logFCs expected from development only (Extended Data Fig. 9). We use squared Pearson correlation between observed and expected logFCs to quantify the variance explained by development in the observed logFC.

Control and post-dauer *C. elegans* samples[48] were germline-staged (see above) on the 'Cel_larv_YA' reference[21], and on the 'Cel_YA_2' reference[22] of the wormRef package for confirmation, as they landed near the edges of the first reference. The number of interpolated time points in the Cel_larv_YA and Cel_YA_2 references were set to 1,000 and 500 respectively. Using the method described above, we quantified the differential expression explained only by difference in developmental stages between the control and post-dauer samples.

We could not compare our results to the original results as we were unable to exactly reproduce the distribution of differential expression and P values of the original t-test analysis. We therefore recalculated differential gene expression using linear models (function 'lmFit' in 'limma' library in R).

Recovering direct perturbation effects using reference data. WT and *xrn-2* time series of *C. elegans* late-larval development[49] were staged on the 'Cel_larv_YA' reference[21] from the wormRef package (Supplementary Table 1), with 500 interpolated time points. We restricted further analysis to the genes with both at least five raw counts for at least one sample, and overlapping with the reference gene set (17,656 genes).

Defining the differential expression gold standard. To establish the gold standard of differentially expressed genes, we selected time points 8–10 of *xrn-2* and WT, as they had the best (estimated) developmental match. We then calculate differential expression fitting a generalized linear model (GLM) on raw counts using the glmFit function of egdeR (v.3.36.0), including only the strain variable (model_1), and considered genes differentially expressed with Benjamini–Hochberg adjusted P values < 0.05 of a likelihood ratio test (glmLRT function of 'edgeR') on the strain coefficient.

Evaluating gold-standard gene detection decrease with age gap. To test how increasing mismatch in developmental time between *xrn-2* and WT impacts differential expression analysis we apply the same GLM used for the gold standard (model_1) to calculate differential expression between the mutant and WT samples shifted by −1, −2, −3, −5, and −7 time points and we estimated expression changes explained by development as detailed above. We then evaluated how well model_1 P values detect gold-standard differentially expressed genes at increasing age gaps by precision–recall (PR) curves and area under PR curves (AUPRC) using the 'prediction' function of the 'ROCR' package (v.1.0.11).

Correcting expression changes from development. To accurately account for developmental changes we combine the samples of interest with the interpolated reference.

For each set of samples (including WT and mutant samples), we define the reference window to include as the range of age estimates widened by a 1-h margin on either side. For example, age estimates of the 'WT-1' set range 51.7–58.3 h. Thus, we include a 50.7–59.3 h window of interpolated reference.

We transform the interpolated reference data to artificial counts assuming a fixed library size of 25×10^6 counts per sample and a fixed number of reads 'per gene length' defined by the median of available gene lengths:

$$\text{ArtificialCounts} = (\text{interpolatedTPM}/(10^6))\times((25\times10^6)/\text{median}(\text{geneLengths}))\times\text{geneLengths}$$

The artificial count matrix is then joined to the sample count matrix and a GLM is fit ('*glmFit*' in '*edgeR*"), including batch (between reference and sample data), the variable of interest (strain) where reference data is grouped together with the control, and developmental time modeled with splines ('ns' function in 'splines'). To select the optimal spline degree of freedom for each window, we minimized the residual SSQ of a linear model fit on the reference window only (Extended Data Fig. 10h). Only model coefficients of the variable of interest (strain logFCs) are considered.

We first evaluated how well strain logFCs detects differentially expressed genes from the gold standard using PR curves and AUPRC ('prediction' function in 'ROCR'). We then defined an age-corrected classifier (ACC) as the weighted mean of the model_1 *P* value and strain logFC of the model including the reference:

$$\text{ACC} = w\times\text{strainLogFC} + (1-w)\times(-\log_{10}(\text{model\_1Pval}))$$

with w, the weight ratio of either classifier. We defined the optimal w as the value for which the AUPRC is maximal, and estimated it for each set of WT shifts. At optimal w, we then reported the AUPRC of our ACC and compared it to the standard model.

As the optimal w cannot usually be estimated in this way, we explored the relationship between optimal w and observed/expected logFC correlation (as defined above) calculated for a larger amount of WT three-sample sets (Supplementary Table 13).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data for all figures is provided. Source data are provided with this paper.

Code availability

The code to download and (pre)process the data, perform the analyses and generate the figures of this paper can be found at https://gitbio.ens-lyon.fr/LBMC/qrg/raptor-analysis

References

50. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
51. Borchani, H., Varando, G., Bielza, C. & Larrañaga, P. A survey on multi-output regression. *WIRES Data Min. Knowl.* **5**, 216–233 (2015).
52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
53. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
54. Hill, M. A. Mouse Stages. *Embryology* https://embryology.med.unsw.edu.au/embryology/index.php/Main_Page (2022).
55. Michaelson, J. J., Alberts, R., Schughart, K. & Beyer, A. Data-driven assessment of eQTL mapping methods. *BMC Genomics* **11**, 502 (2010).
56. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
57. Llobat, L. Pluripotency and growth factors in early embryonic development of mammals: a comparative approach. *Vet. Sci.* **8**, 78 (2021).

Acknowledgements

We are grateful to S. E. Hall, M. Sémon, and S. Pantalacci for providing data from their profiling experiments. We are also grateful to G. Yvert, D. Jost, M. Sémon, A. Piazza, S. Pantalacci, and B. Lehner for their critical reading of the manuscript. M.F. is supported by INSERM. Work in the laboratory of M.F. is supported by a grant from the Agence Nationale pour la Recherche (ANR-19-CE12-0009 'InterPhero'), Université de Lyon (IDEX IMPULSION G19002CC) and ENS-Lyon (Projet emergent 2019). R.B. PhD fellowship is funded by the French Ministry of Research.

Author contributions

M.F. and R.B. conceived the method; R.B. developed the computational framework and performed the analyses; and M.F. and R.B. wrote the manuscript.

Competing interests

The authors report no competing interests.

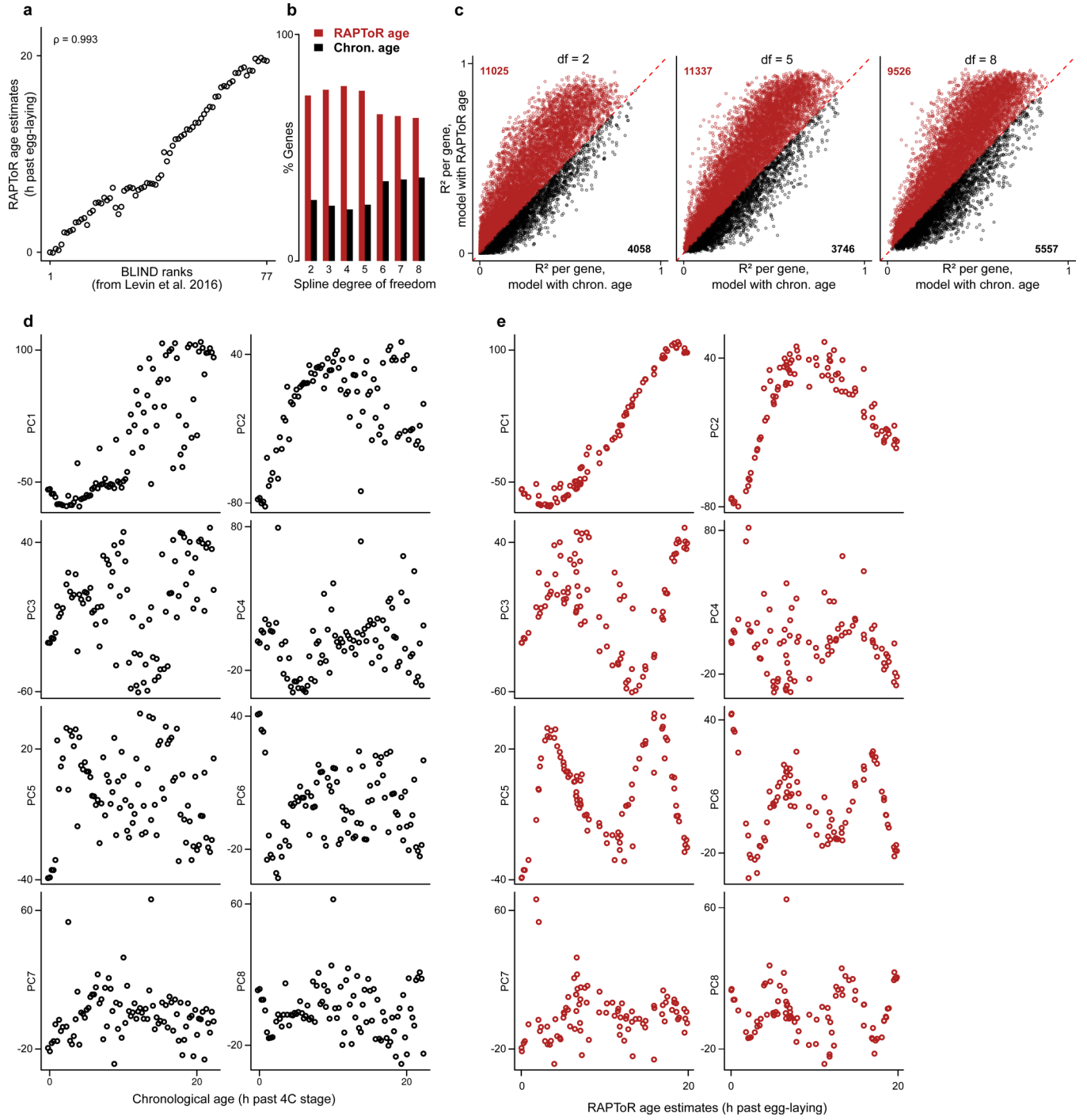
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01540-0.

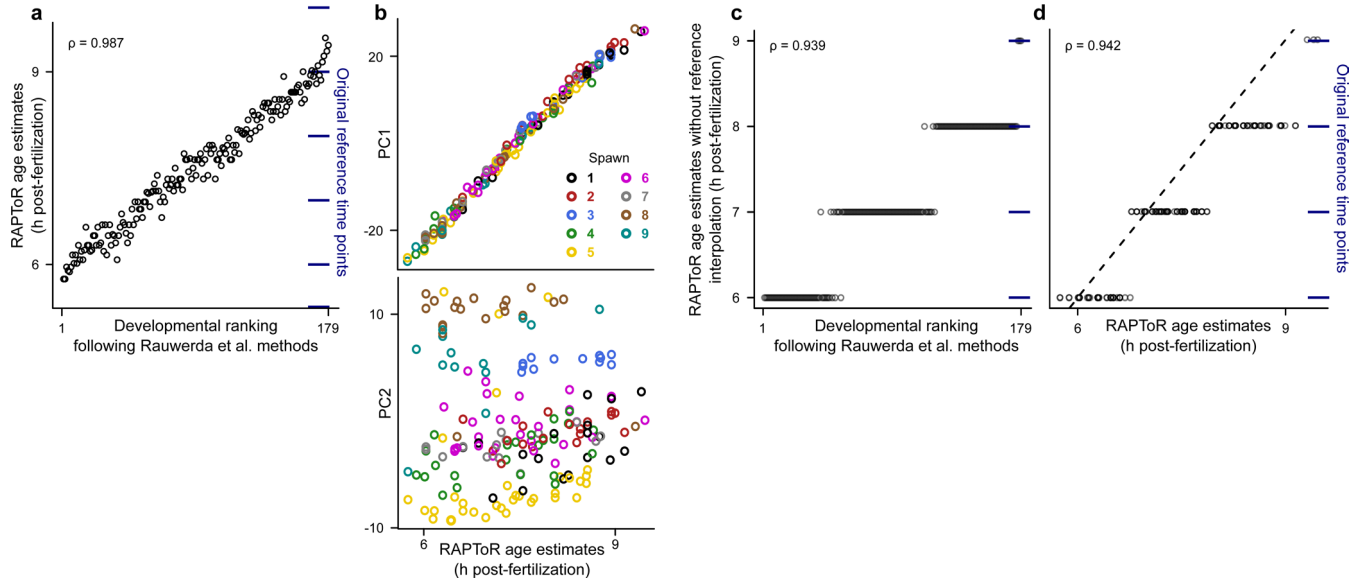
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01540-0.

Correspondence and requests for materials should be addressed to Mirko Francesconi.

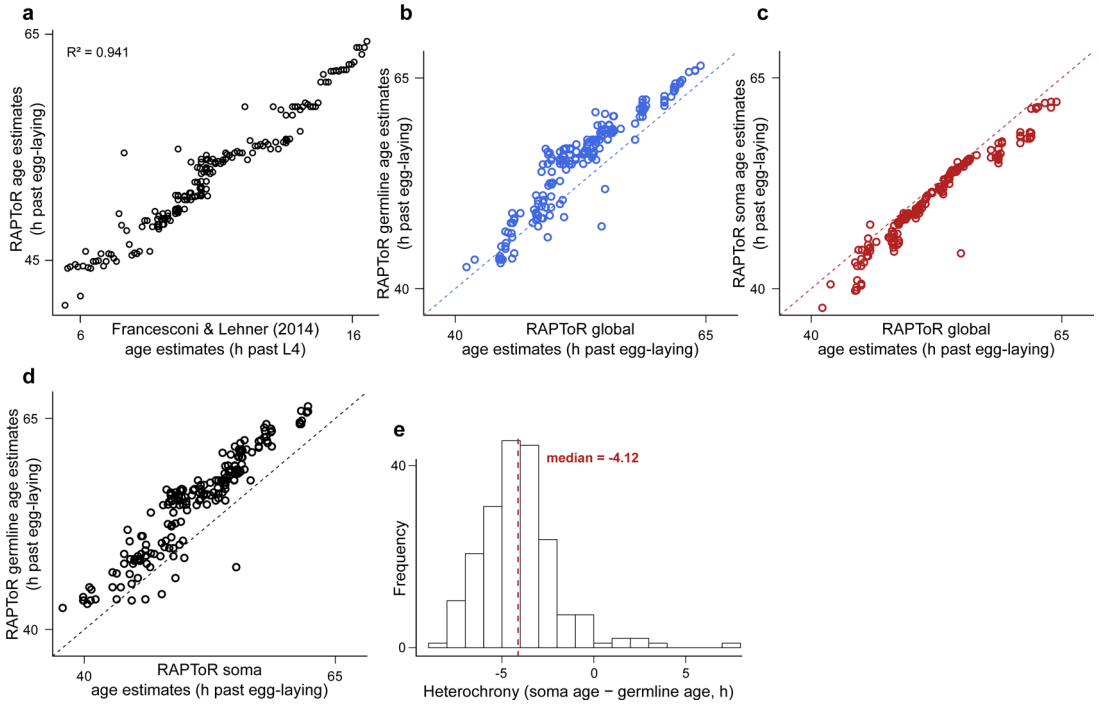
Peer review information *Nature Methods* thanks Helge Grosshans, Adam Alexander Thil Smith and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.



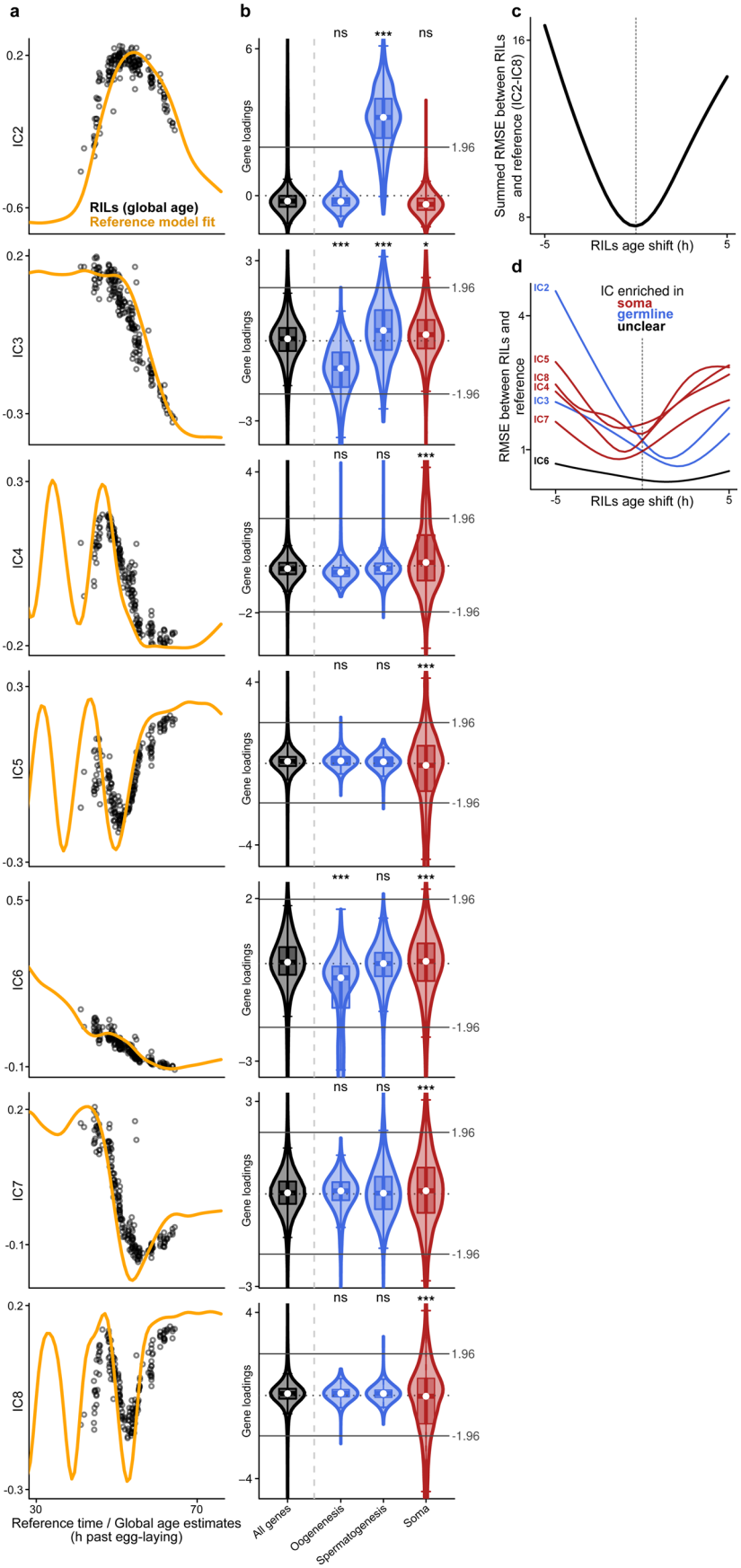
Extended Data Fig. 1 | RAPToR estimates fit gene expression data better than chronological age. a, RAPToR estimates of *D. melanogaster* single-embryo samples[27] staged on a reference built from bulk data[25] plotted against established BLIND ranks[27]. **b**, Percentage of genes better fitted by either RAPToR estimates or chronological age modeled with splines using 2-8 degrees of freedom in otherwise identical models. **c**, R^2 of models from (**b**) gene count in each half of the plot is indicated in the corners. **d,e**, Principal components plotted along chronological age (**d**), and RAPToR estimates (**e**) (as in Fig. 2d-f).



Extended Data Fig. 2 | Reference interpolation allows RAPToR estimates at high resolution. a, RAPToR estimates of a zebrafish embryonic time-series from 9 spawns[28] staged on a reference built from Domazet et al. data[23] plotted against original developmental ranks[28]. **b**, First 2 principal components of the zebrafish time-series plotted against RAPToR age estimates. Spawns are color-coded. **c,d**, RAPToR estimates of the zebrafish time-series on the non-interpolated reference (*i.e* the sampling time of the reference sample with the highest correlation) vs. original developmental ranks (**c**) and vs. standard RAPToR estimates (as in **a**) (**d**). In **a,c,d**, original reference time points within the plot area are shown on the right, in blue.

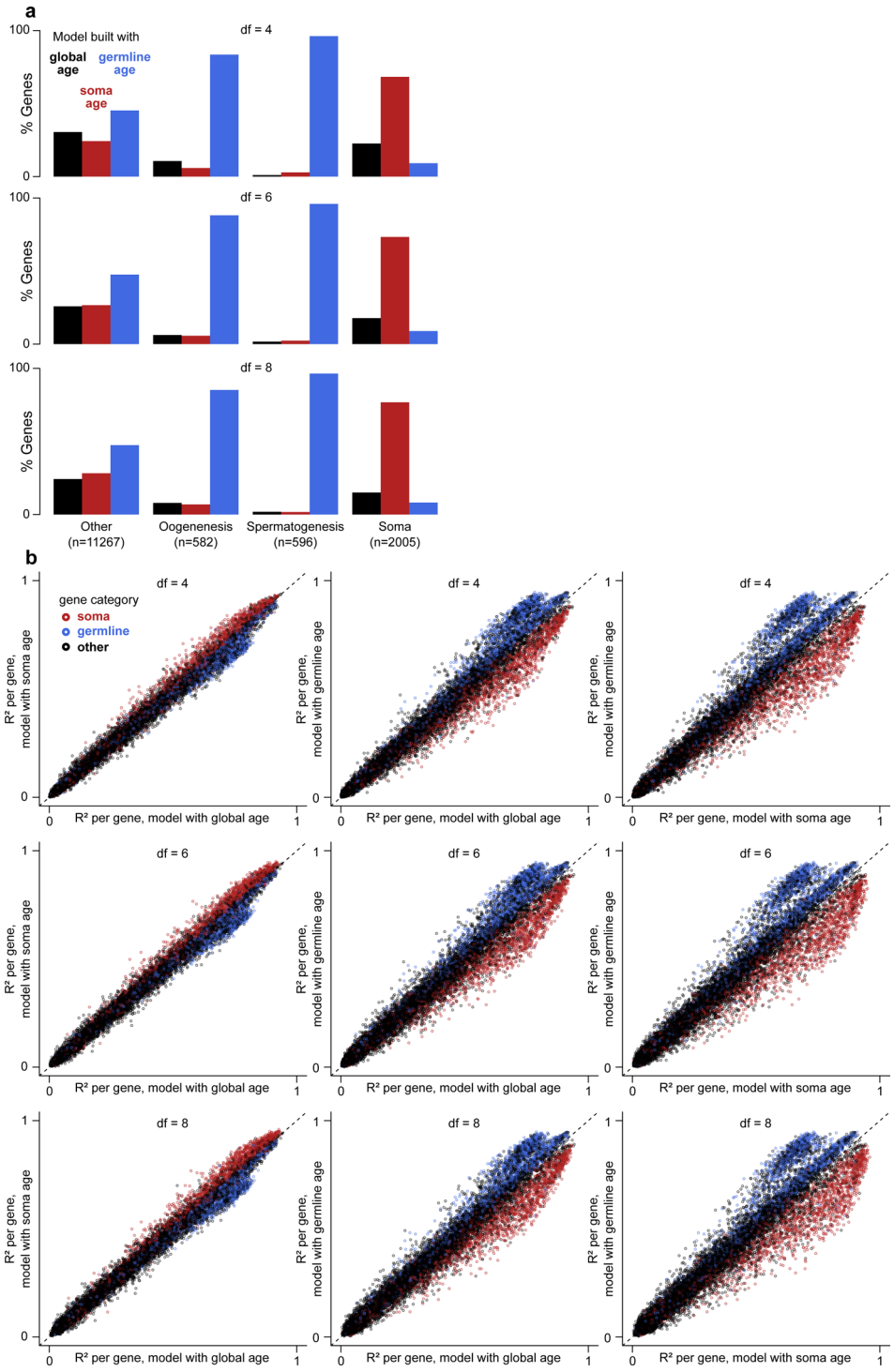


Extended Data Fig. 3 | Tissue-specific staging yields soma and germline ages. a, RAPToR estimates of *C. elegans* Recombinant Inbred Lines (RILs)[11] staged on the larval to young-adult reference built from Meeuse et al.[21] vs. Francesconi & Lehner[12] estimates. **b-d**, Comparison of RAPToR estimates of global age vs. germline age (**b**), global age vs. soma age (**c**), and soma age vs. germline age (**d**). **e**, Distribution of soma–germline heterochrony.

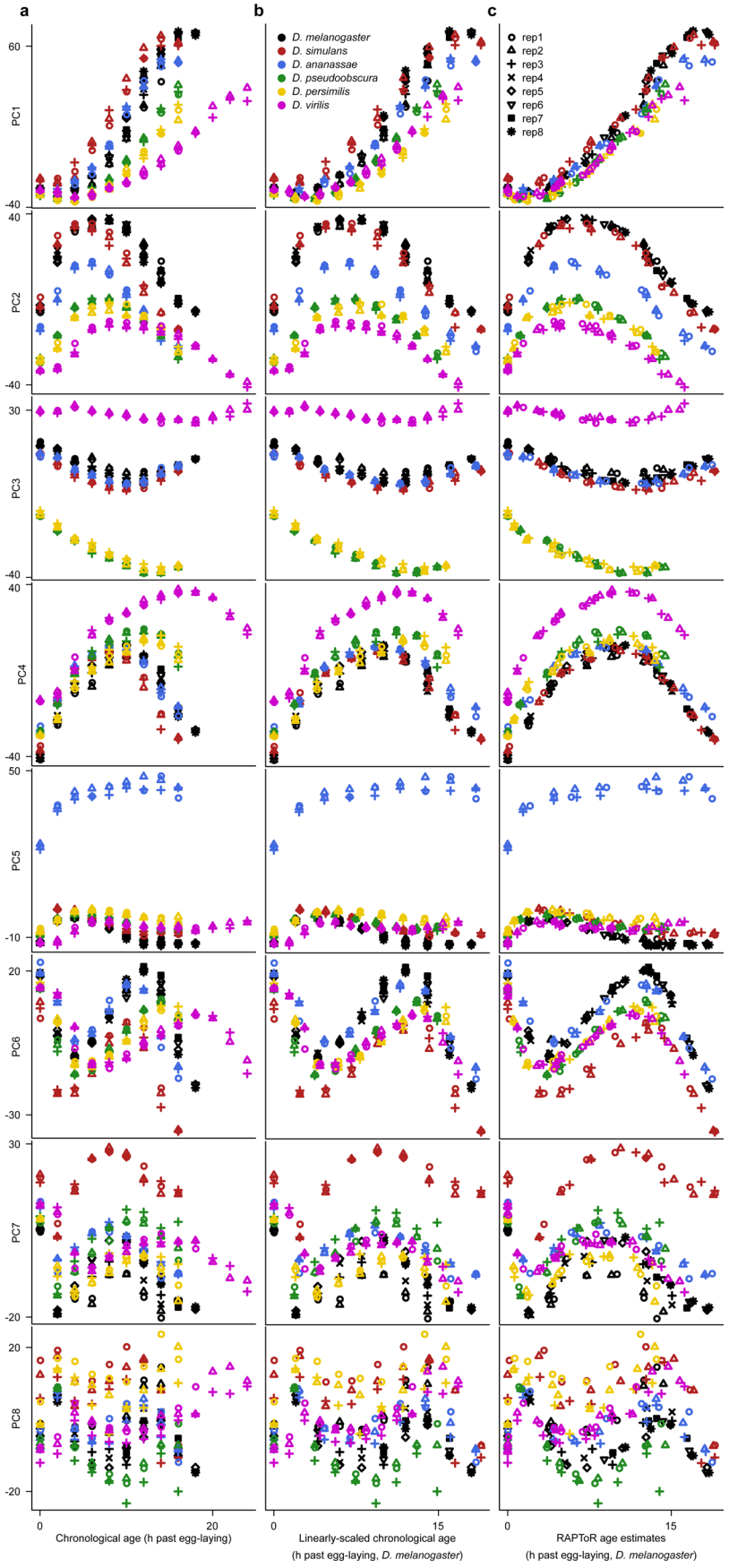


Extended Data Fig. 4 | See next page for caption.

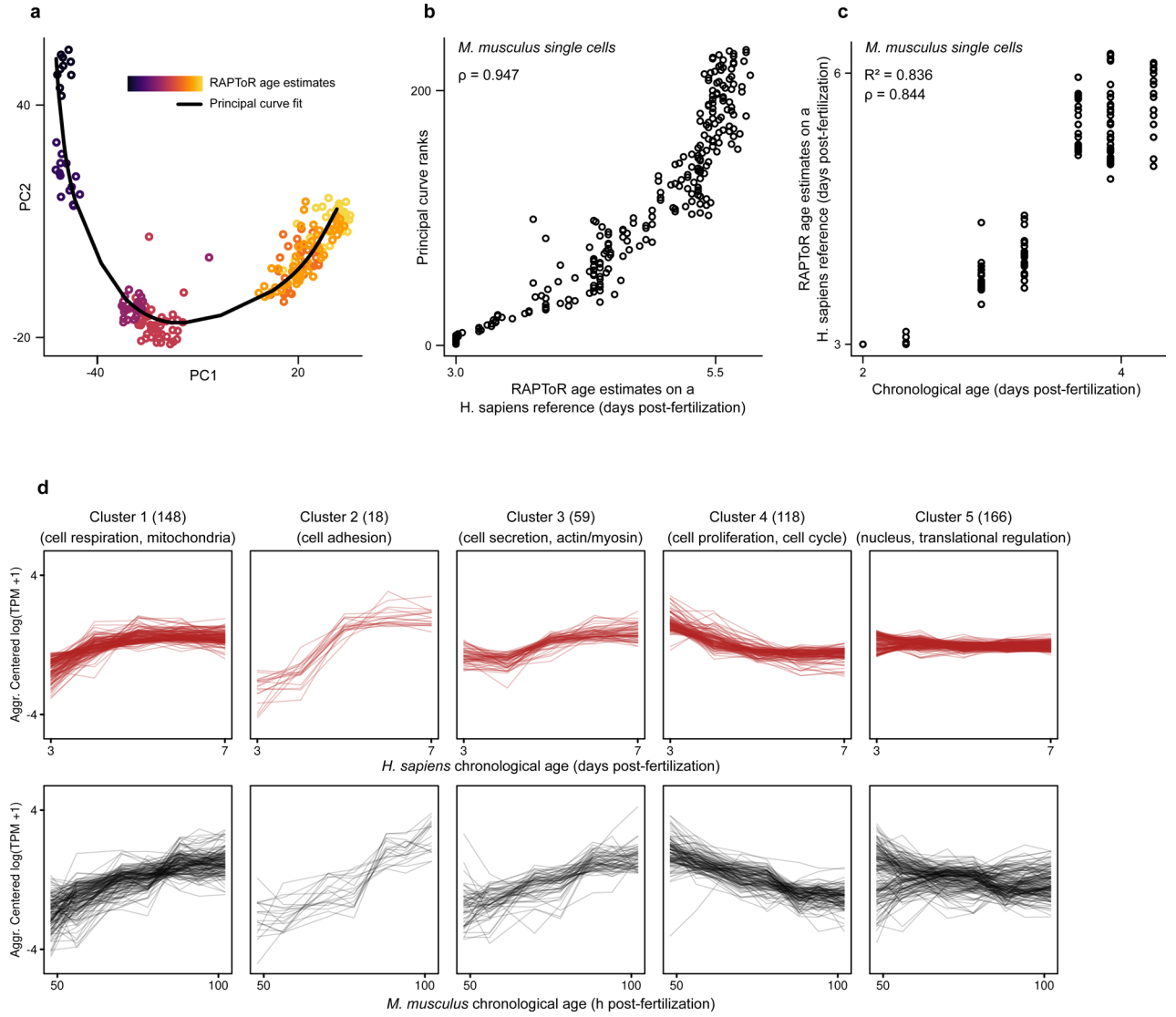
Extended Data Fig. 4 | A delayed germline and an advanced soma. a, Independent Components from ICA on *C. elegans* Recombinant Inbred Lines (RILs)[11] joined to the (non-interpolated) reference data[21] plotted along chronological age and RAPToR global estimates for the reference (orange) and RILs (black) respectively. **b**, Gene loadings on ICA components for all genes (n = 14132), germline genes (oogen. n = 582, sperm. n = 596) and soma (n = 2005) categories. Each box within violins spans the interquartile range (IQR), the central white dot denotes the median, and whiskers extend to 1.5×IQR in either direction. Category enrichment p-values derive from a two-sided hypergeometric test on genes with absolute loadings above 1.96. From left to right, p-values are IC2: $p > 0.99$, $p < 1e\text{-}10$, and $p > 0.99$; IC3: $p < 1e\text{-}10$, $p = 2.66e\text{-}06$, and $p = 0.022$; IC4: $p > 0.99$, $p > 0.99$, and $p < 1e\text{-}10$; IC5: $p > 0.99$, $p > 0.99$, and $p < 1e\text{-}10$; IC6: $p < 1e\text{-}10$, $p > 0.99$, and $p = 6.54e\text{-}04$; IC7: $p > 0.99$, $p > 0.99$, and $p < 1e\text{-}10$; IC8: $p > 0.99$, $p > 0.99$, and $p < 1e\text{-}10$. **c,d**, Summed (**c**) and per-component (**d**) Root Mean Square Error (RMSE) between RILs and reference fit on IC2-IC8 when shifting RIL (global) age estimates. RMSE per-component shows heterochrony, with soma dynamics of RILs matching younger reference time and the reverse for germline dynamics. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



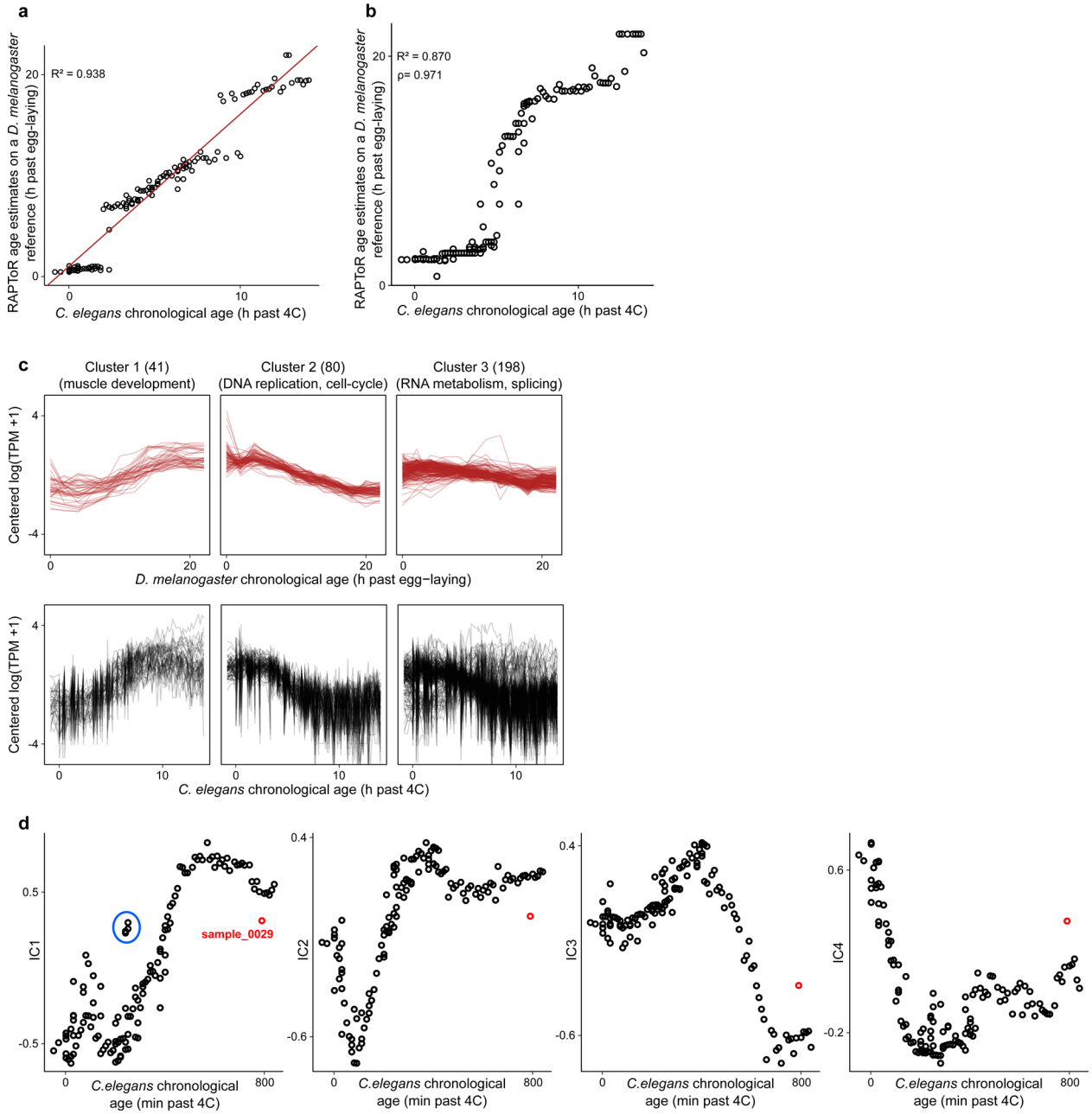
Extended Data Fig. 5 | Soma–germline heterochrony among C. elegans recombinant lines. Recombinant Inbred Lines (RILs)[11] are staged on the larval to young-adult reference built from Meeuse et al. samples[21]. **a**, Percentage of genes better fitted by either RAPToR global, soma, or germline age estimates, modeled with splines with 4, 6, or 8 degrees of freedom in otherwise identical models. Genes are classified into spermatogenesis, oogenesis, somatic, or other (see methods). **b**, R^2 per gene of models with global, soma, or germline age estimates as predictors for 4, 6, and 8 spline degrees of freedom.



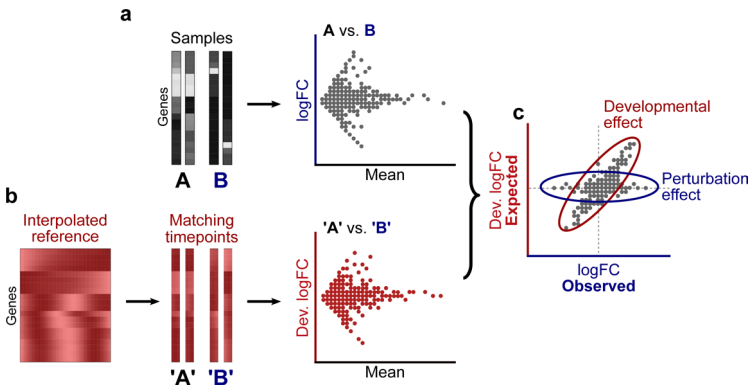
Extended Data Fig. 6 | RAPToR age estimates synchronize expression dynamics across species. a-c, Principal components of *Drosophila* embryogenesis in 6 species[41] plotted along chronological age (**a**), linearly scaled chronological age[41] (**b**), and RAPToR age estimates on a *D. melanogaster* reference[25] (**c**).



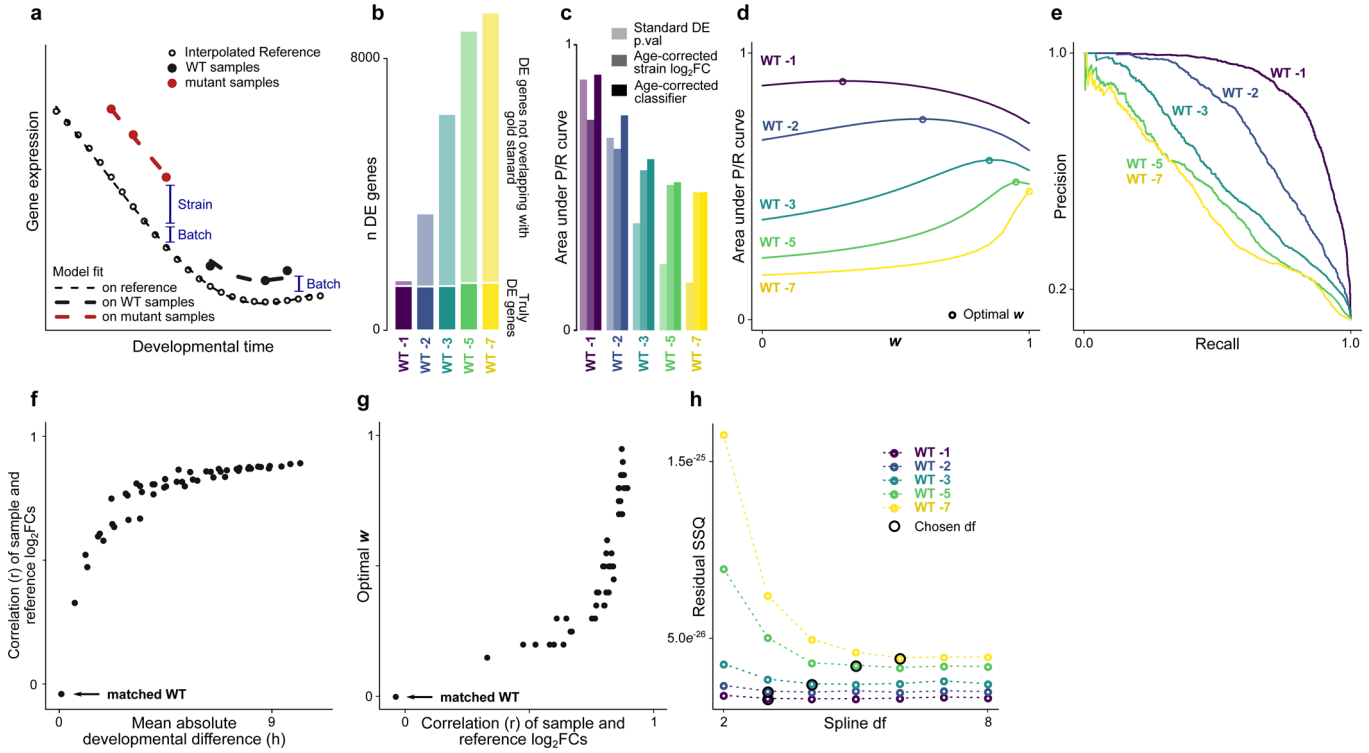
Extended Data Fig. 7 | Staging M. musculus single cells on H. sapiens reference. Single cells from *M. musculus* embryos[42] were staged on a *H. sapiens* single-cell embryogenesis reference[39] using orthologs. **a**, First 2 principal components of a PCA done on the 1000 most variable genes. A principal curve is fit on the first 3 components. Cells are colored by RAPToR age estimate on the *H. sapiens* reference. **b**, RAPToR age estimates of *M. musculus* single cells on *H. sapiens* reference vs. cell ranks along principal curve (a). **c**, Chronological age of *M. musculus* single cells vs. RAPToR age estimates on *H. sapiens* reference using top 10% most correlated genes between mouse and human for staging (see methods). **d**, *H. sapiens* (red) and *M. musculus* (black) clustered gene expression profiles (aggregated per time point) of highest-correlated genes between both species (see methods).



Extended Data Fig. 8 | Staging C. elegans embryogenesis with D. melanogaster. a, *C. elegans* embryo samples from Levin et al.[27] staged on the *D. melanogaster* reference built from Graveley et al.[25] samples. Gaps appear in the estimates, likely at points where fly expression dynamics are incompatible with those of worms. **b**, As in (**a**), staging on the adjusted fly reference and using top 10% most correlated genes between fly and worm embryogenesis (see methods). **c**, *D. melanogaster* (red) and *C. elegans* (black) clustered gene expression profiles of highest-correlated genes between both species (see methods). **d**, ICA components of the *C. elegans* embryo time course plotted along sampling time. Both the red highlighted outlier and 4 samples with erroneous chronological age (circled in IC1) are omitted from analysis (see methods).



Extended Data Fig. 9 | Estimating the impact of development by integrating reference data. a-c, Cartoon detailing how the log-fold-changes (logFCs) of a differential expression analysis between two sample groups (**a**) and the logFCs of their matching time points in the RAPToR interpolated reference (**b**) can be compared to quantify the impact of development (**c**).



Extended Data Fig. 10 | Correcting the effect of development by integrating reference data. Samples from *C. elegans* time-course experiments of wildt-type (WT) and *xrn-2* mutants, profiled by Miki et al.[49], and staged on the larval to young-adult reference built from Meeuse et al. samples[21], are used to validate developmental correction approach (see also Fig. 5f-i). **a**, Cartoon of a model integrating a window of reference data, with Strain and Batch coefficients shown in blue. **b**, Number of DE genes found by a standard differential expression model (FDR < 0.05) increases with the age gaps between compared groups, with a quasi-constant fraction of truly DE genes. **c**, Area under PR curves (AUPRC) in detecting gold-standard DE genes for standard differential expression model p-value, age-corrected logFCs, or the age-corrected classifier for each shifted WT subset. **d**, w parameter optimization for shifted WT sets, by maximizing area under the PR curves. **e**, PR curves of gold-standard gene detection by the age-corrected classifier for each shifted WT subset. **f**, Correlation of expected development logFCs and observed logFCs between the *xrn-2* subset and combinations of 3-sample WT sets (note these are not the "WT -n" subsets, see Supplementary Table 13). **g**, Relationship between optimal w and sample-reference logFC correlation, as in (**f**). **h**, Optimal spline degree-of-freedom (df) selection for the different WT shifted sets by reaching a residual Sum of Square (SSQ) plateau. The selected df increases with the shift, which is expected since the reference window to include gets larger and may thus contain more complex dynamics. DE, Differentially Expressed. logFC, log2 fold-change. FDR, false discovery rate, PR: Precision-Recall.

nature research

Corresponding author(s): Mirko FRANCESCONI

Last updated by author(s): Apr 22, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

R (v4.1.2) packages used for data collection/loading:
- GEOquery (v2.62.2)
- readxl (v1.3.1)
- utils (base)
- biomaRt (v2.50.1)
- affy (v1.72.0)
- gcrma (v2.58.0)
- Biobase (v2.54.0)

Software used for M. musculus tooth count mapping:
- trimmomatic (v0.39)
- salmon (v0.14.1)

All code for data download and processing is available at
https://gitbio.ens-lyon.fr/LBMC/qrg/raptor-analysis

Data analysis

All analyses were performed in R (v4.1.2), with the help of the following packages
- splines (base)
- stats (base)
- parallel (base)
- ica (v1.0.2)
- randomForest (v4.6.14)
- ROCR (v1.0.11)
- mgcv (v1.8.39)

All code for data analysis and figure plotting is available at
https://gitbio.ens-lyon.fr/LBMC/qrg/raptor-analysis

We share our tool as an open-source R package on github:
https://github.com/LBMC/RAPToR

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The full list of datasets and accession numbers is given in Supplementary Table 12.
All the data used in this study were previously published and are available on GEO, GSE49043 (C. elegans larval development time course, 20C), GSE52861 (C. elegans late larval development time course, 25C), GSE60471 (D. melanogaster embryonic development time course), GSE60755 (C. elegans embryonic development time course), GSE60619 (D. rerio embryonic development time course), GSE83395 (D. rerio early embryonic samples), GSE39897 (M. musculus embryonic development time course), GSE76316 (M. musculus first molar embryonic development, replicate 1), GSE23857 (C. elegans late larval/young adult recombinant inbred lines), GSE696 (C. elegans young adult to adult development time course), GSE130811 (C. elegans larval to adult time course), GSE12298 (C. elegans young adult toxicity assay of 3 drugs), GSE97775 (C. elegans larval development time course of WT and xrn-2 mutant), GSE45719 (M. musculus single-cell early-embryo development), GSE93826 (C. elegans aging time course), GSE77110 (C. elegans aging time course), GSE12290 (C. elegans single-worm aging time course), GSE71620 (H. sapiens profiling of brain tissues), GSE178436 (B. taurus early-embryo development time course), and GSE29397 (H. sapiens early-embryo development time course), available on ArrayExpress, E-MTAB-404 (Embryonic development time course of 6 Drosophila species, E-MTAB-1333 (C. elegans young adult to adult mutant vs. wild type experiment, and E-MTAB-3929 (H. sapiens single-cell early-embryo development), available on fruitfly.org (D. melanogaster embryonic development time course), in the supplementary data of the original publication (M. musculus embryonic development time course), on request to the authors and now in our code repository (C. elegans post-dauer vs. control, and D. melanogaster aging time course). The data from Sémon et al. (M. musculus first molar embryonic development, replicate 2) is, at the time of writing, in the process of acquiring a GEO accession ID.
Source data for all plots is provided.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by the previously published data used for this study. Datasets were chosen to showcase the various functionalities of RAPToR.
Data exclusions	As described in methods, poor quality samples were filtered out per dataset by spearman correlation with other samples. Please refer to methods for full details. Further reasons for omitting samples are also detailed in the methods section of relevant analyses (e.g. keeping samples within appropriate developmental range for staging, or removing a clear outlier from ICA components).
Replication	All attempts at replication of data analysis were successful. No replication of experimental data was performed since we did not collect any experimental data.
Randomization	Randomization of samples is not applicable to our study as we do not collect any experimental data.
Blinding	Blinding was not relevant to our study as we report a new software as main finding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

1.1.3 Supplementary notes, figures, and tables

Note 1: Supplementary figures and their legends were collapsed together to make reading easier, and a trailing reference line was moved to save space. The document is otherwise unchanged from the published version at
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-022-01540-0/MediaObjects/41592_2022_1540_MOESM1_ESM.pdf

Note 2: Supplementary tables were fit to PDF format to be included in this manuscript when their size permits. Enrichment result tables (*i.e.* Sup. Tables 4-11) are too large for this, but can be donwloaded from the online article at
https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-022-01540-0/MediaObjects/41592_2022_1540_MOESM4_ESM.xls

nature portfolio

https://doi.org/10.1038/s41592-022-01540-0

Supplementary information

Real age prediction from the transcriptome with RAPToR

In the format provided by the authors and unedited

Supplementary note 1

Computational Requirements

RAPToR is an R package. We have tested and confirmed the package works in R 3.6.3 and 4.1, under Windows 7 & 10, Ubuntu 18.04 & 20.04, and macOS (10.14). The package installs under 20s once dependencies are met. Standard datasets can easily run with 4Gb of RAM and 2 CPU cores. For reference, the dataset used for demo in the main documentation vignette of the package (43 samples by ~19500 genes, GSE80157 [1]) can be both downloaded and staged (including the default 30 bootstraps with n/3 genes) with RAPToR in under 30 seconds, using less than 2Gb of RAM.

Reference interpolation

For reference interpolation, we take advantage of the redundancy of gene expression. We use either Principal or Independent Component Analysis (PCA, ICA) to decompose the expression matrix into components (sample scores, or eigengenes) that summarize the gene expression dynamics, and gene loadings that are gene contributions to each dynamic [2,3]. We interpolate the components with respect to time (Fig 1e), and reconstruct the full interpolated gene expression by the matrix product of interpolated components and the gene loadings (Fig 1f, methods). In this way, we simplify model building and validation across thousands of genes to a few components. We select a number of components by a threshold on cumulative variance explained (see methods), keeping components with 'intelligible' dynamics which we defined as those with a spline fit explaining > 0.5 of the deviance (Sup. Fig. 6). However, staging results are robust to the variation number of components used (Sup. Fig. 5a-b, see below).

In order to validate an interpolated reference for staging with RAPToR, we first stage the original reference data on its interpolated version. We fit a linear model predicting RAPToR age estimates with chronological age expecting a near-perfect match, with adjusted $R^2 > 0.99$, and non-significant intercepts. This was the case for all references presented throughout the study, aside from "Cel_YA_2" which is built from old data[4] (2004) and only had an R^2 of 0.901. Then, when possible we stage independent time course experiments and

expect very good fits of linear models ($R^2 \geq 0.9$, as shown in Fig. 2). We note that increased variability is expected from single-organism profiling data (such as Fig. 2d and Fig. 2h) as well as for aging time-series (such as Fig. 2j and Sup. Fig. 8), due to their inherent biological heterogeneity.

Interpolation errors or bias are more frequent at the edges of a time series (e.g. splines are known to behave erratically at edges). In addition, when the real age of a sample is outside the reference range, it will likely match a time point close to the edge, making estimates close to the edges less reliable. Therefore, when age estimates are found near the edges, we suggest using another overlapping reference to confirm the estimates.

Evaluating RAPToR performance

Reference interpolation effectively increases estimates temporal resolution

RAPToR uses correlation with a reference to estimate age. If we were to use a non-interpolated reference (ie. the expression matrix in Fig. 1d), this would only allow estimates at the sampled time points of the reference. Reference interpolation enables age estimates between the original reference time points.

We successfully staged worm[5], fly[6], and mouse[7] time-series (Fig. 2a, 2d, 2c) with 3, 8, and over 18 times the temporal resolution of their respective original references (see methods). This clearly shows that interpolation allowed us to accurately reconstruct gene expression dynamics of the reference time-series at an increased temporal resolution and consequently improve the accuracy of the age estimates.

To further test this, we staged another zebrafish time-course consisting of 180 embryos spread in a very short developmental window around gastrulation[8]. Despite having over 40 times the estimated temporal resolution of the reference before interpolation, the data is accurately staged by RAPToR, not only matching the established ranking[8] (ρ=0.99, Extended Data Fig. 2a) and expression dynamics (Extended Data Fig. 2b), but giving absolute times that are comparable to any estimate obtained with the same reference.

We additionally staged the same zebrafish samples on the non-interpolated reference data to explicitly show the gain from interpolation. Samples have the highest correlation with the expected reference time points, still matching the aforementioned ranks (ρ=0.94, Extended Data Fig. 2c) and our previous estimates (ρ=0.94,

Extended Data Fig. 2d), but the staging is imprecise due to the reference data only having one expression profile per hour of development.

Effect of gene set size and data quality on age estimates and bootstrap intervals

We tested RAPToR robustness to changes in the gene set size by staging an independent time series[5] of C. elegans larval development on the reference[9] we built (as in Fig. 2a). With random gene sets of 4000 genes, estimates have a median absolute deviation under 10 minutes from full gene set estimates (18718 genes); with sets of 1000, under 20 minutes (Sup. Fig. 1a-b). With smaller sets (<1000), estimates are unreliable likely because repeated expression patterns – such as oscillations or pulses – create multiple correlation peaks with the reference (Sup. Fig. 1c). For example, the repeated molts of *C. elegans* larval development generate oscillations in the correlation profiles of staged samples (Sup. Fig. 2). To address this, we implemented the possibility to include a prior for estimates (see methods). With the prior, no estimates get misplaced to other correlation peaks, even for sets of a few hundred genes (Sup. Fig. 1c-d). We note that priors must be given in reference time (though approximate timings are enough, within a few hours), which is why we provide correspondence between developmental stages and chronological time in the references through RAPToR (*plot_refs* function).

Potential sources of uncertainty for staging can be technical (profiling quality, number of genes available), or biological (developmental spread of individuals within a bulk sample, heterochrony between tissues). We explored how these factors could influence the confidence intervals (CI) of RAPToR estimates using the *C. elegans* time-series[5] and reference[9] mentioned above. We find that noise in reference data gets largely filtered out by the PCA, and thus does not impact the CI size (Sup. Fig. 4a). Staged samples of poor quality have wider CIs, but estimates stay strikingly accurate with intense noise (Sup. Fig. 4c). Indeed, even when noisy expression profiles have low correlation scores with the original data ($\rho<0.8$, Sup. Fig. 4d), staging accuracy is not diminished ($R^2=0.99$, Sup. Fig. 4c). As discussed above, with less information (fewer genes) the reliability of staging decreases, which leads to increasing CIs. To simulate developmental spread within a sample, we averaged the expression values of 2, 3, 5, and 7 consecutive time points in the staged time-series; similarly, we simulated heterochrony by randomly selecting one of the expression values between

2, 3, 5, and 7 consecutive time points (i.e. mixing). As expected, CI size increases in both scenarios (Sup. Fig. 4e, 4f).

We conclude our bootstrapping approach generates confidence intervals that reflect the multiple technical or biological sources of variability in the data.

Confidence intervals of RAPToR estimates for *C. elegans* samples[5], built from bootstrap estimates with 6239 genes (see methods), are extremely small – between 5 and 20 minutes (Sup. Fig. 3a). As expected, staging zebrafish embryos[6] on a reference[10] with lower gene overlap (8662 genes) results in larger confidence intervals built from bootstrap estimates with 2887 genes – on average, slightly over 2h across the full time series, and under 50 minutes for samples before 30h of development (Sup. Fig. 3b).

Of note, the combination of interpolation accuracy and estimate precision for small gene sets means that even time-series transcriptomic data with low sampling rate or gene coverage can be exploited to build interpolated references.

Effect of interpolation parameters on age estimates

We tested whether reference-building is robust to parameter changes. As expected, selecting more components leads to a decrease in prediction error of the model (Sup. Fig. 5a-b), and a slight increase in the correlation between staged samples and reference at estimate (Sup. Fig. 5c-d). However, the age estimates and bootstrap intervals were mostly unperturbed by the changes in number of components and were also robust to choosing PCA or ICA for interpolation (Sup. Table 2, Sup. Fig. 4b).

Scenarios such as aging and cross-species staging can require building references with few, or even a single component. In this case, expression dynamics are still accurately interpolated when they can be reconstructed from linear combinations of the selected components. For example, using a single monotonous component to build a reference will result in genes with complex expression dynamics being poorly modeled, while those with monotonous dynamics will have accurate fits.

Within-study aging expression dynamics allow robust RAPToR staging

While there is a known increase of gene expression noise with aging[11], the heterogeneity of aging also greatly depends on environmental factors. Indeed, we find that even though “standard” (whole-transcriptome) RAPToR references poorly stage independent experiments (data not shown), they accurately stage independent samples from the same study. We show this in *C. elegans* (R^2>0.99, Sup. Fig. 7a) and even in dissected tissues of wild-type and transgenic mice[12] (R^2>0.99, Sup. Fig. 7b). Together with the fact that RAPToR estimates stay accurate in very noisy expression data (see above, Sup. Fig. 4c), this suggests that the reason for poor staging performance across studies is likely due to heterogeneous environmental conditions across studies.

RAPToR stages human adults from dissected brain tissue

RAPToR aging references built with monotonous genes allow us to infer the age of adults accurately despite heterogeneous environments and genetic backgrounds. In humans – where neither of these factors is controlled – we could accurately stage individuals from the transcriptome of dissected tissues of two neighboring brain regions[13]: Brodmann Area BA11 (R^2=0.74, Sup. Fig. 8a) and BA47 (R^2=0.68, Sup. Fig. 10d). We could even stage BA11 samples on a BA47 reference (R^2=0.61, Sup. Fig. 10b) and vice versa (R^2=0.71, Sup. Fig. 8e) with comparable accuracy. Importantly, BA11 sample age estimates on the BA11 reference finely match those acquired on the BA47 reference (R^2=0.93, Sup. Fig. 8c) and the same goes for BA47 samples (R^2=0.93, Sup. Fig. 8f). This suggests that RAPToR captures genuine variability between chronological and physiological age, since the same age is given to a sample with both references.

In summary, RAPToR can infer the age of adult individuals with heterogeneous environmental and genetic backgrounds from the transcriptome of dissected tissues, and do so reliably from a reference built with a similar tissue.

Inferring developmental speed factors

Beside the expected 1.5 fold increase in developmental speed due to temperature we observed in *C. elegans* (Fig. 2a), we also observe a difference between chronological and estimated times in the independent zebrafish developmental time series[6] we staged on the zebrafish reference[10] determining a developmental

speed factor of 0.7. We suspect this speed factor is also due to a temperature difference with the reference, as growth speed scales with temperature also in zebrafish. While the reference data embryos developed at 28.5°C[10] , we were unable to confirm at which temperature the staged data embryos developed but we presume a lower one.

Soma-germline heterochrony between C. elegans experiments

To confirm the presence of soma-germline heterochrony between *C. elegans* Recombinant Inbred Lines[14] (RILs) and the reference[15] they were staged on, we compared the expression dynamics of both datasets (Extended Data Fig. 4a). While the overall Root Mean Square Error (RMSE) between the RILs and the reference fit is minimal at the RIL global estimated age (Extended Data Fig. 4c), that is not the case for single components. Indeed, reference dynamics match RILs better when shifting the RIL age estimates back for soma-enriched components and forward for germline-enriched components (Extended Data Fig. 4d). Soma- and germline-specific age estimates using tissue-specific genes then further improve the match between the RILs and reference for dynamics of these respective tissues (Fig. 3g).

Soma-germline heterochrony between the reference and RILs only explains that the dynamics of the soma be shifted along germline age and vice versa. However, the clear noise increase we see in germline dynamics along soma age (Fig. 3c) and vice versa (Fig. 3f), also implies heterochrony among RILs as it shows the soma-germline age difference with the reference varies from sample to sample.

Quantitative Trait Loci analysis on soma-germline heterochrony

In our QTL analysis of soma-germline heterochrony in *C. elegans* Recombinant Inbred Lines[14], Random Forest prediction of the trait was poor and non-significantly correlated with the trait ($r = 0.12$, $p = 0.09$) and we found no significant hits out of the 1,455 markers, even at FDR of 0.5. Removing the batch covariate from the analysis results in even poorer predictions ($r = 0.08$, $p = 0.2$), suggesting uncontrolled environmental factors may be driving heterochrony.

Staging samples on references of a different specie

When we first staged a *C. elegans* embryo development time course[6] on the *Drosophila* reference[16], we noted two breaks in the age estimates, possibly due to heterochrony of developmental processes between the 2

species (R^2 = 0.938, Sup. Fig. [[19]]a). Staging was notably improved after rebuilding the reference with only 2 components which have broad dynamics (R^2 = 0.958, Fig 4d), further suggesting that sharper expression dynamics (like pulses or oscillations) diminished staging performance.

To explore the reason behind successful cross-species staging, we analyzed the 319 genes (10%) with highest correlation between *C. elegans* and *D. melanogaster* during embryogenesis, as well as the 509 genes (10%) with highest correlation between *H. sapiens* and *M. musculus* early embryo development in single-cells (see methods), which suffice to stage the embryos well despite their small number (R^2 = 0.87 and R^2 = 0.84, Extended Data Fig. 8b and 7c). We found the fly-worm gene set clusters into an ascending gene expression signature of muscle development (cluster 1, Extended Data Fig. 8c, Sup. Table 4), and two decreasing signatures of cell proliferation split between DNA replication (cluster 2, Extended Data Fig. 8c, Sup. Table 5) and splicing (cluster 3, Extended Data Fig. 8c, Sup. Table 6) respectively. Similarly, the mouse-human gene set consists of clear ascending expression signatures of cell respiration, adhesion and secretion (clusters 1-3, Extended Data Fig. 7d, Sup. Tables 7-9) and of a strong decreasing signature of cell proliferation (cluster 4, Extended Data Fig. 7d, Sup. Table 10). Other translational-related processes are grouped without a clear trend (cluster 5, Extended Data Fig. 7d, Sup. Table 11)

Expression dynamics match well between fly and worm embryogenesis (Sup. Fig. [[19]]c), and between human and mouse early-embryo development, consistent with previous work showing widespread conservation of decrease in cell proliferation during embryogenesis[6] .

Supplementary note 2

Exploiting inferred age in genome-wide expression studies

Inferring the impact of environmental or genetic perturbations on development

When staging C. elegans exposed to increasing concentrations of mefloquine, dichlorvos, and fenamiphos[17], we noted that beyond the germline developmental delay induced by all three drugs, dichlorvos also showed a significant and opposite effect of dose on somatic age. However, the scale of the effect is a fraction of the one observed on the germline (Sup. Fig. 12b).

Exploiting developmental variation to increase power to detect differential expression

After comparing chronological age and RAPToR age estimates as predictors for *C. elegans* control and *pash-1* mutants profiled at 4 time points of late development[18] and finding better model fits (Fig. 5c), we further tested whether random perturbations on age could induce a similar result. Thus, we generated age sets of similar deviation from chronological age than RAPToR estimates (see methods) and found that these consistently decreased model fits and DE gene detection, confirming that precise age estimates increase the power of the analysis (Sup. Fig. 13a-c, methods).

We also found a curious batch effect on development: mutants are systematically older than controls in the first two replicates while it was the opposite in the third (Sup. Fig. 13d).

Detecting and correcting expression changes caused by development using reference data

When samples are few and experimental groups have little or no overlap in development, the information available in the experiment is not sufficient to separate the effects of development from those of interest. To overcome this, we developed an approach using RAPToR interpolated references to quantify and correct for development in genome-wide expression data.

Quantifying developmental expression changes

To quantify the impact of development in DE analysis, we compare observed $\log_2$-fold changes (observed logFC) between the two groups (i.e. mutant and wt) with changes expected purely from developmental differences between groups (expected logFC) which we estimate comparing age-matched interpolated reference profiles (Extended Data Fig. 9). We quantify development impact using Pearson correlation between observed and expected logFCs (or its square). We use Transcripts Per Million (TPMs) to compute the logFCs, as they are more comparable across samples and datasets.

Development can completely confound DE analysis leading to erroneous conclusions

Not accounting for confounding developmental variation in DE analysis can lead to erroneous conclusions. Comparing young adult *C. elegans* that developed through dauer state (post-dauer) to controls that did not, Hall et al. [19] conclude that post-dauer animals have reduced spermatogenesis from a down-regulation of spermatogenesis-associated genes and an up-regulation of oogenesis-associated genes in their DE analysis. However, this could simply be explained by post-dauer samples being older than controls as *C. elegans* naturally switch from sperm to egg production during development. Moreover the increased brood size in post-dauer worms described by the authors[19] would even suggest the opposite as sperm number limits brood size in *C. elegans*: post-dauer animals would have up- and not down-regulation of spermatogenesis genes.

To rigorously test if these expression changes are caused by development, we estimated the global and tissue-specific age of samples with our best-quality reference, and found post-dauer samples were systematically older. However, while germline age estimates were reliable, global and soma-specific staging put some samples at the edge of the reference, indicating development beyond reference bounds (Sup. Fig. 14a). Thus, we validated the age estimates with an older and lower-quality reference spanning a few hours further and found the same divide in global, soma, and germline age between groups (Sup. Fig. 14b).

RAPToR age estimates show the control samples are in late spermatogenesis, while the post-dauer samples are 5-10 hours older, fully switched to oogenesis (Fig 5c, Sup. Fig. 14a). A DE analysis between the two conditions does recapitulate reported divide between spermatogenesis and oogenesis genes (Sup. Fig. 14c). However, this is also fully recapitulated by the expected developmental changes (r = 0.82, Fig 5d, Sup. Fig.

14d). Correlation also stands with all genes (r = 0.44, Sup. Fig. 14e). Repeating the logFC comparison with the older reference yielded similar results (germline r = 0.74, all genes r = 0.41 Sup. Fig. 14f-g).

This supports our hypothesis that the expression changes between groups, particularly in germline genes, is due to a difference in development between the samples, rather than a direct effect of the post-dauer condition.

Recovering direct perturbation effects using reference data

To recover the effect of a perturbation confounded by development we propose a model in which we include all reference expression profiles in a time window spanning the development of the selected samples (see methods) and model expression dynamics as a spline of developmental time, batch between reference and sample data, and the perturbation of interest. We first convert interpolated reference data from TPM to counts, assuming a fixed library size (see methods) to ensure compatibility with tools that require counts to calculate differential expression. Including reference data with artificially low dispersion invalidates models statistics such as p-values. We can however rely on the model coefficients (logFCs) estimated with the reference data to account for development. The perturbation (strain) logFC coefficients estimated by this reference integrated model are controlled for developmental changes shared by the samples and the reference (Extended Data Fig. 10a).

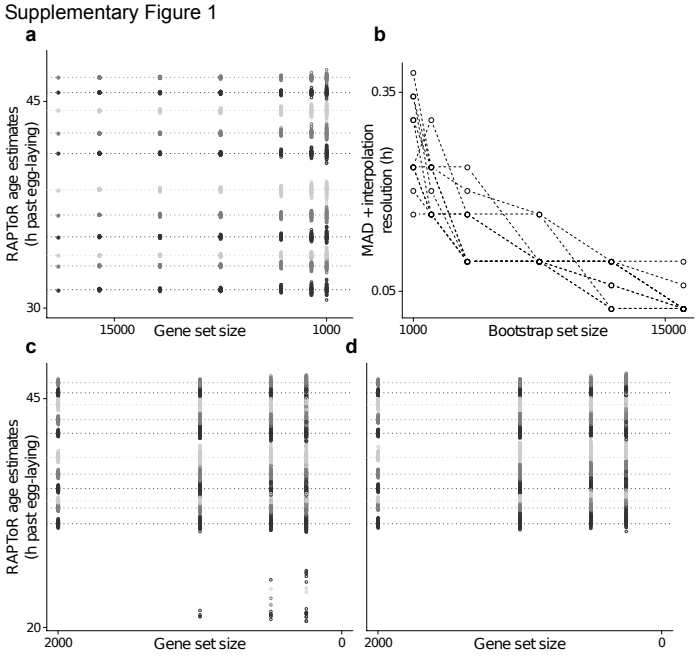
To evaluate how effectively our approach recovers truly DE genes, we exploit time series data of *C. elegans xrn-2* and WT late larval development sampled every hour at 25°C[20] . We first define a gold standard of truly DE genes by comparing three mutant and three WT samples with the best developmental match. Next we evaluate the effect of increasing developmental difference on the DE analysis by comparing the same three mutant samples with three increasingly mismatching WT samples (Fig. 5f, methods). We shifted WT samples back by 1, 2, 3, 5 and 7 time points (corresponding roughly to 1, 2, 3, 5, and 7 hours of development at 25°C). Correlation between expected and observed logFC quickly increases with increasing developmental shifts up to 0.9, meaning that around 80% of the variance in logFCs is explained by development at 7 hours of time shift. At the same time, the performance of a standard linear model p-value in identifying truly DE genes

quickly drops (Fig. 5h) as more and more expression changes are due to development (Extended Data Fig. 10b).

We show that the strain logFC from the reference integrated model already performs better in detecting truly DE genes than the standard analysis p-value for developmental shifts of 3 or more hours (at 25°C, Extended Data Fig. 10c). However, we propose an integrated predictor including the weighted mean between the standard analysis p-value and the strain logFC of the reference integrated where the optimal weight **w** is proportional to the variance of standard logFC explained by development (Extended Data Fig. 10d-e). At the optimal **w** (see methods, Extended Data Fig. 10e), our integrated predictor outperforms the standard analysis p-value for all time-shifts considered, with larger time-shifts showing the strongest improvements (Extended Data Fig. 10 c-d). For the largest shift (WT -7), **w** = 1 meaning that no information from the standard DE p-value is used to get the best results.

As no gold-standard is usually available to guide the choice of **w**, we explored the relationship between the optimal **w** and the correlation of logFCs with the reference. Sampling more WT subsets including non-contiguous sets (see methods, Sup. Table 13) reveals a tight relationship between optimal ***w*** and the correlation between observed and expected logFCs (Extended Data Fig. 10 f-g) which can therefore suggest the appropriate value of ***w***.

Supplementary Figure 1



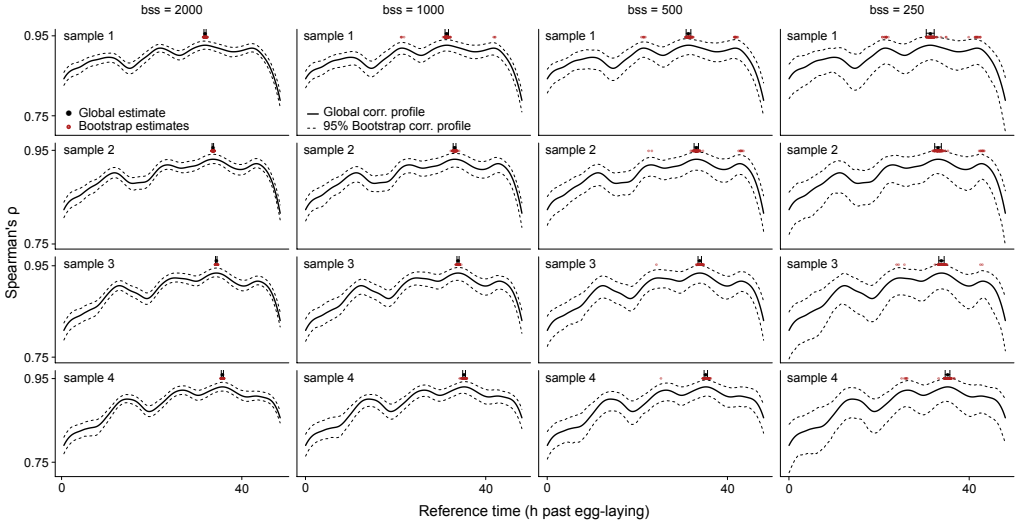
Supplementary Figure 1 – Effect of gene set size on RAPToR estimates

a, Effect of gene set size on RAPToR estimate of a *C. elegans* larval development time series[5] staged on a reference built from Kim et al.[9] data. Leftmost points and dashed lines indicate estimates on all genes (18718).

b, Median Absolute Deviation of bootstrap estimates from global estimate + interpolation resolution (i.e. half a confidence interval) by bootstrap set size for 50 bootstrap estimates.

c,d, Effect of small gene sets on RAPToR estimates without prior (**c**), or with prior (**d**). Some samples are staged to a previous molt (see also Sup. Fig. 2) when the gene set is small, which is solved by including a prior. Horizontal dashed lines indicate estimates of the samples using the full available gene set (18718).

Supplementary Figure 2

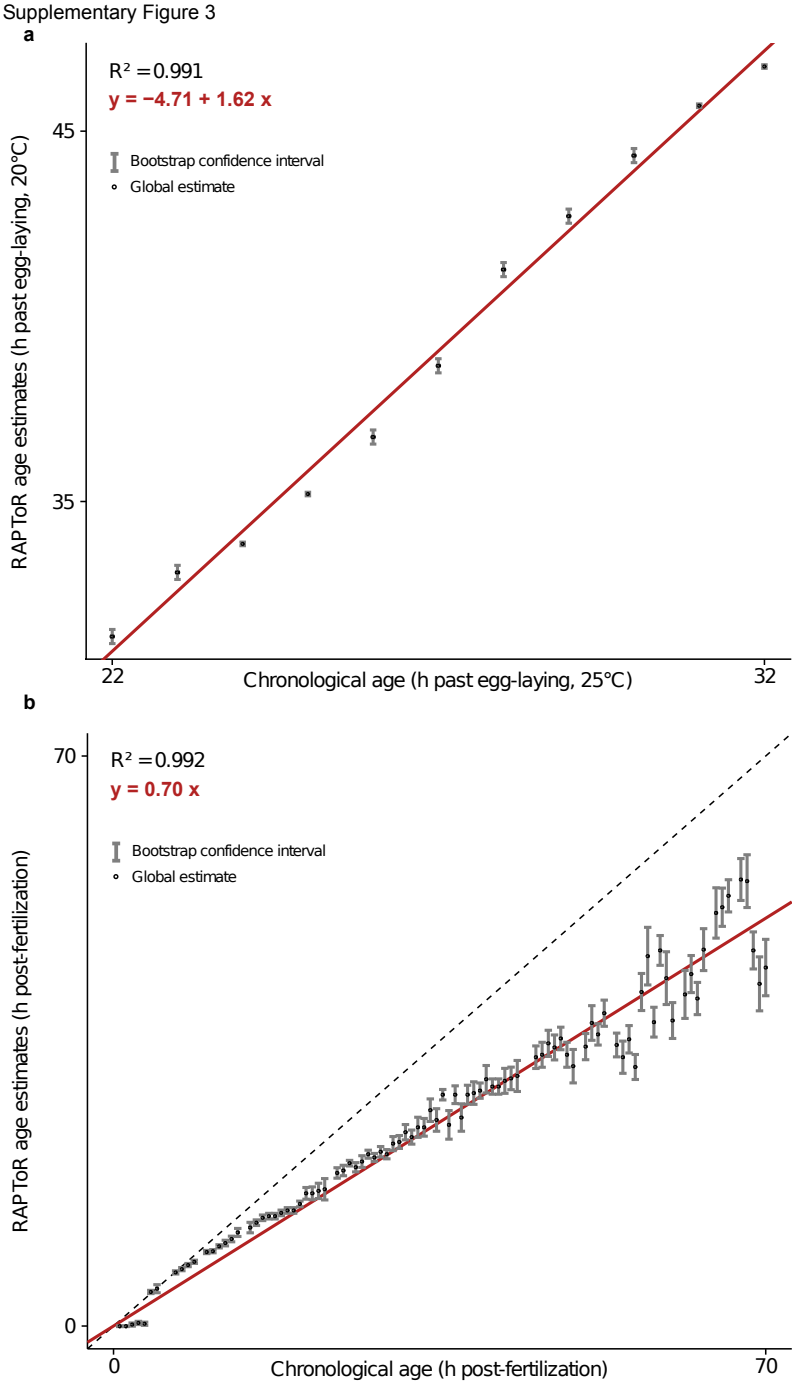


Supplementary Figure 2 – Effect of bootstrap (or gene) set size on RAPToR correlation profiles

Correlation profiles of the first 4 samples of the Hendriks et al.[5] time course staged on the reference built from Kim et al.[9] samples. Bootstrap gene set size (bss) was set to 2 000, 1 000, 500 or 250, with 100 bootstraps. With small gene sets, samples can be staged to other points in the transcriptomic landscape with similar expression (e.g, the four maxima in the profiles shown here are in phase with the oscillatory expression pattern of the four successive larval molts of *C. elegans*).

Global and bootstrap estimates, as well as the confidence interval, are shown above each profile. Dotted lines around the global correlation profiles correspond to 0.025 and 0.975 quantiles of the bootstrap correlation profiles. As expected, this interval gets larger for smaller bootstrap gene set sizes.

Supplementary Figure 3

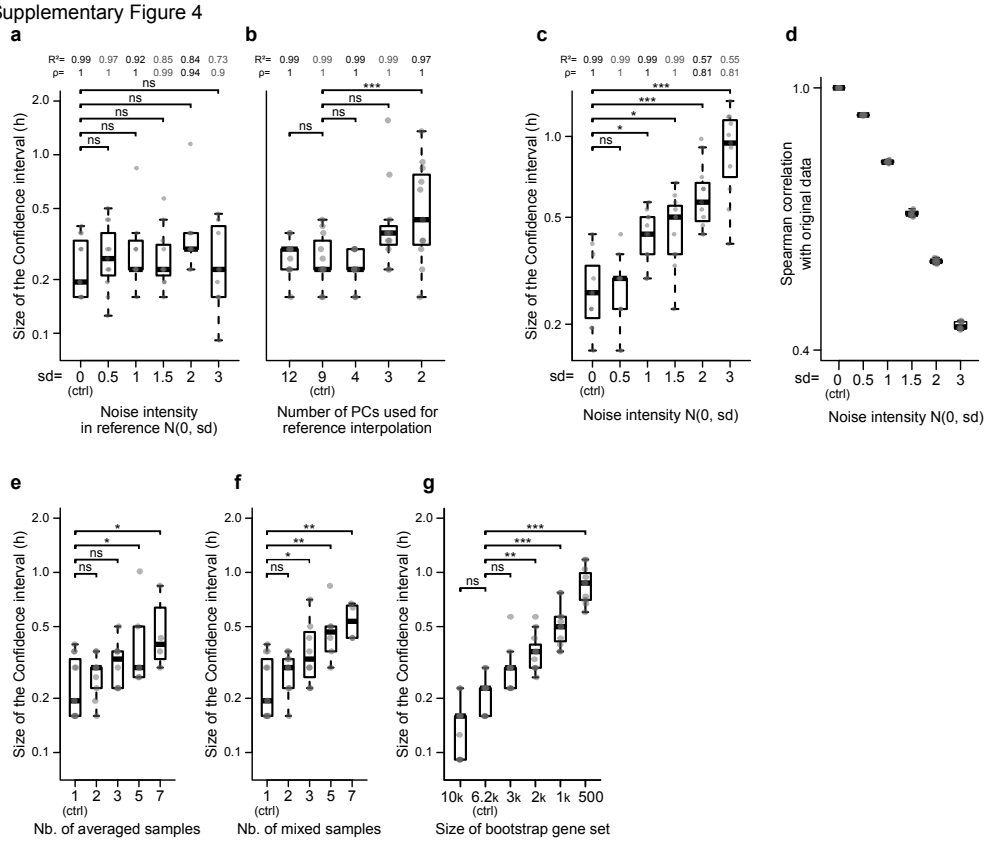


Supplementary Figure 3 – Minute-scale precision staging of development

a,b, Chronological age vs. RAPToR age estimates (black dots) and their bootstrap confidence intervals (grey bars) of *C. elegans* late larval development[5] (**a**), and *D. rerio* embryo development[6] (**b**) staged on appropriate references[9,10] (as in Fig 2a, 2b).

Bootstrap confidence intervals were computed with 50 bootstrap sets by RAPToR as part of the staging process (see methods).

Supplementary Figure 4



Supplementary Figure 4 – Effect of data quality on staging and confidence intervals

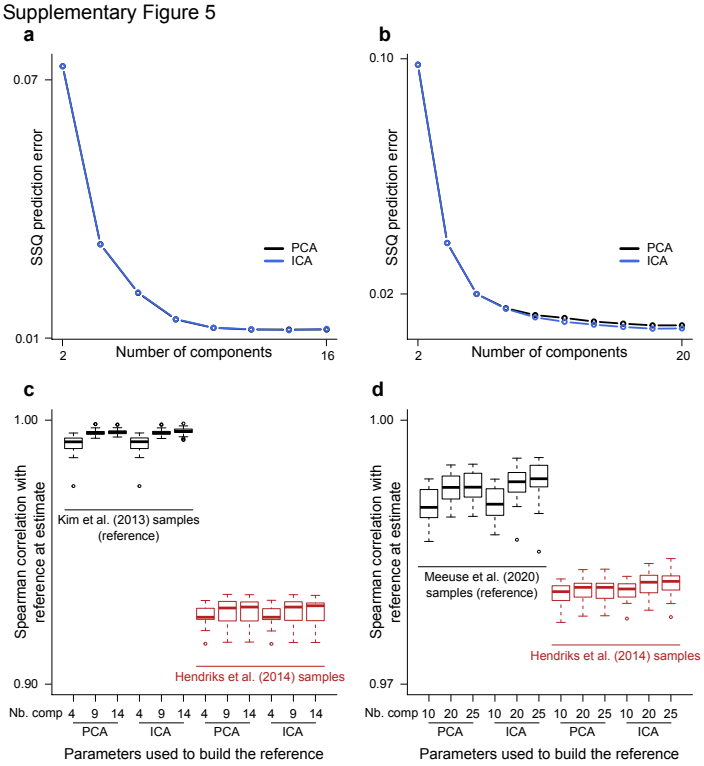
A bulk RNA-Seq *C. elegans* larval development time-series[5] is staged on a reference built from an independent time series[9] . RNA-seq samples were staged using 10,000 genes, with bootstrap confidence intervals (CI) of the age estimates built from 50 bootstrap estimates.

a,b, CI size when adding gaussian noise to log(TPM+1) reference data before interpolation (**a**), and when changing the number of components used for interpolation (**b**). From top to bottom, p-values are (**a**): p=0.73, p=0.73, p=0.98, p=0.73, and p=0.98; (**b**): p=8.35e-04, p=0.40, p=0.91, and p=0.91.
c, CI size when adding gaussian noise to log(TPM+1) sample expression data. From top to bottom, p-values are p<1e-10, p=2.05e-05, p=0.016, p=0.036, and p=0.96.
d, Spearman correlation between original data and samples+noise (as in **c**).
e,f, CI size when averaging samples together to mimic developmental spread of individuals (**e**) and when mixing samples together (randomly picking a gene expression value between n samples) to mimic heterochrony (**f**). From top to bottom, p-values are (**e**): p=0.048, p=0.048, p=0.44, and p=0.74; (**f**): p=2.00e-04, p=3.68e-03, p=0.25, and p=0.57.
g, CI size when decreasing bootstrap gene set size. From top to bottom, p-values are p<1e-10, p=2.29e-07, p=5.88e-03, p=0.12, and p=0.17.

In **a-c**, R^2 and Spearman correlation between chronological age and RAPToR estimates for each condition are shown above plots.
In **a-c,e-g**, significance of mean difference with the control condition (noted “(ctrl)”) is tested with a linear model. P-values are FDR-adjusted within each panel.

In **a-d,g**, each box is n=11 RNA-seq samples; in **e,f**, boxes are n=11, 9, 6, and 4 sample combinations from left to right respectively. Each box spans the interquartile range (IQR), the central bar denotes the median, and whiskers extend to 1.5×IQR in either direction.

Supplementary Figure 5



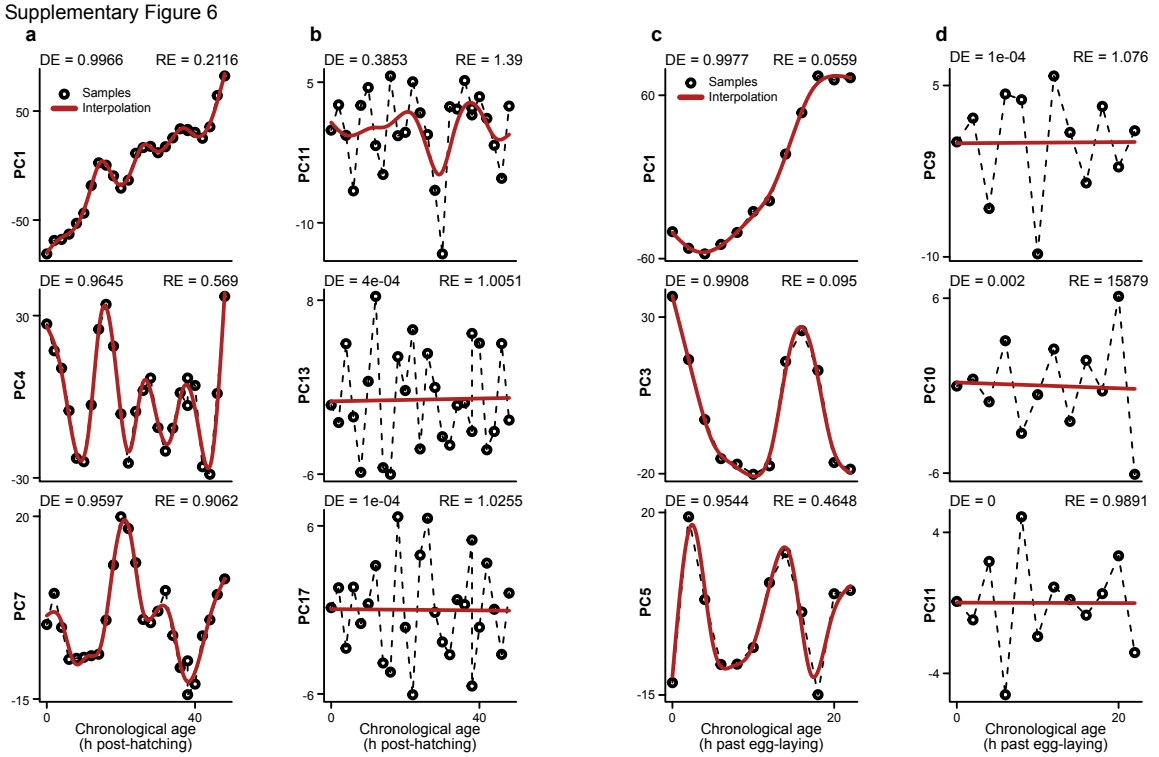
Supplementary Figure 5 – Robustness of reference-building to parameter change

a,b, Sum Squared (SSQ) prediction error of gene expression interpolation at known time points by number of components and dimension-reduction method used for interpolation in the Kim et al.[9] (**a**), and Meeuse et al.[15] (**b**) references of *C. elegans* larval development.

c,d, Spearman correlation between interpolated reference and (non-interpolated) reference and independent time series[5] samples at their age estimates when varying dimension reduction method and component number used for reference-building, using Kim et al. samples (**c**) or Meeuse et al. (**d**) samples to build a reference (see also Sup. Table 2).

In **c**, each box is n=26 for Kim et al.[9] samples and n=12 for Hendriks et al.[5] samples; in **d**, each box is n=44 for Meeuse et al.[15] samples and n=16 for Hendriks et al. samples. Each box spans the interquartile range (IQR), the central bar denotes the median, and whiskers extend to 1.5×IQR in either direction.

Supplementary Figure 6

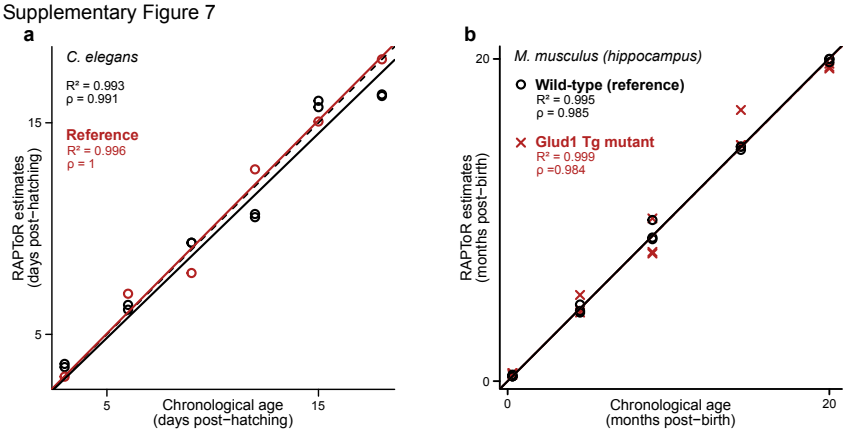


Supplementary Figure 6 – Defining intelligible dynamics with spline fits

Expression profiling time series are projected into principal component space and fit with their respective reference interpolation models (see Sup. Table 1). The interpolation of each component is then evaluated with the deviance explained (DE) and relative error (RE) of the fit.

a,b, *C. elegans* larval development[9] selected components with (**a**) and without (**b**) intelligible dynamics (with respect to time), that are kept or dropped for reference building respectively.
c,d, *D. melanogaster* embryo development[16] selected components with (**c**) and without (**d**) intelligible dynamics, that are kept or dropped for reference building respectively.

Supplementary Figure 7

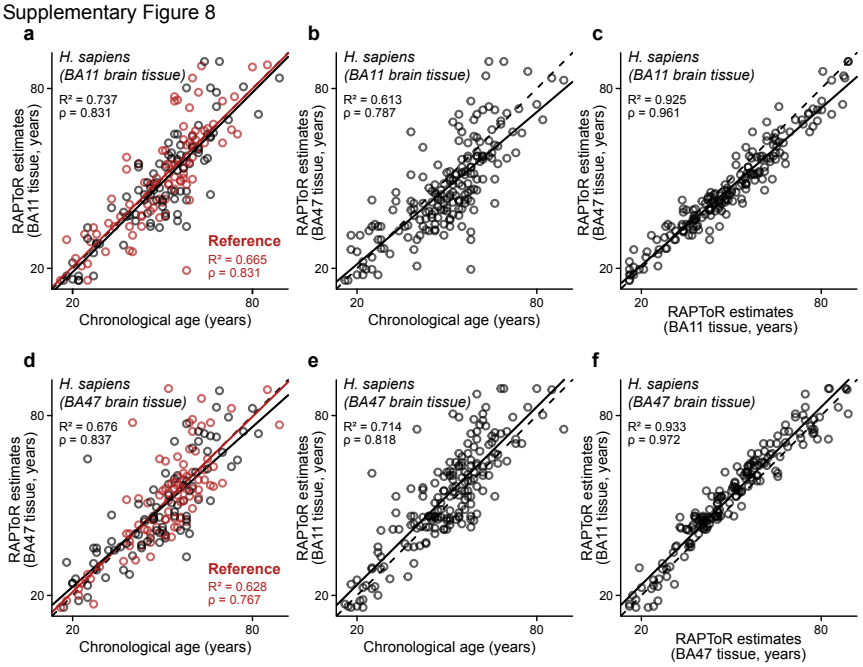


Supplementary Figure 7 – RAPToR standard reference stage samples well within experiments

a, Chronological age vs. RAPToR age estimates of a *C. elegans* aging time series (Byrne et al. 2020, unpublished). The samples from one of 3 replicates were used to build the reference (in red).

b, Chronological age vs. RAPToR age estimates of dissected hippocampus tissue from *M. musculus* across the entire lifespan of mice in wild-type (black) and *Glud1* transgenic (Tg) animals (red)[12]. Wild-type samples were used to build the reference.

Supplementary Figure 8



Supplementary Figure 8 – RAPToR stages adult human brain tissue samples

a,b, Chronological age vs. RAPToR estimates of human BA11 brain tissue samples on (**a**) a reference built from half the samples (in red) and (**b**) a reference built from human BA47 brain tissue samples.

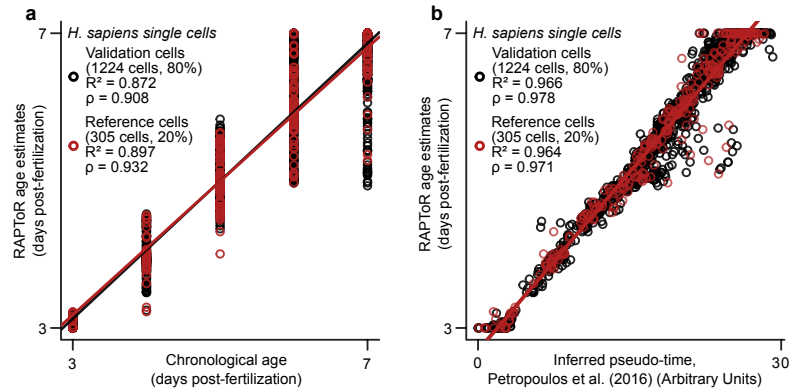
c, RAPToR estimates of BA11 samples on BA11 reference (as in **a**) vs. RAPToR estimates on a BA47 reference (as in **b**).

d,e, Chronological age vs. RAPToR estimates of human BA47 brain tissue samples on (**d**) a reference built from half the samples (in red) and (**e**) a reference built from human BA11 brain tissue samples.

f, RAPToR estimates of BA47 samples on BA47 reference (as in **d**) vs. RAPToR estimates on a BA11 reference (as in **e**).

All samples are from Chen et al.[13]

Supplementary Figure 9



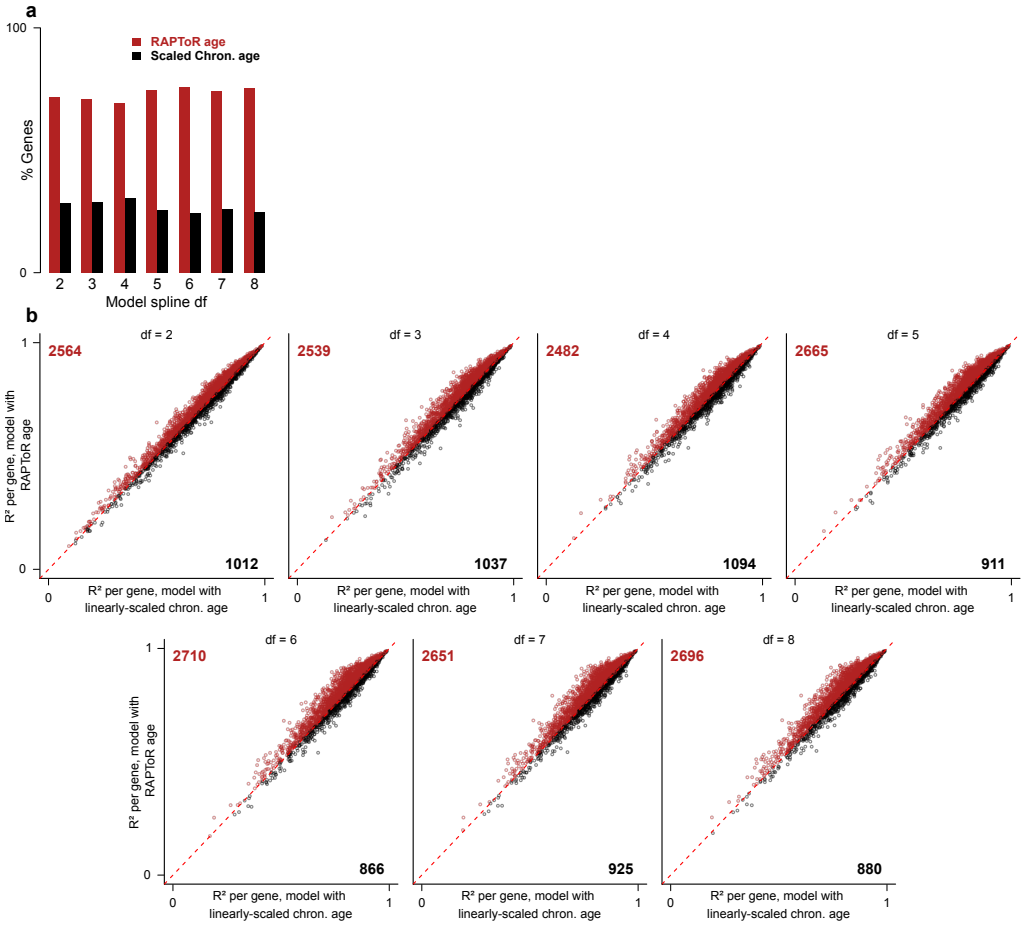
Supplementary Figure 9 – Staging human single cells with RAPToR

20% of a human early-embryogenesis single cell dataset[21] were used to build a reference, on which all cells were then staged. Metrics are split between reference and validation cell subsets.

a, Chronological age vs. RAPToR age estimates of single-cells.

b, Inferred pseudotime from the authors vs. RAPToR age estimates of single cells.

Supplementary Figure 10

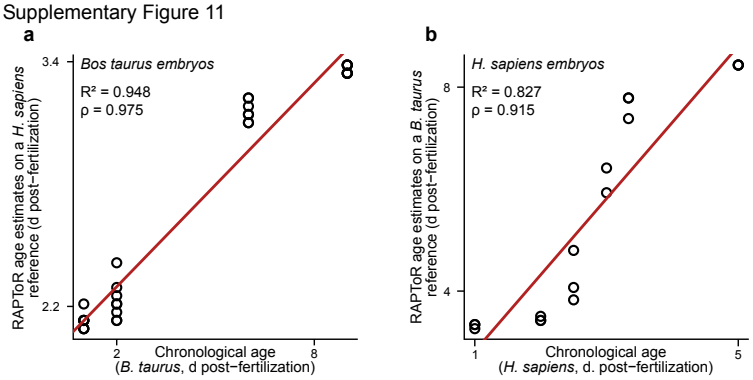


Supplementary Figure 10 – RAPToR age estimates improve model fits over linear age scaling
Drosophila embryo samples from Kalinka et al.[23] of 6 species are staged on a *D. melanogaster* reference built from Graveley et al.[16] samples (as in Fig 3a).

a, Choice between identical models fit on gene expression with linearly-scaled chronological age or RAPToR estimates as predictors.

b, R^2 per gene of models with chronological age as predictor vs. the same model using RAPToR estimates, across 2-8 of spline degrees of freedom.

Supplementary Figure 11

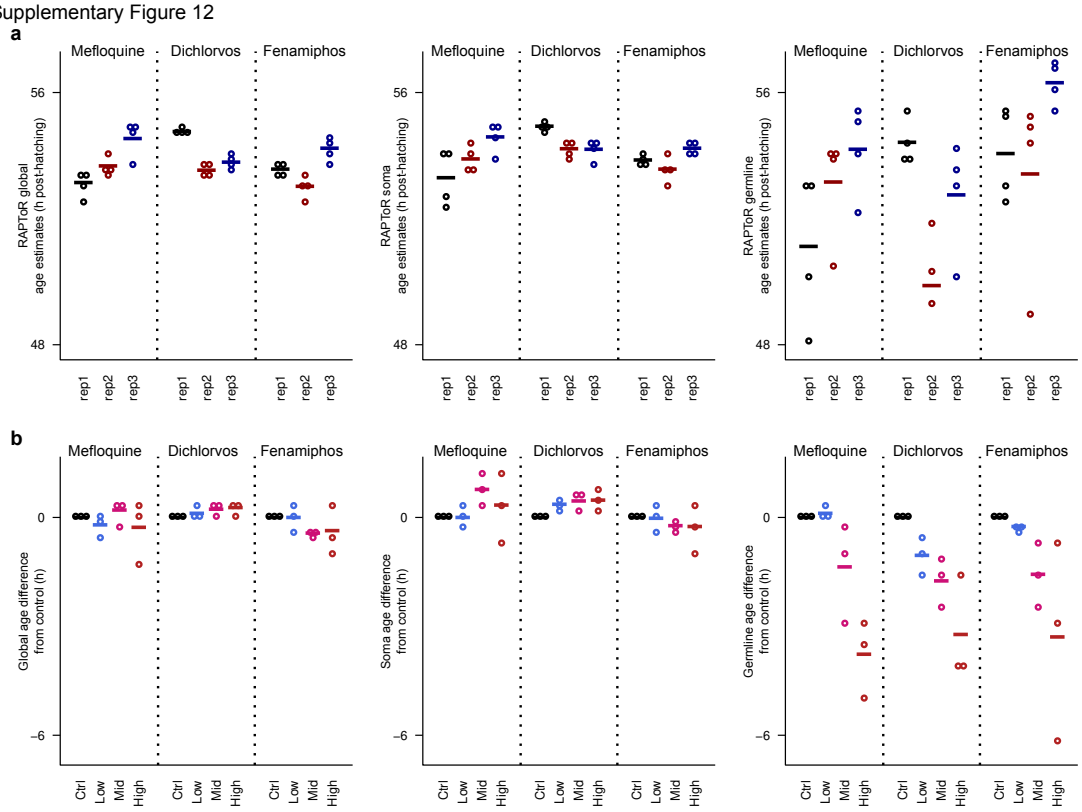


Supplementary Figure 11 – Staging cow embryos on human embryos and vice versa

a, Staging early-embryo development of *B. taurus*[24] on a human early-embryo development reference[25].

b, Staging early-embryo development of *H. sapiens*[25] on a cow early-embryo development reference[24].

Supplementary Figure 12



Supplementary Figure 12 – Impact of drugs on C. elegans germline development

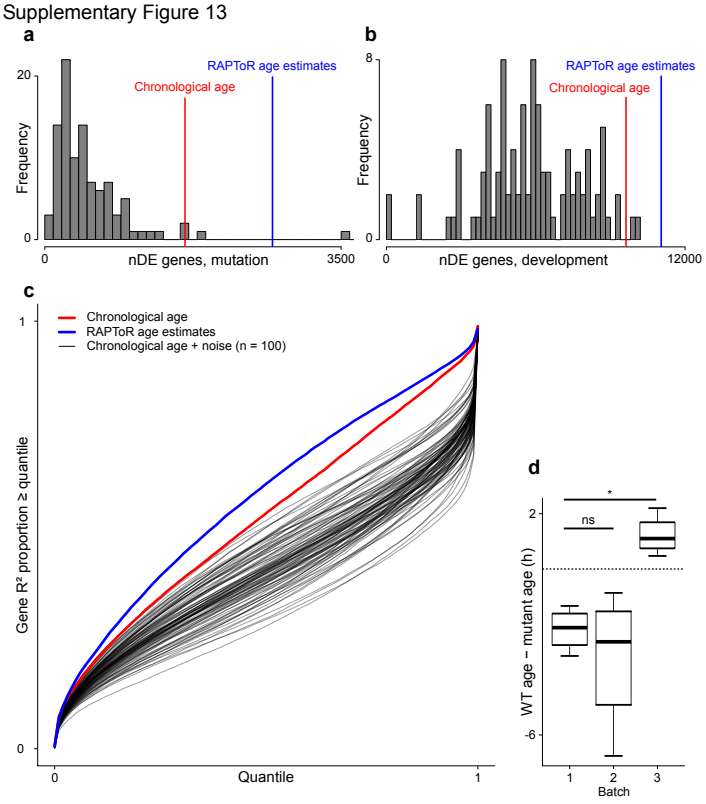
C. elegans samples exposed to 3 doses of 3 different drugs profiled by Lewis et al.[17] are staged on the reference built from Meeuse et al.[15] samples.

a, Impact of batch on global, soma and germline age.

b, Impact of drug dose on global, soma and germline age, normalized per batch. Age difference is computed by subtracting the age of the control sample within each batch.

In **a,b**, bars indicate group mean.

Supplementary Figure 13



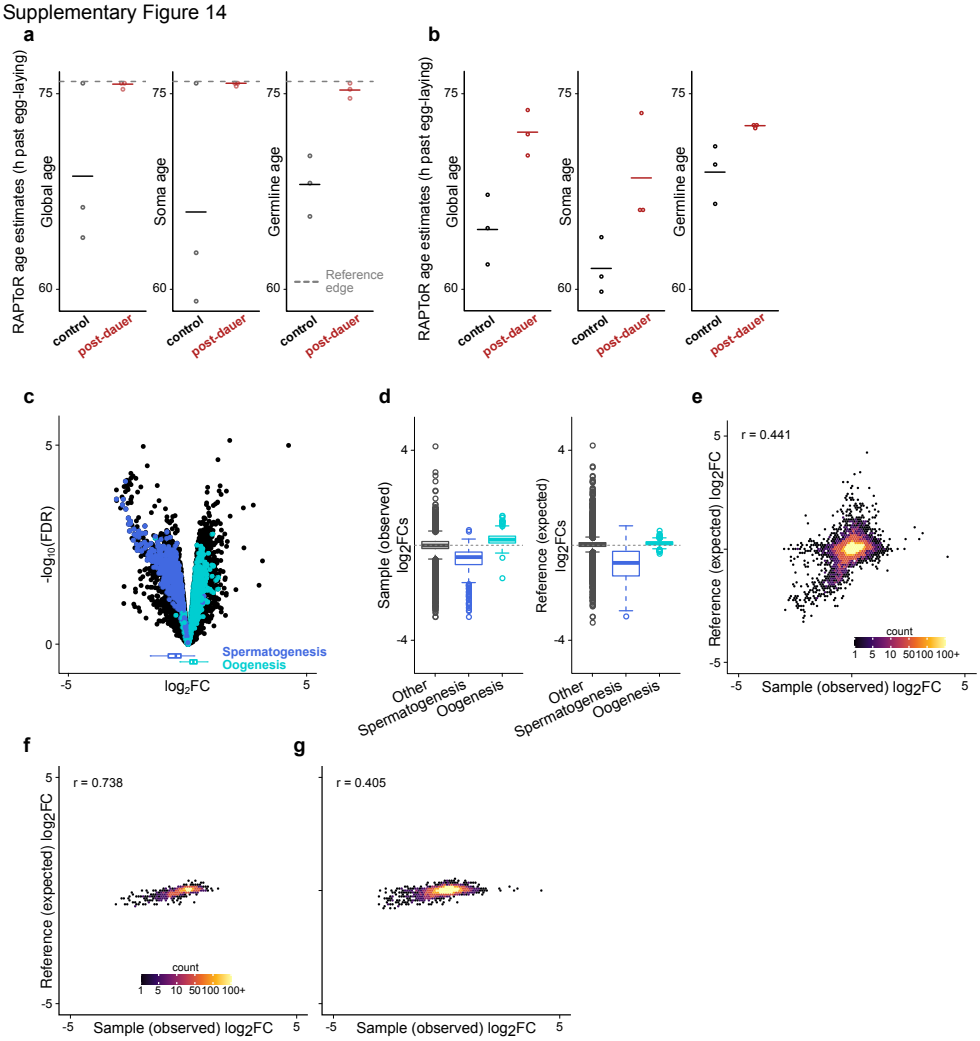
Supplementary Figure 13 – Increasing the power of DE gene detection with RAPToR estimates

*pash-1*ts or wild-type C*. elegans* RNA-seq bulk samples profiled by Lehrbach et al.[18] are staged on the reference built from Reinke et al.[4] samples.

a-c, Random perturbations on chronological age (n=100 sets, see methods) produce overall fewer DE genes for strain (**a**), and development (**b**) than either chronological age (red) or RAPToR estimates (blue). These random perturbations further cause poorer model fits than chronological age whereas RAPToR age estimates systematically outperform chronological age (**c**).

d, Batch effect on developmental difference between control and *pash-1*ts samples. Each box is n = 4 comparisons between WT and mutant time points, and spans the interquartile range (IQR), the central bar denotes the median, and whiskers extend to 1.5×IQR in either direction. Exact p-values, derived from two-sided t-tests on linear model coefficients, are p = 0.0139, and p = 0.3755 from top to bottom respectively, no correction was applied.

Supplementary Figure 14



Supplementary Figure 14 – Germline expression changes recapitulated by a developmental shift

a,b, Global, soma-specific and germline-specific RAPToR age estimates of *C. elegans* control and post-dauer (PD) samples[19] staged on a reference built from Meeuse et al.[15] (**a**), and Reinke et al.[4] (**b**). Each group is n=3, with bars indicating mean age.

c, Volcanoplot of control vs. PD groups. Spermatogenesis and oogenesis genes are color-coded and logFCs of both categories are shown in boxplots at the bottom.

d, logFCs of control vs. PD for spermatogenesis and oogenesis genes observed in samples (left) and expected from development in the reference built from Meeuse et al., (right). Boxes are n=15482, n=596, and n=875 for other, sperm, and oogenesis respectively. Each box spans the interquartile range (IQR), the central bar denotes the median, and whiskers extend to 1.5×IQR in either direction..

e, Observed expression logFCs between control and PD samples vs. expected developmental expression logFCs from the reference built from Meeuse et al. (as in Fig 5h, but for all genes).

f,g, Observed expression logFCs genes between control and PD samples vs. expected developmental expression logFCs from the reference built from Reinke et al., for germline genes (**e**), and all genes (**f**).

logFC: log2 fold-change, FDR: false discovery rate.

References (Supplementary)

1. Aeschimann, F. *et al.* LIN41 post-transcriptionally silences mRNAs by two distinct and position-dependent mechanisms. *Mol. Cell* **65**, 476–489 (2017).
2. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* **97**, 10101–10106 (2000).
3. Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. & Davis, R. W. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci.* **102**, 12837–12842 (2005).
4. Reinke, V., San Gil, I., Ward, S. & Kazmer, K. Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development* **131**, 311–323 (2004).
5. Hendriks, G.-J., Gaidatzis, D., Aeschimann, F. & Großhans, H. Extensive oscillatory gene expression during C. elegans larval development. *Mol. Cell* **53**, 380–392 (2014).
6. Levin, M. *et al.* The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637 (2016).
7. Collins, J. E. *et al.* Common and distinct transcriptional signatures of mammalian embryonic lethality. *Nat. Commun.* **10**, 1–16 (2019).
8. Rauwerda, H. *et al.* Transcriptome dynamics in early zebrafish embryogenesis determined by high-resolution time course analysis of 180 successive, individual zebrafish embryos. *BMC Genomics* **18**, 1–15 (2017).
9. Kim, D. hyun, Grün, D. & van Oudenaarden, A. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. *Nat. Genet.* **45**, 1337–1344 (2013).
10. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
11. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* **441**, 1011–1014 (2006).
12. Wang, X. *et al.* Gene expression patterns in the hippocampus during the development and aging of Glud1 (Glutamate Dehydrogenase 1) transgenic and wild type mice. *BMC Neurosci.* **15**, 1–17 (2014).
13. Chen, C.-Y. *et al.* Effects of aging on circadian patterns of gene expression in the human prefrontal

cortex. *Proc. Natl. Acad. Sci.* **113**, 206–211 (2016).

14. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330**, 372–376 (2010).
15. Meeuse, M. W. *et al.* Developmental function and state transitions of a gene expression oscillator in Caenorhabditis elegans. *Mol. Syst. Biol.* **16**, e9498 (2020).
16. Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473 (2011).
17. Lewis, J. A., Szilagyi, M., Gehman, E., Dennis, W. E. & Jackson, D. A. Distinct patterns of gene and protein expression elicited by organophosphorus pesticides in Caenorhabditis elegans. *BMC Genomics* **10**, 202 (2009).
18. Lehrbach, N. J. *et al.* Post-developmental microRNA expression is required for normal physiology, and regulates aging in parallel to insulin/IGF-1 signaling in C. elegans. *Rna* **18**, 2220–2235 (2012).
19. Hall, S. E., Beverly, M., Russ, C., Nusbaum, C. & Sengupta, P. A cellular memory of developmental history generates phenotypic diversity in C. elegans. *Curr. Biol.* **20**, 149–155 (2010).
20. Miki, T. S., Carl, S. H. & Großhans, H. Two distinct transcription termination modes dictated by promoters. *Genes Dev.* **31**, 1870–1879 (2017).
21. Petropoulos, S. *et al.* Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
22. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505**, 208–211 (2014).
23. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).
24. Cuthbert, J. M. *et al.* Comparing mRNA and sncRNA profiles during the maternal-to-embryonic transition in bovine IVF and scNT embryos. *Biol. Reprod.* **105**, 1401–1415 (2021).
25. Vassena, R. *et al.* Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699–3709 (2011).
26. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).

List of Supplementary Tables

*: Tables omitted as they are too large for PDF format. All supplementary tables are accessible online at `https://www.nature.com/articles/s41592-022-01540-0#Sec53`

Dataset	Organism	Reference name	Data-package	Formula	Dim. red, nc*	Notes
Kim et al. (2013)	C. elegans	NA	NA	`X ~s(age, bs='cr', k=14)`	PCA, 9	This reference is not a part of a data-package because wormRef has a reference built from the joint Kim and Hendriks data (`Cel_larval`). This reference was built to demonstrate that staging could capture developmental speed differences. Only the 20C samples of the study were used.
Meeuse et al. (2020)	C. elegans	`Cel_larv_YA`	wormRef	`X ~s(age, k=25, bs='cr')`	ICA, 20	Age is scaled to "20C development" by staging overlapping samples on the `Cel_larval` reference in wormRef and computing the developmental speed difference.
Reinke et al. (2004)	C. elegans	`Cel_YA_2`	wormRef	`X ~s(age, bs='tp') + cov`	PCA, 14	'cov' is the batch variable. Age is scaled to "20C development" by staging on the `Cel_larval` reference in wormRef.
Graveley et al. (2011)	D. melanogaster	`Dme_embryo`	drosoRef	`X ~s(age, bs='cr')`	PCA, 8	Timing of the samples was considered as the lower bound of the given time interval (e.g. "em6.8hr" was considered at 6h of development).
Domazet et al. (2010)	D. rerio	`Dre_emb_larv`	zebraRef	`X ~s(age,bs ='cr')`	PCA, 12	Mixed sex samples from 0 to 1080 hours post-fertilization were used.
Lu et al. (2013)	M. musculus	`Mmu_embryo`	mouseRef	`X ~s(age, bs='cr')`	ICA, 8	Only samples older than 0.5 day post coitus were used (large gap between 0.5dpc and next timepoints at 6.5dpc, 7.5dpc, ...)
Pantalacci et al. (2017) & Sémon et al. (bioRXiv)	M. musculus (first lower molars)	NA	NA	`X ~s(age, bs='ts', k=4)+rep`	PCA, 5	Only lower molar samples were used to build the reference.
Petropoulos et al. (2016)	H. Sapiens (single-cells)	NA	NA	`X ~s(age, bs='cs', k=5)`	PCA, 6	Only 20% of cells at each time point were used to build the reference (see methods)
Byrne et al. (unpublished)	C. elegans	NA	NA	`X ~s(age, bs='cr', k=4)`	PCA, 1	All samples were used to build the reference. For Sup. Fig.9, a third of samples were randomly selected to build the reference, using the model formula and 3 components.
Pletcher et al. (2002)	D. melanogaster	NA	NA	`X ~s(age_days, bs='tp', k=3)`	PCA, 1	Only control (food ad libitum) samples were used to build the reference.
Chen et al. (2016)	H. Sapiens (BA47 brain tissue)	NA	NA	`X ~s(age, bs='cr')`	PCA, 1	Half of the samples were randomly sampled to use as reference.
Chen et al. (2016)	H. Sapiens (BA11 brain tissue)	NA	NA	`X ~s(age, bs='cr')`	PCA, 1	Half of the samples were randomly sampled to use as reference.
Wang et al. (2014)	M. musculus (hippocampus)	NA	NA	`X ~s(age, bs='cs', k=5)`	PCA, 4	Only Wild-type samples were used to build the reference
Cuthbert et al. (2021)	B. taurus	NA	NA	`X ~s(age, bs='cr', k=4)`	PCA, 2	Only "somatic cell nuclear transfer" embryos (half) were used to build the reference. Sample timings are from the literature (see methods)
Vassena et al. (2011)	H. sapiens	NA	NA	`X ~s(age, bs='cr', k=6)`	PCA, 2	Only replicates 2 and 3 were used to build the reference. Sample timings are from the literature (see methods)

*: Dimension reduction, number of components

Supplementary Table 1 – RAPToR References used in this study

	Reference built from Kim et al. (2013)					
Nb. components	4		9		14	
	PCA	ICA	PCA	ICA	PCA	ICA
Kim et al. (2013) samples*	0.999	0.999	0.999	0.999	0.999	0.999
Hendriks et al. (2014) samples	0.991	0.993	0.991	0.992	0.992	0.994
	Reference built from Meeuse et al. (2020)					
Nb. components	10		20		25	
	PCA	ICA	PCA	ICA	PCA	ICA
Meeuse et al. (2013) samples*	0.999	0.999	1	1	1	1
Hendriks et al. (2014) samples	0.998	0.999	0.999	0.999	0.998	0.999

*: samples used to build the reference

Supplementary Table 2 – Robustness of Reference-building to parameter changes (R^2 of RAPToR estimates vs. Chronological age)
Samples of the 20°C time-series profiling experiments from Kim et al. (2013) were used to build a reference, changing the number of components used and between the use of PCA/ICA for interpolation. The same samples were then staged with RAPToR on the references, and R^2 value of Chronological age vs. RAPToR estimates are reported. Another time-course (Hendriks et al. 2014) was used as an external dataset validation and similarly staged on the references. The same was done using samples from the Meeuse et al. (2020) dataset to build the references. See also Sup. Fig 7c, 7d.

Species	Developmental speed factor computed from RAPToR estimates	Developmental speed factor from Kalinka et al. (2010)
D. melanogaster	1.060	1.00
D. simulans	1.185	1.18
D. ananassae	1.190	1.15
D. pseudoobscura	0.901	0.93
D. persimilis	0.880	0.98
D. virilis	0.696	0.70

Supplementary Table 3 – Developmental speed factors of *Drosophila* species compared to *D. melanogaster*
Briefly, Kalinka et al. (2010) established their scaling factors by interpolating on the gene expression with respect to time and computing the sum square difference between a specie and *D. melanogaster* at equal time, for a range of scaling values. The minimum of the scaling parameter curves became the chosen factor.

Table too large for PDF (can be downloaded from `https://www.nature.com/articles/s41592-022-01540-0#Sec53`)

Supplementary Table 4 – * Enrichment analysis of Cluster 1 of highest-correlated *C. elegans* and *D. melanogaster* genes

Table too large for PDF (can be downloaded from `https://www.nature.com/articles/s41592-022-01540-0#Sec53`)

Supplementary Table 5 – * Enrichment analysis of Cluster 2 of highest-correlated *C. elegans* and *D. melanogaster* genes

Table too large for PDF (can be downloaded from `https://www.nature.com/articles/s41592-022-01540-0#Sec53`)

Supplementary Table 6 – * Enrichment analysis of Cluster 3 of highest-correlated *C. elegans* and *D. melanogaster* genes

Table too large for PDF (can be downloaded from https://www.nature.com/articles/s41592-022-01540-0#Sec53)

Supplementary Table 7 – * Enrichment analysis of Cluster 1 of highest-correlated *M. musculus* and *H. sapiens* genes

Table too large for PDF (can be downloaded from https://www.nature.com/articles/s41592-022-01540-0#Sec53)

Supplementary Table 8 – * Enrichment analysis of Cluster 2 of highest-correlated *M. musculus* and *H. sapiens* genes

Table too large for PDF (can be downloaded from https://www.nature.com/articles/s41592-022-01540-0#Sec53)

Supplementary Table 9 – * Enrichment analysis of Cluster 3 of highest-correlated *M. musculus* and *H. sapiens* genes

Table too large for PDF (can be downloaded from https://www.nature.com/articles/s41592-022-01540-0#Sec53)

Supplementary Table 10 – * Enrichment analysis of Cluster 4 of highest-correlated *M. musculus* and *H. sapiens* genes

Table too large for PDF (can be downloaded from https://www.nature.com/articles/s41592-022-01540-0#Sec53)

Supplementary Table 11 – * Enrichment analysis of Cluster 5 of highest-correlated *M. musculus* and *H. sapiens* genes

Publication	Database	Accession	Description
Kim et al. (2013)	GEO	GSE49043	*C. elegans* larval development time course (20C).
Hendriks et al. (2014)	GEO	GSE52861	*C. elegans* late larval development time course (25C).
	GEO	GSE60471	*D. melanogaster* embryonic development time course.
Levin et al. (2016)	GEO	GSE60755	*C. elegans* embryonic development time course.
	GEO	GSE60619	*D. rerio* embryonic development time course.
Graveley et al. (2011)	-	fruitfly.org	*D. melanogaster* embryonic development time course.
Rauwerda et al. (2017)	GEO	GSE83395	*D. rerio* early embryonic samples.
Lu et al. (2013)	GEO	GSE39897	*M. musculus* embryonic development time course.
Collins et al. (2019)	-	NA[1]	*M. musculus* embryonic development time course.
Pantalacci et al. (2017)	GEO	GSE76316	*M. musculus* first molar embryonic development (replicate 1)
Sémon et al. (biorXiv)	-	NA[2]	*M. musculus* first molar embryonic development (replicate 2)
Kalinka et al. (2010)	ArrayExpress	E-MTAB-404	Embryonic development time course of 6 *Drosophila* species.
Rockman et al. (2010)	GEO	GSE23857	*C. elegans* late larval/young adult recombinant inbred lines.
Reinke et al. (2004)	GEO	GSE696	*C. elegans* young adult to adult development time course.
Meeuse et al. (2020)	GEO	GSE130811	*C. elegans* larval to adult time course.
Lewis et al. (2009)	GEO	GSE12298	*C. elegans* young adult toxicity assay of 3 drugs
Lehrbach et al. (2012)	ArrayExpress	E-MTAB-1333	C. elegans young adult to adult mutant vs. wild type experiment.
Hall et al. (2010)	-	NA[3]	*C. elegans* young adult post-dauer vs. control experiment.
Miki et al. (2017)	GEO	GSE97775	*C. elegans* larval development time course of WT and xrn-2 mutant.
Petropoulos et al. (2016)	ArrayExpress	E-MTAB-3929	*H. sapiens* single-cell early-embryo development
Deng et al. (2014)	GEO	GSE45719	*M. musculus* single-cell early-embryo development
Byrne et al. (unpublished)	GEO	GSE93826	*C. elegans* aging time course
Hou et al. (2016)	GEO	GSE77110	*C. elegans* aging time course
Golden et al. (2008)	GEO	GSE12290	*C. elegans* single-worm aging time course
Pletcher et al. (2002)	-	NA[3]	*D. melanogaster* aging time course in caloric restriction and control conditions
Chen et al. (2016)	GEO	GSE71620	*H. sapiens* profiling of BA47 and BA11 brain tissues in individuals aged 16-89
Wang et al. (2014)	GEO	GSE48911	*M. musculus* hippocampus aging time course in wild-type and Glud1 mutants
Cuthbert et al. (2021)	GEO	GSE178436	*B. taurus* early-embryo development time course.
Vassena et al. (2011)	GEO	GSE29397	*H. sapiens* early-embryo development time course.

[1]: data is available in supplementary of the publication (`https://ndownloader.figshare.com/files/11864189`).
[2]: data is awaiting publication.
[3]: the authors kindly provided us with their data (available in our analysis repo at `https://gitbio.ens-lyon.fr/LBMC/qrg/raptor-analysis`).

Supplementary Table 12 – List of datasets used in this study with accession IDs

Combination of WT timepoints			Note
8	9	10	Gold Standard
7	9	10	
7	8	10	
6	9	10	
6	8	10	
6	7	10	
5	9	10	
5	8	10	
5	7	10	
5	6	10	
7	8	9	WT -1
6	8	9	
6	7	9	
5	8	9	
5	7	9	
5	6	9	
4	8	9	
4	7	9	
4	6	9	
4	5	9	
6	7	8	WT -2
5	7	8	
5	6	8	
4	7	8	
4	6	8	
4	5	8	
3	7	8	
3	6	8	
3	5	8	
3	4	8	
5	6	7	WT -3
4	6	7	
4	5	7	
3	6	7	
3	5	7	
3	4	7	
2	6	7	
2	5	7	
2	4	7	
2	3	7	
4	5	6	
3	5	6	
3	4	6	
2	5	6	
2	4	6	
2	3	6	
1	5	6	
1	4	6	
1	3	6	
1	2	6	

Supplementary Table 13 – List of WT time point combinations used to estimate the relationship between optimal $\boldsymbol{w}$ and sample-reference logFC correlation

Using the expression data from Miki et al. (2017), each set of 3 WT samples is compared to the gold standard set of xrn-2 samples (8,9,10), see Figure 5f-i.

1.2 Further improvements of RAPToR

1.2.1 Correcting for age in differential expression analysis

1.2.1.1 Introduction

In fast-growing organisms like *C. elegans*, even a subtle influence of experimental conditions on developmental speed can significantly impair gene expression analysis. Indeed, samples are usually collected at precise timings, aiming for as little developmental spread as possible within each group. Small growth speed differences between groups are therefore sufficient for the variable of interest to be fully confounded by development (Fig. 1.1a). Because of this, gene expression changes normally seen during development are easily misattributed to the variable of interest (Fig. 1.1b).

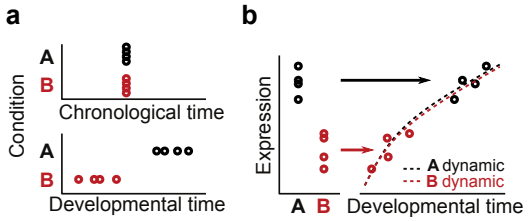


Figure 1.1 – Effect of a condition delaying development on differential expression analysis.
a,b, Cartoons illustrating a confusion of age and condition (A/B) due to the latter delaying development (**a**), and subsequent differential expression analysis situation where the developmental dynamic fully explains the difference between A and B groups (**b**). Adapted from BULTEAU & FRANCESCONI, 2022.

We showed that using RAPToR, we can precisely infer age from gene expression to detect previously unknown age differences between samples (including in a tissue-specific manner), detect and quantify the variance of expression changes due to these age differences, and recover perturbation-specific expression changes even when these age differences completely confound the variable of interest (BULTEAU & FRANCESCONI, 2022).

However, the method we initially proposed to recover perturbation-specific expression changes has two major drawbacks: 1) meaningful p-values can not be derived from the model, thus requiring a different classifier which 2), relies on an extra parameter that needs to be estimated from the data with an empirical relationship (Extended Data Fig 10g of the article). This requires advanced knowledge of Differential Expression analysis (DE analysis), making our method difficult to use, which is not ideal for such a widespread problem.

In this section, I improve upon this method and further characterize the effects of age differences between groups on DE analysis.

1.2.1.2 How to correct for age in differential expression analysis

In theory

The principles behind the age correction method remain the same as initially presented. That is, RAPToR reference data can bridge the age gap between non-overlapping sample groups, and rescue otherwise impossible DE analyses. Using a model that includes both (a window of) reference data and samples of interest, the developmental dynamic is correctly inferred with the reference and serves as the baseline for the difference between the groups of interest (Fig. 1.2).

The model integrating the reference (mref) can be formally defined as follows:

$$Y \sim \beta_0 + s(age) + \beta_{batch} * I_{ref} + \beta_{condition} * I_{condition} \quad \text{(mref)}$$

with

Y, the observations (*i.e.* expression values) of a gene;
β_0, the intercept;
$s()$, a polynomial spline to fit nonlinear dynamics along age;
β_{batch}, a batch term between the reference and samples of interest and
I_{ref}, its associated binary indicator (0 for reference, 1 for samples);
$\beta_{condition}$, the effect of the condition of interest and
$I_{condition}$, its associated indicator (1 for condition B samples, else 0).

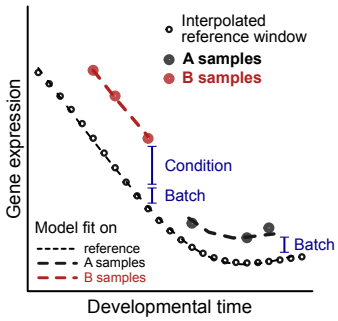


Figure 1.2 – Cartoon of DE analysis correction by integrating reference data
Adapted from BULTEAU & FRANCESCONI, 2022.

In practice

Most DE analysis tools (such as DESeq2, LOVE et al., 2014, or edgeR, ROBINSON et al., 2010) input raw counts for their particular statistical properties. As RAPToR references are in transcript per million (TPM), the interpolated reference data must first be converted from TPM to artificial counts assuming a fixed arbitrary library size (see Methods). Because these counts derive from interpolated reference data, they are essentially noiseless, thus invalidating the gene dispersions (or variances) estimated by DE analysis tools and subsequent statistical testing. This is the source of the aforementioned problems.

To bypass the issue, we now infer gene dispersions from a model without reference data and inject them into the final model with reference data. In this way, the model coefficients (log2-fold-changes, hereafter logFCs) between sample groups are corrected for development by the reference dynamic, and their respective statistical tests use dispersion values inferred only from samples, resulting in valid p-values.

1.2.1.3 Effect of age differences and DE correction in a controlled case

Analysis strategy

To test this new strategy, we define a gold-standard of truly Diffferentially Expressed genes (DE genes) between *C. elegans* wild-type (WT) and *xrn-2* mutants by comparing samples in a window of matching physiological age from published time-series expression data (MIKI et al., 2017). We then compare DE genes obtained when shifting the window of WT samples by **1**, **2**, **3**, **5**, and **7** time points with the gold-standard (Fig. 1.3). This allows us to evaluate first, the effect of increasing age mismatch between the mutant and WT on the performance of a standard DE analysis, and second, the performance of our method to recover truly DE genes of the gold-standard compared to the standard DE analysis.

We consider two scenarios for DE analysis.

$$Y \sim \beta_0 + \beta_{strain} * I_{strain} \quad \text{(model 1)}$$
$$Y \sim \beta_0 + age + \beta_{strain} * I_{strain} \quad \text{(model 2)}$$

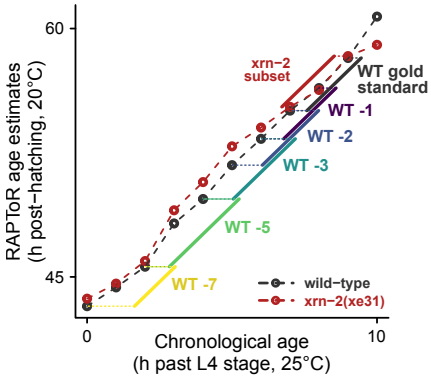


Figure 1.3 – Selected subsets of *C. elegans* WT and *xrn-2* mutant samples
Chronological age vs. RAPToR age estimates of expression data from MIKI et al. (2017).

In the first (model 1), age is not specified. Therefore, we evaluate the effect of age as an "unknown" confounder on DE analysis, which is the most common scenario. In the second (model 2), inferred age is included. This should result in better model fits and thus increase statistical power for DE gene detection as long as the effects of age and strain are separable. However, as the age difference between groups increases, so does the collinearity between age and strain variables, which decreases statistical power to detect the specific effect of each individual predictor (GRAHAM, 2003).

Both the presence of confounders and of collinearity between variables result in debatable model and statistical validity (CLARKE & GREEN, 1988; GRAHAM, 2003) and, as shown below, they have different effects on the analysis.

Age differences between groups decrease DE analysis performance

Using the age-matched set of samples, we find 955 and 1125 truly DE genes with model 1 and model 2 respectively (FDR < 0.01, $|log\text{FC}| > 1$), which are our gold standard. More genes are found with model 2 thanks to better fits, but the gene sets largely overlap (943).

As expected, the performance of both models to detect their respective gold-standard genes drops sharply with increasing developmental shifts between WT and mutant samples (Fig. 1.4a,b, 1.4d-e), particularly once there is no more age overlap between the compared groups (starting at **WT-3**). In model 1, where age is an unknown confounder, this drop in performance corresponds to skyrocketing rates of false positives (Fig. 1.4c) due to developmental changes being misattributed to the strain effect. In contrast, model 2 keeps a low false positive rate (< 5%) until the age overlap between groups disappears, and we observe a steady decline in the number of true positives (Fig. 1.4f) which likely corresponds to the loss of statistical power caused by the collinearity of age and strain variables (GRAHAM, 2003).

Integrating reference data restores analysis performance

As we previously reported (BULTEAU & FRANCESCONI, 2022), age correction by integrating reference data in the model strongly rescues DE analysis performance in cases where the compared groups have no overlap (Fig. 1.5). The improved method shown here recovers the analysis to the same extent as the one presented in our publication.

In sample sets with overlapping age (**GS**, **WT-1**, **WT-2**, Fig. 1.5), we see similar performance of DE analyses with or without reference data. Slight decreases (**GS** for both models) or increases (**WT-1**, **WT-2** in model 1) in performance are explained by reference batch effect or the addition of age to the model respectively.

We conclude that our updated method works without needing the previously introduced classifier, thus greatly simplifying its use.

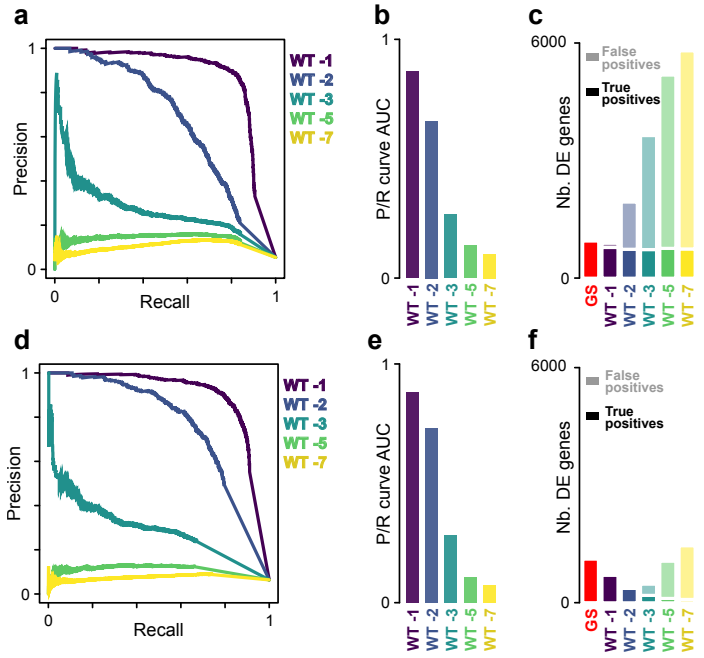


Figure 1.4 – Differential expression analysis performance drops with increasing age difference between compared groups.
Differential expression analysis is performed between mutant and WT sample groups with increasing age differences, as defined in Fig. 1.3.
a-c, Precision-Recall curves (**a**) and their AUC (**b**) of the strain effect in model 1 (no age), and true and false positives of DE genes with False Discovery rate (FDR) < 0.01 and $|logFC| > 1$ (**c**).
d-f, idem for model 2 (with age).
DE: differentially expressed; AUC: area under curve; logFC: log2-fold-change; GS: gold standard.

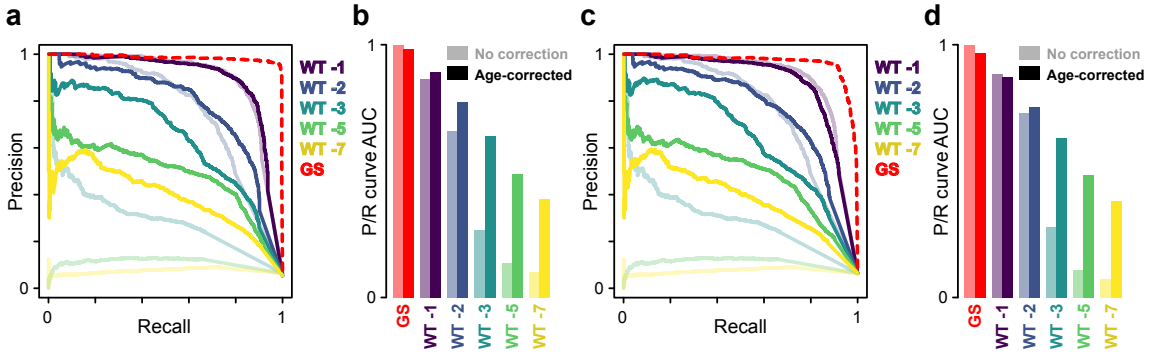


Figure 1.5 – Integrating reference data restores differential expression analysis performance
Differential expression analysis is performed between mutant and WT sample groups with increasing age differences, as defined in Fig. 1.3.
a,b, Precision-Recall curves (**a**) and their AUC (**b**) of the strain effect in mref in finding model 1 gold-standard genes. model 1 curves and AUC are shown in lighter color.
c,d, idem for mref in finding model 2 gold-standard genes, with model 2 curves and AUC shown in lighter color.
DE: differentially expressed; AUC: area under curve; GS: gold standard.

1.2.1.4 Conclusions

The first approach we proposed for age correction in DE analysis was difficult to use because it relied on an unconventional classifier, rather than standard p-values. Here, I presented a simpler method with the same principles and performance that produces meaningful p-values.

We show how a standard analysis becomes dominated by false positives with increasing (unknown) age differences between groups. Including age as a covariate is a simple solution to keep a low false positive rate when there is overlap between the compared groups, at the cost of reduced statistical power. When age is fully confounded with the variable of interest however, the sharp drop in performance of a DE analysis could only be recovered by integrating external reference data despite the potential batch effects and complexity this adds to the analysis.

Our approach works under the assumption that batch effects and, most importantly, development are conserved between the reference and all samples. Because of this, interaction between the variable of interest and development cannot be considered in the corrected models. The most common "3 controls vs. 3 perturbed" scenario is however unlikely to have the power to detect meaningful interaction if age and the variable of interest are fully confounded.

This strategy therefore provides a precious opportunity to re-analyze data where a standard DE analysis provided completely meaningless results due to strong unintended physiological differences between conditions. Given how prevalent this is in published expression data (Snoek et al., 2014), we believe our approach will lead to the discovery of many previously missed effects, as well as the debunking of erroneous findings.

1.2.2 Building more robust aging references

1.2.2.1 Introduction

Development is a robust process, with conserved gene expression across individuals, and to some extent, species (Levin et al., 2016). In stark contrast, aging is highly variable; individuals age at different rates but also in different ways due to the influence of diverse factors such as the diet (Hou et al., 2016), perception of the environment (B. Kim et al., 2020), parental effects (Greer et al., 2011), as well as unknown causes (W. B. Zhang et al., 2016). This makes predicting age from gene expression along aging more difficult than along development.

Nevertheless, we showed that RAPToR could accurately infer age from gene expression in aging worms, flies, and humans, recapitulating known biological effects (Bulteau & Francesconi, 2022). We achieved this by selecting informative genes to build the reference, empirically defined as those with a monotonous trend along aging. This gene selection approach is however very sensitive to the number of time points and samples in the reference, and thresholds for gene selection are defined arbitrarily. Therefore, we searched for a more data-driven method to select informative genes.

Here, I briefly show that combining *C. elegans* aging time series datasets of different profiling technologies, strains, and culture conditions allows us to find a common set of aging genes sufficient for staging with RAPToR.

1.2.2.2 Few shared dynamics among aging time-series

Table 1.1 lists aging time-series datasets used in this section and their characteristics. We remark that samples from Golden et al., 2008 distinguish themselves from the others, as gene expression from fertile, unperturbed, single wild-type worms was profiled along the whole lifespan of *C. elegans*. Others sterilize wild-type worms by adding FUDR (fluroxidine) to the culture medium (Hou et al., 2016), or work using sterile strains to avoid signal from embryos: *glp-1* is a heat-sensitive germline-less mutant (Suriyalaksh et al., 2022), *gon-2/gem-1* have degenerate gonads while *fem-3* mutants produce no oocytes (Hastings et al., 2019), and *rrf-3* is a heat-sensitive

Source	Tag[1]	Genotype	Culture condition	T (°C)	Chron. age span	Profiling technology	Accession
Golden et al., 2008	gol	Wild-type	plate	20	4-24 days post egg lay	microarray	GSE12290
Hou et al., 2016	hou	Wild-type	plate + FUDR	20	2-10 days of adulthood	microarray	GSE77111
Hastings et al., 2019	has	*gon-2(q388);gem-1(bc364)* and *fem-3(q20)*	plate	25	2-10 days post L1 feeding	RNA-seq	GSE124994
Byrne et al., 2020	byr	*rrf-3(pk1426)*	liquid	20	3-18 days of adulthood	RNA-seq	GSE93826
Suriyalaksh et al., 2022	sur	*glp-1(e2144)*	liquid	25	2-10 days post hatching	RNA-seq	GSE166512

Table 1.1 – *C. elegans* aging time-series profiling datasets used in this section
[1]: Used in subsequent plots to identify datasets.

sterile strain (Zhuang & Hunter, 2011). We note that *rrf-3* mutants are also RNAi hypersensitive (but under no RNAi treatment), and that the data is unpublished (Byrne et al., 2020).

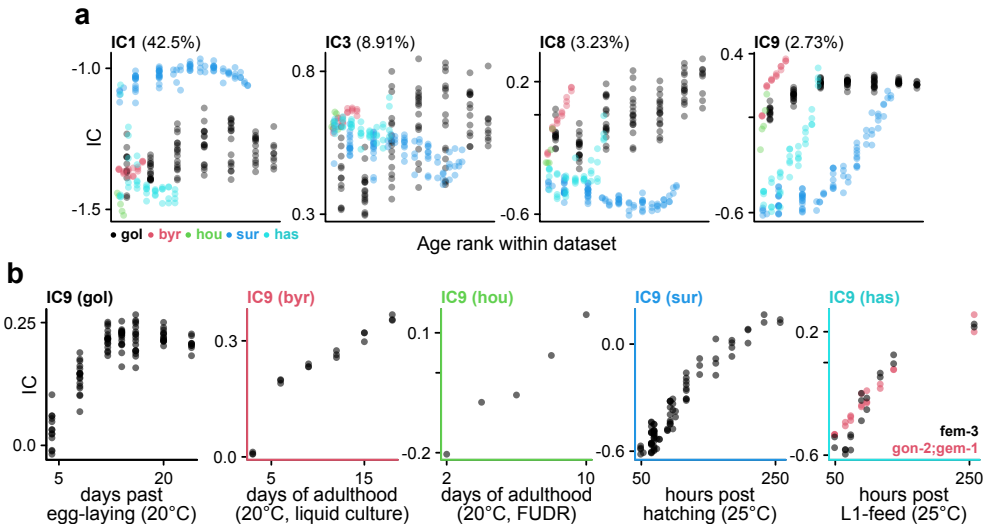


Figure 1.6 – Few shared aging expression dynamics across datasets
a, Selected independent components (IC) from a joint Independent Component Analysis (ICA) of aging time-series (Table 1.1). Percentage of total variance explained is indicated for each component. See also Sup. Fig. A.1.
b, IC9 loadings of samples from each dataset along their respective time axes.

An Independent Component Analysis (ICA) with all 5 datasets, shows very different expression dynamics between datasets (Fig. 1.6a, Sup. Fig. A.1). Beyond the expected batch effects, we find few components where most time-series have comparable dynamics, and a single component with a universally monotonic dynamic (IC9, Fig. 1.6b). Therefore, we selected the genes contributing highly to this component as the shared set of informative aging genes (see Methods).

The resulting set of 1652 genes (653 increasing, and 999 decreasing) is enriched in diverse biological processes and tissues (Appendix A, Sup. Fig. A.2), a lack of specificity suggesting we capture an overall aging signature as intended, and not process- or tissue-specific aspects that would be less robust to perturbation or across datasets.

1.2.2.3 A core set of informative genes stages aging across datasets

Staging single-individual data with the new gene set matches our published analysis, with individuals staged younger and older than their chronological ages behaving accordingly (Fig. 1.7a, Bulteau & Francesconi, 2022). Beyond behavior, age differences captured by staging within chronological time points, notably earlier ones, agree with age estimates of these individuals on

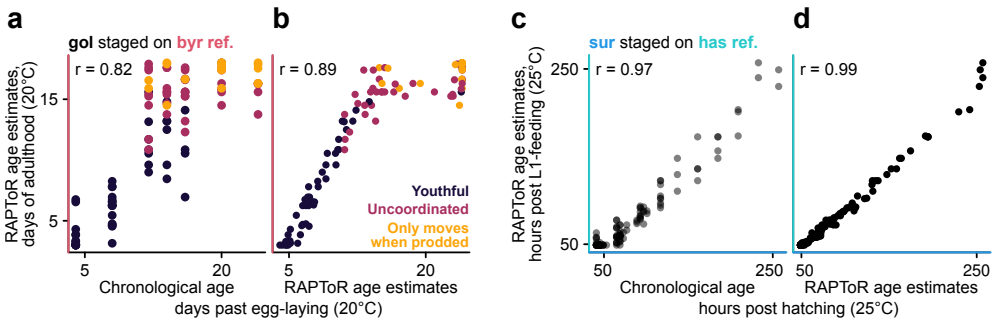


Figure 1.7 – RAPToR stages aging with the shared aging gene signature
a,b, Biologically meaningful differences in age are captured when staging single *C. elegans* individuals (gol, GOLDEN et al., 2008) on an independent reference (byr, BYRNE et al., 2020). Behavioral differences within chronological time points are explained by age estimates (**a**), and overall age differences largely correspond to those found by age estimates of samples on their own reference (**b**).
c,d, Bulk samples (sur, SURIYALAKSH et al., 2022) are well staged on an independent reference (has, HASTINGS et al., 2019, **c**), and remaining variability matches age differences captured by their own reference (**d**).

their own reference (Fig. 1.7b), further showing that RAPToR captures genuine differences in biological age between samples. Despite this variability being largely averaged out in bulk data, we find that the remaining differences between chronological and estimated age of samples (Fig. 1.7c). are similarly captured by their own reference (Fig. 1.7d).

To conclude, despite relatively few genes and dynamics matching between datasets, RAPToR stages samples well, and recapitulates the effects shown in our publication.

1.2.2.4 Conclusions

In this section, I defined a robust set of informative genes for staging aging with RAPToR through a joint analysis of multiple references, thus improving upon our previous empirical approach to select monotonous genes from a single time-series. Despite few shared expression dynamics across datasets, we found the selected gene set allows proper and precise staging of aging from gene expression. Although selecting informative genes in this way requires several independent datasets of aging transcriptomes and may therefore be limited to a few model organisms, selection is data-driven and the comparative analysis could thus also bring insights into understanding aging processes.

RAPToR currently stages individuals along a single aging trend, which is helpful to detect differences in aging speed along that trajectory. However, individuals might also age in different ways, along several different gene expression trajectories. While staging would be hindered by this with the current method, a version of RAPToR capable of staging across multiple trajectories could theoretically determine both the age of an individual, and *how* that individual is aging.

1.2.3 Multi-Trajectory RAPToR

1.2.3.1 Introduction

The timing of events is a fundamental and carefully-controlled aspect of development. In *C. elegans*, several molecular timing mechanisms are encoded by heterochronic genes that (among other roles) control the sequence of larval stages (MOSS, 2007). Modifications to heterochronic genes can cause developmental events to occur prematurely or late. For example, JOHNSON et al., 2009 characterized a *lin-14* mutation that causes a delay in vulva development resulting in asynchrony with germline maturation.

The relative timing of soma and germline tissues (soma-germline heterochrony) is not only genetically controlled, but is also a plastic trait in healthy worms (PEREZ et al., 2017; BULTEAU &

FRANCESCONI, 2022), that can be influenced by parental effects (PEREZ et al., 2021), or experimental conditions like drug exposure (BULTEAU & FRANCESCONI, 2022). Therefore, much like development, heterochrony between tissues can be a major confounder in gene expression analysis.

We showed that, RAPToR can infer a tissue-specific age from whole-organism profiling data by staging only with tissue specific genes (BULTEAU & FRANCESCONI, 2022). However, this only works if the relative development of tissues is simply shifted, and we cannot choose amongst different developmental trajectories. Furthermore, handling multiple trajectories would significantly improve the usability of RAPToR in single-cell data, where branching expression dynamics are common. Therefore, we investigated whether a "multi-trajectory RAPToR" could stage and select among multiple developmental trajectories from the gene expression of samples.

Below, I present a proof-of-concept that a simple strategy based on the principles of RAPToR is capable of staging early *C. elegans* embryo single cells, and correctly assigning them to their lineage.

1.2.3.2 Distinct trajectories in a single component space

HASHIMSHONY et al., 2015 profiled gene expression of single-cells from *C. elegans* embryos, tracing the main cell lineages (AB, C, D, E, MS, and P) from 2-cell to the end of gastrulation (330 min past 4-cell). An independent component analysis (ICA) on this data shows that the lineages branch into distinct trajectories along development in multiple components (Fig. 1.8). This means that expression differences between lineages are captured by the components, and therefore that reference interpolation for the lineages can be done within the same dimensionally-reduced space (see Appendix A, Sup. Fig. A.3, Methods).

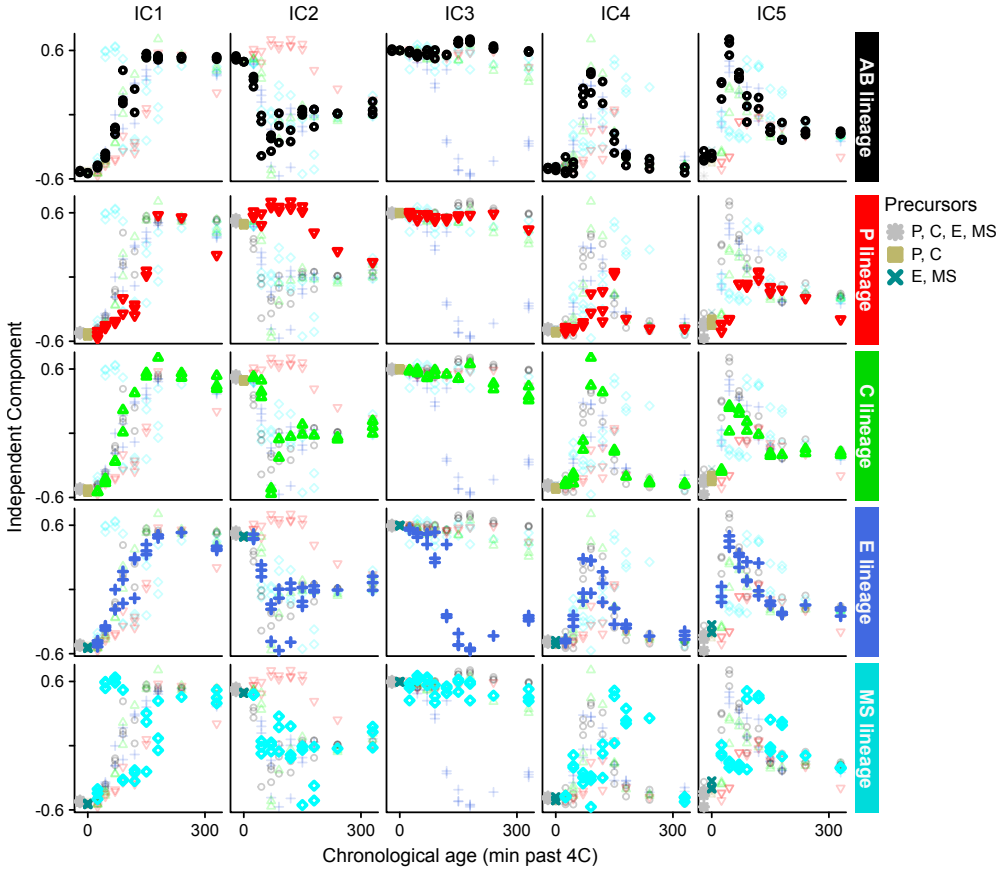


Figure 1.8 – ICA distinguishes cell lineages along early embryo development
AB, P, C, E, and MS lineages highlighted on the 5 first Independent Components (IC) of an ICA extracting 40 components. Data from HASHIMSHONY et al., 2015.

1.2.3.3 Cells from developing embryos are properly recognized and staged

After building an interpolated reference per lineage, we staged each cell on the reference of each lineage using standard RAPToR (i.e. spearman correlation between the cell and reference transcriptomes), and confirmed that cells are properly staged on their respective lineages (R^2 of chronological vs. inferred age > 0.9, Appendix A Sup. Fig. A.4). Then, for each cell, we simply assign the reference with the highest correlation at estimate (i.e. maximum correlation) as the inferred lineage. We find that most cells are indeed more highly correlated with their respective lineages than others. Furthermore, the preference of a cell's lineage over others (correlation difference) increases with developmental time, likely because cell types become more defined (Fig. 1.9).

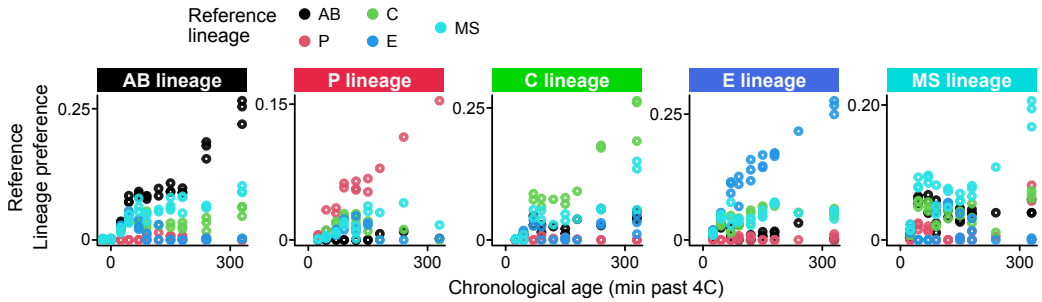


Figure 1.9 – Cells increasingly prefer the reference of their lineage along development
In each subplot, cells from a single lineage indicated in the top label are staged on references built from AB, P, C, E, and MS lineages (dot color). The reference lineage preference for each cell is defined as the correlation at estimate with each reference standardized to its lowest value (ie. $cor.est - min(cor.est)$).

1.2.3.4 Discussion

I have briefly shown that the RAPToR strategy can in principle be adapted to build references and stage distinct developmental trajectories from gene expression. Despite its crude implementation, this proof-of-concept shows that cell lineages in developing *C. elegans* embryos can be distinguished using ICA to build a reference per lineage, and that cells can be staged and assigned to their respective lineages with increasing clarity along age (i.e. differentiation).

Of course, we must properly verify these results using several independent datasets and other data types like whole-individual profiling data, as well as the feasibility of applying this approach to more than 5 different lineages with potentially subtler differences (e.g. distinguishing neuron types).

Rethinking the design and implementation of the tool could also substantially improve usability and performance. For example, although branching splines exist (Silverman & Wood, 1987), we could not find proper implementation in R and thus simply treated lineages as separate data. As a result, precursor cells (e.g. C-P, E-MS, and C-P-E-MS) are used several times for interpolation across references of the corresponding lineages and are thus redundant in early time points. Furthermore, given the typical size of single-cell datasets and the diversity of cell types, computing correlations across all cells and lineages is extremely inefficient and could surely be improved by a multi-step staging process starting with coarser references.

Nevertheless, the results shown here are promising and could be used in tandem with methods that identify *de novo* trajectories from expression data – e.g. Slingshot (Street et al., 2018), Monocle (Cao et al., 2019) – to build references capable of precisely staging the age and characterizing the type of a single cell from its transcriptome.

1.2.4 Staging tissue samples on whole-organism data

1.2.4.1 Introduction

There are abundant time-series profiling experiments of whole organisms, especially models, in the literature (BAUGH et al., 2003; REINKE et al., 2004; D. H. KIM et al., 2013; HENDRIKS et al., 2014; LEVIN et al., 2016; HASHIMSHONY et al., 2015; MEEUSE et al., 2020; SURIYALAKSH et al., 2022 to cite a few for just *C. elegans*). However, the same cannot be said for dissected tissues or dissociated cells from tissues (FOX et al., 2007 and EDWARDS et al., 2021 are rare examples). Therefore, although we showed RAPToR works in dissected tissue samples when using a reference of similar nature (BULTEAU & FRANCESCONI, 2022), actual use in this context is unlikely, and staging dissected tissue samples on whole-organism data is liable to several biases. First, although tissues have specific transcriptomic signatures, few genes are uniquely expressed in a given tissue (KALETSKY et al., 2018), meaning that the detected expression level of most genes is the combination of multiple tissues in whole-organism data. Then, dissecting or dissociating biological material applies a stress on cells that is reflected in gene expression (BRINK et al., 2017; MACHADO et al., 2021). Furthermore, cell types within a tissue may have distinct sensitivities and responses to the dissociation and collection process (BRINK et al., 2017; DENISENKO et al., 2020). Therefore, it is not trivial to assume that gene expression from dissected or dissociated cells from tissues is appropriate for staging on whole-organism data.

Despite this, in a collaboration with (now Dr.) Charline Roy and her supervisor Dr. Florence Solari, I had the opportunity to stage dissociated muscle cells from *C. elegans* young adults on whole-organism data with surprising accuracy using RAPToR. With their permission, I briefly describe the insights gained from this below.

1.2.4.2 Muscle cells from long-lived daf-2 mutants appear younger than wild-type

In the context of their research on aging muscles (ROY et al., 2022; ROY, 2022), Dr. Roy and Dr. Solari dissociated and collected *C. elegans* muscle cells, and profiled their gene expression in bulk in wild-type (WT) and long-lived *daf-2* mutant adults at days 0, 1, and 6 of adulthood (D0, D1, D6 respectively) in triplicate. Given the lifespan extension of *daf-2* mutants, our aim was to ensure that the developmental age of WT and *daf-2* samples matched within each time point to avoid measuring differential expression caused by developmental differences.

We staged the D0 and D1 samples that are within the range of our developmental references (see Methods) using all available genes, and found that both mutant and WT groups have the expected age at D0 (45-50 h post-hatching, Fig. 1.10). However at D1, although WT cells are properly staged 24 h later, daf-2 muscle cells appear nearly 10 hours younger, suggesting a slowed aging rate similar to the expected lifespan extension of *daf-2(e1370)* (GEMS et al., 1998). We could not successfully stage D6 samples on aging references (data not shown) to confirm this delay at later time points, perhaps because the combination of staging aging and tissue samples on whole-organism data is testing the limits of our approach.

When collecting dissociated muscle cells, contamination by sperm cells is difficult to avoid due to their small size and adhesive properties. As our collaborators had already noticed sperm-specific genes in their data, notably at D1 which are collected during spermatogenesis in hermaphrodites (see IC1 reference dynamic, Fig. 1.10), we were worried this could bias staging since the extent of contamination (and therefore, gene expression signal) can vary from sample to sample. Therefore, we confirmed the age of samples by staging without germline genes (REINKE et al., 2004) with very similar results (Fig. 1.10).

1.2.4.3 Staging muscle-cell bulk samples from sperm contamination

To our surprise, we could even stage samples using only germline, and even sperm-specific genes (Fig. 1.10). Genes related to spermatogenesis have a characteristic bell-curve dynamic, so we wondered how RAPToR could determine "which side" of the bell curve the samples lie on. By

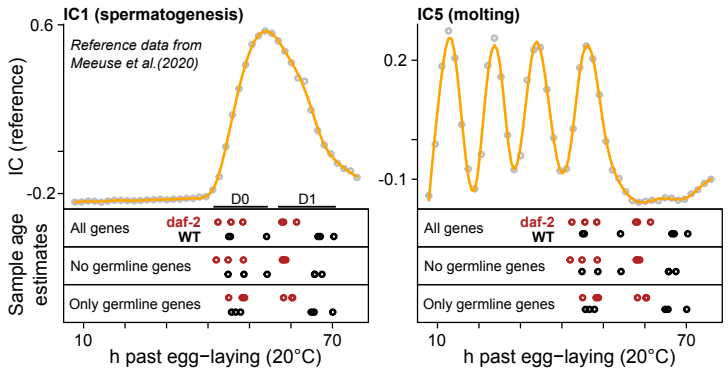


Figure 1.10 – RAPToR sucessfully stages dissociated muscle cell samples on a whole-organism reference
Selected Independent Components (IC) from an ICA on reference data plotted along reference time, indicating landmark gene expression dynamics of spermatogenesis (IC1) and molting cycle (IC5). Below ICs, age estimates of *daf-2* and WT bulk dissociated muscle cell samples at D0 and D1, when staged with all genes, without and restricting to germline genes (as defined as defined in REINKE et al., 2004).

clustering spermatogenesis genes in the reference (Fig. 1.11), we found that subtle timing differences in the bell-curve are sufficient to allow staging. Indeed, staging with sperm genes from a single cluster does not work, while a random gene set of similar size across all clusters does (Table 1.2). We therefore believe that regardless of the amount of sperm contamination, the relative expression level of genes within each sample is sufficient for staging.

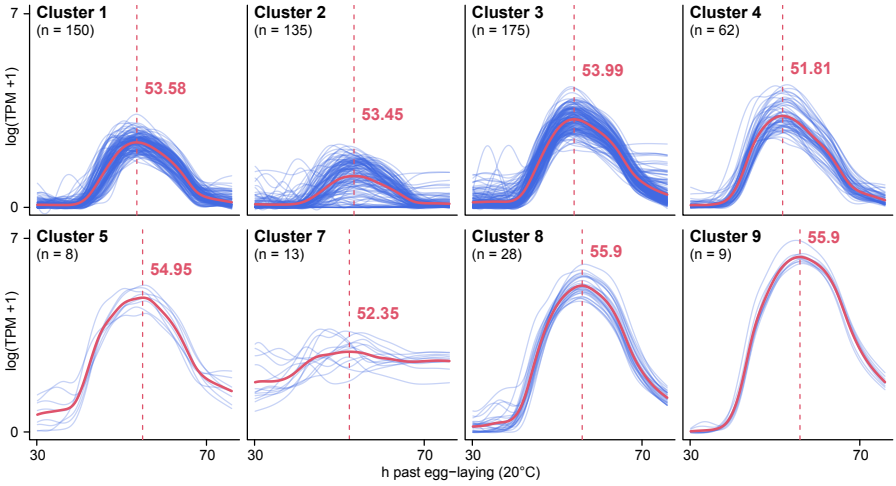


Figure 1.11 – Subtle differences in the timing of expression peak in spermatogenesis genes
Clustering of spermatogenesis genes, as defined by REINKE et al., 2004, in the interpolated reference (built from data by MEEUSE et al., 2020) shows slight shifts in peak timing of the bell-curve. Blue lines are individual genes, red line is the average per cluster. Timing of the peak of average expression (vertical dashed line) is indicated for each cluster.

1.2.4.4 Conclusions

I have shown that despite potential biases, we could successfully stage bulk dissociated muscle cell samples against a whole-organism reference from their gene expression (and that of their sperm contaminants). RAPToR could therefore be useful even when a reference of matching sample type is not available. Furthermore, we note that staging allowed us to notice that the labels of two samples were inadvertently swapped by the sequencing platform (F. Solari, personal communication, 2022).

We remark that D0 and D1 samples were collected during the young-adult to adult transition, a developmental window characterized by particularly intense gene expression changes including

Cluster	Gene set size	Correlation (r) with age estimates using all sperm genes			
		Cluster genes	Random sperm genes		
			set 1	set 2	set 3
1	150	0.39	0.98	0.97	0.98
2	129*	-0.04	0.97	0.96	0.84
3	175	0.18	0.98	0.98	0.84
4	62	0.12	0.48	0.61	0.51

Table 1.2 – Genes across multiple spermatogenesis clusters are required for staging
Correlation between age estimates using all spermatogenesis genes and age estimates using either spermatogenesis genes of a single cluster (1-4, see Fig. 1.11) or random spermatogenesis gene sets of equal size sampled across all clusters.
*: 129 out of 135 genes overlap between the data and reference.

(but not limited to) spermatogenesis and the onset of germline proliferation (SNOEK et al., 2014; FRANCESCONI & LEHNER, 2014). It is therefore possibly more challenging to stage tissue samples from other development stages on whole-organism references, and may partly explain why we failed to stage D6 samples.

As such, validating other sample types (e.g. dissected tissue), other tissues, and other organisms would be required to generalize whether RAPToR is capable of staging tissue samples on whole-organism data, or at least delimit the cases where it is possible. Nevertheless, the working example shown here is encouraging and suggests we can further broaden the already large scope of application of our method.

1.3 Discussion

Throughout this chapter, I have shown how age can be predicted from the transcriptome with RAPToR. I demonstrated the usefulness of this ability through several applications, relying on the fact that the method is extremely robust and can be applied in diverse contexts.

Indeed, we showed that RAPToR can infer age in all tested organisms, with bulk, single-individual, dissected-tissue, or single-cell expression profiles, by staging on appropriate references. To an extent, references can also be used to stage samples of different species and sample types, for example staging tissue samples on whole-organism reference data.

RAPToR age estimates make it possible to detect the effects of perturbations on developmental speed, and take advantage of age variation among samples that would hinder analyses to instead improve statistical power to detect differential expression (Bulteau & Francesconi, 2022). Here, we characterized how the power of a differential expression analysis decreases with growing age differences between compared groups, and how this can be salvaged by including inferred age as a covariate while age groups still overlap and by integrating reference data when it is no longer the case.

Both development and aging can be staged with RAPToR, although the latter requires a selection of informative genes to build references. We further investigated how to select these informative genes, improving upon our initially fragile threshold-based approach by combining data from several sources, and could reliably stage *C. elegans* samples aging in diverse conditions and genetic backgrounds. Taking advantage of these improvements, my supervisor Mirko Francesconi is also currently collaborating with Dr. Nicholas Stroustrup and his team at the Centre for Regulatory Genomics (CRG), to understand the differences between slow and fast-aging individuals amongst isogenic *C. elegans* populations.

We showed that single cells can in principle be staged and assigned to their respective differentiation trajectories during embryogenesis with a "multi-trajectory RAPToR" proof-of-concept. Although preliminary, these results could pave the way for a tool capable of using gene expression to infer precise age and cell type from single cells, as well as determine the age and developmental or aging path of individuals among several possible trajectories. Being able to capture and stage cell differentiation also supports the idea that RAPToR can be adapted to other processes with robust gene expression dynamics. We further point out that we have yet to explore the limits of the flexibility of RAPToR with data types other than the transcriptome, such as the methylome which has already proven predictive of aging (Wang et al., 2020; Bell et al., 2019), or the proteome for which data collection is gaining in throughput (Cui et al., 2022).

With RAPToR, my work integrates and reuses often under-exploited published data to offer solutions for addressing developmental bias in gene expression studies, an issue that has largely been ignored. In this context, I developed RAPToR as an `R` package [2], a programming language widely-used by biologists and bioinformaticians to analyze expression data, aiming for minimal `R` or coding experience requirements, and provide extensive documentation vignettes detailing general usage and application examples. Positive feedback on the package and its documentation, as well as growing usage of RAPToR in *C. elegans* (G. Zhang et al., 2022; Bell et al., 2023; Kim et al., 2023), humans (Hagan et al., 2022), and non-models (Sinigaglia et al., 2022) is therefore encouraging.

Beyond its benefits for analyzing existing data, staging post-profiling with RAPToR also allows us to consider novel experimental designs. Notably, as shown in following chapters, it eliminates the need for synchronization or for tedious and potentially difficult steps of accurate staging before profiling, greatly simplifying collection in large-scale single-organism profiling experiments.

[2] available at `https://www.github.com/LBMC/RAPToR`

1.4 Methods

All the analyses were performed in R 4.1.2.

1.4.1 Data loading and pre-processing

Expression data and sample metadata were downloaded from GEO (accession codes specified where relevant) using the GEOquery R package (v2.62.2). Transcript or probe IDs were converted to WormBase gene IDs, and counts to transcript per million (TPM) as previously described (BULTEAU & FRANCESCONI, 2022). Expression values (TPM or microarray expression) are log-transformed $log(X+1)$, and quantile-normalized with *normalizeBetweenArrays()* function of `limma` (v3.56.1), unless otherwise specified.

1.4.2 DE correction

Code to reproduce DE correction is available in the "*DE-correction*" vignette of the RAPToR package (v1.2.0) (BULTEAU & FRANCESCONI, 2022) upon which the DE correction section is largely based. With RAPToR installed, the vignette can be accessed with `vignette('RAPToR-DEcorrection')`, or `vignette('RAPToR-DEcorrection-pdf')` in an R console.

1.4.2.1 Sample subset definition

Age of sample from *C. elegans* wild-type and *xrn-2* mutant time series (MIKI et al., 2017, accession GSE97775) was estimated the *ae()* function of RAPToR, on the `'Cel_larv_YA'` reference from the wormRef data-package (v0.5) interpolated to 600 time points.

The gold-standard samples were chosen as the set of 3 time points with closest-matching age estimates between WT and mutants, and "shifted" subsets by sliding the window of WT time points back by 1, 2, 3, 5, and 7 time points from the gold standard.

1.4.2.2 Differential expression analysis

Genes with no ID match in the reference, or with less than 5 counts in all samples were dropped, leaving 17659 for analysis. Non-corrected DE analysis is performed on all sample subsets defined above on raw counts using the standard DESeq2 (LOVE et al., 2014) (v1.34.0) Wald test workflow, specifying models with strain only, or strain and age for model 1 and model 2 respectively, and `fitType='local'`. Genes are considered differentially expressed with FDR < 0.01, and $|log\text{FC}| \geq 1.0$.

1.4.2.3 Precision-recall assessment

For each DE analysis, FDR values of the strain coefficient were set to 1 for genes with $|log\text{FC}| < 1.0$, and the performance of the resulting classifier to detect genes in the appropriate gold-standard set (model 1 or model 2) was evaluated using precision ($\frac{true\ positives}{true\ positives + false\ positives}$) and recall ($\frac{true\ positives}{true\ positives + false\ negatives}$) with the *predict()* function of the ROCR package (v1.0-11).

1.4.2.4 Age correction

For each set of samples (including WT and mutant samples), we define the reference window to include as the range of age estimates widened by a 1h margin on either side. We transform the interpolated reference data to artificial counts assuming a fixed library size of 25×10^6 counts per sample and a fixed number of reads 'per gene length' defined by the median of available gene lengths:

$$artificialCounts = \frac{interpolatedTPM}{10^6} \times \frac{25 \times 10^6}{median(geneLengths)} \times geneLengths \quad (1.1)$$

The artificial count matrix is then rounded to the nearest integer and joined to the sample count matrix and a DE analysis is performed by fitting a DESEq2 model including batch (between reference and sample data), the variable of interest (strain) where reference data is grouped together with the control, and age modeled with splines (*ns()* function in the splines package). To select the optimal spline degree of freedom for each window, we minimized the residual SSQ of a linear model fit on the reference window only, resulting in 3, 4, 4, 4, 5, and 5 degrees of freedom for the gold standard, WT-1, -2, -3, -5, and -7 sample sets respectively. Gene dispersions are inferred using the *estimateDispersions()* function with `fitType='local'` on a DESeq2 model without reference data and with only a strain term, before being injected into the full model and running *nbinomWaldTest()*.

1.4.3 Aging

1.4.3.1 Selecting informative genes in aging data

Aging time-series datasets specified in Table 1.1 were pre-processed as described above, joined with overlapping IDs, leaving 9523 genes, and quantile-normalized together using *normalizeBetweenArrays()* from `limma`. Only ad-libitum fed samples from HOU et al., 2016 were used.

We then performed a centered PCA (*prcomp*) to determine that 14 components are sufficient to explain at least 90% of variance in the data, and performed an independent component analysis (ICA) extracting 14 components with the *icafast* function of the `ica` package (v1.0-3).

Informative genes were defined as those with absolute loading on Independent Component 9 > 1, resulting in 1652 genes.

Gene ontology and tissue enrichment of these monotonically increasing (653) and decreasing (999) genes against the overlapping gene background was performed using the local version of the wormbase `tea` tool (ANGELES-ALBORES et al., 2018), with a q-value threshold of 0.05.

1.4.3.2 Staging aging datasets

Aging references were built with the *ge_im()* and *make_ref()* functions of RAPToR for each time-series, restricting to the informative gene set and using a single PCA component for interpolation, with formulas as follows. `"X~(s(age, bs='cr', k=k")`, with k set to 5, 3, 4, 5, and 4 for `gol`, `byr`, `hou`, `sur`, and `has` datasets respectively, with genotype included as a covariate for `has`.

Each dataset was then staged on each reference (including the one built from their own samples), and age estimates were compared along with chronological age with pearson correlation to assess staging performance. We remark that differing time units, starting times, age spans, and aging speeds (temperatures) of the time-series, complicate further comparison between age values.

1.4.4 Multi-trajectory RAPToR

Single-cell expression profiles (HASHIMSHONY et al., 2015, accession GSE50548) were filtered to remove a poor-quality sample with 99th quantile of spearman correlation with others below $mean(cor) - 2 \times sd(cor)$.

We defined timings for each blastomere stage based on the original publication, stages 1-11 correspond to -20, 0, 25, 45, 70,90, 120, 150, 180, 240, and 330 minutes past 4-cell respectively. Cell lineages reported by the authors were used.

40 components were extracted with and ICA using *icafast()*. We then built RAPToR references for each lineage (AB, C, P, E, MS) by interpolating on their respective samples (including precursors) within this component space. A single formula was used for all lineages and components, `"X~s(time, bs='cr', k=6)"`.

All cells were then staged on all lineage references. We then compared the correlation score at estimate of each cell against each lineage reference, noting the increasing correlation gap between the correct lineage and others along development.

1.4.5 Tissue sample staging

Dr. Roy and Dr. Solari kindly provided us with the RNA-seq count data of their WT and *daf-2(e1370)* dissociated muscle cell bulk samples, collected at D0, D1, and D6 of adulthood. We processed the data as described above, keeping only D0 and D1 samples for the analyses below, and staged them as is with *ae()* of RAPToR on the `Cel_larv_YA` reference from wormRef (data from MEEUSE et al., 2020) interpolated to 500 time points, using either all available genes, only germline genes (defined as the union of 'germline_intrinsic', 'spermatogenesis_enriched', and 'oogenesis_-enriched' gene sets defined in REINKE et al., 2004), or no germline genes.

ICA components showing landmark expression dynamics and reference interpolation were directly extracted from the `Cel_larv_YA` RAPToR reference.

1.4.5.1 Staging with sperm genes

Selecting interpolated expression from the reference after 30h post-hatching, we clustered sperm genes ('spermatogenesis_enriched', REINKE et al., 2004) by expression dynamic using *hclust()* on a distance matrix computed by *dist()* without centering expression values. 12 clusters were kept to have at least 5 genes per cluster. Then, we computed the average expression dynamic of each cluster and noted the timing of its maximum.

To test whether slight shifts in dynamics could explain successful staging with sperm genes, we successively restricted the gene set to each cluster with at least 50 genes (clusters 1-4) for staging, and compared results with staging using 3 randomly selected sets of sperm genes with the same size. We then reported pearson correlation between the resulting estimates and age inferred using all sperm genes.

References

ANGELES-ALBORES, DAVID et al. (2018). "Two new functions in the WormBase Enrichment Suite". In: *microPublication Biology* 2018.

BAUGH, L RYAN et al. (2003). "Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome". In.

BELL, AVERY DAVIS et al. (2023). "Beyond the reference: gene expression variation and transcriptional response to RNA interference in Caenorhabditis elegans". In: *G3: Genes, Genomes, Genetics* 13.8, jkad112.

BELL, CHRISTOPHER G et al. (2019). "DNA methylation aging clocks: challenges and recommendations". In: *Genome biology* 20, pp. 1–24.

BRINK, SUSANNE C VAN DEN et al. (2017). "Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations". In: *Nature methods* 14.10, pp. 935–936.

BULTEAU, ROMAIN and MIRKO FRANCESCONI (Aug. 2022). "Real age prediction from the transcriptome with RAPToR". en. In: *Nature Methods* 19.8, pp. 969–975. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01540-0.

BYRNE, J et al. (Jan. 2020). *Gene changes over aging in the C.elegans rrf-3(pk1426) mutant*. unpublished. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93826 (visited on 09/22/2022).

CAO, JUNYUE et al. (2019). "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745, pp. 496–502.

CLARKE, KR and RH GREEN (1988). "Statistical design and analysis for a'biological effects' study". In: *Marine Ecology Progress Series*, pp. 213–226.

CUI, MIAO, CHAO CHENG, and LANJING ZHANG (2022). "High-throughput proteomics: a methodological mini-review". In: *Laboratory Investigation* 102.11, pp. 1170–1181.

DENISENKO, ELENA et al. (2020). "Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows". In: *Genome biology* 21.1, pp. 1–25.

EDWARDS, STACEY L et al. (2021). "Insulin/IGF-1 signaling and heat stress differentially regulate HSF1 activities in germline development". In: *Cell reports* 36.9.

FOX, REBECCA M et al. (2007). "The embryonic muscle transcriptome of Caenorhabditis elegans". In: *Genome biology* 8, pp. 1–20.

FRANCESCONI, MIRKO and ROMAIN BULTEAU (2022). "Inferring biological age from the transcriptome with RAPToR". In: *Nature Methods* 19.8, pp. 936–937. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01542-y.

FRANCESCONI, MIRKO and BEN LEHNER (2014). "The effects of genetic variation on gene expression dynamics during development". In: *Nature* 505.7482, pp. 208–211.

GEMS, DAVID et al. (1998). "Two pleiotropic classes of daf-2 mutation affect larval arrest, adult behavior, reproduction and longevity in Caenorhabditis elegans". In: *Genetics* 150.1, pp. 129–155.

GOLDEN, TAMARA R et al. (2008). "Age-related behaviors have distinct transcriptional profiles in Caenorhabditis elegans". In: *Aging cell* 7.6, pp. 850–865.

GRAHAM, MICHAEL H (2003). "Confronting multicollinearity in ecological multiple regression". In: *Ecology* 84.11, pp. 2809–2815.

GREER, ERIC L et al. (2011). "Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans". In: *Nature* 479.7373, pp. 365–371.

HAGAN, THOMAS et al. (2022). "Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses". In: *Nature Immunology* 23.12, pp. 1788–1798.

HASHIMSHONY, TAMAR et al. (2015). "Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer". In: *Nature* 519.7542, pp. 219–222.

HASTINGS, JANNA et al. (2019). "Multi-omics and genome-scale modeling reveal a metabolic shift during C. elegans aging". In: *Frontiers in molecular biosciences* 6, p. 2.

HENDRIKS, GERT-JAN et al. (2014). "Extensive oscillatory gene expression during C. elegans larval development". In: *Molecular cell* 53.3, pp. 380–392.

HOU, LEI et al. (2016). "A systems approach to reverse engineer lifespan extension by dietary restriction". In: *Cell metabolism* 23.3, pp. 529–540.

JOHNSON, RYAN W et al. (2009). "The Caenorhabditis elegans heterochronic gene lin-14 coordinates temporal progression and maturation in the egg-laying system". In: *Developmental Dynamics* 238.2, pp. 394–404.

KALETSKY, RACHEL et al. (2018). "Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression". In: *PLoS genetics* 14.8, e1007559.

KIM, BYOUNGHUN et al. (2020). "Regulatory systems that mediate the effects of temperature on the lifespan of Caenorhabditis elegans". In: *Journal of Neurogenetics* 34.3-4, pp. 518–526.

KIM, DONG HYUN, DOMINIC GRÜN, and ALEXANDER VAN OUDENAARDEN (2013). "Dampening of expression oscillations by synchronous regulation of a microRNA and its target". In: *Nature genetics* 45.11, pp. 1337–1344.

KIM, EUNAH et al. (2023). "Mitochondrial aconitase suppresses immunity by modulating oxaloacetate and the mitochondrial unfolded protein response". In: *Nature Communications* 14.1, p. 3716.

LEVIN, MICHAL et al. (2016). "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596, pp. 637–641.

LOVE, MICHAEL I, WOLFGANG HUBER, and SIMON ANDERS (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12. Publisher: BioMed Central, pp. 1–21.

MACHADO, LÉO et al. (2021). "Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation". In: *Cell Stem Cell* 28.6, pp. 1125–1135.

MEEUSE, MILOU WM et al. (2020). "Developmental function and state transitions of a gene expression oscillator in Caenorhabditis elegans". In: *Molecular systems biology* 16.7, e9498.

MIKI, TAKASHI S, SARAH H CARL, and HELGE GROSSHANS (2017). "Two distinct transcription termination modes dictated by promoters". In: *Genes & development* 31.18, pp. 1870–1879.

MOSS, ERIC G (2007). "Heterochronic genes and the nature of developmental time". In: *Current Biology* 17.11, R425–R434.

PEREZ, MARCOS FRANCISCO et al. (2017). "Maternal age generates phenotypic variation in Caenorhabditis elegans". In: *Nature* 552.7683, pp. 106–109.

PEREZ, MARCOS FRANCISCO et al. (2021). "Neuronal perception of the social environment generates an inherited memory that controls the development and generation time of C. elegans". In: *Current Biology* 31.19, pp. 4256–4268.

REINKE, VALERIE et al. (2004). "Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans". In.

ROBINSON, MARK D, DAVIS J MCCARTHY, and GORDON K SMYTH (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1. Publisher: Oxford University Press, pp. 139–140.

ROY, CHARLINE (2022). "Rôle du récepteur de l'insuline/IGF-1 DAF-2 dans le contrôle du vieillissement musculaire et son impact sur le transcriptome". PhD thesis. Lyon 1.

ROY, CHARLINE et al. (2022). "DAF-2/insulin IGF-1 receptor regulates motility during aging by integrating opposite signaling from muscle and neuronal tissues". In: *Aging Cell* 21.8, e13660.

SILVERMAN, BW and JT WOOD (1987). "The nonparametric estimation of branching curves". In: *Journal of the American Statistical Association* 82.398, pp. 551–558.

SINIGAGLIA, CHIARA et al. (2022). "Distinct gene expression dynamics in developing and regenerating crustacean limbs". In: *Proceedings of the National Academy of Sciences* 119.27, e2119297119.

SNOEK, L BASTEN et al. (2014). "A rapid and massive gene expression shift marking adolescent transition in C. elegans". In: *Scientific reports* 4.1, pp. 1–5.

STREET, KELLY et al. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC genomics* 19, pp. 1–16.

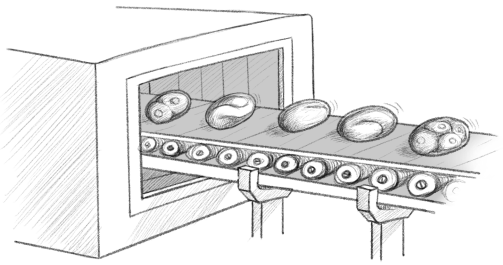
SURIYALAKSH, MANUSNAN et al. (2022). "Gene regulatory network inference in long-lived C. elegans reveals modular properties that are predictive of novel aging genes". In: *Iscience* 25.1.

WANG, TINA et al. (2020). "Quantitative translation of dog-to-human aging by conserved remodeling of the DNA methylome". In: *Cell systems* 11.2, pp. 176–185.

ZHANG, GAOTIAN et al. (2022). "The impact of species-wide gene expression variation on Caenorhabditis elegans complex traits". In: *Nature communications* 13.1, p. 3462.

ZHANG, WILLIAM B et al. (2016). "Extended twilight among isogenic C. elegans causes a disproportionate scaling between lifespan and health". In: *Cell Systems* 3.4, pp. 333–345.

ZHUANG, JIMMY J and CRAIG P HUNTER (2011). "Tissue specificity of Caenorhabditis elegans enhanced RNA interference mutants". In: *Genetics* 188.1, pp. 235–237.



2. Streamlining high-throughput RNA-sequencing of single individuals

Contents

Abstract

The single-cell rush has pushed profiling technologies to the point where low quantities of input RNA and large-scale sample collection are no longer a bottleneck for RNA-sequencing. We sought to apply these recent developments to whole-individual profiling to enable cost-efficient

large scale study of inter-individual variation. In this chapter, I demonstrate that high throughput RNA-seq of *C. elegans* single embryos can be achieved with minimal adjustments to existing protocols developed for single-cell. First I show that single live wild-type embryos can be sorted with standard FACS. Second, I show that it is also possible to use FACS select embryos at specific stages from a mixed population using only physical parameters and autofluorescence without the need for fluorescent markers. Third, I adapted the Smart-seq3 protocol to profile single embryos at minimal cost and high complexity enabling cost effective scaling up of single individual profiling.

2.1 Introduction

The last decade has seen incredible achievements in single-cell technologies. Thanks to microfluidics and improvements in profiling sensitivity, quality, and throughput (STARK et al., 2019; ZIEGENHAIN et al., 2017), samples with low RNA input can be sorted in large numbers and profiled at high complexity and quality. RNA-seq library preparation and sequencing have also become much cheaper. For example, reagent costs between Smart-seq2 (PICELLI et al., 2014, 35€ per library) and Smart-seq3 (HAGEMANN-JENSEN et al., 2020, <1€) protocols have been divided 30-fold. Despite this however, the most recent efforts in high-throughput single animal profiling still do not include the latest developments in low input RNA library preparation, and still rely on manual sorting (LEVIN et al., 2016; MACCHIETTO et al., 2017; PEREZ-MOJICA et al., 2023).

Standard Fluorescence-Activated Cell Sorting (FACS) instruments have been used to sort single cells (JAITIN et al., 2014) and could in theory be used to efficiently sort small single animals such as *C. elegans*. FACS has indeed previously been used to collect large populations of live synchronized *C. elegans* embryos at precise stages (STOECKIUS et al., 2009), as well as larvae mutants among a mixed population (FERNANDEZ et al., 2010) by using fluorescent markers, though not to isolate single individuals. More recent studies have taken to high-throughput technologies (such as the COPAS Biosort [1]) to study single individuals, but they largely focus on imaging (O'REILLY et al., 2014; KWON et al., 2018) or measuring other phenotypes such as life span (STROUSTRUP et al., 2013), and to our knowledge never on single-worm collection to profile gene expression.

Rather than using the FACS to get synchronized populations for expression profiling we can now simply sort single animals and precisely infer their age from gene expression (BULTEAU & FRANCESCONI, 2022). Inferring age on large numbers of single individuals profiles can then be used to compare expression dynamics between conditions and genetic backgrounds (FRANCESCONI & LEHNER, 2014) with significant improvements over the standard "3 vs. 3 controls". Furthermore, inter-individual variation in gene expression can also be studied, while controlling for variation due to interindividual differences in physiological age.

In this chapter, I demonstrate that high throughput RNA-seq of *C. elegans* single embryos can be achieved with minor adjustments to existing protocols developed for single-cell. First, I show single live wild-type embryos can be efficiently sorted using standard FACS. Second, thanks to flow cytometry coupled to bright field imaging, I developed a robust marker-free strategy to sort *C. elegans* embryos of any developmental stage just using a combination of physical and autofluorescence parameters. Third, I adapt the Smart-seq3 protocol (HAGEMANN-JENSEN et al., 2020) to scale up gene expression profiling of single embryos with full transcript length and high complexity at minimal cost.

2.2 Flow cytometry with *C. elegans* embryos

As noted by STOECKIUS et al., 2009, *C. elegans* embryos (eggs) are particularly suited to flow cytometry and sorting. Embryos are 50 µm in length and 30 µm in diameter, fitting snugly within the 10-100 µm size range of mammalian cells (Fig. 2.1, GINZBERG et al., 2015), their size stays constant during embryogenesis (unlike e.g. for zebrafish, KIMMEL et al., 1995), constrained by a resilient egg-shell which also makes them easy to isolate and purify with bleach. Starting from plates with egg-laying wild-type adults, we can therefore prepare embryo suspensions for sorting using a standard nematode bleaching protocol (see Methods), and directly pass the output through FACS instruments (Fig. 2.2a).

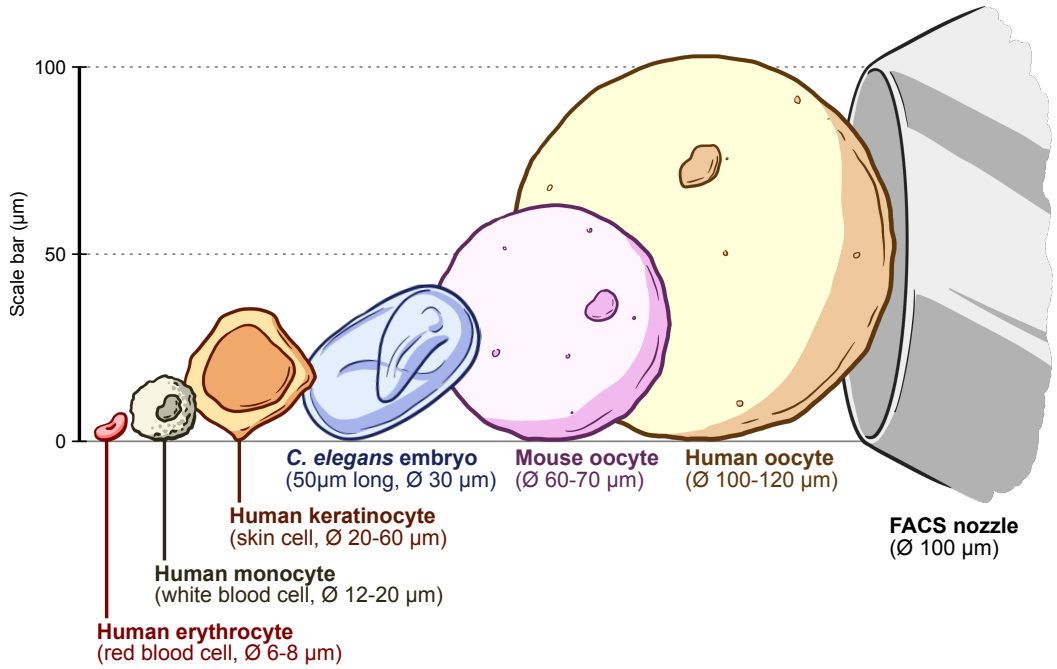


Figure 2.1 – Size comparison of *C. elegans* embryo and mammalian cells
Size measurements of human red and white blood cells from PRINYAKUPT & PLUEMPITIWIRIYAWEJ, 2015, human keratinocyte from HUZAIRA et al., 2001, *C. elegans* embryo from RIDDLE et al., 1997, and humans and mouse oocyte from GRIFFIN et al., 2006.

2.2.1 Embryos are reliably identified across instruments

Using an Attune^{TM} CytPix^{TM} flow cytometer capable of taking brightfield images, we could identify and separate embryos from debris and dissociated cells in the solution (Fig. 2.2b) using Forward and Side Scatter (FSC and SSC, respectively). We then distinguished between dead, unfertilized, and live embryos with their autofluorescence on blue color channels (Fig. 2.2c). The resulting population spans the whole of embryo development, from 1-cell stage to hatching (Fig. 2.2c-d). We then reproduced this gating strategy on a FACSAria IIµ to sort live embryos. Despite the differing channels and measurements between both instruments, we could reliably find similar clusters (Fig. 2.2e) with high reproducibility (Appendix B) and select live embryos for sorting without further adjustments. The amount of live embryos we recovered from a sample depended on the success of the bleach and amount of starting material. When bleaching a single plate with 100-150 egg-laying adults we could generally identify 500 to 1200 live embryos, at a rate of 1 to 2 per second (Appendix B).

2.2.2 Sorting efficiency in low volumes

If sorting many embryos in large buffer volumes and tubes is straightforward and efficient (FERNANDEZ et al., 2010), recent single-cell RNA-seq library preparation protocols use very low volumes to minimize costs, which could jeopardize this efficiency. Indeed, we initially noticed poor yield in collection volumes of 2.5 µL(around 10%, Fig. 2.3), and reasoned this was due to droplets with embryos drying out after landing on the wall of the tubes. We therefore tested ways to improve embryo recovery after sorting (see Methods).

Although efficiency wasn't significantly improved by longer post-sorting centrifugation or adding detergent to the sample, sorting in PCR strip caps (which have a wider surface area for 2.5 µL) improved efficiency to 30-60% (Fig. 2.3), thus confirming droplets must reach the buffer to recover embryos. Sorting in caps is however impractical, so we explored other solutions and

[1] now called COPAS FP-250, www.unionbio.com/copas/fp-250.aspx

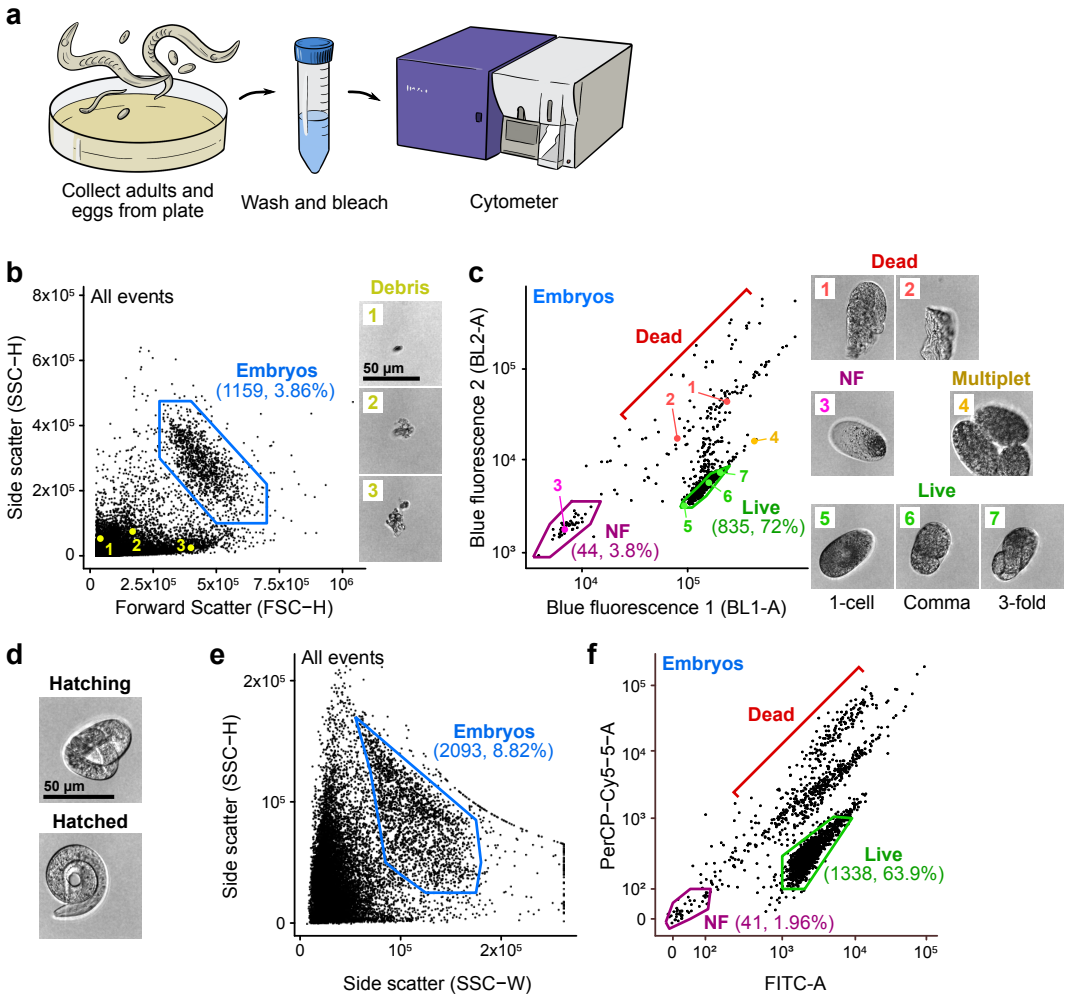


Figure 2.2 – FACS with *C. elegans* embryos
a, Cartoon of workflow for sorting *C. elegans* embryos.
b, With an Attune CytPix, embryos are separated from debris by gating on scatter parameters. Debris examples are numbered (1-3), with corresponding brightfield images shown on the right.
c, Dead, non-fertilized (NF), and live embryos gated from (**c**) are distinguished using autofluorescence. All embryo developmental stages can be found in the resulting "Live" embryo population. Examples of dead (1,2), unfertilized (3), multiplets (4) and single live embryos of various stages (5-7) are marked, with corresponding brightfield images shown on the right.
d, Brightfield images of hatching and hatched larvae taken with the CytPix (independent sample from b,c)
e,f, Equivalent selection of embryos with scatter parameters (**e**, as in **b**), and live embryos with autofluorescence (**f**, as in **c**) using a FACSAria IIµ.

In b-d, all images are 72 µm in total width.

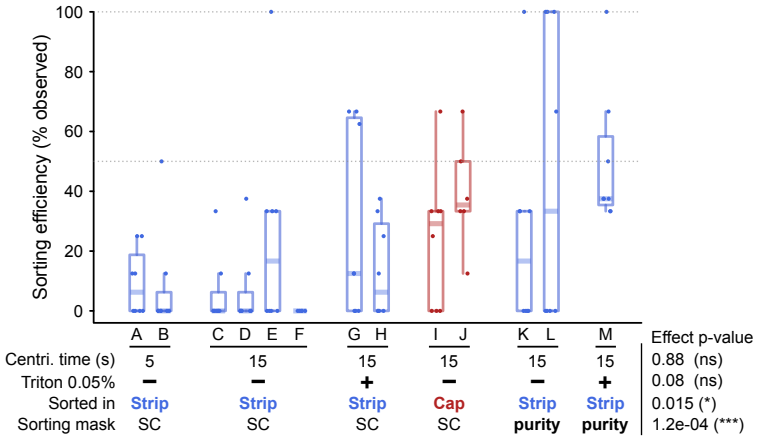


Figure 2.3 – Sorting efficiency of embryos in 2.5 µL
Effect of varying centrifugation time, adding detergent (Triton) in the solution, sorting in strips or caps, and sorting mask on sorting efficiency, defined as the ratio of sorted vs. observed embryos per well. Embryos were sorted in PCR strips with 2.5 µLof buffer using a FACSAria IIµ, and significance of the effects is evaluated with a linear model including all data (see Methods). .
Each box corresponds to a strip sorted independently, with n=8 wells except F and J where n=5 and n=6 respectively. Boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction. SC, single-cell; ns, non-significant.

were able to achieve similar efficiency by using a different sorting mask dubbed 'purity' (Fig. 2.3), which sorts two consecutive droplets per well. Although this mask theoretically increases the risk of sorting doublets, our rate of events (rarely above 50/s, of which embryos only constitute a small percentage) and the droplet rate (10,000/s) are orders of magnitude apart, making this essentially impossible. Furthermore, we never found extra embryos when sorting few or single embryos per well. We believe the increase in efficiency may instead result from the second droplet pushing embryos down into the bottom of the wells.

Sorting efficiency can likely still improve with further optimization. For example, increasing the nozzle diameter from 100 µm to 130 µm or decreasing the jet pressure, would result in bigger droplets that are less likely to stick to the edge of a tube during sorting. Robustly exceeding 50% yield should therefore be possible to make sorting viable.

2.2.3 Embryo development can be inferred from FACS measurements

Collecting embryos at a precise developmental stage currently requires fluorescent markers under control of stage-specific promoters, such as the promoter of *oma-1* (required for oocyte maturation) to select 1-cell stage embryos (Stoeckius et al., 2009). However, this requires mutant lines that complicate experimental designs, and can alter signals of interest e.g. due to the metabolic cost and potential toxicity of fluorescence for the embryo (Ansari et al., 2016).

We initially noticed a link between embryo developmental stage and multiple scatter and fluorescence channel measurements (Fig. 2.4a-b). Therefore, we explored the possibility of a marker-free sorting strategy targeting embryos at specific developmental stages by taking advantage of the brightfield images coupled with flow cytometry parameters provided by the CytPix. Classifying embryos from their images into 1-cell, 2-cell, 4-cell, 4-8-cell, 8-32-cell, >32-cell, Comma, 2-fold, and Late categories (Fig. 2.4b, see Methods) revealed that multiple scatter and fluorescence measurements are indeed strongly correlated with development (Fig. 2.4c).

Blue autofluorescence of accumulating gut lipid granules have previously been reported in *C. elegans* larvae (Clokey & Jacobson, 1986) and hinted at in mid to late embryo stages (Bossinger & Schierenberg, 1992), which is consistent with the positive correlation we observe between embryo age and blue and violet fluorescence measurements (BL1-A and VL3-A, respectively, Fig. 2.4c). However, the dynamics of embryo autofluorescence along development and across different wave-

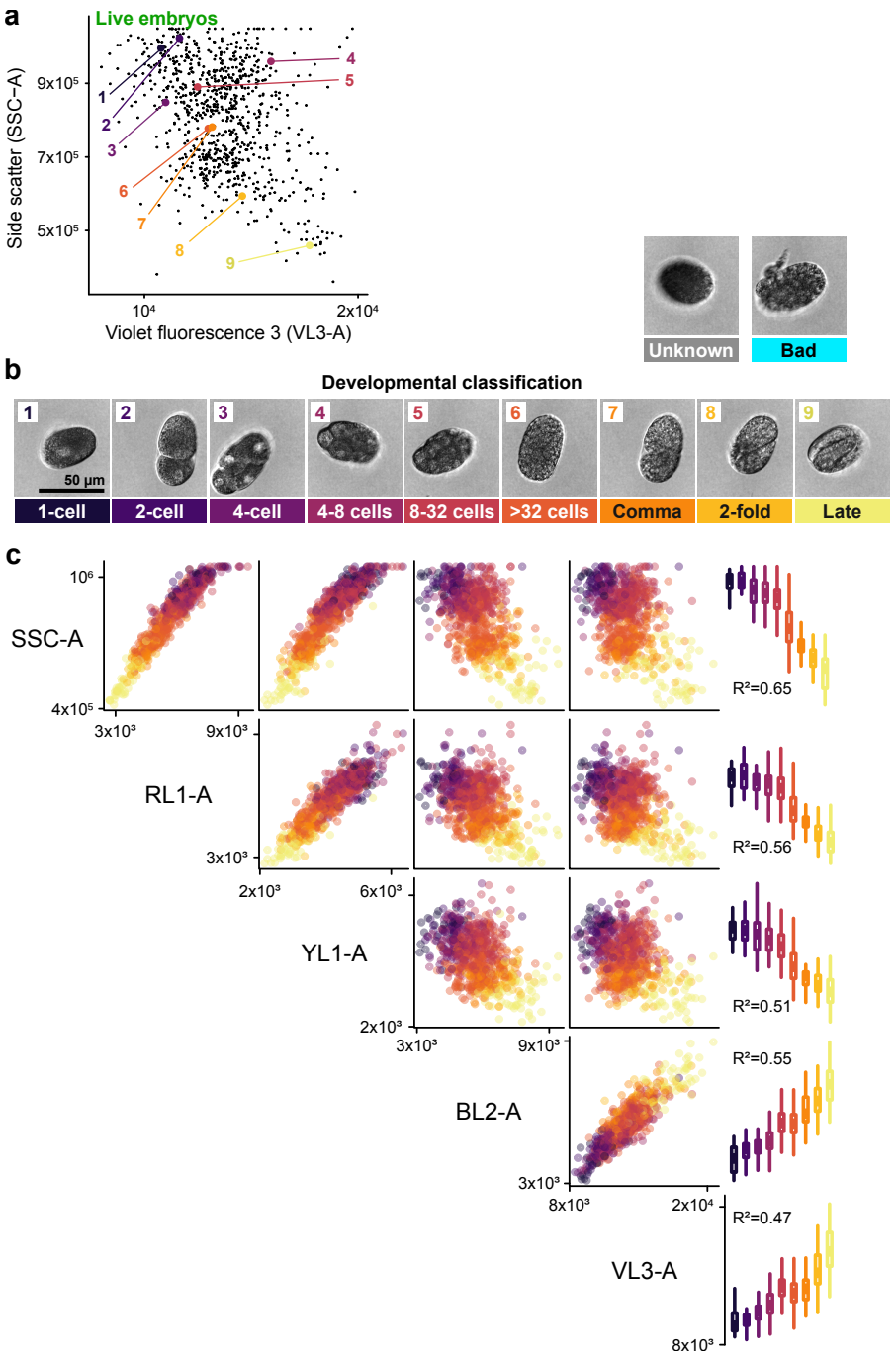


Figure 2.4 – Embryo development correlates with FACS measurements
a,b, Staging embryos from brightfield images (**b**) reveals that embryo development is correlated with scatter and fluorescence measurements (**a**).
c, Strong parameter predictors for embryo development. Colors and stages defined in (**b**), R^2 from linear model fits of each parameter by the developmental classification. Boxplots are n=22, 35, 48, 80, 157, 207, 67, 37, and 62 for 1-cell, 2-cell, 4-cell, 4-8 cells, 8-32-cells, >32 cells, Comma, 2-fold, and late respectively. Boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction.

lengths have yet to be properly characterized, perhaps due to autofluorescence being mostly considered as a hindrance rather than a proper subject of study in microscopic studies (HEPPERT et al., 2016; RODRIGUES et al., 2022). Our data therefore also provides the first multi-color (and scatter) description of embryo autofluorescence along the whole of embryogenesis

We then quantified how well combinations of FACS parameters could predict development, and found that nearly 100% of embryo age variation in a sample can be explained using all measured channels in a random forest model (R^2 = 0.97, Fig. 2.5a). The age of embryos from an independent sample is also reliably inferred from the same model (R^2 = 0.83, Fig. 2.5b), thus showing the relationship between development and FACS measurements is robust across experiments. Crucially, a simple linear model including side scatter and violet fluorescence (selected as the two most important predictors by the random forest model see Methods) can reliably distinguish embryo age within (R^2 = 0.79, Fig. 2.5c) and across (R^2 = 0.82, Fig. 2.5d) experiments, closely matching age inferred from the random forest model (R^2 = 0.90, Fig. 2.5g). As a result, embryo age progression can be mapped to the plane described by these two parameters (Fig. 2.5f-h, Methods), which makes it possible to sort embryos of targeted stages with simple gates.

To summarize, we developed a simple marker-free sorting strategy to reliably sort wild-type embryos at any developmental stage using side scatter and autofluorescence measurements of standard flow cytometers and FACS.

2.2.4 Conclusions

In this section, we have first shown that single live embryos can be easily collected and sorted with standard FACS in low volumes required for downstream RNA-seq profiling with recent protocols at around 50% efficiency, and with plenty of room to improve this efficiency. Second, we demonstrated that standard FACS measurements can be used to accurately predict embryo age, and therefore to obtain synchronized single-individual or bulk samples at any stage of embryo development without the need for fluorescent markers. Thus, standard flow cytometry and FACS coupled to transcriptomic profiling or bright field imaging provide simple and powerful marker-free strategies for large scale studies of inter-individual variation.

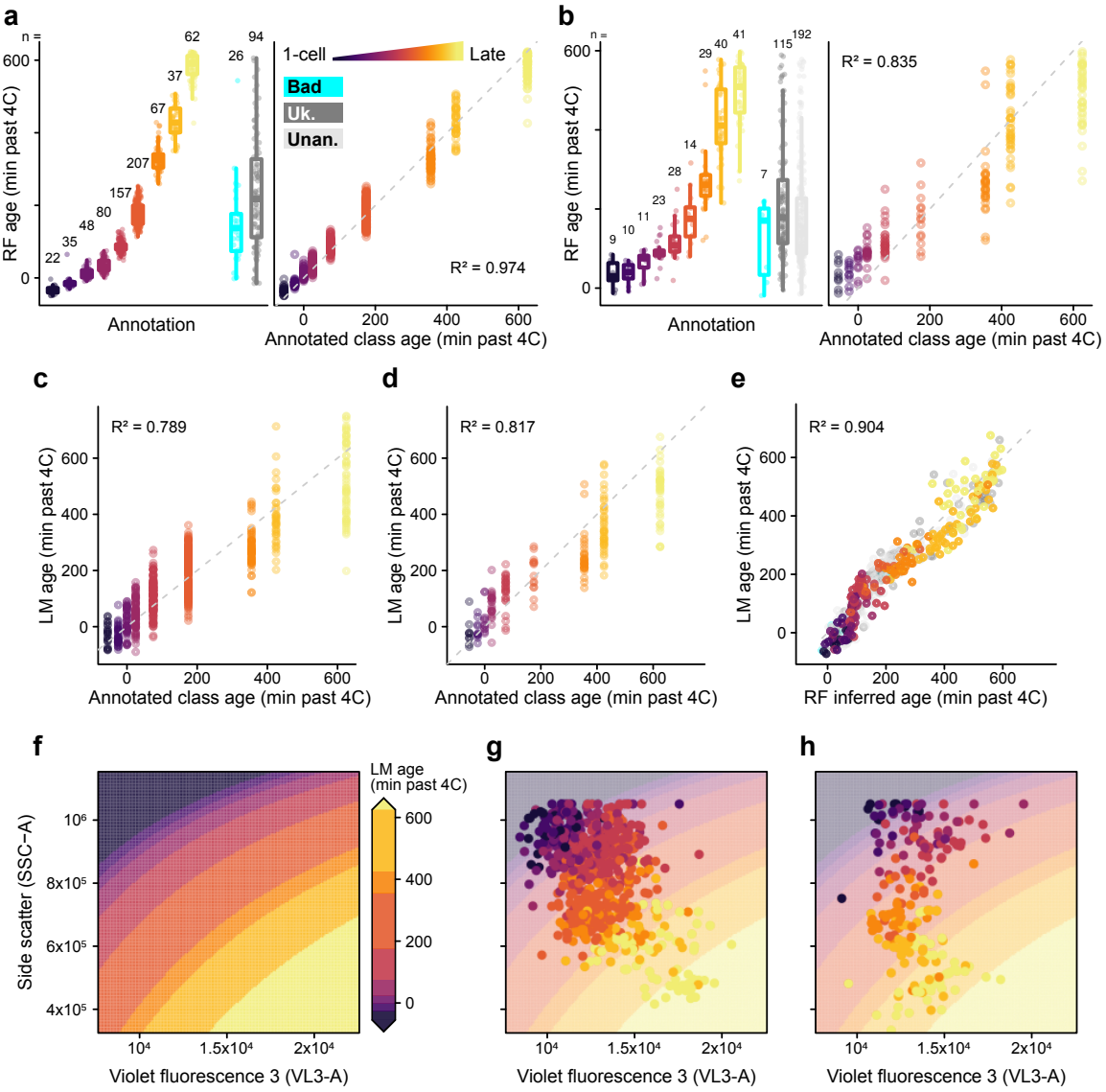


Figure 2.5 – Embryos can be staged and gated using scatter and fluorescence
a,b, Embryo age can be inferred from a random forest (RF) including all FACS measurements in training (**a**) and independent (**b**) data from distinct samples, matching stage annotation from images (left), and their corresponding age values (right).
c-e, A linear model (LM) including only side scatter (SSC-A) and violet fluorescence (VL3-A) with interaction can also properly predict age in training data (**c**), as well as for an independent sample (**d**) where it agrees closely with the more complex RF model (**e**).
f-h, Embryo age progression on the side scatter-violet fluorescence plane as described by the linear model above (**f**), followed by the training (**g**) and validation (**h**) data.

Embryo annotation and color-coding are as described in Fig. [[FACS.4]]b. In a-e, the dashed grey line is x=y. In **a-b** (right), **c**, **d**, **g**, and **h**, only annotated embryos are shown. In **a**, **b**, (left), boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction. Uk, unknown; Unan, unannotated

2.3 Adapting Smart-seq3 to profile single embryos

Smart-Seq2 is a single-cell RNA-seq library preparation protocol developed by PICELLI et al., 2013 to provide unmatched sensitivity (ZIEGENHAIN et al., 2017) and full transcriptome coverage with short-read sequencing (PICELLI et al., 2014) rather than throughput (SVENSSON et al., 2018). This makes the protocol very flexible and resulted in adaptations to other sample types, notably single-worm (or single-embryo) developed in following years (SERRA et al., 2018; CHANG et al., 2021) and applied to *C. elegans* and other nematodes (MACCHIETTO et al., 2017; REY et al., 2022).

The recently-improved Smart-seq3[2] further enhances sensitivity, reduces costs, and introduces a Unique Molecular Identifier (UMI) to account for PCR amplification bias (HAGEMANN-JENSEN et al., 2020). Furthermore, it allows computational reconstruction of RNA molecules, and thus of alternatively spliced transcripts (as well as assigning alleles) without resorting to long-read sequencing that generally lacks in throughput (OIKONOMOPOULOS et al., 2020).

We therefore sought to adapt Smart-seq3 for single embryo profiling to improve upon current single-worm adaptations of Smart-seq2, notably in cost and sensitivity.

2.3.1 Overview of the original Smart-seq3 protocol

The robustness and sensitivity of Smart-seq3 are the result of hundreds of optimization experiments performed by HAGEMANN-JENSEN et al., 2020, while recovering full-transcripts from short-read sequencing data is mainly a computational innovation. As a result, the main steps of the protocol [3] described below follow standard RNA-seq library preparation guidelines.

Workbench cleanup to clear RNAse and potential contaminants.

Lysis buffer preparation which includes the oligodT primer that will target mature mRNAs for reverse transcription and dNTPs required for the reaction.

Sample collection followed, if necessary, by storage at -80°C.

Cell lysis and RNA denaturation Lysis is a simple heat treatment, as cells are fragile. The oligodT also hybridizes with the polyA tail of mRNAs at this step.

Reverse Transcription (RT) PCR selects mRNAs through the oligodT primer, and uses template switching to incorporate a UMI (and PCR primer) on the 5' end of the molecule (Fig. 2.6a)

Preamplification PCR to amplify RT-PCR products with UMIs (Fig. 2.6b).

Purification of cDNA with magnetic beads gets rid of leftover reaction components or other elements from the input sample.

Quality control and normalization of cDNA Control for both cDNA quantity and quality. For example, poor quality samples with signs of RNA degradation can be detected with Bioanalyzer or Tapestation cDNA fragment length distribution profiles, and discarded (see Appendix C). Normalization of the purified cDNA concentration ensures an even tagmentation reaction and final cDNA amount across samples.

Tagmentation in which cDNA is (essentially) randomly cut by a tagmentase enzyme (Tn5) and the resulting fragments flanked with PCR primers (Fig. 2.6c). After this step, the library consists of UMI-tagged fragments and internal fragments, the ratio of which can be tweaked with input Tn5 and cDNA concentration (but is also influenced by the sequencer type,

[2] and even more recent Smart-seq3xpress, HAGEMANN-JENSEN et al., 2022.

[3] accessible in its latest version (v3 at the time of writing) on protocols.io at https://dx.doi.org/10.17504/protocols.io.bcq4ivyw

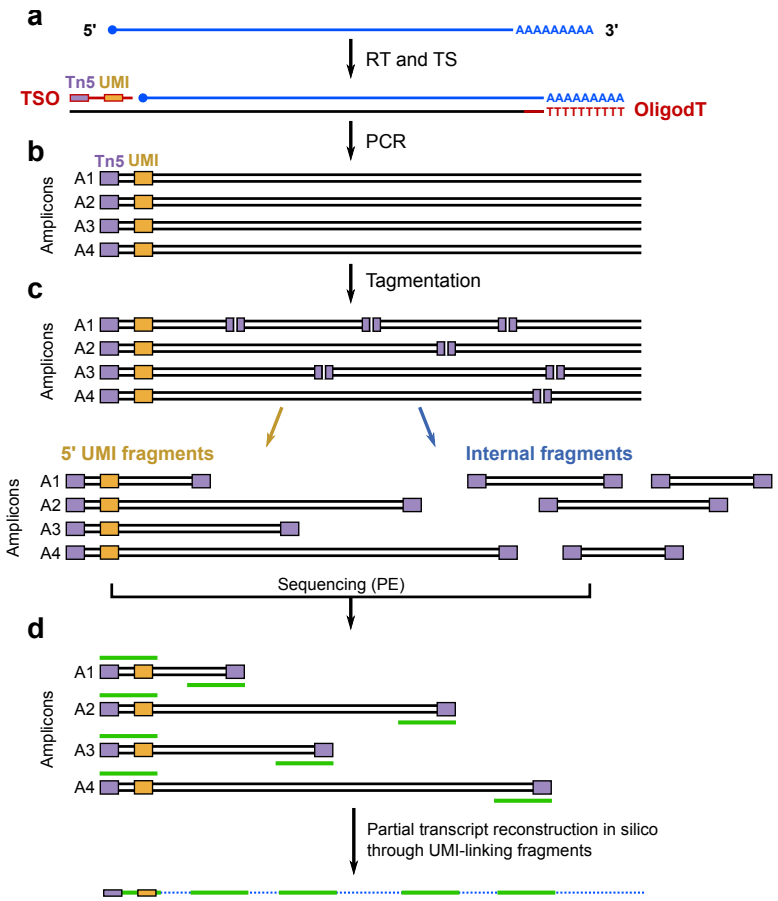


Figure 2.6 – Journey of an RNA molecule through Smart-seq3
a, RNA molecules with a polyA tail (in blue) are reverse transcribed from the oligodT, and template switching is carried out at the 5' end, introducing a Unique Molecular Identifier (UMI).
b, PCR preamplification produces multiple amplicons of the same UMI-tagged initial molecule
c, Tagmentation then randomly cuts the amplicons, producing 5'-UMI-tagged fragments and internal fragments.
d, After paired-end (PE) sequencing (reads shown in green), variable 3' ends of fragments with the same UMI allow partial reconstruction of the initial RNA molecule.

Adapted from Figures 1a and 3a of HAGEMANN-JENSEN et al., 2020.

HAGEMANN-JENSEN et al., 2020). Unlike most short-read protocols, Smart-seq3 benefits from a tagmentation output with varied and large fragment sizes. Indeed, copies of the same initial cDNA (i.e. RNA molecule, with the same UMI) will have variable 3' ends due to tagmentation. After paired-end sequencing, different transcript regions spanned by the 3' sequences (read 2) can therefore be computationally linked to a single molecule with the UMI in 5' (read 1) and enable what the authors call *parallel reconstruction of the RNA molecule* (Fig. 2.6d).

Tagmentation PCR to amplify and barcode tagmented fragments.

Pooling and final purification Barcoded samples can now be pooled, and purified to remove any leftover reaction components.

Final quality control to check fragment length distribution after tagmentation.

Sequencing on any Illumina-compatible sequencer, short read and paired end.

Data processing using the zUMIs (PAREKH et al., 2018) pipeline to demultiplex and process sample barcodes and UMIs.

2.3.2 Protocol modifications for *C. elegans* single embryos

Perhaps as a testament to the robustness of the protocol achieved by HAGEMANN-JENSEN et al., 2020, adapting Smart-seq3 from single-cell to single-embryo required fairly minor adjustments and optimization. In this section, I describe the main changes to the protocol, which concern sample collection (and potential subsequent freezing), lysis, and the PCR and purification steps following tagmentation. The full modified protocol is described in Appendix C.

2.3.2.1 Sample collection and lysis

A pure solution of *C. elegans* embryos of all stages can be collected by bleaching the contents of a plate (ie. adults + laid eggs, see Methods). Individual embryos can then be pipetted into the lysis buffer by hand or, as seen above, sorted using a FACS instrument.

More robust than isolated cells, embryos have a protective eggshell and later stages also form a cuticle, so they require a rougher and more active lysis. Standard practice for worm RNA-seq is to start by lysing with proteinase-K, as done by SERRA et al., 2018 in their single-worm adaptation of Smart-seq2. We therefore added a working concentration of the enzyme (1µg/µL) to the lysis buffer, and introduced its corresponding incubation cycle of 10 min at 65°C, followed by 1 min at 85°C for inactivation. We removed the oligodT and dNTPs from the buffer to avoid damaging them during lysis, adding them afterwards, followed by 5 minutes of incubation at 72°C to denature and hybridize mRNAs with the oligo. Though the eggshells of *C. elegans* embryos are also often lysed with chitinase to dissociate cells (ZHANG & KUHN, 2018; PACKER et al., 2019), we found this to be unnecessary and could prepare clean libraries from all developmental stages without it (Fig. 2.7).

Crucially, we found that any kind of freezing (close to 0,-20 or snap freezing) before the lysis step damages the samples (Fig. 2.8a-d), while freezing after lysis doesn't (Fig. 2.8e). We don't know why, but one hypothesis is that the RNAse inhibitor can not diffuse and perform properly before the embryo is fully degraded. We found no negative effects on sample quality when freezing post-lysis at -20°C for 1.5 hr, which left ample time for our sample collection needs. The effect of longer freezing remains to be tested.

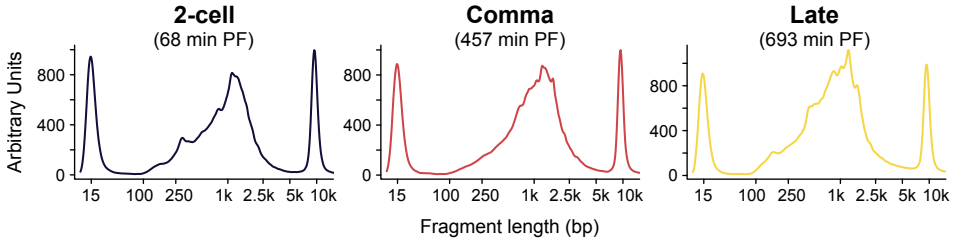


Figure 2.7 – Example cDNA libraries spanning all *C. elegans* embryo development
Embryos were roughly staged under a binocular before collection, and a precise timing in minutes post-fertilization (min PF) was acquired post-profiling with RAPToR, which confirms the initial staging. Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

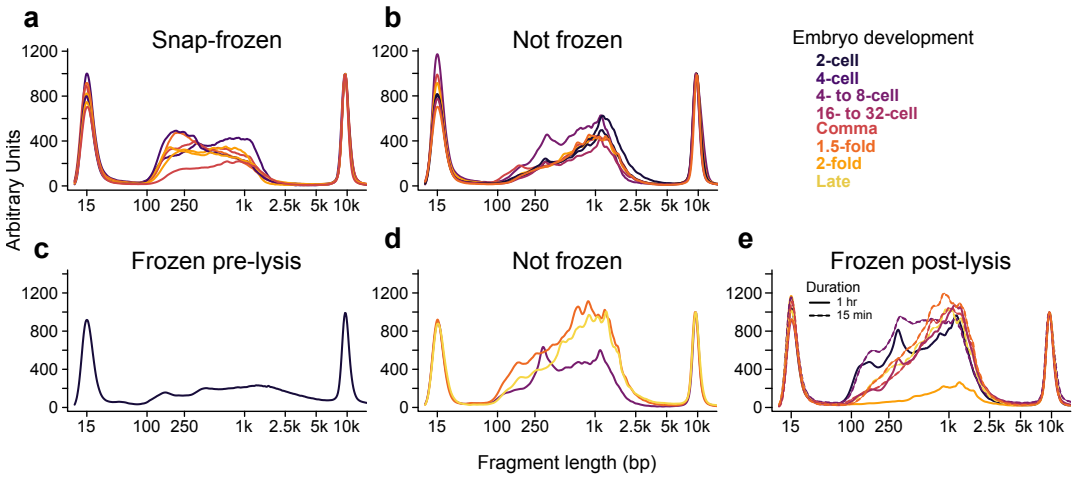


Figure 2.8 – Freezing before lysis damages samples
a,b, Fragment length distribution of cDNA libraries from samples snap-frozen at collection (**a**) or not (**b**).
c-e, Fragment length distribution of cDNA libraries from samples frozen before lysis (**c**), not frozen (**d**), or frozen after lysis for 15 minutes or 1 hour (**e**).
Samples are color-coded by embryo development. **a**,**b** and **c-e** are two separate experiments; samples were processed together within each experiment. In **c**, the only sample with cDNA yield (out of 4) is shown. Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

2.3.2.2 Tagmentation and following steps

In our hands, the original tagmentation and following PCR initially yielded no detectable amplification of material. Following a lead from a Smart-Seq3 author [4], we found the main cause to be the SDS added to stop tagmentation, which severely inhibited the PCR reaction. To mediate this effect, I first doubled the PCR reaction volume (from 7µL to 14µL) to stabilize the reaction and dilute the SDS. Then, the PCR reaction was further stabilized by adding DMSO (at 2.5%, according to manufacturer instructions) and Tween-20 (at 0.01%), known to counteract PCR inhibitors (LORENZ, 2012), which we confirmed (Fig. 2.9a). Finally, the enzyme concentration was increased to match the manufacturer instructions (0.02U/µL), and the number of PCR cycles raised from 12 to 15.

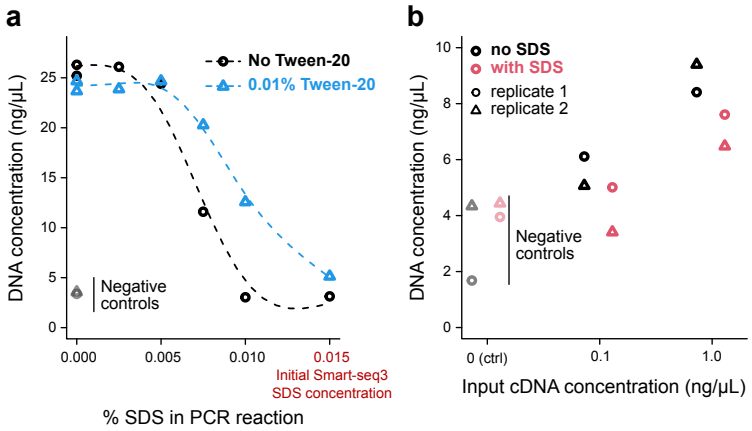


Figure 2.9 – Countering the inhibitory effect of SDS on the PCR reaction
a, Tween-20 counteracts SDS, which totally inhibits the PCR at the concentration suggested in the original Smart-seq3 protocol. Input DNA concentration is 10 ng/µL. Smoothing splines are fit per condition.
b, Low PCR amplification even after adjustments. The PCR reaction includes Tween-20 (0.1%) and SDS (0.0075%).
DNA concentration was measured with a Qubit instrument after 30 PCR cycles, with no purification (thus also measuring the primers). Input DNA concentrations refer to their respective solutions, not the PCR reaction concentration. Negative controls replace input DNA with water.

Despite these adjustments, PCR amplification is still clearly not exponential, and negatively impacted by the presence of SDS (Fig. 2.9b). At the same time, further increasing the number of PCR cycles leads to the amplification of undesirable long chimeric fragments, easily identified when longer than the starting material and often "bleeding through" the fragment length profile (Fig. 2.10a-c, see also Appendix C). We had little success in removing long chimeric fragments through magnetic bead size-selection (Fig. 2.10c), and therefore instead focused on reconcentrating the low PCR output at the purification step. Since libraries are barcoded and pooled at this point, the final elution of the purification can be done in a fraction of the large starting volume to concentrate the final library up to or above sequencing requirements. We found that eluting in 1/5 of the starting volume (feasible with at least 3 samples in the pool) is sufficient to reach comfortable cDNA concentrations, above 4.0 ng/uL.

A final issue to be resolved is the adapter-dimer contamination (peak around 120bp in Fig. 2.10b-d). Like the chimeras, these fragments compete with the actual cDNA library for sequencing and therefore decrease the final amount of usable reads. As I was unable to remove the adapters, even with multiple rounds of purification selecting longer fragments (Fig. 2.10d, and given that over half of the already low starting material is lost at each round, we decided to accept the read losses of small contaminations. Although it should be possible to rid the libraries of adapter dimers with a gel size-selection step, this practice is not recommended for low

[4] commented on the online protocol, see
https://www.protocols.io/view/smart-seq3-protocol-36wgq5rjxgk5/v3/comments?q=sds

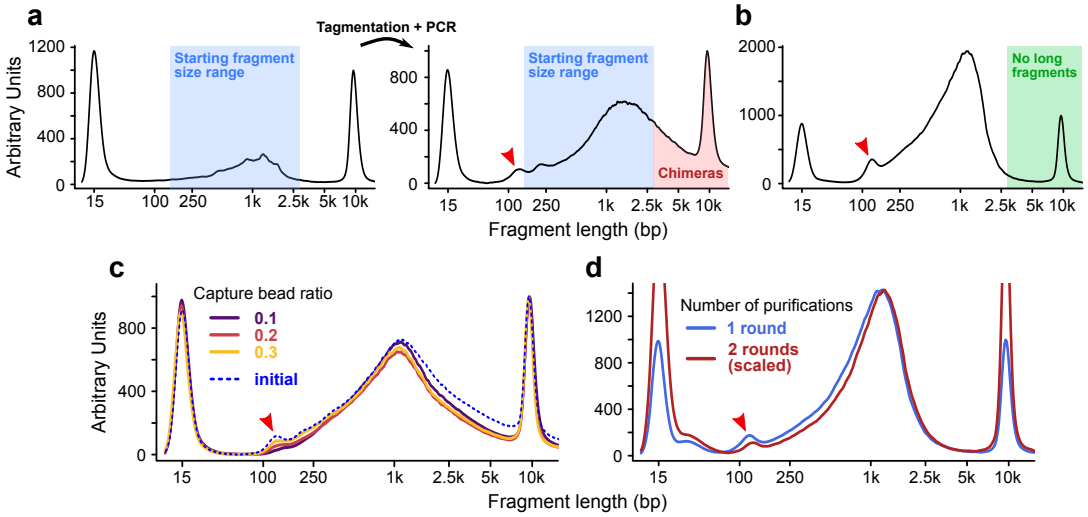


Figure 2.10 – Chimeras and adapter-dimers in tagmented libraries
a, cDNA fragment length distribution of a single sample before (left) and after tagmentation (right), where long chimeric fragments appear after PCR amplification with 30 cycles.
b, Example of a cDNA fragment length profile after tagmentation without chimeras (pool of multiple input samples)
c, Magnetic bead capture of large fragments only slightly depletes chimeras in a tagmented pool of samples. Beads are added to the pool in the specified ratio (e.g. 0.1:1 beads to sample), left to settle on the magnet, and the supernatant is collected for analysis. The supernatant should in theory be depleted of longer fragments which have higher affinity with the beads.
d, Selecting long fragments with one (blue) or two (red) rounds of purification (0.8:1 beads to sample) fails to remove adapters-dimers (arrow) in tagmented pool of samples. Profiles are scaled to their main peak for comparison.
In **a-d**, arrows around 120 bp indicate adapter-dimer contamination. Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

concentration inputs (HOEIJMAKERS et al., 2013), even with more recent automated instruments (e.g. the BluePippin [5]), and would result in further significant loss of material.

To summarize, we introduced an active lysis step to properly degrade embryo tissues, and optimized the tagmentation PCR and following purification to reach acceptable library quantity and quality for sequencing.

2.3.3 A missing link

In full transparency, the combination of FACS-sorting embryos and RNA-seq with Smart-Seq3 has yet to be completely realized. The initially poor sorting efficiency of embryos we encountered was unexpected, and we resorted to manual collection for the pilot experiments presented below. We did however successfully prepare cDNA libraries from the few sorted embryos (Fig. 2.11), showing there are no problems with processing samples after FACS sorting.

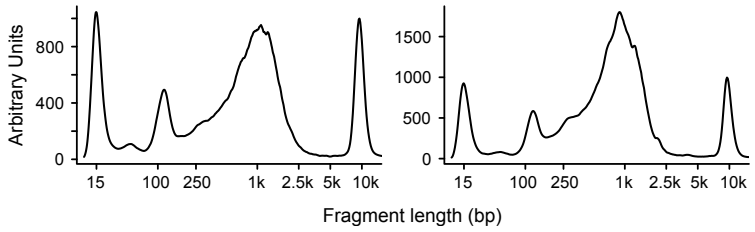


Figure 2.11 – cDNA libraries from FACS-sorted single embryos
Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

We expect complementary experiments to optimize sorting efficiency (ensuring at least a robust 50% yield) are the last requirement for the full streamlined protocol to be viable. Empty wells can be detected during cDNA quantification after the preamplification PCR, and thus discarded before the most expensive part of the protocol (tagmentation).

2.4 Quality RNA-sequencing data from single-embryos

In this section, we evaluate the quality of RNA-seq libraries made with our adapted Smart-seq3 protocol and when appropriate, compare our data with two published single-embryo RNA-seq time-series (MACCHIETTO et al., 2017; LEVIN et al., 2016) (Table 2.1) that also used library preparation protocols adapted from single-cell and span all of embryogenesis.

Source	Library preparation protocol	Sequencing	Avg. depth (M reads)	Number of samples
This study	Adapted Smart-seq3 (this study)	150 PE	4.5	63
MACCHIETTO et al., 2017	Adapted Smart-seq2 (SERRA et al., 2018)	43 PE	10	48
LEVIN et al., 2016	CEL-seq (HASHIMSHONY et al., 2012)	50 SE*	1	110

*: CEL-seq is technically PE, with a 15 base read 1 sequence only containing the barcode. After demultiplexing, only the 50 base read 2 is used for mapping and counting.

Table 2.1 – List of compared *C. elegans* single-embryo RNA-seq datasets

[5] https://sagescience.com/products/bluepippin/

2.4.1 High-complexity libraries saturated for gene detection

Using our adaptation of Smart-seq3, we prepared 63 libraries of single embryos spanning all developmental stages, and sequenced them at 150 bp paired-end, aiming for an average sequencing depth of 4.5 million pairs of reads (M reads). Overall sequencing yield was good despite the presence of adapter dimer and chimeras (around 3% loss, see Methods), and 79% of reads mapped uniquely to the *C. elegans* genome.

Mapped library sizes have a median of 3.5M and span 0.3-12M counts, which is coherent with our sequencing depth and the variability of other single-individual data (Fig. 2.12a). Despite this variation in size most of our libraries are saturated in gene detection, meaning we reach a plateau where deeper sequencing is not worth the resulting slight increase in library complexity (Fig. 2.12b). We estimate that 93% of genes are detected at median library size, and that doubling it would only increase this to 96% (Fig. 2.12b, Methods). Sequenced at over twice the depth, libraries from MACCHIETTO et al., 2017 also show high variability in size (Fig. 2.12a) but good saturation (Fig. 2.12c). A significant fraction of the libraries from LEVIN et al., 2016 could however have benefitted from further sequencing (Fig. 2.12d), as doubling the median library size (0.85M) would increase gene detection from 82% to 90%.

Consistent with studies in *C. elegans* (TINTORI et al., 2016) and zebrafish embryos (WHITE et al., 2017) showing an increase in the diversity of expressed genes along embryo development, we find the large differences in complexity between saturated libraries are explained by embryo age (Fig. 2.12e). In fact, a linear combination of log(library size) and embryo age explains over 99% of variance in the number of detected genes in all 3 datasets (Fig. 2.12f).

The sensitivity of a protocol can be measured by the number of detected genes at fixed depth (ZHANG et al., 2019). Although both Smart-seq protocols are clearly more sensitive than CEL-seq (Fig. 2.13a), in line with previous observations (ZIEGENHAIN et al., 2017), Smart-seq3 and Smart-seq2 show comparable results, and accounting for embryo age also reveals no consistent difference in sensitivity between them (Fig. 2.13b-d). As the increase in sensitivity of Smart-seq3 over Smart-seq2 reported by HAGEMANN-JENSEN et al., 2020 strongly depended on cell type, the lack of improvement seen here could be explained by our peculiar sample type, where starting material is more abundant than in single cells.

In summary, our Smart-seq3 single-embryo libraries have comparable or higher gene complexity than similar existing data, and saturate gene detection at our average sequencing depth of 4.5M reads.

2.4.2 Sources of variability for RNA molecule detection

Counting RNA molecules with Unique Molecular Identifiers (UMIs) should correct for amplification biases introduced by PCR and give a direct and biologically meaningful scale of gene expression (ISLAM et al., 2014; KIVIOJA et al., 2012). Therefore, we were surprised to find the number of UMIs (i.e. molecules) detected per sample spanned an order of magnitude, from 140,000 to 1.64M, and searched for the reasons behind this variation.

2.4.2.1 UMI library size depends on preamplification PCR yield and sequencing depth

Unlike for gene detection, UMI complexity is not explained by embryo age (r=0.18, p=0.14 Fig. 2.14a). Instead, we find that two obvious limitations are sufficient to explain most of the differences in UMI library size: the captured RNA pool (i.e, the amount of material after RT and preamplification) and the total library size (i.e, sequencing depth). Indeed, both the yield of the preamplification PCR and the total library size are well correlated with UMI library size (r=0.48, Fig. 2.14b, and r=0.56, Fig. 2.14c, respectively), despite being non-correlated with each other (r=-0.1, p=0.43, Fig. 2.14d), and a log-linear combination of both yield and total library size explains

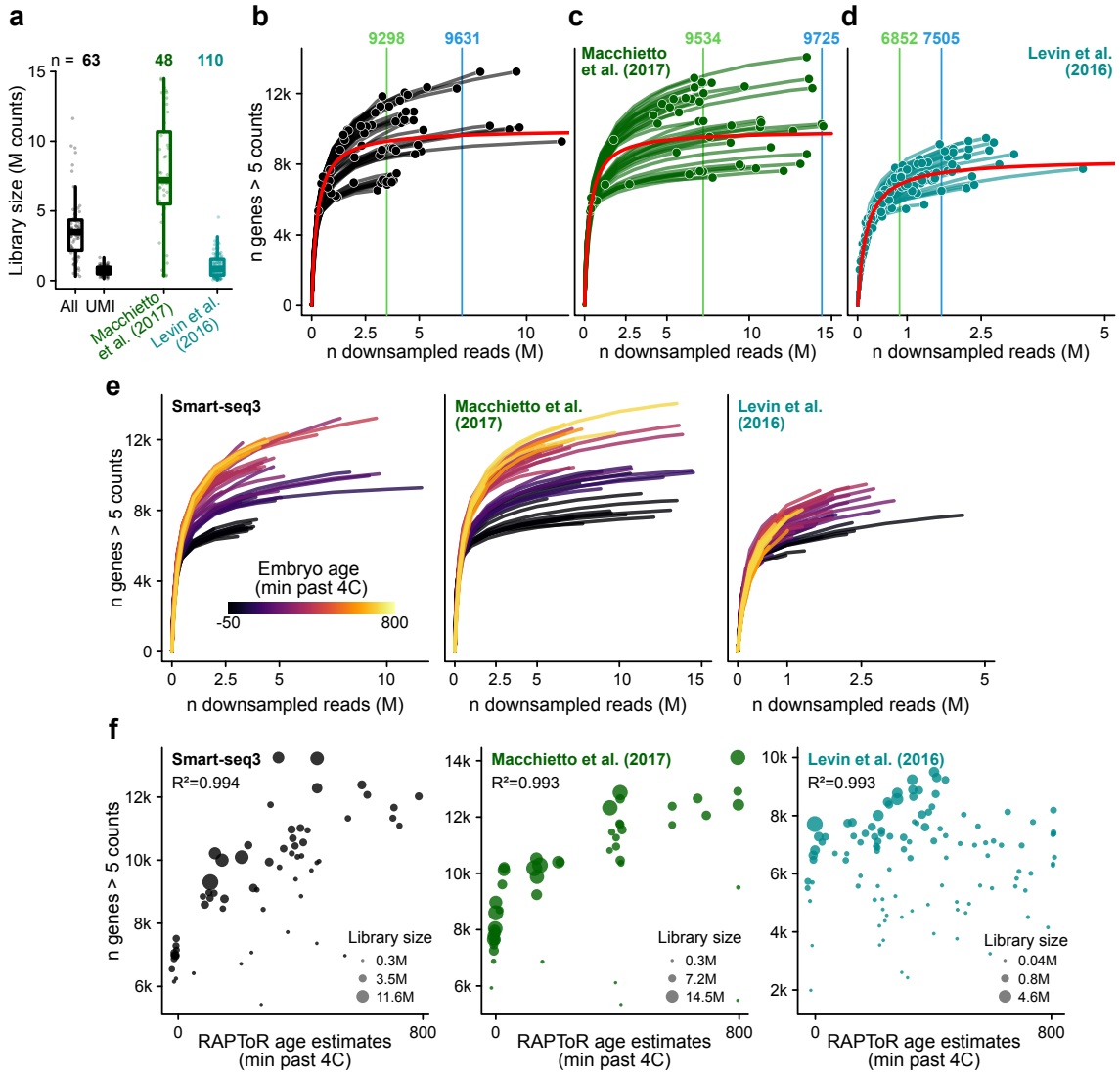


Figure 2.12 – Gene detection saturation and complexity of single-embryo RNA-seq
a, Library sizes of *C. elegans* single-embryo RNA-seq samples from this study (All mapped reads, and UMI-corrected), MACCHIETTO et al., 2017, and LEVIN et al., 2016.
b-d, Saturation of gene detection per library inferred by read downsampling. Saturation fits per dataset in red, with the corresponding number of genes detected at median library size (in green) and twice the median library size (blue).
e, Maximum library gene complexity is dictated by embryo age. Curves identical to **b-d**.
f, Complexity is a function of library size and embryo age. R^2 values from linear model fits on library complexity by age and $log(library\_size)$ with interaction and no intercept.
In **a**, boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction. UMI, Unique Molecular Identifier; M, Million.

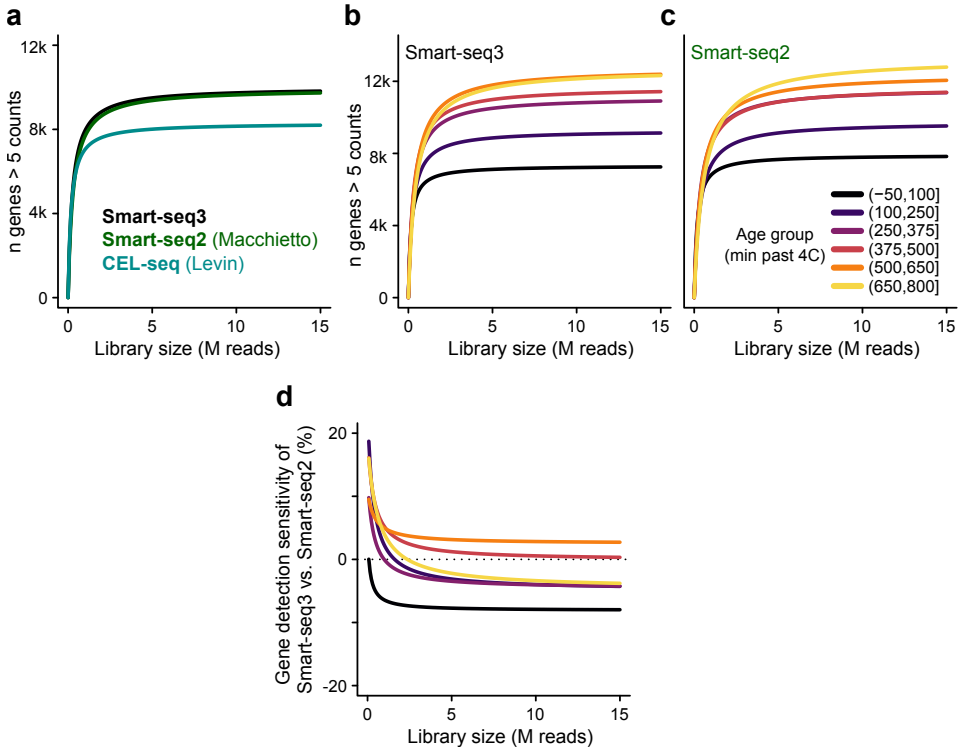


Figure 2.13 – Smart-seq3 sensitivity for single embryos compared to other protocols
a, Saturation curve fit comparison between Smart-seq3 (this study), Smart-seq2 (MACCHIETTO et al., 2017), and CEL-seq (LEVIN et al., 2016). Fits per dataset from Fig 2.12b-d.
b-c, Saturation curve fits per embryo age group for Smart-seq3 (**b**), and Smart-seq2 (**c**).
d, Gene detection sensitivity difference per age group between Smart-seq3 and Smart-seq2. Positive values mean better gene detection with Smart-seq3.

over 95% of variance in the number of detected RNA molecules. If sequencing depth can be readily increased, we have yet to understand the origins of the PCR yield variation.

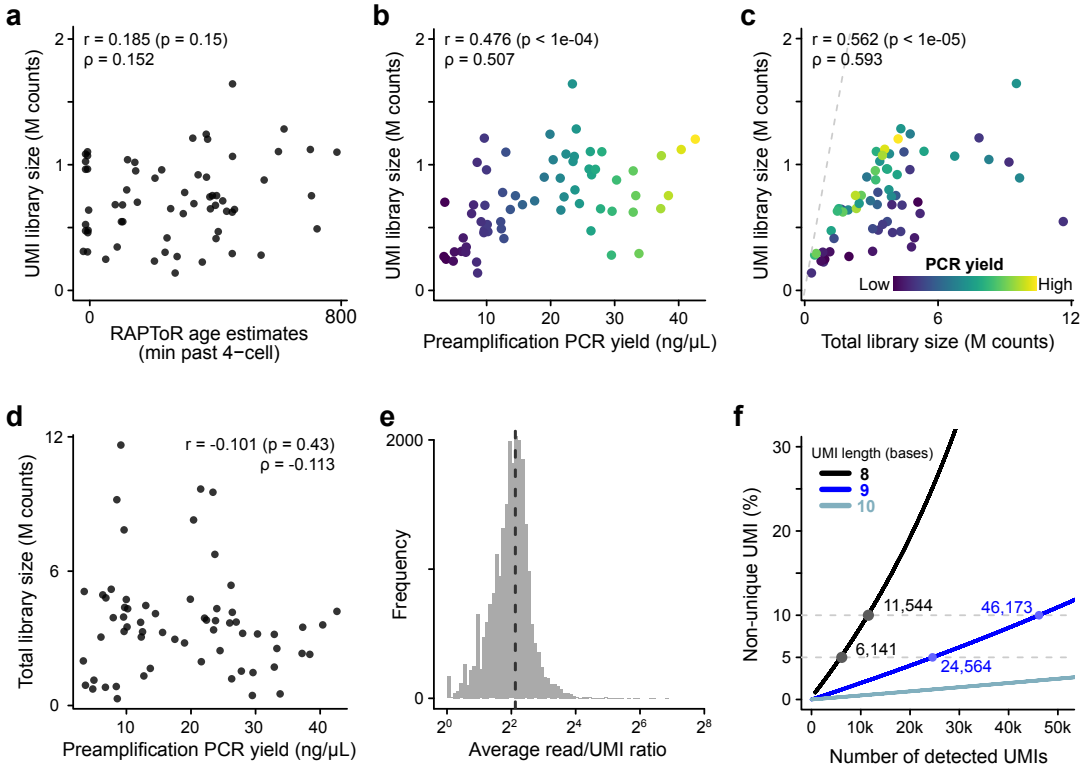


Figure 2.14 – UMI detection depends on PCR yield and sequencing depth
a, UMI library size is uncorrelated with embryo development.
b-d, Preamplification PCR yield (**a**) and total library size (**b**) explain most of the variance in the number of detected molecules per sample, and are not correlated with each other (**d**).
e, Number of reads per UMI averaged per gene. The vertical dashed line denotes the mean (4.36).
f, Estimated fraction of non-unique UMIs per number of detected UMIs for a gene when assuming random sampling, for an 8-, a 9- and a 10-base sequence. Numbers of UMIs to reach 5 and 10% non-unique barcodes are indicated on the plot.
In **a-d**, Pearson correlation is indicated alongside the p-value of its association test. In **c**, the dashed grey line denotes $x = y$. UMI, Unique Molecular Identifier; M counts, Million counts.

After accounting for yield and depth, residual variance in molecule count is also slightly, but significantly, explained by the sample batch (PCR strip) during library preparation (p<0.01 ANOVA of batch on residuals, see Methods), thus also suggesting UMI library size can still be influenced by other technical and experimental factors.

The clear limit imposed by sequencing depth on UMI detection, particularly in high-yield samples (Fig. 2.14c) indicates these libraries have not reached saturation (unlike for gene detection), with each UMI on average sequenced 4 times (Fig. 2.14e). Single-cell RNA-seq comparative studies and benchmarks agree that 1M reads per cell is generally sufficient to saturate UMI detection (Vieth et al., 2019; Svensson et al., 2017; Zhang et al., 2019). Although such information is lacking for whole-organism profiling where UMI-based protocols are seldom used, this number depends on the initial amount of RNA and it is therefore not surprising for 3.5M reads to be insufficient in our case. To our knowledge, a single other work profiled single (*Drosophila*) embryos using a UMI-based protocol (Perez-Mojica et al., 2023). While no saturation analysis was performed by the authors, they mention an average library size of 6M reads and median detection of 600,000 UMIs, which is slightly below the complexity we achieve at lower depth with Smart-seq3 (690,000 UMIs, with 3.5M reads).

2.4.2.2 Higher UMI complexity will require longer barcodes

Since the starting pool of RNA for whole embryos is larger than for single cells, there is valid concern that the diversity of 8b UMI tags (65,536 possibilities) could be limiting. Assuming equal capture of UMI tags, we can infer that a capture of 6,448 molecules for a gene will on average have 5% non-unique UMIs, resulting in a UMI count of 6,141 (Fig. 2.14f, see Methods). Only 61 genes (0.25%) pass this UMI count threshold in our libraries, suggesting the 8-base UMI is sufficient here, but this may not be the case if libraries reach saturation. Using a 9- (or even 10-) base UMI would dramatically reduce the chance of non-unique UMI tags (Fig. 2.14f) at virtually no extra cost and no hindrance to the template-switching reaction (in fact, a similar adjustment has been made to the latest Smart-seq3 installment by HAGEMANN-JENSEN et al., 2022). Furthermore, given how strongly preamplification PCR yield influences the size of the captured RNA pool (Fig. 2.14b), we expect that very high-UMI-complexity libraries can be achieved by selecting informative (i.e. high-yield, >20 ng/µL) samples, and that sequencing such libraries at a depth exceeding 10M reads would likely still not saturate UMI detection.

Importantly, saturating the detection of UMIs is not always necessary depending on the study. Expression can be quantified with high accuracy at much lower depths (SVENSSON et al., 2017; ZHANG et al., 2019), sufficient for differential expression analysis, especially considering the added power from UMI correction (ZIEGENHAIN et al., 2017; PAREKH et al., 2016). Furthermore, although saturated UMI libraries should provide an absolute quantification of expression, SVENSSON et al., 2017 show this is not exactly true.

To summarize, we find that UMI-corrected library sizes measure the amount of unique RNA molecules we could capture and sequence from a sample, rather than a true amount of starting mRNA. Then, although an 8b UMI is sufficient here, increasing to 9 or 10 bases would be safer for highly-expressed genes and deeper sequencing. Finally, high-yield samples tend to make more complex libraries, which can therefore be selected early-on during library preparation.

2.4.3 Accurate embryo development and gene expression dynamics

We find that age inferred from embryo gene expression using RAPToR (BULTEAU & FRANCESCONI, 2022) matches both our manual staging and timed egg-lays done at sample collection (Fig. 2.15a, Methods), thus exemplifying that such tedious manual staging of embryos is no longer necessary.

Grouping samples from all datasets together, we perform an Independent Component Analysis (ICA) to summarize the main signals of gene expression and find remarkably little batch effect and variation between the three datasets (Fig. 2.16). This is particularly the case between the Smart-seq protocols, where developmental dynamics in are tightly matched in most Independent Components (ICs), and batch effects usually seen in the first components only appear in ICs explaining a low percentage of variance in the data (e.g. IC13, 1.81% or IC15, 1.57%, Fig. 2.16).

Embryos from LEVIN et al., 2016 show more typical batch effects, such as their clear segregation in IC5 (5.95%), and differences in the amplitude of expression dynamics (ICs 3, 4, 7, and 9, Fig. 2.16), despite normalization (see Methods). Amplitude differences can not be explained only by lower sequencing depth, as an ICA on libraries subsampled to 1M reads still shows the artefact (Fig. 2.17), so they could be a consequence of the different sensitivity of the CEL-seq protocol. Other differences such as diverging dynamics (e.g. IC10, IC11, Fig. 2.16) could be biological, or the result of differing sample collection protocols (LEVIN et al., 2016 bleached early embryos and waited for the desired timing, whereas MACCHIETTO et al., 2017 and we directly collected targeted stages).

We conclude our data closely matches expected expression dynamics along embryo development.

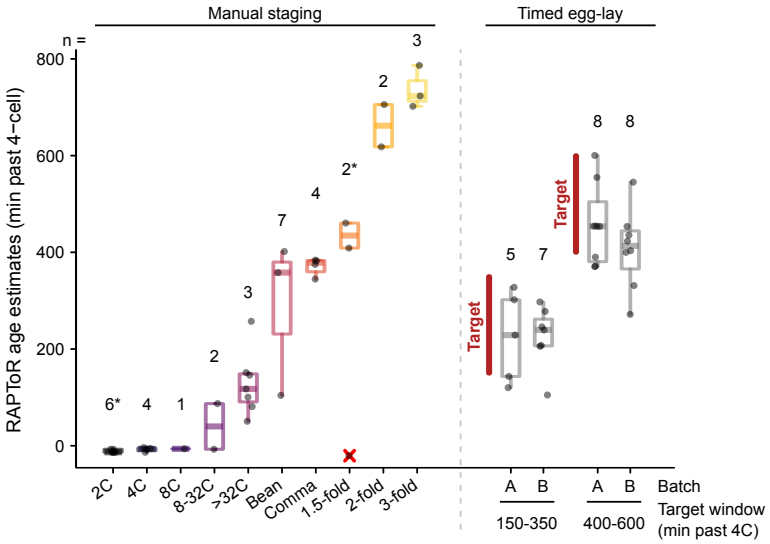


Figure 2.15 – Inferred age matches expected embryo development of samples
RAPToR age estimates of manually staged (left) and timed egg-lay (right) single embryos. Timed egg lays targeted the windows of development indicated in red (150-350, and 400-600 min past 4C), in two batches each (A, B). A red cross indicates an embryo mis-staged as 1.5-fold during collection, which is actually a 2-cell stage embryo according to RAPToR staging and correlation with other 2-cell samples.
Boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction. *: the number of samples indicated above each box doesn't include the mis-staged sample.

2.4.4 Conclusions

We have successfully profiled single *C. elegans* embryos of all stages with our adapted Smart-seq3 protocol, and generated high quality libraries. The cost of profiling is drastically reduced over the previously adapted Smart-seq2 (from 35€ to around 1.5€ per sample in reagent costs alone), but also compared to state-of-the-art protocols. Indeed, a recent single-embryo study reports a total cost per sample (including sequencing) of 36€ (PEREZ-MOJICA et al., 2023), while we achieve greater UMI complexity for under half that price (see Methods) Our samples match expected developmental stages, as well as the expression dynamics of previously established studies. Although library complexity can still be improved by extending the UMI barcode and selecting high preamplification-PCR yield samples, we conclude our adapted protocol can already produce exploitable libraries for gene expression analysis.

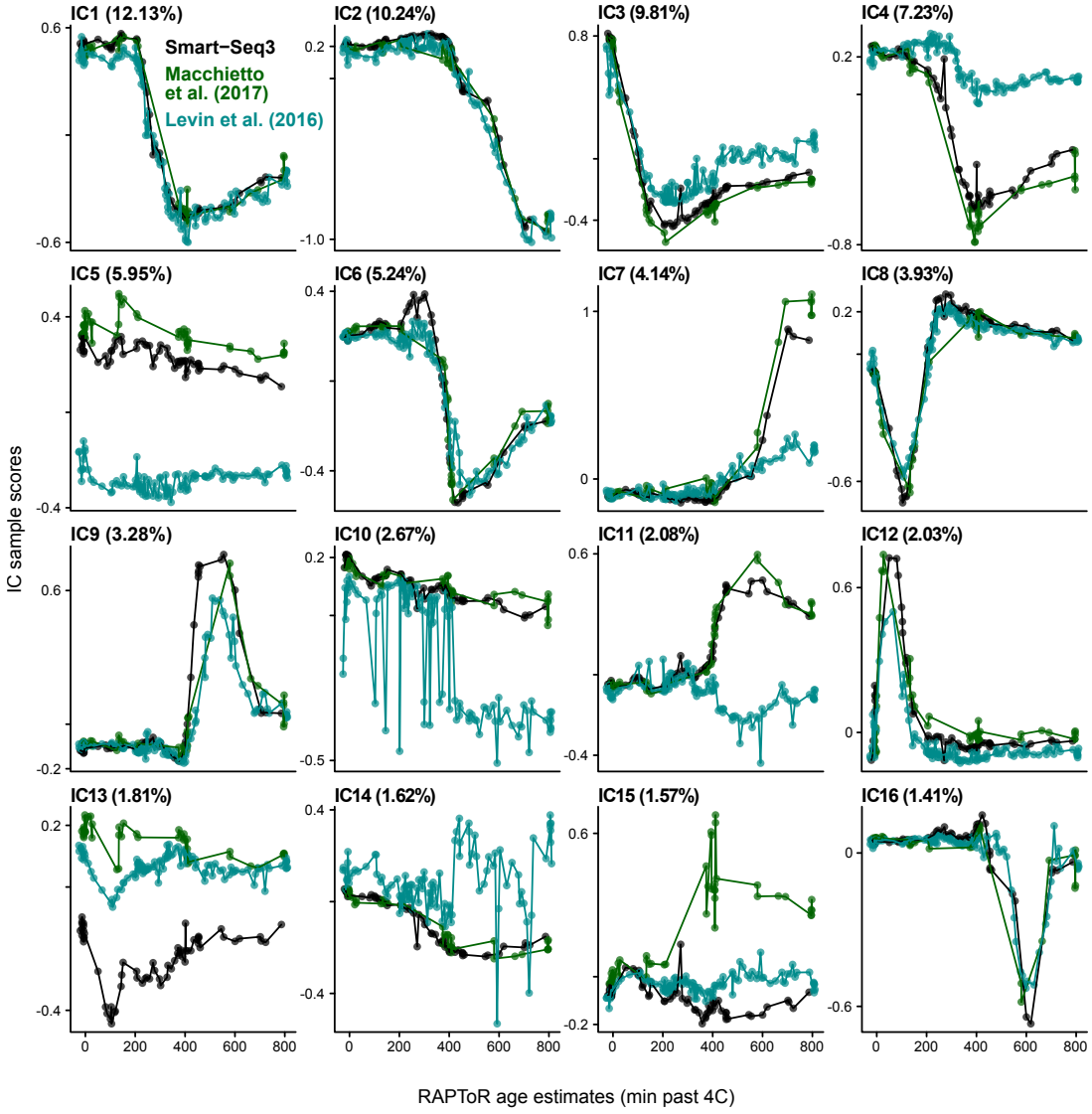


Figure 2.16 – Gene expression dynamics embryogenesis across 3 single-embryo RNA-seq datasets
Independent Components (ICs) of an ICA joining single-embryo RNA-seq samples from this study (Smart-seq3), MACCHIETTO et al., 2017, and LEVIN et al., 2016 plotted along embryo age. The percentage of total variance explained per component is indicated next to IC number.

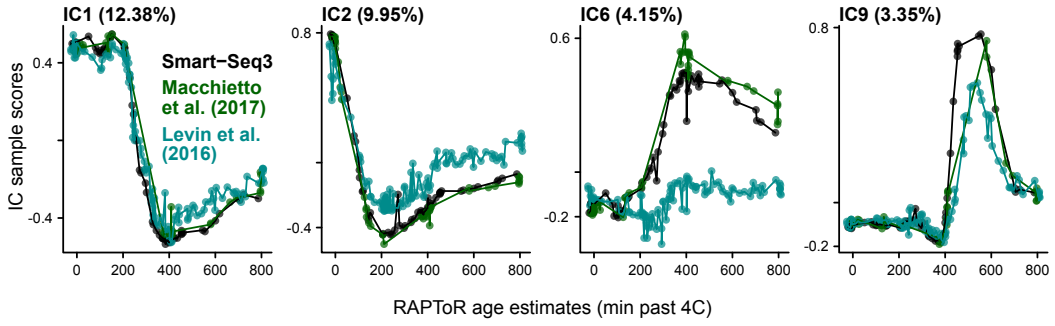


Figure 2.17 – Amplitude differences of expression dynamics remain at equal library size
Selected Independent Components (ICs) of an ICA joining single-embryo RNA-seq samples downsampled to 1M reads, from this study (Smart-seq3), MACCHIETTO et al., 2017, and LEVIN et al., 2016 plotted along embryo age. The percentage of total variance explained is indicated for each IC.

2.5 Discussion

I have shown here that single-individual profiling of small organisms can take full advantage of single-cell technologies. First, we reaffirm that flow cytometry equipment can be used to characterize and sort single live *C. elegans* embryos, improving upon a seldom-used technique from over a decade ago (STOECKIUS et al., 2009). Thanks to a cytometer capable of taking brightfield images of events, we found that unfertilized, dead, or live embryos can be reliably selected without using any markers, which is transposable to other instruments. We further established and characterized a clear link between embryo autofluorescence in multiple wavelengths and developmental progression, which can therefore be used to enrich in developmental stages of interest during sorting. Our findings make flow cytometry a promising technique not only to sort individuals at high throughput, but also potentially to study interindividual variation at scale in *C. elegans*, for example by understanding how variation in genome wide gene expression relates to phenotypic variation measured by multiple cytometer parameters (autofluorescence, scatter etc..) or features extracted from images.

Then, we adapted the Smart-seq3 single-cell RNA-seq library preparation protocol for single embryos, and show they can be profiled across development at high gene and UMI complexity at much reduced cost from previous (PICELLI et al., 2013) and current (PEREZ-MOJICA et al., 2023) protocols. Furthermore, we have yet to explore RNA molecule reconstruction and isoform assignment, which add advantages and possibilities for analysis. Coupled to post-profiling staging from the transcriptome provided by RAPToR (BULTEAU & FRANCESCONI, 2022), these breakthroughs enable a streamlined high-throughput and cost effective way to profile single nematodes across embryo development.

Although we have reached a point where both flow cytometry, sorting, and library preparation of single embryos can be done successfully, further optimization is still needed to make the most of this work.

To begin, FACS sorting efficiency of embryos must robustly reach at least 50% to truly improve throughput. Next, we have yet to determine a proper method to remove adapter and chimera contaminations in pooled tagmented libraries. Finally, UMI complexity of libraries can be substantially increased by combining a longer UMI barcode with deeper sequencing of high-yield samples. We also expect that "hands-on" time, plastics use, and errors due to pipetting low volumes can be significantly reduced by using automated liquid dispensers (such as the MANTIS® robot [6]), as we found multichannel pipettes particularly imprecise at low volumes and resorted to using single-channel pipettes for most of the library preparation steps.

Our sorting method is restricted to smaller individuals like the embryos (or L1 larvae, FERNANDEZ et al., 2010) of *C. elegans* because we rely on flow cytometry equipment. However, because such equipment is more widespread than specialized instruments capable of sorting larger, adult-*C. elegans*-sized particles like biosorters, we note this also makes our approach more accessible to potentially interested labs. Large particle sorters can still likely be used in a similar way, but this and the efficiency of our library preparation protocol on adults and larger animals remains to be tested.

Moving away from the traditional "3 perturbed vs. 3 control" profiling experimental design to a larger, less synchronized number of single individuals per condition is now possible at equal or lower cost to analyze gene expression changes between conditions, taking into account not only inter-individual variation but also dynamic changes that could have otherwise been missed.

My work focuses on *C. elegans*, but we believe that other small organisms can also benefit from our findings. For example, we successfully prepared a cDNA library from the embryo of a distant nematode cousin, *Mesorhabditis Belari*, in a small pilot experiment with the lab of Dr. Marie Delattre. There is also no reason why Smart-seq3 would not work on other sample types, such as dissected tissues of larger organisms, provided RNA can be extracted from the material.

[6] see formulatrix.com/liquid-handling-systems/mantis-liquid-handler/

Going forward, we envision that integrating several levels of information from many single individuals – such as images, fluorescence and scatter measurements, and gene expression – will enable us to gain a more complete picture of individuals and their differences within and between populations.

2.6 Methods

All data analyses were performed with R (v4.3.0) unless otherwise noted.

2.6.1 Nematode culture and handling

C. elegans worms of the standard "N2" laboratory strain were kept at 20°C in roughly synchronized populations on 60mm NGM agar plates seeded with OP50. 100 eggs are transferred to new plates every 3-4 days to maintain populations.

2.6.1.1 Bleaching protocol

To collect embryos at all stages, we wash a plate with egg-laying adults, laid eggs, and hatched larvae (indicating that late embryo stages are present) using M9 into a 15mL falcon tube, taking care to scrape the agar and collect as many eggs as possible. After collection, the tube is filled to 15 mL with M9, vortexed, and centrifuged for 1 min at 5000 RCF. During centrifugation of the collection tube, another tube is prepared with 250 µLNaOH (1M) + 150 µL Bleach (NaClO, 13%). 75 µL of the accumulated worm pellet is then transferred to the bleach tube, which is vortexed thoroughly and checked frequently under a binocular until adult worms are dissolved (generally 2-3 min). M9 is then added to 15 mL to dilute bleach and stop the reaction, and centrifuge the tube 1 min at 5000 RCF. We check for the presence of a pellet and remove as much supernatant as possible without disturbing it. This M9 wash is repeated twice before collecting 1 mL of bleach output with embryos. As this protocol only takes only 10-15 min, our bleach output still contains early 1-cell embryos. Centrifugation and M9 washes can also be done at 4°C to slow development and collect early stages.

2.6.2 Flow cytometry experiments and analysis

2.6.2.1 Data acquisition and analysis

Bleached embryo suspensions were directly passed through an Attune^{TM} CytPix^{TM} flow cytometer, acquiring up to 30,000 event images alongside fluorescence and scatter measurements. Embryos-sized events are gated with FSC-H and SSC-H (as shown in Fig. 2.2b, and live embryos with fluorescence channels BL1-A and BL2-A (Fig. 2.2c). Similarly, embryos were sorted with a BD FACSARIA IIµ (with 3 lasers and 10 fluorescence channels), with FSC and blue laser channel area scaling factors set to 0.7 and 0.5 respectively. Embryos-sized events are gated with SSC-W and SSC-H, and live embryos with fluorescence channels FITC-A and PerpCP-Cy5-5-A (Fig. 2.2e-f, see also Appendix B).

A total of 8 runs were acquired on the Cytpix (2 of which are shown in this chapter, with the remainder in Appendix B), and 10 runs on the FACSAria.

Flow Cytometry Standard (FCS) files were exported from both instruments (alongside images for the CytPix), and re-analyzed in R using the `flowcore` (v2.12.0) and `ggcyto` (v1.28.0) libraries.

2.6.2.2 Sorting efficiency experiments

To evaluate sorting efficiency, 1, 3, or 8 live embryos were sorted using a FACSAria into PCR strips or strip caps containing 2.5uL of M9 buffer. We tested a centrifugation time of 5 or 15 seconds (using a small bench centrifuge) after collection, adding Triton at 0.05% in the sorted

solution, sorting in strips or strip caps, as well as applying either “single-cell” (0:32:16) or “purity” (32:32:0) sorting masks. After sorting, embryos were counted in wells under a binocular, and double-checked on slides. We then fitted a linear model including all the parameters (with no interaction) on the ratio of observed over expected embryos per well to evaluate their effects, reporting the significance of model coefficients in Fig. 2.3.

2.6.2.3 Staging embryos from autofluorescence

835 and 327 live embryos from two independent samples were manually staged from their images using the classification shown in (Fig. 2.3b), with timings in minutes past 4-cell assigned per class as follows: -55 (1-cell), -25 (2-cell), 0 (4-cell), 25 (4-8 cells), 75 (8-32 cells), 175 (>32 cells), 355 (Comma), 425 (2-fold), and 625 (Late). Embryos classified as ‘Bad’ or ‘Unknown’ were not used for model training or validation.

Using the first sample, we fit a random forest on embryo age using all 18 channel measurements of the CytPix (FSC, SSC, BL1, BL2, YL1, YL3, RL1, VL1, and VL3, in area (-A, eg. FSC-A), height (-H), and width (-W)) with the *randomForest()* function of the homonym package (v4.7-1.1), fitting 1000 trees with 6 variables. This allowed us to determine the strongest age predictors as those with the highest average decrease in mean squared error per tree (SSC-A, RL1-A, YL1-A, BL2-A, and VL3-A). We then used the random forest model to predict the age of embryos from the second sample.

To prove embryos can be staged using few parameters, we selected two parameters: SSC-A (the strongest predictor), and VL3-A (which was among the best predictors, and least correlated with SSC-A) to fit a linear model with interaction on age in the first sample. We then used the model to infer the age of embryos from the independent second sample, and to establish a map of development progression along the two parameters.

2.6.3 Smart-seq3 protocol optimization

For all optimization experiments, embryos of different developmental stages spanning at least 4-cell to 3-fold stage were collected by hand after bleaching, as described above and processed as described in the adapted Smart-seq3 protocol in Appendix C unless otherwise stated. All electropherogram data was acquired using an Agilent Tapestation, HS-D5000 tapes and reagents, and plotted with `R`.

2.6.3.1 Freezing experiments

To test the effect of snap-freezing on sample quality, PCR strips with samples were placed into liquid nitrogen immediately after collection, and taken out after 15 min for lysis together with the unfrozen samples, and processed together for subsequent library preparation steps.

To test the effect of freezing before and after lysis, embryos were collected and either immediately frozen (“Frozen pre-lysis”) for 1h, immediately lysed and stored at -20°C for 1h or 15 min (“Frozen post-lysis”), or immediately lysed and followed by the next protocol steps. All conditions were processed together from pre-RT incubation onwards.

2.6.3.2 Tagmentation PCR

To evaluate the effect of sodium dodecyl sulfate (SDS) and Tween-20 on the post-tagmentation PCR, 14µL mixes with reaction concentrations of 0%, 0.0025%, 0.005%, 0.0075%, 0.01%, and 0.015% SDS were prepared with and without Tween-20 at 0.01%, with 2.8 µL of input gDNA at 10 ng/µL (or water for controls) and 0.7µL of primers at 10µM, and amplified for 30 PCR cycles. DNA concentration was then measured with a dsDNA High Sensitivity quantification assay on a Qubit.

We evaluated the effect of SDS on the adjusted PCR after tagmentation, with 0, 0.1, and 1.0 ng/µL input cDNA pre-tagmentation. Tagmentation and PCR were performed as described in

Appendix C, adding either 0.5µL SDS 0.2% or 0.5µL water, after tagmentation incubation, and amplifying for 30 PCR cycles.

2.6.3.3 Chimera and adapter dimer contamination removal

For Fig. 2.10a, a 1.0 ngµL cDNA input library was tagmented followed by a 30-cycle PCR, and purified with 0.8:1 magnetic bead ratio. In Fig. 2.10b, multiple libraries were tagmented and pooled, following the final version of the adapted protocol. In Fig. 2.10c, large fragments of a post-tagmentation library with chimeras were captured with 0.1, 0.2, or 0.3 bead to sample ratios. Beads were left to settle on the magnet, and the supernatant (in theory depleted of longer fragments attached to the beads) collected for analysis. In Fig. 2.10d, 1 or 2 consecutive rounds of purification (0.8:1 beads to sample) were done on a library with adapter dimer contamination.

2.6.4 RNA-seq library preparation and pre-processing

2.6.4.1 Smart-seq3 libraries

A first set of multiple batches totalling 32 single embryos were manually staged and picked from bleach outputs. Timed egg-lays were done to collect a second set of batches of 31 embryos targeting 150-300 and 400-600 minutes past 4-cell windows of development. Libraries were prepared following the adapted Smart-seq3 protocol described in Appendix C. Preamplification PCR yield was measured after purification using an Invitrogen™ High-sensitivity dsDNA detection kit (Q32851) and a Qubit instrument. Both sets of samples (hereafter pool 1 and pool 2) were processed, pooled, and sequenced separately.

Pools 1 and 2, with tagmented library profiles shown in Fig. 2.10c (initial) and Fig. 2.10b respectively, were sent for paired-end (PE) 150 bp sequencing on an Illumina Novaseq 6000 instrument. Targeting 150 million pairs of reads (M) for each pool, yields were 128M and 163M respectively, resulting in an overall loss of reads of $(128 + 163)/300 = 3\%$. Lower yield in the first pool may be caused by the presence of chimeras (which are absent from the second pool), since adapter dimer is present in both pools, and mapping rate of both pools is identical: 79.2% and 79.4%, respectively. Raw fastq data was processed with the zUMIs pipeline (Parekh et al., 2018) using the parameters provided by Smart-seq3 authors at `protocols.io`, applying no UMI sequence error correction (Hamming distance parameter set to 0), mapping to the *C. elegans* genome (WBcel235, annotation v109) using STAR (v2.5.4b, as given in the zUMIs conda environment). The output count matrices for UMIs and reads uniquely mapping to intron+exon were used in all analyses (intron reads account for under 0.5% of mapped reads).

2.6.4.2 Published single-embryo libraries

Raw reads from Macchietto et al., 2017 *C. elegans* embryo samples were downloaded from the SRA (accession: SRP084244) using the sra-toolkit (v2.10.7), processed with Trimmomatic (v0.36) to remove adapters and low-quality reads with fastQC, mapped to the *C. elegans* genome (WBcel235) with STAR (v2.5.4b), and expression quantified with featureCounts (v1.6.0).

Counts from Levin et al., 2016 samples were directly downloaded from GEO (accession: GSE60755), as provided by the authors. As described in Bulteau & Francesconi, 2022, samples poorly correlated with others and clear outliers in an ICA were filtered out, leaving 110 samples for analysis.

2.6.4.3 Inferring embryo age

We used *ae()* from `RAPToR` (v1.2.0) to infer the age of all samples against the "`Cel_embryo`" reference from `wormRef` (v0.5). Age is reported in minutes past the 4-cell stage (and can thus be negative for earlier samples).

2.6.5 Analysis of library quality and properties

2.6.5.1 Gene detection saturation

Gene detection saturation was estimated by down-sampling read counts of libraries at 1k, 5k,10k, 50k, 100k, 250k, 500k, 1M, 2M, 3M, 4M, 5M, 7.5M, 10M, and 12.5M reads within the available library size in triplicate, and reporting the number of genes with at least 5 counts. Curves fit in Figures 2.12-2.13 are Michaelis-Menten saturation equations fit using the *nls()* function of the stats core package. We estimated the average number of genes detected for median and 2x median library sizes using the fits.

To assess sensitivity differences between Smart-seq2 and 3 libraries taking into account embryo development, we binned samples by (inferred) age as shown in Table 2.2 and fit curves as described above. Fig. 2.13d shows the percentage increase in gene detection of fits per age bin as follows: $100 \times \frac{SS3-SS2}{SS3}$.

Age bin (min past 4-cell)	**Smart-seq3 libraries (this study)**	**Smart-seq2 libraries** (MACCHIETTO et al., 2017)
(-50,100]	16	17
(100,250]	13	8
(250,375]	12	2
(375,500]	14	12
(500,650]	4	2
(650,800]	4	7

Table 2.2 – Number of samples per age bin for library saturation curves

Linear models to explain library complexity differences were fit for all 3 datasets with the following formula, using the lm function of R `complexity ~0 + age * log(libsize)` .

2.6.5.2 Explaining UMI library size variation

We fit a linear model using the lm function of R to explain UMI library size by preamplification PCR yield and its interaction with total library size with the following formula: `UMI_libsize ~0+log(yield)+log(yield):log(libsize)` We then tested if variance in the residuals of the model above could be significantly explained by sample batch (PCR strip) using an ANOVA, p<0.01.

2.6.5.3 Estimating duplicate UMI capture

Assuming we select k UMI tags among n possible values with equal probability of being picked (with replacement) $\frac{1}{n}$, then the expected number of distinct tags $\mathbb{E}(D_k)$ is

$$\mathbb{E}(D_k) = n\left[1-\left(1-\frac{1}{n}\right)^k\right]$$

which corresponds to an observed UMI count. Therefore, we can infer the ratio of duplicated molecules in an observed UMI count with $\frac{k-\mathbb{E}(D_k)}{k}$ duplicates, for a given number of possible UMIs. An 8-base UMI has $n = 4^8 = 65,536$ possible sequences. 9 and 10 base UMIs similarly have $n = 4^9 = 262,144$ and $n = 4^{10} = 1,048,576$ possible sequences respectively.

2.6.5.4 Protocol cost comparisons

For Smart-seq2, PICELLI et al., 2014 give an estimated cost of reagents per library of 36€, and HAGEMANN-JENSEN et al., 2020 of under 1€ per library for Smart-seq3. We estimate the cost of reagents per library of our adapted Smart-seq3 protocol for single embryos to under 1.5€. Sequencing costs come to 15-16€ per sample with an average depth of 4.5M reads, bringing the total cost per embryo to 17€.

PEREZ-MOJICA et al., 2023 mention a total cost per embryo of 36€, sequenced with identical parameters to ours (150bp, PE, Novaseq 6000 sequencer) aside from depth, at 6.5M, resulting in a median UMI complexity of 600,000. Despite lower depth, our libraries already have higher median UMI complexity (690,000) at half the cost. Deeper sequencing of our libraries to match 6.5M would still result in a significantly cheaper cost per sample, at under 25€. $\frac{6.5M}{4.5M} \times 16 + 1.5 = 24.6€$.

Note that sequencing depth (number of sequenced reads) rather than library size (number of reads mapped to transcripts) is used in the calculations above.

2.6.5.5 Comparison of gene expression dynamics

Expression values of the 3 single-embryo datasets, TPM for Smart-seq3 and MACCHIETTO et al., 2017 samples and CPM for LEVIN et al., 2016 (as gene length is already accounted for by CEL-seq RNA capture), were joined with matching gene IDs, logged ($log(X+1)$), and quantile-normalized with the *normalizeBetweenArrays()* function of `limma` (v3.56.1). We then performed a centered PCA (*prcomp()*) to determine that 22 components are sufficient to explain at least 80% of variance in the data, and performed an independent component analysis (ICA) extracting 22 components with the *icafast()* function of the `ica` package (v1.0-3).

In order to check for expression dynamic amplitude difference at equivalent library size, we down-sampled libraries to 1M reads as described above (*Gene detection saturation*), and similarly joined, logged, and normalized before extracting 22 components with an ICA.

References

ANSARI, AMIR MEHDI et al. (2016). "Cellular GFP toxicity and immunogenicity: potential confounders in in vivo cell tracking experiments". In: *Stem cell reviews and reports* 12, pp. 553–559.

BOSSINGER, OLAF and EINHARD SCHIERENBERG (1992). "Transfer and tissue-specific accumulation of cytoplasmic components in embryos of Caenorhabditis elegans and Rhabditis dolichura: in vivo analysis with a low-cost signal enhancement device". In: *Development* 114.2, pp. 317–330.

BULTEAU, ROMAIN and MIRKO FRANCESCONI (Aug. 2022). "Real age prediction from the transcriptome with RAPToR". en. In: *Nature Methods* 19.8, pp. 969–975. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01540-0.

CHANG, DENNIS et al. (2021). "A revised adaptation of the smart-Seq2 protocol for single-nematode RNA-seq". In: *RNA Abundance Analysis: Methods and Protocols*, pp. 79–99.

CLOKEY, GEORGE V and LEWIS A JACOBSON (1986). "The autofluorescent "lipofuscin granules" in the intestinal cells of Caenorhabditis elegans are secondary lysosomes". In: *Mechanisms of ageing and development* 35.1, pp. 79–94.

FERNANDEZ, ANITA G et al. (2010). "Automated sorting of live C. elegans using laFACS". In: *Nature methods* 7.6, pp. 417–418.

FRANCESCONI, MIRKO and BEN LEHNER (2014). "The effects of genetic variation on gene expression dynamics during development". In: *Nature* 505.7482, pp. 208–211.

GINZBERG, MIRIAM B, RAN KAFRI, and MARC KIRSCHNER (2015). "On being the right (cell) size". In: *Science* 348.6236, p. 1245075.

GRIFFIN, JEANINE et al. (2006). "Comparative analysis of follicle morphology and oocyte diameter in four mammalian species (mouse, hamster, pig, and human)". In: *Journal of experimental & clinical assisted reproduction* 3, pp. 1–9.

HAGEMANN-JENSEN, MICHAEL, CHRISTOPH ZIEGENHAIN, and RICKARD SANDBERG (2022). "Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress". In: *Nature Biotechnology* 40.10, pp. 1452–1457.

HAGEMANN-JENSEN, MICHAEL et al. (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In: *Nature Biotechnology* 38.6, pp. 708–714.

HASHIMSHONY, TAMAR et al. (2012). "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification". In: *Cell reports* 2.3, pp. 666–673.

HEPPERT, JENNIFER K et al. (2016). "Comparative assessment of fluorescent proteins for in vivo imaging in an animal model system". In: *Molecular biology of the cell* 27.22, pp. 3385–3394.

HOEIJMAKERS, WIETEKE AM, RICHÁRD BÁRTFAI, and HENDRIK G STUNNENBERG (2013). "Transcriptome analysis using RNA-Seq". In: *Malaria: Methods and Protocols*, pp. 221–239.

HUZAIRA, MISBAH et al. (2001). "Topographic variations in normal skin, as viewed by in vivo reflectance confocal microscopy". In: *Journal of investigative dermatology* 116.6, pp. 846–852.

ISLAM, SAIFUL et al. (2014). "Quantitative single-cell RNA-seq with unique molecular identifiers". In: *Nature methods* 11.2, pp. 163–166.

JAITIN, DIEGO ADHEMAR et al. (2014). "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types". In: *Science* 343.6172, pp. 776–779.

KIMMEL, CHARLES B et al. (1995). "Stages of embryonic development of the zebrafish". In: *Developmental dynamics* 203.3, pp. 253–310.

KIVIOJA, TEEMU et al. (2012). "Counting absolute numbers of molecules using unique molecular identifiers". In: *Nature methods* 9.1, pp. 72–74.

KWON, YOUNG JOON et al. (2018). "High-throughput BioSorter quantification of relative mitochondrial content and membrane potential in living Caenorhabditis elegans". In: *Mitochondrion* 40, pp. 42–50.

LEVIN, MICHAL et al. (2016). "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596, pp. 637–641.

LORENZ, TODD C (2012). "Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies". In: *JoVE (Journal of Visualized Experiments)* 63, e3998.

MACCHIETTO, MARISSA et al. (2017). "Comparative transcriptomics of Steinernema and Caenorhabditis single embryos reveals orthologous gene expression convergence during late embryogenesis". In: *Genome biology and evolution* 9.10, pp. 2681–2696.

O'REILLY, LINDA P et al. (2014). "C. elegans in high-throughput drug discovery". In: *Advanced drug delivery reviews* 69, pp. 247–253.

OIKONOMOPOULOS, SPYROS et al. (2020). "Methodologies for transcript profiling using long-read technologies". In: *Frontiers in genetics* 11, p. 606.

PACKER, JONATHAN S et al. (2019). "A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution". In: *Science* 365.6459, eaax1971.

PAREKH, SWATI et al. (2016). "The impact of amplification on differential expression analyses by RNA-seq". In: *Scientific reports* 6.1, p. 25533.

PAREKH, SWATI et al. (2018). "zUMIs-A fast and flexible pipeline to process RNA sequencing data with UMIs". In: *Gigascience* 7.6, giy059.

PEREZ-MOJICA, J EDUARDO et al. (2023). "Continuous transcriptome analysis reveals novel patterns of early gene expression in Drosophila embryos". In: *Cell genomics* 3.3.

PICELLI, SIMONE et al. (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nature methods* 10.11, pp. 1096–1098.

PICELLI, SIMONE et al. (2014). "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature protocols* 9.1, pp. 171–181.

PRINYAKUPT, JAROONRUT and CHARNCHAI PLUEMPITIWIRIYAWEJ (2015). "Segmentation of white blood cells and comparison of cell morphology by linear and naive Bayes classifiers". In: *Biomedical engineering online* 14, pp. 1–19.

REY, CARINE et al. (2022). "Programmed-DNA Elimination in the free-living nematodes Mesorhabditis". In: *bioRxiv*, pp. 2022–03.

RIDDLE, DONALD L et al. (1997). "C. elegans ii". In.

RODRIGUES, NELIO TL et al. (2022). "SAIBR: A simple, platform-independent method for spectral autofluorescence correction". In: *Development* 149.14, dev200545.

SERRA, LORRAYNE et al. (2018). "Adapting the smart-seq2 protocol for robust single worm RNA-seq". In: *Bio-protocol* 8.4, e2729–e2729.

STARK, RORY, MARTA GRZELAK, and JAMES HADFIELD (2019). "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11, pp. 631–656.

STOECKIUS, MARLON et al. (2009). "Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression". In: *Nature methods* 6.10, pp. 745–751.

STROUSTRUP, NICHOLAS et al. (2013). "The Caenorhabditis elegans lifespan machine". In: *Nature methods* 10.7, pp. 665–670.

SVENSSON, VALENTINE, ROSER VENTO-TORMO, and SARAH A TEICHMANN (2018). "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nature protocols* 13.4, pp. 599–604.

SVENSSON, VALENTINE et al. (2017). "Power analysis of single-cell RNA-sequencing experiments". In: *Nature methods* 14.4, pp. 381–387.

TINTORI, SOPHIA C et al. (2016). "A transcriptional lineage of the early C. elegans embryo". In: *Developmental Cell* 38.4, pp. 430–444.

VIETH, BEATE et al. (2019). "A systematic evaluation of single cell RNA-seq analysis pipelines". In: *Nature communications* 10.1, p. 4667.

WHITE, RICHARD J et al. (2017). "A high-resolution mRNA expression time course of embryonic development in zebrafish". In: *elife* 6, e30860.

ZHANG, SIHUI and JEFFREY R KUHN (2018). "Cell isolation and culture". In: *WormBook: The Online Review of C. elegans Biology [Internet]*.

ZHANG, XIANNIAN et al. (2019). "Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems". In: *Molecular cell* 73.1, pp. 130–142.

ZIEGENHAIN, CHRISTOPH et al. (2017). "Comparative analysis of single-cell RNA sequencing methods". In: *Molecular cell* 65.4, pp. 631–643.

3. Parental pheromone perception alters gene expression of the developing nervous system

Contents

Abstract

Social cues perceived by sensory neurons in *C. elegans* can impact the generation time of their progeny. Gene expression of single embryos across development reveals that besides delaying the germline of progeny, parental pheromone exposure likely alters the development of their nervous system, and particularly of their sensory organs. We hypothesize that these changes could alter perception of and reaction to the environment in the progeny, possibly influencing important life decisions such as *dauer* entry.

3.1 Introduction

The environment is a major source of phenotypic variation in organisms. Although some changes can be passive, for example reduced growth due to lack of nutrients (VAN KLEUNEN & FISCHER, 2005), organisms also actively react through sensory perception of their environment often causing substantial changes.

In *C. elegans* nematodes, the perception of ascaroside pheromones can influence behavior such as aggregation (SRINIVASAN et al., 2012) or foraging (GREENE et al., 2016b; CHUTE et al., 2019), but can also regulate developmental speed (LUDEWIG et al., 2019) and even lifespan (LUDEWIG et al., 2013; MAURES et al., 2014). Furthermore, high concentrations of pheromone (signaling a poor environment with high population density) cause immature larvae to enter a developmental arrest state known as *dauer*, that is highly resistant to various stresses and can go months without food (GOLDEN & RIDDLE, 1984; BUTCHER et al., 2007). Pheromone perception thus not only influences the behavior of *C. elegans*, but also causes lasting developmental and physiological changes.

Pheromones can also elicit changes across generations. PEREZ et al., 2021a recently showed that dauer pheromone exposure quantitatively controls the generation time of progeny by delaying development of their germline. They demonstrate pheromone perception (implicating chemosensory neurons such as ASI and AWC) and TGF-β ligand DAF-7 (produced in ASI) are required in the parents for the progeny phenotype. They further show that the main downstream effector of the TGF-β pathway, DAF-3, is required in the progeny but dispensable in the parents. However, a systematic characterization of the molecular changes that occur in the progeny upon parental perception of pheromone is still lacking.

In this chapter, I describe transcriptional changes induced by parental exposure to pheromone. With preliminary gene expression profiling and analysis of single embryos across development, we find exciting leads into potential physiological and behavioral changes induced by parental perception of the social environment, as well as hints to signal transmission.

3.2 Experimental design, data collection and analysis

To better understand the changes occurring in the progeny of pheromone-exposed parents, we profiled the transcriptome of single individuals from pheromone-exposed (PHE) and control (CTR) parents across embryo development (Fig. 3.1a,b), using an adapted Smart-seq3 RNA-seq protocol (HAGEMANN-JENSEN et al., 2020, see Chapter 2). Parents were exposed to control- or crude pheromone-treated plates as previously described (Fig. 3.1a, PEREZ et al., 2021a), and the effect of pheromone was confirmed through the soma-germline delay in the progeny of pheromone-exposed parents (Fig. 3.1c, Methods).

Inferring age from gene expression with RAPToR confirms our samples span most of embryogenesis in both control and pheromone conditions (Fig. 3.1b). Of note, embryos from pheromone-exposed parents collected after a timed egg-lay targeting 400-600 min past 4-cell were on average older than controls, though non-significantly ($p>0.1$, Fig. 3.1d). This was not the case for an earlier developmental window target, and suggests that embryo development could be slightly accelerated by parental exposure to pheromone.

We find developmental expression dynamics match previously published single-embryo profiling data (see Chapter 2) and conclude our data accurately reflects gene expression.

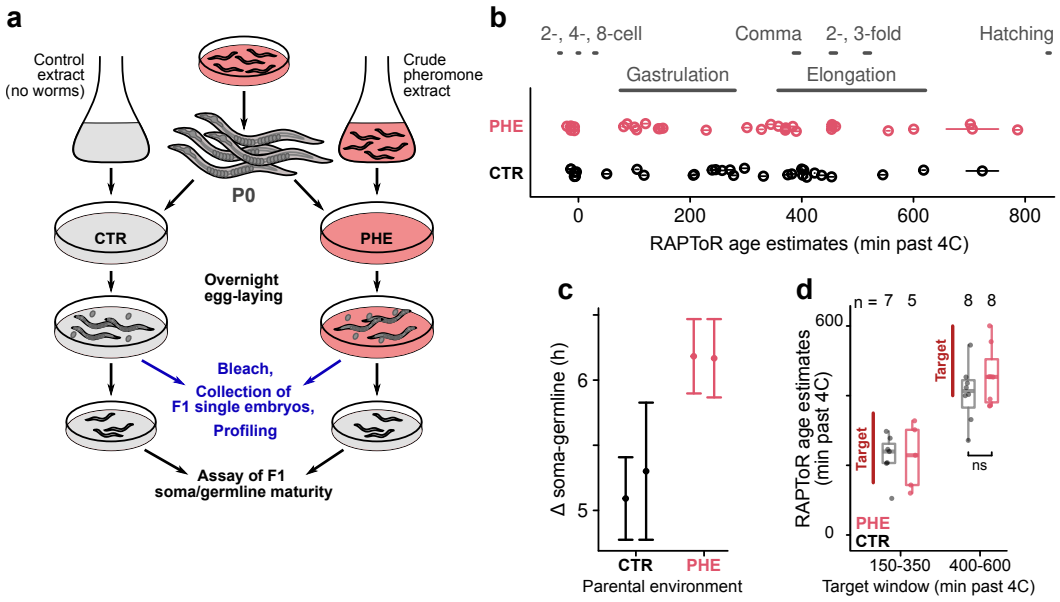


Figure 3.1 – Single-embryo profiling to characterize the molecular changes caused by parental exposure to pheromone

a, Experimental schematic showing the division of a synchronized adult population to plates conditioned by either control or crude pheromone extracts, produced by filtering liquid media in which worms had been cultured for 3-4 days (or no worms, for control). The parental worms (P0) were then left overnight to lay eggs, before the plate contents were bleached to collect F1 single embryos for gene expression profiling. A portion of the F1 eggs were left to hatch on a fresh plate and assayed for soma-germline delay.

b, RAPToR age estimates of profiled embryos from pheromone-exposed parents (PHE, $n = 32$) or controls (CTR, $n = 31$). Horizontal segments denote RAPToR bootstrap confidence intervals. Landmark embryogenesis events are indicated above for reference, timing in minutes past 4-cell stage at 20°C (min past 4C).

c, Time between L4/young-adult molt and appearance of the first embryo (Δ soma-germline), as described in PEREZ et al., 2021a. Error bars denote 95% confidence intervals.

d, RAPToR age estimates for embryos collected after a timed egg-lay. One-sided t-test between age estimates of CTR and PHE embryos for the second window, p=0.13, non-significant (ns).

3.3 Parental exposure to pheromone induces neuron-related changes in gene expression during embryo development

3.3.1 Overall expression changes point to the developing nervous system

We find a total of 1317 DE genes at a false discovery rate (FDR) of 0.1 (see Methods), with overall tissue enrichment in the nervous system (Table 3.1). Gene ontology (GO) enrichment shows a strong enrichment in chromatin-related elements due to the presence of many differentially expressed histone genes, alongside several categories consistent with the nervous system (synapse-related elements). Of note, we find no significant difference in interindividual variability of gene expression between conditions (see Methods), or particular bias in genomic position or GC content for DE genes (Appendix D, Sup. Fig. D.1).

	Term	Observed	Enrichment FC	Q.value
Tissue	nervous system	594	1.24	2.08E-05
GO	structural constituent of chromatin	29	4.81	9.29E-12
	protein heterodimerization activity	30	3.78	3.04E-09
	nucleosome	22	4.27	4.04E-08
	postsynaptic membrane	19	2.05	5.63E-02
	regulation of postsynaptic membrane potential	16	2.12	5.63E-02
	extracellular ligand-gated ion channel activity	18	2.12	5.63E-02
	chemical synaptic transmission postsynaptic	17	2.10	5.63E-02

Table 3.1 – Global enrichment of differentially expressed genes
GO, Gene ontology; Enrichment FC, Enrichment Fold-Change

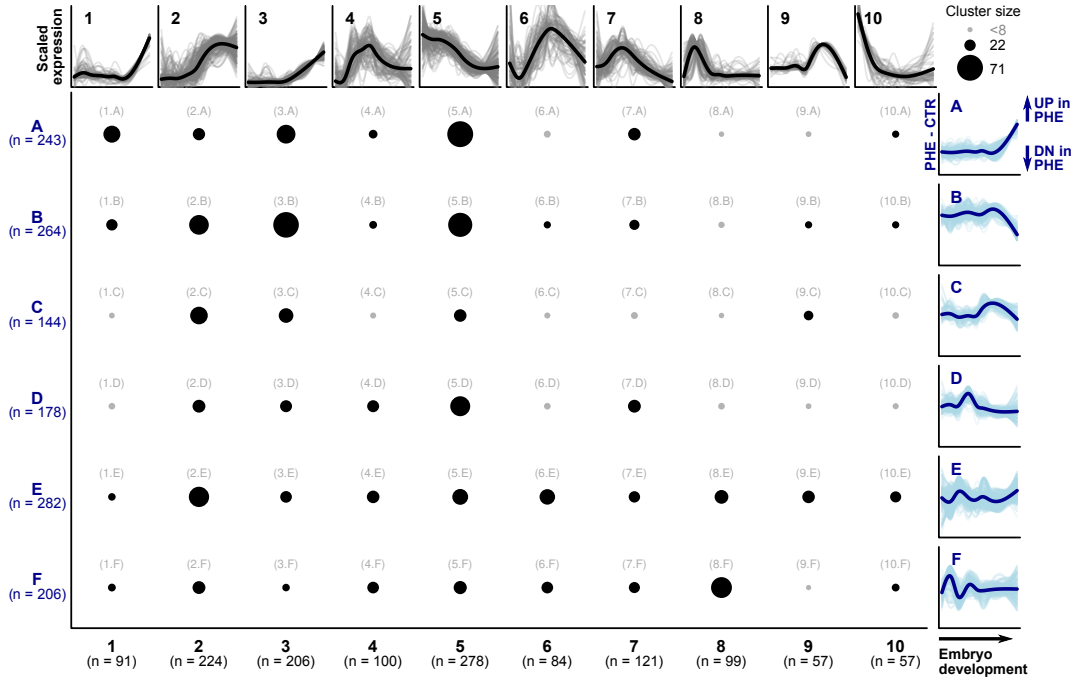


Figure 3.2 – Clustering genes according to development dynamic and differential expression
1317 DE Genes are separately clustered according to their expression dynamic along embryo development (1-10, dynamics are fit using all samples regardless of condition), and according to differential expression (A-F). Grey dots indicate sub-clusters with < 8 genes, for which enrichment was not tested. See Tables 3.2, 3.3, 3.4 for tissue enrichment and Appendix D, Sup. Fig. D.2, D.3, D.4, D.5, D.6 for GO and phenotype enrichment.

We then cluster DE genes by developmental dynamic and differential expression to better understand the processes and tissues affected by parental exposure to pheromone (Fig. 3.2). Expression dynamic clusters (1-10) allow us to see that genes expressed along the whole of embryo development are impacted. Clusters 1, 2, 3, and 9 group genes expressed in the late embryo, clusters 4, 6, and 7, genes peaking during mid-embryo development, while clusters 5, 8, and 10 involve genes expressed in early stages.

Further clustering according to differences between conditions (see Methods), then helps better understand the processes and tissues that are impacted.

3.3.2 Late embryo development

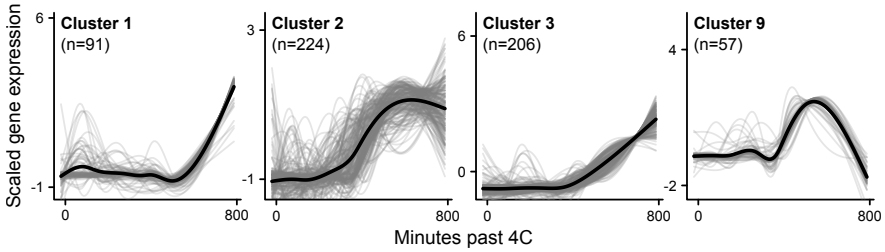


Figure 3.3 – Differentially expressed genes peaking in late embryogenesis
Expression dynamic clusters of DE genes, as defined in Fig. 3.2.

We find that clusters corresponding to expression in later embryo development (1, 2, 3, and 9, Fig. 3.3) are strongly enriched in the nervous system, notably the lateral and retrovesicular ganglia, where the cell bodies of chemosensory organs are located (Table 3.2). Indeed, amongst enriched tissues, we find sensory neurons from amphids (AWC, ASI, ADL ASE, ADF) alongside interneurons known to specifically process their inputs (AIA, AIB, AIY, AUA, RIA, RIZ, AIN precursor), as well as the amphid sheath cells.

Genes involved in steps from sensory perception to signal transduction in various neurons contribute to these enrichments (Fig. 3.4). For example, SRG-53 is a G-protein coupled receptor that binds signal ligands and changes conformation, passing the signal to a G-protein α subunit such as GPA-11. Then, receptor Guanyl Cyclases (rGC) such as GCY-7 receive this signal and transmit them to Transient Receptor Potential (TRP) channels such as OCR-1, activating ion transfers to generate electric potential differences within neurons. Ligand-gated channels like LGC-51 can then transmit this potential to downstream neurons via neuropeptides such as NLP-5 that are received by neuropeptide receptors like NPR-14. We also remark that DAF-9, the enzyme responsible for production of dafachronic acid, downstream of the DAF-7/TGF-β signaling pathway, is up-regulated in the progeny of pheromone-exposed worms (Fig. 3.4).

Amphids are sensory organs that secrete factors regulating many traits, including fat storage (Hussey et al., 2017), foraging (Greene et al., 2016a), and lifespan (Ludewig et al., 2013). ASI and ADF in particular also regulate dauer-entry, through pheromone and temperature perception respectively, with ASI at the top of the signaling pathway as the main producer of DAF-7/TGF-β in *C. elegans*. Finding that these clusters are also associated with phenotypes such as variants in fat content, foraging, body shape, dauer, chemosensory behavior, or lifespan further supports the fact that the progeny response affects the amphids (Appendix D, Sup. Fig. D.2, D.3). We also remark that ASI and AWC neurons found enriched here are required for pheromone perception to delay progeny germline (Perez et al., 2021a).

After amphids, we note an enrichment of other head sensory neurons, such as outer and inner labial neurons (IL1, IL2), oxygen sensors (URX), putative chemosensors (URY, URA, URB), and proprioceptors (SMD, SAA). Then, multiple interneurons from the nerve ring (RIM, RIG, RIV, RIS) and surrounding head ganglia (AVE, AVD, AVH, AVK, AVB), are also strongly enriched. Together

Tissue	1	1.A	1.E	2	2.B	2.C	2.E	3	3.A	3.B	3.C	3.D	3.E	9	9.B	9.E
lateral ganglion				**101**		**16**	**29**	**86**		**29**			**15**			
somatic nervous system				**107**				**92**		**32**						
nervous system	**40**							**98**						**20**		
preanal ganglion				**38**			**13**	**28**		**12**						
DA neuron				**44**				**28**		**13**						
VA neuron				40				27		13						
outer labial neuron				**66**												
head mesodermal cell				65												
retrovesicular ganglion				**47**												
RIM				**21**			**8**	**12**		**6**						
RIS			**3**	**20**				**14**		**7**						
epithelial system		20												22		
ASE								**33**					**8**			
amphid sheath cell	8	5						16	7							
SMD				**15**			**6**	**10**				**4**				
AIY				**15**				**11**	**4**							
URX				**14**	**7**			**9**								
AWC				**14**				**9**		**4**						
ALM				**18**						**7**						
DB neuron				**25**												
AVE								**18**				**6**				
ADF								**18**					**5**			
PVW				9				8				5				
RIV				**16**	**6**											
RIA				**14**			**7**									
AVD				**20**												
ASI								**14**					**5**			
PVP				**12**			**6**									
RIB				**16**												
RID				**16**												
RIG				**13**								**3**				
M1 neuron	**4**							**7**		**4**						
RMD								**9**		**6**						
ADL								**14**								
BDU				**14**												
ALA				**13**												
AIB				**11**												
AVB				**11**												
AVH				**11**												
AVK										**11**						
AIA								**10**								
g1						**5**								**5**		
DVA								**9**								
rectal gland cell											**8**				**1**	
RIP							**3**	**4**	**2**							
VC neuron								9								
ADA								**5**				**3**				
anterior arcade cell					**8**											
HSN								**8**								
I2 neuron								**5**	**3**							
SAA				**6**						**2**						
URA								**5**		**3**						
URY								**5**				**3**				
g2														**4**		**3**
nociceptor neuron										**7**						
osmosensory neuron										**7**						
PVN								**5**					**2**			
RMG								**4**				**3**				
IL1 neuron										**6**						
AUA										**5**						
I4 neuron							**5**									
PDA			**2**							**3**						
URB								**5**								
DVC												**4**				
I3 neuron							**4**									
IL2													**4**			
K/K' cell														**4**		
MC neuron												**4**				
PVQ		**2**										**2**				
tail hypodermis														**4**		
CEM													**3**			
hook sensillum													**3**			
posterior arcade cell														**3**		
ray													**3**			
AVF												**2**				
RMH											**2**					
hyp4														**2**		
hyp5														**2**		
hyp6														**2**		

Table 3.2 – Tissue enrichment of late embryo development clusters
Clusters as defined in Fig. 3.2, with enrichment for the whole expression dynamic cluster (1, 2, 3, 9), and sub-cluster with ≥ 8 genes. Cell content and color corresponds the number of annotated genes in each category. Color code shows **neurons** in bold, sensory organs/neurons highlighted in yellow, interneurons processing sensory input highlighted in green, **excretory system cells** in blue, and **cells born post-embryonically** in grey. See also Appendix D, Sup. Fig. D.2, D.3.

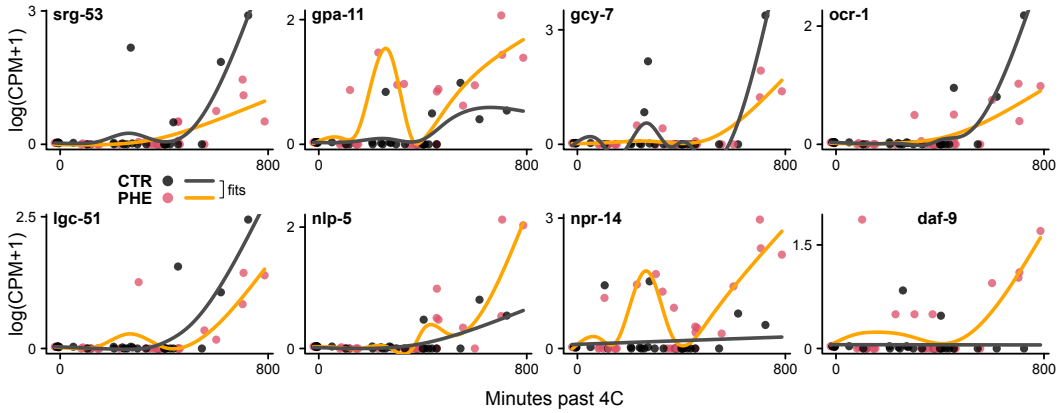


Figure 3.4 – Selected differentially expressed genes involved in neuron signaling
srg-53, *gpa-11*, *gcy-7*, *ocr-1*, *lgc-51*, *nlp-5*, *npr-14*, and *daf-9* expression in embryos from pheromone-exposed parents (PHE) and controls (CTR), with their respective fits.

with pharyngeal sensory neurons (I2, I3, I4, MC), this suggests that changes occur not just in genes specific to amphids, but also in other sensilla.

The excretory system also appears implicated, as excretory cells g1 and g2 are found amongst enriched tissues in different clusters, with excretion also appearing in gene ontology (Sup. Fig. D.2,D.3).

Multiple clusters are enriched in dorsal and ventral motor neurons (DA, DB) and the epithelial system, with a few specific cell types located in the head (mesodermal and arcade cells). We note that genes related to the epithelial system, extracellular region, and cell body are upregulated in late embryo development of pheromone-exposed parents (cluster 1.A).

Lastly, although most of the affected tissues are located in the head, genes differing in late embryo development are also associated to a few neurons located in the tail ganglia of the worm (PVP, PVQ, DVA, DVC, preanal ganglion).

We raise caution with respect to enrichments in post-embryonically born neurons (e.g., VA, VC), that are most likely due to gene sets overlapping with neuron classes born earlier (DA, HSN, respectively).

Overall gene ontology enrichment for these clusters (Appendix D, Sup. Fig. D.2, D.3) strongly supports neuron activity in both chemical and electrical synapse signaling, with hints to specific neuron function (taxis, chemosensation) and neuron development (cell projection). Given the enrichment in motor neurons, it is also possible that part of the genes related to neuron activity implicate embryo twitching (that starts around 400 min past 4-cell).

3.3.3 Mid embryo development

Unlike in later embryo stages, cell types are less defined during mid embryo development. Clusters of genes peaking during this period (4, 6, and 7, Fig. 3.5a) are however enriched in specific cell lineages that give rise to neurons (Table 3.3).

3.3.3.1 Neuron precursors are affected prior to differentiation

We find enrichment in precursors to ASI neurons, that notably only produce ASI neurons as the other daughter cells are programmed to die after cell division (SULSTON et al., 1983). Contributing to this enrichment, the gene encoding TGF-β/DAF-7 itself is downregulated in the progeny of pheromone-exposed individuals (Fig. 3.5b). Precursors to URX and HSN neurons, that were enriched in later embryo development expression changes, are also enriched here. This suggests changes occur in neurons also before their differentiation.

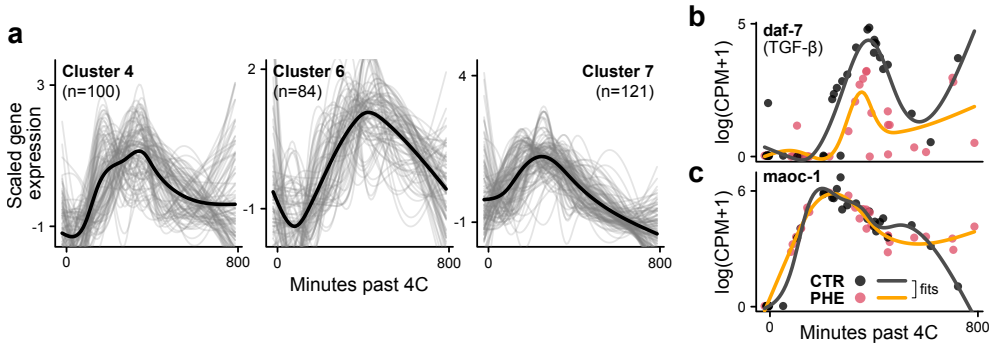


Figure 3.5 – Differentially expressed genes peaking during mid-embryogenesis
a,Expression dynamic clusters of DE genes, as defined in Fig. 3.2.
b,c, *daf-7* (**b**) and *maoc-1* (**c**) expression in embryos from pheromone-exposed parents (PHE) and controls (CTR), with their respective fits.

Tissue	Final cell types (if lineage)	4.B	4.D	6	7	7.D
ABplaapappp	**ASIL**			**8**		
ABpraapappp	**ASIR**			**8**		
ABarpaapap	hyp7				8	
ABarpaappa	hyp7				8	
ABarpaappp	hyp7				8	
ABarppaapa	hyp7				8	
ABarpppapa	hyp7				8	
ABplaapppa	hyp7				8	
ABplaapppp	hyp7				8	
ABplappppa	hyp7				8	
ABpraapppa	hyp7				8	
ABpraapppp	hyp7				8	
ABprappppa	hyp7				8	
ABplaaaaapp	**CEPDL, URXL**			**7**		
ABarpapaapp	**CEPDR, URXR**			**7**		
ABplapppapp	**HSNL, PHBL**			**7**		
ABprapppapp	**HSNR, PHBR**			**7**		
AVK			**4**			**3**
ADE sheath cell		3				
ADE socket cell		3				
amphid sheath cell		3				
amphid socket cell		3				
CEP socket cell		3				
IL sheath cell		3				
IL socket cell		3				
OL sheath cell		3				
PDE sheath cell		3				
PDE socket cell		3				
phasmid sheath cell		3				
phasmid socket cell		3				
anterior arcade cell		3				
B cell		3				
hyp4		3				
hyp5		3				
hyp6		3				
tail precursor cell		3				
Y cell		3				
ABaraapapaa	**NSML**		**2**			
ABaraapppaa	**NSMR**		**2**			
G cell		2				
hyp1		2				
hyp2		2				
tail hypodermis		2				
W cell		2				
XXX cell		2				

Table 3.3 – Tissue enrichment of mid embryo development clusters
Clusters as defined in Fig. 3.2, with enrichment for the whole expression dynamic cluster (6, 7, no enrichment found at dynamic level for cluster 4), and sub-cluster with ≥ 8 genes. Cell content and color corresponds the number of annotated genes in each category. Color code shows **neurons** in bold, sensory organs/neurons highlighted in yellow, and **excretory system cells** in blue. See also Appendix D, Sup. Fig. D.4.

3.3.3.2 A peak of expression around gastrulation in the progeny of pheromone-exposed individuals

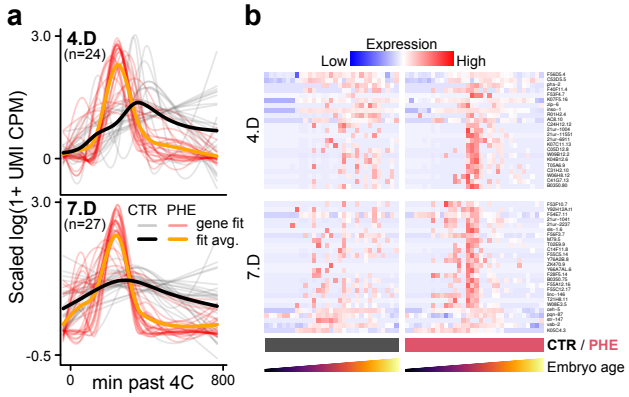


Figure 3.6 – Mid-embryogenesis expression peak in the progeny of pheromone-exposed worms
a,b, Scaled expression dynamics (**a**) and heatmap of underlying data (**b**) per gene and condition in subclusters 4.D and 7.D, as defined in Fig. 3.2.

We notice that pheromone induces a clear expression peak around 250 min past 4C (end of gastrulation) in clusters 4.D and 7.D (Fig. 3.6). Both of these clusters are enriched in genes associated with the AVK neuron class, and cluster 7.D is also enriched in synaptic signaling, indicating neuron activity (Sup. Fig. D.4). This is surprising given AVK interneurons are born precisely at the timing of this peak (265 min past 4C, SULSTON et al., 1983), with activity (suggested by *flp-1* neuropeptide expression, (HUMS et al., 2016)) and axon development starting at elongation (around 75 min later, MUCH et al., 2000). In mature animals, AVK neurons have cell bodies posterior to the pharynx and an axon extending towards the anterior before looping around the nerve ring and running along the full length of the worm; they are involved in food-related locomotion behavior, notably dispersal (HUMS et al., 2016; ORANTH et al., 2018).

Of note, although the majority of genes are tissue-annotated in both clusters, few are annotated for GO or phenotypes. This may be due to the fact most of these genes are non-coding elements, including 5 piRNAs, (21-ur transcripts), 2 long intervening non-coding RNAs (linc) *linc-146* and *linc-147*, as well as trans-spliced leader sequences (sls, 2) *sls-1.6* and *sls-2.4*, a tRNA, and multiple other small ncRNAs and snoRNA. Together, these results suggest regulatory expression changes implicating or affecting AVK neuron development.

We note the continued presence of the epithelial (hyp7 precursors) and excretory (G cell) systems amongst enriched tissues, and point out that *maoc-1*, a gene implicated in ascaroside synthesis (VON REUSS et al., 2012), is upregulated in the progeny of pheromone-exposed worms (Fig. 3.5c).

3.3.4 Early embryo development

Clusters of DE genes expressed in the earliest embryo stages (clusters 5, 8, 10, Fig. 3.7a), are distinct both in dynamics and enriched terms (Table 3.4). Cluster 10 exclusively groups maternally-contributed genes (expressed in 4-cell and earlier embryos) whose expression sharply decreases, and is enriched in germline (and precursors, Z2, Z3), oocyte differentiation and metabolism regulation, consistent with early-embryo development. Cluster 8 mainly corresponds to genes that are expressed by the embryo itself at the onset of zygotic transcription, and appears enriched in intestine-related genes, as well as immune response and dauer metabolism (Sup. Fig. D.5, D.6).

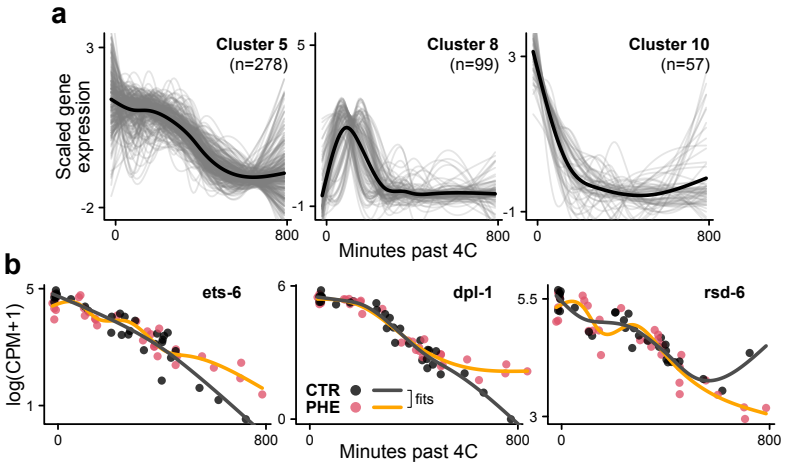


Figure 3.7 – Differentially expressed genes peaking at the start of embryogenesis
a,Expression dynamic clusters of DE genes, as defined in Fig. 3.2.
b, *ets-6*, *dpl-1*, and *rsd-6* expression in embryos from pheromone-exposed parents (PHE) and controls (CTR), with their respective fits.

Tissue	Final cell types (if lineage)	5	5.A	5.B	5.D	5.E	5.F	8	8.F	10	10.B	10.E	10.F
reproductive system		242	67	63	36	29	22			42			10
germ line		237	66	61	36	28	21			41			10
gonadal primordium		102		24	33	14							
nervous system		**144**											
thermosensory neuron		**58**			**16**								
intestine			40						19				
striated muscle		45			13								
AVA		**28**											
Caapa	**DVC**	**16**			**7**								
male								13	9				
ABplapappp	**ALNL, PLML**	**15**		**7**									
ABprapappp	**ALNR, PLMR**	**15**		**7**									
ABalappppa	**IL1DR, IL2DR**	**15**		**7**									
ABarappppp	**IL1VR, IL2VR**	**15**		**7**									
ABalapaapp	**IL1L, IL2L**	**14**		**7**									
ABalaapppp	**IL1R, IL2R**	**14**		**7**									
ABalppappp	**IL1VL, IL2VL**	**14**		**7**									
ABaraappaa	MI, pm1DR	14			7								
ABplapaaap	**AIZL, FLPL, RMGL**	**12**			**6**								
ABprapaaap	**AIZR, FLPR, RMGR**	**12**			**6**								
ABalpapaap	arc ant V	12		6									
ABarappaap	hyp1	12		6									
head mesodermal cell										17			
ABarpaaa	**CEPshDL, CEPshDR, OLQshDL, OLQshSR, OLQsoDL, OLQsoDR**	**12**			**5**								
MSpaaapa	**I6, M5**	**12**			**5**								
ABalpppppa	**ADFL, AWBL**	**11**		**5**									
ABpraaappa	**ADFR, AWBR**	**11**		**5**									
ABplaaaaaa	**CEMDL, URADL**	**12**				**4**							
ABarpapaaa	**CEMDR, URADR**	**12**				**4**							
ABplpaaapa	**CEPshVL, URAVL**	**12**				**4**							
ABprpaaapa	**CEPshVR, URAVR**	**12**				**4**							
MSpaaaaa	**M4**	**10**		**6**									
ABalapapap	**OLQDL, URYDL**	**12**			**4**								
ABalapppap	**OLQDR, URYDR**	**12**			**4**								
ABplpaaapp	**OLQVL, URYVL**	**12**			**4**								
ABprpaaapp	**OLQVR, URYVR**	**12**			**4**								
ABalpppapa	**AFDL, RMDL**	**10**			**5**								
ABpraaaapa	**AFDR, RMDR**	**10**			**5**								
MSaapaapa	g1AL	10			5								
MSpapaapa	g1AR	10			5								
ABprpppppap	**B, DVA**	**10**				**4**							
ABplpppppap	F, U	10				4							
ABalapaaaa	**AVHL**	**12**											
ABalappapa	**AVHR**	**12**											
ABplpppaapa	**LUAL, PVCL**	**12**											
ABprpppaapa	**LUAR, PVCR**	**12**											
ABarappppa	**BAGR, SMDVR**	**11**											
ABalpppppp	**ASEL, ASJL, AUAL**	**10**											
ABpraaappp	**ASER, ASJR, AUAR**	**10**											
ABalpppppppp	**ASJL, AUAL**	**10**											
ABpraaapppp	**ASJR, AUAR**	**10**											
ABalppappa	**BAGL, SMDVL**	**10**											
ABplpappaa	**RMEV, exc_cell**	**10**											
ABplapaaaa	**ADAL, ADEL**			**6**									
ABprapaaaa	**ADAR, ADER**			**6**									
ABplpapapa	**AVKL, exc_gl_L**			**5**									
ABprpapapa	**AVKR, exc_gl_R**			**5**									
MSpaapaa	**M1**			**5**									
somatic cell		4											

(Continued on next page)

Tissue	Final cell types (if lineage)	5	5.A	5.B	5.D	5.E	5.F	8	8.F	10	10.B	10.E	10.F
ABprpppppp	hyp10				4								
ABplpppppp	hyp10, spike				4								
MSpapp	mu_bod	3										1	
ABalapaapa	**RIAL**				**4**								
ABalaapppa	**RIAR**				**4**								
ABalaaaarl	**RMEL**				**4**								
ABalaaaarr	**RMER**				**4**								
Z2										2			1
Z3										2			1
ABalpppa	**AFDL, ASKL, AVEL, OLLshL, OLQshVL, RMDL**	**3**											
ABplppaa	**AIAL, DB6, RICL, RIML, SIBDL**	**3**											
ABplaapap	**AIBL, ASGL, ASIL, AWAL**	**3**											
ABplpapa	**AIYL, AVKL, DB5, SIADL, SIAVL, SIBVL, SMDDL, exc_gl_L**	**3**											
ABplapapp	**ALNL, P11, PLML**	**3**											
ABplaapaa	**AMshL, ILsoDL, URBL, hyp3**	**3**											
ABalpaap	arc ant DL, arc post DL, arc post VL, e2DL, hyp1, hyp2, pm3L, mc1DL	3											
ABalppaa	**AVAL, CEPsoVL, OLQsoVL**	**3**											
ABalppap	**BAGL, IL1VL, IL2VL, ILshVL, ILsoVL, RMDVL, SAADL, SMDVL**	**3**											
ABplppap	**DA2, DA4, DD1, DD3, DD5, RIFL, RIGL, SABD, SABVL**	**3**											
ABplpppa	**DA6, DA9, LUAL, PHAL, PHshL, PVCL, hyp8/9**	**3**											
ABplaaapp	HOL, H1L	3											
ABplaappa	hyp4, hyp6	3											
ABplaappp	hyp7	3											
Ea	int							2				1	
Eal	int							2				1	
Ep	int							2				1	
ABalpaaa	MCL, e3VL, pm1VL, pm2L, pm2VL	3											
ABalpapa	**RIPL, RMDDL, SMBDL, SMBVL, arc ant V, hyp2**	**3**											
ABplaaapa	XXXL, hyp5	3											

Table 3.4 – Tissue enrichment of early embryo development clusters
Clusters as defined in Fig. 3.2, with enrichment for the whole expression dynamic cluster (5, 8, 10), and sub-cluster with ≥ 8 genes. Cell content and color corresponds the number of annotated genes in each category. Color code shows **neurons** in bold, sensory organs/neurons highlighted in yellow, interneurons processing sensory input highlighted in green, **excretory system cells** in blue, and **cells born post-embryonically** in grey. See also Appendix D, Sup. Fig. D.5, D.6.

The largest group, cluster 5, corresponds to monotonically decreasing genes along embryogenesis, and is largely enriched in the germline. This includes germline-related (but not specific) genes such as *ets-6* and *dpl-1* (transcription factors negatively regulating vulva development), or *rsd-6*, involved in P-granule organization (Fig. 3.7b). However, germline enrichment is also largely explained by genes related to cell proliferation that are expressed in the germline of mature worms, and are expected to decrease along embryogenesis, as supported by GO and phenotype enrichment in mitosis-related categories (Sup. Fig. D.5, D.6)). The large overlap between neuron-related and germline categories further shows that many genes contributing to germline enrichment are non-specific (e.g. in cluster 5, n=278, of which 237 contribute to germline enrichment and 144 to the nervous system, Table 3.4).

3.3.4.1 Sensory organ lineages are affected in early development

Consistent with enrichment results in late and mid embryo development clusters, we find that genes in these early clusters are associated with precursors of sensory organs and neurons and their downstream interneurons. Notably, precursors to ASE, ASI, ASJ, ASK, ADF, AWA, and AWB chemosensors, the AFD thermosensor, and AIA, AIZ, AVH, RIA, RIG interneurons processing sensory input. We further remark enrichment in genes linked to precursors of several sheath and socket cells, including amphids.

Precursors to inner and outer labial neurons (IL, OL), as well as several putative sensory neurons (URA, URB, URY, AUA) seen in enrichments of late embryo development are also found here, as well as both the AVK neuron class, and the excretory system.

Together, these enrichments suggest that development of the nervous system, particularly sen-

sory neurons and organs, is affected starting from early cell lineages and throughout embryogenesis.

3.3.4.2 Prolonged expression of several histones

We find 29 histones amongst DE genes, most (24) of which are in cluster 5.D (Fig. 3.8). In progeny from control worms, the expression of these genes decreases after peaking around 100 min past 4C, while it persists longer in progeny of pheromone-exposed individuals, before sharply decreasing, often to a lower level than controls. Histones genes explain the enrichment of chromatin and nucleosome-related GO and phenotype categories seen in this cluster (Sup. Fig. D.5, D.6). Of note, some of these histones are known as specific to head neurons (*his-4* and *his-13*).

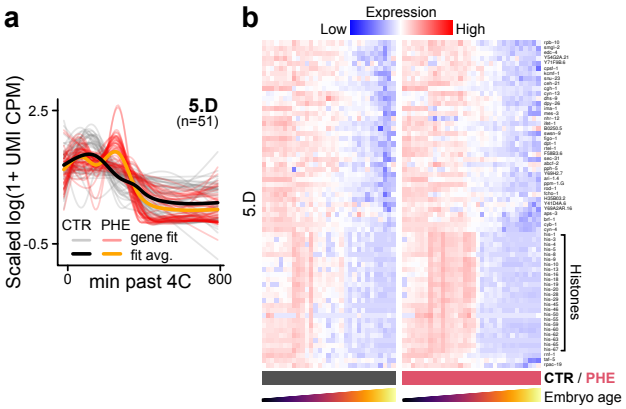


Figure 3.8 – Prolonged expression of histones in the progeny of pheromone-exposed worms
a,b, Scaled expression dynamics (**a**) and heatmap of underlying data (**b**) per gene and condition in subcluster 5.D as defined in Fig. 3.2. Histone genes are indicated on heatmap.

3.3.5 Insights into signal transmission

Transcripts prior to the start of transcription during the 4-cell stage are maternally contributed, and their differential expression is therefore caused by differences in maternal RNA loading and could give hints as to transmitted signals from parents. To investigate this, we tested for differences specifically in the samples staged at 4-cell and below amongst our DE genes (see Methods), resulting in 18 up-regulated and 17 down-regulated genes in the progeny of pheromone-exposed individuals (Fig. 3.9).

Surprisingly, several of these genes (up- or down-regulated) are also neuron- (even amphid) related, which is supported by significant enrichment of neuron precursors and neuron-related GO categories (Table 3.5, Fig. 3.9). For example, *lgc-29* and *unc-38* are both ligand-gated channels involved in synaptic signaling, while *Y71F9AL.7* and *glb-29* are mainly expressed in sensory neurons. We also remark that *W09H1.1* was recently implicated in genetic variation of ascaroside biosynthesis LEE et al., 2023. Of note, we find no apparent changes in maternal loading of translational machinery, in contrast to a previously reported intergenerational pheromone effect (WASSON et al., 2021).

These results suggest either differential production of neuron-related transcripts in the maternal germline for embryo loading in response to neuron input, or transfer of neuronal RNAs directly to the germline and embryos. The latter has yet to be explicitly shown for endogenous products, but exogenous dsRNA produced in neurons can enter the germline in *C. elegans*, even provoking transgenerational silencing effects (DEVANAPALLY et al., 2015), while extracellular (exogenous) RNAs can be loaded into oocytes together with yolk (MARRÉ et al., 2016). The export of endogenous neuronal RNAs to the germline and to embryos is thus within the realm of possibilities. However, given that both require the dsRNA-selective importer SID-1, which is not the

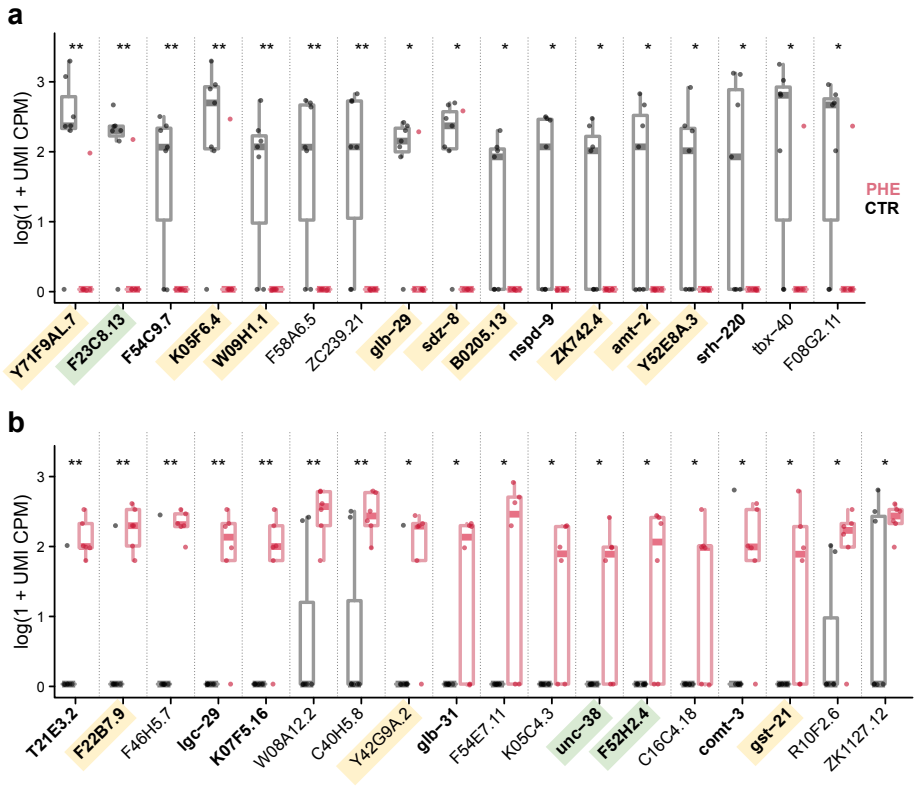


Figure 3.9 – Differential maternal transcript loading due to pheromone perception
a,b, DE genes with significant down (**a**) and up (**b**) regulation in embryos staged under 10 min past 4C of pheromone-exposed parents (red) compared to controls (black).
Labels show genes expressed mainly in **neurons** in bold, sensory organs/neurons highlighted in yellow, and interneurons processing sensory input highlighted in green. Boxes span the interquartile range (IQR), the central bar dot denotes the median, and whiskers extend to 1.5 × IQR in either direction, $n = 7$ and $n = 6$ for CTR and PHE respectively. **: $p < 0.01$, *: $p < 0.05$ of a t-test.

	Enriched Term	Final cell type (if lineage)	Observed	Enrichment FC	Q.value
Tissue	AB arappppp	IL1 / IL2 (VR)	3	11.16	0.092
	AB alapaapp	IL1 / IL2 (L)	3	11.27	0.092
	AB alaapppp	IL1 / IL2 (R)	3	11.08	0.092
	AB alppappp	IL1 / IL2 (VL)	3	11.04	0.092
	AB alappppa	IL1 / IL2 (DR)	3	11.04	0.092
	AB prapaaaa	ADA / ADE (R)	2	7.44	0.092
	AB plapaaaa	ADA / ADE (L)	2	7.41	0.092
GO	extracellular ligand-gated ion channel activity		2	23.8	0.019
	ligand-gated channel activity		2	17.8	0.019
	chemical synaptic transmission postsynaptic		2	25.0	0.019
	postsynaptic membrane		2	21.8	0.019
	regulation of postsynaptic membrane potential		2	26.8	0.019
	synaptic signaling		2	9.9	0.029
	passive transmembrane transporter activity		2	6.9	0.065
	transporter activity		3	4.3	0.088
	transmembrane transport		3	4.3	0.088

Table 3.5 – Enrichment of maternally loaded up-regulated genes
Only categories with at least 2 contributing genes are shown. No significant enrichments were found for down-regulated genes. GO, Gene ontology; Enrichment FC, Enrichment Fold-Change.

case for the intergenerational germline delay (Fig. In.6d, PEREZ et al., 2021b), then signal transmission of pheromone perception from parents to progeny either does not involve its canonical machinery, or operates through distinct pathways to generate the germline delay and transcriptional changes described here.

To summarize, we find differences in maternally-loaded transcripts related to neurons, that are distinct from a previous report of intergenerational effect of pheromone.

3.4 Discussion

In this chapter, I have described transcriptional changes induced by parental exposure to pheromone in developing *C. elegans* embryos, implicating many changes to the developing nervous system, and notably sensory organs.

We consistently found genes specific for amphids and sensory neurons, particularly ASI and ADF, and their downstream interneurons differentially expressed along embryo development. Ontology enrichments also suggest that both neuron development and (possibly as a consequence) function, notably signaling, are largely altered by parental pheromone exposure. We further reported intriguing effects implicating the AVK neuron and early expression of histone genes, as well as differences in neuron-specific maternally-loaded transcripts.

Although these results are convincing, we acknowledge they should be reinforced by collecting more data. Indeed, further sampling will add weight to many changes we find in late embryo development that currently rely on few data points, and fill in the gaps of early development before gastrulation. The increase in sample size will perhaps also allow us to find differences in interindividual variability between conditions that we are currently unable to detect. We also remark that many genes we found differentially expressed have little to no functional annotation, meaning we are likely missing part of the picture. Nevertheless, given the tissues and processes we find affected at the transcriptional level, our characterization suggests several potential phenotypes to score in response to parental pheromone exposure.

First, given the implication of AS and AW chemosensory neuron classes, progeny responses to known attractant or repellent solubles and volatiles (including dauer pheromones) could be altered, possibly in a cross-generational feedback loop similar to previously-described olfactory imprinting (REMY, 2010). This can be tested through chemotaxis assays (e.g. BARGMANN et al., 1993). Then, AVK neurons are known to be involved in food-related locomotion (HUMS et al., 2016; ORANTH et al., 2018), and therefore assaying changes in foraging and roaming behavior (e.g. GREENE et al., 2016b; GREENE et al., 2016a) could help understand the consequences of the observed expression changes. Together, these tests would help understand how progeny behavior is affected by the parental experience.

Then, since ASI, ADF, and dauer-related categories were recurrent in enrichment results, testing if the propensity of progeny to enter dauer in unfavorable environments is increased by parental pheromone exposure would help understand whether the phenomenon is adaptive. Preliminary experiments from our lab done by Marie-Alice Miniassian (data not shown) suggest this could be the case.

Although challenging, characterizing the excretome of young individuals from pheromone-exposed or control worms could also shed light onto the implication of the excretory system, and of genes required for ascaroside production like *maoc-1* (VON REUSS et al., 2012). Perhaps not only pheromone perception, but also production could be altered in the progeny.

Differential maternal transcript contributions to embryos lack the clear translational elements that were reported in another study (WASSON et al., 2021), despite the fact that crude pheromone extract should also bear the non-dauer metabolites responsible for the effect. A possible explanation could be that WASSON et al., 2021 profiled wild-type and *flp-21* mutant embryos, and therefore did not measure the direct effect of parental pheromone perception on gene expression but the

absence of the pathway responsible for parental signal transduction (mediated by *flp-21*). We also do not exclude the possibility of the effects reported here acting antagonistically or synergistically.

Although several genes are differentially contributed by mothers to embryos, we have yet to find convincing candidates for signal transmission, but this could be further investigated by profiling mutants of interest. DAF-7/TGF-β is required in parents to transmit the signal (PEREZ et al., 2021a), and is down-regulated in the progeny according to our data. Furthermore, maternal *daf-7* can rescue constitutive dauer entry in *daf-7*-null progeny (KLABONSKI et al., 2016), even over several generations according to experiments from our lab performed by Noémie Brisemeur (data not shown). Perhaps, DAF-7/TGF-β itself could be the transmitted signal. Germline delay is difficult to assay in *daf-7* mutant progeny due to strongly unsynchronized populations, but we could test this by profiling (and staging) *daf-7* null mutants born from either *daf-7* null or heterozygous mothers. If maternal DAF-7/TGF-β provided to early embryos rescues gene expression to a state comparable to controls in this study, while non-rescued mutants resemble progeny of pheromone-exposed worms, this would show DAF-7/TGF-β is required in the progeny for a control phenotype, thus bearing the signal.

To summarize, I have shown that parental perception of pheromones influences gene expression throughout embryogenesis, notably impacting the developing nervous system and sensory organs. These results suggest that parental perception of the social environment may alter how the progeny perceives and reacts to their environment and suggest specific behavioral and physiological changes to experimentally test.

3.5 Methods

3.5.1 Sample collection

3.5.1.1 Nematode cultures

C. elegans worms of the standard "N2" laboratory strain were kept at 20°C in roughly synchronized populations on 60mm NGM agar plates seeded with OP50. 100 eggs are transferred to new plates every 3-4 days to maintain populations.

4 days prior to sample collection, synchronized worms (the P0 generation) were collected with a 1-hour L1 hatch, and seeded to fresh plates at a density of 100 worms per plate. On the day prior to collection, half of the (now adult) worms on each P0 plate were transferred to fresh plates treated with 0.3 mL of either control or crude pheromone extract. Plate contents were washed and bleached on the next day (as described in Chapter 2) to collect F1 embryos for profiling or seeding to assay germline delay.

3.5.1.2 Pheromone extract preparation and germline delay assay

Control and crude pheromone extract were produced as described in PEREZ et al., 2021a. Briefly, worms were grown in large-scale liquid NGM cultures at 20°C with OP50-1 (streptomycin resistant) *E. coli* bacteria for 4 days before removing worms, centrifuging, and filtering the clear fraction to obtain crude pheromone extract. Control extract was prepared in the same way without worms.

Soma-germline delay was scored as described in PEREZ et al., 2017. Briefly, the young-adult to adult molt (soma transition) and the appearance of the first embryo (germline transition) were monitored in worms on agar plates with a standard brightfield binocular, and the fraction of the population past each developmental transition was estimated by counting worms before and after the transition at time points before and after 50% of the population underwent transition.

3.5.1.3 RNA-seq library preparation and sequencing

A first set of 16 PHE and 16 CTR single embryos were manually staged to ensure collection across all embryo development. Timed egg-lays were done to further collect 15 CTR and 16 PHE

single embryos targeting 150-300 and 400-600 minutes past 4-cell windows of development. Libraries were prepared with the adapted Smart-seq3 protocol described in Appendix C. Both sets of samples were processed, pooled, and sequenced separately at 150 bp PE on an Illumina Novaseq 6000 instrument, with an average 4.5 million pairs of reads (M) per sample.

3.5.2 Data pre-processing

Raw fastq data was processed with the zUMIs pipeline (PAREKH et al., 2018) using the parameters provided by Smart-seq3 authors at `protocols.io`, applying no UMI sequence error correction (Hamming distance parameter set to 0), mapping to the *C. elegans* genome (WBcel235, annotation v109) using STAR (v2.5.4b, as provided in the zUMIs conda environment), with > 79% of uniquely mapped reads. UMI-corrected counts uniquely mapping to intron+exon were converted to counts per million (CPM) to account for UMI library size, and used in all following analyses.

$$\mathrm{CPM}_{\mathrm{UMI}} = \frac{\mathrm{count}_{\mathrm{UMI}} \times 1e^{6}}{\text{library size}_{\mathrm{UMI}}}$$

We used `RAPToR` (v1.2.0) (BULTEAU & FRANCESCONI, 2022) to estimate the age of all samples against the "`Cel_embryo`" reference from `wormRef` (v0.5). Age is reported in minutes past the 4-cell stage (min past 4C).

3.5.3 Gene expression analysis

We filtered out lowly expressed genes, keeping only those with at least 3 counts in 3 samples, leaving 15727 genes for analysis. UMI CPMs were then log-transformed ($log(X+1)$), and quantile-normalized with the *normalizeBetweenArrays* function of `limma` (v3.56.1).

3.5.3.1 Differential expression analysis

Time-series data is more complex to analyze than static perturbation experiments, and the usual tools used for analysis may not be appropriate (BAR-JOSEPH et al., 2012). Although DEseq (LOVE et al., 2014) or edgeR (ROBINSON et al., 2010) allow the use of splines to fit nonlinear dynamics, they require matching sample timings or constrain the spline dynamic across conditions. Therefore, we tested for differential expression by comparing Generalized Additive Models (GAMs) that include the pheromone condition variable (m1) to a null-hypothesis model which does not (m0) for each gene. Similarly to "between-class temporal differential expression" discussed by STOREY et al., 2005, a gene is considered differentially expressed (DE) if the fit of m1 is significantly better than m0, i.e. when pheromone explains a significant amount of variance in expression.

We fit GAMs using the *gam()* function of the `mgcv` package (v1.8-42), modeling development with a cubic regression spline on age, and with the following formulas:

```
(m0) : "~s(age, bs='cr', k=8)"
(m1) : "~s(age, bs='cr', k=8, by = cond) + cond"
```

The value of k (similar to spline degree of freedom) was defined by reaching a plateau in overall goodness of fit across all genes with m0 after testing values ranging 4-12.

A gene was considered differentially expressed (DE) when both FDR of an ANOVA between m0 and m1 was below 0.1, and m1 has a better fit (AIC) than m0, resulting in 1317 DE genes.

3.5.3.2 Clustering DE genes

We used m0 fits (predictions at 100 evenly spaced time points between earliest and oldest embryo timings, -20.6 and 786.4 min past 4C respectively) to cluster genes according to their expression dynamic along embryogenesis. A distance matrix was computed from the (gene-wise) scaled

m0 fits with the base R *dist()* function, on which we applied hierarchical clustering (base *hclust()* function, with method="ward.D2"). 10 clusters (1-10) were determined sufficient from silhouette and average dissimilarity indices ('avg.silwidth' and 'dunn2', respectively) computed with the *cluster.stats()* function of the `fpc` package (v2.2-10).

Similarly, m1 fits of pheromone and control conditions were subtracted (PHE – CTR), and the resulting matrix scaled gene-wise to cluster genes according to their differential expression using the same approach. 6 clusters (A-E) were determined sufficient, with the same indices as above.

Sub-clusters (e.g. "1.A" or "5.D") are simply the overlap between both clustering results, as shown in Fig. 3.2.

3.5.3.3 Enrichment analysis

Tissue, gene ontology (GO), and phenotype enrichments were performed using the local version of the wormbase `tea` tool (ANGELES-ALBORES et al., 2018), a q-value threshold of 0.1, and specifying the background as the 15727 expressed genes selected above.

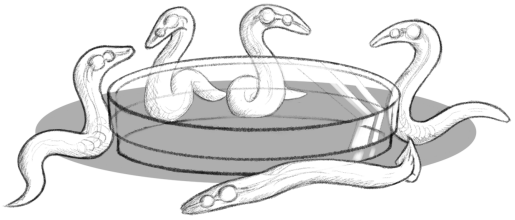
We tested for enrichment per dynamic cluster (1-10) and per subcluster with at least 8 genes, discarding enriched categories with less than 2 genes across clusters. Redundant tissue categories with identical gene sets were collapsed (e.g. IL2L and IL2R collapsed into IL2). Affected tissues are AWC (on/off, L/R), RMD (L/R), IL2 (L/R), and g1 (L/R).

Enriched cell lineages were assigned to cell types using tables available at `www.wormatlas.org/celllistsulston.htm`, derived from SULSTON et al., 1983.

3.5.3.4 Selecting maternally-contributed transcripts

To select genes with significant differences maternal contribution between conditions, we computed differences between pheromone and control model fits on [-20;10] min past 4C within the list of DE genes, and kept those with an average absolute fit difference $> 1 log(\mathrm{CPM}_{\mathrm{UMI}} + 1)$. The resulting genes, were tested for significant differences with a t.test on $log(\mathrm{CPM}_{\mathrm{UMI}} + 1)$ expression values restricted to samples staged under 10 min past 4C (7 CTR and 6 PHE), leaving 17 down and 18 up-regulated genes with p<0.05, shown in Fig. 3.9.

Discussion



To conclude, I have characterized the molecular changes in the developing *C. elegans* embryo caused by parental exposure to pheromone, revealing differences in gene expression of the developing nervous system, and particularly sensory organs, throughout embryogenesis. These results suggest that parental perception of the social environment may in turn alter how the progeny perceive and react to their own environment.

To reach this end, I have improved and developed methods from sample collection and library preparation to data analysis and integration that all have potential well beyond the study of intergenerational effects of pheromone and of *C. elegans*. Indeed, single embryo collection and study with FACS, profiling whole individuals with Smart-Seq3, and inferring age from the transcriptome provide accessible solutions to study single individuals at high-throughput. Indeed, despite growing interest and method development for such single-animal studies, notably in microfluidics (Frey et al., 2022; Bhattacharjee et al., 2016), requiring specialized equipment has slowed their spread to other labs (Bhattacharjee et al., 2016; Wan & Lu, 2020). The large portion of my work dedicated to experimental and computational method development is thus certainly relevant for future research in this area.

With the increase in throughput, better tools to analyze and interpret the resulting data will also likely emerge. In my last chapter, for example, I developed a custom approach to analyze and classify relevant differences of gene expression dynamics between conditions, which can likely be improved.

Despite its recent publication, RAPToR has already been employed in several studies to infer age in models and humans (Zhang et al., 2022; Bell et al., 2023; Kim et al., 2023; Sinigaglia et al., 2022; Hagan et al., 2022). Therefore, I believe that staging samples post-profiling (with RAPToR or otherwise) will become standard practice in gene expression studies, improving results and conclusions gained from profiling data. Looking back on already-published experiments and findings with RAPToR will also likely debunk false discoveries and lead to new discoveries using existing data. Nearly 10 years ago, Snoek et al., 2014 gave us a glimpse of developmental bias in one expression database. Since then, data collection has skyrocketed while little has been done to address the issue which in our experience is likely still widespread. Given the reliance of annotations on expression data and, in turn, of studies on such annotations (this work included), the importance of removing developmental bias from databases cannot be understated.

Going forward, we envision inferring age from gene expression will help us dissect the intricacies of developmental timing regulation and aging, and that analogous strategies to monitor disease progression will be of clinical significance. Using population-scale studies at single-individual resolution will also bring us closer to truly understanding why individuals are different, and perhaps at some point, to even stop biological material from doing '*whatever it damn well pleases*'.

References

BELL, AVERY DAVIS et al. (2023). “Beyond the reference: gene expression variation and transcriptional response to RNA interference in Caenorhabditis elegans”. In: *G3: Genes, Genomes, Genetics* 13.8, jkad112.

BHATTACHARJEE, NIRVEEK et al. (2016). “The upcoming 3D-printing revolution in microfluidics”. In: *Lab on a Chip* 16.10, pp. 1720–1742.

FREY, NOLAN et al. (2022). “Microfluidics for understanding model organisms”. In: *Nature communications* 13.1, p. 3195.

HAGAN, THOMAS et al. (2022). “Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses”. In: *Nature Immunology* 23.12, pp. 1788–1798.

KIM, EUNAH et al. (2023). “Mitochondrial aconitase suppresses immunity by modulating oxaloacetate and the mitochondrial unfolded protein response”. In: *Nature Communications* 14.1, p. 3716.

SINIGAGLIA, CHIARA et al. (2022). “Distinct gene expression dynamics in developing and regenerating crustacean limbs”. In: *Proceedings of the National Academy of Sciences* 119.27, e2119297119.

SNOEK, L BASTEN et al. (2014). “A rapid and massive gene expression shift marking adolescent transition in C. elegans”. In: *Scientific reports* 4.1, pp. 1–5.

WAN, JASON and HANG LU (2020). “Enabling high-throughput single-animal gene-expression studies with molecular and micro-scale technologies”. In: *Lab on a Chip* 20.24, pp. 4528–4538.

ZHANG, GAOTIAN et al. (2022). “The impact of species-wide gene expression variation on Caenorhabditis elegans complex traits”. In: *Nature communications* 13.1, p. 3462.

Appendices

Contents

A Supplementary information on RAPToR improvements

This appendix contains supplementary figures for Chapter 1, sections 1.2.2 and 1.2.3.

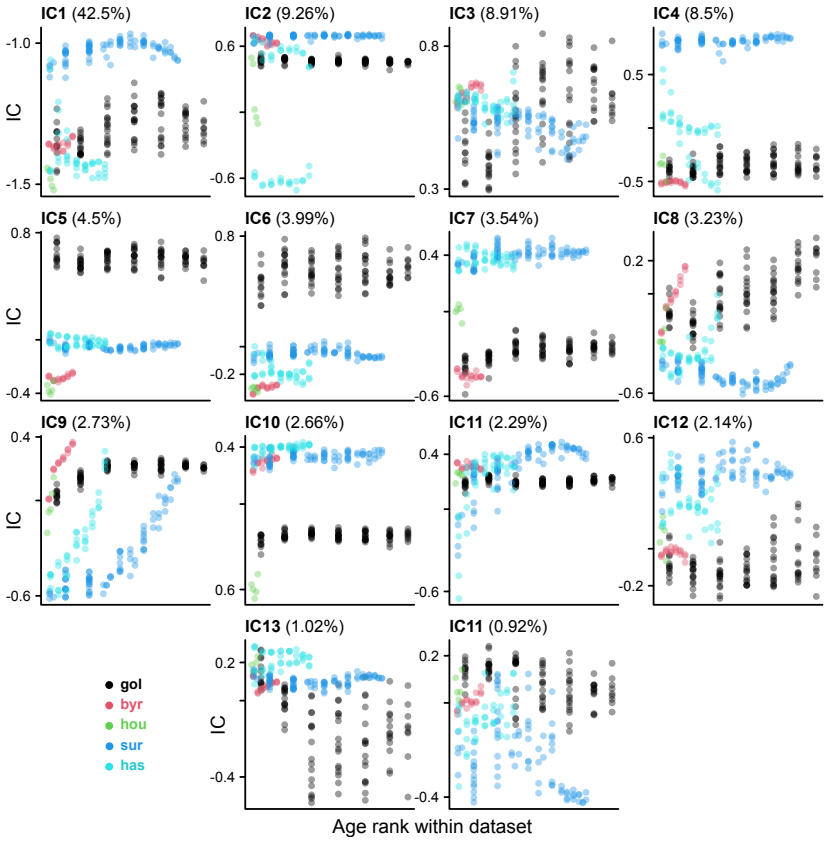


Figure A.1 – Joint ICA on gene expression of 5 aging time-series
Independent components (IC) from a joint Independent Component Analysis (ICA) of aging time-series (see Table 1.1). Percentage of total variance explained is indicated for each component.

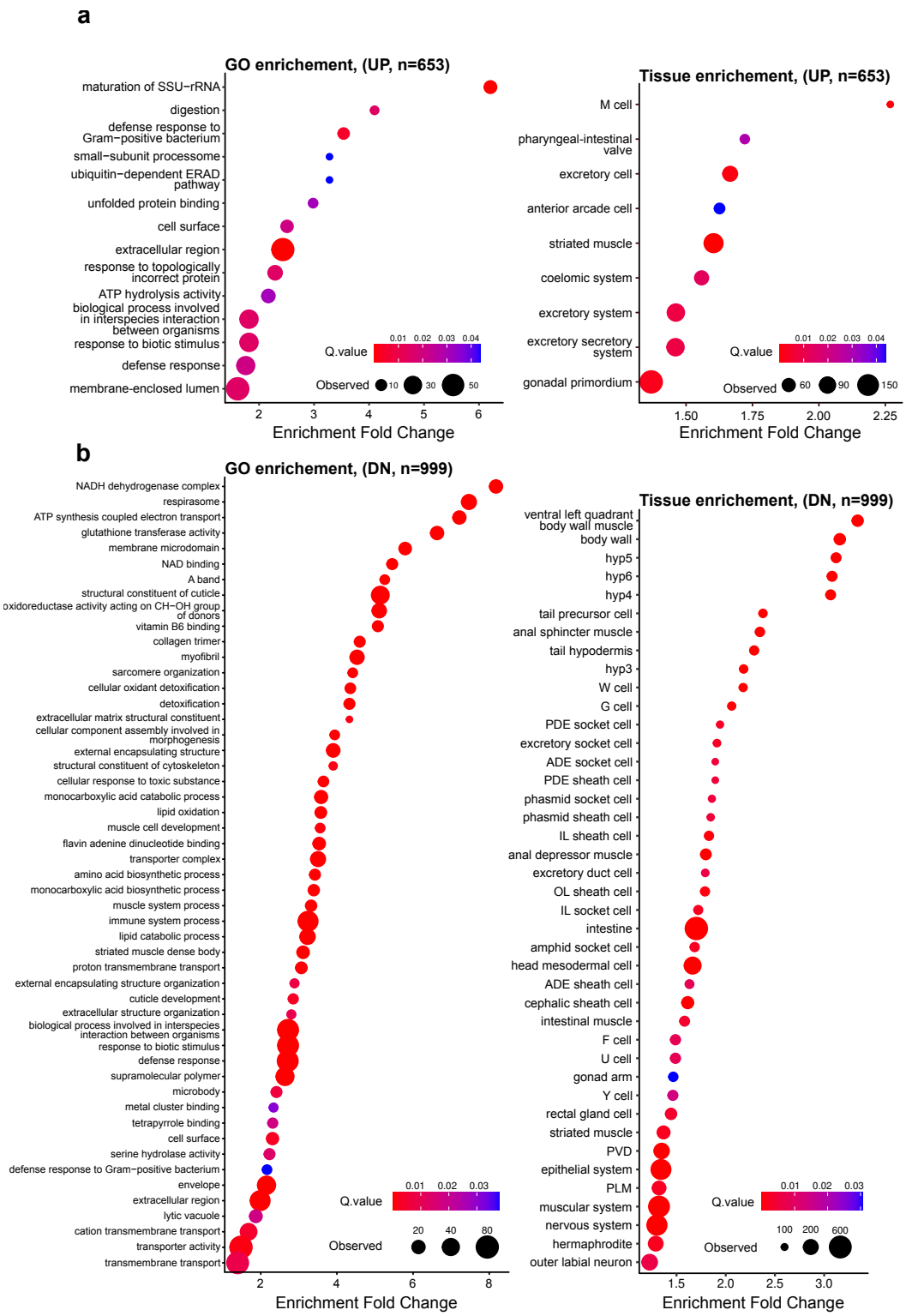


Figure A.2 – Enrichment of the informative aging gene set
a,b, Gene Ontology (GO, left) and tissue (right) enrichment of gene increasing (**a**) or decreasing (**b**) monotonously with age.

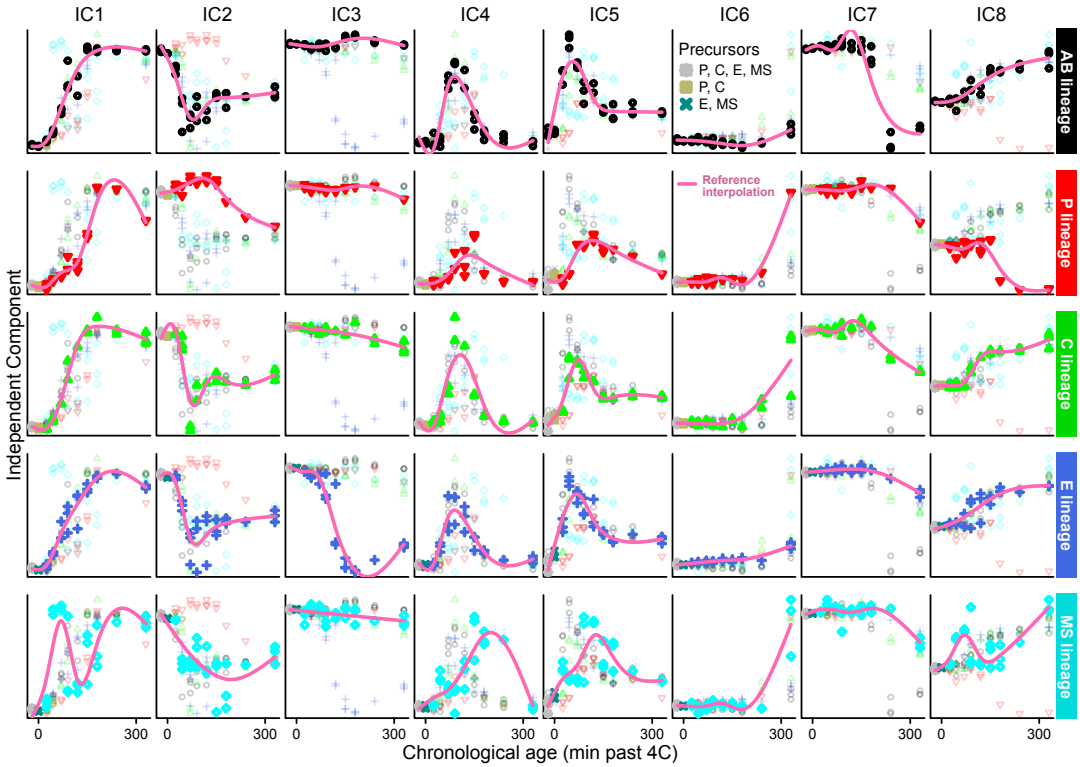


Figure A.3 – Reference interpolation of separate trajectories within a shared component space
First 8 Independent Components (ICs) of an ICA extracting 40 ICs for reference interpolation. Each cell lineage (including their precursors) is treated as a separate reference to interpolate within the component space.

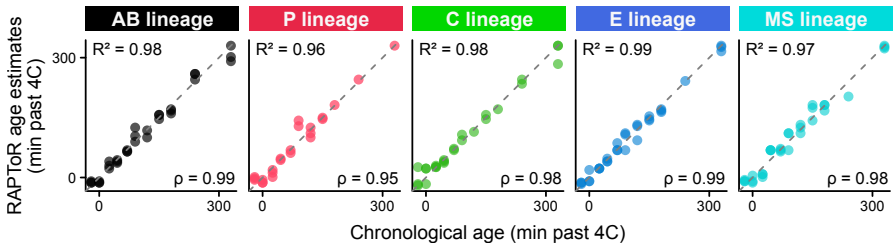


Figure A.4 – RAPToR age estimates of single cells on their respective lineage references
Squared pearson correlation (R^2) and Spearman's correlation (ρ) are given for each lineage.

B Supplementary information on FACS with *C. elegans* embryos

This appendix contains supplementary material for Chapter 2, section 2.2.

B.1 Embryos can be robustly selected across experiments and instruments

After bleaching the contents of a plate (see Chapter 2 Methods), we pass the resulting embryo suspension through an Attune^{TM} CytPix^{TM} cytometer, and reliably find a cluster of embryos, which are large events with high Side and Forward SCatter (FSC-H, SSC-H, Fig. B.1).

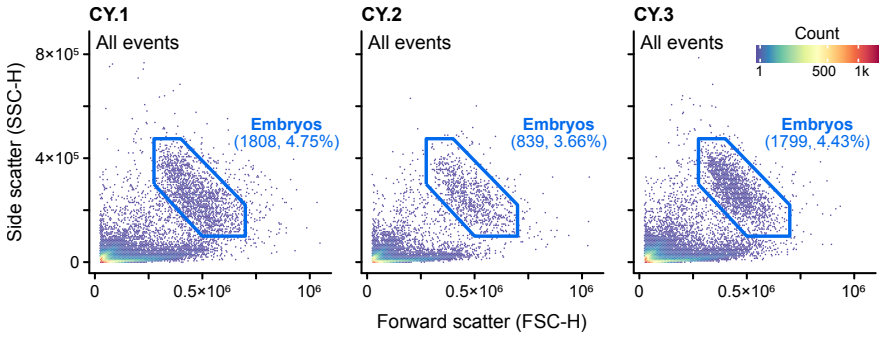


Figure B.1 – Gating emrbyos on forward and side scatter with a CytPix
CY.1 is the annotated sample shown in Fig. 2.2b, and CY.2 the independent sample used for staging validation in Fig. 2.5. Plots include events without images, and thus numbers differ from the figures cited above.

On a FACSAria IIμ, forward scatter measurements differed and resulted in less discernible clusters than with the CytPix (data not shown). Therefore, we relied on side scatter Width and Height to gate embryos (SSC-W, SSC-H, Fig. B.2), which is more consistent across both instruments and also adequately segregates embryos from debris (Fig. B.3)

Selecting live embryos from the dead or unfertilized (NF) ones is then identical in both instruments, using fluorescence channels (Fig. B.4, B.5).

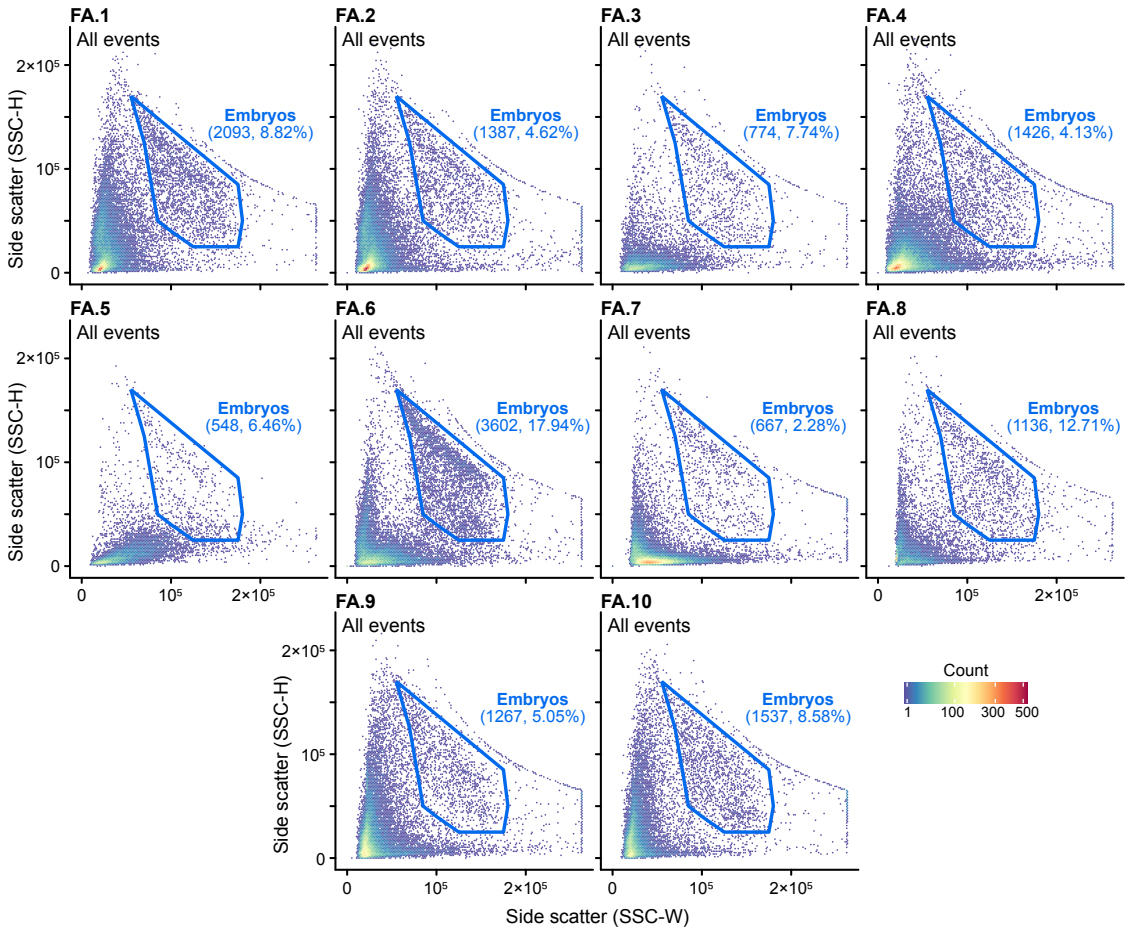


Figure B.2 – Gating embryos on side scatter width and height with a FACSAria
FA.1 is the sample shown in Fig. 2.2e-f.

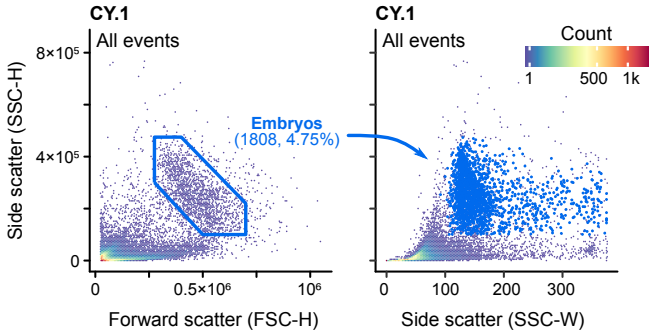


Figure B.3 – Side scatter width and height also separate embryos from debris

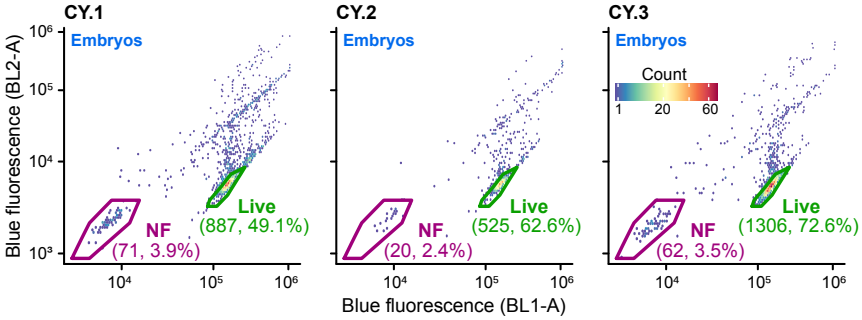


Figure B.4 – Gating live embryos from debris using autofluorescence with a CytPix

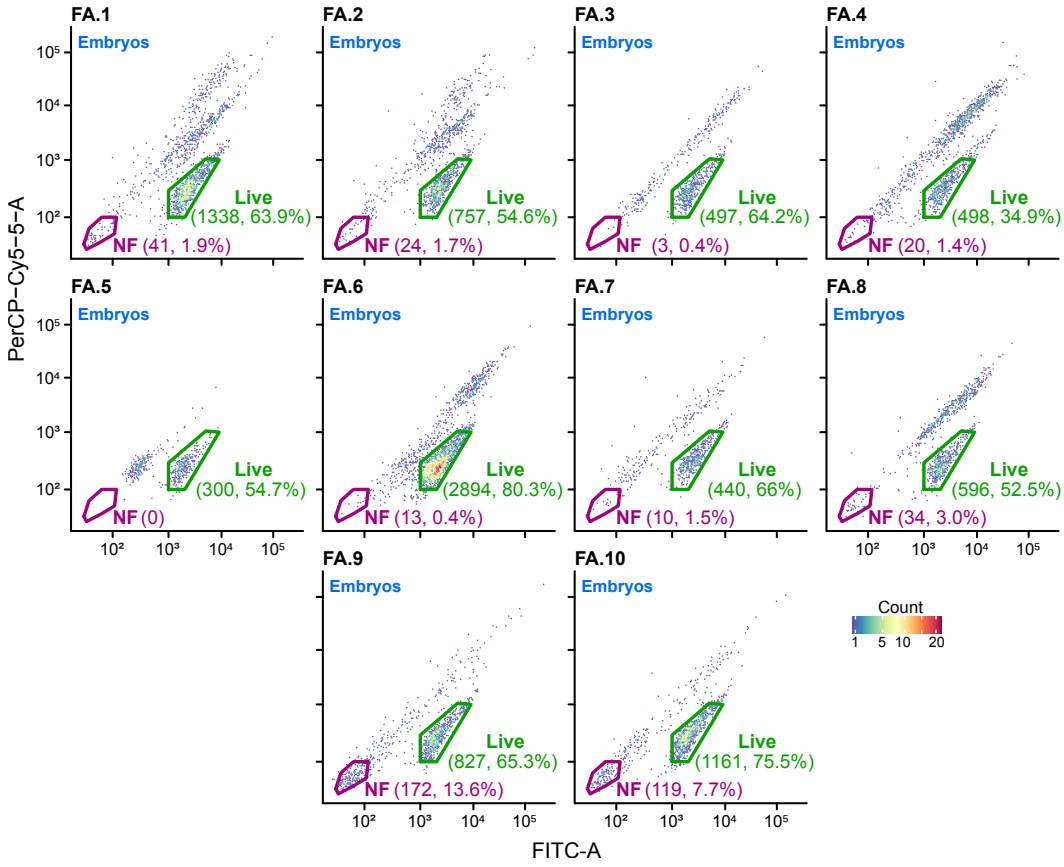


Figure B.5 – Gating live embryos from debris using autofluorescence with a FACSAria

B.2 Embryos survive sorting (better than bleach)

To ensure that sorting with the FACSAria doesn't damage embryos, we first sorted live embryos (FA.1, FA.2, Fig. B.5), and then passed the collected solution in the CytPix (Fig. B.6a). We note a significant reduction in cellular debris (and therefore higher embryo ratio), indicating that bleaching is the main cause of damage to the embryos rather than their journey through the cytometers.

The total disappearance of unfertilized embryos confirms that they are properly removed with the FACSAria by fluorescence gating (Fig. B.6b). However, remaining debris and dead embryos also suggest that either passing through the instruments, sorting, or time elapsed since the bleach causes embryos to die (if bleach washes are insufficient).

Finally, we ran part of a sample through the CytPix right after the bleach (CY.7, Fig. B.7a), and the remainder 30 minutes later (CY.8, Fig. B.7a). This confirmed that the fluorescence allowing us to differentiate live embryos from others does not change with time (Fig. B.7b). We also note that the ratio of live embryos is stable between the runs, thus ruling out the possibility of embryos dying post-bleach.

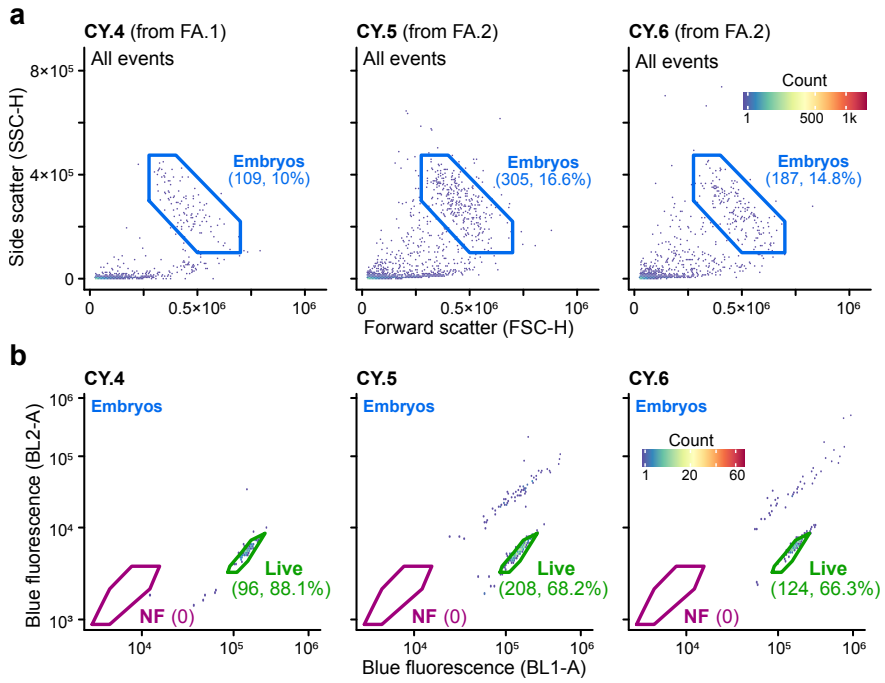


Figure B.6 – Live embryos can be sorted multiple times
a, Gating embryos from debris, as in Fig. B.1.
b, Gating live from dead embryos, as in Fig. B.4.
Live embryos (FA.1, FA.2, Fig. B.5) were collected in 500 µL of M9, and passed to the CytPix.

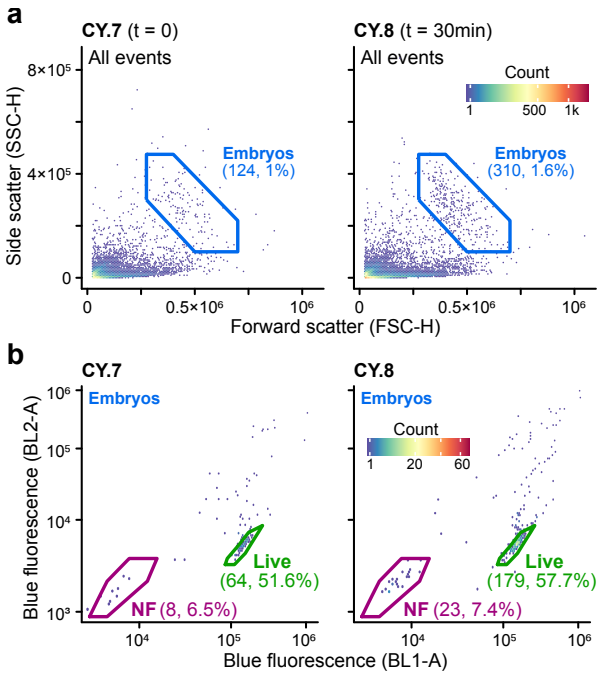


Figure B.7 – Embryo fluorescence and survival in the suspension are stable in time
a, Gating embryos from debris, as in Fig. B.1.
b, Gating live from dead embryos, as in Fig. B.4.
The same embryo suspension was run in part just after bleaching, and the remainder 30 min later.

C Single-embryo Smart-seq3 detailed protocol

This is the detailed RNA-seq library preparation protocol for single embryos of C. elegans I adapted from the Smart-seq3 protocol (HAGEMANN-JENSEN et al., 2020) (accessible in its latest version on protocols.io, at dx.doi.org/10.17504/protocols.io.bcq4ivyw), also taking into account elements from the previous version Smart-seq2 (PICELLI et al., 2014), and its adaptation for single-worm RNA-seq (SERRA et al., 2018).

C.1 Materials

The tables below list the necessary reagents (Table C.1), oligos (Table C.2), and barcodes (Table C.3) for the protocol. Reagents for quantification and quality control of cDNA libraries assume the use of Qubit and Tapestation instruments respectively.

Name	Vendor	Catalog ref.	Steps/mixes requiring reagent
Proteinase K	Sigma Aldrich	P2308-25MG	Lysis Buffer
Triton X-100	Sigma Aldrich	T8787-50ML	Lysis Buffer
dNTP Set (100mM)	Thermo Fisher	R0182	Lysis Buffer, Preamp. PCR, Tagmentation
Poly Ethylene Glycol (PEG) 8000	Sigma Aldrich	89510-250G-F	Lysis Buffer, Purification
Recombinant RNAse Inhibitor	Takarabio	2313A	Lysis Buffer, RT
UltraPure^{TM} DNase/RNase-Free Distilled Water	ThermoFisher	10977035	Most
KAPA HiFi Hotstart PCR kit	Roche	KK2502	Preamp. PCR
Sera-Mag Speed Beads	Ge Healthcare	65152105050250	Purification
Sodium Azide	Sigma Aldrich	S2002-100G	Purification
IGEPAL CA-630	Sigma Aldrich	I8896	Purification
EDTA (0.5M, pH 8.0, RNase-free)	ThermoFisher	AM9260G	Purification, QC
Qubit HS dsDNA assay	ThermoFisher	Q32854	QC
Agilent High Sensitivity D5000 ScreenTape	Agilent	5067-5592	QC
Agilent HSD5000 Reagents, Ladder	Agilent	5067-5594, 5067-5593	QC
GTP (Tris-buffered solution 100mM)	Thermo Scientific	R1461	RT
Dithiothreitol (DTT)	ThermoFisher	707265ML	RT
Maxima H Minus Reverse Transcriptase (200U/µL)	ThermoFisher	EP0751	RT
Magnesium Chloride (1M)	Invitrogen	AM9530G	RT, Preamp. PCR, Tagmentation
Sodium Chloride (5M)	Invitrogen	AM9760G	RT, Purification
Trizma-base	Sigma Aldrich	T6791-100G	RT, Purification
Nextera XT DNA Library Preparation Kit	illumina	FC-131-1096	Tagmentation
SDS (10% Solution, RNase-free)	ThermoFisher	AM9822	Tagmentation
Phusion High-Fidelity DNA Polymerase (2U/µL)[1]	Thermo Scientific	F530L	Tagmentation
NN-Dimethylformamide	Sigma Aldrich	D4551	Tagmentation
Tween 20	Sigma Aldrich	P9416	Tagmentation

[1]: Phusion Kit also contains DMSO

Table C.1 – Reagents used for single-embryo Smart-seq3.

Oligo	Purification	Sequence
Smartseq3_ OligodT30VN	HPLC	/5Biosg/ACGAGCATCAGCAGCATACGATTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Smartseq3_N8 _TSO	RNase-Free HPLC	/5Biosg/AGAGACAGATTGCGCAATGNNNNNNNNrGrGrG
Fwd_PCR_primer	HPLC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCGCAA*T*G
Rev_PCR_primer	HPLC	ACGAGCATCAGCAGCATAC*G*A
Nextera^{TM} compatible primers with barcodes	Standard desalting	see Table C.3 for sequences and 96 unique i5-i7 ID pairs

Table C.2 – Oligos used in Smart-seq3
All oligos were purchased from IDT.

Oligo	Sequence
Nextera^{TM} s5 primer	AATGATACGGCGACCACCGAGATCTACAC[8-bp i5 index]TCGTCGGCAGCGTC
Nextera^{TM} s7 primer	CAAGCAGAAGACGGCATACGAGAT[8-bp i7 index]GTCTCGTGGGCTCGG

Barcode ID/well	i5 index	i7 index
01 A01	ATATGCGC	CTGATCGT
02 A02	TGGTACAG	ACTCTCGA
03 A03	AACCGTTC	TGAGCTAG
04 A04	TAACCGGT	GAGACGAT
05 A05	GAACATCG	CTTGTCGA
06 A06	CCTTGTAG	TTCCAAGG
07 A07	TCAGGCTT	CGCATGAT
08 A08	GTTCTCGT	ACGGAACA
09 A09	AGAACGAG	CGGCTAAT
10 A10	TGCTTCCA	ATCGATCG
11 A11	CTTCGACT	GCAAGATC
12 A12	CACCTGTT	GCTATCCT
13 B01	ATCACACG	TACGCTAC
14 B02	CCGTAAGA	TGGACTCT
15 B03	TACGCCTT	AGAGTAGC
16 B04	CGACGTTA	ATCCAGAG
17 B05	ATGCACGA	GACGATCT
18 B06	CCTGATTG	AACTGAGC
19 B07	GTAGGAGT	CTTAGGAC
20 B08	ACTAGGAG	GTGCCATA
21 B09	CACTAGCT	GAATCCGA
22 B10	ACGACTTG	TCGCTGTT
23 B11	CGTGTGTA	TTCGTTGG
24 B12	GTTGACCT	AAGCACTG
25 C01	ACTCCATC	CCTTGATC
26 C02	CAATGTGG	GTCGAAGA
27 C03	TTGCAGAC	ACCACGAT
28 C04	CAGTCCAA	GATTACCG
29 C05	ACGTTCAG	GCACAACT
30 C06	AACGTCTG	GCGTCATT
31 C07	TATCGGTC	ATCCGGTA
32 C08	CGCTCTAT	CGTTGCAA
33 C09	GATTGCTC	GTGAAGTG
34 C10	GATGTGTG	CATGGCTA
35 C11	CGCAATCT	ATGCCTGT
36 C12	TGGTAGCT	CAACACCT
37 D01	GATAGGCT	TGTGACTG
38 D02	AGTGGATC	GTCATCGA
39 D03	TTGGACGT	AGCACTTC
40 D04	ATGACGTC	GAAGGAAG
41 D05	GAAGTTGG	GTTGTTCG
42 D06	CATACCAC	CGGTTGTT
43 D07	CTGTTGAC	ACTGAGGT
44 D08	TGGCATGT	TGAAGACG
45 D09	ATCGCCAT	GTTACGCA
46 D10	TTGCGAAG	AGCGTGTT
47 D11	AGTTCGTC	GATCGAGT
48 D12	GAGCAGTA	ACAGCTCA

Barcode ID/well	i5 index	i7 index
49 E01	ACAGCTCA	GAGCAGTA
50 E02	GATCGAGT	AGTTCGTC
51 E03	AGCGTGTT	TTGCGAAG
52 E04	GTTACGCA	ATCGCCAT
53 E05	TGAAGACG	TGGCATGT
54 E06	ACTGAGGT	CTGTTGAC
55 E07	CGGTTGTT	CATACCAC
56 E08	GTTGTTCG	GAAGTTGG
57 E09	GAAGGAAG	ATGACGTC
58 E10	AGCACTTC	TTGGACGT
59 E11	GTCATCGA	AGTGGATC
60 E12	TGTGACTG	GATAGGCT
61 F01	CAACACCT	TGGTAGCT
62 F02	ATGCCTGT	CGCAATCT
63 F03	CATGGCTA	GATGTGTG
64 F04	GTGAAGTG	GATTGCTC
65 F05	CGTTGCAA	CGCTCTAT
66 F06	ATCCGGTA	TATCGGTC
67 F07	GCGTCATT	AACGTCTG
68 F08	GCACAACT	ACGTTCAG
69 F09	GATTACCG	CAGTCCAA
70 F10	ACCACGAT	TTGCAGAC
71 F11	GTCGAAGA	CAATGTGG
72 F12	CCTTGATC	ACTCCATC
73 G01	AAGCACTG	GTTGACCT
74 G02	TTCGTTGG	CGTGTGTA
75 G03	TCGCTGTT	ACGACTTG
76 G04	GAATCCGA	CACTAGCT
77 G05	GTGCCATA	ACTAGGAG
78 G06	CTTAGGAC	GTAGGAGT
79 G07	AACTGAGC	CCTGATTG
80 G08	GACGATCT	ATGCACGA
81 G09	ATCCAGAG	CGACGTTA
82 G10	AGAGTAGC	TACGCCTT
83 G11	TGGACTCT	CCGTAAGA
84 G12	TACGCTAC	ATCACACG
85 H01	GCTATCCT	CACCTGTT
86 H02	GCAAGATC	CTTCGACT
87 H03	ATCGATCG	TGCTTCCA
88 H04	CGGCTAAT	AGAACGAG
89 H05	ACGGAACA	GTTCTCGT
90 H06	CGCATGAT	TCAGGCTT
91 H07	TTCCAAGG	CCTTGTAG
92 H08	CTTGTCGA	GAACATCG
93 H09	GAGACGAT	TAACCGGT
94 H10	TGAGCTAG	AACCGTTC
95 H11	ACTCTCGA	TGGTACAG
96 H12	CTGATCGT	ATATGCGC

Table C.3 – Nextera^{TM} compatible primers and Unique Dual Index (UDI) barcode sequences
IDT references for the barcode oligos are IDT8_UDI_1-96. i5 index is given as forward sequence.

C.2 Protocol

C.2.1 General guidelines

- Prepare workbench and tools by cleaning with 100% ethanol and RNAseZAP, RNAse-Away, DNA-OFF or similar to remove RNAse (5% SDS in a spray bottle also works fine).
- Work with gloves and change them frequently.
- Work quickly, and on ice or refrigerated plate holders.
- Prepare master mixes right before use. Some components can be mixed in advance to save time, but oligos and enzymes should be added at the last minute.
- Count at least 10% extra for mixes (with 1µL minimum margin).
- Do not pipet up and down to avoid loss of material in low-volume steps (< 10 µL).
- Keep tubes/strips sealed as mush as possible to avoid evaporation and contamination.

C.2.2 Lysis buffer and oligodT+dNTP mix

1. Prepare the following lysis buffer solution:

Reagent	**Reaction concentration**	**Volume** (µL) **for**	
		1 reaction	**96 samples**
Poly-ethylene Glycol 8000 (50% solution)	5%	0.40	44.00
Triton X-100 (10% solution)	0.1%	0.03	3.30
RNAse Inhibitor (40 U/µL)	1 U/µL	0.08	8.80
Protease-K (20 mg/mL)	1 μg/µL	0.13	14.08
Nuclease free water		1.86	204.82
Total (µL)		**2.50**	**275.00**

   Note that the 50% PEG solution should be prepared beforehand, ensuring that PEG is fully mixed into solution.

2. Add 2.5 µL of lysis buffer in each tube of a PCR strip (or well of a plate) and do a quick centrifugation to collect the buffer at the bottom of the wells.

   Lysis buffer can be stored (at -20°C or -80°C, for at least a month according to the authors of Smart-seq3).

3. Prepare the following oligodT+dNTP mix:

Reagent	**Reaction concentration**	**Volume** (µL) **for**	
		1 reaction	**96 samples**
OligodT30VN (100 µM)	0.5 µM	0.02	2.20
dNTPs (25 mM each)	0.5 mM each	0.08	8.80
Nuclease free water		0.40	44.00
Total (µL)		**0.50**	**55.00**

   This mix can be stored *ad libitum* at -20°C, as per the contents.

C.2.3 Sample collection

1. Prepare samples. Assuming manual collection of *C. elegans* single embryos, wash and bleach the contents of a plate with egg-laying adults, and place bleach output in an empty petri dish under a binocular. 0.05% Triton can be added to the bleach output to prevent embryos from stricking to the dish.
2. Place and keep the prepared strip(s)/plates of lysis buffer on ice, or appropriate cooling block. Be careful of lysis buffer freezing, which will damage the samples.
3. Pipet single embryos (with 0.400 µLor less) into each well/tube, placing the pipet tip at the very bottom of the well at a slight angle to release the embryo within the lysis buffer.

   Keep strips/tubes sealed as much as possible to avoid evaporation/contamination.
4. After a strip (8 samples) has been collected, centrifuge and run lysis (see below) in a hot-started thermocycler before storing at -20°C (if needed). **Freezing samples before lysis will damage them.**

   If collecting multiple strips, launch lysis after each one and collect the next strip in the meantime. After lysis of each strip, centrifuge and store samples at -20°C. Collecting a strip takes roughly as much time as lysis, so many strips can be continuously collected, lysed, and stored by using a 2-block thermocycler. We see no signs of RNA degradation after storing samples post-lysis for 1.5 h.

C.2.4 Lysis

1. Incubate collected samples for proteinase-K lysis in a hot-started thermocycler with the following program:

Temperature	Time
65°C	10 min
85°C	1 min
4°C	Hold

2. If collecting multiple strips, centrifuge and store samples at -20°C as soon as lysis is done.

C.2.5 Reverse-Transcription

1. Add 0.5 µL of the oligodT+dNTP mix in each tube, and incubate at 72°C for 5 minutes, followed by a 4°C hold.
2. During incubation, prepare the following RT mix:

Reagent	**Reaction concentration**	**Volume (µL) for**	
		1 reaction	**96 samples**
Tris-HCl pH 8.3 (1 M)	25 mM	0.10	11.00
NaCl (1 M)	30 mM	0.12	13.20
MgCl2 (100 mM)	2.5 mM	0.10	11.0
GTP (100 mM)	1 mM	0.04	4.40
DTT (100 mM)	8 mM	0.32	35.20
RNase Inhibitor (40 U/µL)	0.5 U/µL	0.05	5.50
TSO (100 µM)	2 µM	0.08	8.80
Maxima H-minus RT enzyme (200 U/µL)	2 U/µL	0.04	4.40
Nuclease free water	–	0.15	16.50
Total (µL)		**1.00**	**110.00**

3. Add 1 µL of RT mix to each tube.
4. Do a quick centrifugation to collect reaction at the bottom, before placing samples in a thermocycler with the following program:

Temperature	Time	Repeats
42°C	90 min	1x
50°C	2 min	14x
42°C	2 min	
85°C	5 min	1x
4°C	Hold	–

C.2.6 Preamplification PCR

1. When the RT incubation is nearing completion, prepare the following PCR mix. **Only add polymerase just before using the master-mix**, as the Kapa DNA polymerase has a 3-5' exonuclease activity that is not HotStart.

Reagent	Reaction concentration	Volume (µL) for	
		1 reaction	96 samples
Kapa HiFi HotStart buffer (5X)	1X	2.00	220.00
dNTPs (25 mM each)	0.3 mM each	0.12	13.20
MgCl2 (100 mM)	0.5 mM	0.05	5.50
Fwd Primer (100 µM)	0.5 µM	0.05	5.50
Rev Primer (100 µM)	0.1 µM	0.01	1.10
Kapa Polymerase (1 U/µL)	0.02 U/µL	0.20	22.00
Nuclease free water	–	3.57	392.70
Total (µL)		**6.00**	**660.00**

2. Add 6 µL of PCR mix to each tube.
3. Do a quick centrifugation to collect reaction at the bottom, before placing samples in a thermocycler with the following PCR program:

Temperature	Time	Repeats
98°C	3 min	1x
98°C	20 sec	20x
65°C	30 sec	
72°C	4 min	
72°C	5 min	1x
4°C	Hold	–

C.2.7 cDNA purification

C.2.7.1 Prepare 22% PEG Clean-up Beads solution These beads perform similar to Ampure XP beads, and are prepared as per mcSCRB-seq protocol[1].

1. Prepare 10mM Tris-HCl, pH 8.0, 1 mM EDTA buffer (TE buffer).

[1]: see https://dx.doi.org/10.17504/protocols.io.p9kdr4w

2. Prepare the following PEG buffer in a 50mL falcon tube, but don't add all the water until PEG is fully solubilized. To help solubilize PEG, incubate at 40°C and vortex regularly.

Reagent	**Amount/Volume**
Poly-ethylene Glycol 8000	11 g
NaCl (5M)	10 mL
Tris-HCl pH 8.0 (1M)	500 µL
EDTA, 0.5M	100 µL
IGEPAL, 10% solution	50 µL
Sodium Azide, 10% solution	250 µL
UltraPure Water	up to 49 mL
Total	**49 mL**

   Buffers can be stored at 4°C.

3. To prepare 1mL of 22% PEG Clean-up Beads, resuspend Sera-Mag Speed Beads stock carefully
4. Pipet 20 µL of bead stock into a 1.5 mL eppendorf tube.
5. Place tube on magnet stand and wait a few minutes for beads to pellet.
6. Remove supernatant.
7. Add 20 µL of TE buffer, and resuspend beads off magnet by pipetting up and down.
8. Place back on magnet stand.
9. Remove supernatant and repeat wash one more time.
10. Add 18 µL of TE buffer, and resuspend beads off magnet.
11. Add 980 µL of PEG solution above and mix well.

C.2.7.2 Purify cDNA

1. Add 0.8 : 1 ratio of 22% PEG beads to sample (samples shoulb be 10 µLby this point, so 8 µLbeads), and mix by gently pipetting up and down 10x.
2. Incubate at room temperature for 8 minutes.
3. Place on magnet and allow beads to settle, roughly 5 minutes.
4. Remove supernatant[2].
5. Wash beads once with 100 µLof freshly prepared 85% ethanol, incubate 30s.
6. Remove ethanol and repeat ethanol wash and incubation.
7. Centrifuge 10 seconds at low speed to collect all liquids on the tube walls.
8. Place back on magnet, wait 30 seconds for beads to collect on the side, and remove leftover ethanol.
9. Let the beads air dry 3-5 minutes (don't wait too long, ideally just before cracks appear).
10. Elute beads in 12 µLof UltraPure Water, resuspend beads off magnet, mixing thoroughly by pipetting up and down at least 20x and scraping tube with the pipet tip.

[2]: Contrary to remarks from the Smart-seq2 adaptation to single-worm (SERRA et al., 2018), the supernatant **can not** be re-incubated with the beads if the cDNA yield is insufficient in Smart-seq3. In our hands, this only yields primers and primer dimers. This could be due to the smaller volumes used here compared to Smart-seq2.

11. Incubate for 8 minutes.
12. Place on magnet for 2-3 minutes until beads collect and solution is clear.
13. Transfer 10 µLof supernatant to fresh PCR strips (free from nuclease), being careful to minimize bead carryover.

C.2.8 cDNA quality control and normalization

1. Follow manufacturer instructions of Qubit dsDNA HS kit or similar to measure cDNA concentration of all samples.

   Samples with a cDNA concentration under 1 ng/µLshould be considered as failed, those with 1-5 ng/µLare passable, and those over 10 ng/µLhave good yield. Samples with higher yields should be prioritized, as they have a more diverse RNA molecule pool and thus larger UMI-corrected library sizes; high yield samples also tend to have higher complexity.
2. Dilute all samples to 0.5 ng/µLin fresh PCR strip. Prepare at least 4 µLof each to have enough for quality control and tagmentation.
3. Follow manufacturer instructions of Tapestation High-Sensitivity D5000 ScreenTape Assay or similar to assess sample quality.

   Figure C.1 gives examples of good and bad profiles. Proceeding with a few subpar samples is fine, but having a majority of them should be cause for concern.

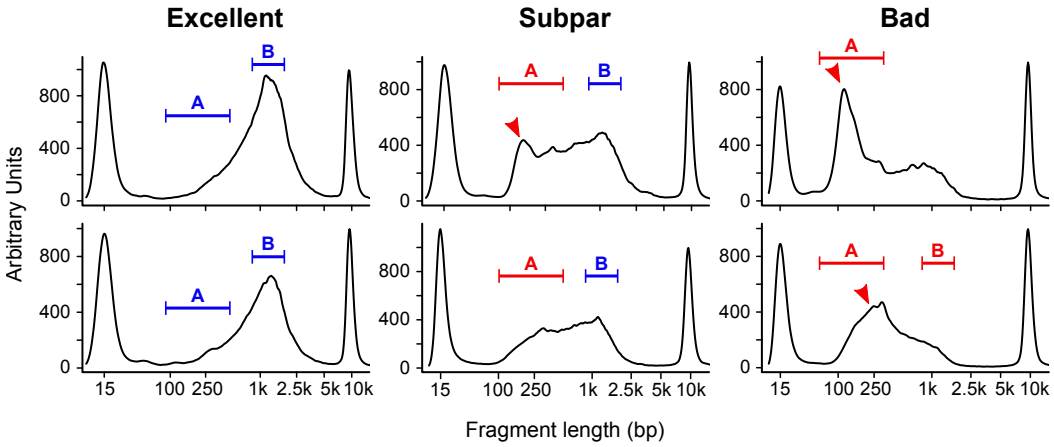


Figure C.1 – Example profiles of amplified cDNA of varying quality
The ideal sample profile should have no peaks in region A, and a strong main peak in region B (1 – 1.5 kb). Peaks in region A correspond to small fragments of cDNA and indicate a sample with degraded RNA. Arrows indicate a peak of amplified primer dimers, usually around 120 bp, which doesn't fail the sample unless it accounts for a large fraction of cDNA (e.g. top-right profile).
Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

C.2.9 Tagmentation

1. Prepare the following 4X tagmentation buffer. **Dimethylformamide (DMF) should be handled in a fume hood** and according to local safety regulations.

Reagent	**Concentration (4X)**	**Volume (µL) for**	
		1 reaction	**96 samples**
Tris-HCl pH 7.5 (1 M)	40 mM	0.02	2.20
MgCl2 (100 mM)	20 mM	0.10	11.00
Dimethylformamide (DMF)	20%	0.10	11.00
Nuclease free water	–	0.28	30.80
Total (µL)		**0.50**	**55.00**

Aliquots of 4X tagmentation buffer can be stored at 4°C for later use. The TD buffer (2x) from Nextera Kits can also be used, however with the current small amount of ATM (Tn5 tagmentase) used, the Illumina TD buffer will at some point run out.

2. Ensure Nextera™ compatible primer (barcodes) are at 0.5 µM.

   Orders from IDT (Table C.3) are packaged at 10 µM, we recommend diluting a full plate by adding 8 µL of primers to 152 µL of UltraPure water. Store at -20°C.

3. Add 1 µL of normalized cDNA (500 pg/µL) to a new PCR strip/plate.

4. Prepare the following tagmentation mix.

Reagent	**Reaction concentration**	**Volume (µL) for**	
		1 reaction	**96 samples**
Tagmentation Buffer (4X)	1X	0.50	55.00
Amplicon Tagmentation Mix (ATM, Tn5)		0.08	8.80
UltraPure water	–	0.42	46.20
Total (µL)		**1.00**	**110.00**

   The given ATM amount, with 500 pg/µLcDNA input yields around 80% UMI-reads with a Novaseq6000. Some optimization might be necessary to reach a desired UMI-read to Internal-read ratio depending on the sample type and sequencer. (See Figure 1c and Extended data Fig. 3b of HAGEMANN-JENSEN et al., 2020 for effects of cDNA concentration, ATM amount, and sequencer bias on UMI-read to internal-read ratio.)

5. Dispense 1 µL of Tagmentation mix in each tube.

6. Do a quick centrifugation before incubation in a hot-started thermocycler at 55°C for 10 minutes.

7. Add 0.5 µL of 0.2% SDS to each well to strip off Tn5 from the DNA. Ensure that SDS is not too concentrated, as it will inhibit the PCR reaction.

8. Quick centrifugation before incubation for 5 minutes at Room temperature.

9. Add 1.5 µL of (different) primers (0.5 µM) to each well.

C.2.10 Tagmentation PCR

1. Prepare the folowing PCR mix. We recommend making the mix without enzyme during the incubation steps of the tagmentation, and adding enzyme at the last minute to avoid leaving the samples too long at room temperature.

Reagent	**Concentration (4X)**	**Volume (µL) for**	
		1 reaction	**96 samples**
Phusion HF buffer (5X)	1X	2.80	308.00
dNTPs (25 mM each)	0.2 mM each	0.14	15.40
DMSO (50%)	2.50%	0.50	55.00
Tween 20 (1%)	0.01%	0.10	11.00
Phusion HF (2 U/µL)	0.02 U/µL	0.14	15.40
Nuclease free water	–	6.32	695.20
Total (µL)		**10.00**	**1100.00**

2. Dispense 10 uL of PCR mix to each tube. If reusing pipet tips for multiple wells, flip the strip/plate around to ensure that the tip does not come in contact with the area where primers were deposited and avoid cross-well contamination of barcodes.
3. Do a quick centrifugation and place samples in a hot-started thermocycler with the following program:

Temperature	Time	Repeats
72°C	3 min	1x
98°C	5 min	1x
98°C	20 sec	15x
55°C	30 sec	
72°C	30 sec	
72°C	5 min	1x
4°C	Hold	–

C.2.11 Library clean-up

1. Pool the samples in a large enough tube.
2. Add 0.6 : 1 ratio of 22% PEG beads to the final volume of pooled tagmented cDNA.
3. Mix by gently pipeting 10x and incubate 8 minutes at room temperature.
4. Place on magnet and allow beads to settle 5 minutes.
5. Discard supernatant and wash twice with >= 1000µLof freshly prepared 85% ethanol (ethanol wash amount should exceed the initial volume of the pool).
6. Remove ethanol and let the beads air dry for a few minutes. As for the previous purification (see C.2.7), a light centrifugation step can be done to ensure no ethanol remains.
7. Elute cDNA in a volume of UltraPure Water of at most $\frac{1}{5}$ of the pool input volume to ensure library concentration requirements.
8. Mix well by pipetting to resuspend beads, and incubate 10 minutes.
9. Place on magnet for 2-3 minutes unitl beads collect and solution is clear.
10. Transfer [Elution volume −2 uL] of the supernatant to a fresh tube, being careful to minimize bead carryover.

C.2.12 Final library quality control

1. Follow manufacturer instructions of Qubit dsDNA kit or similar to measure library cDNA concentration.
2. Dilute library to 2 ng/µL if necessary.
3. Follow manufacturer instructions of Tapestation High-Sensitivity D5000 ScreenTape Assay or similar to assess final library quality.

   Figure C.2 gives examples of good and bad profiles.

   Filtering out adapters can be attempted with a second purification, although this will also remove large amounts of material.

   Although adapter dimers and chimeric fragments are undesirable, they can be filtered out during data processing. Adapter dimer reads will fail length and complexity QC, and chimeric reads will be filtered out during mapping. Contaminated libraries can therefore still be sequenced if the loss of reads is acceptable.

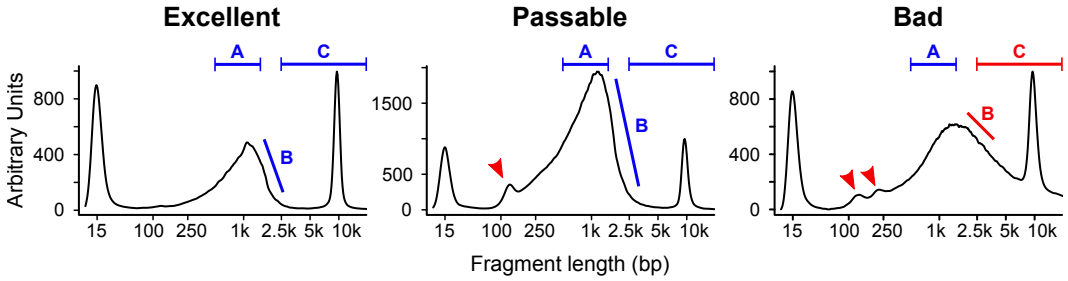


Figure C.2 – Example profiles of tagmented cDNA libraries
The ideal library profile should have no adapter dimer peak (arrows), a main peak in region A (600 – 1 kb) followed by a sharp decrease in region B and no fragments in region C.
Fragments should be should be overall shorter than before tagmentation. Thus, fragments in region C (> 2.5 kb) are likely to be chimeric and caused by PCR over-amplification. Both chimeric fragments and adapter dimers (arrows around 120 bp) and multimers (*e.g.* right profile, at 240 bp) should also be avoided as they will compete with the library for sequencing, lowering the yield.
Electropherogram data was acquired on a tapestation using HS-D5000 tapes and reagents.

C.2.13 Sequencing and data processing

1. Sequencing should be done on any Illumna-compatible sequencer, with parameters appropriate for the need (length, single-end/paired-end). Note that paired-end (PE) sequencing is required to computationally reconstruct RNA molecules and thus enable isoform assignment. The Smart-seq3 authors also remark that NovaSeq/HiSeq sequencers are more tolerant than NextSeq towards larger fragment distributions.

   We successfully seqeunced single-embryo libraries with 150PE on a NovaSeq6000.

2. Raw fastq files can then be processed using zUMIs (Parekh et al., 2018) (`https://github.com/sdparekh/zUMIs`).

D Supplementary information on the effects of parental pheromone exposure

This appendix contains supplementary material for Chapter 3.

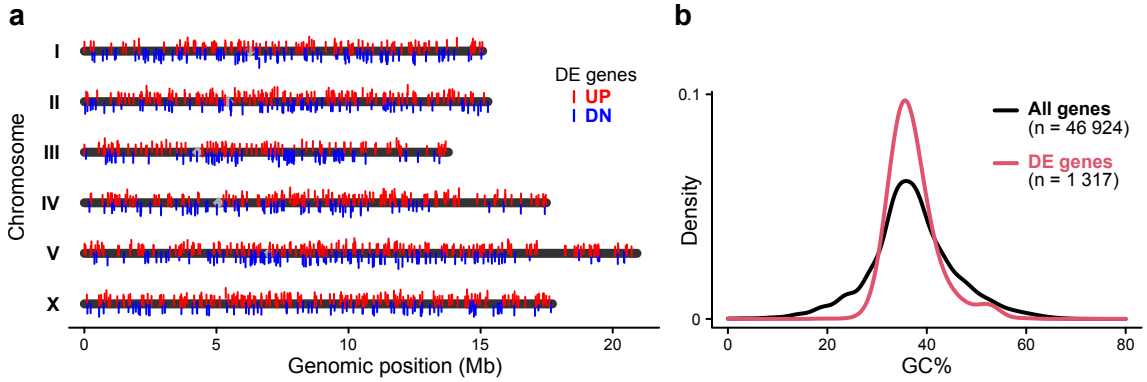


Figure D.1 – No genomic position or GC% bias of differentially expressed genes
a, Genomic position of DE genes. Color and distance from the chromosome bar correspond to the sign and size of maximum difference of PHE and CTR model fits (PHE – CTR) respectively.
b, GC% distribution for all *C. elegans* genes, and DE genes. Smoothed density from the $density()$ R function, with bandwidth parameter set to 2.
Genomic position and GC% from the Parasite biomart, for *C. elegans* PRJNA13758 (v. WS285).

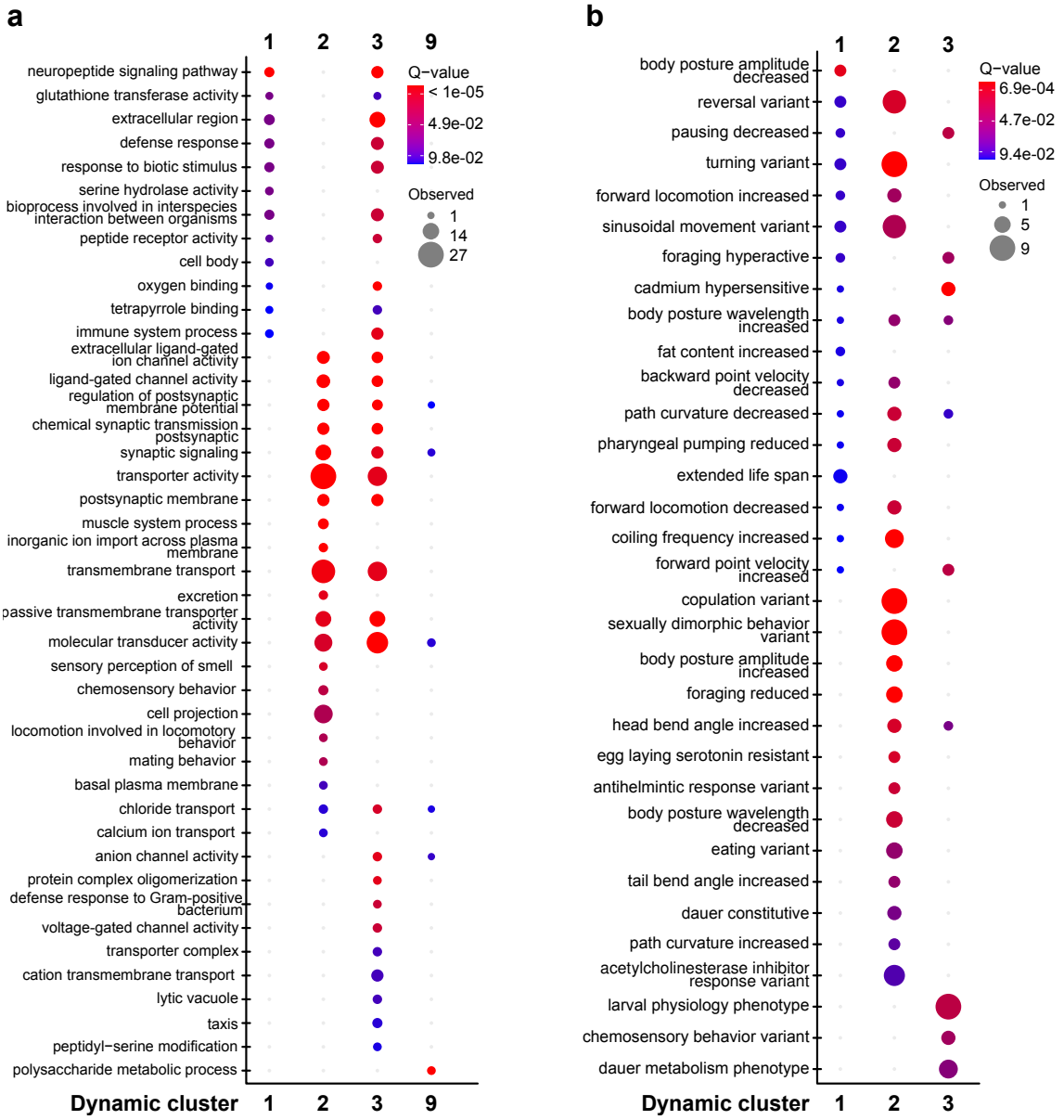


Figure D.2 – GO and phenotype enrichment of late embryo development clusters
a,b Gene ontology (**a**) and phenotype (**b**) enrichment of dynamic clusters 1, 2, 3, and 9, as defined in Fig. 3.2. Only clusters with at least 2 genes contributing to enrichments are shown.

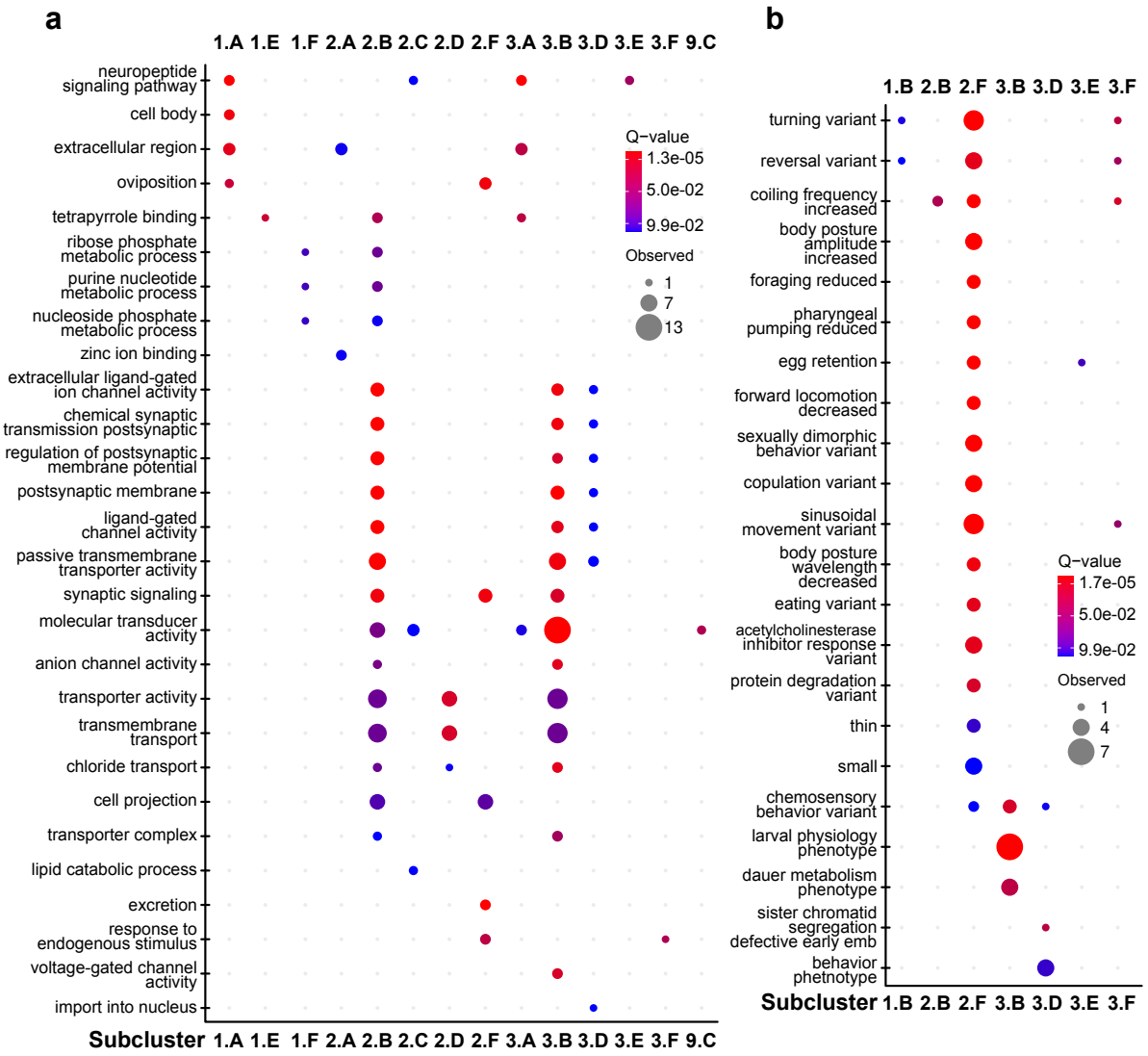


Figure D.3 – GO and phenotype enrichment of late embryo development subclusters
a,b Gene ontology (**a**) and phenotype (**b**) enrichment of subclusters from dynamic clusters 1, 2, 3, and 9, as defined in Fig. 3.2, with ≥ 8 genes. Only clusters with at least 2 genes contributing to enrichments are shown.

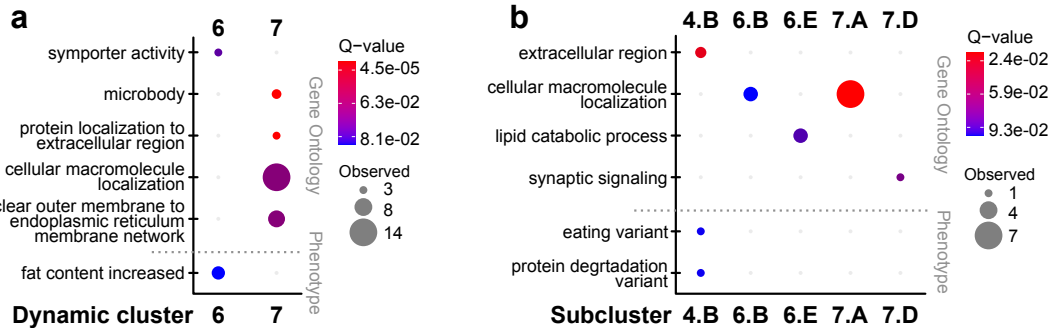


Figure D.4 – GO and phenotype enrichment of mid embryo development clusters
a,b Gene ontology and phenotype enrichment of mid-development clusters 4, 6, and 7 (**a**), and sub-clusters with ≥ 8 genes(**b**). Clusters as defined in Fig. 3.2.
Only clusters with at least 2 genes contributing to enrichments are shown.

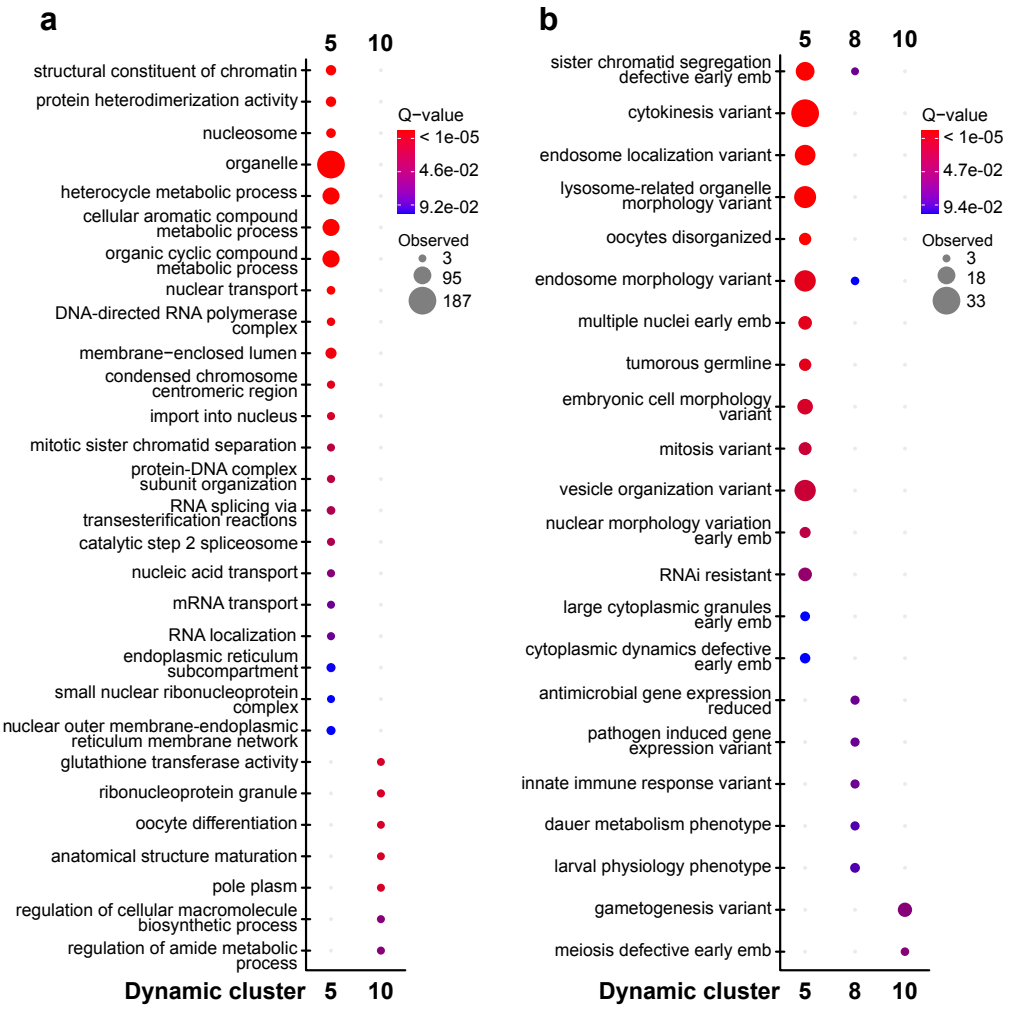


Figure D.5 – GO and phenotype enrichment of early embryo development clusters
a,b Gene ontology (**a**) and phenotype (**b**) enrichment of dynamic clusters 5, 8 and 10, as defined in Fig. 3.2. Only clusters with at least 2 genes contributing to enrichments are shown.

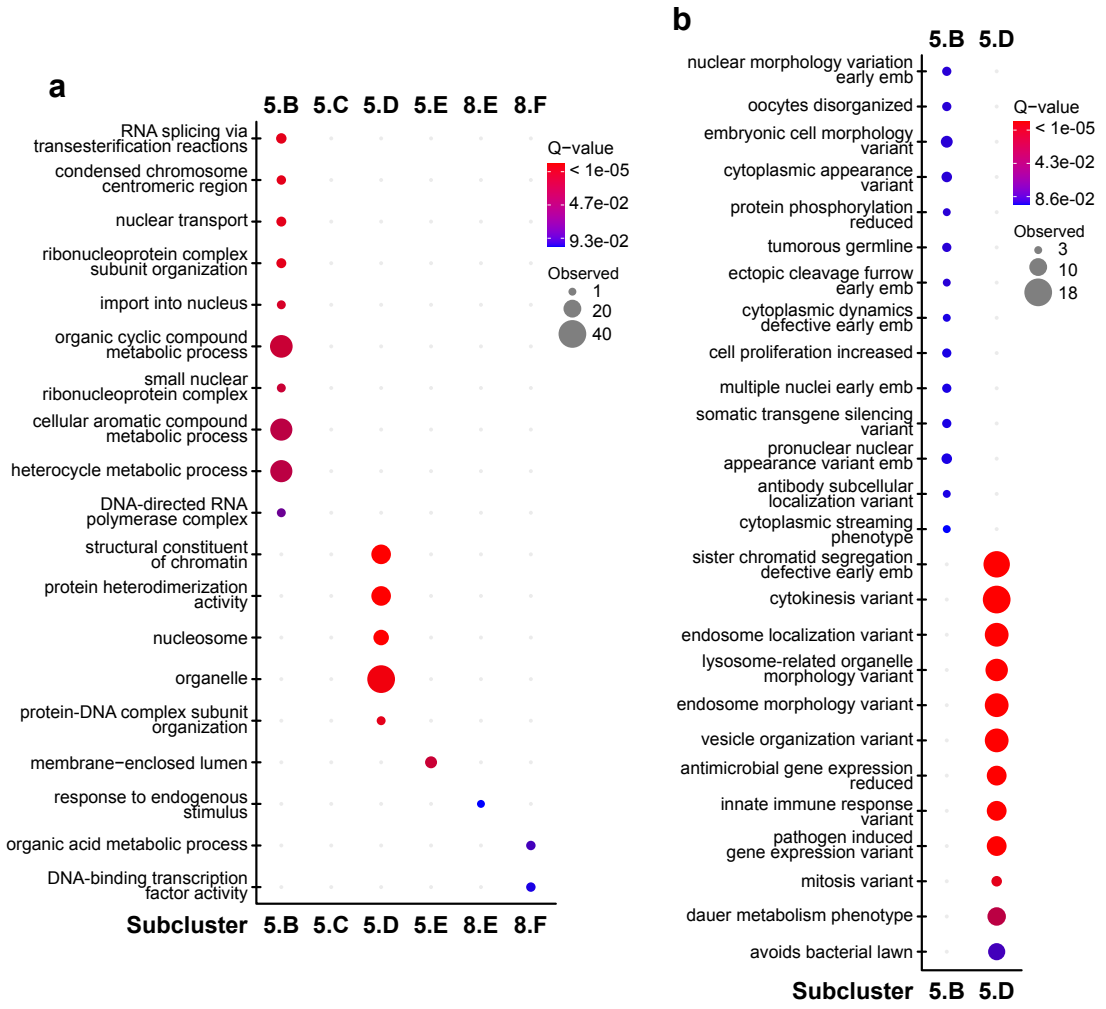


Figure D.6 – GO and phenotype enrichment of early embryo development subclusters
a,b Gene ontology (**a**) and phenotype (**b**) enrichment of subclusters from dynamic clusters 5, 8, and 10, as defined in Fig. 3.2, with ≥ 8 genes Only clusters with at least 2 genes contributing to enrichments are shown.

E Collaborations

E.1 Single-cell data analysis to support the role of IL-6 in cell differentiation

During my PhD, I collaborated with the team of Pr. Thomas Graf at the Centre for Genomic Regulation (CRG), in the Barcelona Institute of Science and Technology (BIST).

Pr. Graf and my supervisor Mirko Francesconi had recently partnered on a project to uncover the heterogeneity of cell differentiation and reprogramming efficiency using single-cell RNA-seq (FRANCESCONI et al., 2019). In the continuity of this project, Pr. Graf and Marcos Plana-Carmona were interested in further understanding the regulation and roles of InterLeukin 6 (IL-6) in cell differentiation. To this end, I re-analyzed the single-cell data from the previous collaboration to support experimental evidence from Pr. Graf's lab. Taking advantage of both the time-series design of the initial experiment and the similarity of cells to a cell type atlas, my work mostly confirmed expression trends of key genes and signatures.

This collaboration led to the two publications below, of which my supervisor and I are co-authors.

MARCOS PLANA-CARMONA, GREGOIRE STIK, **ROMAIN BULTEAU**, CAROLINA SEGURA-MORALES, NOELIA ALCÁZAR, CHRIS D. R. WYATT, ANTONIOS KLONIZAKIS, LUISA DE ANDRÉS-AGUAYO, MAXIME GASNIER, TIAN V. TIAN, GUILLEM TORCAL GARCIA, MARIA VILA-CASADESÚS, NICOLAS PLACHTA, MANUEL SERRANO, MIRKO FRANCESCONI, and THOMAS GRAF (2022). "The trophectoderm acts as a niche for the inner cell mass through C/EBPα-regulated IL-6 signaling". In: *Stem Cell Reports* 17.9, pp. 1991–2004

GUILLEM TORCAL GARCIA, ELISABETH KOWENZ-LEUTZ, TIAN V TIAN, ANTONIS KLONIZAKIS, JONATHAN LERNER, LUISA DE ANDRES-AGUAYO, VALERIIA SAPOZHNIKOVA, CLARA BERENGUER, MARCOS PLANA CARMONA, MARIA VILA CASADESUS, **ROMAIN BULTEAU**, MIRKO FRANCESCONI, SANDRA PEIRO, PHILIPP MERTINS, KENNETH ZARET, ACHIM LEUTZ, and THOMAS GRAF (2023). "Carm1-arginine methylation of the transcription factor C/EBPα regulates transdifferentiation velocity". In: *eLife* 12, e83951

E.2 Staging *C. elegans* aging muscle cells with RAPToR

As discussed in Chapter 1, section 1.2.4, I also collaborated with Dr. Florence Solari, in the context of Dr. Charline Roy's thesis. I could use RAPToR to confirm the biological age of bulk dissociated muscle cells from *C. elegans*, revealing not only that cells from long-lived *daf-2* mutants were delayed with respect to controls, but also that two sample IDs had been swapped after sequencing.

This work allowed us to prove that RAPToR could stage tissue samples on whole-organism reference data.

List of acronyms

DE Differentially Expressed, (DE analysis, Differential Expression Analysis).

FACS Fluorescence-Activated Cell Sorting.

FSC Forward Scatter.

GAM Generalized Additive Model.

GO Gene ontology.

ICA Independent Component Analysis (IC, Independent Component).

NGM Nematode Growth Medium.

PCA Principal Component Analysis (PC, Principal Component).

PCR Polymerase Chain Reaction.

RAPToR **R**eal **A**ge **P**rediction by **T**ranscriptome staging **o**n **R**eference.

RNA-seq RNA sequencing.

RNAi RNA interference.

RT-PCR Reverse-Transcription PCR.

SSC Side Scatter.

TGF-β Transforming Growth Factor β.

TPM Transcripts Per Million.

UMI Unique Molecular Identifier.

WT Wild-Type.

Full list of references

ACAR, MURAT, JEROME T METTETAL, and ALEXANDER VAN OUDENAARDEN (2008). "Stochastic switching as a survival strategy in fluctuating environments". In: *Nature genetics* 40.4, pp. 471–475.

AGRAWAL, ANURAG A, CHRISTIAN LAFORSCH, and RALPH TOLLRIAN (1999). "Transgenerational induction of defences in animals and plants". In: *Nature* 401.6748, pp. 60–63.

ALCAZAR, ROSA M, RUEYLING LIN, and ANDREW Z FIRE (2008). "Transmission dynamics of heritable silencing induced by double-stranded RNA in Caenorhabditis elegans". In: *Genetics* 180.3, pp. 1275–1288.

ALTUN, ZF and DH HALL (2003). "WormAtlas Hermaphrodite Handbook-Nervous System-Neuronal Support Cells". In: *WormAtlas. doi* 10.

ALTUN, ZF, DH HALL, and LA HERNDON (2006). "WormAtlas Hermaphrodite Handbook-Introduction". In: *WormAtlas*.

ANDERSON, LUCY M et al. (2006). "Preconceptional fasting of fathers alters serum glucose in offspring of mice". In: *Nutrition* 22.3, pp. 327–331.

ANGELES-ALBORES, DAVID et al. (2018). "Two new functions in the WormBase Enrichment Suite". In: *microPublication Biology* 2018.

ANSARI, AMIR MEHDI et al. (2016). "Cellular GFP toxicity and immunogenicity: potential confounders in in vivo cell tracking experiments". In: *Stem cell reviews and reports* 12, pp. 553–559.

ARTYUKHIN, ALEXANDER B, FRANK C SCHROEDER, and LEON AVERY (2013). "Density dependence in Caenorhabditis larval starvation". In: *Scientific reports* 3.1, p. 2777.

ARTYUKHIN, ALEXANDER B et al. (2018). "Metabolomic "dark matter" dependent on peroxisomal β-oxidation in Caenorhabditis elegans". In: *Journal of the American Chemical Society* 140.8, pp. 2841–2852.

BANSAL, ANKITA et al. (2015). "Uncoupling lifespan and healthspan in Caenorhabditis elegans longevity mutants". In: *Proceedings of the National Academy of Sciences* 112.3, E277–E286.

BAR-JOSEPH, ZIV, ANTHONY GITTER, and ITAMAR SIMON (2012). "Studying and modelling dynamic biological processes using time-series gene expression data". In: *Nature Reviews Genetics* 13.8, pp. 552–564.

BARGMANN, CORNELIA I (2006). "Chemosensation in C. elegans". In: *WormBook: The online review of C. elegans biology [Internet]*.

BARGMANN, CORNELIA I, ERIKA HARTWIEG, and H ROBERT HORVITZ (1993). "Odorant-selective genes and neurons mediate olfaction in C. elegans". In: *Cell* 74.3, pp. 515–527.

BARGMANN, CORNELIA I and H ROBERT HORVITZ (1991). "Control of larval development by chemosensory neurons in Caenorhabditis elegans". In: *Science* 251.4998, pp. 1243–1246.

BAUGH, L RYAN and TROY DAY (2020). "Nongenetic inheritance and multigenerational plasticity in the nematode C. elegans". In: *Elife* 9, e58498.

BAUGH, L RYAN et al. (2003). "Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome". In.

BELL, AVERY DAVIS et al. (2023). "Beyond the reference: gene expression variation and transcriptional response to RNA interference in Caenorhabditis elegans". In: *G3: Genes, Genomes, Genetics* 13.8, jkad112.

BELL, CHRISTOPHER G et al. (2019). "DNA methylation aging clocks: challenges and recommendations". In: *Genome biology* 20, pp. 1–24.

BHATTACHARJEE, NIRVEEK et al. (2016). "The upcoming 3D-printing revolution in microfluidics". In: *Lab on a Chip* 16.10, pp. 1720–1742.

BOSSINGER, OLAF and EINHARD SCHIERENBERG (1992). "Transfer and tissue-specific accumulation of cytoplasmic components in embryos of Caenorhabditis elegans and Rhabditis dolichura: in vivo analysis with a low-cost signal enhancement device". In: *Development* 114.2, pp. 317–330.

BRENNER, SYDNEY (1974). "The genetics of Caenorhabditis elegans". In: *Genetics* 77.1, pp. 71–94.

BRINK, SUSANNE C VAN DEN et al. (2017). "Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations". In: *Nature methods* 14.10, pp. 935–936.

BRITTIN, CHRISTOPHER A et al. (2021). "A multi-scale brain map derived from whole-brain volumetric reconstructions". In: *Nature* 591.7848, pp. 105–110.

BUCKLEY, BETHANY A et al. (2012). "A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality". In: *Nature* 489.7416, pp. 447–451.

BULTEAU, ROMAIN and MIRKO FRANCESCONI (Aug. 2022). "Real age prediction from the transcriptome with RAPToR". en. In: *Nature Methods* 19.8, pp. 969–975. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01540-0.

BURTON, NICHOLAS O and ERIC L GREER (2022). "Multigenerational epigenetic inheritance: Transmitting information across generations". In: *Seminars in cell & developmental biology*. Vol. 127. Elsevier, pp. 121–132.

BUTCHER, REBECCA A et al. (2007). "Small-molecule pheromones that control dauer development in Caenorhabditis elegans". In: *Nature chemical biology* 3.7, pp. 420–422.

BYRNE, J et al. (Jan. 2020). *Gene changes over aging in the C.elegans rrf-3(pk1426) mutant*. unpublished. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93826 (visited on 09/22/2022).

CAO, JUNYUE et al. (2019). "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745, pp. 496–502.

CASANUEVA, M OLIVIA, ALEJANDRO BURGA, and BEN LEHNER (2012). "Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans". In: *Science* 335.6064, pp. 82–85.

CHANG, DENNIS et al. (2021). "A revised adaptation of the smart-Seq2 protocol for single-nematode RNA-seq". In: *RNA Abundance Analysis: Methods and Protocols*, pp. 79–99.

CHEN, YAN-TING et al. (2021). "Imprinted lncRNA Dio3os preprograms intergenerational brown fat development and obesity resistance". In: *Nature communications* 12.1, p. 6845.

CHISHOLM, ANDREW D and SUHONG XU (2012). "The Caenorhabditis elegans epidermis as a model skin. II: differentiation and physiological roles". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 1.6, pp. 879–902.

CHOE, ANDREA et al. (2012). "Ascaroside signaling is widely conserved among nematodes". In: *Current Biology* 22.9, pp. 772–780.

CHUTE, CHRISTOPHER D et al. (2019). "Co-option of neurotransmitter signaling for inter-organismal communication in C. elegans". In: *Nature Communications* 10.1, p. 3186.

CLARKE, KR and RH GREEN (1988). "Statistical design and analysis for a'biological effects' study". In: *Marine Ecology Progress Series*, pp. 213–226.

CLOKEY, GEORGE V and LEWIS A JACOBSON (1986). "The autofluorescent "lipofuscin granules" in the intestinal cells of Caenorhabditis elegans are secondary lysosomes". In: *Mechanisms of ageing and development* 35.1, pp. 79–94.

COOK, STEVEN J, CRISTINE A KALINSKI, and OLIVER HOBERT (2023). "Neuronal contact predicts connectivity in the C. elegans brain". In: *Current Biology* 33.11, pp. 2315–2320.

CUI, MIAO, CHAO CHENG, and LANJING ZHANG (2022). "High-throughput proteomics: a methodological mini-review". In: *Laboratory Investigation* 102.11, pp. 1170–1181.

DALFÓ, DIANA, DAVID MICHAELSON, and E JANE ALBERT HUBBARD (2012). "Sensory regulation of the C. elegans germline through TGF-β-dependent signaling in the niche". In: *Current Biology* 22.8, pp. 712–719.

DENISENKO, ELENA et al. (2020). "Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows". In: *Genome biology* 21.1, pp. 1–25.

DEVANAPALLY, SINDHUJA, SNUSHA RAVIKUMAR, and ANTONY M JOSE (2015). "Double-stranded RNA made in C. elegans neurons can enter the germline and cause transgenerational gene silencing". In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2133–2138.

DIAS, BRIAN G and KERRY J RESSLER (2014). "Parental olfactory experience influences behavior and neural structure in subsequent generations". In: *Nature neuroscience* 17.1, pp. 89–96.

DOROQUEZ, DAVID B et al. (2014). "A high-resolution morphological and ultrastructural map of anterior sensory cilia and glia in Caenorhabditis elegans". In: *Elife* 3, e01948.

EDWARDS, STACEY L et al. (2021). "Insulin/IGF-1 signaling and heat stress differentially regulate HSF1 activities in germline development". In: *Cell reports* 36.9.

FERNANDEZ, ANITA G et al. (2010). "Automated sorting of live C. elegans using laFACS". In: *Nature methods* 7.6, pp. 417–418.

FITZ-JAMES, MAXIMILIAN H and GIACOMO CAVALLI (2022). "Molecular mechanisms of transgenerational epigenetic inheritance". In: *Nature Reviews Genetics* 23.6, pp. 325–341.

FOX, REBECCA M et al. (2007). "The embryonic muscle transcriptome of Caenorhabditis elegans". In: *Genome biology* 8, pp. 1–20.

FRANCESCONI, MIRKO and ROMAIN BULTEAU (2022). "Inferring biological age from the transcriptome with RAPToR". In: *Nature Methods* 19.8, pp. 936–937. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01542-y.

FRANCESCONI, MIRKO and BEN LEHNER (2014). "The effects of genetic variation on gene expression dynamics during development". In: *Nature* 505.7482, pp. 208–211.

FRANCESCONI, MIRKO et al. (2019). "Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming". In: *Elife* 8, e41627.

FREY, NOLAN et al. (2022). "Microfluidics for understanding model organisms". In: *Nature communications* 13.1, p. 3195.

FRÉZAL, LISE and MARIE-ANNE FÉLIX (2015). "C. elegans outside the Petri dish". In: *elife* 4, e05849.

FRIDMANN-SIRKIS, YAEL et al. (2014). "Delayed development induced by toxicity to the host can be inherited by a bacterial-dependent, transgenerational effect". In: *Frontiers in genetics* 5, p. 27.

GEMS, DAVID et al. (1998). "Two pleiotropic classes of daf-2 mutation affect larval arrest, adult behavior, reproduction and longevity in Caenorhabditis elegans". In: *Genetics* 150.1, pp. 129–155.

GINZBERG, MIRIAM B, RAN KAFRI, and MARC KIRSCHNER (2015). "On being the right (cell) size". In: *Science* 348.6236, p. 1245075.

GOLDEN, JAMES W and DONALD L RIDDLE (1982). "A pheromone influences larval development in the nematode Caenorhabditis elegans". In: *Science* 218.4572, pp. 578–580.

GOLDEN, JAMES W and DONALD L RIDDLE (1984). "The Caenorhabditis elegans dauer larva: developmental effects of pheromone, food, and temperature". In: *Developmental biology* 102.2, pp. 368–378.

GOLDEN, TAMARA R et al. (2008). "Age-related behaviors have distinct transcriptional profiles in Caenorhabditis elegans". In: *Aging cell* 7.6, pp. 850–865.

GOODMAN, MIRIAM B and PIALI SENGUPTA (2019). "How Caenorhabditis elegans senses mechanical stress, temperature, and other physical stimuli". In: *Genetics* 212.1, pp. 25–51.

GRAHAM, MICHAEL H (2003). "Confronting multicollinearity in ecological multiple regression". In: *Ecology* 84.11, pp. 2809–2815.

GRANDJEAN, VALÉRIE et al. (2015). "RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders". In: *Scientific reports* 5.1, p. 18193.

GREENE, JOSHUA S et al. (2016a). "Balancing selection shapes density-dependent foraging behaviour". In: *Nature* 539.7628, pp. 254–258.

GREENE, JOSHUA S et al. (2016b). "Regulatory changes in two chemoreceptor genes contribute to a Caenorhabditis elegans QTL for foraging behavior". In: *elife* 5, e21454.

GREENGARD, PAUL (2001). "The neurobiology of slow synaptic transmission". In: *Science* 294.5544, pp. 1024–1030.

GREER, ERIC L et al. (2011). "Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans". In: *Nature* 479.7373, pp. 365–371.

GRIFFIN, JEANINE et al. (2006). "Comparative analysis of follicle morphology and oocyte diameter in four mammalian species (mouse, hamster, pig, and human)". In: *Journal of experimental & clinical assisted reproduction* 3, pp. 1–9.

HAGAN, THOMAS et al. (2022). "Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses". In: *Nature Immunology* 23.12, pp. 1788–1798.

HAGEMANN-JENSEN, MICHAEL, CHRISTOPH ZIEGENHAIN, and RICKARD SANDBERG (2022). "Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress". In: *Nature Biotechnology* 40.10, pp. 1452–1457.

HAGEMANN-JENSEN, MICHAEL et al. (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In: *Nature Biotechnology* 38.6, pp. 708–714.

HALLGRIMSSON, BENEDIKT and BRIAN K HALL (2011). *Variation: a central concept in biology*. Elsevier.

HARDO, GEORGEOS and SOMENATH BAKSHI (2021). "Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments". In: *Essays in Biochemistry* 65.1, pp. 67–79.

HASHIMSHONY, TAMAR et al. (2012). "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification". In: *Cell reports* 2.3, pp. 666–673.

HASHIMSHONY, TAMAR et al. (2015). "Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer". In: *Nature* 519.7542, pp. 219–222.

HASTINGS, JANNA et al. (2019). "Multi-omics and genome-scale modeling reveal a metabolic shift during C. elegans aging". In: *Frontiers in molecular biosciences* 6, p. 2.

HENDRIKS, GERT-JAN et al. (2014). "Extensive oscillatory gene expression during C. elegans larval development". In: *Molecular cell* 53.3, pp. 380–392.

HEPPERT, JENNIFER K et al. (2016). "Comparative assessment of fluorescent proteins for in vivo imaging in an animal model system". In: *Molecular biology of the cell* 27.22, pp. 3385–3394.

HIBSHMAN, JONATHAN D, ANTHONY HUNG, and L RYAN BAUGH (2016). "Maternal diet and insulin-like signaling control intergenerational plasticity of progeny size and starvation resistance". In: *PLoS genetics* 12.10, e1006396.

HIRSH, DAVID, DANIEL OPPENHEIM, and MICHAEL KLASS (1976). "Development of the reproductive system of Caenorhabditis elegans". In: *Developmental biology* 49.1, pp. 200–219.

HODGKIN, JONATHAN and THOMAS M BARNES (1991). "More is not better: brood size and population growth in a self-fertilizing nematode". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 246.1315, pp. 19–24.

HOEIJMAKERS, WIETEKE AM, RICHÁRD BÁRTFAI, and HENDRIK G STUNNENBERG (2013). "Transcriptome analysis using RNA-Seq". In: *Malaria: Methods and Protocols*, pp. 221–239.

HOU, LEI et al. (2016). "A systems approach to reverse engineer lifespan extension by dietary restriction". In: *Cell metabolism* 23.3, pp. 529–540.

HUH, DANN and JOHAN PAULSSON (2011). "Random partitioning of molecules at cell division". In: *Proceedings of the National Academy of Sciences* 108.36, pp. 15004–15009.

HUMS, INGRID et al. (2016). "Regulation of two motor patterns enables the gradual adjustment of locomotion strategy in Caenorhabditis elegans". In: *Elife* 5, e14116.

HUSSEY, ROSALIND et al. (2017). "Pheromone-sensing neurons regulate peripheral lipid metabolism in Caenorhabditis elegans". In: *PLoS genetics* 13.5, e1006806.

HUZAIRA, MISBAH et al. (2001). "Topographic variations in normal skin, as viewed by in vivo reflectance confocal microscopy". In: *Journal of investigative dermatology* 116.6, pp. 846–852.

INSALL, ROBERT H (2001). "The Whole Organism, and nothing but the Organism". In: *Cell* 107.3, pp. 279–281.

ISLAM, SAIFUL et al. (2014). "Quantitative single-cell RNA-seq with unique molecular identifiers". In: *Nature methods* 11.2, pp. 163–166.

JABLONKA, EVA, MICHAEL LACHMANN, and MARION J LAMB (1992). "Evidence, mechanisms and models for the inheritance of acquired characters". In: *Journal of theoretical biology* 158.2, pp. 245–268.

JAITIN, DIEGO ADHEMAR et al. (2014). "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types". In: *Science* 343.6172, pp. 776–779.

JOHNSON, RYAN W et al. (2009). "The Caenorhabditis elegans heterochronic gene lin-14 coordinates temporal progression and maturation in the egg-laying system". In: *Developmental Dynamics* 238.2, pp. 394–404.

JORDAN, JAMES M et al. (2019). "Insulin/IGF signaling and vitellogenin provisioning mediate intergenerational adaptation to nutrient stress". In: *Current Biology* 29.14, pp. 2380–2388.

JUANG, BI-TZEN et al. (2013). "Endogenous nuclear RNAi mediates behavioral adaptation to odor". In: *Cell* 154.5, pp. 1010–1022.

KALETSKY, RACHEL et al. (2018). "Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression". In: *PLoS genetics* 14.8, e1007559.

KALETSKY, RACHEL et al. (2020). "C. elegans interprets bacterial non-coding RNAs to learn pathogenic avoidance". In: *Nature* 586.7829, pp. 445–451.

KIM, BYOUNGHUN et al. (2020). "Regulatory systems that mediate the effects of temperature on the lifespan of Caenorhabditis elegans". In: *Journal of Neurogenetics* 34.3-4, pp. 518–526.

KIM, DONG HYUN, DOMINIC GRÜN, and ALEXANDER VAN OUDENAARDEN (2013). "Dampening of expression oscillations by synchronous regulation of a microRNA and its target". In: *Nature genetics* 45.11, pp. 1337–1344.

KIM, EUNAH et al. (2023). "Mitochondrial aconitase suppresses immunity by modulating oxaloacetate and the mitochondrial unfolded protein response". In: *Nature Communications* 14.1, p. 3716.

KIMMEL, CHARLES B et al. (1995). "Stages of embryonic development of the zebrafish". In: *Developmental dynamics* 203.3, pp. 253–310.

KIRKWOOD, THOMAS BL et al. (2005). "What accounts for the wide variation in life span of genetically identical organisms reared in a constant environment?" In: *Mechanisms of ageing and development* 126.3, pp. 439–443.

KISHIMOTO, SAYA et al. (2017). "Environmental stresses induce transgenerationally inheritable survival advantages via germline-to-soma communication in Caenorhabditis elegans". In: *Nature communications* 8.1, p. 14031.

KIVIOJA, TEEMU et al. (2012). "Counting absolute numbers of molecules using unique molecular identifiers". In: *Nature methods* 9.1, pp. 72–74.

KLABONSKI, LAUREN et al. (2016). "A bystander mechanism explains the specific phenotype of a broadly expressed misfolded protein". In: *PLoS Genetics* 12.12, e1006450.

KLOSIN, ADAM et al. (2017). "Transgenerational transmission of environmental information in C. elegans". In: *Science* 356.6335, pp. 320–323.

KWON, YOUNG JOON et al. (2018). "High-throughput BioSorter quantification of relative mitochondrial content and membrane potential in living Caenorhabditis elegans". In: *Mitochondrion* 40, pp. 42–50.

LEE, DAEHAN et al. (2023). "Natural genetic variation in the pheromone production of C. elegans". In: *Proceedings of the National Academy of Sciences* 120.26, e2221150120.

LEVIN, MICHAL et al. (2016). "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596, pp. 637–641.

LIKITLERSUANG, JIRAPAT et al. (2012). "C. elegans tracking and behavioral measurement". In: *JoVE (Journal of Visualized Experiments)* 69, e4094.

LORENZ, TODD C (2012). "Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies". In: *JoVE (Journal of Visualized Experiments)* 63, e3998.

LOVE, MICHAEL I, WOLFGANG HUBER, and SIMON ANDERS (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12. Publisher: BioMed Central, pp. 1–21.

LUDEWIG, ANDREAS H and FRANK C SCHROEDER (2018). "Ascaroside signaling in C. elegans". In: *WormBook: The Online Review of C. elegans Biology [Internet]*.

LUDEWIG, ANDREAS H et al. (2013). "Pheromone sensing regulates Caenorhabditis elegans lifespan and stress resistance via the deacetylase SIR-2.1". In: *Proceedings of the National Academy of Sciences* 110.14, pp. 5522–5527.

LUDEWIG, ANDREAS H et al. (2019). "An excreted small molecule promotes C. elegans reproductive development and aging". In: *Nature chemical biology* 15.8, pp. 838–845.

MACCHIETTO, MARISSA et al. (2017). "Comparative transcriptomics of Steinernema and Caenorhabditis single embryos reveals orthologous gene expression convergence during late embryogenesis". In: *Genome biology and evolution* 9.10, pp. 2681–2696.

MACHADO, LÉO et al. (2021). "Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation". In: *Cell Stem Cell* 28.6, pp. 1125–1135.

MACOSKO, EVAN Z et al. (2009). "A hub-and-spoke circuit drives pheromone attraction and social behaviour in C. elegans". In: *Nature* 458.7242, pp. 1171–1175.

MARRÉ, JULIA, EDWARD C TRAVER, and ANTONY M JOSE (2016). "Extracellular RNA is transported from one generation to the next in Caenorhabditis elegans". In: *Proceedings of the National Academy of Sciences* 113.44, pp. 12496–12501.

MAURES, TRAVIS J et al. (2014). "Males shorten the life span of C. elegans hermaphrodites via secreted compounds". In: *Science* 343.6170, pp. 541–544.

MCKNIGHT, KATHERINE et al. (2014). "Neurosensory perception of environmental cues modulates sperm motility critical for fertilization". In: *Science* 344.6185, pp. 754–757.

MEEUSE, MILOU WM et al. (2020). "Developmental function and state transitions of a gene expression oscillator in Caenorhabditis elegans". In: *Molecular systems biology* 16.7, e9498.

MIKI, TAKASHI S, SARAH H CARL, and HELGE GROSSHANS (2017). "Two distinct transcription termination modes dictated by promoters". In: *Genes & development* 31.18, pp. 1870–1879.

MOORE, REBECCA S, RACHEL KALETSKY, and COLEEN T MURPHY (2019). "Piwi/PRG-1 argonaute and TGF-β mediate transgenerational learned pathogenic avoidance". In: *Cell* 177.7, pp. 1827–1841.

MORIMOTO, JULIANO, STEPHEN J SIMPSON, and FLEUR PONTON (2017). "Direct and trans-generational effects of male and female gut microbiota in Drosophila melanogaster". In: *Biology letters* 13.7, p. 20160966.

MOSS, ERIC G (2007). "Heterochronic genes and the nature of developmental time". In: *Current Biology* 17.11, R425–R434.

MUCH, JASON W et al. (2000). "The fax-1 nuclear hormone receptor regulates axon pathfinding and neurotransmitter expression". In: *Development* 127.4, pp. 703–712.

O'REILLY, LINDA P et al. (2014). "C. elegans in high-throughput drug discovery". In: *Advanced drug delivery reviews* 69, pp. 247–253.

OIKONOMOPOULOS, SPYROS et al. (2020). "Methodologies for transcript profiling using long-read technologies". In: *Frontiers in genetics* 11, p. 606.

ORANTH, ALEXANDRA et al. (2018). "Food sensation modulates locomotion by dopamine and neuropeptide signaling in a distributed neuronal network". In: *Neuron* 100.6, pp. 1414–1428.

PACKER, JONATHAN S et al. (2019). "A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution". In: *Science* 365.6459, eaax1971.

PAREKH, SWATI et al. (2016). "The impact of amplification on differential expression analyses by RNA-seq". In: *Scientific reports* 6.1, p. 25533.

PAREKH, SWATI et al. (2018). "zUMIs-A fast and flexible pipeline to process RNA sequencing data with UMIs". In: *Gigascience* 7.6, giy059.

PEMBREY, MARCUS, RICHARD SAFFERY, LARS OLOV BYGREN, et al. (2014). "Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research". In: *Journal of medical genetics* 51.9, pp. 563–572.

PEMBREY, MARCUS E et al. (2006). "Sex-specific, male-line transgenerational responses in humans". In: *European journal of human genetics* 14.2, pp. 159–166.

PEREZ, MARCOS FRANCISCO and BEN LEHNER (2019). "Intergenerational and transgenerational epigenetic inheritance in animals". In: *Nature cell biology* 21.2, pp. 143–151.

PEREZ, MARCOS FRANCISCO et al. (2017). "Maternal age generates phenotypic variation in Caenorhabditis elegans". In: *Nature* 552.7683, pp. 106–109.

PEREZ, MARCOS FRANCISCO et al. (2021a). "Neuronal perception of the social environment generates an inherited memory that controls the development and generation time of C. elegans". In: *Current Biology* 31.19, pp. 4256–4268.

PEREZ, MARCOS FRANCISCO et al. (2021b). *Unpublished data related to Curr. biology article*. unpublished.

PEREZ-MOJICA, J EDUARDO et al. (2023). "Continuous transcriptome analysis reveals novel patterns of early gene expression in Drosophila embryos". In: *Cell genomics* 3.3.

PERKINS, LIZABETH A et al. (1986). "Mutant sensory cilia in the nematode Caenorhabditis elegans". In: *Developmental biology* 117.2, pp. 456–487.

PICELLI, SIMONE et al. (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nature methods* 10.11, pp. 1096–1098.

PICELLI, SIMONE et al. (2014). "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature protocols* 9.1, pp. 171–181.

PLANA-CARMONA, MARCOS et al. (2022). "The trophectoderm acts as a niche for the inner cell mass through C/EBPα-regulated IL-6 signaling". In: *Stem Cell Reports* 17.9, pp. 1991–2004.

POSNER, RACHEL et al. (2019). "Neuronal small RNAs control behavior transgenerationally". In: *Cell* 177.7, pp. 1814–1826.

PRINYAKUPT, JAROONRUT and CHARNCHAI PLUEMPITIWIRIYAWEJ (2015). "Segmentation of white blood cells and comparison of cell morphology by linear and naive Bayes classifiers". In: *Biomedical engineering online* 14, pp. 1–19.

RECHAVI, ODED et al. (2014). "Starvation-induced transgenerational inheritance of small RNAs in C. elegans". In: *Cell* 158.2, pp. 277–287.

REINKE, VALERIE et al. (2004). "Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans". In.

REMY, JEAN-JACQUES (2010). "Stable inheritance of an acquired behavior in Caenorhabditis elegans". In: *Current Biology* 20.20, R877–R878.

REMY, JEAN-JACQUES and OLIVER HOBERT (2005). "An interneuronal chemoreceptor required for olfactory imprinting in C. elegans". In: *Science* 309.5735, pp. 787–790.

REN, PEIFENG et al. (1996). "Control of C. elegans larval development by neuronal expression of a TGF-β homolog". In: *Science* 274.5291, pp. 1389–1391.

REY, CARINE et al. (2022). "Programmed-DNA Elimination in the free-living nematodes Mesorhabditis". In: *bioRxiv*, pp. 2022–03.

RIDDLE, DONALD L et al. (1997). "C. elegans ii". In.

ROBINSON, MARK D, DAVIS J MCCARTHY, and GORDON K SMYTH (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1. Publisher: Oxford University Press, pp. 139–140.

RODRIGUES, NELIO TL et al. (2022). "SAIBR: A simple, platform-independent method for spectral autofluorescence correction". In: *Development* 149.14, dev200545.

ROMÁN, ANGEL-CARLOS et al. (2018). "Histone H4 acetylation regulates behavioral inter-individual variability in zebrafish". In: *Genome biology* 19.1, pp. 1–21.

ROY, CHARLINE (2022). "Rôle du récepteur de l'insuline/IGF-1 DAF-2 dans le contrôle du vieillissement musculaire et son impact sur le transcriptome". PhD thesis. Lyon 1.

ROY, CHARLINE et al. (2022). "DAF-2/insulin IGF-1 receptor regulates motility during aging by integrating opposite signaling from muscle and neuronal tissues". In: *Aging Cell* 21.8, e13660.

SCHOTT, DANIEL, ITAI YANAI, and CRAIG P HUNTER (2014). "Natural RNA interference directs a heritable response to the environment". In: *Scientific reports* 4.1, p. 7387.

SEQUENCING CONSORTIUM, C. ELEGANS (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology". In: *Science* 282.5396, pp. 2012–2018.

SERRA, LORRAYNE et al. (2018). "Adapting the smart-seq2 protocol for robust single worm RNA-seq". In: *Bio-protocol* 8.4, e2729–e2729.

SILVERMAN, BW and JT WOOD (1987). "The nonparametric estimation of branching curves". In: *Journal of the American Statistical Association* 82.398, pp. 551–558.

SINIGAGLIA, CHIARA et al. (2022). "Distinct gene expression dynamics in developing and regenerating crustacean limbs". In: *Proceedings of the National Academy of Sciences* 119.27, e2119297119.

SMITH, STEPHEN and RAMON GRIMA (2018). "Single-cell variability in multicellular organisms". In: *Nature communications* 9.1, p. 345.

SNOEK, L BASTEN et al. (2014). "A rapid and massive gene expression shift marking adolescent transition in C. elegans". In: *Scientific reports* 4.1, pp. 1–5.

SRINIVASAN, JAGAN et al. (2008). "A blend of small molecules regulates both mating and development in Caenorhabditis elegans". In: *Nature* 454.7208, pp. 1115–1118.

SRINIVASAN, JAGAN et al. (2012). "A modular library of small molecule signals regulates social behaviors in Caenorhabditis elegans". In: *PLoS biology* 10.1, e1001237.

STARK, RORY, MARTA GRZELAK, and JAMES HADFIELD (2019). "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11, pp. 631–656.

STERKEN, MARK G et al. (2015). "The laboratory domestication of Caenorhabditis elegans". In: *Trends in Genetics* 31.5, pp. 224–231.

STOECKIUS, MARLON et al. (2009). "Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression". In: *Nature methods* 6.10, pp. 745–751.

STOREY, JOHN D et al. (2005). "Significance analysis of time course microarray experiments". In: *Proceedings of the National Academy of Sciences* 102.36, pp. 12837–12842.

STREET, KELLY et al. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC genomics* 19, pp. 1–16.

STROUSTRUP, NICHOLAS et al. (2013). "The Caenorhabditis elegans lifespan machine". In: *Nature methods* 10.7, pp. 665–670.

SULSTON, JOHN E et al. (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans". In: *Developmental biology* 100.1, pp. 64–119.

SURIYALAKSH, MANUSNAN et al. (2022). "Gene regulatory network inference in long-lived C. elegans reveals modular properties that are predictive of novel aging genes". In: *Iscience* 25.1.

SVENSSON, VALENTINE, ROSER VENTO-TORMO, and SARAH A TEICHMANN (2018). "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nature protocols* 13.4, pp. 599–604.

SVENSSON, VALENTINE et al. (2017). "Power analysis of single-cell RNA-sequencing experiments". In: *Nature methods* 14.4, pp. 381–387.

TAUFFENBERGER, ARNAUD and J ALEX PARKER (2014). "Heritable transmission of stress resistance by high dietary glucose in Caenorhabditis elegans". In: *PLoS genetics* 10.5, e1004346.

THATTAI, MUKUND and ALEXANDER VAN OUDENAARDEN (2004). "Stochastic gene expression in fluctuating environments". In: *Genetics* 167.1, pp. 523–530.

TINTORI, SOPHIA C et al. (2016). "A transcriptional lineage of the early C. elegans embryo". In: *Developmental Cell* 38.4, pp. 430–444.

TOKER, ITAI ANTOINE et al. (2022). "Transgenerational inheritance of sexual attractiveness via small RNAs enhances evolvability in C. elegans". In: *Developmental Cell* 57.3, pp. 298–309.

TORCAL GARCIA, GUILLEM et al. (2023). "Carm1-arginine methylation of the transcription factor C/EBPα regulates transdifferentiation velocity". In: *eLife* 12, e83951.

UPHOFF, STEPHAN et al. (2016). "Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation". In: *Science* 351.6277, pp. 1094–1097.

VAN KLEUNEN, MARK and MARKUS FISCHER (2005). "Constraints on the evolution of adaptive phenotypic plasticity in plants". In: *New phytologist* 166.1, pp. 49–60.

VASTENHOUW, NADINE L et al. (2006). "Long-term gene silencing by RNAi". In: *Nature* 442.7105, pp. 882–882.

VIETH, BEATE et al. (2019). "A systematic evaluation of single cell RNA-seq analysis pipelines". In: *Nature communications* 10.1, p. 4667.

VON REUSS, STEPHAN H et al. (2012). "Comparative metabolomics reveals biogenesis of ascarosides, a modular library of small-molecule signals in C. elegans". In: *Journal of the American Chemical Society* 134.3, pp. 1817–1824.

WALSH, MATTHEW R et al. (2015). "Predator-induced phenotypic plasticity within-and across-generations: a challenge for theory?" In: *Proceedings of the Royal Society B: Biological Sciences* 282.1798, p. 20142205.

WAN, JASON and HANG LU (2020). "Enabling high-throughput single-animal gene-expression studies with molecular and micro-scale technologies". In: *Lab on a Chip* 20.24, pp. 4528–4538.

WANG, TINA et al. (2020). "Quantitative translation of dog-to-human aging by conserved remodeling of the DNA methylome". In: *Cell systems* 11.2, pp. 176–185.

WASSON, JADIEL A et al. (2021). "Neuronal control of maternal provisioning in response to social cues". In: *Science Advances* 7.34, eabf8782.

WEBER, KATHERINE P et al. (2010). "Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans". In: *PloS one* 5.11, e13922.

WHITE, JOHN G et al. (1986). "The structure of the nervous system of the nematode Caenorhabditis elegans". In: *Philos Trans R Soc Lond B Biol Sci* 314.1165, pp. 1–340.

WHITE, RICHARD J et al. (2017). "A high-resolution mRNA expression time course of embryonic development in zebrafish". In: *elife* 6, e30860.

WITVLIET, DANIEL et al. (2021). "Connectomes across development reveal principles of brain maturation". In: *Nature* 596.7871, pp. 257–261.

ZHANG, GAOTIAN et al. (2022). "The impact of species-wide gene expression variation on Caenorhabditis elegans complex traits". In: *Nature communications* 13.1, p. 3462.

ZHANG, SIHUI and JEFFREY R KUHN (2018). "Cell isolation and culture". In: *WormBook: The Online Review of C. elegans Biology [Internet]*.

ZHANG, WILLIAM B et al. (2016). "Extended twilight among isogenic C. elegans causes a disproportionate scaling between lifespan and health". In: *Cell Systems* 3.4, pp. 333–345.

ZHANG, XIANNIAN et al. (2019). "Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems". In: *Molecular cell* 73.1, pp. 130–142.

ZHUANG, JIMMY J and CRAIG P HUNTER (2011). "Tissue specificity of Caenorhabditis elegans enhanced RNA interference mutants". In: *Genetics* 188.1, pp. 235–237.

ZIEGENHAIN, CHRISTOPH et al. (2017). "Comparative analysis of single-cell RNA sequencing methods". In: *Molecular cell* 65.4, pp. 631–643.

PhD thesis summary

The environment influences not only the behavior and physiology of an organism, but can also impact its descendents. In the nematode model *C. elegans*, perception of social cues (pheromones) elicits such intergenerational effects, notably increasing generation time of the progeny. Here, I characterize the molecular changes in embryos caused by parental pheromone exposure by profiling gene expression in single individuals.

To achieve this, I first developed a robust computational method that infers age from the transcriptome in diverse organisms and sample types, makes it possible to detect and correct for developmental bias in gene expression data, and allows us to bypass synchronization and staging challenges for embryo collection. Then, I adapted experimental techniques used for sorting and profiling single cells to single embryos in order to improve throughput, revealing great potential for accessible and cost-efficient studies at large scale. Armed with these methods, I could then profile genome-wide gene expression across embryo development in the progeny of pheromone-exposed and control parents. I show that the developing nervous system and sensory organs are influenced by parental neuronal perception of the environment, likely changing how progeny will experience their own surroundings.

Résumé de la thèse

L'environnement n'influence pas seulement le comportement et la physiologie d'un organisme, mais peut également avoir un impact sur sa descendance. Dans le nématode modèle *C. elegans*, la perception de l'environnement social (phéromones) déclenche de tels effets intergénérationnels, augmentant notamment le temps de génération de la progéniture. Dans mes travaux, je caractérise les changements moléculaires dans les embryons causés par l'exposition des parents aux phéromones en profilant l'expression des gènes à l'échelle de l'individu.

Pour y parvenir, j'ai d'abord développé une méthode computationnelle robuste capable d'estimer l'âge à partir du transcriptome dans divers organismes et types d'échantillons, qui permet de détecter et de corriger les biais liés au développement dans les données d'expression génique, et nous permet de contourner les défis de synchronisation et de stadification pour la collecte des embryons. J'ai ensuite adapté des techniques expérimentales initialement utilisées pour trier et profiler les cellules uniques (*single-cell*) aux embryons individuels pour permettre un haut débit, révélant un important potentiel pour mener des études à grande échelle de manière accessible et à moindre coût. Armé de ces méthodes, j'ai ensuite pu profiler l'expression des gènes à l'échelle du génome tout au long du développement de l'embryon chez la progéniture de parents exposés aux phéromones et de témoins. Je montre que l'expression génique du système nerveux et des organes sensoriels est influencée au cours de leur développement par la perception neuronale de l'environnement des parents, ce qui change certainement la manière dont la progéniture percevra son propre environnement.