1 **Control of tissue development and cell diversity by cell cycle dependent transcriptional filtering**

2

3 Maria Abou Chakra, Ruth Isserlin, Thinh Tran and Gary D. Bader*

4 The Donnelly Centre, University of Toronto, Toronto, M5S 3E1, Canada.

5 *Correspondence: gary.bader@utoronto.ca

6

7 **Abstract**

8 Cell cycle duration changes dramatically during development, starting out fast to generate cells quickly

9 and slowing down over time as the organism matures. The cell cycle can also act as a transcriptional

10 filter to control the expression of long gene transcripts which are partially transcribed in short cycles.

11 Using mathematical simulations of cell proliferation, we identify an emergent property, that this filter

12 can act as a tuning knob to control gene transcript expression, cell diversity and the number and

13 proportion of different cell types in a tissue. Our predictions are supported by comparison to single-cell

14 RNA-seq data captured over embryonic development. Additionally, evolutionary genome analysis

15 shows that fast developing organisms have a narrow genomic distribution of gene lengths while slower

16 developers have an expanded number of long genes. Our results support the idea that cell cycle

17 dynamics may be important across multicellular animals for controlling gene transcript expression and

18 cell fate.

19

20 **Keywords**

21 cell cycle, development, cell differentiation, gene regulation, computational model, single cell,

22 transcriptome, transcriptional filter, gene length

23

24

Introduction

A fundamental question in biology is how a single eukaryotic cell (e.g. zygote, stem cell) produces the complexity required to develop into an organism. A single cell will divide and generate many progeny, diversifying in a controlled and timely manner (Mueller et al., 2015) to generate cells with very different functions than the parent, all with the same genome (Wilmut et al., 1997). Many regulatory mechanisms coordinate this process, but much remains to be discovered about how it works (Zoller et al., 2018). Here, we explore how cell cycle regulation can control gene transcript expression timing and cell fate during tissue development.

The canonical view of the cell cycle is a timely stepwise process. Typically, the complete cell cycle is divided into four phases: first gap phase (G1), synthesis phase (S), second gap phase (G2) and mitotic phase (M). The length of each phase determines how much time a cell allocates for processes associated with growth and division. However, the amount of time that is spent in each phase frequently differs from one cell type to another within the same organism. For example, some cells experience fast cell cycles, especially in early embryogenesis. Organisms such as the fruit fly *(Drosophila melanogaster)* and the worm (*Caenorhabditis elegans*) exhibit cell cycle durations as short as 8-10 mins (Edgar et al., 1994; Foe, 1989). Cell cycle duration also changes over development (Figure 1 and Supplementary File 1). For example, it increases in mouse (*Mus musculus*) brain development from an average of 8 hours at embryonic day 11 (E11) to an average of 18 hours by E17 (Furutachi et al., 2015; Takahashi et al., 1995a).

Interestingly, cell cycle duration can act as a transcriptional filter that constrains transcription (Rothe et al., 1992; Shermoen and O'Farrell, 1991). In particular, if the cell cycle progresses relatively fast, transcription of long genes will be interrupted. In typical cells, the gene transcription rate is between 1.4-3.6 kb per minute (Ardehali and Lis, 2009). Thus, an 8 minute cell cycle would only allow transcription of the shortest genes, on the order of 10 kb measured by genomic length, including introns and exons, whereas a 10 hour cell cycle would allow transcription of genes as long as a megabase on the genome.

Cell cycle dependent transcriptional filtering has been proposed to be important in cell fate control (Bryant and Gardiner, 2016; Swinburne and Silver, 2008). Most multicellular eukaryotic animals start

embryogenesis with short cell cycle durations and a limited transcription state (O'Farrell et al., 2004) with typically short zygotic transcripts (Heyn et al., 2014). These cells allocate the majority of their cycle time to S-phase (synthesis), where transcription is inhibited (Newport and Kirschner, 1982a), and M-phase (division), with little to no time for transcription in the gap phases. However, as the cell cycle slows down, time available for transcription increases (Edgar et al., 1986; Newport and Kirschner, 1982a, 1982b), enabling longer genes to be transcribed (Djabrayan et al., 2019; Shermoen and O'Farrell, 1991; Yuan et al., 2016).

We asked what effects cell cycle dependent transcriptional filtering may have over early multicellular organism development. Through extensive mathematical simulations of developmental cell lineages, we identify the novel and unexpected finding that a cell cycle dependent transcriptional filter can directly influence the generation of cell diversity and can provide fine-grained control of cell numbers and cell type ratios in a developing tissue. Our computational model operates at single-cell resolution, enabling comparison to single-cell RNA-seq data captured over development, supporting our model by showing similar trends. Our model also predicts genomic gene length distribution and gene transcript expression patterns that are consistent with a range of independent data. Our work provides new insight into how cell cycle parameters may be important regulators of cell type diversity over development.

Results

Computational model of multicellular development

We model multicellular development starting from a single totipotent cell that gives rise to many progeny, each with its own transcriptome (Figure 2). We developed a single cell resolution agent-based computational model to simulate this process (see materials and methods). Each cell behaves according to a set of rules, and cells are influenced solely by intrinsic factors (e.g. number of genes in the genome, gene length, transcript levels and transcription rate). We intentionally start with a simple set of rules, adding more rules as needed to test specific mechanisms. Our analysis is limited to pre-mRNA transcript expression and we do not consider other gene expression-related factors, such as splicing, translation or gene-gene interactions. We also omit external cues (e.g. intercellular signaling or environmental gradients) to focus on the effects of intrinsic factors.

86   In our model, each cell is characterised by a fixed genome containing a set of G genes
87   $(gene_1, gene_2, ... gene_G),$ shown in Figure 2A. Each $gene_i$ is defined by a length, L_i (in kb), and in all
88   our simulations each gene is assigned a different length $(L_1 < L_2 < ... L_G)$. Since each $gene_i$ has a unique
89   length, L_i, we label genes by their length $(gene_i^{L_i} = gene^{L_i};$ e.g. $gene^3$ is a gene of length 3 kb). We
90   assume transcription time for $gene_i$ is directly proportional to its length, L_i. In the model, each $cell_j$ is
91   initialized with a cell cycle duration (Γ_{cell}) which represents the total time available for gene
92   transcription (see materials and methods). For example, we can initialise $cell_1$ with a three-gene genome
93   $(gene^1, gene^2, gene^3),$ where L = (1 kb, 2 kb, 3 kb) and a cell cycle duration Γ_1 of 1 hr. We fix
94   transcription rate, λ, to 1 kb/hr for all genes (though this assumption can be relaxed without changing
95   our results, Figure 3 and 3-figure supplement 1). As transcription progresses for all genes, $cell_1$ will only
96   express $gene^1$. Increasing cell cycle duration, Γ_{cell}, will allocate more time for transcription, allowing
97   longer genes to be transcribed. For example, if we initialise $cell_2$ with a cell cycle duration $\Gamma_2=3$ hrs, $cell_2$
98   will express all three genes, with time to make three copies of $gene^1$ (Figure 2B). We assume RNA
99   polymerase II re-initiation occurs along the gene, a distance Ω apart (Figure 3-figure supplement 1).

100

101   Once transcription is complete, the cell enters M-phase, during which it divides, and expressed
102   transcripts are randomly distributed to the two progeny cells (Figure 2C). This is the main stochastic
103   component in our model. We assume that transcription begins anew at the start of the cell cycle (i.e. all
104   transcripts from a gene that can't be finished in one cycle are eliminated), modeling the known
105   degradation of incomplete nascent transcripts in M-phase (Shermoen and O'Farrell, 1991). Relaxing our
106   assumption to include parental transcript inheritance and decay (Sharova et al., 2009), where a
107   proportion of inherited parental transcripts remain after each cell division does not change our overall
108   results (Figure 3-figure supplement 2). All individual cells and their transcriptomes are tracked over the
109   course of the simulation, enabling single cell resolution analysis. Transcriptomes are stored as vectors
110   containing the total number of transcripts per gene. For instance, $cell_2$ may have a transcriptome of
111   (3,1,1), indicating that three genes are expressed, with $gene^1$ expressed at three transcripts per cell and
112   the other two genes expressed at one transcript per cell (Figure 2D).

113

114   *Model prediction: Cell cycle duration influences transcript count - short genes generate more*
115   *transcripts than longer genes*

116   We begin by examining how a transcriptional filter impacts transcript counts, as controlled by cell cycle

117   duration. Shorter cell cycles will interrupt long gene transcription resulting in relatively high expression

118   of short gene transcripts and low expression of long gene transcripts. Our computational simulations

119   generate this expected pattern (Figure 3A). Each simulated cell transcriptome is divided into three bins

120   containing short, medium and long genes and then each bin is summarized with an average transcript

121   count. In simulations, bins with short genes exhibit the highest average transcript count levels. As cell

122   cycle duration increases, more cells show an increase in transcript count of longer genes; the trend is

123   consistent for various genome sizes and gene length distributions (Figure 3A and 3-figure supplement

124   3).

125

126   Single cell RNA-seq (scRNA-seq) has recently been used to profile mRNA expression of thousands of

127   cells for one cell type (microglia) across multiple species (Geirsdottir et al., 2019) or for multiple

128   embryonic developmental time points in one species, such as *Xenopus tropicalis* (Briggs et al., 2018)

129   and *Danio rerio* (Kimmel et al., 1995; Wagner et al., 2018), or tissue, such as mouse neural cortex

130   (Yuzwa et al., 2017). We analyzed these data in the same manner as our model (Figure 3B and 3-figure

131   supplement 4) and found that, in general, short genes have a higher mRNA expression level than longer

132   genes within a cell. Thus, gene mRNA expression patterns from a range of scRNA-seq data sets,

133   including developmental time courses, are compatible with our model prediction.

134

135   *Model prediction: Cell cycle duration can control cell diversity*

136   We next asked how three major model parameters (cell cycle duration, maximum gene length, and

137   number of genes in the genome) can influence the generation and control of cell diversity observed

138   during normal multicellular development. We conducted simulations for a single cell division step for

139   simplicity, but these were repeated thousands of times to model cell population effects. We compute cell

140   diversity in two ways; first, as the number of distinct transcriptomes in the cell population

141   (transcriptome diversity); and second, as the number of distinct transcriptomic clusters, as defined using

142   standard single cell transcriptomic analysis techniques (Satija et al., 2015) (see methods). Both measures

143   model real cell types and states that are distinguished by their transcriptomes, with transcriptome

144   diversity as an upper bound on cell type number, and cluster number approximating a lower bound. We

145   first ran simulations with an active transcriptional filter by varying only the cell cycle duration, Γ, for a

146   genome with 10 genes, with genes ranging in size from 1 to 10 kb, such that it satisfies $L_1 = 1 \leq$

147   …Γ…≤ L_G = 10. Short cell cycle duration parameter values generated a homogenous population of

148   cells because only short transcripts can be transcribed. As cell cycle duration was increased,

149   transcriptome diversity also increased. Longer cell cycle duration values generated heterogeneous

150   populations, because a range of transcripts can be expressed (Figure 4A, brown line). Interestingly, cell

151   cluster diversity peaks at intermediate cell cycle duration parameter values (Figure 4B, brown line; 4C),

152   because new genes are introduced with increasing cell cycle lengths, but eventually long cell cycles

153   provide sufficient time for cells to make all transcripts, which leads to reduced variance between the

154   progeny. We next repeat this experiment by turning off the transcriptional filter by reducing the

155   maximum gene length such that L_G < Γ, (Figure 4A,B, blue line). In this case, cell diversity can be

156   generated, but it quickly saturates (Figure 4B, blue line), as all transcripts are expressed, given a cycle

157   duration allowing the expression of the longest transcript. Thus, while cellular diversity can be generated

158   with an active or inactive transcriptional filter, diversity is more easily controlled by cell cycle duration

159   when the transcriptional filter is active.

160

161   In general, transcriptome diversity increases as a function of cell cycle duration (Γ), transcription rate

162   (λ), and number of genes in the genome (G). In particular, transcriptome diversity = $n\prod_{i=1}^{G}(T/L_i + 1)$,

163   where n is the genome ploidy level, $T = \sum_{a=0}^{\frac{Li}{\Omega}-1} f(a)$, $\forall\, f(a) \geq 0, f(a) = \Gamma * \lambda - \frac{a\Omega}{\lambda}$ (i.e. the maximum

164   transcribed gene length, T, is restricted by the product of cell cycle duration, Γ, transcription rate, λ and

165   RNA polymerase II re-initiation, Ω), and L_i is the length of gene$_i$. This analytical solution of cell

166   transcriptome diversity was validated by comparing it to simulations (Supplementary File 2). While the

167   number of genes and their length distribution can change over the course of evolution, these numbers are

168   constant for a given species, and transcription rate is likely highly constrained (Ardehali and Lis, 2009),

169   leaving only cell cycle duration as a controllable parameter of cell diversity during development,

170   according to our model.

171

172   *Model prediction: Varying cell cycle duration over developmental time controls tissue cell proportions*

173   *and number*

174   During multicellular organism development, it is essential to generate the correct numbers of cells and

175   cell types. Cell cycle duration changes dramatically during development, generally starting out fast to

176   generate cells quickly and slowing down over time as the organism matures (Supplementary File 1 and

177   Figure 1) (Farrell and O'Farrell, 2014; O'Farrell et al., 2004). Clearly, cells with short cell cycles

178   generate more progeny compared to those with longer cell cycles. However, we propose that a trade-off

179   exists, balancing the generation of diversity (longer cell cycle durations) with the fast generation of cells

180   (shorter cell cycle durations; **Figure 4**Figure 4B). To study this trade-off, we simulated cell propagation

181   under a "mixed lineage" scenario where, after the first division, one child cell and its progeny maintains

182   a constant cell cycle duration ($\Gamma_1 = 1$ hr) and the second child cell and its progeny maintains an equal or

183   longer constant cell cycle duration over a lineage with 20 cell division events (Figure 5**,** grey and blue

184   lineages, respectively). We initialize the starting cell with no prior transcripts (naïve theoretical state)

185   and a genome containing five genes ranging from length one to two kb (gene^1, gene$^{1.25}$, gene$^{1.5}$, gene$^{1.75}$,

186   gene^2), setting cell cycle duration in the second lineage to range between one and two, controlling the

187   transcriptional filter threshold in this lineage only. We considered three scenarios: 1) both cell lineages

188   cycle at the same rate (Fast-Fast, Figure 5A); 2) the first (blue) lineage is slower than the second (gray)

189   (Slow-Fast, Figure 5B); 3) both slow and fast lineages divide asymmetrically, producing one slow and

190   one fast cell (Slow-Fast, Figure 5C).

191

192   In the simulation where both cell lineages cycle at the same rate (Figure 5A), both lineages generate the

193   same number of progeny with the same level of diversity (Figure 5D). When cell cycle duration for the

194   second (blue) lineage is increased across simulations (Figure 5E), the transcriptional filter acts to

195   generate more diverse progeny, but with fewer cell numbers and progressively smaller population

196   proportions due to the slower cell cycle (Figure 5E, blue bars). Meanwhile, the short cell cycle lineage

197   maintains a steady, low level of diversity generation (Figure 5E, gray bars). When a fast cell can divide

198   asymmetrically, generating one slow and one fast cell at each division, the number of slow cells in the

199   population can increase, however this comes with a reduction of the number of fast cells in the

200   population (Figure 5F). Thus, our simulations show how the cell cycle duration parameter can impose a

201   trade-off between cell proportion and diversity generation, and mixing lineages with different cell cycle

202   durations can generate mixed cell populations each with their own diversity levels.

203

204   To more faithfully simulate multicellular animal development where cell cycle duration increases over

205   time, we next allowed progeny cells to differ in their cell cycle duration from their parents in each

206   generation (Figure 6A). Increasing the cell cycle duration over time reveals that cell cycle dynamics can

207   alter the number and proportions of cells as a function of time (cell generations; Figure 6B and 6-figure

208 supplement 1). To compare with a real system, we explore single cell transcriptomics data measured
209 over four timepoints of mouse cortex development (Yuzwa et al., 2017). Average cell cycle duration
210 over mouse neural cortex development is known to increase from 8 hours at embryonic day 11 (E11) to
211 an average of 18 hours by E17 (Furutachi et al., 2015; Takahashi et al., 1995a). Within this range,
212 progenitor cells are, in general, expected to be characterized by fast cycles with short G1 duration and
213 neurons by slower cell cycles with long G1 duration (Calegari et al., 2005). In our analysis of the mouse
214 cortex scRNA-seq data, we find genes with increasing transcript expression across the time course
215 (E11.5 < E13.5 < E15.5 < E17.5) are associated with neural developmental (maturing cell) pathways
216 whereas the genes with decreasing transcript expression across time (E11.5 > E13.5 >E15.5 > E17.5) are
217 associated with transcription and proliferation (stem and progenitor cell) pathways (Figure 6-figure
218 supplement 2). Furthermore, we observe an overall pattern of an increasing number of cells with long
219 cell cycle duration and a decrease in fast cycling cells (Figure 6C) following the same general trend as
220 observed in our simulations (Figure 6A), supporting the idea that cell cycle duration dynamics could
221 play a role in controlling cell proportions and cell diversity in a developing tissue.

222
223

224 *Hypothesis: A cell cycle dependent transcriptional filter may help control cell proportion and diversity*
225 *in tissue development*
226 Our theoretical model and agreement with general trends in scRNA-seq data supports the hypothesis that
227 a cell cycle dependent transcriptional filter has the potential to control cell proportion and diversity in
228 tissue development. In this section, we use the model to generate specific questions that can be checked
229 in real data, further supporting our model.

230

231 *Organismal Level*
232 Our model suggests that organisms with long genes will need to maintain long cell cycle durations
233 during development. Cell cycle duration measurements are not widely available which makes directly
234 testing this hypothesis difficult. Instead, we explored related questions. We started by asking if
235 organisms with longer genes would also take longer to develop. We analyze gene length distributions for
236 twelve genomes spanning budding yeast to human with a diverse range of developmental durations, as
237 shown in Figure 7 and Supplementary File 3 (Gilbert and Barresi, 2016; Jukam et al., 2017). Non-
238 mammalian species that we analyze are relatively fast developing, ranging from approximately two

239 hours (e.g. *Saccharomyces cerevisiae*) to a few days (e.g. *Xenopus tropicalis* and *Danio rerio*), while

240 mammals (*Mus musculus*, *Sus scrofa*, *Macaca mulatta*, and *Home sapiens*) are relatively slow

241 developing (20, 114, 168 and 280 days, respectively, Supplementary File 3). These species also have

242 different gene length distributions; to illustrate this quantitatively, using a typical transcription rate of

243 1.5 kb/min (Ardehali and Lis, 2009), a cell cycle duration of 1 hr can exclude up to 20% total genes

244 found in relatively slow developers and not exclude any genes in fast developers (Figure 7A). In

245 agreement with our hypothesis, the gene length distribution is narrower and left shifted (shorter genes)

246 for fast developers and broader and right shifted (longer genes) for slower developing species.

247 Interestingly, one seeming exception to the overall gene length distribution trend in multicellular

248 animals is the tunicate *Oikopleura dioica*, which has relatively short genes, but also has a rapid gestation

249 period of four hours to hatched tadpole (approximately twice as fast as *C. elegans* and six times faster

250 than *D. melanogaster*), supporting our analysis. Broadening this analysis to 101 species, we again find

251 an association (r=0.74) between estimated developmental time and median gene length (Figure 7B and

252 7-figure supplement 1).

253

254 Our model suggests that short genes will be enriched in pathways that can function independently from

255 long genes, and that long genes may be enriched in pathways related to mature, differentiated cell types

256 with slower cell cycles (Figure 8). We examined the functions of short and long genes by conducting a

257 pathway enrichment analysis on all genes in a genome ranked by their length. In the human genome, the

258 longest genes are enriched in processes such as, neural development, muscle control, cytoskeleton, cell

259 polarity and extracellular matrix and the shortest genes are enriched in processes that presumably need

260 to be quickly activated transcriptionally (e.g. immune, translation, and environment sensing; Figure 8-

261 figure supplement 1). We performed a similar pathway analysis for human (Figure 8 and 8-figure

262 supplement 2-3) and 12 other species (Figure 9) and found general agreement with these patterns,

263 finding the longest genes (gene length in the 95% quantile) enriched for genes involved in mature cell

264 related processes (e.g. brain and muscle development), whereas the shortest genes (gene length in the

265 5% quantile) are enriched for genes involved in core processes (e.g. immune, RNA processing, and

266 olfactory receptors).

267

268 *Spatial Level*

269 Within an organism, cell cycle duration and transcript expression vary across many factors, including

270 spatially. We hypothesize that spatial transcript expression patterns can be initially organized by gene

271 length. To explore this, we study the developing fruit fly embryo (*D. melanogaster*) where the average

272 cell cycle rates differ spatially (Foe, 1989). At the onset of cell cycle 14, cells in different embryo

273 regions start to divide at different rates, caused by an increase in their gap phase length, varying from 30

274 mins to 170 mins (Foe, 1989; Foe and Alberts, 1983). Cell cycle duration lengthening is spatially

275 organized, with anterior regions dividing faster than posterior regions, with the mid-ventral region being

276 the slowest (Figure 10). The embryo also exhibits spatial segregation patterns due to Hoxd gene family

277 transcript expression (Mallo and Alonso, 2013). Overlaying the spatial patterns of hox gene family

278 transcript expression and cell cycle duration obtained from independent studies, we observe that fast

279 cycling regions express the shortest hox genes (Dfd 10.6kb, lab 17.2kb) and slow cycling regions

280 express the longest hox genes (Ubx 77.8kb and Antp 103.0kb)(Foe, 1989; Lemons and McGinnis,

281 2006) in agreement with our model.

282

283 **Discussion**

284 How cellular processes support the carefully orchestrated timing of tissue development that results in a

285 viable multicellular organism is still unclear. While a combination of many potential cell autonomous

286 and non-autonomous mechanisms, such as cytoplasmic molecules and gradients, cell-cell

287 communication, microenvironment signals, and effective cell size (Edgar et al., 1986; Mukherjee et al.,

288 2020; Tabansky et al., 2013; Yoon et al., 2017), are likely important, one hypothesis is that gene length

289 can be used as a mechanism to control transcription time in this process (Artieri and Fraser, 2014; Gubb,

290 1986; Keane and Seoighe, 2016; Swinburne et al., 2008). Bryant and Gardiner further hypothesize that

291 cell cycle duration may play a role in filtering genes that influence pattern formation and regeneration

292 (Bryant and Gardiner, 2018; Ohsugi et al., 1997), as cell cycle lengthens over development (Figure 1

293 and Supplementary File 1) (Foe, 1989; Foe and Alberts, 1983; Newport and Kirschner, 1982b;

294 Takahashi et al., 1995b). Early experiments using embryos suggested that cell cycle duration has a role

295 in transcription initiation, however these experiments lacked the temporal resolution necessary to

296 dissociate the effects of cell cycle duration and transcriptional control from other mechanisms (Edgar et

297 al., 1986, 1994; Kimelman et al., 1987; Newport and Kirschner, 1982b, 1982a). It is also well known

298 that cell cycle length changes can control cell fate and development (Coronado et al., 2013; Mummery et

299 al., 1987; Pauklin and Vallier, 2013; Singh et al., 2013), however, this has remained observational and

300 not linked to a mechanism. To help address these limitations, we developed an *in silico* cell growth

301 model to directly study the relationship between cell cycle duration and gene transcription in a

302 developmental context. The new discovery we make is that a transcriptional filter can be controlled by
303 cell cycle duration and used to simultaneously control the generation of cell diversity, the overall cell
304 growth rate and cellular proportions during development (defining an emergent property of our
305 computational model – see Appendix). Genomic information (gene number and gene length distribution)
306 and cell cycle duration are critical parameters in this model. Across evolutionary time scales, cell
307 diversity can be achieved by altering gene length (Keane and Seoighe, 2016), however, in terms of
308 developmental time scales, we propose that cell cycle duration is an important factor that may control
309 cell diversity and proportions within a tissue.

310

311 We predict that increasing the gene length distribution across a genome over evolution can provide more
312 cell cycle dependent transcriptional control in a developing system, leading to increased cellular
313 diversity. Examining a range of genomes and associated data provides support for this novel idea. We
314 observe that fast developing organisms have shorter median gene lengths relative to the broad
315 distributions, including many long genes, exhibited by slow developers (mammals). This aspect of
316 genome structure may help explain the observed rates of cell diversity and organism complexity, as
317 measured by number of different cell types, over a wide range of species, Figure 7-figure supplement 1
318 (Valentine et al., 1994; Vogel and Chothia, 2006).

319

320 While we hypothesize that a cell cycle dependent transcriptional filter is a fundamental regulatory
321 mechanism operating during development (because gene length is fixed in the genome and transcription
322 rate is expected to lie in a narrow range), multiple other regulatory mechanisms could modulate its
323 effects. Furthermore, exploring these mechanism may even result in similar conclusions, as it can be
324 evolutionary advantageous to have multiple paths to the same outcome; These include, but are not
325 limited to, silencing or deactivating genes, gene regulatory networks, blocking gene clusters, e.g. Hoxd
326 (Rodríguez-Carballo et al., 2019) changing the transcription or re-initiation rate of RNA polymerase II
327 (Figure 3-figure supplement 1), or inheriting long transcripts maternally at the zygote stage (Figure 3-
328 figure supplement 2). Our current model only explores the effects of transcription and re-initiation rates
329 of RNA polymerase II, mRNA transcript degradation rates, and maternally introduced transcripts. For
330 the latter mechanism, we expect longer transcripts to be major contributors during the early maternal
331 phase (Jukam et al., 2017), which agrees with zebrafish (*D. rerio*) experiments showing that maternal
332 transcripts are longer and have evolutionary conserved functions (Heyn et al., 2014). Indeed, if we add

333 maternal transcript inheritance to our model, we see the same pattern of a small number of long
334 transcripts present early, as expected (Figure 3-figure supplement 2). Future work would entail curating
335 experimental data about more regulatory mechanisms in cell systems and testing their association with
336 cell cycle duration.

337

338 Our analysis raises interesting directions for future work. We focus on development, but transcriptional
339 filtering may be important in any process involving cell cycle dynamics, such as regeneration (Bryant
340 and Gardiner, 2018), wound repair, immune activation and cancer. We must also more carefully
341 consider cell cycle phase, as transcription mainly occurs in the gap phases (Bertoli et al., 2013; Newport
342 and Kirschner, 1982b). Experiments indicate that a cell will have different fates depending on its phase
343 (Dalton, 2013; Pauklin and Vallier, 2013; Vallier, 2015). This agrees with our model, as a cell at the
344 start of its cell cycle will have a different transcriptome in comparison to the end of the cell cycle.
345 Induced pluripotent cell state is also associated with cell cycle phases (Dalton, 2015) and efficient
346 reprogramming is only seen in cell subsets with fast cell cycles (Guo et al., 2014). Our model could
347 explain these observations, as slower cycling cells could express long genes that push a cell to
348 differentiate rather than reprogram. However, our model is limited to total transcription duration for
349 interphase (G1, S and G2), thus a future direction would be to explore different durations for each cell
350 cycle phase. Collecting more experimental data about cell phase in developing systems will help explore
351 these effects. Further, it will be important to explore how cell cycle duration is controlled. Molecular
352 mechanisms of cell cycle and cell size (Liu et al., 2018) control could be added to our model to provide
353 a more biochemically realistic perspective on this topic. Ultimately, a better appreciation of the effects
354 of cell cycle dynamics will help improve our understanding of a cell's decision-making process during
355 differentiation, and may prove useful for the advancement of tools to control development, regeneration
356 and cancer. Finally, it is important to note that we have not provided experimental model support, only
357 analyses that do not disagree with model predictions. We have also not proven the generality of the
358 results across species. However, we hope the hypotheses we explore here motivate new experimental
359 studies to directly test the validity and generality of our model.

360

361 **Materials and Methods**

362 *Mathematical model*

363   Our mathematical model is agent and rule-based. A single cell behaves and interacts according to a fixed

364   set of rules. Our major rule involves a gene-length mechanism, where each cell is defined by a genome

365   and a cell cycle duration. The cell cycle duration determines which gene transcripts are expressed within

366   the cell, based on the transcription rate. All decisions are based on a cell's autonomous information and

367   we deliberately choose to consider this simple baseline setup to clarify the

368   contribution of cell cycle duration to overall cell population growth.

369

370   Each cell is defined by a genome G (containing a set of genes), cell cycle duration in hours and the

371   transcripts inherited or recently transcribed. In the genome, each gene is defined by a length, $gene^{Length}$.

372   For example, in a genome with three genes, ($gene^1$, $gene^2$, $gene^3$) represents genes of length 1, 2, 3 kb,

373   respectively.

374

375   Each cell can divide and make two progeny cells. This process can continue many times to simulate the

376   growth of a cell population and we keep track of the entire simulated cell lineage. For each cell division

377   (one time step in the simulation): each $Cell_i$ will transcribe its genes based on the time available, defined

378   by the cell cycle duration. We assume the time it takes to transcribe a gene depends on its length and a

379   fixed transcription rate, although a simplification, there are examples where this occurs, for instance the

380   human dystrophin gene is 2241765bp long and takes about 16 hrs to transcribe (Tennyson et al., 1995).

381   Once a cell cycle is finished, the cell divides. When cells are synchronized, the first cell division T= Γ_i.

382   When the cells are asynchronized then the algorithm identifies the time allocated as the shortest cell

383   cycle duration in the population as the time step and each cell division will have a different duration. In

384   this case, we keep track of the exact duration such that cells with short cell cycles, for example Γ=1 hr,

385   will register 10 divisions in 10 hrs while cells with long cell cycles, for example Γ=10 hrs, will register 1

386   division in over the same time. We limited the model to two modes of division, symmetric (where the

387   cell gives rise to identical cells, e.g. Figure 5A) and asymmetric (where the cell gives rise to a fast and

388   slow cell, e.g. Figure 5C). We do not consider mechanisms that reduce cell numbers (cell death). For

389   certain experiments (e.g. Figure 6), the cell cycle duration for each progeny is allowed to diverge from

390   the parental duration using a monotonic function (increasing or decreasing) and a stochastic variable

391   based on a Gaussian distribution with a mean equal to Γ_i (parental cell cycle duration). This models a

392   more realistic noisy distribution of cell cycle durations in the simulated cell population. The cell cycle

393 and division rules are repeated for all cells in the population until a set number of cell divisions have
394 been reached.

395

396 During a cycle, each cell contains a certain number of transcripts. The number of transcripts for each
397 gene is calculated by a function of cell cycle duration, Γ, transcription rate, λ, re-initiation distance, Ω,
398 and gene length, L: $\sum_{a=0}^{\frac{\text{gene}^{L_i}}{\Omega}-1} \Gamma * \lambda - \frac{a\Omega}{\lambda}$. If the cell does not divide, then the number of transcripts
399 reflects the current cell cycle phase, which is computed and stored. If the cell can divide within the time
400 T=$(\Gamma * \lambda)$, then it will randomly, according to a uniform distribution, assign its transcripts between its
401 two progeny cells. Typically simulations were conducted with $\lambda = 1$, simplifying the analysis to
402 $(\Gamma - a\Omega)/\text{gene}^{L_i}$, however, we also explored the effects of transcript re-initiation and transcription rate
403 on the system as shown in Figure 3-figure supplement 1.

404

405 Our model tracks single cells, with each cell identified by a transcriptome and cell cycle duration. The
406 transcriptome data resemble a single cell RNA-seq matrix to aid comparison between simulation and
407 experimental data. We allow cells without any transcripts e.g. (0,0,0) to exist – due to the low numbers
408 of genes considered in our simplified model and results, and that parental transcripts are distributed
409 between progeny, there is a probability of 2/(the total number of transcripts) that all the transcripts will
410 end up in only one of the new cells, leaving the other one empty (Zhou et al., 2011). Theoretically we
411 have no reason to omit these cells and they may represent the most naïve theoretical state of a cell
412 without any prior information. Early embryos, such as in xenopus stages that lack zygotic transcription,
413 may be similar real systems to such a state (Newport and Kirschner, 1982b).

414

415 Parameters tracked for each cell_i= (number of divisions, current cell cycle phase, current time in cell
416 cycle, length until next division, relative time passed, total cell cycle duration, transcriptome list, cell
417 name, and lineage history). All cells are set with the same genome, ploidy level and RNA polymerase II
418 transcription rate and RNA polymerase II re-initiation distance.

419

420 Our model was developed and simulated using Mathematica (Wolfram-Research, 2017).

421

422 *Quantification and statistical analysis*

423   *Gene length analysis*

424   All protein coding genes were downloaded from Ensembl genome database version 95 or 100 (Yates et

425   al., 2016) using the R (3.6.1) Biomart package version 3.10 (Durinck et al., 2005). The length of each

426   gene was calculated using start_position and end_position for each gene, as extracted from the Ensembl

427   database (Yates et al., 2016).

428

429   *Single cell analysis pipeline*

430   Simulated data sets were pre-processed and clustered in R using the standard workflow implemented in

431   the Seurat package version 3.1.2 (Satija et al., 2015). We used default parameters unless otherwise

432   stated. Data were log-normalized and scaled before principal component analysis (PCA) was used to

433   reduce the dimensionality of each data set. Due to the small number of simulated genes in our

434   experiments, the maximum number of PCs (one fewer than the number of genes dims=1:3) was

435   calculated and used in clustering. FindVariableFeatures was used with loess.span set to 0.3 unless the

436   number of genes were less than 5, then (0.4, 0.7 and 1 was used for simulations with 4, 3 and 2 genes,

437   respectively). Cells were clustered using a shared nearest neighbor (SNN)-based 'Louvain' algorithm

438   implemented in Seurat with reduction set as "pca". The clustering resolution was set to 1 for all

439   experiments, and all calculated PCs were used in the downstream clustering process using the Louvain

440   algorithm accessed via Seurat. Data was visualized with t-SNE after clustering.

441

442   *Developmental time curation*

443   Estimated developmental time was curated from encyclopedia of life or PubMed accessible articles

444   (Supplementary File 3). We used gestation time for mammals and hatching time for species who lay

445   eggs (since it is difficult to accurately define a comparative stage for all species). Species were grouped

446   based on their taxonomic class and their developmental time was estimated by calculating the average

447   number of days from zygote to birth or hatching.

448

449   *Pathway Enrichment Analysis*

450   We used Gene Set Enrichment Algorithm (GSEA version 4.0.2), in pre-ranked analysis mode, to

451   identify pathways enriched among all genes in a genome ranked by gene length (Subramanian et al.,

452   2005). Gene ranks started at (number of genes)/2 to its negative equivalent and were normalised such

453   that we generated a ranked list from 1 to -1, with 1 specifying the shortest gene and -1 the longest. The

454 ranked gene length list was analyzed for pathway enrichment GSEA with parameters set to 1000 gene
455 set permutations and gene set size between 15 and 200. Pathways used for the analysis were from Gene
456 Ontology biological process (Ashburner et al., 2000), MSigDB c2 (Ashburner et al., 2000),
457 WikiPathways (Slenter et al., 2017), Panther (Mi et al., 2005), Reactome (Croft et al., 2011), NetPath
458 (Kandasamy et al., 2010), and Pathway Interaction database (Schaefer et al., 2009) downloaded from the
459 Bader lab pathway resource (http://baderlab.org/GeneSets). An enrichment map, created using the
460 EnrichmentMap Cytoscape app version 3.3.0 (Merico et al., 2010), was generated using Cytoscape
461 (version 3.8.0) using only enriched pathways with p value of 0.05 and FDR threshold of 0.01 (Reimand
462 et al., 2019). Cross talk (shared genes) between pathways was filtered by Jaccard similarity greater than
463 0.25. Pathways were automatically summarized using the AutoAnnotate App to assign pathways to
464 themes (Kucera et al., 2016). Themes were further summarized by grouping pathways into more general
465 themes with a mixture of automatic classification using key words and manual identification.
466
467 *Pathway word cloud analysis*
468 All Gene Ontology pathway (GO biological processes) downloaded from the Ensembl genome database,
469 version 100, (Yates et al., 2016) using the R Biomart package version 3.5 (Durinck et al., 2005). We
470 restricted analysis to pathways with at least three genes. We grouped genes based on their gene length
471 (see *Gene length analysis* for details) and identified the pathways associated with each gene. The
472 description of each pathway was collected and the frequency of each word within the pathway name was
473 calculated. We defined themes (Supplementary File 5-6) for all *Homo sapiens* available pathways (using
474 only GO biological processes). Common, generic and uniformly distributed themes (such as cellular-
475 response, metabolic-biosynthesis, protein-processes, signalling, and transcription) were manually
476 removed from the list. The frequencies were visualised as word clouds using Mathematica (Wolfram-
477 Research, 2017).
478
479 ***Data and code availability***
480 Our simulation code is available at https://github.com/BaderLab/Cell_Cycle_Theory
481
482 ***Key Resources Table***

Key Resources Table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	(Wolfram-Research, 2017)			Mathematica (Wolfram Research Inc, Mathematica Versions 11.0-12, Champlain, IL, 2017) http://www.wolfram.com/mathematica/
Software, algorithm		This paper		Cell developmental model https://github.com/BaderLab/Cell_Cycle_Theory
Software, algorithm	(Satija et al., 2015)	PMID: 25867923		Seurat (3.1.2) https://satijalab.org/seurat/
Software, algorithm	(Yates et al., 2016)	PMID: 26687719		Ensembl (95) and (100) https://useast.ensembl.org/index.html
Software, algorithm	(Ashburner et al., 2000)	PMID: 10802651		Gene Ontology http://geneontology.org/
Software, algorithm	(Durinck et al., 2005)	PMID: 16082012		BioMart (3.10) http://useast.ensembl.org/biomart/martview/
Software, algorithm	(Merico et al., 2010)	PMID: 21085593		Enrichment Map software (3.3.0) https://www.baderlab.org/Software/EnrichmentMap
Software, algorithm	(Kucera et al., 2016)	PMID: 14597658		AutoAnnotate App https://baderlab.org/Software/AutoAnnotate
Software, algorithm	(Shannon et al., 2003)	PMID: 14597658		Cytoscape (3.8.0) https://cytoscape.org/

Software, algorithm	(Reimand et al., 2019)	PMID: 30664679		Baderlab pathway resource (updated June 01,2020) http://download.baderlab.org/EM_ Genesets/

Appendix

Why is cell cycle duration changing?

While defining a general mathematical representation of cell cycle kinetics for a developing system, we assembled available cell cycle length measurements from published studies for various species and tissues. Figure 1 shows measurements obtained from *Mus musculus*. For other data see Supplementary File 1. The data motivated us to ask "why is cell cycle duration changing over development?" and propose that changes in cell cycle duration can be used to guide the progression of cell development.

Theoretically we devised a simple model that can test this idea by assuming:

- Cell cycle duration can change across developmental time
- Gene length distribution is constant among all cells in the same organism, such that we can denote the length by L
- The difference in cell cycle can affect the time a cell spends transcribing genes
- All active genes are transcribed and transcription rate is constant in a cell

The novel aspect of our work is the proposal that a cell cycle dependent transcriptional filter can control cellular diversity within a tissue over development. However, some of the concepts that we build on are known and are recognized in the community to varying degrees. We bring these together for the first time to support the model and generate predictions. In particular, we list these concepts below and clarify our novel contribution.

Prior contributions:

- Cell cycle lengthens over development
  - "The Xenopus embryo undergoes 12 rapid synchronous cleavages followed by a period of slower asynchronous divisions more typical of somatic cells." after which the cell cycle duration continues to increase. (Newport and Kirschner, 1982) https://pubmed.ncbi.nlm.nih.gov/6183003

- In D. melanogaster early development, the first 10 cell divisions are fast and synchronous, then progressively increase in cell cycle duration. (Foe, 1989; Foe and Alberts, 1983)
  https://pubmed.ncbi.nlm.nih.gov/6411748

  https://pubmed.ncbi.nlm.nih.gov/2516798

- The cell cycle lengthens during *Mus musculus* brain development. "The length of the cell cycle increases from 8.1 to 18.4 hr, which corresponds to a sequence of 11 integer cell cycles over the course of neuronal cytogenesis in mice. The increase in the length of the cell cycle is due essentially to a fourfold increase in the length of G1 phase which is the only phase of the cell cycle which varies systematically." (Takahashi et al., 1995)
  https://pubmed.ncbi.nlm.nih.gov/7666188

- We also compiled cell cycle duration from 25 papers, which clearly support this statement (see Figure 1and Supplementary File 1).
- Gene length controls transcription timing. Short cell cycles limit transcription and long cell cycles allow transcription of longer genes
  - Cell cycle duration can limit transcripts based on their size
    - Short cell cycles can constrain transcription in D. melanogaster. "the length of mitotic cycles provides a physiological barrier to transcript size, and is therefore a significant factor in controlling developmental gene activity during short 'phenocritical' periods." (Rothe et al., 1992)
      https://pubmed.ncbi.nlm.nih.gov/1522901
  - Zygotic transcripts are encoded by short genes and start being expressed when cell cycle lengthens.
    - "We propose that early development in Drosophila operates according to a hierarchy of events. The first 13 division cycles are driven by a maternal mechanism which responds to the increasing nuclear density by extending the interphase periods of successive cycles. This lengthening of interphases allows transcriptional activation, and the expression of new zygotic gene products triggers events such as cellularization of the blastoderm, gastrulation, and further rounds of mitosis." (Edgar et al., 1986)
      https://pubmed.ncbi.nlm.nih.gov/3080248
    - *Danio rerio* zygotic transcript lengths are shorter than maternally provided ones; The earliest zygotic genes are without introns. (Heyn et al., 2014; Kwasnieski et al., 2019; Shermoen and O'Farrell, 1991)
      https://pubmed.ncbi.nlm.nih.gov/1680567
      https://pubmed.ncbi.nlm.nih.gov/24440719
      https://pubmed.ncbi.nlm.nih.gov/31235656
  - Longer genes, with larger introns, take longer to transcribe ("intron delay")
    - Intron delay and transcriptional timing can affect development. (Artieri and Fraser, 2014; Gubb, 1986; Swinburne and Silver, 2008)
      DOI:10.1002/dvg.1020070302
      https://pubmed.ncbi.nlm.nih.gov/18331713
      https://pubmed.ncbi.nlm.nih.gov/2506953
- Cell-cycle dependent transcriptional filter is a mechanism for gene transcript expression regulation

558   o   Hypothesized in (Bryant and Gardiner, 2016), but no analysis or experimental data to
559   support this statement is provided in this publication.
560   •   Cell cycle length changes can control cell fate and development
561   o   In cell lines
562   ▪   Differentiation can be induced in G1-phase isolated pluripotent embryonal
563   carcinoma cells by treating with retinoic acid (RA) while other cell cycle phases
564   are not RA stimulated. (Mummery et al., 1987)
565   https://pubmed.ncbi.nlm.nih.gov/2883052
566   ▪   "A short G1 phase is an intrinsic determinant of naïve embryonic stem cell
567   pluripotency" (Coronado et al., 2013)
568   https://pubmed.ncbi.nlm.nih.gov/23178806
569   ▪   "The cell-cycle state of stem cells determines cell fate propensity" (Pauklin and
570   Vallier, 2013)
571   https://pubmed.ncbi.nlm.nih.gov/24074866
572   ▪   Embryonic stem cells are more responsive to differentiation signals in G1 than in
573   other phases of the cell cycle. (Singh et al., 2013)
574   https://pubmed.ncbi.nlm.nih.gov/24371808
575   o   In an organism
576   ▪   Alteration of cell cycle length can cause changes in *Gallus gallus* limb pattern.
577   Gene transcripts normally expressed in mesenchyme cells are sensitive to cell
578   cycle length. (Ohsugi et al., 1997)
579   https://pubmed.ncbi.nlm.nih.gov/9281333
580   •   Transcription rate and duration
581   o   Transcription elongation rate is about 1.4kb/min
582   ▪   Transcript elongation rates tend to be uniform within a cell type. (Ardehali and
583   Lis, 2009)
584   https://pubmed.ncbi.nlm.nih.gov/19888309
585   ▪   Transcription of human dystrophin gene requires 16 hours (Tennyson et al. 1995)
586   https://pubmed.ncbi.nlm.nih.gov/7719
587   o   Transcription is repressed in S phase.
588   ▪   Early evidence that transcription is repressed in synthetic phase (S). (Newport and
589   Kirschner, 1982b)
590   https://pubmed.ncbi.nlm.nih.gov/7139712
591   ▪   "Upon G1–S transcriptional activation, cells progress to S phase, initiate DNA
592   replication and subsequently inactivate transcription." (Bertoli et al., 2013)
593   https://pubmed.ncbi.nlm.nih.gov/23877564
594

595   Our novel contributions

596   •   Our main novel claim: We are the first to link cell cycle duration to control of cell diversity and
597   proportions of cells in tissues
598   •   We are the first to support the idea that a cell-cycle dependent transcriptional filter is a
599   mechanism for gene transcript expression regulation that affects development using quantitative
600   modeling
601   •   First to link gene length distribution in genomes of multiple species to length of organism
602   development

- First to show major functional differences between short and long genes in animal genomes
- Our single cell transcriptomic mathematical model is novel and shared as a community resource

Acknowledgements

Declaration of Interests

The authors declare no competing interests

Figure Legends

Figure 1: Cell cycle duration changes during mouse development. The data was curated from several publications (PubMed identifiers: 5859018, 14105210, 5760443, 5542640, 4041905, 7666188, 12151540, 18164540), shown in the legend as authors and (year). For other species and tissues see Supplementary File 1.

Figure 2: A novel mathematical model of cell lineage generation. A) a single cell is defined by a given number of genes in its genome as well as their gene lengths (e.g. three genes, $gene^1 < gene^2 < gene^3$). Cell cycle duration defines the time a cell has available to transcribe a gene. B) For example, a cell with cell cycle duration = 1 hour will only enable transcription of $gene^1$; cell cycle duration = 2 hours will enable transcription of $gene^1$ and $gene^2$; cell cycle duration = 3 hours enables transcription of all three genes. C) Our model assumes transcripts passed from parental cell to its progeny will be randomly distributed during division (M-phase). D) Each cell is characterized by its transcriptome, represented as a vector.

Figure 3: Short genes produce more transcripts than longer genes at multiple cell cycle duration lengths. The transcriptome for each cell is subdivided into short, medium, and long gene bins and transcript counts are averaged per bin per cell. A) Simulations predict that short gene transcripts will be more highly expressed than long gene transcripts, irrespective of the genome size. The top panel shows the total number genes expected from each genome per bin. Simulation results are shown for cell cycle

634 durations of 1, 5 and 10 hours and gene lengths (gene^{L}$_1$$^{-L}$$_{10}$), see Figure 3-figure supplement 3 for

635 additional simulations (other parameters ploidy=1, one cell division, iterations = 5000000, genome G =

636 10, gene^{L}$_1$$^{-L}$$_{10}$, transcription rate, λ = 1 kb/hr, RNA polymerase II re-initiation, $\Omega = 0.25$kb). Bins are

637 defined such that genes are evenly distributed across them. B) Single cell microglia data obtained from

638 GSE134707 (Geirsdottir et al., 2019) displaying expected patterns where short genes (lengths <10 kb)

639 have a higher transcript expression than both medium genes (lengths > 10kb) and longer genes (lengths

640 >25 kb) -- Kolmogorov-Smirnov test p $<10^{-16}$, the upper bound p-value for all short-medium and short-

641 long comparisons -- across nine different species (age): *Macaca fascicularis* (3 years), *Callithrix jacchus*

642 (7 years), *Mus musculus* (8-16 weeks), *Rattus norvegicus* (11-14 weeks), *Mesocricetus auratus* (8-16

643 weeks), *Nannospalax galili* (2-4 years), *Ovis aries* (18-20 months), *Gallus gallus* (24 weeks) and *Danio*

644 *rerio* (4-5 months). The top part of the plot shows the total number genes possible in each bin, given the

645 gene length distribution of each genome. Bins are defined such that they are both consistent across all

646 species and also approximately evenly filled with genes.

647 **Figure 4: Cell cycle duration can control cell diversity.** Simulations explore the effects of cell cycle

648 duration, Γ, gene number, G, and gene length distribution. A) Simulations show that cell diversity

649 (transcriptome diversity) increases as a function of cell cycle duration. Short cell cycle durations can

650 constrain the effects of gene number as long as a transcriptional filter is active (gene length distributions

651 are broad, $L_1 < ... (\Gamma * \lambda) ... < L_G$). When $L_G < (\Gamma * \lambda)$, cell cycle duration does not control cell

652 diversity. Cell cycle duration effects are relative to the gene length distribution in the genome. B) We

653 use Seurat to cluster the simulated single cell transcriptomes (10,000 cells) using default parameters and

654 report the number of cell clusters over the simulations. This shows that cell diversity increases with gene

655 number but the number of clusters identified decreases when all the gene transcripts can be expressed

656 similarly among all cells. C) Representative examples (10,000 cells) of t-SNE visualizations (RunTSNE

657 using Seurat version 3.1.2) are shown for simulations with cell cycle durations 2, 6 and 10 hours

658 (genome G = 10, gene^{L}$_1$$^{-L}$$_{10}$, ploidy n = 1 and transcription rate, λ = 1 kb/hr, RNA polymerase II re-

659 initiation, $\Omega = 0.25$kb).

660

661 **Figure 5: Cell cycle duration can control the generation of cell proportions and cell types within a**

662 **population.** Simulations start with two cells and run for 18 divisions (generating 2^{19} cells when cell

663 cycles are the same). Cell$_1$ is initialized with cell cycle duration Γ_1 = 1 hour, Cell$_2$ has cell cycle

664   duration, Γ_2, ranging from 1 to 2 hours. All progeny are tracked based on their cell cycle duration

665   (lineage $\Gamma_1 = 1$ cell cycle duration, gray, or lineage Γ_2 cell cycle duration, blue). Tree plot depicting

666   lineages when the cell cycle duration A) are the same, $\Gamma_1 = \Gamma_2$ (scenario 1), or B-C)differ, $\Gamma_1 < \Gamma_2$

667   (scenarios 2 and 3). Scenario 2 captures a situation when the cell cycle is determined by the parental

668   lineage, while scenario 3 captures a situation when a cell splits asymmetrically into a fast and slow cell,

669   resulting with the fast lineage having just one cell. D-F) Müller visualizations show that when the cell

670   cycle duration is the same, both cells contribute the same number of progeny and cell proportions (%)

671   are 50:50 (bottom left panel). The visualization is stacked, down-scaling the blue lineage slightly to

672   reduce occlusion of the grey lineage. Cells with longer cell cycle duration (blue lineage) generate fewer

673   progeny with respect to the cells with a short cell cycle duration of 1 hour (gray lineage). However, the

674   slower cells contribute more to the diversity observed in the population, shown as the blue and grey

675   transcriptome diversity bars. Thus, increasing cell cycle duration increases cell diversity, but also limits

676   the number of progeny generated. The system can overcome the limit on cell number by using scenario

677   3, where more slow cells can be generated (other parameters G=5, gene lengths (gene$^{L}_1$-$^{L}_2$),

678   genome={1,1.25,1.5,1.75,2} and ploidy =1, RNA polymerase II re-initiation, $\Omega = 0.25$kb).

679

680   **Figure 6: Varying cell cycle duration across time affects cell type proportions.** A) Cell cycle

681   duration increases after each cell division, with amount of increase defined using a Gaussian

682   distribution. B) Simulation of gradually increasing cell cycle duration over time, such that Γ=Gaussian

683   (mean Γparent $\pm$ 6, standard deviation σ=0.06), affects the relative proportion of cells with different cell

684   cycle durations (pie charts). All cell progeny are labeled based on their cell cycle duration (inherited

685   from parent). See Figure 6-figure supplement 1 for results using other increment rates. Parameters:

686   genome=10, gene lengths (gene$^{L}_1$-$^{L}_{10}$), λ=1 kb/hr, 18 cell divisions, iterations= 500, ploidy n=1, , RNA

687   polymerase II re-initiation, $\Omega = 0.25$kb. C) Single cell transcriptomics data from GSE107122 (Yuzwa et

688   al., 2017) for embryonic mouse cortex development, known to exhibit increasing cell cycle duration

689   over time. This data includes identified cell types, is a time series, and we know the average cell cycle

690   duration at each time point; At E11.5 the average cell cycle duration is 8 hrs and by E17.5 it is 18 hrs

691   (Furutachi et al., 2015; Takahashi et al., 1995a). Cells were defined as relatively fast cycling cells

692   (apical progenitors), relatively medium cycling (intermediate progenitors) and relatively slow cycling

693   (neurons), with cell type annotation based on cell clustering analyses conducted in (Yuzwa et al., 2017).

694   We show how cell proportions (pie charts) change across time, with apical progenitors (relatively fast

695   cycling cells) decreasing in frequency as the average cell cycle duration increases.

696

697   **Figure 7: Gene length distribution and developmental time are correlated.** A) Model organisms

698   exhibit a large diversity in gene length distributions over their genomes. Species that have narrower gene

699   length distributions tend to develop faster, while slow developers (mammals) exhibit broad and right

700   shifted gene length distributions. Demarcating a 1 hour cell cycle duration using an average transcription

701   rate of 1.5 kb/min illustrates the proportion of genes that would be interrupted before transcript

702   completion for each organism. *Saccharomyces cerevisiae* (budding yeast), *Caenorhabditis elegans*

703   (worm), *Drosophila melanogaster* (fruit fly), *Oikopleura dioica* (tunicate), *Danio rerio* (zebrafish),

704   *Takifugu rubripes* (fugu), *Xenopus tropicalis* (frog), *Gallus gallus* (chicken), *Mus musculus* (mouse),

705   *Sus scrofa* (pig), *Macaca mulatta* (monkey), *Homo sapiens* (human). B) There is a clear positive

706   correlation between developmental time and median gene length (101 species, Figure 7-figure

707   supplement 1). Estimated developmental time was curated from encyclopedia of life or articles found in

708   PubMed (Supplementary File 3). We used gestation time for mammals and hatching time for species

709   who lay eggs (since it is difficult to accurately define a comparative stage for all species). We analyzed

710   the data using a Pearson correlation test, shown as r. For each species we calculated median gene length:

711   All protein coding genes were downloaded from Ensembl version 95 (Yates et al., 2016) using the R

712   Biomart package (Durinck et al., 2009, 2005). The length of each gene was calculated using

713   start_position and end_position for each gene as extracted from Ensembl data.

714

715   **Figure 8: Short genes and long genes participate in different pathways.** The plot shows the *H.*

716   *sapiens* gene length distribution. We selected the shortest 5% quantile as a list of short genes and the

717   95% quantile as a list of long genes. Short genes < 1.6Kb (n=1124) are involved in immune defense,

718   environment-sensing, and olfactory, and long genes >243kb (n=1125) are represented in processes

719   involving muscle and brain development, as well as morphogenesis. For each gene group we identified

720   all corresponding Gene Ontology (Ashburner et al., 2000) biological process terms downloaded from the

721   Ensembl genome database version 100 (Yates et al., 2016), grouped the terms into themes

722   (Supplementary File 5-6) and visualized the resulting term frequencies as word clouds using

723   Mathematica. Refer to Figure 8-figure supplement 2-3 for a more detailed analysis of the themes across

724   all gene groups.

725

Figure 9: Short genes exhibit different pathways than long genes and this trend is consistent across a wide species range. We selected the shortest 5% quantile as a list of short genes (top panels in blue) and genes above the 95% quantile to define a list of long genes (bottom panels in gray. *Saccharomyces cerevisiae* (short<0.24kb long>3.5kb), *Ashbya gossypii* (short<0.36kb long>3.5kb), *Komagataella pastoris* (short<0.37kb long>3.3kb), *Yarrowia lipolytica* (short<0.39kb long>3.5kb), *Caenorhabditis elegans* (short<0.47kb long>9.6kb), *Drosophila melanogaster* (short<0.56kb long>29kb), *Danio rerio* (short<1.3kb long>127kb), *Takifugu rubripes* (short<0.72kb long>27kb), *Xenopus tropicalis* (short<0.93kb long>83kb), *Gallus gallus* (short<0.67kb long>104kb), *Mus musculus* (short<1.2kb long>183kb), *Sus scrofa* (short<0.57kb long>197kb). For each gene group we identified all corresponding Gene Ontology biological process terms from the Ensembl genome database (100) and visualized the resulting term frequencies as word clouds using Mathematica.

Figure 10: Hox gene length is correlated with spatial expression and cell cycle duration in the *D. melanogaster* embryo. Drosophila Hoxd family genes are each represented by a colored rectangle, containing the length of the gene in base pairs. Spatial expression of a gene transcript is marked by its corresponding color on the Drosophila embryo map. Hoxd gene length is correlated with the cell cycle duration of the embryo location where the gene transcript is expressed, with short Hox gene transcripts expressed in regions with short mitotic cycles and long Hox gene transcripts expressed in regions of long mitotic cycles. Spatial map of cell cycle duration from (Foe, 1989; Foe and Alberts, 1983) and gene transcript expression from (Mallo and Alonso, 2013).

Figure 3-figure supplement 1: Simulations exploring the effects of cell cycle duration and RNA Polymerase II (rnaPol II) for different re-initiation distances, Ω and transcription rates, λ. A) We conducted simulations for different re-initiation distances (Default assumes that re-initiation only occurs one rnaPol II reaches the end of the gene). B) We conducted simulations for different transcription rate. Rates were randomized using a Gaussian distribution (mean=λ, standard deviation=λ) such that transcription time T=$\Gamma\lambda$/L. We generally assumed a fixed rate λ=1, where all genes exhibit the same rate. Changing parameters in A and B does not alter the overall trend we observe, that short genes produce more transcripts than longer genes. Other parameters for panels A-B were: ploidy=1, Li={1,2,3,4,5,6,7,8,9} and G=9, one cell division, iterations=1000. We also show in the case where all

755 genes are the same length, L=9, and cell cycle duration is constant (panels C-D) that when cell cycle

756 durations are short, Γ=1 hour, only high transcription rates can quench the constraint of short cell cycle.

757 E-F) When cell cycle duration is long, Γ=10 hours, changing transcription parameters can directly affect

758 transcript number. Other parameters for panels C-F were: ploidy=1, Li={9}, G=9, genome={$gene_1^9$,

759 $gene_2^9$ …, $gene_9^9$}, one cell division, iterations=1000 per panel.

760

761 **Figure 3-figure supplement 2: Effects of maternal transcript inheritance.** A) Simulation of expected

762 transcript count when cells are initiated without maternal transcripts in comparison to cells with

763 maternal transcripts. Parameters: gene lengths ($gene^{L_1}$-$^{L_{10}}$), ploidy=1, one cell division, 10 hours cell

764 cycle duration, iterations=10000. B) Simulation of expected transcript count where a proportion (zero or

765 half) of inherited parental transcripts remain after each cell division. Parameters: gene lengths ($gene^{L_1}$-

766 $^{L_{10}}$), ploidy=1, two cell divisions, 10 hours cell cycle duration, iterations=1000.

767

768 **Figure 3-figure supplement 3: Simulations exploring the effects of cell cycle duration on transcript**

769 **count per cell.** We find that short genes produced more transcripts than longer genes. We conducted

770 simulations for different gene numbers (10, 100, and 1000) and distributions across various cell cycle

771 durations (1-10 hours), always resulting in the same overall trend being observed. The transcriptome for

772 each cell is subdivided into short, medium, and long gene bins. Transcript counts in each bin are

773 averaged for each cell. Prediction from simulation shows that cells will have higher expression of short

774 gene transcripts and longer genes irrespective the number of genes in each bin. However, simulation

775 shows that longer cycle durations will increase relative transcript count per cell (other parameters

776 ploidy=1, one cell division, iterations=1000, RNA polymerase II re-initiation,

777 Ω occurs at the end of the gene).

778

779 **Figure 3-figure supplement 4: Single cell data exploring the effects of cell cycle duration on**

780 **transcript count per cell.** We find that short genes produce more transcripts than longer genes. A)

781 *Xenopus tropicalis* (Xenopus) single cell data obtained from GSE113074 (Lib2) B) *Danio rerio*

782 (zebrafish) single cell data obtained from GSE112294 and C) *Mus Musculus* (mouse) cortex single cell

783 data obtained from GSE107122, displaying predicted patterns where short genes have a higher average

784 transcript expression than longer genes.

785

786 **Figure 6-figure supplement 1: Varying cell cycle duration across time affects cell type proportions.**

787 All cell progeny are labeled based on their cell cycle duration (inherited from parent). Each line

788 represents a different rate of increase in average cell cycle duration, following rate increments 0

789 (shallowest constant curve), 1, 3, 6 and 9 (steepest curve). Cell cycle duration increases after each

790 division using a Gaussian distribution, such that Γ=Gaussian(mean Γ parent ± rate increment, standard

791 deviation σ=0.06). Gradual cell cycle duration changes affect the relative proportion of cells with

792 different cell cycle durations (pie charts). Parameters genome=10, gene lengths (gene$^L_1$-$^L_{10}$), λ=1 kb/hr,

793 15 cell divisions, iterations=500, ploidy n=1, RNA polymerase II re-initiation, $\Omega = 0.25$kb.

794

795 **Figure 6-figure supplement 2: Genes with increasing transcript expression are associated with**

796 **neuronal and synaptic pathways.** We used the mouse cortex time series single cell transcriptomics

797 data, obtained from GSE107122 (Yuzwa et al., 2017), to identify gene transcripts that increase or

798 decrease in expression level over development (with gene transcript expression following the pattern:

799 E11.5 < E13.5 < E15.5 < E17.5 for increasing and E11.5 > E13.5 >E15.5 > E17.5 for decreasing genes).

800 To summarize pathway annotation information for each gene, we identified all corresponding Gene

801 Ontology biological process terms for each gene from the Ensembl genome database and visualized term

802 frequencies as word clouds using Mathematica. We find the genes (n=2186) that increase across time are

803 associated with neural developmental (brain, neuron and synapse) pathways whereas the genes (n=1834)

804 that decrease across time are associated with DNA-repair and proliferation pathways.

805

806 **Figure 7-figure supplement 1: Illustrating the association between developmental time and median**

807 **gene length across 101 species, grouped by taxonomy class** (Supplementary File 3). A) The

808 association between number of cell types and species appearance in the fossil record according to

809 (Valentine et al., 1994) B) The association between median gene length and emergence time. C) The

810 association between median gene length and developmental time across 12 species taxonomy classes

811 (Saccharomycetes, Chromadorea, Ascidiacea, Insecta, Hyperoartia, Amphibia, Actinopterugii, Aves,

812 Reptilia, Mammalia, Sarcopterygii and Myxini). For each species we calculated median gene length:

813 The length of each gene was calculated using start and end positions for each gene as extracted from the

814 Ensembl genome database (version 95). Estimated developmental time was curated from the

815 encyclopedia of life or articles found in PubMed (Supplementary File 3). We used gestation time for

816   mammals and hatching time for species who lay eggs (since it is difficult to accurately define a

817   comparative stage for all species). We analyzed the data using a Pearson correlation test shown as r.

818

819   **Figure 8-supplement 1: Enriched pathways in short(A) and long(B) genes in Human.** Enrichment

820   map (Merico et al., 2010) showing the pathways enriched in short genes compared to long gene. Circles

821   represent pathway gene sets. Lines connecting circles represent overlap between pathways. Similar

822   pathways are grouped in larger bubbles and manually labeled using the AutoAnnotate (Kucera et al.,

823   2016), Cytoscape app (Reimand et al., 2019) and custom scripts (Supplementary File 4). Blue pathways

824   (nodes) are enriched in long genes and red pathways are enriched in short genes (p value < 0.05, FDR <

825   0.01, Jaccard coefficient > 0.25). Long genes were enriched in many more functional groups than short

826   genes with 798 and 152 enriched pathways, respectively.

827

828   **Figure 8-supplement 2: Genes participate in different pathways.** The plot shows the *H. sapiens*

829   genes divided into 20 groups from shortest genes to the longest genes. For each gene group we identified

830   all corresponding Gene Ontology (Ashburner et al., 2000) biological process terms downloaded from the

831   Ensembl genome database version 100 (Yates et al., 2016), grouped the terms into themes

832   (Supplementary File 5-6) and visualized the resulting term frequencies as word clouds using

833   Mathematica.

834   **Figure 8-supplement 3: Moving average across gene length.** *H. sapiens* genes were divided into 20

835   groups from shortest genes to longest genes. For each gene group, we identified all corresponding Gene

836   Ontology (Ashburner et al., 2000) biological process terms downloaded from the Ensembl genome

837   database version 100 (Yates et al., 2016) and grouped the terms into themes (Supplementary File 5-6).

838   We calculated a moving average to explore the trend across gene groups by theme. For example, we

839   identified the 15 most frequent themes between the shortest (blue) and longest (black) gene groups. We

840   found themes such as 'environment-sensing', 'immune', and 'olfactory' show a trend with a decreasing

841   average and are found only in the shortest gene group (blue). On the other hand, themes such as 'brain',

842   'muscle', 'neuron', and 'synapse', show an increasing trend and are found only in the longest genes

843   group (black).

844

845 **Figure 8-supplement 4: Pathway themes are associated with gene length.** The plot shows the *H.*
846 *sapiens* genes divided into 20 groups from shortest genes to the longest genes. For each gene group we
847 identified all corresponding Gene Ontology (Ashburner et al., 2000) biological process terms
848 downloaded from the Ensembl genome database version 100 (Yates et al., 2016), grouped the terms
849 into themes (Supplementary File 5-6) and visualized the resulting term relative frequencies in a matrix
850 plot using Mathematica. Darker shading indicates higher term frequency.
851
852 **Supplementary Files**
853
854 Supplementary File 1: Curated cell cycle duration data
855
856 Supplementary File 2: Simulations supporting transcriptome diversity analytical solution
857
858 Supplementary File 3: Curated developmental time for species and their corresponding median gene
859 length. The length of each gene was calculated using start and end positions for each gene as extracted
860 from the Ensembl genome database (version 95). Estimated developmental time was curated from the
861 encyclopedia of life or articles found in PubMed
862
863 Supplementary File 4: General pathway themes from Figure 8-figure supplement 1 generated by using a
864 mixture of automatic classification applying key words and manual identification.
865
866 Supplementary File 5: General pathway themes their corresponding the list of words that were used to
867 manually classify the pathways in Figure 8 and 8- figure supplement 2-3.
868
869 Supplementary File 6: General pathway themes in Figure 8 and 8- figure supplement 2-3 applied to H.
870 sapiens and their corresponding and Gene Ontology identifiers descriptions extracted from the Ensembl
871 genome database.
872
873 **References**

874 Ardehali MB, Lis JT. 2009. Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol*
875 **16**:1123–1124. doi:10.1038/nsmb1109-1123

876 Artieri CG, Fraser HB. 2014. Transcript Length Mediates Developmental Timing of Gene Expression
877   Across Drosophila. *Mol Biol Evol* **31**:2879–2889. doi:10.1093/molbev/msu226

878 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS,
879   Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE,
880   Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. *Nat*
881   *Genet* **25**:25–29. doi:10.1038/75556

882 Bertoli C, Skotheim JM, Bruin RAM de. 2013. Control of cell cycle transcription during G1 and S
883   phases. *Nat Rev Mol Cell Bio* **14**:518–528. doi:10.1038/nrm3629

884 Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM. 2018. The
885   dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*
886   **360**:eaar5780. doi:10.1126/science.aar5780

887 Bryant SV, Gardiner DM. 2018. Regeneration: sooner rather than later. *Int J Dev Biology* **62**:363–368.
888   doi:10.1387/ijdb.170269dg

889 Bryant SV, Gardiner DM. 2016. The relationship between growth and pattern formation. *Regen* **3**:103–
890   122. doi:10.1002/reg2.55

891 Calegari F, Haubensak W, Haffner C, Huttner WB. 2005. Selective Lengthening of the Cell Cycle in the
892   Neurogenic Subpopulation of Neural Progenitor Cells during Mouse Brain Development. *J Neurosci*
893   **25**:6533–6538. doi:10.1523/jneurosci.0778-05.2005

894 Coronado D, Godet M, Bourillot P-Y, Tapponnier Y, Bernat A, Petit M, Afanassieff M, Markossian S,
895   Malashicheva A, Iacone R, Anastassiadis K, Savatier P. 2013. A short G1 phase is an intrinsic
896   determinant of naïve embryonic stem cell pluripotency. *Stem Cell Res* **10**:118–131.
897   doi:10.1016/j.scr.2012.10.004

898 Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal
899   B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney
900   E, Hermjakob H, D'Eustachio P, Stein L. 2011. Reactome: a database of reactions, pathways and
901   biological processes. *Nucleic Acids Res* **39**:D691–D697. doi:10.1093/nar/gkq1018

902 Dalton S. 2015. Linking the Cell Cycle to Cell Fate Decisions. *Trends Cell Biol* **25**:592–600.
903   doi:10.1016/j.tcb.2015.07.007

904 Dalton S. 2013. G1 Compartmentalization and Cell Fate Coordination. *Cell* **155**:13–14.
905   doi:10.1016/j.cell.2013.09.015

906   Djabrayan NJ-V, Smits CM, Krajnc M, Stern T, Yamada S, Lemon WC, Keller PJ, Rushlow CA,

907   Shvartsman SY. 2019. Metabolic Regulation of Developmental Cell Cycles and Zygotic

908   Transcription. *Curr Biology Cb* **29**:1193-1198.e5. doi:10.1016/j.cub.2019.02.028

909   Durinck S, Moreau Y, Kasprzyk A, Davis S, Moor BD, Brazma A, Huber W. 2005. BioMart and

910   Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinform*

911   *Oxf Engl* **21**:3439–40. doi:10.1093/bioinformatics/bti525

912   Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic

913   datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**:1184–1191.

914   doi:10.1038/nprot.2009.97

915   Edgar BA, Kiehle CP, Schubiger G. 1986. Cell cycle control by the nucleo-cytoplasmic ratio in early

916   Drosophila development. *Cell* **44**:365–372. doi:10.1016/0092-8674(86)90771-3

917   Edgar LG, Wolf N, Wood WB. 1994. Early transcription in Caenorhabditis elegans embryos. *Dev Camb*

918   *Engl* **120**:443–51.

919   Farrell JA, O'Farrell PH. 2014. From Egg to Gastrula: How the Cell Cycle Is Remodeled During the

920   Drosophila Mid-Blastula TransitionUndefined, Annual Review of Genetics. pp. 1–26.

921   doi:10.1146/annurev-genet-111212-133531

922   Foe VE. 1989. Mitotic domains reveal early commitment of cells in Drosophila embryos. *Dev Camb*

923   *Engl* **107**:1–22.

924   Foe VE, Alberts BM. 1983. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles

925   that precede gastrulation in Drosophila embryogenesis. *J Cell Sci* **61**:31–70.

926   Furutachi S, Miya H, Watanabe T, Kawai H, Yamasaki N, Harada Y, Imayoshi I, Nelson M, Nakayama

927   KI, Hirabayashi Y, Gotoh Y. 2015. Slowly dividing neural progenitors are an embryonic origin of

928   adult neural stem cells. *Nat Neurosci* **18**:657–665. doi:10.1038/nn.3989

929   Geirsdottir L, David E, Keren-Shaul H, Weiner A, Bohlen SC, Neuber J, Balic A, Giladi A, Sheban F,

930   Dutertre C-A, Pfeifle C, Peri F, Raffo-Romero A, Vizioli J, Matiasek K, Scheiwe C, Meckel S, Mätz-

931   Rensing K, Meer F van der, Thormodsson FR, Stadelmann C, Zilkha N, Kimchi T, Ginhoux F,

932   Ulitsky I, Erny D, Amit I, Prinz M. 2019. Cross-Species Single-Cell Analysis Reveals Divergence of

933   the Primate Microglia Program. *Cell* **179**:1609-1622.e16. doi:10.1016/j.cell.2019.11.010

934   Gilbert SF, Barresi MJ. 2016. Developmental BiologyUndefined. pp. 1–810.

935   Gubb D. 1986. Intron-delay and the precision of expression of homoeotic gene products inDrosophila.

936   *Dev Genet* **7**:119–131. doi:10.1002/dvg.1020070302

937   Guo S, Zi X, Schulz VP, Cheng J, Zhong M, Koochaki SHJ, Megyola CM, Pan X, Heydari K,
938   Weissman SM, Gallagher PG, Krause DS, Fan R, Lu J. 2014. Nonstochastic Reprogramming from a
939   Privileged Somatic Cell State. *Cell* **156**:649–662. doi:10.1016/j.cell.2014.01.020

940   Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, Neugebauer KM. 2014. The Earliest
941   Transcribed Zygotic Genes Are Short, Newly Evolved, and Different across Species. *Cell Reports*
942   **6**:285–292. doi:10.1016/j.celrep.2013.12.030

943   Jukam D, Shariati SAM, Skotheim JM. 2017. Zygotic Genome Activation in Vertebrates. *Dev Cell*
944   **42**:316–332. doi:10.1016/j.devcel.2017.07.026

945   Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, Venugopal AK, Telikicherla D,
946   Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H,
947   Subbannayya Y, Goel R, Jacob HK, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R,
948   Ramachandra Y, Rahiman BA, Prasad TK, Lin J-X, Houtman JC, Desiderio S, Renauld J-C,
949   Constantinescu SN, Ohara O, Hirano T, Kubo M, Singh S, Khatri P, Draghici S, Bader GD, Sander
950   C, Leonard WJ, Pandey A. 2010. NetPath: a public resource of curated signal transduction pathways.
951   *Genome Biol* **11**:R3. doi:10.1186/gb-2010-11-1-r3

952   Keane PA, Seoighe C. 2016. Intron Length Coevolution across Mammalian Genomes. *Mol Biol Evol*
953   **33**:2682–2691. doi:10.1093/molbev/msw151

954   Kimelman D, Kirschner M, Scherson T. 1987. The events of the midblastula transition in Xenopus are
955   regulated by changes in the cell cycle. *Cell* **48**:399 407. doi:10.1016/0092-8674(87)90191-7

956   Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic
957   development of the zebrafish. *Dev Dynam* **203**:253–310. doi:10.1002/aja.1002030302

958   Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. 2016. AutoAnnotate: A Cytoscape app for
959   summarizing networks with semantic annotations. *F1000research* **5**:1717.
960   doi:10.12688/f1000research.9090.1

961   Lemons D, McGinnis W. 2006. Genomic Evolution of Hox Gene Clusters. *Science* **313**:1918–1922.
962   doi:10.1126/science.1132040

963   Liu S, Ginzberg MB, Patel N, Hild M, Leung B, Li Z, Chen Y-C, Chang N, Wang Y, Tan C, Diena S,
964   Trimble W, Wasserman L, Jenkins JL, Kirschner MW, Kafri R. 2018. Size uniformity of animal cells
965   is actively maintained by a p38 MAPK-dependent regulation of G1-length. *Elife* **7**:e26947.
966   doi:10.7554/elife.26947

967   Mallo M, Alonso CR. 2013. The regulation of Hox gene expression during animal development.

968   *Development* **140**:3951–3963. doi:10.1242/dev.068346

969   Merico D, Isserlin R, Stueker O, Emili A, Bader GD. 2010. Enrichment Map: A Network-Based Method

970   for Gene-Set Enrichment Visualization and Interpretation. *Plos One* **5**:e13984.

971   doi:10.1371/journal.pone.0013984

972   Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A,

973   Doremieux O, Campbell MJ, Kitano H, Thomas PD. 2005. The PANTHER database of protein

974   families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**:D284–D288.

975   doi:10.1093/nar/gki078

976   Mueller WA, Hassel M, Grealy M. 2015. Development and Reproduction in Humans and Animal Model

977   Species, Development and Reproduction in Humans and Animal Model Species. Springer.

978   doi:10.1007/978-3-662-43784-1

979   Mukherjee RN, Sallé J, Dmitrieff S, Nelson KM, Oakey J, Minc N, Levy DL. 2020. The Perinuclear ER

980   Scales Nuclear Size Independently of Cell Size in Early Embryos. *Dev Cell*.

981   doi:10.1016/j.devcel.2020.05.003

982   Mummery CL, Brink CE van den, Laat SW de. 1987. Commitment to differentiation induced by retinoic

983   acid in P19 embryonal carcinoma cells is cell cycle dependent. *Dev Biol* **121**:10–19.

984   doi:10.1016/0012-1606(87)90133-3

985   Newport J, Kirschner M. 1982a. A major developmental transition in early xenopus embryos: II. control

986   of the onset of transcription. *Cell* **30**:687–696. doi:10.1016/0092-8674(82)90273-2

987   Newport J, Kirschner M. 1982b. A major developmental transition in early xenopus embryos: I.

988   characterization and timing of cellular changes at the midblastula stage. *Cell* **30**:675–686.

989   doi:10.1016/0092-8674(82)90272-0

990   O'Farrell PH, Stumpff J, Su TT. 2004. Embryonic Cleavage Cycles: How Is a Mouse Like a Fly? *Curr*

991   *Biol* **14**:R35–R45. doi:10.1016/j.cub.2003.12.022

992   Ohsugi K, Gardiner DM, Bryant SV. 1997. Cell Cycle Length Affects Gene Expression and Pattern

993   Formation in Limbs. *Dev Biol* **189**:13–21. doi:10.1006/dbio.1997.8665

994   Pauklin S, Vallier L. 2013. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell*

995   **155**:135–147. doi:10.1016/j.cell.2013.08.031

996   Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong

997   J, Xu C, Merico D, Bader GD. 2019. Pathway enrichment analysis and visualization of omics data

998   using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**:482–517.

999   doi:10.1038/s41596-018-0103-9

1000   Rodríguez-Carballo E, Lopez-Delisle L, Yakushiji-Kaminatsui N, Ullate-Agote A, Duboule D. 2019.

1001   Impact of genome architecture on the functional activation and repression of Hox regulatory

1002   landscapes. *Bmc Biol* **17**:55. doi:10.1186/s12915-019-0677-x

1003   Rothe M, Pehl M, Taubert H, Jäckle H. 1992. Loss of gene function through rapid mitotic cycles in the

1004   Drosophila embryo. *Nature* **359**:359156a0. doi:10.1038/359156a0

1005   Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene

1006   expression data. *Nat Biotechnol* **33**:495–502. doi:10.1038/nbt.3192

1007   Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the Pathway

1008   Interaction Database. *Nucleic Acids Res* **37**:D674–D679. doi:10.1093/nar/gkn653

1009   Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T.

1010   2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction

1011   Networks. *Genome Res* **13**:2498–2504. doi:10.1101/gr.1239303

1012   Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, Ko MSH. 2009. Database for mRNA Half-

1013   Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating

1014   Mouse Embryonic Stem Cells. *Dna Res* **16**:45–58. doi:10.1093/dnares/dsn030

1015   Shermoen AW, O'Farrell PH. 1991. Progression of the cell cycle through mitosis leads to abortion of

1016   nascent transcripts. *Cell* **67**:303–310. doi:10.1016/0092-8674(91)90182-x

1017   Singh AM, Chappell J, Trost R, Lin L, Wang T, Tang J, Matlock BK, Weller KP, Wu H, Zhao S, Jin P,

1018   Dalton S. 2013. Cell-cycle control of developmentally regulated transcription factors accounts for

1019   heterogeneity in human pluripotent cells. *Stem Cell Rep* **1**:532–44. doi:10.1016/j.stemcr.2013.10.009

1020   Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL,

1021   Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L,

1022   Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. 2017. WikiPathways: a

1023   multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*

1024   **46**:gkx1064-. doi:10.1093/nar/gkx1064

1025   Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy

1026   SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based

1027   approach for interpreting genome-wide expression profiles. *P Natl Acad Sci Usa* **102**:15545–15550.

1028   doi:10.1073/pnas.0506580102

1029 Swinburne IA, Miguez DG, Landgraf D, Silver PA. 2008. Intron length increases oscillatory periods of
1030   gene expression in animal cells. *Gene Dev* **22**:2342–2346. doi:10.1101/gad.1696108

1031 Swinburne IA, Silver PA. 2008. Intron Delays and Transcriptional Timing during Development. *Dev*
1032   *Cell* **14**:324–330. doi:10.1016/j.devcel.2008.02.002

1033 Tabansky I, Lenarcic A, Draft RW, Loulier K, Keskin DB, Rosains J, Rivera-Feliciano J, Lichtman JW,
1034   Livet J, Stern JNH, Sanes JR, Eggan K. 2013. Developmental Bias in Cleavage-Stage Mouse
1035   Blastomeres. *Curr Biol* **23**:21–31. doi:10.1016/j.cub.2012.10.054

1036 Takahashi T, Nowakowski R, Caviness V. 1995a. The cell cycle of the pseudostratified ventricular
1037   epithelium of the embryonic murine cerebral wall. *J Neurosci* **15**:6046–6057.
1038   doi:10.1523/jneurosci.15-09-06046.1995

1039 Takahashi T, Nowakowski R, Caviness V. 1995b. Early ontogeny of the secondary proliferative
1040   population of the embryonic murine cerebral wall. *J Neurosci* **15**:6058–6068.
1041   doi:10.1523/jneurosci.15-09-06058.1995

1042 Tennyson CN, Klamut HJ, Worton RG. 1995. The human dystrophin gene requires 16 hours to be
1043   transcribed and is cotranscriptionally spliced. *Nat Genet* **9**:ng0295-184. doi:10.1038/ng0295-184

1044 Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans.
1045   *Paleobiology* **20**:131–142. doi:10.1017/s0094837300012641

1046 Vallier L. 2015. Cell Cycle Rules Pluripotency. *Cell Stem Cell* **17**:131–132.
1047   doi:10.1016/j.stem.2015.07.019

1048 Vogel C, Chothia C. 2006. Protein Family Expansions and Biological Complexity. *Plos Comput Biol*
1049   **2**:e48. doi:10.1371/journal.pcbi.0020048

1050 Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. 2018. Single-cell mapping of
1051   gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**:981–987.
1052   doi:10.1126/science.aar4362

1053 Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KHS. 1997. Viable offspring derived from fetal
1054   and adult mammalian cells. *Nature* **385**:810–813. doi:10.1038/385810a0

1055 Wolfram-Research. 2017. Mathematica version 11.0–12.0.

1056 Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P,
1057   Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann
1058   T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A,
1059   Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder

1060 SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ,

1061 Cunningham F, Aken BL, Zerbino DR, Flicek P. 2016. Ensembl 2016. *Nucleic Acids Res* **44**:D710–

1062 D716. doi:10.1093/nar/gkv1157

1063 Yoon K-J, Ringeling FR, Vissers C, Jacob F, Pokrass M, Jimenez-Cyrus D, Su Y, Kim N-S, Zhu Y,

1064 Zheng L, Kim S, Wang X, Doré LC, Jin P, Regot S, Zhuang X, Canzar S, He C, Ming G, Song H.

1065 2017. Temporal Control of Mammalian Cortical Neurogenesis by m6A Methylation. *Cell* **171**:877–

1066 889.e17. doi:10.1016/j.cell.2017.09.003

1067 Yuan K, Seller CA, Shermoen AW, O'Farrell PH. 2016. Timing the Drosophila Mid-Blastula

1068 Transition: A Cell Cycle-Centered View. *Trends Genet* **32**:496–507. doi:10.1016/j.tig.2016.05.006

1069 Yuzwa SA, Borrett MJ, Innes BT, Voronova A, Ketela T, Kaplan DR, Bader GD, Miller FD. 2017.

1070 Developmental Emergence of Adult Neural Stem Cells as Revealed by Single-Cell Transcriptional

1071 Profiling. *Cell Reports* **21**:3970–3986. doi:10.1016/j.celrep.2017.12.017

1072 Zhou C, Slaughter BD, Unruh JR, Eldakak A, Rubinstein B, Li R. 2011. Motility and Segregation of

1073 Hsp104-Associated Protein Aggregates in Budding Yeast. *Cell* **147**:1186–1196.

1074 doi:10.1016/j.cell.2011.11.002

1075 Zoller B, Little SC, Gregor T. 2018. Diverse Spatial Expression Patterns Emerge from Unified Kinetics

1076 of Transcriptional Bursting. *Cell* **175**:835–847.e25. doi:10.1016/j.cell.2018.09.056

1077

1078

1079

1080

1081

1082

M. musculus

Cell Cycle Duration (hrs) vs Days

Legend:
- Altas M., Bond VP. (1965)
- Cameron IL. (1964)
- Kauffman SL. (1968)
- Solter D. et al (1971)
- Young RW. (1985)
- Takahashi T. (1995)
- Lukaszewicz A. et al (2002)
- Artus J., Cohen−Tannoudji M. (2008)

Labels: cortical stem cell, retina, embryo, cerebral wall, trophoblast, brain, retina, neural tube, mesoderm, ectoderm

A A Cell

Transcript

Genome

Gene Length = introns + exons

Cell Cycle Duration

B

Expected Expression

gene

Cell Cycle Duration 1

gene

Cell Cycle Duration 2

gene

Cell Cycle Duration 3

C Cell Division

Random Distribution of Transcripts

D

(2,1,0)

Transcriptome Vector

(1,0,1)

Transcriptome Vector

gene Short

gene Medium

gene Long

A

Simulations

B

Species

A

Parameters			
Filter	L_1	L_G	G
On	1	10	10
Off	0.1	1	10

B

Parameters			
Filter	L_1	L_G	G
On	1	10	10
Off	0.1	1	10

C

2 Clusters

92 Clusters

74 Clusters

Γ hrs
2.00

Γ hrs
6.00

Γ hrs
10.00

tsne

tsne

tsne

A

Scenario 1: Fast–Fast

Fast

Parental Cell

Fast

Cell Divisions

Scenario 2: Slow–Fast

Slow

Parental Cell

Fast

Cell Divisions

B

Fast–Fast

$\Gamma_1=1.00$
$\Gamma_2=1.00$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptor Diversity

0.5

Slow–Fast

$\Gamma_1=1.00$
$\Gamma_2=1.50$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptor Diversity

94.1

Slow–Fast

$\Gamma_1=1.00$
$\Gamma_2=1.75$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptor Diversity

97.

Slow–Fast

$\Gamma_1=1.00$
$\Gamma_2=2.00$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptor Diversity

99.8

A Cell Divisions

Time

B Simulations

C M. musculus cortex

Cell cycle duration, hrs

Relative cell cycle duration

Slow
Medium
Fast

63.0%

21.0%

6.0%

8.8%

A

H. sapiens
M. mulatta
S. scrofa
M. musculus
G. gallus
X. tropicalis
T. rubripes
D. rerio
O. dioica
D. melanogaster
C. elegans
S. cerevisiae

1hr

Gene Length Log$_{10}$(bp)

Percent of Genes restricted

B

r=0.74

E. burgeri
M. nemestrina
H. sapiens
M. spretus
M. murinus
M. caroli
N. galili
G. gorilla
L. chalumnae
M. auratus
D. rerio
M. domestica
P. abelii
C. chok1gshd
L. africana
A. mexicanus
X. tropicalis
A. melanoleuca
F. albicollis
O. aries
E. caballus
G. gallus
M. lucifugus
C. savignyi
O. niloticus
P. marinus
X. maculatus
T. rubripes
G. morhua
G. aculeatus
D. melanogaster
T. nigroviridis
C. elegans
C. intestinalis

Median Gene Length (kb)

Estimated Developmental Time (days)

H. sapiens

Abd-B 45025

Fast mitotic region

lab 17164

pb 34263

Dfd 10593

Scr 26862

Antp 102976

Slow mitotic region

abd-A 22836

Ubx 77804

A

Simulations

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

Avg. Transcript Count per Cell

Cell Cycle Duration, Γ

B

Species

Short ≤ 10.kb
Medium ≤25.kb
Long ≤5000.kb

Gene 10^3

Avg. Transcript Count per Cell

H. sapiens
M. fascicularis
C. jacchus
M. musculus
R. norvegicus
M. auratus
N. galili
O. aries
G. gallus
D. rerio

A

Simulation

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

Avg. Transcript Count per Cell

Without   With

Maternal transcripts

B

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

Avg. Transcript Count per Cell

0   0.5

Proportion of inherited parental transcripts

A

Simulations G=10

Avg. Transcript Count per Cell

Cell Cycle Duration, Γ

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

B

Simulations G=100

Avg. Transcript Count per Cell

Cell Cycle Duration, Γ

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

C

Simulations G=1000

Avg. Transcript Count per Cell

Cell Cycle Duration, Γ

Short ≤ 3.3
Medium ≤ 6.7
Long ≤ 10.0

A *X. tropicalis*

B *D. rerio*

C *M. musculus*

A

B

Parameters			
Filter	L_1	L_G	G
On	1	10	10
Off	0.1	1	10

C

2 Clusters

40 Clusters

25 Clusters

Γ hrs
2.00

Γ hrs
6.00

Γ hrs
10.00

tsne

tsne

tsne

A Scenario 1: Fast–Fast

Parental Cell

Fast

Fast

Cell Divisions

B Scenario 2: Slow–Fast

Parental Cell

Slow

Fast

Cell Divisions

C Scenario 3: Slow–Fast

Parental Cell

Slow

Fast

Cell Divisions

D Fast–Fast

$\Gamma_1 = 1.00$
$\Gamma_2 = 1.00$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptom Diversity

E Slow–Fast

$\Gamma_1 = 1.00$
$\Gamma_2 = 2.00$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptom Diversity

F Slow–Fast

$\Gamma_1 = 1.00$
$\Gamma_2 = 2.00$

Cell Number : Maximum

Cell Divisions

Cell %

Transcriptom Diversity

A Cell Divisions

B Simulations

C M. musculus cortex

Simulations

Increasing
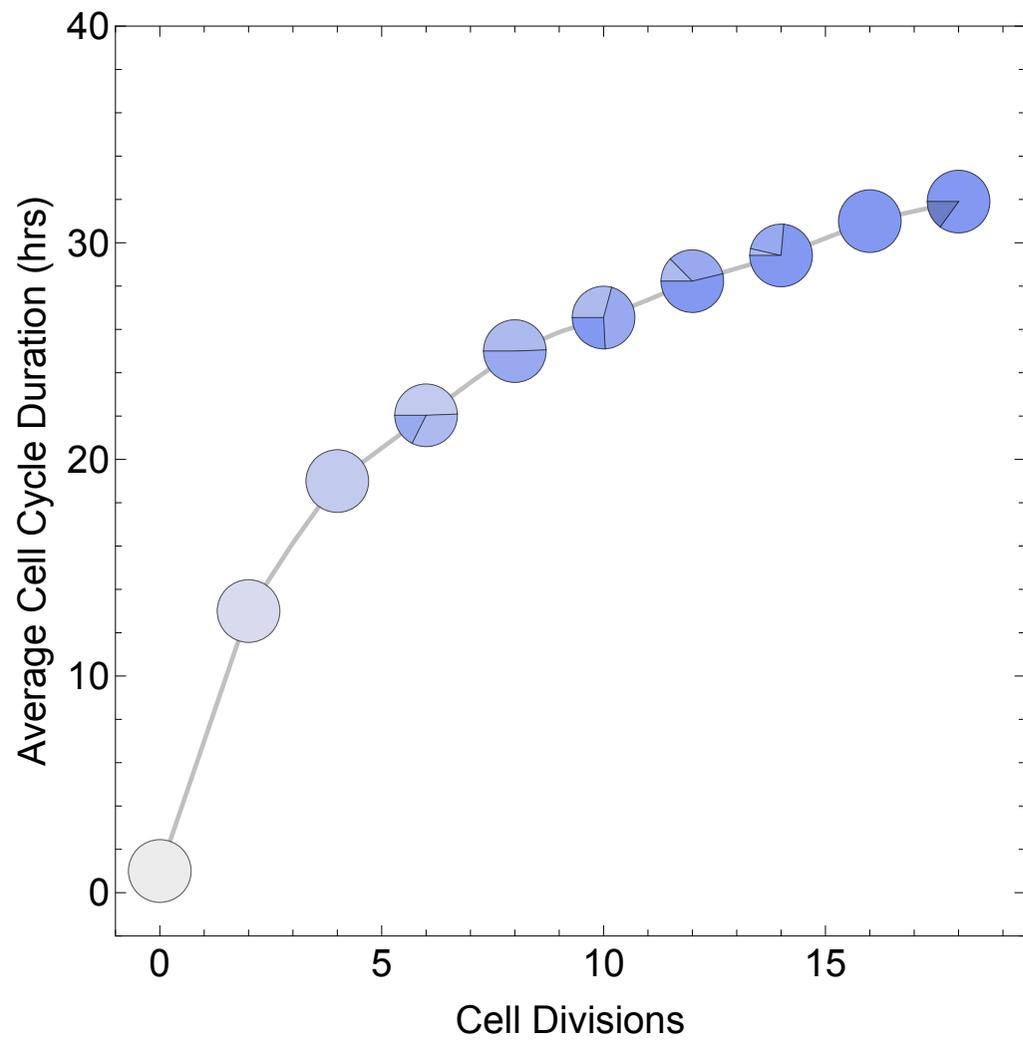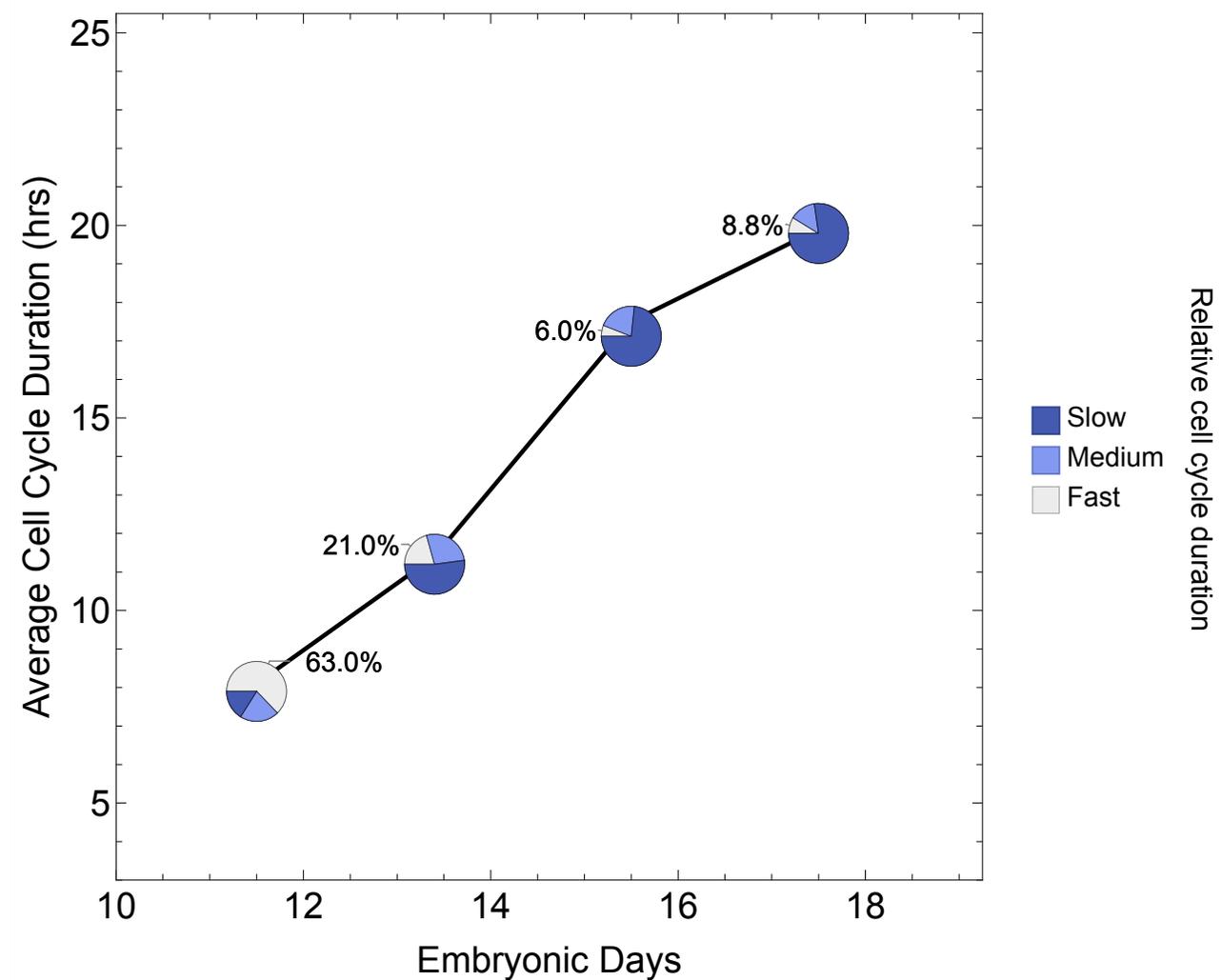E11.5 <E13.5 <E15.5 <E17.5

Decreasing
E11.5 >E13.5 >E15.5 >E17.5

A

H. sapiens
M. mulatta
S. scrofa
M. musculus
G. gallus
X. tropicalis
T. rubripes
D. rerio
O. dioica
D. melanogaster
C. elegans
S. cerevisiae

1hr

Gene Length Log$_{10}$(bp)

Percent of Genes restricted

B

r=0.74

E. burgeri
M. nemestrina
H. sapiens
M. spretus
M. murinus
M. caroli
N. galili
G. gorilla
L. chalumnae
M. auratus
D. rerio
M. domestica
P. abelii
C. chok1gshd
L. africana
X. tropicalis
A. mexicanus
F. albicollis
A. melanoleuca
O. niloticus
G. gallus
O. aries
E. caballus
C. savignyi
P. marinus
M. lucifugus
T. rubripes
X. maculatus
G. morhua
D. melanogaster
G. aculeatus
T. nigroviridis
C. elegans
C. intestinalis

Median Gene Length (kb)

Estimated Developmental Time (days)

A r=0.84

Cell Types vs Million Years Before Present

Labeled points: Mammalia, Aves, Sarcopterygii, Amphibia, Myxini, Insecta, Hyperoartia, Ascidiacea, Chromadorea, Saccharomycetes

B r=0.75

Median Gene Length (kb) vs Million Years Before Present

Labeled points: Myxini, Sarcopterygii, Mammalia, Actinopterygii, Aves, Ascidiacea, Insecta, Chromadorea, Saccharomycetes

C r=0.94

Median Gene Length (kb) vs Estimated Developmental Time (days)

Labeled points: Myxini, Sarcopterygii, Mammalia, Amphibia, Reptilia, Hyperoartia, Aves, Actinopterygii, Ascidiacea, Insecta, Chromadorea, Saccharomycetes

H. sapiens

A

immune

protein translation

mitochondrion

environment sensing

cell death

development

cell communication

B

growth regulation

circadian rythm

ecm

neural development

polarity cell organization

muscle control

immune

development

environment sensing

cell death

transcription

cytoskeleton

cell communication

76–1576bp
n=1124

1576–3060bp
n=1127

3060–4617bp
n=1125

4617–6374bp
n=1125

6374–8429bp
n=1126

8429–10672bp
n=1124

10672–12993bp
n=1127

12993–15600bp
n=1126

15600–18695bp
n=1125

18695–23063bp
n=1126

23063–28047bp
n=1126

28047–33933bp
n=1124

33933–41750bp
n=1127

41750–50709bp
n=1125

50709–63569bp
n=1126

63569–81294bp
n=1125

81294–107125bp
n=1126

107125–149292bp
n=1125

149292–244058bp
n=1126

244058–2473539bp
n=1125

Abd-B
45025

Fast mitotic region

lab
17164

pb
34263

abd-A
22836

Slow mitotic region

Ubx
77804

Dfd
10593

Scr
26862

Antp
102976