

1   **The regulatory landscape of *Arabidopsis thaliana* roots at single-cell resolution**

2  
3   **Authors:** Michael W. Dorrity<sup>1</sup>, Cris Alexandre<sup>1</sup>, Morgan Hamm<sup>1</sup>, Anna-Lena Vigil<sup>3</sup>,  
4   Stanley Fields<sup>1,2</sup>, Christine Queitsch<sup>1</sup> and Josh Cuperus<sup>1</sup>

5  
6   **Affiliations:**

7   <sup>1</sup> Department of Genome Sciences, University of Washington, Seattle, WA 98195

8   <sup>2</sup> Department of Medicine, University of Washington, Seattle, WA 98195

9   <sup>3</sup> School of Life Sciences, University of Nevada, Las Vegas, NV 89154

10  
11   <sup>†</sup> Correspondence to Christine Queitsch (queitsch@uw.edu) and Josh Cuperus  
12 (cuperusj@uw.edu)

13   **Abstract:** In plants, chromatin accessibility – the primary mark of regulatory DNA – is  
14   relatively static across tissues and conditions. This scarcity of accessible sites that are  
15   dynamic or tissue-specific may be due in part to tissue heterogeneity in previous bulk  
16   studies. To assess the effects of tissue heterogeneity, we apply single-cell ATAC-seq  
17   to *A. thaliana* roots and identify thousands of differentially accessible sites, sufficient to  
18   resolve all major cell types of the root. However, even this vast increase relative to bulk  
19   studies in the number of dynamic sites does not resolve the poor correlation at  
20   individual loci between accessibility and expression. Instead, we find that the entirety  
21   of a cell's regulatory landscape and its transcriptome each capture cell type identity  
22   independently. We leverage this shared information on cell identity to integrate  
23   accessibility and transcriptome data in order to characterize developmental  
24   progression, endoreduplication and cell division in the root. We further use the  
25   combined data to characterize cell type-specific motif enrichments of large  
26   transcription factor families and to link the expression of individual family members to  
27   changing accessibility at specific loci, taking the first steps toward resolving the direct  
28   and indirect effects that shape gene expression. Our approach provides an analytical  
29   framework to infer the gene regulatory networks that execute plant development.

30  
31   **Introduction**

32  
33       Single-cell genomics allows an unbiased sampling of cells during development,  
34   with the potential to reveal the order and timing of gene regulatory and gene  
35   expression events that specify cell identity and lineage. An ideal system to test the  
36   ability of single-cell genomics to provide novel insights into development is the  
37   *Arabidopsis thaliana* root: along its longitudinal axis, a single, radially-symmetric root  
38   captures developmental trajectories for several radially-symmetric cell types.  
39   Approaches in this organism have included single-cell RNA-seq to transcriptionally  
40   profile individual root cell types along this developmental axis<sup>1–6</sup> and with respect to  
41   their ploidy.

42

43       Studies of chromatin accessibility in samples enriched for specific plant cell  
44 types have revealed: (i) the existence of cell type-specific regulatory elements; (ii) the  
45 relative scarcity of such elements compared to their prevalence in animals or humans;  
46 (iii) the expected enrichment of transcription factor binding sites within these elements;  
47 and (iv) a higher frequency of dynamic regulatory elements upstream of  
48 environmentally-responsive genes than constitutively expressed genes.<sup>7,8</sup> Although the  
49 correlation between chromatin accessibility and nearby gene expression is generally  
50 weak in both plants and animals,<sup>9</sup> this correlation improves for regulatory elements that  
51 show dynamic changes in chromatin accessibility, for example in response to an  
52 environmental stimulus or developmental signal.<sup>7,9–11</sup> In contrast to animals, however,  
53 the majority of chromatin-accessible sites in plants show little change across tissues,  
54 conditions, or even genetic backgrounds, raising the possibility that cell and tissue  
55 identity is less rigidly engrained in the chromatin landscape in plants than in animals.<sup>7</sup>  
56 Alternatively, cell type-specific regulatory elements and gene expression in plants may  
57 have been obscured by tissue heterogeneity in bulk tissue studies.

58  
59       Cell type-specific chromatin-accessible landscapes are also of interest for  
60 addressing other fundamental biological questions. General transcription decreases  
61 along a cell type's developmental trajectory while expression of cell type-specific  
62 genes increases,<sup>2,12,13</sup> in agreement with Waddington's predictions on epigenetic  
63 landscapes.<sup>14</sup> In the *A. thaliana* root, the increasing maturity of certain cell layers is  
64 accompanied by endoreduplication. The presence of additional gene copies may  
65 contribute to the observed increase in the expression of cell type-specific genes;  
66 alternatively, the initial gene copies may increase their transcription. Although  
67 endoreduplication is a common mechanism to regulate cell size and differentiation in  
68 plants and some human and animal tissues,<sup>15–17</sup> the influence of this phenomenon on  
69 gene regulation and expression has been largely overlooked. In plants,  
70 endoreduplication generally enhances transcription,<sup>17,18</sup> in particular of cell-wall-related  
71 genes<sup>19</sup> and genes encoding ribosomal RNA,<sup>20</sup> hinting at a role for this process in  
72 driving increased translation.

73  
74       Here, we provide the first single-cell resolution maps of open chromatin in the *A.*  
75 *thaliana* root to address the issue of tissue heterogeneity and to detect likely  
76 endoreduplication events. We use a droplet-based approach to profile over 5000 nuclei  
77 for chromatin accessibility and identify 8000 regulatory elements that together define  
78 most cell types of the root. We describe an analytical framework that links patterns of  
79 open chromatin with transcriptional states to predict the identity, function and  
80 developmental stage of individual cells in the *A. thaliana* root. We integrate the single-  
81 cell ATAC-seq (scATAC-seq) data with published single-cell RNA-seq (scRNA-seq)  
82 profiles of the same tissue to obtain automated cell annotations of scATAC cells. Using  
83 the integrated dataset, we link individual scATAC cells with their nearest neighbors in  
84 scRNA space to define relative developmental progression, level of endoreduplication  
85 and the genes differentially expressed in these nearest neighbors. This approach  
86 allows the identification of three distinct developmental states of endodermis cells that

87 had escaped detection using scRNA-seq alone. Using integrated scRNA-seq data, we  
88 predict individual members of large transcription factor families that play a role in  
89 epidermis development, pinpointing individual regulatory events that link peak  
90 accessibility and transcription factor expression in these cells. The combination of  
91 binding motifs, transcription factor expression and chromatin accessibility provides a  
92 basis for predicting the gene regulatory events that underlie development.  
93

## 94 **Results**

### 95 **scATAC-seq identifies known root cell types**

96 We first asked if ATAC-seq profiles at the single-cell level were capable of  
97 capturing known root cell types. We profiled 5283 root nuclei, at a median of 7290  
100 unique ATAC inserts per cell. A high fraction of these inserts occurred in one of the  
101 21,889 open chromatin peaks (FRIP score = 0.71) based on pseudo-bulk peak calling  
102 (Cellranger v3.1, 10X Genomics); this fraction is similar to that seen in high-quality bulk  
103 accessibility studies (**Figure S1A, S1B**).<sup>9</sup> We used UMAP dimensionality reduction of  
104 the peak by cell matrix to build a two-dimensional representation grouping of cells with  
105 similar accessibility profiles (**Figure 1A**). Subsequent cluster assignment by Louvain  
106 community detection identified nine distinct cell clusters.<sup>21</sup> Across all cell types, we  
107 identified 7910 peaks (ranging from 939 – 2065 per cell type) with significant differential  
108 accessibility, suggesting that around a third of all accessible sites contain some  
109 information on cell type (**Supplementary Table 1**). To assign cell type annotations to  
110 each of these clusters, we generated “gene activity” scores that sum all ATAC inserts  
111 within each gene body and 400 bp upstream of its transcription start site. This  
112 approach rests on the assumption that a chromatin-accessible site in the compact *A.*  
113 *thaliana* genome tends to be associated with regulation of its most proximal gene.<sup>22</sup>  
114 While this assumption may not hold universally, gene activity scores offer the  
115 advantage of allowing a direct comparison to bulk ATAC-seq and single-cell RNA-seq  
116 datasets through a matched feature set. In this way, we identified genes whose  
117 accessibility signal specifically marks each cell cluster. We visualized peaks with cell  
118 type-specific accessibility by grouping cells of a similar type and “pseudobulking” their  
119 insert counts at each position in the genome (**Figure 1B**). Cell type-specific ATAC  
120 tracks that resemble those obtained in prior whole tissue and cell type enrichment-  
121 based ATAC-seq studies for the root (**Figure 1B**).<sup>11</sup>  
122

123 We used comparisons to tissue-specific genes that were identified from single-  
124 cell RNA-seq studies of the *A. thaliana* root to assign a cell type to each cluster defined  
125 by ATAC markers from “gene activity” scores.<sup>2,5,6</sup> We identified 210 genes with unique  
126 accessibility patterns across all cell types (**Supplementary Table 2**); FRIP scores,  
127 fragment lengths, and total read counts did not vary greatly across cell types (**Figure**  
128 **S1C, S1D, S1E**). For each cell type, the median number of genes with tissue-specific  
129 accessibility was 20 (range 5 to 53) (**Figure 1C**). This small number of genes is

130 consistent with earlier studies that show few open chromatin sites that define cell type  
131 identity in *A. thaliana*.<sup>7,23</sup> Although thousands of differentially accessible sites have been  
132 found across tissue types,<sup>7</sup> accessibility differences between more closely related cell  
133 types remains largely unexplored, with the exception of root hair vs non-hair, in which  
134 very few differences were found.<sup>7,11</sup> For three cell clusters (959 cells, or 18% of cells),  
135 we could not identify a coherent set of markers and therefore could not annotate  
136 them (grey points, **Figure 1A**). However, all other cell clusters were manually annotated  
137 and corresponded to the major cell layers of the root: outer layers including epidermis  
138 cortex, and a precursor of endodermis and cortex (ec pre); endodermal layers  
139 comprised of three distinct types (endo 1, 2, and 3); and the stele comprised of two  
140 main types along with a phloem type (stele phloem). In general, ATAC marker genes  
141 did not show a strong overlap with RNA-based marker genes. Endodermis cells were  
142 an exception, as several of their ATAC marker genes (AT3G32980, AT1G61590,  
143 AT1G14580, AT3G22600, AT5G66390) were also found to be marker genes in single-  
144 cell RNA-seq studies.<sup>24</sup> While this lack of overlap makes annotation more challenging,  
145 it is consistent with the reported weak correlation of chromatin accessibility with gene  
146 expression.<sup>23,25</sup> Moreover, the finding that expression levels are not precisely predicted  
147 by nearby accessible sites suggests that accessibility can add orthogonal information  
148 about cell identity to further stratify cell types into distinct subtypes.  
149

150 **Sequences motifs of transcription factor families associate with cell type-specific**  
151 **sites of open chromatin**

152 Accessibility at regulatory sites is driven by transcription factor binding and  
153 modification of local chromatin.<sup>26</sup> We examined if any of the cell type-specific  
154 accessible sites were associated with the presence of transcription factor binding  
155 motifs. To do so, we used a set of representative motifs for all *A. thaliana* transcription  
156 factor families and nearly every individual transcription factor<sup>27</sup> to tally these motif  
157 counts within all 21,889 peaks in the full scATAC-seq dataset to build a peak-by-motif  
158 matrix. As each peak can be described in terms of its relative accessibility in each of  
159 the identified cell types, we performed a linear regression for each motif to test for  
160 significant association of accessibility and motif presence. Relative accessibility values  
161 were calculated by first pseudo-bulking all peak counts by cell type and then  
162 normalizing these cell type-specific peak accessibility scores to a background peak  
163 accessibility of all cells pooled together. By testing the association of motif counts and  
164 cell type-specific accessibility, we identify transcription factor binding motifs whose  
165 presence is correlated with more accessibility in each cell type.  
166

167 We found significant associations with motifs from at least one transcription  
168 factor family in all cell types (**Figure 1D**). For example, relative chromatin accessibility  
169 in epidermal cells was strongly associated (q-values ranging from 1e-24 to 1e-133)  
170 with the presence of motifs from the WRKY transcription factor family; this family  
171 includes *TTG2*, which, along with *TTG1* and *GL2*, has important roles in atrichoblast

173 fate in the epidermis.<sup>28</sup> Furthermore, the effects of each motif family on relative  
174 accessibility was sufficient to hierarchically cluster cell types according to broad tissue  
175 classes (**Figure 1D**). Based on similarities in motif associations, hierarchical clustering  
176 grouped all stele clusters (1, 2, and 11), epidermis and cortex (clusters 0 and 3), two  
177 endodermis clusters (4 and 10), and another endodermis cluster with epidermal  
178 precursor cells (clusters 7 and 8). That motif associations alone can distinguish among  
179 clusters and group similar ones together provides independent verification of the cell  
180 type-specific nature of the chromatin-accessible sites detected in the scATAC-seq  
181 data.

182

### 183 **Epidermal cell layers show increased levels of endoreduplication**

184

185 In contrast to scRNA-seq data, scATAC-seq data can provide insight into DNA  
186 copy number and its impact on gene regulation. DNA copy number is of special  
187 relevance in the *A. thaliana* root, as each cell layer undergoes different rates of  
188 endoreduplication.<sup>19</sup> In a diploid cell, a single accessible locus tends to show 1 or 2  
189 transposition events. In polyploid cells with higher DNA copy number, a single  
190 accessible locus could show 4, 8, or even 16 transpositions. Therefore, cells containing  
191 a large number of peaks with >1 transposition event are likely to represent  
192 endoreduplicated cells. To identify such cells, we classified each cell by the mean  
193 number of cuts it contained per peak and examined the distribution of this metric to  
194 draw a threshold above which cells were classified as likely endoreduplicated (**Figure**  
195 **S5A, S5B**). We examined the fraction of likely endoreduplicated cells per cell type and  
196 compared these fractions to orthogonal measurements of endoreduplication. We found  
197 the expected trend of higher endoreduplication in the outermost cell files, with reduced  
198 prevalence in the stele (**Figure S5C**). Endoreduplicated cells also showed less total  
199 complexity in accessible genes, consistent with their increased developmental  
200 progression (**Figure S3G, S3H**).<sup>2</sup>

201

202

### 203 **Integration of scATAC and scRNA-seq data improves cell type annotation**

204

205 Because scATAC-seq data both identified known root cell types and provided  
206 novel cell identity assignments not identifiable through scRNA-seq, we addressed  
207 whether combining these two data sets results in additional insights than what could  
208 be gained from either alone. We first addressed whether both data types could be  
209 embedded in the same low-dimensional space in a manner that maintains the cell  
210 identities defined by both scATAC-seq and scRNA-seq. Such embedding assumes  
211 that the underlying cell identities represented in each dataset are similar. In this case,  
212 the root tissue sampled for the scATAC-seq experiment and previous scRNA-seq  
213 experiments was similar and therefore should represent similar numbers and types of  
214 cells. Moreover, the data generated by both methods share “gene” as a feature, *i.e.*  
215 accessibility near or within a given gene; expression of a given gene.

216

217 We used the anchor-based multimodal graph alignment tool from the Seurat  
218 package to find nearest-neighbor scRNA-seq matches for each cell in the scATAC-seq  
219 data.<sup>29,30</sup> In short, the tool identifies representative features (shared “anchor” genes in  
220 our case) in each dataset and looks for underlying correlation structure of those  
221 features to group similar cells in a co-embedded space. We plotted all cells within the  
222 resulting co-embedded space with cell type labels from each dataset separately. Cells  
223 derived from scRNA-seq and scATAC-seq experiments were well mixed (**Figure 2A**).  
224 Moreover, we found that cells of the same type were co-localized independent of the  
225 source data (**Figure 2B, 2C**), though some separation by data type was apparent, likely  
226 owing to the imputation step of dataset integration.<sup>29</sup> This result suggests that RNA and  
227 ATAC signals, which are only poorly correlated in bulk studies, are capable of grouping  
228 cell identities when determined in individual cells of a complex tissue. We further used  
229 this co-embedded space to refine our earlier manual cell type annotations by  
230 transferring labels of neighboring scRNA cells onto the scATAC cells (**Figure S2B**);  
231 while most of these labels matched, the greatest number of mismatches was seen in  
232 endodermis sub-type 3. The transferred labels matched our manual annotations, and,  
233 in the case of epidermal cells, allowed us to separate a single ATAC cluster into hair  
234 and non-hair cells (**Figure 2A, Figure S2A**). The three distinct ATAC clusters that were  
235 assigned an “endodermis” label with this approach are a striking example of scATAC  
236 data yielding greater stratification of cell types than the generally richer scRNA data.  
237

### 238 **scATAC-seq captures three distinct endodermis types representing different 239 developmental stages**

240

241 We dissected the three endodermis clusters in greater detail using three  
242 approaches: (i) by identifying differentially accessible sites among subtypes; (ii) by  
243 aligning these subtypes to scRNA-seq data that have been annotated for  
244 endoreduplication and developmental progression; and (iii) by determining differentially  
245 expressed genes in the nearest-neighbors to each of these endodermis subtypes in  
246 scRNA-seq space (**Figure 3A**).  
247

248

249 We identified few differentially accessible peaks genes (adjusted p-value < 0.05  
250 and at least 2-fold change in accessibility) in each endodermis subtype: 25 for the first  
251 subtype, 24 for the second, and 17 for the third (**Figure 3A**). The low number of  
252 associated genes precluded gene set enrichment analyses, but genes uniquely  
253 accessible in subtype 1 included transcription factors *NAC010* (AT1G28470) and  
254 *MYB85* (AT4G22680) as well as genes involved in suberization (*FAR1*, *FAR4*, *FAR5*).  
255 Endodermis subtype 2 showed increased accessibility at *ANAC038* (AT2G24430),  
256 *HIPP04* (AT1G2900), encoding a heavy metal-associated protein, and phenylpropanoid  
257 metabolism genes. Endodermis subtype 3 showed strong accessibility at the *BLUEJAY*  
258 (AT1G14580) locus encoding a C2H2 transcription factor implicated in endodermis  
differentiation (**Figure 3B, S6A**), as well as at genes for phenylpropanoid biosynthesis.

259 We addressed whether these differentially-accessible genes show different expression  
260 patterns in endodermis cells in scRNA-seq space by mapping expression of each gene  
261 onto a subclustered set of endodermis cells combined from several scRNA-seq studies  
262 of the *A. thaliana* root. The small set of marker genes for each scATAC subtype  
263 showed no consistent pattern in the scRNA-seq data (**Figure S3C**), suggesting that  
264 some other feature distinguished these three subtypes.  
265

266 Structure within two-dimensional embeddings of scRNA-seq and scATAC-seq  
267 data derived from developing tissues is often associated with developmental  
268 progression or other asynchronous processes like the cell cycle. Furthermore, root  
269 tissue has the unique feature of being highly endoreduplicated, which could also  
270 account for differences among the subtypes. To assess whether the endodermal  
271 subtypes were associated with these features, we added annotations for cell cycle,  
272 developmental progression and endoreduplication to the combined root scRNA-seq  
273 data and used data integration (as in **Figure 2**) to test whether cells from the  
274 endodermal subtypes were associated with any of these features (**Figure S2C**).  
275

276 We used a list of known cell-cycle marker genes to generate a signature score  
277 marking proliferating cells (Arabidopsis.org). This signature score identified cycling  
278 cells in other cell types, such as early epidermis cells near the quiescent center (**Figure**  
279 **S4A, S4B**), but showed no difference in the nearest-neighbor cells corresponding to  
280 each epidermis subtype (**Figure S4C**). We conclude that cell cycle does not distinguish  
281 the epidermis subtypes.  
282

283 We assessed developmental progression with two orthogonal methods: (i)  
284 correlation with published bulk expression data taken along longitudinal sections of the  
285 root;<sup>1</sup> and (ii) a modified measure of loss in transcriptional diversity (see Methods),  
286 which correlates strongly with developmental progression in a large number of scRNA-  
287 seq datasets, including of the *Arabidopsis* root.<sup>2,31</sup> We found that the developmental  
288 progression metric as measured by loss in transcriptional diversity was strongly  
289 associated with the orthogonal correlation-based classification (**Figure S3A**).<sup>31</sup> For  
290 each cell of the endodermal subtypes, we calculated the average developmental  
291 progression of its 25 nearest neighbors among root scRNA-seq cells (**Figure S3H, S3J**)  
292 and found, assigning this average to each ATAC endodermis cell, a trend of  
293 developmental progression among the endodermis sub-types (**Figure 3C**). This result  
294 was robust to changes in the number of neighbors used to identify similar cells from  
295 scRNA-seq data (**Figure S3D**). This trend was the same if we calculated the  
296 developmental progression metric based on scATAC-seq data alone (**Figure S3F**).<sup>31</sup>  
297 Cells from subtype 1 were the least developed, while cells from subtype 3 tended to  
298 co-occur with the most mature endodermal cells in the co-embedded graph (**Figure**  
299 **3C**). We conclude that the three endodermal subtypes primarily represent cells of  
300 differing developmental progression and that differences in chromatin accessibility are  
301 able to capture this stratification of endodermis maturity.

302

303 Developmental progression in the root is often associated with increased ploidy  
304 through endoreduplication. To identify endoreduplicated cells in scRNA-seq data, we  
305 used a published set of marker genes for ploidy to generate signature scores for 2n,  
306 4n, 8n and 16n ploidies.<sup>19</sup> With these scores, we predicted endoreduplicated cells by  
307 calculating, for each cell, the ratio of the 8n signature relative to the diploid signature.  
308 Similar to the DNA-based metric, this transcriptional approach identified  
309 endoreduplicated root cells in the expected pattern (**Figure S3B, S3E**), with higher  
310 fractions in the epidermis cell layer and diminished levels in the stele (**Figure S5D**).  
311 Because the DNA-based metric showed poorer correlation to prior data and was less  
312 sensitive (**Figure S3F, S3G**), we used the transcriptionally-based metric in subsequent  
313 analyses. This metric captured an abundance of tetraploid xylem cells in the stele  
314 (**Figure S5E**), consistent with previous findings.<sup>19</sup> With confidence in this classifier of  
315 endoreduplicated cells, we examined the predicted ploidy for the nearest RNA-seq  
316 neighbors of each endodermis subtype (**Figure S3I**). We found that the younger  
317 endodermis subtype 1 cells had mostly 2n neighbor cells, while the more mature  
318 subtypes 2 and 3 had mostly endoreduplicated neighbor cells, with similar levels in  
319 each (**Figure 3D**).  
320

321

322 To better understand the differing transcriptional and chromatin accessibility  
323 patterns among endodermis subtypes, we predicted differentially expressed genes  
324 for each endodermis subtype (**Figure S2B**). The early endodermis type, which is not yet  
325 endoreduplicated showed an enrichment of genes (**Supplementary Table 3**) involved  
326 in Caspary strip formation (CASP3, CASP5) and wax biosynthesis (HHT1). The  
327 intermediate subtype 2 also showed enrichment for genes involved in Caspary strip  
328 formation (CASP3, CASP4, CASP5, GSO1), as well as mechanosensitive ion channels  
329 (MSL4, MSL6, MSL10) (**Supplementary Table 4**). The most advanced endodermis  
330 subtype 3 showed enrichment for stress responses and metabolism of toxic  
331 compounds, kinase activity, and high levels of aquaporin water channels  
332 (**Supplementary Table 5**), consistent with this mature endodermis cell type modulating  
333 water permeability via aquaporins as well as through suberization.<sup>32</sup> We also identified  
334 putative regulators of these stages by looking for transcription factors among the  
335 genes that showed specificity for each endodermis cluster. The early endodermis type  
336 showed a single upregulated transcription factor, ERF54, while the intermediate  
337 subtype showed 14 upregulated transcription factors, including KNAT7, SOMNUS, and  
338 HAT22. MYB36, which was found expressed in the late endodermis type, activates  
339 genes involved in Caspary strip formation and regulates a crucial transition toward  
340 differentiation in the endodermis.<sup>33</sup>

341

342 Overall, the combined information gained from transcriptional signatures of  
343 developmental progression and endoreduplication highlights the importance of  
344 integrating both open chromatin and transcriptional profiling to identify cell types or  
cell states that may have otherwise been obscured in a single data type.

345

## 346 Predicting regulatory events using integrated scRNA and scATAC data

347

348 We previously identified transcription factor binding motifs that were enriched at  
349 cell type-specific peaks in the root (**Figure 1D**). While individual motifs may be  
350 associated with binding and activation by transcription factors, a sequence-level  
351 analysis cannot distinguish among the many members of plant transcription factor  
352 families that share near-identical sequence preferences. For example, WRKY family  
353 motifs were highly enriched among epidermis and cortex accessible sites, but this  
354 family contains >50 individual genes. In order to narrow down this list of genes to a few  
355 possible candidates, we leveraged our nearest-neighbor annotation approach (**Figure**  
356 **S2C**) to examine expression levels of all WRKY family transcription factors in the  
357 scATAC data (**Figure 4A**). Overall, we found that the majority of WRKY members  
358 showed expression in the epidermis, cortex or epidermal precursor cells (**Figure 4A**),  
359 though some members showed stele-specific expression. To identify the most likely  
360 members to bind the abundance of motifs in epidermis-specific peaks, we ranked  
361 these genes by their specificity in the epidermis. The top four most specific genes,  
362 *WRKY75*, *WRKY9*, *WRK6*, and *TTG2*, have documented roles in root development.<sup>28,34-</sup>  
363 <sup>36</sup> *TTG2* shows strong specificity for the epidermis, but we also predict expression in  
364 some cortex and precursor cells (**Figure 4B**). Two key interacting factors of *TTG2* that  
365 also contribute to epidermis development, *GL2* and *TTG1*,<sup>37,38</sup> showed epidermis  
366 expression and had correlated (Pearson correlation with *TTG2* across cells for *GL2* =  
367 0.91, and *TTG1* = 0.47) patterns across all cells (**Figure S6B, S6C**).  
368

369

370 Given the important role of *TTG2* in specification of atrichoblast fate in the  
371 epidermis, we examined the consequences of its expression on accessibility of  
372 individual peaks. Inference of individual regulatory events, particularly those involving  
373 transcription factors, has long been a goal of studies that profile accessibility at  
374 regulatory sites in bulk tissue. The varied cell states revealed by single-cell profiling  
375 data, even those within a cell type, allow higher-resolution inference of these events.  
376 To identify accessible sites that showed altered accessibility as a function of  
377 transcription factor expression, we used a linear regression approach. We identified  
378 617 peaks that showed significant ( $q\text{-value} < 0.05$ ) associations with *TTG2* expression  
379 levels (**Supplementary Table 6**). To visualize these associations using scATAC data,  
380 we pseudobulked epidermis, cortex, and c/e precursor cells into four equal-sized bins  
381 based on their level of *TTG2* expression (**Figure 4C**). Most significant associations  
382 were positive, such that increased *TTG2* expression led to increased peak accessibility  
383 (**Figure 4C**, top and lower-left panels), though negative associations could also be  
384 identified (**Figure 4C**, lower-right panel). Positive associations occurred whether or not  
385 a WRKY binding motif was present in the associated peak (**Figure 4C**), suggesting that  
386 the role of WRKY transcription factors in specification of the epidermis likely requires  
387 both direct and indirect regulatory events. Of peaks with significant ( $q\text{-value} < 0.05$ )  
388 positive associations with *TTG2* expression, 80% of these contained a WRKY binding

388 motif, while only 38% of the peaks with negative associations contained a binding  
389 motif (**Figure 4D**). Overall, this analysis identifies transcription factors and putative  
390 target sites that constitute regulatory events important for specifying cell types; these  
391 genes and regulatory sites are good candidates for further functional studies.  
392

### 393 Discussion

394

395 By profiling chromatin accessibility in the *A. thaliana* root at single-cell  
396 resolution, we assessed cell types, developmental stages, the transcription factors  
397 likely driving these stages and DNA copy number changes. We assigned over 5,000  
398 root cells to tissues and cell types, demonstrating that these assignments are  
399 concordant with single-cell transcriptomic studies. These results answer an unresolved  
400 question in plant gene regulation: does the paucity of dynamic open chromatin sites  
401 seen in bulk profiling experiments represent an accurate reflection of uniform gene  
402 regulation in *A. thaliana* or does it reflect a confounding effect of bulk studies? We  
403 found that distinct root cell types show unique patterns of open chromatin sites, with  
404 approximately 1/3 of all accessible sites showing cell type-specific patterns. This  
405 estimate greatly exceeds the earlier estimates from bulk studies of only 5-10% of  
406 accessible sites showing tissue- or condition-specificity,<sup>9</sup> presumably due in part to  
407 tissue heterogeneity.  
408

409 Although this single-cell ATAC study discovered many more dynamic accessible  
410 sites, the correlation between dynamic accessibility and gene expression in single cells  
411 remained poor, reminiscent of the equally poor correlation seen in bulk studies. We  
412 argue that the poor correlation between chromatin accessibility and gene expression is  
413 not a function of data quality. Instead, we propose that this weak correlation reflects  
414 the complex nature of regulatory processes underlying development. Although the  
415 correlation of chromatin accessibility and gene expression is weak at the level of  
416 individual loci, either the entirety of a cell's regulatory landscape or its transcriptome  
417 can independently capture its cell identity. It is this feature that allows joint co-  
418 embedding of both data types and the use of scRNA-seq data to annotate scATAC  
419 cells.  
420

421 Thus, while the patterns of both chromatin accessibility and gene expression  
422 contain information on cell identity and development, the relationships between these  
423 patterns are not well-ordered or parsimonious. For the many cells belonging to a  
424 distinct cell type, gene expression results from direct and indirect regulatory events  
425 involving tens or hundreds of transcription factors and chromatin remodelers that do  
426 not necessarily act in concert. For any individual locus, then, the expectation that  
427 average accessibility predicts average expression breaks down. Without a simple one-  
428 to-one model to explain regulatory output, we are left with significant heterogeneity  
429 within and between cell types, and a subset of convergent expression or accessibility  
430 patterns that define cell type specificity. Alternative explanations for the discrepancy in  
431 accessibility and expression include: (1) maintenance of cell identity requires that a

432 cell's accessibility and expression profile stably reflect the convergent pattern for that  
433 cell type only a fraction of the time; and/or (2) cells have multiple accessibility and  
434 expression patterns that are sufficient to maintain cell identity and together constitute  
435 the convergent patterns we observe. In both scenarios, the heterogeneity in cell type  
436 specification will be buffered by factors outside chromatin accessibility or gene  
437 expression, such as spatial location in tissue, metabolic determinants of cell function or  
438 developmental age.

439

440 We posit that scATAC-seq data combined with scRNA-seq data will ultimately  
441 resolve these alternatives by enabling mechanistic models of gene regulatory  
442 networks. scATAC-seq data alone are sufficient to identify the full set of accessible  
443 sites in the *Arabidopsis* genome, and examination of the transcription factor motifs  
444 within these sites can enable predictions of regulatory networks. However, many plant  
445 transcription factor families are large, some containing over fifty members that  
446 recognize near identical motifs. Thus, the accessibility data must be integrated with  
447 single-cell expression data that capture cell type-specific expression of transcription  
448 factors in order to narrow down the most probable transcription factors that are  
449 enacting individual regulatory events. Building high resolution models of key regulatory  
450 events will require the expression level of individual transcription factors in a cell type,  
451 the accessibility of individual peaks in this cell type and the presence of binding motifs  
452 corresponding to the relevant transcription factors. Theoretically, a comprehensive  
453 capture of cell states with both open chromatin and transcriptional profiling will allow  
454 the ordering of gene regulatory events and the larger scale ordering of regulatory  
455 programs that underlie development. The ability to take single-cell measurements over  
456 distinct developmental stages will also increase the sampling of key regulatory events.  
457 Ultimately, achieving the goal of building models of gene regulatory events underlying  
458 development will require ever larger datasets to fully capture the range of possible cell  
459 states.

460

461 In the future, single-cell studies of more complex plant tissues in crops and  
462 other species will necessitate larger numbers of profiled cells and higher numbers of  
463 cuts per cell. In this way, approaches that maximize the number of cells profiled at low  
464 cost, such as single-cell combinatorial indexing,<sup>39</sup> will be critical. Annotation in future  
465 studies will also present a substantial challenge if a rich literature and genomic  
466 analyses, including single-cell transcriptome profiles, are not available. Nevertheless,  
467 as shown in this proof-of-principle study of the well-characterized *A. thaliana* root, the  
468 knowledge gained should eventually allow us to manipulate gene expression and  
469 organismal phenotype in a targeted manner.

470

471

## 472 **Methods**

473

### 474 *Plant Material*

475 *Genotype: Arabidopsis thaliana* ecotype Col-0 INTACT line *UBQ10:NTF::ACT2:BirA*  
476 (available from ABRC, stock CS68649). *Growth conditions:* LD (16h light/8h dark), 22C,  
477 ~100 µmol m<sup>2</sup>s, 50% RH. *Sample:* whole roots, harvested 12 days after germination,  
478 from seedlings grown vertically on MS + 1% sucrose, atop filter paper (to facilitate root  
479 harvesting).

480

481 *Nuclei Isolation and snATAC-seq*

482 Nuclei were isolated following a modified version of the protocol described in Giuliano  
483 et al., 1988, as follows: 1g of roots was split in two batches of 0.5g, and each batch  
484 chopped with a razor blade in 1 ml of Buffer A (0.8M sucrose, 10mM MgCl<sub>2</sub>, 25mM  
485 Tris-HCl pH 8.0 and 1x Protease Inhibitor Tablet).<sup>40</sup> Extracts were combined, final  
486 volume increased to 5ml with Buffer A, and incubated on ice for 10min, with gentle  
487 swirling. The combined extract was filtered through miracloth, passed through a 26ga  
488 syringe five times and re-filtered through a 40um cell strainer (BD Falcon). After  
489 centrifugation at 2,000g 5min, the pellet was resuspended in 1ml Buffer B (0.4M  
490 sucrose, 10mM MgCl<sub>2</sub>, 25mM Tris-HCl pH 8.0, 1x Protease Inhibitor Tablet, 1% Triton  
491 X - 100) and loaded atop a 2-step 25/75 Percoll gradient ( 1 volume 25% Percoll in  
492 Buffer B over 1 volume 75% Percoll in Buffer B). After centrifugation at 2,500g for  
493 15min, nuclei were collected either at the 25/75 interface or in the subjacent 75  
494 fraction, washed with 5 vols of Buffer B and recovered by centrifugation at  
495 1,700g 5min. The nuclei pellet was resuspended in 100ul Buffer B + 1% BSA and any  
496 nuclei clumps broken down by pipetting up and down multiple times. Nuclei yield with  
497 this protocol was ~ 94,000 nuclei per gram of roots (fresh weight).

498 snATAC-seq libraries were built using the 10x Genomics Chromium Single Cell ATAC  
499 Solution platform, following manufacturer's recommendations. Before transposition,  
500 nuclei were spun 5min at 1,500g and resuspended in 10x Genomics Diluted Nuclei  
501 Buffer, at a concentration of 3,200 nuclei/ul. 5ul of nuclei suspension were used for  
502 transposition (16,000 nuclei being the maximum input recommended for 10x  
503 Chromium, and 10,000 nuclei being the expected recovery).

504 *Combining and processing of root scRNA-seq data*

505 Samples were processed using the CellRanger vX.X pipeline from 10X Genomics,  
506 including updated filtering of "halflet" cells that emerge due to multiply-barcoded  
507 droplets.

508

509 *Integration of scRNA and scATAC data*

510 The R package Seurat version 3.1.5 was used to align and co-embed the scATAC-seq  
511 data with scRNA-seq data published by Ryu et al. 2019, and to transfer cell type labels  
512 from the scRNA data to the scATAC data.<sup>30,41</sup>

513

514 The standard workflow and default parameters as described in the Seurat vignette  
515 "PBMC scATAC-seq Vignette" ([satijalab.org/seurat/v3.1/atacseq\\_integration\\_vignette](http://satijalab.org/seurat/v3.1/atacseq_integration_vignette))  
516 were used with the exception that all features (genes) were used when identifying  
517 transfer anchors and performing the co-embedding rather than a set of "variable"  
518 features as used in the vignette. Briefly this workflow is as follows:

519 An anchor set was established with the function FindTransferAnchors() linking the two  
520 datasets . Cell type annotations were transferred from the scRNA-seq data to the  
521 scATAC data using the function TransferData(). Pseudo RNA-seq count data was  
522 generated for the scATAC cells, again using the TransferData() function. The pseudo  
523 RNA data was then merged with the true scRNA-seq dataset and embedded in 2D  
524 UMAP space using Seurat functions.  
525

526 A co-embedding was performed with a super-set of scRNA-seq data published by  
527 Jean-Baptiste et al. 2019, Shulse et al. 2019, Ryu et al. 2019. In the co-embedded  
528 space the scATAC-seq were found to be most closely co-located with data from  
529 Ryu2019. Based on this observation co-embedding was performed with the Ryu2019  
530 dataset on its own.  
531

532  
533 *Nearest neighbor analysis for transcriptional characterization of cells identified in*  
534 *scATAC assay*  
535

536 To annotate cells from the scATAC-seq assay with transcriptional features, we used  
537 average feature values from the nearest RNA neighbors in our co-embedded data  
538 (**Figure 2A**). In short, the ‘distances’ package in R was used to extract cell labels for  
539 the 25 nearest neighbors of each scATAC cell. For a feature of interest (individual gene  
540 expression, cell-cycle signature score, endoreduplication signature score,  
541 developmental progression signature), we calculated the mean expression from the 25  
542 scRNA cells, and assigned that mean score to each ATAC cell (**Figure S2C**).  
543

544 *Motif analysis*  
545

546 Position weight matrices from the comprehensive DAP-seq dataset<sup>27</sup> were used as  
547 input into FIMO<sup>42</sup> to search for significant matches for each motif (adjusted p-value  
548 threshold < 1e-5) in each of the scATAC peaks. With the output of this motif scan, we  
549 generated a matrix that tallied counts of each motif within each peak. To identify motifs  
550 whose counts were significantly associated with cell type-specific accessibility, we first  
551 generated, for each peak, a relative accessibility score by taking the mean accessibility  
552 of that peak in each cell cluster relative to the overall accessibility of that peak in all  
553 clusters. Next, we used a linear regression framework within Monocle3<sup>43</sup> to identify  
554 individual motifs whose counts showed strong positive or negative correlations with  
555 the cell type-specific accessibility score in each cell cluster. The effect size of each  
556 motif’s contribution to cell type-specific accessibility is given as the  $\beta$  of the linear  
557 regression, shown as a mean across all transcription factors in the same family.  
558

559  
560 **Data Availability**

561 Source data for all figures are available via Dryad (accession number pending).  
562 Expression data are available at the Gene Expression Omnibus (GEO number:  
563 pending).

564 **Acknowledgements**

565 We thank Dr. Ken Jean-Baptiste and Dr. Kerry Bubb for valuable discussions on ATAC-  
566 seq analysis. We also thank Xavi Guitart for helpful discussions on endoreduplication.  
567 This work was supported by the National Science Foundation (RESEARCH-PGR grant  
568 17488843) to S.F. and C.Q. This work was also supported by NIH grant  
569 1RM1HG010461 to C.Q. and S.F.

570

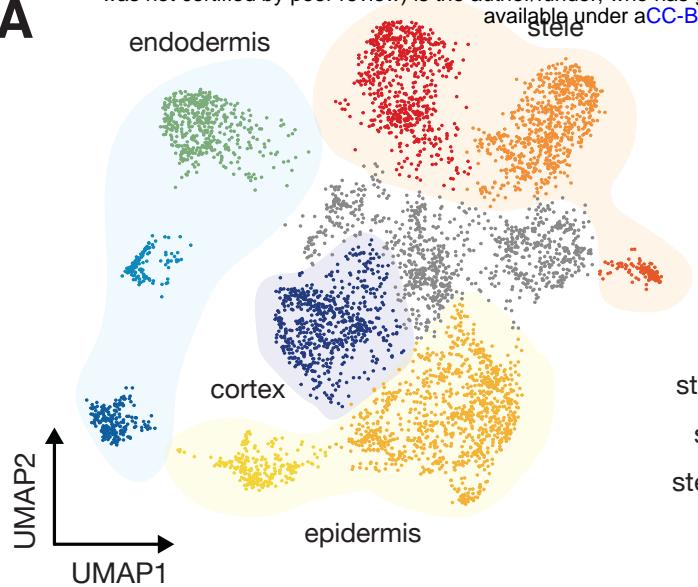
571 **References**

- 573 1. Brady, S. M. *et al.* A High-Resolution Root Spatiotemporal Map Reveals  
574 Dominant Expression Patterns. *Science* (80-. ). **318**, 801–806 (2007).
- 575 2. Jean-Baptiste, K. *et al.* Dynamics of gene expression in single root cells of *A.*  
576 *thaliana*. *Plant Cell* **31**, tpc.00785.2018 (2019).
- 577 3. Shulse, C. N. *et al.* High-Throughput Single-Cell Transcriptome Profiling of Plant  
578 Cell Types. *Cell Rep.* **27**, 2241-2247.e4 (2019).
- 579 4. Zhang, T. Q., Xu, Z. G., Shang, G. D. & Wang, J. W. A Single-Cell RNA  
580 Sequencing Profiles the Developmental Landscape of Arabidopsis Root. *Mol.*  
581 *Plant* **12**, 648–660 (2019).
- 582 5. Ryu, K. H., Huang, L., Kang, H. M. & Schiefelbein, J. Single-cell RNA sequencing  
583 resolves molecular relationships among individual plant cells. *Plant Physiol.* **179**,  
584 1444–1456 (2019).
- 585 6. Denyer, T. *et al.* Spatiotemporal Developmental Trajectories in the Arabidopsis  
586 Root Revealed Using High-Throughput Single-Cell RNA Sequencing. *Dev. Cell*  
587 **48**, 840-852.e5 (2019).
- 588 7. Sullivan, A. M. *et al.* Mapping and Dynamics of Regulatory DNA in Maturing  
589 Arabidopsis thaliana Siliques. *Front. Plant Sci.* **10**, 1–16 (2019).
- 590 8. Alexandre, C. M. *et al.* Complex relationships between chromatin accessibility,  
591 sequence divergence, and gene expression in arabidopsis thaliana. *Mol. Biol.*  
592 *Evol.* **35**, 837–854 (2018).
- 593 9. Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A. &  
594 Queitsch, C. DNase I hypersensitivity mapping, genomic footprinting, and  
595 transcription factor networks in plants. *Curr. Plant Biol.* **3–4**, 40–47 (2015).
- 596 10. Reynoso, M. A. *et al.* Evolutionary flexibility in flooding response circuitry in  
597 angiosperms. *Science* (80-. ). **365**, 1291–1295 (2019).
- 598 11. Maher, K. A. *et al.* Profiling of accessible chromatin regions across multiple plant  
599 species and cell types reveals common gene regulatory principles and new  
600 control modules. *Plant Cell* **30**, 15–36 (2018).
- 601 12. Saunders, L. M. *et al.* Thyroid hormone regulates distinct paths to maturation in

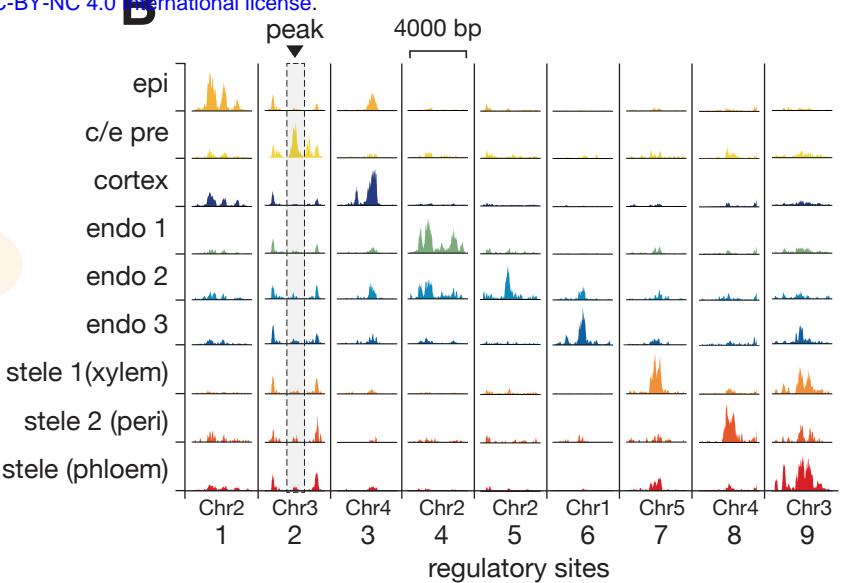
- 602 pigment cell lineages. *Elife* **8**, (2019).
- 603 13. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of  
604 developmental potential. *Science* (80-. ). **367**, 405–411 (2020).
- 605 14. Waddington, C. H. The Strategy of the Genes. (1959).
- 606 15. Orr-Weaver, T. L. When bigger is better: The role of polyploidy in organogenesis.  
607 *Trends Genet.* **31**, 307–315 (2015).
- 608 16. Derkx, W. & Bergmann, O. Polyploidy in cardiomyocytes: Roadblock to heart  
609 regeneration? *Circ. Res.* **126**, 552–565 (2020).
- 610 17. Lang, L. & Schnittger, A. Endoreplication — a means to an end in cell growth and  
611 stress response. *Curr. Opin. Plant Biol.* **54**, 85–92 (2020).
- 612 18. Pirrello, J. et al. Transcriptome profiling of sorted endoreduplicated nuclei from  
613 tomato fruits: how the global shift in expression ascribed to DNA ploidy  
614 influences RNA-Seq data normalization and interpretation. *Plant J.* **93**, 387–398  
615 (2018).
- 616 19. Bhosale, R. et al. A spatiotemporal dna endopolyploidy map of the arabidopsis root  
617 reveals roles for the endocycle in root development and stress adaptation. *Plant  
618 Cell* **30**, 2330–2351 (2018).
- 619 20. Robinson, D. O. et al. Ploidy and Size at Multiple Scales in the Arabidopsis  
620 Sepal. *Plant Cell* **30**, 2308–2329 (2018).
- 621 21. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of  
622 communities in large networks. *J. Stat. Mech. Theory Exp.* **10**, P10008 (2008).
- 623 22. Shu, H., Wildhaber, T., Siretskiy, A., Grussem, W. & Hennig, L. Distinct modes of  
624 DNA accessibility in plant chromatin. *Nat. Commun.* **3**, (2012).
- 625 23. Sullivan, A. M. et al. Mapping and dynamics of regulatory DNA and transcription  
626 factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).
- 627 24. McFaline-Figueroa, J. L., Trapnell, C. & Cuperus, J. T. The promise of single-cell  
628 genomics in plants. *Curr. Opin. Plant Biol.* **54**, 114–121 (2020).
- 629 25. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in  
630 thousands of single cells. *Science* (80-. ). **1385**, 1380–1385 (2018).
- 631 26. Spitz, F. & Furlong, E. E. M. Transcription factors: From enhancer binding to  
632 developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
- 633 27. O'Malley, R. C. et al. Cistrome and Epicistrome Features Shape the Regulatory  
634 DNA Landscape. *Cell* **165**, 1280–1292 (2016).
- 635 28. Johnson, Cameron S.; Kolevski Ben, and S. D. R. TRANSPARENT TESTA  
636 GLABRA2 , a Trichome and Seed Coat Development Gene of Arabidopsis,  
637 Encodes a WRKY Transcription Factor. *Plant Cell* **14**, 1359–1375 (2017).
- 638 29. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–  
639 1902.e21 (2019).
- 640 30. Butler, A., Hoffman, P., Smibert, P., Papalex, E. & Satija, R. Integrating single-  
641 cell transcriptomic data across different conditions, technologies, and species.  
642 *Nat. Biotechnol.* **36**, 411–420 (2018).
- 643 31. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of  
644 developmental potential. *Science* (80-. ). **367**, 405–411 (2020).

- 645 32. Kreszies, T., Schreiber, L. & Ranathunge, K. Suberized transport barriers in  
646 Arabidopsis, barley and rice roots: From the model plant to crop species. *J. Plant*  
647 *Physiol.* **227**, 75–83 (2018).
- 648 33. Liberman, L. M., Sparks, E. E., Moreno-Risueno, M. A., Petricka, J. J. & Benfey,  
649 P. N. MYB36 regulates the transition from proliferation to differentiation in the  
650 Arabidopsis root. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12099–12104 (2015).
- 651 34. Devaiah, B. N., Karthikeyan, A. S. & Raghothama, K. G. WRKY75 transcription  
652 factor is a modulator of phosphate acquisition and root development in  
653 Arabidopsis. *Plant Physiol.* **143**, 1789–1801 (2007).
- 654 35. Chen, Y. F. et al. The WRKY6 transcription factor modulates PHOSPHATE1  
655 expression in response to low pi stress in arabidopsis. *Plant Cell* **21**, 3554–3566  
656 (2009).
- 657 36. Zheng, X. et al. MdWRKY9 overexpression confers intensive dwarfing in the M26  
658 rootstock of apple by directly inhibiting brassinosteroid synthetase MdDWF4  
659 expression. *New Phytol.* **217**, 1086–1098 (2018).
- 660 37. Long, Y. & Schiefelbein, J. Novel TTG1 Mutants Modify Root-Hair Pattern  
661 Formation in Arabidopsis. *Front. Plant Sci.* **11**, 1–12 (2020).
- 662 38. Schiefelbein, J., Kwak, S. H., Wieckowski, Y., Barron, C. & Bruex, A. The gene  
663 regulatory network for root epidermal cell-type pattern formation in arabidopsis.  
664 *J. Exp. Bot.* **60**, 1515–1521 (2009).
- 665 39. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular  
666 organism. *Science (80-. ).* **357**, 661–667 (2017).
- 667 40. Giuliano, G. et al. An evolutionarily conserved protein binding sequence  
668 upstream of a plant light-regulated gene. *Proc. Natl. Acad. Sci. U. S. A.* **85**,  
669 7089–7093 (1988).
- 670 41. Stuart, T. et al. Comprehensive Integration of Single-Cell Data Resource  
671 Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
- 672 42. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a  
673 given motif. *Bioinformatics* **27**, 1017–1018 (2011).
- 674 43. Cao, J. et al. The single-cell transcriptional landscape of mammalian  
675 organogenesis. *Nature* **566**, 496–502 (2019).
- 676

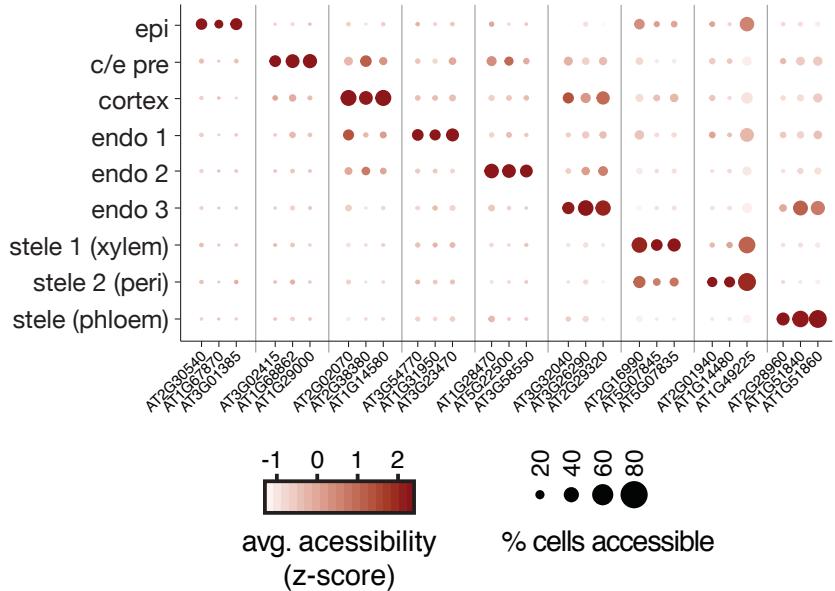
**A**



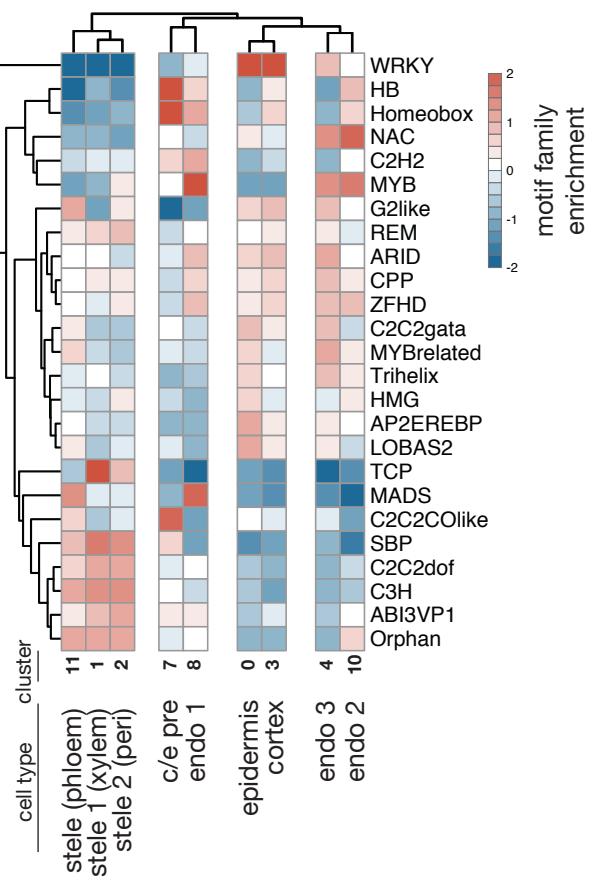
**B**



**C**



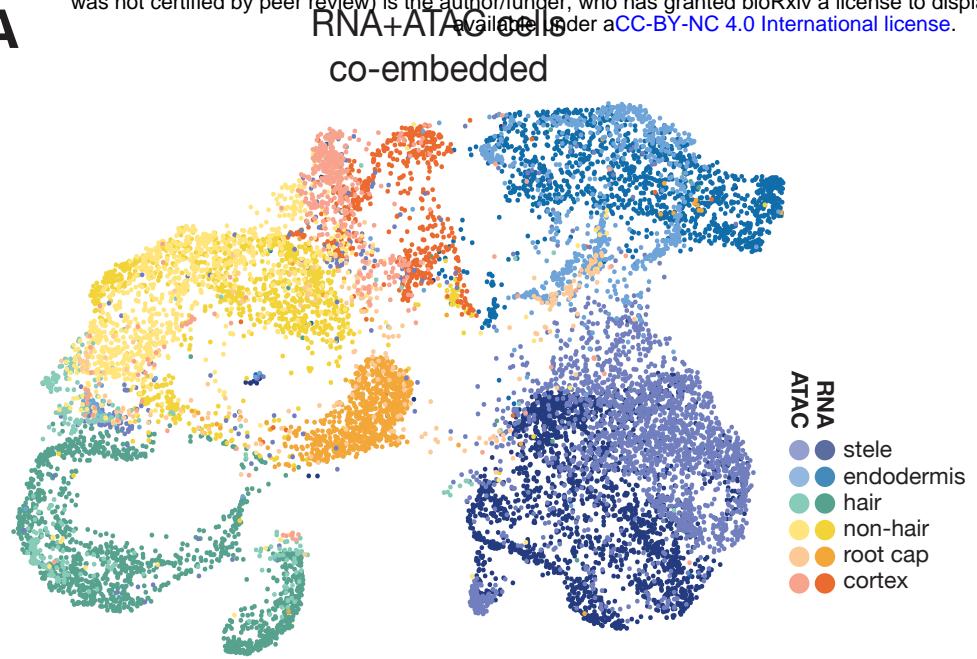
**D**



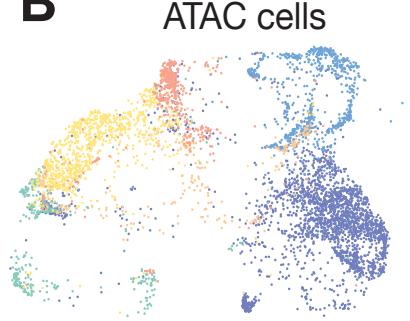
**Figure 1. scATAC-seq identifies known root cell types.**

**Figure 1. scATAC-seq identifies known root cell types.** (A) UMAP dimensionality reduction plot of root cells using peak-level scATAC data. Cells are colored according to Louvain clusters, and broad tissue types are indicated with transparent blobs. (B) Pseudo-bulked peak tracks generated by combining ATAC data from all cells within a cluster. Each column represents a single locus in the genome that shows cell type-specific accessibility; one example is shown for each cell type. Colors match those in previous panel. (C) Dotplot showing marker genes for each cell type cluster. Each column represents a single gene's activity score, the summed accessibility of its gene body and promoter sequence (-400bp from transcription start site). The color of each point indicates the magnitude of accessibility and the size of each point represents the fraction of cells in each type showing accessibility at that gene. (D) Heatmap showing the predicted effect, across all peaks, of motifs from each *Arabidopsis* transcription factor family on cell type specific accessibility. Darker shades of red indicate that presence of the motif is correlated with increased accessibility in that cell type, while shades of blue indicate that the motif is anti-correlated with accessibility. The mean effect all transcription factors within a given family are shown as rows, and each column represents a cell type.

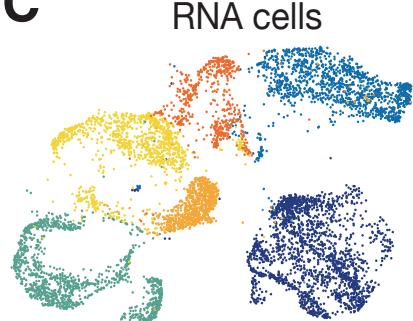
A



B



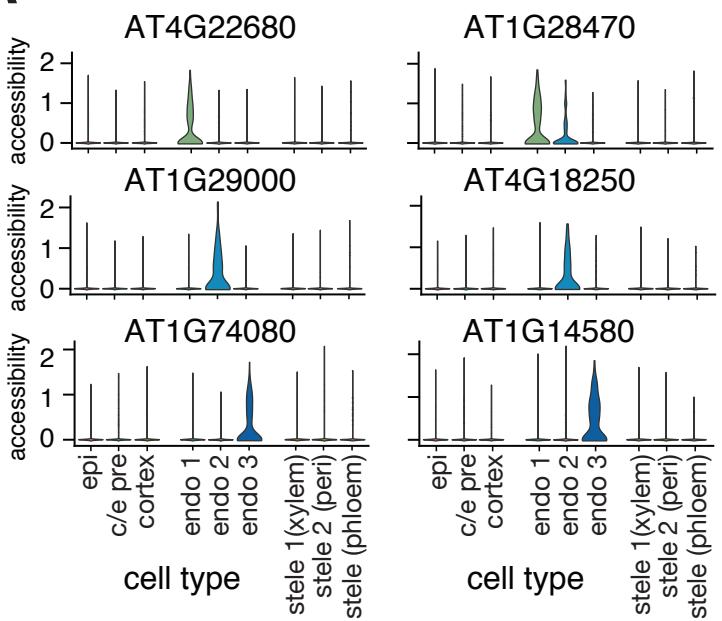
C



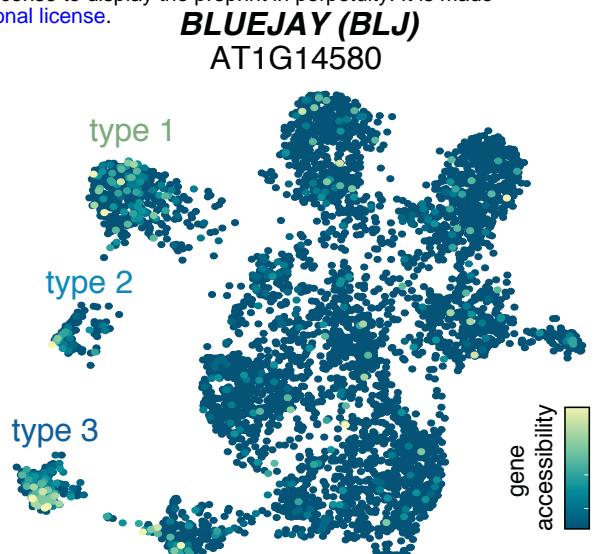
**Figure 2. scATAC-seq data can be integrated with scRNA-seq data to identify cell types.**

**Figure 2. scATAC-seq data can be integrated with scRNA-seq data to identify cell types.** (A) UMAP co-embedding of root scATAC cells alongside root scRNA cells (Schief et al). Cells are colored by broad tissue type, with scATAC cells colored in lighter shades and scRNA cells in darker shades. (B) UMAP from (A), but showing only cells from the scATAC-seq experiment; (C) shows only cells from the scRNA-seq experiment.

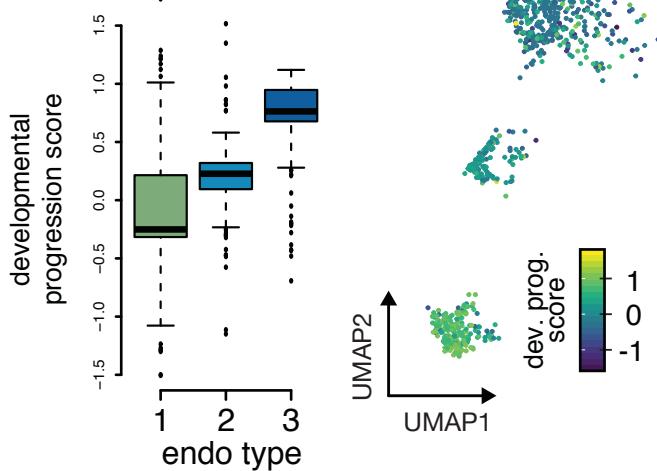
**A**



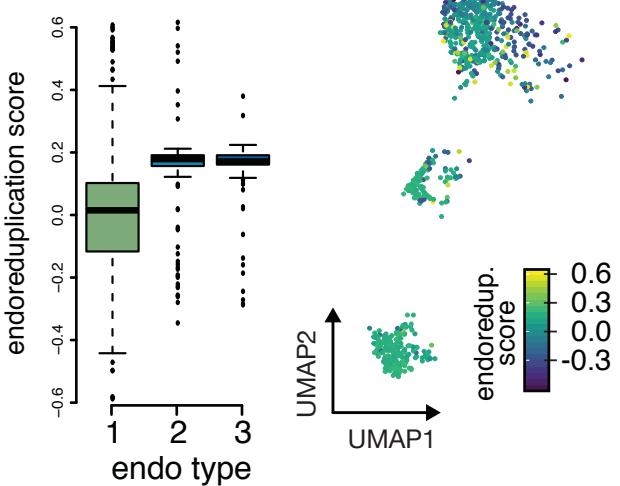
**B**



**C**



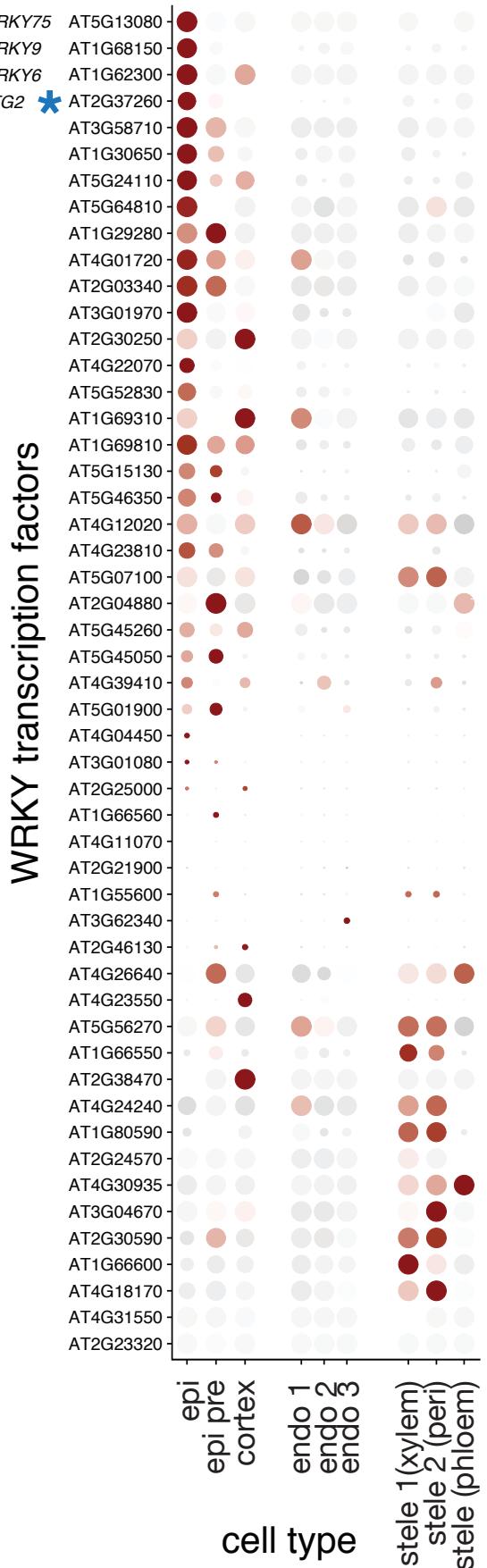
**D**



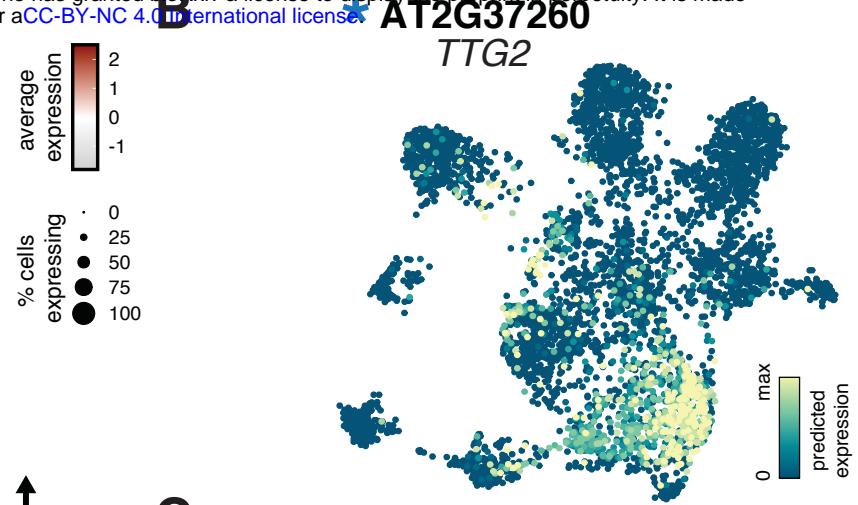
**Figure 3. scATAC-seq identifies distinct sub-types of endodermal cells.**

**Figure 3. scATAC-seq identifies distinct sub-types of endodermal cells.** (A) Violin plots showing specific patterns of accessible genes that mark each endodermal type. Two examples are given for each endodermal type, and gene-level accessibility scores are shown additionally for all other cell types. (B) UMAP of all cells colored by accessibility of the *BLUEJAY* gene, which marks endodermal type 3; corresponding violin plot for this gene in lower right panel in (A). (C) Boxplot showing an increase in median developmental progression of each endodermal type, as determined by average transcriptional complexity in the nearest 25 scRNA neighbors of each scATAC cell in the co-embedded representation from Fig. 2A; right inset shows UMAP of endodermal cells with each cell colored by the average developmental progression of its scRNA neighbors, mirroring the gradual increase seen in left panel. (D) Boxplot showing an increase in median levels of endoreduplication across endodermal types, ascertained as in (C), but instead using a gene expression signature of endoreduplication; right inset shows UMAP of endodermal cells with each cell colored by the average endoreduplication score of its scRNA neighbors, with highest levels seen in endodermal types 2 and 3.

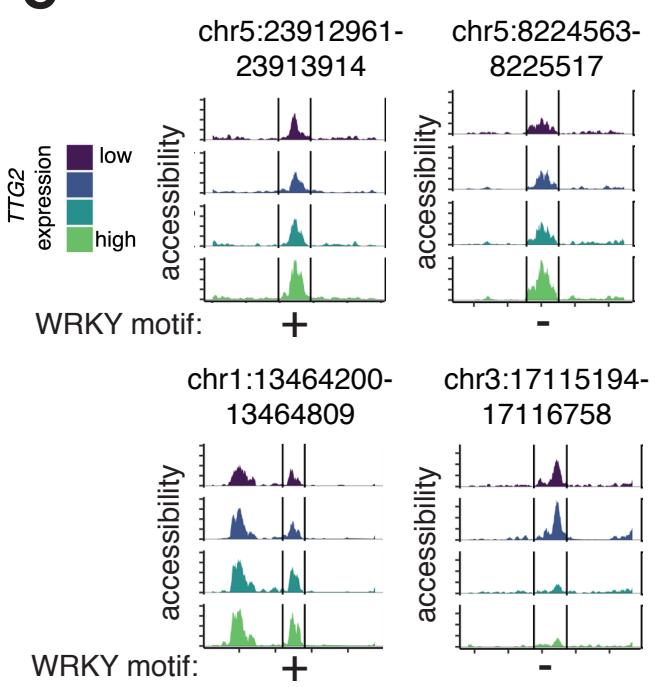
**A**



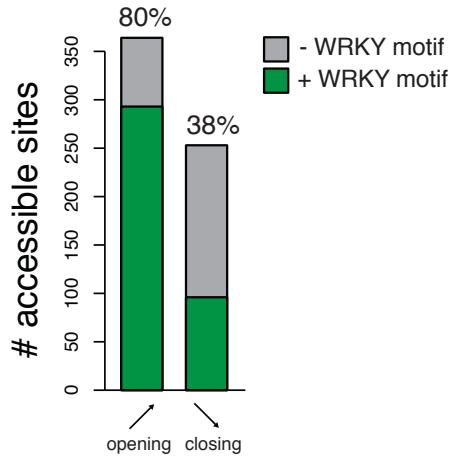
**B**



**C**



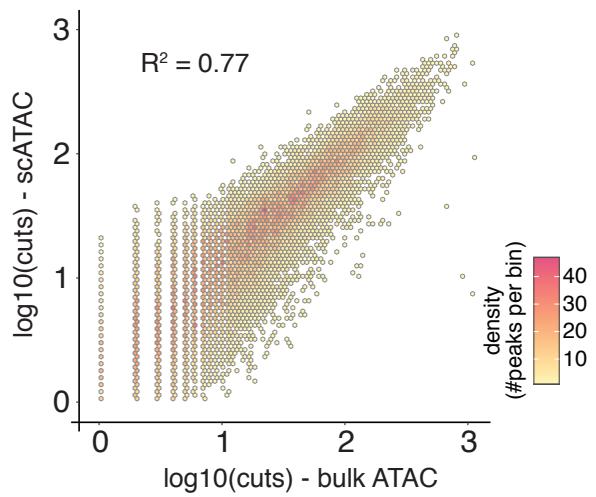
**D**



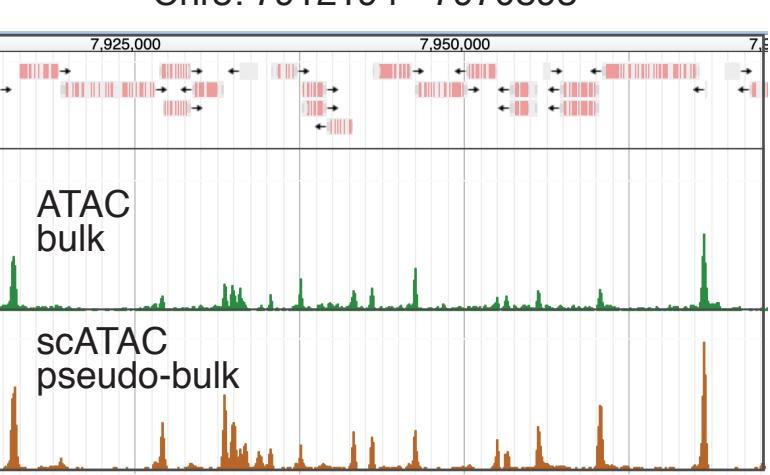
**Figure 4. Integration of scATAC and scRNA-seq data allows prediction of candidate regulatory TFs and genes.**

**Figure 4. Prediction of candidate regulatory transcription factors from integrated scATAC and scRNA data.** (A) Dotplot heatmap showing predicted expression of all WRKY family transcription factors across all cells; rows are ordered by the specificity of their epidermis expression. (B) UMAP plot of cells derived from scATAC experiment, but colored by predicted expression of an epidermis-specific WRKY transcription factor, *TTG2*. (C) Pseudobulked accessibility tracks of epidermis peaks whose accessibility shows a significant association with predicted *TTG2* expression. Cells with higher *TTG2* expression are shown in lighter shades. All panels show examples of significant ( $q < 0.05$ ) positive associations of *TTG2* expression with peak accessibility except the lower right panel. In each case, the presence of a WRKY binding motif is indicated below the peak. (D) Barplot showing fraction of WRKY binding motifs in peaks of the epidermis, cortex, and pre-cursor type that showed significant association with *TTG2* expression. Peaks whose accessibility showed positive associations with expression are labelled as “opening” and those with negative associations with expression are labeled as “closing.”

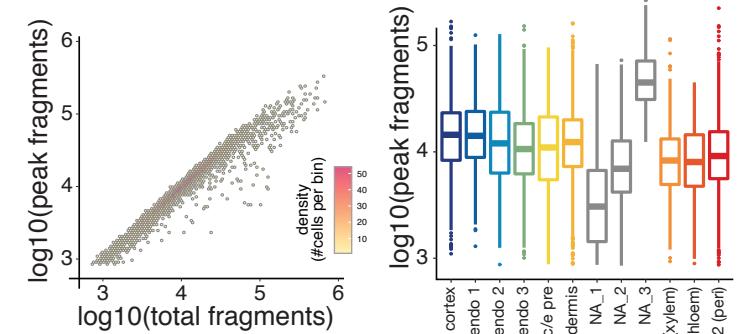
**A**



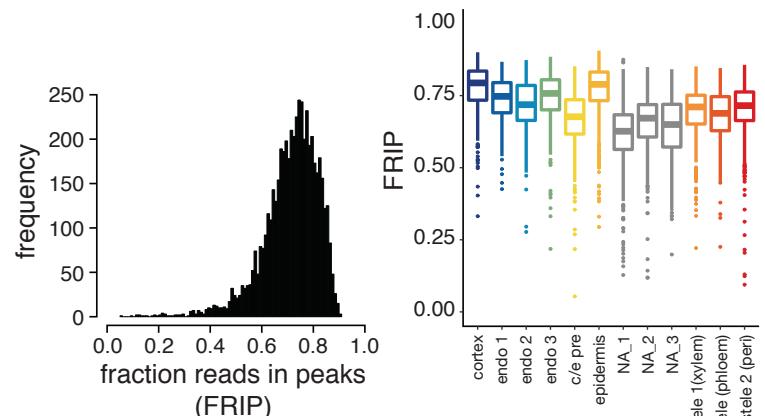
**B**



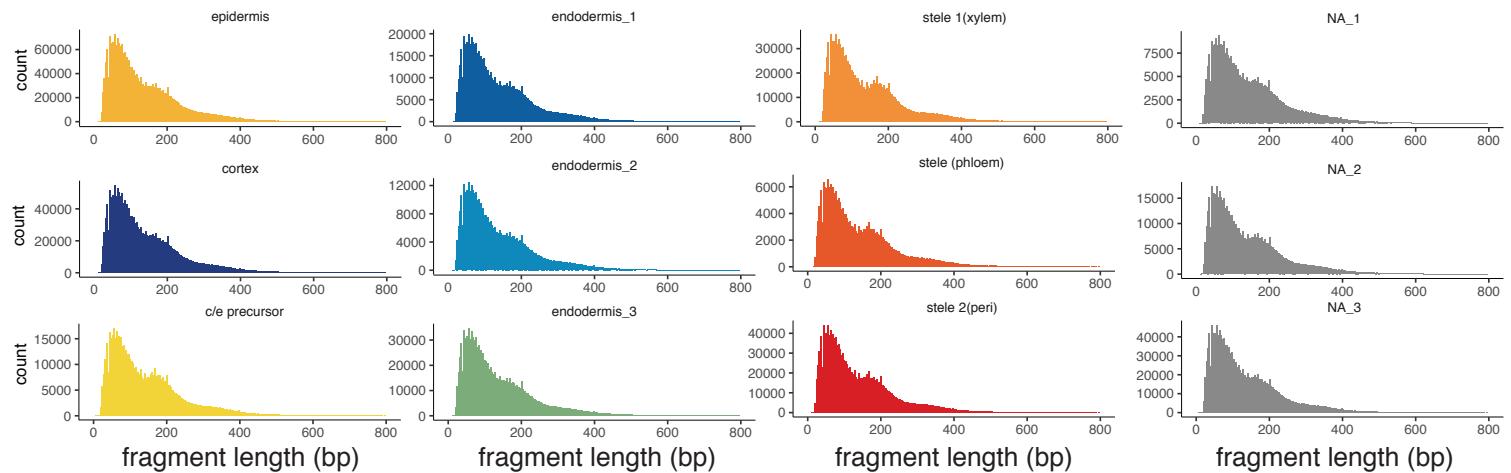
**C**



**D**



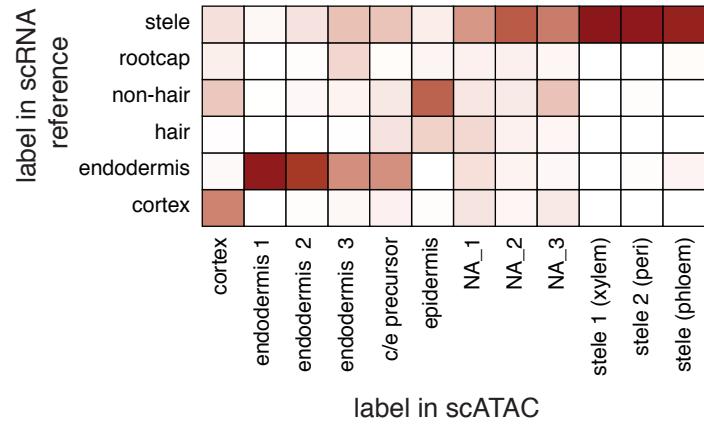
**E**



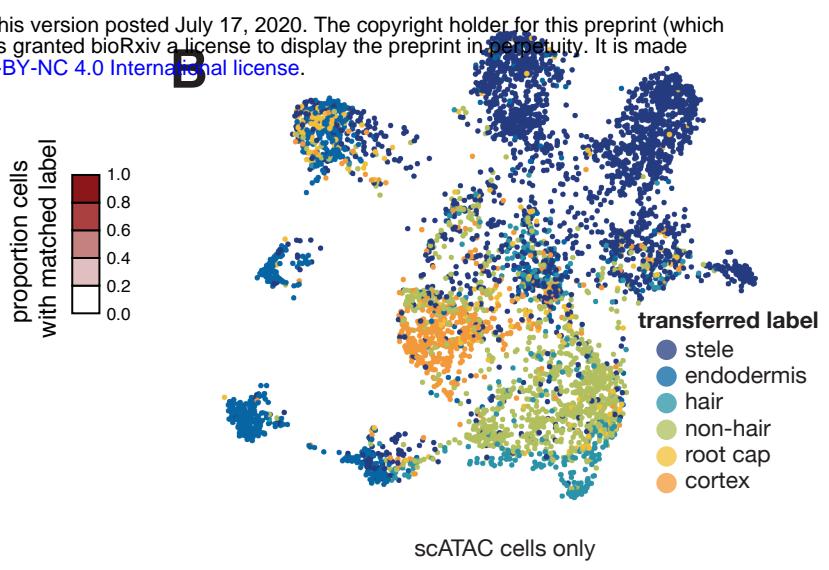
**Supplementary Figure 1. Quality of scATAC-seq data is comparable to bulk ATAC-seq data.**

**Supplementary Figure 1. Quality of scATAC-seq data is comparable to bulk ATAC-seq data.** (A) Scatterplot where each point represents peak defined in the scATAC data, and the x-axis shows the total cutcount within those peaks in bulk ATAC-seq and the y-axis shows the total cutcount within those peaks in scATAC-seq. Point density is indicated by increasing shades of red. (B) Example genomic region showing bulk ATAC accessibility (green) and pseudobulked scATAC accessibility (brown). Gene models are indicated above. (C) Read recovery per cell: Left panel shows relationship between total reads recovered per cell (x-axis) and reads in peaks (y-axis). Areas with higher point density are shown as in (A). Right panel shows boxplots of total number of reads in peaks recovered for each cell type. (D) ATAC quality per cell: Left panel shows overall distribution of fraction reads in peaks (FRIP) across all cells, right panel shows distribution of FRIP for each cell type. (E) Read length distributions for all fragments separated by cell type.

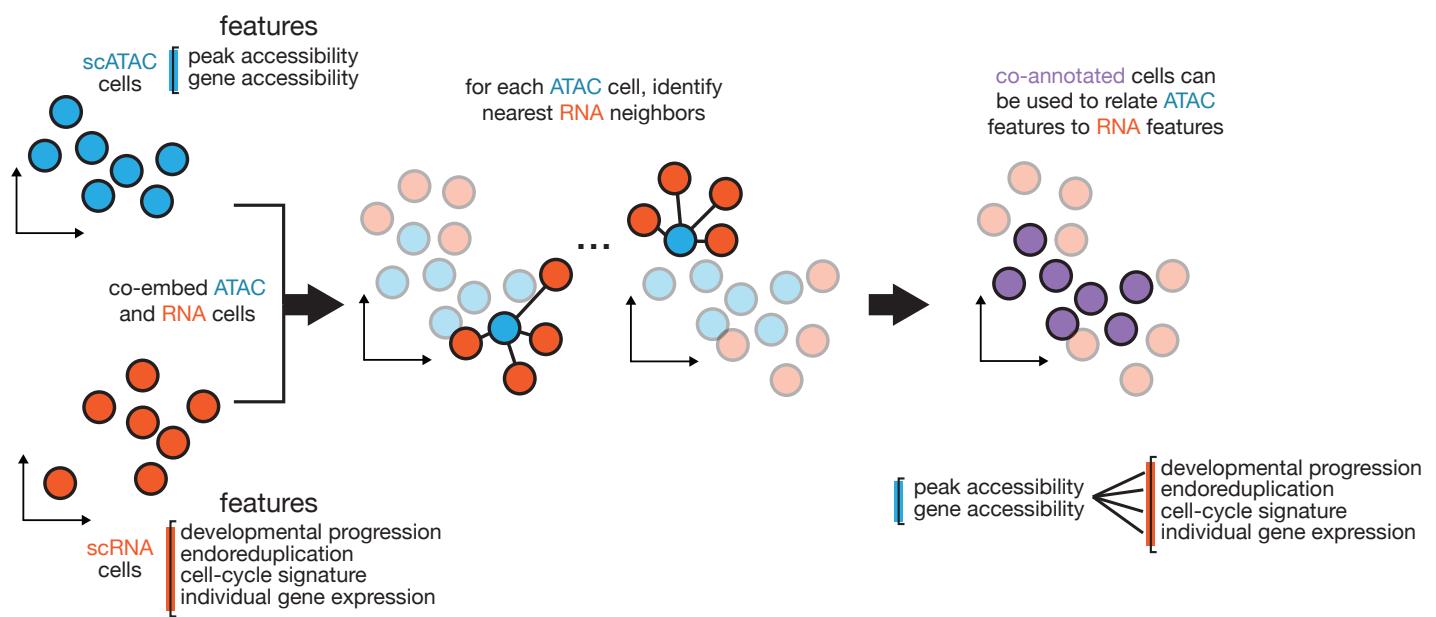
A



B



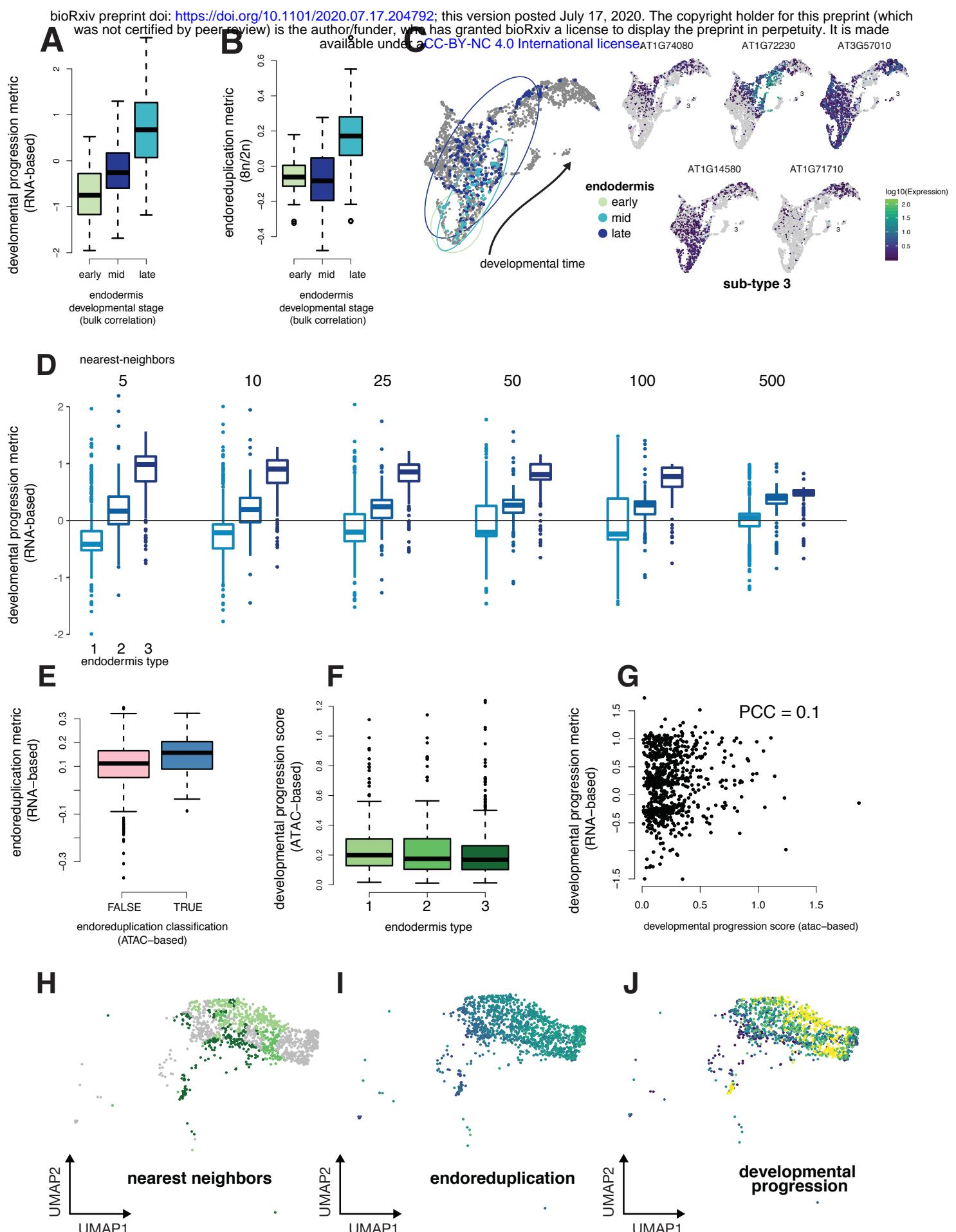
C



Supplementary Figure 2. Co-embedding-derived cell-type labels match manual annotations.

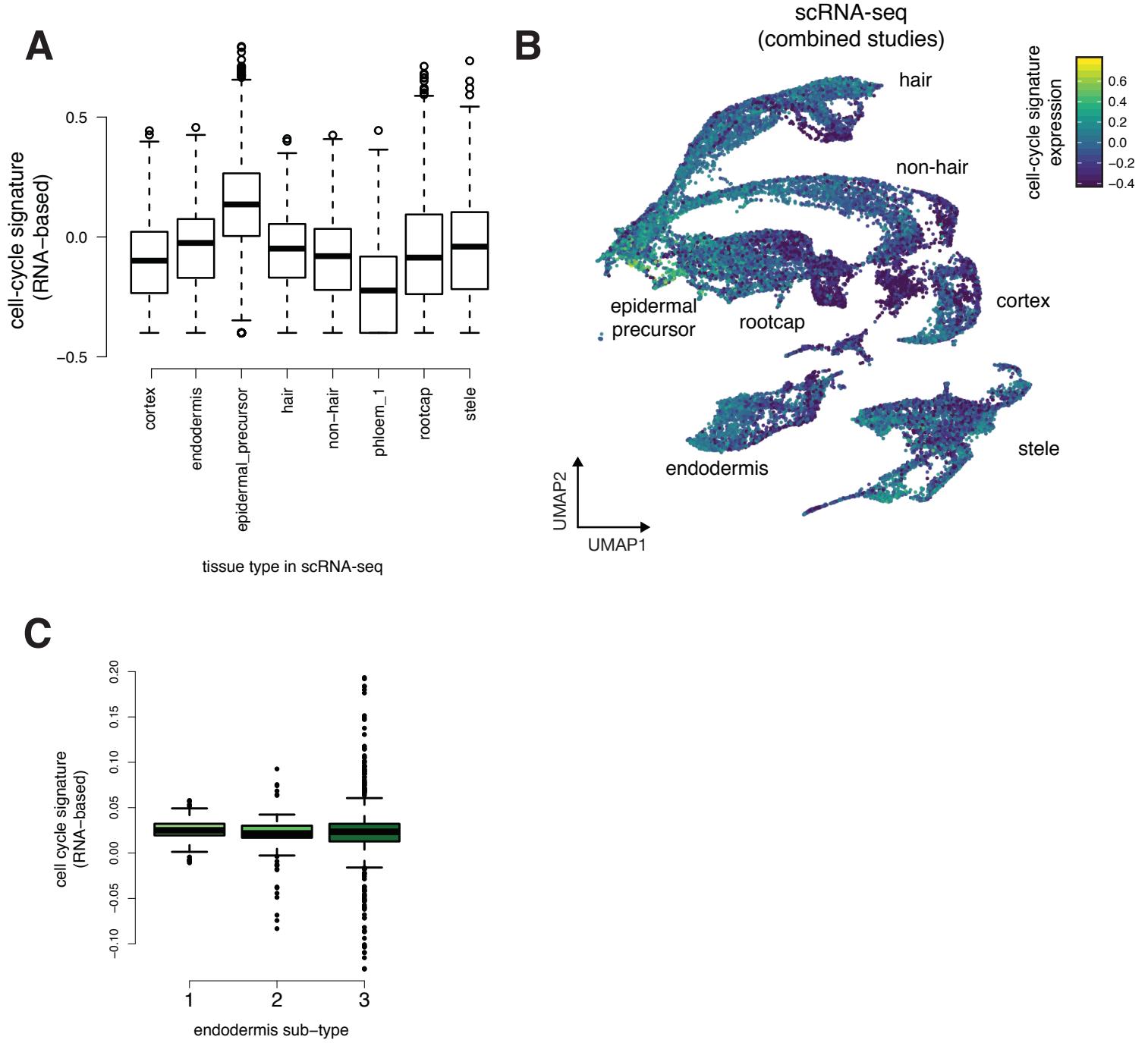
**Supplementary Figure 2. Co-embedding of scATAC and scRNA data allows validation of cell type labels and annotation by RNA-derived features. (A)**

Confusion matrix showing the correspondence of manual cell annotations with those derived from the label-transfer from RNA to ATAC cells. (B) UMAP of scATAC cells as in Fig. 1A, but cells are colored by the cell type label predicted from annotations of scRNA nearest neighbors. These cell types labels broadly match those predicted by manual annotation, and separate the epidermis cluster into hair and non-hair cells. (C) Workflow schematic for annotation of scATAC-cells with transcriptional data. In short, the 25 nearest RNA neighbors from each ATAC cell in the co-embedded graph (**Figure 2A**) were identified, and average expression of individual genes and signatures scores were computed and assigned to each ATAC cell.



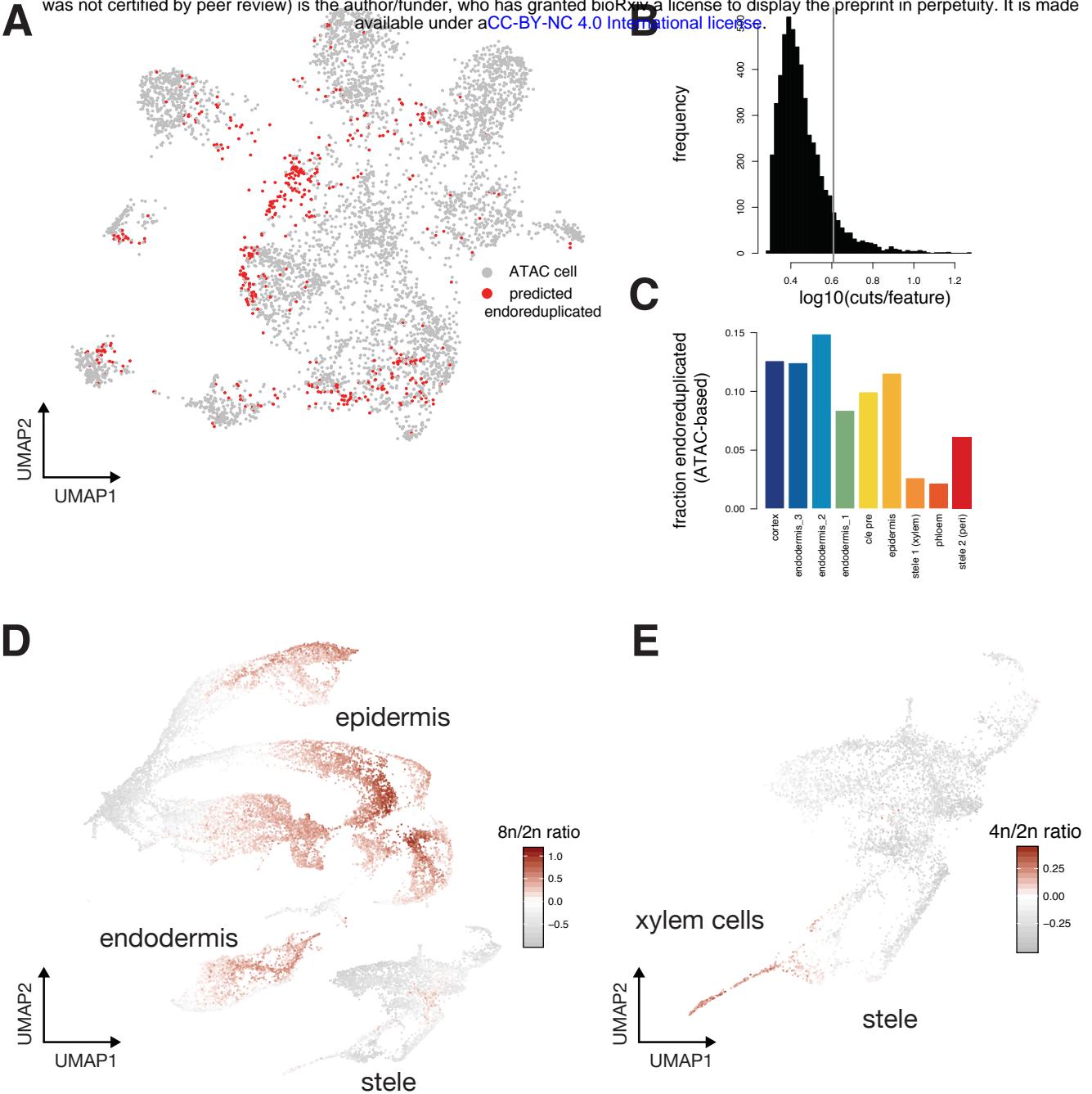
**Supplementary Figure 3. scATAC-seq identifies distinct sub-types of endodermal cells.**

**Supplementary Figure 3. Characterization of endodermal sub-types with combined scATAC and scRNA-seq data.** Boxplots showing that developmental progression scores (A) and endoreduplication scores (B) are consistent with a previously described annotations of developmental progression of the endodermis (Jean-Baptiste et al.). Cells were grouped into early, middle and late (left to right) developmental stages by Jean-Baptiste et al. (C) UMAP of endodermal cells from multiple scRNA-seq studies, with developmental stage designations of cells from the Jean-Baptiste et al study highlighted. Inset shows variable expression patterns of genes with highly specific accessibility patterns in scATAC data. (D) Boxplots showing the transcriptional-signature-based endoreduplication metric compared to a binary classification of endoreduplication cells using scATAC data. scATAC cells with high levels of cutcounts at a single locus (suggesting endoreduplication) were analyzed in the co-embedded graph with scRNA-seq cells to calculate the average level of the endoreduplication signature among each scATAC cell's 25 nearest neighbors. The overall trend shows that the cutcount-based classification of endoreduplication is consistent with the transcriptional-signature-based metric. (E.) Boxplots showing levels of accessible genes (analogous to transcriptional complexity metric from Fig. 3C, only computed as total number of accessible genes rather than total number of transcribed genes). The overall trend remains the same, with progressive loss of complexity in the later endodermal types, but the ATAC-based metric shows less sensitivity than the RNA-based one. (F) Scatterplot showing poor correlation of ATAC-based developmental progression score and the RNA-based score. (G-I) Subset of co-embedded UMAP from Figure 2A showing only endodermal cells; nearest RNA neighbors for each endodermal type are shown in (G); (H) shows RNA cells colored by transcriptional-signature-based endoreduplication metric; (I) shows RNA cells colored by transcriptional complexity metric.



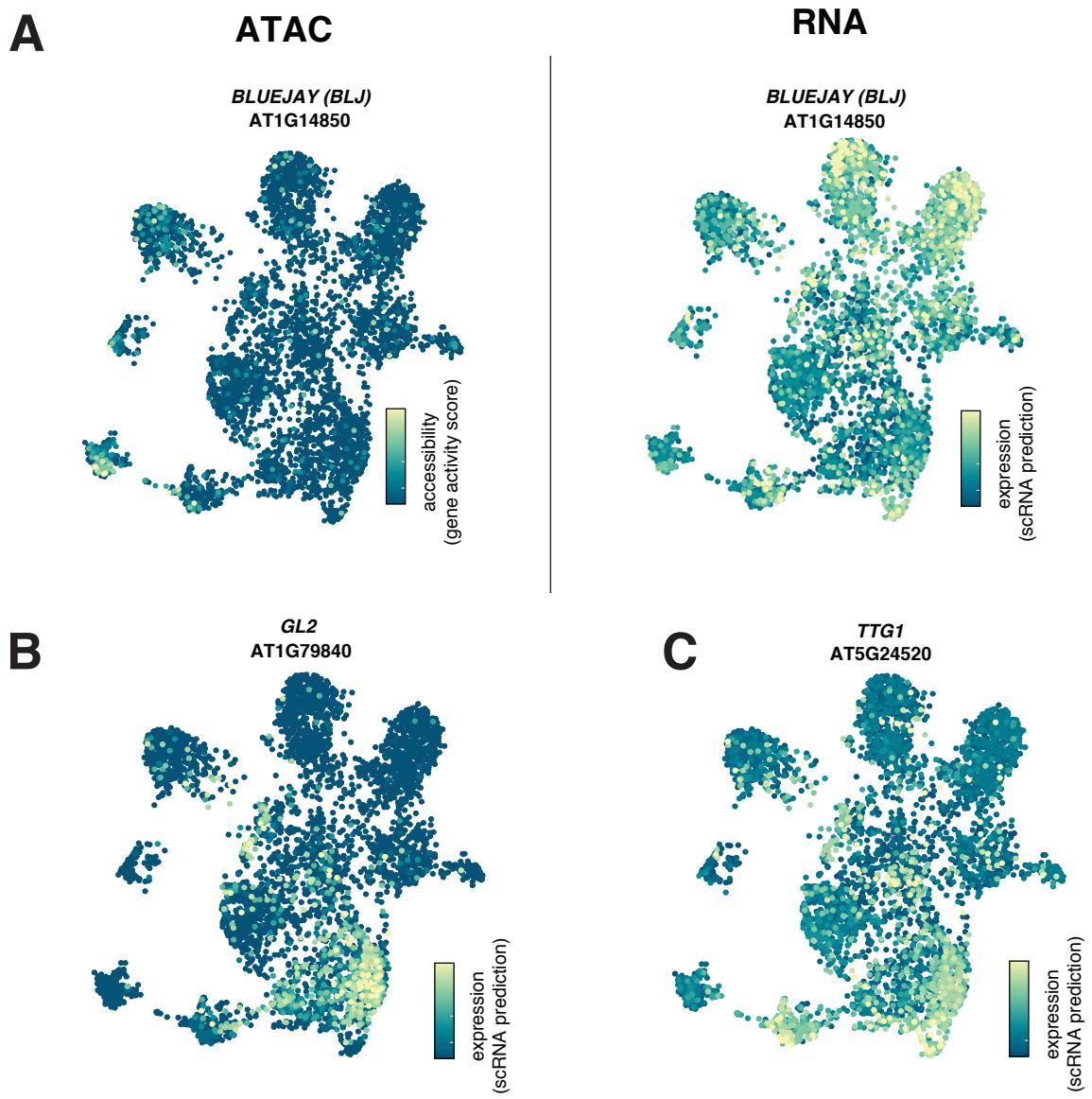
**Supplementary Figure 4. Dividing cells are present in the root, but do not distinguish endodermis sub-types**

**Supplementary Figure 4. Dividing cells are present in the root, but are not a distinguishing feature of endodermal types.** (A) Boxplots showing levels of a cell-cycle signature in each scRNA-seq root cell type. (B) UMAP plot of combined root scRNA-seq studies with each cell colored by its expression the cell cycle signature. (C) Cell cycle signature predicted from nearest neighbors of endodermis types (as in **Figure 3C, 3D**) shows that proliferation is not a strongly distinguishing feature between types.



**Supplementary Figure 5. Endoreduplicated cells can be identified in both scATAC and scRNA-seq data**

**Supplementary Figure 5. Approaches for identifying endoreduplicated cells in both scATAC and scRNA-seq data.** (A) UMAP plot of root scATAC cells, each colored based on whether that cell contains a threshold level of cuts per site. (B) Histogram showing the log cuts per site across all cells, with the threshold used to color cells in (A) shown as a vertical grey line. (C) Barplot showing the fraction of cells in each type that show putative endoreduplication, as determined by the threshold cuts per site drawn in (B). In general, cell layers nearer the epidermis show higher fractions of endoreduplicated cells, while cell layers of the stele showed lower levels. (D) UMAP of root scRNA cells, each colored based on the expression level of a transcriptional signature for endoreduplication, as determined by a ratio of expression levels in genes previously determined as enriched in 8n cells over those enriched in 2n cells. (E) A known instance of endoreduplication in the stele is identified by a metric similar to (D), except that cells are colored by signature for 4n cells (ratio of 4n-specific genes to 2n-specific genes).



**Supplementary Figure 6. Identifying transcription factors involved epidermal specification.**

**Supplementary Figure 6. Identifying transcription factors involved in tissue specification.** (A) Left panel shows UMAP of scATAC cells colored by level of accessibility at the *BLUEJAY*, and right panel shows the same cells colored by predicted expression level of *BLUEJAY*. (B) UMAP of scATAC cells colored by predicted expression level of epidermal specification factor *GL2*. (C) UMAP of scATAC cells colored by predicted expression level of epidermal specification factor *TTG1*.