

File S1: Supplemental methods for “Tissue-specific transcriptomes reveal gene expression trajectories in two maturing skin epithelial layers in zebrafish embryos”

To analyze *Salmon*'s per-experiment per-transcript quantifications, we follow the general outline of bootstrap-aware R Bioconductor package *Sleuth* (as if the *Wasabi* package was importer), but perform all steps manually to have better control and incorporate some *DESeq2* package features that *Sleuth* does not support.

Experiment scale factors (SFs). We need to estimate a per experiment SF to correct for depth-of-sequencing variation. Thousands of transcripts are expected to not truly vary in expression across experiments (having observed values that vary due to biological and technical sampling variation, but not from condition-to-condition effects; indeed, at the end of a typical generic differential expression analysis, most objects analyzed are not considered differentially expressed). If $X(t, e)$ is estimated number of reads for transcript t in experiment e (for SF estimation, we restrict to transcripts with $X(t, e) \geq 15.0$ for all e), and t and u are two such transcripts, then the

“instability” of t to $u :=$ standard deviation over experiments e of $\log_2 [X(t, e)/X(u, e)]$

is expected to be relatively low (compared to cases involving a transcript that does differentially express over conditions), as $X(t, e)/X(u, e)$ should be the relatively constant expression ratio between the transcripts (and the unknown depth-of-sequencing factors do not need to be included, since each ratio involves the same experiment and the common unknown factor cancels). By finding a large group of transcripts where pairwise instability within the group is low, we find transcripts that are well-suited for SF estimation.

The “ i -th instability” for transcript t is the i -th smallest pairwise instability of t to all the transcripts, and the “ j -th mean instability” is the mean of the j -th instability over all transcripts. Find j (the smallest, if tied) for which the difference in mean instability between j and $j + 1$ is minimum (expected to be close to the point of highest density in the distribution of mean instabilities). The transcripts to use for SF estimation are the j ones with smallest j -th instability. We now have j “invariant” transcripts t whose expression can be modeled as

$$X(t, e) = \text{SF}(e) \cdot E(t) \cdot \text{residual}(t, e)$$

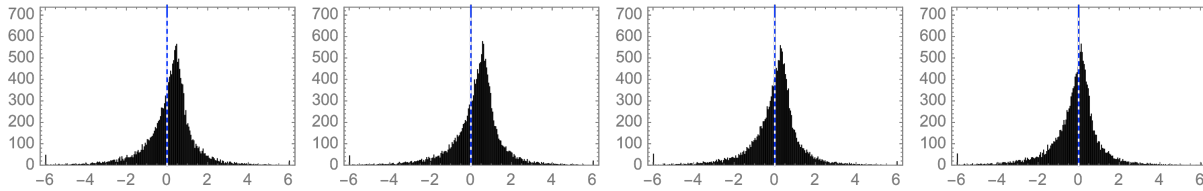
or (equivalently): $\text{residual}'(t, e) = X'(t, e) - \text{SF}'(e) - E'(t)$

where $\text{SF}(e)$ is the scale factor of experiment e , $E(t)$ is the expression level of transcript t , $\text{residual}(t, e)$ entries are “near” 1.0, and primed $X'(t, e)$ is $\log_2 X(t, e)$, etc. We choose to minimize the Frobenius norm of $\text{residual}'$, and resolve the single linear degree of ambiguity of adding a constant to all entries of SF' and subtracting the same constant from all entries of E' by adding the natural requirement that SF' have mean zero over its entries (i.e., that the geometric mean of SFs is 1.0, so post-SF [i.e., normalized] expression is on a scale of read counts of an “average” original experiment). The solution is

$$\text{SF}'(e) = [\text{mean over } t \text{ of } X'(t, e)] - [\text{mean over all entries of } X'].$$

We do this process once for *Salmon* transcript estimates (where 4,802 transcripts are selected), and once for gene estimates (where $X(g, e)$ for a gene g is the sum of $X(t, e)$ for transcripts t belonging to g , and 4,513 genes are selected), and the final SF estimate of an experiment is the geometric mean of these two (which were quite close). Final SFs vary 0.45 to 1.83, and ~93% / 92% of invariant genes / genes with at least one invariant isoform ended up in flow NNN.

The SF estimation of *Sleuth* is simpler (take $X(t$ [or g], e), in each row multiply elements by a constant to make the row geometric mean 1.0, replace each column by its median, and multiply the tentative SFs so obtained by a constant so their geometric mean is 1.0). It is somewhat robust due to its use of geometric means and medians, but when there are large condition-to-condition changes (such as here between nonskin and skin), these SFs can be somewhat distorted by their inclusion of many condition-varying genes.

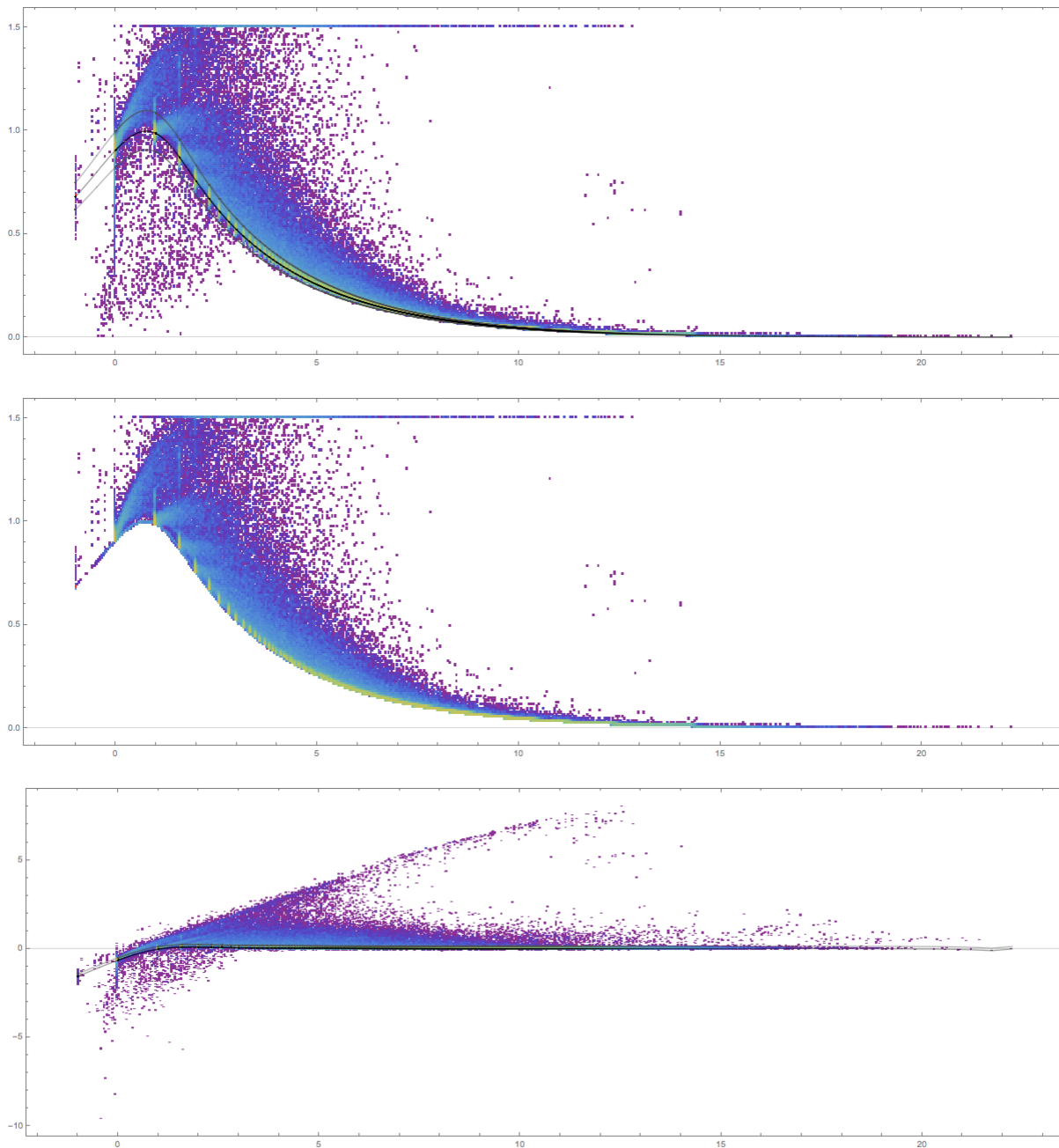


Comparison of estimated counts for “A” := 20 SS nonskin₁ replicate b vs. “B” := 20 SS all skin replicate b , as normalized by different methods: left to right is unnormalized, followed by normalized via total counts, *Sleuth*, and our “low instability” method. Histograms (bins of 0.05) are of $\log_2(A/B)$ over all 31,901 genes, restricted to genes with $A \geq 10$ and $B \geq 10$. The mode is closest to zero with our scale factor method.

Technical variance baseline. Following *Sleuth*, we model “transformed” (log-scale) expression per gene per condition as normally distributed. Rather than a $\log_2(0.5 + \text{estimated reads})$ scale, we use $\log_2(\max[0.5, \text{estimated reads}])$. As in *Sleuth*, we decompose modeled expression variance into a technical assay component (which is large for genes with low counts or those hit by many reads of high mapping ambiguity) plus a presumed independent biological component. Technical variance is informed by *Salmon* bootstrap samples, while biological variance is informed by replicate experiments within conditions; shrinkage procedures help to overcome the small (as is typical) number of replicates per condition (which is two in this study for all but one condition, where it is one), so that direct estimation is too unstable (or not possible). We do not filter out genes of low expression (e.g., *Sleuth*’s genes with $< 47\%$ of experiments having ≥ 5.0 estimated reads) early in the analysis.

Unlike *Sleuth*, we estimate a “baseline” (minimum) technical variance (as ambiguity due to low counts is always present, even if not apparent in some bootstrap samples) given a gene’s transformed expression level x in some experiment as follows. Let y , the gene’s “technical standard deviation” in this experiment, be the standard deviation of the transformed *Salmon* Gibbs bootstraps for the gene in the experiment*, and collect (x, y) pairs where $y > 0$ over all genes and experiments. Note, as the number of reads goes to infinity, that $\log_2(\max[0.5, \text{Poisson}(\text{reads})])$ has mean $\approx \log_2(\text{reads})$, variance $\approx 2/\text{reads}$, and $\log_2(y)$ is expected for large x to tend to $(1-x)/2$. We thus quantile regress x vs. $z := \log_2(y) - (1-x)/2$ [using R packages *quantreg* 5.36 and *splines*: `rq(z ~ bs(x, degree=3, knots=c(1.0, log2(3.0), 2.0, 3.0, 4.0, 8.0, 12.0, 16.0, 20.0, 22.0)), tau=0.25, method="fn")`], then we evaluate x for each gene and experiment, converting the resulting z' to y' in y scale. The gene’s new “ y with technical baseline imposed” for the experiment is $\max(y, y')$. If y is in $(y'/1.1, 1.1 \cdot y')$, we consider the observation to be “uncomplicated by technical quantification issues”, and if every experiment of a gene is uncomplicated then we consider the gene as “uncomplicated”, which is $\sim 82\%$ of genes (with strong representation at all expression levels).

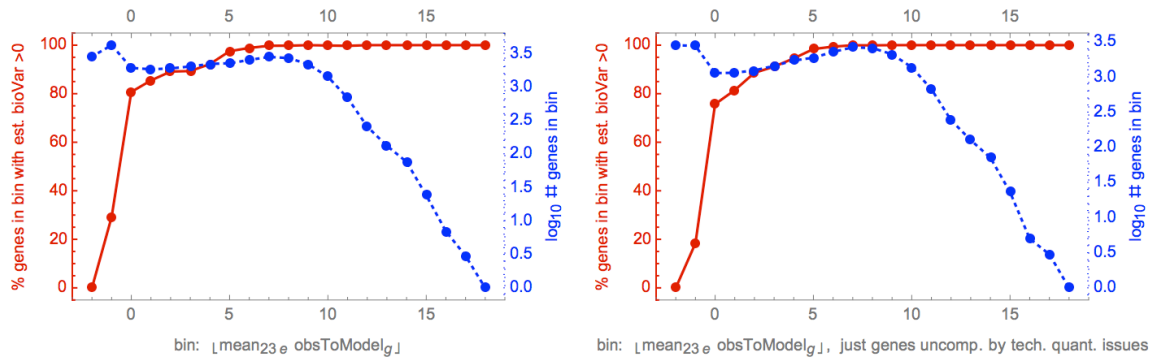
(*For a gene–experiment pair with *Salmon* expected reads < 0.5 , we take $y = 0.68 \approx$ standard deviation of $\log_2[\max(0.5, \text{Poisson}[0.5])]$.)



Imposition of technical baseline. All three panels are 2-D histograms over all genes and experiments, with rainbow hues purple-to-red indicating log-scale tallies low-to-high. In all three, the horizontal axis is the variable x discussed above (i.e., \log_2 estimated counts), and the vertical axis in the top two is variable y , i.e., technical standard deviation: before baseline imposition in the top panel, and afterwards in the middle panel. The vertical axis in the bottom panel is variable z (i.e., y de-trended against large count theoretical expectation; before imposition), with which regression operates. The middle, top, and bottom black curves in the top and bottom panels show the regressed baseline, and it multiplied by 1.1 and divided by 1.1 (the “uncomplicated” thresholds discussed above).

Biological variance baseline. Back to paralleling *Sleuth*, we take, for each gene, observed variance across experiments within a condition to be an independent sum of technical variance due to the assay plus biological variance; some genes are more biologically noisy than others. It is important to estimate total variance affecting a gene as accurately as possible, as this directly affects judgement of statistical significance of observed differences for the gene across conditions. If replicates within conditions were abundant, total variance could simply be taken as observed variance and estimated gene-by-gene (and even per condition) directly; however, as is typical, replicates are not plentiful, and so shrinkage is employed so that genes of similar expression levels can contribute to each other and regularize estimates.

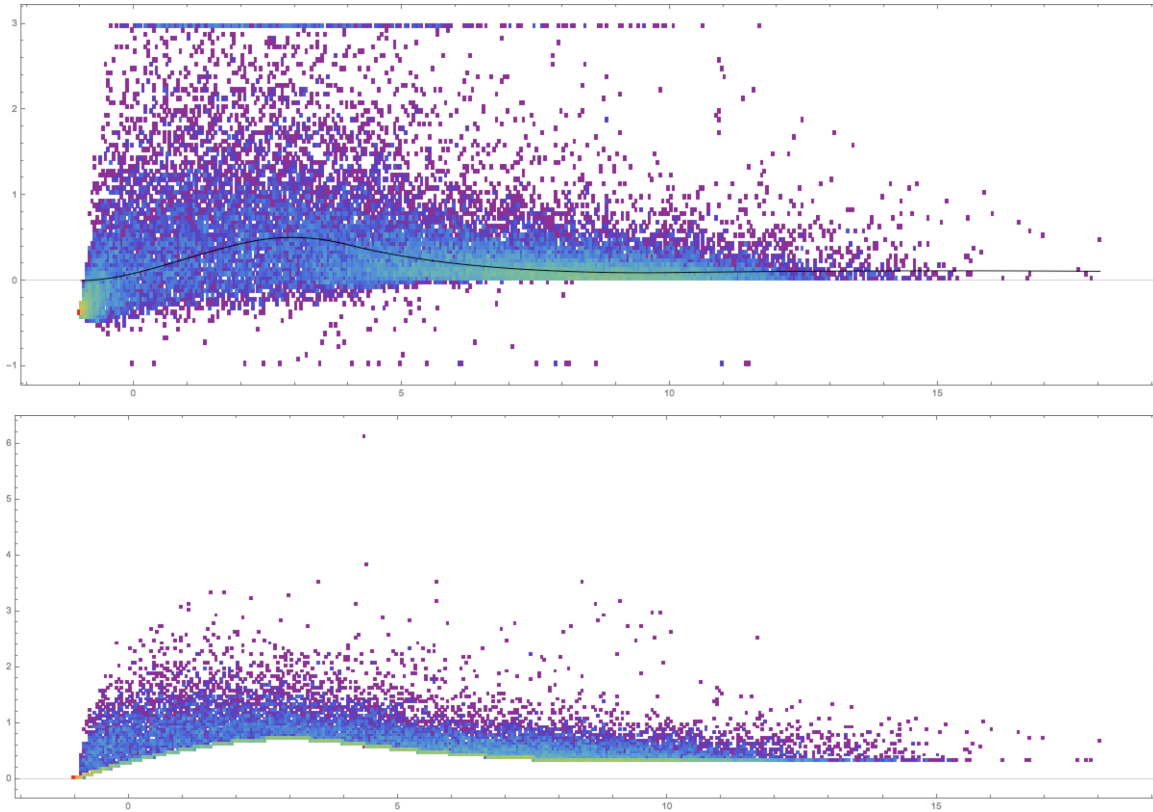
Starting with transformed *Salmon* estimated read counts, we apply per experiment SFs to obtain normalized transformed expression $W(g, e)$ for gene g in experiment e . For gene g , we take $s := \text{mean of } W(g, e) \text{ over all } e$ as its “overall expression level”. The mean σ^2_{total} , taken over the eleven conditions c with two replicate experiments each, of the sample variance of $W(g, e)$ for the two e in c is an estimator of the per condition total variance of g . (Note that sample variance of a two-observation sample a and b is just $(a-b)^2/2$.) The mean across experiments of the squares σ^2_{tech} of the gene’s technical standard deviations with technical baseline imposed is our estimator of the technical variance contribution to σ^2_{total} . Hence, $\sigma^2_{\text{bio}} := \sigma^2_{\text{total}} - \sigma^2_{\text{tech}}$ (which may be below zero; for genes with low counts, for example, technical variance tends to dominate biological variance) is an estimate of the per condition biological variance for the gene.



Expression level uncertainty in low expression genes is dominated by technical variance, and the genes uncomplicated by technical quantification issues well-represent the expression level distribution of all genes. All four histograms (blue and red in left and right panels) are gene-based, and horizontal axis is variable s discussed above (i.e., mean normalized transformed [log₂-scale] expression over experiments), in bins of size 1.0. Vertical blue axis is number of genes in the bin (which falls off rapidly at high expression). Vertical red axis is percentage of genes in the bin for which $\sigma^2_{\text{bio}} > 0$ (positive biovariance excess observed); left panel is for all 31,901 genes, right panel is for those that are uncomplicated. As expression falls below ≈ 30 normalized counts, the fraction of genes for which technical variance overwhelms biological variance grows (and technical variance is the major estimation obstacle for un-/lowly-expressed genes). The right panel’s similarity to the left indicates that restriction to uncomplicated genes (primarily those genes with negligible fractions of reads having read-to-gene assignment ambiguity) does not appreciably change the distribution of expression levels.

Over genes uncomplicated by technical quantification issues (rather than all genes), we quantile regress (rather than LOESS) s vs. $b := \text{the square root of } \max(0.0, \sigma^2_{\text{bio}})$ [with $\text{rq}(b \sim \text{bs}(s, \text{degree}=2, \text{knots}=\text{c}(2.0, 4.0, 8.0, 10.0)), \text{tau}=0.50)$]. (Transforms of the b -axis are less

important for quantile regression compared to LOESS, so we do not use the fourth root as *Sleuth* does, just square root [so s and b are on the same scale], and we do not entirely suppress cases where $\sigma_{\text{bio}}^2 < 0$; we are not trying to estimate “true” biological variance, but rather biological “excess” variance beyond technical variance, and zero excess is commonly expected for genes as s gets low, and is perfectly fine.) For each gene, we evaluate its s to get b' , and the gene’s new “ σ_{bio}^2 with biological baseline imposed” is $\max(\sigma_{\text{bio}}^2, b'^2)$. The gene’s “final per condition σ_{total}^2 ” is then [its σ_{tech}^2 with technical baseline imposed] + [its σ_{bio}^2 with biological baseline imposed].



Imposition of biological baseline. Both panels are 2-D histograms over all genes, with rainbow hues purple-to-red indicating log-scale tallies low-to-high. In both panels, the horizontal axis is variable s discussed above (mean normalized transformed [log₂-scale] expression over experiments). In the top panel, the vertical axis is σ_{bio}^2 , and the regressed biological baseline is shown as a black curve. In the bottom panel, the vertical axis is σ_{bio} with biological baseline imposed.

Expression and contrast models. For condition c of gene g , the model of normalized transformed expression $W(g, c)$ is a normal distribution $\mathbf{N}[\mu, \sigma^2]$ with mean μ = the average of $W(g, e)$ over experiments e in c , and variance σ^2 = the final σ_{total}^2 for g times either $\frac{1}{2}$ or 1 according to whether c has two or one replicate experiments, respectively; models for different conditions are independent. Note that a ratio of two normalized linear-scale expression levels becomes a simple subtraction of the corresponding normalized log-scale expression levels. Hence, for a contrast that is $\mathbf{N}[\mu_B, \sigma_B^2]$ subtracted from (independent) $\mathbf{N}[\mu_A, \sigma_A^2]$ (corresponding to linear-scale ratio A/B)[†], the model is again normal: $\mathbf{N}[\mu_A - \mu_B, \sigma_A^2 + \sigma_B^2]$. This is all essentially the same as *Sleuth* (although its operation in terms of non-singular design matrices and lack of explicit linear combination contrast support may obscure the simple cases here).

([†]Call a gene a “zero” in a condition if expected reads for it in every experiment in that condition are ≤ 0.5 . Since $[\max \text{SF}] / [\min \text{SF}]$ is \approx four-fold, we do not want zero genes to show as differentially expressed due to normalization. Hence, in models $\mathbf{N}[\mu, \sigma^2]$ for ratio condition A over condition B, we adjust μ for some cases involving genes that are zero in A and/or B: If a gene is zero in both A and B, we set μ to 0.0 [no difference expected]. If a gene is zero in A but not B, we change μ to $[\log_2 0.5 \text{ normalized by mean } \log_2 \text{SF for experiments in B (rather than in A)}] - [\mu \text{ for current gene in B}]$ if this moves μ strictly closer to no difference expected; and vice-versa if a gene is zero in B but not A. Note the next section does an independent filtering before constructing q -values; contrast instances with zero genes are not unlikely to be dropped there.)

#Abbreviating all skin / nonskin₁ / periderm / basal cells / nonskin₂ as ‘M’ / ‘C’ / ‘P’ / ‘B’ / ‘c’ and 20 SS / 52 hpf / 72 hpf as ‘2’ / ‘5’ / ‘7’, the 36 contrasts considered were C5/C2, C7/C2, C7/C5, M5/M2, M7/M2, M7/M5, c5/C2, c7/C2, c7/c5, B5/M2, B7/M2, B7/B5, P5/M2, P7/M2, P7/P5, M2/C2, M5/C5, B5/c5, P5/c5, P5/B5, M7/C7, B7/c7, P7/c7, P7/B7, C5/c5, M5/c5, B5/C5, P5/C5, B5/M5, P5/M5, C7/c7, M7/c7, B7/C7, P7/C7, B7/M7, and P7/M7.

Differential expression p - and q -values for conditions pairwise. For the contrast that is the ratio of condition A over B, the \log_2 -scale ratio for each gene has a normal model $\mathbf{N}[\mu, \sigma^2]$ as described in the previous section. Preferring to avoid point comparisons for any detectable difference, we desire p -values with the flexibility of *DESeq2*’s `altHypothesis=` alternative hypothesis and `lfcThreshold=` log-scale fold change threshold options (which *Sleuth* does not support). For example, the master Excel workbook in the NCBI GEO submission contains two-sided p -values for expression level change of more than 1.5x-fold either up or down (*DESeq2* `altHypothesis=`“greaterAbs” and `lfcThreshold =` essentially $f := \log_2 1.5$): $p = \min(1, 2 \cdot [1 - \text{cdf of } \mathbf{N}[f, \sigma^2] \text{ at } |\mu|]) = \min(1, 2 \cdot [\text{cdf of } \mathbf{N}[\mu, \sigma^2] \text{ at } f], 2 \cdot [1 - \text{cdf of } \mathbf{N}[\mu, \sigma^2] \text{ at } -f])$. Other *DESeq2* p -value styles are similarly easy. For example, `altHypothesis=`“greater” with the same `lfcThreshold` is $p = (1 - \text{cdf of } \mathbf{N}[f, \sigma^2] \text{ at } \mu) = (\text{cdf of } \mathbf{N}[\mu, \sigma^2] \text{ at } f)$.

We conduct an independent filtering (as does *DESeq2*) on contrasts before applying Benjamini–Hochberg (BH) False Discovery Rate (FDR) correction to the p -values (rather than dropping low expression genes entirely early in analysis, as in *Sleuth*). Gene g in contrast A/B is retained if and only if the maximum of normalized transformed expression $W(g, e)$ over experiments e in condition A or B is at least a threshold. The threshold chosen, 2.72 (equivalent to $\geq \sim 6.6$ linear scale normalized counts in at least one experiment in the contrast), was optimized (similar to as in *DESeq2*) to maximize the number of significant q -values below 0.02 (for all 36 contrasts of the paragraph marked (#) above considered at once) after FDR correction.

Approximate Transcripts Per Million (TPMs). The *Salmon*/*Sleuth*-like model and statistics we follow are primarily focused on and operate with normalized transformed estimated read counts (per transcript/gene, per experiment/condition), as these are the values essentially measured by RNA-Seq assays. Hence, these are also the values that we focus on in reports and figures. However, these values differ from absolute expression levels in that, e.g., transcripts/genes with longer mRNAs tend to have higher counts. *Salmon* estimates an “effective length” $L(t, e)$ for each transcript t and experiment e (that accounts for transcript annotated length vs. RNA-Seq library insert length distribution and various RNA-Seq biases *Salmon* empirically considers) to be used in combination with estimated read counts $X(t, e)$ if, e.g., TPMs are desired.

To get $\text{TPM}(g, e)$ for gene g in fixed experiment e , form ratios $X(t, e) / L(t, e)$ for all transcripts t , multiply by $S(e) := 1 \text{ million over the sum of these ratios}$, and then sum the resulting $\text{TPM}(t, e)$ over t in g . (Note that per-experiment scale factors $\text{SF}(e)$ do not matter, as they would cancel; TPMs are normalized by simplistic totals.) We summarize the combined effect of $L(t$ in $g, e)$

and $S(e)$ by $S^*(g, e) := \text{TPM}(g, e) / (\text{sum } X(t, e) \text{ over } t \text{ in } g)$ when the denominator is ≥ 0.1 (we do not summarize if the gene is unexpressed in e). We then summarize the effect for this gene over experiments by $S^*(g) :=$ the geometric mean of summarized $S^*(g, e)$ (and we do not summarize if the gene is unexpressed in all experiments). Thus, for an expressed gene g , $\text{TPMs} \approx S^*(g) \cdot \text{counts}$ (approximate up to variation over experiments and isoforms). Using this relationship, we are able to, e.g., dual label counts-focused axes with approximate TPMs, as seen in Figure 2A and the expression profile plots on the website.

Over genes g expressed in at least one experiment, $\sim 75.6\%$ have $0.0113 < S^*(g) < 0.113$ (the decade centered around the densest part of the distribution of $S^*(g)$ over genes). This supports the TPM labeling of the “mean across 12 conditions” colorbar of Figure 5.

Classification of genes into flows. Placement of genes into exactly one “flow”, summarized in Figure 3, is done by independently classifying the gene at 20 SS, 52 hpf, and 72 hpf. Using the condition abbreviations of the paragraph marked ^(#) above, 20 SS is straightforward as it involves only a single pairwise comparison between M2 and C2; once M2 “ $>$ ” / “ \leq ” C2 is resolved, the flow class at 20 SS is ‘S’ / ‘N’, respectively. M2 “ $>$ ” vs. “ \leq ” C2 is taken to be when *DESeq2* greater-style p -values with 1.5x fold change threshold are $<$ vs. ≥ 0.02 .

52 and 72 hpf are more complex, as they involve at least partially ordering three expression levels, rather than two: $x/y/z = c5/B5/P5$ for 52 hpf and $c7/B7/P7$ for 72 hpf, respectively. To order the three values, note that it suffices to consider $a := y - x$ and $b := z - y$ (as the remaining difference, $z - x$, is $a + b$). Further, note that if $x \sim \mathbf{N}[\mu_x, \sigma_x^2]$, $y \sim \mathbf{N}[\mu_y, \sigma_y^2]$, and $z \sim \mathbf{N}[\mu_z, \sigma_z^2]$ are independent normals (as here), that while $a \sim \mathbf{N}[\mu_y - \mu_x, \sigma_x^2 + \sigma_y^2]$ and $b \sim \mathbf{N}[\mu_z - \mu_y, \sigma_y^2 + \sigma_z^2]$ are normals, they are *not* independent, having non-zero covariance $-\sigma_y^2$. Hence, to classify we do not (as can be common) use a decision chain based on pairwise comparisons of conditions; p -value computation should be aware of all three variables at once, treating bivariate (a, b) as a 2-D binormal with mean $(\mu_y - \mu_x, \mu_z - \mu_y)$, variance $(\sigma_x^2 + \sigma_y^2, \sigma_y^2 + \sigma_z^2)$, and covariance $-\sigma_y^2$.

To work toward the needed classification with fold change threshold $f := \log_2 1.5$, we partition $\mathbf{R}^3 \ni (x, y, z)$ into pieces: $(x < y < z)$ vs. $(x < z < y)$ vs. $(y < x < z)$ vs. $(y < z < x)$ vs. $(z < x < y)$ vs. $(z < y < x)$. (The three values are distinct with probability 1 in the model, so we are free to ignore the probability zero cases where two or more of the variables are equal, and other probability zero cases.) If $c < d$ are two of x, y, z ordered in agreement with the current piece, we write “ $c \ll d$ ” if $d - c > f$ (i.e., “ d is significantly larger than c ”), and otherwise “ $c \leq d$ ” ($0 < d - c < f$). If $c < d < e$ are x, y, z ordered in agreement with the current piece, we subpartition the current piece into cases: [I] ($c \ll d \ll e$) vs. [II] ($c \ll d \leq e$) vs. [III] ($c \leq d \ll e$) vs. [IV] ($c \leq d \leq e$ and $c \leq e$) vs. [V] ($c \leq d \leq e$ and $c \ll e$); these correspond to various subsets of $(a, b) \in \mathbf{R}^2$ given by conjunctions of affine inequalities, e.g., for part $x < y < z$: [I] ($a > f$ and $b > f$) vs. [II] ($a > f$ and $0 < b < f$) vs. [III] ($0 < a < f$ and $b > f$) vs. [IV] ($0 < a < f$ and $0 < b < f - a$) vs. [V] ($0 < a < f$ and $f - a < b < f$). The model probability of each case is then the 2-D integral of the probability density function of binormal (a, b) over the case’s subset of \mathbf{R}^2 . We compute these using numerical integration in *Mathematica*; among other strategies, integration of one dimension (e.g., b) is easily analytically expressed in terms of standard $\text{Erf}[\cdot]$ and $\text{Erfc}[\cdot]$ special functions, so that the numerical integrations only need be 1-D (and we provide the integrator appropriate hints to where bulk of density lies). Case [V] is interpreted to be the three values weakly distinct (enough that $c \ll e$ but not $c \ll d$ or $d \ll e$); its mass is redistributed to Cases [II] and [III] in proportion of those two cases to their sum.

We currently have probability 1.0 partitioned across 24 cases: [I] to [IV] for six strict orderings of x, y, z . If x, y , and z were three not-necessarily-distinct real numbers, there would be thirteen

weak orderings of them: $(x < y < z)$, $(x < z < y)$, $(y < x < z)$, $(y < z < x)$, $(z < x < y)$, $(z < y < x)$, $(x = y < z)$, $(z < x = y)$, $(x = z < y)$, $(y < x = z)$, $(y = z < x)$, $(x < y = z)$, and $(x = y = z)$. We re-partition the probability 1.0 to these thirteen weak orderings by interpreting “<” as “<” and “=” as “≤”, and summing over all permutations of variables connected by “=”.

We now have probability 1.0 partitioned into the thirteen weak orderings of x , y , z . If, say, the probability of exactly one of these thirteen was very close to 1.0, then one would accept the classification of genes x , y , and z as statistically ordered in that way. For 52 and 72 hpf flow placement, we do not need distinctions this fine, however; by re-partitioning probabilities a last time, we avoid some problems with being unable to classify well because more than one of the thirteen weak orderings has non-negligible probability, but the split is across cases we do not need to distinguish for flow placement. With $(x, y, z) = (c5, B5, P5)$ or $(c7, B7, P7)$, we gather $\underline{b} :=$ “ y is highest” = $(x < z < y) + (z < x < y) + (x = z < y)$, $\underline{p} :=$ “ z is highest” = $(x < y < z) + (y < x < z) + (x = y < z)$, $\underline{g} :=$ “ y and z equal but above x ” = $(x < y = z)$, and $\underline{n} :=$ sum of the remaining six = $(y < z < x) + (z < y < x) + (y = z < x) + (z < x = y) + (y < x = z) + (x = y = z)$. If probability $\underline{n} < 0.02$ and probability $\underline{g} < 0.02$, then the flow placement is ‘P’/‘B’ if probability \underline{p} is $\geq / < 0.5$, respectively (and no cases actually empirically arise where probability \underline{p} is at all close to the boundary 0.5); otherwise, the flow placement is ‘G’/‘N’ if \underline{n} is $< / \geq 0.02$, respectively.

Genes ranked within flows and combinations of flows. Individual flows are named by combinations of three characters (the first character for 20 SS, the second for 52 hpf, and the third for 72 hpf) determined in the previous section: ‘N’ or ‘S’ at 20 SS, and ‘N’ or ‘G’ or ‘B’ or ‘P’ at each of 52 and 72 hpf. We also consider certain combinations of flows: a first character ‘*’ combines ‘N’ and ‘S’ at 20 SS; a second and/or third character ‘*’ combines ‘N’ and ‘G’ and ‘B’ and ‘P’ at 52 and/or 72 hpf; and a second and/or third character ‘S’ combines ‘G’ and ‘B’ and ‘P’ at 52 and/or 72 hpf.

For reports and Gene Ontology analyses, it is useful to have the genes within each flow and combination of flows ranked, with genes early in a ranked list strongly exhibiting the pattern and those of low rank being weaker examples, getting closer to classification as another pattern. Fix a flow or flow combination; for each gene, we assign an ordering value a , b , and c in $[0.0, 1.0]$ to 20 SS, 52 hpf, and 72 hpf, and then genes are ranked by descending $\min(a, b, c)$. For 20 SS: for class ‘N’, $a :=$ the p -value used to classify (i.e., *DESeq2* greater-style with 1.5x threshold for $M2 > C2$); for class ‘S’, $a := 1.0 -$ that p -value; and for class ‘*’, $a := 1.0$. For 52 hpf: with $(\underline{n}, \underline{g}, \underline{b}, \underline{p})$ being the partition of probability 1.0 as described in the previous section, for classes ‘N’/‘G’/‘B’/‘P’/‘S’/‘*’, then $b := \underline{n}/\underline{g}/\underline{b}/\underline{p}/(\underline{g}+\underline{b}+\underline{p})/1.0$; and 72 hpf assigns its c similarly.

Gene Ontology (GO) gene-term and term-term associations. Compilation of gene-term associations began with extraction of all Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) GO cross-references (from all external DBs, info-type/text combinations, and evidence codes) for genes/transcripts/proteins (ENSDARG/T/P’s, including non-protein coding genes and genes on non-primary assembly components) from the database files of Ensembl release 92 for zebrafish (*danio_rerio_core_92_11*). Associations for transcripts and proteins were taken as for the parent gene. (Here, all uses of the evidence code “ND” = “[N]o biological [D]ata available” are only to BP/MF/CC root GO terms.) Term-term associations began with the GO Consortium basic release (<http://purl.obolibrary.org/obo/go/go-basic.obo>) dated 2018-09-23.

Gene-term associations involving out-of-date GO ids were updated: GO:0000989 → (GO:0008134 + GO:0140110); GO:0000990 → (GO:0043175 + GO:0140110); GO:0000991 →

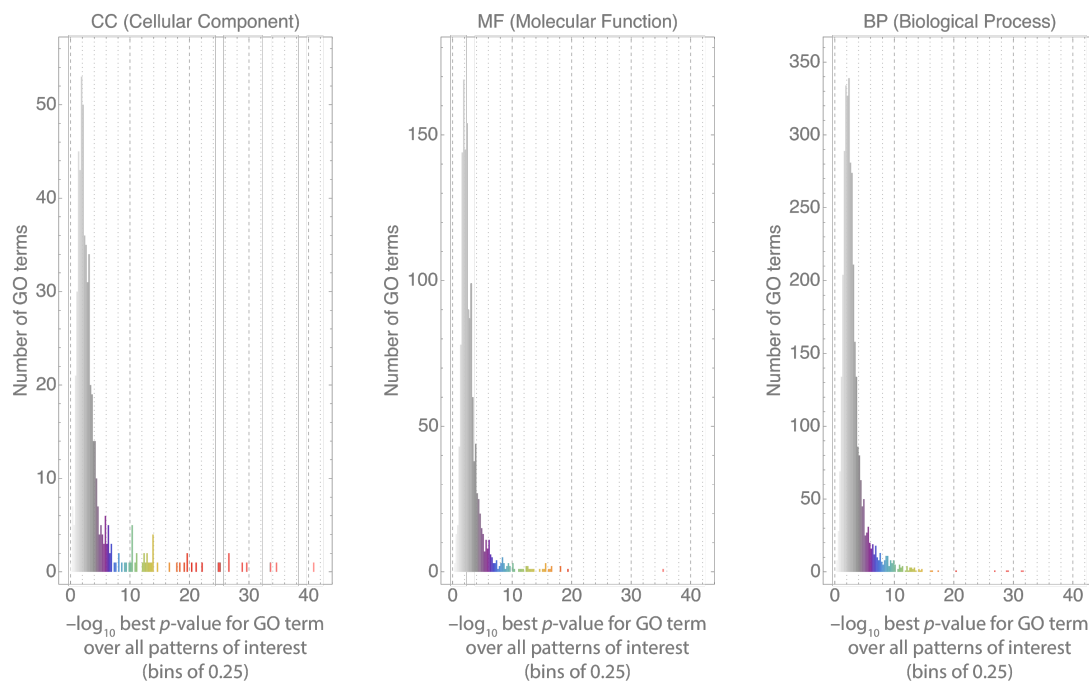
(GO:0000993 + GO:0140110); GO:0001076 → (GO:0001085 + GO:0140110); GO:0001129 → (GO:0001085 + GO:0140110 + GO:0017025 + GO:0051123); GO:0001191 → (GO:0001085 + GO:0001227); GO:0098811 → (GO:0001102 + GO:0001227); GO:0004871 → (GO:0060089 + GO:0007165); GO:0005057 → (GO:0060089 + GO:0007165 + GO:0035556); GO:0004716 → (GO:0060089 + GO:0035556 + GO:0004713 + GO:0023014); GO:0004702 → (GO:0060089 + GO:0035556 + GO:0004674 + GO:0023014); GO:0030818 → (GO:0006171 + GO:1900372); GO:0030819 → (GO:0006171 + GO:1900373); GO:0051436 → (GO:1904667 + GO:1903047); GO:0097033 → (GO:0034551); GO:0097034 → (GO:0033617); GO:0001007 → (GO:0008134 + GO:0140110 + GO:0006359); GO:0001026 → (GO:0000995); GO:0021865 → (GO:0051301 + GO:0021846); GO:0031659 → (GO:1900087 + GO:0045737); GO:0044376 → (GO:0006606 + GO:0000993); and GO:1990022 → (GO:0006606 + GO:0000994); followed by use of the OBO file to replace alternate GO ids by primary ids, and then primary ids by replaced-by ids. This nets 21,381 genes (all but 22 of which have Ensembl biotype `protein_coding`) associated to at least one non-root GO term (in 130,112 distinct pairs), with 7,927 distinct non-root GO terms involved.

Gene-term associations were expanded toward the BP/MF/CC roots to be inferentially closed (by following all relations in the OBO file; while one should not propagate CC across `regulates` / `positively_regulates` / `negatively_regulates` links, no such links exist here). This expands the 21,381 genes to be associated to 11,977 distinct non-root GO terms in 922,075 distinct pairs.

GO enrichment analyses in flows and flow combinations. For GO enrichment in flows and flow combinations of interest — these being all (NS*)(NGBPS*)(NGBPS*) with at least one 'G'/'B'/'P'/'S' — we first determine the subset of genes and subset of GO terms to analyze. Given the lack of gene-term associations for non-protein coding genes (non-“PCG”s), we restrict genes to PCGs. For terms, start with PCGs from all flows and flow combinations of interest. We thin this term set to eliminate trivial redundancies of more general terms that do not encompass more genes: collect all terms *a* for which there exists another term *b* that *a* is a direct parent of, but *a* is not associated to any more genes than *b*. Number current term layers by 1=leaves, 2=direct parents of layer 1, 3=direct parents of 2, and so on; if there are any *a*, delete all *a* on the lowest numbered layer that has any *a* and continue thinning. The terms remaining minus the BP, CC, and MF root terms are the “terms to be analyzed”. This gives 3,432 terms in the BP GO aspect, 559 in the CC GO aspect, and 1,393 in the MF GO aspect.

To analyze one flow or flow combination in a GO aspect (BP, MF, or CC), the “background” (or “universal”) set of associations are the gene-term associations of the previous section, restricted to PCGs, the GO terms to be analyzed, and the current aspect. The “background genes” and “background terms” are the distinct genes and terms mentioned in the background set of associations. Consider the ranked list *L* of genes (including non-PCGs) for the current flow or flow combination, and the ranked subset *L'* of it that is *L* restricted to the background genes. Enrichment is considered for each distinct GO term mentioned in background associations involving genes in *L'*, where the enrichment for a specific term is determined as follows: all non-empty prefixes of *L'* are considered, with the *p*-value for a prefix being the hypergeometric probability that a random subset of size *a* from a universal set of size *u*, and an independently random subset of size *b* from the universal set, have an intersection of size $\geq c$, where $u :=$ the number of background genes, $a :=$ the number of background genes associated to the current term in the background associations, $b :=$ the number of genes in the current prefix of *L'*, and $c :=$ the number of genes in the current prefix of *L'* that are associated to the current term in the background associations. The earliest prefix that minimizes *p* is taken.

Due to the complex structure of these p -values (e.g., from the GO directed acyclic graph-imposed interdependencies on terms), False Discovery Rate correction by the Benjamini–Hochberg procedure is not directly applicable. However, in practice, BH FDR correction often does little more than assist in choosing a threshold on q -values for “significance”, which — since BH FDR does not change rank order of values in the p -to- q transformation — is equivalent to another threshold and the already-in-hand rank order on p -values. Similar to the underlying assumption in typical differential expression analyses that most genes are not differentially expressed between most conditions, here it is reasonable to assume that most GO terms are not enriched in most flows and flow combinations. Thus, we may simply histogram these p -values over a wide range of terms and flows/combinations and empirically determine what p -value ranges are common and, hence, uninteresting/insignificant.



Empirical determination of significance for GO enrichment analyses p -values. For each GO aspect, all p -values computed (in $-\log_{10}$ scales on horizontal axes; bins of size 0.25) are histogrammed. Vertical axes are number of GO terms in the bin. Assuming most GO terms are not enriched in most flows/combinations, the rapid falloff of the distributions from ≈ 4.0 to 5.0 is indicative of passage from insignificance to significance, and was reflected in the colors used for the colorbars of Figure 4 and File S2, with grays below this point transitioning to saturated colors above this point (then progressing in hue from purple-to-red-to-pink as p -values continue to approach zero).

Thus, the p -value-to-color mapping used in Figure 4 (and File S2) was chosen with the assistance of the histograms above. For each GO aspect (BP / MF / CC), all pairs of flows / flow combinations and GO terms that had p -value ≤ 0.0001 , i.e., $P := -\log_{10} p \geq 4.0$, were identified, and then full heatmaps of all flows / flow combinations involved and all GO terms involved composed. Rows and columns were each clustered with Euclidean distance, complete linkage, and optimal swiveling to minimize sum of distances of adjacent leaves, where p -values were temporarily transformed by $\arctan(P - 5.0)$ to soft-threshold the insignificant-to-significant transition near $P \approx 5.0$. After hand inspection of the results (File S2), representative rows and columns were chosen to be highlighted in Figure 4.

Compilation of type I and type II keratins. Identification of type I and type II keratins in the Ensembl release 92 protein coding gene models (we did not consider potential repairs to existing gene models or a deep scan of the reference genome sequence or *de novo* RNA-Seq to model new loci) started with a case-insensitive search for “krt” and “keratin” against much of the database files of Ensembl release 92 for zebrafish (danio_rerio_core_92_11 tables gene, gene_attrib, transcript, transcript_attrib, translation, translation_attrib, protein_feature, xref, object_xref, ontology_xref, dependent_xref, and external_synonym), including gene descriptions, synonyms, alternate long names, remarks, and hidden remarks; protein features/domains (as Ensembl uses InterPro, from CDD 3.14, Gene3D 4.1.0, HAMAP 2017-01, Panther 12.0, PFAM 31.0, PIRSF 3.02, PRINTS 42.0, ProDom 2006.1, PROSITE 20.132, SFLD 3, SMART 7.1, and SuperFam 1.75); transcript synonyms, alternate long names, remarks, and hidden remarks; and all external database cross-references (e.g., into NCBI RefSeq and EMBL-EBI UniProt). These were then hand-filtered to remove nominal false positives from keratin-associated proteins (non-keratins involved in keratinization [e.g., periplakin], filament binding proteins, and proteins involved in keratin metabolism), keratinocyte-associated proteins, and proteins involved in keratinocyte differentiation. “krt” and “keratin” matched gene symbols, gene names / short descriptions (including names of homologous genes in external databases), PRINTS signatures PR01248 (“TYPE1KERATIN”) and PR01276 (“TYPE2KERATIN”), PFAM family PF16208 (“Keratin_2_head”), various Panther subfamilies of PTHR23239 (“INTERMEDIATE FILAMENT”) including “KERATIN, TYPE I” and “KERATIN, TYPE II” CYTOSKELETALS, Gene Ontology term GO:0045095 (“keratin filament”), and Reactome R-DRE-6805567 (“Keratinization”).

This gave 41 ENSDARG gene loci that included numerous well-known type I and type II keratins, with the status of a number of other loci less clear. ENSDARG isoforms of all were multiply aligned (*Geneious* global alignment using BLOSUM62 and gap open / extend 12 / 3 scoring, with five refinement iterations) and a phylogenetic tree constructed (*Geneious* tree builder using neighbor-joining over Jukes–Cantor distances with no outgroup, forming a consensus tree from 100 bootstrap resamples with 50% support threshold). Thinning each gene locus to the isoform most homologous to isoforms of other analyzed genes, there were three clear protein groups — 23 putative type I keratins, six putative type II keratins, and seven putative other intermediate filaments (*nefma*, *nefmb*, *neflb*, *prph*, *gfap*, zgc:65851, and si:dkey-27m7.4) — as well as five proteins at relatively large distances: (1)–(2) *thread keratins alpha / gamma* (“T.K.A.” / “T.K.G.”) and (3) *krt222*, plus two we rejected: (4) *bfs2* (*beaded filament structural protein 2* a.k.a. *phakinin*, an intermediate filament-like eye lens component considered by some literature to be a “beaded filament” protein, hence taken as not a type I or type II keratin, although InterPro family IPR027694 “Phakinin” is a subfamily of IPR002957 “Keratin, type I” and our Figure S1 also suggests it is a reasonable alternative to consider *bfs2* as a type I keratin), and, finally, (5) *fam83hb* (a FAM83-family oncogene pulled in by Ensembl cross-reference UniProt Q1LVV0 as “colocalizes_with” GO:0045095, and not a keratin itself). The working pool at the end of this stage thus contained 23 + 6 + 7 + (5–2) = 39 genes.

InterProScan 5.34–73.0 was run on the 39 genes to examine their domain structure in a wide range of databases (File S4). The universal feature was InterPro domain IPR039008 “Intermediate filament, rod domain”, supported by all of SMART SM01391 “Filament”, Pfam PF00038 “Filament”, and PROSITE PS51842 “IF_ROD_2” (where all three hit once each on top of each other and occupied the majority of the protein’s length, except for *krt222* where the PROSITE hit was longest, the SMART hit somewhat shorter, and the Pfam hit less than half as long). InterPro family IPR002957 “Keratin, type I” supported by Panther PTHR23239 “INTERMEDIATE FILAMENT” (in one full/near-full length hit) corresponded perfectly to putative type I status — and was missing from *T.K.A.*, *T.K.G.*, and *krt222* — and both family IPR003054 “Keratin, type II”

supported by PRINTS PR01276 “TYPE2KERATIN” (in six small hits spread over the rod domain) and domain IPR032444 “Keratin type II head” supported by Pfam PF16208 “Keratin_2_head” (occupying the majority of the residues upstream of the rod domain) corresponded perfectly to putative type II status — and was *also* missing from *T.K.A.*, *T.K.G.*, and *krt222*. Among these 39 genes, this analysis supports the putative type I keratins as precisely the type I’s, the putative type II keratins as precisely the type II’s, and *T.K.A.*, *T.K.G.*, and *krt222* as neither type I nor type II keratins.

To expand the 39 gene list to sequence-similar genes possibly missed by the keyword search, for each of six groups — (1) the 23 putative type I keratins, (2) the six putative type II keratins, (3) the seven putative other intermediate filaments, (4) the single *T.K.A.* gene, (5) the single *T.K.G.* gene, and (6) the single *krt222* gene — we collected over genes in the group the single chosen isoform per gene, and multiple aligned them with *Geneious* as in the paragraph before the previous paragraph. In each multiple alignment, there was a single large region of high conservation; multiple alignments were restricted to these regions and, for each, *HMMer* 3.1b2 *hmmbuild* used to build a Hidden Markov Model (HMM), after which *hmmscan* was used with default parameters to scan all Ensembl peptides (ftp://ftp.ensembl.org/pub/release-92/fasta/danio_rerio/pep/Danio_rerio.GRCz11.pep.all.fa.gz) for occurrences of the HMMs.

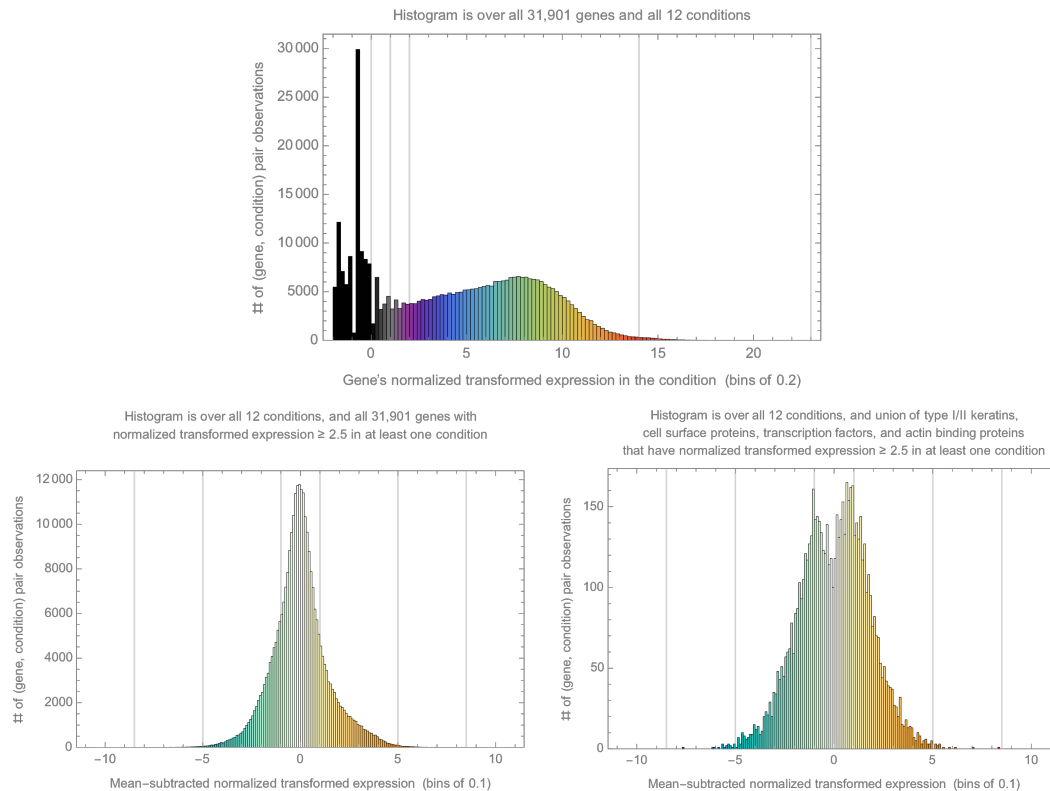
The resulting 627 distinct ENSDARPs contained hits at a wide range of E-values (up to the extremely marginal E-value score 5.0). For each HMM, all hits for members of the HMM’s group had lower E-values than for any hit to that HMM from other groups; the point at which hits by ascending E-value first appeared for a member of another group definitely of the nominal type of that other group established a cut ($\sim 10^{-98}$ to 10^{-72} , except $\sim 10^{-23}$ for *krt222*) beyond which ENSDARPs were discarded. This gave only eight new genes to add from the pile of 627: *nefla*, *desma*, *desmb*, *inaa*, *inab*, *vim*, *viml*, and *si:dkey-33c12.3* (most of which are well-known definite other types of intermediate filaments and not type I or type II keratins).

To provide further context, another round of HMM formation and scanning was done to gather zebrafish intermediate filament proteins more generally. This brought in *lmna*, *lmnb1*, *lmnb2*, *lmnl3*, *ngs*, *nes*, *synm*, *vimr1*, *vimr2*, *iffo1a*, *iffo1b*, *iffo2a*, *iffo2b*, and *zgc:172323*. The resulting $39 + 8 + 14 + bfsp2 = 62$ genes were multiply aligned with *ClustalW* 2.1 (BLOSUM costs and gap open/extend 10/0.1, without free end gaps) and analyzed with *MrBayes* 3.2.6 *aamodelpr=mixed* (for 32 chains and 2.5 million Markov Chain Monte Carlo [MCMC] generations, with average standard deviation of split frequencies descending to ≈ 0.01 – 0.02 for most generations), exporting a 50% majority rule tree (Figure S1). This analysis also supports taking the 23 putative type I keratins (plus *bfsp2* as already mentioned) as the complete list of zebrafish type I keratins, the six putative type II keratins as the complete list of zebrafish type II keratins, and not including *T.K.A.*, *T.K.G.*, or *krt222* in either list (including them instead as other keratins / intermediate filament proteins).

Figure 5, File S5, and gene sets of special interest. The gene sets contributing to Figure 5 and File S5 were determined as follows. The 23 type I and six type II keratins were from the previous section (without *bfsp2* as a type I keratin). The cell surface proteins, transcription factors, and actin binding proteins were each determined by all those genes associated to a single MF GO term in the GO inferential closure as described in an earlier section of this File, the term being GO:0004888 “transmembrane signaling receptor activity”, GO:0003700 “DNA-binding transcription factor activity”, or GO:0003779 “actin binding”, respectively (resulting in 1,346 and 773 and 350 genes). Term choice was based on what terms are actually used in quantity in the extracted inferential closure, in a compromise between the biological function of the term being too general for what was desired vs. having unreasonably few genes

associated. In the master Excel workbook deposited with NCBI GEO, reporting of the GO-term based gene sets extends to genes on alternate chromosomes.

For File S5 and Figure 5, the GO-term based gene sets were intersected with the genes in the union of all flows except NNN before plotting (resulting in 291, 183, and 127 genes for cell surface proteins, transcription factors, and actin binding proteins, respectively); for type I and type II keratins, all genes were retained. In Figure 5, as described in main text Methods, only the top 20 genes (as ordered as described there) in each GO-term based gene set are shown. In both Figure 5 and File S5, genes are clustered by Manhattan (L_1) distance on vectors of mean-subtracted normalized transformed counts $\in \mathbf{R}^{12}$ (that is, rows of the 12 condition values that get transformed to colors in the mean-subtracted colorbar) with average linkage and optimal swiveling to minimize sum of adjacent leaf distances.



Colors for the mean expression colorbar were chosen after inspection of the histogram above of mean expression over all 31,901 genes, using black for effectively unexpressed genes and grays for poorly-expressed genes, reserving rainbow colors for genes non-negligibly expressed. Similarly, colors for the mean-subtracted colorbar were chosen after inspection of the histograms above of mean-subtracted values, using hue to indicate above vs. below mean and white to very desaturated colors for insignificant differences, reserving increasingly saturated colors for increasing differences.

Comparison to de la Garza et al. (2013). The data found available from the de la Garza et al. (2013) study (“DLG”) was the periderm profile (DLG Supplemental Table S1, giving 1,369 Ensembl ENSDART transcript accessions/names and, for each, a single linear scale summary microarray expression value for each of GFP+ and GFP–, along with a binary [boldface] indication of whether the transcript was considered to be from a transcription factor or not); the *dnIrf6*-inhibited profile (DLG Supplemental Table S2, giving 385 ENSDART accessions/names);

and the profile that is the intersection of the first two (DLG Supplemental Table S3, giving 92 ENSDART accessions/names, with the set of accessions indeed exactly those common to the periderm and *dn1rf6*-inhibited profiles). The only archival Ensembl release (<https://ensembl.org/info/website/archives/index.html>) still online that contains all 1,662 ENSDARTs in the union of the profiles (the “DLG Ts”) is release 54 from 2009 based on the Zv8 zebrafish genome. (“090505_Zv7_EXPR_HX12”, the name of the NimbleGen chip layout used by DLG, suggests Zv7, but the Ensembl archive no longer contains any release based on that genome.) Our study is based on Ensembl release 92 (GRCz11). Below, we use “Ens54T/G” and “Ens92T/G” to refer to the ENSDART/G transcripts/genes of the two releases, and interpret DLG Ts as a subset of Ens54Ts. The DLG Ts represent 1,659 distinct Ens54Gs (the “DLG Gs”), and only ~75.0% (1,244) of DLG G serials (without model version number suffix) still exist as Ens92Gs. Using Ensembl’s web ID History Converter (https://ensembl.org/Danio_rerio/Tools/IDMapper) maps only slightly higher DLG Gs (~76.5%: 1,269) to at least one Ens92G (falling to ~75.4% for those mapping to a single Ens92G, with two mapping to the same single Ens92G).

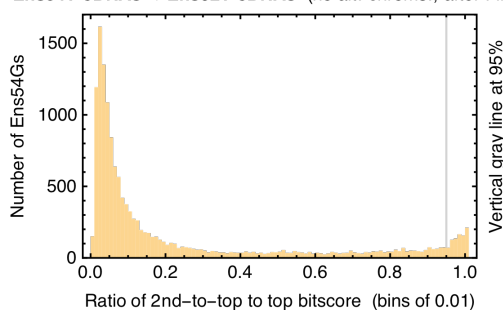
As the precise rules for Ensembl’s maintenance and mapping of ENSDART/G stable identifiers are somewhat mysterious, and as ~75% mappability from DLG Gs to Ens92Gs was suspected low, we used cDNA sequence similarity to establish a new mapping from Ens54Gs to Ens92Gs, as follows. (Mapping based on cDNA alignments is reasonable as the original assignments of microarray probes to ENSDARTs are likely to have been based on sequence identity/similarity.) We started with the cDNA nucleotide sequences for 28,717 Ens54Ts covering 24,233 Ens54Gs and 51,745 Ens92Ts covering 25,906 Ens92Gs obtained after removal of sequences residing on alternate chromosomes from Ensembl FTP files

ftp://ftp.ensembl.org/pub/release-54/fasta/danio_rerio/cdna/Danio_rerio.Zv8.54.cdna.all.fa.gz

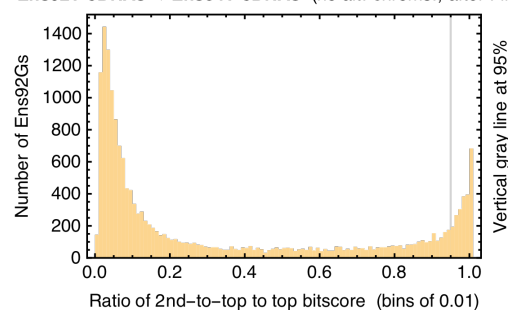
ftp://ftp.ensembl.org/pub/release-92/fasta/danio_rerio/cdna/Danio_rerio.GRCz11.cdna.all.fa.gz

(FASTA entry titles in these files include ENSDART and G serial numbers, and chromosome/scaffold names. Ignoring model version number suffixes, here only 13,826 T and 16,180 [≈two-thirds] G serials are in both releases. Also note that here Ensembl 92 has ~7% more Gs over Ensembl 54, but ~80% more Ts; as we desire one-to-one relationships where possible, we focus on mapping gene identifiers rather than transcript identifiers.) BLASTN 2.2.26 was run twice, once to align Ens54T cDNAs to Ens92T cDNAs (obtaining ≈0.8M alignments), and once in the reciprocal direction (≈1.6M alignments; both runs with E-value threshold 10^{-5} , DUST filtering, 11-mer words, two-stranded search, and up to 99 hits per query). In each direction, for each distinct Ens54G–Ens92G pair, only a single alignment of top bitscore was retained (“Filter 1”, dropping number of alignments to ≈218k and ≈455k). Histograms over query-side Gs with more than one remaining alignment of ratio of next-to-top-to-top bitscore were examined.

Ens54T cDNAs → Ens92T cDNAs (no alt. chroms., after Filter 1)



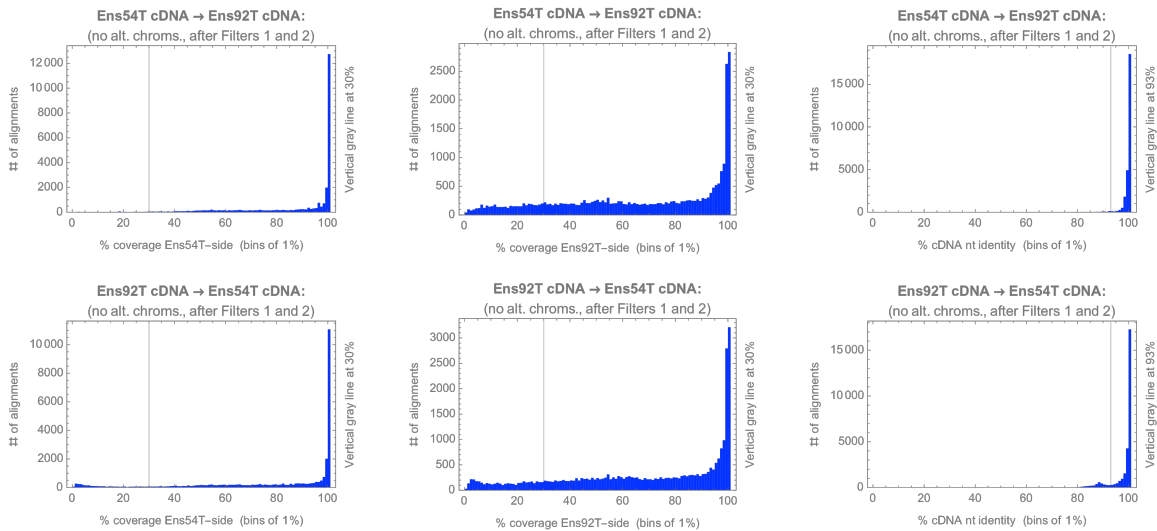
Ens92T cDNAs → Ens54T cDNAs (no alt. chroms., after Filter 1)



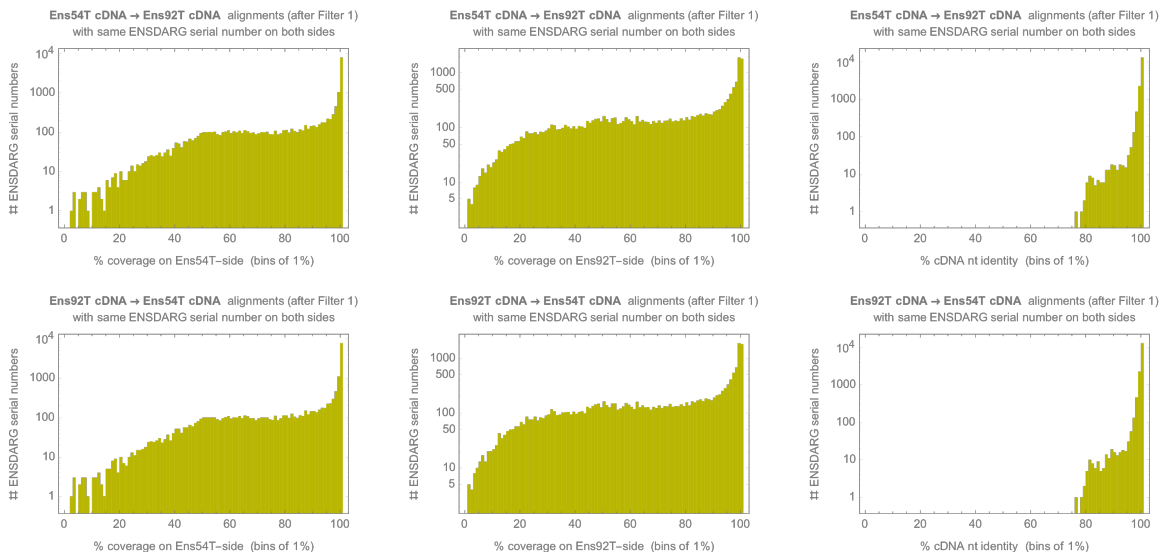
In each direction, for each query-side G, only alignments with bitscore $\geq 95\%$ of the top bitscore were kept (“Filter 2”). For the Ens54T-to-Ens92T direction, ~27k alignments involving 23,746 Ens54Gs survive, and 22,881 Ens54Gs have only a single alignment; for the Ens92T-to-Ens54T

direction, ~29k alignments involving 24,359 Ens92Gs survive, and 22,132 Ens92Gs have only a single alignment.

Before final acceptance of a particular Ens54G–Ens92G pairing, it is reasonable to require a minimum on the percentages (“coverages”) of all nucleotides on each side (Ensembl 54 and 92) that are actually aligned in the supporting transcript–transcript cDNA–cDNA alignment, and that the supporting alignment be of a minimum nucleotide percent identity. Aided by examination of histograms, thresholds of $\geq 30\%$ coverage on each side and $\geq 93\%$ identity were selected, and alignments failing either or both of these criteria were dropped (“Filter 3”).



Interestingly, not all of the 16,180 G serial numbers that have at least one Ens54T with a cDNA and at least one Ens92T with a cDNA have an alignment (e.g., after Filter 1) in both directions (74 are missing at least one direction), or an alignment with high coverages (as apparent from the histograms below). Hence, as numerical identity of G serial number does not always imply a high degree of cDNA sequence similarity between some cross-release pair of transcripts for the nominal single gene, we decided to not treat G or T serials that happen to exist in both releases as special in any way — instead, for such, each side’s instance is taken to operate independently of the other side’s and is filtered/processed just as generic instances of its side.



After Filter 3, the Ens54T-to-Ens92T direction involves 20,299 Ens54Gs (19,706 with a single alignment) and the reverse direction involves 19,797 Ens92Gs (18,492 with a single alignment). There are 19,897 Ens54G–Ens92G pairs (involving 19,130 Ens54Gs and 18,722 Ens92Gs) that have an alignment passing all filters in both directions (our overall strategy being one of “reciprocal near-best hits”); form the undirected bipartite graph with these pairs as edges (using distinct vertices for serial numbers that happen to exist in both releases). Of the 18,388 connected components of this graph, 18,360 (involving 19,020 Ens54Gs and 18,600 Ens92Gs) are complete bipartite, with 17,673 being just a single Ens54G incident to a single Ens92G, i.e., Ens54G:Ens92G 1:1. The other complete bipartite components are variously $n:1$ with $n=2$ to 5; $1:m$ with $m=2$ to 18; or $n:m$ with $n=2$ to 9 and $m=2$ to 8. The 28 non-complete components involve 2 to 19 Ens54Gs each (total 110) and 2 to 35 Ens92Gs each (total 122). We accept the complete bipartite connected components as our mapping between Ens54Gs and Ens92Gs: within every such component, we consider every Ens54G in it to map to every Ens92G in it, and every Ens92G in it to map to every Ens54G in it, and these to be our only mappings. (As our primary focus here is to enable basic comparison of our study with DLG, we do not try to, e.g., recover additional mappings from the non-complete bipartite connected components, or try to resolve $n:m$ components with $n > 1$ and/or $m > 1$ in more detail.)

We finally convert each DLG T to an Ens54G and apply our mapping to Ens92Gs, thereby retaining higher numbers of genes than converting by either retaining only unchanged serial numbers, or by using the Ensembl ID History Converter, while also guaranteeing a minimal degree of cDNA sequence similarity:

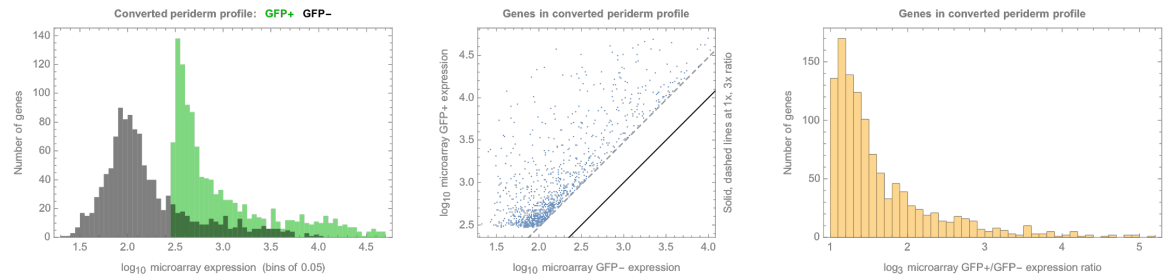
Profile	1:1	($n \geq 2$):1	1:($m \geq 2$)	($n \geq 2$):($m \geq 2$)	Unmapped	Total DLG Ts
periderm	1,128 ~82%	26 ~2%	10 ~1%	2 ~0.1%	203 ~15%	1,369
<i>dnlnf6</i> -inhibited	333 ~86%	13 ~3%	4 ~1%	0 0 %	35 ~ 9%	385
intersection	85 ~92%	2 ~2%	1 ~1%	0 0 %	4 ~ 4%	92
union	1,376 ~83%	37 ~2%	13 ~1%	2 ~0.1%	234 ~14%	1,662.

With our goal being transition to Ens92Gs, we focus on the ($n \geq 1$):1 mappings, and take this opportunity to resolve DLG Ts that collide by being isoforms of the same Ens54G and/or mapping to the same Ens92G as follows:

	DLG T ENSDART#	Profiles	Ens54G ENSDARG#	Ens92G ENSDARG#	Resolution
(1a)	00000074125	<i>dnlnf6</i> -in.	00000002172	00000002172	Drop (1b) as, all else equal, prefer unchanged serial #s (and (1a) cDNA has better alignment)
(1b)	00000114915	<i>dnlnf6</i> -in.	00000078356	00000002172	
(2a)	00000109977	periderm	00000075638	00000087584	Drop (2b) as microarray expr. is higher in (2a) (with similar expr. ratio) + (2a) cDNA aligns better
(2b)	00000111361	periderm	00000079218	00000087584	
(3a)	00000027701	all three	00000036834	00000090268	Drop (3b) as microarray expression is much higher in (3a) (but similar GFP+/GFP– ratio)
(3b)	00000066620	periderm	00000036834	00000090268	
(4a)	00000106367	periderm	00000071790	00000094736	Drop (4a) as microarray expression is higher in (4b) (but with similar GFP+/GFP– ratio)
(4b)	00000106595	periderm	00000071790	00000094736	
(5a)	00000056398	<i>dnlnf6</i> -in.	00000033485	00000103254	Drop (5b) as no microarray expr. available, but choices in same profiles and (5a) aligns better
(5b)	00000090984	<i>dnlnf6</i> -in.	00000067766	00000103254	
(6a)	00000105024	<i>dnlnf6</i> -in.	00000036832	(multiple)	Drop (6a) (multiple Ens92Gs, so already will not use these DLG Ts, but (6b) has very high microarray expr. whereas (6a) expr. unknown).
(6b)	00000105036	all three	00000036832	(multiple)	

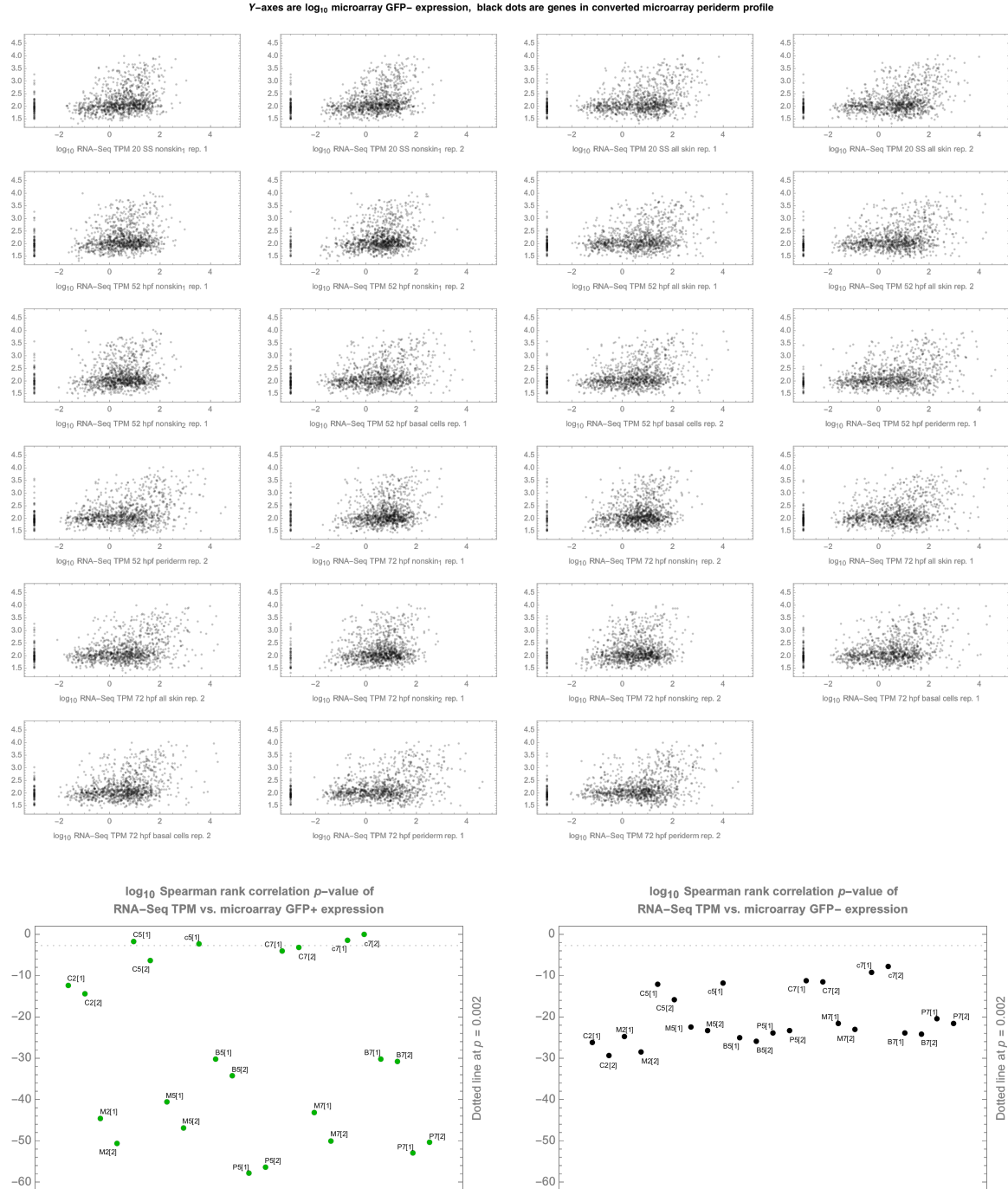
We thus arrive at our final conversions of the DLG profiles (where surviving DLG Ts, Ens54Gs, and Ens92Gs for the union of the profiles are in 1:1:1 correspondence), and the periderm, *dnlnf6*-inhibited, and intersection profiles now contain 1,151; 344; and 87 Ens92Gs, respectively.

File S7 gathers mapping details, the converted profiles, and selected data for mapped Ens92Gs from the master Microsoft Excel workbook included in our NCBI GEO submission (GSE132304), as well as the distributions of our flow classifications across the converted profiles (including comparisons of these to the flow frequencies over our whole Ensembl 92 analysis), these being briefly discussed in the main text. We close this document by examining the microarray expression values (available only in the periderm profile), as these are not discussed elsewhere.



The microarray expression values / probes are from a much earlier timepoint than our study, are based on different technology (hybridization vs. counting), and are being interpreted indirectly (via our mapping procedure described above) across a decade of zebrafish genome assembly and gene modeling changes. Hence, we do not expect tight correspondence of expression levels, but there are, nevertheless, highly non-random isotonic correlations between the microarray values and our study, as seen below.





(RNA-Seq replicate names are abbreviated following (#) earlier in this document.) All of our skin-related RNA-Seq replicates (M, B, P) are monotonically closer to the microarray GFP+ periderm expression than any of our non-skin replicates (C, c), with the closest being P5[1] and P5[2], our earliest periderm replicates. Monotonic correspondence of our RNA-Seq to GFP- microarray expression is, perhaps unsurprisingly, not as strong (being limited to genes DLG determined as enriched in periderm), with C2[2], M2[2] — the closest replicates — being from our closest timepoint. Using our per-condition normal distribution model means instead of TPMs produces similar comparisons (not shown), and with RNA-Seq M2 the best overall condition matching the available microarray GFP- expression.

Fittingly, the monotonically closest RNA-Seq contrast to the microarray GFP+/GFP- expression ratios is M2/C2, our skin-to-nonskin ratio at our closest timepoint to the microarray study. ■

