Supplementary Appendix

For Kavanagh et al. Developmental Bias in the Evolution of Phalanges

List of Supplementary Material:

SOM 1: Tests of distribution within the morphospace of phalanges proportions.

1A. A test for determining how well the data is distributed on a line.

1B. A test for determining how well the data is distributed on a plane.

1C. Tests for whether a 2-trait dataset stems from the same distribution as another dataset

1D. Results for Bird populations (Chicken and Zebrafinch)

1E. Results for Major vertebrate taxonomic groups

1F. Results for Bird species for Digit IV P1-P4

1G. Statistical significance of digit 4 data being on a plane

1H. Results for Darwin's finches

SOM 2: Developmental methods and analyses

2A. Time for formation of a phalanx from proximal to distal joint interzone.

2B. Proliferation study of developing phalanges

2C. Barrier experiments: Metatarsal vs Phalanges

SOM 1: Tests of distribution within the morphospace of phalanges proportions.

By Oren Shoval

1A. A test for determining how well the data is distributed on a line.

Criterion for linear relationship between traits

Here we present a statistical test of whether a dataset in two dimensions is well described by a line. Principal component analysis (PCA) is used to measure the 'linearity' of the data: PCA returns the variance of the data along the first and second principal components. The ratio between these two variances, $v_r=var(PC_2)/var(PC_1)$, is a measure of the correlation of the data. The lower v_r, the more the data is distributed along a line. As an example, consider the data for the birds dataset Fig. S1. The percent variance of PC1 and PC2 is 94.4% and 5.6%, respectively, yielding $v_r=0.059$.

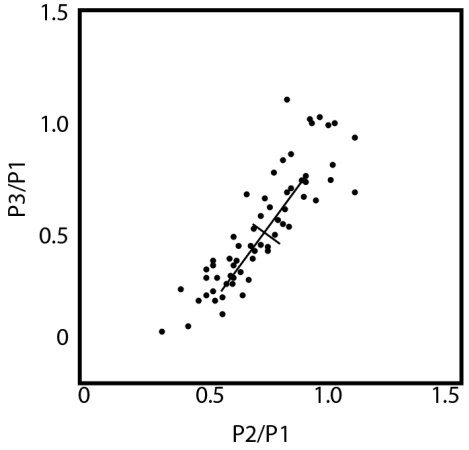

Figure S1. Principal component analysis of the birds dataset. Black lines depict the two principal components. Line length corresponds to the standard deviation (STD) of each component. Crossing point is at average of data in both axes.

Generating a randomized data set

To obtain a statistical significance for the linear-relationship criterion of a dataset, we compare it to a null model - made of an ensemble of suitably randomized datasets. We chose for this purpose a null model that preserves the statistics of each trait, but that reflects a situation where the traits are independent of each other. The null model thus assumes that the two coordinates of the data (x,y) are independent. We generated a large number (10^4) of randomized datasets as follows: each dataset is comprised of the same number of points N as the original dataset. Each point has an x value drawn from the CDF (cumulative distribution function) of the original data's x values, and a y value drawn from the CDF of the original data's y values (Figure S2). We repeat this process until we have the number of points as in the original data set. In this method the null model's x and y CDFs coincide with the CDFs of the x and y of the original data, but we eliminate the relationship between the x and y value (Figure S2). For the randomized dataset 63.8% of the variance is explained by PC_1, 36.2% by PC_2, yielding v_r=0.57, showing that it is significantly less correlated than the original dataset.

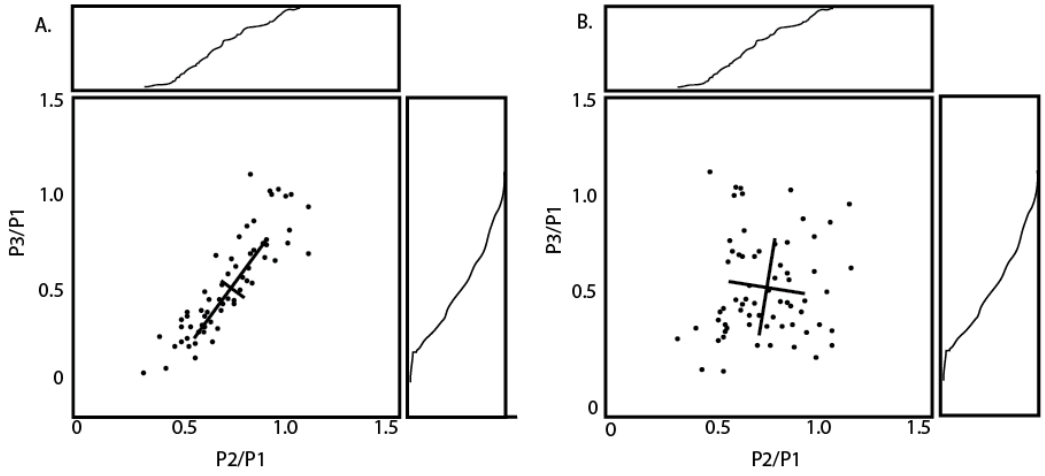

Figure S2. A. Cumulative distribution functions (CDF) of x (P2/P1) and y (P3/P1) values of birds dataset. B. Randomized data set, with CDFs equal to those of the original dataset.

Computing the p-value

To find the p-value for the linearity of a dataset, we first compute v_r - the ratio of variances of the original dataset. We then generate random datasets as described above. For each random dataset we calculate v_r. The resulting p-value is the fraction of randomized datasets for which v_r is lower than the original dataset's v_r. Statistics for 10,000 randomized datasets based on the birds dataset, are shown in Figure S3. Since all 10,000 randomized datasets have a higher v_r, the p-value is smaller than 10^{-4}.

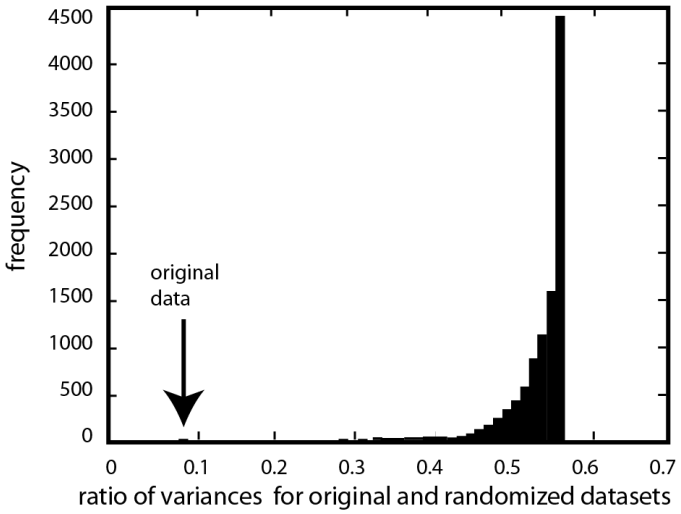

Figure S3. Histogram of v_r – the ratio between variances of the principal components for 10,000 randomized datasets. The original dataset has a v_r value of 0.059, which is lower than the values for all randomized datasets, leading to a p-value < 10^{-4}.

1B. A test for determining how well the data are distributed on a plane.

Criterion for planar relationship between traits

In this case there are three traits – P4/P1, P3/P1, and P2/P1, leading to a three dimensional morphospace. Here, we use a similar method to the one described above in order to analyze how well the data falls on a plane. As a measure, we use the ratio of variances between the 3rd and 2nd PCA components - $v_r=var(PC_3)/var(PC_2)$. As an

example, consider the data for the birds dataset (Figure S4). The percent variance of PC1, PC2, and PC3 is 92.2%, 7%, and 0.8%, respectively, yielding v_r=0.12.

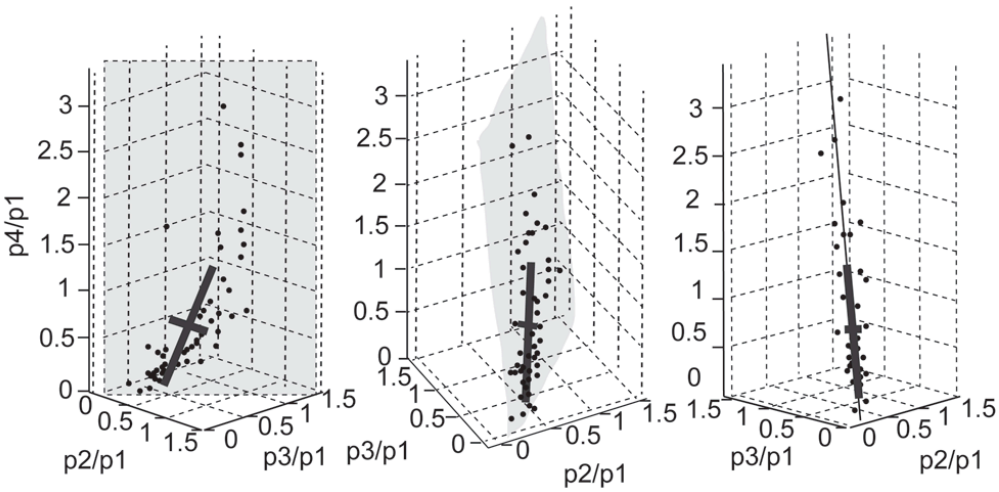

Figure S4. Three dimensional morphospace of digit 4 phalanges. Principal component analysis of the birds dataset. Blue lines depict the three principal components; line length corresponds to the standard deviation (STD) of each component. Crossing point is at average of data in the three axes.

Generating a randomized data set

In a similar fashion to the line criteria discussed above, we obtain a significance measure for the plane criterion of a dataset, by comparison to a null model made of an ensemble of suitably randomize datasets. Each randomized data set preserves the statistics of each trait, but that reflects a situation where the traits are independent of each other. Again, each dataset is comprised of the same number of points N as the original dataset, where x, y, and z values are drawn from the corresponding CDFs. An example of a randomized dataset produced with this procedure for the birds dataset is shown in Figure S5. For each randomized data set, the variance ratio of components 3 and 2 is computed. v_r of the randomized dataset is 0.58, showing that it is less planer than the original dataset.

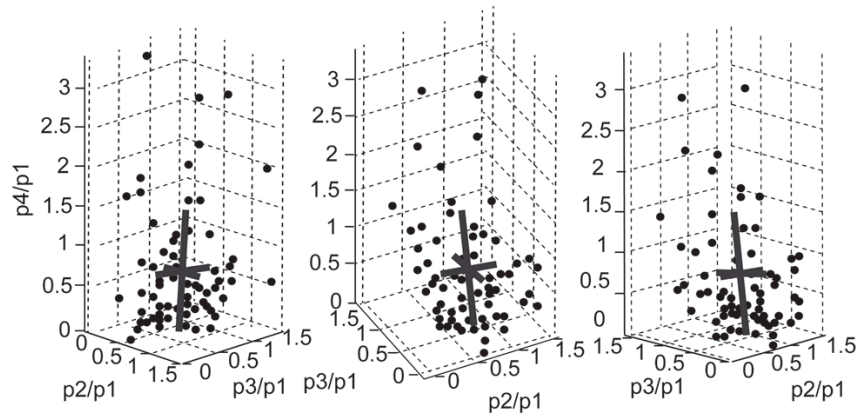

Figure S5. Randomized dataset. Blue lines depict the three principal components of the randomized dataset.

Computing the p-value

To find the p-value for how well the data falls on a plane, we compute v_r of the original dataset, and compare to 10,000 randomized datasets. The z-score is 5.4.

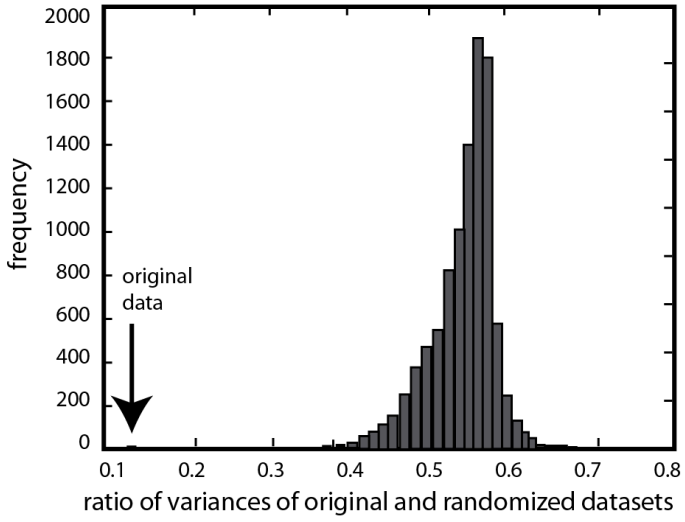

Figure S6. Histogram of v_r for 10,000 randomized data sets. The original dataset has a v_r lower than all randomized dataset, leading to a p-value $< 10^{-4}$.

Extension: calculating whether the Z plane is off the vertical

If the value of the third trait (p4/p1), which is displayed on the z-axis, is not dependent on the other two traits (p2/p1, p3/p1), that represent the x-y axes, we would expect the best-fit plane to be vertical. This would imply that p4/p1 varies independently from p2/p1 and p3/p1. Here we provide a test that examines whether the best-fit plane is off the vertical, and provides a p-value.

First, we define a measure of how vertical is the best-fit plane, which is calculated using principal component analysis. The best-fit plane is found using the first two components. Consider the 3rd principal component, which is perpendicular to the best-fit plane (principal components are always perpendicular to each other). Note that if the 3rd component is parallel with the x-y plane, then the best-fit plane is perpendicular to the x-y plane. Thus, we can use the ratio of the z-value and the x-y values of the 3rd principal component, to determine how vertical is the plane.

In order to test for a p-value, we create randomized datasets as described above. For each dataset we compute the ratio *l*, and compare with the ratio of the original dataset, to get the p-value.

1C. Tests for whether a 2-trait dataset stems from the same distribution as another dataset

Consider the birds (primary) and cetacean (secondary) datasets depicted in Figure S7. Here we present several methods for testing whether the cetacean dataset stems from the same distribution of the birds dataset.

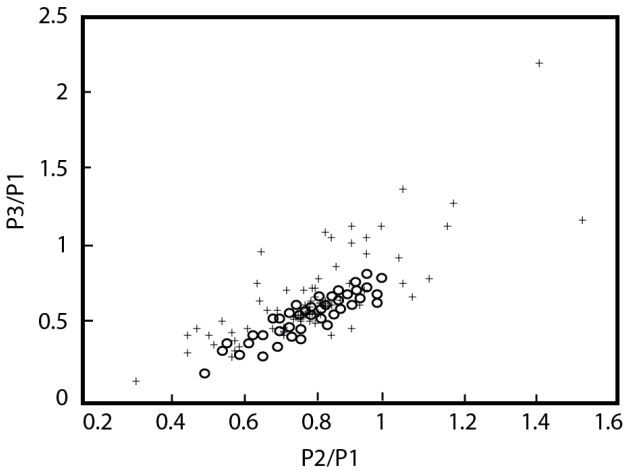

Figure S7. Digit III ratios for the birds dataset (grey dots), and the cetacean dataset (black circles).

A test of whether the secondary dataset is aligned with the primary dataset

We begin by performing principal component analysis of the main dataset (birds in our example, Figure S8A). The percent variance of the second dataset along the first principal component axis of the main dataset is a measure of the alignment of the two datasets (Figure S8B). Next we create randomized datasets based on the statistics of the second dataset (in a similar fashion to the randomizations described above). Each randomized dataset has the same number of measurements as the original dataset, and the same x and y distributions. For each randomized dataset the percent variance along the first principal component of the main dataset is computed (Figure S8C). These results are compared with the values for the original dataset to compute the p-value.

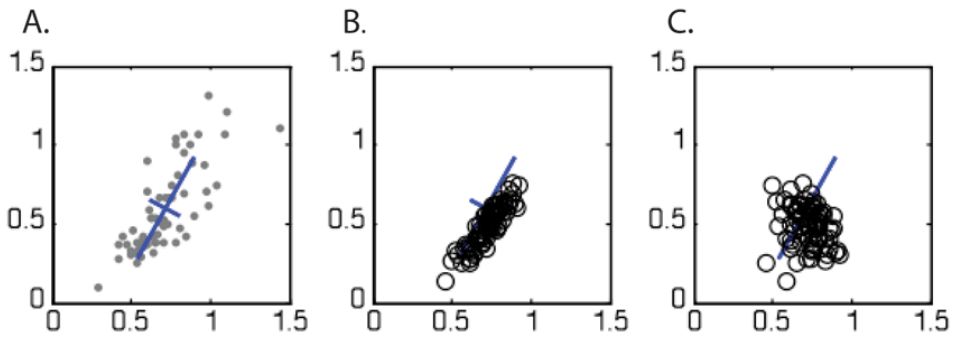

Figure S8. A. Main dataset, and the principal components (blue lines). B. Secondary dataset, with principal components of main dataset. C. Randomized dataset with principal components of main dataset.

A test of whether the secondary dataset is not centered along the primary dataset

The two datasets might be aligned, but parallel, in the sense the intercept is significantly different. Here we present a method of testing whether the secondary dataset is not centered along the primary dataset, in a statistically significant manner.

First the principal components of the primary dataset are computed. For each datapoint j in the secondary dataset, we find the distance d_j from the line defined by the 1st principal component of the main dataset (Figure S9, black lines). The distribution of d_j for the birds and cetaceans example is shown in Figure S9. Note that the distribution is not centered at zero - if the secondary dataset was centered on the main dataset, the mean of the distances would be zero. We now perform a t-test, which tests whether the data in vector d_j has a mean that is not zero. The t-test returns the confidence level, where the standard threshold used is 5%. Thus, if the t-test returns a value higher lower than 5%, than we can reject the hypothesis that the secondary dataset is centered with the main dataset. In this case, the t-test returns 0 – indicating that there is a high probability that the mean of the dataset is not zero, which implies that the two datasets are not centered on the same line.

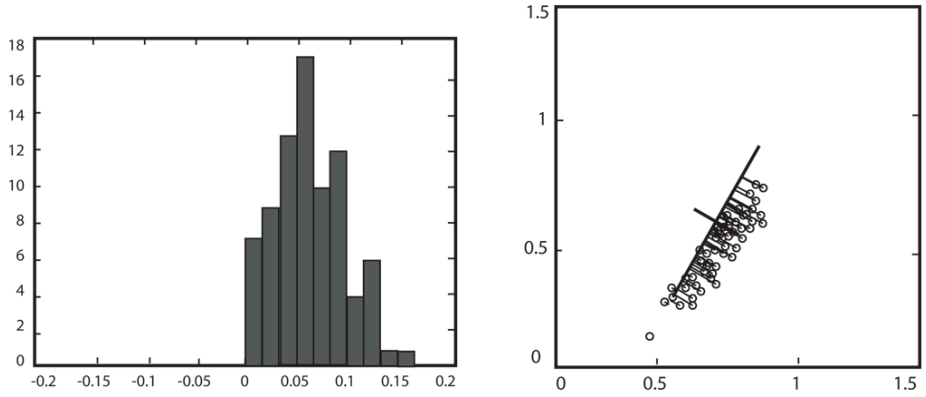

Figure S9. A. Cetaceans dataset, and principal components of the birds dataset. For each datapoint the distance from the 1st principal component is shown by a black line. B. Distribution of distances of datapoints from 1st principal components.

Conducting these Tests on particular datasets:

1D. Results for Bird populations (Chicken and Zebrafinch) for Digit IV, P1-3

Do phalanges vary independently in bird populations?

Figure S10 depicts the variation of the phalanges' proportions of the bird populations. Using the statistical method described above, we find that the phalanges ratios are dependent (p-value < 10^{-4}).

Among chick individuals, using the same test, we find that variation in proportions is not random (p-value < $2*10^{-4}$).

Among zebrafinch individuals, similarly, we find that the variation is not random, with p-value < 10^{-4}.

Are proportion variants within populations similar to proportion variants among species?

Among chick individuals, variation is not different from dataset among all birds: both data sets are aligned (p-value < $2*10^{-4}$), and their means cannot be distinguished (ttest>0.05).

Among zebrafinch individuals, variation is not different from among all birds: both data sets are aligned (p-value < 10^{-4}), and their means cannot be distinguished (ttest>0.05).

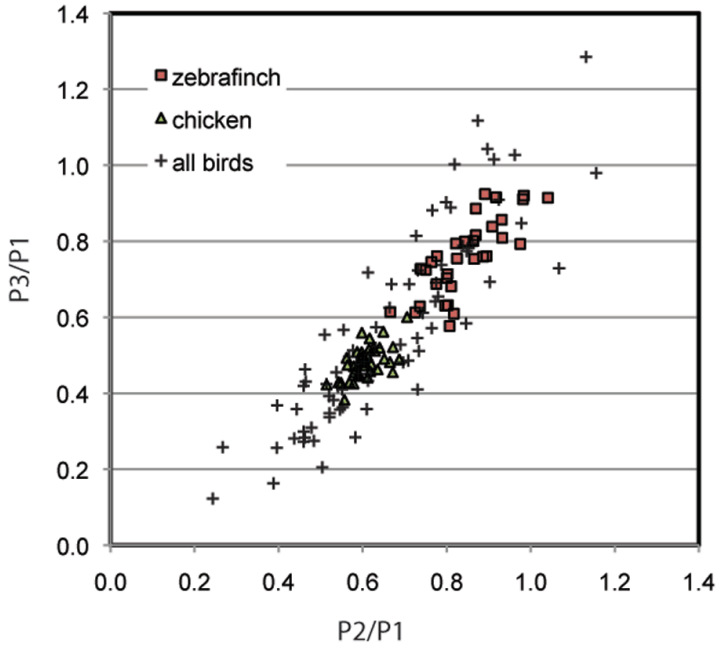

Figure S10. Phalanges proportions for Digit IV, P1-P3 for zebrafinch and chicken populations and means of 76 species of birds.

1E. Results for Major taxonomic group proportion variants

Do phalanges proportions vary independently in vertebrate groups?

No, variation is not random. They all fall on a line from proximo-distal gradient to equal-sized (see p-values in table below).

Is the range of phalanges proportion variants in all major vertebrate taxonomic groups similar?

Among birds, lepidosaur reptiles, lissamphibians, ichthyosaurs, plesiosaurs, and cetaceans, there is no detectable difference in slope. In some groups there is a significant difference in intercepts (see table below with statistical results).

Taxon	Distribution not random p-value	Aligned with the SARC group p-value (similar slope)	Mean distinguishable from SARC group (similar intercept)
Birds	$<10^{-4}$	$<10^{-4}$	Yes
lepidosaur reptiles	$<10^{-4}$	$<10^{-4}$	Yes
lissamphibians	$<10^{-4}$	$<10^{-4}$	No
ichthyosaurs	$<10^{-4}$	$<10^{-4}$	Yes
plesiosaurs	$<10^{-4}$	$<10^{-4}$	No
cetaceans	$<10^{-4}$	$<10^{-4}$	Yes
early SARC	$<10^{-4}$	-	-

1F. Results for Bird species P1-P4

Do phalanges proportions in birds fall along a plane in morphospace?

Yes (p-value$<10^{-4}$).

Statistical significance of digit IV data being on a plane (test 1B)

Fig. 5 in the main text depicts the ratios of phalanges' areas relative to the area of phalanx 1. In three perspectives of this 3 dimensional plot we see that the data is limited to a plane. Using principal component analysis we find the relative variances of the three principal components are 85.2%, 12.4%, and 2.4%. The low variance of the 3rd component mathematically shows that the data fall on a plane. Here we examine the statistical significance of this finding, and evaluate what is the probability that if the

different phalanges were drawn from unrelated distributions, we would get such a result. In summary, using the bootstrapping method, we build a new dataset of the safe size, with 65 samples, where for each sample, the phalanges area ratios are drawn randomly from their distribution. See detailed explanation below.

After normalization by the first phalanx, the data has three variables: p2/p1, p3/p1, and p4/p1. We denote them by A, B and C respectively. There are a total of 65 sample points in the data, each one with a value for A, B and C. in order to compare the three variables in the same scale, we normalize each by its standard deviation (z-score). Using the bootstrapping method, we draw by random a value from A, a value from B and a value from C. This creates a new sample that has phalanges' ratios chosen from three random birds' measurements. This step is repeated 65 times in order to create a data set the same size as the original one. In order to examine if the data in this randomized data set also falls on a plane, principal component analysis is used, and the variance of the third component is analyzed. The above analysis is repeated 10,000 times to get a distribution of the measure of the variance of the third principal component. This distribution has a mean value of 24%, and a standard deviation of 3%. The original data set gives us a variance of 2.4% for the 3rd principal component, which is 7 standard deviations from the mean. For a normal distribution the corresponding p-value is $6.5*10^{-13}$.

In summary, if there was no relationship between phalanges' areas, the probability of finding the data set falling on a plane is extremely low.

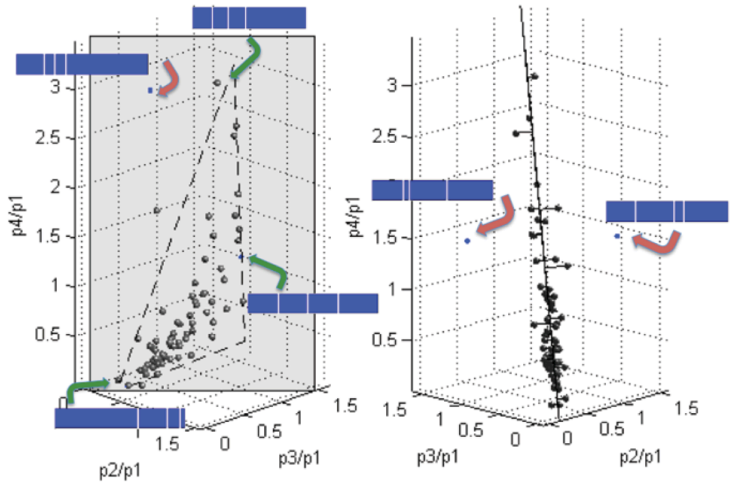

Figure S11. Two rotations of 3D morphospace showing proportions of bird Digit IV phalanges P1-P3 (dots). Dashed triangle shows area generally covered by the range of variation in proportions for this digit. Blue rectangles illustrate proportions; left is proximal. Green arrows point to proportions of some observed species. Red arrows point to proportions that are off the plane or outside the observed range, thus apparently not found in nature.

1G. Statistical significance of digit 4 data being on a plane (test 1B)

Is the Z-plane off the vertical?

Yes, slightly but significantly, indicating P4 variation is also linked to the size proportions of the other phalanges.

1H. Results for Darwin's finches

Arboreal species of Darwin's finches have significantly more elongated P4 phalanges than ground species. The overall proportions of Darwin's finches fall within the plane defined by proportions from all birds.

SOM 2: Developmental methods and analyses

Kathryn Kavanagh, Benjamin Winslow, and Akinori Kan

2A. Time for formation of a phalanx from proximal to distal joint interzone.

The final proportions of the phalanges are established during the period of sequential joint formation in the embryo, occurring over about three days of development in the chick. Final proportions of Emu, Chick, and Barn Owl, three species with very different proportions, are observable at the time of tip formation (Figure S12). Individual phalanges are established between the time when the proximal joint interzone forms and when the distal joint interzone appears. Regulation of the time between sequential joint interzone formation during this period is thus potentially one of the developmental mechanisms that regulates final proportions in the digit.

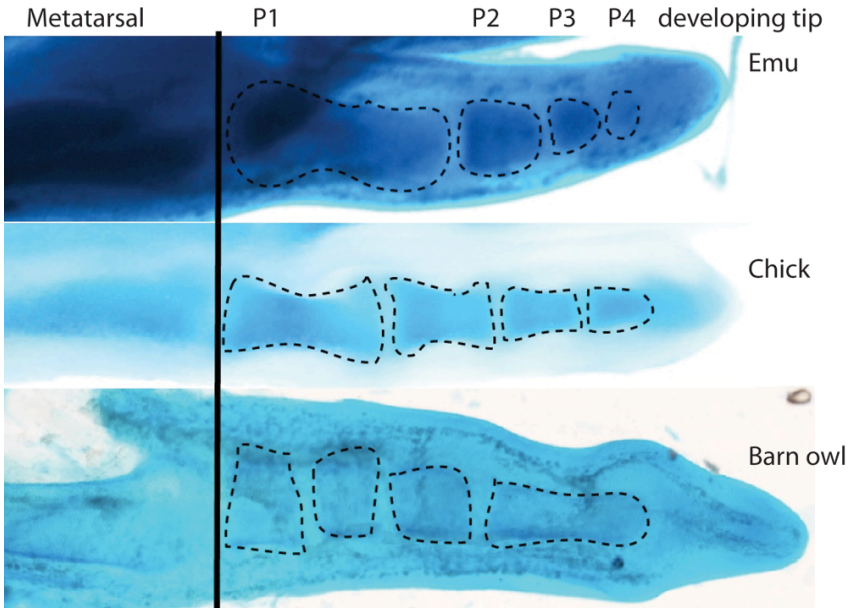

Figure S12: Late phalanx development in emu, chick, and barn owl embryos, just before tip is formed. This demonstrates that the proportions are determined during this morphogenetic phase and not due to post-morphogenetic growth differences.

To calculate number of hours from formation of proximal joint to distal joint in a phalanx, which is the period of patterning and developmental plasticity for that phalanx, we performed the following steps:

Method 1: Time series

1. Over several months, we obtained batches of fertilized chick eggs (>20 batches; Charles River Labs), which had been collected from nests over several hours and cooled to 18°C. Soon after they were received, the entire batch began 38°C incubation to initiate synchronous development.
2. We collected and fixed groups of three embryos at intervals (usually 2-hrs) throughout digit developmental stages.
3. Embryos were KOH cleared and Alcian stained. Distal limbs were removed from the embryos and then photographed from a dorsal (top) view.
4. The 2D area of each developing phalanx (Alcian stained area) was measured using ImageJ. Stage of joint formation in Digit IV recorded as a way of staging digit development so that equivalent stages could be compared.
5. Average condensation area for each phalanx was calculated.
6. All condensation sizes were aligned to find the smallest initial condensation size; the sample size was determined to be sufficient if the smallest 5 condensations do not differ more than ~5%.
7. In order to find the condensation size at which the distal joint of a given phalanx is formed, we determined the size of a given phalanx at the time when the next phalanx has the smallest initial condensation observed (since that is immediately after the joint interzone is formed).
8. The growth rate of a given phalanx was determined by examining phalanx condensation size increase over time in our series, and dividing by the number of hours between collections.
9. The number of hours to form a given phalanx was then calculated by dividing the growth rate by the difference in size between initial condensation and the condensation at the time of distal joint formation.

[(Area at t2) – (area at t1)] / (growth/hr) = # hrs

t=time

Method 2: Chick embryo cut-foot pairs

1. Chicks were incubated to day 7 or 8.
2. A window was opened in the egg and amnion, avoiding blood vessels. One hindlimb autopod was removed with micro-scissors for fixation, and the egg was returned to the incubator with tape over the window.
3. The embryo was allowed to incubate an additional 6-48 hrs before collecting/fixing the other hindlimb autopod.
4. Limbs were Alcian stained and KOH cleared.
5. The number of additional joints was determined by counting the areas of clear tissue (no Alcian stain) indicating the developing joint interzones. The difference in condensation size for a given phalanx between first and second collection was measured.
6. The maximum number of hours before a new joint is observed in a digit was determined as an estimate of the number of hours to form a given phalanx.

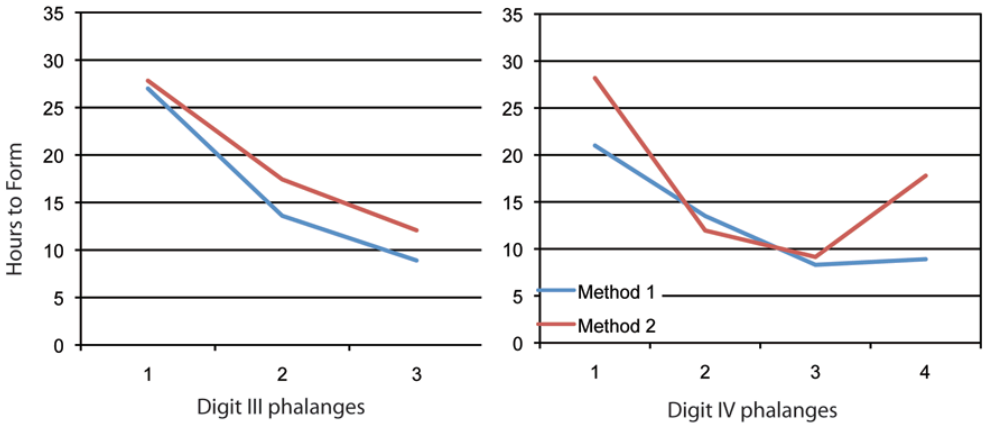

Figure S13: Average time of formation (number of hours between proximal and distal joint interzone formation) for phalanges of Digit III (left) and Digit IV (right).

2B. Proliferation study

We injected 400µl of 1mM EdU solutions into amniotic fluid, and harvested embryos 6 hours after the injection. We made a paraffin section and used Click-iT EdU Alexa Fluor 555 Imaging kit (Invitrogen) for imaging as previously reported (Dev Dyn. 238:944–949, 2009). Proliferation rates were calculated by counting numbers of labeled cells in 200um quadrants.

2C. Barrier experiments

Pre-cut tantalum foil barriers were implanted into digit IV of the right hindlimb using forceps during day 5 (metatarsal barriers) or days 6-7 (phalanges barriers) of development. To perform the microsurgeries eggs were windowed, the amniotic sac was opened with forceps, and the hindlimb was placed on a dark paper stage to provide contrast. Barriers were inserted through the distal end of the digit condensation, after which the limb was returned to the amniotic sac. Penicillin/streptomycin was then added, the egg was sealed with tape, and returned to the incubator. Eggs were incubated until day 10 or 11, when embryos were collected and fixed in formalin over night. Feet were removed and stained for cartilage with Alcian blue, then cleared in KOH.

Wound controls were conducted exactly as above, except that foil barriers were inserted and then removed after ~ 1 minute.

Cleared and stained feet were photographed, and the fourth digits from the experimental and contralateral feet were aligned using Adobe Photoshop. For metatarsal barriers, pairs of digits were visually inspected to determine if the metatarsal was noticeably shortened, and if clear changes occurred to the phalanges.

To determine if phalanges were affected in wound control barriers, first the amount of variation between the same phalanx on the left and right foot in normal embryos was assessed. The area of each phalanx was measured using ImageJ software, and the percent difference of each phalanx was determined for 19 normal chick embryos at day 10-11. From these measurements the average percent difference for all pairs of phalanges was calculated (9.4%), as was the standard deviation (7.7%). The percent

difference between experimental and contralateral control phalanges sizes were then compared to left-right variation observed in normal embryos. Specimens were scored based on if phalanges size differences exceeded the average amount by more than 1, or more than 2 standard deviations. 33/45 (73%) of the experimental digits contained phalanges where the percent difference from the contralateral side exceeded the average plus 1 standard deviation, and 29/45 (64%) exceeded the average percent difference by more than 2 standard deviations. In wound controls, only 4/13 (31%) digits contained phalanges that exceeded the average percent difference plus 1 or 2 standard deviations. Images of wound controls and phalanges experimental barriers were also aligned in Photoshop and assessed visually.