Supplemental Materials for

Early embryonic mutations reveal dynamics of somatic and germ cell lineages in mice

Arikuni Uchimura*, Hirotaka Matsumoto, Yasunari Satoh, Yohei Minakuchi, Sayaka Wakayama, Teruhiko Wakayama, Mayumi Higuchi, Masakazu Hashimoto, Ryutaro Fukumura, Atsushi Toyoda, Yoichi Gondo, Takeshi Yagi

*Correspondence to: uchimura@rerf.or.jp

This PDF file includes:

Supplemental Text
Supplemental Figs. S1 to S9
Supplemental Tables S1 to S5

Other Supplemental Materials for this manuscript include the following:

Supplemental Data 1 to 3 (Excel)
Supplemental Code (Zip file)

Supplemental Text

Here, we describe the detailed procedures and algorithm for lineage reconstruction and lineage origin inference of offspring. The scripts and variant allele frequency (VAF) data used for lineage reconstruction are shown in Supplemental Code.

Reconstruction of cell lineage tree from VAF data

Our goal was to trace mosaic mutation occurrence and reconstruct the cell lineage tree. In the following section, we explain the way by which the lineage tree was reconstructed using VAF values of mosaic mutations.

Outline of the lineage reconstruction algorithm

We introduce the outline of the lineage reconstruction algorithm with small-scale data consisting of the VAF values of six mosaic mutations for three tissues (Supplemental Text Fig. A). We describe the VAF value of a mutation m for a tissue n as $\mathbf{X}_{m,n}$.

First, we searched a partial lineage relationship corresponding to a parent node *i* and two daughter nodes *j* and *k* that satisfied,

$$\mathbf{X}_{i,n} = \mathbf{X}_{j,n} + \mathbf{X}_{k,n}. \quad (1)$$

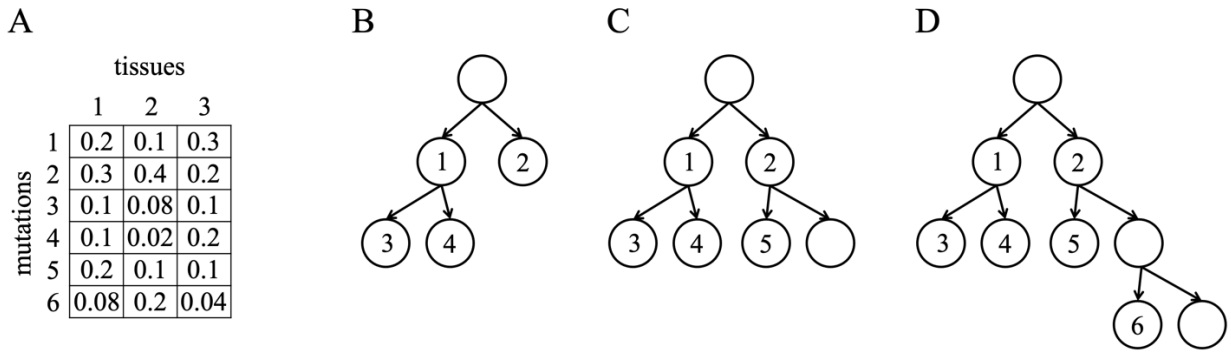
In this case, $0.5 = \mathbf{X}_{1,n} + \mathbf{X}_{2,n}$ is satisfied for all tissues, implying that mutations 1 and 2 correspond to two daughter cells of the zygote. $\mathbf{X}_{1,n} = \mathbf{X}_{3,n} + \mathbf{X}_{4,n}$ is also satisfied for all tissues, and mutations 3 and 4 correspond to two daughter cells of mutation 1 cell. Consequently, the lineage tree was constructed, as shown in Supplemental Text Fig. B.

Next, we searched the additional parent–daughter relationship based on the magnitude relation of the VAF values. If a mutation *i* is in the downstream lineage of mutation *j*, the following relationship is satisfied,

$$\mathbf{X}_{i,n} \leq \mathbf{X}_{j,n}. \quad (2)$$

This means that *i* must not be in the downstream lineage of *j* if $\mathbf{X}_{i,n} > \mathbf{X}_{j,n}$ for some tissues. Therefore, if *i* satisfies the relationship $\mathbf{X}_{i,n} \leq \mathbf{X}_{j,n}$ only for terminal node *j*, it can be considered that *i* is in the downstream lineage of j. In the case of mutation 5, Eq. (2) is always satisfied only for terminal mutation 2, and it will be in the downstream lineage of mutation 2 (Supplemental Text Fig. C). We also added pseudo node *p* corresponding to the other daughter cell and defined the VAF values of the pseudo node as $\mathbf{X}_{p,n} = \mathbf{X}_{2,n} - \mathbf{X}_{5,n}$. Using the above strategy recursively in decreasing order of averaged VAF values of unassigned mutations, the entire lineage tree could be reconstructed (Supplemental Text Fig. D).

Hereafter, we describe the detailed procedure of each step, including the data preprocessing.



Supplemental Text Fig. Outline of the lineage reconstruction algorithm

VAF data pre-processing

Even if the individual does not have a mosaic mutation, the VAF values of the mutation may show small values owing to an artifact-like misalignment and sequencing error. To correct the overestimated VAF values, we used the VAF values of the mouse that does not have the mutations as control data as follows. We describe the VAF value of a mosaic mutation m for a tissue n as $\mathbf{X'}_{m,n}$ and that for a control sample as $\mathbf{X'}_{m,\mathrm{Ctrl}}$. In the case of a mutation in the sex chromosome, we divided the frequency by 2. Then, we used the following corrected VAF value for lineage reconstruction,

$$\mathbf{X}_{m,n} = \mathbf{X'}_{m,n} - \mathbf{X'}_{m,\mathrm{Ctrl}}. \tag{3}$$

Because mutations with low VAFs were difficult to use in lineage reconstruction, we removed the mutations that satisfied $\max_n \mathbf{X}_{m,n} < 0.05$. In addition, we added mutation m that corresponded to the root node and satisfied $\mathbf{X}_{m,n} = 0.5$ for all tissues.

Then, we clustered the mutations that occurred in the same stage in the developmental process. We quantified the difference between two mutations i and j as follows,

$$\mathrm{d}(i,j) \quad = \frac{\sum_{n=1}^{N}\left(\mathbf{X}_{i,n} - \mathbf{X}_{j,n}\right)^2}{N \times 0.5 \times \left(\overline{\mathbf{X}}_i + \overline{\mathbf{X}}_j\right)}, \tag{4}$$

where $\overline{\mathbf{X}}_i$ is the mean value of $\mathbf{X}_{i,n}$. We regarded the mutation i and j as occurring in the same stage if $-\ln\left(\mathrm{d}(i,j)\right) \geq \theta$ was satisfied. In this study, we checked the results with different θ and used $\theta =$ 5, 5, 5, 6, and 6 for ConB23, ConC31, ConD31, ConE29, and ConJ12, respectively. Then, we represented such relationships with a graph (nodes represent mutations and edges represent the above connections) and defined the mutation clusters as connected components of the graph. Hereafter, we call clustered mutation sets "mutation nodes" and represent mutations in the i-th mutation node as M_i. Further, we used the following averaged VAF values over mutations in M_i hereafter,

$$\mathbf{X}_{M_i,n} = \frac{1}{|M_i|}\sum_{m \in M_i} \mathbf{X}_{m,n}\,, \tag{5}$$

where $|M_i|$ is the number of mutations in M_i.

Lineage reconstruction

Inference of relationship between one parent cell and two daughter cells

First, we reconstructed the partial lineage relationship that corresponds to a parent node M_i and two daughter nodes M_j and M_k. For M_i, M_j, and M_k, the following relationship will be satisfied,

$$\mathbf{X}_{M_i,n} \simeq \mathbf{X}_{M_j,n} + \mathbf{X}_{M_k,n}. \tag{6}$$

Therefore, we quantified the validity of the parent-daughter relationships as follows,

$$\mathrm{d}\left(M_i, M_j, M_k\right) = \frac{\sum_{n=1}^{N}\left(\mathbf{X}_{M_i,n} - \mathbf{X}_{M_j,n} - \mathbf{X}_{M_k,n}\right)^2}{N \times \left(\overline{X}_{M_j} + \overline{X}_{M_k}\right)}. \tag{7}$$

If $\mathrm{d}\left(M_i, M_j, M_k\right)$ was lower than ϵ_1, we regarded M_i as the parent node and M_j and M_k as the two daughter nodes. We calculated the above score for all combinations of mutation nodes and reconstructed the entire lineage relationship.

In the above procedure, setting an appropriate value of ϵ_1 was necessary because we would estimate false-positive relationships with large ϵ_1 and we would overlook genuine relationships with small ϵ_1. If the reconstructed lineage contains false-positive relationships, there must be inconsistent relationships such that a node

has more than three daughter nodes or more than two parent nodes. Therefore, we used the following procedure to define the appropriate ϵ_1 and reconstruct the lineage tree.

1. Initialize $\epsilon_1 = 0.005$.
2. Enumerate all relationships that satisfy $\mathrm{d}(M_i, M_j, M_k) < \epsilon_1$.
3. If there are some inconsistent relationships within the entirety of the relationships, we update $\epsilon_1 = \epsilon_1 - 0.00025$ and return to step 2.
4. Output the entire relationship as mutation lineage tree.

Additional lineage inference

The above procedure could reconstruct the whole lineage tree if all developmental stages contain mosaic mutations, and it would overlook lineage relationships if no mutations occurred at some developmental stages. Next, we inferred the relationship of the parent-daughter nodes based on magnitude relation of the VAF values. If a mutation node M_i is in the downstream lineage of mutation node M_j, the following relationship is satisfied,

$$\mathbf{X}_{M_i,n} \leq \mathbf{X}_{M_j,n}. \quad (8)$$

It means that M_i must not be in the downstream lineage of M_j if $\mathbf{X}_{M_i,n} > \mathbf{X}_{M_j,n}$ for some tissues. Therefore, if M_i satisfies the Eq. (8) only for mutation node M_j, we can estimate that M_i is in the downstream lineage of M_j. From the above strategy, we inferred additional lineages as follows.

1. We describe the mutation nodes that correspond to the tip nodes of lineage tree that contain the root node as "leaf nodes" L.
2. We describe the mutation nodes that are not reachable from the root and do not have parent node as "orphaned nodes" O.
3. Choose an orphaned node o that has the largest $\sum_n \mathbf{X}_{o,n}$.
4. For all $l \in L$, we investigate that $\mathbf{X}_{l,n} - \mathbf{X}_{o,n} > -\epsilon_2$ is satisfied for all tissues n (We used $\epsilon_2 = 0.01$ in this study.). If this relationship is satisfied, we add l for the parent candidate of o.
5. If there is only one parent candidate l, we regard l as the parent node of o. Then, we add another daughter node of l as pseudo node p. We define the VAF values of p by $\mathbf{X}_{p,n} = \mathbf{X}_{l,n} - \mathbf{X}_{o,n}$. Then, we remove o from O, l from L, and add o and p to L.
6. Otherwise, we remove o from O.
7. If O is not empty, return to step 3.

Inference of lineage origin of offspring

In the previous section, we reconstructed the mutation lineage tree of a male mouse. Next, we investigated the genotype of the offspring of the mouse. In this section, we explain the probabilistic model to evaluate the tree and infer the lineage composition of the germline cells of the male mouse and the paternal developmental lineage origin of the offspring.

Notation

We describe the number of lineage paths for the reconstructed tree as L and the number of mutations in the tree as M, respectively. We also describe the mutation information of the lineage as $L \times M$ matrix $\mathbf{X}$, and the mutation pattern of a lineage path $l \in (1, \ldots, L)$ as $\mathbf{X}_l = (\mathbf{X}_{l,1}, \mathbf{X}_{l,2}, \ldots, \mathbf{X}_{l,M})$, where $\mathbf{X}_{l,m} = 1$ if the lineage path l includes mutation m and $\mathbf{X}_{l,m} = 0$ otherwise. In addition, we describe the genotype information of offspring by $N \times M$ matrix $\mathbf{Y}$, where N is the number of offspring mice. The mutation pattern of an offspring n

is represented with $\mathbf{Y}_n = (\mathbf{Y}_{n,1}, \mathbf{Y}_{n,2}, \ldots, \mathbf{Y}_{n,M})$, where $\mathbf{Y}_{n,m} = 1$ if the offspring n has the mutation m and $\mathbf{Y}_{n,m} = 0$ otherwise.

Mixture model for offspring genotype

We represent the probability of a lineage path l with π_l ($\sum_{l=1}^{L} \pi_l = 1$) that corresponds to the proportion of germline cells originating from the lineage path l. We describe the lineage origin of the offspring n with the latent variable $\mathbf{Z}_n = (\mathbf{Z}_{n,1}, \ldots, \mathbf{Z}_{n,L})$, where only one of them is 1 and all the rest are 0 (1-of-L representation). The probability of the latent variable for offspring n is described as follows,

$$p(\mathbf{Z}_n) = \prod_{l=1}^{L} \pi_l^{\mathbf{Z}_{n,l}}. \tag{9}$$

We assume that each mutation is inherited independently and randomly, and therefore, the probability that offspring n has mutation m is given by

$$p(\mathbf{Y}_{n,m} = 1 | \mathbf{Z}_{n,l} = 1, \mathbf{X}) = \begin{cases} 0.5 & \text{if } \mathbf{X}_{l,m} = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

and the probability that offspring n does not have mutation m is given by

$$p(\mathbf{Y}_{n,m} = 0 | \mathbf{Z}_{n,l} = 1, \mathbf{X}) = \begin{cases} 0.5 & \text{if } \mathbf{X}_{l,m} = 1 \\ 1.0 & \text{otherwise} \end{cases}. \tag{11}$$

Then, the probability for all offspring and mutations is given by

$$p(\mathbf{Y}_n | \mathbf{Z}_n, \mathbf{X}) = \prod_{l=1}^{L} \left(\prod_{m=1}^{M} p\left(\mathbf{Y}_{n,m} | \mathbf{Z}_{n,l} = 1, \mathbf{X}\right) \right)^{z_{n,l}}. \tag{12}$$

However, some mutations are included in the maternal genotype for some offspring. In such cases, we did not use such mutation information. Moreover, we did not use mutations on Chromosome X for male offspring and used the following probability.

$$p(\mathbf{Y}_n | \mathbf{Z}_{n,l} = 1, \mathbf{X}) = \prod_{m \in Mo(n)} p\left(\mathbf{Y}_{n,m} | \mathbf{Z}_{n,l} = 1, \mathbf{X}\right), \tag{13}$$

where $Mo(n)$ represents the effective mutation set of an offspring n.

Parameter optimization with EM algorithm

We optimized the mixture model based on the EM algorithm and calculated the following E-step and M-step iteratively.

E-step

First, we calculated $\gamma(\mathbf{Z}_{nl})$, which is called responsibility. It corresponds to the expected value of $p(\mathbf{Z}_{n,l} = 1 | \mathbf{Y}_n, \mathbf{X})$, which represents that an offspring n originates from lineage l.

$$\gamma(\mathbf{Z}_{n,l}) = \frac{\pi_l \prod_{m \in Mo(n)} p\left(\mathbf{Y}_{n,m} | \mathbf{Z}_{n,l} = 1, \mathbf{X}\right)}{\sum_{l'=1}^{L} \left(\pi_{l'} \prod_{m \in Mo(n)} p\left(\mathbf{Y}_{n,m} | \mathbf{Z}_{n,l'} = 1, \mathbf{X}\right) \right)} \tag{14}$$

M-step

Next, we optimized π_l with the above $\gamma(\mathbf{Z}_{n,l})$.

$$\begin{aligned} N_l &= \sum_{n=1}^{N} \gamma\left(\mathbf{Z}_{n,l}\right) \\ \pi_l^{\text{new}} &= N_l/N \end{aligned} \tag{15}$$

Evaluation of the reconstructed tree

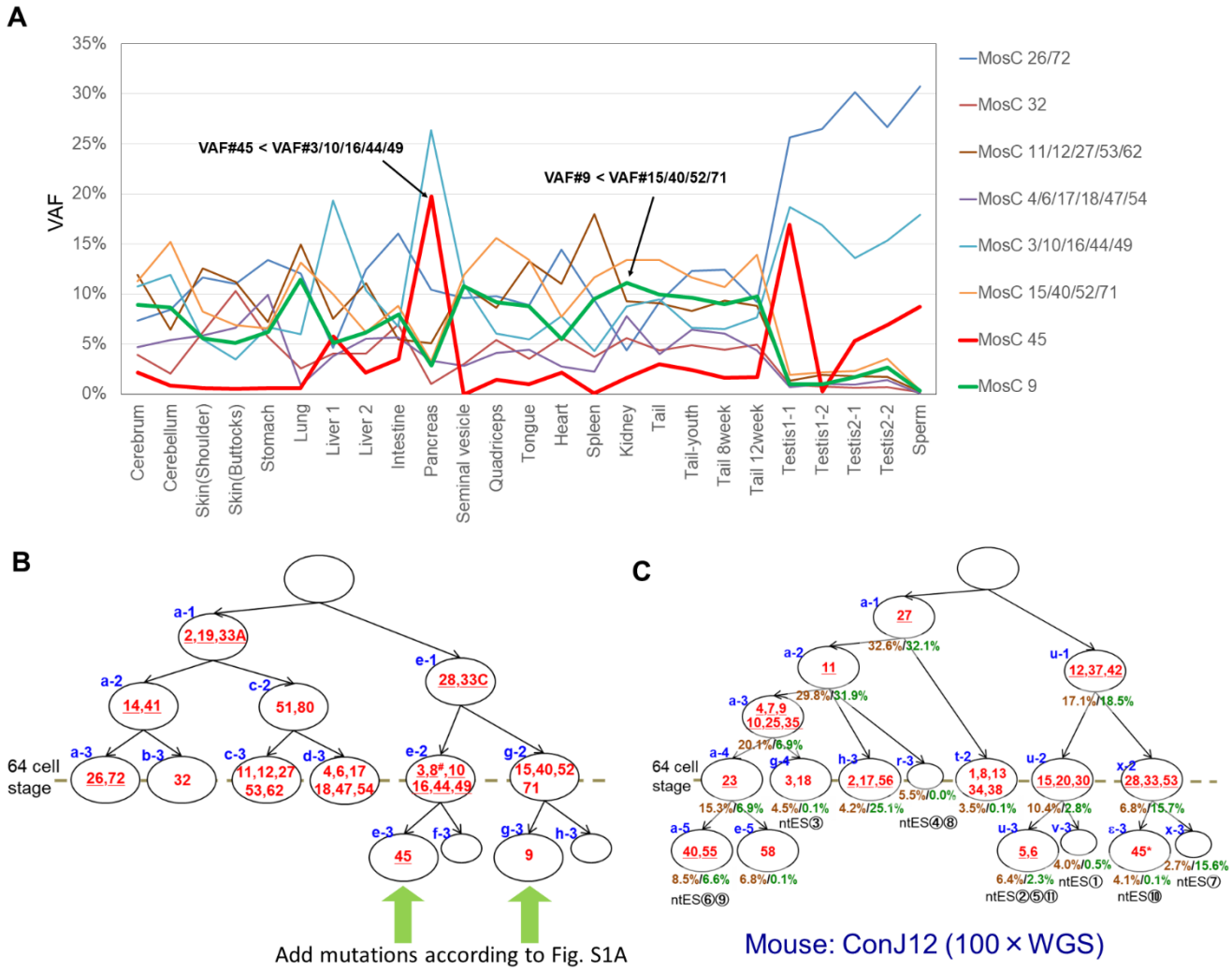
When the reconstructed tree contains the wrong lineage, there will be some inconsistent genotypes among the offspring, such as when an offspring has two mosaic mutations in different lineage paths of the reconstructed tree. In such cases, the probability of the genotype becomes 0.0. We calculated the probabilities of the genotype of offspring data for reconstructed trees described in the main text and confirmed that none of probabilities of genotypes for offspring were 0.

Inference of unassigned mutation lineage

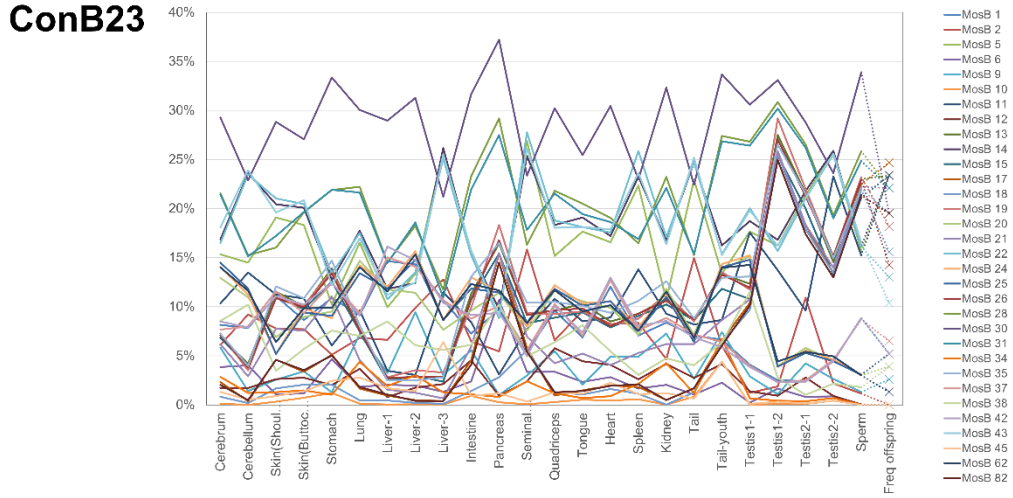
In our probabilistic model, we can derive the expected value that offspring n is derived from lineage l with $\gamma(\mathbf{Z}_{nl})$. Therefore, we regarded that offspring n as being uniquely assigned to lineage l if $\gamma(\mathbf{Z}_{nl}) > 0.99$ was satisfied. If the offspring n has unassigned mutation m', the mutation occurred downstream of the lineage l. Because some of the unassigned mutations may be false-positive cases, we used the following strategy.

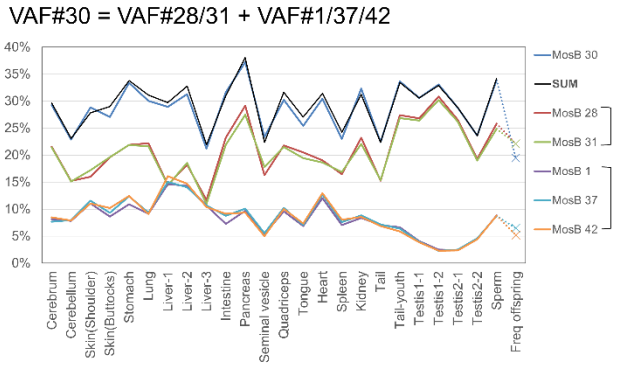
1. We run the following procedure for all offspring n.
    a. If an offspring n shows $\gamma(\mathbf{Z}_{nl}) > 0.99$ for a lineage l and the offspring has an unassigned mutation m', we set $\mathbf{F}_{lm'} = 1$.
2. When an unassigned mutation m' has $\mathbf{F}_{lm'} = 1$ only for lineage l, we estimate that mutation m' occurred downstream of lineage.

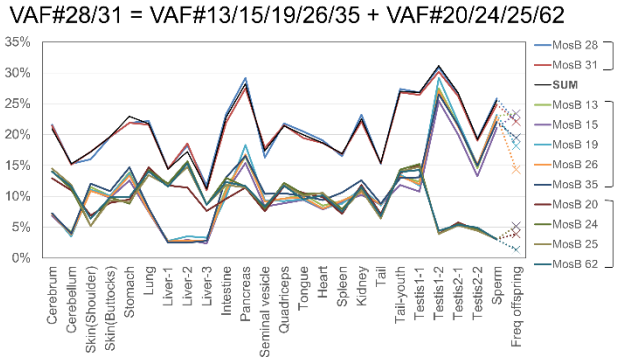
Supplemental Figures

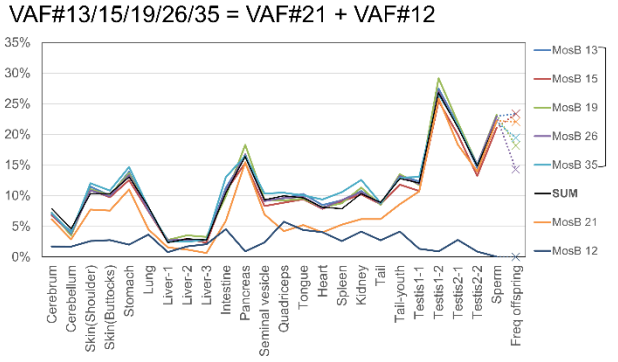


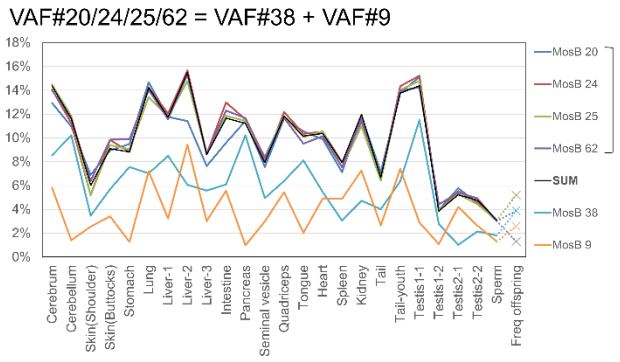
Supplemental Fig. S1. Reconstruction of lineage trees, especially based on the relationship of variant allele frequency (VAF) magnitude. (*A*) Here, we explain the procedure when not finding the summation relationship between the VAFs of mutations. For example, in mouse ConC31, mosaic mutation #45 did not have any counterpart mutations for the summation relationship. In this case, we compared VAF#45 with the VAFs of the mutations belonging to the tips of the lineage tree branch and placed the mutation on a sub-branch of the mutation group #3/10/16/44/49, because only VAF#3/10/16/44/49 were higher than VAF#45 in all tissue samples. Similarly, mosaic mutation #9 was placed on a sub-branch of the mutation group #15/40/52/71. As a counterpart to the mutation that connected the sub-branch, the cell position without the mutation was placed. For the tentative VAF of this cell position, the difference in the VAFs between the mother cell and one daughter cell with a unique mutation was used. In this manner, by sequentially determining the cell positions of each lineage-unassigned mutation, a trichotomy (three-way split, an example is the first cleavage in ConD31) often occurred. (*B*) A schematic of mutation occurrence history reconstructed based on the result of (*A*). (*C*) Reconstructed lineage trees of the ConJ12 mouse based on mosaic mutations originally detected from only the whole-genome sequencing (WGS) with 100× coverage. This tree shows results from the same individual shown in Fig. 2E. (*D*) (shown in the next pages). The tissue-specific VAFs of mosaic mutations (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown. The mutations with asterisk (*) are excluded from the calculation of the VAF at each cell position due to inaccurate measurement.

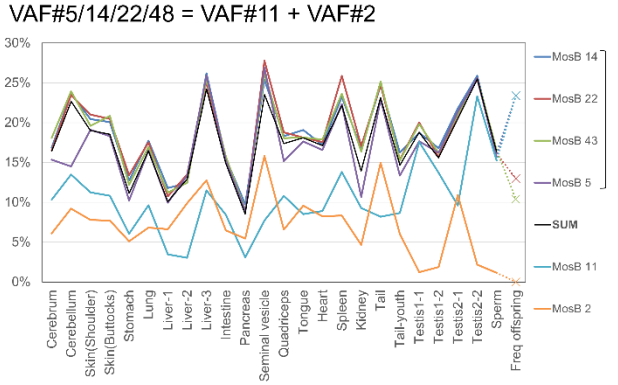




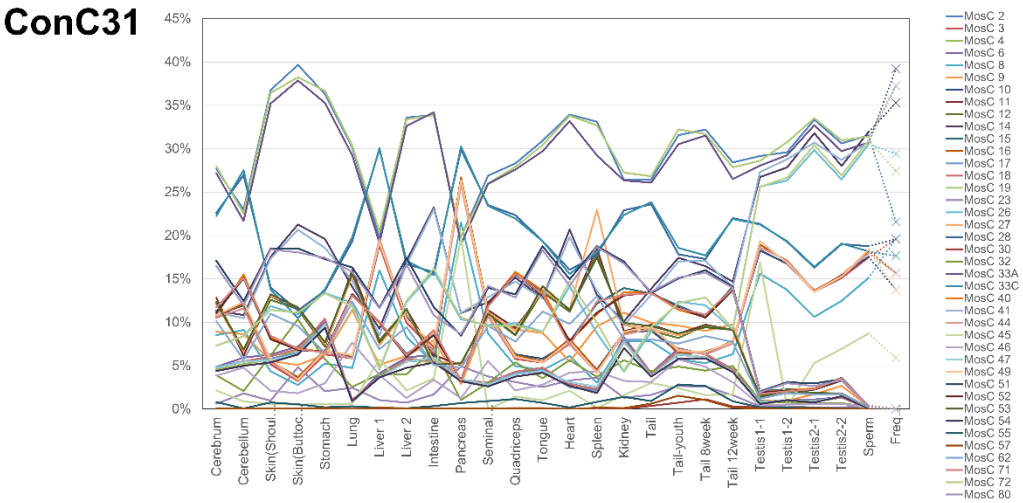


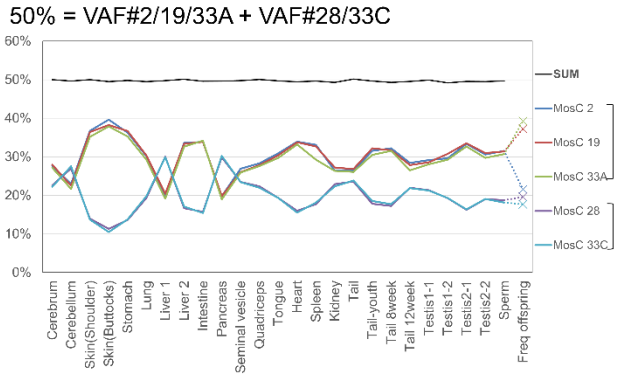


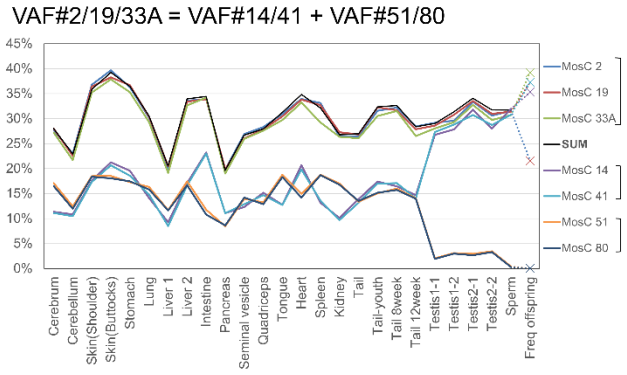


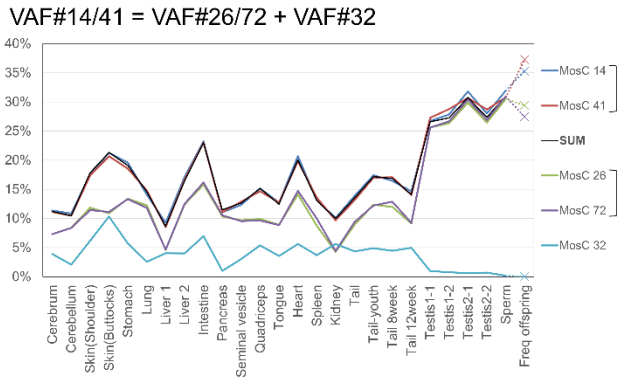


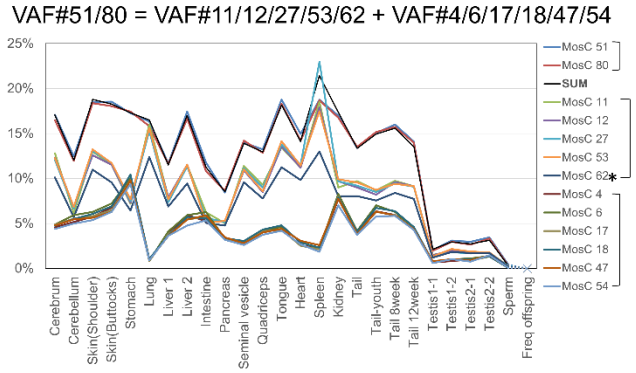
Supplemental Fig. S1D (1/6). The tissue-specific VAFs of mosaic mutations in mouse ConB23 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.

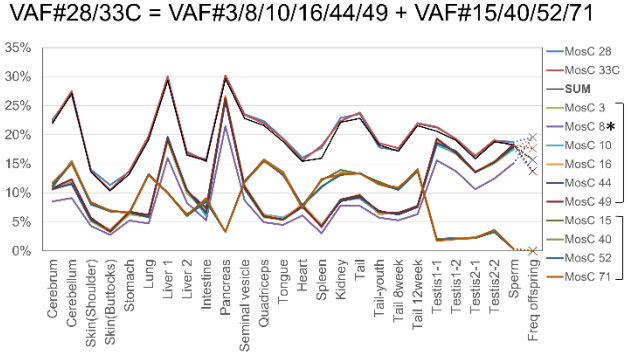




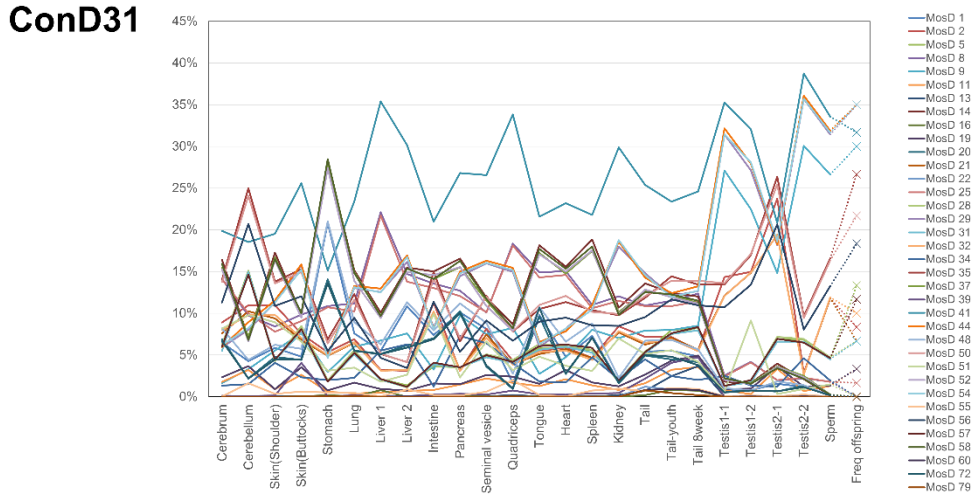


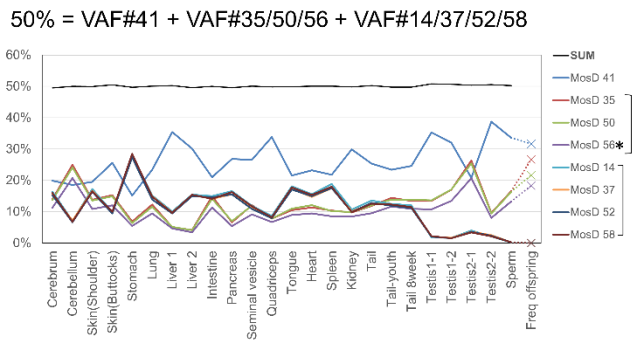


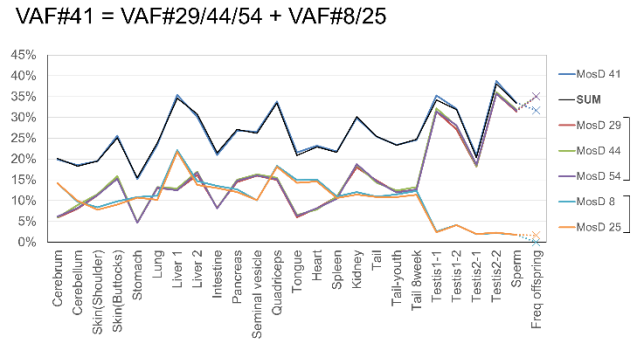


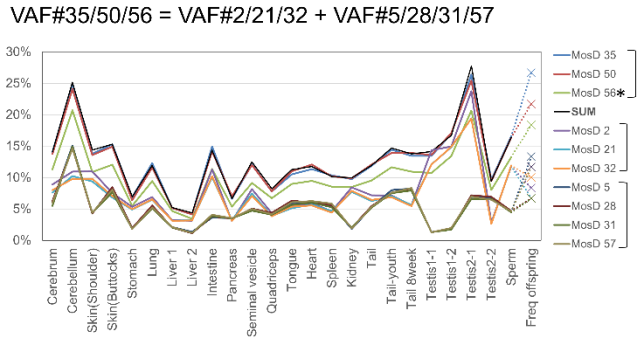


Supplemental Fig. S1D (2/6). The tissue-specific VAFs of mosaic mutations in mouse ConC31 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.

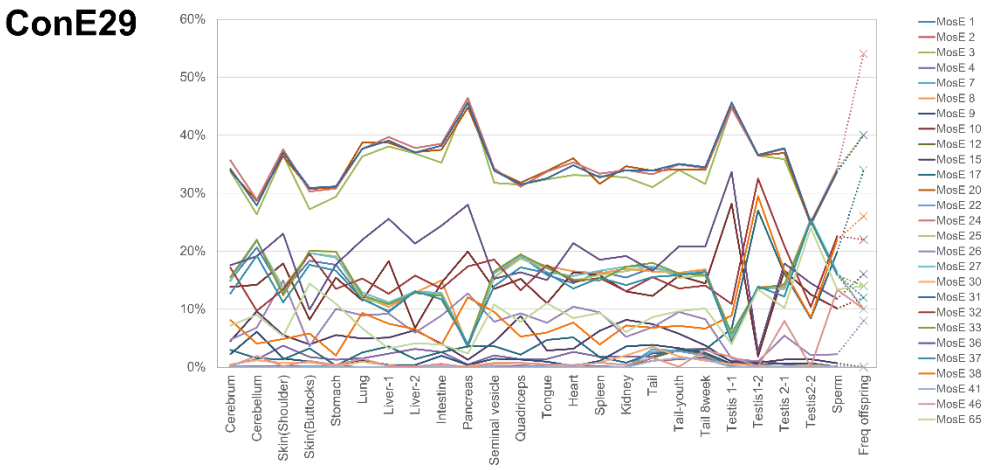




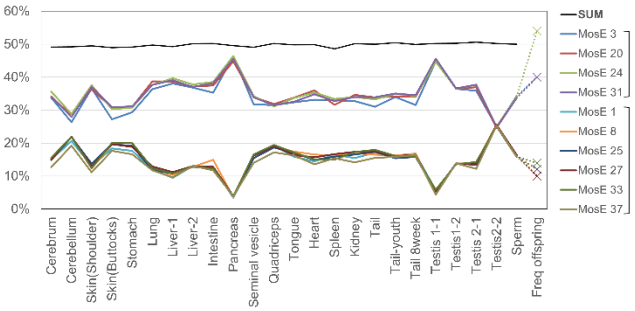




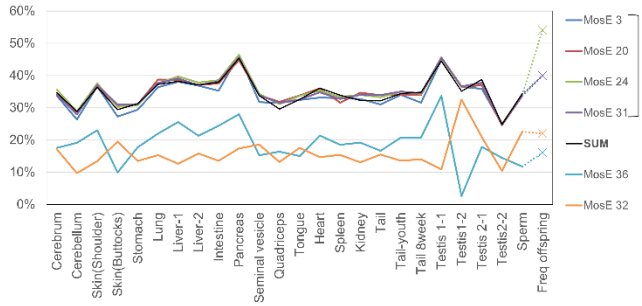
Supplemental Fig. S1D (3/6). The tissue-specific VAFs of mosaic mutations in mouse ConD31 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.



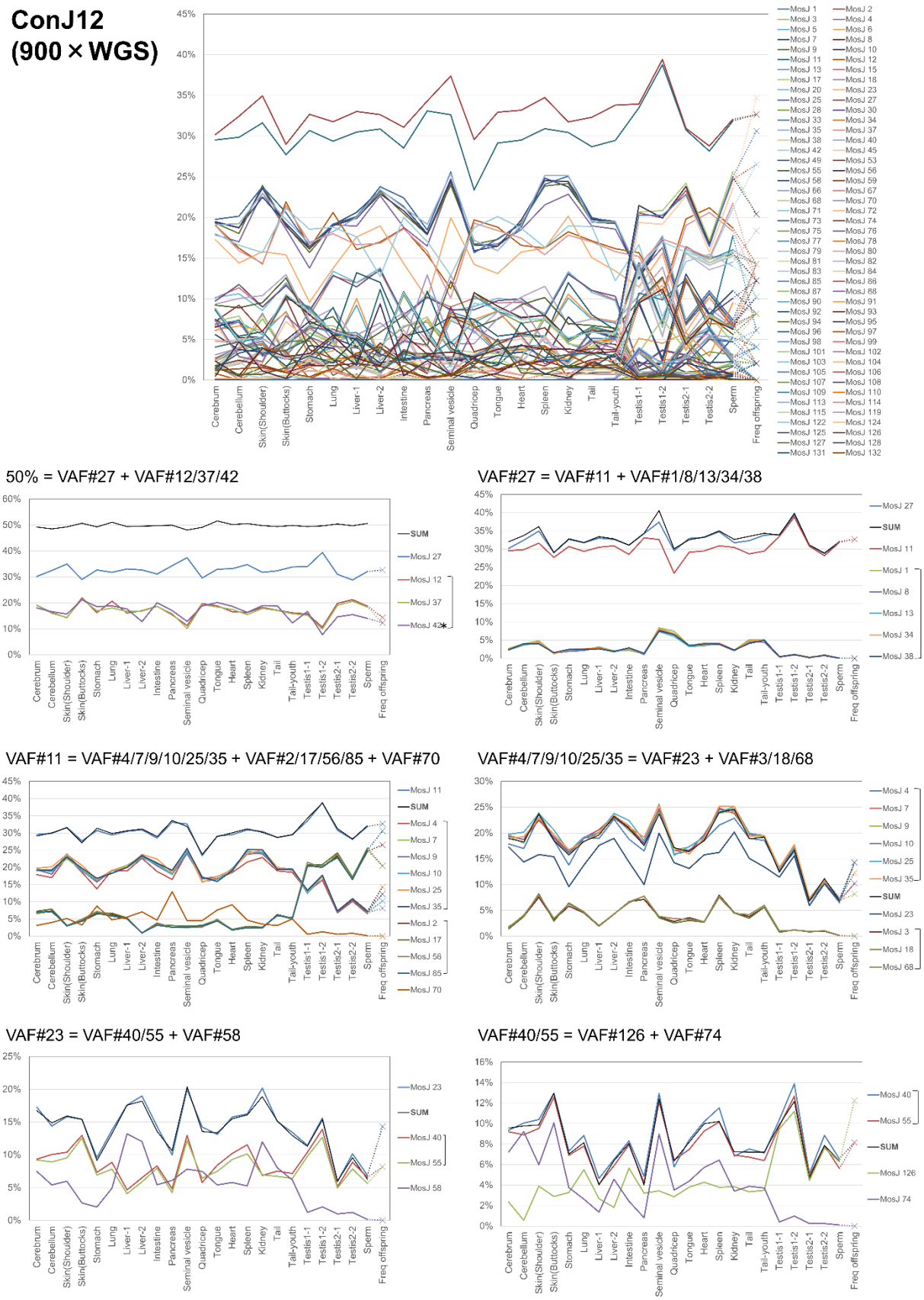
50% = VAF#3/20/24/31 + VAF#1/8/25/27/33/37



VAF#3/20/24/31 = VAF#36 + VAF#32

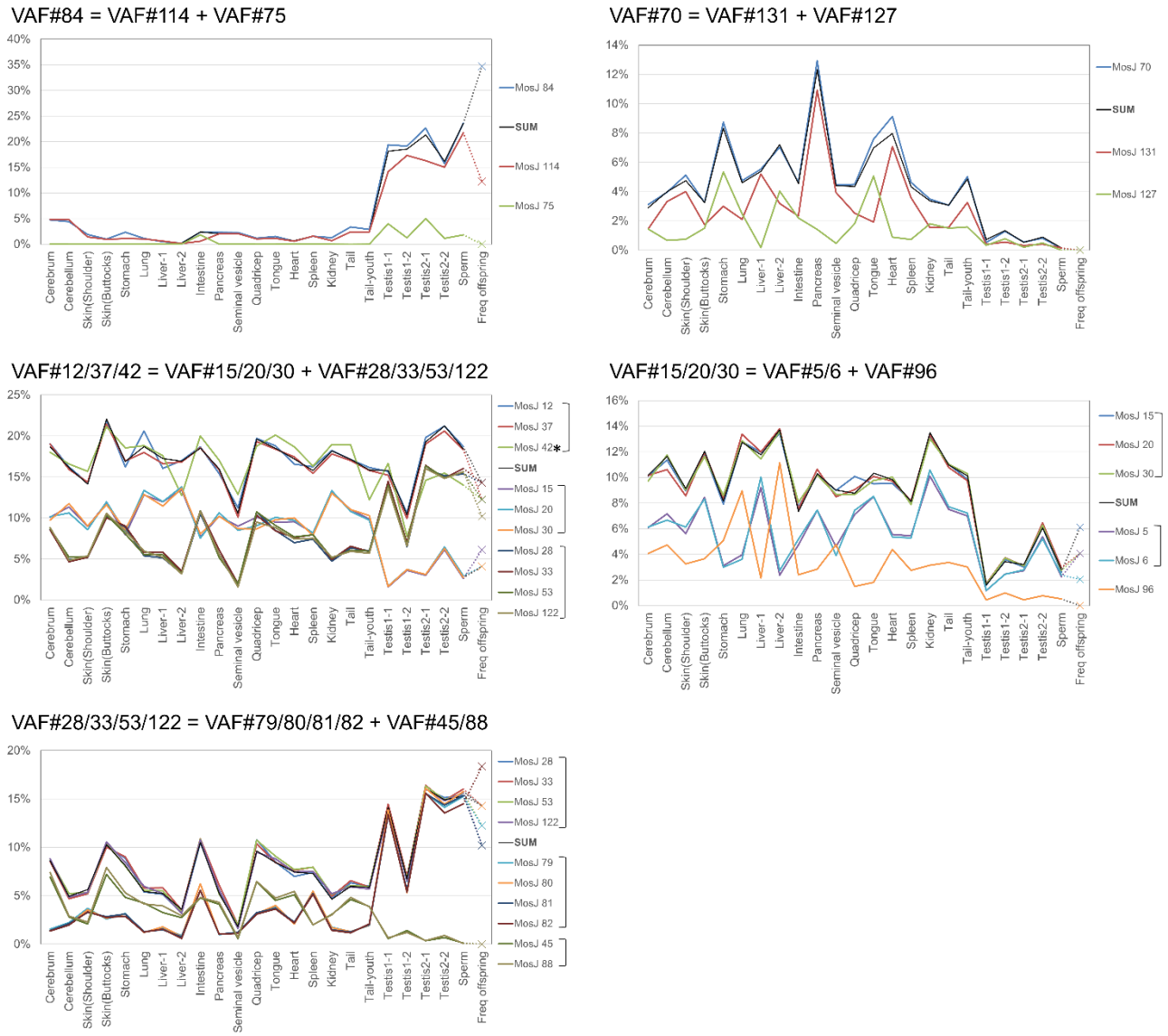


Supplemental Fig. S1D (4/6). The tissue-specific VAFs of mosaic mutations in mouse ConE29 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.

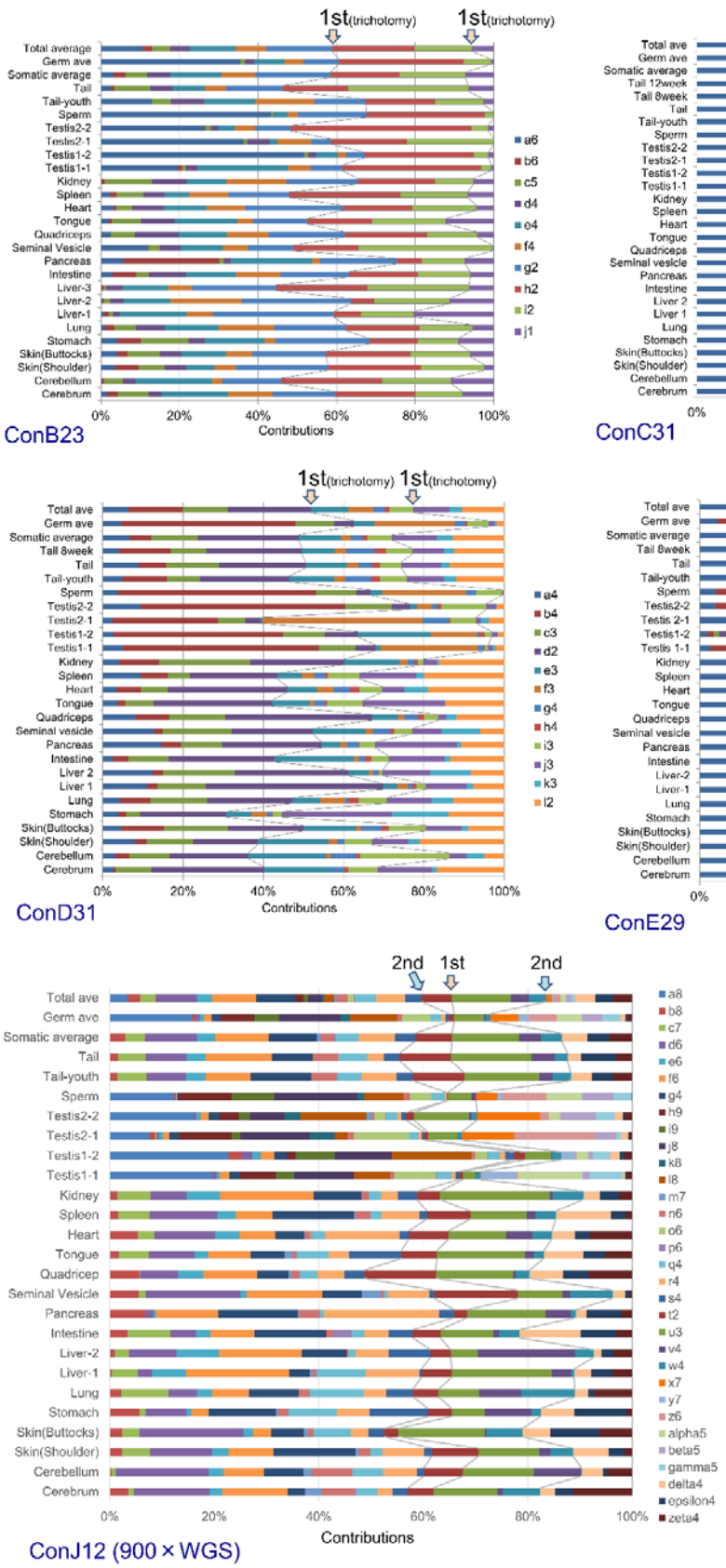


Supplemental Fig. S1D (5/6). The tissue-specific VAFs of mosaic mutations in mouse ConJ12 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.

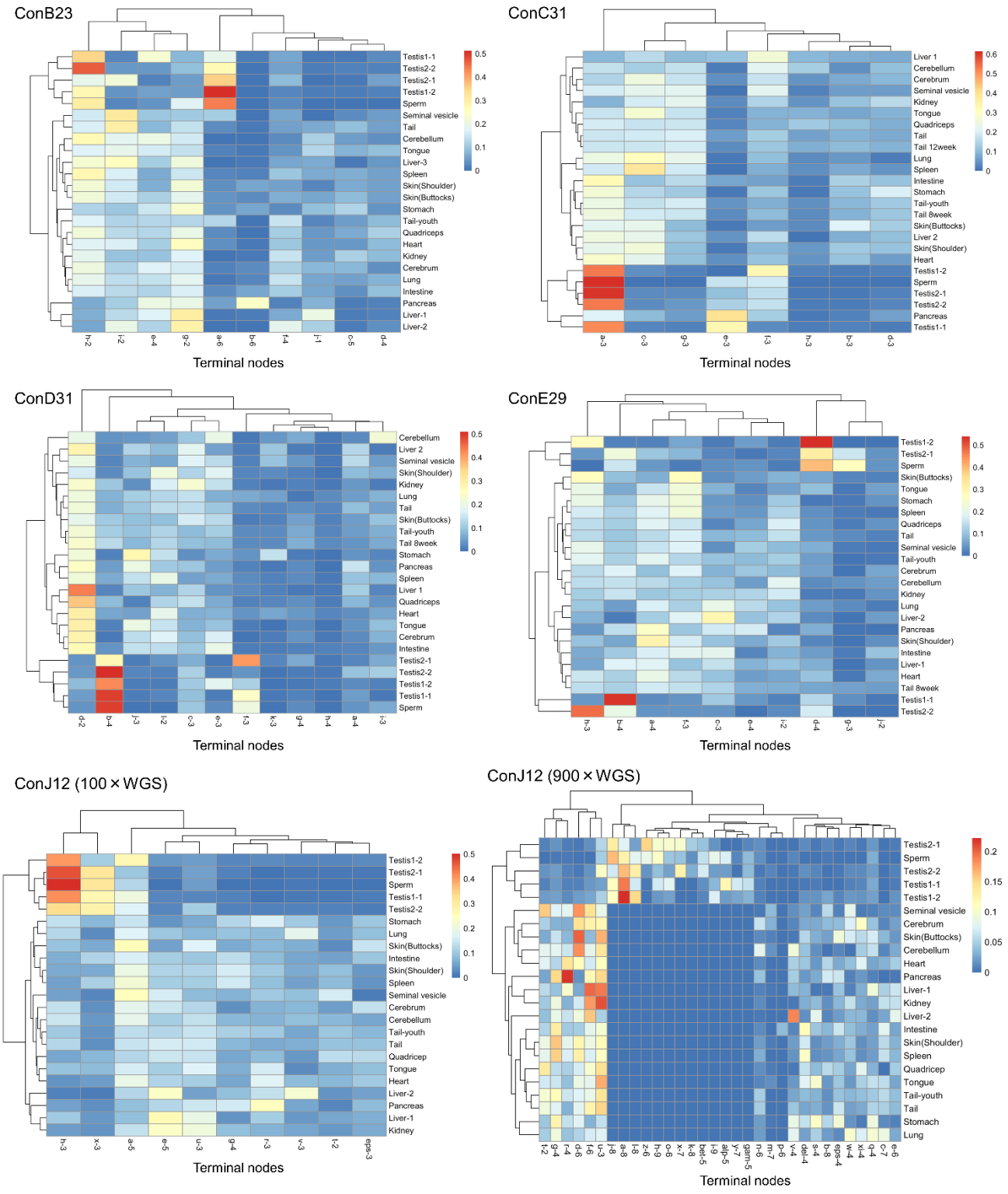
ConJ12 (900 × WGS)



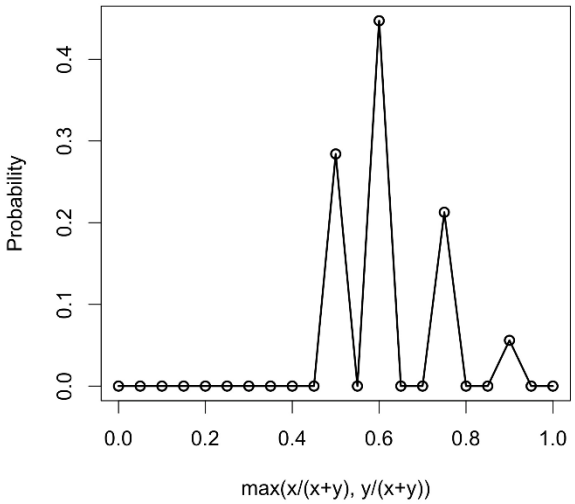
Supplemental Fig. S1D (6/6). The tissue-specific VAFs of mosaic mutations in mouse ConJ12 (including the heritability results for the offspring not involved in genealogy reconstruction), and all summation relationships of mosaic mutations are shown.



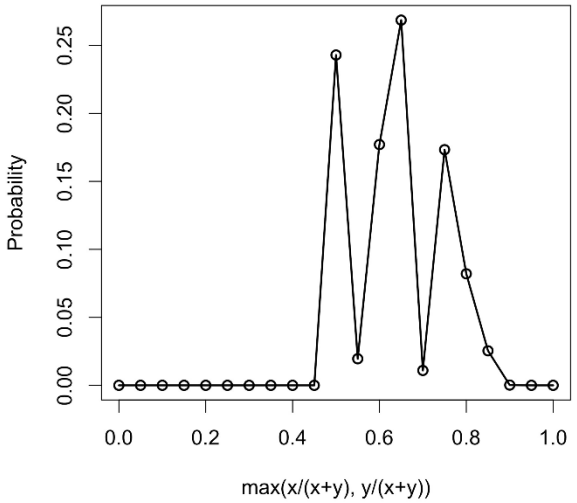
Supplemental Fig. S2. Contribution ratios of each cell lineage into tissues. Each panel corresponds to its respective phylogenetic tree shown in Fig. 2. Arrow with '1st' indicates the position of the first cleavage after the egg is fertilized. Arrow with '2nd' indicates the position of the secondary cleavages after the egg is fertilized. In cases of trichotomy (three-way split) at the first cleavage, two arrows with '1st' are displayed.



Supplemental Fig. S3. Heatmap of contributions of each lineage to adult tissues. The contribution of the terminal nodes in the reconstructed trees (Fig. 2 and Supplemental Fig. S1C) for each tissue is visualized as the heatmap with the pheatmap package in R. The rows and columns were hierarchically clustered based on the Euclidean distance and hclust function with the complete linkage method. For accurate analysis, the heatmap of ConJ12 (100× whole-genome sequencing [WGS] data) was made using the values from the same positions of the tree from 900× WGS data.



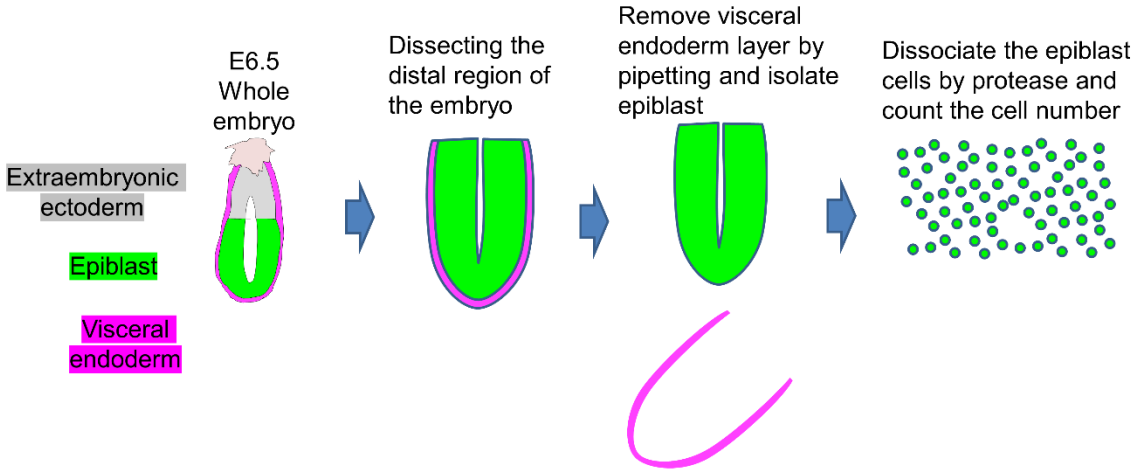
k=1, Expected value of the contribution ratio
= 0.648 (*n*=7)
0.630 (*n*=8)
0.631 (*n*=9)
0.618 (*n*=10)



k=2, Expected value of the contribution ratio
= 0.637 (*n*=7)
0.643 (*n*=8)
0.645 (*n*=9)
0.646 (*n*=10)

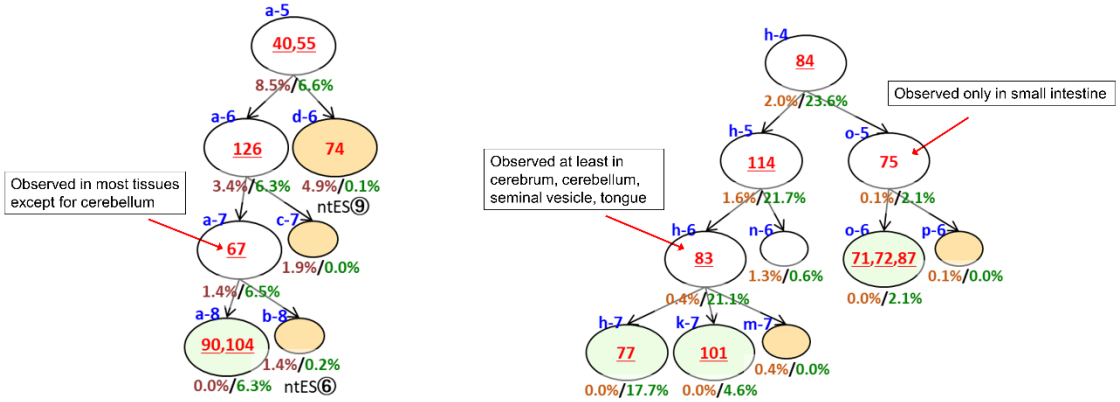
Average (*k*=1,2): 0.637

Supplemental Fig. S4. Simulation analysis of contribution ratios between two daughter cells in a postzygotic cell division. It is known that approximately eight cells contribute to the whole body as epiblasts in the 128-cell stage of the early embryo. Thus, we assumed the process that *n* branches (*n* = 7–10) are stochastically selected as epiblast cells, each with equal contribution, from 128 total branches of the tree. The detailed procedure is shown in the Methods section. Here, each left figure represents the result of the *k*th branch from the fertilized egg in the case of *n* = 8.



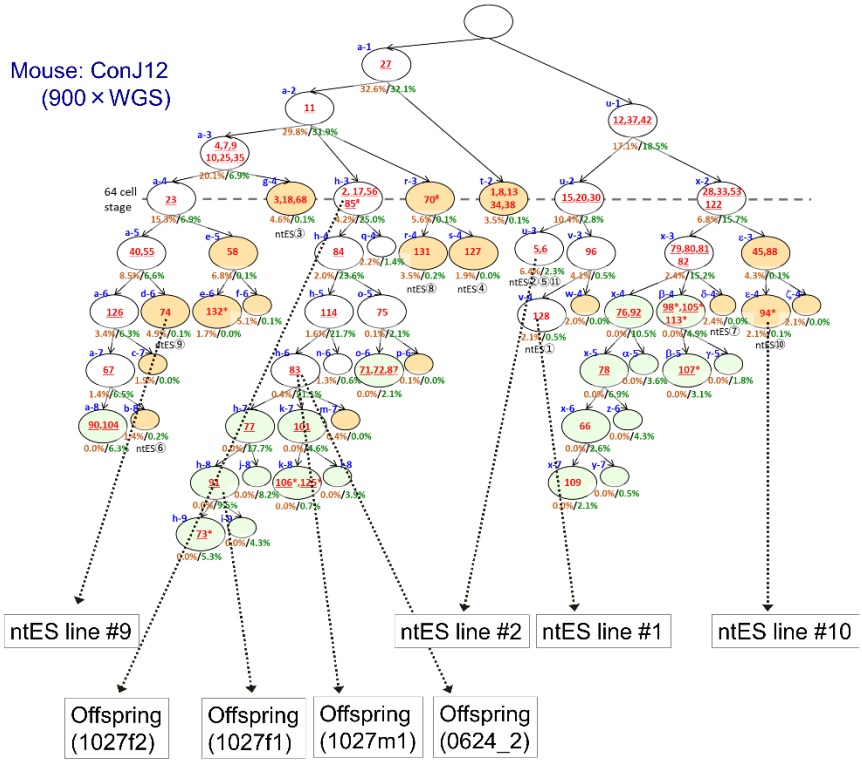
Embryo	Cell number
#1	440
#2	328
#3	597
#4	748
#5	340
#6	633
Mean (±SD)	514 (±171)

Supplemental Fig. S5. Epiblast cell count at embryonic day 6.5. Embryos were harvested and cut at the border of the extraembryonic ectoderm and epiblast with fine forceps. The distal region of the embryos, including the epiblasts, was pipetted several times with glass capillaries in PBS to remove the visceral endoderm layer. Epiblasts were cultured in protease and dissociated into single cells. The collected epiblast cells were manually counted using a stereomicroscope.



Cell Position		Cerebrum	Cerebellum	Skin(Shoul	Skin(Butto	Stomach	Lung	Liver−1	Liver−2	Intestine	Pancreas	Seminal ve	Quadricep	Tongue	Heart	Spleen	Kidney	Tail	Tail−youth	Testis1−1	Testis1−2	Testis2−1	Testis2−2	Sperm	No offspring	No ntES
a5	#40	9.3%	10.0%	10.4%	13.0%	7.3%	8.8%	4.6%	6.5%	8.3%	4.8%	12.9%	5.8%	8.4%	10.2%	11.5%	6.8%	7.5%	7.1%	10.4%	13.9%	5.1%	8.9%	6.6%	4	2
a6	#126	2.3%	0.6%	3.9%	2.5%	3.3%	5.5%	2.7%	1.8%	5.6%	3.2%	3.4%	2.9%	3.8%	4.3%	3.8%	3.8%	3.3%	3.5%	9.5%	11.2%	4.4%	7.6%	6.3%	6	1
a7	#67	1.8%	0.1%	1.1%	1.2%	2.7%	1.1%	0.2%	0.5%	1.7%	3.4%	2.8%	2.8%	0.9%	2.7%	0.8%	0.7%	0.8%	0.7%	9.5%	12.4%	4.3%	8.0%	6.5%	4	1
a8	#90/#104	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	9.5%	11.1%	3.8%	8.1%	6.3%	4.5	0
b8	a7−a8	1.8%	0.2%	1.1%	1.2%	2.7%	1.1%	0.2%	0.5%	1.7%	3.4%	2.8%	2.8%	0.9%	2.7%	0.8%	0.7%	0.8%	0.7%	−0.1%	1.3%	0.5%	−0.1%	0.2%	nd	1
h4	#84	4.8%	4.4%	1.9%	1.0%	2.3%	1.2%	0.5%	0.2%	2.3%	2.3%	2.3%	1.2%	1.5%	0.6%	1.5%	1.3%	3.4%	2.9%	19.4%	19.1%	22.6%	15.7%	23.6%	17	0
h5	#114(modified)	4.8%	4.8%	1.4%	1.0%	1.1%	1.0%	0.6%	0.2%	0.6%	2.0%	2.1%	1.0%	1.1%	0.6%	1.6%	0.7%	2.4%	2.4%	14.1%	17.3%	16.3%	15.0%	21.7%	6	0
h6	#83	1.6%	0.8%	0.3%	0.4%	0.2%	0.0%	0.2%	0.0%	0.0%	0.0%	1.8%	0.3%	0.7%	0.0%	0.1%	0.0%	0.0%	0.0%	13.8%	17.0%	15.7%	14.8%	21.1%	13	0
h7	#77	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	10.7%	9.9%	12.4%	6.3%	17.7%	3	0
k7	#101	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.4%	7.5%	3.6%	8.0%	4.6%	4	0
m7	h6−h7−k7	1.6%	0.8%	0.3%	0.4%	0.2%	0.0%	0.2%	0.0%	0.0%	0.0%	1.8%	0.2%	0.7%	0.0%	0.1%	0.0%	0.0%	0.0%	−0.4%	−0.4%	−0.3%	0.5%	−1.2%	nd	0
h4	#84	4.8%	4.4%	1.9%	1.0%	2.3%	1.2%	0.5%	0.2%	2.3%	2.3%	2.3%	1.2%	1.5%	0.6%	1.5%	1.3%	3.4%	2.9%	19.4%	19.1%	22.6%	15.7%	23.6%	17	0
o5	#75	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	4.0%	1.2%	5.0%	1.1%	1.8%	0	0
o6	#71/#72/#87	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.9%	1.2%	5.2%	1.1%	2.1%	1	0
p6	o5−o6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	−0.2%	0.0%	−0.3%	nd	0

Supplemental Fig. S6. Development of the germ cell-specific lineages in mouse ConJ12. Some of the lineages shown in Fig. 2E and the allele frequencies of the mutations present at each cell position in the tissues are shown. The cell positions indicated by red arrows show the positions immediately before arising cell positions found only in the testis and spermatozoa. The brown and green numbers (%) below each cell position represent the average variant allele frequency (VAF) of the mutations in somatic tissues and the VAF in a sperm sample, respectively. In sperm or all somatic tissue samples, cells with the VAF values comparable to the background were colored orange and light green as somatic cell-specific lineages and germ cell-specific lineages, respectively.

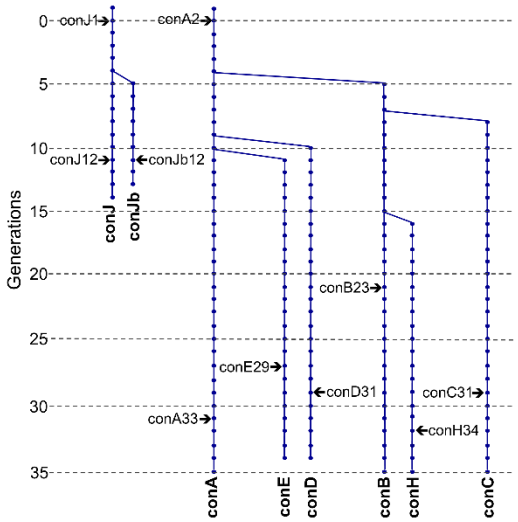


	Cell position	Total	Postzygotic mutations	Mutations other than postzygotic ones
ntES line #1	v-4	773 (SNV:757, indel:10, multisite:6)	8 (SNV:6, indel:2)	765 (SNV:751, indel:8, multisite:6)
ntES line #2	u-3	517 (SNV:498, indel:13, multisite:6)	8 (SNV:6, indel:2)	509 (SNV:492, indel:11, multisite:6)
ntES line #9	d-6	886 (SNV:860, indel:20, multisite:6)	12 (SNV:11, indel:1)	874 (SNV:849, indel:19, multisite:6)
ntES line #10	ε-4	529 (SNV:508, indel:18, multisite:3)	10 (SNV:7, indel:3)	519 (SNV:501, indel:15, multisite:3)
Mean		676.3 (SNV:655.8, indel:15.3, multisite:5.3)	9.5 (SNV:7.5, indel:2.0)	666.8 (SNV:648.3, indel:13.3, multisite:5.3)

	Cell position	Total	Paternal postzygotic mutations	Mutations other than the postzygotic ones
Offspring 1027m1	Downstream of h-6	26 (SNV:23, indel:3)	3 (SNV:3)	23 (SNV:20, indel:3)
Offspring 1027f1	Downstream of h-8	30 (SNV:29, indel:1)	9 (SNV:9)	21 (SNV:20, indel:1)
Offspring 1027f2	Downstream of h-3	28 (SNV:27, multisite:1)	5 (SNV:5)	23 (SNV:22, multisite:1)
Offspring 0624_2	Downstream of h-6	27 (SNV:23, indel:2, multisite:2)	3 (SNV:3)	24 (SNV:20, indel:2, multisite:2)
Mean		27.5 (SNV:25.3, indel:1.5, multisite:0.8)		

Supplemental Fig. S7. Number of accumulated *de novo* mutations in nuclear transfer embryonic stem (ntES) cell lines and offspring. We conducted whole-genome sequencing (WGS) of four ntES cell lines and four offspring (and their mother) for mouse ConJ12 to identify *de novo* mutations. Here, the mutations described in the lineage tree of ConJ12 (900× WGS data, Fig. 2E) were designated as postzygotic mutations. Except for postzygotic mutations, the number of mutations within the effective whole-genome coverage (EWC) region is shown for the ntES cell analysis, and the number of autosomal mutations including mutations outside the EWC region is shown for the offspring analysis. When comparing the mutations in the ntES cell lines to postzygotic mutations, an estimated number of mutations outside the EWC region (76.7 mutations, referring to Supplemental Table S4) was added to the value of 666.8. A list of the mutations is shown in Supplemental Data 2.

A



C

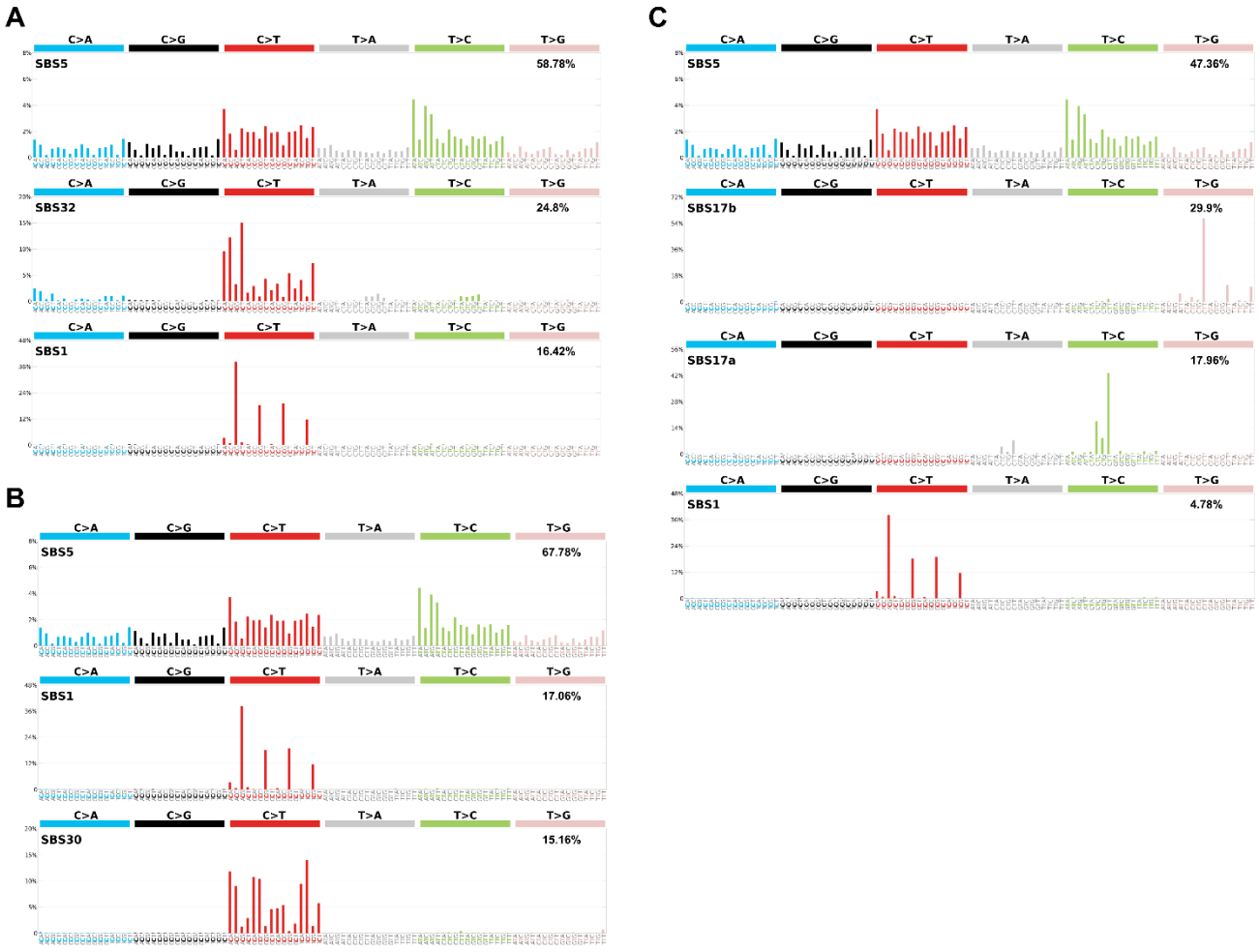
	Mutations	Cell divisions	Mutations per cell division
Whole germline (per generation)	26.3	33	0.80
Zygote ~ E6.5	12.0	12	1.0
E6.5 ~ gamete	14.3	21	0.68

B

			SNVs									Indels + Multisites						
	No. generations of breeding	No. accumulated mutations	No. accumulated mutations				No. mutations per generation			Rate (per generation per nucleotide, $\times 10^{-9}$)		No. accumulated mutations				No. mutations per generation		
			Total	Autosome		Chr. X	Total	Auto.	Chr. X			Total	Autosome		Chr. X	Total	Auto.	Chr. X
				Homo	Hetero		(Auto + 1.5 × ChrX)	(/diploid)	(/haploid)	Auto.	Chr. X		Homo	Hetero		(Auto + 1.5 × chrX)	(/diploid)	(/haploid)
conA33	31	428	405	256	134	15	21.6	21.0	0.4	5.36	4.3	23	16 (7)*	7 (2)	0	1.2	1.2	0.0
conB23	21	307	281	128	140	13	20.6	19.9	0.5	5.07	5.6	26	14 (3)	12 (2)	0	1.9	1.9	0.0
conC31	29	423	385	246	120	19	21.7	20.9	0.5	5.32	5.9	38	28 (7)	10	0	2.2	2.2	0.0
conD31	29	409	379	244	115	20	21.3	20.5	0.6	5.22	6.2	30	17 (1)	11 (2)	2 (2)	1.7	1.6	0.06
conE29	27	406	380	207	153	20	22.7	21.8	0.6	5.55	6.7	26	16 (2)	10 (2)	0	1.6	1.6	0.0
conH34	32	474	434	260	161	13	22.6	22.1	0.3	5.64	3.7	40	24 (7)	15 (2)	1	2.1	2.0	0.03
conJ12	11	212	199	46	144	9	24.0	23.2	0.5	5.91	7.4	13	6	7 (1)	0	1.6	1.6	0.0
conJb12	11	182	164	60	99	5	19.8	19.4	0.3	4.94	4.1	18	5	12 (2)	1 (1)	2.2	2.1	0.06
Total		2,841	2,627	1,447	1,066	114	21.8	21.1	0.5	5.38	5.4	214	126 (27)	84 (13)	4 (3)	1.8	1.8	0.02
Unique		2,456	2,270	2,168		102						186	182 (37)		4 (3)			

* The parentheses indicate the number of multisites.

Supplemental Fig. S8. Analysis of *de novo* germline mutations using long-term breeding mouse lines. We have developed mutation accumulation (MA) mouse lines (wild-type, C57BL/6J) for long-term analysis of the associations between phenotypes and genotypes beyond generations; some of them were reported in our previous paper (Uchimura A. *et al*, 2015). (*A*) Pedigrees of the MA lines. All breeding lines subjected to whole-genome sequencing (WGS) in this study are shown. Red arrows indicate WGS performed on mice from the indicated generation. Accumulated *de novo* germline mutations were identified by comparing WGS results of mice belonging to each breeding line with the results of their ancestral male and female mouse pairs. (*B*) Number of mutations accumulated in effective whole-genome coverage (EWC) regions of each MA line and the mutation rates estimated from them is shown. The rates were estimated according to a previous paper (Uchimura A. *et al*, 2015). (*C*) Estimation of per-cell division mutation rates from (*B*). For an accurate comparison with postzygotic mutations, we added the estimated number of mutations outside the EWC region (2.7 mutations, referring to Supplemental Table S4) to 23.6 (= 21.8 + 1.8), shown in (*B*).



Supplemental Fig. S9. Mutational spectra of postzygotic mutations, germline mutations, and somatic (tail fibroblast) mutations. Mutational spectra can be decomposed into known mutational signatures of single-base substitutions (SBSs) observed in the mutation catalog of humans (Cosmic mutational signature v3.2, Alexandrov *et al.* 2020). (*A*) Postzygotic mutations (validated mosaic mutations detected in deep-coverage whole-genome sequencing on five mice), (*B*) germline mutations (*de novo* mutations accumulated in long-term breeding mouse lines), and (*C*) somatic mutations (*de novo* mutations accumulated in nuclear transfer embryonic stem [ntES] cell lines derived from mouse tail fibroblasts) are shown.

Supplemental Tables

Supplemental Table S1. Estimation of mutation rates during early development

We estimated the rate by focusing on the mosaic mutations with variant allele frequencies (VAFs) ≥6% in tail samples, because such mutations were expected to have a low missing variant rate in 100× whole-genome sequencing (WGS) data. In cases of trichotomy (three-way split), we added one tentative daughter cell position with no unique *de novo* mutations for calculation. This means that the number of each branch of trichotomy is regarded as 1.33. Considering the missing variant rate estimated from the comparison of 100× WGS data and 900× WGS data of mouse ConJ12 into consideration, we estimated that 2.18 mutations occur per branch. To estimate the per-nucleotide mutation rate, we used only mutations occurring in the effective whole-genome coverage (EWC) regions, which comprise two autosomes and one Chromosome X. The per-nucleotide mutation rate (μ) was calculated as $\mu = m/G$, where m is the total number of mutations and G is the size of the analyzed genome in base pairs.

	Mutations (VAF: >6%)	Branch number	Mutations (/branch)	Mutations (/mitosis)	Substitutions in EWC	Substitutions (/branch in EWC)	Substitutions (×10-10/bp branch)
ConB23	27	13.67	1.98	0.9	24	1.76	4.35
ConC31	33	12	2.75	1.25	30	2.5	6.2
ConD31	18	10	1.8	0.82	18	1.8	4.46
ConE29	16	12	1.33	0.6	15	1.25	3.1
ConJ12 (100×)	26	11.67	2.23	1.01	19	1.63	4.01
Total (using ConJ 100×)	120	59.33	2.02	0.92	106	1.79	4.42
ConJ12 (900×)	28	11.67	2.4	1.09	21	1.8	4.43
Total (with correction for missed variants)	129.2	59.33	2.18	0.99	117.2	1.98	4.89

Supplemental Table S2. Information for mice analyzed for the reconstruction of cell lineage trees. Some of the offspring were produced through in vitro fertilization (IVF). Whole-genome sequencing (WGS) coverage represent the values of raw read data (for 100× WGS) and of highly reliable (HR) read data (after removal of duplicate reads for 900× WGS).

	ConB23 (1122m1)	ConC31 (1012m3)	ConD31 (0828m2)	ConE29 (0821m1)	ConJ12 (0718m1)
Age (days) at sampling	587	635	683	690	356
# Tissue types	16	16	16	16	16
# Offspring	77 (IVF: 56)	51 (IVF: 34)	60 (IVF: 29)	50 (IVF: 17)	49 (IVF not performed)
# ntES cell lines	4	-	-	-	11
Coverage (100× WGS)	102.0	103.2	101.8	101.0	106.2
Coverage (900× WGS)	-	-	-	-	454.1 (tail) 456.0 (testis)

Supplemental Table S3. Summary of whole-genome sequencing (WGS)

WGS read coverage representing the values of raw read data except for *900× WGS data (HR [highly reliable] read data after removing duplicate reads for 900× WGS). WGS was conducted at the National Institute of Genetics (the top 14 samples: from 'ConB23' to 'Ancestor [ConJ1] female') and at Macrogen Japan (the bottom nine samples: from 'ntES cell line#1' to 'Spouse #0718f1').

Sample	Tissue	DNA extract	Platform	Library preparation	Read length	Read coverage	
ConB23	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	102.0	For mosaic mutations and accumulated germline mutations
ConC31	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	103.2	For mosaic mutations and accumulated germline mutations
ConD31	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	101.8	For mosaic mutations and accumulated germline mutations
ConE29	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	101.0	For mosaic mutations and accumulated germline mutations
ConJ12 (100×)	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	106.2	For mosaic mutations and accumulated germline mutations
ConJ12 (450×, tail)	Tail	Smart DNA	NovaSeq	PCR free	PE, 150 bp	454.1*	For mosaic mutations
ConJ12 (450×, testis)	Testis	Smart DNA	NovaSeq	PCR free	PE, 150 bp	456.0*	For mosaic mutations
ConA33	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	100.5	For accumulated germline mutations
ConH33	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	103.3	For accumulated germline mutations
ConJb12	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	107.2	For accumulated germline mutations
Ancestor (ConA2) male	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	65.0	For reference of ConA to ConH mice
Ancestor (ConA2) female	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	65.3	For reference of ConA to ConH mice
Ancestor (ConJ1) male	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	68.7	For reference of ConJ mice
Ancestor (ConJ1) female	Tail	DNeasy	HiSeq	PCR free	PE, 250 bp	65.6	For reference of ConJ mice
ntES cell line #1	cells	Smart DNA	HiSeq	PCR free	PE, 150 bp	49.7	For somatic mutations in fibroblast
ntES cell line #2	cells	Smart DNA	HiSeq	PCR free	PE, 150 bp	53.5	For somatic mutations in fibroblast
ntES cell line #9	cells	Smart DNA	HiSeq	PCR free	PE, 150 bp	53.5	For somatic mutations in fibroblast

ntES cell line #10	cells	Smart DNA	HiSeq	PCR free	PE, 150 bp	54.0	For somatic mutations in fibroblast
Offspring #1027m1	Tail	Smart DNA	NovaSeq	TruSeq Nano	PE, 150 bp	35.9	For germline mutations in offspring
Offspring #1027f1	Tail	Smart DNA	NovaSeq	TruSeq Nano	PE, 150 bp	37.2	For germline mutations in offspring
Offspring #1027f2	Tail	Smart DNA	NovaSeq	TruSeq Nano	PE, 150 bp	43.0	For germline mutations in offspring
Offspring #0624-2	Tail	Smart DNA	NovaSeq	PCR free	PE, 150 bp	33.0	For germline mutations in offspring
Spouse #0718f1	Tail	Smart DNA	NovaSeq	TruSeq Nano	PE, 150 bp	34.8	For analysis of germline mutations

Supplemental Table S4. Breakdown of mosaic mutations assessed in this study

Note: The number in parentheses represents the number of mutations outside the effective whole-genome coverage (EWC) region. We confirmed mosaic mutations by checking the consistency of the variant allele frequency (VAF) values based on a dilution series using negative control samples. The number of mutations in the difference between "Successful PCR amplification" and "Confirmed mosaic mutations" represents the number of false-positive calls, including constitutive (non-mosaic) variants.

		Candidates	Successful PCR amplification	Confirmed as mosaic mutations	Lineage assigned	Lineage unassigned	Lineage assigned only by VAF	Lineage assigned with genotyping
ConB23	SNV	46 (3)	41 (3)	30 (2)	27 (2)	3 (0)	27 (2)	0 (0)
	INDEL etc.	13 (0)	12 (0)	2 (0)	1 (0)	1 (0)	1 (0)	0 (0)
ConC31	SNV	62 (5)	59 (3)	39 (2)	34 (2)	5 (0)	34 (2)	0 (0)
	INDEL etc.	9 (0)	9 (0)	1 (0)	1 (0)	0 (0)	1 (0)	0 (0)
ConD31	SNV	78 (16)	74 (14)	36 (2)	29 (1)	7 (1)	27 (1)	2 (0)
	INDEL etc.	12 (1)	12 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ConE29	SNV	41 (3)	39 (3)	26 (2)	19 (2)	7 (0)	18 (1)	1 (1)
	INDEL etc.	9 (1)	7 (1)	1 (0)	0 (0)	1 (0)	0 (0)	0 (0)
ConJ12 (100×)	SNV	52 (18)	48 (14)	30 (5)	29 (4)	1 (1)	29 (4)	0 (0)
	INDEL etc.	20 (5)	15 (3)	5 (0)	5 (0)	0 (0)	4 (0)	1 (0)
ConJ12 (900×)	SNV	129 (47)	110 (29)	82 (14)	70 (11)	12 (3)	61 (10)	9 (1)
	INDEL etc.	21 (5)	16 (3)	6 (0)	6 (0)	0 (0)	6 (0)	0 (0)
Total (Only 100×WGS)	SNV	279 (45)	261 (37)	161 (13)	138 (11)	23 (2)	135 (10)	3 (1)
	INDEL etc.	63 (7)	55 (5)	9 (0)	7 (0)	2 (0)	6 (0)	1 (0)
Total (Includig 900×WGS)	SNV	356 (74)	323 (52)	213 (22)	179 (18)	34 (4)	167 (16)	12 (2)
	INDEL etc.	64 (7)	56 (5)	10 (0)	8 (0)	2 (0)	8 (0)	0 (0)

Supplemental Table S5. Breakdown of the sample fraction for DNA extraction used in amplicon-seq. The number of cells was calculated under the assumption that the weight of a haploid genomic DNA was 3.3 pg. Except for 'Tail-8week' and 'Tail-youth' samples, they were sampled on the date shown in Supplemental Table S2. 'Tail-8week' was sampled at 8 weeks of age and 'Tail-youth' was sampled at 3–6 weeks of age. The exact amounts of these tail tissue fragments were missed because these samples were used for another purpose. However, 5–10 µg of genomic DNA was extracted from each fragment.

	ConB23		ConC31		ConD31		ConE29		ConJ12	
	DNA µg	Cells (× 10^4)	DNA µg	Cells (×10^4)	DNA µg	Cells (×10^4)	DNA µg	Cells (×10^4)	DNA µg	Cells (×10^4)
Cortex	5.6	84.8	7.4	112.7	14.6	221.8	6.4	97.0	1.3	20.2
Cerebellum	28.9	437.5	21.6	327.3	32.4	490.9	46.2	700.0	26.4	400.0
Skin, shoulder	10.7	162.4	10.9	164.8	5.4	81.2	10.5	159.4	8.2	124.5
Skin, buttock	13.2	200.6	5.3	80.9	6.1	92.1	6.2	93.9	5.0	75.8
Stomach	20.4	308.8	33.4	506.1	13.2	199.4	19.3	291.8	15.4	233.9
Lung	1.8	26.7	12.6	191.5	5.7	86.7	26.8	406.1	9.4	141.8
Liver #1	10.6	161.2	12.2	185.5	23.2	351.5	13.4	202.4	10.1	152.7
Liver #2	11.5	174.5	20.4	309.1	36.8	557.6	13.0	197.0	15.4	233.3
Intestine	17.4	263.6	20.6	312.1	28.4	430.3	97.2	1472.7	23.9	361.4
Pancreas	24.8	375.8	10.3	156.4	19.6	296.4	19.4	294.5	7.3	110.3
Seminal vesicle	2.0	29.8	0.4	5.5	14.8	223.6	1.7	25.8	12.3	186.1
Skeletal muscle	1.4	21.1	5.0	75.8	3.3	49.2	6.6	100.6	0.5	7.6
Tongue	12.5	188.7	12.4	188.5	7.7	117.0	12.8	193.3	6.9	104.2
Heart	9.1	138.3	5.6	85.2	8.2	124.2	4.9	73.9	8.8	133.3
Spleen	65.6	993.2	103.6	1569.7	48.8	739.4	69.6	1054.5	63.6	963.6
Kidney	12.0	182.4	17.4	263.6	26.0	393.9	20.2	306.1	14.5	219.5
Testis1-1	19.2	290.6	26.6	403.0	13.2	200.6	30.3	459.1	27.4	415.2
Testis1-2	11.2	170.3	16.4	248.5	12.5	189.7	11.8	179.4	7.9	120.0
Testis2-1	13.9	210.9	33.8	512.1	8.3	125.5	17.6	266.7	5.4	82.4
Testis2-2	15.4	233.9	9.8	149.1	24.6	372.7	7.9	120.0	17.1	258.8
Sperm	0.8	11.7	0.9	13.9	1.2	18.5	0.3	4.7	0.8	12.7
Tail	10.4	158.2	16.1	243.6	11.1	168.5	15.0	227.9	14.6	220.6
Tail 8week			6.8	102.7	7.1	107.9	5.5	83.6		