Characterizing Mutagenesis Across Developmental Time with single-cell indexing (sci) ATAC-seq

Yu-Chen Pan

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Kelley Harris

Nasa Sinnott-Armstrong

Alison Feder

Program Authorized to Offer Degree:

Institute of Public Health Genetics

University of Washington

Abstract

Characterizing Mutagenesis Across Developmental Time

with single-cell indexing (sci) ATAC-seq

Yu-Chen Pan

Chair of the Supervisory Committee:

Kelley Harris

Genome Sciences

Mutations in DNA are caused by replication errors or exposure to mutagen that damage the DNA repair mechanisms. Germline variants are mutations that occur in germ cells and can be passed onto offspring, ultimately becoming polymorphic sites; while somatic variants are mutations that arise spontaneously in the soma cells during growth and aging. Somatic mutations have been traditionally studied in cancer due to their natural clonal expansion. However, recent work has described an association of the accumulation of somatic mutations in healthy tissues with aging and age-related diseases. The current methodologies for obtaining clonal sequences from healthy and developing tissues are either costly or laborious, limiting scalability. Therefore, in this thesis, we explored the feasibility of using single-cell combinatorial indexing (sci) ATAC-seq data to

identify somatic and germline mutations and to study mutational processes across tissues during embryogenesis. Leveraging available sci-ATAC-seq datasets from fruit fly embryo and human fetal samples, we were able to identify mutations and differentiate them into germline polymorphisms and somatic mutations. In the fruit fly embryo dataset, we detected population structure based on the extracted germline polymorphisms, while in human fetal samples, we observed an increase in somatic mutation burden over developmental time. We also performed mutational signature extraction, finding that the activities of clock-like signatures, such as SBS1 and SBS5, positively correlated with developmental time. This indicates a potential association between somatic mutation accumulation and aging during embryogenesis. We also observed variations in the mutational processes across tissues. Finally, we reconstructed the main tissue layers of early development from the detected somatic mutations, suggesting these datasets may help validate transcriptionally based lineage predictions. Our analysis showed that the reanalysis of available sci-ATAC-seq datasets can be an alternative solution to study somatic mutagenesis at a more affordable cost and enhanced scalability. These findings also have provided insights into developmental processes and cell lineage tracing during embryonic development.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have supported me throughout the process of completing my thesis.

First and foremost, I extend my sincere thanks to my committee chair, Dr. Kelley Harris, for believing in me and giving me the opportunity to contribute to this project. Her continuous encouragement and unwavering guidance have been fundamental in shaping both my scientific journey and personal growth. I am also extremely appreciative of the incredible guidance and support from Dr. David Mas-Ponte. His mentorship, patience, and insightful feedback helped me overcome many challenges during the journey. A heartfelt thanks to Dr. Annabel Beichman for her valuable insights and steadfast support.

I would also like to thank the members of my thesis committee, Dr. Nasa Sinnott-Armstrong and Dr. Alison Feder for their valuable insights and support throughout the process. Special thanks to Dr. Diego Calderon for providing access to the datasets used in this thesis, as well as for his assistance and advice throughout the research process.

I am deeply grateful to my colleagues and friends for their discussions and assistance. The collaborative environment greatly enriched my research experience. I am also indebted to my family for their unconditional love and encouragement.

This thesis is a testament to the collective support and contributions of all these individuals. I am sincerely grateful to each one of you.

INTRODUCTION

BACKGROUND

Germline variants are mutations that occur in the germ cells and are inherited to the offspring; most of these mutations are imprinted in all cells of an individual (Yu et al., 2024). In contrast, somatic variants are mutations that occur in cells other than germ cells; these mutations usually originate from a single cell and propagate as a result of erroneous DNA replication, endogenous process and exposure to external mutagens that cause errors in DNA repair mechanisms (Alexandrov et al., 2013; Beichman, 2023; Yu et al., 2024; Harris & Pritchard, 2017). Both germline polymorphism and somatic mutations can be classified into a mutation spectrum by their trinucleotide context, the 5' and 3' bases adjacent to the mutated loci. In the past, there has been an extensive effort to identify and catalog these patterns into mutational signatures, especially in cancer studies. Mutational signatures help unravel the biochemical processes that lead to mutations underlying the observed mutation spectra (Alexandrov et al., 2013; Islam et al., 2022).

Somatic mutations were initially studied in the context of cancer, but some studies have shown that the accumulation of somatic mutations is associated with aging and this genomic instability may contribute to age-related diseases (Beichman et al., 2024; López-Otín et al., 2023). And with the advent of new sequencing technologies, there has been a growing interest to study somatic mutation accumulations in normal cells to better understand its potential associations with age-related diseases and the possibility of tracing cell lineages during development (Dou et al., 2018; Manders et al., 2021; Solís-Moruno et al., 2023, Cagan et al., 2022). For example, recent research has used whole-genome sequencing to study the mutagenesis in tissues during

development and reconstruct phylogenetic trees for cell lineage through somatic mutations (Coorens et al., 2021; Spencer Chapman et al., 2021; Park et al., 2021).

However, unlike tumors, healthy tissues are composed of heterogeneous populations of cells and will require special equipment to expand the specific cell types before sequencing their DNA. The current methodologies to obtain clonal sequences include single cell expansion in cell culture, laser capture microdissection (LCM), whole genome sequencing and duplex sequencing. *In vitro* clonal expansion is isolating a single stem cell and expanding it in culture until there is enough DNA to harvest for bulk sequencing, but this is limited to the cell types that can be expanded in tissue culture; LCM involves dissecting small groups of cells that arise from a single natural clonal expansion under microscopic visualization (Espina et al., 2006), but this depends on finding at least semi clonal population in the tissue sample; whole genome sequencing (Olafsson & Anderson, 2021) involves sequencing the complete DNA sequence of the genome, but is restricted detect mutations with relatively high variant allele frequency (VAF) (Alioto et al., 2015; Dou et al., 2018); duplex sequencing employs tag-based error correction method to detect very rare mutations with improved accuracy, but it suffers from low sequencing depth (Schmitt et al., 2012). In general, these technologies are either expensive or labor intensive, making them difficult to scale up in high-throughput experiments.

Given this constraint, our goal is to investigate the possibility of using *single-cell combinatorial indexing* (sci) ATAC-seq data to extract somatic and germline mutations. The Single-cell Combinatorial Indexing (sci) methodology uses the combination of molecular indexes and the "split-pool" method to label a large quantity of cells or nuclei with unique nucleic acid barcodes,

allowing the study of single cells with enhanced scalability and at reduced cost compared to the conventional single cell techniques (Cusanovich et al., 2015; Martin et al., 2023). In their most recent study, the sci methodology was applied to study the single cell gene expression profiles in mouse embryonic development and was able to generate a single dataset equivalent to 30% of the Human Cell Atlas at a relatively modest cost (approximately 370K USD for extraction, reagents and sequencing costs combined) (Qiu et al., 2024). In other studies, sci method has been applied to ATAC-seq to investigate the chromatin accessibility landscapes during tissue development in different organisms like worms, fruit fly, mouse and humans. (Cao et al., 2017, Calderon et al., 2022; Cusanovich et al., 2015; Domcke et al., 2020).

AIMS

In this thesis, we aim to explore the feasibility of identifying variant sites in reads derived from sci-generated chromatin-accessible regions with the available fruit fly embryos and human fetal samples sci-ATAC-seq datasets. From the obtained variants, we aim to explore the mutational processes across tissues during development. Finally, we aim to explore the potential of using acquired somatic mutations to reconstruct and validate cell lineages, which were traditionally inferred from transcriptional profiles and chromatin accessibility.

RESULTS

DETECTING MUTATIONS FROM SCI-ATAC-SEQ3 DATA

For the fruit fly embryonic development, we were able to retrieve 43,382,175 variants for AD1, 1,891,765 variants for AD2 and 13,942,684 variants for AD3+. We categorized the obtained mutations according to their allelic depth in AD1, AD2 or AD3+ variant sets (see Methods).

Within each variant set, we explored their 1-mer mutational spectrum, which classifies the mutations according to their reference and alternative bases. For the analyses, we focused on the subset of AD3+ mutations because AD1 and AD2 are likely sequencing artifacts (Supplementary Fig. 1). In the 1-mer mutation spectrum among AD3+ variants, we observed the highest relative frequency of C>T mutations followed by T>C mutations (Fig. 1). To further investigate the consistency of this mutational profile with the previous study, we computed their trinucleotide context, which resulted in 96 mutation types. The trinucleotide profile was enriched in the TCN>T (where N is A, C, G or T) within the mutation class C>T and in TTN>C peaks within the mutation class T>C. We compared this 3-mer mutation spectrum in our datasets with the spectrum in Asaaf, 2017, which is constructed from whole genome sequencing of mutation accumulation experiments of 17 *Drosophila melanogaster* lines (Fig. 2). Overall, similar patterns of mutation class frequencies are observed between the two spectra, especially in the C>T and T>G (TTN > G) mutation classes. However, mutations in the T>C component of the profile are significantly enriched in our dataset compared to the mutation accumulation experiments. We suspect that this may be a characteristic of polymorphism in our data, possibly due to mis-polarization from the differences between the experimental strain Canton-S and the reference strain dm6.

In addition, we also measured the distribution of variant allele frequencies in our samples for AD3+ variants (Fig 3). The distribution shows a decaying pattern enriched in young or low frequency mutations (Beichman et al., 2024). Of note, as the fruit fly embryos used in this study come from a diverse population (Calderon et al., 2022; Cusanovich et al., 2018), the allele frequency extracted in this analysis will likely represent the frequency within that population.

The resulting plot mimics a site frequency spectrum (SFS) in a neutrally evolving population, suggesting the reliability of the extracted polymorphisms from sci-ATAC-seq3 data is sufficient to recreate a population structure.

INVESTIGATING SOMATIC MUTAGENESIS ACROSS TISSUE DEVELOPMENT DURING HUMAN EMBRYOGENESIS

From the human fetal sample dataset, we obtained a total of 2,332,288 mutations. We then split these mutations according to their cell allele frequency with a hard cutoff at 25%. We also removed samples (n=9) with fewer than 101 somatic mutations in both somatic and germline variants set for all the downstream analysis (Fig. 4, see Methods).

Mutations in somatic cells show distinct mutational processes than extracted polymorphisms

We first obtained the trinucleotide spectra of mutations in both the somatic and germline variant set and performed a principal component analysis (PCA) on the frequency of each 3-mer. We observe a clear differentiation between germline polymorphisms and somatic variant sets based on their trinucleotide context, as demonstrated by the PCA (Fig. 5a). However, when grouping the samples based on the germ layers from which their tissues derived (Supplementary Table 1), there isn't a distinct separation (Fig 5b).

To better understand specific mutation processes that we might be capturing in our data and to identify possible artifacts, we further detangled the mutation profiles of each sample into mutational signatures. Mutational signatures can reveal specific patterns associated with particular mutational mechanisms, providing insights for mutagenesis during tissue development.

We ran SigProfilerExtractor (Alexandrov et al., 2013) on our dataset (see Methods) and extracted 7 *de novo* signatures (Supplementary Fig. 2) based on the selection plot that gives us the most sample reconstruction with a local maximum in average stability (Fig. 6). From the extracted *de novo* signatures, we performed receiver operating characteristic (ROC) analysis to select the most predictive signature in terms of distinguishing germline and somatic variants (Fig 7a). The corresponding AUC values for the ROC curve are listed in Table 2. Notably, all mutational signatures were able to differentiate the mutation class to certain extent, implying mutation patterns occurring in the putative-somatic class were distinct from the germline polymorphism class. We used the two signatures with the highest AUC, SBS96D followed by SBS96A, and plotted the proportion of the signature's activity found in each of the samples (Fig 7b). Most of the germline mutation samples have higher SBS96A signature activity, while most of the somatic mutations have higher SBS96D signature activity (Supplementary Fig. 3). These results suggest that the polymorphisms in the germline and the somatic cells are involved in independent mutational processes during tissue development.

Mutational processes during embryogenesis

The mutation profile of the most predictive de novo signature distinguishes between variant sets might reveal the fundamental mutational mechanisms that differ in germline and somatic cells during development, we therefore delved deeper into the two most predictive de novo signatures, SBS96A and SBS96D, to uncover the specific mutational mechanisms underlying tissue development (Fig. 8). We first decomposed the *de novo* 3-mer mutation spectra into mutational signatures from the COSMIC database (Sondka et al., 2024), most of which have annotated etiologies and associated molecular mechanisms. SBS96A is decomposed into SBS1 (70.14%),

SBS5 (17.4%), and SBS54 (12.46%) (Fig. 8a). SBS1 and SBS5 are considered ubiquitous signatures because they are present in most tissues including *de novo* mutations, polymorphisms and somatic mutations (Alexandrov et al., 2020; Blokzijl et al., 2016; Beichman, 2023). These signatures are also coined as clock-like because their behaviors are found to increase linearly with age (Alexandrov et al., 2015). We also observe the spike in T>C mutations at TpG sites corresponding to the artifactual signature of SBS54. This signature in the cancer analysis is likely a result from a mis-polarization of the ubiquitous signature SBS1, which is an indication of germline polymorphisms. On the other hand, SBS96D is decomposed into SBS4 (51.84%), SBS43 (30.18%), SBS5 (14.5%) and SBS1 (3.48%) (Fig. 8b). Although SBS4 is associated with exposure to tobacco carcinogens in cancer tissues, its profile enriched in C>A mutations is shared with other signatures that capture naturally occurring oxidative DNA damage (Poetsch, 2020). The SBS43 currently has an unknown etiology but is suggested as a potential sequencing artifact. Dissecting the mutation spectrum into mutational signatures gives us an understanding of the underlying mutational processes taking place during embryonic development. The rest of the decomposed *de novo* signatures and their corresponding etiologies can be found in Supplementary Fig. 2 and Table 2.

Mutational processes across developmental time

To study the mechanisms of mutagenesis in tissues during embryogenesis, we plotted the mutational load of each sample normalized by the accessible genome across days of pregnancy (Fig 9). We fitted a regression line and hypothesized that the mutational burdens should increase over time in soma cells but remain constant in the germline polymorphisms since these mutations are inherited at birth. Although it shows an increase in mutational burden across developmental

time in somatic cells, contrary to our expectations, germline mutations also show an increase over time. We suspected the increase in the mutations could be due to the confounder of total read coverage. Therefore, we plotted the residuals from the fitted linear model of mutation counts against coverage, with days of pregnancy (Supplementary Fig. 4). Although there is a slight correlation between the residuals and days or pregnancy, they are not significant (adjusted r-squared value of 0.218 and 0.0145, p-values are 0.138 and 0.17 for germline and somatic respectively).

Mutational processes across tissues

Similar to the increase for mutational load, we also observed an increase in the mutational signature activities for both SBS1 and SBS5 over developmental time (Fig. 10a). We are interested to see if SBS1 and SBS5 activities might differentiate between tissues at this early stage as it is observed in adult tissues. In the adult tissue, the replicating or mitotic cells, such as those in the colon and small bowel, show a more rapid increase in SBS1 activity while SBS5 activity remains relatively constant over time. However, this relationship is reversed in noon actively replicating post-mitotic cells such as neurons, liver, and lung (Blokzijl et al., 2016; Abascal et al., 2021; Spisak et al., 2023). It is hypothesized that SBS1 is associated with or driven by rounds of DNA replication, whereas SBS5 is independent of DNA replication cycles and increases with biological time. Therefore, we analyzed the activities of these signatures across time in each tissue type for our somatic mutation class (Fig. 10b), but we did not observe a significant distinction between SBS1 and SBS5 activities in different tissues; all the regression lines are parallel. Nevertheless, we do see variations of these activities across developmental time. In general, adrenal, brain, heart and intestine exhibit upward mutational signature activities,

while kidney, liver, lung, and placenta showed reduced activities across the developmental time span. It is important to note that these trends are extrapolated from sparse data points and could be sensitive to noise.

BUILDING CELL PHYLOGENIES FROM SOMATIC MUTATIONS DURING DEVELOPMENT

One potential application of detecting somatic mutations during development is to use these somatic mutations (extracted from sci-ATAC-seq3) to reconstruct cell lineages that recapitulate the developmental lineage of each tissue. We used the algorithm Unweighted Pair Group Method with Arithmetic Mean (UPGMA) to build a phylogeny tree for each donor. We calculated the phylogenetic distance of each pair of samples by counting unique mutations present in each tissue within the shared accessible regions. This ensured sufficient coverage in both tissues and comparability between them (see Methods). The resulting trees represent the cell lineage of the available tissues in each donor. The tissue is color coded with its corresponding germ layer. For instance, for donor H27431 shown in Fig 11, tissues originating from the same germ layer are found together within the same clade in the soma cells. However, the phylogenetic tree that was reconstructed with germline variants displays more dispersed groupings. This is expected for germline mutations because they are inherited at birth and should not differentiate during tissue development. Other donors in our dataset present groupings to a lesser extent (Supplementary Fig. 5).

In order to obtain a quantifiable measure for the phylogenetic distance for the tissue pairs across all donors, we then classified each tissue pair into three categories based on their phylogenetic distance: technical replicates, tissues derived from the same germ layer, and tissues derived from

different germ layers (Fig 12). We hypothesized that the phylogenetic distance between the pairs in the technical replicates group should be the lowest, followed by the same germ layers, and the different germ layers be the highest. In the putative somatic variant set, we show that the median of the distances among technical replicates are the shortest with a fold change difference against the distinct germ layer of 0.466 (p-value = 0.016). Tissue pairs in the same germ layer also show lower distance in medians compared to distinct germ layers (f.c. 0.958), although this difference is not significant (p-value = 0.95). However, when computing the mean distance in each category, we see that the mean distance in the same germ layer is slightly higher than the distinct germ layer, but is also not significant (f.c. 1.038, t.test p-value = 0.777). In the germline polymorphism sets, which should not reconstruct a representative phylogenetic tree, it also shows the shortest phylogenetic distances in the technical replicates. But the highest phylogenetic distance is found in the pairs within the same germ layer. This suggests we should be mindful of making further conclusions inferred from the obtained trees (Fig 11).

Overall, here we show the potential of utilizing the somatic mutations detected from sci-ATAC-seq to build a phylogenetic tree that recapitulates cell lineage during embryonic development, but it is important to note that there are still caveats and limitations with this approach.

DISCUSSION

We have shown that it is possible to detect single nucleotide variants from sci-ATAC seq data and extract a putative set of somatic variants. Contrary to other methods available (Dou et al., 2018; Muyas et al., 2023), our approach does not require a direct tissue matching between subpopulations of cells nor a direct linkage to any common polymorphisms. We have also

explored two datasets, from fruit fly (Calderon et al., 2022) and human (Domcke et al., 2020) that were not previously surveyed for mutations. Both principal component analysis (PCA) on the mutation spectra and an accuracy test using ROC curve analysis of mutational signature activities provide evidence that we can differentiate the extracted mutations further based solely on allele frequency, and that these groups have distinct patterns of trinucleotide accumulation. We also presented that the underlying mutational processes present in both germline polymorphisms and soma cells by extracting the mutational signatures from the stratified mutation sets. Among the decomposed signatures, SBS96A is particularly relevant as it contains SBS1 and SBS5, together with the mispolarization signature (SBS54). SBS96A shows higher activity in the germline polymorphisms set (see Fig. 7b); and thus the presence of SBS54 is most likely originating from the mispolarization of the CpG deamination associated mutations, indicating that our variants in the germline variant set display expected profile for a polymorphism loci profile (Alexandrov et al., 2020; Beichman, 2023.; Mathieson & Reich, 2017). We also identified a somatic specific mutational signature, SBS96D, which is also able to classify among somatic and germline variants. The C>A mutation component in our spectrum appears authentic, i.e. does not seem like any obvious artifactual signature from COSMIC, could represent unrepaired oxidative damage that occurs during the rapid division rate in development (Poetsch, 2020). We note that previous studies (Spencer Chapman et al., 2021) using deep targeted sequencing also observed a C>A mutational process occurring during development with a limited similarity at the trinucleotide level. However, other components of the extracted signatures do resemble potential sequencing artifacts and should reserve caution upon interpretation.

In terms of mutation accumulation during embryonic development, we see an increase in mutation burden and activities of the clock-like signatures SBS1 and SBS5 in all samples in the soma cells. This could potentially imply that the association between the accumulation of somatic mutations and aging (Abascal et al., 2021; Alexandrov et al., 2015) begins to unfold during embryogenesis. However, we also observed this increasing trend in the germline polymorphism variant set, which is not expected because the amount of inherited polymorphic sites should remain constant across the development and aging of the individual. This suggests there might be noise coming from the sequencing artifacts in the germline variant sets and requires caution when interpreting the results in the putative somatic set. Additionally, the narrow time span in our data, compared to other aging studies (Abascal et al., 2021; Cagan et al., 2022) also restricted our ability to investigate the effect of aging in our dataset.

We also explored if different tissues undergo distinct mutational processes during development. However, there is no clear clustering or separation between tissues or germ layers, and there are no significant variations in the clock-like signature activities within tissues. This could be because the mutations accumulated in the tissues were not differentiated enough at this early developmental time. Some of the samples exhibit data sparsity and the developmental time span is limited. Additional data is thus needed to conclude how mutational processes differ between tissues during development.

Lastly, while the previous literature demonstrated the usage of somatic mutations reconstruct cell lineage with whole genome sequencing (Coorens et al., 2021, 2024; Spencer Chapman et al., 2021), we demonstrated the possibility of reconstructing the tissue development from the

somatic mutations detected from sci-ATAC-seq using UPGMA. To further support this, we found that technical pairs have the shortest phylogenetic distance among the detected variants and that tissues from the same germ layer are located within clades in the resulting trees (Fig. 11 and Supplementary Fig. 5). Although this is still preliminary data and will require future research to increase the resolution, it still suggests that our methodology could be used to provide orthogonal evidence for the cell lineages derived from transcriptomic and chromatin accessible data (Domcke et al., 2020; Qiu et al., 2024). Overall, these results open the doors to future refinement of the cell lineage mapping.

CONCLUSION AND FUTURE DIRECTION

In this thesis, we have developed a new method that aims at extracting variants from sci-ATAC-seq3 data and is able to differentiate between somatic variants and germline polymorphisms. Additionally, we investigated mutagenesis during tissue development in embryogenesis. Finally, we demonstrated the potential of using these detected mutations to reconstruct a cell lineage. We believe this innovative method could serve as a cost-effective alternative for studying mutagenesis in tissue development.

There are limitations in our study and will require future improvements. It is important to interpret the results with consideration since most of the samples on COSMIC come from tumor samples whereas our samples are from healthy tissues. For example, SBS4 observed in our data might be due to other factors that have similar etiologies to tobacco carcinogens that occur during normal development, such as oxidative stress. With the current COSMIC signatures and literature, it is difficult for us to draw a conclusion about the etiology of the detected mutational

signatures as the current data on human fetal studies is limited. Furthermore, the phylogenetic trees for some of the donors do not exhibit expected clades; and it is difficult to interpret the length of the tree branch. A recently developed method to reconstruct phylogeny tree capturing cell lineage, Sequoia, can be applied on genome-wide somatic mutation sets from clonal samples and single-cell genomic data (Coorens et al., 2024). We are interested in applying this method to our data to see if we can obtain a more accurate phylogenetic tree representation for tissue development during embryogenesis.

METHODS

DATA

Fruit fly embryos data

sci-ATAC-seq3 data from synchronized Canton-S fruit fly embryos was obtained from (Calderon et al., 2022) and it is publicly available in (Data). In brief, embryos were collected across 11 windows with each window spaced 2 to 4 hours apart. Each collection of embryos was then pooled to be sequenced via sci-ATAC-seq3. The sequencing reads were mapped to the reference genome dm6 and were processed using a standardized pipeline for quality control, filtering low read depths and reads mapping to mitochondria or ribosomal genes. This dataset comprises a total of 1.5 million cells.

sci-ATAC-seq3 Human fetal samples data

Raw sci-ATAC-seq3 data from human fetal samples was obtained from (Domcke et al., 2020). In brief, 59 fetal samples, representing 15 organs, were obtained from 23 fetuses with an estimated

post-conceptual age ranging from 89 to 125 days. The raw reads were re-mapped to the GRCh38 genome assembly. In total, approximately 1.6 million cells were captured.

DETECTING VARIANTS

Detecting mutations from sci-ATAC-seq data

First, we processed the BAM files to filter out higher quality reads based on various criteria such as mapping quality, insert size, and number of mismatch positions. Then, we utilize samtools rmdup (version 1.19) to eliminate duplicate marks, accounting for artifacts in our base calling procedure. We then employed bcftools mpileup (version 1.19) to call variants. Upon obtaining an initial set of mutations, we revisited the read data (BAM files) to extract cell information including tissue, sample, and donor for each variant. We aggregate and pseudo-bulk the cells at the sample level and further refine our selection by focusing on mutations present in at least two cells, with a total coverage of 34, to minimize artifacts and sequencing errors.

Classifying variants from the detected mutations

In the fruit fly dataset, we used allelic depth (AD), measured as the number of alternative alleles at a given locus, to classify mutations in likely artifacts (AD1, AD = 1), putative somatic or rare variants (AD2, AD = 2) and common polymorphisms (AD3+, AD = 3+). In the human fetal data dataset, we used cell allele frequency distribution (CAF) to distinguish the mutations present across all cells. We expect variants with a CAF of <25% to be putative somatic mutations, and those with a CAF of >50% are likely germline mutations.

There is discrepancy in variant classification between the two datasets because of their experimental setup. In the fruit fly data, multiple embryos were pooled into a sample according to the collection windows; therefore, the germline polymorphisms allele frequency that are exclusive to smaller subset of flies within the sampled population is not expected to be 50%. In contrast, in the human fetal data, each embryo was pooled into a sample based on individual donor; therefore, the germline polymorphism allele frequency was expected to be 50% for heterozygous variant and 100% for homozygous variant. The cutoff of 25% was selected based on the histogram of variant allele frequency, where it clearly separated the major peak of heterozygous variants.

STATISTICAL ANALYSIS

Excluded samples

In the fetal human sample datasets, to avoid potential noise associated with low counts in samples with low coverage, we filtered out sets of variants with fewer than 101 detected somatic mutations and removed the corresponding sets of variants in germline variants. In total, we removed 9 samples: run3.H27423.brain, run3.H27469.lung, run3.H27791.bonemarrow, run3.H27820.pancreas, run3.H27846.eye, run3.H27847.thymus, run3.H27875.spleen, run3.H27962.gonad and run1.H27477.kidney.

PCA

We performed the PCA on the trinucleotide context of the mutations using R function prcomp (version 4.3.2). The first two principal components that captured the most variation (PC1: 27.9%,

PC2: 17.62%) were plotted. The data points were then colored by either their mutation types or the germ layers.

Mutation burden over time

We calculated the total mutations corrected by accessible genome regions. We then mapped the samples to their corresponding donor's days of pregnancy by the donor ID and calculated the mutational burden across developmental time. For the residual plot, we fitted a linear model of total mutation counts and coverage and plotted the residuals of the fitted model against days of pregnancy.

Signature proportion

We selected the signatures with the highest AUC and plotted the proportion of each *de novo* signature for all the samples. We divide SBS96D's activity across the total signature activity for every sample. We performed the same calculations for SBS96A. We plotted the signature SBS96A's activity proportion on x-axis and signature SBS96D's activity proportion on the y-axis. And then we colored the samples by the mutation types.

Signature across time

For the SBS1 and SBS5 signature across developmental time, the samples were grouped by the donor ID and were matched to the donor's pregnancy. For the signature stratified by tissue types, tissues with less than 3 samples were removed. A total of 6 samples were removed: run3.H27791.gonad.somaticvr, run3.H27791.gonad.germlinevr, run3.H27895.spleen.somaticvr ,

run3.H27895.spleen.germlinevr, run3.H27917.pancreas.somaticvr and, run3.H27917.pancreas.germlinevr.

Germline versus somatic classification by de novo signatures

From the SigProfiler Extractor output, we used the samples' *de novo* signature activities (total of 7 signatures, SBSA to SBSG) to predict if the inputted mutation set was attributed to the germline or somatic set. The most predictive de novo signature will indicate the specific mutational mechanisms that are fundamentally different in germline and somatic cells during development. Therefore, we used the R library pROC (version 1.18.5) to predict the class of the mutation set, either somatic or germline. From each prediction, we calculated a ROC curve, and AUC (area under the curve) to assess the performance and accuracy of each de novo signature.

MUTATIONAL SIGNATURE ANALYSIS

Since sci-ATAC-seq3 only captures sequencing reads within open chromatin regions, promoters, enhancers and other regulatory elements, the captured reads resulted in a significant divergence in trinucleotide content compared to the rest of the genome. Therefore, we need to adjust for this bias when comparing the mutation signatures in the cosmic database, which is obtained from whole genome sequencing.

In order to achieve this, we derive the correction factor by obtaining the ratio of trinucleotide context frequencies, comparing those obtained from the read coverage of our samples to those of the UCSC hg38 reference genome. We then multiply the number of mutations in each

trinucleotide context by the correction factor to get the number of mutations corrected for the background genomic composition.

Mutational signatures were then extracted from all available variant sets, including both the polymorphism and the somatic set within each individual sample. Due to the abundance of polymorphisms compared to somatic mutations in our data, we also needed to correct the total number of germline mutations to match the number of somatic mutations in each sample to ensure comparability between sets of variants. To do so, we divided the number of mutations in each trinucleotide by the ratio of total germline to somatic variants and rounded the resulting fractional values to simulate actual counts. This process resulted in an equal number of germline and somatic mutations, while maintaining the overall trinucleotide profile within each sample.

After the mutation profile matrixes were constructed, we run the SigProfiler Extractor (version 1.1.23) from 1 to to 10 signatures using default settings and parameters, which include both a *de novo* extraction and a decomposition of the extracted signatures to known COSMIC signatures (COSMIC v99). For each extracted signature, we plot the sample cosine distance and the average stability, as provided in the output of the SigProfiler Extractor (see Fig. 6). We selected 7 signatures for our analysis, which is the optimal number of signatures that maximize the stability of the solution while minimizing the reconstruction error.

PHYLOGENETIC ANALYSIS

Phylogenetic tree

We divided all of our sample variant sets according to their donor and to their variant classification (putative-somatic or germline). Within each donor and variant set, we computed all possible tissue pair combinations and calculated a phylogenetic distance. To calculate this distance, we first identified shared open chromatin regions that are common in each tissue pair by finding the number of shared regions of the genome (measured in nucleotides) where both samples were covered with at least 10 reads between their accessible genomes. We then extracted any mutations that are both present and available in these shared chromatin regions for each tissue pair. Next, we detected the mutations unique to each tissue among the available mutations. We compiled and tallied the total count of unique mutations found in each tissue separately and divide it by the total number of shared nucleotides for each tissue pair. Finally, we used the computed distance to infer a phylogenetic tree of the distinct tissues using the UPGMA algorithm from the R function hclust (version 4.3.2).

Phylogenetic distance category

We further grouped each tissue pair based on their phylogenetic distance into three categories: technical replicates, tissues derived from the same germ layer, and tissues derived from different germ layers. The table for the germ layers and their derived tissues can be found in Supplementary Table 1. We used the median to calculate the fold change difference between the distance of each category and used Wilcoxon rank sum test to calculate the p-value between groups. For the mean, we used student student's t test to calculate the p-value.

FIGURES AND TABLES

Table 1. | Summary of the main differences between the datasets. The table compares the key differences between the fruit fly embryo datasets and human fetal samples datasets.

	Fruit fly embryos	Human fetal samples
Reads	Each read is from different fly individual, treating all the cells as a whole population	Each read is from the same tissue, treating cells in each tissue as a population
Coverage	Lower coverage per sample	Higher coverage per sample per individual
Time points	0-20 hours, over lapping 2-hour collection window	89 -125 days
Advantages	Smaller sample size, Includes developmental times	Higher coverage, Single sample per individual
Goal	Mutation loads across developmental time	Mutational processes across tissue types

Fig 1. | 1-mer mutation spectrum of AD3+ mutations in fruit fly embryos dataset. Relative mutation frequency (which sums to 1) for the six mutation classes in AD3+ variants.

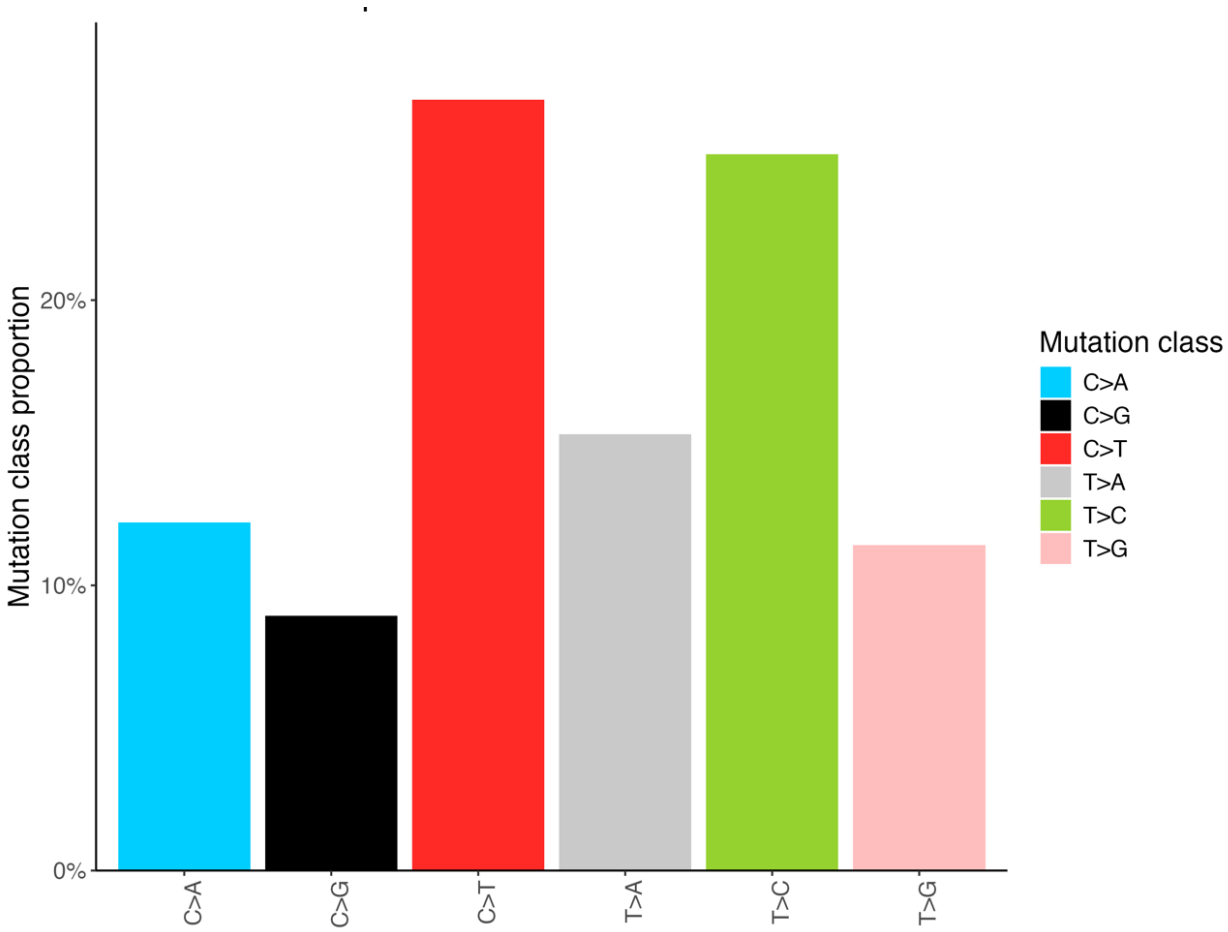
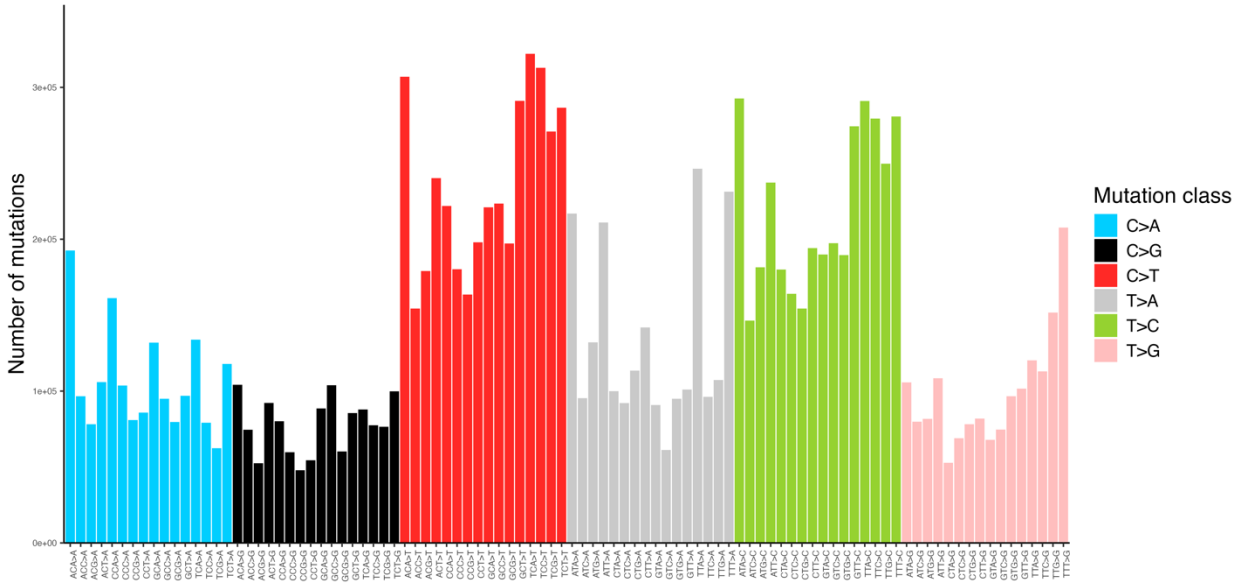


Fig. 2 | Comparison of mutation spectrum with previous literature. **a)** 3-mer mutation spectrum from AD3+ variants in fruit fly embryo dataset. The number of mutations counts for the 96 mutation types. The 96 mutation types were color labeled by the six mutation classes. **b)** 3- mer mutation spectrum retrieved from Assaf, 2017.

2a.



2b.

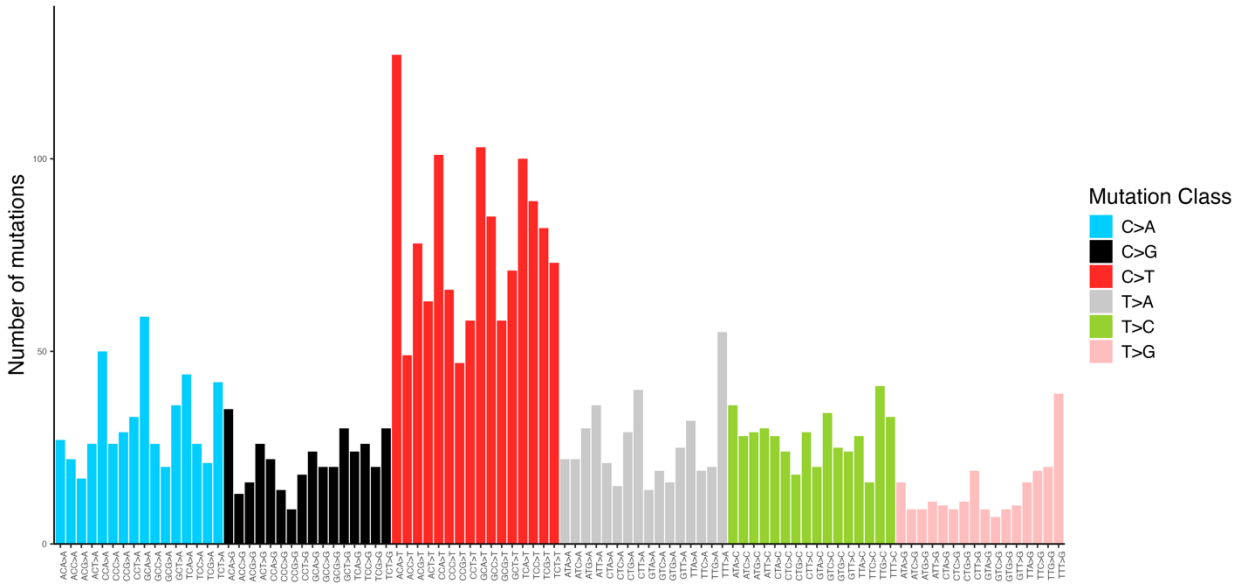


Fig 3. | **Histogram of variant allele frequency (VAF) distribution for the AD3+ variants**. The x-axis is the VAF distribution percentage and y-axis is the number of variants (in millions) corresponding to the derived allele frequency bin.

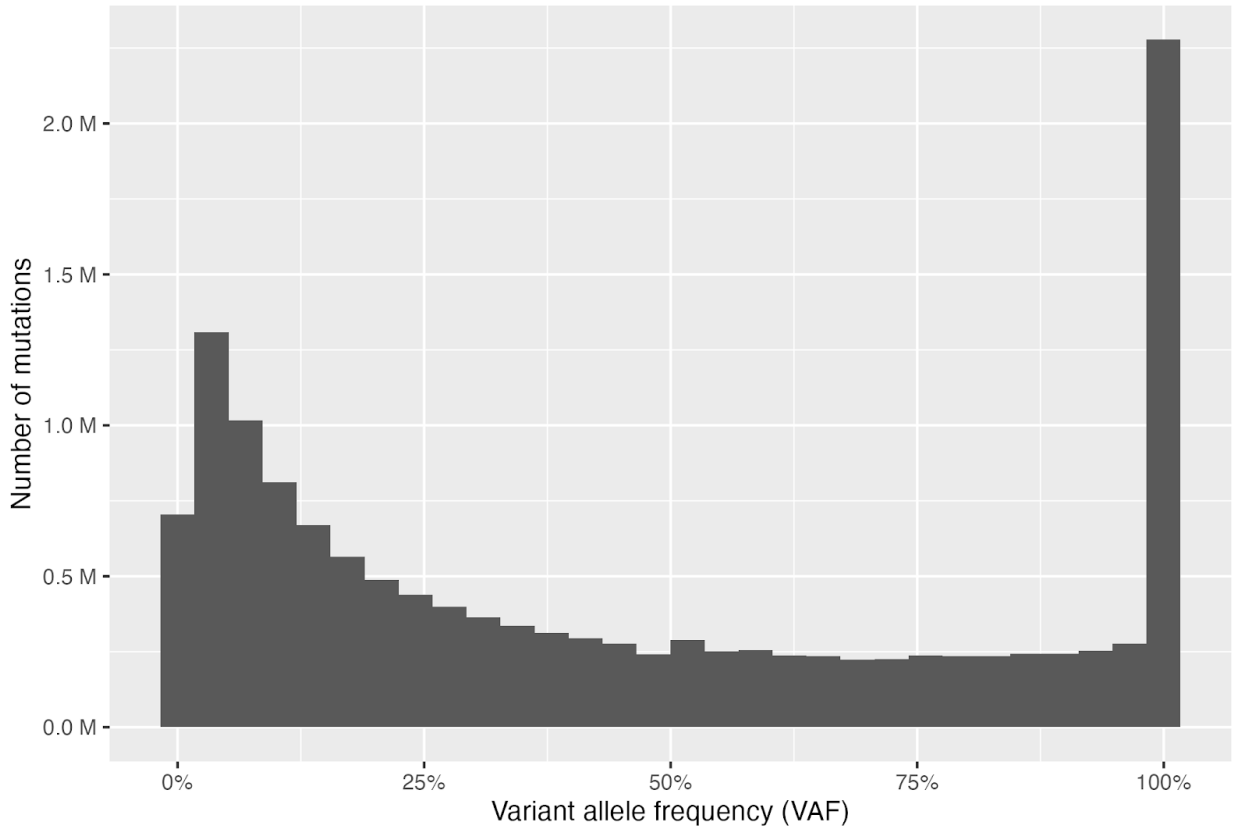


Fig. 4 | Total mutations in all the samples. Histogram showing the total mutation load for samples in inferred germline variants (CAF > 0.25) and inferred somatic variants (CAF <= 0.25). The line indicates the threshold at mutation counts of 100.

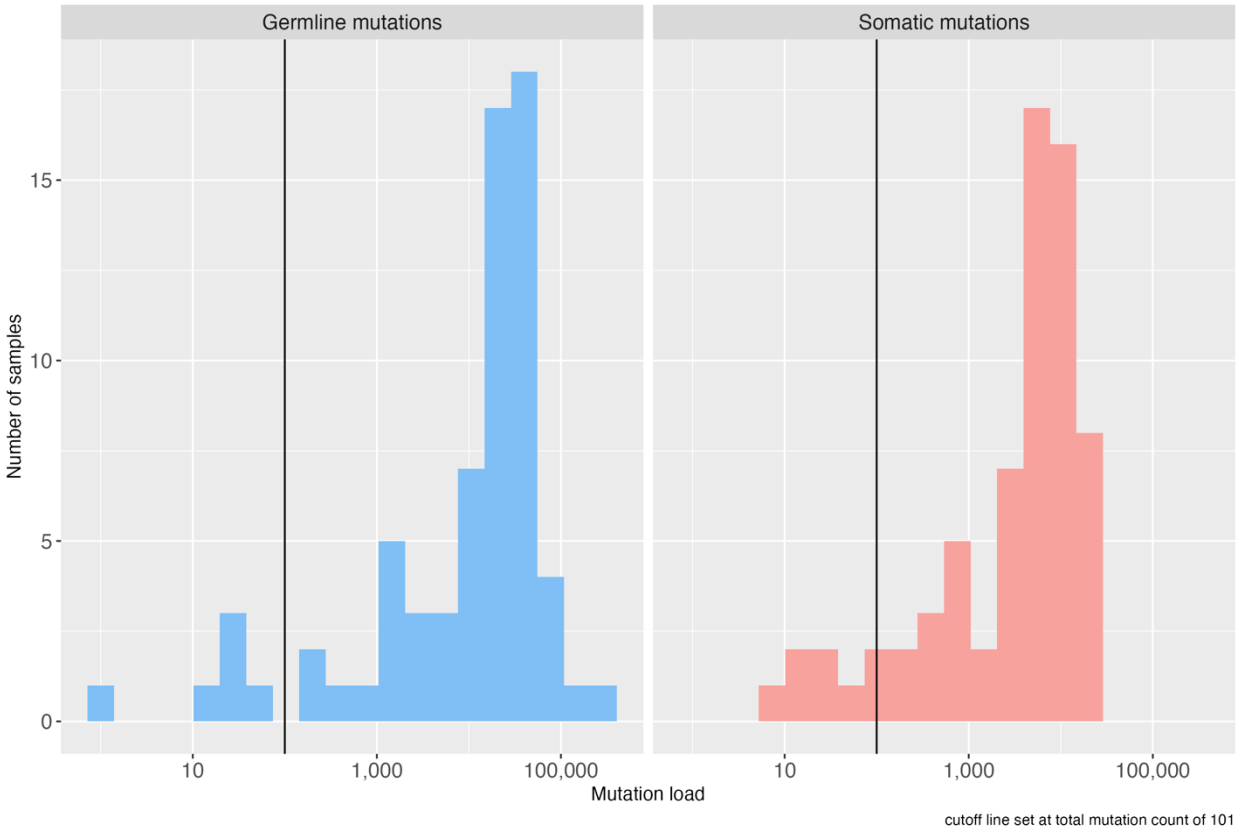
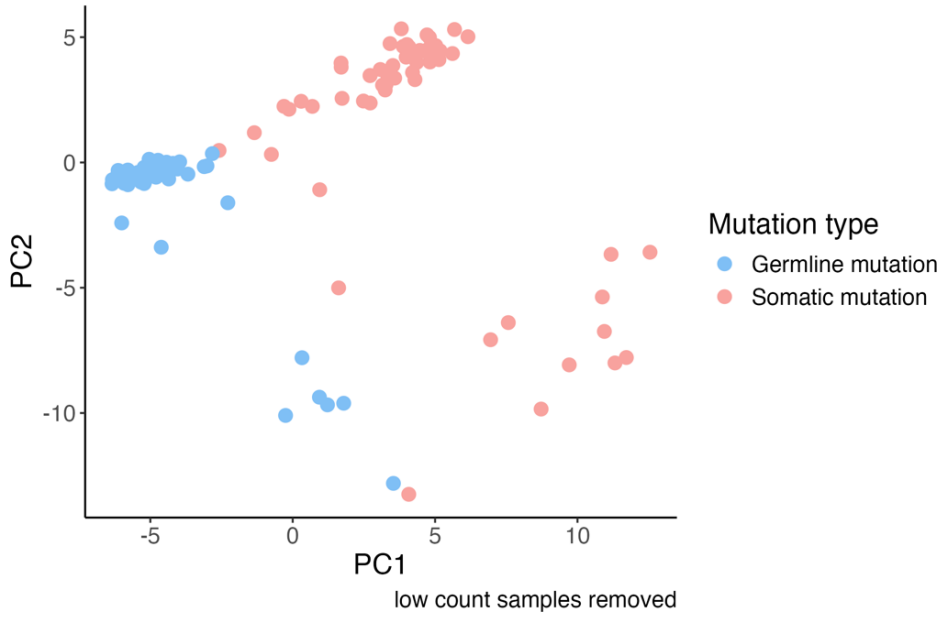


Fig. 5 | Differentiation of mutations explained by the trinucleotide context. Principal component analysis, centered and scaled, performed using mutations in the trinucleotide context. The clusters displayed the results with respect to the first two principal components. **a)** capacity of trinucleotide context to distinguish the germline variants and somatic variants. **b)** analysis as in **a** but with three germ layers: mesoderm, endoderm, and ectoderm.

5a.



5b.

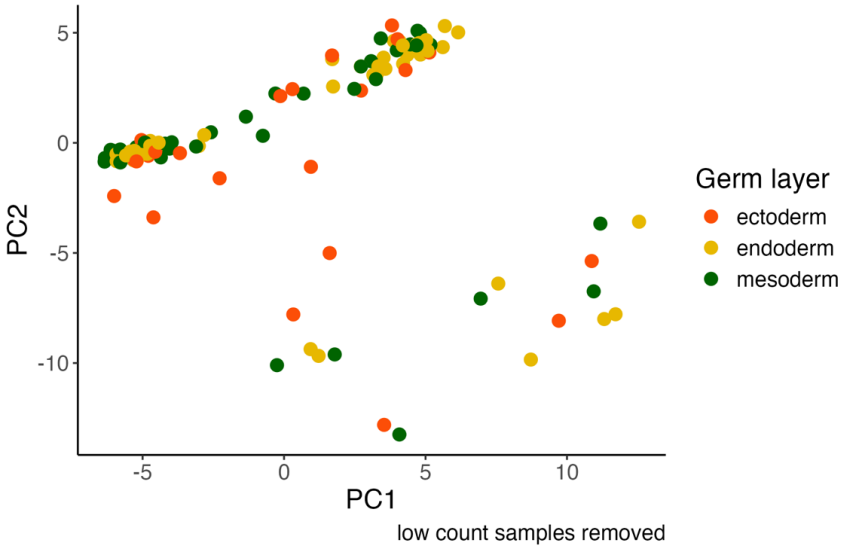


Fig. 6 | Selection plot for all the extracted signature from Sigprofile Extractor. The mean sample cosine distance is shown on the left y-axis and the average stability is shown on the right y-axis for each of the decomposition.

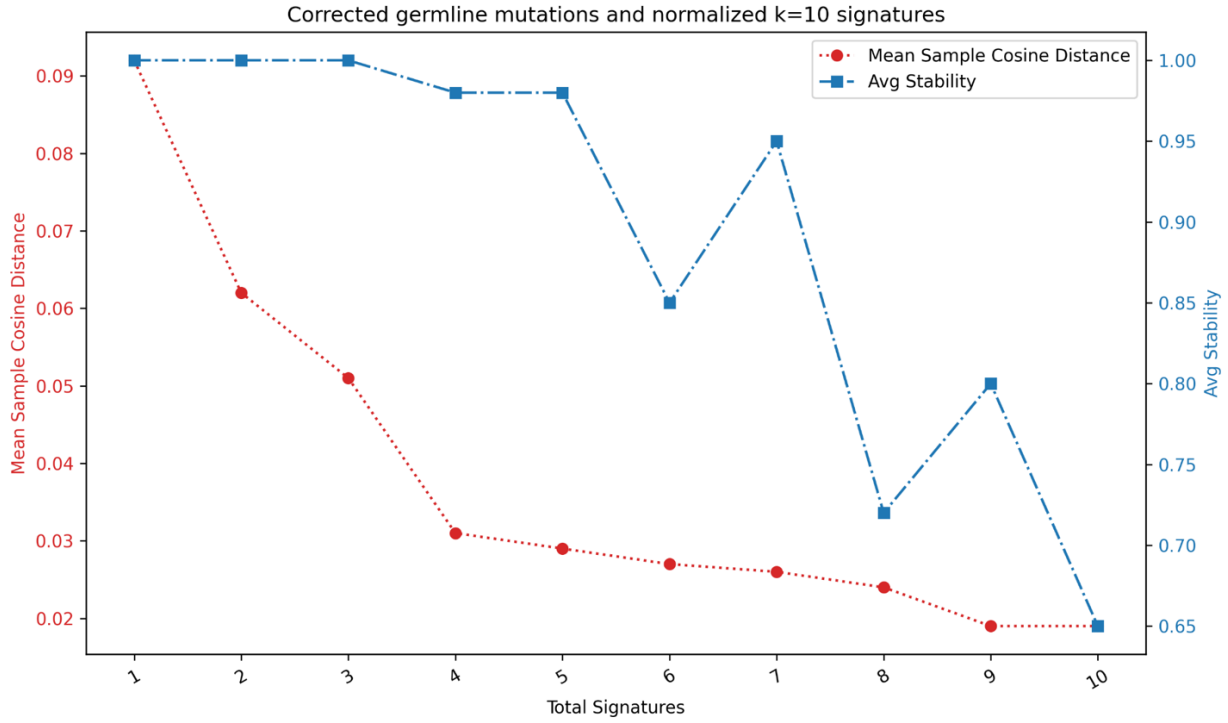
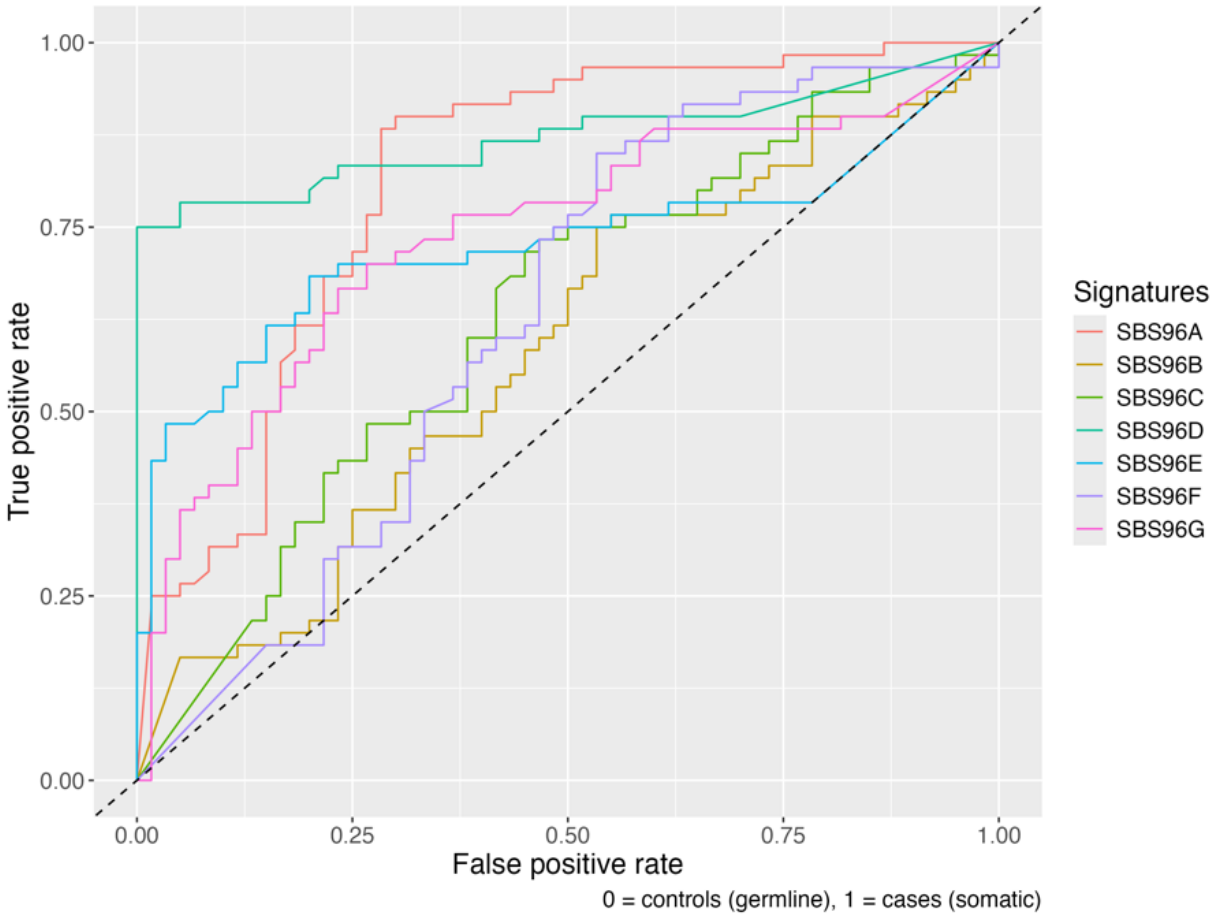


Fig. 7 | A summary of evaluating the most effective reconstructed de novo signatures to differentiate between germline and somatic variants. a) ROC curve to show the each of the reconstructed de novo signature's ability to differentiate between germline and somatic variants. **b)** distinguishing somatic and germline variants using the two de novo signatures with the highest AUC, SBS96D and SBS96A. The x-axis is the proportion of SBS96A, and y-axis is the proportion SBS96 in each sample. Each data point, representing each sample, was labeled for germline or somatic variants.

7a.



7b.

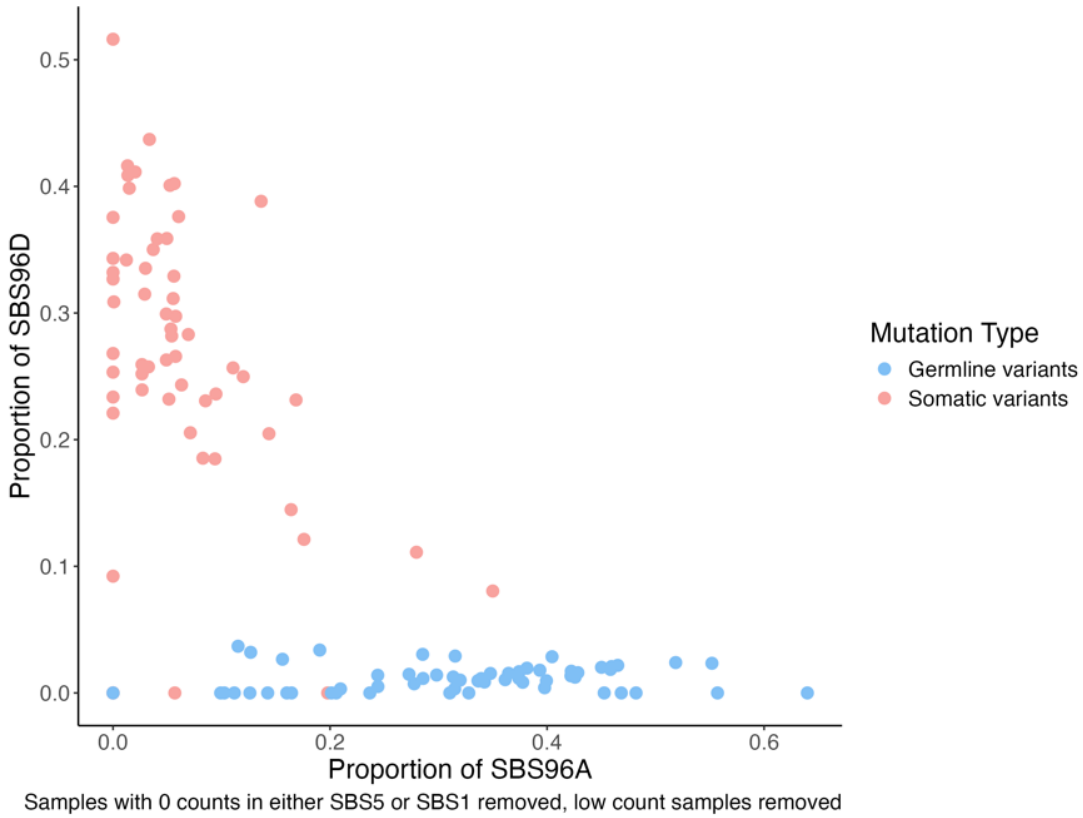
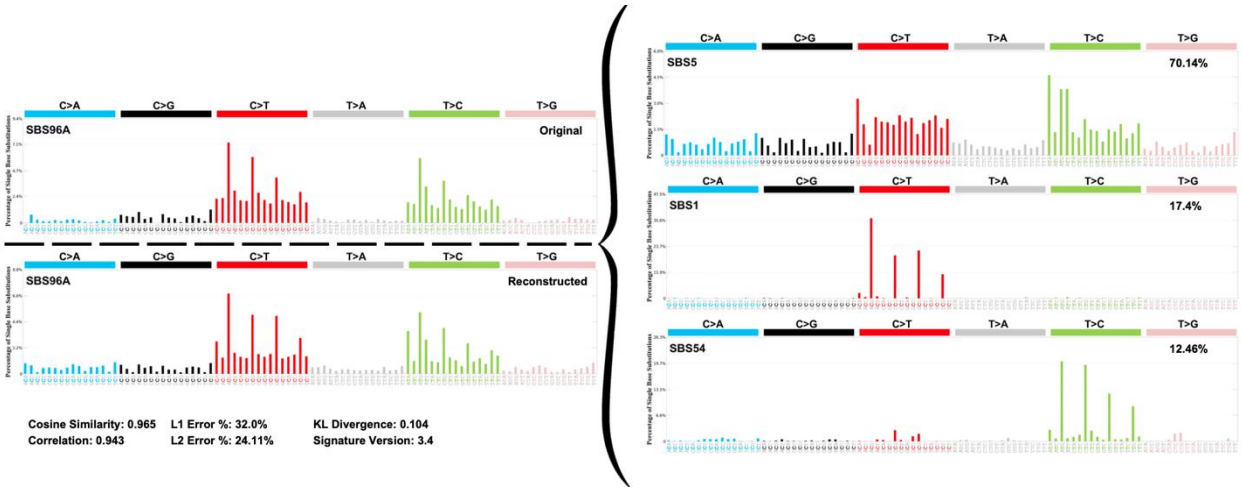


Table 2. | The AUC value for each ROC curve sorted from highest to lowest.

	SBS96D	SBS96A	SBS96G	SBS96E	SBS96C	SBS96F	SBS96B
AUC	0.8729167	0.8147222	0.7390278	0.7268056	0.6318056	0.6266667	0.5827778

Fig. 8 | Decomposition of 3-mer mutation spectrum into COSMIC mutational signatures. Original spectrum (top left) is the spectrum we obtained from our data; Reconstructed spectrum (bottom left) is the spectrum reconstructed from the COSMIC signatures. **a)** Reconstructed SBS96A spectrum and its decomposed mutational signatures. The percentage of the signature's contribution is shown on the top right corner. **b)** analysis as in **a** but for SBS96D. The rest of the decompositions are in supplementary Fig. 2.

8a.



8b.

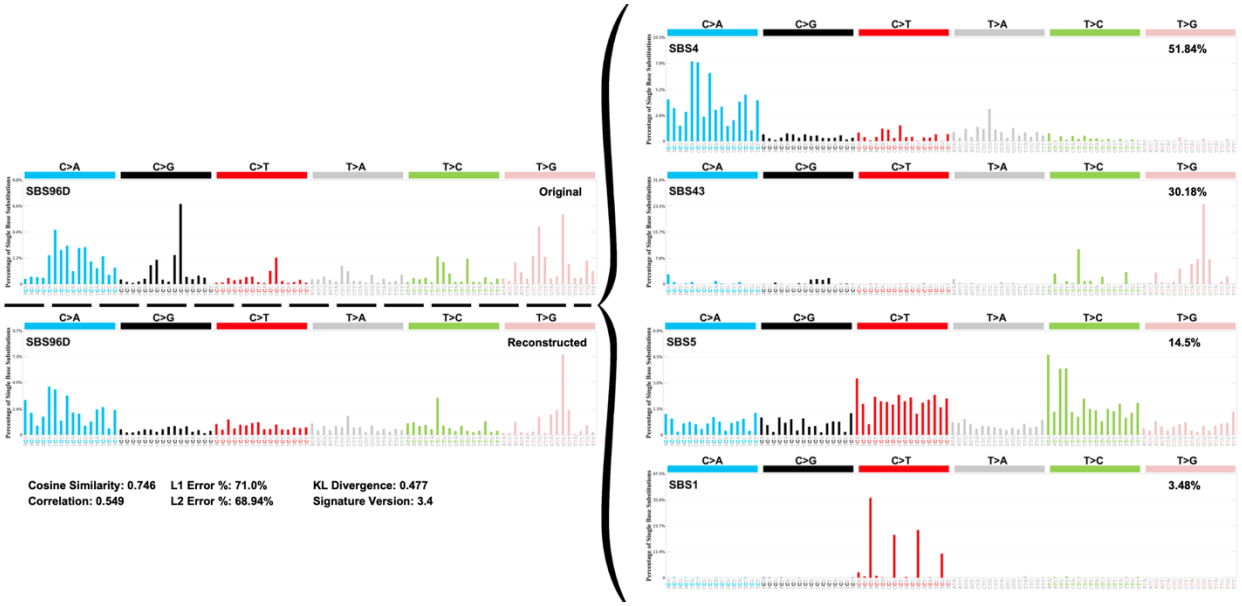


Fig. 9 | Mutation load across developmental time. The mutation burden across developmental time in soma and germline cells. The x-axis is the days of pregnancy, representing developmental time, and y-axis is the mutation counts normalized by sample coverage.

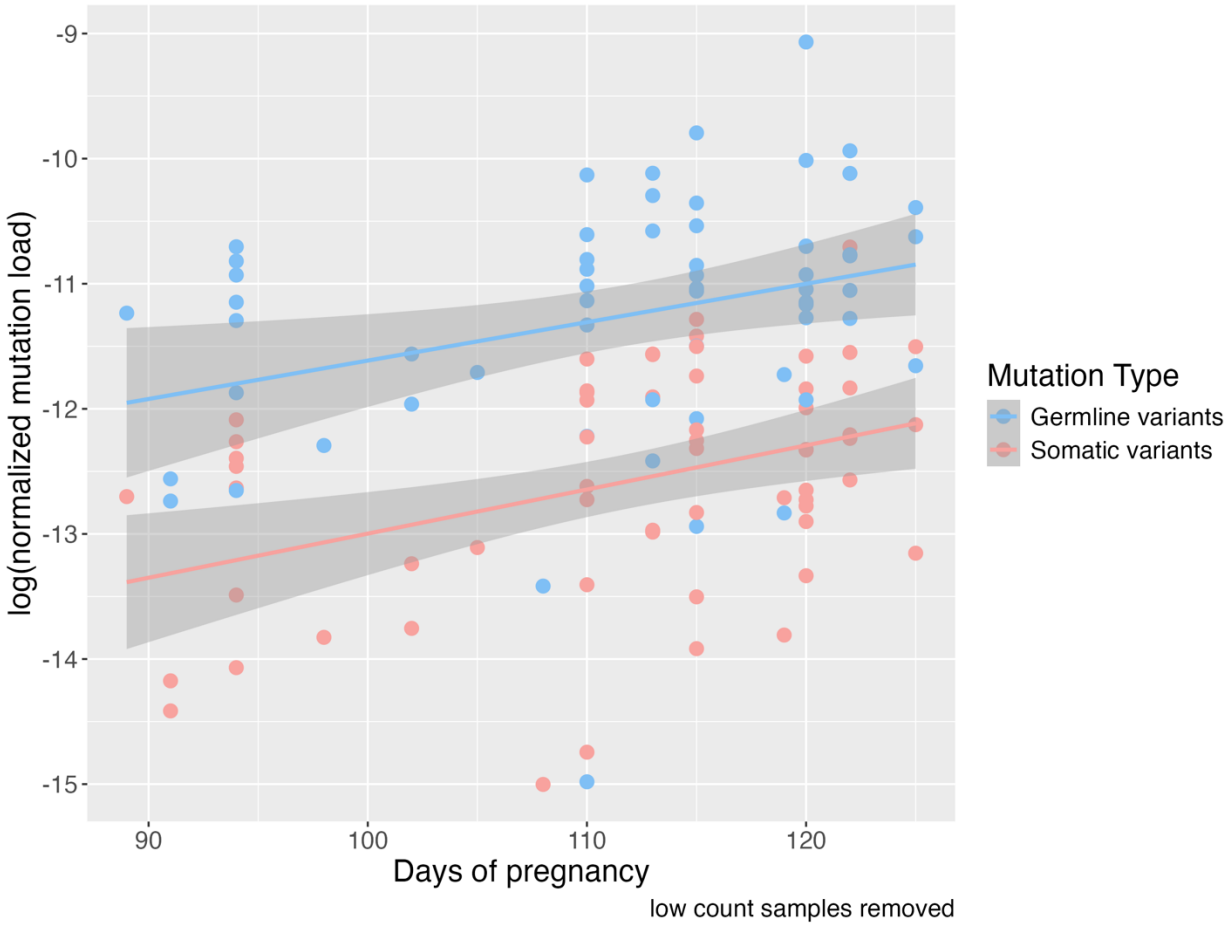
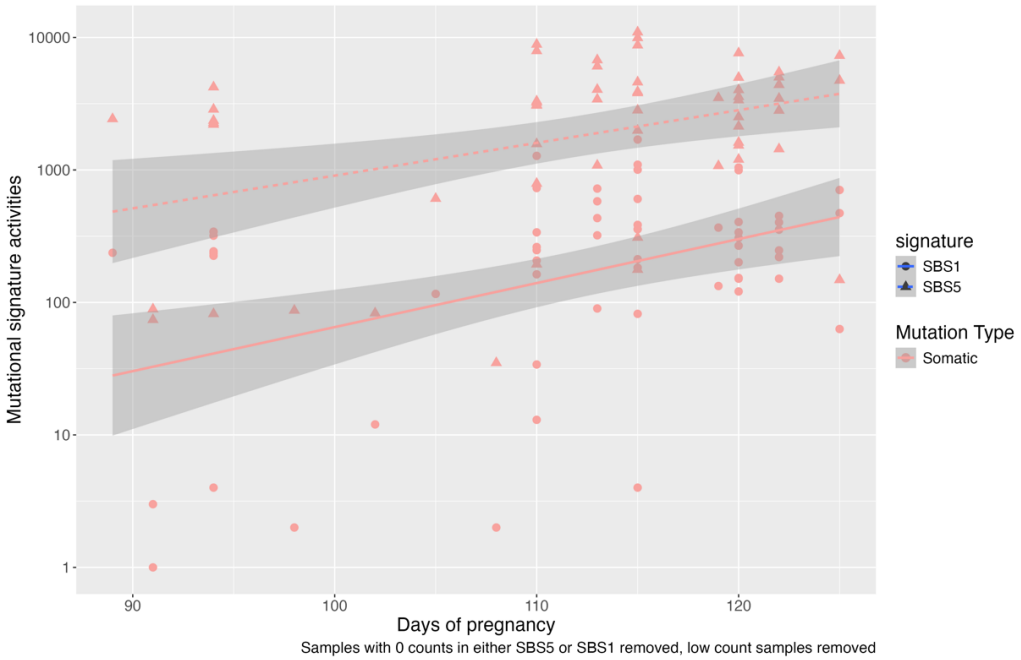


Fig. 10 | Activity of signature SBS1 and SBS5 across developmental time in soma cells. a) Activity of signature SBS1 and SBS5 in somatic mutations across days of pregnancy. Linear regression lines show the linear relationship between signature activities and developmental time, along with confidence intervals. **b)** analysis as in **a** but stratified by tissues

10a.



10b.

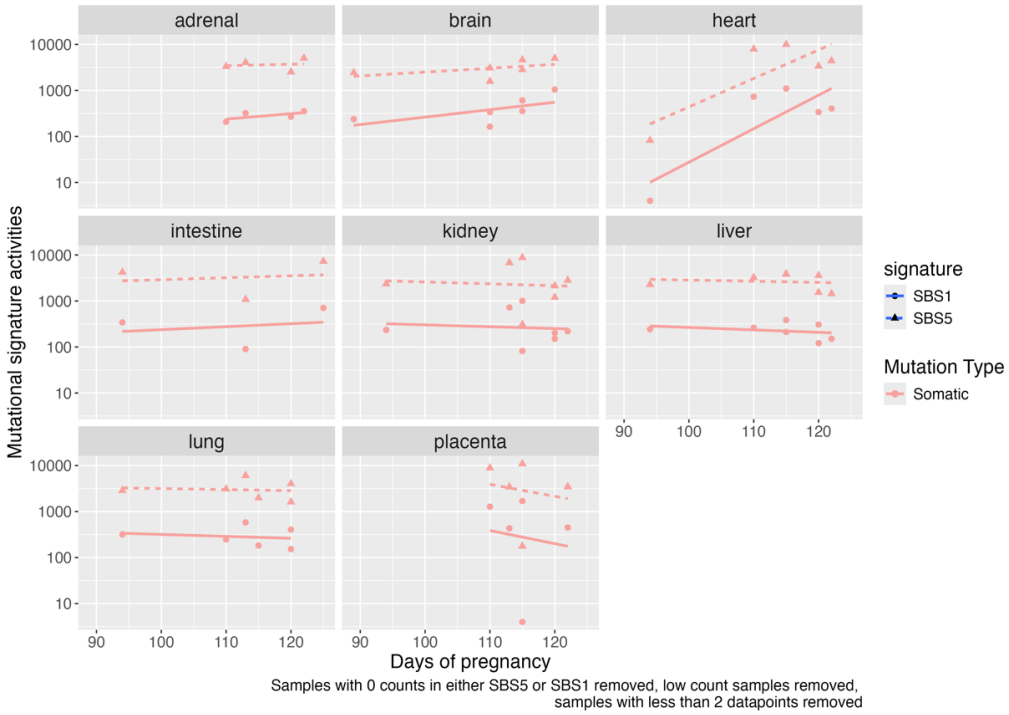


Fig. 11| Reconstructed phylogenetic tree representing cell lineage during embryonic development using the extracted mutations. The phylogenetic tree reconstructed from somatic mutations for donor H27431 (left). Each branch is color-labeled by tissue's corresponding germ layer. The same analysis but for germline mutations (right). Additional phylogenetic trees for all the donors can be found in the supplementary Fig. 5.

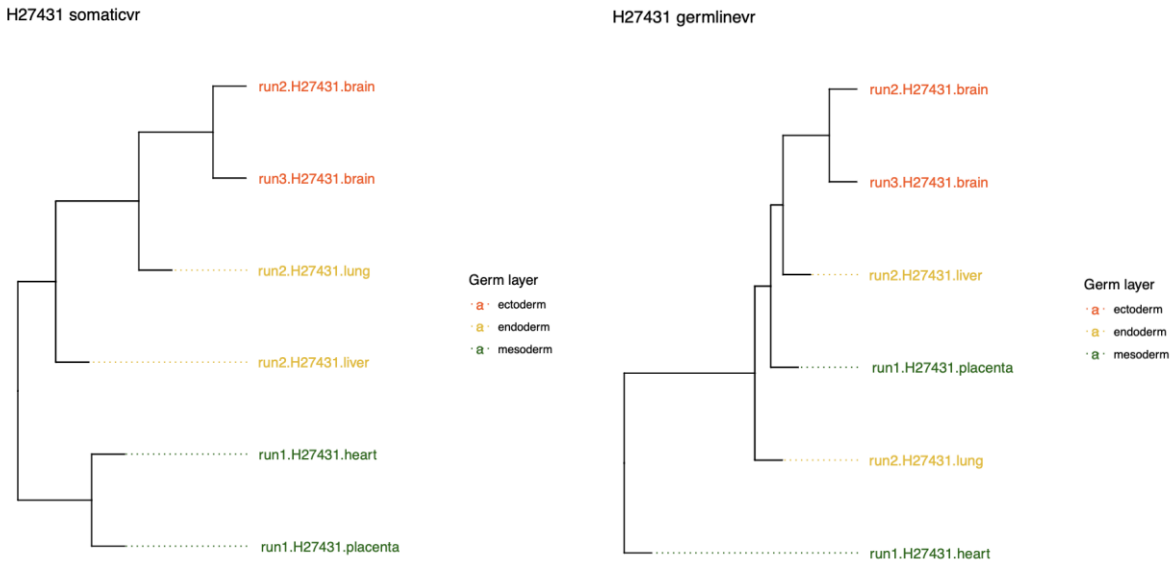
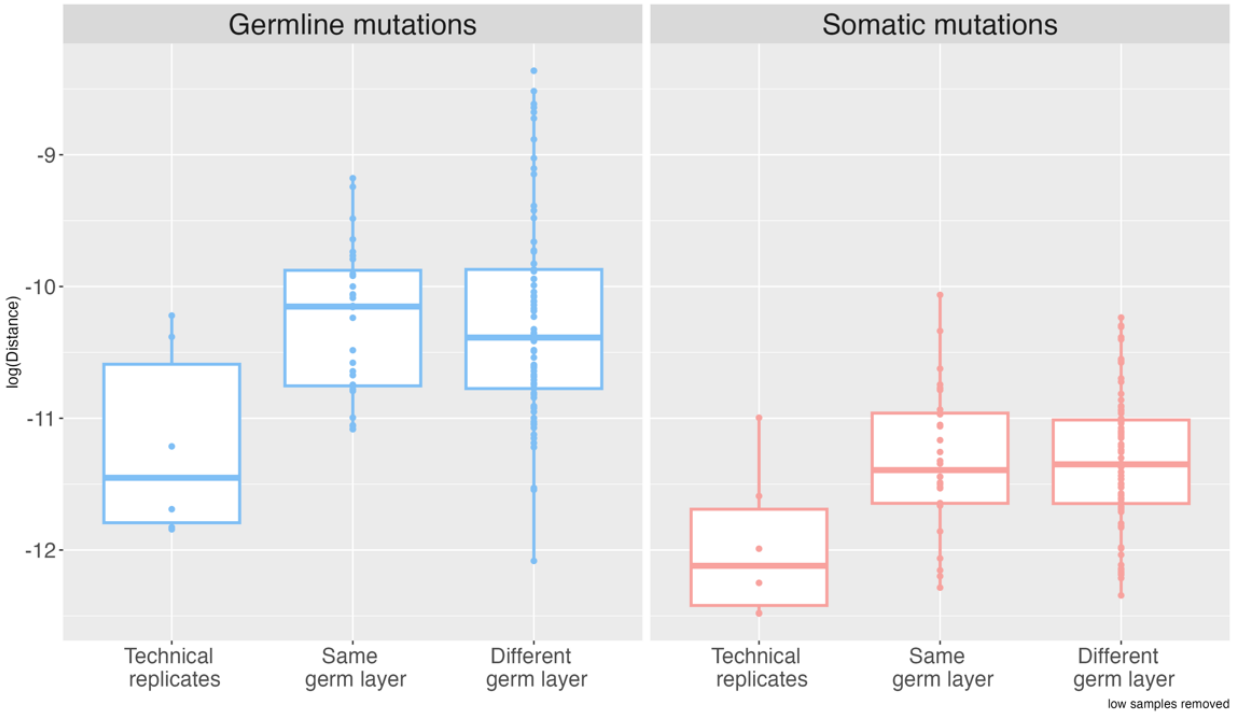


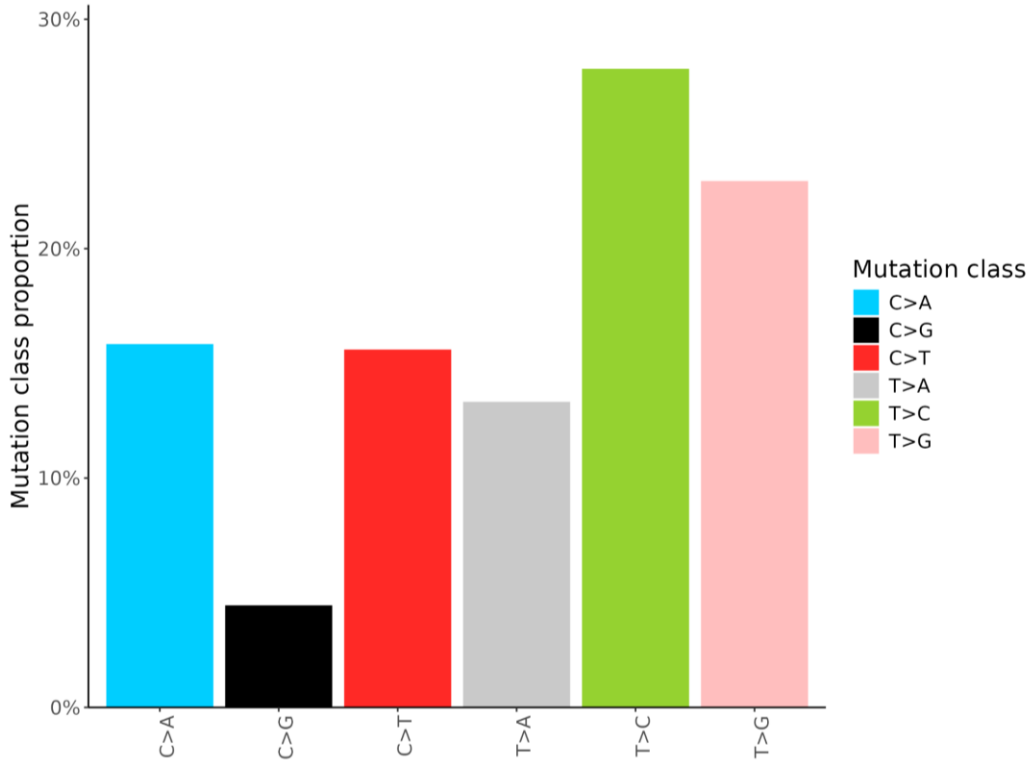
Fig. 12 | Distance for each tissue pair based on their category. Each data point, representing the distance of each tissue pair, was categorized into either technical replicate, same germ layer, or different germ layers (See distance calculation in Method). The line indicates the median. Same germ layer represents the tissues in the tissue pair are derived from the same germ layer and different germ layers represent tissues are derived from different germ layers.



SUPPLEMENTARY MATERIALS

Fig 1. | 1-mer mutation spectra for AD1 and AD2 in fruit fly embryo dataset. a). For AD1. **b).** Analysis as in **a** but for AD2

1a.



1b.

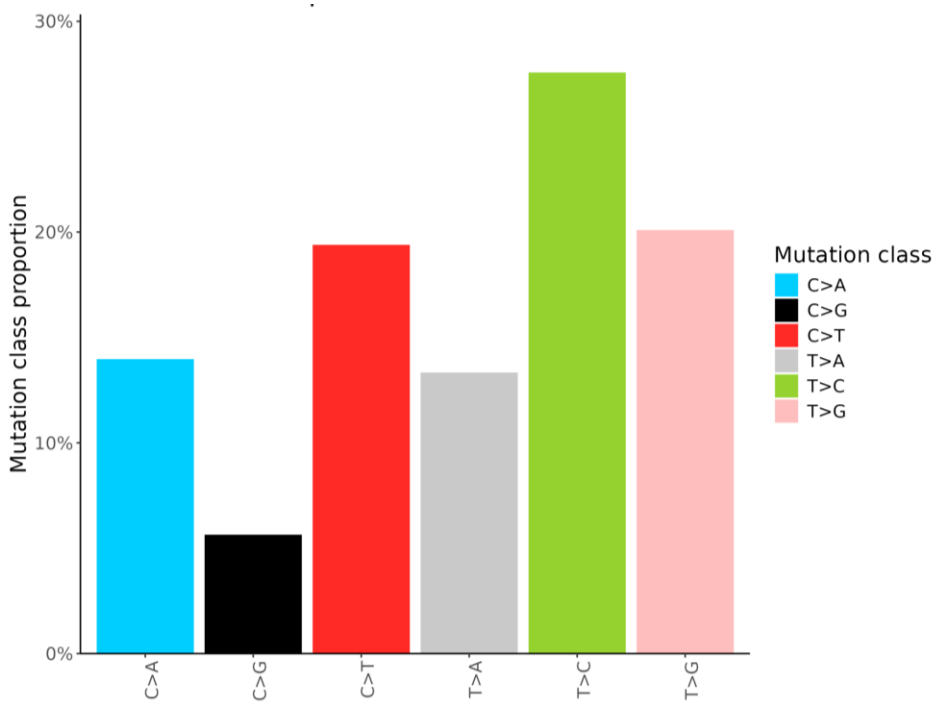
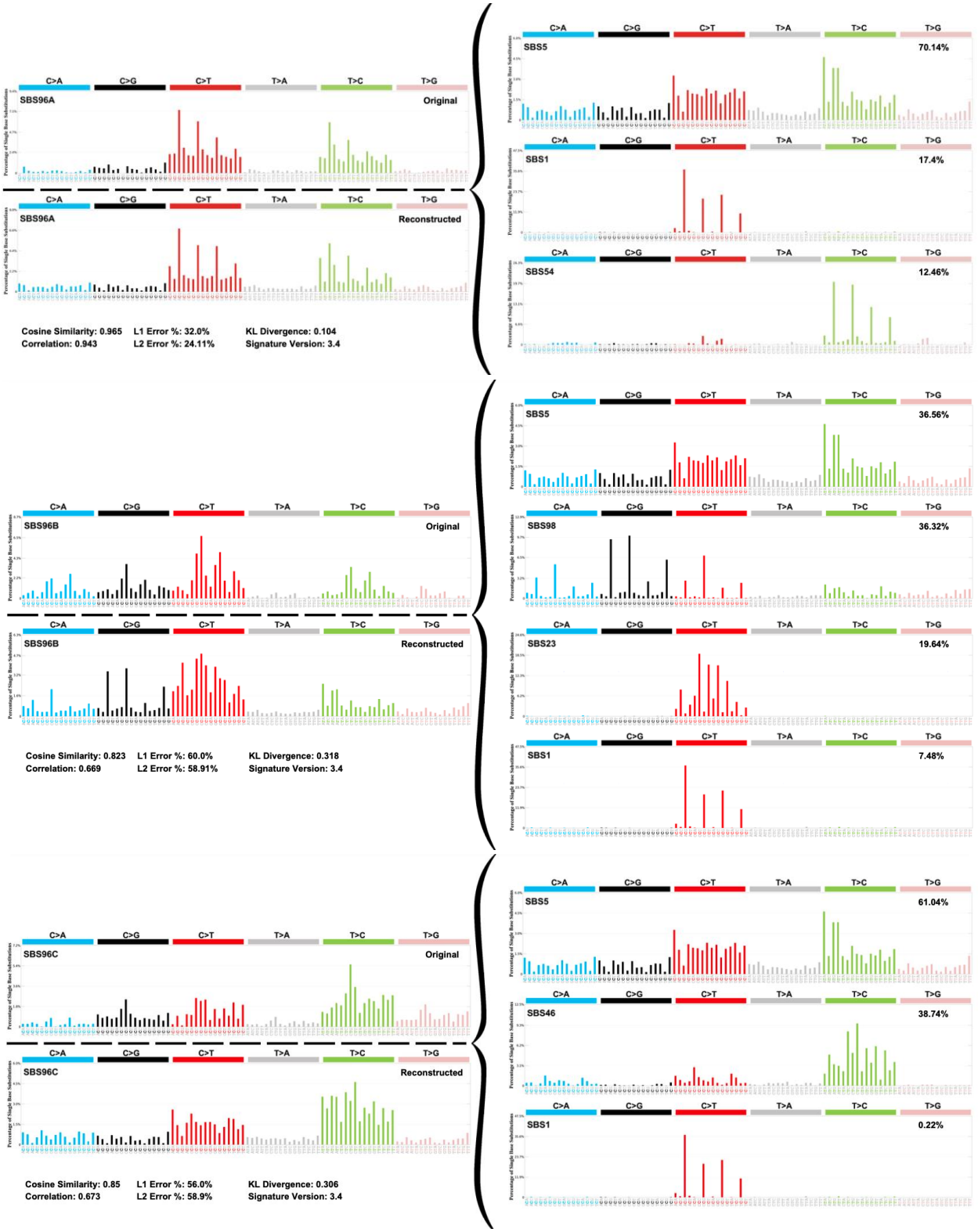
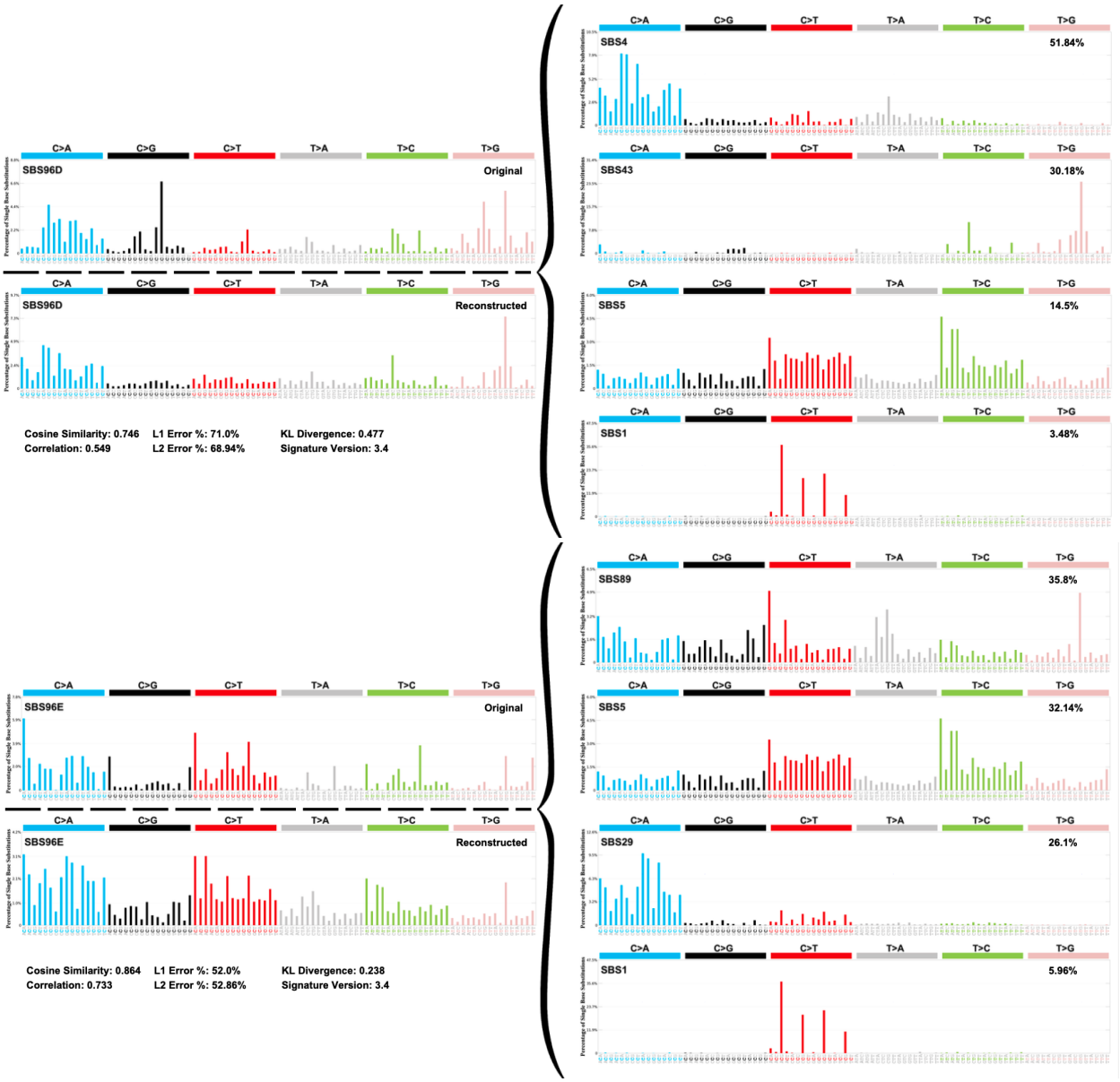


Fig. 2 | All of the mutational signatures extracted from Sigprofiler Extractor (SBSA-SBSG).




C>A C>G C>T T>A T>C T>G
SBS96D
Original
SBS96D
Reconstructed
Cosine Similarity: 0.746
Correlation: 0.549
L1 Error %: 71.0%
L2 Error %: 68.94%
KL Divergence: 0.477
Signature Version: 3.4
SBS4
51.84%
SBS43
30.18%
SBS5
14.5%
SBS1
3.48%
SBS96E
Original
SBS96E
Reconstructed
Cosine Similarity: 0.864
Correlation: 0.733
L1 Error %: 52.0%
L2 Error %: 52.86%
KL Divergence: 0.238
Signature Version: 3.4
SBS89
35.8%
SBS5
32.14%
SBS29
26.1%
SBS1
5.96%

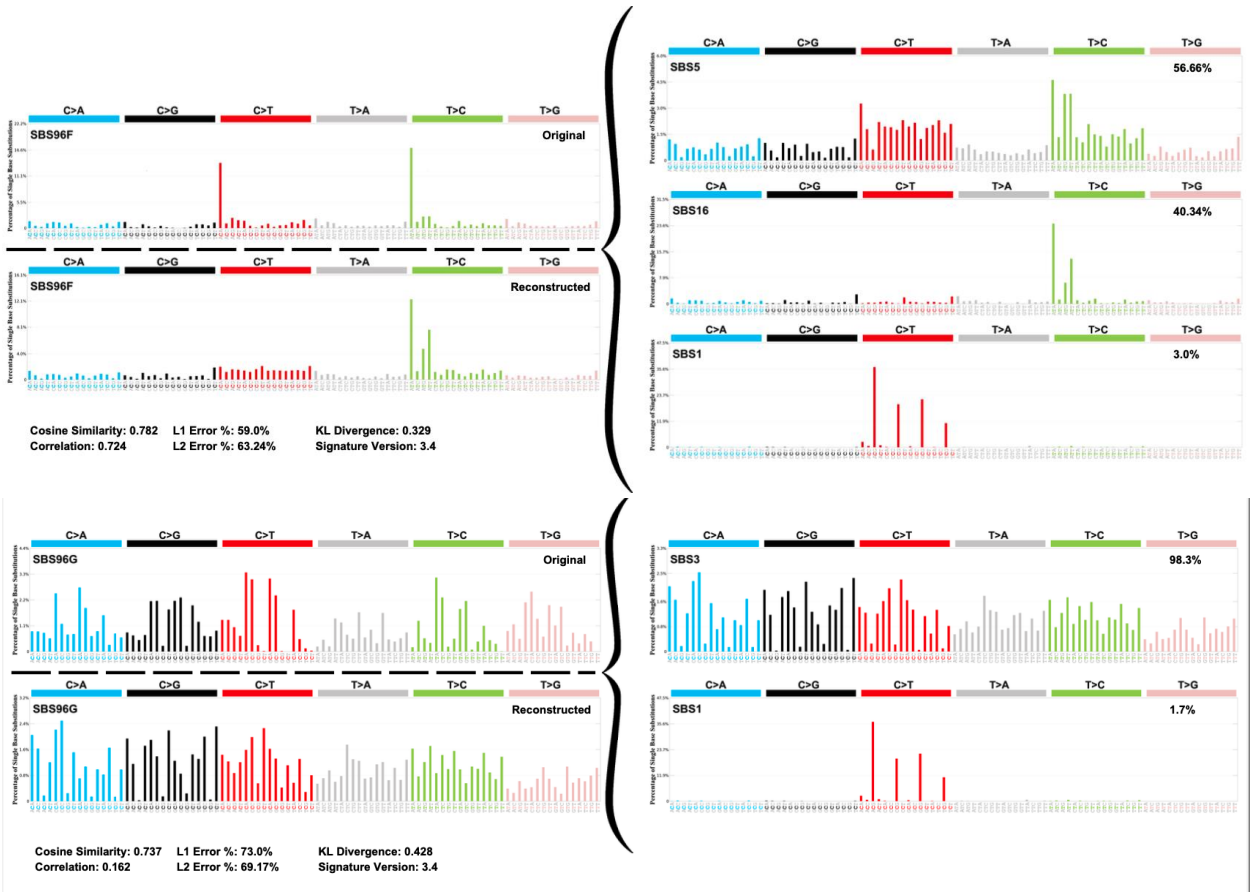
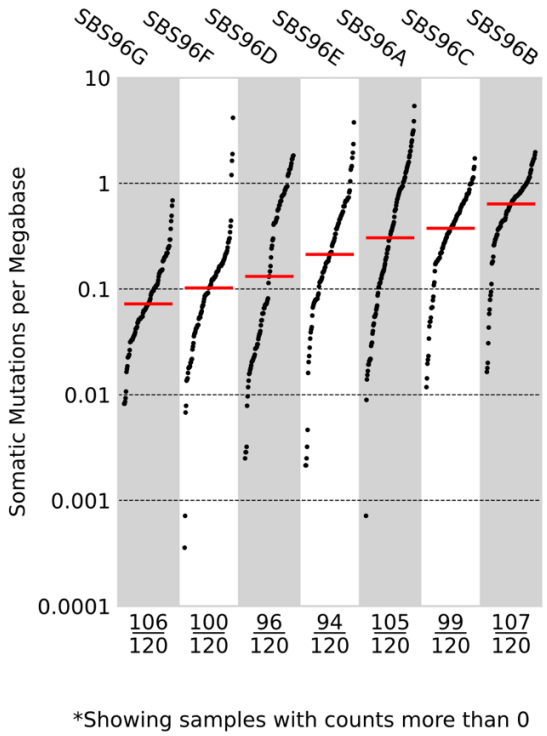

C>A
C>G
C>T
T>A
T>C
T>G
SBS96F
Original
SBS96F
Reconstructed
Cosine Similarity: 0.782
L1 Error %: 59.0%
KL Divergence: 0.329
Correlation: 0.724
L2 Error %: 63.24%
Signature Version: 3.4
SBS5
56.66%
SBS16
40.34%
SBS1
3.0%
SBS96G
Original
SBS96G
Reconstructed
Cosine Similarity: 0.737
L1 Error %: 73.0%
KL Divergence: 0.428
Correlation: 0.162
L2 Error %: 69.17%
Signature Version: 3.4
SBS3
98.3%
SBS1
1.7%

Fig 3. | Mutational burden for extracted mutational signature from SigProfiler Extractor. **a)** The x-axis is the number of samples plotted over total number of samples; y-axis is the somatic mutations per megabase. Each column represents the decomposed *de novo* mutational signature. **b)** analysis as in **a** for COSMIC mutational signatures. The plot is ordered by the mean somatic mutations per megabase.

3a.



3b.

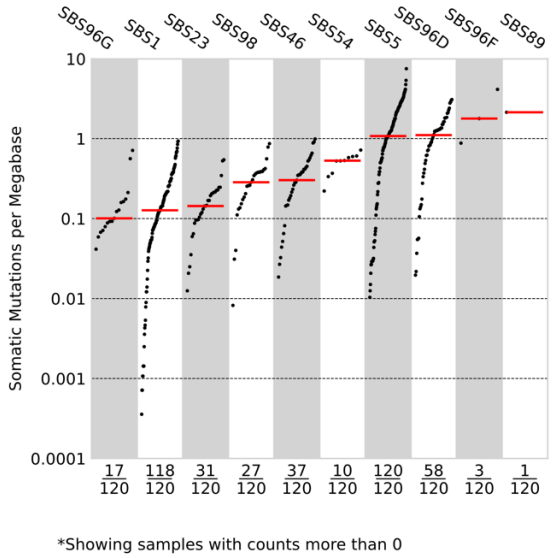
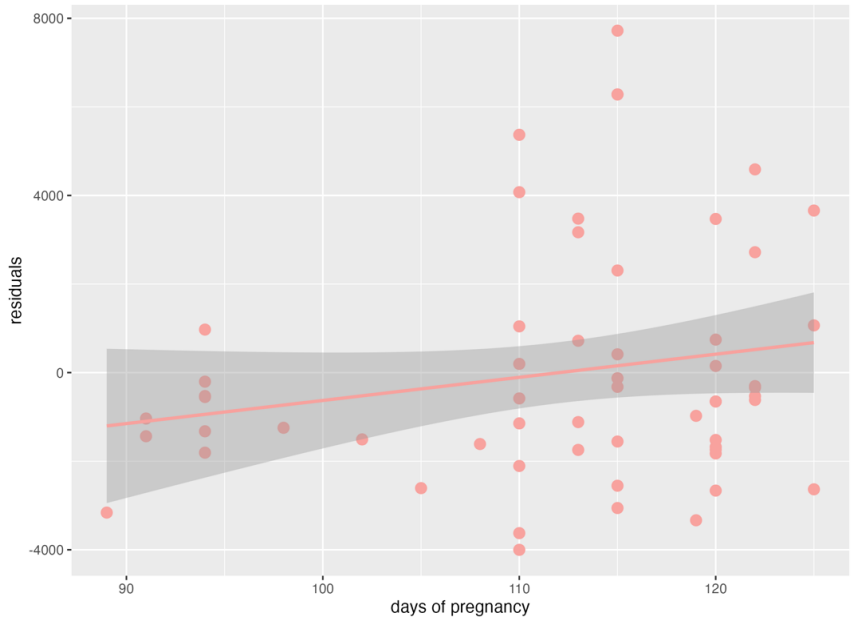


Fig 4. | Residuals plot for the fitted model of mutation counts ~ coverage against days of pregnancy. a) For somatic variants, the adjusted r-squared value is 0.2177 and the p-value for the correlation is 0.1377 **b)** for germline variants, the adjusted r-squared value is 0.01462 and the p-value for correlation is 0.17.

4a.



4b.

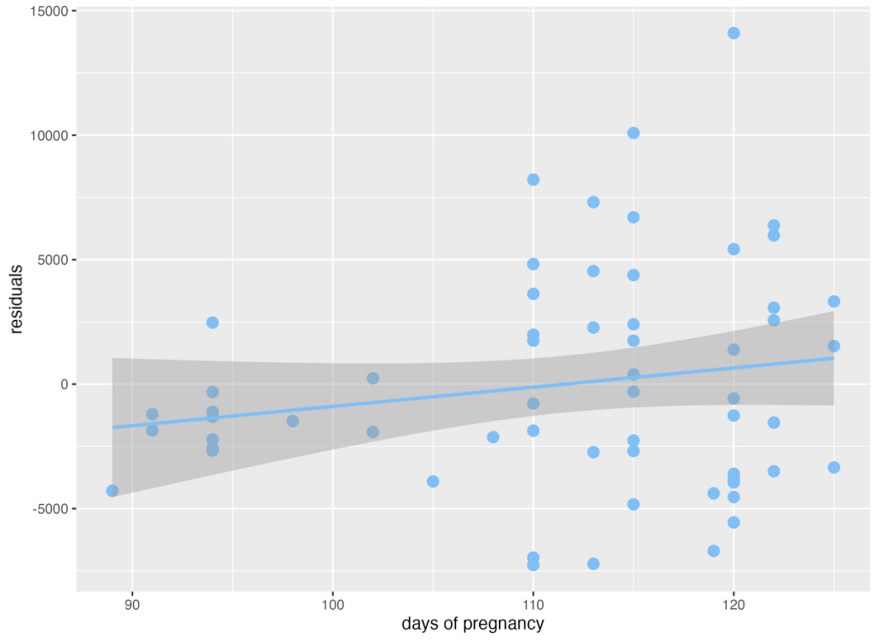
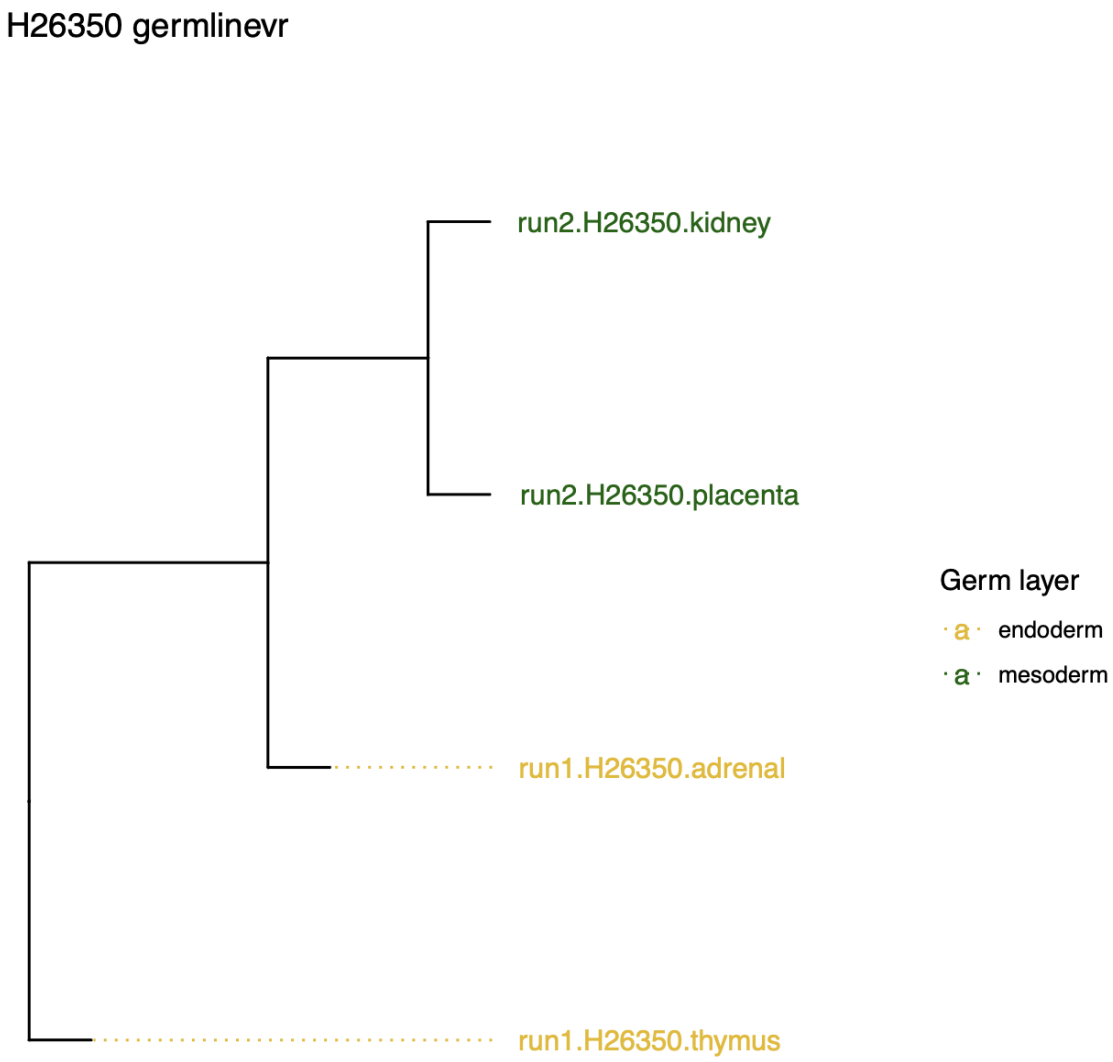
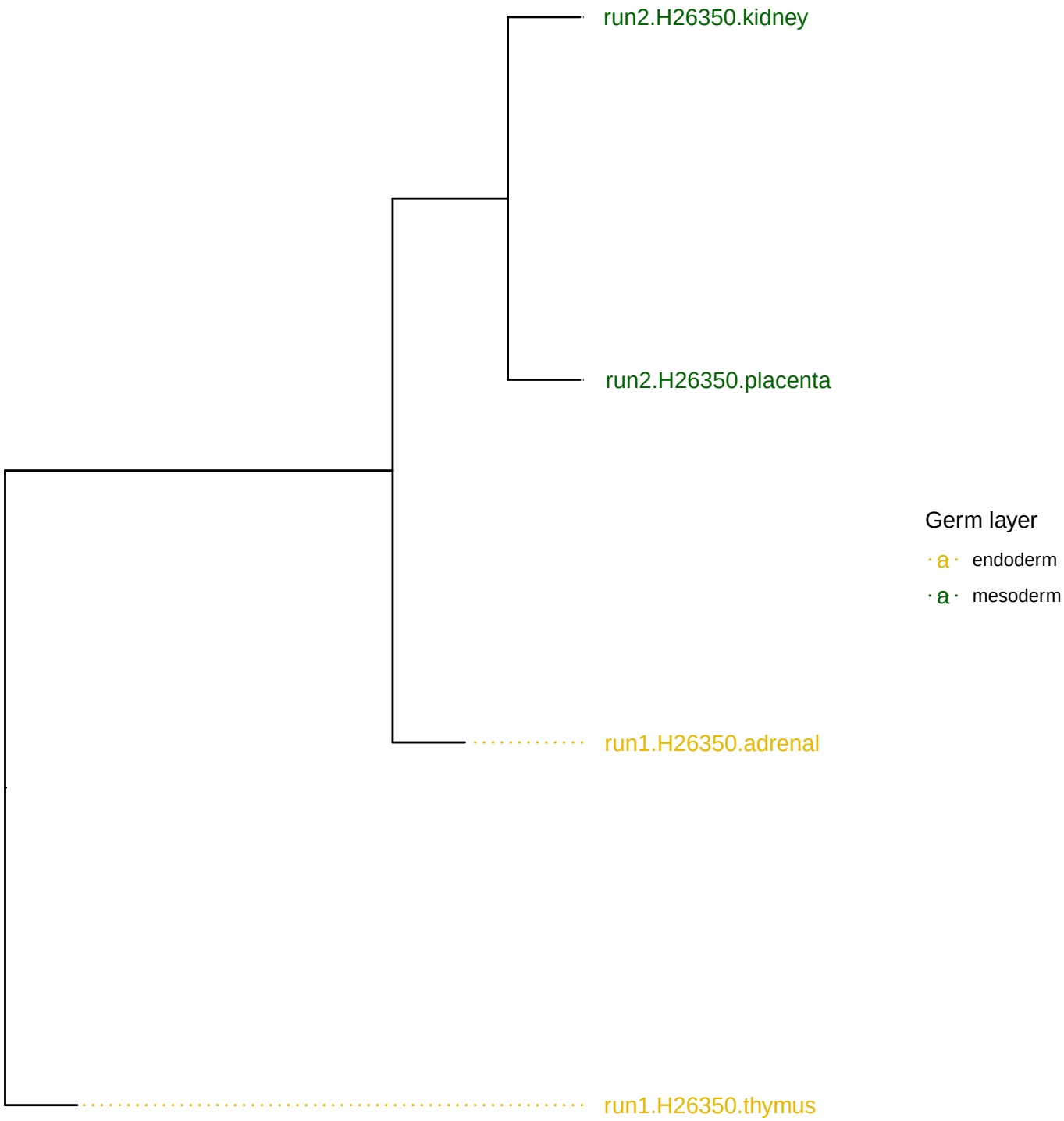


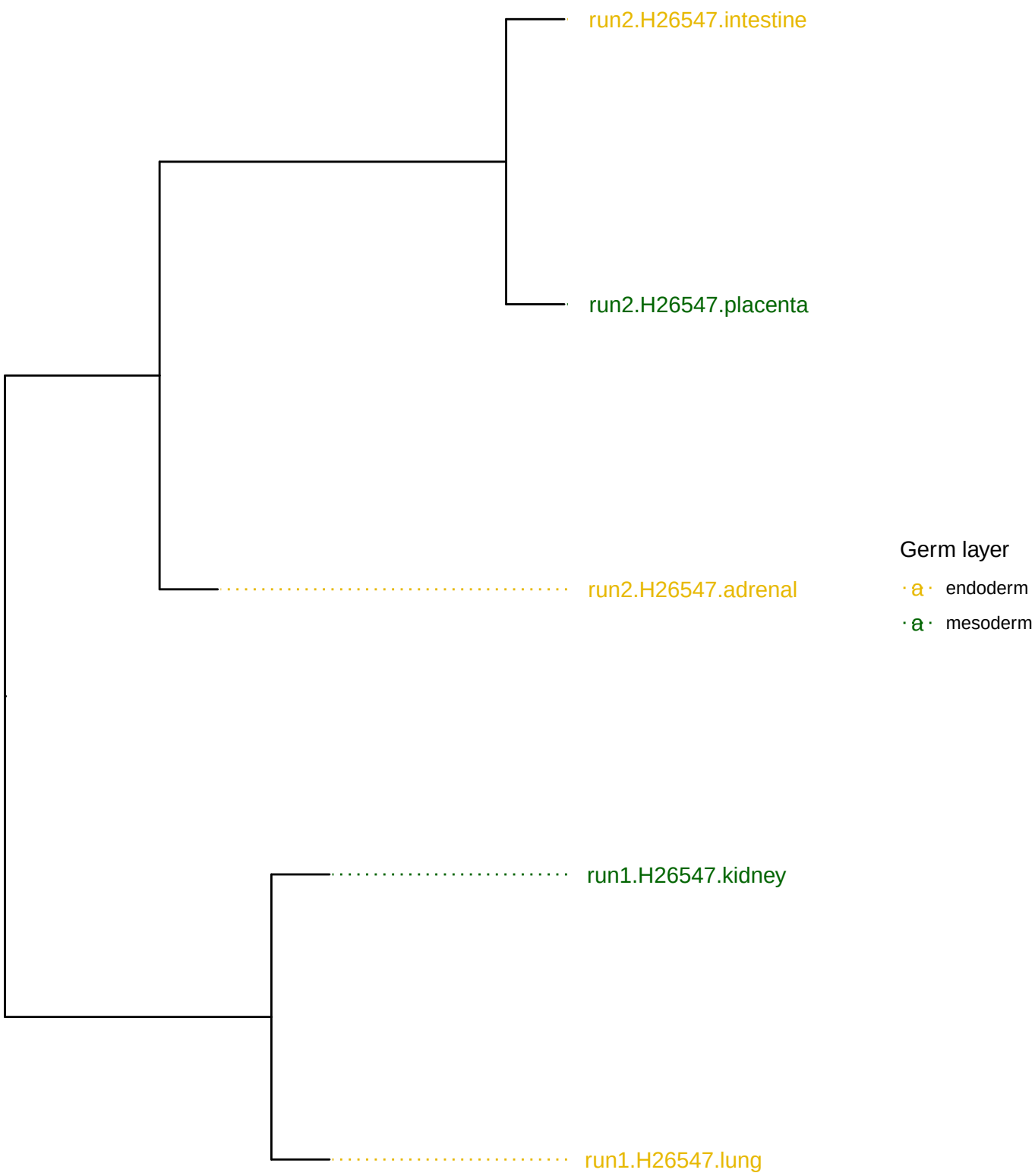
Fig 5. | Reconstructed phylogenetic tree representing cell lineage during embryonic development using the extracted mutations for all the donors.



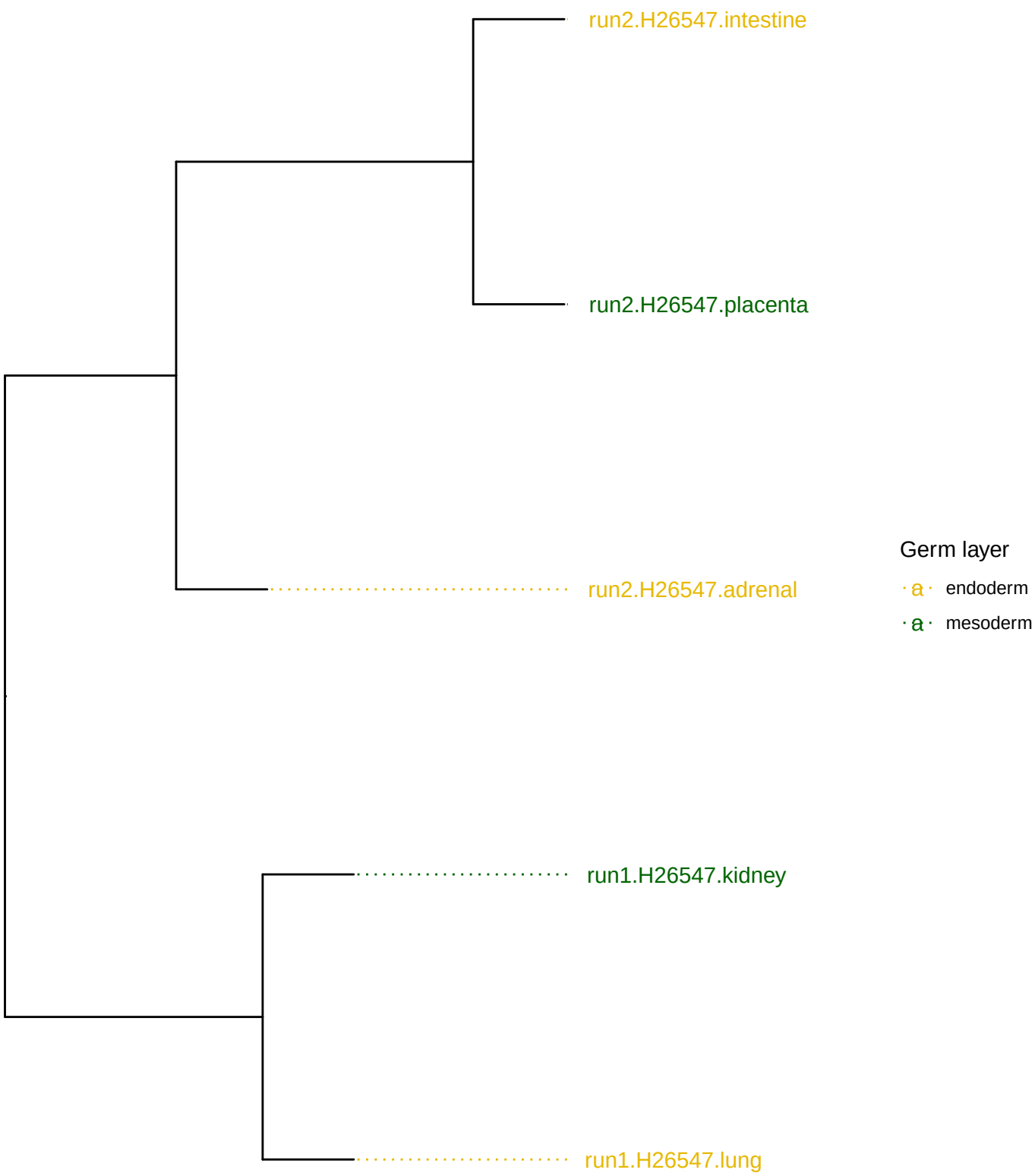
H26350 somaticvr



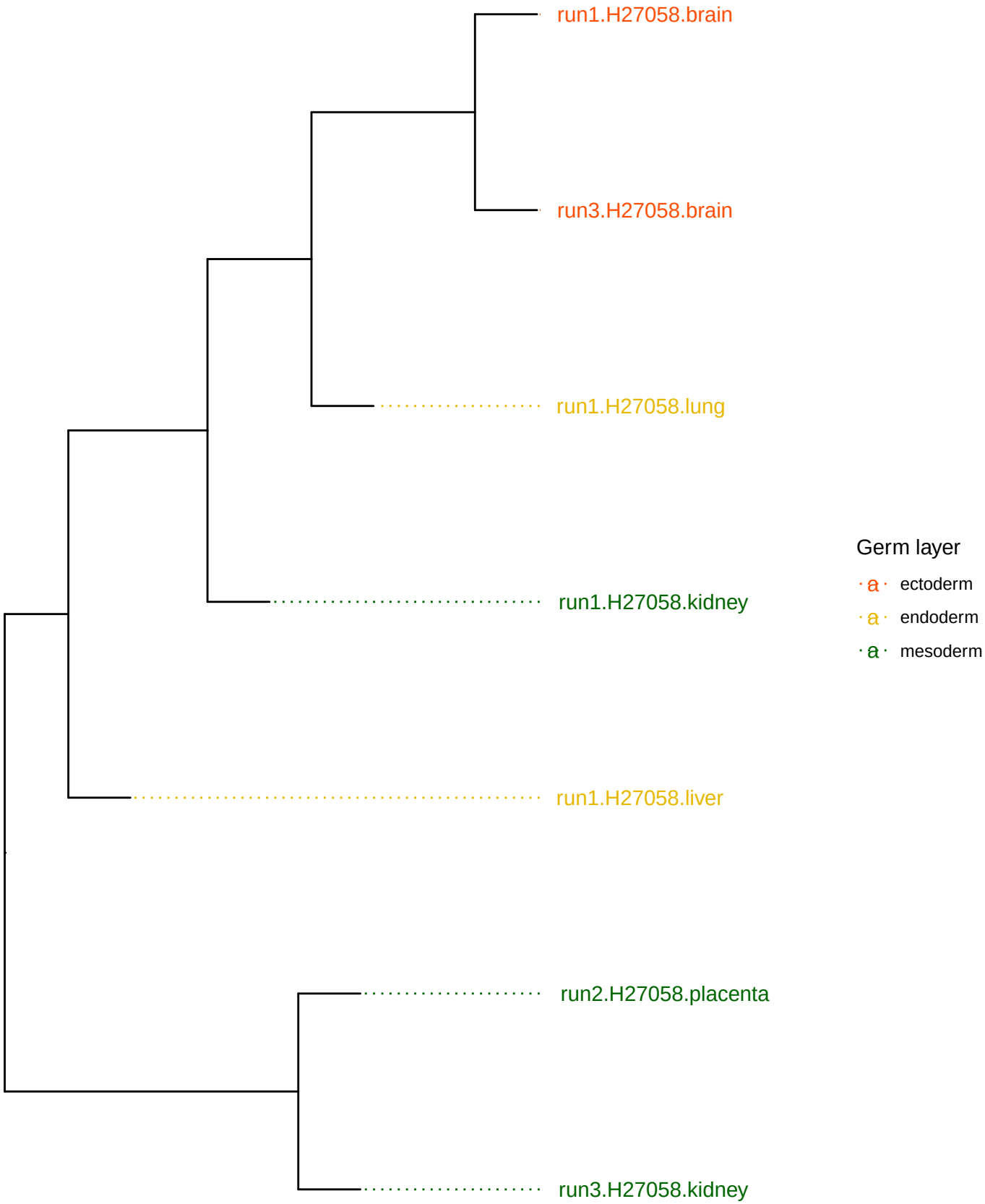
H26547 germlinevr



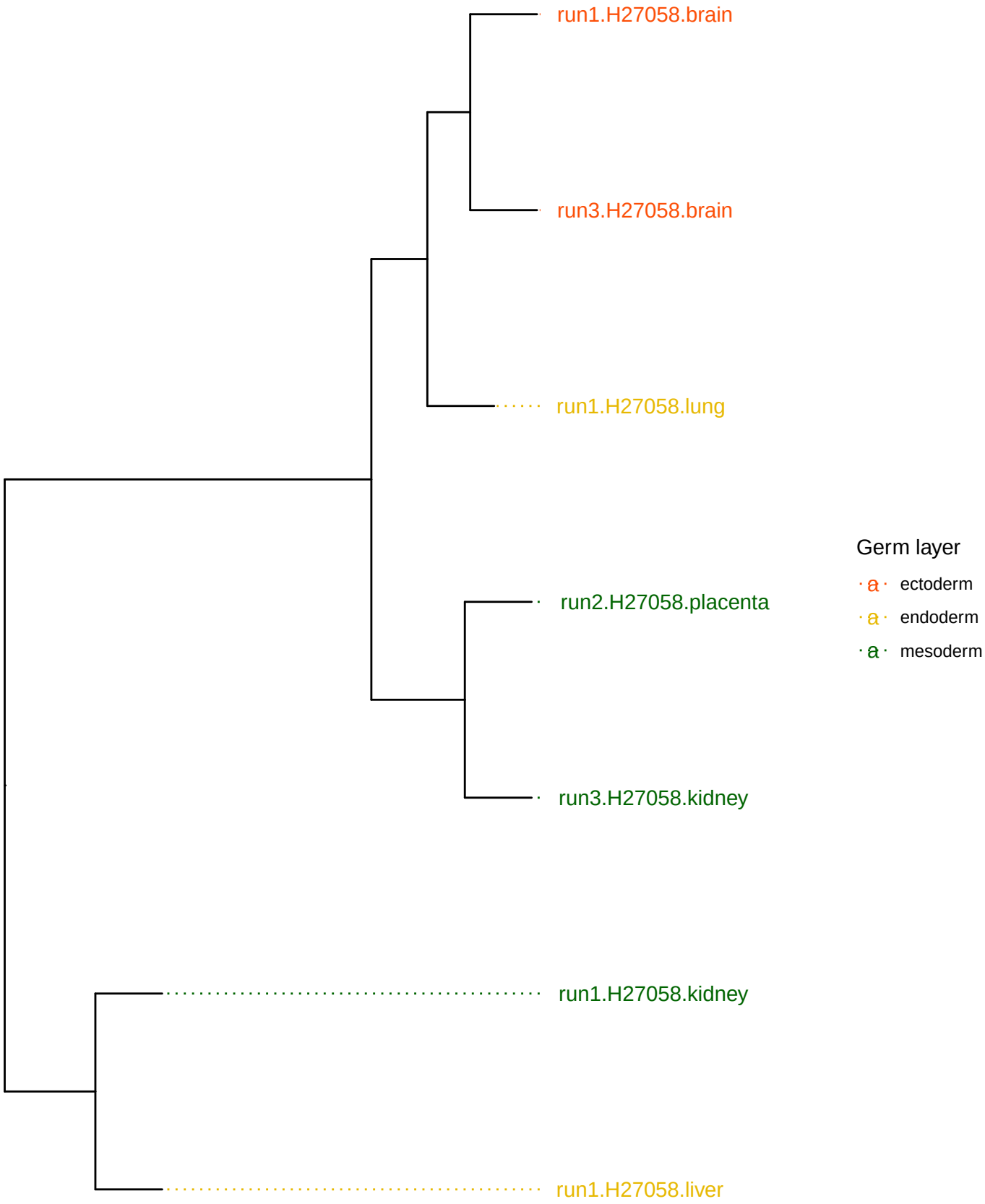
H26547 somaticvr



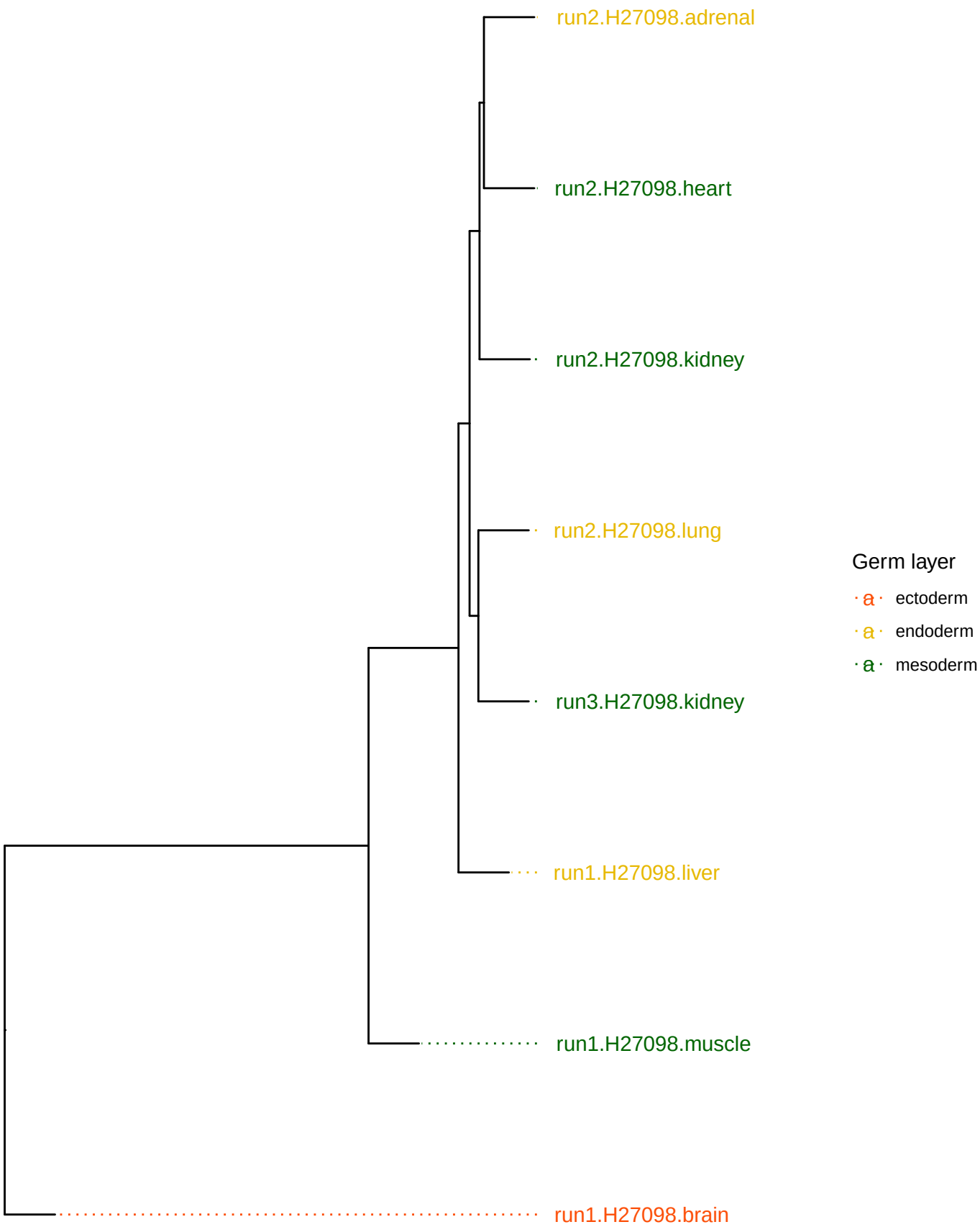
H27058 germlinevr



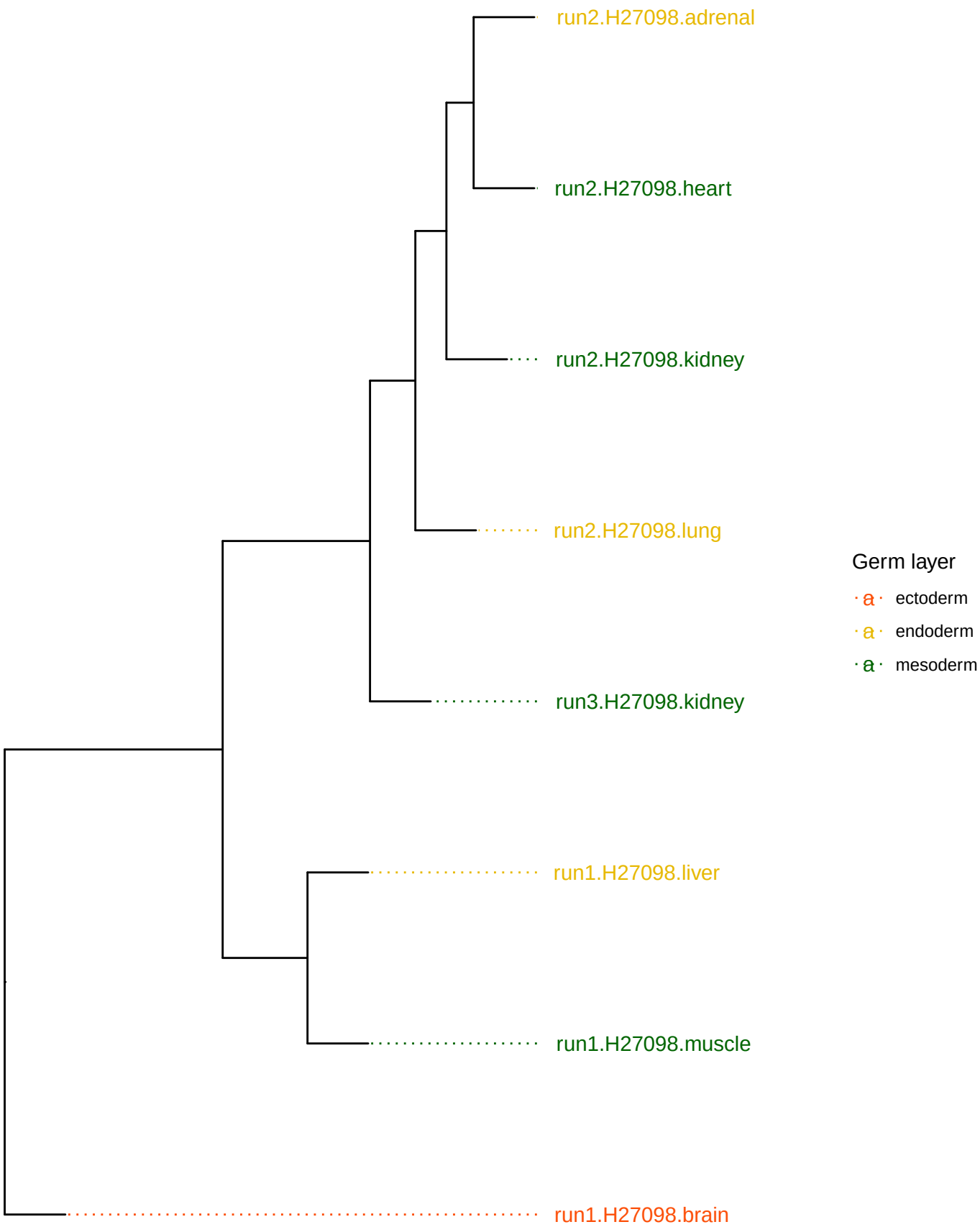
H27058 somaticvr



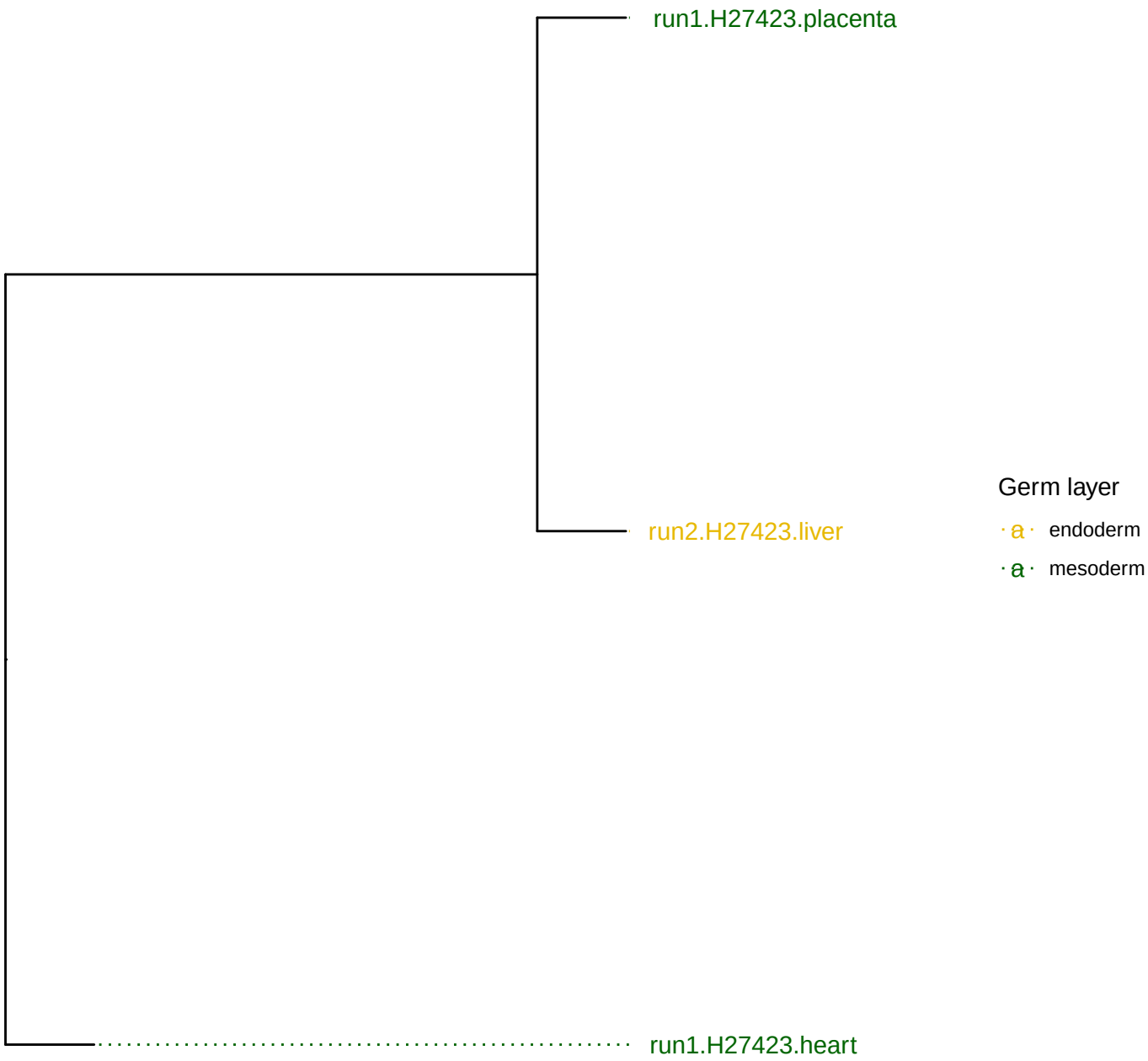
H27098 germlinevr



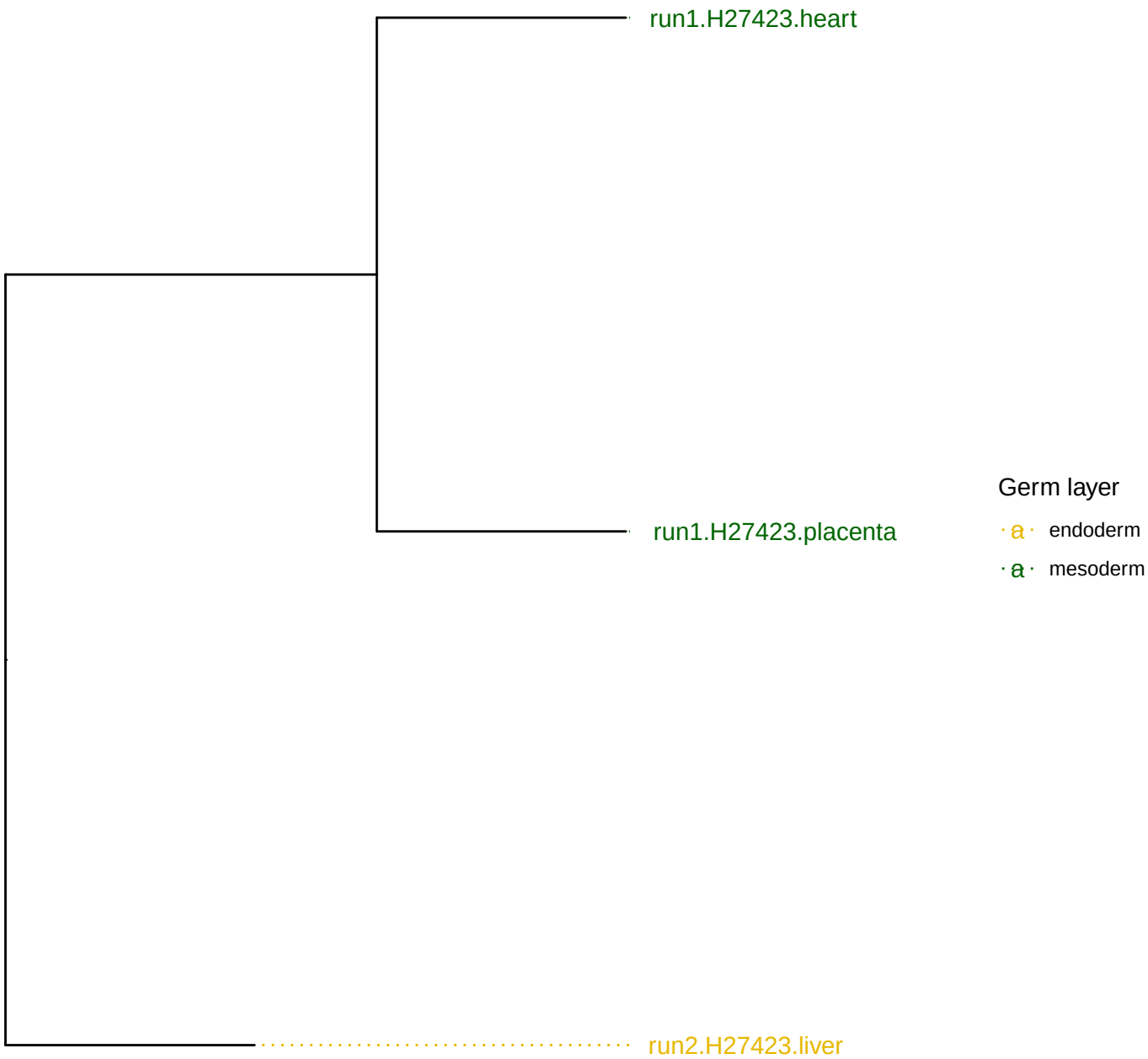
H27098 somaticvr



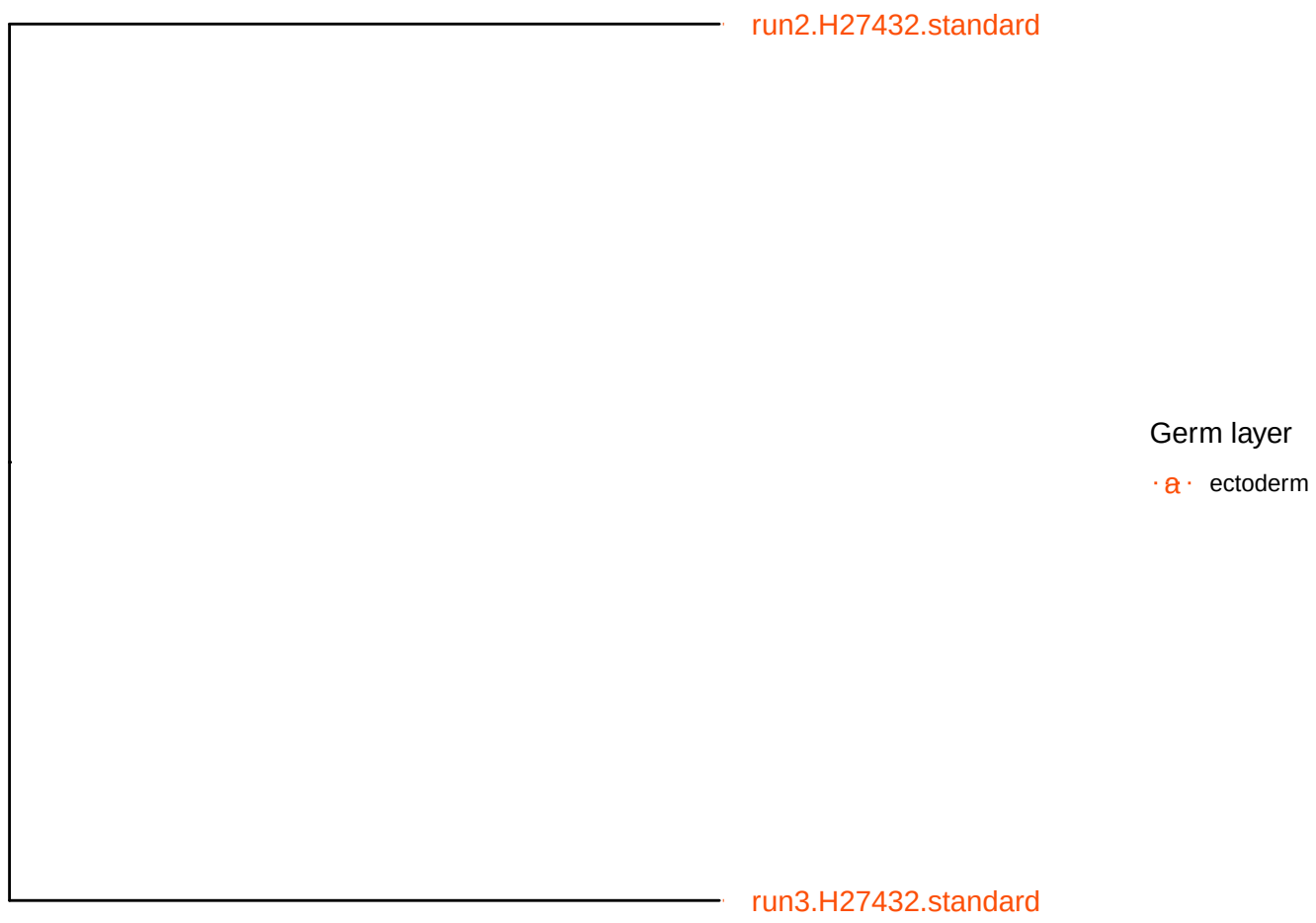
H27423 germlinevr



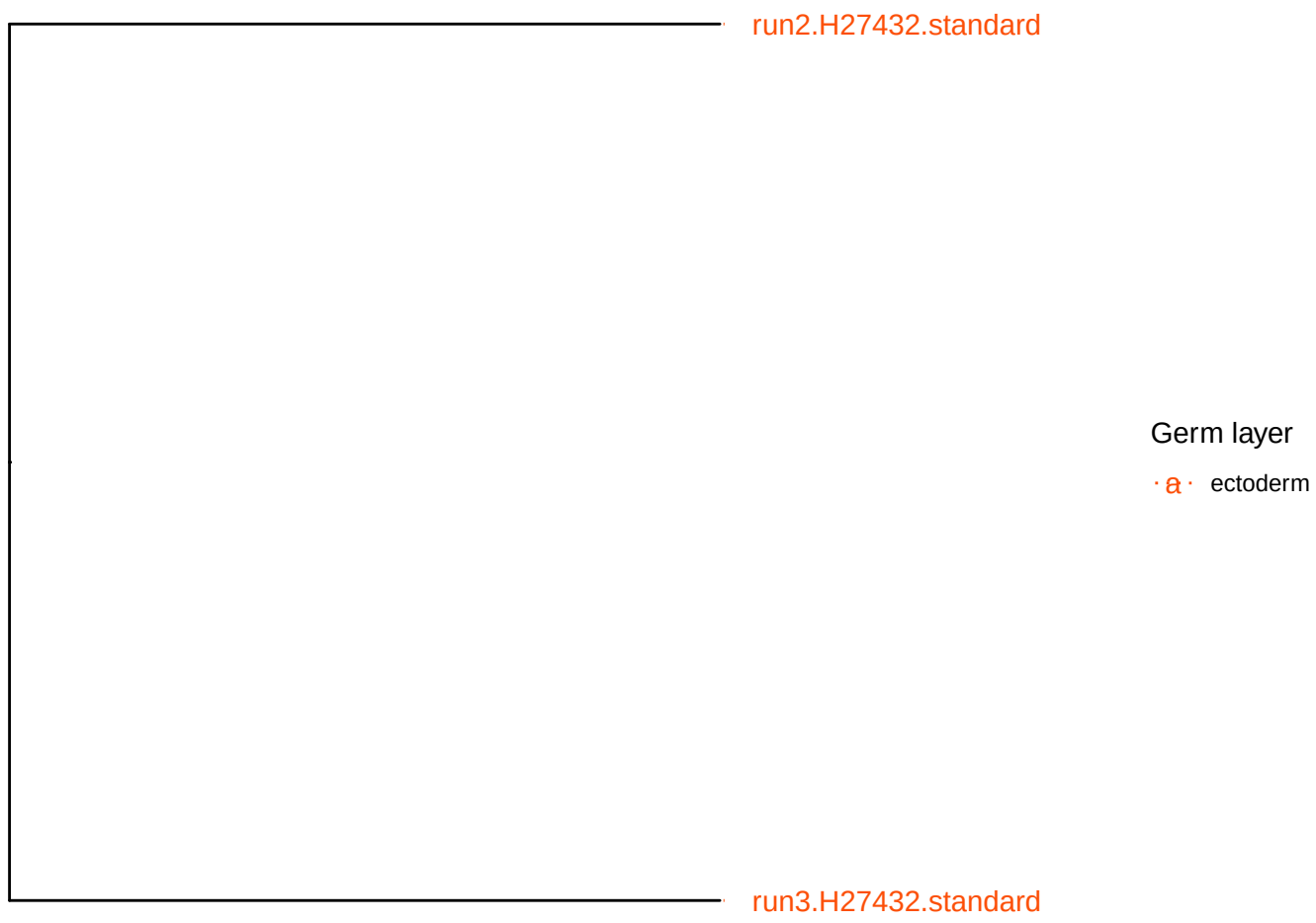
H27423 somaticvr



H27432 germlinevr



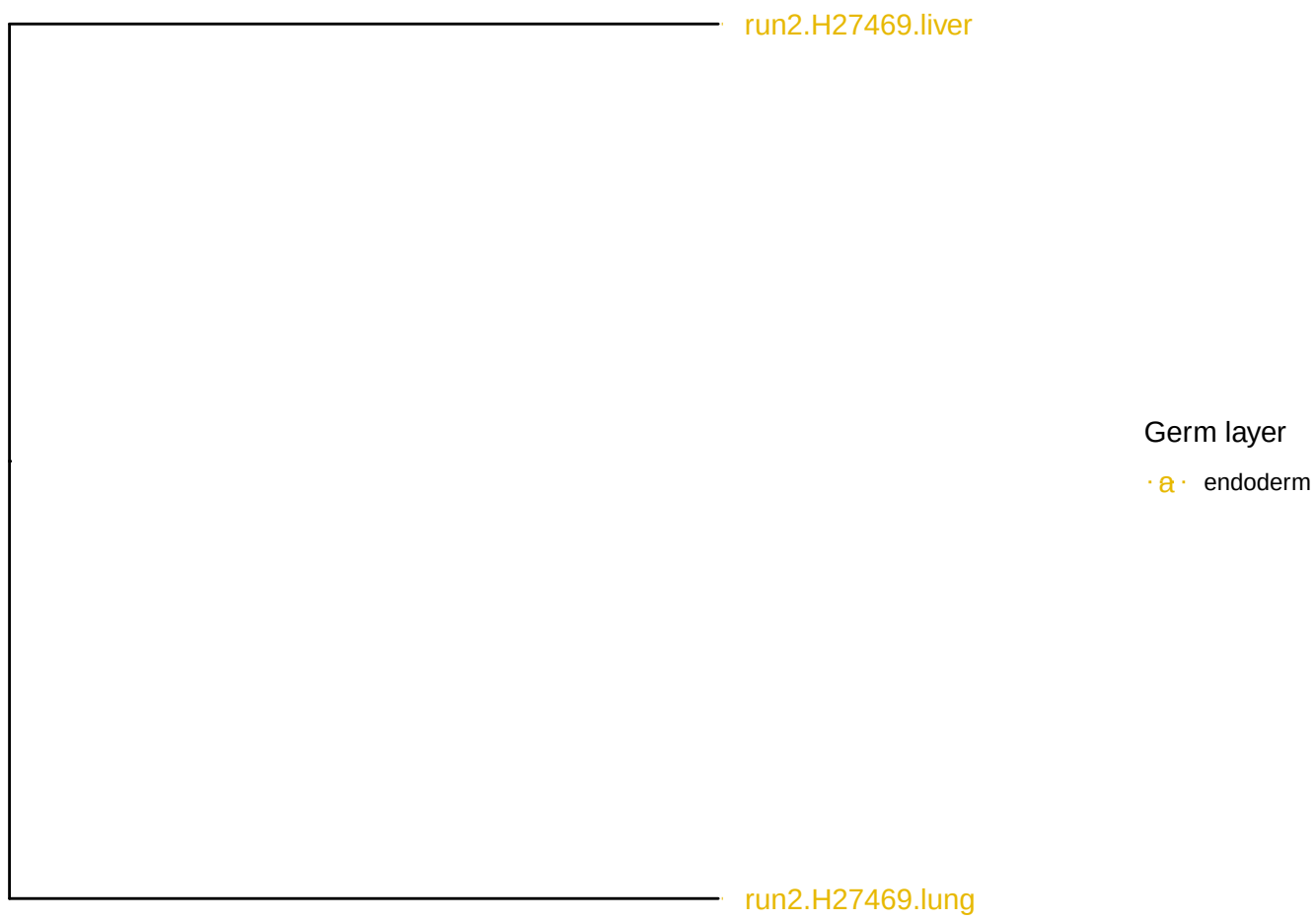
H27432 somaticvr



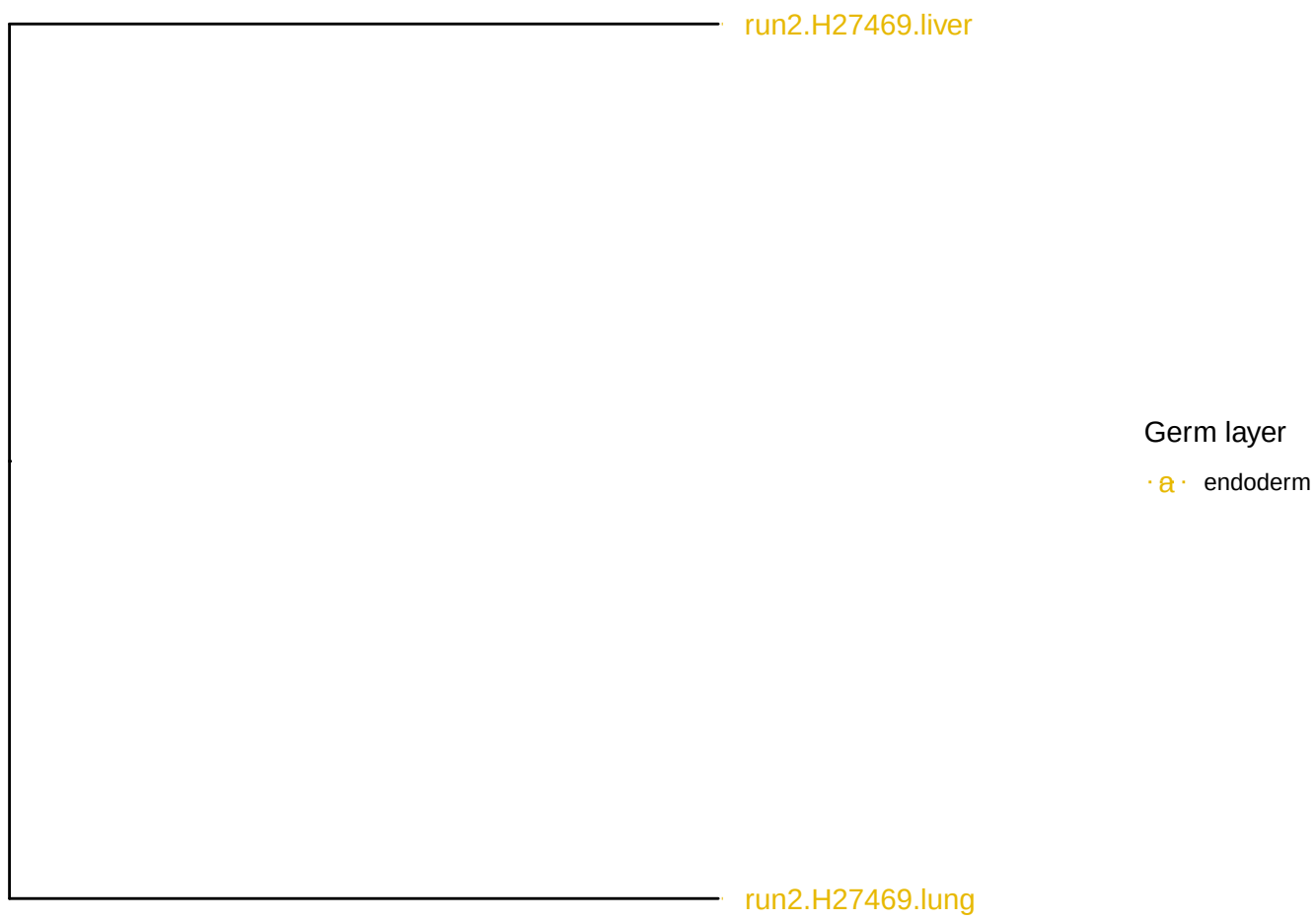
H27464 germlinevr

H27464 somaticvr

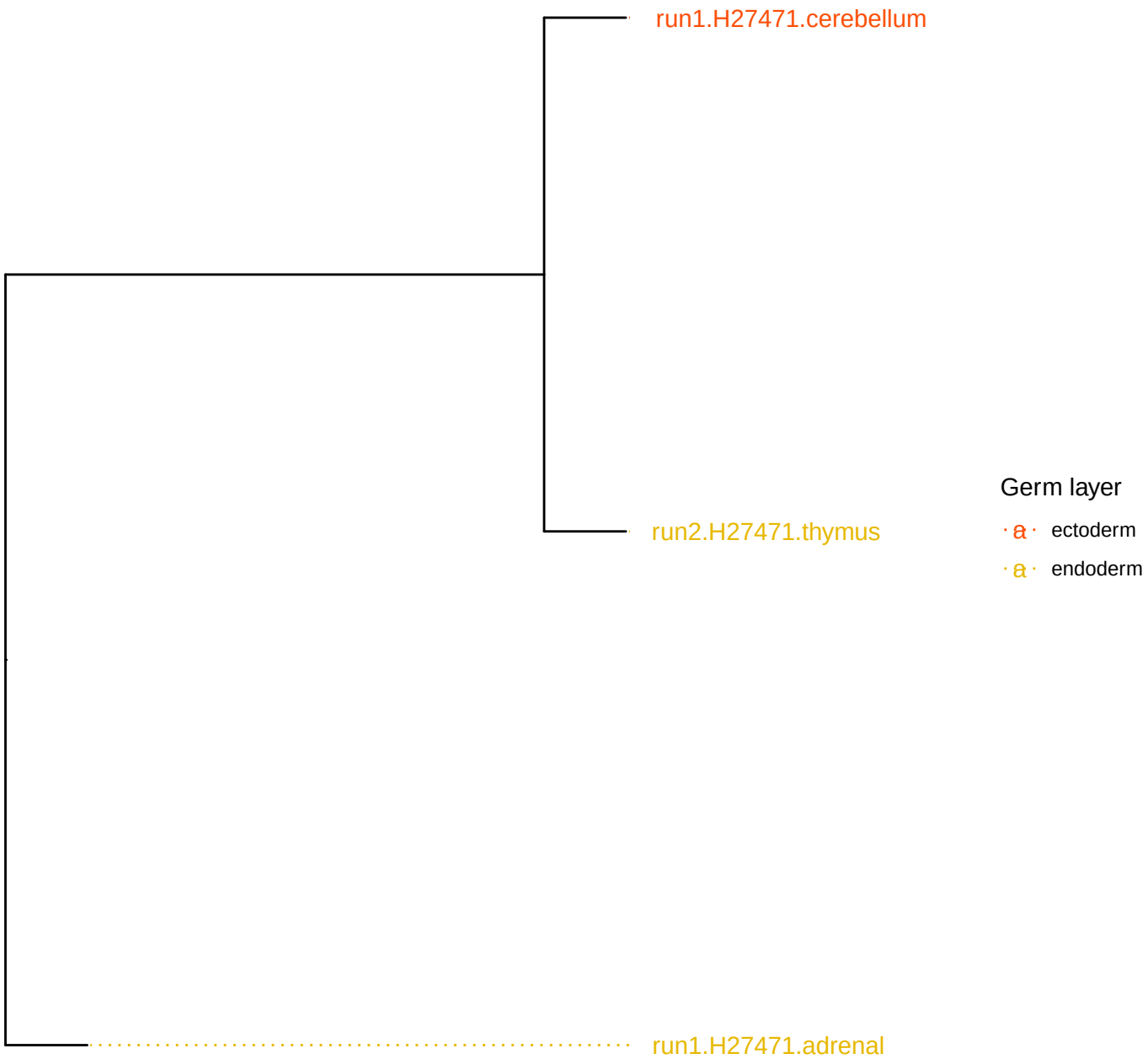
H27469 germlinevr



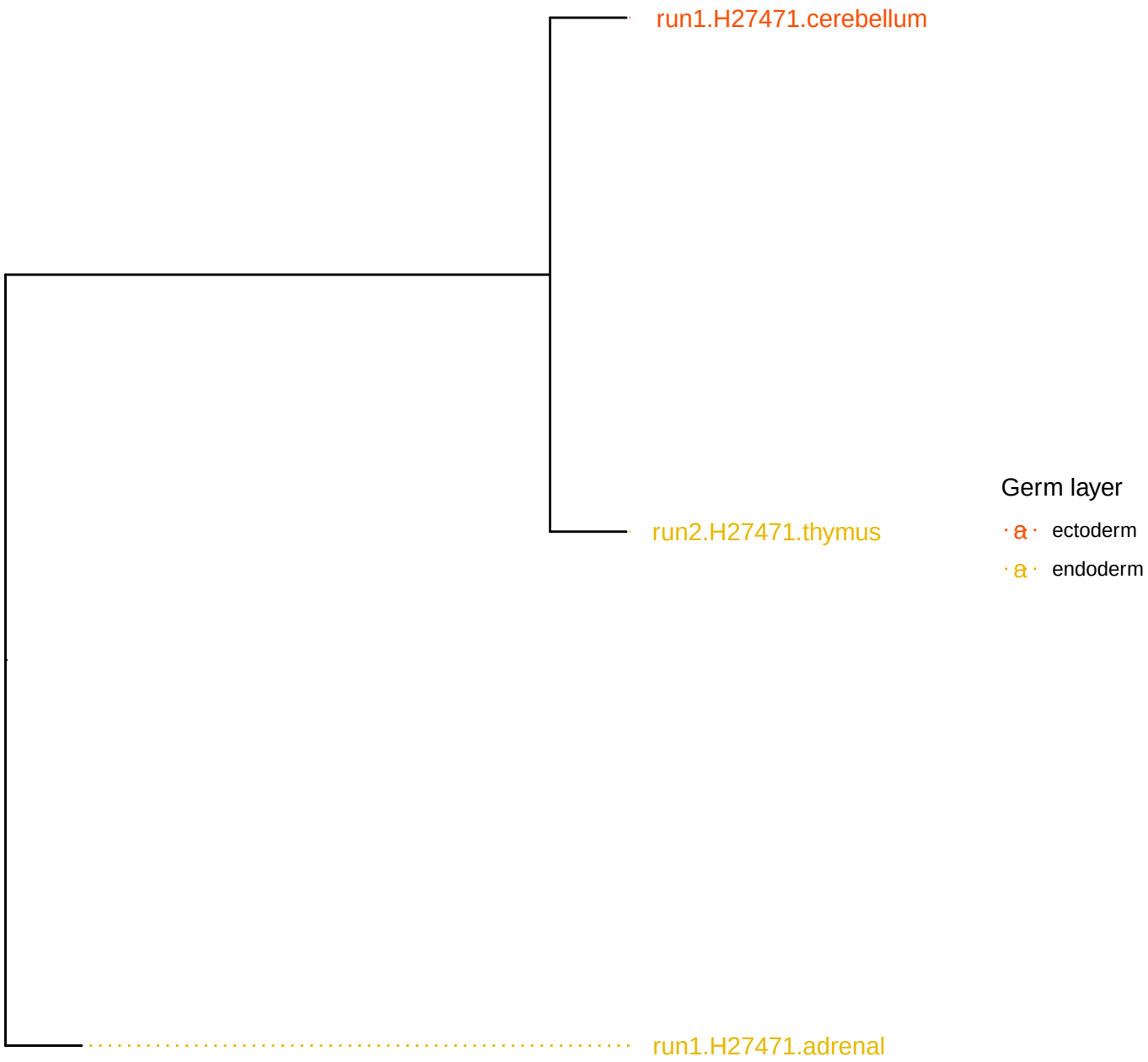
H27469 somaticvr



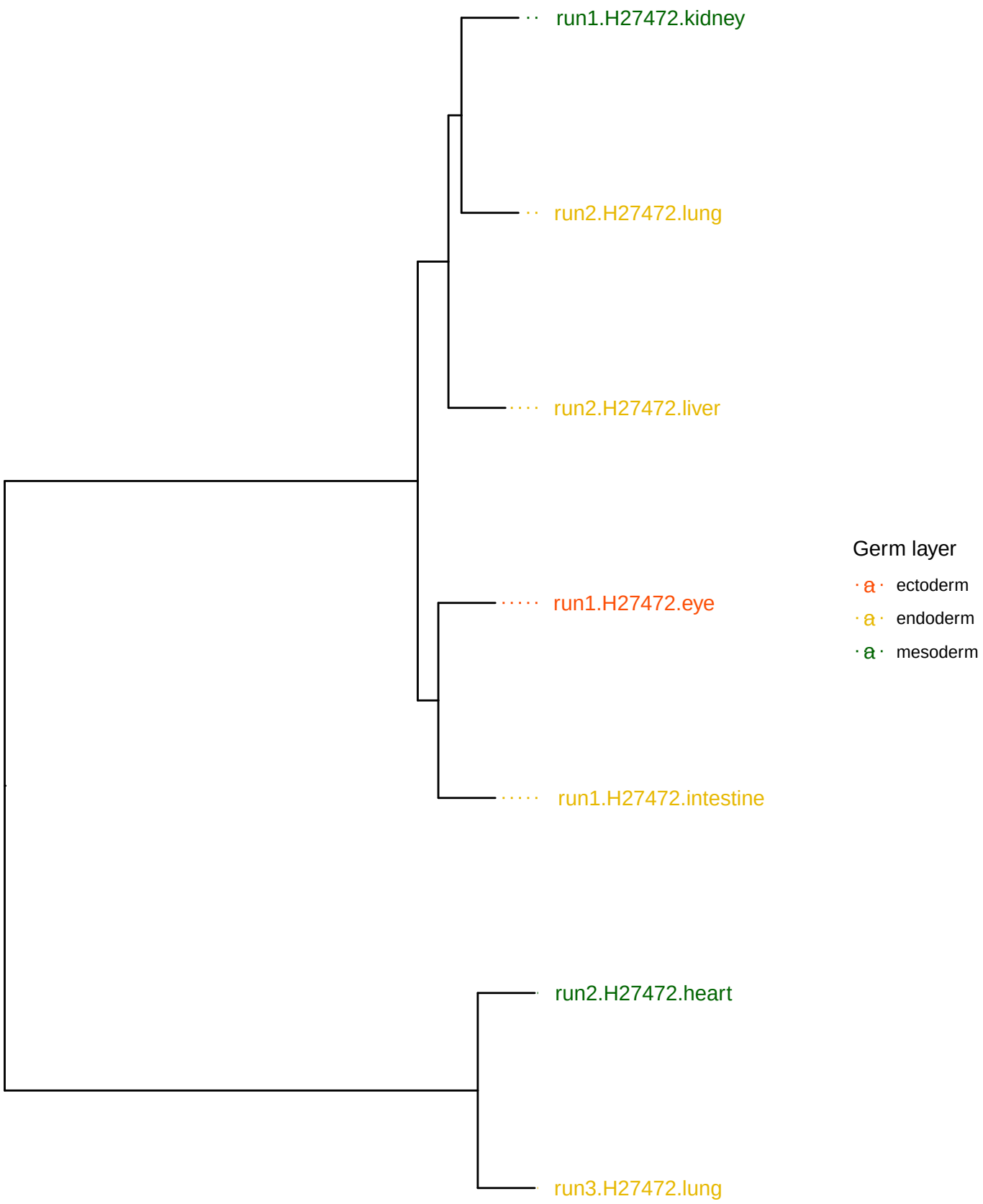
H27471 germlinevr



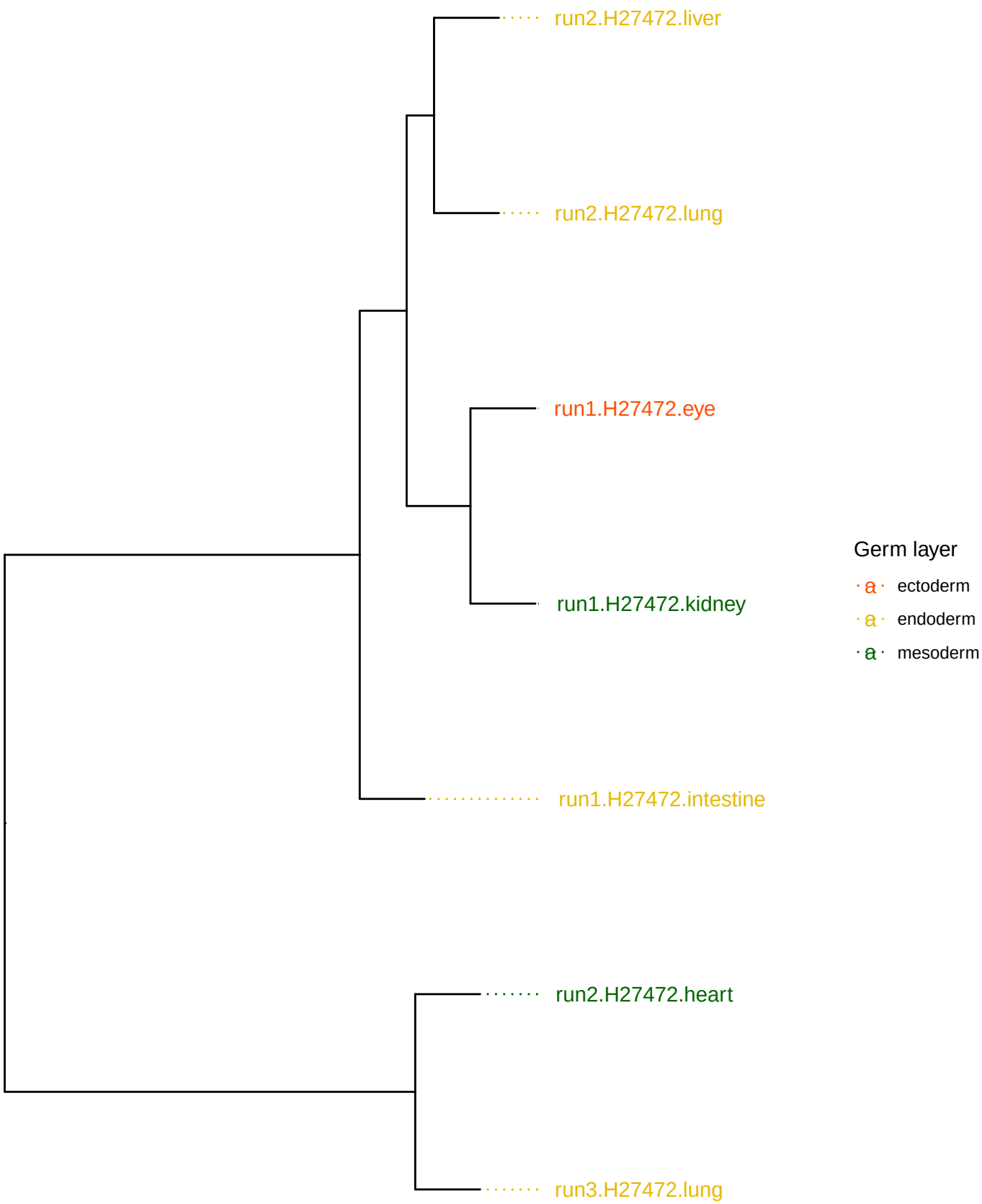
H27471 somaticvr



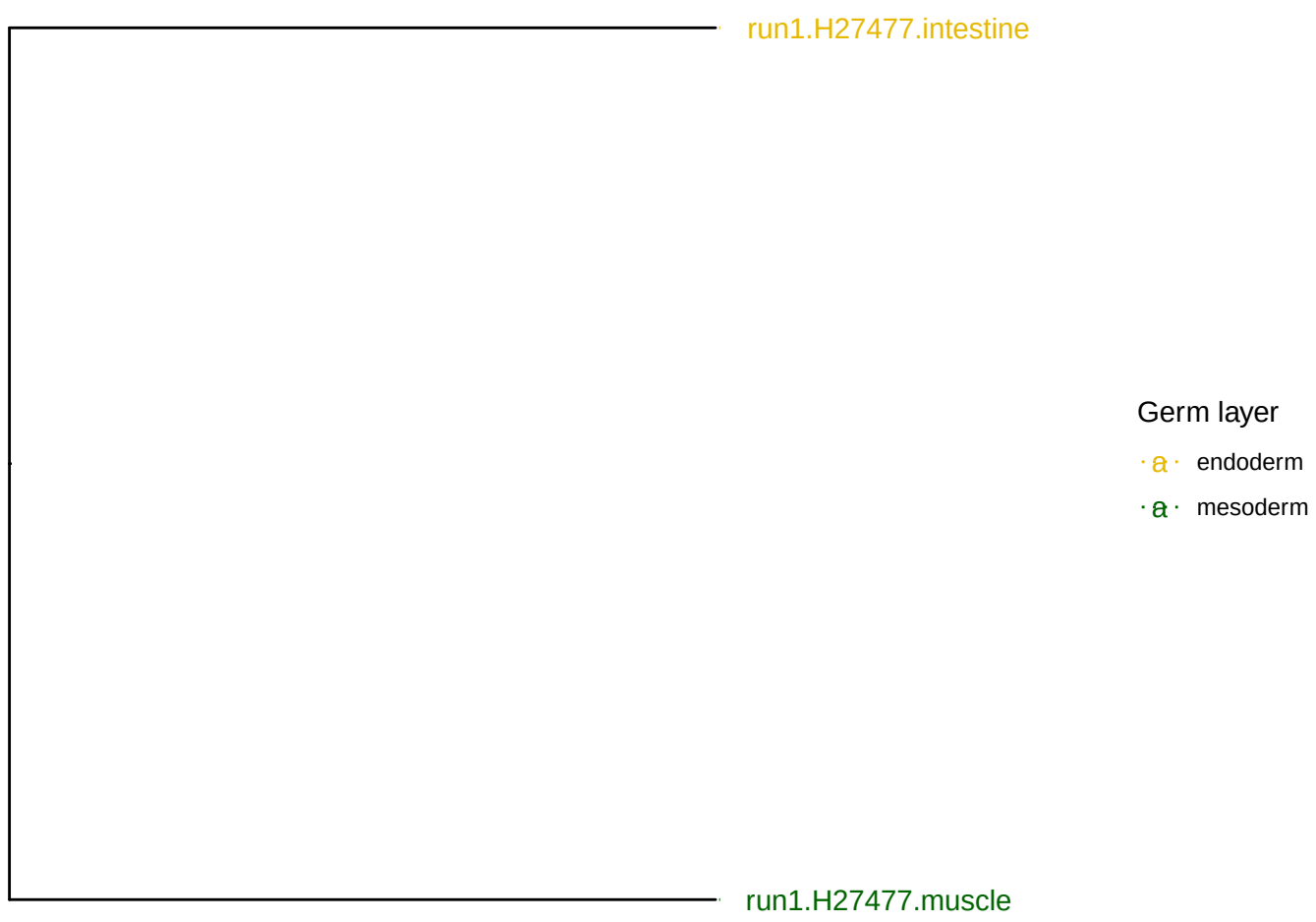
H27472 germlinevr



H27472 somaticvr



H27477 germlinevr



H27477 somaticvr

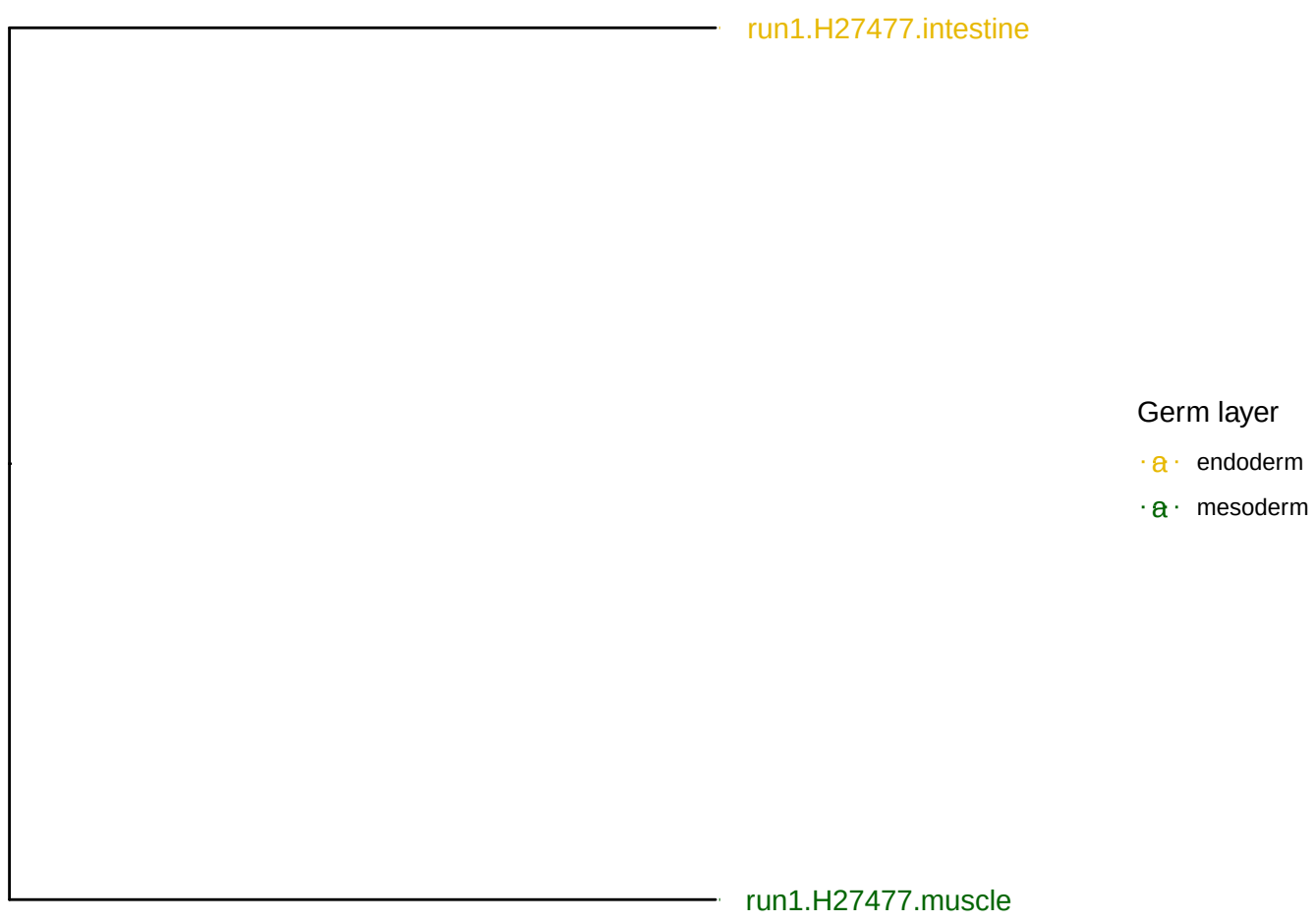


Table 1. | Germ layers and its derived tissue

Mesoderm	Endoderm	Ectoderm
Muscle	Intestine	Brain
Heart	Stomach	Cerebllum
Kidneys	Adrenal	Eye
Gonads	Lung	
Placenta	Pancreas	
Spleen	Liver	
Bonemarrow	Thymus	

Table 2. | COSMIC signatures and the proposed etiology

Signature	Proposed etiology
SBS1	An endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates G:T mismatches in double stranded DNA. Failure to detect and remove these mismatches prior to DNA replication results in fixation of the T substitution for C.
SBS3	Defective homologous recombination-based DNA damage repair which manifests predominantly as small indels and genome rearrangements due to abnormal double strand break repair but also in the form of this base substitution signature.
SBS4	Associated with tobacco smoking. Its profile is similar to the mutational spectrum observed in experimental systems exposed to tobacco carcinogens such as benzo[a]pyrene. SBS4 is, therefore, likely due to direct DNA damage by tobacco smoke mutagens.
SBS5	Unknown. SBS5 mutational burden is increased in bladder cancer samples with ERCC2 mutations and in many cancer types due to tobacco smoking.
SBS54	Possible sequencing artefact. Possible contamination with germline variants.
SBS16, SBS23, SBS43, SBS46, SBS89, SBS96,SBS98,	Unknown

REFERENCES

Abascal, F., Harvey, L. M. R., Mitchell, E., Lawson, A. R. J., Lensing, S. V., Ellis, P., Russell, A. J. C., Alcantara, R. E., Baez-Ortega, A., Wang, Y., Kwa, E. J., Lee-Six, H., Cagan, A., Coorens, T. H. H., Chapman, M. S., Olafsson, S., Leonard, S., Jones, D., Machado, H. E., … Martincorena, I. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature*, *593*(7859), 405–410. https://doi.org/10.1038/s41586-021-03477-4

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, *47*(12), 1402–1407. https://doi.org/10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, … Von Mering, C. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Heisler, L. E., Beck, T. A., Simpson, J. T., Tonon, L., Sertier, A.-S., Patch, A.-M., Jäger, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*(1), 10001. https://doi.org/10.1038/ncomms10001

Assaf, Z. J., Tilk, S., Park, J., Siegal, M. L., & Petrov, D. A. (2017). Deep sequencing of natural and experimental populations of Drosophila melanogaster reveals biases in the spectrum of new mutations. *Genome Research*, *27*(12), 1988–2000. https://doi.org/10.1101/gr.219956.116

Beichman, A. C. (2023). *Evolution of the Mutation Spectrum Across a Mammalian Phylogeny*.

Beichman, A. C., Zhu, L., & Harris, K. (2024). The Evolutionary Interplay of Somatic and Germline Mutation Rates. *Annual Review of Biomedical Data Science*. https://doi.org/10.1146/annurev-biodatasci-102523-104225

Blokzijl, F., De Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., Nijman, I. J., Martincorena, I., Mokry, M., Wiegerinck, C. L., Middendorp, S., Sato, T., Schwank, G., Nieuwenhuis, E. E. S., Verstegen, M. M. A., … Van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, *538*(7624), 260–264. https://doi.org/10.1038/nature19768

Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T. H. H., Sanders, M. A., Lawson, A. R. J., Harvey, L. M. R., Bhosle, S., Jones, D., Alcantara, R. E., Butler, T. M., Hooks, Y., Roberts, K., Anderson, E., Lunn, S., Flach, E., Spiro, S., Januszczak, I., … Martincorena, I. (2022). Somatic mutation rates scale with lifespan across mammals. *Nature*, *604*(7906), 517–524. https://doi.org/10.1038/s41586-022-04618-z

Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R. M., Martin, B., Dulja, A., Schaub, C., Trapnell, C., Larschan, E., O'Connor-Giles, K. M., Furlong, E. E. M., & Shendure, J. (2022). The continuum of Drosophila embryonic development at single-cell resolution. *Science (New York, N.Y.)*, *377*(6606), eabn5800. https://doi.org/10.1126/science.abn5800

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F. J., Glass, I. A., Trapnell, C., & Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science (New York, N.Y.)*, *370*(6518), eaba7721. https://doi.org/10.1126/science.aba7721

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667. https://doi.org/10.1126/science.aam8940

Coorens, T. H. H., Moore, L., Robinson, P. S., Sanghvi, R., Christopher, J., Hewinson, J., Przybilla, M. J., Lawson, A. R. J., Spencer Chapman, M., Cagan, A., Oliver, T. R. W., Neville, M. D. C., Hooks, Y., Noorani, A., Mitchell, T. J., Fitzgerald, R. C., Campbell, P. J., Martincorena, I., Rahbari, R., & Stratton, M. R. (2021). Extensive phylogenies of human development inferred from somatic mutations. *Nature*, *597*(7876), 387–392. https://doi.org/10.1038/s41586-021-03790-y

Coorens, T. H. H., Spencer Chapman, M., Williams, N., Martincorena, I., Stratton, M. R., Nangalia, J., & Campbell, P. J. (2024). Reconstructing phylogenetic trees from genome-wide somatic mutations in clonal samples. *Nature Protocols*. https://doi.org/10.1038/s41596-024-00962-8

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science (New York, N.Y.)*, *348*(6237), 910–914. https://doi.org/10.1126/science.aab1601

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., & Furlong, E. E. M. (2018). The cis-regulatory dynamics of embryonic development at single cell resolution. *Nature*, *555*(7697), 538–542. https://doi.org/10.1038/nature25981

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., Zager, M. A., Glass, I. A., Steemers, F. J., Doherty, D., Trapnell, C., Cusanovich, D. A., & Shendure, J. (2020). A human cell atlas of fetal chromatin accessibility. *Science (New York, N.Y.)*, *370*(6518), eaba7612. https://doi.org/10.1126/science.aba7612

Dou, Y., Gold, H. D., Luquette, L. J., & Park, P. J. (2018). Detecting somatic mutations in normal cells. *Trends in Genetics : TIG*, *34*(7), 545–557. https://doi.org/10.1016/j.tig.2018.04.003

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., Geho, D. H., Petricoin, E. F., & Liotta, L. A. (2006). Laser-capture microdissection. *Nature Protocols*, *1*(2), 586–603. https://doi.org/10.1038/nprot.2006.85

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, *6*, e24284. https://doi.org/10.7554/eLife.24284

Islam, S. M. A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., He, Y., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Steele, C. D., Otlu, B., … Alexandrov, L. B. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics*, *2*(11), 100179. https://doi.org/10.1016/j.xgen.2022.100179

Kakiuchi, N., & Ogawa, S. (2021). Clonal expansion in non-cancer tissues. *Nature Reviews Cancer*, *21*(4), Article 4. https://doi.org/10.1038/s41568-021-00335-3

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2023). Hallmarks of aging: An expanding universe. *Cell*, *186*(2), 243–278. https://doi.org/10.1016/j.cell.2022.11.001

Manders, F., van Boxtel, R., & Middelkamp, S. (2021). The Dynamics of Somatic Mutagenesis During Life in Humans. *Frontiers in Aging*, *2*, 802407. https://doi.org/10.3389/fragi.2021.802407

Martin, B. K., Qiu, C., Nichols, E., Phung, M., Green-Gladden, R., Srivatsan, S., Blecher-Gonen, R., Beliveau, B. J., Trapnell, C., Cao, J., & Shendure, J. (2023). Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nature Protocols*, *18*(1), 188–207. https://doi.org/10.1038/s41596-022-00752-0

Mathieson, I., & Reich, D. (2017). Differences in the rare variant spectrum among human populations. *PLOS Genetics*, *13*(2), e1006581. https://doi.org/10.1371/journal.pgen.1006581

Muyas, F., Sauer, C. M., Valle-Inclán, J. E., Li, R., Rahbari, R., Mitchell, T. J., Hormoz, S., & Cortés-Ciriano, I. (2023). De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01863-z

Olafsson, S., & Anderson, C. A. (2021). Somatic mutations provide important and unique insights into the biology of complex diseases. *Trends in Genetics*, *37*(10), 872–881. https://doi.org/10.1016/j.tig.2021.06.012

Park, S., Mali, N. M., Kim, R., Choi, J.-W., Lee, J., Lim, J., Park, J. M., Park, J. W., Kim, D., Kim, T., Yi, K., Choi, J. H., Kwon, S. G., Hong, J. H., Youk, J., An, Y., Kim, S. Y., Oh, S. A., Kwon, Y., … Ju, Y. S. (2021). Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature*, *597*(7876), 393–397. https://doi.org/10.1038/s41586-021-03786-8

Poetsch, A. R. (2020). The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Computational and Structural Biotechnology Journal*, *18*, 207–219. https://doi.org/10.1016/j.csbj.2019.12.013

Qiu, C., Martin, B. K., Welsh, I. C., Daza, R. M., Le, T.-M., Huang, X., Nichols, E. K., Taylor, M. L., Fulton, O., O'Day, D. R., Gomes, A. R., Ilcisin, S., Srivatsan, S., Deng, X., Disteche, C. M., Noble, W. S., Hamazaki, N., Moens, C. B., Kimelman, D., … Shendure, J. (2024). A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature*. https://doi.org/10.1038/s41586-024-07069-w

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14508–14513. https://doi.org/10.1073/pnas.1208715109

Solís-Moruno, M., Batlle-Masó, L., Bonet, N., Aróstegui, J. I., & Casals, F. (2023). Somatic genetic variation in healthy tissue and non-cancer diseases. *European Journal of Human Genetics*, *31*(1), 48–54. https://doi.org/10.1038/s41431-022-01213-8

Sondka, Z., Dhir, N. B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M., Argasinska, J., Beare, D., Chawla, M. S., Duke, S., Fasanella, I., Neogi, A. G., Haller, S., Hetenyi, B., Hodges, L., … Teague, J. (2024). COSMIC: A curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, *52*(D1), D1210–D1217. https://doi.org/10.1093/nar/gkad986

Spencer Chapman, M., Ranzoni, A. M., Myers, B., Williams, N., Coorens, T. H. H., Mitchell, E., Butler, T., Dawson, K. J., Hooks, Y., Moore, L., Nangalia, J., Robinson, P. S., Yoshida, K., Hook, E., Campbell, P. J., & Cvejic, A. (2021). Lineage tracing of human development through somatic mutations. *Nature*, *595*(7865), 85–90. https://doi.org/10.1038/s41586-021-03548-6

Spisak, N., de Manuel, M., Milligan, W., Sella, G., & Przeworski, M. (2023). Disentangling sources of clock-like mutations in germline and soma. *bioRxiv*, 2023.09.07.556720. https://doi.org/10.1101/2023.09.07.556720

Wang, R., Yang, X., Chen, J., Zhang, L., Griffiths, J. A., Cui, G., Chen, Y., Qian, Y., Peng, G., Li, J., Wang, L., Marioni, J. C., Tam, P. P. L., & Jing, N. (2023). Time space and single-cell resolved tissue lineage trajectories and laterality of body plan at gastrulation. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41482-5

Yu, Z., Coorens, T. H. H., Uddin, M. M., Ardlie, K. G., Lennon, N., & Natarajan, P. (2024). Genetic variation across and within individuals. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-024-00709-x