# Convergent Usage of Amino Acids in Human Cancers as a Reversed Process of Tissue Development

Yikai Luo[1,2,a], and Han Liang[2,3,1,*,b]

[1] *Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston TX 77030, USA*

[2] *Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA*

[3] *Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA*

[*] Corresponding author.

E-mail: hliang1@mdanderson.org (Liang H)

**Running title:** ***Luo Y*** */ Convergent usage of amino acids in cancer*

[a]ORCID: 0000-0001-7589-7981.

[b]ORCID: 0000-0001-7633-286X.

Total word counts: 4053

Total figures: 6

Total tables:0

Total supplementary figures: 7

Total supplementary tables: 0

Total supplementary files: 0

## Abstract

Genome and transcriptome-wide amino acid usage preference across different species is a well-studied phenomenon in molecular evolution, but its characteristics and implication in cancer evolution and therapy remain largely unexplored. Here, we analyzed large-scale transcriptome/proteome profiles such as TCGA, GTEx, and CPTAC and found that compared to normal tissues, different cancer types showed a convergent pattern towards using biosynthetically low-cost amino acids. Such a pattern can be accurately captured by a single index based on the average biosynthetic energy cost of amino acids, termed Energy Cost Per Amino Acid (ECPA). With this index, we further compared the trends of amino acid usage and the contributing genes in cancer and tissue development and revealed their reversed patterns. Finally, focusing on the liver, a tissue with a dramatic increase in ECPA during development, we found that EPCA represented a powerful biomarker that could distinguish liver tumors from normal liver samples consistently across 11 independent patient cohorts (AUROC = ~0.99) and outperformed any index based on single genes. Our study reveals an important principle underlying cancer evolution and suggests the global amino acid usage as a system-level biomarker for cancer diagnosis.


**KEYWORDS:** Amino acid usage; Tissue development; Biosynthetic energy; Diagnostic biomarker

## Introduction

Amino acids are the basic building blocks of a cell. Coding sequences and gene expression profiles are two key factors determining the overall amino acid usage of a cell. Through analysis of the genomes or transcriptomes of many species, preferred amino acid usage is a well-studied topic in macroevolution. The universal trend of "Cost-Usage anticorrelation" suggests that the relative abundance of amino acids, quantified as the number of codons encoding a specific amino acid in the genome of a species, is mainly driven by their biosynthetic energy costs [1–5]. However, it remains unclear how amino acid usage of cancer cells deviates from normal tissues and evolve in different tumor contexts.

From an evolutionary point of view, cancer cells are characterized by a low degree of divergence from its tissue of origin, measured by the limited amount of somatic changes, which is in contrast to the macroevolution that happens across different taxa or even the microevolution existing between within-species individuals [6]. However, such trifling transformation does yield a wide range of phenotypic commonalities shared by distinct cancer types, including activated proliferative signaling, resistance to programmed cell death, induction of angiogenesis, and metastatic capability [7]. Among many theories proposed to understand such convergence, one appealing concept is that cancer cells bear a set of genomic, transcriptomic, and epigenomic features that can be summed up as "stemness," [8–11] which in the context of ontogeny, defines the level of reprogramming/dedifferentiation of adult tissue cells. The underlying mechanistic links between cancer evolution and tissue development have been hinted at by the observations of frequent mutations leading to reactivation of stem cell-related pathways in cancer [12,13]. However, little effort has been made to examine a potential association between these two seemingly non-overlapping processes with respect to amino acid usage.

Characterizing the amino acid usage of cancer cells not only helps us understand the evolutionary constraints in the tumor microenvironment but may also have clinical utility. In recent years, tremendous efforts have been made to identify gene expression-based biomarkers for cancer diagnosis, outcome prediction, and treatment selection, but successful cases with proven clinical values are still limited [14–16]. One factor that determines the feasibility of such biomarkers in clinical practice, the robustness, is rarely satisfied, meaning that a threshold chosen based on limited data is usually not generalizable to unseen scenarios. In contrast to conventional biomarkers based on individual genes, the amino acid usage represents a holistic property of a cellular state. Therefore, there is a possibility that its related

indices represent more robust biomarkers for clinical applications. To fill these knowledge gaps, here we performed a systematic analysis of the amino acid usage profiles across many cohorts of tumor and normal tissue samples.

## Results

### A convergence of amino acid usage across cancer types

Since gene expression levels are largely associated with amino acid usage in a cell, we first examined the gene expression patterns of 30 tissue types in the Genotype-Tissue Expression (GTEx) cohort [17] (Figure S1A) and 31 cancer types in The Cancer Genome Atlas (TCGA) cohort [18] (Figure S1B). Using the t-distributed stochastic neighborhood embedding (t-SNE)[19] projection, we found that samples of a common tissue origin largely formed a single cluster regardless of being normal or cancerous. In addition, cancer types with the same tissue origin, such as brain cancers (glioblastoma multiforme [GBM] and lower grade glioma [LGG]), kidney cancers (kidney renal clear cell carcinoma [KIRC] and kidney renal papillary cell carcinoma [KIRP]), lung cancers (lung adenocarcinoma [LUAD] and lung squamous cell carcinoma [LUSC]), and liver cancers (hepatocellular carcinoma [LIHC] and cholangiocarcinoma [CHOL]), tended to be mingled or closer to each other than to other cancer types. We observed similar patterns in two other large, pan-cancer cohorts, PCAWG [20], and MET500 [21] (Figure S1C and D). Consistent with previous studies [18,22], these results indicate that cancer cells largely retain their tissue-specific gene expression profiles.

To study whether this tissue-specific pattern holds for amino acid usage, we calculated the similarity of transcriptome-based amino acid usage by integrating the gene expression profiles and the amino acid frequencies of protein-coding genes (**Figure 1**A) and visualized their patterns in the same way. Similar to the strong tissue specificity observed in the gene expression analysis, we found that normal tissues of the GTEx cohort still had distinct amino acid usage patterns (Figure 1B). We further confirmed this result by co-clustering amino acid usage profiles of the Human Protein Atlas (HPA) cohort [23] with corresponding GTEx tissue types (Figure S2A). More intriguingly, samples of a multi-species multi-tissue cohort [24] were principally separated by tissue type rather than by species, suggesting that tissue-specific amino acid usage is highly conserved across mammals (Figure 1C).

In sharp contrast to normal tissues, when clustered by amino acid usage, samples of different cancer types were much less separated and did not segregate on the basis of tissue origins (Figure 1D). To further confirm this observation, we clustered amino acid usage

profiles of two other cancer cohorts, PCAWG and MET500, and observed a dramatic loss of tissue-specificity relative to the patterns observed in the gene expression-based analysis (Figure S1C and D and Figure S2B and C). To ensure that the detected pattern was not due to a disparity in sample size or unmatched tissue types, we leveraged a conservative GTEx-TCGA mapping to only include normal and tumor samples whose tissue origins are matched without ambiguity, then performed down-sampling within individual tissue-specific cohorts, and finally, applied t-SNE to redo a supervised clustering. The results remained the same for the comparison between down-sampled GTEx and TCGA samples (Figure S2D and E) as well as for that between TCGA tumor samples and the normal adjacent to tumor (NAT) (Figure S2F and G). This observation is important since, evaluating tumor purity and gene signatures, recent studies have shown that NAT samples reside in an intermediate state between healthy and tumor [25,26].

The observation that amino acid usage for cancer cells failed to preserve their distinct tissue origins raised two possibilities: (i) cancer cells evolved to possess highly stochastic amino acid usage profiles both within and between cancer types; or (ii) they went through convergence of amino acid usage, thereby losing the constraint of the original tissue specificity. To identify the correct hypothesis, we simply asked whether, in the 20-dimensional space (each dimension representing the frequency of specific amino acid), the distances between samples of different cancer types were shorter than those among samples of different normal tissues. Based on Pearson's distance, for each sample, we defined an amino acid usage convergence index that measured its distance to all other samples of different tissue or cancer types. Through a comparative analysis of GTEx normal vs. TCGA tumor and TCGA NAT vs. tumor, we found that tumor samples showed significantly increased convergence than normal samples, a pattern consistently observed across all surveyed cancer types (Figure 1E and F). Furthermore, we compared the variations of amino acid frequencies across NAT samples and tumor samples of different cancer types based on the same set of standard deviations. Indeed, the extent to which amino acids are differentially used in tumors was markedly reduced than that in NATs (Figure S4A and B). Collectively, these results indicated a strong convergence rather than a stochastic transformation of amino acid usage across cancer types, supporting our second hypothesis.

**Cancer cells tend to use biosynthetically low-cost amino acids**

To understand how such a convergent pattern occurs, we quantified the differential usage of each amino acid in tumors vs. normal tissues and found no highly consistent trend across

cancer types in terms of increased or decreased usage (Figure S4C). However, when taking a higher view of the heatmap, structurally complex amino acids, such as tryptophan and cysteine, tended to be significantly depleted in most cancer types, whereas those with relatively simpler structures tended to be significantly enriched in a majority of cancers. Because the structural complexity of the amino acids correlates well with the energy cost of their biosynthesis [1], we hypothesized an association between the biosynthetic energy cost of amino acid and its usage tendency in cancers. Indeed, we observed a strong negative correlation between the biosynthetic energy cost and the net number of cancer types in which the usage of an amino acid was significantly increased (**Figure 2**A**,** Rs = -0.56, p = 0.01), suggesting that cancer cells prefer amino acids with a lower biosynthetic energy cost. We previously introduced two indices, $ECPA_{gene}$, and $ECPA_{cell}$, which quantify the average biosynthetic energy cost per amino acid for a gene and a cell (or a sample), respectively [27] (Figure 2B). $ECPA_{gene}$ is based on the amino acid frequency encoded in a gene, and $ECPA_{cell}$ considers the expression levels and amino acid frequencies of all the genes in a cell. A high ECPA value indicates that the gene or the cell tends to use biosynthetically expensive amino acids. We found that compared to NAT samples, $ECPA_{cell}$ of the tumor samples became significantly lower for 9 out of the 15 tested cancer types, while no significantly opposite patterns were observed (Figure 2C). To confirm this pattern at the proteomic level, we extended these analyses to six cancer proteomics datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [28] and others [29,30], covering five cancer types. Strikingly, in all the cases, proteins that were significantly up-regulated in tumor samples ($log_2FC > 0$, FDR < 0.05) had significantly lower $ECPA_{gene}$ than the proteins that were significantly down-regulated ($log_2FC < 0$, FDR < 0.05) (Figure 2D). These results indicate that cancer cells reshaped their gene/protein expression programs to use biosynthetically inexpensive (or structurally simpler) amino acids, thereby losing their original tissue-specific amino acid usage profiles. Finally, we sought to test if our ECPA index is insensitive to the expression of genes with extremely high abundance, including those encoding certain housekeeping proteins as well as tissue-specific proteins. After removal of all genes that either encode cytoplasmic and mitochondrial ribosome proteins or rank within top 200 in median TPM of the same cancer type, we recalculated the ECPA index for each sample and found that the pattern of consistent decrease of $ECPA_{cell}$ in tumor samples across multiple cancer types was almost perfectly reproduced (Figure S3).

We next tested whether the amino acid usage convergence level of a tumor was correlated with its $ECPA_{cell}$. Indeed, we found a strong inverse relationship for seven out of the nine

cancer types where $ECPA_{cell}$ was significantly lower in tumors (Figure 2E). Thus, the more a tumor follows a convergent path to a common state of amino acid usage, the higher the bias it has toward using biosynthetically low-cost amino acids. These results also suggest that $ECPA_{cell}$ is a simple, informative, interpretable index that effectively captures the overall preference of amino acid usage for a specific sample. Therefore, we focused on this index in subsequent analyses.

**Biosynthetically expensive amino acids are increasingly used during tissue development**

To elucidate the underlying cause for the convergence of amino acid usage in cancer, we first sought to understand how tissue-specific amino acid usage patterns are established during development. Using the $ECPA_{cell}$ index, we quantified the overall amino acid usage of liver and kidney tissues across different development stages in mammals, including humans, mice, rats, rabbits, and opossums. Intriguingly, both tissues showed an increasing trend of $ECPA_{cell}$ along their developmental trajectories in all five mammals (**Figure 3**A and B). A closer inspection of the $ECPA_{cell}$ trend lines led to two observations: i) key turning points of $ECPA_{cell}$ in different species tend to happen at corresponding developmental stages; and ii) the rise of $ECPA_{cell}$ in the liver takes concave trajectories while that in the kidney takes convex trajectories, suggesting that the establishment of high $ECPA_{cell}$ status is driven by evolutionarily conserved synchronous molecular events that possess strong tissue specificity. To confirm this pattern, we collected another three independent RNA-seq datasets on mouse liver development and found a consistent $ECPA_{cell}$ increase along the developmental paths in all three cases (Figure 3C-E).

To pinpoint which gene modules are responsible for the tissue-specific build-up of a high $ECPA_{cell}$ status, we first defined a "$\Delta ECPA_{cell}$ contribution index" for each gene, which quantified the contribution of the gene to the global shift of $ECPA_{cell}$ (see Materials and methods). We then divided all genes into 15 equal bins based on their index values and employed a mutual information-based enrichment identification algorithm called iPAGE [31] to detect the enrichment of these gene groups with well-established functional gene modules. We noted that genes contributing to the $ECPA_{cell}$ increase were conserved among mammals but were tissue-specific. For the liver, the enriched modules included glucuronosyltransferase activity and complement activation (Figure 3F, Figure S5A, C and E); and for the kidney, the enriched modules included sphingolipid biosynthetic process and zinc/calcium ion homeostasis (Figure 3G, Figure S5B, D, and F).

Development-related cellular states that are instituted in adulthood can be prone to significant transformation or even complete collapse during aging [32]. To further understand how tissue-specific amino acid usage patterns alter when the tissue undergoes senescence, we gathered independent transcriptome profiles of aging livers and kidneys in humans, mice, and rats, and characterized the $ECPA_{cell}$ patterns. Both tissues showed a stable pattern of high $ECPA_{cell}$ status with reasonable fluctuations (Figure S6A-C). We concluded that tissue-specific, preferred usage of biosynthetically expensive (or structurally complex) amino acids, characterized by a high-$ECPA_{cell}$ status, was gradually formed during development and remained largely unchanged in aging.

**Amino acid usage convergence of tumor follows a reversed path of tissue development**

The strong convergence of amino acid usage across different cancer types is reminiscent of the "reverse-evolution" concept for tumorigenesis. As demonstrated above, this idea is well illustrated by the observation that there is a consistent decline of $ECPA_{cell}$ in tumors, whereas there is a gradual increase of $ECPA_{cell}$ during tissue development. To test the hypothesis that cancer evolution and tissue development move in opposite directions with respect to amino acid usage, we assessed whether the genes that boosted $ECPA_{cell}$ in tissue development were overlapped with those that reduced $ECPA_{cell}$ in tumors of the corresponding tissue origin and vice versa. Following the same method of computing $\Delta ECPA_{cell}$ contribution index for tissue development, we measured the contribution of individual genes to $\Delta ECPA_{cell}$ in cancer evolution for three cancer types for which gene expression profiles of normal developing tissues are available, namely LIHC, KIRC, and KIRP. Based on their contributions to $\Delta ECPA_{cell}$ in either development or tumorigenesis, we divided individual genes into four quadrants with zero as the cutoff. We then used Fisher's exact test to analyze the overlap of developmental $\Delta ECPA_{cell}$-positive-contributing genes with tumorigenic $\Delta ECPA_{cell}$-negative-contributing genes and vice versa. We observed that genes indeed tended to make opposite contributions to $\Delta ECPA_{cell}$ in tumorigenesis and tissue development (**Figure 4**A-C, Fisher's exact test, LIHC, $p = 1.6\times10^{-156}$; KIRC, $p = 1.9\times10^{-39}$; KIRP, $p = 8.9\times10^{-30}$). Furthermore, for the genes reducing $ECPA_{cell}$ in tumorigenesis and increasing $ECPA_{cell}$ in development, their $\Delta ECPA_{cell}$ contribution index in these two processes were significantly negatively correlated (Figure 4D-F).

While the gene-level analyses above were possibly hindered by the fact that cancer progression is highly heterogeneous even within the same cancer type [33,34], we can expect that a sample-level analysis would be more efficient to detect potential reverse relationships

between cancer evolution and tissue development regarding amino acid usage. To this end, we defined the "developmental reversal index" for each tumor sample, which quantifies how strongly its gene expression pattern reversed what was instituted in tissue development. Specifically, we first calculated the gene-expression fold change of each tumor sample in terms of that averaged over the adjacent normal samples in order to measure the transcriptomic shift during tumorigenesis. We then measured the strength of anti-correlation between such a shift and the expression changes of the same gene set along the developmental trajectories of matched tissues (see Materials and methods). Interestingly, using this index to stratify cancer patients in terms of overall survival time, we found that a higher developmental reversal value was consistently associated with a worse prognosis (Figure 4G-I), suggesting that more aggressive tumors tend to have gene expression profiles more reversed in the tissue development trajectory.

Finally, we employed a multivariate linear regression model to clarify the associations between how biased a tumor sample tends to be in using biosynthetically inexpensive amino acids (represented by $ECPA_{cell}$), how far it travels on the path of amino acid usage convergence relative to other cancer types (represented by amino acid usage convergence index), and how strongly its gene expression pattern reversed from what was instituted in tissue development (represented by the developmental reverse index). Remarkably, both the convergence level and the developmental reversal level were strongly anti-correlated with $ECPA_{cell}$ across cancer types (Figure 4J-L). We, therefore, put forward an integrated model in which cancer cells initiated from distinct tissue origins converge into a common state favoring the use of biosynthetically inexpensive amino acids through reversed paths of tissue development (Figure 4M).

**The amino acid usage index, $ECPA_{cell}$, is a robust biomarker for liver cancer diagnosis**

Among different cancer types in our $ECPA_{cell}$ analysis, the difference between liver normal and liver tumor samples was striking, making this tissue stand out from others (Figure 2C). Indeed, by quantifying the downward shift of $ECPA_{cell}$ ($\Delta ECPA_{cell}$) between tumor and the matched NAT pairs, the top two cancers were CHOL and LIHC, both of which originate from the liver (**Figure 5**A**)**. We suspected that such a striking pattern could be attributed to liver-specific gene expression. To test this, we calculated $ECPA_{cell}$ of both GTEx normal samples and TCGA NAT samples based only on tissue-specific genes [35] and ranked the tissues by their average $ECPA_{cell}$. Indeed, the liver $ECPA_{cell}$ level was higher than almost all

other tissues (Figure 5B and C) (although the pancreas showed an even higher $ECPA_{cell}$ according to the GTEx samples, the pattern did not hold for TCGA NAT samples). Of note, while the sample size of LIHC-NAT was as large as 50, the variation of their $ECPA_{cell}$ based on tissue-specific genes was low. Furthermore, a comparison of the developmental $ECPA_{cell}$ trend lines for different human tissues revealed that a fast and early build-up of a high-$ECPA_{cell}$ status only existed for the liver (Figure 5D). We observed similar patterns in other mammals as well (Figure S7A-D). These results suggest that during development, the liver acquires a very high $ECPA_{cell}$ state, and the liver-specific genes are the underlying contributing factor.

Given (i) the extremely high $ECPA_{cell}$ level of liver tissue, and (ii) the dramatic difference between liver tumor and matched normal samples, we speculated whether $ECPA_{cell}$ could be utilized as a novel biomarker for detecting liver cancer. To this end, we first collected 11 independent liver-cancer RNA-seq datasets (including TCGA LIHC and CHOL) where matched tumor and adjacent normal biopsies were simultaneously collected, thereby enabling a direct comparison of $ECPA_{cell}$ between these conditions. In all cases, the tumor samples showed significantly reduced $ECPA_{cell}$ with large effect sizes (**Figure 6**A).

To evaluate more rigorously the capacity of $ECPA_{cell}$ to serve as a diagnostic marker in discriminating liver tumors from normal tissues, we employed the area under the receiver operating characteristic curve (AUROC) as a performance metric. To ensure the robustness of our analyses, we only included six datasets with sample size ≥12. The $ECPA_{cell}$ index was able to separate tumor vs. normal samples with very high ROC scores (median value = 0.993, range = 0.982 - 1.00, Figure 6B). To compare the predictive power of $ECPA_{cell}$ relative to individual gene-based biomarkers, we calculated the average AUROC of all detectable genes across the six datasets and assessed their performance in the same way. Among 9,559 genes assessed, only three genes (*CCT3*, *DDX39A*, and *FLAD1*) showed slightly better performance than $ECPA_{cell}$ (0.992), but none of them had statistically significant superiority (Figure 6C and D). In addition, $ECPA_{cell}$ showed significantly higher discriminating power than the usage of any single amino acid (Figure 6E). Along with accuracy, a key feature of a successful biomarker is its robustness. To assess this feature, we computed the coefficient of variation (CV) for the optimal thresholds of $ECPA_{cell}$ and individual genes across different datasets as an indicator of robustness. $ECPA_{cell}$ showed exceptionally high robustness with its CV as low as $7.9\times10^{-3}$, about 5× smaller than the lowest CV of any single gene-based biomarker (Figure 6F). Notably, the three genes that had a statistically insignificant advantage over $ECPA_{cell}$ by AUROC had extremely unstable optimal cutoffs among different

datasets, suggesting their limited power in detecting liver cancer across diverse clinical scenarios. Collectively, these results suggest that, as a system-level feature capturing the global usage of amino acids in a sample, $ECPA_{cell}$ represents a promising biomarker for liver cancer diagnosis, and possesses both high accuracy and exceptional robustness.

## Discussion

Here we performed a systematic analysis on transcriptome and proteome-based amino acid usage across a broad range of cancer types. Using a previously introduced index, $ECPA_{cell}$, our results revealed, for different tumors, a convergent pattern toward a cellular state of using more biosynthetically low-cost amino acids. In parallel, we studied the amino acid usage in the developmental trajectories of multiple organs and uncovered diverse paths into a tissue-specific high-$ECPA_{cell}$ status that were evolutionarily conserved across mammals. Thus, a reverse relationship existed between cancer evolution and tissue development, which can be viewed as reminiscent of the widely accepted concept of the cancer cell "stemness." Furthermore, given the long-standing parallels between phylogeny and ontogeny [36], supported by recent evidence [24,37,38], it would be reasonable to interpret cancer evolution as a reversed process of not only the development of an organism or its tissues but also the evolution of species. It has been argued that one key mechanism adopted by cancer cells to obtain fitness in spite of the diversity of the microenvironments is to unleash the force that is suppressed in multicellular organisms but is borne by unicellular organisms that are at the very bottom of the evolutionary hierarchy [39–43]. Thus, amino acid usage, a key aspect of cellular metabolism, may provide a unique perspective to understand the fundamental principles governing cancer progression, tissue development, and macroevolution, three evolutionary processes on different scales.

With the advances in transcriptome profiling technology, gene expression-based biomarkers have attracted wide attention for tumor detection and patient stratification. However, due to the high heterogeneity of cancer and intrinsically stochastic nature of gene expression, biomarkers based on either a single gene or a set of genes tend to suffer from numerical instability, thereby performing poorly. As demonstrated for liver cancer diagnosis, our $ECPA_{cell}$ index represents a system-level biomarker that has at least three remarkable advantages. First, $ECPA_{cell}$ captures a global cellular state by retaining the entire transcriptome as its information source, thereby conferring unparalleled robustness. Second, $ECPA_{cell}$ was derived *de novo* from the gene expression profile of a sample, thus independent of external reference, which might introduce large noise predominantly attributable to batch

effect. Third, in contrast to data-driven metrics, $ECPA_{cell}$ has a well-defined biological meaning, the biosynthetic energy cost of amino acids. Because of these properties, $ECPA_{cell}$ is an extremely robust diagnostic biomarker for liver cancer with a nearly constant threshold for tumor-normal segregation. Further efforts are warranted to assess the utility of this index in other cancer types and clinical applications.

## Materials and methods

### Data acquisition and processing

We obtained the gene-level expression values (e.g., fragments per kilobase per million [FPKM] or transcripts per million [TPM]) of the TCGA cancer sample cohorts, the GTEx normal tissue cohort, and the MET500 metastatic tumor cohort, from the Xena data portal (https://xenabrowser.net/datapages/); the HPA cohort from the HPA data portal (http://www.proteinatlas.org/); and the PCAWG cohort from the ICGC data portal (https://dcc.icgc.org/releases/PCAWG/transcriptome/). We also obtained the gene expression values of the mammalian tissue development cohorts from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under the accession IDs E-MTAB-6769 (chicken), E-MTAB-6782 (rabbit), E-MTAB-6798 (mouse), E-MTAB-6811 (rat), E-MTAB-6813 (rhesus macaque), E-MTAB-6814 (human), and E-MTAB-6833 (opossum); and two independent RNA-seq datasets of mouse liver development from the Gene Expression Omnibus (GEO) under the accession IDs GSE58733 and GSE58827, as well as from ArrayExpress under the accession ID E-MTAB-2328. Finally, we obtained RNA-seq datasets of aging mouse liver and kidney from GEO under the accession ID GSE132040.

To convert gene-level FPKM values to TPM [44] values for a gene $g_i$ in a sample $s_k$, we used the formula:

$$TPM_{g_i,s_k} = \frac{FPKM_{g_i,s_k}}{\sum_{j=1}^{n} FPKM_{g_j,s_k}} \times 10^6$$

where the denominator on the right side is the sum of FPKM values of all the genes for an individual sample.

We downloaded raw RNA-seq fastq files of human liver cancer from GEO under the accession IDs GSE65485, GSE119336, GSE77314, GSE77509, GSE63863, GSE94660, GSE25599, GSE124535, and GSE55758; files of aging rat liver from the Sequence Read Archive (SRA) under the accession ID PRJNA516151, and files of TCGA LIHC and CHOL cohorts from the GDC Data Portal (https://portal.gdc.cancer.gov/). MultiQC [45] was used to

assess the quality of the sequencing files and the performance of the preprocessing steps. Transcript-level abundances were quantified by Salmon [46] using the GRCh38 transcriptome as the reference. Gene-level TPM values were aggregated from transcript-level TPM values by tximport [47].

We obtained the proteomics datasets of KIRC, COAD, LUAD, and OV patient cohorts from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/). We obtained two proteomics datasets of liver cancer from the NODE data portal (https://www.biosino.org/node/index/) and the CNHPP data portal (http://liver.cnhpp.ncpsb.org/), respectively.

**Calculation of transcriptome-based amino acid usage**

We used the following formula to compute the amino acid frequency matrix given an RNA-seq dataset (see also Fig. 1a):

$$F_{m\times 20} = E_{m\times n} A_{n\times 20}{}^{T}$$

where $E$ is a matrix of genes $g_1, g_2, \ldots, g_n$ by samples $s_1, s_2, \ldots, s_m$ with entries as TPM values, and A is a matrix of genes $g_1, g_2, \ldots, g_n$ by amino acids $a_1, a_2, \ldots, a_{20}$ with entries as relative frequencies of amino acids computed using the protein sequences annotated in the Swiss-Prot and TrEMBL databases hosted by the UniProt website (https://www.uniprot.org/). When a gene has multiple isoforms, we used its canonical sequence, as defined by UniProt based on criteria such as transcript length, relative abundance, and evolutionary conservation, in our analyses. We also repeated our analyses using transcript-level TPM data, where all isoforms annotated by ENSEMBL were included and had nearly identical results.

**Variation analysis of amino acid usage for TCGA samples**

To illustrate the variation of amino acid usage of NAT samples from different tissues, we computed z-scores based on the average frequencies for individual amino acids across tissues. To compare these with the variations in amino acid usage of tumor samples across cancer types, instead of using *de novo* standard deviations to compute z-scores, we used the set of standard deviations derived for the NAT samples to obtain z-scores for the tumor samples. We used hierarchically clustered heatmaps with Euclidean distance as the distance metric to visualize the tissue-specificity of amino acid usage. To identify differential amino acid usage between tumor and NAT samples, we performed the Wilcoxon rank-sum test for frequencies of individual amino acids using paired tumor and NAT samples and used an FDR-adjusted p-

value of 0.05 as the threshold for significance. Similarly, a hierarchically clustered heatmap was used to display amino acid de-regulation patterns across cancer types.

**Calculation of $ECPA_{gene}$ and $ECPA_{cell}$**

We calculated two indices of amino acid usage, $ECPA_{gene}$, and $ECPA_{cell}$, representing the average biosynthetic energy cost per amino acid of a gene and a cell, respectively, as described previously [27]. Briefly, the biosynthetic costs of amino acids are based on the amount of high-energy phosphate bond equivalents required for amino acid biosynthesis in yeast and are normalized by amino acid decay rates (the biosynthetic costs of amino acids are highly correlated between different species). We then calculated $ECPA_{gene}$ and $ECPA_{cell}$ by multiplying the biosynthetic energy costs with the relative amino acid frequency of a gene or a cell (sample).

**Quantification of amino acid usage convergence for TCGA samples**

To quantify the similarity of NAT or tumor samples in the TCGA cohort in terms of their amino acid usage patterns, we applied the Pearson's distance metric to the amino acid frequency profiles, derived as described above. We also employed the Spearman rank correlation coefficient as an alternative metric and obtained the same results. Specifically, to capture the convergent pattern of amino acid usage across cancer types, we defined, for a sample $s_i$ of cancer type $X$, the amino acid usage convergence index as:

$$1 - \frac{\sum_{j=1}^{N} d_{s_i,s_j}}{N} (s_j \notin X)$$

where $d_{s_i,s_j}$ is the Pearson's distance between sample $s_i$ from cancer type $X$ and sample $s_j$ not from cancer type $X$.

**Calculation of $\Delta ECPA_{cell}$ contribution index**

To estimate the contribution of individual genes to the alteration of $ECPA_{cell}$ in a specific biological process, we considered both how different the ECPA of a gene is from the baseline $ECPA_{cell}$, as well as how much its expression level has changed. Formally, we defined the $\Delta ECPA_{cell}$ contribution index of a gene $g_i$ as:

$$(ECPA_{g_i} - ECPA_{baseline}) \times I_{g_i}$$

where $I_{g_i}$ is an importance score that describes the extent of deregulation of $g_i$. In tumorigenesis, we employed the $\log_2$ fold-change of average expression level between tumor

and NAT samples as the importance score. In tissue development, we employed a different importance score that was not based on binary comparison as in tumorigenesis since the nature of the dataset is time-course measurements. Specifically, we applied an R package designed for transcriptomic time courses, maSigPro [48], to build a polynomial regression model (degree = 3) for each gene using its expression level as the response variable and the log-transformed post-conception days as the independent variable. Such models yielded the goodness-of-fit ($R^2$) values that were then signed by the corresponding Spearman correlation coefficients and were finally used as the importance score.

**Pathway analysis of $\Delta ECPA_{cell}$ contribution in mammalian tissue development**

We employed an information-theoretic framework [31] to reveal gene modules or regulatory pathways that were enriched in genes with a significant contribution to the increase of $ECPA_{cell}$ during tissue development. First, we focused on down-regulated genes with lower-than-baseline $ECPA_{gene}$ and up-regulated genes with higher-than-baseline $ECPA_{gene}$, both of which could contribute to the increase of developmental $ECPA_{cell}$. Second, we distinguished these two groups of genes by signing the index of down-regulated genes as negative, followed by rank-transforming all retained genes, and dividing the genes into equal bins. Third, we used the iPAGE algorithm that calculated the mutual information between the gene ranks and the pathway memberships (the number of genes belonging to a pathway in each bin) for every Gene Ontology term. A random-permutation test was used to estimate the significance of these mutual information (MI) values so that significantly informative pathways were identified with high MI values and low p values. Finally, the hypergeometric test was used to determine whether a specific pathway was over- or under-represented in each bin. For visualization, heatmaps of pathways by bins were drawn using log-transformed p values.

**Calculation of developmental reversal index of tumor samples**

To assess the level of developmental reversal for tumor samples of TCGA LIHC, KIRC, and KIRP cohorts, we asked how greatly the shift of a tumor transcriptome from a mega NAT reference (averaging gene expressions over all NAT samples of a certain cancer type) had reversed the shift of the transcriptome along the developmental trajectory of a corresponding tissue. Formally, we defined, for a sample $s_i$, the developmental reversal index as:

$$\rho(\log_2(\overrightarrow{e_{s_i}} \oslash (\overrightarrow{Em^{-1}})), \vec{r})$$

where ⊘ is element-wise division, $\rho$ is the Spearman correlation coefficient, $e_{s_i}$ is a vector of $n$ gene expressions for sample $s_i$, $E$ is a matrix of genes $g_1, g_2, \ldots, g_n$ by NAT samples $s_1, s_2, \ldots, s_m$ of a certain cancer type with entries as expression level, $\overrightarrow{m^{-1}}$ is a normalization vector of constant $m^{-1}$, and $\vec{r}$ is a vector of signed goodness-of-fit values of genes $g_1, g_2, \ldots, g_n$ derived from the developmental RNA-seq data of a matched tissue type. We examined the association of this index with patients' overall survival times in TCGA LIHC, KIRC, and KIRP cohorts using log-rank tests, where patients were split into two equal groups based on the median value of developmental reversal index.

**Evaluation of the utility of $ECPA_{cell}$ as a diagnostic biomarker**

To quantify the performance of $ECPA_{cell}$ in differentiating tumors from related normal samples, we used the AUROC metric to compare it with those of all detectable individual genes (TPM ≥ 1 in ≥50% of samples in the cohort). To determine the optimal threshold of $ECPA_{cell}$ or gene expression level for tumor-normal separation, we chose the value that maximizes Youden's J statistic, which equals to (sensitivity + specificity – 1). If multiple optimal cutoffs existed for a biomarker whose average level was higher in NAT than in tumors, the one with the highest value was picked and vice versa.

## CRediT author statement

**Yikai Luo**: Conceptualization, Formal Analysis, Visualization, Writing – Original Draft.

**Han Liang**: Conceptualization, Supervision, Writing - Review & Editing, Funding acquisition. All authors read and approved the final manuscript.

## Competing interests

H.L. is a shareholder and scientific advisor to Precision Scientific Ltd.

## Acknowledgements

We thank H. Zhang, H. Chen, and other members of the Liang lab for helpful discussions. We thank H. Goodarzi for critical review of the manuscript. We also thank K. Mojumdar for editorial assistance. This work was supported by the US National Institutes of Health (U24CA209851 and Cancer Center Support Grant P30CA016672 to H.L.), an MD Anderson Faculty Scholar Award (to H.L.), and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (to H.L.).

## Authors' ORCID IDs

0000-0001-7589-7981 (Yikai Luo)

0000-0001-7633-286X (Han Liang)

## References

[1] Seligmann H. Cost-minimization of amino acid usage. J Mol Evol 2003;56:151–61. https://doi.org/10.1007/s00239-002-2388-z.

[2] Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller R V., Krane DE. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: A whole-genome analysis. Mol Biol Evol 2006;23:1670–80. https://doi.org/10.1093/molbev/msl029.

[3] Raiford DW, Heizer EM, Miller R V., Akashi H, Raymer ML, Krane DE. Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? J Mol Evol 2008;67:621–30. https://doi.org/10.1007/s00239-008-9162-9.

[4] Harrison RJ, Charlesworth B. Biased gene conversion affects patterns of codon usage and amino acid usage in the saccharomyces sensu stricto group of yeasts. Mol Biol Evol 2011;28:117–29. https://doi.org/10.1093/molbev/msq191.

[5] Krick T, Verstraete N, Alonso LG, Shub DA, Ferreiro DU, Shub M, et al. Amino acid metabolism conflicts with protein diversity. Mol Biol Evol 2014. https://doi.org/10.1093/molbev/msu228.

[6] Wu C-I, Wang H-Y, Ling S, Lu X. The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process. Annu Rev Genet 2016;50:347–69. https://doi.org/10.1146/annurev-genet-112414-054842.

[7] Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011;144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

[8] Miranda A, Hamilton PT, Zhang AW, Pattnaik S, Becht E, Mezheyeuski A, et al. Cancer stemness, intratumoral heterogeneity, and immune response across cancers. Proc Natl Acad Sci U S A 2019;116:9020–9. https://doi.org/10.1073/pnas.1818210116.

[9] Saygin C, Matei D, Majeti R, Reizes O, Lathia JD. Targeting Cancer Stemness in the Clinic: From Hype to Hope. Cell Stem Cell 2019;24:25–40. https://doi.org/10.1016/j.stem.2018.11.017.

[10] Milanovic M, Fan DNY, Belenki D, Däbritz JHM, Zhao Z, Yu Y, et al. Senescence-

associated reprogramming promotes cancer stemness. Nature 2018. https://doi.org/10.1038/nature25167.

[11] Peiris-Pagès M, Martinez-Outschoorn UE, Pestell RG, Sotgia F, Lisanti MP. Cancer stem cell metabolism. Breast Cancer Res 2016. https://doi.org/10.1186/s13058-016-0712-6.

[12] Bellacosa A. Developmental disease and cancer: Biological and clinical overlaps. Am J Med Genet Part A 2013;161:2788–96. https://doi.org/10.1002/ajmg.a.36267.

[13] Aiello NM, Stanger BZ. Echoes of the embryo: Using the developmental biology toolkit to study cancer. DMM Dis Model Mech 2016;9:105–14. https://doi.org/10.1242/dmm.023184.

[14] Kamel HFM, Al-Amodi HSAB. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics, Proteomics Bioinforma 2017;15:220–35. https://doi.org/10.1016/j.gpb.2016.11.005.

[15] Alkhateeb A, Rezaeian I, Singireddy S, Cavallo-Medved D, Porter LA, Rueda L. Transcriptomics Signature from Next-Generation Sequencing Data Reveals New Transcriptomic Biomarkers Related to Prostate Cancer. Cancer Inform 2019;18. https://doi.org/10.1177/1176935119835522.

[16] Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med 2019;25:751–8. https://doi.org/10.1038/s41591-019-0424-4.

[17] Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204–13. https://doi.org/10.1038/nature24277.

[18] Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 2018;173:291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

[19] Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–625.

[20] Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature 2020;578:82–93. https://doi.org/10.1038/s41586-020-1969-6.

[21] Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature 2017;548:297–303. https://doi.org/10.1038/nature23306.

[22] Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014;158:929–44. https://doi.org/10.1016/j.cell.2014.06.049.

[23] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science (80- ) 2017;357. https://doi.org/10.1126/science.aan2507.

[24] Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature 2019;571:505–9. https://doi.org/10.1038/s41586-019-1338-5.

[25] Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:1–11. https://doi.org/10.1038/ncomms9971.

[26] Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun 2017;8:1–13. https://doi.org/10.1038/s41467-017-01027-z.

[27] Zhang H, Wang Y, Li J, Chen H, He X, Zhang H, et al. Biosynthetic energy cost for amino acids decreases in cancer evolution. Nat Commun 2018;9:1–15. https://doi.org/10.1038/s41467-018-06461-1.

[28] Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, et al. Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. Cancer Discov 2013;3:1108–12. https://doi.org/10.1158/2159-8290.CD-13-0219.

[29] Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature 2019;567:257–61. https://doi.org/10.1038/s41586-019-0987-8.

[30] Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell 2019;179:561-577.e22. https://doi.org/10.1016/j.cell.2019.08.052.

[31] Goodarzi H, Elemento O, Tavazoie S. Revealing Global Regulatory Perturbations across Human Cancers. Mol Cell 2009;36:900–11. https://doi.org/10.1016/j.molcel.2009.11.016.

[32] López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell 2013;153:1194. https://doi.org/10.1016/j.cell.2013.05.039.

[33] Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of

606 connectivity in cancer co-expression networks. PLoS One 2014;9.
607 https://doi.org/10.1371/journal.pone.0087075.
608 [34] Han R, Huang G, Wang Y, Xu Y, Hu Y, Jiang W, et al. Increased gene expression
609 noise in human cancers is correlated with low p53 and immune activities as well as
610 late stage cancer. Oncotarget 2016;7:72011–20.
611 https://doi.org/10.18632/oncotarget.12457.
612 [35] Yang RY, Quan J, Sodaei R, Aguet F, Segrè A V., Allen JA, et al. A systematic survey
613 of human tissue-specific gene expression and splicing reveals new opportunities for
614 therapeutic target identification and evaluation. BioRxiv 2018:311563.
615 https://doi.org/10.1101/311563.
616 [36] Gould SJ. Ontogeny and phylogeny--revisited and reunited. Bioessays 1992;14:275–9.
617 https://doi.org/10.1002/bies.950140413.
618 [37] Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, et al. Gene
619 expression divergence recapitulates the developmental hourglass model. Nature
620 2010;468:811–6. https://doi.org/10.1038/nature09634.
621 [38] Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors
622 ontogenetic divergence patterns. Nature 2010;468:815–9.
623 https://doi.org/10.1038/nature09632.
624 [39] Davies PCW, Lineweaver CH. Cancer tumors as Metazoa 1.0: Tapping genes of
625 ancient ancestors. Phys Biol 2011;8. https://doi.org/10.1088/1478-3975/8/1/015001.
626 [40] Lineweaver CH, Davies PCW, Vincent MD. Targeting cancer's weaknesses (not its
627 strengths): Therapeutic strategies suggested by the atavistic model. BioEssays
628 2014;36:827–35. https://doi.org/10.1002/bies.201400070.
629 [41] Chen H, Lin F, Xing K, He X. The reverse evolution from multicellularity to
630 unicellularity during carcinogenesis. Nat Commun 2015;6:1–9.
631 https://doi.org/10.1038/ncomms7367.
632 [42] Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Altered interactions between
633 unicellular and multicellular genes drive hallmarks of transformation in a diverse range
634 of solid tumors. Proc Natl Acad Sci U S A 2017;114:6406–11.
635 https://doi.org/10.1073/pnas.1617743114.
636 [43] Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Somatic mutations in early
637 metazoan genes disrupt regulatory links between unicellular and multicellular genes in
638 cancer. Elife 2019;8:1–28. https://doi.org/10.7554/eLife.40947.
639 [44] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression

estimation with read mapping uncertainty. Bioinformatics 2009;26:493–500. https://doi.org/10.1093/bioinformatics/btp692.

[45] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8. https://doi.org/10.1093/bioinformatics/btw354.

[46] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14:417–9. https://doi.org/10.1038/nmeth.4197.

[47] Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. F1000Research 2016;4:1521. https://doi.org/10.12688/F1000RESEARCH.7563.2.

[48] Nueda MJ, Tarazona S, Conesa A. Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics 2014;30:2598–602. https://doi.org/10.1093/bioinformatics/btu333.

## Figure legends

**Figure 1 Pan-cancer convergence of transcriptome-based amino acid usage**

**A**. Schematic diagram showing the computation of amino acid usage frequency based on the gene expression profile derived from an RNA-seq sample. t-SNE projection of the GTEx (**B**), developing mammalian tissue (**C**), and TCGA tumor samples (**D**) based on their amino acid frequency profiles. Samples are color-coded based on tissue or cancer types. Marker shapes correspond to species. Developmental stages were classified into three categories and indicated by marker size. All t-SNE projections were generated using sklearn TSNE, with perplexity as 30, learning rate as 200, and the number of iterations as 1,000. Comparison of amino acid usage convergence index between tumor samples and either matched down-sampled normal samples (**E**) or adjacent normal samples (**F**) across multiple cancer types. Box plots show the quartiles, and the whiskers indicate quartile ± 1.5 × interquartile range. A two-sided Mann-Whitney U-test was used to calculate the p-value. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

**Figure 2 Amino acid usage preference in tumor evolution as quantified by $ECPA_{cell}$**

**A**. Correlation between the biosynthetic energy cost of an amino acid and the net number of cancer types with significantly increased usage across 20 amino acids. The net number is

673 defined as the number of cancer types with significantly increased usage of the amino acid
674 minus the number with significantly decreased usage. The colored region around the
675 regression lines indicates a 95% confidence interval. **B**. Schematic diagram showing the
676 computation of $ECPA_{gene}$ and $ECPA_{cell}$ based on the gene expression profile derived from
677 RNA-seq data. **C**. $ECPA_{cell}$ of tumor samples and matched normal tissue samples across
678 TCGA cancer types. A paired two-sided Wilcoxon signed-rank test was used to calculate the
679 p values. **D**. Bar plots showing $ECPA_{gene}$ values of significantly down- and up-regulated
680 proteins in several cancer proteomics datasets. Error bars denote 95% confidence intervals. A
681 two-sided Mann–Whitney U-test was used to calculate the p values. **E**. Correlation between
682 $ECPA_{cell}$ and amino acid usage convergence index across samples in nine cancer types. The
683 colored regions around the regression lines indicate 95% confidence intervals. *p < 0.05, **p
684 < 0.01, ***p < 0.001.
685
686 **Figure 3 The increasing trend of $ECPA_{cell}$ throughout mammalian organogenesis**
687 Trend lines of $ECPA_{cell}$ during the development of the liver (**A**), and the kidney (**B**) across
688 five mammalian species. Developmental stages of non-mouse species correspond to the
689 mouse stages shown in brackets. Error bars denote 95% confidence intervals. The trend line
690 of $ECPA_{cell}$ along the developmental trajectory of the mouse liver across three independent
691 datasets (**C**-**E**). Error bars denote 95% confidence intervals. Heatmaps showing enrichment
692 patterns of gene modules that contribute to $\Delta ECPA_{cell}$ during the development of the human
693 liver (**F**) and the human kidney (**G**). The red stripes embedded in the black background on
694 top of each heatmap designate the range of $\Delta ECPA_{cell}$ contribution index within every bin.
695
696 **Figure 4 A proposed model unifying developmental reversal, amino acid usage**
697 **convergence, and $ECPA_{cell}$ decline of cancer samples**
698 Stacked bar plots showing the proportion of genes that positively or negatively contribute to
699 $\Delta ECPA_{cell}$ in either tumorigenesis or development for LIHC-liver (**A**), KIRC-kidney (**B**), and
700 KIRP-kidney (**C**). Scatter plots showing, for genes with negative $\Delta ECPA_{cell}$ contribution
701 index in tumorigenesis and positive $\Delta ECPA_{cell}$ contribution index in tissue development,
702 scaled $\Delta ECPA_{cell}$ contribution index in tumorigenesis versus scaled $\Delta ECPA_{cell}$ contribution
703 index in tissue development for LIHC-liver (**D**), KIRC-kidney (**E**), and KIRP-kidney (**F**).
704 Colored regions around the regression lines indicate 95% confidence intervals. Kaplan-Meier
705 plots show the overall survival for patients with LIHC (**G**), KIRC (**H**), or KIRP (**I**) stratified
706 by developmental reversal index into two equal groups, respectively. The p values were

calculated from two-sided log-rank tests. Multivariate linear regression of $ECPA_{cell}$ with developmental reversal index and amino acid usage convergence index as dependent variables for LIHC-liver (**J**), KIRC-kidney (**K**), and KIRP-kidney (**L**). **M**. Cartoon depicting a conceptual model in which cancer evolution is accompanied by the convergence of amino acid usage and decrease of $ECPA_{cell}$, which is a reversal of the tissue development process.

**Figure 5 The liver shows the most dramatic $ECPA_{cell}$ reduction in tumorigenesis**

Distributions of $\Delta ECPA_{cell}$ between tumor samples and paired NAT samples across multiple cancer types (**A**), tissue-specific genes-based $ECPA_{cell}$ of normal samples across multiple tissues (**B**), tissue-specific genes-based $ECPA_{cell}$ of adjacent normal samples across multiple cancer types (**C**), ranked by the median values. The box plots show the quartiles. The whiskers indicate quartile ± 1.5× interquartile range. The horizontal dashed line indicates the level of $\Delta ECPA_{cell} = 0$. **D**. Trend lines of $ECPA_{cell}$ of multiple tissues across human developmental stages. Error bars denote 95% confidence interval. wpc, weeks post conception.

**Figure 6 $ECPA_{cell}$ is a robust diagnostic biomarker for liver cancer**

**A**. $ECPA_{cell}$ of tumor samples and matched normal tissue samples in 11 independent RNA-seq datasets of liver cancer and their matched normal samples. A paired two-sided Wilcoxon signed-rank test was used to calculate the p values. **B**. ROC curves of $ECPA_{cell}$ as a diagnostic biomarker in six independent liver cancer cohorts with sample size ≥12. Dashed lines indicate the lines of identity. ROC, receiver operating characteristic; AUC, area under the ROC curve. **C**. Histogram showing the distribution of the average AUC across the six cohorts for tumor-normal segregation using the mRNA expression level of each of the 9,559 detectable genes. The vertical dashed line corresponds to the average AUC of $ECPA_{cell}$. **D**. Box plots showing the AUC of the top four metrics, including three genes and $ECPA_{cell}$, in discriminating tumor samples from normal samples across the six cohorts. A paired two-sided Wilcoxon signed-rank test was used to calculate the p values. **E**. Box plots showing the AUC of $ECPA_{cell}$ and the frequency of each amino acid in detecting tumors across the six cohorts. The box plots show the quartiles. The whiskers indicate quartile ± 1.5× interquartile range. A paired two-sided Wilcoxon signed-rank test was used to calculate the p values. **F**. Histogram showing the distribution of coefficients of variation (CV) of the optimal thresholds in using individual genes for tumor-normal segregation. The vertical red dashed

740 line indicates the CV of $ECPA_{cell}$. Vertical lines in three other colors indicate the CV of three
741 genes whose average AUCs are higher than $ECPA_{cell}$. *p < 0.05, **p < 0.01, ***p < 0.001.
742
743
744

# Supplementary materials

**Figure S1 t-SNE projection of samples based on gene expression**

t-SNE projection of the GTEx (**A**), TCGA (**B**), PCAWG (**C**), and MET500 (**D**) samples based on their gene expression profiles.

**Figure S2 t-SNE projection of samples based on amino acid frequency**

t-SNE projection of the GTEx & HPA (**A**), PCAWG (**B**), MET500 (**C**), down-sampled GTEx (**D**), down-sampled TCGA tumor (**E**), matched TCGA NAT (**F**), and matched TCGA tumor (**G**) samples based on their amino acid frequency.

**Figure S3 Re-calculation of $ECPA_{cell}$ in TCGA samples without highly-expressed genes**

**A**. $ECPA_{cell}$ of tumor samples and matched normal tissue samples across TCGA cancer types after removal of genes encoding high-abundance housekeeping or tissue-specific proteins

**Figure S4 Differential amino acid usage within and between tumor and NAT samples**

Heatmaps showing the average frequency of individual amino acids for NAT samples (**A**) and tumor samples (**B**), normalized as z-scores, across 15 cancer types. **C**. Heatmap showing significantly increased or decreased usage of individual amino acids between tumor and NAT samples across 15 cancer types.

**Figure S5 Functional enrichment of genes with a positive contribution to $\Delta ECPA_{cell}$ in non-human tissue development**

Heatmaps showing enrichment patterns of well-defined gene modules that contribute to $ECPA_{cell}$ increase during the development of the mouse liver (**A**) and kidney (**B**), the rat liver (**C**) and kidney (**D**), and the rabbit liver (**E**) and kidney (**F**). Red stripes embedded in the black background on top of each heatmap designate the range of $\Delta ECPA_{cell}$ contribution index within every bin.

**Figure S6 Variations of $ECPA_{cell}$ during tissue aging**

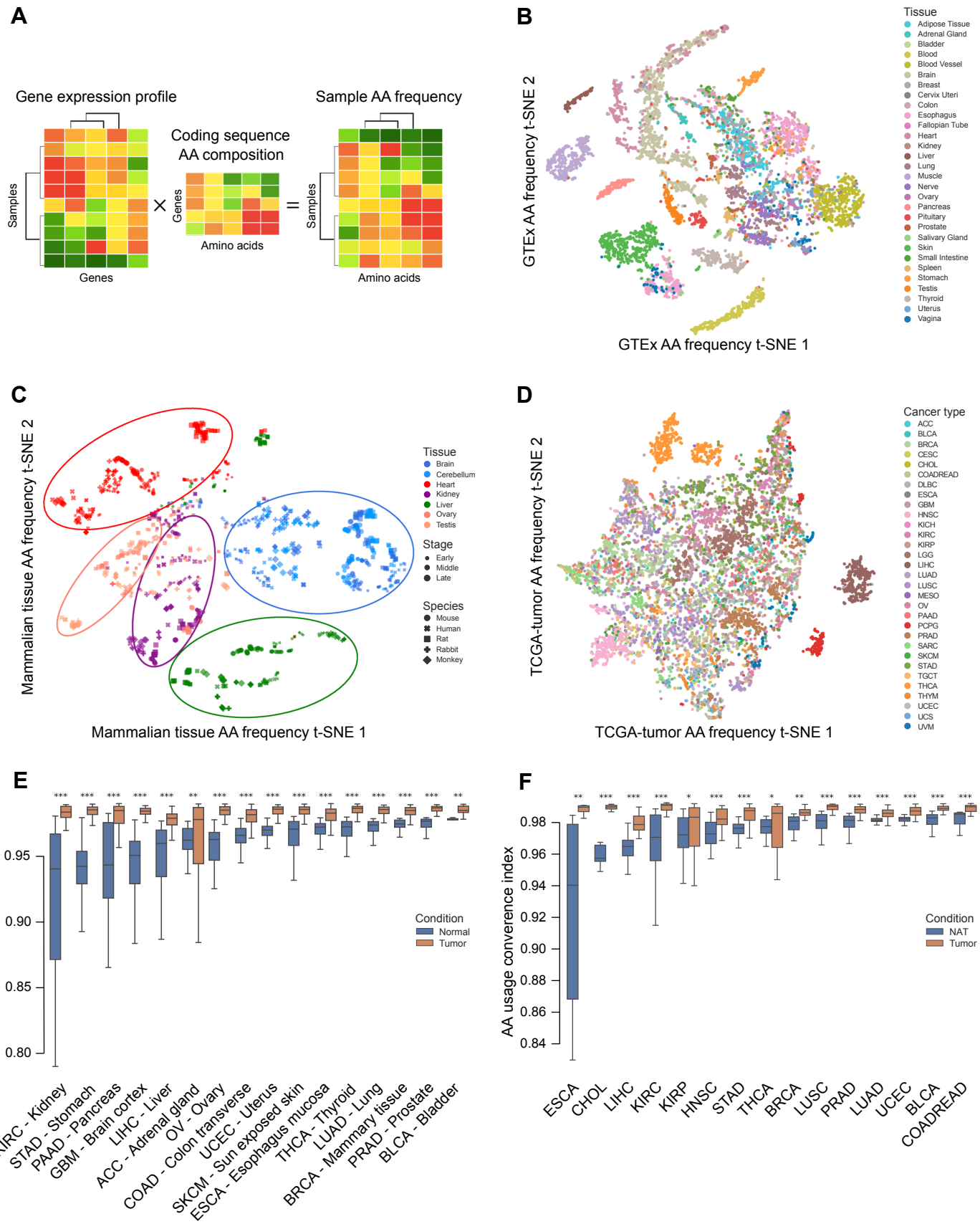
Trend lines of $ECPA_{cell}$ during aging of the liver and the kidney in humans (**A**), mice (**B**), and rats (**C**). Blue and orange dashed lines indicate the average levels of $ECPA_{cell}$ across age groups for the liver and the kidney, respectively. Error bars denote 95% confidence intervals.

778 **Figure S7 Variations of $ECPA_{cell}$ during the development of multiple tissues in non-**
779 **human mammals**

780 Trend lines of $ECPA_{cell}$ in seven tissues across four mammals, including mice (**A**), rats (**B**),
781 rabbits (**C**), and opossums (**D**). Error bars denote 95% confidence intervals.

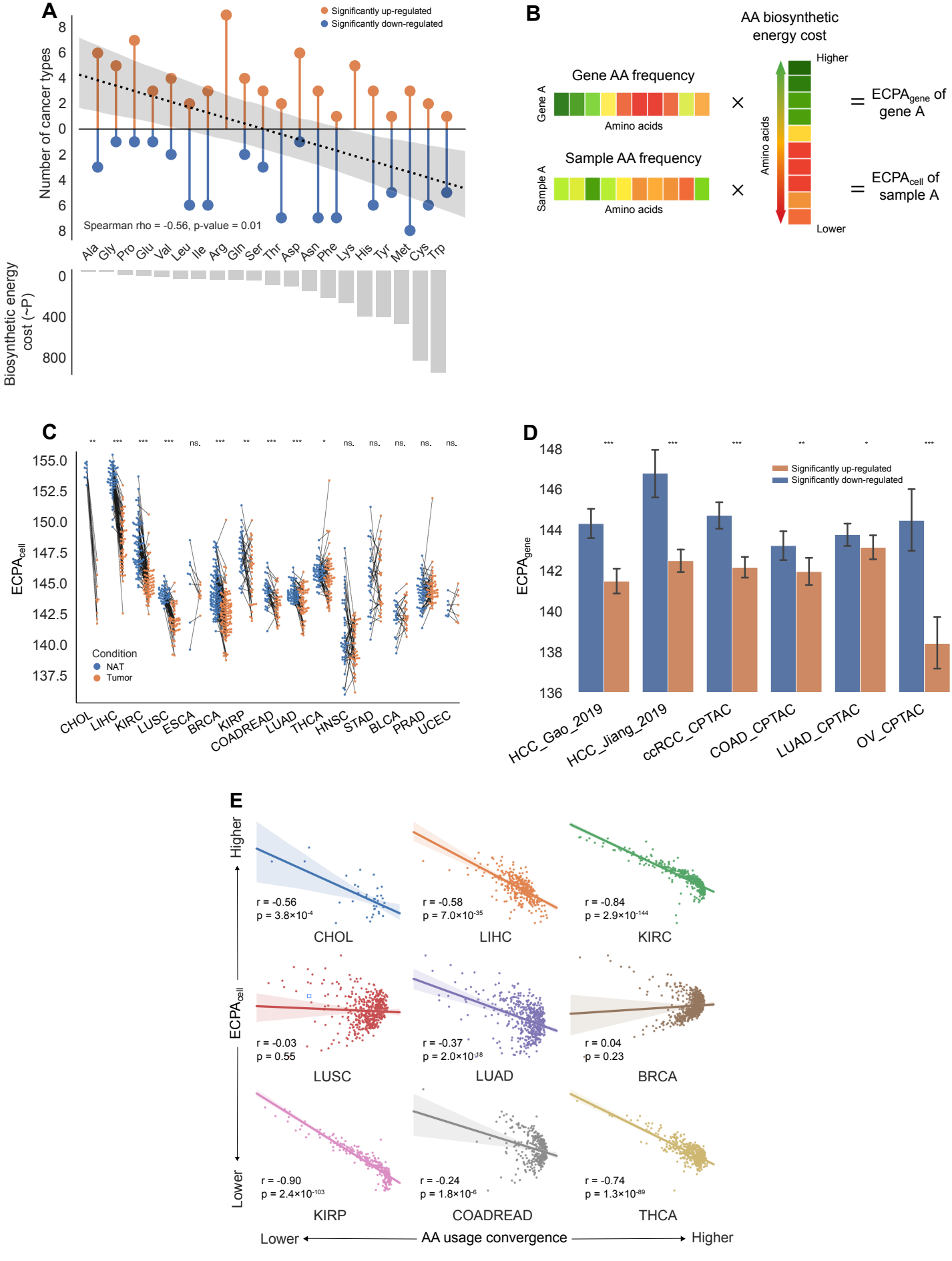
# Figure 1





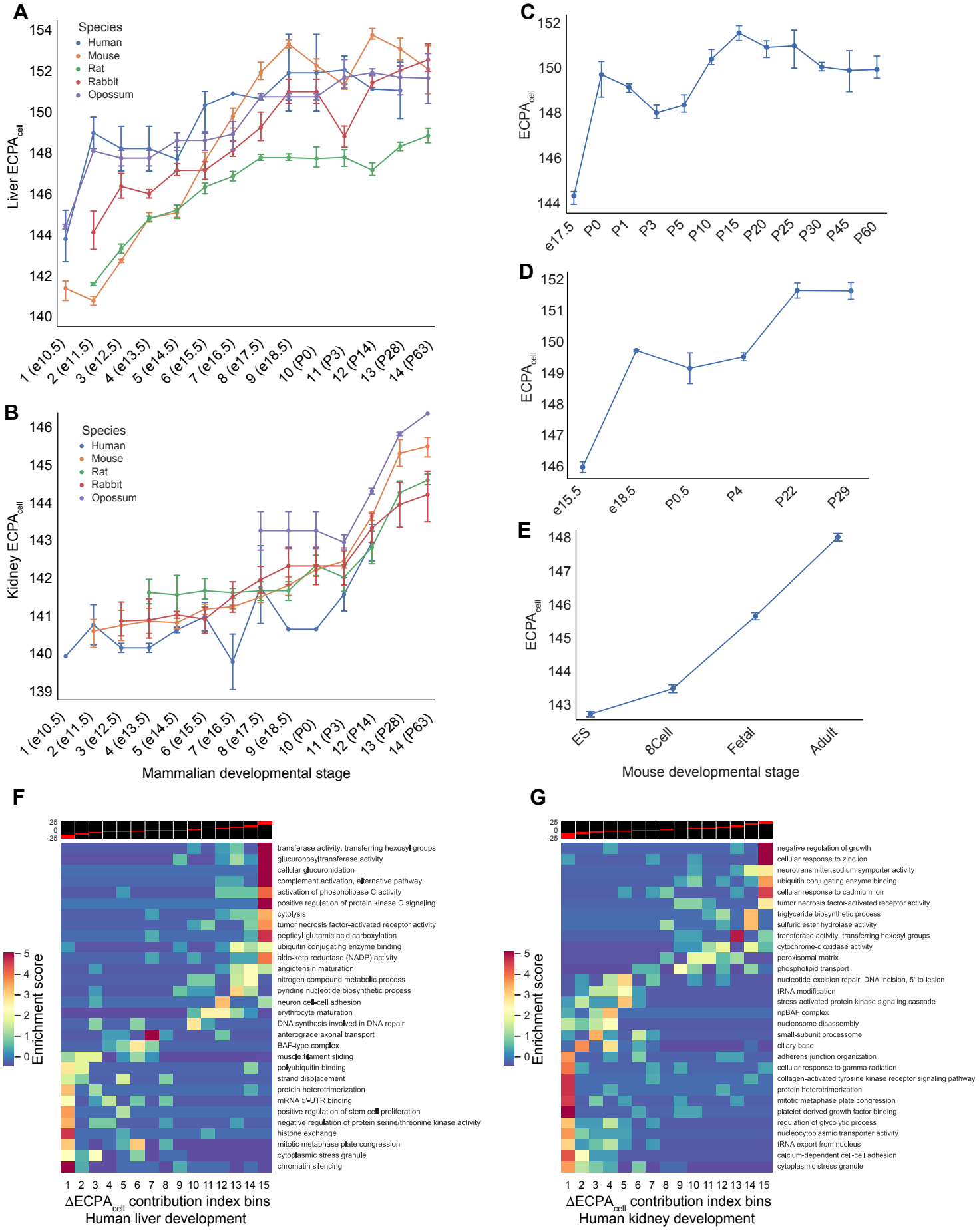
# Figure 2





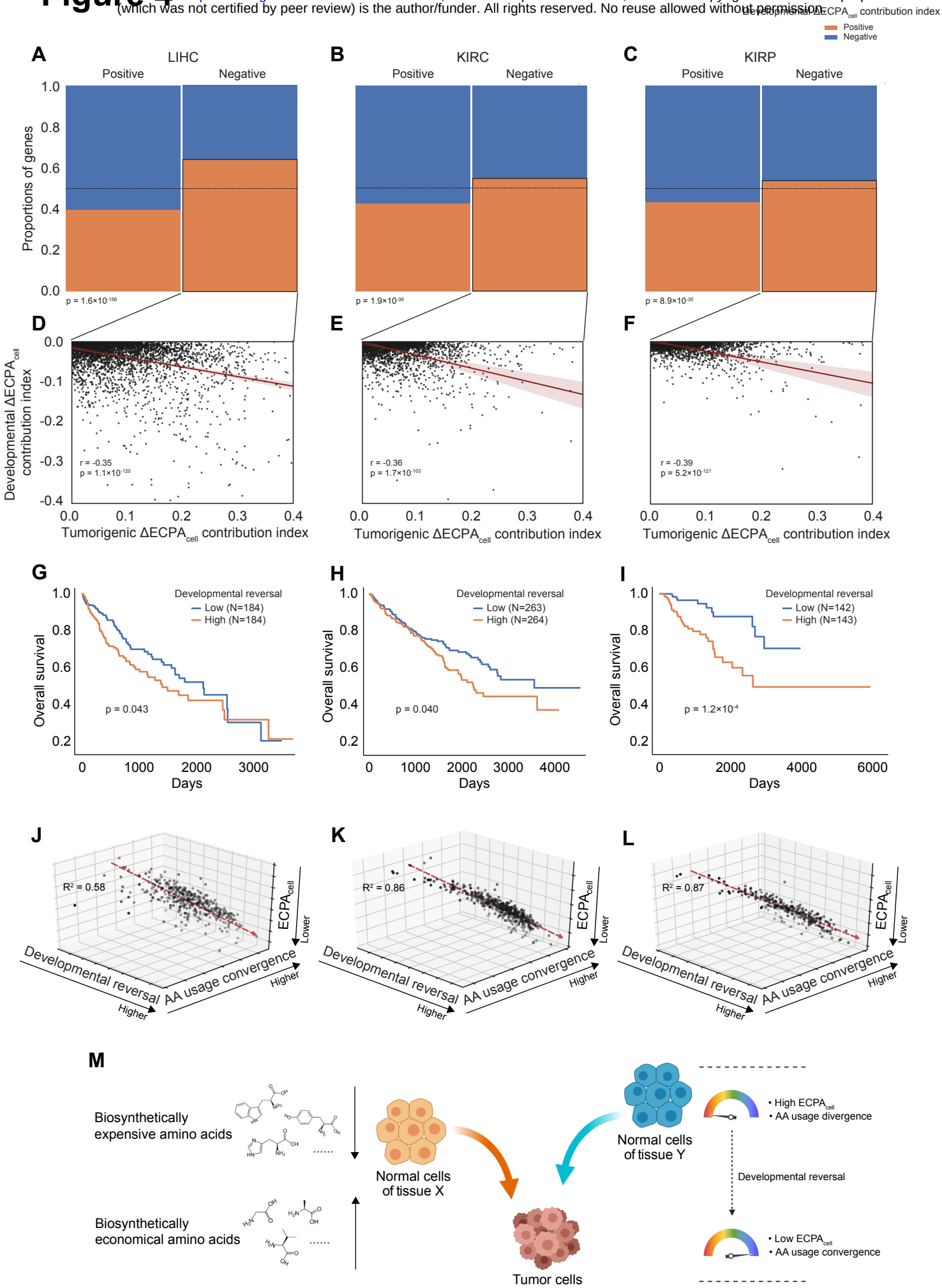
# Figure 3





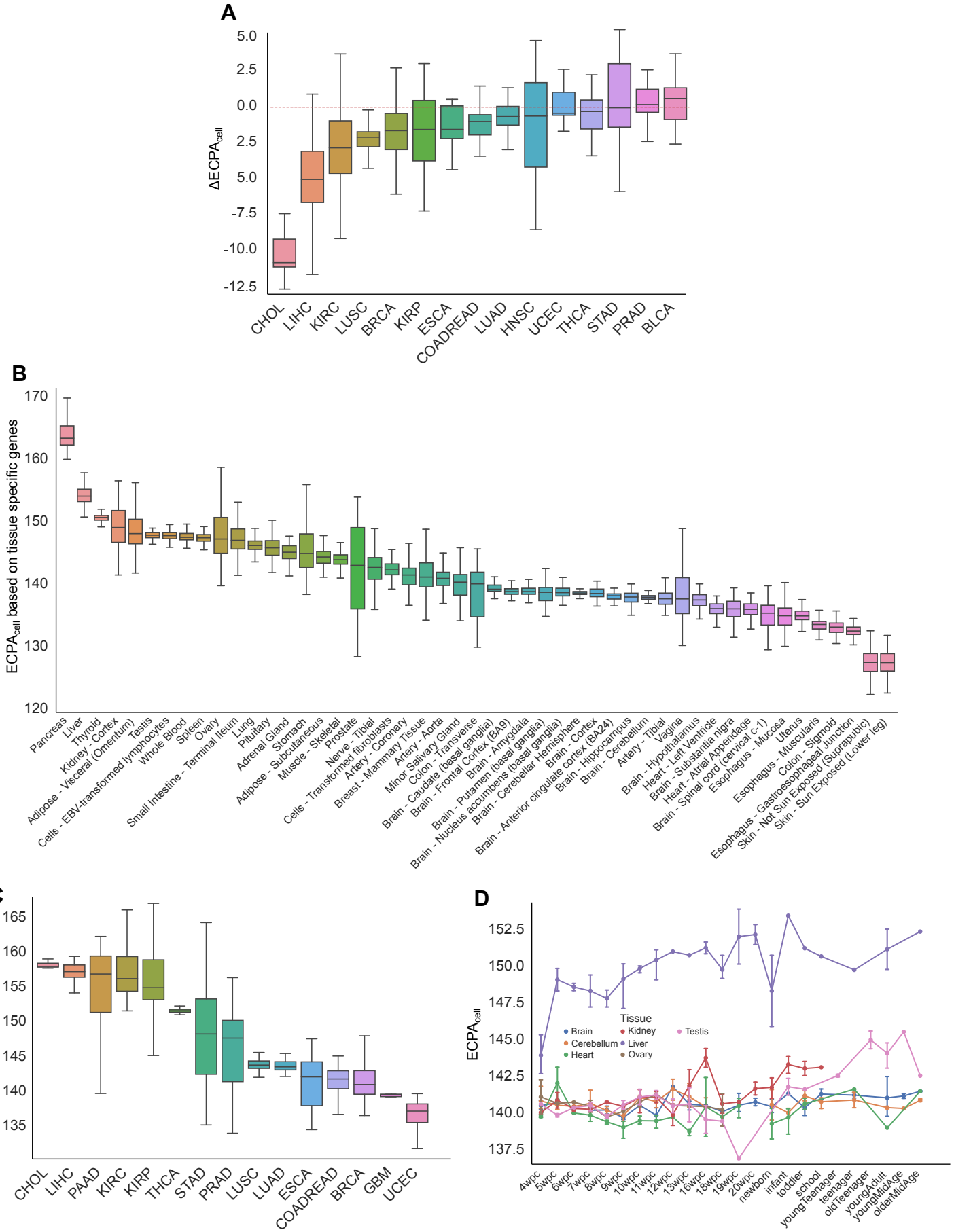
# Figure 4





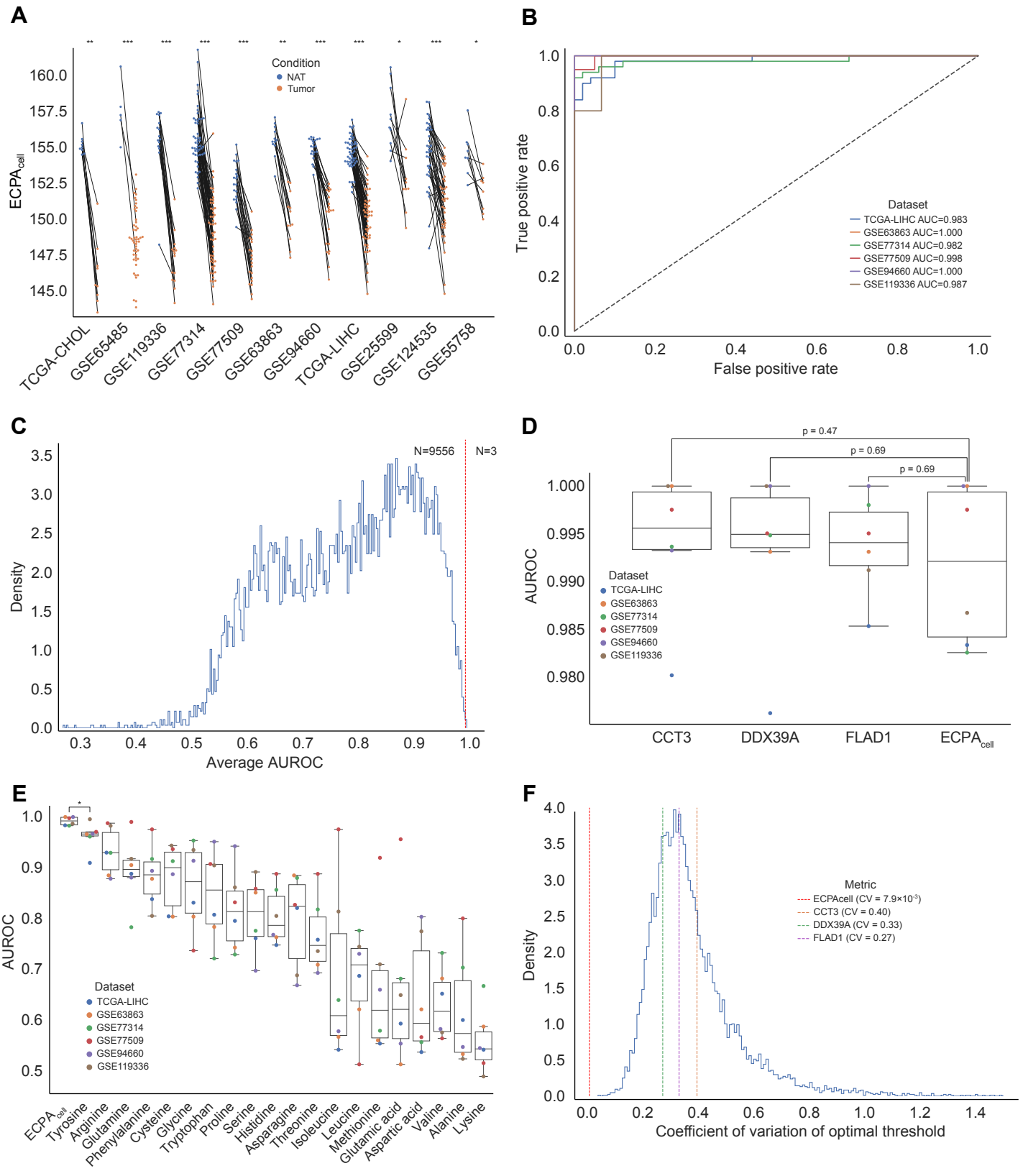
# Figure 5





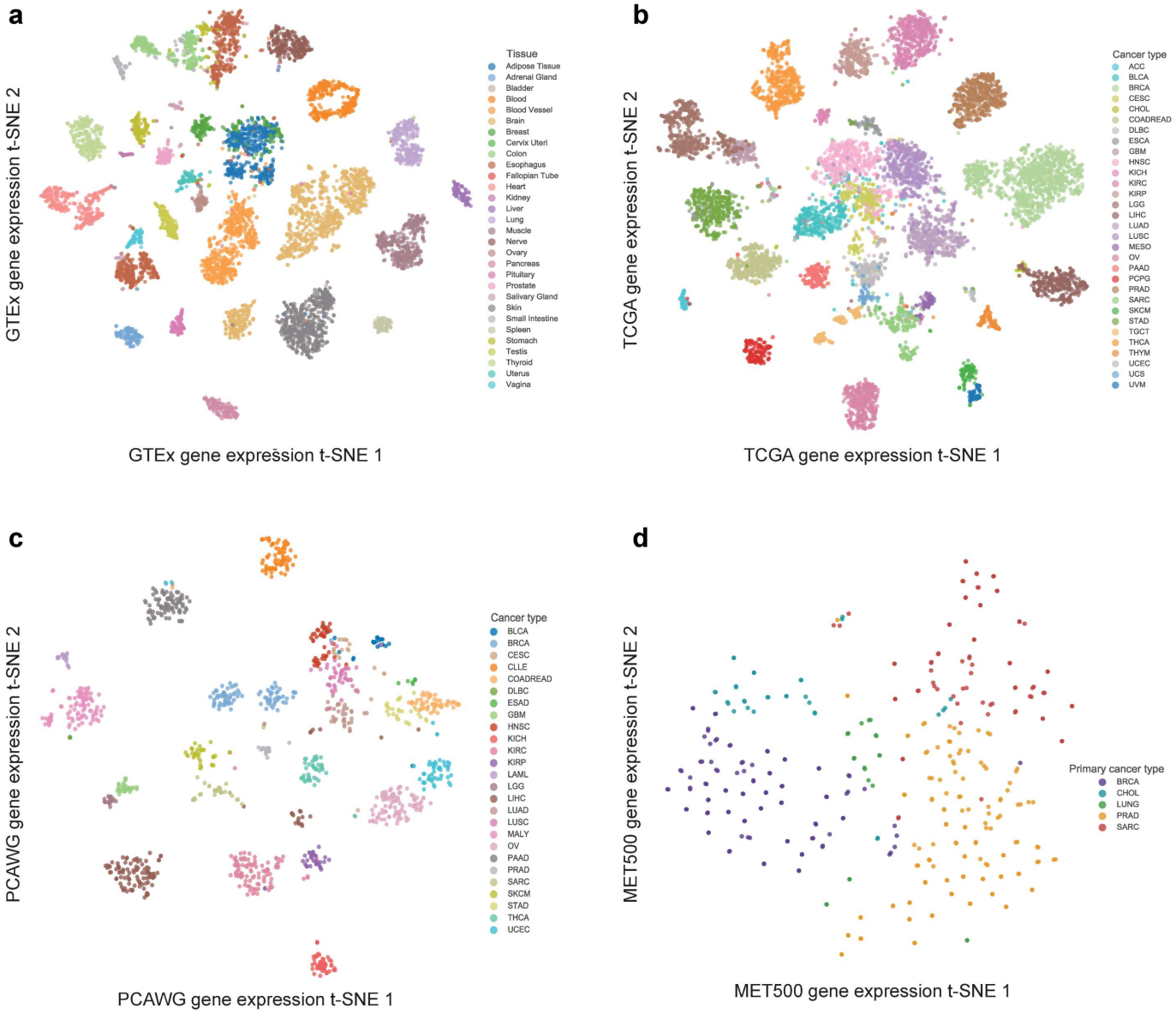
# Figure 6





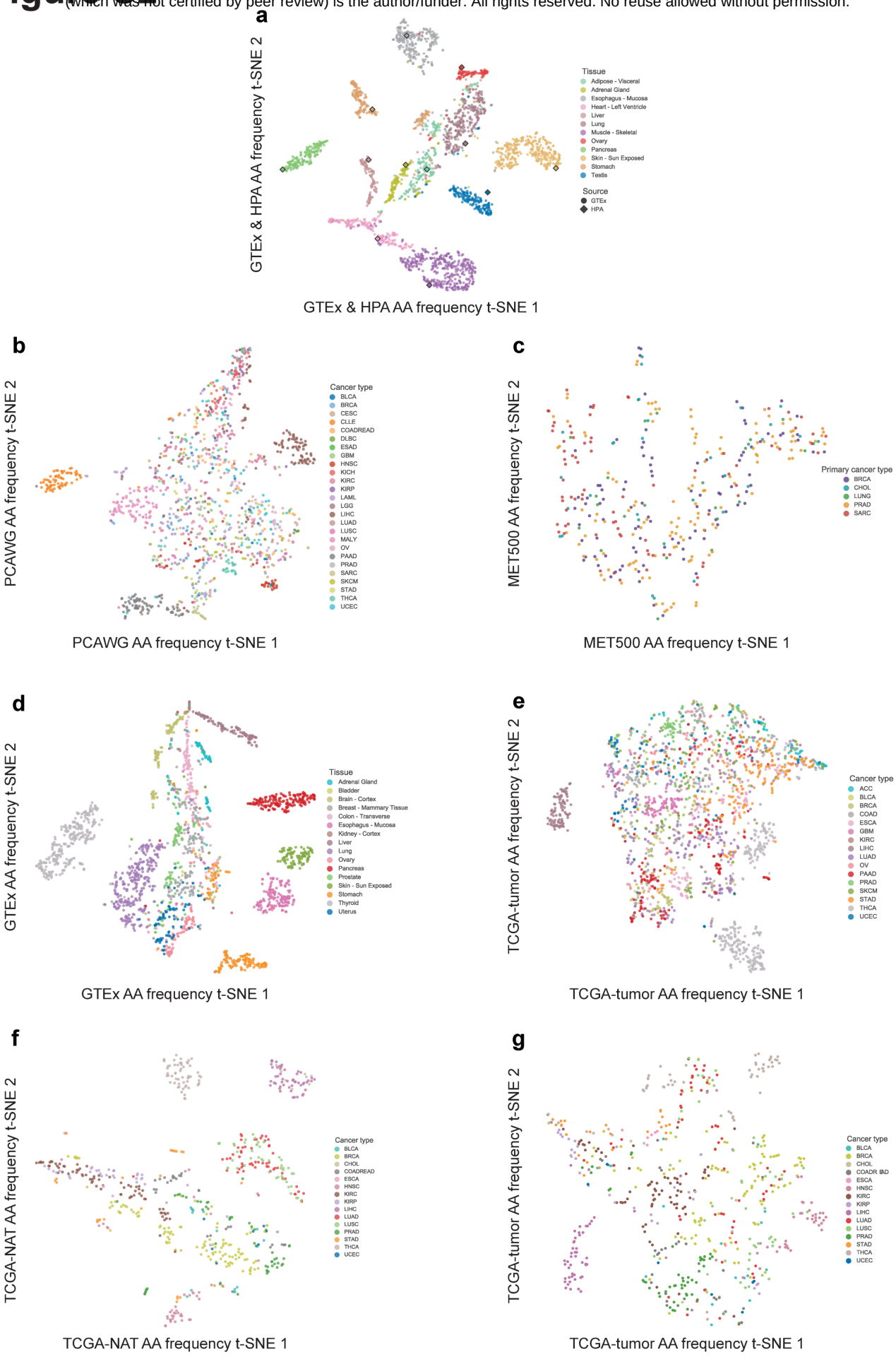
# Figure S1





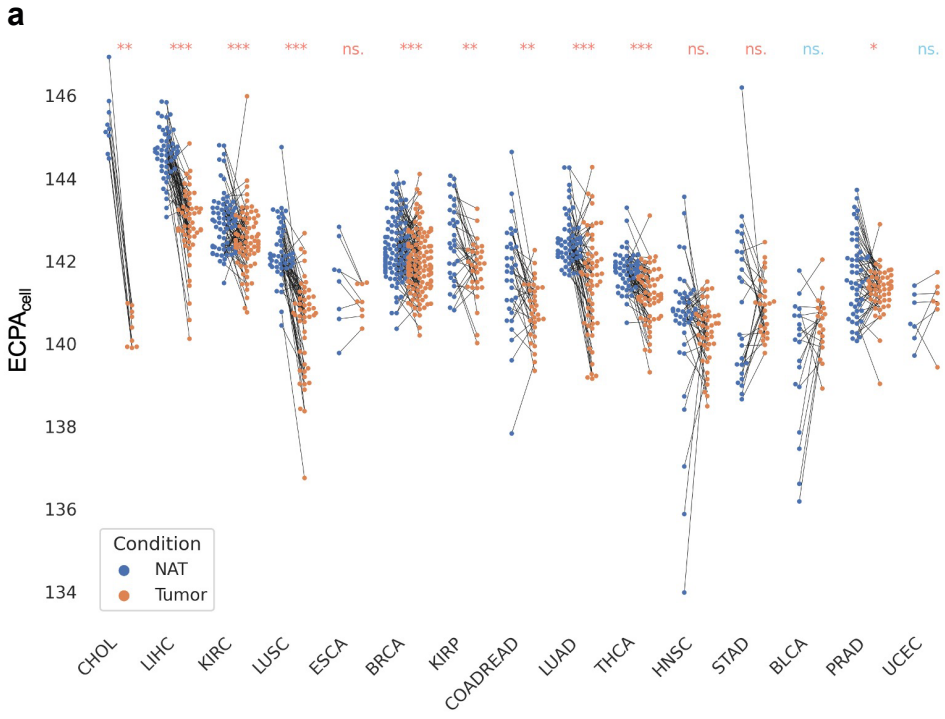
# Figure S2



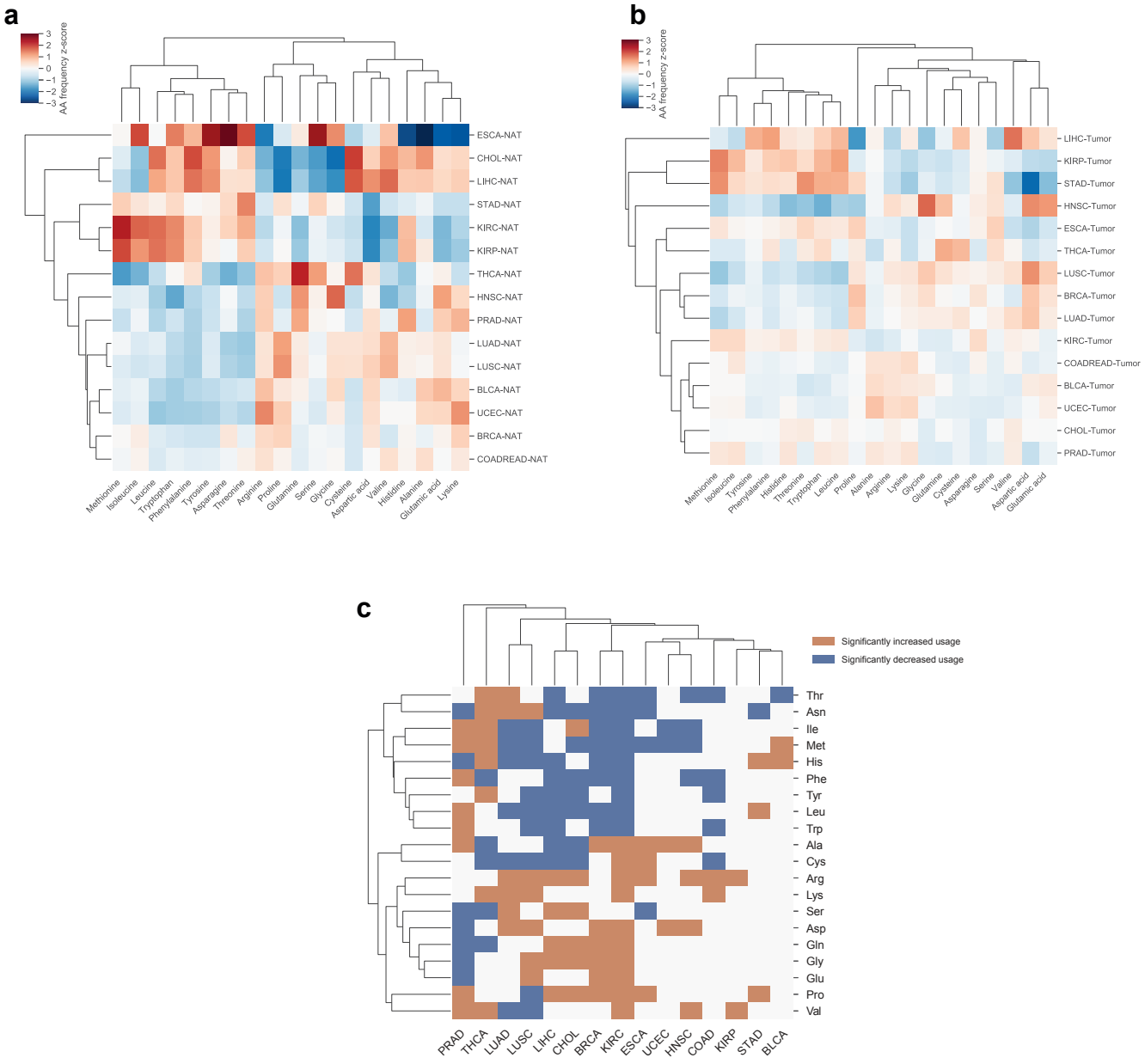


# Figure S3

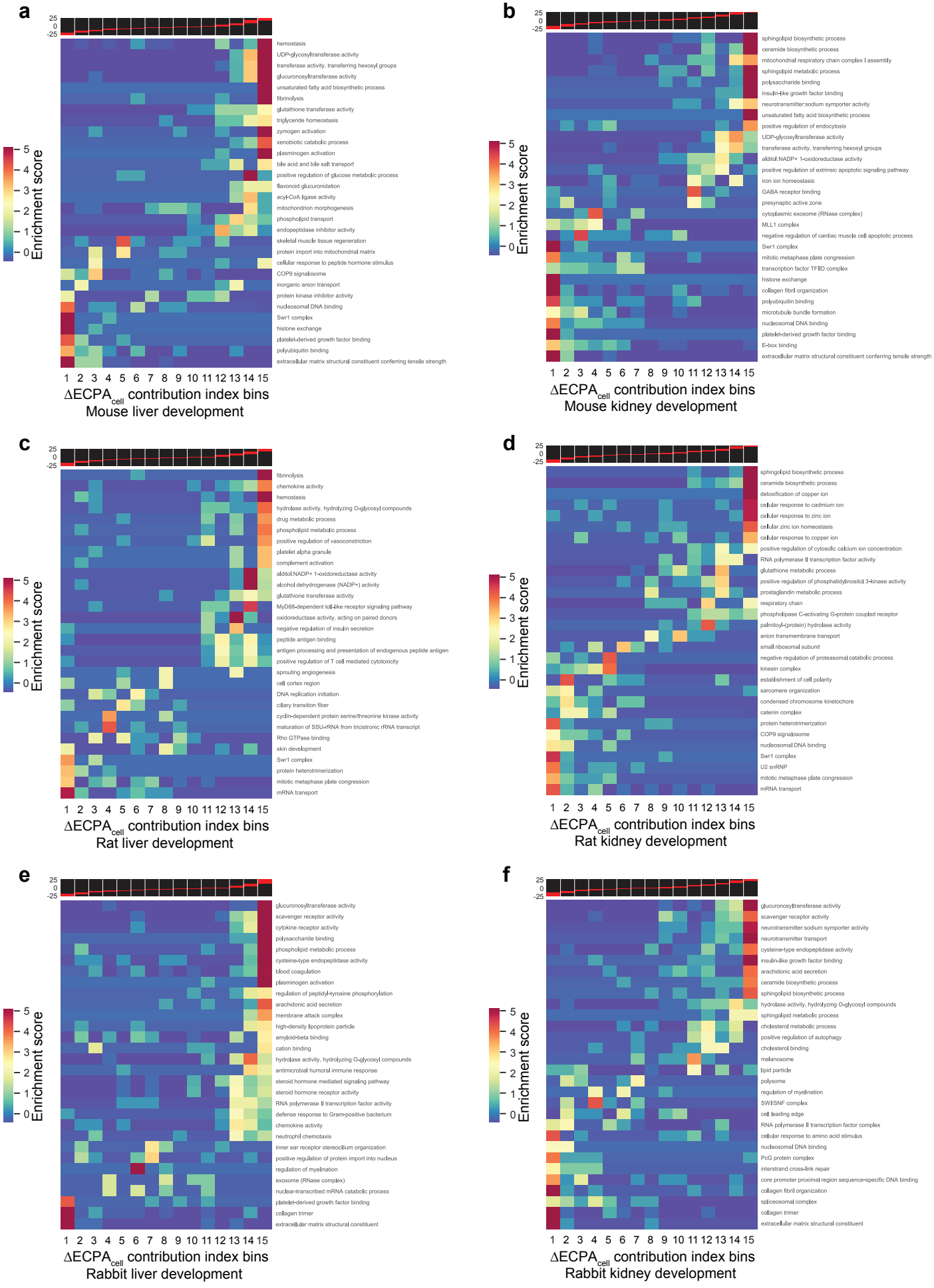




# Figure S4

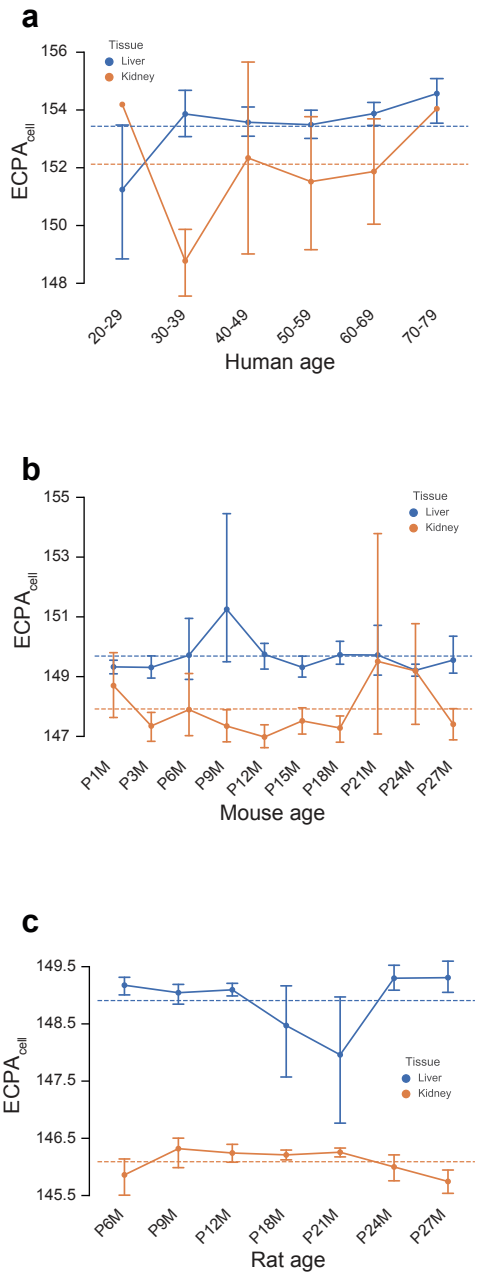


# Figure S5



# Figure S6





# Figure S7