Genomics Proteomics Bioinformatics 20 (2022) 147–162

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

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

Yikai Luo [1,2], Han Liang [1,2,3,*]

[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*

Received 21 December 2020; revised 13 July 2021; accepted 26 August 2021
Available online 4 September 2021

Handled by Zhang Zhang

KEYWORDS

Amino acid usage;
Tissue development;
Biosynthetic energy;
Diagnostic biomarker

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 The Cancer Genome Atlas (TCGA), the Genotype-Tissue Expression (GTEx), and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), and found that compared to normal tissues, different cancer types showed a convergent pattern toward 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 ECPA represents a powerful biomarker that could distinguish liver tumors from normal liver samples consistently across 11 independent patient cohorts and outperforms 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.

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 analyses of the genomes or transcriptomes of many species,

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

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.08.004

amino acid usage preference is a well-studied topic in macroevolution. The universal trend of "cost–usage anti-correlation" 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 evolves 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 in 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, 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

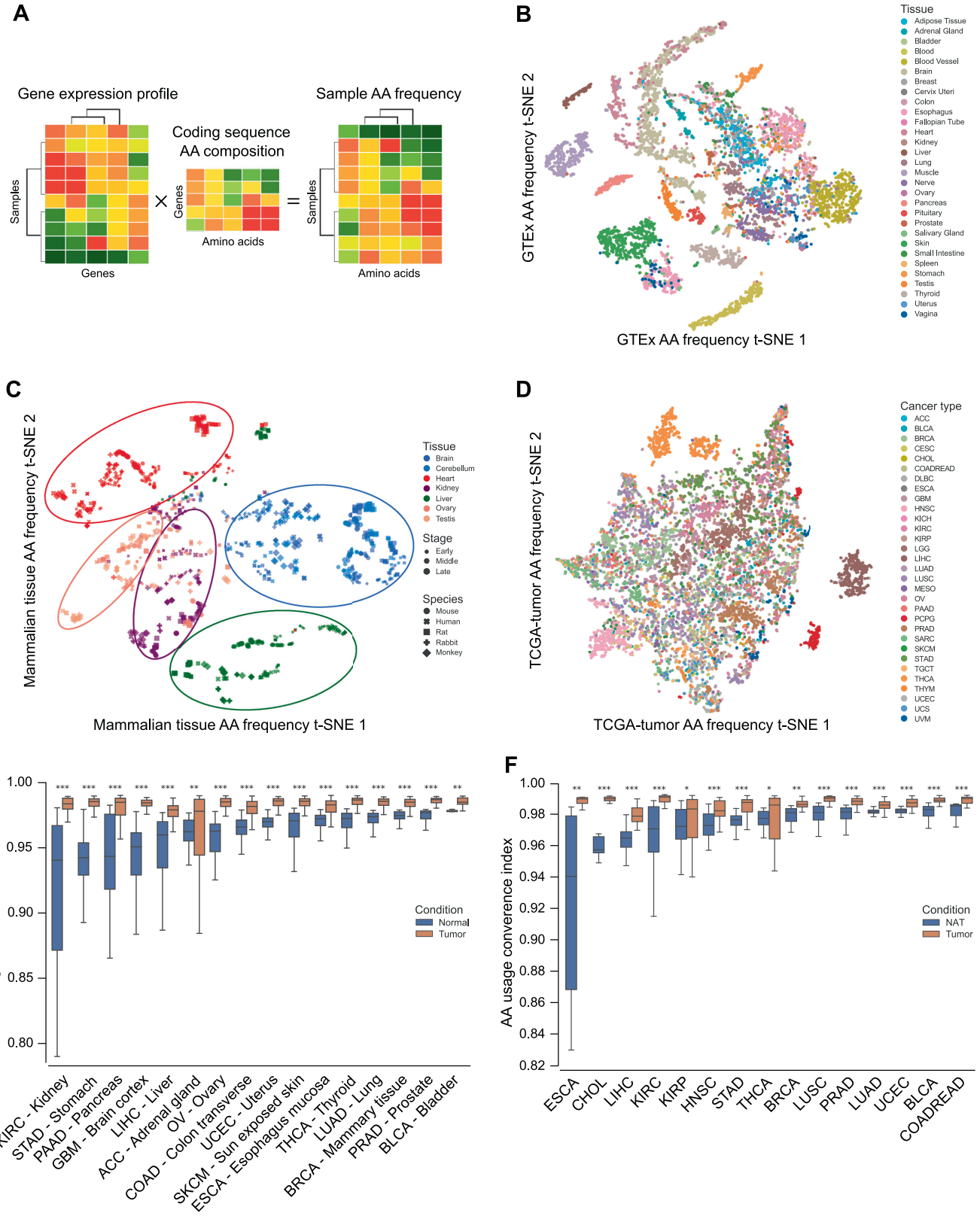
A convergence of amino acid usage across cancer types

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, 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) samples (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 samples [25,26].

The observation that amino acid usage for cancer cells failed to preserve their distinct tissue origins raised two possibilities: 1) cancer cells evolved to possess highly stochastic amino acid usage profiles both within and between cancer types; or 2) they went through a 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


A
Gene expression profile
Coding sequence AA composition
Sample AA frequency
Samples
Genes
Genes
Amino acids
Samples
Amino acids
B
GTEx AA frequency t-SNE 2
GTEx AA frequency t-SNE 1
Tissue
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Brain
Breast
Cervix Uteri
Colon
Esophagus
Fallopian Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
C
Mammalian tissue AA frequency t-SNE 2
Mammalian tissue AA frequency t-SNE 1
Tissue
Brain
Cerebellum
Heart
Kidney
Liver
Ovary
Testis
Stage
Early
Middle
Late
Species
Mouse
Human
Rat
Rabbit
Monkey
D
TCGA-tumor AA frequency t-SNE 2
TCGA-tumor AA frequency t-SNE 1
Cancer type
ACC
BLCA
BRCA
CESC
CHOL
COADREAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS
UVM
E
AA usage convergence index
1.00
0.95
0.90
0.85
0.80
Condition
Normal
Tumor
KIRC - Kidney
STAD - Stomach
PAAD - Pancreas
GBM - Brain cortex
LIHC - Liver
ACC - Adrenal gland
OV - Ovary
COAD - Colon transverse
UCEC - Uterus
SKCM - Sun exposed skin
ESCA - Esophagus mucosa
THCA - Thyroid
LUAD - Lung
BRCA - Mammary tissue
PRAD - Prostate
BLCA - Bladder
F
AA usage convergence index
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
Condition
NAT
Tumor
ESCA
CHOL
LIHC
KIRC
KIRP
HNSC
STAD
THCA
BRCA
LUSC
PRAD
LUAD
UCEC
BLCA
COADREAD

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 S3A 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 S3C). 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 acids and their 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, rho = −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 ($\log_2$ FC > 0, FDR < 0.05) had significantly lower $ECPA_{gene}$ than the proteins that were significantly down-regulated ($\log_2$ FC < 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 was insensitive to the expression of genes with extremely high abundance, including those encoding certain housekeeping proteins as well as tissue-specific proteins. After removing all genes that either encode cytoplasmic and mitochondrial ribosome proteins or rank among the top 200 genes in median transcripts per million (TPM) of the same cancer type, we recalculated the ECPA index for each sample and found that the decreasing pattern of $ECPA_{cell}$ in tumor samples across multiple cancer types was almost perfectly reproduced (Figure S4).

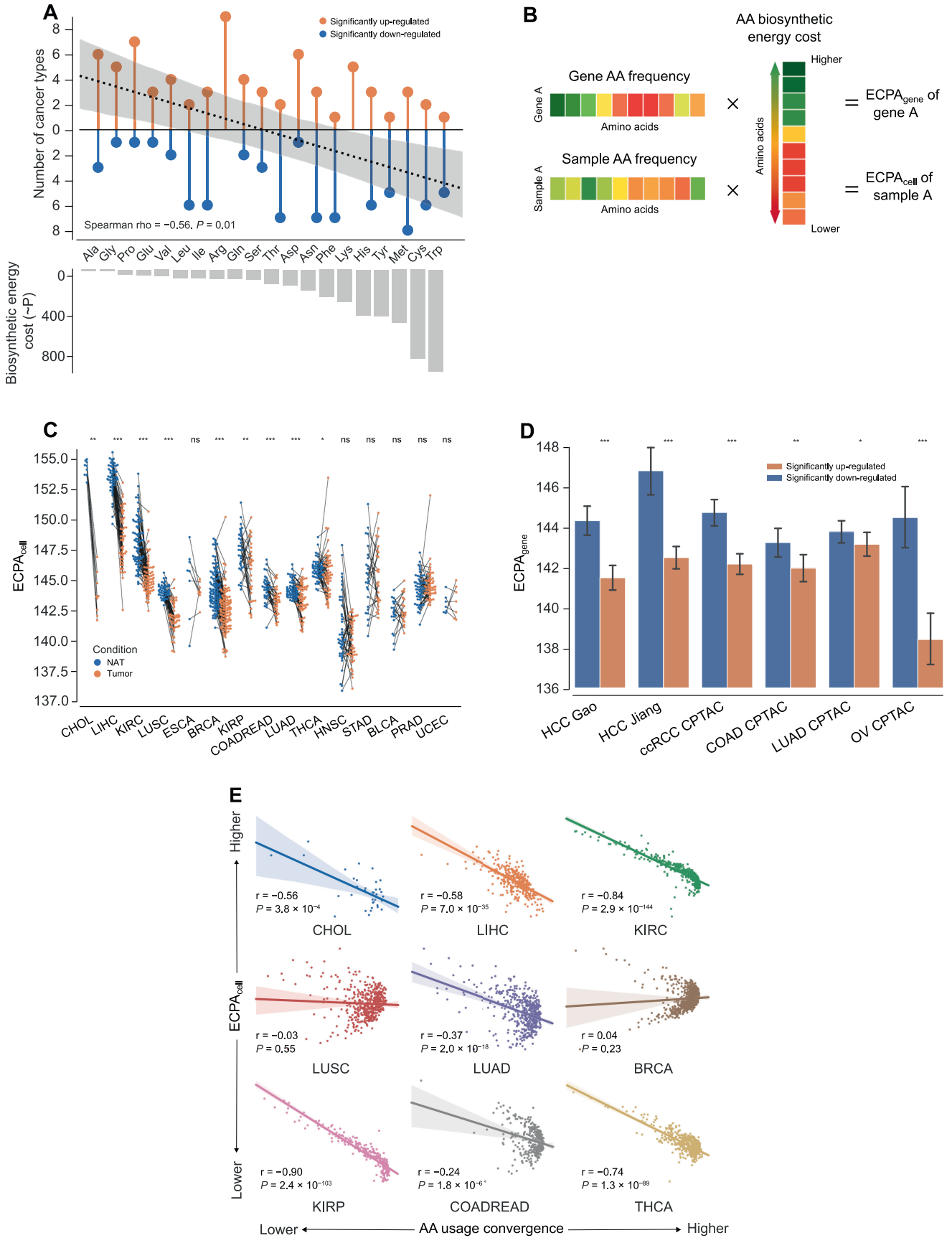
We next tested whether the amino acid usage convergence level of a tumor was correlated with its $ECPA_{cell}$. Indeed, we found a strong inverse relationship for seven out of the nine cancer types where $ECPA_{cell}$ was significantly lower in tumors (Figure 2E). Thus, the more a tumor follows a convergent 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, and 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

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. **B.**–**D.** t-SNE projection of the GTEx (B), developing mammalian tissue (C), and TCGA tumor (D) samples based on their amino acid frequency profiles. Samples are color-coded based on tissue or cancer types. In (C), 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 1000. **E.** Comparison of amino acid usage convergence index between tumor samples and matched down-sampled normal samples across multiple cancer types. **F.** Comparison of amino acid usage convergence index between tumor samples and adjacent normal samples 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* values. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. AA, amino acid; GTEx, Genotype-Tissue Expression; TCGA, The Cancer Genome Atlas; t-SNE, t-distributed stochastic neighborhood embedding.


A
Significantly up-regulated
Significantly down-regulated
Number of cancer types
Spearman rho = −0.56. P = 0.01
Ala Gly Pro Glu Val Leu Ile Arg Gln Ser Thr Asp Asn Phe Lys His Tyr Met Cys Trp
Biosynthetic energy cost (~P)
B
AA biosynthetic energy cost
Higher
Lower
Gene AA frequency
Gene A
Amino acids
Sample AA frequency
Sample A
Amino acids
Amino acids
= ECPA$_{gene}$ of gene A
= ECPA$_{cell}$ of sample A
C
ECPA$_{cell}$
Condition
NAT
Tumor
CHOL LIHC KIRC LUSC ESCA BRCA KIRP COADREAD LUAD THCA HNSC STAD BLCA PRAD UCEC
D
ECPA$_{gene}$
Significantly up-regulated
Significantly down-regulated
HCC Gao
HCC Jiang
ccRCC CPTAC
COAD CPTAC
LUAD CPTAC
OV CPTAC
E
Higher
Lower
ECPA$_{cell}$
r = −0.56
P = 3.8 × 10^{-4}
CHOL
r = −0.58
P = 7.0 × 10^{-35}
LIHC
r = −0.84
P = 2.9 × 10^{-144}
KIRC
r = −0.03
P = 0.55
LUSC
r = −0.37
P = 2.0 × 10^{-18}
LUAD
r = 0.04
P = 0.23
BRCA
r = −0.90
P = 2.4 × 10^{-103}
KIRP
r = −0.24
P = 1.8 × 10^{-6}
COADREAD
r = −0.74
P = 1.3 × 10^{-89}
THCA
Lower
AA usage convergence
Higher

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: 1) key turning points of $ECPA_{cell}$ in different species tend to happen at corresponding developmental stages; and 2) 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.

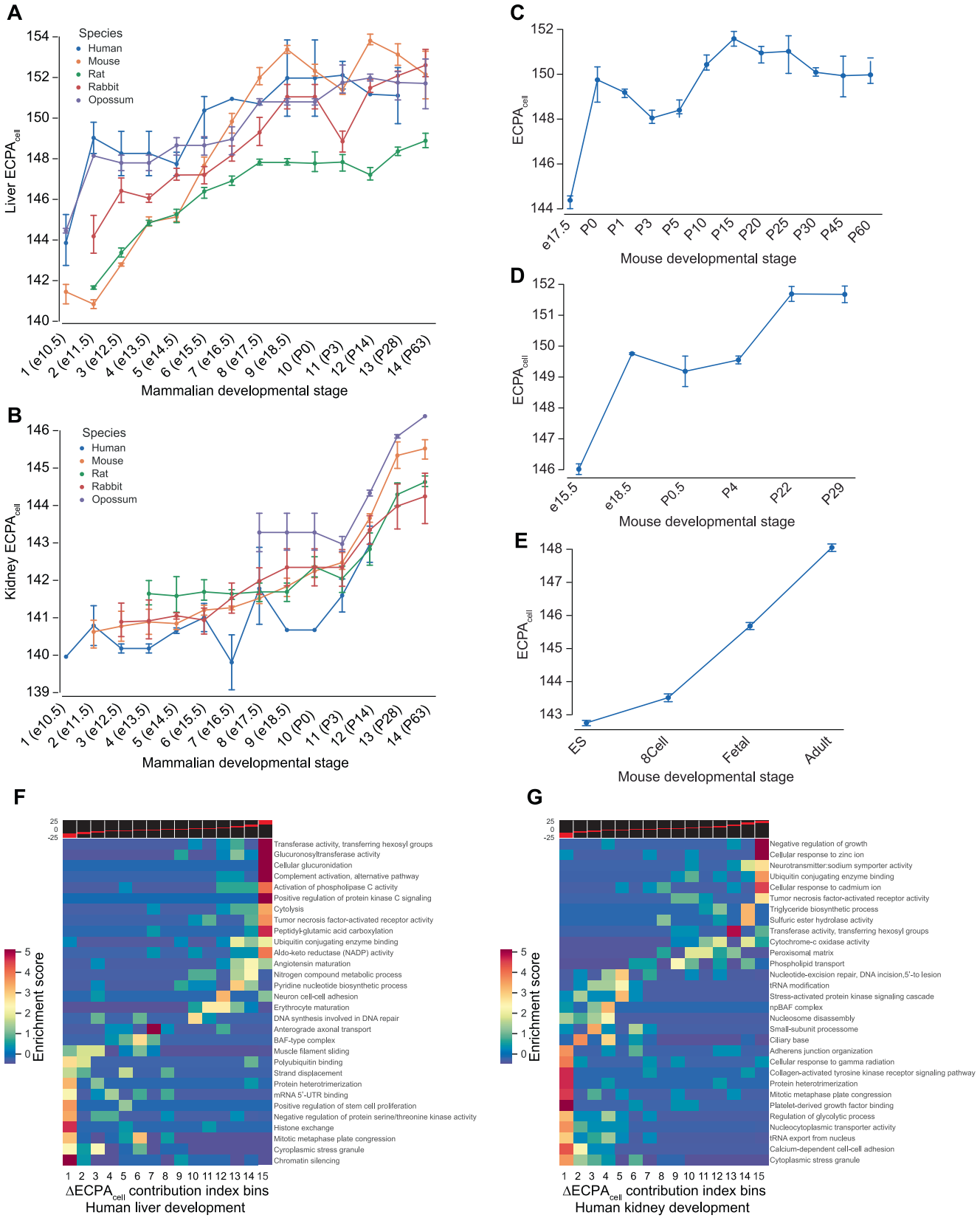
Amino acid usage convergence of tumor follows a reversed path of tissue development

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, $P = 1.6 \times 10^{-156}$; KIRC, $P = 1.9 \times 10^{-39}$; KIRP, $P = 8.9 \times 10^{-30}$). Furthermore, for the genes reducing $ECPA_{cell}$ in tumorigenesis and increasing $ECPA_{cell}$ in development, their $\Delta ECPA_{cell}$ contribution indexes 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.

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 defined as the number of cancer types with significantly increased usage of the amino acid minus the number with significantly decreased usage. The colored region around the regression lines indicates a 95% confidence interval. **B.** Schematic diagram showing the computation of $ECPA_{gene}$ and $ECPA_{cell}$ based on the gene expression profile derived from RNA-seq data. **C.** $ECPA_{cell}$ of tumor samples and matched normal tissue samples across TCGA cancer types. A paired two-sided Wilcoxon signed-rank test was used to calculate the *P* values. **D.** Bar plots showing $ECPA_{gene}$ values of significantly down- and up-regulated proteins in several cancer proteomics datasets. Error bars denote 95% confidence intervals. A two-sided Mann-Whitney U-test was used to calculate the *P* values. **E.** Correlation between $ECPA_{cell}$ and amino acid usage convergence index across samples in nine cancer types. The colored regions around the regression lines indicate 95% confidence intervals. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ns, not significant. $ECPA_{gene}$, energy cost per amino acid per gene; $ECPA_{cell}$, energy cost per amino acid per cell.


A
Species
Human
Mouse
Rat
Rabbit
Opossum
Liver ECPA$_{cell}$
Mammalian developmental stage
B
Kidney ECPA$_{cell}$
Mammalian developmental stage
C
ECPA$_{cell}$
Mouse developmental stage
D
ECPA$_{cell}$
Mouse developmental stage
E
ECPA$_{cell}$
Mouse developmental stage
F
Enrichment score
Transferase activity, transferring hexosyl groups
Glucuronosyltransferase activity
Cellular glucuronidation
Complement activation, alternative pathway
Activation of phospholipase C activity
Positive regulation of protein kinase C signaling
Cytolysis
Tumor necrosis factor-activated receptor activity
Peptidyl-glutamic acid carboxylation
Ubiquitin conjugating enzyme binding
Aldo-keto reductase (NADP) activity
Angiotensin maturation
Nitrogen compound metabolic process
Pyridine nucleotide biosynthetic process
Neuron cell-cell adhesion
Erythrocyte maturation
DNA synthesis involved in DNA repair
Anterograde axonal transport
BAF-type complex
Muscle filament sliding
Polyubiquitin binding
Strand displacement
Protein heterotrimerization
mRNA 5′-UTR binding
Positive regulation of stem cell proliferation
Negative regulation of protein serine/threonine kinase activity
Histone exchange
Mitotic metaphase plate congression
Cyroplasmic stress granule
Chromatin silencing
ΔECPA$_{cell}$ contribution index bins
Human liver development
G
Enrichment score
Negative regulation of growth
Cellular response to zinc ion
Neurotransmitter:sodium symporter activity
Ubiquitin conjugating enzyme binding
Cellular response to cadmium ion
Tumor necrosis factor-activated receptor activity
Triglyceride biosynthetic process
Sulfuric ester hydrolase activity
Transferase activity, transferring hexosyl groups
Cytochrome-c oxidase activity
Peroxisomal matrix
Phospholipid transport
Nucleotide-excision repair, DNA incision,5′-to lesion
tRNA modification
Stress-activated protein kinase signaling cascade
npBAF complex
Nucleosome disassembly
Small-subunit processome
Ciliary base
Adherens junction organization
Cellular response to gamma radiation
Collagen-activated tyrosine kinase receptor signaling pathway
Protein heterotrimerization
Mitotic metaphase plate congression
Platelet-derived growth factor binding
Regulation of glycolytic process
Nucleocytoplasmic transporter activity
tRNA export from nucleus
Calcium-dependent cell-cell adhesion
Cytoplasmic stress granule
ΔECPA$_{cell}$ contribution index bins
Human kidney 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

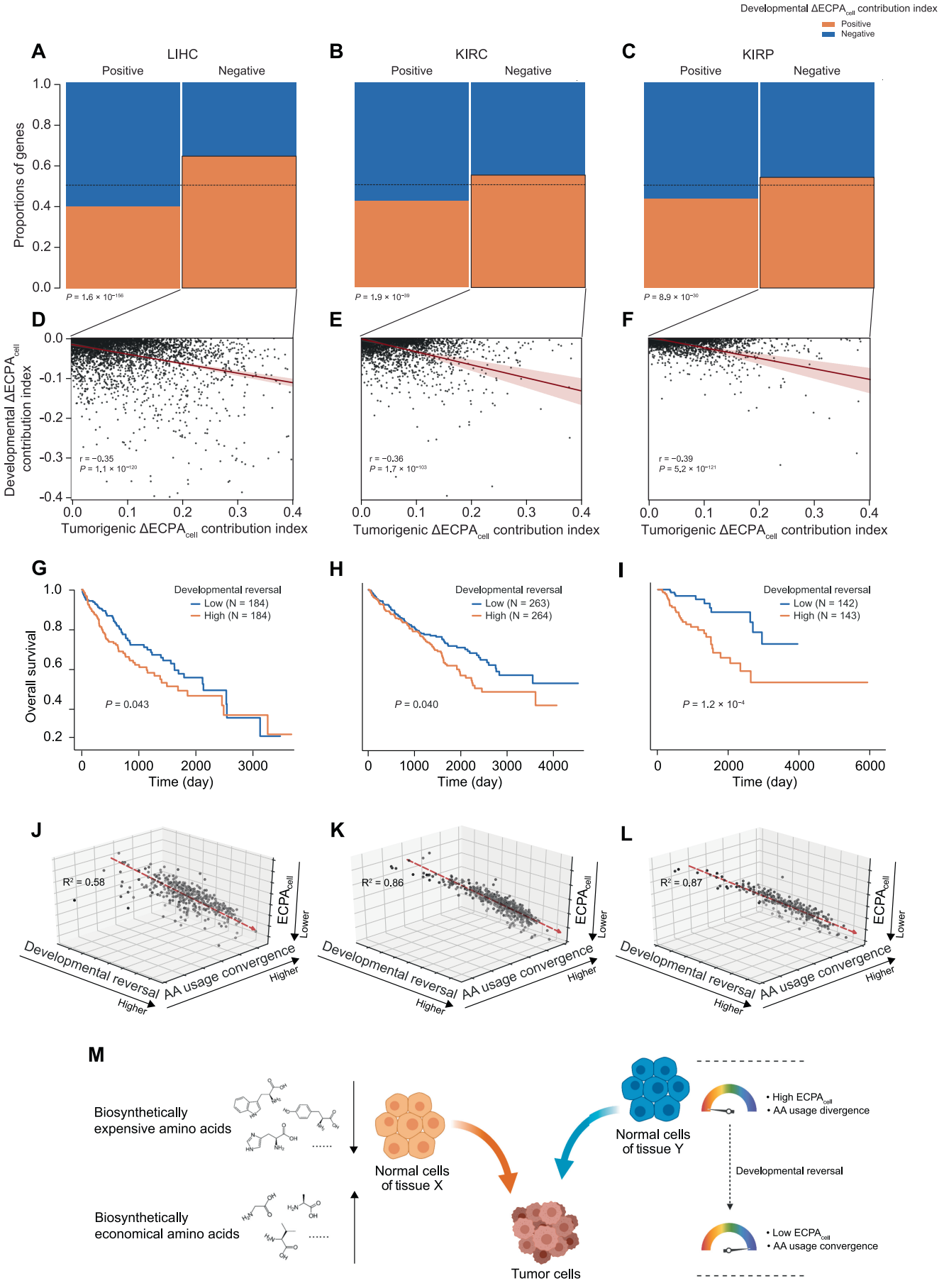
Among different cancer types in our $ECPA_{cell}$ analysis, the difference between liver normal and liver tumor samples was striking, making this tissue stand out from others (Figure 2C). Indeed, by quantifying the downward shift of $ECPA_{cell}$ ($\Delta ECPA_{cell}$) between tumor and the matched NAT 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 1) the extremely high $ECPA_{cell}$ level of liver tissue and 2) 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 a sample size great than or equal to 12. The $ECPA_{cell}$ index was able to separate tumor *vs.* normal samples with very high AUROC 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 9559 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 7.9×10^{-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,

Figure 3 The increasing trend of $ECPA_{cell}$ throughout mammalian organogenesis

A. and **B.** Trend lines of $ECPA_{cell}$ during the development of the liver (A) and the kidney (B) across five mammalian species. Developmental stages of non-mouse species correspond to the mouse stages shown in brackets. Error bars denote 95% confidence intervals. **C.–E.** Trend lines of $ECPA_{cell}$ along the developmental trajectory of the mouse liver across three independent datasets. Error bars denote 95% confidence intervals. **F.** and **G.** Heatmaps showing enrichment patterns of gene modules that contribute to $\Delta ECPA_{cell}$ during the development of the human liver (F) and kidney (G). The red stripes embedded in the black background on top of each heatmap designate the range of $\Delta ECPA_{cell}$ contribution index within every bin.


Developmental ΔECPA$_{cell}$ contribution index
Positive
Negative
A
LIHC
Positive
Negative
Proportions of genes
1.0
0.8
0.6
0.4
0.2
0.0
P = 1.6 × 10^{-156}
B
KIRC
Positive
Negative
P = 1.9 × 10^{-39}
C
KIRP
Positive
Negative
P = 8.9 × 10^{-30}
D
Developmental ΔECPA$_{cell}$ contribution index
0.0
-0.1
-0.2
-0.3
-0.4
r = −0.35
P = 1.1 × 10^{-120}
0.0 0.1 0.2 0.3 0.4
Tumorigenic ΔECPA$_{cell}$ contribution index
E
r = −0.36
P = 1.7 × 10^{-103}
0.0 0.1 0.2 0.3 0.4
Tumorigenic ΔECPA$_{cell}$ contribution index
F
r = −0.39
P = 5.2 × 10^{-121}
0.0 0.1 0.2 0.3 0.4
Tumorigenic ΔECPA$_{cell}$ contribution index
G
Overall survival
1.0
0.8
0.6
0.4
0.2
Developmental reversal
Low (N = 184)
High (N = 184)
P = 0.043
0 1000 2000 3000
Time (day)
H
Developmental reversal
Low (N = 263)
High (N = 264)
P = 0.040
0 1000 2000 3000 4000
Time (day)
I
Developmental reversal
Low (N = 142)
High (N = 143)
P = 1.2 × 10^{-4}
0 2000 4000 6000
Time (day)
J
R^2 = 0.58
ECPA$_{cell}$
Lower
Developmental reversal
Higher
AA usage convergence
Higher
K
R^2 = 0.86
ECPA$_{cell}$
Lower
Developmental reversal
Higher
AA usage convergence
Higher
L
R^2 = 0.87
ECPA$_{cell}$
Lower
Developmental reversal
Higher
AA usage convergence
Higher
M
Biosynthetically expensive amino acids
Biosynthetically economical amino acids
Normal cells of tissue X
Normal cells of tissue Y
Tumor cells
• High ECPA$_{cell}$
• AA usage divergence
Developmental reversal
• Low ECPA$_{cell}$
• AA usage convergence

$ECPA_{cell}$ 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 despite 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 the 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 a highly 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 TPM] of 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 (https://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/), and two independent RNA-seq datasets of mouse liver development from the Gene Expression Omnibus (GEO), as well as from ArrayExpress. Finally, we obtained RNA-seq datasets of aging mouse liver and kidney from GEO.

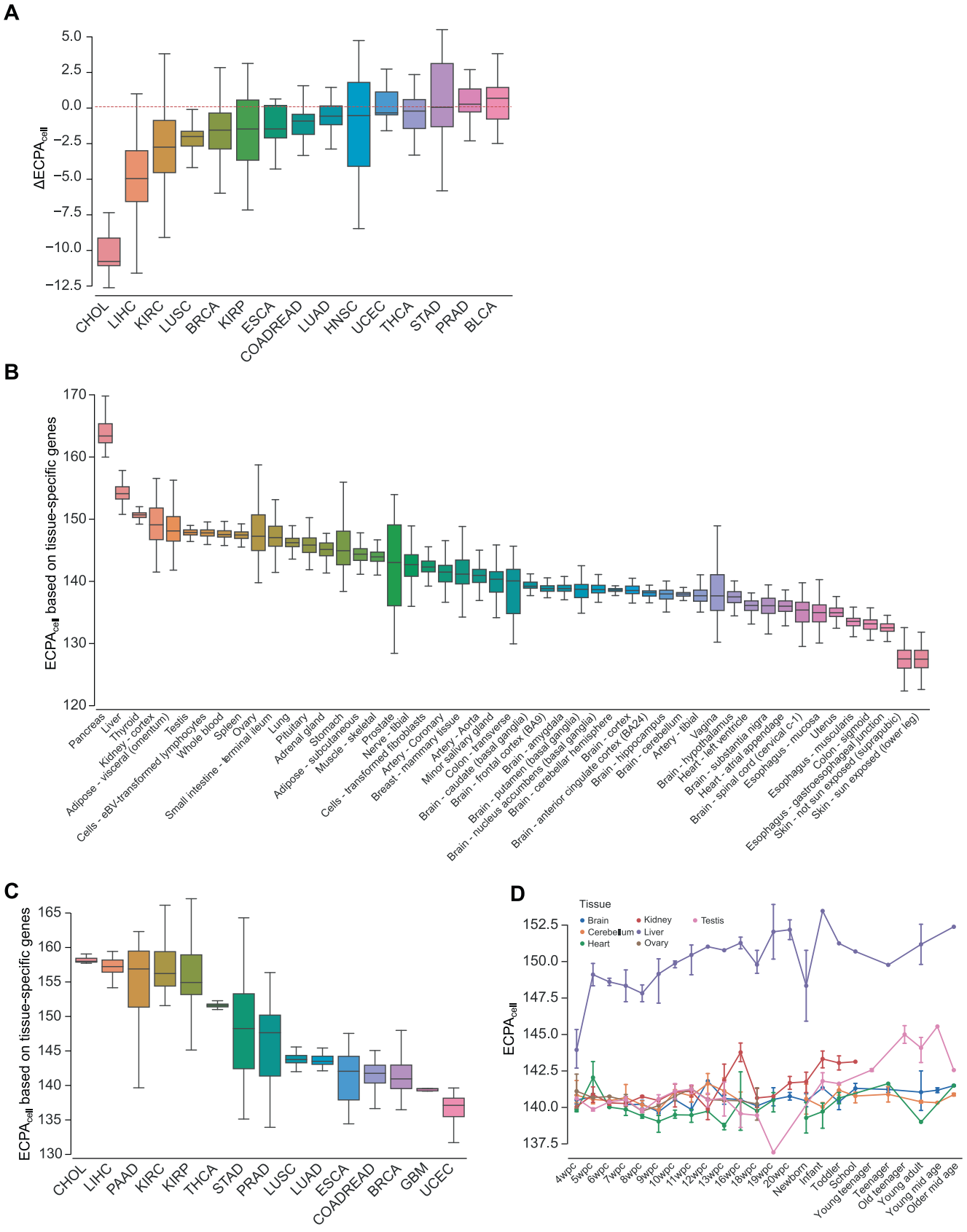
To convert gene-level FPKM values to TPM [44] values for a gene g_i in a sample s_k, we used the equation:

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

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, files of aging rat liver from the Sequence Read Archive (SRA), 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.

Figure 4 A proposed model unifying developmental reversal, amino acid usage convergence, and $ECPA_{cell}$ decline of cancer samples
A.–C. Stacked bar plots showing the proportion of genes that positively or negatively contribute to $\Delta ECPA_{cell}$ in either tumorigenesis or development for LIHC-liver (A), KIRC-kidney (B), and KIRP-kidney (C). **D.–F.** Scatter plots showing, for genes with negative $\Delta ECPA_{cell}$ contribution index in tumorigenesis and positive $\Delta ECPA_{cell}$ contribution index in tissue development, scaled $\Delta ECPA_{cell}$ contribution index in tumorigenesis *vs.* scaled $\Delta ECPA_{cell}$ contribution index in tissue development for LIHC-liver (D), KIRC-kidney (E), and KIRP-kidney (F). Colored regions around the regression lines indicate 95% confidence intervals. **G.–I.** Kaplan-Meier plots showing the overall survival for patients with LIHC (G), KIRC (H), and KIRP (I) stratified by developmental reversal index into two equal groups, respectively. The *P* values were calculated from two-sided log-rank tests. **J.–L.** 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. LIHC, liver hepatocellular carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma.


A
ΔECPA$_{cell}$
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
CHOL
LIHC
KIRC
LUSC
BRCA
KIRP
ESCA
COADREAD
LUAD
HNSC
UCEC
THCA
STAD
PRAD
BLCA
B
ECPA$_{cell}$ based on tissue-specific genes
170
160
150
140
130
120
Pancreas
Liver
Thyroid
Kidney - cortex
Adipose - visceral (omentum)
Testis
Cells - eBV-transformed lymphocytes
Whole blood
Spleen
Ovary
Small intestine - terminal ileum
Lung
Pituitary
Adrenal gland
Stomach
Adipose - subcutaneous
Muscle - skeletal
Prostate
Nerve - tibial
Cells - transformed fibroblasts
Artery - Coronary
Breast - mammary tissue
Artery - Aorta
Minor salivary gland
Colon - transverse
Brain - caudate (basal ganglia)
Brain - frontal cortex (BA9)
Brain - amygdala
Brain - putamen (basal ganglia)
Brain - nucleus accumbens (basal ganglia)
Brain - cerebellar hemisphere
Brain - anterior cingulate cortex (BA24)
Brain - cortex
Brain - hippocampus
Brain - cerebellum
Artery - tibial
Vagina
Brain - hypothalamus
Heart - left ventricle
Brain - substantia nigra
Heart - atrial appendage
Brain - spinal cord (cervical c-1)
Esophagus - mucosa
Uterus
Esophagus - muscularis
Colon - sigmoid
Esophagus - gastroesophageal junction
Skin - not sun exposed (suprapubic)
Skin - sun exposed (lower leg)
C
ECPA$_{cell}$ based on tissue-specific genes
165
160
155
150
145
140
135
130
CHOL
LIHC
PAAD
KIRC
KIRP
THCA
STAD
PRAD
LUSC
LUAD
ESCA
COADREAD
BRCA
GBM
UCEC
D
Tissue
Brain
Cerebellum
Heart
Kidney
Liver
Ovary
Testis
ECPA$_{cell}$
152.5
150.0
147.5
145.0
142.5
140.0
137.5
4wpc
5wpc
6wpc
7wpc
8wpc
9wpc
10wpc
11wpc
12wpc
13wpc
16wpc
18wpc
19wpc
20wpc
Newborn
Infant
Toddler
School
Young teenager
Teenager
Old teenager
Young adult
Young mid age
Older mid age

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.

All data used in this study are publicly available through consortia websites and are listed in Table S1.

Calculation of transcriptome-based amino acid usage

We used the following equation to compute the amino acid frequency matrix given an RNA-seq dataset (see also Figure 1A):

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

where E is a matrix of genes $g_1, g_2, \cdots, g_n$ by samples $s_1, s_2, \cdots, s_m$ with entries as TPM values, and A is a matrix of genes $g_1, g_2, \cdots, g_n$ by amino acids $a_1, a_2, \cdots, a_{20}$ 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 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

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

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

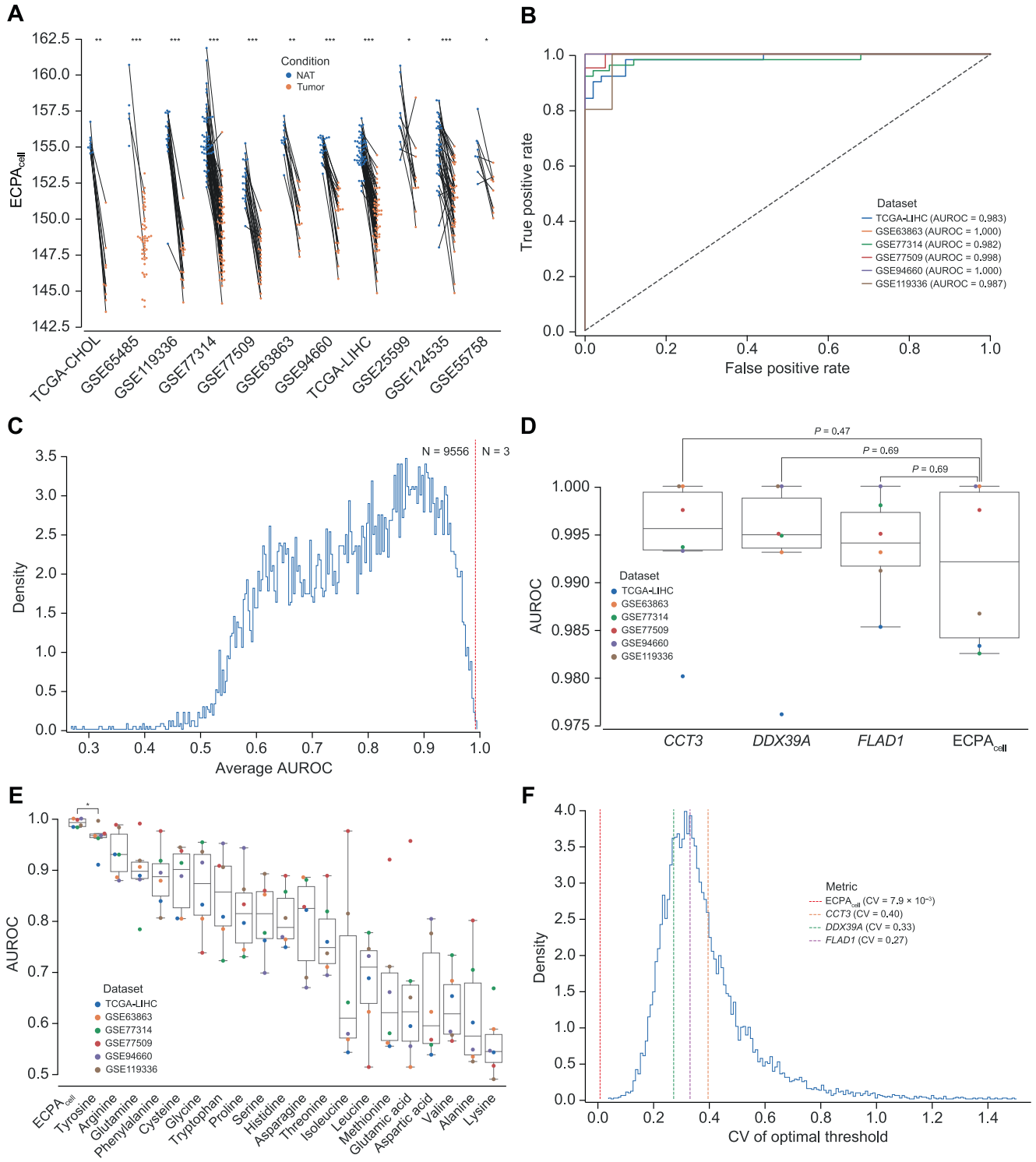


Figure 6 ECPA$_{cell}$ is a robust diagnostic biomarker for liver cancer

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. Colorful lines indicate the lines of identity. **C.** Density plot showing the distribution of the average AUROC across the six cohorts for tumor-normal segregation using the mRNA expression level of each of the 9559 detectable genes. The vertical dashed line corresponds to the average AUROC of ECPA$_{cell}$. **D.** Box plots showing the AUROC 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 AUROC 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 $\pm$ 1.5$\times$ interquartile range. A paired two-sided Wilcoxon signed-rank test was used to calculate the *P* values. **F.** Density plot showing the distribution of CV of the optimal thresholds in using individual genes for tumor-normal segregation. The vertical red dashed line indicates the CV of ECPA$_{cell}$. Vertical lines in three other colors indicate the CV of three genes whose average AUROCs are higher than ECPA$_{cell}$. ROC, receiver operating characteristic; AUROC, area under the ROC curve; CV, coefficient of variation. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

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 the 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 expression 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:

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

where $\oslash$ is element-wise division, ρ 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 $g_1, g_2, \cdots, g_n$ by NAT samples $s_1, s_2, \cdots, s_m$ of a certain cancer type with entries as expression level, $\overrightarrow{m^{-1}}$ is a normalization vector of constant m^{-1}, and $\overrightarrow{r}$ is a vector of signed goodness-of-fit values of genes $g_1, g_2, \cdots, g_n$ 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 $\geq$ 1 in $\geq$ 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 (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*.

Code availability

The source code and documentation for all analyses conducted in this study have been deposited in Zenodo (https://doi.org/10.5281/zenodo.5055829) in the format of Jupyter Notebooks.

CRediT author statement

Yikai Luo: Conceptualization, Formal analysis, Visualization, Writing - original draft. **Han Liang:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition. Both authors have read and approved the final manuscript.

Competing interests

Han Liang is a shareholder and scientific advisor to Precision Scientific Ltd.

Acknowledgments

We thank Hong Zhang, Han Chen, and other members of the Liang lab for helpful discussions. We thank Hani Goodarzi for his critical review of the manuscript. We also thank Kamalika Mojumdar for editorial assistance. This work was supported by the US National Institutes of Health (Grant No. U24CA209851 to HL), the Cancer Center Support Grant (Grant No. P30CA016672 to HL), an MD Anderson Faculty Scholar Award (to HL), and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (to HL).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.08.004.

ORCID

ORCID 0000-0001-7589-7981 (Yikai Luo)
ORCID 0000-0001-7633-286X (Han Liang)

References

[1] Seligmann H. Cost-minimization of amino acid usage. J Mol Evol 2003;56:151–61.

[2] Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller RV, Krane DE. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol 2006;23:1670–80.

[3] Raiford DW, Heizer EM, Miller RV, Akashi H, Raymer ML, Krane DE. Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*? J Mol Evol 2008;67:621–30.

[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.

[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;31:2905–12.

[6] Wu CI, Wang HY, Ling S, Lu X. The ecology and evolution of cancer: the ultra-microevolutionary process. Annu Rev Genet 2016;50:347–69.

[7] Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.

[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.

[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.

[10] Milanovic M, Fan DNY, Belenki D, Däbritz JHM, Zhao Z, Yu Y, et al. Senescence-associated reprogramming promotes cancer stemness. Nature 2018;553:96–100.

[11] Peiris-Pagès M, Martinez-Outschoorn UE, Pestell RG, Sotgia F, Lisanti MP. Cancer stem cell metabolism. Breast Cancer Res 2016;18:55.

[12] Bellacosa A. Developmental disease and cancer: biological and clinical overlaps. Am J Med Genet A 2013;161:2788–96.

[13] Aiello NM, Stanger BZ. Echoes of the embryo: using the developmental biology toolkit to study cancer. Dis Model Mech 2016;9:105–14.

[14] Kamel HFM, Al-Amodi HSAB. Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. Genomics Proteomics Bioinformatics 2017;15:220–35.

[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: 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.

[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.

[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.

[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.

[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.

[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.

[23] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science 2017;357:eaan2507.

[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.

[25] Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971.

[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;8:1077.

[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;9:4124.

[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.

[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.

[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;179:561–77.

[31] Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell 2009;36: 900–11.

[32] López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell 2013;153:1194–217.

[33] Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of connectivity in cancer co-expression networks. PLoS One 2014;9:e87075.

[34] Han R, Huang G, Wang Y, Xu Y, Hu Y, Jiang W, et al. Increased gene expression noise in human cancers is correlated with low p53 and immune activities as well as late stage cancer. Oncotarget 2016;7:72011–20.

[35] Yang RY, Quan J, Sodaei R, Aguet F, Segrè AV, Allen JA, et al. A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. bioRxiv 2018;311563.

[36] Gould SJ. Ontogeny and phylogeny–revisited and reunited. Bioessays 1992;14:275–9.

[37] Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, et al. Gene expression divergence recapitulates the developmental hourglass model. Nature 2010;468:811–4.

[38] Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 2010;468:815–8.

[39] Davies PCW, Lineweaver CH. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. Phys Biol 2011;8:015001.

[40] Lineweaver CH, Davies PCW, Vincent MD. Targeting cancer's weaknesses (not its strengths): therapeutic strategies suggested by the atavistic model. Bioessays 2014;36:827–35.

[41] Chen H, Lin F, Xing K, He X. The reverse evolution from multicellularity to unicellularity during carcinogenesis. Nat Commun 2015;6:6367.

[42] Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. Proc Natl Acad Sci U S A 2017;114:6406–11.

[43] Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer. Elife 2019;8:e40947.
[44] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 2010;26:493–500.
[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.
[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.
[47] Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 2015;4:1521.
[48] Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics 2014;30:2598–602.