Computational and Structural Biotechnology Journal 23 (2024) 2534–2547

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Gene choice in cancer cells is exclusive in ion transport but concurrent in DNA replication

Samuel Mondal, Attila Becskei *

Biozentrum, University of Basel, Spitalstrasse 41, Basel 4056, Switzerland

ARTICLE INFO

Keywords:
Co-occurrence
Stochastic gene choice
Carbonic anhydrase
Protocadherin
Odorant receptor
Tumorigenesis

ABSTRACT

Cancers share common cellular and physiological features. Little is known about whether distinctive gene expression patterns can be displayed at the single-cell level by gene families in cancer cells. The expression of gene homologs within a family can exhibit concurrence and exclusivity. Concurrence can promote all-or-none expression patterns of related genes and underlie alternative physiological states. Conversely, exclusive gene families express the same or similar number of homologs in each cell, allowing a broad repertoire of cell identities to be generated. We show that gene families involved in the cell-cycle and antigen presentation are expressed concurrently. Concurrence in the DNA replication complex MCM reflects the replicative status of cells, including cell lines and cancer-derived organoids. Exclusive expression requires precise regulatory mechanism, but cancer cells retain this form of control for ion homeostasis and extend it to gene families involved in cell migration. Thus, the cell adhesion-based identity of healthy cells is transformed to an identity based on migration in the population of cancer cells, reminiscent of epithelial-mesenchymal transition.

1. Introduction

Despite the diversity of cancers, they have common metabolic, histopathologic and genetic properties [1]. For example, the metabolism of cancer cells is characterized by a hyperactive glucose uptake, followed by a preferential fermentation into lactate, a phenomenon known as the Warburg effect [2]. In terms of cell behavior, six hallmarks distinguish cancer cells from healthy cells: they grow independently of growth signals (i.e. cells can grow without growth factors), resist anti-growth signals from contact inhibition and cell adhesion, avoid apoptosis, possess limitless replicative potential, induce angiogenesis, and invade tissues [3].

Genetic alterations, such as point mutations and copy number alterations, affect only a few common oncogenes and proto-oncogenes, which act as early clonal drivers [4]. Half of early clonal driver mutations are located in less than ten driver genes, whereas subclonal mutations affect four times more genes, indicating a progressive increase in genetic heterogeneity during tumorigenesis [4]. Many mutations are mutually exclusive or co-occurrent (concurrent) in pathways or gene families of individual patients [5,6], indicating the importance of heterogeneity in the genetic signatures of cancer.

Non-genetic processes also contribute to heterogeneity. For example, tumor expansion leads to cell crowding, which in turn alters gene expression by increasing heterogeneity (noise) in the expression of genes involved in the epithelial-mesenchymal transition [7]. Phenotypic heterogeneity can have multiple sources [8,9], making it challenging to identify differentially expressed genes in tumors in bulk assays [10,11]. Therefore, spatial or single-cell analysis is likely to help identify relevant alterations. Gene expression heterogeneity in cancer cells can affect their response to therapy [12], underscoring the importance of analyzing heterogeneity in single-cell expression, which we explored in this study.

To identify gene expression patterns in populations of single cancer cells, we analyzed stochastic gene choice in gene families, focusing on exclusivity and concurrence [13]. Gene families are groups of homologous genes within a species, originating from a common gene ancestor, typically through gene duplications. This process can occur surprisingly rapidly during evolution [14,15]. Homologous genes may have distinct tissue-specific expression, or encode enzymes with different catalytic activities and substrate specificities. For instance, the largest gene family in mammals encodes a thousand or more odorant receptor homologs [16], each recognizing a specific odor. Each olfactory neuron expresses a single odorant receptor homolog, exemplifying exclusive stochastic gene choice. Similarly, each T-cell chooses one gene homolog from the

* Corresponding author.
E-mail address: attila.becskei@unibas.ch (A. Becskei).

https://doi.org/10.1016/j.csbj.2024.06.004
Received 29 February 2024; Received in revised form 4 June 2024; Accepted 4 June 2024
Available online 10 June 2024

repertoire of T-cell receptor genes, determining the type of antigen the cell will react to [17–19].

Exclusive gene choice from a large repertoire requires precision and complex control mechanisms that enable the simultaneous activation of a single gene, its long-term stability of expression, and the suppression of all other related genes [20]. Other gene (sub)families, like the clustered protocadherin-alpha, also achieve exclusivity but typically two gene homologs are chosen for expression, which shows that exclusivity is not necessarily synonymous with the expression of a single gene homolog [13]. Furthermore, not all cells express the same number of homologs. In a broader probabilistic (stochastic) sense, exclusivity implies that most cells express a specific number of gene homologs, with some cells deviating from this number. The result is a small variance in the number of expressed homologs (Fig. 1a, left panel).

In contrast, a large variance in the number of expressed gene homologs in the cell population indicates concurrence (co-occurrence) (Fig. 1a, right panel). Few families display exclusive gene choice, as most homologous genes retain co-regulation [17]. If the genes are co-regulated, they will tend to display concurrent, and in extreme cases, all-or-none expression patterns, which can create alternating physiological states. On the other hand, exclusive expression serves as a source of phenotypical diversity in the context of a single physiological state. For example, all clustered protocadherins share the same physiological role of mediating cell-cell adhesion, but the expression of different homologs in each cell and their varying binding with neighboring cells confer distinct cell identities, which underlies the formation of neuronal networks.

In adult mice, gene families other than the above prototypical gene families can exhibit exclusivity, and they typically function in cell adhesion and ion homeostasis. Little is known about how stochastic gene choice is controlled in cancer cells.

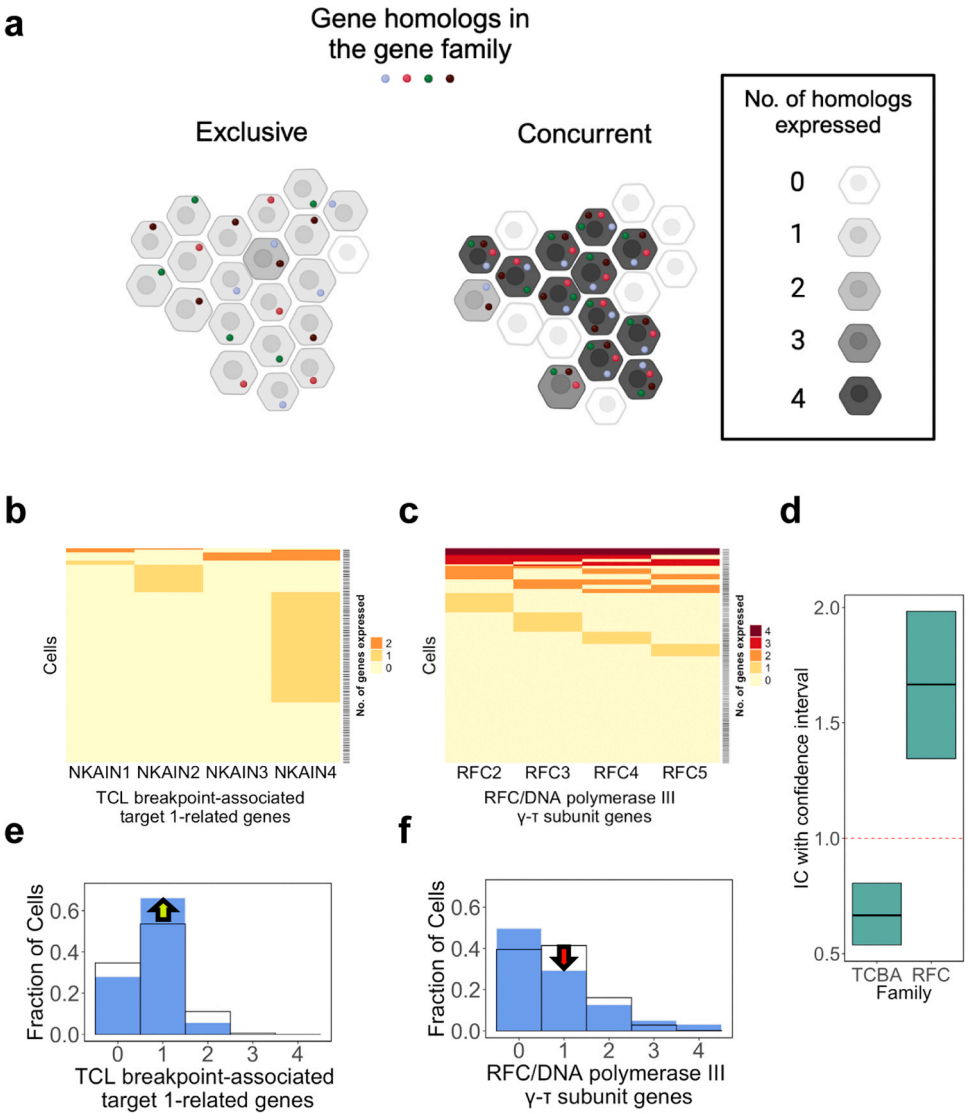

Fig. 1. Patterns of stochastic gene choice in cancers. a, Gene choice in a gene family with four homologs. Each homolog's expression is indicated by a circle in one of four colors. The cell's background color (gray scale) shows the number of expressed homologs per cell. The exclusive pattern mainly features cells expressing a single gene. The concurrent pattern shows most cells expressing either no genes or all four, with a few cells expressing three genes (created with biorender.com). **b,** Each black line corresponds to a single cell, with cells expressing the same number of homologs sharing the same field color in the heatmap for an exclusive gene family in the Glioblastoma MGH151 dataset. Less than 5 % of the cells express more than two genes (typically the fixed pair of NKAIN3 and 4). **c,** Heatmap of a concurrent family. **d,** The 95 %CI of the IC of the TCL breakpoint-associated target 1-related (TCBA) and the RFC/DNA polymerase III γ-τ subunit (RFC) families. The horizontal dashed line denotes IC = 1. **(e-f)**, The number of expressed homologs per cell is calculated from (b, c) (blue histogram). The histogram with the black edges denotes the Poisson-binomial distribution calculated from the gene expression frequency. The arrows show the difference between the modes of the two distributions.

2. Materials and methods

2.1. Datasets and cell classification

Cancer datasets, along with the tissue of origin and the condition (cancer/normal), are described in Table S1 [21–32]. It contains also datasets of two melanoma cell lines [33,34]. Cells were classified with MarkerCount [35], which requires a marker gene set. For this purpose, CellMarker 2.0 was used [36]. Cells not endogenous to the cancer and unidentified cell were filtered out (Table S2).

Each dataset must contain a total number of $N_t \geq 100$ cells after the eliminations. To ensure robust inference, we excluded datasets from further data analysis if the 10th percentile of the number of the detected genes per cell (dgpc) did not exceed 2000 genes. If dgpc <2000, we eliminated cells with the lowest coverage from the dataset to reach the 2000 dgpc threshold.

2.2. Dichotomization of gene expression

After the cell classification and filtering, a specific threshold was calculated for each gene family, which was used to dichotomize the RNA count into binary on/off expression states [17]. The gene family list was obtained from PANTHER 15.0 [37].

First, we assembled an RNA count distribution that is composed of all genes in the specific gene family. The geometric trimmed mid-extreme (GTME) threshold was applied for datasets with TPM units:

$$GTME = \begin{cases} \sqrt{x_{0.025} \cdot x_{0.975}}, & N \geq 120 \\ \sqrt{\sum_{i=1}^{3} x_i \cdot \sum_{i=N-2}^{N} x_i}, & N < 120 \end{cases}$$

if g is the set of all genes in a gene family (GF), then, $x_{0.975} = \max(x_{g.0.975} : g \epsilon GF)$, where $x_{g,\ 0.975}$ is the 97.5 percentile expression value of a gene in a cell population. Correspondingly, $x_{0.025} = \min(x_{g.0.025} : g \epsilon GF)$. N is the number of cells with TPM >0.5.

The fraction of the maximum (FM) threshold was applied for datasets with Unique Molecular Identifier (UMI) counts since it approximates better the fitted probability mass functions [17].

$$FM = \begin{cases} \frac{x_{0.975}}{10}, & N \geq 120 \\ \frac{\sum_{i=N-2}^{N} x_i}{10}, & N < 120 \end{cases}$$

N is the number of cells with count >0.

When $N < 120$, x_i is the expression value of the cells with ordered TPM / UMI values.

2.3. Interdependence coefficient (IC) calculation and classification of gene families

IC is the ratio of the observed variance of the number of gene homologs chosen to be expressed in a family to the variance of the Poisson-binomial distribution expected from the on-state frequencies [13,38]:

$$IC = \frac{\sigma^2_{OBS}}{\sigma^2_{PB}}, where\ \sigma^2_{PB} = \sum_{i=1}^{N_\alpha} (1 - p_i)p_i$$

The Poisson-binomial distribution is also known as the generalized binomial distribution. p_i is approximated by the on-state frequency for $N_\alpha > 100$, where $N_\alpha$is the number of cells in the sample.

Bootstrapping was performed to calculate the 95 % confidence interval (CI) of IC by resampling of cells with replacement. Resampling was performed 10,000 times for each family and IC was calculated for each resampling.

Gene families that fulfilled two criteria, non-zero genes >3 and mean gene per cell (mgpc) > 0.1, were selected for further analysis. The first criterion ensures sufficient diversity in the gene repertoire, and it is motivated by the four-color theorem according to which four genes are sufficient to impart unique identities among neighboring cells in a plane [17]. The second criterion excludes gene families with scarce expression.

Gene families with a 95 %CI of the IC less than 1 were classified as exclusive. Gene families with a 95 %CI of the IC greater than one were classified as base concurrent.

The gene families were ranked according to their mgpc values to select families with excess concurrence. The ten nearest neighbors of every candidate family, 5 on each side, were analyzed. If at least nine of the ten nearest neighbor families have non-overlapping 95 %CIs, the candidate family is considered concurrent in excess. For the five families with the lowest mgpc levels, the candidate families had more neighbors with higher ranks than with lower ranks. The five families with the highest ranks were not included in the analysis due to the concomitant surge in both mgpc and IC values. Further subclasses of concurrent families are described in Table S3.

2.4. Statistical analysis of contingency tables

The odds ratio and the P-value for the Fisher's exact test were calculated with the R function *fisher.test*.

In the first type of contingency table, two different properties (categories) were compared to assess, for example, whether exclusivity (exclusive vs non-exclusive families) is associated with pathology (families in healthy versus tumor cells). Thus, the contingency table is defined by two categories.

In the second type of contingency table, the same property (category) was compared in two different samples or conditions. For instance, it was used to assess whether exclusive families in two different tissues co-occur beyond random coincidence.

The two-tailed Fisher test was used and the significance level was $\alpha = 0.05$ for both cases.

2.5. Differential gene expression analysis

In differential gene expression analysis, the expression of genes was compared between tumor and healthy samples after cell type classification and selection. Wilcoxon test (Mann-Whitney U) was used for asserting significant change, following the method described in [39]. For this purpose, TPM values were normalized resulting in TMM normalized TPM (Supplementary Methods), before applying the Wilcoxon test.

The log2 fold change is equal to:

$$\log_2\left(\frac{\text{mean}(tumor\ TMM\ normalized\ TPM) + 1}{\text{mean}(periphery\ TMM\ normalized\ TPM) + 1}\right)$$

2.6. Multiple testing correction for differential gene expression analysis

For multiple testing correction, we utilized a step-down multiple testing procedure, termed minP [40], which yields adjusted P-values.

After obtaining the original Wilcoxon (Mann-Whitney U) two-tailed P-values, they are arranged in ascending order as $p_1, p_2, \ldots, p_k$, where k stands for the total number of genes. Next, the cells were shuffled randomly between the two conditions and a permuted healthy and tumor expression data was obtained. After calculating Wilcoxon (Mann-Whitney U) two-tailed P-values for the permuted set, they are arranged in order of the original ascending P-values. Since these permuted P-values are out of order, q-values are obtained by defining successive minima as follows

p_r^* are the permuted P − values

$$q_k^* = p_{r_k}^*$$

$$q_{k-1}^* = \min\left(q_k^*, p_{r_{k-1}}^*\right)$$

$$q_{k-2}^* = \min\left(q_{k-1}^*, p_{r_{k-2}}^*\right)$$

$$\vdots$$

$$q_1^* = \min\left(q_2^*, p_{r_1}^*\right)$$

The q-values for all the k genes are obtained for N = 10,000 permutations of the cells and whenever, $q_i^* \leq p_i$, $\mathrm{COUNT_i}$ is increased by 1. Then, another set of P-values (denoted as $\widetilde{p}_{(i)}^{(N)}$ for gene i) is obtained by dividing the final value of $\mathrm{COUNT_i}$ by N (for all the genes). To obtain the final adjusted P-value, successive maximization is used to enforce monotonicity as follows

$$\widetilde{p}_{(1)}^{(N)} = \widetilde{p}_{(1)}^{(N)}$$

$$\widetilde{p}_{(2)}^{(N)} \leftarrow \max\left(\widetilde{p}_{(1)}^{(N)}, \widetilde{p}_{(2)}^{(N)}\right)$$

$$\vdots$$

$$\widetilde{p}_{(k)}^{(N)} \leftarrow \max\left(\widetilde{p}_{(k-1)}^{(N)}, \widetilde{p}_{(k)}^{(N)}\right)$$

2.7. Permutation test for the change in IC between healthy and tumor samples

We carried out a permutation test to assess if the change in IC was purely by chance (null hypothesis). Dichotomized data were shuffled between the two conditions, and IC was recalculated for the permuted tumor and healthy samples.

100,000 permutations were performed. The two-tailed P-values for the IC change were calculated based on the Monte Carlo sampling permutation test method [41]:

$$p = \frac{1 + \sum_{i=1}^{N} [|t_i - \bar{t}| \geq |t^* - \bar{t}|]}{1 + N}$$

The squared brackets denote the index function, where t* is the original IC change, t_i is the IC change of the i^{th} permutation, $\bar{t}$ is the mean of all the shuffled IC changes, and $N = 100{,}000$. A pseudo-count of 1 is used to prevent $p = 0$. Therefore, the smallest P-value is 0.00001.

After obtaining the P-value using the permutation test, multiple testing correction was applied using the Bonferroni-Holm method to control the family-wise error rate.

2.8. Assessment of overrepresentation of concurrent and exclusive gene families

The binomial test for overrepresentation was employed to identify conserved families across all tumor samples or within specific cancer types. The P-value was calculated from the one-tailed binomial test, with the significance level described below:

$$P - value = \sum_{i=k}^{n} \Pr(X = i) =$$

$$= \sum_{i=k}^{n} \binom{n}{i} {\pi_o}^i (1 - \pi_o)^{n-i}$$

n is the total number of cancer samples. π_o is the average frequency of the concurrent / exclusive families across the cancer samples:

$$\pi_o = \frac{d}{f_{tot} n}$$

d is the total number of detected exclusive or concurrent families, whereas f_{tot} is the total number of families with more than 3 genes in the PANTHER dataset [37]; $f_{tot} = 1144$.

For all cancer samples, $n_{\mathrm{tot}} = 76$. When enrichment in specific types of cancer was calculated, then the number of samples in that cancer type was utilized.

Each exclusive/concurrent family has a number of occurrence, k. The maximum possible $k = n$. For all possible values of k, the P-value (p) is calculated. Since this is equivalent to performing the calculation $f_{tot} = 1144$ times, the significance level is adjusted with Bonferroni multiple testing correction: $\alpha = 0.05/1144 = 4.3{\cdot}10^{-5}$. This correction for multiple testing is very stringent [42].

2.9. Gene ontology (GO) enrichment

We obtained GO enrichment with the *enrichGO* command in ClusterProfiler version 4.8.2 in R (Bioconductor version 3.17) [43]. We used Human annotation found in org.Hs.eg.db [Carlson M (2019). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.], along with GO information from GO.db [Carlson M (2019). GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.8.2.] for the *enrichGO* function. False discovery rate (FDR) is calculated using the Benjamini-Hochberg method [44]. After obtaining the results from *enrichGO*, filtering was applied to obtain a subset of interest for this study.

GO enrichment was calculated for the families that were overrepresented in either all cancers or a specific cancer type (Section 2.8). Any gene that is not expressed in any cell is excluded from the GO analysis (e.g. OPN1MW3 in exclusive families). Separate GO enrichment analyses were run for exclusive and concurrent families. After obtaining the GeneRatio and BgRatio from enrichGO, fold enrichment is calculated by dividing the GeneRatio (ratio of the number of genes in the list associated with a GO term to the total number of genes in the list) by the BgRatio (ratio of the number of genes in a GO term to the total number of genes in the genome). In Tables S6 and S7, GO terms with adjusted $p < 0.05$ are listed that meet the criteria regarding the number of families associated with a GO (*fago*).

3. Results

3.1. Detection of exclusive and concurrent gene families in single-cell RNA-seq datasets

Single-cell RNA-seq data were pre-processed to select the appropriate cells for the data analysis. Cells not endogenous to the cancer, such as immune cells, endothelial cells, fibroblasts, etc. were removed. Some of the cells have low gene coverage, and consequently, the number of detected marker genes is insufficient to identify the cell type. The exclusion of unidentifiable cells can improve the gene coverage in the remaining population (Supplementary Fig. 1), and they were also excluded. After these exclusions, 76 higher quality datasets remained that met the conditions for sufficient cell number and gene coverage (Section 2.1); most of these datasets were obtained from brain tumor samples.

After dichotomizing the RNA counts, each gene was assigned to a binary (on/off) expression state from which the interdependence coefficient (IC) was calculated (Sections 2.2, 2.3, Fig. 1b-c, Supplementary Fig. 2) [13,38]. The 95 % confidence interval (95 % CI) of the IC is below one for exclusive gene families and above one for concurrent families (Fig. 1d). For example, gene choice in the T-cell lymphoma breakpoint-associated (TCBA) family is exclusive in a glioblastoma dataset: IC = 0.66 (95 % CI, 0.53–0.80). This family encodes four gene

homologs, NKAIN1–4 [45]. Most cells express one homolog, approximately one-third of the cells express none, and less than 5 % of the cells express more than two homologs (Fig. 1b). The number of homologs each cell expresses is more narrowly distributed than expected from a Poisson-binomial distribution (Fig. 1e). This peaked distribution indicates the precision of regulation in this gene family, ensuring that each cell expresses a similar number of gene homologs with a mean number close to one (mean gene per cell = 0.77).

A similar mean number of homologs is expressed by the RFC/DNA polymerase III family (mean gene per cell = 0.82), but the cells express a broadly varying number of homologs, with many cells expressing either all four RFC genes or none at all (Fig. 1c, f). Consequently, gene choice is

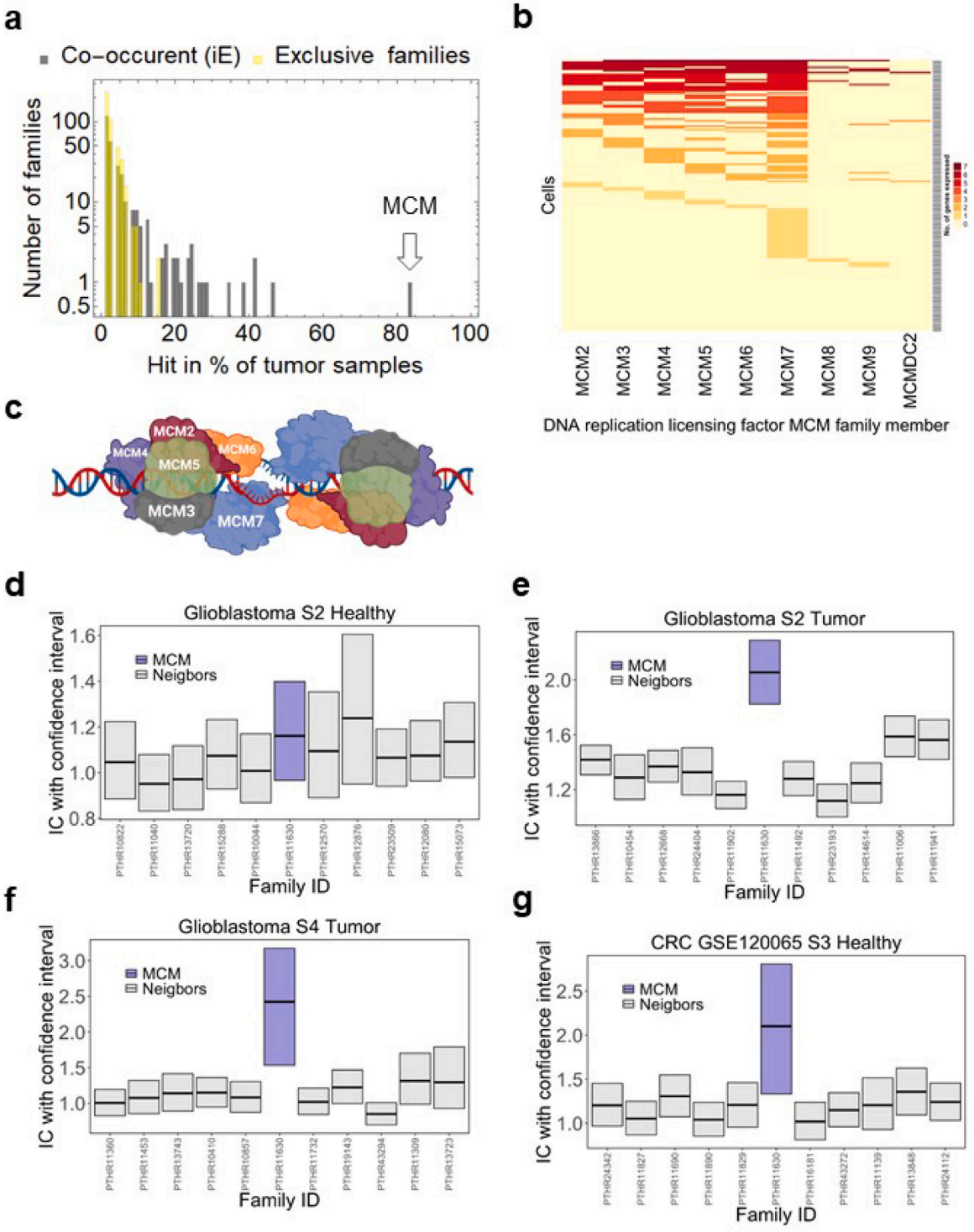

Fig. 2. Concurrence in the DNA replication licensing factor MCM family. a, The MCM family (arrow) displays excess concurrence in 61 out of 76 cancer datasets (frequency = 80 %). The histogram shows the binned frequencies of all concurrent (gray) and exclusive (yellow) families. **b,** Single-cell expression heatmap of the MCM family in the MGH151 glioblastoma dataset **c,** The MCM2–7 genes of the family encode the proteins that form the hexa-dimeric replication complex. (**d-g)** The 95 %CI of the IC of the MCM (blue) and the 5–5 neighboring families (gray) ranked according to an ascending value of mean gene per cell. CRC S3 healthy (g) stands for healthy colon tissue obtained from a patient with colorectal cancer.

concurrent: IC = 1.66 (95 % CI, 1.34–1.98).

These two families well illustrate the biological basis of concurrence and exclusivity, as detailed below. The expression of all homologs of the RFC family in a single cell is meaningful since all homologs of the Replication Factor C comprise the hetero-pentameric protein complex, consisting of RFC2 (40 kDa), RFC3 (38 kDa) RFC4 (37 kDa) RFC5 (36 kDa) and the non-family-member RFC1 (140 kDa) [46]. The formation of a single protein complex from five different proteins defines a single physiological state, the ability of the cell to replicate. In contrast, the NKAIN genes encode a protein that interacts with the b1 subunits of Na^+/K^+-ATPase, and only one protein interacts with the ATPase, modulating its activity. Thus, the complexation of NKAIN genes serves to increase the diversity of ATPase function in the cell population.

Most gene families are concurrent, which is likely explained by two factors, specific and inherent correlations. Firstly, homologs derived from a single gene ancestor often retain their original regulation, and genes and mRNAs regulated by the same pathway show coherent expression states [17,47–49]. Secondly, homologs have similar lengths, which entail common modes of control in transcription or mRNA degradation [47,50], as well as similar susceptibility to technical noise in the measurement of RNA [50], which introduce inherent correlations to most homologs. Therefore, we aimed here to identify true concurrent families that surpass the base concurrence due to the inherent correlation of homologous genes.

Inherent correlations of the homologs add up, as evidenced by the positive correlation between the mean gene per cell and the IC of each family (Pearson correlation, $\rho = 0.76$ and 0.54 for the S2 glioblastoma tumor and healthy cells, Supplementary Fig. 3 and Text). Therefore, we introduced the concept of excess concurrence that indicates the degree by which the IC of a family exceeds the IC of families with a similar number of expressed homologs. In the glioblastoma tumor and healthy cells, 52 % and 27 % of families displayed base concurrence (95 % CI of $IC > 1$), respectively. Next, we ranked families according to their mean number of expressed homologs, and compared the 95 % confidence intervals of the ICs of the 10 nearest neighbors (Section 2.3). If the 95 %CI of a given family surpassed that of all or all but one of its neighbors, we classified them as exhibiting excess concurrence. In this manner, only 1.9 % and 1.4 % of the families showed excess concurrence (Supplementary Fig. 3e, f), representing a small fraction of families with baseline concurrence.

3.2. The excess concurrence of the MCM gene family reflects the replicative status of cells

We identified gene families with excess concurrence across all cancer samples. The replication-licensing factor MCM family emerged as the most prevalent concurrent family, exhibiting excess concurrence in 80 % of the samples (Fig. 2a-c). Therefore, we conducted a more detailed analysis of this family. One of the glioblastoma datasets presents a unique opportunity to compare cancer cells with healthy cells in the tumor periphery in two different patients (S2, S4) [21]. In healthy cells (S2), the MCM family displayed an IC comparable to neighboring families, indicating no excess concurrence (Fig. 2d). In contrast, the MCM family exhibited strong concurrence in tumor cells, with no overlap of confidence intervals relative to neighboring families (Fig. 2e). Another glioblastoma (S4) had an intermediate concurrence (Fig. 2f) with two overlaps, which may indicate a lower grade of this tumor.

Since neurons and glia in adult human do not replicate or do so minimally, while tumor cells replicate frequently [51–53], the above results suggest that the excess concurrence reflects the replicative potential of the cells. We examined other cell types, as well. Healthy colon cells had an IC 1.8 times higher than the average of the neighboring families, but the confidence intervals overlapped with more than three families, indicating a borderline concurrence (Fig. 2g, Table S3). Thus, the concurrence level of the MCM family in the normally proliferating colon cells is intermediate between postmitotic neurons and rapidly proliferating tumor cells in glioblastoma. While the MCM family was concurrent in most tumors, it was not so in half of the oligodendrogliomas. This aligns with the fact that patients with oligodendroglioma have the longest survival rates among the brain tumors analyzed [54]. Thus, excess concurrence is a useful metric, reflecting the replication potential of cells.

In the subsequent text in this section, excess concurrence is simply referred to as concurrence. To identify all other conserved concurrent families, we calculated the average frequency of concurrent families across all cancer samples. We applied the binomial test to define the overrepresentation threshold, which required each concurrent family to be present in no fewer than eight cancer samples. Additionally, we identified families overrepresented in specific cancers. In this way, 43 out of the 286 concurrent families were overrepresented (Section 2.8, Table S4, Supplementary Fig. 4). Given that the MCM proteins are involved in the replication phase of the cell cycle, we quantified the enrichment of the cell cycle and related functions (DNA replication) in these concurrent families. Interestingly, the above cell cycle related functions were present in around half of the families (20/43, Fig. 3a). The concurrent MCM and RFC gene families display a significant overlap across the cancer samples (OR=4.74, $p = 0.04$, Fisher-exact test), indicating the coordinated concurrence of these two families.

While the MCM family ranked first and the Metallothionein family second in frequency of concurrence, the most enriched biological process among the conserved concurrent families was antigen processing and presentation (Fig. 3b). Antigen presentation is mediated by both MHC class I and class II molecules. Interestingly, the concurrent MHC I and Metallothionein families overlap significantly in the cancer samples (OR=4.47, $p = 0.002$, Fisher-exact test). Additionally, extracellular matrix disassembly, which is crucial for cell invasion and metastasis, was significantly enriched due to its association with multiple families.

3.3. Excess concurrence as a metric to assess the extent to which cell lines and organoids represent the original tissue

Interestingly, hierarchical clustering of cancers based on the frequency of concurrence in gene families arranges the cancers correctly according to their histological origin, such as estrogen dependent tissues (endometrial, ovarian cancers, DCIS) and brain tumors (oligodendroglioma, astrocytoma, glioma, glioblastoma) (Supplementary Fig. 4). Therefore, we used concurrence to assess how cancer cell lines and cancer organoids represent the cellular heterogeneity of the original cancers. In the colorectal cancer datasets, cancers and cancer organoids were compared (Section 2.1). In two of the four pairs, the concurrent families significantly overlapped between the cancer and organoids (Supplementary Fig. 5a). In the melanoma dataset, two distinct cell lines had a significant overlap with distinct cancer samples (Supplementary Fig. 5b).

3.4. Healthy cells in the tumor periphery express cell adhesion and ion homeostasis genes exclusively

Following the analysis of concurrence, we turned our attention to exclusivity. In the glioblastoma dataset, a significant number of exclusive families ($n_t = 5$), are conserved between the healthy peripheral tissues of the two patients (Odds ratio (OR) = 12, P-value, $p = 0.0002$, two-tailed Fisher exact test, Fig. 4a). Two out of the five families are involved in ion transport and homeostasis, including the Sodium/Potassium transporting ATPase unit Gamma, and the carbonic anhydrase. Further two families, the fibrillin and the multicopper oxidase related genes are involved in cell adhesion.

The above findings are in agreement with the finding that cell adhesion and ion homeostasis are highly enriched among the exclusive families in healthy mouse cells [17]. Just as in mouse cells, the gene families can contribute to cell adhesion directly or indirectly. The binding of fibrillins to integrins promotes cell adhesion in a variety of

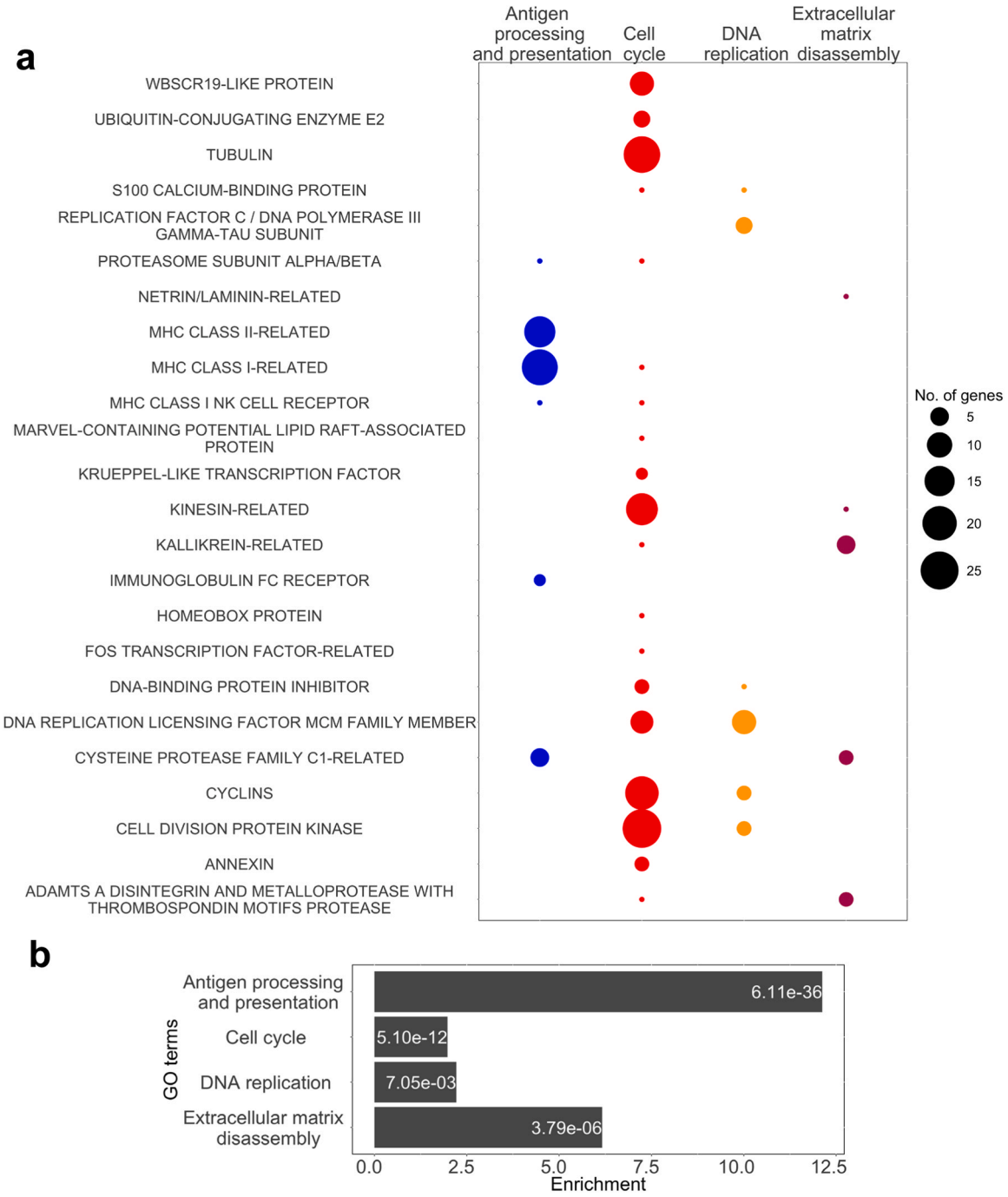

Fig. 3. Enrichment of biological processes in concurrent gene families. a, 24 of the 43 concurrent families that contribute to the selected GO terms are shown (Table S7). The size of the circles in the bubble plot represents the number of genes in each family associated with the specific GO term. **b,** The P-values are displayed (fold enrichment in parentheses) for antigen processing and presentation (12.1), cell cycle (2.0), DNA replication is (2.2), and extracellular matrix disassembly (7.9).

tissues [55], while fibulin-2, which also belongs to this family, binds to proteoglycans in the extracellular matrix and mediate synapse formation [56,57]. Among the multicopper oxidase related genes, EDIL3 binds integrin [58], and retinoschisin, a lectin, mediates cell-cell adhesion [59]. Several genes couple cell adhesion with ion homeostasis. For instance, retinoschisin binds to the extracellular domain of Na/K-ATPase subunit β2 [60].

There was no significant overlap between the exclusive families of the periphery and tumors within the same patient (Supplementary Fig. 6a, b), which suggests that gene families can lose or gain exclusivity during tumorigenesis. The number of exclusive families reduces in the tumors of both patients, by a factor of 2 and 1.4, with the changes being marginally or not significant (Fig. 4c). To assess this trend in more detail, we analyzed the direction of changes in the ICs of the families. The families have a significantly higher ICs in the tumors than in healthy cells (sign-test; $p = 2.2{\cdot}10^{-16}$ (S2), $6.2{\cdot}10^{-4}$ (S4); Fig. 4e, f). For example, antigen processing by the MHC class II-related family profits from this change displaying excess concurrence in tumor cells (S4, IC (healthy)= 1.51 and IC(tumor)= 3.87, permutation test $p = 0.0015$).

We examined the changes of specific gene families that show exclusivity in many cell types in mouse [17]: the Na/K transporting ATPase subunit gamma and the protocadherin (Fig. 4d). Both of these families shift toward concurrence in tumors. For the protocadherin alpha-array, the IC is significantly higher in tumor (S2) than in the

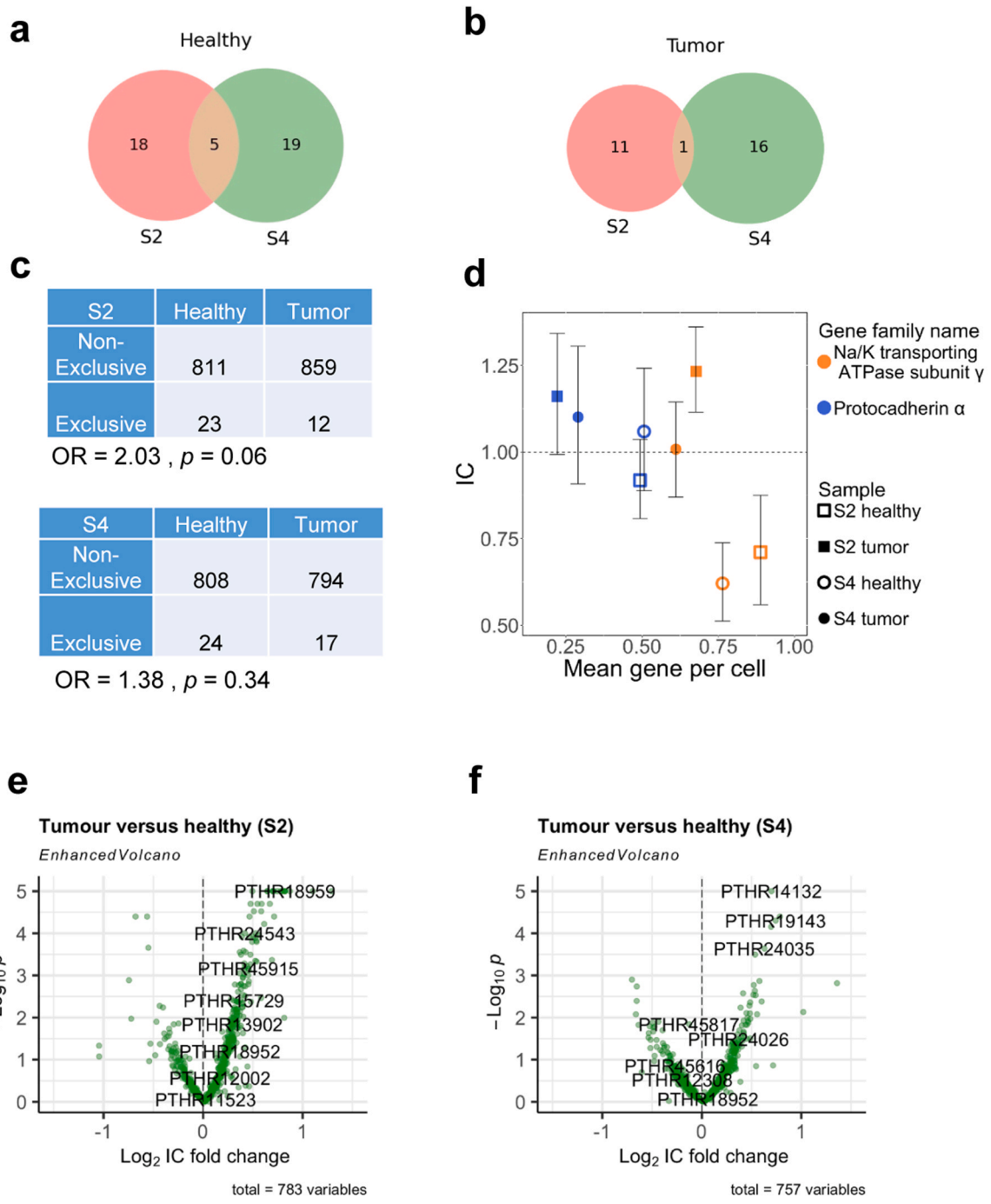

Fig. 4. Exclusive families in the cancer and healthy cell populations of the glioblastoma dataset. a, There is a significant number of families shared between the tumors of the two patients (S2, S4): Carbonic anhydrase, Metabotropic glutamate receptor, Sodium/Potassium-transporting ATPase unit Gamma, Multicopper oxidase-related and Fibrillin-related. Number of nonexclusive families (*N*) = 837. **b,** Only the Guanylate cyclase soluble subunit beta-2 family is shared between the tumors of the two patients; $N = 880$. **c,** The P-value for the association between the exclusivity and pathology (glioblastoma tumor/healthy cells) is calculated with the two-tailed Fisher exact test for patients S2 and S4. **d,** IC as a function of mean gene per cell for two families in samples from Glioblastoma patients S2 and S4. The error bars represent the 95 %CI of the IC. The P-values (permutation test) for the IC fold change (fc) between tumor and healthy samples are given in parentheses along with stars: Na/K transporting ATPase subunit γ, fc = 1.73 (10^{-5})* ** for S2 and fc = 1.62 (10^{-5})* ** for S4. For Protocadherin-α, fc = 1.26 (0.01)* * for S2 and fc = 1.03 (0.85) for S4. **e,** For patient S2, 25 families had a significantly higher *IC* in the tumor of than in the healthy cells, and only two families had a lower IC. The P-value is calculated with permutation test. The significance level is determined with the Bonferroni-Holm correction ($\alpha = 6{\cdot}10^{-5}$). The families exclusive in either healthy or tumor cells are indicated by Panther numbers. **f,** For patient S4, only 3 families have a significantly higher IC in cancer cells, and none of them lower.

healthy periphery despite the fact that the expression declines (from mgpc = 0.49 to mgpc = 0.22). Thus, a shift toward concurrence is coupled with a reduction in the number of expressed gene homologs, showing that exclusivity does not necessarily profit from a lower number of chosen genes. In addition to the involvement of Na/K transporting ATPase gamma in ion homeostasis, it acts also in cell adhesion, indicating a decline of exclusivity in both gene families involved in cell adhesion in glioblastoma.

Not only do gene choice patterns change, but also the expression of individual genes change in each family, exemplified by the carbonic

anhydrases (Car). Many catalytic Car genes increase their expression in cancer (Car2, 3, 9, 12). Some Cars are acatalytic (Car10, Car11) and Car10 is involved in cell-cell adhesion. Car10 expression drops to zero in tumors of both patients (Table S5), suggesting that the cell adhesion property of this family is lost in cancer.

3.5. Changes in gene expression drive changes in gene choice pattern in some gene families

In essence, stochastic gene choice can change with or without significant changes in average gene expression. In the latter case, the expression states of the genes are simply rearranged among the individual cells. To explore these scenarios, we compared cancer and healthy cells, using the Wilcoxon test to calculate the P-value for the differential expression of each gene. To assess the significance level due to multiple testing of genes, we calculated the minP adjusted P-values because it accounts for the single-cell distribution of gene expression (Section 2.6).

For patients S2, 18 % of the genes showed significant changes in expression. For the other patient (S4), only 8 % of the genes changed (Fig. 5a-d), indicating a cancer with less progression, which may explain its lower replicative potential (Fig. 2e, f, Table S3). We used the Fisher exact test for the association between significant changes in gene expression and IC. The odds ratio (OR) is around two for both patients, which was significant for patient S2 (Fig. 5e, f). This shows that alterations in the bulk gene expression are significantly associated with changes in IC. The expression changes do not affect exclusivity more than other forms of stochastic gene choice (Supplementary Fig. 6c).

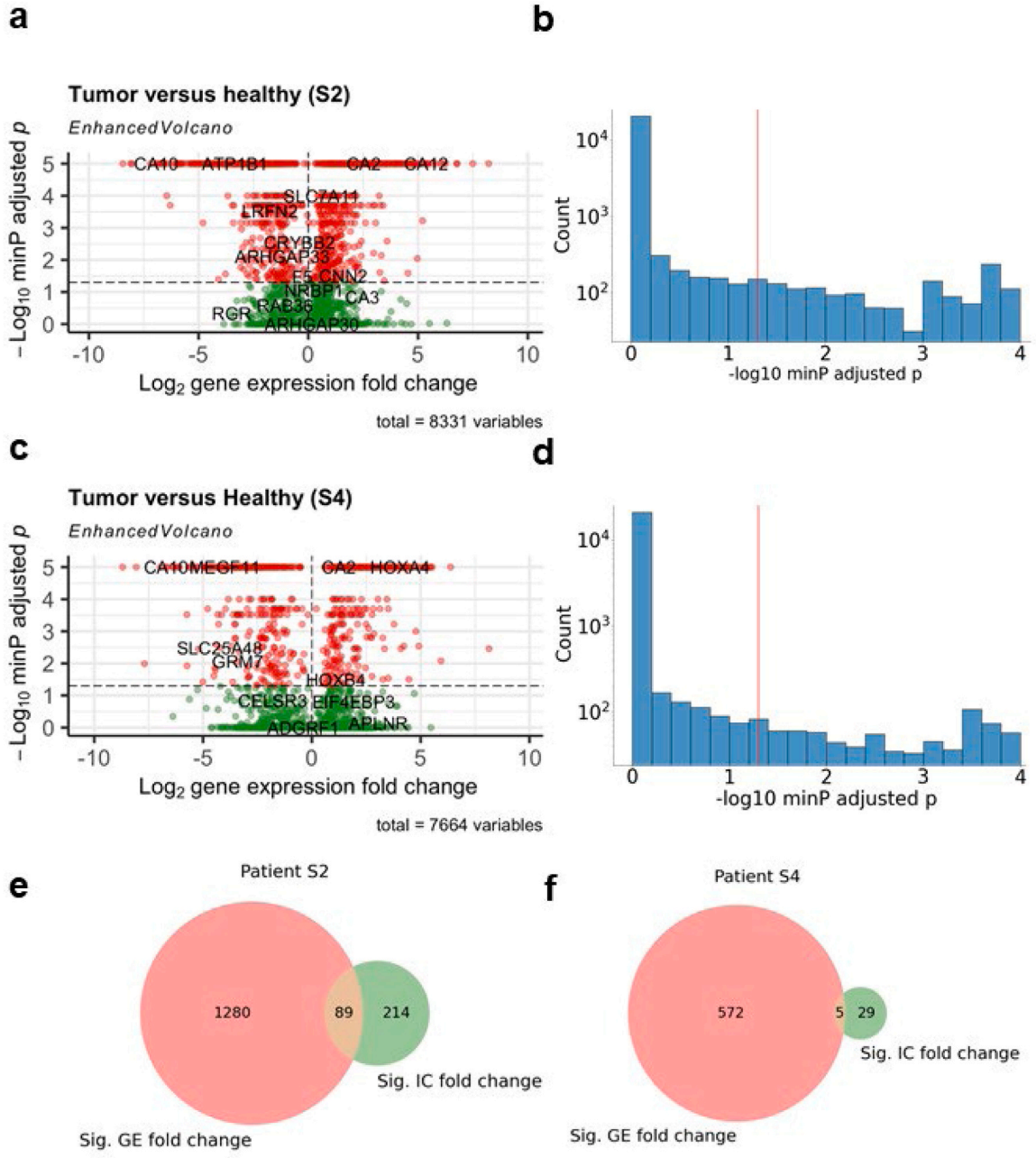

Fig. 5. Change in gene expression and its association with stochastic gene choice. The significance of fold change in gene expression (minP adjusted) was calculated for genes that belong to families with mean gene per cell > 0.1 and non-zero genes > 3. **(a-b)** 1517 out of 8331 (18.2 %) genes show a significant change in expression **(c,d)** is 637 out of 7664 (8.31 %) genes show a significant change in expression. **e,** Association is calculated with two-tailed Fisher exact test; OR $= 1.94$, $p = 9.0{\cdot}10^{-7}$, $N = 5981$ genes change neither the expression nor the IC significantly. **f,** OR $= 1.93$, $p = 0.20$, $N = 6391$ (as in e).

Notwithstanding, there are gene families with significant IC change in which none of the genes showed a significant change in gene expression, exemplified by the ATP-dependent RNA helicase (S2), and the sodium/potassium-transporting ATPase unit gamma (S4) families. In summary, changes in average gene expression and gene choice patterns are significantly associated, but examples without such association are also relevant for exclusive families.

3.6. Cancer cells express genes involved in ion transport, cell motility and migration exclusively

While exclusive gene families in healthy cells are associated with cell adhesion and ion homeostasis [17], their biological role in tumors is unclear. The exclusive gene families in the tumors of the two glioblastoma patients do not show significant overlap (OR = 4.98, $p = 0.20$, Fig. 4b), partly due to the slightly lower number of exclusive families in cancer samples. Therefore, we analyzed comprehensively all tumor datasets to identify conserved exclusive families using the binomial test. This analysis identified eight exclusive families present in at least seven datasets. We also included exclusive families overrepresented in specific cancers, adding four other families (Fig. 6a).

Subsequently, we performed a gene ontology enrichment analysis of these families (Section 2.9). We found enrichment in transport, including transmembrane and metal ion transport in most families. Seven families are associated with metal ion transport, such as the NKAINS (Na^+/K^+ transporting ATPase interacting 2) and the Sodium glucose transporters (SGLTs) belonging to the mammalian solute carrier family SLC5 (Fig. 6b). Furthermore, biological processes related to cell migration, including actin-filament based movement, and regulation of anatomical structure size, were enriched.

The lack of conservation of cell adhesion across the tumor samples supports the observation that specific families involved in cell adhesion increase their IC or lose exclusivity altogether in glioblastoma (Fig. 4d, Table S6). The loss of cell adhesion and the increased motility of cells in tumors are key aspects of the epithelial-mesenchymal transition (EMT). Therefore, we examined the Glial fibrillary acidic protein (GFAP) family, which includes important EMT marker genes, such as vimentin (VIM) [61,62]. VIM participates in cell migration of mesenchymal cells. VIM expression is higher in glioblastoma than in healthy periphery (Table S5). Interestingly, the IC of the GFAP family decreases significantly in the tumor ($IC_{healthy} = 1.31$ and $IC_{tumor} = 1.07$, $p = 0.014$, permutation-test, S2), in agreement with the enrichment of exclusive families involved in migration and-actin based motility, in tumors.

.

4. Discussion

Concurrence is concentrated in specific families, such as the MCM family (Table S4). The MCMs form a protein complex that licenses replication during the G1/S transition [63]. DNA replication, cell cycle and division are in fact biological processes associated with about half of the overrepresented concurrent families, like the replication factor C, cell division protein kinase (CDK), cyclins, tubulin, and the recently identified WBSCR19 family. The latter encodes the speedy/RINGO cell cycle regulator, which can bind and activate the Cdk directly [64]. The Ubiquitin-conjugating enzymes act in proteolysis, which drives the cell cycle.

How do the above families benefit from concurrence? The cells with zero and many expressed genes most likely correspond to non-dividing and replicating cells, respectively. Alternatively, the all-or-none expression can reflect mechanisms that drive alternating expression of these genes during the different phases of the cell cycle. Concurrence in gene expression is common across organisms, including even single-celled organisms like yeast, where genes activated by the same pathway exhibit correlated stochastic gene expression [65]. In this study, we analyzed concurrence that exceeds inherent correlations, which may require additional mechanisms for coordination beyond the co-regulation of shared targets. The pronounced concurrence seen in the MCM family likely reflects a control mechanism that maintains a proper stoichiometry of MCM components post-transcriptionally at both mRNA and protein levels [66]. This form of dosage compensation ensures a proper ratio of subunits in the replication licensing protein complex.

Replication rates vary widely in human cells. For example, brain cells such as neurons, oligodendrocytes and astrocytes undergo little to no detectable division in adults and are considered post-mitotic [51,52,67]. In contrast, the enterocyte population renews every 2–5 days [68]. Glioblastoma cells can divide very rapidly, with a rate of 0.1 to 1 division per day [53]. Higher division rates may worsen the prognosis [69]. We found the following ascending order of concurrence in the MCM family: healthy post-mitotic brain cells, healthy dividing cells (colon epithelium), brain tumors with long survival (oligodendroglioma) and brain tumors with lower survival rates (glioma, glioblastoma). Thus, MCM concurrence is a good indicator of the replication potential and cancer progression. Furthermore, concurrence can be used to cluster cell types in functional groups and to decide if cell lines and organoids reproduce the heterogeneity of the original cell population of tumors (Supplementary Figs. 4 and 5).

Antigen processing and presentation is the most enriched biological process among the concurrent families, partly because multiple gene families contribute to this gene ontology. While MHC-I homologs, such as HLA-A, HLA-B and HLA-C, present antigens in somatic cells, including groups of neurons [70], cancer cells of various origins can also present antigens by MHC-II, which is normally restricted to professional immune cells, such as dendritic cells [71]. The mechanism underlying concurrence in antigen presentation remains to be determined. Although a loss of HLA heterozygotic expression, i.e. allele-specific expression loss [72], has been observed in cancer cells, it is unclear if this contributes to the concurrence.

The complex regulatory schemes that enable exclusive gene choice are expected to be disrupted in the dysfunctional cells of cancers. Indeed, exclusive patterns decline during tumorigenesis and exclusivity is less concentrated in specific families than concurrence (Fig. 2a). However, cancer cells maintain exclusive gene choice in transmembrane ion transport but the involved families change during tumorigenesis. The γ-subunit of the Na^+/K^+ ATPase family (Fxyd genes) is one of the most exclusive gene families in murine cells [17], and was also detected in healthy cells in the tumor periphery (Fig. 4d). This family is, however, replaced during tumorigenesis by a functionally similar exclusive family, the T-cell lymphoma breakpoint associated (TCBA) family, encoding the NKAIN proteins that interact with the b1 subunits of Na^+/K^+-ATPase [73]. Thus, they both function in ion transport. As its name suggests, the TCBA family is frequently affected by chromosomal rearrangements. The NKAIN genes may be oncogenic [45], just like another exclusive family, the opsins [74].

Furthermore, exclusive patterns shift from cell adhesion of healthy cells to migration and regulation of anatomical structure size. The loss of cell adhesion is promoted by both the replacement of cell-adhesion specific forms of some molecules, like carbonic anhydrase, as well as by the reduction of exclusivity in the protocadherin alpha-array. Since the homophilic interaction of protocadherin homologs leads to repulsion [75], a shift toward concurrence can lead to the loss of cell adhesion (Fig. 6c). The loss of cell adhesion and acquisition of migratory ability of cells is a key change during tumorigenesis that is termed the epithelial-mesenchymal transition. Thus, this transition is regulated by stochastic exclusive gene choice [62,76]. This transition can be further promoted by the shift toward exclusivity in EMT families, like the GFAP family. The role of ion transport remains to be determined but it may contribute to epithelial-mesenchymal transition. Both cell adhesion and cell migration can be affected by ion-mediated regulation of protein interactions and cell protrusions [77]. Through the local regulation of osmotic pressure, cell protrusions can develop, and ion homeostasis also interacts with actin to facilitate actin-mediated protrusions (Fig. 6d)

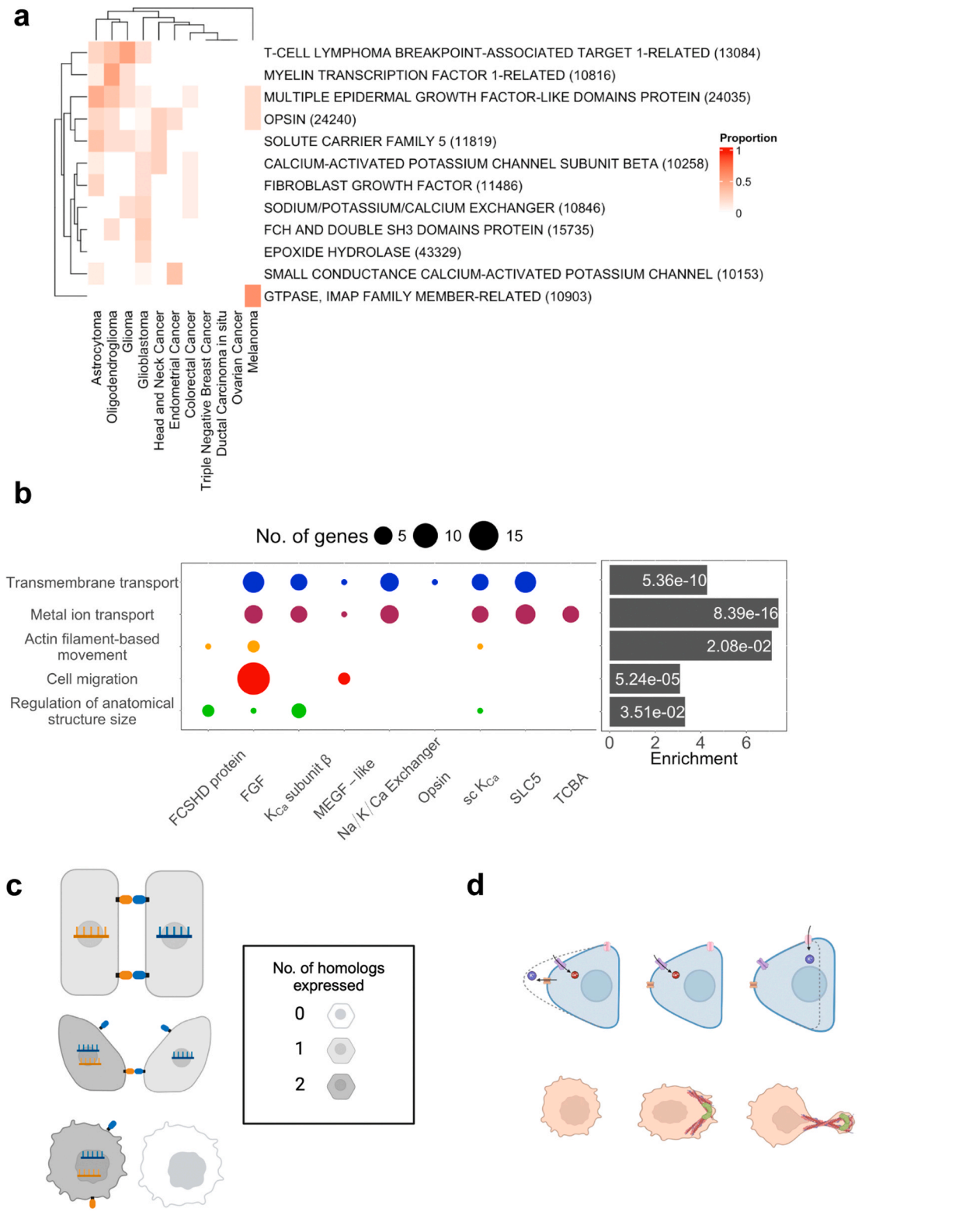

Fig. 6. Gene families with exclusive gene choice in cancer cells. a, The heatmap shows the frequency of exclusive gene families in different cancer types. **b,** Association of GOs with the exclusive families. The size of the circles in the bubble plot denotes the number of genes in each family that is associated with the specific GO. 9 out of the 12 overrepresented exclusive contribute to the indicated GOs: FCH and double SH3 domains protein (FCHSD protein), fibroblast growth factor (FGF), calcium-activated potassium channel subunit beta (KCa subunit β), Multiple epidermal growth factor-like domains protein (MEGF-like), sodium/potassium/calcium exchanger (Na/K/Ca Exchanger), opsin, small conductance calcium-activated potassium channel (sc KCa), solute carrier family 5 (SLC5) and t-cell lymphoma breakpoint-associated target 1-related (TCBA). The P-values associated with the fold-enrichment are displayed. **c,** A scheme showing how loss of exclusivity can lead to a loss of cell adhesion during tumorigenesis. Exclusive expression leads to the adhesion (top panel) mediated between different homologs. A higher variability leads to the expression of variable number of homologs, and hemophilic interaction (blue-blue) leads to a partial repulsion (middle panel). Large variability (concurrence) can results in all-or-none response when the cells cannot interact (bottom panel). **d,** Scheme depicting how differential ion transport leads to cell shrinkage and extension (top panels), which leads to cell migration, which is also supported by actin-filament based movement (bottom panels).

[77,78].

What are the advantages of exclusivity for the above families? In higher organisms, prototypical exclusive families, like the odorant and T-cell receptors, are involved in sensing external or internal environment. Likewise, ion channels may sense the local environment, assisting migrating cells in adapting to the ever-evolving surroundings.

Our analysis is based on datasets primarily containing brain tumors, so our conclusions will have to be evaluated in the context of other cancer types since gene families can behave differently in different cell types. For example, the protocadherin-alpha array is exclusive in some cell types but not in others [17]. Additionally, we had only two datasets that allowed for a comparison between healthy and tumor cells. However, the findings in healthy cells align well with those observed in mouse cells. Further datasets that include both healthy and tumor cells will be important to understand the details of how concurrence and exclusivity change during tumorigenesis.

Since it can be advantageous to design therapies that take into account heterogeneity in cancer cell populations [79,80], gene choice patterns are relevant in this respect. In a concurrent biological process, such as replication, cell cycle or antigen presentation, it may be sufficient to target a single component of the gene family / protein complex, because the all-or-none response implies only two states and hitting a single component can fully inhibit the process, such as replication or cell division. By hitting this component, only cells not expressing the complex or not replicating at the time of drug administration remain unaffected. This cell subpopulation can be targeted by repeated application of the drug so that ultimately most cells enter the relevant stages of the cell cycle or when tumor cells present antigens. Indeed, such a dosing constitutes the cornerstone of classical chemotherapy, which is typically aimed at eliminating replicating cells. At the same time, exclusive gene choice generates a larger repertoire of cell identities during epithelial-mesenchymal transition, suggesting that a combinatorial therapy with simultaneous targeting of multiple genes/proteins is likely to be needed to block the transition to cell identities defined by migratory ability.

5. Conclusions

Concurrence in the gene choice of the MCM and RFC families, which underlie DNA replication, is prevalent in cancer samples. This pattern can ensure proper stoichiometry of replication components, correlating with higher replication rates and cancer progression. Antigen processing and presentation are also enriched among concurrent families, though the underlying mechanism is unclear. In tumor cells, ion transport maintains its exclusivity, albeit with changes in the genes and gene families involved. This resulting diversity may help migrating cancer cells cope with their continually evolving surroundings.

Code availability

The algorithms used in this work are available in the GitHub repository: https://github.com/SM205/stochastic-cancer.

Author statement

A.B. conceived and supervised the study, S.M. developed the new algorithms, S.M. and A.B. performed the data analysis, A.B and S.M. wrote and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Samuel Mondal: Data curation, Formal analysis, Software, Visualization, Writing – original draft, Writing – review & editing. **Attila Becskei:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that the publication of this paper has no conflict of interest.

Acknowledgement

We thank S. Faravelli for the help with the initial dataset collection. This work was supported by the Swiss National Foundation (310030_185001).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.06.004.

References

[1] Lin JR, Chen YA, Campton D, Cooper J, Coy S, et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. Nat Cancer 2023;4:1036–52. https://doi.org/10.1038/s43018-023-00576-1.
[2] Liberti MV, Locasale JW. The Warburg effect: how does it benefit cancer cells? Trends Biochem Sci 2016;41:211–8. https://doi.org/10.1016/j.tibs.2015.12.001.
[3] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.
[4] Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med 2023;29:554–66. https://doi.org/10.1016/j.molmed.2023.03.007.
[5] Raimondi F, Inoue A, Kadji FMN, Shuai N, Gonzalez JC, et al. Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm. Oncogene 2019;38:6491–506. https://doi.org/10.1038/s41388-019-0895-2.
[6] Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;47:106–14. https://doi.org/10.1038/ng.3168.
[7] Zhao X, Hu J, Li Y, Guo M. Volumetric compression develops noise-driven single-cell heterogeneity. Proc Natl Acad Sci USA 2021;118. https://doi.org/10.1073/pnas.2110550118.
[8] Saxena K, Subbalakshmi AR, Kulkarni P, Jolly MK. Cancer: More than a geneticist's Pandora's box. J Biosci 2022;47.
[9] Lee J, Lee J, Farquhar KS, Yun J, Frankenberger CA, et al. Network of mutually repressive metastasis regulators can promote cell heterogeneity and metastatic transitions. Proc Natl Acad Sci USA 2014;111:E364–73. https://doi.org/10.1073/pnas.1304840111.
[10] Jha A, Quesnel-Vallieres M, Wang D, Thomas-Tikhonenko A, Lynch KW, Barash Y. Identifying common transcriptome signatures of cancer by interpreting deep learning models. Genome Biol 2022;23:117. https://doi.org/10.1186/s13059-022-02681-3.
[11] Roelands J, van der Ploeg M, Ijsselsteijn ME, Dang H, Boonstra JJ, et al. Transcriptomic and immunophenotypic profiling reveals molecular and immunological hallmarks of colorectal cancer tumourigenesis. Gut 2023;72:1326–39. https://doi.org/10.1136/gutjnl-2022-327608.
[12] Teruya N, Inoue H, Horii R, Akiyama F, Ueno T, et al. Intratumoral heterogeneity, treatment response, and survival outcome of ER-positive HER2-positive breast cancer. Cancer Med 2023;12:10526–35. https://doi.org/10.1002/cam4.5788.
[13] Wada T, Wallerich S, Becskei A. Stochastic gene choice during cellular differentiation. Cell Rep 2018;24:3503–12. https://doi.org/10.1016/j.celrep.2018.08.074.
[14] Dornburg A, Mallik R, Wang Z, Bernal MA, Thompson B, et al. Placing human gene families into their evolutionary context. Hum Genom 2022;16:56. https://doi.org/10.1186/s40246-022-00429-5.
[15] Demuth JP, Hahn MW. The life and death of gene families. Bioessays 2009;31:29–39. https://doi.org/10.1002/bies.080085.
[16] Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian gene families. e85 PLoS One 2006;1. https://doi.org/10.1371/journal.pone.0000085.
[17] Iakovlev M, Faravelli S, Becskei A. Gene families with stochastic exclusive gene choice underlie cell adhesion in mammalian cells. Front Cell Dev Biol 2021;9:642212. https://doi.org/10.3389/fcell.2021.642212.
[18] Zhou Y, Xu S, Zhang M, Wu Q. Systematic functional characterization of antisense eRNA of protocadherin alpha composite enhancer. Genes Dev 2021;35:1383–94. https://doi.org/10.1101/gad.348621.121.
[19] Wu Q, Jia Z. Wiring the brain by clustered protocadherin neural codes. Neurosci Bull 2021;37:117–31. https://doi.org/10.1007/s12264-020-00578-4.

[20] Tian XJ, Zhang H, Sannerud J, Xing J. Achieving diverse and monoallelic olfactory receptor selection through dual-objective optimization design. Proc Natl Acad Sci USA 2016;113:E2889–98. https://doi.org/10.1073/pnas.1601722113.
[21] Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, et al. Single-Cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Rep 2017;21:1399–410. https://doi.org/10.1016/j.celrep.2017.10.030.
[22] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. e21 Cell 2019;178:835–49. https://doi.org/10.1016/j.cell.2019.06.024.
[23] Cochrane DR, Campbell KR, Greening K, Ho GC, Hopkins J, et al. Single cell transcriptomes of normal endometrial derived organoids uncover novel cell type markers and cryptic differentiation of primary tumours. J Pathol 2020;252:201–14. https://doi.org/10.1002/path.5511.
[24] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 2021;39:599–608. https://doi.org/10.1038/s41587-020-00795-2.
[25] Bian S, Hou Y, Zhou X, Li X, Yong J, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science 2018;362:1060–3. https://doi.org/10.1126/science.aao3791.
[26] Tirosh I, Izar B, Prakadan SM, Wadsworth 2nd MH, Treacy D, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189–96. https://doi.org/10.1126/science.aad0501.
[27] Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. e24 Cell 2017;171:1611–24. https://doi.org/10.1016/j.cell.2017.10.044.
[28] Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 2018;9:3588. https://doi.org/10.1038/s41467-018-06052-0.
[29] Wang R, Mao Y, Wang W, Zhou X, Wang W, et al. Systematic evaluation of colorectal cancer organoid system by single-cell RNA-Seq analysis. Genome Biol 2022;23:106. https://doi.org/10.1186/s13059-022-02673-3.
[30] Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science 2017;355. https://doi.org/10.1126/science.aai8478.
[31] Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science 2018;360:331–5. https://doi.org/10.1126/science.aao4750.
[32] Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539:309–13. https://doi.org/10.1038/nature20123.
[33] Gerber T, Willscher E, Loeffler-Wirth H, Hopp L, Schadendorf D, et al. Mapping heterogeneity in patient-derived melanoma cultures by single-cell RNA-seq. Oncotarget 2017;8:846–62. https://doi.org/10.18632/oncotarget.13666.
[34] Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, et al. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. e5 Cell Syst 2018;6:171–9. https://doi.org/10.1016/j.cels.2018.01.014.
[35] Kim H, Lee J, Kang K, Yoon S. MarkerCount: A stable, count-based cell type identifier for single-cell RNA-seq experiments. Comput Struct Biotechnol J 2022;20:3120–32. https://doi.org/10.1016/j.csbj.2022.06.010.
[36] Hu C, Li T, Xu Y, Zhang X, Li F, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. D870-D6 Nucleic Acids Res 2023;51. https://doi.org/10.1093/nar/gkac947.
[37] Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci 2022;31:8–22. https://doi.org/10.1002/pro.4218.
[38] Hong YL. On computing the distribution function for the Poisson binomial distribution. Comput Stat Data 2013;59:41–51. https://doi.org/10.1016/j.csda.2012.10.006.
[39] Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 2018;15:255–61. https://doi.org/10.1038/nmeth.4612.
[40] Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment. New York: Wiley; 1993. xvii, 340 p. p.
[41] Ernst MD. Permutation methods: a basis for exact inference. Stat Sci 2004;19:676–85. https://doi.org/10.1214/088342304000000396.
[42] Menyhart O, Weltz B, Gyorffy B. MultipleTesting.com: a tool for life science researchers for multiple hypothesis testing correction. PLoS One 2021;16:e0245824. https://doi.org/10.1371/journal.pone.0245824.
[43] Wu T, Hu E, Xu S, Chen M, Guo P, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innov (Camb) 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141.
[44] Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300. doi: DOI 10.1111/j.2517-6161.1995.tb02031.x.
[45] Zhao SC, Zhou BW, Luo F, Mao X, Lu YJ. The structure and function of NKAIN2-a candidate tumor suppressor. Int J Clin Exp Med 2015;8:17072–9.
[46] Li Y, Gan S, Ren L, Yuan L, Liu J, et al. Multifaceted regulation and functions of replication factor C family in human cancers. Am J Cancer Res 2018;8:1343–55.
[47] Jaquet V, Wallerich S, Voegeli S, Turos D, Viloria EC, Becskei A. Determinants of the temperature adaptation of mRNA degradation. Nucleic Acids Res 2022;50:1092–110. https://doi.org/10.1093/nar/gkab1261.
[48] Baudrimont A, Jaquet V, Wallerich S, Voegeli S, Becskei A. Contribution of RNA degradation to intrinsic and extrinsic noise in gene expression. e5 Cell Rep 2019;26:3752–61. https://doi.org/10.1016/j.celrep.2019.03.001.
[49] Ghoshdastider U, Sendoel A. Exploring the pan-cancer landscape of posttranscriptional regulation. Cell Rep 2023;42:113172. https://doi.org/10.1016/j.celrep.2023.113172.
[50] Rahaman S, Faravelli S, Voegeli S, Becskei A. Polysome propensity and tunable thresholds in coding sequence length enable differential mRNA stability. Sci Adv 2023;9:eadh9545. https://doi.org/10.1126/sciadv.adh9545.
[51] Sorrells SF, Paredes MF, Cebrian-Silla A, Sandoval K, Qi D, et al. Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. Nature 2018;555:377–81. https://doi.org/10.1038/nature25975.
[52] Colodner KJ, Montana RA, Anthony DC, Folkerth RD, De Girolami U, Feany MB. Proliferative potential of human astrocytes. J Neuropathol Exp Neurol 2005;64:163–9. https://doi.org/10.1093/jnen/64.2.163.
[53] Larsson I, Dalmo E, Elgendy R, Niklasson M, Doroszko M, et al. Modeling glioblastoma heterogeneity as a dynamic network of cell states. e10105 Mol Syst Biol 2021;17. https://doi.org/10.15252/msb.202010105.
[54] Visser O, Ardanaz E, Botta L, Sant M, Tavilla A, et al. Survival of adults with primary malignant brain tumours in Europe; results of the EUROCARE-5 study. Eur J Cancer 2015;51:2231–41. https://doi.org/10.1016/j.ejca.2015.07.032.
[55] Zeyer KA, Zhang RM, Kumra H, Hassan A, Reinhardt DP. The Fibrillin-1 RGD integrin binding site regulates gene expression and cell function through microRNAs. J Mol Biol 2019;431:401–21. https://doi.org/10.1016/j.jmb.2018.11.021.
[56] Olin AI, Morgelin M, Sasaki T, Timpl R, Heinegard D, Aspberg A. The proteoglycans aggrecan and Versican form networks with fibulin-2 through their lectin domain binding. J Biol Chem 2001;276:1253–61. https://doi.org/10.1074/jbc.M006783200.
[57] Patel MR, Weaver AM. Astrocyte-derived small extracellular vesicles promote synapse formation via fibulin-2-mediated TGF-beta signaling. Cell Rep 2021;34:108829. https://doi.org/10.1016/j.celrep.2021.108829.
[58] Ozawa A, Sato Y, Imabayashi T, Uemura T, Takagi J, Sekiguchi K. Molecular basis of the ligand binding specificity of alphavbeta8 integrin. J Biol Chem 2016;291:11551–65. https://doi.org/10.1074/jbc.M116.719138.
[59] Tolun G, Vijayasarathy C, Huang R, Zeng Y, Li Y, et al. Paired octamer rings of retinoschisin suggest a junctional model for cell-cell adhesion in the retina. Proc Natl Acad Sci USA 2016;113:5287–92. https://doi.org/10.1073/pnas.1519048113.
[60] Plossl K, Royer M, Bernklau S, Tavraz NN, Friedrich T, et al. Retinoschisin is linked to retinal Na/K-ATPase signaling and localization. Mol Biol Cell 2017;28:2178–89. https://doi.org/10.1091/mbc.E17-01-0064.
[61] Groves SM, Panchy N, Tyson DR, Harris LA, Quaranta V, Hong T. Involvement of epithelial-mesenchymal transition genes in small cell lung cancer phenotypic plasticity. Cancers (Basel) 2023;15. https://doi.org/10.3390/cancers15051477.
[62] Majc B, Sever T, Zaric M, Breznik B, Turk B, Lah TT. Epithelial-to-mesenchymal transition as the driver of changing carcinoma and glioblastoma microenvironment. Biochim Biophys Acta Mol Cell Res 2020;1867:118782. https://doi.org/10.1016/j.bbamcr.2020.118782.
[63] Noguchi Y, Yuan Z, Bai L, Schneider S, Zhao G, et al. Cryo-EM structure of Mcm2-7 double hexamer on DNA suggests a lagging-strand DNA extrusion model. Proc Natl Acad Sci USA 2017;114:E9529–38. https://doi.org/10.1073/pnas.1712537114.
[64] Gonzalez L, Nebreda AR. RINGO/Speedy proteins, a family of non-canonical activators of CDK1 and CDK2. Semin Cell Dev Biol 2020;107:21–7. https://doi.org/10.1016/j.semcdb.2020.03.010.
[65] Stewart-Ornstein J, Weissman JS, El-Samad H. Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell 2012;45:483–93. https://doi.org/10.1016/j.molcel.2011.11.035.
[66] Chuang CH, Yang D, Bai G, Freeland A, Pruitt SC, Schimenti JC. Post-transcriptional homeostasis and regulation of MCM2-7 in mammalian cells. Nucleic Acids Res 2012;40:4914–24. https://doi.org/10.1093/nar/gks176.
[67] Yeung MS, Zdunek S, Bergmann O, Bernard S, Salehpour M, et al. Dynamics of oligodendrocyte generation and myelination in the human brain. Cell 2014;159:766–74. https://doi.org/10.1016/j.cell.2014.10.011.
[68] Darwich AS, Aslam U, Ashcroft DM, Rostami-Hodjegan A. Meta-analysis of the turnover of intestinal epithelia in preclinical animal species and humans. Drug Metab Dispos 2014;42:2016–22. https://doi.org/10.1124/dmd.114.058404.
[69] Tini P, Yavoroska M, Mazzei MA, Miracco C, Pirtoli L, et al. Low expression of Ki-67/MIB-1 labeling index in IDH wild type glioblastoma predicts prolonged survival independently by MGMT methylation status. J Neurooncol 2023;163:339–44. https://doi.org/10.1007/s11060-023-04342-2.
[70] Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci USA 2015;112:7285–90. https://doi.org/10.1073/pnas.1507125112.
[71] Johnson AM, Bullock BL, Neuwelt AJ, Poczobutt JM, Kaspar RE, et al. Cancer cell-intrinsic expression of MHC Class II regulates the immune microenvironment and response to Anti-PD-1 therapy in lung adenocarcinoma. J Immunol 2020;204:2295–307. https://doi.org/10.4049/jimmunol.1900778.
[72] Filip I, Wang A, Kravets O, Orenbuch R, Zhao J, et al. Pervasiveness of HLA allele-specific expression loss across tumor types. Genome Med 2023;15:8. https://doi.org/10.1186/s13073-023-01154-x.
[73] Gorokhova S, Bibert S, Geering K, Heintz N. A novel family of transmembrane proteins interacting with beta subunits of the Na,K-ATPase. Hum Mol Genet 2007;16:2394–410. https://doi.org/10.1093/hmg/ddm167.
[74] de Assis LVM, Lacerda JT, Moraes MN, Dominguez-Amorocho OA, Kinker GS, et al. Melanopsin (Opn4) is an oncogene in cutaneous melanoma. Commun Biol 2022;5:461. https://doi.org/10.1038/s42003-022-03425-6.

[75] Rubinstein R, Goodman KM, Maniatis T, Shapiro L, Honig B. Structural origins of clustered protocadherin-mediated neuronal barcoding. Semin Cell Dev Biol 2017; 69:140–50. https://doi.org/10.1016/j.semcdb.2017.07.023.
[76] Haerinck J, Goossens S, Berx G. The epithelial-mesenchymal plasticity landscape: principles of design and mechanisms of regulation. Nat Rev Genet 2023;24: 590–609. https://doi.org/10.1038/s41576-023-00601-0.
[77] Stock C, Schwab A. Ion channels and transporters in metastasis. Biochim Biophys Acta 2015;1848:2638–46. https://doi.org/10.1016/j.bbamem.2014.11.012.
[78] Turner KL, Sontheimer H. Cl- and K+ channels and their role in primary brain tumour biology. 20130095 Philos Trans R Soc Lond B Biol Sci 2014;369. https://doi.org/10.1098/rstb.2013.0095.
[79] Farquhar KS, Charlebois DA, Szenk M, Cohen J, Nevozhay D, Balazsi G. Role of network-mediated stochasticity in mammalian drug resistance. Nat Commun 2019; 10:2766. https://doi.org/10.1038/s41467-019-10330-w.
[80] Wan Y, Mu Q, Krzyszton R, Cohen J, Coraci D, et al. Adaptive DNA amplification of synthetic gene circuit opens a way to overcome cancer chemoresistance. e2303114120 Proc Natl Acad Sci USA 2023;120. https://doi.org/10.1073/pnas.2303114120.