"In the light of evolution:" keratins as exceptional tumor biomarkers

Işıl Takan[1,2], Gökhan Karakülah[1,2], Aikaterini Louka[3,4] and Athanasia Pavlopoulou[1,2]

[1] Izmir Biomedicine and Genome Center, Izmir, Turkey
[2] Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir, Turkey
[3] DNA Damage Laboratory, Department of Physics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, Greece
[4] Section of Cell Biology and Biophysics, Department of Biology, School of Sciences, National and Kapodistrian University of Athens, Athens, Greece

Corresponding authors
Işıl Takan, isil.takan@ibg.edu.tr
Athanasia Pavlopoulou,
athanasia.pavlopoulou@ibg.edu.tr

ABSTRACT

Keratins (KRTs) are the intermediate filament-forming proteins of epithelial cells, classified, according to their physicochemical properties, into "soft" and "hard" keratins. They have a key role in several aspects of cancer pathophysiology, including cancer cell invasion and metastasis, and several members of the KRT family serve as diagnostic or prognostic markers. The human genome contains both, functional *KRT* genes and non-functional *KRT* pseudogenes, arranged in two uninterrupted clusters on chromosomes 12 and 17. This characteristic renders KRTs ideal for evolutionary studies. Herein, comprehensive phylogenetic analyses of KRT homologous proteins in the genomes of major taxonomic divisions were performed, so as to fill a gap in knowledge regarding the functional implications of keratins in cancer biology among tumor-bearing species. The differential expression profiles of *KRTs* in diverse types of cancers were investigated by analyzing high-throughput data, as well. Several *KRT* genes, including the phylogenetically conserved ones, were found to be deregulated across several cancer types and to participate in a common protein-protein interaction network. This indicates that, at least in cancer-bearing species, these genes might have been under similar evolutionary pressure, perhaps to support the same important function(s). In addition, semantic relations between KRTs and cancer were detected through extensive text mining. Therefore, by applying an integrative *in silico* pipeline, the evolutionary history of KRTs was reconstructed in the context of cancer, and the potential of using non-mammalian species as model organisms in functional studies on human cancer-associated *KRT* genes was uncovered.

INTRODUCTION

Cancer is a leading cause of death, currently accounting for one in six deaths worldwide (World Health Organization, https://www.who.int/news-room/fact-sheets/detail/cancer). It is a group of diseases characterized by abnormal cell growth, potentially invading neighboring tissues and/or spreading to other part(s) of the body (World Health

Organization, https://www.who.int/health-topics/cancer). It is denoted by diversities, complexities and, often, unpredicted dynamics in disease progression or metastasis (*Williams, 2015*; *Williams, Zaidi & Sengupta, 2022*). On the other hand, recent discoveries of earliest hominin malignancies dating millions of years make paleontology, epidemiology, history, systems, and phylogenetic analyses feasible for elucidating underexplored underlying mechanisms and cancer patterns (*Williams, 2015*; *Faltas, 2011*; *Wahba, Herrerín & Sánchez, 2021*; *Dittmar et al., 2020*).

Current anticancer modalities include radiation, chemotherapy, and surgery. Nonetheless, DNA damage and genome instability caused by radiation and chemotherapy have also several adverse effects and in many cases cause toxicity. A percentage of patients may initially respond to cancer, but eventually develop resistance to therapy and disease recurrence, while current treatments may frequently result in a 'tsunami' that kills both cancer and healthy cells non-selectively, often leading to side-effects (*Nikolaou et al., 2018*; *Toy et al., 2021*; *Pavlopoulou et al., 2017*; *Emran et al., 2022*). A deeper understanding of the molecular determinants and mechanisms that govern carcinogenesis would likely enable the identification of reliable diagnostic and prognostic biomarkers, improvement of clinical decision-making as well as the improvement of targeted therapies.

The emerging field of evolutionary medicine or "Darwinian medicine" allows the investigation of human diseases from an evolutionary perspective (*Nesse, 2001*; *Kranke, 2022*; *Logotheti et al., 2022*). Thus, deciphering the genetic and molecular factors/ mechanisms that contribute to shaping disease evolution, by examining their conservation across diverse species, would probably increase our knowledge regarding the etiology, development, progression and treatment of chronic diseases like cancer (*Lineweaver, Davies & Vincent, 2014*; *Marquardt et al., 2021*; *Trigos et al., 2017*). Multigene families like kallikreins (*Pavlopoulou et al., 2010*), cathelicidins (*Kosciuczuk et al., 2012*), *MAPK* (mitogen-activated protein kinases) (*Li, Liu & Zhang, 2011*), carcinoembryonic antigens (*Pavlopoulou & Scorilas, 2014*), *OXPHOS* (oxidative phosphorylation) (*De Grassi, Lanave & Saccone, 2008*) and *T2R* (bitter taste receptor) (*Dong, Jones & Zhang, 2009*), which have undergone a series of gene duplications, are implicated in critical pathophysiological processes. In this regard, herein, we aimed to investigate the role of the extended *keratin* gene family in cancer from an evolutionary point of view.

Keratins (KRTs) are intermediate-filament-forming proteins present in epithelial cells, which are broadly classified into "hard keratins" and "soft keratins", on the basis of their physicochemical properties and their sulfur content (*Bragulla & Homberger, 2009*; *Schweizer et al., 2006*; *Zhang & Fan, 2021*). Hard keratins make up morphological structures in birds (scales, claws and feathers) and mammals (hair and nails). Soft epithelial keratins are highly involved in epithelial cell protection from mechanical and non-mechanical stressors, and regulate a number of cellular processes, such as apical–basal plasma membrane polarity, cell size, cell motility, protein synthesis, membrane trafficking, wound healing, cell growth and cell death (*Moll, Divo & Langbein, 2008*; *Coulombe & Omary, 2002*; *Sarma, 2022*). The human genome encodes both functional keratin genes and non-functional keratin pseudogenes (KRT8/18/19P), which are located in two clusters

on chromosomes 12 (type II keratins except KRT18) and 17 (type I keratins) (*Bowden, 2005*; *Hesse et al., 2004*; *Jacob et al., 2018*).

There is accumulating evidence that keratins are involved in critical aspects of cancer, including invasion and metastasis (*Elazezy et al., 2021*; *Yu et al., 2022*; *Sharma et al., 2019*). In particular, changes in the activity of *E-cadherin*, which plays a critical role in epithelial-mesenchymal transition (EMT), increased the expression of *vimentin*, a well-studied marker in EMT, as well as keratin 17 in skin squamous cell carcinoma (*Lan et al., 2014*). It has also been shown that the interaction between vimentin and keratin 14 is essential for epidermal cell migration in epithelial cells (*Velez-delValle et al., 2016*). Keratin 16 has recently been linked to cancer metastasis and EMT as well (*Elazezy et al., 2021*). Therefore, keratins are used as commercially available markers to diagnose cancer: TPS (KRT18), TPACYK (KRT8/18) and CYFRA 21-1 (KRT19) (*Bodenmuller et al., 1994*; *Lane & Alexander, 1990*; *van Dalen, 1996*). Keratins are also considered as prognostic indicators in several epithelial cancers and they are implicated in treatment responsiveness (*Vaidya, Dmello & Mogre, 2022*; *Welch et al., 2019*; *Werner, Keller & Pantel, 2020*).

In this study, *in silico* systems biology approaches were employed, including comparative genomics, phylogenetics, high-throughput data processing, network-based methods, and natural language processing, in order to enhance our understanding regarding the functional implications of keratins in cancer biology among tumor-bearing species.

MATERIALS AND METHODS

Database searching

The approved symbols and names of *Homo sapiens* keratins were collected from the HUGO Gene Nomenclature Committee (HGNC) database (https://www.genenames.org/; (*Tweedie et al., 2021*)) (Tables S1 and S2).

Retrieval of KRT protein sequences

The human *KRT* gene symbols were used to retrieve the corresponding KRT protein sequences from the sequence database NCBI's RefSeq release 210 (*O'Leary et al., 2016*). Given that the protein sequences are more evolutionarily conserved compared to their corresponding nucleotide sequences, the proteins encoded by the *KRT* genes were used in the phylogenetic analyses.

Cross-genome search for KRT orthologs

The prototypic human type I and type II KRT protein sequences were used to search the well-annotated genomes of the species *Mus musculus* (Mouse), *Bos taurus* (Cow), *Gallus gallus* (Chicken), *Xenopus tropicalis* (Frog) and *Danio rerio* (Zebrafish) for corresponding orthologs in the databases ENSEMBL release 105 (*Cunningham et al., 2022*), NCBI's RefSeq release 210 (*O'Leary et al., 2016*) and UniProtKB release 2021_04 (*The UniProt Consortium, 2019*) in an iterative manner until no novel putative sequences could be detected, by employing reciprocal BLASTP and TBLASTN (*Altschul et al., 1990*).

The Translate program (http://web.expasy.org/translate/) was utilized to translate any nucleotide sequences into amino acid sequences.

Regarding the nomenclature of the KRT sequences in mouse, cow, chicken, frog and zebrafish, they were named based on their homology to their closest related well-annotated human KRT gene. Those KRT homologs with no significant sequence similarity to the fellow human KRTs (*i.e.*, forming separate branches), they were arbitrarily referred to as "orphan" KRTs.

Phylogenetic inference

The corresponding KRT protein sequences were extracted from the relevant databases. Subsequently, alignment of the full-length KRT amino acid sequences was performed using MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/; *Edgar, 2004*) the multiple sequence alignment *clitool* library. Phylogenetic inference based on the multiply aligned KRTs was conducted through phylogenetic tree construction, by employing neighbor-joining (NJ) and maximum-likelihood (ML) methods. The robustness of the inferred phylogenetic trees was evaluated by bootstrapping (100 bootstrap pseudo-replicates).

NJ trees were constructed using the *Bio.Phylo.TreeConstruction* module in Python. The module *Phylo* from Bio (*Cock et al., 2009*) was used to obtain the Newick and PhyloXML trees. The PhyloXML format was selected for storing the style of the phylogenetic trees. NJ phylogenetic trees were visualized using Matplotlib (*Hunter, 2007*), whereas ML trees were visualized with iTOL (*Letunic & Bork, 2021*). The ML trees were generated using the software package MEGA version 10.1 (*Kumar et al., 2018*).

Differential gene expression patterns

RNA sequencing (RNA-Seq) gene expression data for tumor and corresponding normal tissue samples from the TCGA and GTEx databases, respectively, were downloaded from the GEPIA2 (Gene Expression Profiling Interactive Analysis) online web server (http://gepia2.cancer-pku.cn/; *Li et al., 2021*).

The differentially expressed *KRT* genes between tumor and normal samples were identified using one-way analysis of variance (ANOVA), by setting the cut-off value for absolute log2 fold change |log2FC| ≥ 2 and FDR-adjusted *p*-value ≤ 0.05. The *KRT* genes and their corresponding values were stored as a server-less and self-contained database using Python SQLite. In order to visualize the differentially expressed genes (DEGs) between tumor and corresponding normal tissue samples, an R script was implemented. The interactions between the different types of cancers and the DEGs were used to generate a network which was manipulated and visualized by Cytoscape (http://www.cytoscape.org/; *Shannon et al., 2003*).

Protein-protein interactions (PPI)

A functional network of the interactions among the protein products of the identified cancer-relevant differentially expressed *KRT* genes was generated by utilizing STRING (Search Tool for the Retrieval of Interacting Genes) version 11.0b (*Szklarczyk et al., 2021*);

a database of experimental and predicted, direct (physical) or indirect (functional), associations among genes/proteins, derived from different sources such as high-throughput experiments, biomedical text mining, co-expression or gene fusion. The gene/protein associations detected in STRING were provided as input to Cytoscape (*Shannon et al., 2003*) for network construction and visualization.

Natural language processing

Semantic relationships between keratins and cancer for each member of the type I *KRT* gene family were discovered through biomedical literature mining, with the application of Natural Language Processing (NLP) methods (*Lauriola, Lavelli & Aiolli, 2022*). The scientific literature database MEDLINE/PubMed (https://pubmed.ncbi.nlm.nih.gov/) was searched thoroughly with artificial intelligence techniques using "keratin" AND "cancer" keywords to obtain relevant articles. A Python script was implemented to convert the raw text to proper data structure for further processing. Of the 4,950 candidate articles, 1,377 met the inclusion criteria, namely: (a) written in English, (b) including an abstract, (c) containing adequate text for processing. Finally, the Python regular expression library was utilized to distribute each type of keratin into different groups. Only those keratins with an adequate quantity of relevant articles were considered. The freely accessible Python libraries, natural language toolkit (NLTK: https://www.nltk.org/) and spaCy (https://spacy.io/), were used for text processing including tokenization, parsing, lemmatizing and stemming. The Python scispaCy (https://allenai.github.io/scispacy/) package, containing the spaCy "en_core_sci_lg" model, was used for processing biomedical scientific text; scispaCy's parsing tools were utilized to retrieve phrases related with the entities in cancer and keratins.

Word embeddings

Word2Vec is a NLP system that uses neural networks in order to create a distributed word representations in a *corpus* (*Sivakumar et al., 2020*). Word2Vec embeddings module was implemented in the Python library Gensim (https://pypi.org/project/gensim/) to train word vectors of the pre-processed text. A list of all word-to-word distances was extracted. To compute the similarity distances between each pair of terms, the *Word2Vec. most_similar* function in the gensim Word2Vec model was used. A continuous bag-of-words (CBOW) algorithm was used, which forms a new text vector representation for predicting other words in the sentence (*Liu, 2018*). Only those words that appeared more than five times (*i.e.*, minimum frequency threshold) were vectorized; iteration was set at 30 (epochs). Increased number of iterations enhanced the performance of Word2Vec, since the algorithm re-learned the relation between words.

Data visualization

A list of all term (word)-to-term (word) distances was retrieved. To compute the similarity distances between associated *KRT* genes and related words, the *Word2Vec.most_similar* function in the Gensim Word2Vec model was applied. The highest ranking 50 detected entries were included. Entries with word frequency below 10 were excluded. The interactions between the *KRT* genes and their top 50 similar words (frequency

Figure 1 ML radial cladogram of KRT type I protein sequences. The sequences are represented by the species name and the KRT names.

threshold set to 10) were used to generate networks which were manipulated and visualized with Cytoscape (*Shannon et al., 2003*). Also, for this study, Python SQLite, a server-less and self-contained database was created to easily access, manage, and update the collection of the cancer datasets (https://www.sqlite.org/).

RESULTS

Phylogenetic analysis of keratins

In order to infer evolutionary relationships among keratins, phylogenetic trees of KRT protein sequences were generated using both the NJ and ML methods. Several major distinct clades were identified in all trees (Figs. 1–4).

Keratins type I

The two phylogenetic trees (Figs. 1 and 2), reconstructed based on KRT type I proteins, using both methods, are congruent with similar topologies. The zebrafish orphan KRTs

Figure 2 NJ phylogram of KRT type I proteins. Bootstrap values (>50) are shown at the nodes. The branch length at the bottom indicates the length of amino acid substitutions per position.

form their own distinct clade supported by high bootstrap values, a fact that implies that these are likely the members of the *KRT* family that diverged earliest ("proto-KRTs") before the emergence of Amphibia 330 million years ago. One plausible explanation is that
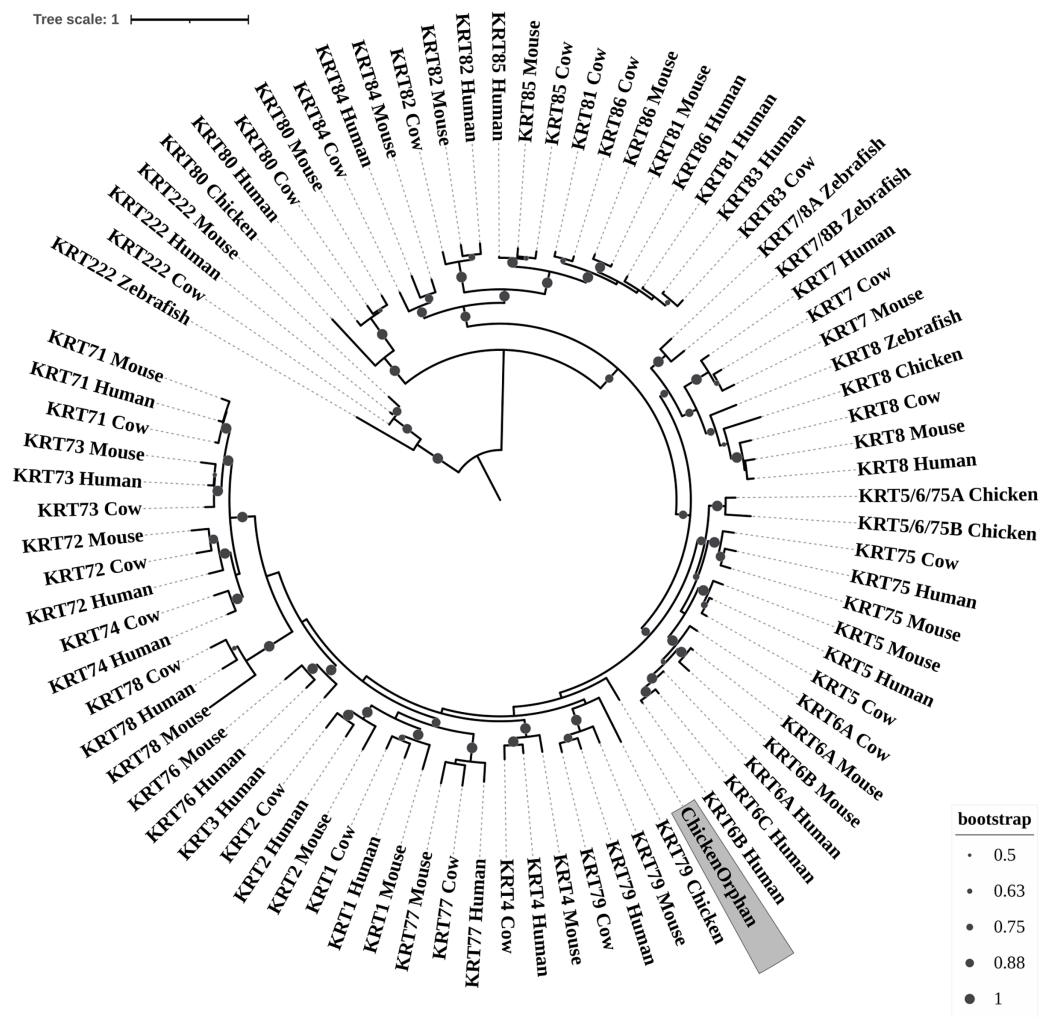
Figure 3 ML radial cladogram of KRT type II protein sequences. The sequences are represented by the species name and the KRT protein names.

a *proto-KRT* gene emerged in an ancestor of Teleostei and a series of lineage-specific gene duplication events gave rise to *KRTorphan1* and *KRTorphan2* found in the contemporary zebrafish genomes.

The primordial gene of the *KRT* family appears to be *KRT18*, since it was detected in zebrafish and frog, leading to the suggestion that it first emerged in an ancestor of Euteleostomi. Interestingly, the branches of the KRT18 clade are exceptionally long, suggesting that the KRT18 members evolved independently and more rapidly compared to the other KRTs.

The largest clade of the reconstructed trees (Figs. 1 and 2) comprises KRT31, KRT32, KRT33, KRT34, KRT35, KRT36, KRT37, KRT38, KRT39 and KRT40 which were identified exclusively in Eutheria, *i.e.*, placental mammals. The subclade consisting of KRT37 and KRT38 is basal to the major clade in both trees (Figs. 1 and 2). This result suggests that their corresponding genes were probably the first to emerge through a recent duplication event that took place in primates, since the human KRT37 and KRT38
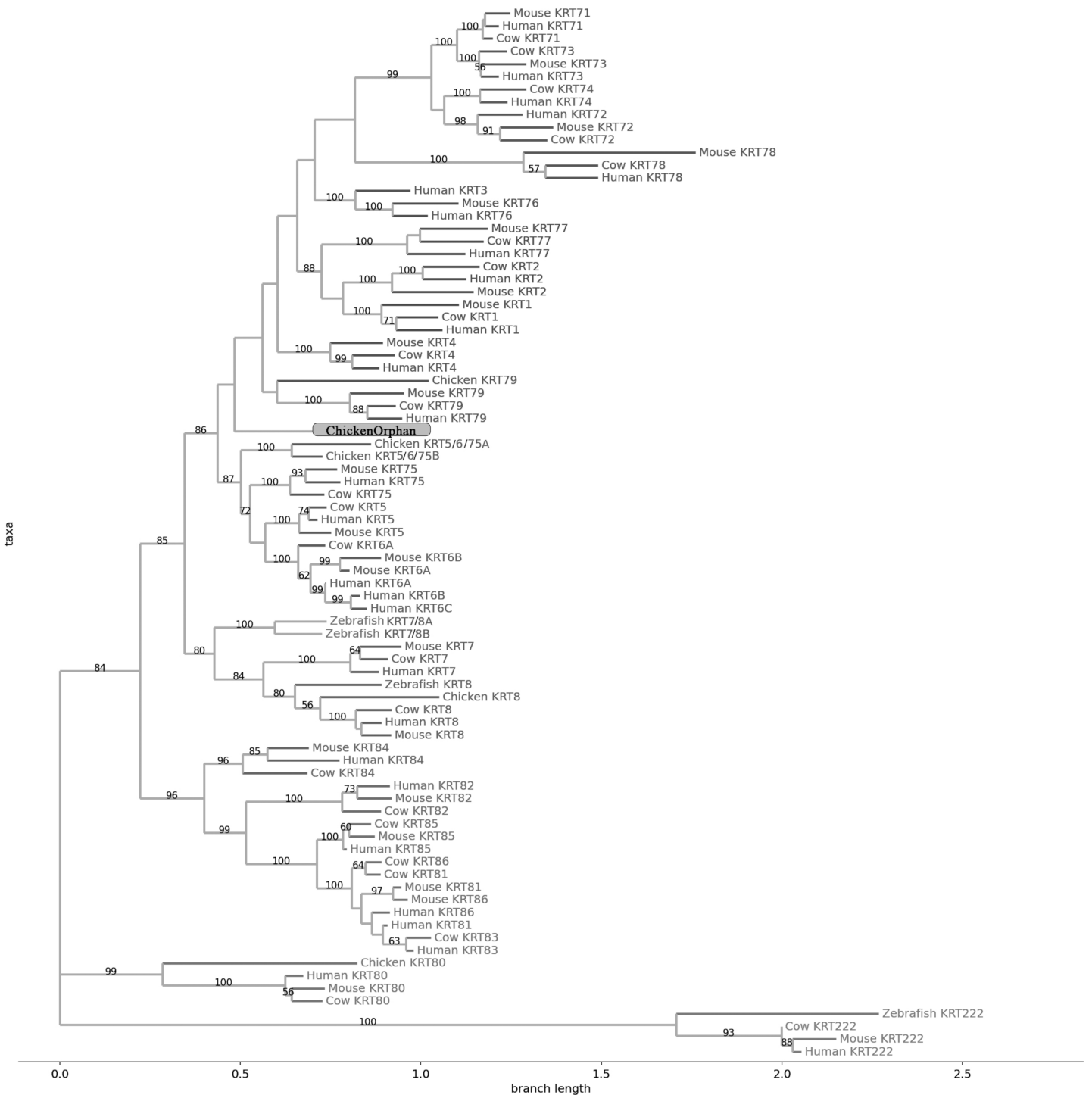
Figure 4 NJ phylogram of KRT type II proteins. Bootstrap values (>50) are shown at the nodes. The branch length at the bottom indicates the length of amino acid substitutions per position.

sequences appear to have strong similarity. Of note, KRT37 and KRT38 murine orthologs were not detected, a fact that implies that the corresponding genes either got lost during evolution or were highly corrupted in glires. The phylogenies indicate that a series of nine

mammalian-specific tandem duplication events took place before the Laurasiatheria–Euarchontoglires divergence, thereby yielding those ten paralogs.

Another set of sister clade is formed by KRT20 and KRT23, where KRT23 appears to be more ancient. KRT23 was first detected in chicken, and thus, it is of Amniota origin, while KRT20 was detected exclusively in Eutheria. *KRT20* appears to be the product of a *KRT23* duplication which occurred in a eutherian ancestor presumably after the Mammalia-Sauropsida split.

Additionally, an ancestral KRT14/16/17 was identified in chicken. A *KRT14/16/17* primordial gene of Amniota origin, through a series of mammalian-specific duplication events, presumably has given rise to the three separate *KRT14*, *KRT16* and *KRT17* orthologs found in the mammalian genomes. *KRT14* and *KRT17* appear to have emerged first during the course of mammalian evolution, while *KRT16* arose later, after the Laurasiatheria–Euarchontoglires split, since it was not detected in the Laurasiatheria.

A sister group to the KRT14/16/17 group consists of the KRT19 homologs. The two groups have strong similarity, as confirmed by relatively high bootstrap values (Figs. 1 and 2). KRT19 was also first identified in chicken, and thus, most likely, it emerged in an ancestor of Amniota. Those findings trigger the speculation that *KRT19* and *KRT14/16/17* occurred due to a duplication event which took place in an ancestor of Amniota before the divergence of Mammalia and Sauropsida.

KRT13 and KRT15 together constitute a sister group to the KRT14/16/17–KRT19 group. Both KRT13 and KRT15 are detected exclusively in Eutheria, hence they might have emerged after a series of species-specific duplication events that yielded those mammalian paralogs in an ancestor of Eutheria after the Laurasiatheria-Euarchontoglires split.

Another KRT member that was first detected in Amniota is KRT9, which most likely appeared earlier in amniotic evolution, relatively to KRT14/16/17 and KRT19, since it does not share a high degree of sequence similarity with its other two paralogs. Moreover, as suggested by the topology and the high bootstrap values, KRT9 appears to have strong similarity with KRT10 (Figs. 1 and 2), the latter of which is found only in Eutheria, triggering the speculation that *KRT10* emerged through a mammalian-specific duplication event of *KRT9* that occured most likely after the divergence of Mammalia and Sauropsida.

Finally, KRT12, KRT24, KRT25, KRT26, KRT27 and KRT28 form a separate, highly-supported clade, restricted to Mammalia (Figs. 1 and 2). *KRT12* is most likely the first of this group to arise during the course of mammalian evolution, while *KRT24*, *KRT28*, *KRT25*, *KRT27* and *KRT26* appeared later in the *KRT* family. The phylogenies suggest that a series of four mammalian-specific tandem gene duplication events might have given rise to the corresponding orthologs, apparently after the Laurasiatheria–Euarchontoglires divergence.

Keratins type II

The two phylogenetic trees (Figs. 3 and 4), reconstructed based on KRT Type II proteins, using both methods, are congruent with similar topologies. On the basis of the reconstructed phylogenies, the primordial gene of the *KRT* family appears to be *KRT222*

("proto-KRTs") since it was detected in zebrafish, leading to the suggestion that it first emerged in an ancestor of Teleostei. The branches of KRT222 form their own distinct clade which arises from the basal node with high bootstrap values, indicating that the corresponding gene diverged earliest. Additionally, the KRT222 branch is exceptionally long, suggesting that the KRT222 members evolved independently and more rapidly compared to the other KRTs, and thus, they are distantly related to the rest of the members of the *KRT* family. Therefore, it is intriguing to speculate that the *KRT222* orthologs in the contemporary genomes are probably the products of a series of duplications of an ancestral *KRT222* gene in Teleostei.

Moreover, a distinct clade comprised of the *KRT80* gene products is distantly related to the other KRTs in the trees. KRT80 was first detected in chicken, and thus, it is of Amniota origin, suggesting that an ancestral *KRT80* gene in Amniota gave rise to the *KRT80* orthologs in placental mammals, which evolved more rapidly compared to the fellow Type II keratins.

Another highly-supported distinct large clade of the trees comprises of the *KRT81*, *KRT82*, *KRT83*, *KRT84*, *KRT85* and *KRT86* gene products which are restricted to Eutheria. The subclade consisting of KRT84 is basal to the major clade in both trees (Figs. 3 and 4), suggesting that the *KRT84* gene probably arose first through a recent duplication event that took place in a eutherian ancestor of the contemporary placental mammals. The next gene to emerge during the course of mammalian evolution, according to the phylogenies, appears to be *KRT82* followed by *KRT85*. The genes *KRT86*, *KRT81* and *KRT83* seem to have emerged later since their products' sequences appear to have strong similarity. Conclusively, the reconstructed phylogenies lead to the speculation that a series of five eutherian-specific tandem gene duplication events might have given rise to those six paralogs.

Also, a set of sister clades is formed by the *KRT7* and *KRT8* gene products. KRT8 appears to be the older of the two, since it was first detected in zebrafish, and thus, it is of Teleostei origin, while KRT7 was identified exclusively in Eutheria. *KRT7* gene appears to be the product of a *KRT8* gene duplication which occurred in a eutherian ancestor presumably after the divergence of Mammalia and Sauropsida, since KRT7 was not detected in chicken.

A clade basal to the KRT7/8 group contains the zebrafish KRT7/8A and KRT7/8B; these two sequences exhibit a strong similarity and they form a discernible highly-supported branch in both trees, suggesting products of a teleost-specific gene duplication event. One plausible explanation is that a *proto-KRT7/8* gene emerged in an ancestor of Teleostei and a series of lineage-specific gene duplication events gave rise to the two paralogous genes found in the contemporary zebrafish genomes. However, those two sequences might have evolved rapidly and separately, since they do not have great similarity with their homologs detected in other species.

In addition, KRT5, KRT6 and KRT75 constitute a distinct group in both trees. KRT6B was identified in mouse for the first time, leading to the suggestion that a gene duplication event, which took place in an early eutherian mammal after the Laurasiatheria – Euarchontoglires divergence, yielded those paralogs. Finally, KRT6C was detected

exclusively in human, which triggers the speculation that it emerged through a recent duplication event that took place in primates, since the human KRT6A, KRT6B and KRT6C sequences appear to have a high degree of similarity, as confirmed by high bootstrap values (Figs. 3 and 4). KRT5 members are more similar to KRT6s, suggesting products of the duplication of a eutherian *KRT5/6* gene. Moreover, the KRT75 clade, which comprises a sister group to KRT5/6, appears to have emerged later in the evolution and includes only eutherian KRT members. Two chicken KRT5/6/75 paralogs, which sort into a distinct, well-supported clade, suggest that an ancestral amniotic *KRT5/6/75* gene has given rise to the KRT5, KRT6 and KRT75 members.

An orphan KRT sequence was detected in *Gallus gallus* (Chicken KRTorphan), which bears strong similarity to the chicken paralogs KRT6/75, based on the short branch lengths connecting them, suggestive of short evolutionary distance. Moreover, KRT4 sequences form a discernible clade which is a sister group to KRT79, the latter of which was first detected in chicken, and thus, it is of Amniota origin.

Another distinct clade is comprised of the *KRT1*, *KRT2* and *KRT77* gene products which are restricted to Eutheria. The subclade consisting of KRT77 is basal to the major clade in both trees (Figs. 3 and 4), suggesting that the *KRT77* gene was probably the first to emerge through a duplication event that occurred in a Eutherian ancestor of the contemporary placental mammals. Another lineage-specific gene duplication seems to have yielded the *KRT1* and *KRT2* paralogs.

One distinct clade is formed by the *KRT76* and *KRT3* gene products, which have strong similarity, as confirmed by relatively high bootstrap values (Figs. 3 and 4). Notably, KRT76 was first identified in mouse, suggesting that the *KRT76* gene first arose in an early eutherian mammal, most likely after the divergence of Laurasiatheria and Euarchontoglires. KRT3, on the other hand, was detected exclusively in human, suggesting that it emerged through a recent duplication event that occurred in primates.

The final highly-supported distinct clade of both trees contains the *KRT78*, *KRT71*, *KRT72*, *KRT73* and *KRT74* gene products, which were identified exclusively in Eutheria. The subclade consisting of KRT78 sequences is basal to the major clade in both trees (Figs. 3 and 4), suggesting that the *KRT78* gene was probably the first to emerge in the course of mammalian evolution through duplications that took place in a eutherian ancestor of the contemporary placental mammals. Two sister groups are comprised of the KRT72 and KRT74, and KRT73 and KRT71 sequences, respectively. The two groups appear to share a high degree of similarity, as confirmed by relatively high support values, thus triggering the speculation that a series of three mammalian-specific gene duplication events yielded the *KRT71*, *KRT72*, *KRT73* and *KRT74* paralogs.

Differential *KRT* gene expression across cancers

To assess the role of keratins in cancer, the differential *KRT* gene expression patterns were investigated across cancers. Several *KRT* genes, including the phylogenetically older *KRT8*, *KRT18*, *KRT19*, *KRT23*, *KRT79*, *KRT80* and *KRT222*, were found to be differentially expressed in diverse types of cancers (Table 1 and Fig. 5). Of note, *KRT222*, which was first detected in *Gallus gallus*, was down-regulated specifically in brain neoplasms (lower grade

Table 1 *KRT* genes, cancer types and gene expression status.

GeneName	TCGA cancer type	Status
KRT1	DLBC, ESCA, SKCM	Down
KRT10	SKCM, LAML	Down
	THYM	Up
KRT13	LUSC	Up
	UCEC, UCS, LUAD, ESCA, HNSC, COAD, ACC, SKCM, STAD, TGCT, READ, PRAD, OV	Down
KRT14	SKCM, BRCA	Down
	BLCA, LUSC, THYM	Up
KRT15	THYM, LUSC, CESC, PAAD	Up
	BRCA, SKCM, TGCT, ESCA, PRAD	Down
KRT16	SKCM	Down
	PAAD, CESC, LUSC	Up
KRT17	LUSC, HNSC, ESCA, COAD, UCEC, BLCA, CESC, STAD, THYM, UCS, PAAD, OV, READ	Up
	SKCM, TGCT, BRCA	Down
KRT18	UCS, THYM, UCEC, READ, OV, TGCT, STAD, CESC, BRCA, COAD, ESCA	Up
KRT19	SKCM, LAML, KICH	Down
	CESC, LUSC, COAD, THCA, TGCT, OV, PAAD, READ, THYM, UCEC	Up
KRT2	SKCM	Down
KRT20	READ, COAD	Up
KRT222	LGG, GBM	Down
KRT23	HNSC, DLBC, SKCM, THYM, TGCT, PRAD	Down
	COAD, READ, PAAD, OV, UCEC	Up
KRT24	ESCA	Down
KRT31	ESCA, SKCM	Down
KRT32	ESCA	Down
KRT33A	TGCT	Down
KRT4	ESCA, HNSC, LUAD	Down
KRT5	LUSC, THYM, CESC	Up
	BRCA, SKCM, ESCA	Down
KRT6A	SKCM	Down
	THYM, PAAD, CESC, LUSC	Up
KRT6B	LUSC, CESC, PAAD	Up
	SKCM	Down
KRT6C	LUSC	Up
	ESCA	Down
KRT7	KIRC, LUSC, SKCM	Down
	BLCA, STAD, THYM, PAAD, OV, CESC, UCS, UCEC	Up
KRT72	SKCM, TGCT	Down
KRT73	SKCM	Down
KRT75	HNSC	Up
KRT77	SKCM	Down
KRT78	SKCM, HNSC	Down

(Continued)

GeneName	TCGA cancer type	Status
KRT78	ESCA	Down
KRT79	SKCM	Down
KRT8	TGCT, STAD, THYM, READ, OV, PAAD, UCS BRCA, CESC, COAD, ESCA, UCEC	Up
	LAML	Down
KRT80	COAD, CESC, LUAD, OV, READ, STAD, THCA	Up
	SKCM	Down
KRT86	TGCT	Down

Note:
ACC, Adrenocortical carcinoma; BLCA, Bladder Urothelial Carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; COAD, Colon adenocarcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney renal clear cell carcinoma; LAML, Acute Myeloid Leukemia; LGG, Brain Lower Grade Glioma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; READ, Rectum adenocarcinoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular Germ Cell Tumors; THCA, Thyroid carcinoma; THYM, Thymoma; UCEC, Uterine *Corpus* Endometrial Carcinoma; UCS, Uterine Carcinosarcoma.
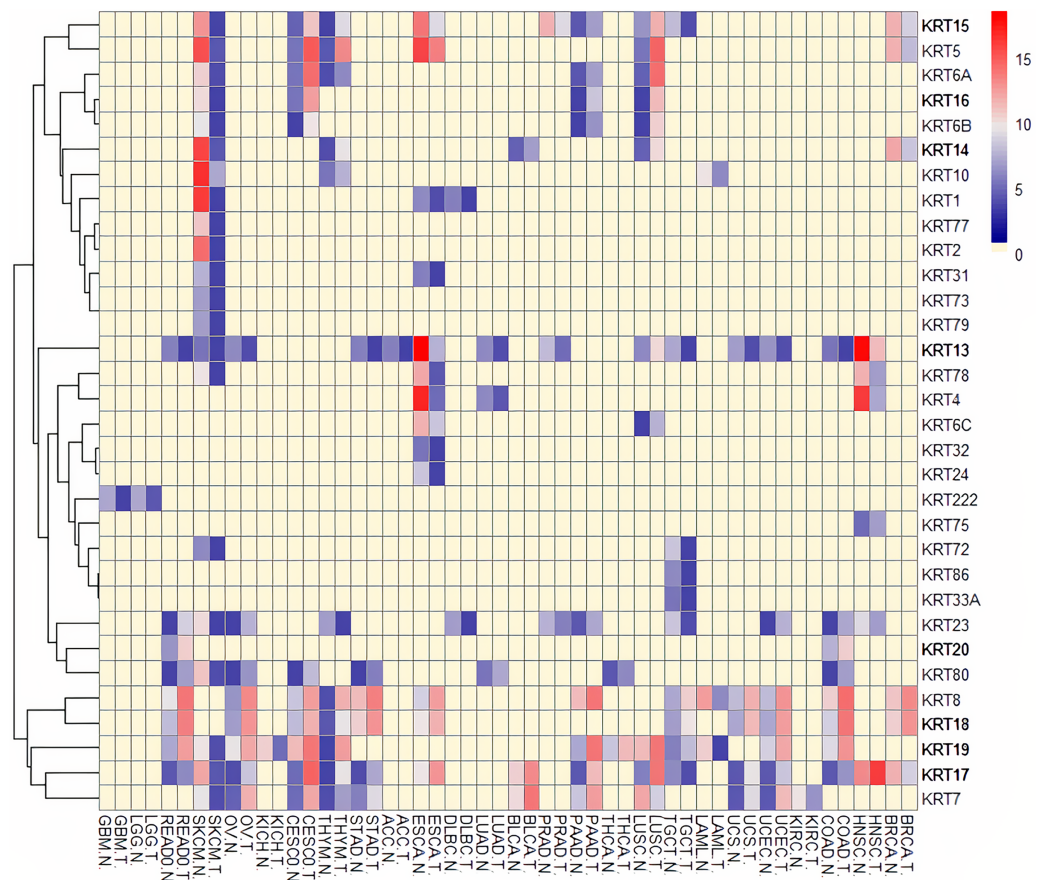
Figure 5 Heatmap of differentially expressed *KRT* genes based on data derived from the cancer SQLite database. Red: up-regulated, Blue: down-regulated, Yellow: no data. T, tumor; N, matched normal.
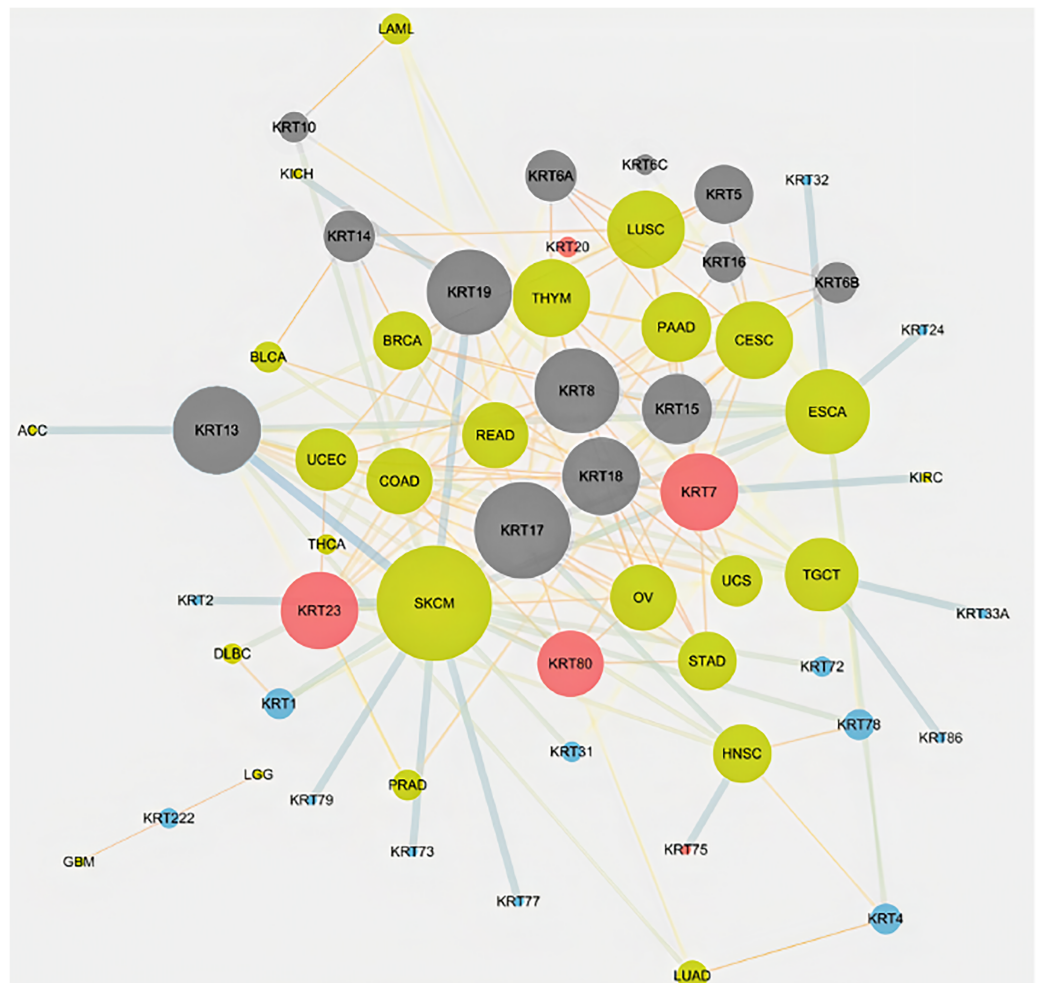
Figure 6 Keratins are linked to multiple cancers. The node size is proportional to the number of direct links. Red and blue color denotes consistent up-regulation and down-regulation across cancers, respectively; gray color indicates, both, up- and down-regulation in the different types of cancers in Table 1.

glioma and glioblastoma multiforme). Moreover, the phylogenetically conserved *KRT5*, *KRT7*, *KRT13*, *KRT14*, *KRT15* and *KRT17*, were differentially expressed in multiple types of cancers (Fig. 5, bold). To visualize the relationships between KRTs and cancers, a bipartite network was constructed displaying TCGA-derived cancer-KRT associations by using information from the cancer SQLite database developed in this study. The size of the nodes is proportional to their degree of connectivity. The network is highly interconnected, suggesting associations among diverse cancer types and both types of *KRT* genes (Fig. 6). Of note, the protein products of those *KRTs* differentially expressed in multiple cancers, including the older ones, form a highly interconnected network, suggesting functional or physical associatiosn among them (Fig. 7).

The aforementioned findings further validate the interrelatedness among keratins and their important implication in multiple cancer pathways.
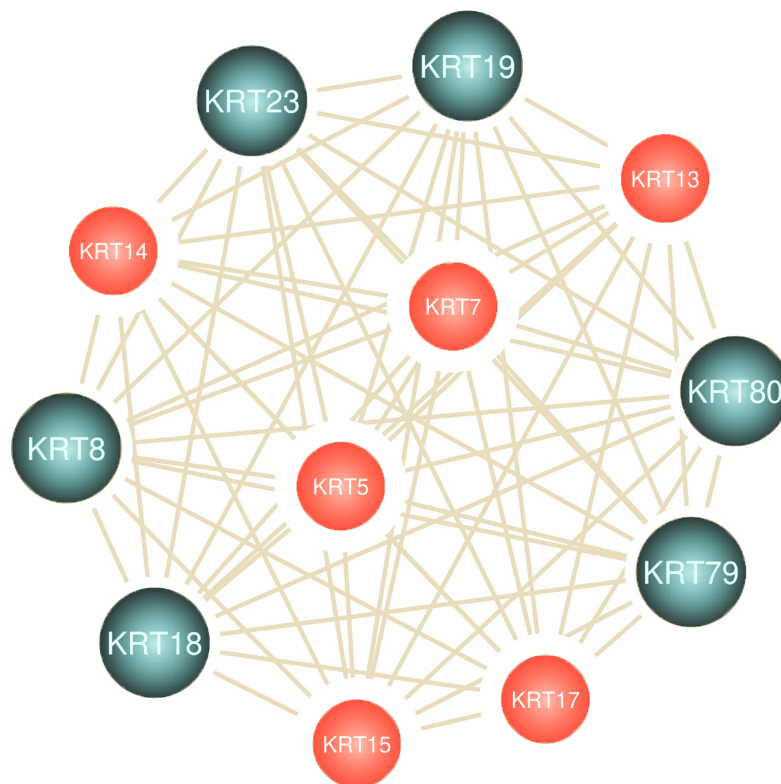
Figure 7 Keratins are linked to multiple cancers. Node size represents the number of direct links.

KRT and cancer semantic relations

NLP-aided literature mining was performed in order to identify important relationships between type I keratins and cancers. This method allowed the systematic, extensive and comprehensive scrutinization of a vast number of scientific articles for extracting relevant information. Networks illustrating the semantic relations of the highest ranking 50 most similar terms (words) to KRT13, KRT14, KRT15, KRT16, KRT17, KRT18, KRT19 and KRT20 were created (Fig. 8). The nodes represent the words, and the edges denote the semantic associations to selected KRTs with adequate texts in their document.

The semantic relations of the top ranking 50 terms (words) were constructed using the similarity distances between each word; a KRT association network was generated which was manipulated and visualized by Cytoscape (http://www.cytoscape.org/; (*Shannon et al., 2003*)). The word nodes on the network were placed according to their similarity with each KRT.

DISCUSSION

The reconstructed phylogenies of keratins allowed us to get a glimpse of the evolutionary history of this extended gene family, and gain an understanding of how members of this family are associated with certain pathophysiological processes. According to the results of the phylogenetic analysis, duplication and mutation events during the course of evolution

Figure 8 Network depicting the semantic relationships of the top 50 most similar terms to KRTs.

gave rise to the *KRT* gene family, essential for developing novel genes; some *KRT* genes were probably converted to pseudogenes and *vice versa*.

Several members of the *KRT* gene family, including the evolutionarily older *KRT8*, *KRT18*, *KRT19*, *KRT23*, *KRT79*, *KRT80* and *KRT222*, were shown to be differentially regulated in diverse cancer types (Table 1). The expression status of the *KRT* genes under study varies in different types of cancers, *i.e.*, some are up-regulated, and, conversely, some are down-regulated in different cancer types. Hence, at least in cancer-bearing species, the evolutionarily conserved genes have likely resist evolutionary changes, so as to carry out
critical functions in carcinogenesis, consistent with the hypothesis postulated by *Makashov, Malov & Kozlov (2019)* that oncogenes and tumor suppressor genes represent the evolutionarily oldest classes of genes in eukaryotic species. Furthermore, the Amniota-specific *KRT8*, *KRT19*, *KRT23* and *KRT80* were shown to be up-regulated in ovarian cancer. Given that chicken is currently the only animal model available to investigate the etiology and progression of human ovarian cancer (*Hawkridge, 2014*), the respective orthologs could be taken into consideration in the study of ovarian cancer. Moreover, the *KRT18* ortholog first appeared in *Danio rerio*. Notably, *KRT18* was under-expressed in skin cutaneous melanoma relative to normal tissue (Table 1). Since zebrafish represents a unique experimental model organism for studying melanoma development, progression and treatment (*Bootorabi et al., 2017*), the teleost *krt18* gene could also be considered in the melanoma clinical research. Furthermore, an amphibian *KRT18* ortholog was detected in frog. According to *Hardwick & Philpott (2018)*, *Xenopus* model systems have diverse applications in cancer research and, particularly, in tumor immunity.

The constructed KRT-oriented networks, using word embedding, provide valuable information on how related words in a given text dataset have semantic and syntactic similarity with a given keratin. Our results support that there is a significant association between keratins and a number of cancer biomarkers (Fig. 8).

Keratin 18 represents a robust diagnostic and prognostic biomarker for human cancers (*Menz et al., 2021*). For instance, the expression patterns of conventional tumor markers, such as the proliferating cell nuclear antigen (PCNA) and the minichromosome maintenance protein 3 (MCM3) in breast cancer were found to be similar to those of *KRT18*. Besides, *KRT18* was significantly correlated with the loss of estrogen and progesterone receptors (*Ha et al., 2011*). In the present study, *KRT18* was over-expressed in breast invasive carcinoma (BRCA) (Table 1), further supporting that KRT18 could represent a candidate biomarker in breast cancer for predicting poor prognosis in breast cancer. Moreover, it has been suggested that caspase-cleaved KRT18, a serum apoptosis product, could be a functional biomarker for predicting the response of breast carcinomas to chemotherapy (*Olofsson et al., 2007*).

It has been demonstrated that the evolutionarily old *KRT19* is differentially expressed in several types of cancers. In the present study, *KRT19* was found overexpressed in breast, colon, lung, liver and thyroid cancer (Table 1), consistently with previous reports; *KRT19* was associated with poor clinical outcomes in cancer patients as well (*Kawai et al., 2015*; *Saha et al., 2019*; *Wang et al., 2019*; *Yuan et al., 2021*). KRT19 represents one of the factors determining tumor response to chemo/radiotherapy (*Bozionellou et al., 2004*; *Saha et al., 2018*).

Low expression of *KRT15* in breast invasive carcinoma (Table 1) was associated with unfavorable prognosis (*Zhong et al., 2021*). Moreover, KRT15 was demonstrated to promote migration and invasion of colorectal cancer cells partly *via* β-catenin-mediated signaling (*Chen & Miao, 2022*); β-catenin, which regulates cell-cell adhesions (*Brembeck, Rosario & Birchmeier, 2006*), was also found to be semantically related with KRT15 (Fig. 8). Furthermore, the phylogenetically (Figs. 1 and 2) and semantically related (Fig. 8)

KRT13 and KRT15 are also directly associated in oral cancers (*Khanom et al., 2012*). *KRT13* is transcriptionally up-regulated by KLF4 to induce differentiation of esophageal squamous cell carcinoma (*He et al., 2015*). Likewise, *KRT15* is over-expressed in esophageal carcinoma (Cancer Stat Facts: Leukemia—Acute Myeloid Leukemia (AML), https://seer.cancer.gov/statfacts/html/amyl.html). However, *KRT13* and *KRT15* were found to be down-regulated in TCGA-derived esophageal carcinoma samples (Table 1). It has also been shown that *KRT13* is transcriptionally suppressed during TGF-β1-induced EMT (*Hatta et al., 2018*), suggesting implication of KRT13 in a hallmark of cancer, *i.e.*, invasion and metastasis.

Cancer-associated fibroblasts were shown to exert a powerful stimulatory effect on the expression of KRT14, which is a basal/myoepithelial marker (*Dvorankova et al., 2012*). KRT14 enhances the metastatic potential of lung cancer cells, promotes cell invasion of salivary adenoid cystic carcinoma, and is also correlated with worse patient prognosis (*Gao et al., 2017*). In addition, keratin 14 is correlated with nodal metastasis and unfavorable prognosis in human lung adenocarcinoma *via* Gkn1 induction (*Yao et al., 2019*). Similarly, in our study, the corresponding *KRT14* transcripts were elevated in lung squamous cell carcinoma (Table 1). Based on the human protein atlas (HPA) (https://www.proteinatlas.org), KRT14 is a favorable prognostic biomarker for breast cancer, consistent with our findings wherein the *KRT14* gene expression is reduced in breast invasive carcinoma (Table 1). Moreover, transcriptional up-regulation of *KRT14*, and down-regulation of *KRT15* and *KRT19* was observed in oral squamous cell carcinoma (OSCC). Also, deregulated *KRT15* and *KRT19* expression was observed in well-differentiated OSCC as compared to moderately/poorly differentiated OSCC (*Khanom et al., 2012*).

Increased *KRT16* expression is highly associated with weak differentiation, augmentation of lymph node metastasis, worse survival outcome and advanced stages of OSCC. Also, inhibition of KRT16 resulted to reduced OSCC progression and chemoresistance, whereas *KRT16* silencing improved chemosensitivity (*Huang et al., 2019*). Moreover, there is a significant correlation between enhanced *KRT16* expression and poor overall survival in metastatic breast cancer patients (*Joosse et al., 2012*). Based on the HPA, keratin 16 is enhanced in cervical cancer and it is a poor prognostic biomarker for pancreatic cancer. Consistent with the latter, in our study, the corresponding *KRT16* gene was found to be significantly up-regulated in pancreatic adenocarcinoma as well as the cervical squamous cell carcinoma and endocervical adenocarcinoma (Table 1).

Keratin 17 has been underscored as an emerging diagnostic, prognostic, and predictive biomarker (*Yang, Zhang & Wang, 2019*), based on preclinical and clinical cancer studies. According to *Baraks et al. (2022)*, KRT17 is implicated in eight out of ten deadly hallmarks of cancer. KRT17 triggers the AKT-mediated signaling pathway and induces EMT, while it is strongly correlated with malignant transformation and worse prognosis in esophageal squamous cell carcinoma (ESCC) patients. Therefore, KRT17 may serve as a therapeutic target for the treatment of ESCC. Increased level of KRT17 is directly correlated with the progression of pancreatic cancer (*Chen et al., 2020*). In addition, keratin 17 is considered a novel cytologic biomarker for accurately distinguishing between recurrent urothelial carcinoma and benign urothelial cells (*Babu et al., 2021*). Moreover, according to HPA

results, KRT17 is over-expressed in cervical and head and neck cancers and represents a favorable prognostic marker for breast cancer. In agreement with the aforementioned findings, in our study, *KRT17* is over-expressed in the head and neck squamous cell carcinoma, esophageal carcinoma, bladder urothelial carcinoma, pancreatic adenocarcinoma as well as cervical squamous cell carcinoma and endocervical adenocarcinoma, whilst it is down-regulated in breast invasive carcinoma (Table 1).

Quantitating *KRT20* expression by RT-PCR is a very sensitive and accurate method to detect unknown lymph node metastases. In addition, by measuring *KRT20* mRNA expression in lymph nodes is essential for exact tumor staging and for postoperative adjuvant treatment of colorectal cancer patients (*Chen et al., 2004*). Studies on pancreatic carcinoma, gastrointestinal cancers, colorectal carcinoma and miscellaneous tumors suggest that it is convenient to detect circulating tumor cells based on KRT20 RT-PCR assays (*Joosse et al., 2012*; *Lukyanchuk et al., 2003*). According to *Eckstein et al. (2018)*, high *KRT5* and low *KRT20* expression defines distinct prognostic subgroups in urothelial bladder cancer. Moreover, according to HPA results, keratin 20 is enriched in colorectal cancer. Consistently, in our study, enhanced expression of *KRT20* was found in colon and rectum adenocarcinoma (Table 1).

The findings of the present study highlight the prominent role of keratins in mammalian cancers. Members of the keratin family could serve as robust diagnostic and prognostic cancer biomarkers as well as potential therapeutic targets. Notably, the evolutionary conservation of several *KRT* genes across taxa points to the importance of using non-mammalian model organisms in functional studies towards investigating the etiology, development, progression, and treatment of cancer.

CONCLUSIONS

In this study, cross-disciplinary systems biology methods were employed by combining phylogenetics, biological literature mining, differentially expressed gene profiles, and biological networks to investigate evolutionarily conserved *KRT* genes implicated in different aspects of cancer. The findings of this study could have potential application in the clinical setting, where the phylogenetically preserved genes can be exploited as diagnostic or prognostic tumor markers. Furthermore, the detection of evolutionarily conserved *KRTs* in non-mammalian species highlights the possible importance of using these organisms as model systems in cancer research.

ACKNOWLEDGEMENTS

We are grateful to Dr. Stella Logotheti, Dr. Stella Geronikolou and Prof. George P. Chroussos for their insightful comments.

Funding

The authors received no funding for this work.

Competing Interests

Gökhan Karakülah is an Academic Editor for PeerJ.

Author Contributions

- Işıl Takan conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Gökhan Karakülah performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Aikaterini Louka performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Athanasia Pavlopoulou conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The approved symbols and names of *Homo sapiens* keratins were collected from the HUGO Gene Nomenclature Committee (HGNC) database (https://www.genenames.org/). The symbols and accession numbers are available in the Supplemental Tables.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.15099#supplemental-information.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403–410 DOI 10.1016/S0022-2836(05)80360-2.

Babu S, Kim NW, Wu M, Chan I, Escobar-Hoyos LF, Shroyer KR. 2021. Keratin 17 is a novel cytologic biomarker for urothelial carcinoma diagnosis. *American Journal of Clinical Pathology* **156(5)**:926–933 DOI 10.1093/ajcp/aqab050.

Baraks G, Tseng R, Pan CH, Kasliwal S, Leiton CV, Shroyer KR, Escobar-Hoyos LF. 2022. Dissecting the oncogenic roles of keratin 17 in the hallmarks of cancer. *Cancer Research* **82(7)**:1159–1166 DOI 10.1158/0008-5472.CAN-21-2522.

Bodenmuller H, Donie F, Kaufmann M, Banauch D. 1994. The tumor markers TPA, TPS, TPACYK and CYFRA 21-1 react differently with the keratins 8, 18 and 19. *The International Journal of Biological Markers* **9(2)**:70–74 DOI 10.1177/172460089400900202.

Bootorabi F, Manouchehri H, Changizi R, Barker H, Palazzo E, Saltari A, Parikka M, Pincelli C, Aspatwar A. 2017. Zebrafish as a model organism for the development of drugs for skin cancer. *International Journal of Molecular Sciences* **18(7)**:1550 DOI 10.3390/ijms18071550.

Bowden PE. 2005. The human type II keratin gene cluster on chromosome 12q13.13: final count or hidden secrets? *The Journal of Investigative Dermatology* **124(3)**:xv–xvii DOI 10.1111/j.0022-202X.2005.23647.x.

Bozionellou V, Mavroudis D, Perraki M, Papadopoulos S, Apostolaki S, Stathopoulos E, Stathopoulou A, Lianidou E, Georgoulias V. 2004. Trastuzumab administration can effectively

target chemotherapy-resistant cytokeratin-19 messenger RNA-positive tumor cells in the peripheral blood and bone marrow of patients with breast cancer. *Clinical Cancer Research* **10(24)**:8185–8194 DOI 10.1158/1078-0432.CCR-03-0094.

Bragulla HH, Homberger DG. 2009. Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. *Journal of Anatomy* **214(4)**:516–559 DOI 10.1111/j.1469-7580.2009.01066.x.

Brembeck FH, Rosario M, Birchmeier W. 2006. Balancing cell adhesion and Wnt signaling, the key role of beta-catenin. *Current Opinion in Genetics & Development* **16(1)**:51–59 DOI 10.1016/j.gde.2005.12.007.

Chen G, McIver CM, Texler M, Lloyd JM, Rieger N, Hewett PJ, Sen Wan D, Hardingham JE. 2004. Detection of occult metastasis in lymph nodes from colorectal cancer patients: a multiple-marker reverse transcriptase-polymerase chain reaction study. *Diseases of the Colon and Rectum* **47(5)**:679–686 DOI 10.1007/s10350-003-0118-2.

Chen W, Miao C. 2022. KRT15 promotes colorectal cancer cell migration and invasion through beta-catenin/MMP-7 signaling pathway. *Medical Oncology* **39(6)**:68 DOI 10.1007/s12032-021-01619-2.

Chen P, Shen Z, Fang X, Wang G, Wang X, Wang J, Xi S. 2020. Silencing of keratin 17 by lentivirus-mediated short hairpin RNA inhibits the proliferation of PANC-1 human pancreatic cancer cells. *Oncology Letters* **19**:3531–3541 DOI 10.3892/ol.2020.11469.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25(11)**:1422–1423 DOI 10.1093/bioinformatics/btp163.

Coulombe PA, Omary MB. 2002. 'Hard' and 'soft' principles defining the structure, function and regulation of keratin intermediate filaments. *Current Opinion in Cell Biology* **14(1)**:110–122 DOI 10.1016/S0955-0674(01)00301-5.

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugan JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Pilizota I, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Perez-Silva JG, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner MM, Szpak M, Thormann A, Tricomi FF, Urbina-Gomez D, Veidenberg A, Walsh TA, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt SE, Gr II, Loveland JE, Martin FJ, Moore B, Mudge JM, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Dyer S, Harrison PW, Howe KL, Yates AD, Zerbino DR, Flicek P. 2022. Ensembl 2022. *Nucleic Acids Research* **50**:D988–D995 DOI 10.1093/nar/gkab1049.

De Grassi A, Lanave C, Saccone C. 2008. Genome duplication and gene-family evolution: the case of three OXPHOS gene families. *Gene* **421**:1–6 DOI 10.1016/j.gene.2008.05.011.

Dittmar JM, Berger ES, Mao R, Wang H, Yeh HY. 2020. A probable case of multiple myeloma from Bronze Age China. *International Journal of Paleopathology* **31(18)**:64–70 DOI 10.1016/j.ijpp.2020.10.003.

Dong D, Jones G, Zhang S. 2009. Dynamic evolution of bitter taste receptor genes in vertebrates. *BMC Evolutionary Biology* **9**:12 DOI 10.1186/1471-2148-9-12.

Dvorankova B, Szabo P, Lacina L, Kodet O, Matouskova E, Smetana K Jr. 2012. Fibroblasts prepared from different types of malignant tumors stimulate expression of luminal marker keratin 8 in the EM-G3 breast cancer cell line. *Histochemistry and cell biology* **137**:679–685 DOI 10.1007/s00418-012-0918-3.

Eckstein M, Wirtz R, Gross-Weege M, Breyer J, Otto W, Stoehr R, Sikic D, Keck B, Eidt S, Burger M, Bolenz C, Nitschke K, Porubsky S, Hartmann A, Erben P. 2018. mRNA-expression of KRT5 and KRT20 defines distinct prognostic subgroups of muscle-invasive urothelial bladder cancer correlating with histological variants. *International Journal of Molecular Sciences* **19(11)**:19 DOI 10.3390/ijms19113396.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32(5)**:1792–1797 DOI 10.1093/nar/gkh340.

Elazezy M, Schwentesius S, Stegat L, Wikman H, Werner S, Mansour WY, Failla AV, Peine S, Müller V, Thiery JP, Ebrahimi Warkiani M, Pantel K, Joosse SA. 2021. Emerging insights into keratin 16 expression during metastatic progression of breast cancer. *Cancers* **13(15)**:3869 DOI 10.3390/cancers13153869.

Emran TB, Shahriar A, Mahmud AR, Rahman T, Abir MH, Siddiquee MF- R, Ahmed H, Rahman N, Nainu F, Wahyudin E, Mitra S, Dhama K, Habiballah MM, Haque S, Islam A, Hassan MM. 2022. Multidrug resistance in cancer: understanding molecular mechanisms, immunoprevention and therapeutic approaches. *Frontiers in Oncology* **12**:891652 DOI 10.3389/fonc.2022.891652.

Faltas B. 2011. Cancer is an ancient disease: the case for better palaeoepidemiological and molecular studies. *Nature Reviews Cancer* **11(1)**:76 DOI 10.1038/nrc2914-c1.

Gao XL, Wu JS, Cao MX, Gao SY, Cen X, Jiang YP, Wang SS, Tang YJ, Chen QM, Liang XH, Tang Y. 2017. Cytokeratin-14 contributes to collective invasion of salivary adenoid cystic carcinoma. *PLOS ONE* **12**:e0171341 DOI 10.1371/journal.pone.0171341.

Ha SA, Lee YS, Kim HK, Yoo J, Kim S, Gong GH, Lee YK, Kim JW. 2011. The prognostic potential of keratin 18 in breast cancer associated with tumor dedifferentiation, and the loss of estrogen and progesterone receptors. *Cancer Biomarkers* **10(5)**:219–231 DOI 10.3233/CBM-2012-0250.

Hardwick LJA, Philpott A. 2018. Xenopus models of cancer: expanding the oncologist's toolbox. *Frontiers in Physiology* **9**:1660 DOI 10.3389/fphys.2018.01660.

Hatta M, Miyake Y, Uchida K, Yamazaki J. 2018. Keratin 13 gene is epigenetically suppressed during transforming growth factor-beta1-induced epithelial-mesenchymal transition in a human keratinocyte cell line. *Biochemical and Biophysical Research Communications* **496**:381–386 DOI 10.1016/j.bbrc.2018.01.047.

Hawkridge AM. 2014. The chicken model of spontaneous ovarian cancer. *PROTEOMICS— Clinical Applications* **8(9–10)**:689–699 DOI 10.1002/prca.201300135.

He H, Li S, Hong Y, Zou H, Chen H, Ding F, Wan Y, Liu Z. 2015. Kruppel-like factor 4 promotes esophageal squamous cell carcinoma differentiation by up-regulating keratin 13 expression. *The Journal of Biological Chemistry* **290(21)**:13567–13577 DOI 10.1074/jbc.M114.629717.

Hesse M, Zimek A, Weber K, Magin TM. 2004. Comprehensive analysis of keratin gene clusters in humans and rodents. *European Journal of Cell Biology* **83(1)**:19–26 DOI 10.1078/0171-9335-00354.

Huang WC, Jang TH, Tung SL, Yen TC, Chan SH, Wang LH. 2019. A novel miR-365-3p/EHF/ keratin 16 axis promotes oral squamous cell carcinoma metastasis, cancer stemness and drug resistance via enhancing beta5-integrin/c-met signaling pathway. *Journal of Experimental & Clinical Cancer Research* **38**:89 DOI 10.1186/s13046-019-1091-5.

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering* **9(3)**:90–95 DOI 10.1109/MCSE.2007.55.

Jacob JT, Coulombe PA, Kwan R, Omary MB. 2018. Types I and II keratin intermediate filaments. *Cold Spring Harbor Perspectives in Biology* **10(4)**:a018275 DOI 10.1101/cshperspect.a018275.

Joosse SA, Hannemann J, Spotter J, Bauche A, Andreas A, Muller V, Pantel K. 2012. Changes in keratin expression during metastatic progression of breast cancer: impact on the detection of circulating tumor cells. *Clinical Cancer Research* **18(4)**:993–1003 DOI 10.1158/1078-0432.CCR-11-2100.

Kawai T, Yasuchika K, Ishii T, Katayama H, Yoshitoshi EY, Ogiso S, Kita S, Yasuda K, Fukumitsu K, Mizumoto M, Hatano E, Uemoto S. 2015. Keratin 19, a cancer stem cell marker in human hepatocellular carcinoma. *Clinical Cancer Research* **21**:3081–3091 DOI 10.1158/1078-0432.CCR-14-1936.

Khanom R, Sakamoto K, Pal SK, Shimada Y, Morita K, Omura K, Miki Y, Yamaguchi A. 2012. Expression of basal cell keratin 15 and keratin 19 in oral squamous neoplasms represents diverse pathophysiologies. *Histology and Histopathology* **27**:949–959 DOI 10.14670/HH-27.949.

Kosciuczuk EM, Lisowski P, Jarczak J, Strzalkowska N, Jozwik A, Horbanczuk J, Krzyzewski J, Zwierzchowski L, Bagnicka E. 2012. Cathelicidins: family of antimicrobial peptides. A review. *Molecular Biology Reports* **39**:10957–10970 DOI 10.1007/s11033-012-1997-x.

Kranke N. 2022. Explanatory integration and integrated explanations in Darwinian medicine and evolutionary medicine. *Theoretical Medicine and Bioethics* **44**:1–20 DOI 10.1007/s11017-022-09594-z.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35(6)**:1547–1549 DOI 10.1093/molbev/msy096.

Lan YJ, Chen H, Chen JQ, Lei QH, Zheng M, Shao ZR. 2014. Immunolocalization of vimentin, keratin 17, Ki-67, involucrin, beta-catenin and E-cadherin in cutaneous squamous cell carcinoma. *Pathology & Oncology Research* **20(2)**:263–266 DOI 10.1007/s12253-013-9690-5.

Lane EB, Alexander CM. 1990. Use of keratin antibodies in tumor diagnosis. *Seminars in Cancer Biology* **1**:165–179.

Lauriola I, Lavelli A, Aiolli F. 2022. An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing* **470(1)**:443–456 DOI 10.1016/j.neucom.2021.05.103.

Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49(W1)**:W293–W296 DOI 10.1093/nar/gkab301.

Li M, Liu J, Zhang C. 2011. Evolutionary history of the vertebrate mitogen activated protein kinases family. *PLOS ONE* **6**:e26999 DOI 10.1371/journal.pone.0026999.

Li C, Tang Z, Zhang W, Ye Z, Liu F. 2021. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. *Nucleic Acids Research* **49(W1)**:W242–W246 DOI 10.1093/nar/gkab418.

Lineweaver CH, Davies PC, Vincent MD. 2014. Targeting cancer's weaknesses (not its strengths): therapeutic strategies suggested by the atavistic model. *BioEssays* **36**:827–835 DOI 10.1002/bies.201400070.

Liu B. 2018. Text sentiment analysis based on CBOW model and deep learning in big data environment. *Journal of Ambient Intelligence and Humanized Computing* **11**:1–8 DOI 10.1007/s12652-018-1095-6.

Logotheti S, Pavlopoulou A, Marquardt S, Takan I, Georgakilas AG, Stiewe T. 2022. p73 isoforms meet evolution of metastasis. *Cancer Metastasis Reviews* **41**(4):853–869 DOI 10.1007/s10555-022-10057-z.

Lukyanchuk VV, Friess H, Kleeff J, Osinsky SP, Ayuni E, Candinas D, Roggo A. 2003. Detection of circulating tumor cells by cytokeratin 20 and prostate stem cell antigen RT-PCR in blood of patients with gastrointestinal cancers. *Anticancer Research* **23**:2711–2716.

Makashov AA, Malov SV, Kozlov AP. 2019. Oncogenes, tumor suppressor and differentiation genes represent the oldest human gene classes and evolve concurrently. *Scientific Reports* **9**(1):16410 DOI 10.1038/s41598-019-52835-w.

Marquardt S, Pavlopoulou A, Takan I, Dhar P, Putzer BM, Logotheti S. 2021. A systems-based key innovation-driven approach infers co-option of jaw developmental programs during cancer progression. *Frontiers in Cell and Developmental Biology* **9**:682619 DOI 10.3389/fcell.2021.682619.

Menz A, Weitbrecht T, Gorbokon N, Büscheck F, Luebke AM, Kluth M, Hube-Magg C, Hinsch A, Höflmayer D, Weidemann S, Fraune C, Möller K, Bernreuther C, Lebok P, Clauditz T, Sauter G, Uhlig R, Wilczak W, Steurer S, Minner S, Burandt E, Krech R, Dum D, Krech T, Marx A, Simon R. 2021. Diagnostic and prognostic impact of cytokeratin 18 expression in human tumors: a tissue microarray study on 11,952 tumors. *Molecular Medicine* **27**(1):16 DOI 10.1186/s10020-021-00274-7.

Moll R, Divo M, Langbein L. 2008. The human keratins: biology and pathology. *Histochemistry and Cell Biology* **129**(6):705–733 DOI 10.1007/s00418-008-0435-6.

Nesse RM. 2001. How is Darwinian medicine useful? *The Western Journal of Medicine* **174**:358–360 DOI 10.1136/ewjm.174.5.358.

Nikolaou M, Pavlopoulou A, Georgakilas AG, Kyrodimos E. 2018. The challenge of drug resistance in cancer treatment: a current overview. *Clinical & Experimental Metastasis* **35**(4):309–318 DOI 10.1007/s10585-018-9903-0.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**:D733–D745 DOI 10.1093/nar/gkv1189.

Olofsson MH, Ueno T, Pan Y, Xu R, Cai F, van der Kuip H, Muerdter TE, Sonnenberg M, Aulitzky WE, Schwarz S, Andersson E, Shoshan MC, Havelka AM, Toi M, Linder S. 2007. Cytokeratin-18 is a useful serum biomarker for early determination of response of breast carcinomas to chemotherapy. *Clinical Cancer Research* **13**:3198–3206 DOI 10.1158/1078-0432.CCR-07-0009.

Pavlopoulou A, Bagos PG, Koutsandrea V, Georgakilas AG. 2017. Molecular determinants of radiosensitivity in normal and tumor tissue: a bioinformatic approach. *Cancer Letters* **403**(Suppl. 2):37–47 DOI 10.1016/j.canlet.2017.05.023.

Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. *PLOS ONE* **5**:e13781 DOI 10.1371/journal.pone.0013781.

Pavlopoulou A, Scorilas A. 2014. A comprehensive phylogenetic and structural analysis of the carcinoembryonic antigen (CEA) gene family. *Genome Biology and Evolution* **6**:1314–1326 DOI 10.1093/gbe/evu103.

Saha SK, Kim K, Yang GM, Choi HY, Cho SG. 2018. Cytokeratin 19 (KRT19) has a role in the reprogramming of cancer stem cell-like cells to less aggressive and more drug-sensitive cells. *International Journal of Molecular Sciences* **19(5)**:1423 DOI 10.3390/ijms19051423.

Saha SK, Yin Y, Chae HS, Cho SG. 2019. Opposing regulation of cancer properties via KRT19-mediated differential modulation of Wnt/beta-catenin/notch signaling in breast and colon cancers. *Cancers* **11(1)**:99 DOI 10.3390/cancers11010099.

Sarma A. 2022. Biological importance and pharmaceutical significance of keratin: a review. *International Journal of Biological Macromolecules* **219**:395–413 DOI 10.1016/j.ijbiomac.2022.08.002.

Schweizer J̈, Bowden PE, Coulombe PA, Langbein L, Lane EB, Magin TM, Maltais L, Omary MB, Parry DAD, Rogers MA, Wright MW. 2006. New consensus nomenclature for mammalian keratins. *The Journal of Cell Biology* **174(2)**:169–174 DOI 10.1083/jcb.200603161.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13(11)**:2498–2504 DOI 10.1101/gr.1239303.

Sharma P, Alsharif S, Fallatah A, Chung BM. 2019. Intermediate filaments as effectors of cancer development and metastasis: a focus on keratins, vimentin, and nestin. *Cells* **8(5)**:497 DOI 10.3390/cells8050497.

Sivakumar S, Videla LS, Rajesh Kumar T, Nagaraj J, Itnal S, Haritha D. 2020. Review on Word2Vec word embedding neural net. In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. 282–290.

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49(D1)**:D605–D612 DOI 10.1093/nar/gkaa1074.

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47(D1)**:D506–D515 DOI 10.1093/nar/gky1049.

Toy HI, Karakulah G, Kontou PI, Alotaibi H, Georgakilas AG, Pavlopoulou A. 2021. Investigating molecular determinants of cancer cell resistance to ionizing radiation through an integrative bioinformatics approach. *Frontiers in Cell and Developmental Biology* **9**:620248 DOI 10.3389/fcell.2021.620248.

Trigos AS, Pearson RB, Papenfuss AT, Goode DL. 2017. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. *Proceedings of the National Academy of Sciences of the United States of America* **114**:6406–6411 DOI 10.1073/pnas.1617743114.

Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, Bruford EA. 2021. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49**:D939–D946 DOI 10.1093/nar/gkaa980.

Vaidya M, Dmello C, Mogre S. 2022. Utility of keratins as biomarkers for human oral precancer and cancer. *Life* **12(3)**:343 DOI 10.3390/life12030343.

van Dalen A. 1996. Significance of cytokeratin markers TPA, TPA (cyk), TPS and CYFRA 21.1 in metastatic disease. *Anticancer Research* **16**:2345–2349.

Velez-delValle C, Marsch-Moreno M, Castro-Munozledo F, Galvan-Mendoza IJ, Kuri-Harcuch W. 2016. Epithelial cell migration requires the interaction between the vimentin and keratin intermediate filaments. *Scientific Reports* **6(1)**:24389 DOI 10.1038/srep24389.

Wahba A, Herrerín J, Sánchez M. 2021. Metastatic carcinoma in human remains from TT110, Luxor, Egypt (ancient Thebes). *HOMO* **72(4)**:307–316 DOI 10.1127/homo/2021/1477.

Wang X, Xu X, Peng C, Qin Y, Gao T, Jing J, Zhao H. 2019. BRAF(V600E)-induced KRT19 expression in thyroid cancer promotes lymph node metastasis via EMT. *Oncology Letters* **18**:927–935 DOI 10.3892/ol.2019.10360.

Welch D, Kahen E, Fridley B, Brohl AS, Cubitt CL, Reed DR. 2019. Small molecule inhibition of lysine-specific demethylase 1 (LSD1) and histone deacetylase (HDAC) alone and in combination in Ewing sarcoma cell lines. *PLOS ONE* **14**:e0222228 DOI 10.1371/journal.pone.0222228.

Werner S, Keller L, Pantel K. 2020. Epithelial keratins: biology and implications as diagnostic markers for liquid biopsies. *Molecular Aspects of Medicine* **72**:100817 DOI 10.1016/j.mam.2019.09.001.

Williams SC. 2015. News feature: capturing cancer's complexity. *Proceedings of the National Academy of Sciences* **112(15)**:4509–4511 DOI 10.1073/pnas.1500963112.

Williams PA, Zaidi SK, Sengupta R. 2022. AACR cancer progress report 2022: decoding cancer complexity, integrating science, and transforming patient outcomes. *Clinical Cancer Research* **28(19)**:4178–4179 DOI 10.1158/1078-0432.CCR-22-2588.

Yang L, Zhang S, Wang G. 2019. Keratin 17 in disease pathogenesis: from cancer to dermatoses. *The Journal of Pathology* **247**:158–165 DOI 10.1002/path.5178.

Yao S, Huang HY, Han X, Ye Y, Qin Z, Zhao G, Li F, Hu G, Hu L, Ji H. 2019. Keratin 14-high subpopulation mediates lung cancer metastasis potentially through Gkn1 upregulation. *Oncogene* **38**:6354–6369 DOI 10.1038/s41388-019-0889-0.

Yu B, Kong D, Cheng C, Xiang D, Cao L, Liu Y, He Y. 2022. Assembly and recognition of keratins: a structural perspective. *Seminars in Cell & Developmental Biology* **128(Pt 21)**:80–89 DOI 10.1016/j.semcdb.2021.09.018.

Yuan X, Yi M, Dong B, Chu Q, Wu K. 2021. Prognostic significance of KRT19 in lung squamous cancer. *Journal of Cancer* **12(4)**:1240–1248 DOI 10.7150/jca.51179.

Zhang W, Fan Y. 2021. Structure of keratin. *Methods in Molecular Biology* **2347**:41–53 DOI 10.1007/978-1-0716-1574-4_5.

Zhong P, Shu R, Wu H, Liu Z, Shen X, Hu Y. 2021. Low KRT15 expression is associated with poor prognosis in patients with breast invasive carcinoma. *Experimental and Therapeutic Medicine* **21(4)**:305 DOI 10.3892/etm.2021.9736.