T.C.
DOKUZ EYLÜL ÜNIVERSITESI
IZMIR INTERNATIONAL
BIOMEDICINE AND GENOME
INSTITUTE

APPLICATION OF COMPUTATIONAL BIOLOGY APPROACHES FOR THE INVESTIGATION OF THE MOLECULAR MECHANISMS OF CANCER ACROSS TAXONOMIC GROUPS

IŞIL TAKAN

MOLECULAR BIOLOGY AND GENETICS

PhD THESIS

IZMIR-2022

THESIS CODE: DEU.HSI.PhD-2017850067

T.C.
DOKUZ EYLÜL ÜNIVERSITESI
IZMIR INTERNATIONAL
BIOMEDICINE AND GENOME

INSTITUTE

APPLICATION OF COMPUTATIONAL BIOLOGY APPROACHES FOR THE INVESTIGATION OF THE MOLECULAR MECHANISMS OF CANCER ACROSS TAXONOMIC GROUPS

MOLECULAR BIOLOGY AND GENETICS

PhD THESIS

IŞIL TAKAN

Supervisor:

Asst. Prof. ATHANASIA PAVLOPOULOU

THESIS CODE: DEU.HSI.PhD-2017850067

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
İZMİR ULUSLARARASI BİYOTIP VE GENOM ENSTİTÜSÜ

DOKTORA TEZ SAVUNMA SINAVI TUTANAĞI

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü Genom Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı Moleküler Biyoloji ve Genetik Doktora Programı 2017850067 numaralı öğrencisi Işıl Takan.'nın, "Farklı Taksonomik Gruplarda Kanserin Moleküler Mekanizmalarının Hesaplamalı Biyoloji Uygulamalarıyla Araştırılması" konulu Doktora tezini 08/06/2022 tarihinde yapılan savunma sınavı sonucunda başarılı olmuştur.

BAŞKAN
Dr. Öğr. Üyesi Athanasia PAVLOPOULOU

JÜRİ ÜYESİ
Dr. Öğr. Üyesi Hani ALOTAİBİ

JÜRİ ÜYESİ
Prof. Dr. Reşat ÜNAL

JÜRİ ÜYESİ
Prof. Dr. Ayse NALBANTSOY

JÜRİ ÜYESİ
Prof. Dr. Ferhat MATUR

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ

İZMİR ULUSLARARASI BİYOTIP VE GENOM ENSTİTÜSÜ

Dokuz Eylül University Izmir International Biomedicine and Genome Institute Department of Genomics and Molecular Biotechnology, Molecular Biology and Genetics PhD program student Işil TAKAN has successfully completed the PhD thesis entitled "Application of Computational Biology Approaches for the Investigation of the Molecular Mechanisms of Cancer across Taxonomic Groups" on the date of 08/06/2022.

HEAD OF THE COMMITTEE
Asst. Prof. Dr. Athanasia PAVLOPOULOU

MEMBER
Asst. Prof. Dr. Hani ALOTAİBİ

MEMBER
Prof. Dr. Reşat ÜNAL

MEMBER
Prof. Dr. Ayse NALBANTSOY

MEMBER
Prof. Dr. Ferhat MATUR

INDEXES

INDEX OF TABLES

INDEX OF FIGURES

LIST OF ABBREVIATIONS

KRT	Keratin
AI	Artificial Intelligence
NLP	Natural Language Processing
ML	Machine Learning
SQL	Structured Query Language
ANSI	American National Standards Institute
TCGA	The Cancer Genome Atlas
GTEx	Genotype Tissue Expression
STRING	Search Tool for the Retrieval of Interacting Genes
MEGA	Molecular Evolutionary Genetics Analysis
NGS	Next generation sequencing

RNA-Seq	RNA sequencing
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangio carcinoma
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
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
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
UVM	Uveal Melanoma

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor Asst. Prof. Dr Athanasia PAVLOPOULOU for her guidance, motivation, immense knowledge, and excellent, endless research support during my doctoral education. I could not have imagined having a better advisor and mentor for my PhD training.

I would like to extend my deep gratitude and appreciation to Asst. Prof. Dr Hani ALOTAİBİ for his valuable guidance and time. I would also like to thank all Pavlopoulou Lab members, especially Halil Ibrahim TOY, Didem ÖKMEN and Hande YILMAZ, for their kind support and friendship. I would also like to thank my friends Esra ESMERAY SÖNMEZ, Tevfik HATİPOĞLU and Deniz DOĞAN, who have always been with me since the beginning of my graduate, for their wonderful friendship and support in many ways.

I would also like to appreciate the scholarship I received from The Council of Higher Education (CoHE) (YÖK 100/2000) for providing financial support for my doctoral education.

I am sincerely appreciative to the rest of my committee, Prof. Dr. Reşat ÜNAL, Prof. Dr. Ferhat MATUR, Prof. Dr. Ayşe NALBANTSOY for their valuable time and contributions.

Also, I would like to take this opportunity to thank my dearest friend Gökay Mehmet BİZ with whom I have shared a lot. I really appreciate his support and understanding in my difficult times.

Above all else, I would like to thank my mother, Birgül TAKAN and my father, Aybey TAKAN, for always being there for me, believing in me and supporting me.

Application of Computational Biology Approaches for the Investigation of the Molecular Mechanisms of Cancer across Taxonomic Groups

Işıl Takan, International Izmir Biomedicine and Genome Institute,

ABSTRACT

Investigating the fundamental mechanisms of tumorigenesis is considered crucial in developing alternative therapies and new diagnostic strategies. These mechanisms, which were investigated in this study, could provide an indispensable guide to identifying alternative treatment methods in cancer research. Data mining methods allow the evaluation of large-scale biological data. To this end, cross-disciplinary *in silico* systems biology methods by combining comparative genomics, differentially expressed gene profiles, biological networks and biological text mining were employed to investigate the molecular mechanisms of carcinogenesis in diverse taxonomic groups. For this purpose, the following steps were carried out: (1) Identification of novel genes/proteins involved in different aspects of cancer through mining publicly accessible primary genomic databases, (2) Conduct comprehensive *in silico* phylogenetic analyses of genes/proteins, covering the major taxonomic divisions, (3) Identification of patterns of differentially expressed genes in different types of cancers through processing high-throughput data derived from public repositories, (4) Construction of functional networks and subsequent identification of gene/protein modules by employing biological network inference techniques, (5) Natural language processing using artificial intelligence techniques for the detection of semantic relations. The findings of this study could have potential application in the clinical setting. In this regard, the corresponding genes can be exploited as diagnostic or prognostic tumor markers, and they could also be utilized in cancer therapeutic decision-making or the rational strategy of future experimental studies for the design of targeted cancer drugs.

Keywords: Cancer; Evolution; Comparative genomics; Phylogeny; Data mining; Interaction network; Natural Language Processing.

Farklı Taksonomik Gruplarda Kanserin Moleküler Mekanizmalarının Hesaplamalı Biyoloji Uygulamalarıyla Araştırılması

Işıl Takan, İzmir Uluslararası Biyotıp ve Genom Enstitüsü,

ÖZET

Tümörogenezin temel mekanizmalarının araştırılmasının, alternatif tedaviler ve yeni teşhis stratejileri geliştirmede çok önemli olduğu kabul edilmektedir. Bu çalışmada araştırılan mekanizmalar, kanser araştırmalarında alternatif tedavi yöntemlerinin belirlenmesi için vazgeçilmez bir kılavuz olması mümkündür. Veri madenciliği yöntemleri, geniş çaplı biyolojik verilerin değerlendirilmesine olanak sağlar. Bu nedenle, karsinogenez moleküler mekanizmalarını araştırmak için karşılaştırmalı genomik, diferansiyel eksprese olan gen profillerinin belirlenmesi, biyolojik ağ analizi, biyolojik metin madenciliği yöntemleri birleştirilerek disiplinler arası sistem biyolojisi yöntemleri kullanılmıştır. In-siliko sistem biyolojisi yaklaşımı sırasıyla; (1) erişime açık birincil genomik veri tabanlarının madencilik yoluyla yeni genlerin/proteinlerin belirlenmesi, (2) Başlıca taksonomik grupları kapsayan, genlerin/proteinlerin kapsamlı filogenetik analizlerinin yürütülmesi, (3) erişime açık veri tabanlarından yüksek verimli veri işleme ile farklı kanser türlerinde diferansiyel olarak eksprese olan genlerin belirlenmesi, (4) işlevsel ağların oluşturulması ve daha sonra biyolojik ağ yorumlama teknikleri kullanılarak gen/protein modüllerinin belirlenmesi, (5) Anlamsal ilişkileri tespit etmek için yapay zeka tekniklerini kullanarak doğal dil işleme metotlarının uygulanması adımlarını içermektedir. Çalışma bulgularının klinik ortamda potansiyel uygulamaya sahip olması mümkündür. Bu bağlamda, ilgili genler, tanısal veya prognostik tümör belirteçleri olarak kullanılabilir; ayrıca kanser tedavisine yönelik karar vermek veya hedefe yönelik kanser ilaçları geliştirmek için gelecekteki deneysel çalışmaların rasyonel stratejisinde kullanılabilmesi mümkündür.

Anahtar kelimeler: Kanser; Evrim; karşılaştırmalı genomik; filogeni; Veri madenciliği; Etkileşim ağları; Doğal Dil İşleme.

1. INTRODUCTION AND OBJECTIVES

1.1. Statement and Importance of the Problem

Cancer is one of the most prevalent chronic diseases worldwide. Current anticancer modalities include radiation, chemotherapy, and surgery. DNA damage and genome instability caused by radiation and chemotherapy have also several adverse effects and usually cause toxicity. In general, current treatments provide a temporary advantage, and frequently result in a 'tsunami' that kills both cancer and healthy cells non-selectively, as well as the development of cancer cell resistance (Jaffray and Gospodarowicz 2015). Better understanding of the molecular mechanisms that underlie cancer would enable the identification of reliable diagnostic and prognostic biomarkers, improve clinical decision-making and lay the foundation for the development of targeted therapies.

1.2. Aims of the Study

The goal of this study is to identify fundamental biological processes and molecular mechanisms/pathways implicated in carcinogenesis, highly conserved across diverse taxonomic groups, which could be utilized in the clinical setting or cancer-oriented experimental studies. To this end, in this study, *in silico* systems biology approaches were employed, including comparative genomics, phylogenetics, high-throughput data processing, statistical models, network-based methods, and natural language processing.

1.3. Hypothesis of the Study

Given the limited benefits of current treatments, there is a need for more effective and efficient approaches, directed towards the selective targeting of cancer cells, and at the same time, sparing the healthy tissue/cells and reducing any adverse effects. The early and accurate diagnosis of cancer is also of great importance (Jaffray and Gospodarowicz 2015; Saville et al. 2019). Therefore, by tracing the origin of carcinogenesis and identifying the fundamental mechanisms of tumorigenesis would likely enable biomedical research to introduce novel diagnostic and alternative therapeutic approaches for cancer.

2. GENERAL INFORMATION

2.1. Unravelling the Evolutionary History of Cancer

An evolutionary perspective of carcinogenesis would enable a better understanding of the underlying mechanisms of cancer. A prominent example is the "atavistic model" of cancer (Davies and Lineweaver 2011; Louka et al. 2021)**,** which is based on the hypothesis that by tracing the origin of carcinogenesis and identifying the fundamental mechanisms of tumorigenesis would likely enable biomedical research to introduce novel diagnostic and alternative therapeutic approaches for cancer. According to this model, cancer results from the activation of highly conserved, primitive genes, leading to molecular phenotypes in single-celled organisms and population dynamics.

The disruption of genes and processes associated with increased intercellular collaboration, which occurred in early metazoan life, is a trigger of carcinogenesis (Bussey et al. 2017; Trigos et al. 2017)**.** The study of the underlying molecular mechanisms of cancer in different taxonomic groups would enable the identification of genes and pathways of high importance in carcinogenesis.

2.1.1. Evolution of jaws propels occurrence of tumor metastasis

Apart from being a disease of genetic deregulation, cancer is an evolutionary and ecological process, and tracing its evolution could indicate routes to metastasis. Darwinian laws apply to both, tumors and the respective organisms on which they grow, since heterogeneous tumor cell population undergoes selection of the fittest within a population of organisms which is also under evolutionary pressure. Moreover, all metazoa develop tumors, but differences in their metastatic potential are observed across phyla. Considering that all animal phyla may have derived from a common progenitor through major macroevolutionary events, how species capable of metastasis descended from nonmetastatic ancestors was investigated. Given that the genetic attributes of a specific species are accessible to the respective tumor, we hypothesized that the same genetic gains through which the species evolved may have offered selective advantages for the evolution of the primary tumor towards aggressive stages.

Cyclostomes are basal to the vertebrate clade and are the only living representatives of jawless vertebrates (agnatha). Cyclostomes is a monophyletic group (Teishima et al. 2014),

comprised of the Petromyzontia (lampreys), which are living fossils from the late Devonian period (Cui et al. 2019), and Myxinoidei (hagfishes).

The occurrence of tumors in Cyclostomes and the histopathological characteristics thereof has been the major focus of Falkmer and colleagues during the seventies (Bhattacharya, Ray Chaudhuri, and Roy 2018; Cui et al. 2019; Jizhong Wang et al. 2020) and, up to this day, these reports are considered the most extensive, detailed and well-controlled studies on this enigmatic basal vertebrate superclass.

Falkmer screened two large populations of hagfishes and lampreys for tumor incidence. Out of 6000 lampreys caught in the Ume and the Ricklean rivers between 1973 and 1974, only one individual (0.017%) presented well differentiated primary hepatocellular carcinoma. In contrast, of the 27,300 hagfishes captured in the Gullmar fjord within five years (1972-1976), up to 5.8% exhibited liver adenomas and carcinomas. This percentage was substantially higher as compared to the tumor-bearing hagfishes (2.8%) in the control group, which was the open-sea population of 1183 individuals that were caught outside the fjord.

Intriguingly, the affected hagfishes developed tumors with high or low differentiation, but no signs of metastasis (Bhattacharya, Ray Chaudhuri, and Roy 2018; Cui et al. 2019; Jizhong Wang et al. 2020). This increased cancer incidence was attributed to a mixture of PCBs (polychlorinated biphenyls) and DDT (dichlorodiphenyltrichloromethylmethane), two anthropogenic organochloride contaminants which were washed down the fjord. Because hagfishes were parasitizing on PCB/DDT-contaminated fishes, these agents bioaccumulated and triggered carcinomas in the liver and, to a lesser extent, in pancreatic islets (Jizhong Wang et al. 2020). The fact that despite their exposure to confirmed carcinogens (T. Chen et al. 2017; Y. Huang et al. 2015; Matsuda et al. 2011, 2012), a certain percentage of hagfishes developed malignant tumors, but none of the 28,483 subjects developed metastasis, suggests that these animals may be metastasis-refractory or metastasis-incapable. Nevertheless, in Chondrichthyes, fifty cases of spontaneous cancer, including invasive (T. Chen et al. 2017) and metastatic (Kim et al. 2019; Ju Wang et al. 2011) tumors, have been reported in the wild since the 1900s, suggesting that the

descendant species of Cyclostomes are capable of metastasis. There is an increased incidence of tumors in Osteichthyes and more frequent metastatic cases/events (Teishima et al. 2014; Ju Wang et al. 2011; Wein et al. 2017). In amphibia (Montazeri and Bellmunt 2020), reptiles and birds (Y. Huang et al. 2015), invasive and metastatic cancers are more common, while mammals show an increased prevalence of metastatic cancers (Brand et al. 2014). Collectively, an overview of the reported tumors within the vertebrate clade suggests the consistent occurrence of metastatic tumors in Chondrichthyes and descendants, thereby determining the origin of metastasis between Cyclostomes and Chondrichthyes (Figure 1). Metazoans occupy sequential node positions in the phylogenetic tree, each revealing for the first time an indispensable macroevolutionary trait that is consistently inherited by descendants (Table 1). In jawless vertebrates, an interplay between cancer and immune cells might have promoted co-evolution between metastatic tumors and the immune system in their gnathostome descendants. It is possible that cancer cells showed a preference for exploiting gnathostome- specific innovations, such as jaws, to increase their fitness into a rapidly changing tumor microenvironment, which became enriched in adaptive immune cells. If this is indeed the case, then perhaps metastasis can be perceived as a collateral effect of the evolution of the immune system in vertebrates.

Preclinical studies using in vitro cell lines, mouse xenografts and/or patient samples underscore that most of the invasive stages in a cancer type-dependent manner are known pro-metastatic drivers. For several of them, molecular modifiers have been developed and entered clinical trials, either as monotherapies or as combination regimens (The ClinicalTrials.gov database Accessed 1 May 2020). Regarding the age of the hubs of s jaw-developmental network which are deregulated in invasive cell lines, our studies indicate that 37.2% (16 out of 43) occurred first in gnathostomes, whereas 16.3% (7 out of 43) were of pre-gnathostome origin (Table 2).

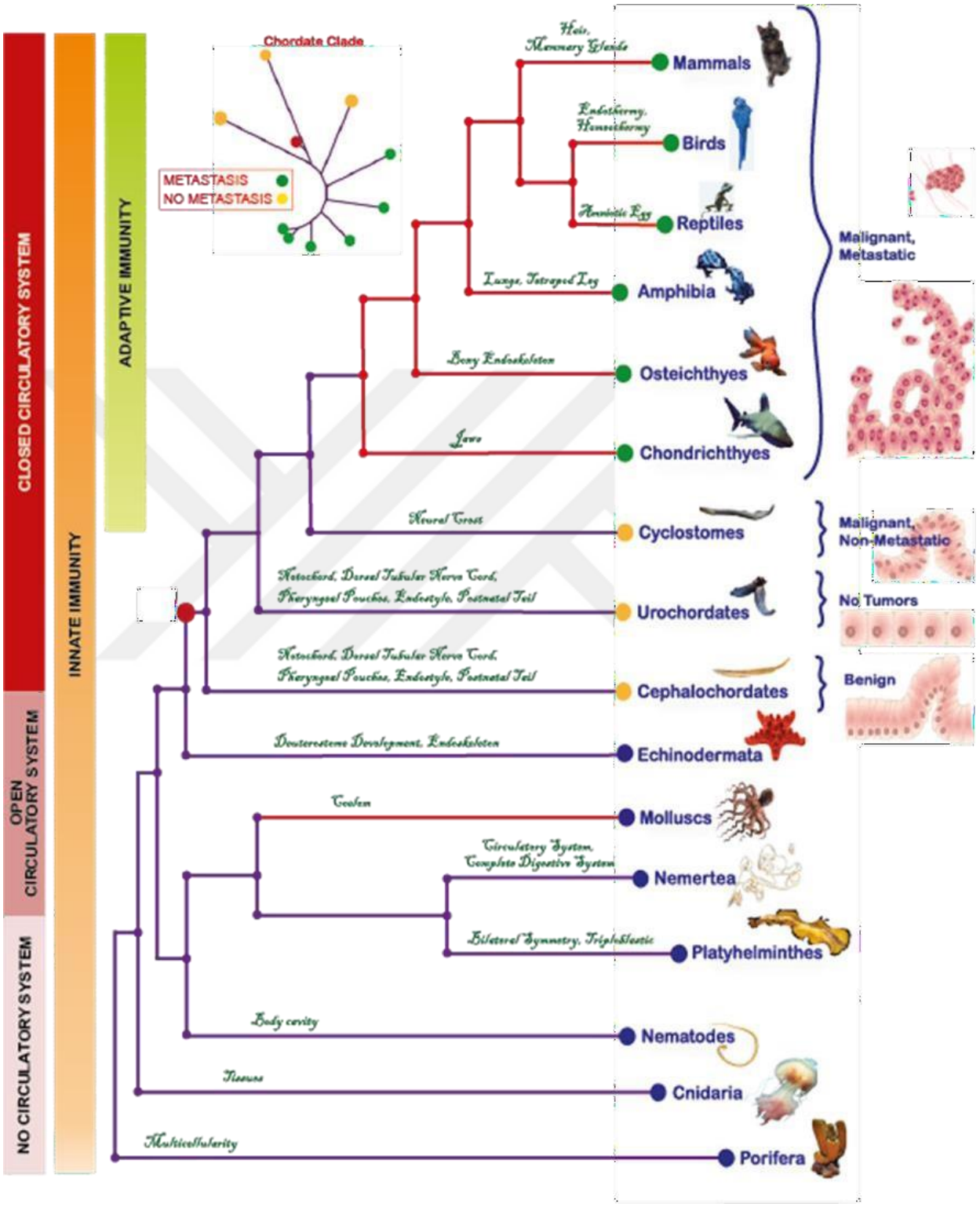


Figure 1: Phylogenetic tree representing the evolutionary relationships of metazoa and the emergence of metastasis in the cyclostome-to-Chondrichthyes transition Each phylogenetic group

has manifested, for the first time, a key innovation that was conserved throughout the lineage. In the case of Protochordates, due to their nodal position between invertebrates and vertebrates, both Cephalochordates and Urochordates were included in the present analysis. Red-colored clades indicate groups with lethal cancer types. Tumors were classified as benign, malignant/non-metastatic, or malignant and metastatic, based on information obtained from histological analyses. Invasive and metastatic tumors have been reported from Chondrichthyes to mammals. Mollusks have infectious and invasive cancers, while the urochordates are resistant to cancer. The red dot on the tree represents the common chordate ancestor, which gave rise to both, metastatic (green dots) and non- metastatic species (yellow dots). There is also the representation of the origins of a closed circulatory system and of acquired immunity. All chordates have a closed circulatory system. Adaptive immunity first occurs in cyclostomes, whereas all their ancestors exhibit only innate immunity. Model of the progression from a normal cell to metastatic cancer based on an original illustration by Iacobuzio-Donahue et al. 2012

Table 1: Taxa under study and hallmark macroevolutionary characteristic(s) that first appeared in each taxon. The circulatory and immune systems of each taxonomic group are also presented.

TAXON	EVOLUTIONARY HALLMARK FIRST APPEARED IN TAXON	CIRCULATORY SYSTEM	IMMUNE DEFENSE
Porifera	multicellularity	no circulatory system	innate immunity
Cnidaria	tissues	no circulatory system	innate immunity
Platyhelminthes	bilateral symmetry, triploblastic (endoderm-mesoderm-exoderm)	no circulatory system	innate immunity
Nemertea	circulatory system, complete digestive system	open circulatory system	Innate immunity
Nematodes	body cavity	no circulatory system	innate immunity
Molluscs	coelom	open circulatory system	innate immunity
Echinodermata	deuterostome development, endoskeleton	Main circulation of coelomic fluids by peritoneal cilia	innate immunity
Cephalochordates	notochord, dorsal tubular nerve cord, pharyngeal pouches, endostyle, postnatal tail	closed circulatory system. Ventral heart, with dorsal and ventral blood vessels	innate immunity
Urochordates	notochord, dorsal tubular nerve cord, pharyngeal pouches, endostyle, postnatal tail	closed circulatory system. Dorsal and ventral blood vessels. No heart	innate immunity
Cyclostomes	neural crest	closed circulatory system, 2-chambered heart	innate and adaptive immunity
Chondrichthyes	cartilaginous jaws	closed circulatory system, 2-chambered heart	innate and adaptive immunity
Osteichthyes	bony endoskeleton	closed circulatory system, 2-chambered heart	innate and adaptive immunity
Amphibia	lungs, tetrapod leg	closed circulatory system, double circulation, 3-chambered heart	innate and adaptive immunity
Reptiles	amniotic egg	closed circulatory system, double circulation, 3-chambered heart	innate and adaptive immunity

Birds	endothermy, homeothermy	closed circulatory system, double circulation, 4-chambered heart	innate and adaptive immunity
Mammals	hair, mammary glands	closed circulatory system, double circulation, 4-chambered heart	innate and adaptive immunity

Table 2: Status of preclinical and clinical cancer research studies on the hub JIGs that are deregulated in invasive cancer cell lines

	Study system(s)	TARGETED THERAPIES	REFs[#]

GENE[$]	PRECLINICAL STUDIES			Clinical trials	Approved/marketed drugs	
	Cancer type(s)	Effect on cell invasion and/or metastasis				
BMP4	breast, colon, esophageal squamous cell carcinoma, glioma, hepatocellular, melanoma	promoting or inhibiting	cancer cell lines, mouse xenografts, patient samples	1 in phase 1	ND*	[1-8]
CDON	prostate, endometrial, colorectal	promoting	cancer cell lines, mouse xenografts, patient samples	ND	ND	[9-11]
EGFR	breast, lung, melanoma, prostate	promoting	cancer cell lines, mouse xenografts, patient samples	15 in early phase 1 549 in phase 1 1071 in phase 2 258 in phase 3 37 in phase 4	cetuximab, dacomitinib, erlotinib, gefitinib, lapatinib, necitumumab, neratinib, osimertinib, panitumumab, vandetanib	[12-20]
FGF18	breast, lung, ovarian, gastric	promoting	cancer cell lines, mouse xenografts, patient samples	1 observational, recruiting	ND	[21-25]
FGF9	gastric, prostate, ovarian	promoting	cancer cell lines, mouse xenografts, patients samples	ND	ND	[26-31]
FGFR2	breast, colorectal, gastric	promoting	cancer cell lines, mouse xenografts, patient samples	9 in phase 1 27 in phase 2 6 in phase 3 1 in phase 4	erdafitinib: FDA-approved for metastatic bladder cancer pemigatinib: FDA approved for locally advanced or metastatic cholangiocarcinoma	[32-38]
FGFR3	bladder, hepatocellular	promoting	cancer cell lines, mouse xenografts	6 in phase 1 22 in phase 2 3 in phase 3	erdafitinib: FDA-approved for metastatic bladder cancer	[37-40]

					pemigatinib: FDA approved for locally advanced or metastatic cholangiocarcinoma	
GLI2	bladder, gastric, glioblastoma, hepatocellular, melanoma, osteosarcoma, prostate	promoting	cancer cell lines, mouse xenografts	ND	ND	[41-48]
GLI3	colorectal, lung, oral squamous cell carcinoma	promoting	cancer cell lines, mouse xenografts	ND	ND	[49-51]
GPC3	breast, gastric, hepatocellular, ovarian	promoting or inhibiting	cancer cell lines, mouse xenografts, patient samples	12 in phase 1, 2 in phase 2	ND	[52-57]
ITGB1	colorectal, esophageal, ovarian	promoting	cancer cell lines, mouse xenografts, patient samples	1 in early phase 1 1 in phase 1 2 in phase 2	ND	[58-62]
MMP2	breast, colorectal, lung	promoting	cancer cell lines, mouse xenografts, knockout mice	3 in Phase 1 1 in Phase 2	ND	[63-66]
SFN	lung, pancreatic	promoting	cancer cell lines, mouse xenografts, patient samples	ND	ND	[67-72]
SHH	bladder, breast, gastric, lung, oral squamous cell carcinoma, osteosarcoma, prostate	promoting	cancer cell lines, mouse xenografts, patient samples	1 in early phase 1 1 in phase 1 7 in phase 2 1 in phase 3	glasdegib: FDA-approved for AML sonidegib: FDA-approved for locally advanced basal cell carcinoma vismodegib: FDA-approved for metastatic basal cell carcinoma	[73-85]
SPRY2	colon, glioblastoma, lung, ovarian, renal cell carcinoma, rhabdomyosarcoma, prostate	promoting or inhibiting	cancer cell lines, mouse xenografts, patient samples	ND	ND	[86-94]

VEGFA	breast, colorectal, glioblastoma, lung, renal carcinoma, ovarian, sarcoma	promoting	cancer cell lines, mouse xenografts, patient samples	13 in early phase 1 145 in phase 1 338 in phase 2 68 in phase 3 8 in phase 4	axitinib, bevacizumab, cabozantinib, lenvatinib, pazopanib, ponatinib, ramucirumab, regorafenib, sorafenib, sunitinib, vandetanib, ziv-aflibercept	[95-98]
DYNC2H1	gastric	frequently mutated in gastric tumors with N3 stage lymph node metastasis	patient samples	ND	ND	[99]
EDNRA	bladder, breast, colon, head and neck, gastric, melanoma, nasopharynx, ovarian, prostate	promoting	cancer cell lines, mouse xenografts	3 in Phase 1 8 in phase 2 7 in phase 3	ND	[100-105]
GAS1	breast, colorectal, gastric, hepatocellular, melanoma	inhibiting	cancer cell lines, mouse xenografts, patient samples	ND	ND	[106-112]
IFT57	breast, prostate	associated with disease recurrence	patient samples	ND	ND	[113, 114]
LRP2	melanoma, renal cell carcinoma, prostate	associated with disease recurrence	patient samples	ND	ND	[115-117]
SMO	breast, head and neck squamous cell carcinoma	promoting	cancer cell lines, mouse xenografts, patient samples	1 in early phase 1 13 in phase 1 10 in phase 2	vismodegib: FDA-approved for metastatic basal cell carcinoma	[85] [118, 119]
WNT3A	colon, lung	promoting	cancer cell lines, mouse xenografts, patient samples	ND	ND	[120, 121]

$ Grey-colored rows highlight gnathostome-specific genes, and white lines genes of pregnathostome origin

*ND - not determined

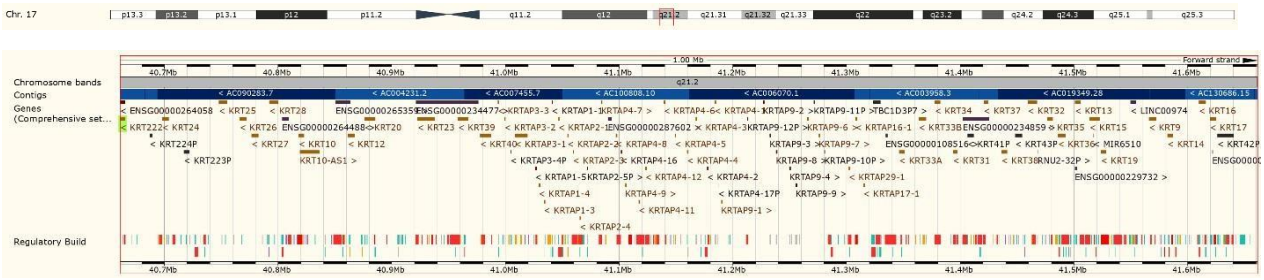
References are provided in ADDENDA 8.3.

2.1.2. Functional implications of keratins in cancer biology

Keratins (KRTs) constitute the intermediate filament-forming proteins of epithelial cells. They are also grouped as "soft" and "hard" keratins on the basis of their physicochemical properties. Hard keratins make up morphological structures such as scales, claws and feathers in birds and hair and nails in mammals. Soft epithelial keratins are highly involved in epithelial cell protection from mechanical and non-mechanical stressors and also play a regulatory role in apical-basal plasma membrane polarity, cell size, cell motility, protein synthesis, membrane trafficking and signaling pathways that regulate cell response to wound healing, cell growth and cell death (Coulombe and Omary 2002; Moll, Divo, and Langbein 2008). Because of these properties, soft keratins are known to have a prominent 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 (Karantza 2011).

The human genome contains both functional KRT genes and nonfunctional KRT pseudogenes, which are arranged in two uninterrupted clusters on chromosomes 12 and 17 (Figure 2 a-b) (Hesse et al. 2004). This exceptional characteristic of KRTs renders them ideal for evolutionary studies in order to infer the timing and direction of gene duplication events. However, there is a lack of a comprehensive *in silico* analysis of keratins which would enhance our understanding regarding the functional implications of keratins in cancer biology among tumor-bearing species.

(a)



(b)

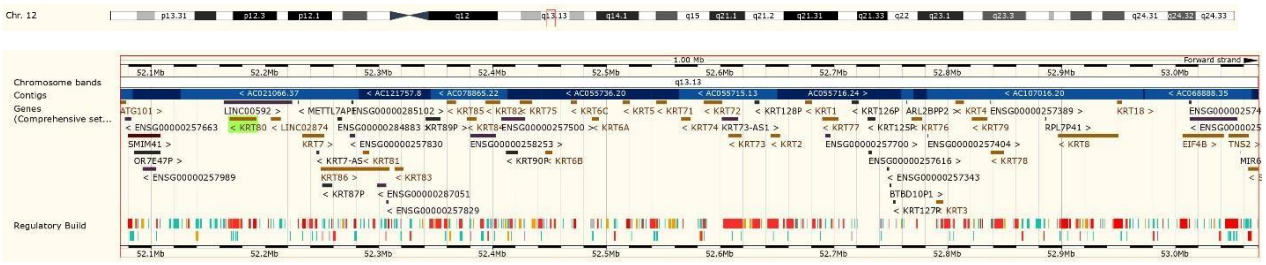


Figure 2: Keratin clusters on chromosomes (a) 12 and (b) 17 (Downloaded from Ensembl genome viewer; http://www.ensembl.org/).

Similar protein sequences tend to share similar biological functions and three-dimensional structures. Therefore, investigating the evolutionary relationships among the putative KRT candidate gene products and the well-annotated ones enabled us to assign roles/functions to the unknown proteins. Duplication and mutation events during evolution gave rise to the *KRT* gene family, which is essential for developing novel genes. It has been shown that kallikreins (Pavlopoulou et al. 2010), oxidative phosphorylation (OXPHOS), keratin-associated proteins (KRTAPs) (Wu, Irwin, and Zhang 2008), and bitter taste receptor (T2R) gene families (de Grassi, Lanave, and Saccone 2008), which have undergone a series of duplications, are implicated in critical physiological processes.

Despite the critical properties of keratins, there is a lack of comprehensive phylogenetic analysis to understand the relationship between keratins and species. Some studies have shown active keratin involvement in cancer cell invasion and metastasis, which are highly critical stages of tumorigenesis and essential in response to treatment. In this study, the evolutionary history of the *KERATIN* gene family was reconstructed.

For this purpose, comprehensive phylogenetic analyses of the corresponding KRT homologous proteins in multiple genomes were performed. To this end, the publicly available genomes, including those completed recently, were searched extensively for keratin homologs. Phylogenetic trees based on the entire length of keratin homologous proteins were reconstructed. Comprehensive phylogenetic analyses of KRT homologous proteins have been performed.

2.2. Bioinformatics approaches

2.2.1. Database Homology Search

Homologous sequences, that is, sequences with shared ancestry tend to exhibit similar sequences, structures and biological functions. Therefore, a putative protein sequence can be compared against well-annotated sequences in databases, like NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/; (Benson et al. 2017)), Ensembl (http://www.ensembl.org/; (Zerbino et al. 2018)) or UniProtKB (https://www.uniprot.org/uniprot/; (UniProt Consortium 2018)), with resolved structure and experimentally verified function, in order to predict its structure and infer its function. This process is called database sequence similarity search, or homology searching, where a query sequence is searched against each sequence in the database for the discovery of similar sequences. BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) is a popular local alignment heuristic algorithm for database homology search.

2.2.2. Sequence Alignment

Sequence alignment is the first step used to quantify similarity and allows to make inferences regarding the evolutionary history of the aligned sequences (Hodgman, 2010).

Pairwise sequence alignment is the alignment of two sequences (RNA, DNA or protein sequences). Pairwise sequence alignments can be interpreted in terms of evolution. For instance, when identical residues are aligned, then these residues are derived from a common ancestral sequence and have remained unchanged (conserved), suggesting structural or functional importance of this residue; conversely, when non-identical residues are aligned, then a residue has been likely substituted during the course of evolution by another one.

Multiple sequence alignment (MSA) is the alignment of three or more RNA, DNA or protein sequences (L. Wang and Jiang 1994). There are several MSA methods (e.g., iterative, progressive, consistency-based) used to maximize the scores and the accuracy of the alignments. MSA illustrate the inter-relationships among sequences and are informative about evolutionary conservation. Conserved residues that are aligned across the length of quite diverse sequences have presumably

descended from a common ancestor and they often play a critical role in the maintenance of structural stability or biological function. Notably, MSA represents the most critical step in phylogenetic analysis.

2.2.3. Phylogenetic Analysis

Phylogenetics is the study of the evolutionary relationships among sequences or taxa. Phylogenetic analysis can be employed to describe relationships of paralogues in a gene family, and the differentiation and cancer development (Salipante and Horwitz 2006; Yang and Rannala 2012).

2.2.3.1. Phylogenetic Tree Construction

Investigating the evolutionary relationships among the putative genes/gene products and the well-annotated ones would allow the assignment of function and structure to the unknown genes/proteins. Phylogenetic trees are used to visualize the inferred evolutionary relationships. The use of alignment-based methodologies in phylogenetic tree construction is common and affects tree structure formation with alignment parameters, including insertions, deletions, and gaps. The methods for reconstructing phylogenetic trees are classified into two major categories, the distance-based and the character-based (Benner, Bačák and Bourguignon 2014, Gonnet 2012).

2.2.3.2. Character-Based Methods of Phylogenetic Inference

Character-based methods such as maximum parsimony and maximum likelihood, which use probability and considerable variation in a set of sequences, compare all sequences simultaneously, consider one character (or column in an alignment) separately, and search for the tree that accommodates best all of this information (Yang and Rannala 2012). They evaluate a subset of all plausible trees, discarding those trees with a topology unlike to fit the observed data. Although maximum parsimony and maximum likelihood approaches consider the tree that requires the least number of changes to perform the alignment as the tree with the best score, the maximum parsimony is adversely affected by long-branch attraction (i.e., sequences with no significant similarity cluster together) and provides the least information about branch lengths (Jäger 2018).

2.2.3.3. Distance-Based Methods of Phylogenetic Inference

Distance-based methods use the dissimilarity (the distance) between the two sequences to construct trees. The computational intensity of these methods is less compared to the character-based methods which are more accurate due to the consideration of evolution at specific sites (e.g., catalytic sites). The evolutionary distance matrix and the tree-building algorithm are components of distance-based methods. The evolutionary distance matrix has the distance between pairs of objects derived from a substitution model, and then a tree is constructed hierarchically from the distance matrix (Jäger 2018).

2.2.3.4. Bootstrap analysis

Bootstrapping is the most common approach to evaluate the reliability of the inferred phylogenetic tree. In bootstrapping, the original dataset is randomly sampled by replacing the columns of a multiple sequence alignment to generate multiple pseudo-datasets of the same size as the original dataset (Soltis 2003). Some columns may be sampled more than once and others not at all. Trees for each pseudo-dataset are constructed. In the generated consensus tree, the frequency with which each node in the tree is reproduced is calculated. Bootstrap values depict the reliability of particular branches in the phylogenetic tree, where values above 70% for any given node provide 95% confidence that the grouping is reliable. In this study, bootstrap analysis was conducted to evaluate the statistical significance of the reconstructed trees.

2.2.4. Database Construction

Structured Query Language (SQL), one of the databases managing language (ANSI: American National Standards Institute), is used to perform database construction, deletion, modifying rows such various operations (Sengupta and Dalkilic 2002). For this thesis, the database was created to easily access, manage, and update the collection of the datasets. Microsoft SQL Server Management Studio was used to manage the cancer database (https://www.microsoft.com/).

2.2.5. Natural Language Processing

Natural language processing (NLP) combines Artificial Intelligence (AI) and computational linguistics to construct logical relations between human language and computers, so as to convert large amounts of text documents into a form that can be processed and be suitable for further

analysis. This thesis focused on establishing semantic relationships among the top ranking 50 similar words implicated in keratins and cancer for each member of the *KRT* gene family through biomedical literature mining, with the application of NLP methods (Lou, et al. 2019).

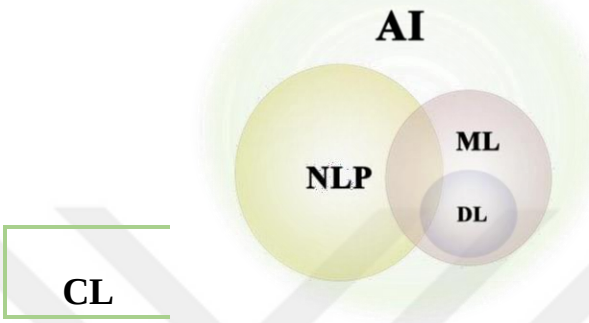


Figure 3: The scheme depicts the association of artificial intelligence (AI), natural language processing (NLP), machine learning (ML), deep learning (DL) and computational linguistics (CL).

2.2.5.1. Tokenization

Tokenization is a method of dividing a text document into smaller units called tokens to be able to recognize the entities. Tokens can consist of one word or more or can be characters or sub-words (Garat and Wonsever 2019).

2.2.5.2. Stop words

Stop words typically apply to the most frequent words in a language; there is no one standardized stop word list used by all the natural language processing tools. Removing stop words in Named entity recognition (NER) (Zhang and Elhadad 2013), categorizing and extraction of key entity, increases the performance rate and affects the ranking, and thus stop words are generally omitted from the datasets.

2.2.5.3. Stemming

Stemming algorithms work by meaningfully trimming inflectional groups at the end or beginning of a word, considering a list of common suffixes in a word to reach a common base word

form. The indiscriminate segment of data preparation is successful most of the time but not always. Therefore, finding the root of words through lemmatization becomes important (Sugumar 2018).

2.2.5.4. Lemmatization

As opposed to stemming, lemmatization takes into account the morphological analysis of words. In order to analyze words morphologically, it is necessary to have detailed dictionaries that the algorithm can utilize. In the natural language processing libraries, stemming algorithms are used in conjunction with lemmatization algorithms to obtain more accurate results (Arora and Kansal 2019).

2.2.5.5. Word embeddings

Word Embeddings are considered the most popular learned representation methods for text, where words have the same meaning in a numerical manner. Word2Vec is one Natural Language Processing system that uses neural networks in order to create a distributed word representations in a corpus (Wu, et al. 2019).

A continuous Bag of Words (CBOW), one of the word representation language model architectures, was constructed, which forms a new text vector representation. Word2Vec training methods include CBOW as a parameter to learn text features in a more effective way (Liu 2018).

3. MATERIALS AND METHODOLOGY

3.1. Type of the Study

This study was conducted based on computational methods.

3.2. Time and Place of the Study

This study was conducted in Izmir Biomedicine and Genome Institute from December 2020 to May 2022.

3.3. Materials and Methods of the Study

In this thesis, • Protein sequences were collected from the:

National Center for Biotechnology Information (NCBI) RefSeq [https://www.ncbi.nlm.nih.gov/refseq/].

ENSEMBL [(http://www.ensembl.org/].

UniProt KnowledgeBase (UniProtKB) (https://www.uniprot.org/uniprot/).

- RNA sequencing (RNA-Seq) gene expression data for tumor and corresponding normal tissue samples from the TCGA (The Cancer Genome Atlas) and GTEx (Genotype- Tissue Expression) databases, respectively, were downloaded from the GEPIA2 (Gene Expression Profiling Interactive Analysis) online web server (Tang et al. 2019) (http://gepia2.cancer-pku.cn/).
- The gene/protein associations were retrieved from STRING (Search Tool for the Retrieval of Interacting Genes) (Szklarczyk et al. 2017) (https://string-db.org/).
- Articles were collected from the scientific literature database MEDLINE/PubMed (https://pubmed.ncbi.nlm.nih.gov/) for the analysis of word embeddings for semantic relationship inference.

3.4. Variables of the Study

The variables of this study are: tumor samples and corresponding normal tissue samples obtained from the TCGA and GTEx databases, respectively; protein sequences collected from

NCBI RefSeq (O'Leary et al. 2016), ENSEMBL (Zerbino et al. 2018), UniProtKB (UniProt Consortium 2018); articles collected from the scientific literature database MEDLINE/PubMed (https://pubmed.ncbi.nlm.nih.gov/) for the analysis of word embeddings for semantic relationship inference.

3.5. Programming Languages and Tools for Data Processing

Programming languages and software are list below:

- R programming language analyses (http://www.r-project.org/) (differential gene expression analysis and visualization)
- Python programming language (https://www.python.org/) (natural language processing, phylogenetic analysis, sequence alignment, data collection and processing, import data to SQL server)
- MSSQL (cancer database construction)
- Bash scripting (multiple sequence alignment, data collection and processing)
- Cytoscape (https://cytoscape.org/) (visualization of semantic relation networks, protein-protein interaction network and gene-cancer network)
- NCBI (https://www.ncbi.nlm.nih.gov/) online tools (BLASTP, TBLASTN)
- STRING (https://string-db.org/) (Protein-Protein Interaction)
- ENSEMBL (http://www.ensembl.org/) (Genome Browser)
- Software package MEGA (https://www.megasoftware.net/) version 10. 1
  (phylogenetic tree construction)
- iTOL (https://itol.embl.de/) online (phylogenetic tree visualization)

3.6. Research plan and timeline

<table>
<tr><td>Type I KRT gene family

• Sequence data collection
• Cross-Genome Search for Orthologs
• Phylogenetic analysis</td><td>December-May 2020</td></tr>
<tr><td>Type II KRT gene family

• Sequence data collection
• Cross-Genome Search for Orthologs
• Phylogenetic analysis</td><td>May-June 2021</td></tr>
<tr><td>Cancer Database Construction</td><td rowspan="4">May-2021 March 2022</td></tr>
<tr><td>Differential Gene Expression Analysis</td></tr>
<tr><td>Protein-Protein Interaction Network Analysis</td></tr>
<tr><td>Natural Language Processing of Keratins</td></tr>
</table>

3.7. Phylogenetic Tree Construction

Phylogenetic trees were constructed in order to infer and visualize phylogenetic relationships. The overall procedure consists of two major steps, homologs search and phylogenetic inference, which are further divided into six major substeps described below.

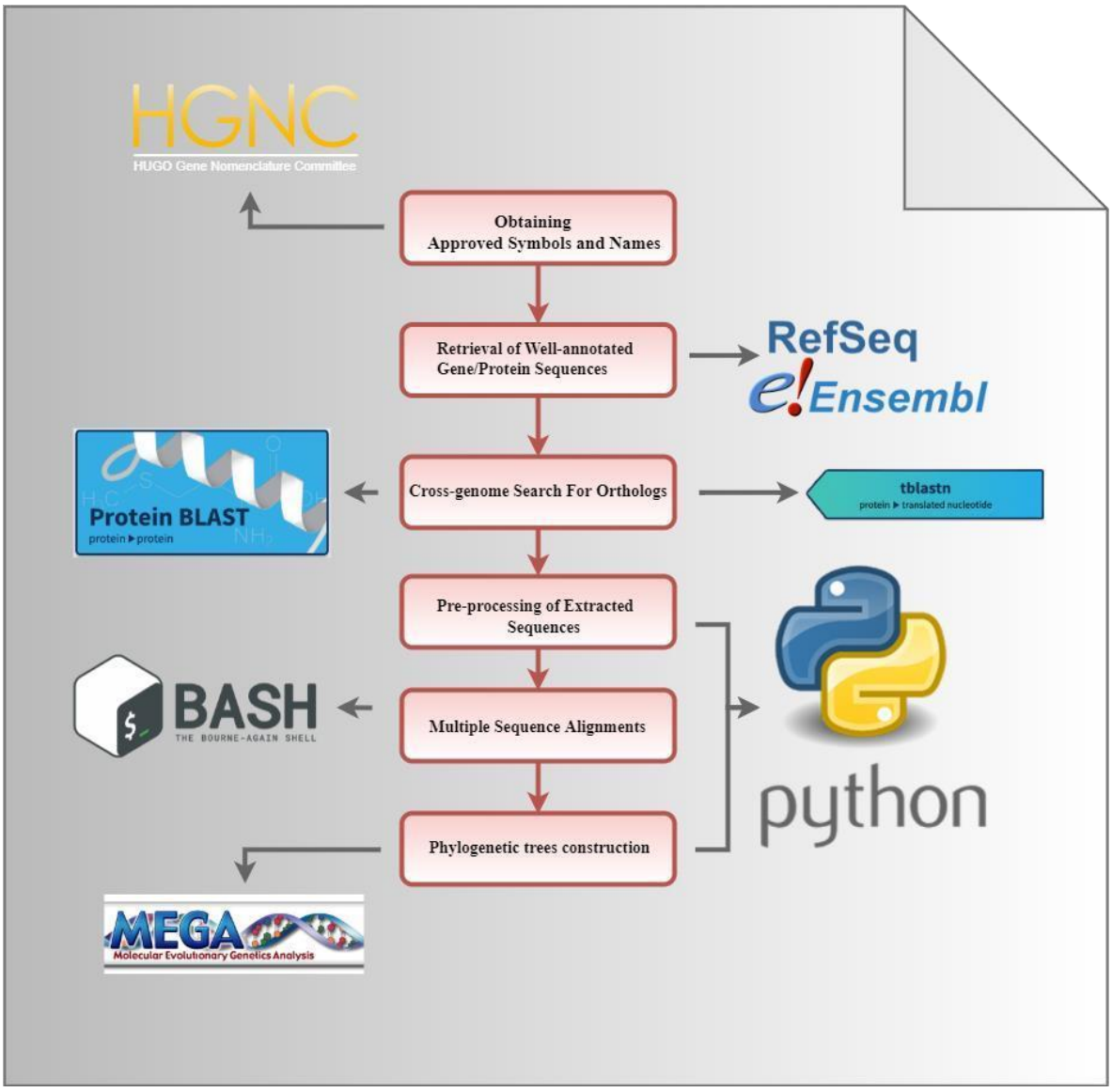


Figure 4: Graphical representation of the workflow for phylogenetic inference.

3.7.1. Database Searching

Approved symbols and names of keratins were retrieved from the HUGO Gene Nomenclature Committee (HGNC) database (https://www.genenames.org/; (Bruford et al. 2008)) (Table 3). The symbols were used to retrieve well-annotated gene/protein sequences.

Table 3: HGNC IDs, approved symbols, names, and synonyms for Type I KRT genes.

HGNC ID (gene)	Approved symbol	Approved name	Previous symbols	Aliases
6412	KRT1	keratin 1	EHK1	KRT1A

6439	KRT2	keratin 2	KRT2A	KRTE
6440	KRT3	keratin 3		CK3, K3
6441	KRT4	keratin 4	CYK4	CK4, K4
6442	KRT5	keratin 5	EBS2	KRT5A
6443	KRT6A	keratin 6A	KRT6C, KRT6D	CK6C, K6C, CK6D, K6D
6444	KRT6B	keratin 6B	KRTL1	
20406	KRT6C	keratin 6C	KRT6E	
6445	KRT7	keratin 7		K7, CK7, K2C7, SCL
6446	KRT8	keratin 8		CARD2, K8, CK8, CYK8, K2C8, KO
28927	KRT71	keratin 71		KRT6IRS, KRT6IRS1, K6IRS1
28932	KRT72	keratin 72		K6IRS2, KRT6IRS2, KRT6, K6irs
28928	KRT73	keratin 73		KRT6IRS3, K6IRS3
28929	KRT74	keratin 74		K6IRS4, KRT5C, KRT6IRS4
24431	KRT75	keratin 75		K6HF
24430	KRT76	keratin 76		HUMCYT2A, KRT2B, KRT2P
20411	KRT77	keratin 77	KRT1B	
28926	KRT78	keratin 78		K5B
28930	KRT79	keratin 79		K6L, KRT6L
27056	KRT80	keratin 80		KB20
6458	KRT81	keratin 81	KRTHB1	Hb-1
6459	KRT82	keratin 82	KRTHB2	Hb-2
6460	KRT83	keratin 83	KRTHB3	Hb-3
6461	KRT84	keratin 84	KRTHB4	Hb-4
6462	KRT85	keratin 85	KRTHB5	Hb-5
6463	KRT86	keratin 86	KRTHB6	MNX, Hb6
28695	KRT222	keratin 222	KRT222P	KA21, MGC45562

Table 4: HGNC IDs, approved symbols, names, and synonyms for type II KRT genes.

HGNC ID (gene)	Approved symbol	Approved name	Previous symbols	Aliases
6447	KRT9	keratin 9		EPPK, K9, CK-9
6413	KRT10	keratin 10	KPP	K10, CK10
6414	KRT12	keratin 12		K12
6415	KRT13	keratin 13		K13, CK13, MGC3781, MGC161462
6416	KRT14	keratin 14	EBS3, EBS4	
6421	KRT15	keratin 15		K15, CK15, K1CO
6423	KRT16	keratin 16		NEPPK
6427	KRT17	keratin 17	PCHC1	
6430	**KRT18**	**keratin 18**		
6436	KRT19	keratin 19		K19, CK19, K1CS, MGC15366
20412	KRT20	keratin 20		CK20, K20, MGC35423
6438	KRT23	keratin 23		K23, DKFZP434G032, HAIK1, CK23, MGC26158
18527	KRT24	keratin 24		FLJ20261, MGC138169, MGC138173
30839	KRT25	keratin 25	KRT25A	
30840	KRT26	keratin 26	KRT25B	
30841	KRT27	keratin 27	KRT25C	
30842	KRT28	keratin 28	KRT25D	
6448	KRT31	keratin 31	KRTHA1	Ha-1
6449	KRT32	keratin 32	KRTHA2	Ha-2
6450	KRT33A	keratin 33A	KRTHA3A	Ha-3I, Krt1-3
6451	KRT33B	keratin 33B	KRTHA3B	Ha-3II
6452	KRT34	keratin 34	KRTHA4	Ha-4
6453	KRT35	keratin 35	KRTHA5	Ha-5
6454	KRT36	keratin 36	KRTHA6	
6455	KRT37	keratin 37	KRTHA7	
6456	KRT38	keratin 38	KRTHA8	

32971	KRT39	keratin 39		KA35
26707	KRT40	keratin 40		FLJ36600, KA36

3.7.2. Retrieval of Annotated Sequences

The databases for acquiring the corresponding sequences of the most well-annotated protein-coding genes, with experimentally verified biological function or resolved protein tertiary structure were:

- NCBI's RefSeq (O'Leary et al. 2016), which provides manually annotated, non-redundant sequences with expert annotation,
- ENSEMBL (Zerbino et al. 2018), which contains eukaryotic genomes,
- UniProtKB (UniProt Consortium 2018), which is the most comprehensive, centralized repository of non-redundant, manually curated protein sequences

The protein sequences are more evolutionarily conserved than their corresponding nucleotide sequences. For this reason, in our comparative genomics and phylogenetic analyses, only proteins encoded by the jaw-indispensable genes (JIGs) and *keratin* genes were used.

The corresponding protein sequences of the most well-annotated protein-coding *KRT* genes were retrieved from NCBI's RefSeq; to obtain the corresponding protein sequence accession numbers, Python Entrez (Buchmann and Holmes 2019), Regex (Regular expressions, called regex patterns) library (van Rossum and Drake Jr 1995) was utilized (Script 1).

The 253 human JIG symbols are listed in the table in Addenda 8.2 were used initially to retrieve the corresponding, well–annotated, gnathostome (*Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio* and *Callorhinchus milii*) protein sequences from the publicly available non-redundant sequence database NCBI RefSeq (O'Leary et al. 2016).

Script 1.1. was used to import required libraries (re, urlib, time, csv, json, glob (van Rossum and Drake Jr 1995) and Bio (Cock et al. 2009)) and to open a session and send a request to NCBI.

Script 1

Script 1.1.

```
# Importing required libraries
Entrez.email = “isiltakan@gmail.com”
# To increase API limits, turn to uncomment the below line and write your own API key.
# Entrez.api_key = ‘your_api_key’
requests.packages.urllib3.disable_warnings(category=InsecureRequestWarning)
# Hyperparameters used on the website
headers = {**the hyperparameters was removed due to the attribute of temporariness of this section’}
# Definitions required to open the session
def init():
    ch = requests.Session()
    adapter = requests.adapters.HTTPAdapter(max_retries=30)
    ch.mount(‘https://’, adapter) ch.mount(‘http://’, adapter)
    cookies = requests.utils.cookiejar_from_dict(requests.utils.dict_from_cookiejar(ch.cookies))
    return [ch, cookies]
# An http function created not to crowd try-except
blocks def cli(addr, ch=None, cookies=None): try:
      grab = ch.get(addr, headers=headers, cookies=cookies, timeout=200, verify=False, allow_redirects=F
alse) except:
       time.sleep(10) ch,
       cookies = init()
      grab = ch.get(addr, headers=headers, cookies=cookies, timeout=200, verify=False, allow_redirects=F
alse) return
   grab
```

Sequences were fetched using *Entrez* from BioPython (Buchmann and Holmes 2019; Cock et al. 2009). The *search* function was used to obtain well annotated RefSeq protein accession codes (Script 1.2.).

Script 1.2.

```
# Get refseq from search term def
search(term, ch=None, cookies=None):
   # regex pattern
   pattern = r"<a href=\"https://www.ncbi.nlm.nih.gov/protein/(.*?)\"
         id=\"more_refseqproteins\""
   grab = cli('https://www.ncbi.nlm.nih.gov/protein/?term=' + urlencode(term),
         ch, cookies)
   grab = grab.text
grab = re.findall(pattern, grab) # scrape ids if
   grab:
      return (grab[0].split(','))[0] # return idlist and count
   else: return
      'None'

# Download sequences using ID def
download(id, x):
   handle = Entrez.efetch(db='protein', id=id, 29etype='fasta',
   retmode='text') f = open('%s.txt' % x, 'w') f.write(handle.read())
   handle.close() f.close()
```

The main function of the Script 1 was used to implement the defined functions *download* and *search* to collect accession IDs for 'Bos taurus', 'Gallus gallus', 'Homo sapiens', 'Canis lupus familiaris', 'Pan troglodytes', 'Macaca mulatta', 'Sus scrofa', 'Rattus norvegicus', 'Xenopus laevis' (Script 1.3.).

Script 1.3.

```
# Search ids def
main(genes, species): ids
= [] ch, cookies = init()
for gene in genes:
    RefSeqID = gene + ' ' + species x
    = search(RefSeqID, ch, cookies) if
    x is not None:
      ids.append(gene + ';' + x)
    else:
      ids.append(gene + ';' + 'None')
  with open('ids_%s.txt' % species, 'w') as f:
    f.write('\n'.join(ids))
    f.close()
genes=['KRT1', 'KRT2', 'KRT3', 'KRT4', 'KRT5', 'KRT6A', 'KRT6B', 'KRT6C', 'KRT7', 'KRT8', 'KRT71',
'KRT72', 'KRT73', 'KRT74', 'KRT75', 'KRT76', 'KRT77', 'KRT78', 'KRT79', 'KRT80', 'KRT81', 'KRT82', '
KRT83', 'KRT84', 'KRT85', 'KRT86', 'KRT222', 'KRT9', 'KRT10', 'KRT12', 'KRT13', 'KRT14', 'KRT15', '
KRT16', 'KRT17', 'KRT18', 'KRT19', 'KRT20', 'KRT23', 'KRT24', 'KRT25', 'KRT26', 'KRT27', 'KRT28', '
KRT31', 'KRT32', 'KRT33A', 'KRT33B', 'KRT34', 'KRT35', 'KRT36', 'KRT37', 'KRT38', 'KRT39', 'KRT4
0']
species = ['Bos taurus', 'Gallus gallus', 'Homo sapiens', 'Canis lupus familiaris', 'Pan troglodytes', 'Macaca
mulatta', 'Sus scrofa', 'Rattus norvegicus', 'Xenopus laevis']
# main function for all
species for sp in species:
main(genes, sp)
```

3.7.3. Cross-Genome Search for Orthologs

The well-annotated gene/protein sequences were used to search for corresponding homologs in the databases ENSEMBL (Zerbino et al. 2018), NCBI's GenBank (Benson et al. 2017) and UniProtKB (UniProt Consortium 2018), as well as species-specific genome sequence databases (van Dijk 2017) in an iterative manner until convergence, that is, no novel putative sequences could be detected, by employing reciprocal BLASTP and TBLASTN (Altschul et al. 1990) for both

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

In order to identify the 'true orthologs' of each jaw-indispensable gene/protein (Addenda 8.2), phylogenetic trees (a total of 253) of the homologous, gnathostome and agnathan, protein sequences were constructed. To select the appropriate outgroup for each ortholog (Table 1), the corresponding *H. sapiens* and *M. musculus* protein sequences were used as seeds in a BLASTP search (with default parameters) to identify the most closely related human and mouse sequence, respectively (shown as the best 'hit' in the BLAST output).

3.7.4. Pre-processing of Extracted Sequences

3.7.4.1. Merging Accession Numbers

After the sequence accession numbers were obtained separately, all sequence IDs that correspond to different KRT proteins were merged using Script 2. Pandas (McKinney et al. 2010) and its functions were used to process those IDs.

Script 2

```
# merge all IDS
# Importing required libraries
# File path path = "PATH/fileName*.txt"
files = []

for filename in glob.glob(path):
    files.append(filename)
```

```
frames = [] cnt
= 0

for file in files: f_name =
        '_'.join(os.path.basename(file).split('_')[1].split('.')[0].split(' '))
   dfall = pd.read_csv(file, sep=';', names=['genes', f_name])
   if cnt == 0:
      frames.append(dfall)
   else:
      frames.append(dfall[f_name])
   cnt += 1

df = pd.concat(frames, axis=1) df.to_csv('fileName.csv',
sep=';',index=False)
```

3.7.4.2. Merging File Content and Downloading Sequences

After the corresponding KRT protein sequences were downloaded with the Script 1 *download* function, the sequences were merged using Script 3.

Script 3

```
# Merge file content
# Importing required libraries
path = "PATH/fileName *.txt"
all = ""
type_1_seqs = temp_1 = "" type_2_seqs
= temp_2 = ""

type_2 = ['KRT1', 'KRT2', 'KRT3', 'KRT4', 'KRT5', 'KRT6A', 'KRT6B', 'KRT6C', 'KRT7', 'KRT8', 'KRT7
1', 'KRT72', 'KRT73', 'KRT74', 'KRT75', 'KRT76', 'KRT77', 'KRT78', 'KRT79', 'KRT80', 'KRT81', 'KRT8
2', 'KRT83', 'KRT84', 'KRT85', 'KRT86']
```

```
type_1 = ['KRT222', 'KRT9', 'KRT10', 'KRT12', 'KRT13', 'KRT14', 'KRT15', 'KRT16', 'KRT17', 'KRT18',
'KRT19', 'KRT20', 'KRT23', 'KRT24', 'KRT25', 'KRT26', 'KRT27', 'KRT28', 'KRT31', 'KRT32', 'KRT33A'
, 'KRT33B', 'KRT34', 'KRT35', 'KRT36', 'KRT37', 'KRT38', 'KRT39', 'KRT40']

for filename in glob.glob(path):
   with open(filename) as f:
      if f_name in type_1:
         type_1_seqs = f.read()
         type_1_seqs += "\n"
         temp_1 += type_1_seqs
      if f_name in type_2:
type_2_seqs = f.read()
type_2_seqs += "\n" temp_2 +=
type_2_seqs all = temp_1 + "\n" +
temp_2

with open ('type_1_KRTs.txt', 'w') as fp:
   fp.write(temp_1)
with open ('type_2_KRTs.txt', 'w') as fp:
   fp.write(temp_2)
```

For obtaining orthologous sequences from the agnatha genera under study (Petromyzon, Branchiostoma, Ciona, Strongylocentrotus and Hydra), the retrieved gnathostome sequences were used as probes to iteratively search the agnathan genomes available in NCBI RefSeq and GenBank (Benson et al. 2017), by applying reciprocal BLASTP and BLASTN (Altschul et al.
1990b) with default parameters; Script 4 was employed:

Script 4

```
# Importing required libraries
Entrez.email = "isiltakan@gmail.com "
# To increase API limits, turn to uncomment the below line and write your own API key.
# Entrez.api_key = 'your_api_key'
```

requests.packages.urllib3.disable_warnings(category=InsecureRequestWarning)

 # Hyperparameters used on the website
headers = {**the hyperparameters was removed due to the attribute of temporariness of this section'}

 # Definitions required to open the session def **init**():
ch = requests.Session() adapter = requests.adapters.HTTPAdapter(max_retries=30) ch.mount('https://', adapter)
ch.mount('http://', adapter)

cookies = requests.utils.cookiejar_from_dict(requests.utils.dict_from_cookiejar(ch.cookies))

return [ch, cookies]

 # An http function created not to crowd try-except blocks def **cli**(addr, ch=None, cookies=None): try:
grab = ch.get(addr, headers=headers, cookies=cookies, timeout=200, verify=False, allow_redirects=False) *# ncbi for scraping refseq*
except:
time.sleep(10) ch, cookies = init()
grab = ch.get(addr, headers=headers, cookies=cookies, timeout=200, verify=False, allow_redirects=False) *# ncbi for scraping refseq*
return grab
 # Get refseq from search term def **search**(term, ch=None, cookies=None):
 # regex pattern

```
pattern = r"<a href=\"https://www.ncbi.nlm.nih.gov/protein/(.*?)\" id=\"more_refseqproteins\"" grab =
cli('https://www.ncbi.nlm.nih.gov/protein/?term=' + urlencode(term), ch, cookies) grab = grab.text
grab = re.findall(pattern, grab) # scrape ids if grab: return (grab[0].split(','))[0] # return idlist and count
else:
return 'None'


# Download sequences by ID def d_ID(id, x):
handle = Entrez.efetch(db='protein', id=id, rettype='fasta', retmode='text') f = open('%s.txt' % x, 'w')
f.write(handle.read()) handle.close() f.close()


# search ids for Blast def main(genes, species):
idsForBlast = [] idsForExcel = [] ch, cookies = init() for gene in genes:
RefSeqID = gene + ' ' + species x = search(RefSeqID, ch, cookies) if x is not None:
idsForBlast.append(x) idsForExcel.append(gene + ';' + x)
else:
idsForExcel.append(gene + ';' + 'None')
with open('idsForBlast.txt', 'w') as f:
f.write('\n'.join(idsForBlast))
```

```
f.close()


with open('idsForExcel.txt', 'w') as f:
f.write('\n'.join(idsForExcel))
f.close()


genes = ['DCTN5', 'DHRS3', 'DISP1', 'DLG1', 'DNAH11', 'DNAH5', 'DNAI1', 'DPH1', 'DRC1', 'ECE1',
'E
GR2', 'EPS8L3','EYA1', 'FAM20C', 'FBLN1', 'FGFRL1', 'FLVCR1', 'FOXC1', 'FOXC2', 'FOXH1',
'FOXI
3', 'FUZ', 'GAD1', 'GLG1','GREM2', 'GSC', 'HAND1', 'HAND2', 'HAPLN1', 'HAT1', 'HHAT', 'HIPK2',
'H OXA2', 'HOXA3', 'IDUA', 'IFITM5','IFT140', 'IFT27', 'IKBKAP', 'IRF6', 'JAG1', 'JAG2', 'KAT14',
'KAT6
A', 'KAT6B', 'LIMK2', 'LMNA', 'LOXL3','LPAR4', 'LRRK1', 'LTBP1', 'LTBP3', 'MAP3K7', 'MAU2',
'ME
CP2', 'MEGF8', 'MMP14', 'NABP2', 'NDST1', 'NFATC1','NFIC', 'NIPBL', 'NPR2', 'OTULIN', 'OTX2',
'P
APPA2', 'PCSK6', 'PDGFRB', 'PDPK1', 'PDS5B', 'PDSS2', 'PGAP1','PHEX', 'PIGV', 'PITX1', 'PKD1',
'PK DCC', 'PLXND1', 'POU3F3', 'PRDM16', 'PRKRA', 'PRRX1', 'PTH1R', 'PTHLH',
'PTK7', 'RFNG', 'ROR2', 'RPGRIP1L', 'RSPO2', 'RYK', 'SC5D', 'SETD5', 'SH3BP2', 'SIRT1', 'SIX1',
'SLC
25A21','SLC26A2', 'SLC27A4', 'SLC39A13', 'SMAD2', 'SMG1', 'SMOC1', 'SOSTDC1', 'SRF', 'TBX1',
'T
BX15', 'TCOF1','TFAP2A', 'TGFBR2', 'TIPARP', 'TNFSF11', 'TRPS1', 'TTC28', 'VPS25', 'WDR19',
'XYL
T1', 'YBX1', 'ZMPSTE24','ZNF366', 'ACVR1', 'ACVR2A', 'ALKBH1', 'ALX1', 'AMBN', 'AMELY', 'AN
KH', 'APC', 'ARID5B', 'ARSB', 'ASXL1','ATMIN', 'ATR', 'BAZ1B', 'BCL11B', 'BNC2', 'BRD4',
'C5orf42'
, 'CACNA1S', 'CASK', 'CCDC39', 'CEP290', 'CHD1','CHRD', 'CHST11', 'CHUK', 'CNBP', 'COL11A1',
'C OL1A1', 'COL27A1', 'COL2A1', 'CREBBP', 'CTNNBIP1', 'CTSK','CYP26B1', 'CYP51A1',
'DAAM2', 'D
KK1', 'DLX2', 'DLX5', 'DLX6', 'FIGN', 'FOXD3', 'FOXF2', 'FRAS1','KCNJ2', 'KDM4B', 'MMP21',
'MNT'
, 'MSX1', 'MYH10', 'NOG', 'OSTM1', 'PAK1IP1', 'PAX1', 'PAX3', 'PAX6', 'PAX7','PAX9', 'PFAS',
'PITX
```

2', 'POSTN', 'ROBO1', 'SATB2', 'SCHIP1', 'SH3PXD2B', 'SKI', 'TMEM67', 'TNFRSF11B','TWSG1']

species = ['Mus musculus', 'Danio rerio', 'Gallus gallus', 'Homo sapiens', 'Callorhinchus milli', 'Xenopus la evis',
'Hydra vulgaris', 'Strongylocentrotus purpuratus', 'Branchiostoma floridae', 'Ciona intestinalis', 'Petromyzon marinus']

```
main(genes, 'Homo sapiens')
##########################################################################
# Processing the results found using the online blast tool ##########
# Download all results as multiple json files
# Get best hits
ids = []


for spe in species:
files = glob.glob('jsonFiles/' + spe + '/*.json')[0] ids.append(files.split('.')[0] + '_')
 for ct in range(len(species)):
sp = species[ct] p = [] for i in range(189):
with open('%s%s.json' % (ids[ct], str(i + 1)), 'r') as f:
a = json.load(f) x = {'q_id': a['BlastOutput2']['report']['results']['search']['query_id'],
    'q_title':                a['BlastOutput2']['report']['results']['search']['query_title']}                if
len(a['BlastOutput2']['report']['results']['search']['hits']):
x['title']    =    a['BlastOutput2']['report']['results']['search']['hits'][0]['description'][0]['title']    x['id']    =
a['BlastOutput2']['report']['results']['search']['hits'][0]['description'][0]['id']
else:
for n in range(len(p)):
if p[n]['q_id'] == x['q_id']:
x = p[n]
else:
x['title'] = 'None' x['id'] = 'None'
p.append(x)
with open('json_species/%s.json' % sp, 'w') as fs:
fs.write(json.dumps(p)) fs.close()
```

```
inputFile = open('json_species/%s.json' % sp) # open json file outputFile = open('csv_species/%s.csv'
% sp, 'w', newline='') # load csv file data = json.load(inputFile) # load json content inputFile.close() #
close the input file output = csv.writer(outputFile) # create a csv.write
output.writerow(list(data[0].keys())) # header row refs = []

 for row in data: row = list(row.values()) if len(row) > 3:
if row[3] != None or row[3] != 'None': if not 'pdb' in row[3]: refid = re.findall('\|(.*)\|', row[3]) if len(refid)
> 0:
row[3] = str(refid[0])
refs.append(row[3])
else:
refs.append('None')
output.writerow(row) # values row
with open('ids/%s.csv' % sp, 'w', newline='') as f:
writer = csv.writer(f) writer.writerow([sp]) for ref in refs:
writer.writerow([ref])
# Merge all csv files
# After control, the results by manually################################ dict = {}


with open('IDsTheAll', 'r', encoding='utf-8') as f:
f = f.read() rows = f.split('\n') for r in rows:
```

```
    rowInrows = r.split(',') if
    rowInrows[0] != '':
        dict[rowInrows[0]] = [] for
        i in rowInrows[1:]:
            if i != 'None' and i != '':
                dict[rowInrows[0]].append(i)

for gene in genes:
  d_ID(dict[gene], gene)
```

3.7.5. Multiple Sequence Alignment

Alignment of the full-length KRT amino acid sequences was performed using *bash* Script 5 with the MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/) multiple sequence alignment (MSA) tool. The amino acid sequences of the homologous jaw indispensable proteins were aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), version 1.2.4 (Sievers et al. 2011).

Script 5

```
files=( 'type_1_KRTs' 'type_2_KRTs' )

for i in ${files[@]} do muscle -in /PATH/"$i".txt -out
/PATH/OUTPUT/"$i"_fin.aln -clwstrict done
```

3.7.6. Phylogenetic Inference

Phylogenetic inference based on the trimmed 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 Python script (Script 6), whereas ML trees were constructed using the software package MEGA version 10. 1.

(Kumar et al. 2018) for the phylogenetic inference of keratins.

The aligned sequences of the homologous jaw indispensable proteins were provided as input to the software package MEGA version 10. 1 (Kumar et al. 2018) in order to perform phylogenetic analyses, by employing a neighbor-joining (NJ) and maximum likelihood (ML) method. Only those agnathan sequences that clustered with the known gnathostome sequences under study were considered as 'true orthologs' (Table 2).

3.7.6.1. Neighbor Joining Tree

Distance-based methods involve agglomerative hierarchical clustering, where a distance matrix of pairwise scores for all the aligned sequences are employed to generate a phylogenetic tree (Gascuel 1997). Python script 6 was implemented for NJ tree construction, and the required libraries (glob, os, re, Bio (Cock et al. 2009), matplotlib (Hunter 2007), seaborn (Waskom et al. 2017), NumPy (Harris et al. 2020)) were imported. Using *AlignIO*, aligned sequences (Script 5) were imported by Script 6.1.

Script 6

Script 6.1.

```
# Importing required libraries
# Reading aligned sequences
files_aln = ['FileName'] path
= 'PATH/FileName' for f in
files_aln:
   fin = path + f + '_fin'+ '.aln' aln =
   AlignIO.read(open(fin), "clustal")
```

NJ trees were constructed using the *Bio.Phylo.TreeConstruction* module.

The module *Phylo* from Bio (Cock et al. 2009) was used to obtain Newick and PhyloXML trees. PhyloXML format was selected for storing the style of the phylogenetic trees. Bootstrap analysis (n=100) was conducted to evaluate the statistical significance of the inferred trees. (Script 6.2)

Script 6.2.

```
   calculator = DistanceCalculator('identity')
   dm = calculator.get_distance(aln) constructor =
   DistanceTreeConstructor(calculator) trees =
   bootstrap_trees(aln, 100, constructor) s_tree =
   constructor.build_tree(aln) s_tree.rooted = True
   tree = [] tree = list(trees) support_tree =
   get_support(tree, tree) fig =
   Phylo.draw(support_tree)
Phylo.write(support_tree, " PATH/FileName%s.xml" % f, "phyloxml")
Phylo.write(support_tree, " PATH/FileName/%s.nwk" % f, "newick")
```

Script 6.3 was utilized for phylogenetic tree visualization. *Matplotlib* (Hunter 2007) *rc* and *pyplot* modules were applied to adjust figure parameters. Finally, the *Phylo* module was utilized to retrieve the final trees.

Script 6.3.

```
dict_color = {} def
get_label_color(leaf):
   dict_color[leaf.name] = 'red'
   return dict_color
path2 = " PATH/FileName/"
for f in files_aln:
```

```
base = path2 + f fin = base + '_bs.xml' newi = base + '_bs.nwk'
support_tree = Phylo.read(fin, "phyloxml") fig = Phylo.draw(support_tree) fig = plt.figure(figsize=(30,50), dpi=300) matplotlib.rc('font', size=10, 'lines', linewidth=1, color='darkgrey')
matplotlib.rcParams['text.color'] = 'red'
colorsL = {"KRT13":"#24C4CB","KRT2":"#520D3C","KRT3":"#502257","KRT4":"#FDA19D","KRT5":"#65F099","KRT6A":"#7671A2",
"KRT6B":"#AF40B2","KRT6C":"#20A95A","KRT7":"#14D1C8","KRT8":"#581E03",
"KRT71":"#53768C","KRT72":"#9927E9", "KRT73":"#CC5BAE","KRT74":"#64272E",
"KRT75":"#65915F","KRT76":"#605122", "KRT77":"#216F36","KRT78":"#54F22D",
"KRT79":"#FB8649","KRT80":"#4B5235", "KRT81":"#91E506","KRT82":"#8BFD05",
"KRT83":"#2B1A4E","KRT84":"#6FBBB6", "KRT85":"#DF1762","KRT86":"#2A03BC",
"KRT222":"#8DCAF8","KRT9":"#EC48B2", "KRT10":"#C9F54D","KRT12":"#CF5EBE",
"KRT1":"#238ACC","KRT14":"#3C8539", "KRT15":"#07A2C5","KRT16":"#73BB8C",
"KRT17":"#8BEAAA","KRT18":"#2DB304", "KRT19":"#BE1CAE","KRT20":"#EE3C8F",
"KRT23":"#69D565","KRT24":"#F3B98F", "KRT25":"#D9DA39","KRT26":"#FC581A",
"KRT27":"#0D78EE","KRT28":"#8A6CEE", "KRT31":"#3A8EC7","KRT32":"#B07EF3",
"KRT33A":"#F63302","KRT33B":"#9A9F86",
```

```
"KRT34":"#18B2FB","KRT35":"#2BD3C8","KRT36":"#7347FD","KRT37":"#AB71CC","KRT38":"#A
70242","KRT39":"#310304","KRT40":"#AEBF02"} axes = fig.add_subplot(1,1,1) plt.axis('off')
support_tree.root.color = (128, 128, 128) for k,clade in enumerate(support_tree.find_clades()):
if 'Inner' in clade.name:
clade.color = 'grey'
else: clname = clade.name.split('|')[0] try:
clade.color = colorsL[clname]
        except:
            clade.color = 'grey'
    clade.width = 1
  for n in support_tree.get_terminals():
    n.name = ( (n.name).split('|')[1] + ' ' + (n.name).split('|')[0] ) def
    labelColor(label):
    if (len(str(label).strip()) > 0): geneName =
      str(label).split('        ')[1].strip()        if
      (geneName in colorsL):
        return colorsL[geneName] return
      'red'
    else:
      return 'black'
  Phylo.draw( support_tree, label_func = lambda x:"" if 'Inner' in x.name else x, axes=axes,
label_colors = labelColor )
```

3.7.6.2. Maximum-likelihood Tree

Maximum-likelihood methods determine the tree topology and branch lengths that have the maximum likelihood of producing the observed data (Guindon et al. 2009). The maximum likelihood-based trees were inferred using the MEGA software package version 10. 1 (Kumar et al. 2018).

3.8. 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 (Tang et al. 2019) (http://gepia2.cancer-pku.cn/).

Bash script was used to manipulate the content of the dataset stored to create the cancer database.

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 log fold change $|\log2FC| \geq 2$ and FDR-adjusted p-value ≤ 0.05. *KRT* genes and their corresponding values were stored in JSON format (Addenda 8.1).

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)).

3.9. Cancer Database Construction

A cancer *database was* constructed based on the data downloaded from GEPIA2 to easily access, manage, and update the collection of the datasets, by utilizing MSSQL. This database has also been used for other research (Theofylaktou et al. 2021).

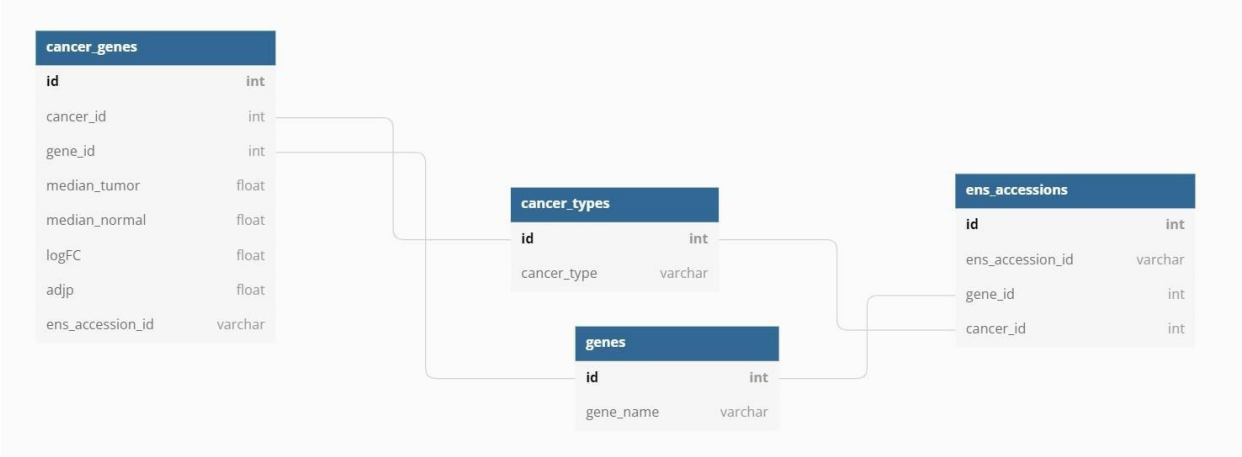


Figure 5: ER Diagram

The ER (entity relationship) diagram (figure 5) shows the database structure, and its table content. Table Scripts were implemented to create the tables of the database.

Table Scripts

```
USE [cancer_database]

GO

/****** Object: Table [dbo].[genes] ******/

SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[genes](
	[id] [int] IDENTITY(1,1) NOT NULL,
	[gene_name] [varchar](50) NOT NULL,
UNIQUE NONCLUSTERED
(	[gene_name] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON,
OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
GO
USE [cancer_database]
GO
```

```
/****** Object: Table [dbo].[ens_accessions] ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[ens_accessions](
	[id] [int] IDENTITY(1,1) NOT NULL,
	[ens_accession_id] [varchar](50) NOT NULL,
	[gene_id] [int] NOT NULL,
```

```
	[cancer_id] [int] NOT NULL
) ON [PRIMARY]
GO
USE [cancer_database]
GO
/****** Object:  Table [dbo].[cancers_genes]  ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[cancers_genes](
	[id] [int] IDENTITY(1,1) NOT NULL,
	[cancer_id] [int] NOT NULL,
	[gene_id] [int] NOT NULL,
	[median_tumor] [float] NULL,
	[median_normal] [float] NULL,
	[logFC] [float] NULL,
	[adjp] [float] NULL,
	[ens_accession_id] [int] NOT NULL
) ON [PRIMARY]
GO
USE [cancer_database]
GO
```

```
/****** Object:  Table [dbo].[cancer_types] ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[cancer_types](
	[id] [int] IDENTITY(1,1) NOT NULL,
	[cancer_type] [varchar](150) NOT NULL
) ON [PRIMARY]
```

```
GO
USE [cancer_database]
GO
```

The data were appended to the *sql* table by using Python scripts (Python SQL Script 1 and Python SQL Script 2).

Required libraries were imported (Python SQL Script 1: *glob, os, pyodbc, function and* Python SQL Script 2: *config, pandas*)).

Also, the Pyodbc library was implemented to connect to SQL server for the cancer database in Python SQL Script 1.

Python_SQL_Script_1

```
# Importing required libraries
# data directory project_path = os.path.dirname(os.path.abspath(__file__)) + '/data/'
# Database connection conn =
pyodbc.connect('Driver={SQL Server};'
               'Server=DESKTOP-2TUV3I4;'
               'Database=cancerDatabase;'
               'Trusted_Connection=yes;')
```

Python SQL Script 2 was implemented to import the data attributes such as "gene_name", "gene_id", "median_tumor", "median_normal", "logFC", "adjp" into database.

Python_SQL_Script_2

```
# importing required libraries cursor =
conn.cursor() file_list =
glob.glob(project_path + '*.tsv') for file in
file_list:
    cancer_name = grab_file_name(file)
```

```
df = pd.read_csv(file, sep="\t")                                    "logFC", "adjp"]
header = ["gene_name", "gene_id", "median_tumor", "median_normal",
for row in range(len(df)):
    dicts = {} for head
    in header:
        if pd.isnull(df[head].iloc[row]) == True:
            continue
        else:
            dicts[head] = df[head].iloc[row]
    sql_insert(cursor, cancer_name, dicts) conn.commit()
```

View_Script was created to obtain full information of all tables.

View_Script

```
/****** Object: View [dbo].[cancers_genes_wth_name] ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON GO
create view [dbo].[cancers_genes_wth_name] as
select g.gene_name, ct.cancer_type, cs.logFC
from cancers_genes cs, cancer_types ct, genes g
where cs.cancer_id=ct.id and cs.gene_id=g.id
GO
```

Finally, Query 1 was used to obtain the corresponding KRTs, cancer types and their expression status. Table 5 shows the query results.

Query 1

```
/****** ******/ select * from cancers_genes_wth_name
where gene_name in (
'KRT1', 'KRT2', 'KRT3', 'KRT4', 'KRT5', 'KRT6A', 'KRT6B', 'KRT6C', 'KRT7', 'KRT8',
'KRT71', 'KRT72', 'KRT73', 'KRT74', 'KRT75','KRT76', 'KRT77', 'KRT78', 'KRT79', 'KRT80',
'KRT81', 'KRT82', 'KRT83', 'KRT84', 'KRT85', 'KRT86', 'KRT222', 'KRT9', 'KRT10', 'KRT12',
'KRT13', 'KRT14', 'KRT15', 'KRT16', 'KRT17', 'KRT18', 'KRT19', 'KRT20', 'KRT23', 'KRT24',
'KRT25', 'KRT26', 'KRT27', 'KRT28', 'KRT31', 'KRT32', 'KRT33A', KRT33B','KRT34','KRT35',
'KRT36', 'KRT37', 'KRT38', 'KRT39', 'KRT40'
)
```

Table 5: *KRT* genes, cancer types and gene expression status.

Gene name	TCGA cancer type	Status
KRT1	DLBC	Down
	ESCA	
	SKCM	
KRT10	SKCM	Down
	LAML	
	THYM	Up
KRT13	LUSC	Up
	UCEC	Down
	UCS	
	LUAD	
	ESCA	
	HNSC	
	COAD	
	ACC	
	SKCM	
	STAD	
	TGCT	
	READ	
	PRAD	
	OV	
KRT14	SKCM	Down
	BRCA	
	BLCA	Up
	LUSC	

	THYM	

KRT15	THYM	Up
	LUSC	
	CESC	
	PAAD	
	BRCA	Down
	SKCM	
	TGCT	
	ESCA	
	PRAD	
KRT16	SKCM	Down
	PAAD	Up
	CESC	
	LUSC	
KRT17	LUSC	Up
	HNSC	
	ESCA	
	COAD	
	UCEC	
	BLCA	
	CESC	
	STAD	
	THYM	
	UCS	
	PAAD	
	OV	
	READ	
	SKCM	Down
	TGCT	
	BRCA	
KRT18	UCS	Up
	THYM	
	UCEC	
	READ	
	OV	
KRT18	TGCT	Up
	STAD	
	CESC	
	BRCA	
	COAD	
	ESCA	
	SKCM	

KRT19	LAML	Down
	KICH	
	CESC	Up
	LUSC	
	COAD	

	THCA	
	TGCT	
	OV	
	PAAD	
	READ	
	THYM	
	UCEC	
KRT2	SKCM	Down
KRT20	READ	Up
	COAD	
KRT222	LGG	Down
	GBM	
KRT23	HNSC	Down
	DLBC	Down
	COAD	Up
	SKCM	Down
	READ	Up
	PRAD	Down
	PAAD	Up
	OV	Up
	THYM	Down
	TGCT	Down
	UCEC	Up
KRT24	ESCA	Down
KRT31	ESCA	Down
	SKCM	
KRT32	ESCA	Down
KRT33A	TGCT	Down
KRT4	ESCA	Down
	HNSC	
	LUAD	
	LUSC	

KRT5	THYM	Up
	CESC	
	BRCA	Down
	SKCM	
	ESCA	
KRT6A	SKCM	Down
	THYM	Up
	PAAD	
	CESC	

	LUSC	
KRT6B	LUSC	Up
	CESC	
	PAAD	
	SKCM	Down
KRT6C	LUSC	Up
	ESCA	Down
KRT7	KIRC	Down
	LUSC	
	SKCM	
	BLCA	Up
	STAD	
	THYM	
	PAAD	
	OV	
	CESC	
	UCS	
	UCEC	
KRT72	SKCM	Down
	TGCT	
KRT73	SKCM	Down
KRT75	HNSC	Up
KRT77	SKCM	Down
KRT78	SKCM	Down
	HNSC	

KRT78	ESCA	Down
KRT79	SKCM	Down
KRT8	TGCT	Up
	STAD	
	THYM	
	READ	
	OV	
	PAAD	
	UCS	
	BRCA	
	CESC	
	COAD	
	ESCA	
	UCEC	
	LAML	Down
KRT80	COAD	Up
	CESC	
	LUAD	
	OV	
	READ	
	STAD	
	THCA	
	SKCM	Down
KRT86	TGCT	Down

3.10. Protein-Protein Interactions (PPI)

Networks of the protein products of the identified cancer DEGs were constructed based on association data derived from various experimental studies. STRING (Search Tool for the Retrieval of Interacting Genes) (Szklarczyk et al. 2017), a database of experimental and predicted, direct (physical) or indirect (functional), associations among genes/protein, derived from different sources such as high-throughput experiments, biomedical text mining, co- expression, gene fusion were used to find any physical/functional associations.

The gene/protein associations detected in STRING were used as input to Cytoscape (Shannon et al. 2003) to construct networks. Functional modules were recognized in the generated network

by employing clusterMaker (Morris et al. 2011), and ‘hubs’ (highly connected nodes) in the functional modules were identified through CytoHubba (Chin et al. 2014).

3.11. 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 NLP methods.

3.11.1. Acquisition of data

The scientific literature database MEDLINE/PubMed (https://pubmed.ncbi.nlm.nih.gov/) was searched with artificial intelligence techniques thoroughly 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 process (Script 7).

Script 7

```
with open('PATH\file', 'r') as f:
f = f.read()
    splitText = f.split('\n\n\n')
    titleAbstDict = {}
    for txt in splitText:
        titleAbstDict[txt.split('\n\n')[1]] = ''.join(txt.split('\n\n')[4].split('\n'))
   with open('PATH\fileOut', 'w') as fout:
        fout.write(str(titleAbstDict))
```

Of the 4950 candidate articles, 1377 met the inclusion criteria:

(a) written in English,

(b) including an abstract, (c) containing adequate text for processing.

Finally, the Python regex library was utilized to distribute each type of keratin into different groups. We only used the keratin types that have an adequate quantity of relevant articles. The counts of relevant articles are shown in Table 6.

Table 6: Counts of KRT-relevant articles. Counts of included articles met the inclusion criteria are shown in yellow background.

Gene Name	Counts of relevant articles
KRT15	**39**
KRT38	1
KRT36	1
KRT37	0
KRT32	1
KRT20	**222**
KRT28	2
KRT39	0
KRT35	1
KRT18	**106**
KRT16	**26**
KRT23	8
KRT31	0
KRT19	**673**
KRT9	9
KRT24	1
KRT27	1
KRT40	1
KRT34	7
KRT25	3
KRT26	0
KRT33A	0
KRT33B	0
KRT17	**81**
KRT14	**166**
KRT13	**28**

3.11.2. Text Processing

The natural language toolkit (NLTK: https://www.nltk.org/) and spaCy (https://spacy.io/), freely accessible Python libraries, were both used for text processing including tokenization, parsing, lemmatizing and stemming (Script 8).

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 keratin types.

Script 8

```
# Importing required libraries nltk_words =
list(stopwords.words('english')) nlp =
spacy.load('en_core_sci_lg') stop =
set(nltk_words)
exclude = set(string.punctuation)
lemmatizer = WordNetLemmatizer()
def clean(doc):
   stop_free = " ".join([i for i in doc.lower().split() if i not in stop]) punc_free =
   "".join(ch for ch in stop_free if ch not in exclude) normalized = "
   ".join(lemmatizer.lemmatize(word) for word in punc_free.split()) return
   normalized
for name in glob.glob('/PATH'):
   fileName = os.path.basename(name).split()[0]
   f = open (name, "r", encoding='utf-8') dictFile
   = ast.literal_eval(f.read()) dictKeywords = {
   "PCNA" : "proliferating cell nuclear antigen",
   "MCM3" : "minichromosome maintenance protein 3",
   "ER" : "estrogen receptor",
   "PR" : "progesterone receptor",
   "EMT" : "epithelial-mesenchymal transition",
   "CKs" : "cytokeratin",
   "GC" : "gastric cancer",
   "SCC" : "squamous cell carcinoma",
   "OSCC" : "oral squamous cell carcinoma",
   "TNF-α" : "Tumor necrosis factor-α",
   "PDXs" : "patient-derived xenografts",
   "CRC" : "colorectal cancer",
```

```
"tKRT18" : "total soluble KRT18",
"nKRT18" : "necrosis related KRT18",
"CRP" : "C-reactive protein",
"OC" : "oblongifolin C",
"TRAIL" : "TNF-related apoptosis inducing ligand",
"DRs" : "death receptor",
"rhTRAIL" : "recombinant human TRAIL",
"AZA" : "agent azacitidine",
"231Br" : "brain colonizing counterpart",
"RS" : "Raman spectroscopy",
"ERalpha" : "estrogen receptor alpha",
"PGR" : "progesterone receptor",
"HER2" : "epidermal growth factor receptor 2",
"CSC" : "cancer stem cell",
"ECM" : "extracellular matrix"}
data = [] for k,v in
dictFile.items():
    doc = nlp(clean(v)) termList = list([d.text.strip() for
    d in list(doc.ents)])
   for i in str(doc).split(' '):
       if i.upper() in (dictKeywords.keys()):
           if i not in termList:
               termList.append(dictKeywords[i.upper()])
      data.append(termList)
with open('PATH/stem_words_{id}'.format(id=fileName) , 'w') as fout:

    fout.write(str(data))
```

3.11.3. Word Embeddings

Word2Vec embeddings module was implemented in the Python library Gensim

(https://pypi.org/project/gensim/) to train word vectors of pre-processed text (Script 9). 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 gensim Word2Vec model was used. We used continuous bag-of-words (CBOW) algorithm which learns representations of words to serve for predicting other words in the sentence. We only vectorized words that appeared more than five times (minimum frequency threshold). Iteration was set at 30 (epochs). Increased number of iterations enhanced the performance of Word2Vec since the algorithm re-learned the relation between words.

Parameter of Word2Vec		
min_count	5	minimum frequency threshold
vector_size	300	layer size
epochs	30	number of iterations of learning
sg	0	the training algorithm is CBOW

Script 9

```
# Importing required libraries def
modelRun(processed_text):
      cores = multiprocessing.cpu_count()
   w2v_model = Word2Vec( min_count=5, window=5, vector_size=300,
                  sample=6e-5,
                  min_alpha=0.0007,
                  negative=20,
                  workers=cores - 1,
                  seed=1    )
   w2v_model.build_vocab( processed_text,
                    progress_per=10000 )
```

```
    w2v_model.train( processed_text,
                    total_examples=w2v_model.corpus_count,
                    epochs=30, report_delay=1 )
    w2v_model.init_sims(replace=True)  word_vectors  =
    w2v_model.wv                                    return
    word_vectors.most_similar(fileName, topn=50)
for name in glob.glob('PATH/file_name'):
    fileName = (os.path.basename(name).split('_')[2]).lower()
    Arr = [] f = open (name, "r", encoding='utf-8')
  processed_text = ast.literal_eval(f.read()) if
  len(processed_text) < 10: continue
    for i in range(1):
        thwords = modelRun(processed_text) for
        th in thwords:
            Arr.append(th[0]) Arr.append(fileName) words = list(set(Arr))
    with open('PATH/file_name_{id}'.format(id=fileName), 'w') as fout:
        fout.write(str(words))
```

3.11.4. Data Visualization

A list of all term (word)-to-term (word) distances was retrieved. To compute the similarity distances between associated keratin gene and words related, the *Word2Vec.most_similar* function in Gensim Word2Vec model was applied (Script 10).

- 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 threshold is 10) were used to generate networks which were manipulated and visualized with Cytoscape (Shannon et al. 2003).

Script 10

```
# Importing required libraries
# Import processed text and corresponding word list
w2v_model.build_vocab( processed_text,
progress_per=10000)
   w2v_model.train( processed_text,
                total_examples=w2v_model.corpus_count,
                epochs=30, report_delay=1 )
   w2v_model.init_sims(replace=True)
   word_vectors = w2v_model.wv conns
   = len(words) * 2
   ## finding distances using gensim to obtain the similarity between the krt and words
      for i1,word1 in enumerate(words): if word1 == fileName: continue
      cosine_similarity = word_vectors.similarity(word1,fileName) cosine_distance = 1
      - cosine_similarity dist = (word1, fileName, cosine_distance) graphText += word1
      + ';' +
                  fileName + ';' +
                  str(cosine_distance)  +  ';'  +
                  str(cosine_similarity) + ';' +
                  str(cosine_similarity) + '\n'
      with open('PATH/file_name_{id}'.format(id=fileName), 'w') as f:
         f.write(graphText)
```

3.12. Limitations of the Study

Due to the computationally intensive data processing, the computation time was excessively extended.

3.13. Ethics Committee Approval

Izmir Biomedicine and Genome Center Noninvasive Researches Ethics Committee Approval carrying protocol number “2020-045” was granted for this thesis on the 25th December 2020.

4. RESULTS

4.1. Phylogenetic analyses

Phylogenetic trees of the KRT and jaw indispensable protein sequences were generated by employing the neighbor joining (NJ) and the maximum likelihood (ML) methods.

4.1.1. Phylogenetic analysis of Keratins

Phylogenetic trees of KRT protein sequences were generated using both the NJ and ML methods. Several major distinct (color-coded) clades were identified in all trees (Figures 6-9).

4.1.1.1. Keratins Type I

The two phylogenetic trees (Figures 6 and 7), reconstructed based on Keratins Type I proteins, using both methods respectively, are congruent with similar topologies.

The phylogenies suggest that, apart from the ‘‘conventional’’ *KRT* family members, two ‘orphan’ *KRTs* are present in zebrafish. The zebrafish *KRT* orphans form their own distinct clades which arise from the basal node with high bootstrap values, leading to the suggestion that they are the members of the *KRT* family that diverged earliest (‘‘*proto-KRTs*’’) before the emergence of Amphibia 330 Mya. One plausible explanation is that the ancestor of the zebrafish orphan *KRTs*, a *proto-orphan KRT* emerged in an ancestor of Teleostei and a series lineage-specific gene duplication events gave rise to *KRT15* and *KRT17* that can be 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, suggesting it first emerged in an ancestor of Euteleostomi. Interestingly, the branches of the *KRT18* clade are exceptionally long, suggesting that the *KRT17* members evolved independently and more rapidly compared to the other *KRTs*.

The largest clade of the tree represents the genes *KRT31*, *KRT32*, *KRT33*, *KRT34*, *KRT35*, *KRT36*, *KRT37*, *KRT38*, *KRT39* and *KRT40* which were identified exclusively in Eutheria i.e. placental mammals. *KRT37* and *KRT38* were probably the first to emerge through a duplication

event and appear to have strong similarity. The next genes that seem to arise during the course of evolution are *KRT33* and *KRT34* along with *KRT31* that appear to be sister groups, suggesting that probably two more duplication events occurred. Later in mammalian evolution, most likely through yet another duplication event, the sister clades of *KRT40* and *KRT32* emerged, followed by the emergence of *KRT35*, *KRT36* and finally *KRT39*. The phylogenies suggest that a series of nine mammalian-specific tandem duplication events took place, in an early eutherian mammal that apparently lived before the Laurasiatheria – Euarchontoglires divergence, yielding those ten paralogs.

Another set of sister clades is formed by the genes *KRT20* and *KRT23*. *KRT23* seems to be the older of the two since it was first detected in chicken and thus is of Amniota origin, while *KRT20* was detected exclusively in Eutheria. More precisely, *KRT20* appears to be the product of a *KRT23* duplication which has occurred in a eutherian ancestor presumably after the Mammalia-Sauropsida split.

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

A sister group to the *KRT14/16/17* group, is consisted of the *KRT19* homologs. The two groups appear to have strong homology, as confirmed by relatively high bootstrap values. *KRT19* was also first identified in chicken and thus most likely first appeared 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 who lived before the divergence of Mammalia and Sauropsida.

A sister group to the *KRT14/16/17* - *KRT19* group, appears to be *KRT13* and *KRT15*. Both of those genes are detected exclusively in Eutheria, leading to the suggestion that they might have

emerged after a series of species-specific duplication events that yielded those mammalian paralogs in an ancestor of Eutheria who lived after the Mammalia-Sauropsida split.

Another gene that was first detected in Amniota is *KRT9*. It 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. On the contrary, as suggested by the topology and the high bootstrap values, *KRT9* appears to have strong homology with *KRT10*, which is found only in Eutheria, consequently, triggering the speculation that *KRT10* emerged through a mammalian-specific duplication event of *KRT9*, most likely in a eutherian ancestor that lived after the divergence of Mammalia and Sauropsida.

Finally, *KRT12*, *KRT24*, *KRT25*, *KRT26*, *KRT27* and *KRT28* are restricted to Mammalia. *KRT12* is most likely the first, out 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 in an early eutherian mammal that lived before the Laurasiatheria – Euarchontoglires divergence.

4.1.1.2. Keratins Type II

On the basis of the reconstructed phylogenies depicted in Figures 8 and 9, *KRT222* is the most 'ancient' clade, basal to the KRT tree, distantly related to the other KRTs; the KRT222 orthologs in the contemporary genomes are probably the products of a series of duplications of an ancestral *KRT222* gene in Teleosts (zebrafish). The genes *KRT71/72/73/78*, *KRT81/82/85/86*, and *KRT1/2/77*, which form highly-supported distinct clades in both trees, are restricted to placental mammals, suggesting that a series of Eutherian-specific gene duplication events might have given rise to the corresponding orthologs. The gene *KRT5* was detected exclusively in Amniota. Moreover, the zebrafish genes *KRT4* and *KRT5* appear to form a separate branch in both trees, suggesting products of a teleost-specific duplication.

The *Gallus gallus* genes *KRT6A* and *KRT75* appear to be basal to a clade which includes

Eutherian *KRT* genes. It is intriguing to speculate that a duplication of a *KRT6A/75* ancestral gene might have existed in Amniota which after a series of species-specific duplications gave rise to the Eutherian *KRT5*, *KRT6A/B/C* and *KRT75* orthologs.

The *KRT80* genes form a distinct clade, distantly related to the other KRTs in the tree, suggesting that an ancestral *KRT80* gene in amniota gave rise to the *KRT80* orthologues which evolved more rapidly compared to the fellow Type II keratins.

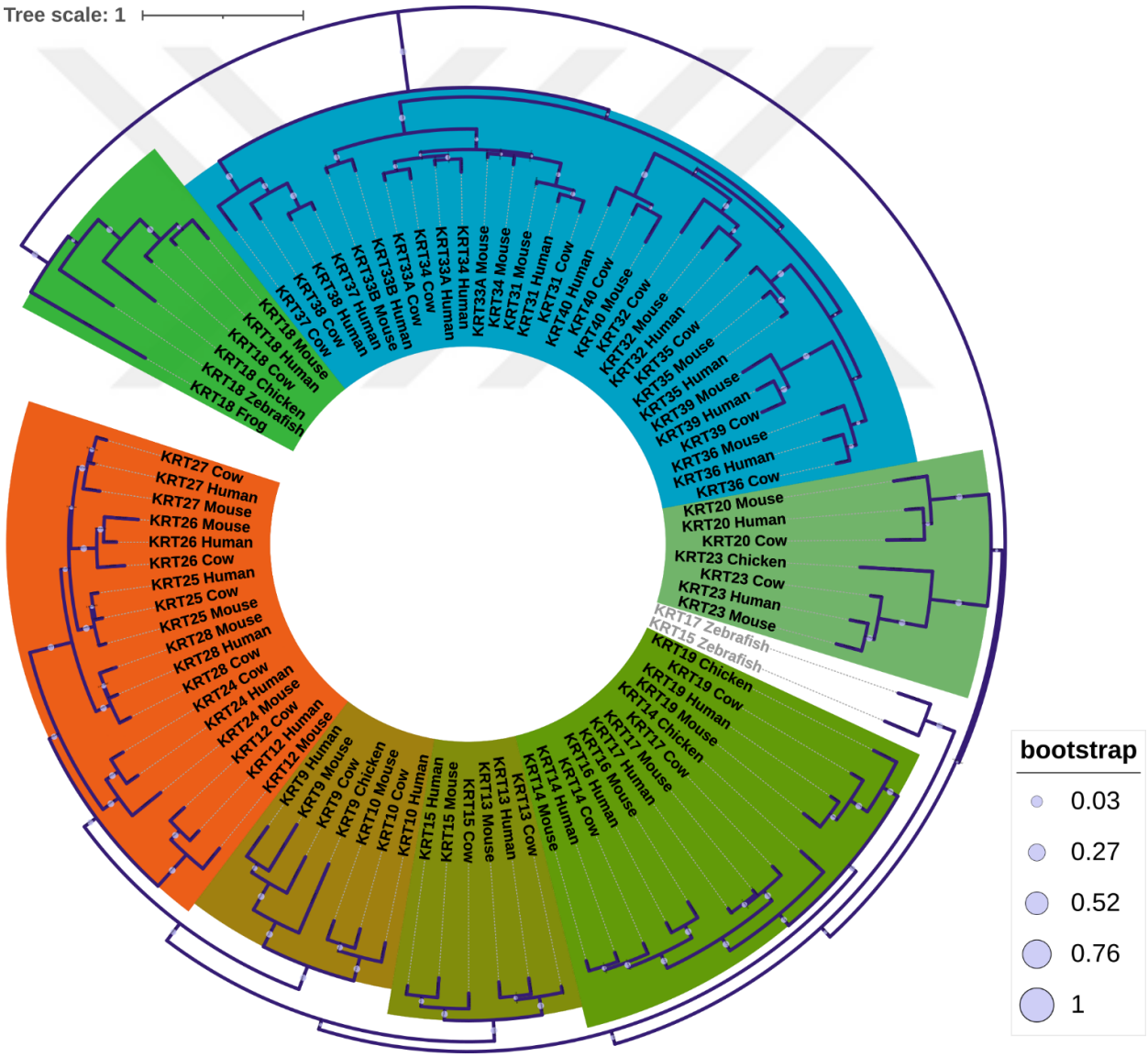


Figure 6: ML radial cladogram of KRT protein sequences. The sequences are represented by the species name and the KRT names (visualization with iTOL).

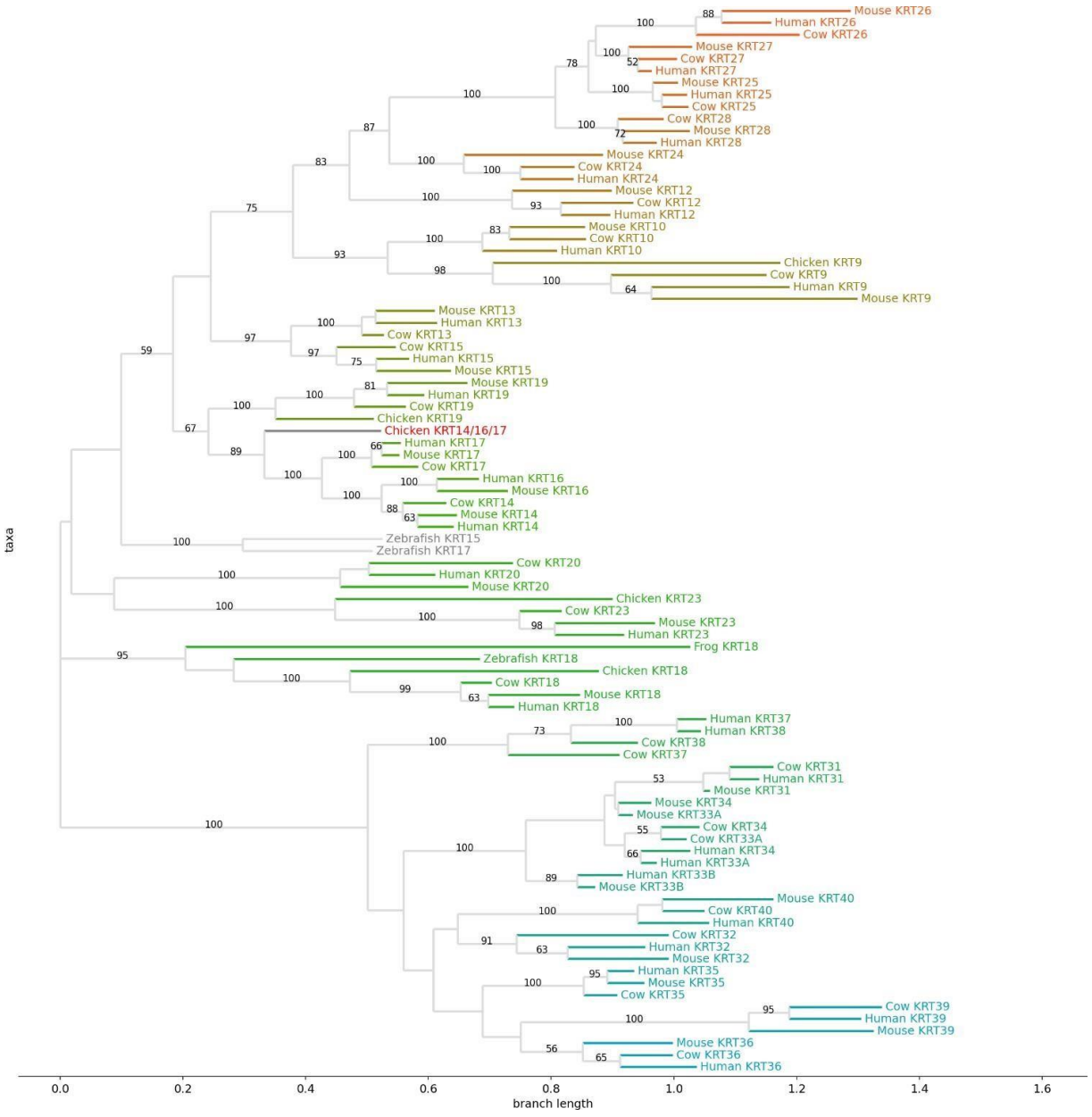


Figure 7: NJ phylogram of representative KRT type I proteins. Bootstrap values (>50) are shown at the nodes. The branch length at the bottom indicates the length of amino acid substitutions per position.

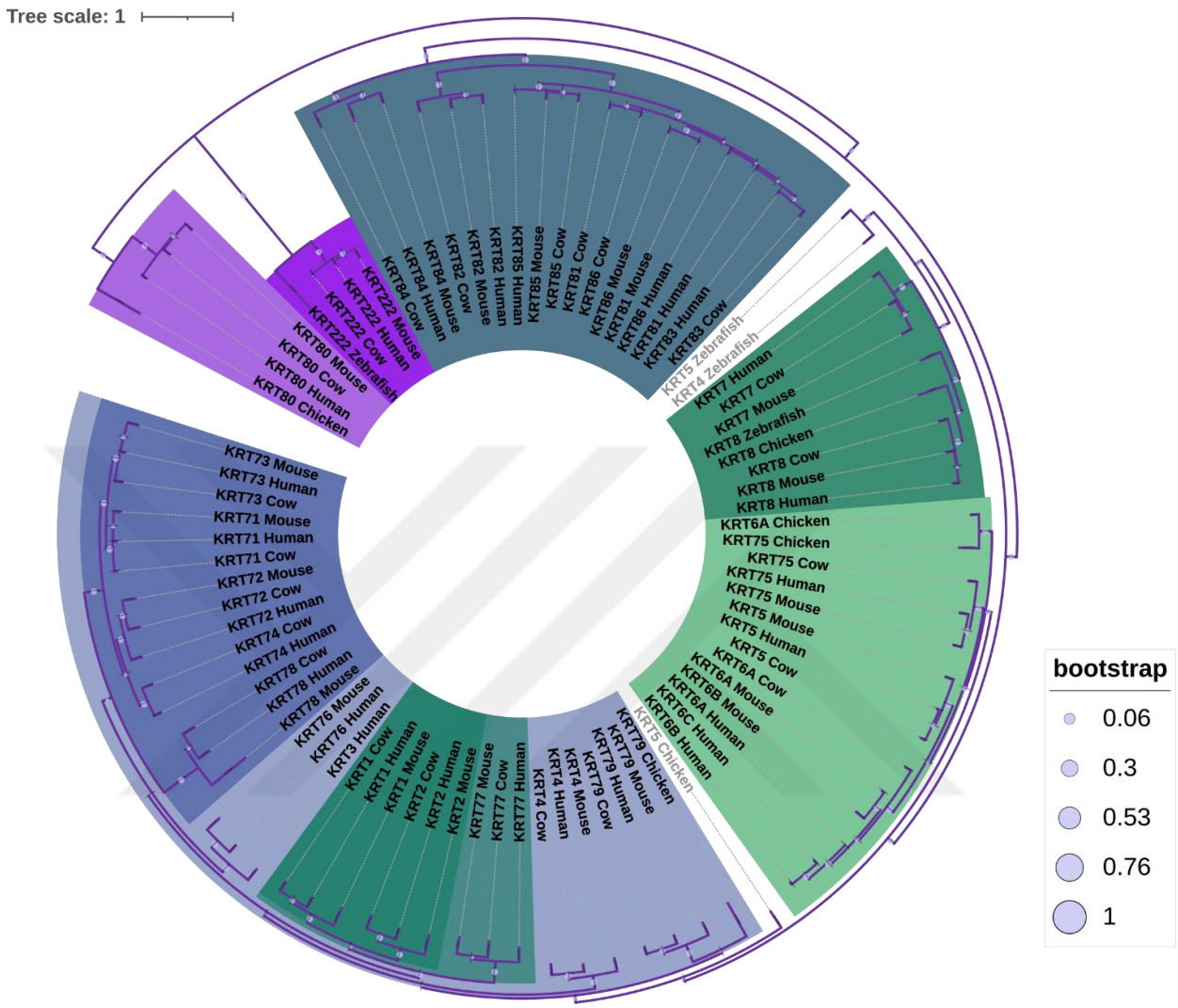


Figure 8: ML radial cladogram of KRT type II protein sequences. The sequences are represented by the species name and the KRT protein names (visualization with iTOL).

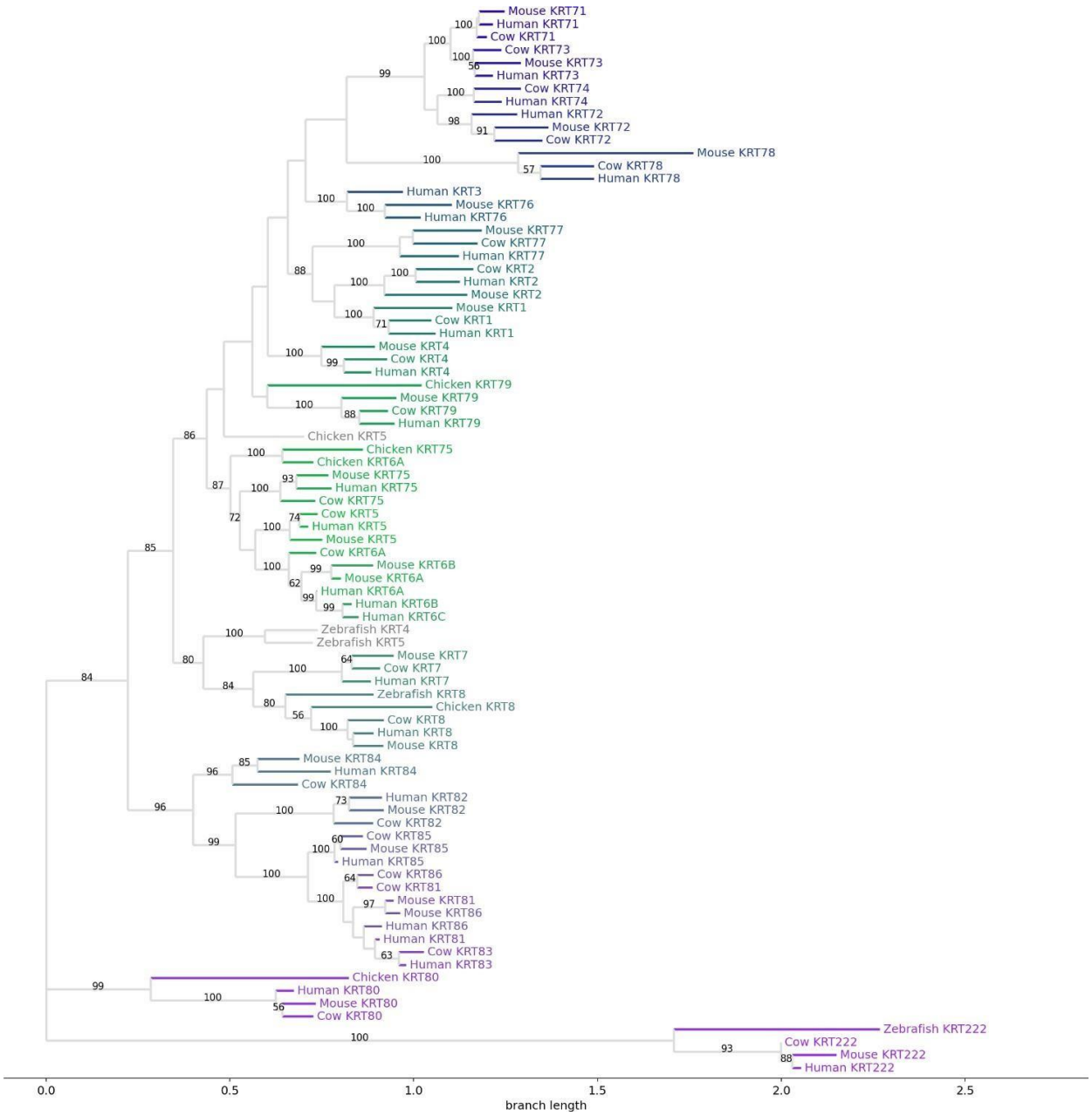


Figure 9: NJ phylogram of representative KRT type II proteins. Bootstrap values (>50) are shown at the nodes. The branch length at the bottom indicates the length of amino acid substitutions per position.

4.1.2. Phylogenetic analysis of Jaw Indispensable Genes

A characteristic example of gnathostome-specific genes is the well-studied *TP53* gene family, where distinct orthologs of the three members of this family (i.e., TP53, TP63 and TP73) were found exclusively in gnathostome species. The agnathan p53 homologs occupy a basal position in this phylogenetic tree, suggesting that a primordial *TP53/63/73* gene might have existed in jawless species which after a series of duplications gave rise to the functionally divergent *TP53*, *TP63* and *TP73* genes in jaw-bearing vertebrates.

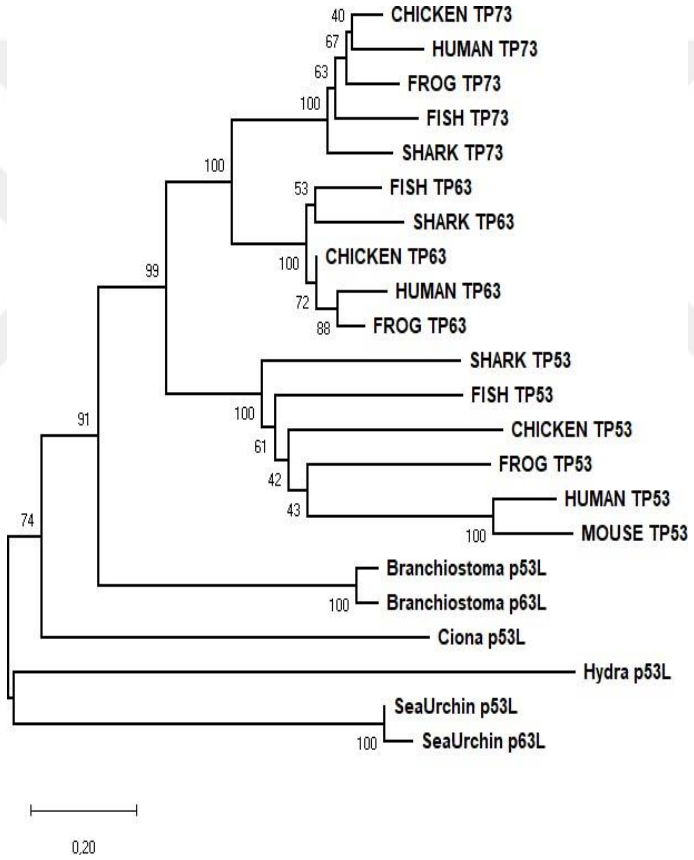


Figure 10: ML-based tree of ther 'true ortholog' TP53. Only agnathan sequences that clustered with the known gnathostome sequences under study were considered as 'true orthologues'. The bootstrap values show the statistical significance of the inferred tree. Numbers above branches are bootstrap values. Note that some nodes with bootstrap values of 100%, indicate strong support for these nodes, whereas other nodes have much weaker support.

4.2. Differential *KRT* gene expression

KRT genes, including the phylogenetically 'older' *KRT13*, *KRT15*, *KRT19*, *KRT23* and *KRT222*, are shown to be differentially expressed in multiple types of cancers. The phylogenetically conserved *KRT17* and *KRT19* were shown (Figure 11) to be differentially expressed in multiple types of cancers. This observation suggests that evolutionary pressure might have been exerted upon those genes (and their protein products) to remain unaltered, at least, in cancer-bearing species.

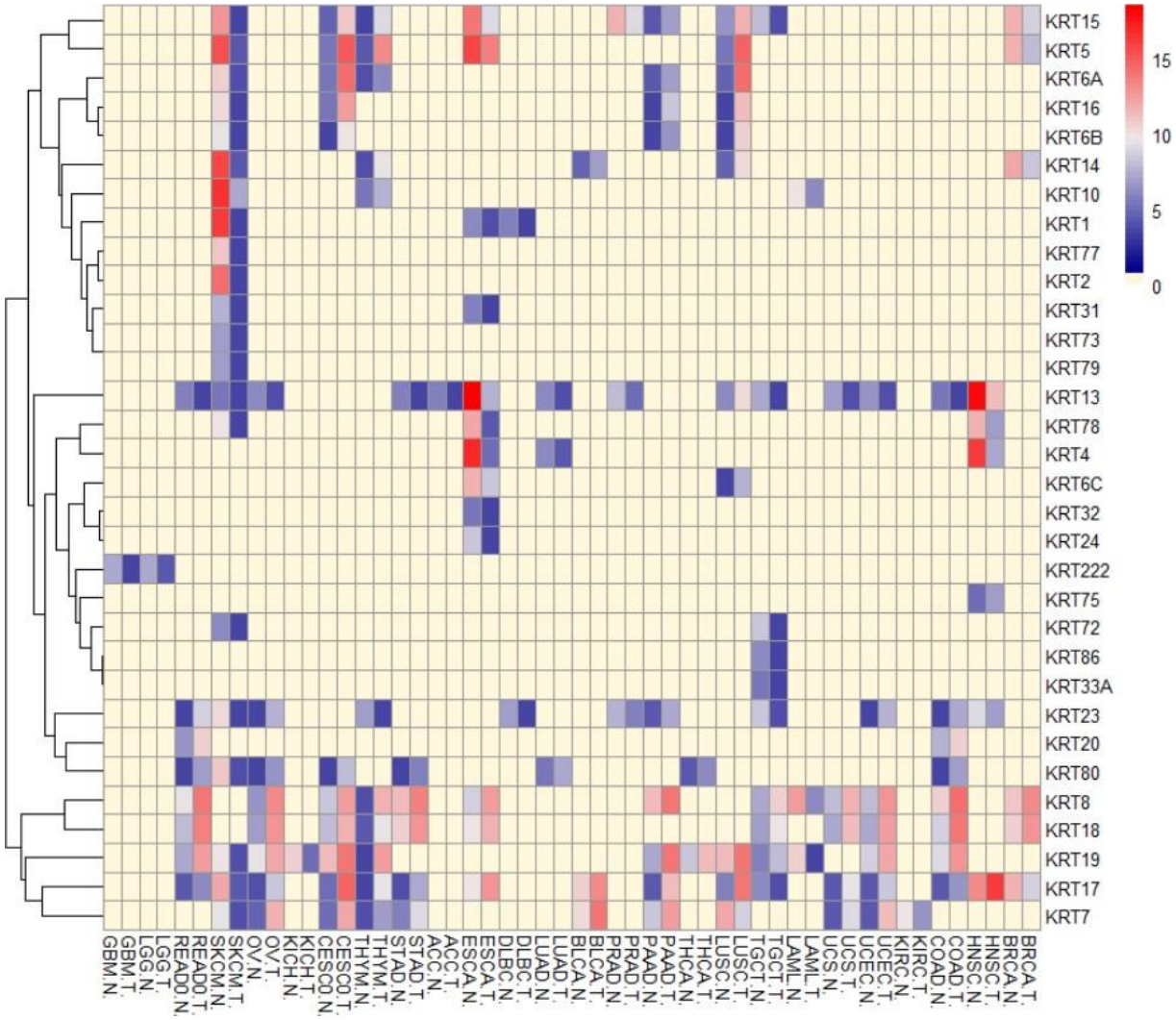


Figure 11: Heatmap of differentially expressed *KRT* genes based on data derived from the cancer database (Red: Up-regulated, Blue: Down-regulated, Yellow: No data).

4.3. Network Analysis

4.3.1. Protein-Protein Interactions of Keratins

The protein products of the differentially expressed KRTs, including the 'older' ones, form a highly interconnected network, suggesting functional or physical association among them.

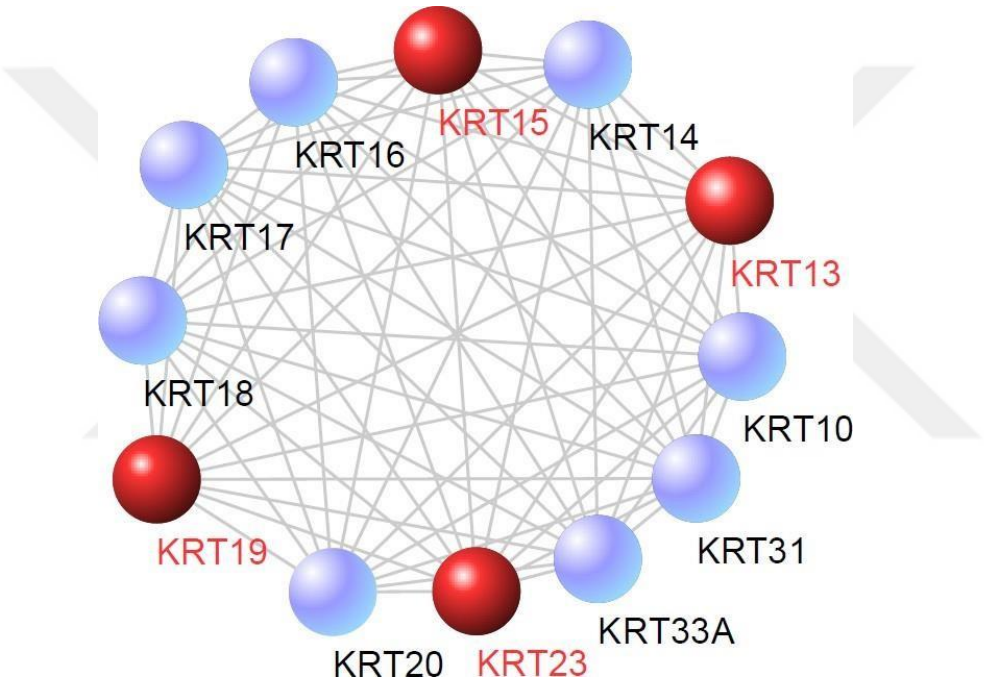


Figure 12: *KRT13, KRT15, KRT19 and KRT23* form a highly interconnected network.

4.3.2. of Keratin-Cancer Associations

A bipartite network was constructed displaying TCGA-derived cancer-*KRT* associations by using information from the cancer database developed in this study. The size of the nodes is proportional to their degree of connectivity. The network is highly interconnected, suggesting that functional or physical association among diverse cancer types and both types of *KRT* genes.

Network

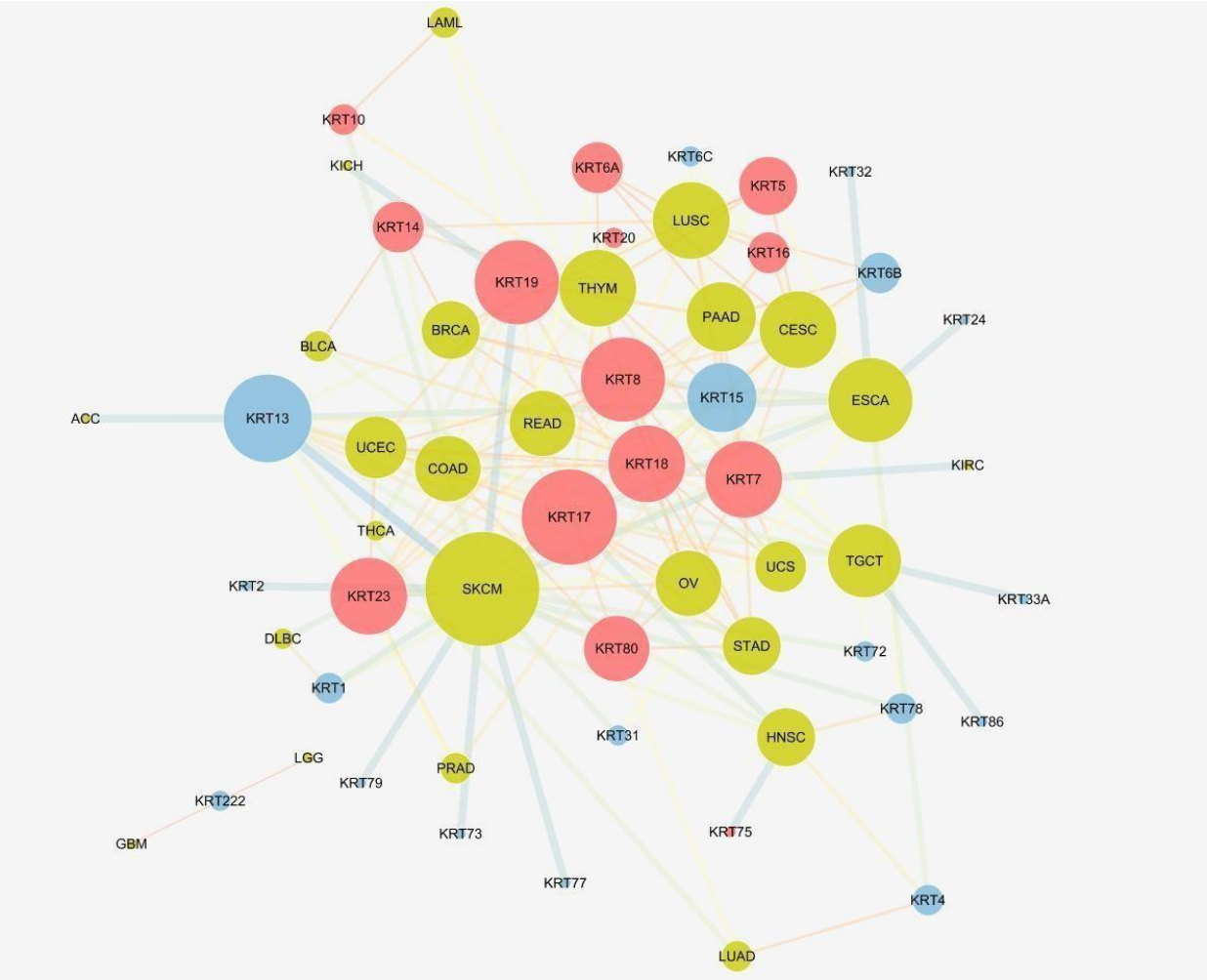


Figure 13: Keratins are linked to multiple cancers.

4.3.3. Analysis of Jaw Indispensable Genes

Our phylogenetic analysis revealed that 138 genes presumably originated in jawed vertebrates, since no ‘true orthologues’ of their corresponding proteins could be detected in jawless species after extensive database searching. In contrast, the other 115 genes are of pre- gnathostome

Network

origin (Addenda 8.2). Of note, these 253 JIG factors appear to form a highly interconnected network (Figure 13).

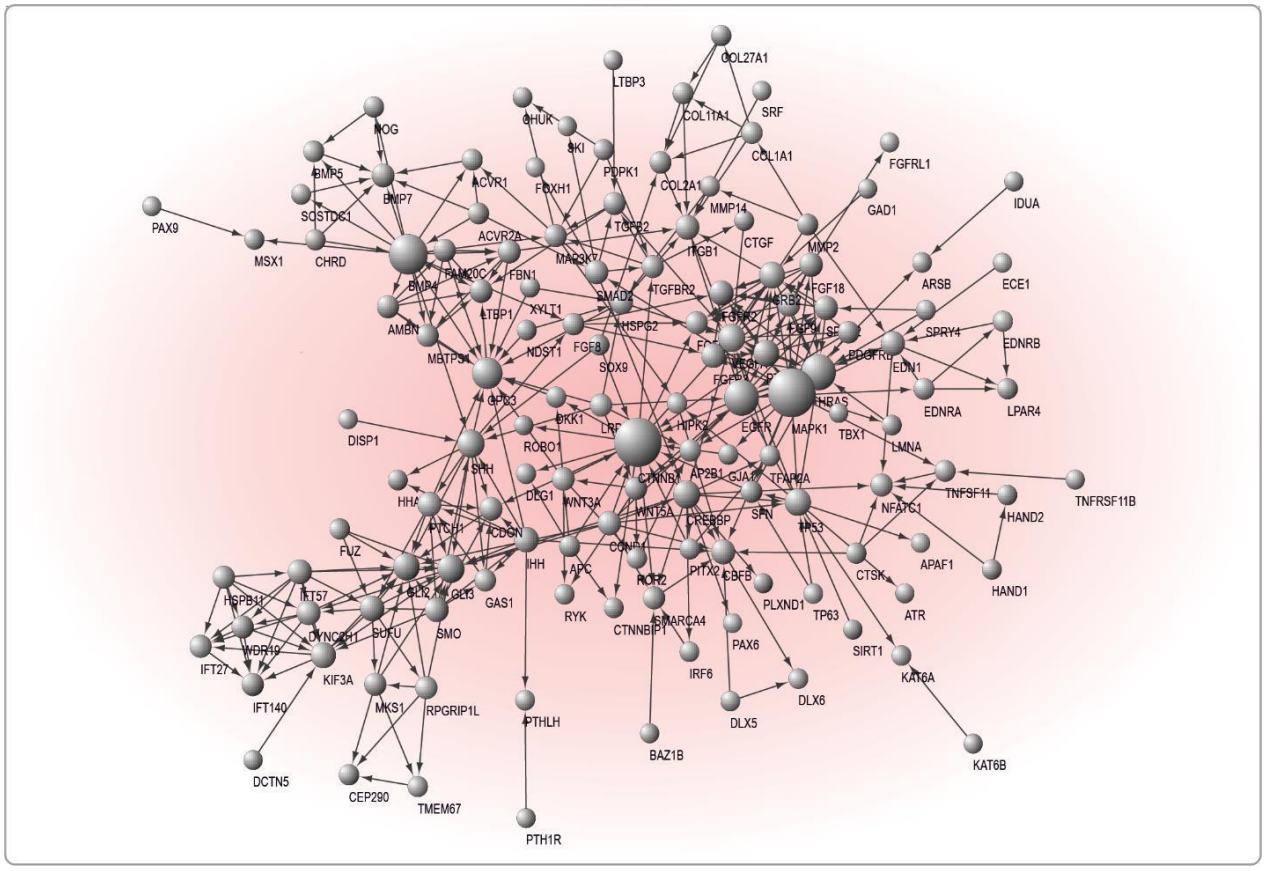


Figure 14: Interaction network of JIGs. A total of 173 of 305 genes/proteins are highly interconnected. The nodes represent genes/proteins and the connecting edges functional associations. Node sizes reflect node degree. Interaction confidence score ≥ 0.9.

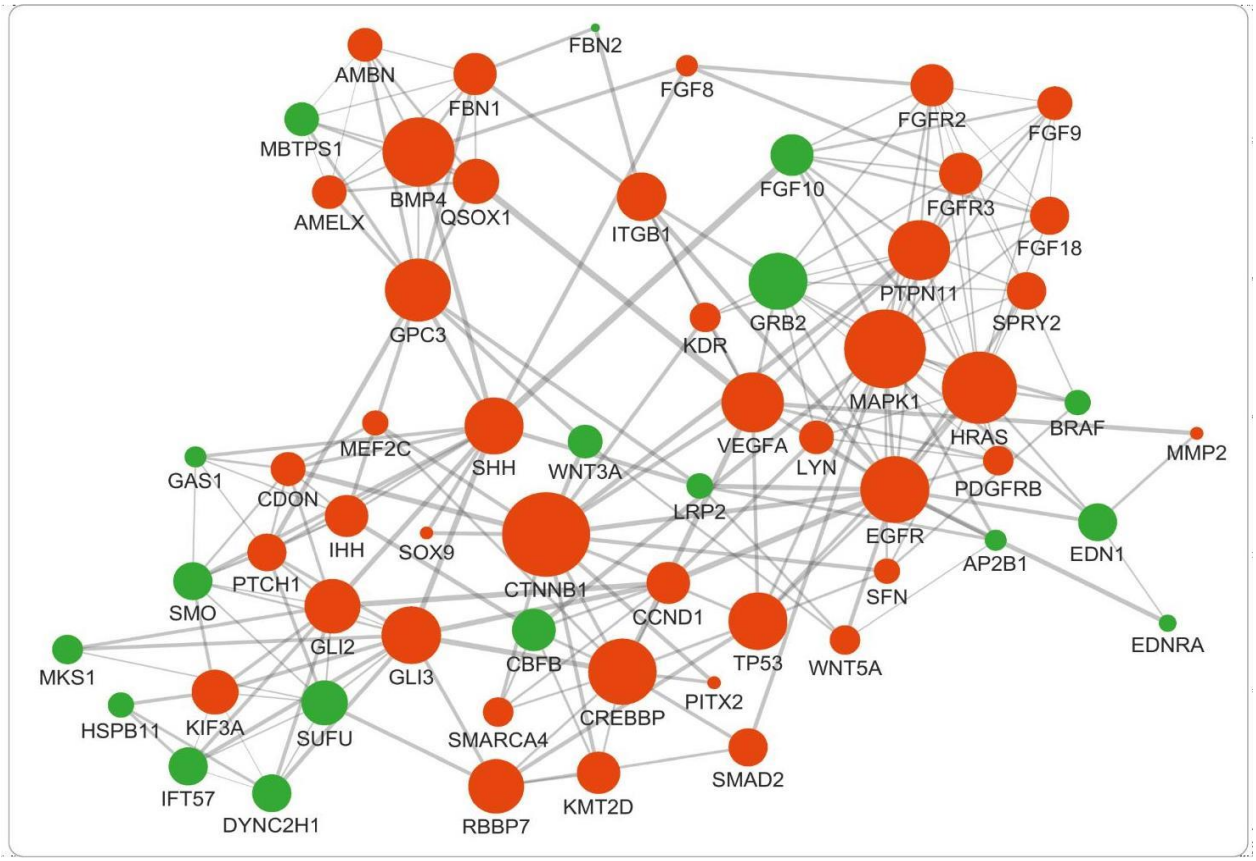


Figure 15: Deregulations of gnathostome-specific hubs of the JDN are more frequently associated with invasive and metastatic stages, in contrast with pre-gnathostome hubs. Functional interaction network (confidence interaction score ≥ 0.9) of the gnathostome-specific and pre- gnathostome JIG hubs. The size of the nodes is proportional to the node degree in the 'original' JDN network in Figure 15. The 60 nodes representing intra- and inter-modular hubs in the original network appear to be interconnected and form a relatively dense network.

In biological networks, the most highly connected nodes, "hubs", are considered to be of biological importance and more pertinent to the overall function of the network in general (Barabási, Gulbahce, and Loscalzo 2011; Kontou et al. 2016; Sethuraman and Sujasree 2020). The intra-modular hubs are central to a given network module, with the highest number of connections to the neighboring nodes. In contrast, inter-modular hubs are intermediate between two or more modules. With this in mind, we sought to investigate whether gnathostome-specific genes deregulated in cancer also occupy hub positions in the jaw-developmental network (JDN).

First, 60 nodes representing intra- and inter-modular hubs were identified through STRING network analysis (Figure 16). By approximating their evolutionary age, these hubs correspond highly to gnathostome-specific versus pre-gnathostome orthologues. Upon comparison of these 60 hubs to CCLE-derived data, it was found that 33 (55.0%) of them are deregulated in highly invasive cancer cells, 25 of which are of gnathostome, and 8 of pregnathostome origin ($\chi 2 = 7.76$, $p < 0.0054$) (Marquardt et al. 2021). Remarkably, all 33 JDN hubs predicted are effectively correlated to invasion and/or metastasis in a cancer type-dependent manner, according to experimental evidence in cancer cell lines, mouse models and tumor samples, whereas drugs targeting several of them are already in clinical trials or even FDA-approved for the treatment of aggressive and metastatic cancers (Marquardt et al. 2021). Taken together, the results are a sign of a prevalence of gnathostome-specific genes in occupying JDN hubs; nevertheless, at the same time, cancer cells preferentially exploited these over pre-gnathostome hubs to promote cancer aggressiveness.

4.4. Natural Language Processing

Networks illustrating the semantic relations of the highest ranking 50 most similar terms (words) to KRT18- KRT 19- KRT 13- KRT 14- KRT 15- KRT 16- KRT 17- KRT 20. The nodes represent the words, and the edges denote the semantic associations to selected KRTs which have adequate texts in their document. The size of the nodes indicates the degree of similarity between the given term and KRTs. The semantic relations of the highest ranking 50 terms (words) were constructed using the similarity distances between each word and were generated a KRT association network which was manipulated and also visualized by Cytoscape (http://www.cytoscape.org/). The word nodes on the network were placed according to their similarity with each KRTs.

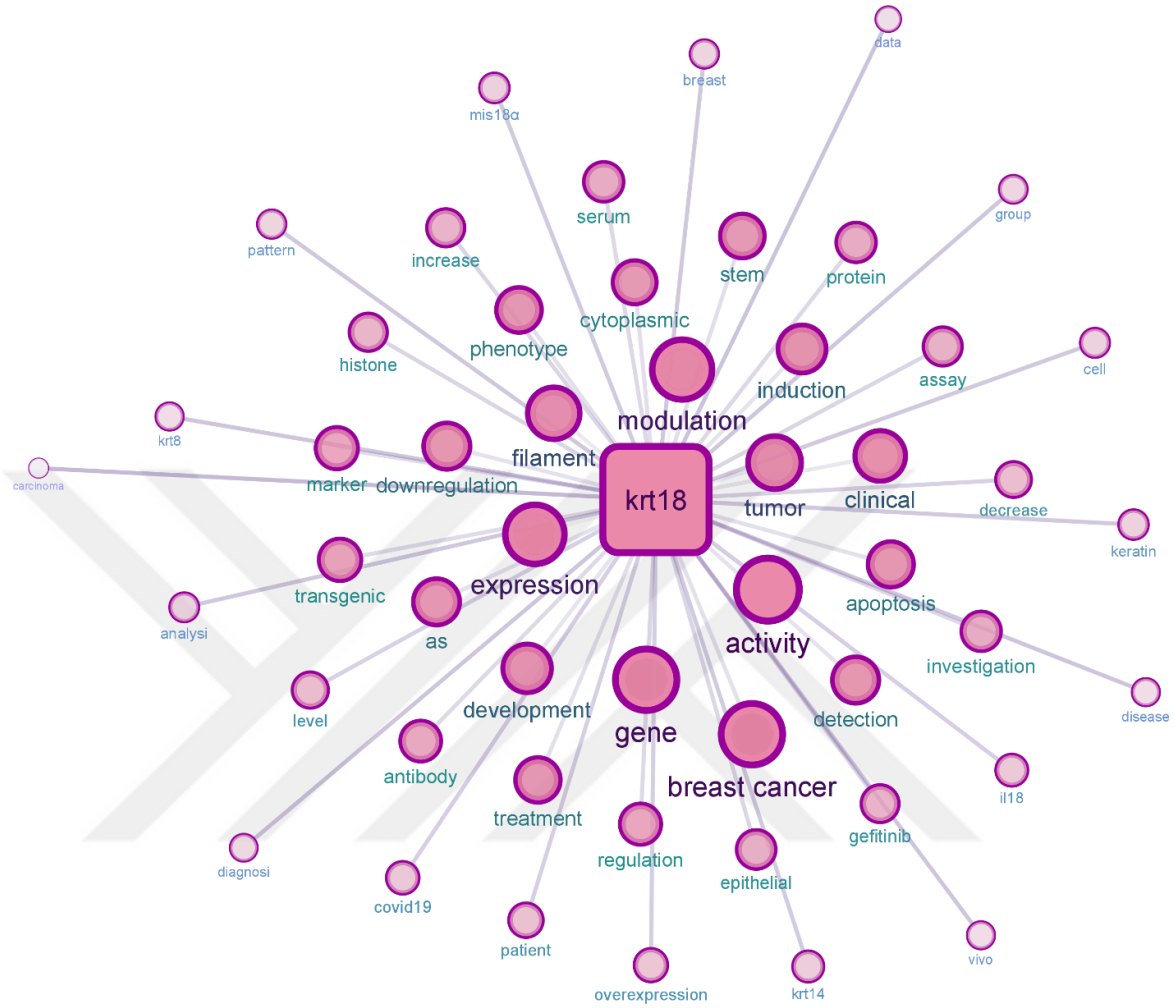


Figure 16: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT18. The nodes represent the words, and the edges indicate the semantic associations to KRT18. The size of the nodes indicates the degree of similarity between the given term and KRT18.

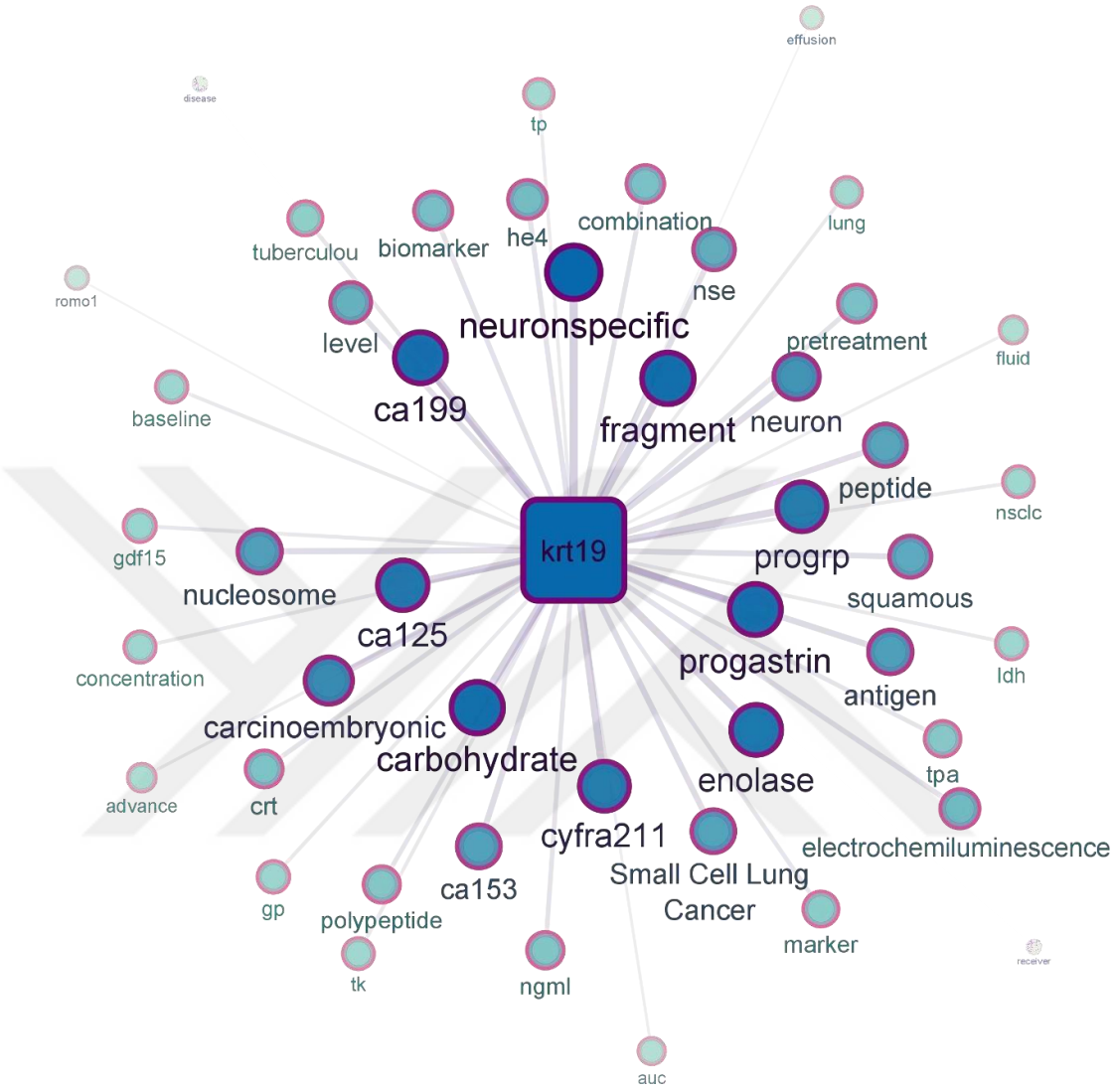


Figure 17: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT19. The nodes represent the words, and the edges represent the semantic associations to KRT19. The size of the nodes indicates the degree of similarity between the given term and KRT19.

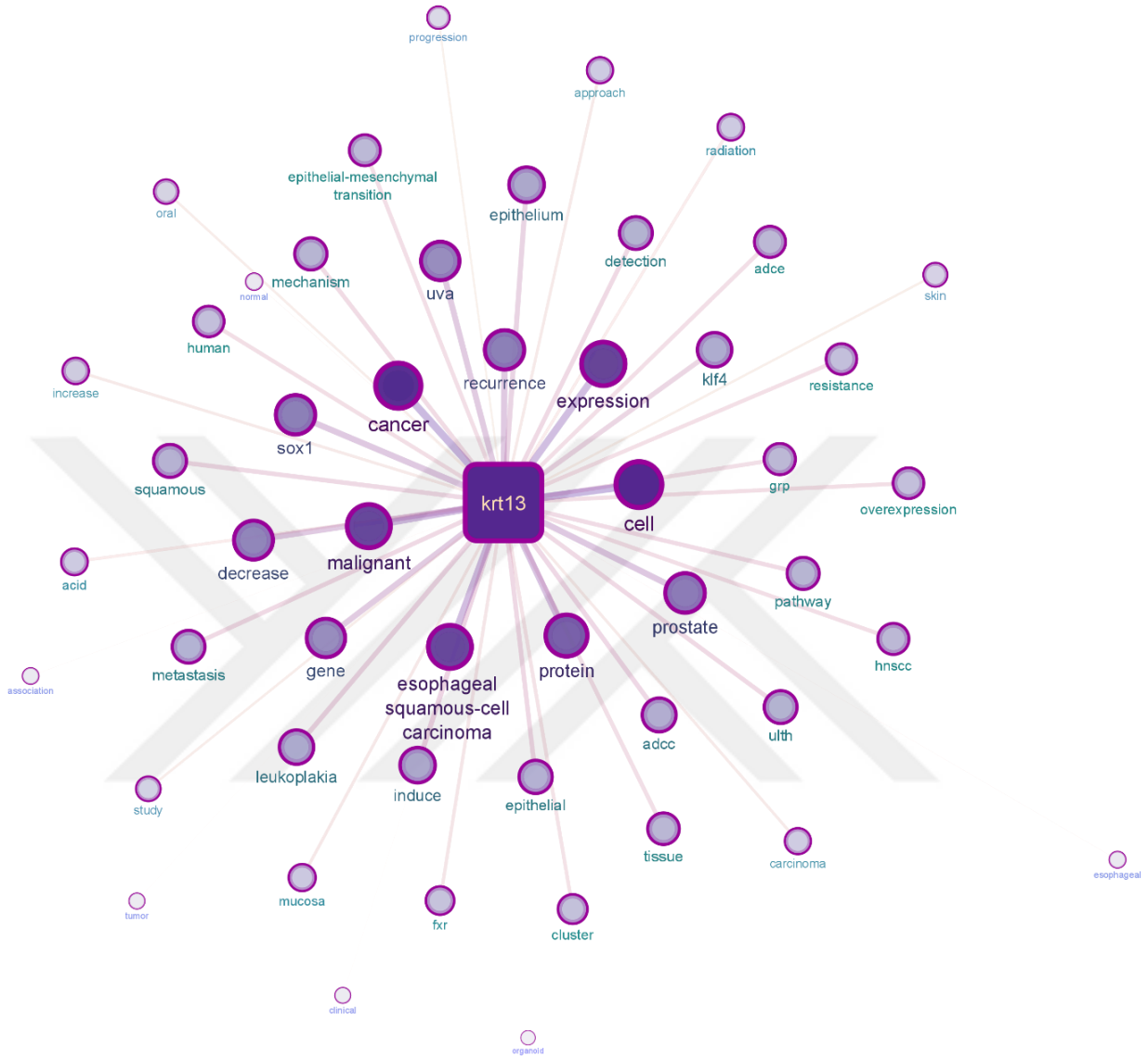


Figure 18: Network *illustrating* the semantic relations of the highest ranking 50 most similar terms to KRT13. The nodes represent the words, and the edges indicate the semantic associations to KRT13. The size of the nodes indicates the degree of similarity between the given term and KRT13.

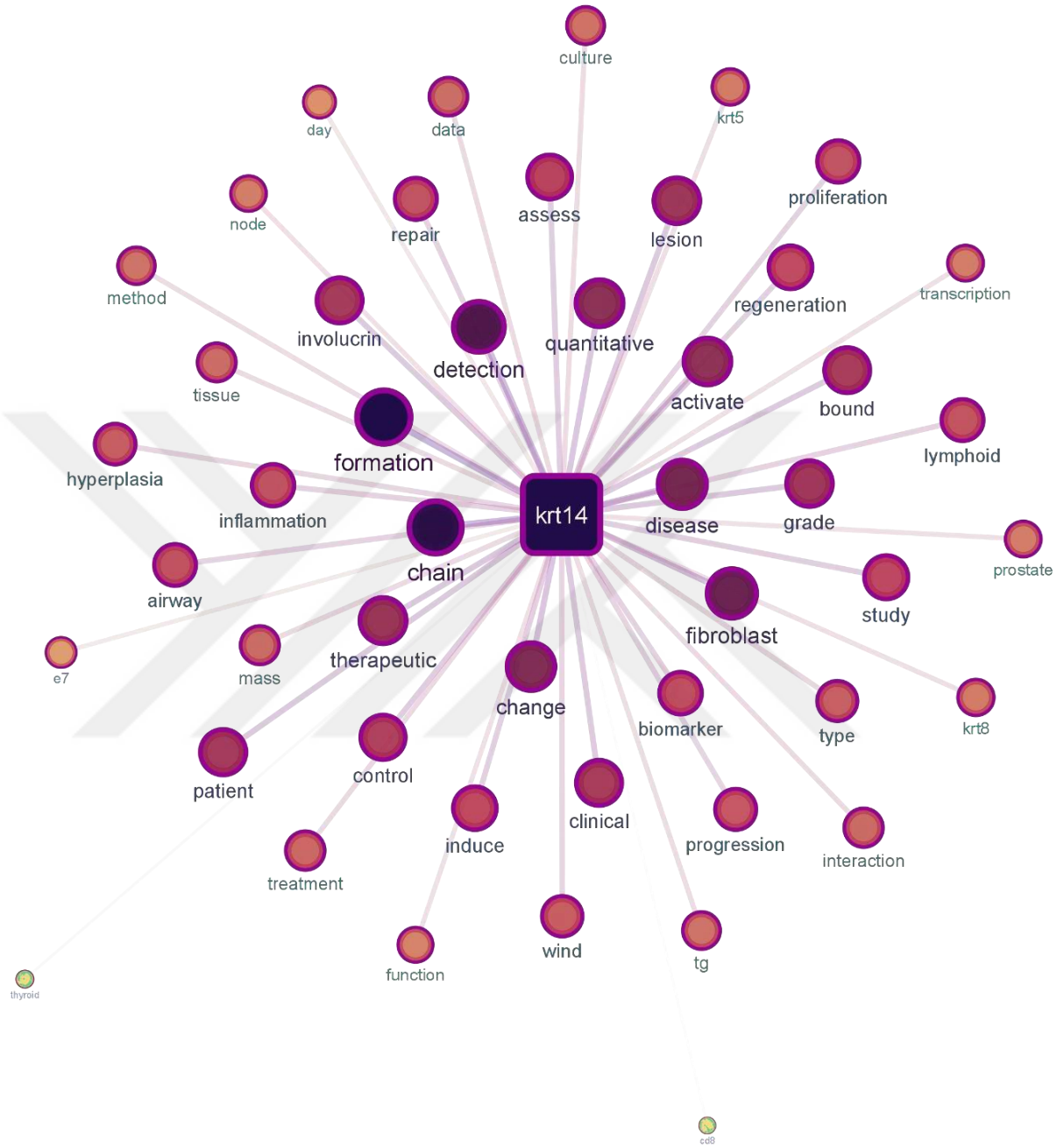


Figure 19: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT14. The nodes represent the words, and the edges represent the semantic associations to KRT14. The size of the nodes indicates the degree of similarity between the given term and KRT14.

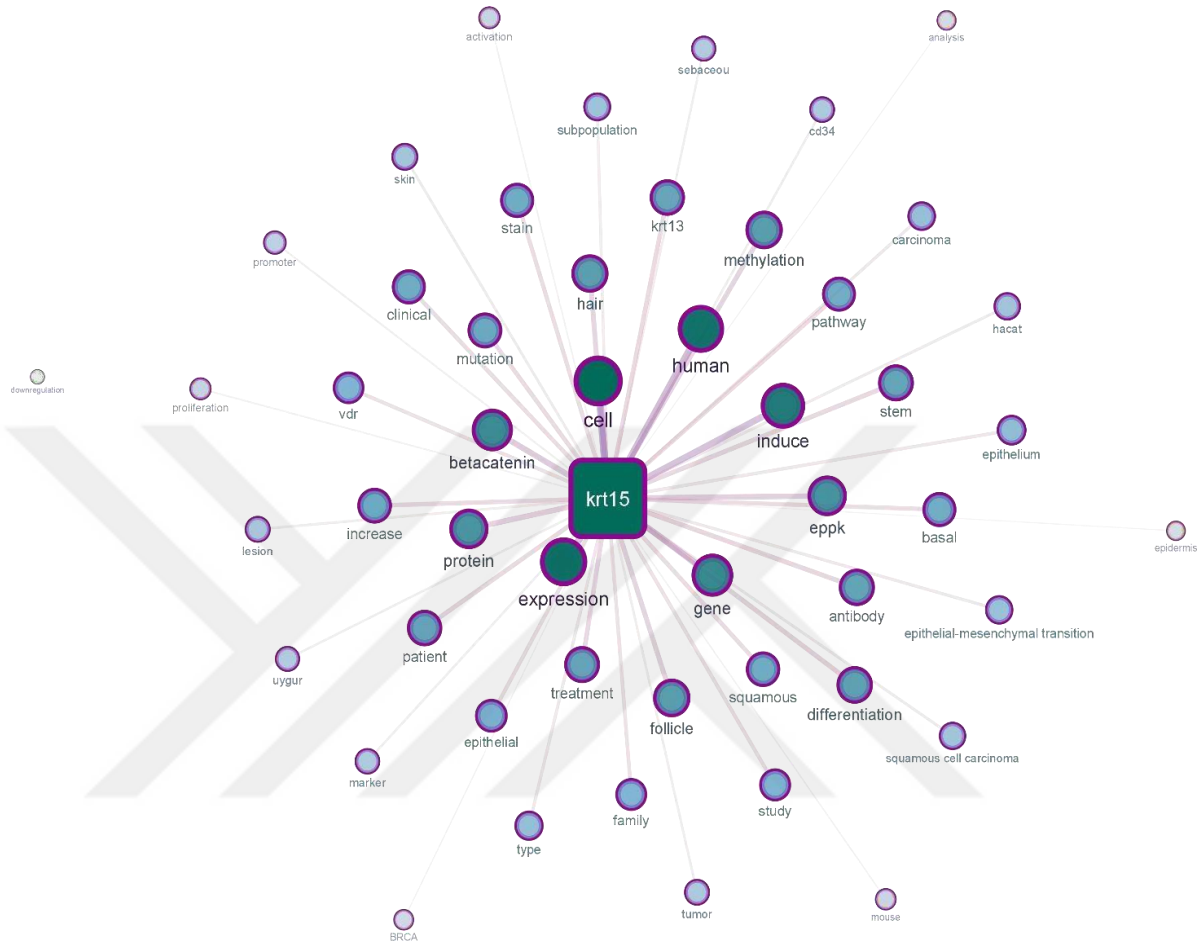


Figure 20: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT15. The nodes represent the words, and the edges indicate the semantic associations to KRT15. The size of the nodes indicates the degree of similarity between the given term and KRT15.

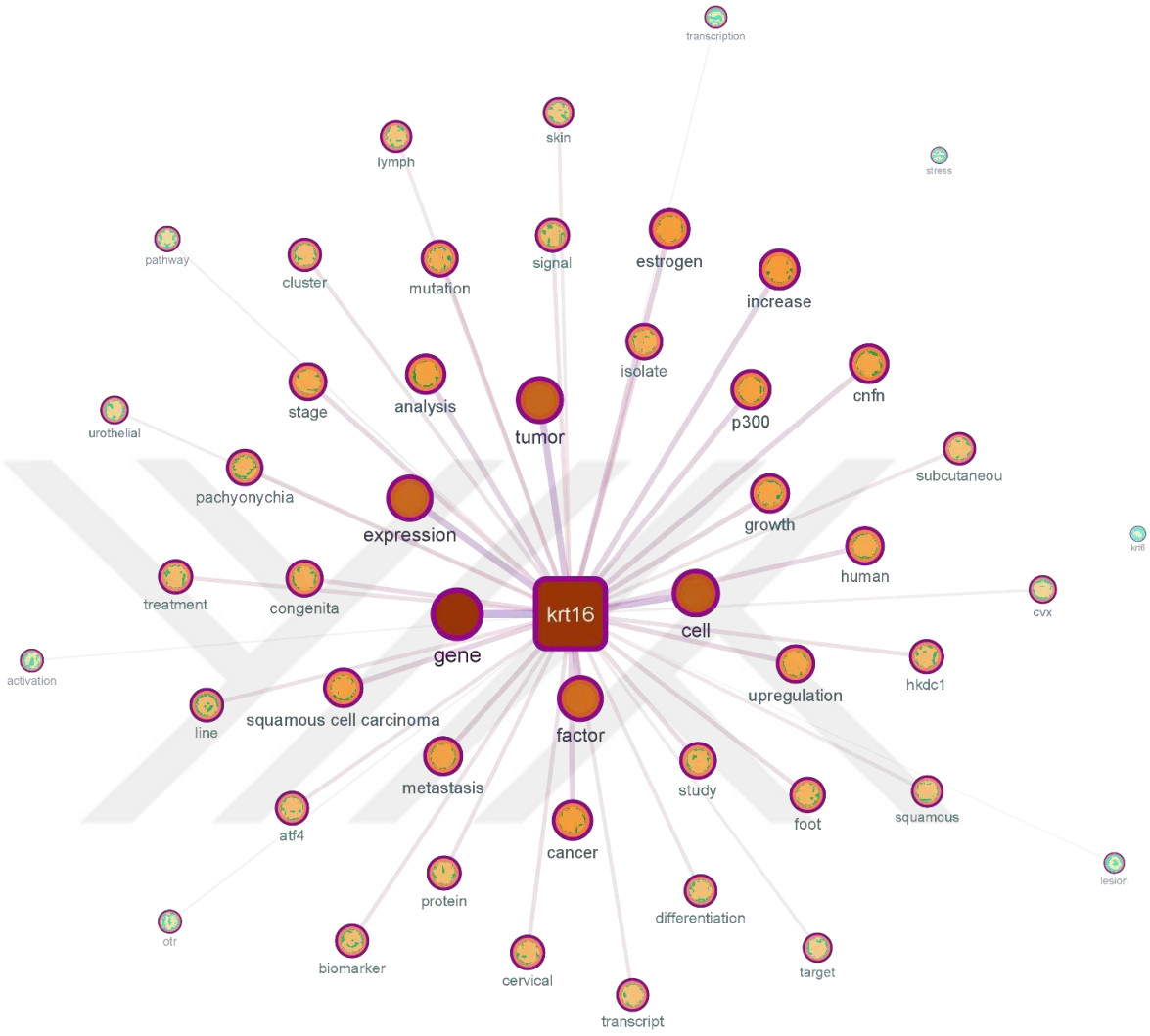


Figure 21: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT16. The nodes represent the words, and the edges represent the semantic associations to KRT16. The size of the nodes indicates the degree of similarity between the given term and KRT16.

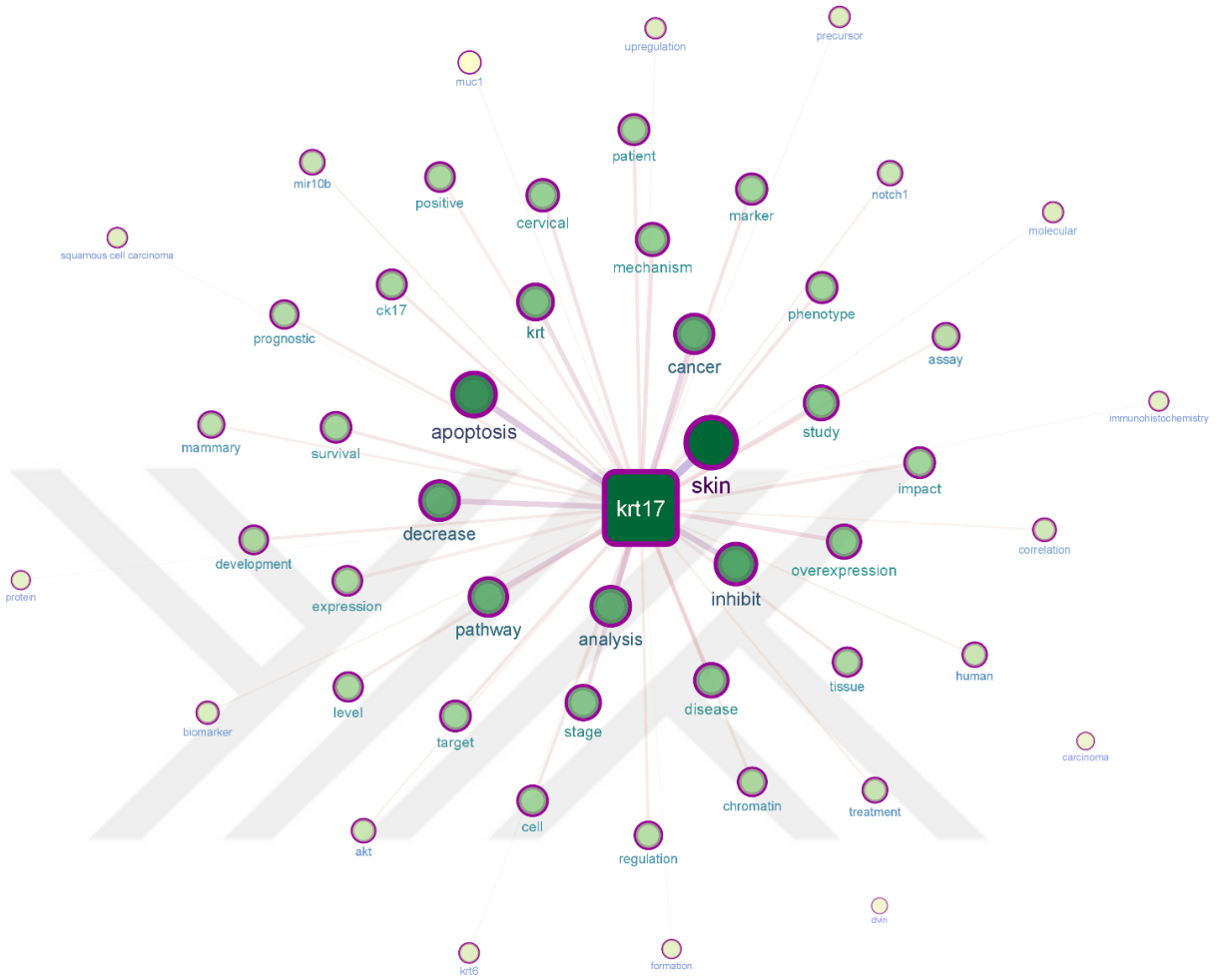


Figure 22: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT17. The nodes represent the words, and the edges indicate the semantic associations to KRT17. The size of the nodes indicates the degree of similarity between the given term and KRT17.

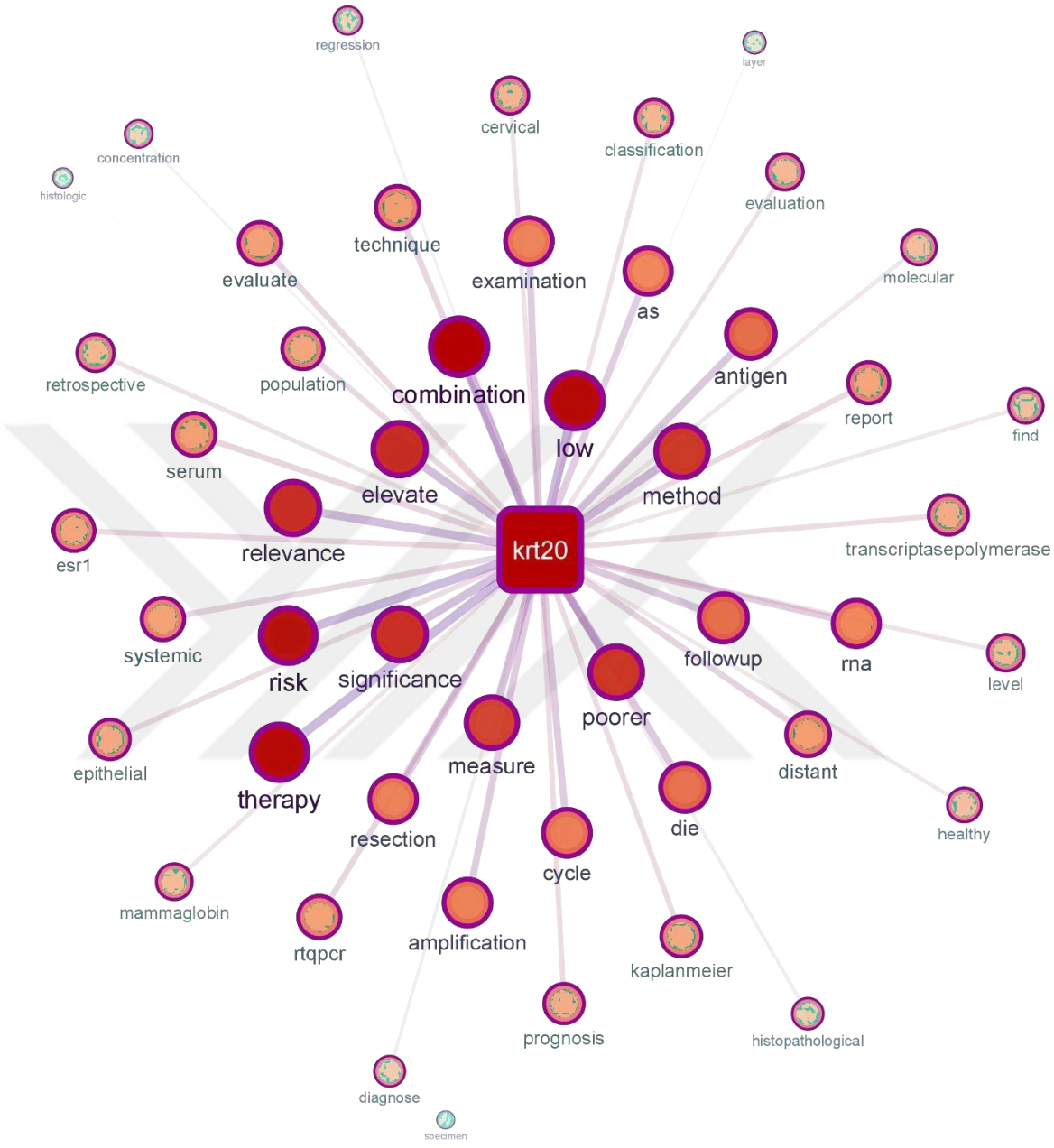


Figure 23: Network illustrating the semantic relations of the highest ranking 50 most similar terms to KRT20. The nodes represent the words, and the edges represent the semantic associations to KRT20. The size of the nodes indicates the degree of similarity between the given term and KRT20.

5. DISCUSSION

The scope of the present thesis was to identify the hallmark macroevolutionary trait(s) and corresponding genes that promoted the manifestation of the molecular mechanisms of

carcinogenesis in diverse taxonomic groups, by employing computational systems biology methods.

5.1. "In The Light of Evolution": Keratins as exceptional tumor biomarkers

The reconstructed phylogenies of keratins allowed us to get a glimpse into the evolutionary history of this extended gene family, and how members of this family are associated with pathophysiological processes. It was found that some *keratin* genes are converted to pseudogenes and *vice versa* during the course of evolution. According to the results of the phylogenetic analysis, the evolutionarily 'older' *KRT13*, *KRT15*, *KRT19*, *KRT23* and *KRT222*, were expressed differentially in diverse cancer types. Of note, *KRT222*, which was first detected in *Gallus gallus*, was found to be down-regulated specifically in brain neoplasms (lower grade glioma and glioblastoma multiforme). Furthermore, *KRT13*, *KRT19* and *KRT23*, first detected in chicken, were shown to be up-regulated in ovarian cancer. This suggests that, at least in cancer-bearing species, these genes (and their corresponding protein products) remained unchanged due to evolutionary pressure. Given that chicken is currently the only animal model available to investigate the etiology and progression of human ovarian cancer, the *krt13*, *krt19* and *krt23* chicken orthologues could be taken into consideration in the study of ovarian cancer.

The constructed KRT-oriented networks using word embedding provide notable information on how related words in our text dataset have semantic and syntactic similarity with the corresponding keratin. Our results support that there is significant association between keratins with cancer biomarkers. For instance, the expression patterns of conventional tumor markers such as proliferating cell nuclear antigen (PCNA) and minichromosome maintenance protein 3 (MCM3) in breast cancer are similar with *KRT18* patterns. Besides, in a majority of breast carcinomas, *KRT18* was mostly upregulated (P.-B. Wang et al. 2020).

Our similarity results also support that KRT18 could be a potential prognostic marker in breast cancer since it is associated with tumor dedifferentiation. It has been also suggested that downregulation of *KRT18* is linked to epithelial-to-mesenchymal transition (EMT), and therefore KRT18 can be considered as a candidate biomarker 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 determining the response of breast carcinomas to chemotherapy in advance (Olofsson et al. 2007).

It has been demonstrated that the levels of KRT19 and carcinoembryonic antigen are increased in carcinomas such as adenocarcinoma, squamous cell carcinoma and small cell lung cancer (Khanom et al. 2012). Reduced level of serum KRT19 fragments during chemoradiotherapy of non-small cell lung cancer (NSCLC) represents one of the prediction factors in tumor response. Also, KRT19 levels are strongly correlated with metastasis occurrence and survival status in all lung cancer patients. Hence, KRT19 is a strong independent prognostic factor for different stages of lung cancer (Juan Wang et al. 2011).

Cancer-associated fibroblasts show a powerful stimulatory effect on the expression of KRT14, which is basal/myoepithelial marker. KRT14 depicts the hyperproliferative responses of keratinocytes which play important role in skin repair (Dvořánková et al. 2012; Zhou et al. 2022). KRT14 could also serve as a biomarker in rare salivary gland neoplasm SACC (salivary adenoid cystic carcinoma) for predicting worse prognosis (Belulescu et al. n.d.; Gao et al. 2017).

According to our findings, KRT15 is intimately linked to molecular factors and pathways that mediate cell-cell interactions. The downregulation of transmembrane receptor β1-integrin allows cells to recognize and communicate with the extracellular matrix by inducing *KRT15* during terminal differentiation (Brooker et al. 2016).

Upregulation of *KRT14*, as well as downregulation of *KRT15* and *KRT19*, were obserserved in oral squamous cell carcinoma. There were considerable variations in the expression of *KRT15* and *KRT19* through all cancer types. Increased level of *KRT15* expression and decreased level of *KRT19* expression are observed in well-differentiated OSCC compared to moderately/poorly differentiated OSCC (Khanom et al. 2012).

Increased *KRT16* expression is highly associated with weak differentiation, augmented lymph nodes metastasis and advanced stages as well. Moreover, inhibition of *KRT16* reduceD migration, invasion, metastasis and chemoresistance, whereas knockdown of *KRT16* improves chemosensitivity in oral squamous cell carcinoma (W.-C. Huang et al. 2019). Also, there are

associations between increased *KRT16* expression and poor survival in metastatic breast cancer (Joosse et al. 2012).

Furthermore, KRT15 is directly associated with the differentiation-related keratin KRT13 (Khanom et al. 2012b). *KRT13* is transcriptionally regulated by KLF4 which is involved in esophageal squamous cell carcinoma cell differentiation. It has been shown that regulation of epithelial and mesenchymal marker genes was INITIATED by the TGF-β1-induced EMT program, and *KRT13* is transcriptionally suppressed via modulation of the state of chromatin at its promoter (He et al. 2015).

KRT17 triggers AKT signaling pathway and induces EMT, and it is highly correlated with progression of malignancy and poor prognosis of esophageal squamous cell carcinoma (ESCC) patients. Therefore, KRT17 may serve as a therapeutic target for the treatment of ESCC (Liu et al. 2020). It has been also demonstrated that increased level of *KRT17* is directly correlated with the progression of pancreatic cancer. Knockdown of *KRT17* led to suppression of cell growth, migration, tumor growth, and also induction of apoptosis and cell cycle arrest (P. Chen et al. 2020).

Quantitating *KRT20* expression by RT-PCR is a very sensitive and accurate method to detect unknown lymph node metastases. Also, measuring *KRT20* mRNA expression in lymph nodes is essential for exact tumor staging and for postoperative adjuvant treatment of colorectal cancer patients (G. Chen et al. 2004). Studies on pancreatic carcinoma, gastric cancer, 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).

The aforementioned findings highlight the prominent role of *KRT* genes in mammalian cancers. Several KRTs 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 on human cancer-associated *KRT* genes.

5.2. The 'Bing Bang' of Cancer Metastasis

The findings of our study suggest that the phenotypic threshold of metastasis coincides with agnatha-to-gnathostome transition and is associated with jaw evolution. Our findings underscore a preference of invasive tumors for usurping evolutionarily newer features. Otherwise, a recent concept, known as “atavistic model”, asserts that neoplasms rely on re-expression of ancestral traits and reverse evolution from multicellularity (MC) to unicellularity (UC) (Bussey et al. 2017), which is likely promoted by upregulation of UC genes, disruption of interconnectedness between UC and MC genes (Trigos et al. 2017) and loss-of-function mutations on MC genes (H. Chen et al. 2015). Furthermore, recent studies show that metastatic competence can arise from heterogeneous cancer cell populations without the need for acquisition of additional mutations and is benefited from further selection of tumor-initiating mutations that seed primary tumorigenesis (Jacob et al. 2015).

We propose that tumor initiation may be triggered by mutations in evolutionarily old genes that control processes at the ‘dawn’ of multicellularity, such as cell proliferation and genomic stability (Lineweaver, Davies, and Vincent 2014). Such alterations in ancient genes may confer genetic heterogeneity (H. Chen et al. 2015), which is the substrate for evolution. Subsequently, tumor progression may be enabled via selection of clones that entail crosstalks between tumor-initiating mutations already acquired at primary tumor cells and evolutionarily-young genes that support gnathostome key innovations. This scenario would, at least in part, explain why, upon comparisons between primary and secondary tumors, only few metastasis-specific mutations have been identified thus far, while activations of multiple transcriptional modules are a common theme (Rodrigues et al. 2018).

Taken together, our findings provide a glimpse of the evolutionary trajectories to metastasis. Identifying genes recurrently hijacked by tumors could inspire anti-metastatic therapeutic strategies towards prevention of cancer evolution.

6. CONCLUSION AND FUTURE ASPECTS

In this thesis, cross-disciplinary systems biology methods were employed by combining comparative genomics, biological literature mining, differentially expressed gene profiles, and biological networks to investigate evolutionarily conserved genes or molecular mechanisms and

pathways implicated in different aspects of cancer. The findings of this study could have potential application in the clinical setting. In this regard, the phylogenetically conserved genes can be exploited as diagnostic or prognostic tumor markers; they could also be utilized in cancer therapeutic decision-making or in future experimental studies for the rational development of targeted antineoplastic drugs.

7. REFERENCES

Arora, Monika, and Vineet Kansal. 2019. "Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis." *Social Network Analysis and Mining* 9 (1): 12. Accessed 5 14, 2022. https://link.springer.com/article/10.1007/s13278-019-0557-y.

Benner, Philipp, Miroslav Bačák, and Pierre-Yves Bourguignon. 2014. "Point estimates in phylogenetic reconstructions." *Bioinformatics* 30 (17): 534-540. Accessed 5 14, 2022. https://ncbi.nlm.nih.gov/pubmed/25161244.

Garat, Diego, and Dina Wonsever. 2019. "Towards De-identification of Legal Texts." *arXiv: Computation and Language.* Accessed 5 14, 2022. https://arxiv.org/abs/1910.03739.

Gonnet, Gaston H. 2012. "Surprising results on phylogenetic tree building methods based on molecular sequences." *BMC Bioinformatics* 13 (1): 148-148. Accessed 5 14, 2022. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-148.

Liu, Bing. 2018. "Text sentiment analysis based on CBOW model and deep learning in big data environment." *Journal of Ambient Intelligence and Humanized Computing* 1-8. Accessed 5 14, 2022. https://link.springer.com/article/10.1007/s12652-018-1095-6.

Lou, Robert, Darco Lalevic, Charles Chambers, Hanna M. Zafar, and Tessa S. Cook. 2019. "Automated Detection of Radiology Reports that Require Follow-up Imaging Using Natural Language Processing Feature Engineering and Machine Learning Classification." *Journal of Digital Imaging* 1-6. Accessed 5 14, 2022. https://ncbi.nlm.nih.gov/pubmed/31482317.

Sengupta, Arijit, and Mehmet M. Dalkilic. 2002. "DSQL — An SQL for Structured Documents Extended Abstract." *Lecture Notes in Computer Science* 757-760. Accessed 5 14, 2022. https://link.springer.com/chapter/10.1007/3-540-47961-9_61.

Soltis, Pamela S. Soltis and Douglas E. 2003. "Applying the Bootstrap in Phylogeny Reconstruction." *Statistical Science* 256–267.

Sugumar, Ramalingam. 2018. "IMPROVED PERFORMANCE OF STEMMING USING EFFICIENT STEMMER ALGORITHM FOR INFORMATION RETRIEVAL." *Journal of Global Research in Computer Sciences* 9 (5): 01-05. Accessed 5 14, 2022. http://jgrcs.info/index.php/jgrcs/article/view/1062.

Wu, Chuanyan, Chuanyan Wu, Rui Gao, Yusen Zhang, and Yang De Marinis. 2019. "PTPD: predicting therapeutic peptides by deep learning and word2vec." *BMC Bioinformatics* 20 (1): 456. Accessed 5 14, 2022. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3006-z.

Zhang, Shaodian, and Noémie Elhadad. 2013. "Unsupervised biomedical named entity recognition." *Journal of Biomedical Informatics* 46 (6): 1088-1098. Accessed 5 14, 2022. http://people.dbmi.columbia.edu/noemie/papers/13jbi_shaodian.pdf.

Altschul, S F et al. 1990a. “Basic Local Alignment Search Tool.” *Journal of molecular biology* 215(3): 403–10.

———. 1990b. “Basic Local Alignment Search Tool.” *Journal of molecular biology* 215(3): 403–10.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. “Network Medicine: A Network-Based Approach to Human Disease.” *Nature Reviews Genetics* 12(1): 56–68.

Belulescu, Iulia Cristiana et al. “Adenoid Cystic Carcinoma of Salivary Gland: A Ten-Year Single Institute Experience.” *Current health sciences journal* 46(1): 56–65.

Benson, Dennis A et al. 2017. “GenBank.” *Nucleic acids research* 45(D1): D37–42.
Bhattacharya, Rahul, Susri Ray Chaudhuri, and Sib S Roy. 2018. “FGF9-Induced Ovarian

Cancer Cell Invasion Involves VEGF-A/VEGFR2 Augmentation by Virtue of ETS1 Upregulation and Metabolic Reprogramming." *Journal of cellular biochemistry* 119(10): 8174–89.

Brand, Toni M et al. 2014. "AXL Mediates Resistance to Cetuximab Therapy." *Cancer research* 74(18): 5152–64.

Brooker, Sarah M, Allison M Bond, Chian-Yu Peng, and John A Kessler. 2016. "B1-Integrin Restricts Astrocytic Differentiation of Adult Hippocampal Neural Stem Cells." *Glia* 64(7): 1235–51.

Bruford, Elspeth A et al. 2008. "The HGNC Database in 2008: A Resource for the Human Genome." *Nucleic acids research* 36(Database issue): D445-8.

Buchmann, Jan P, and Edward C Holmes. 2019. "Entrezpy: A Python Library to Dynamically Interact with the NCBI Entrez Databases." *Bioinformatics* 35(21): 4511–14.

Bussey, Kimberly J, Luis H Cisneros, Charles H Lineweaver, and Paul C W Davies. 2017. "Ancestral Gene Regulatory Networks Drive Cancer." *Proceedings of the National Academy of Sciences of the United States of America* 114(24): 6160–62.

Chen, Gong et al. 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): 679–86.

Chen, Han, Fangqin Lin, Ke Xing, and Xionglei He. 2015. "The Reverse Evolution from Multicellularity to Unicellularity during Carcinogenesis." *Nature communications* 6: 6367.

Chen, Peng et al. 2020. "Silencing of Keratin 17 by Lentivirus-Mediated Short Hairpin RNA Inhibits the Proliferation of PANC-1 Human Pancreatic Cancer Cells." *Oncology letters* 19(5): 3531–41.

Chen, Taotao et al. 2017. “Fibroblast Growth Factor 18 Promotes Proliferation and Migration of H460 Cells via the ERK and P38 Signaling Pathways.” *Oncology reports* 37(2): 1235–42.

Chin, Chia-Hao et al. 2014. “CytoHubba: Identifying Hub Objects and Sub-Networks from Complex Interactome.” *BMC systems biology* 8 Suppl 4: S11.

Cock, P. J. A. et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25(11): 1422–23.

Coulombe, Pierre A, and M Bishr Omary. 2002. “‘Hard’ and ‘Soft’ Principles Defining the Structure, Function and Regulation of Keratin Intermediate Filaments.” *Current opinion in cell biology* 14(1): 110–22.

Cui, Genggang et al. 2019. “The Value of FGF9 as a Novel Biomarker in the Diagnosis of Prostate Cancer.” *Artificial cells, nanomedicine, and biotechnology* 47(1): 2241–45.

Davies, P C W, and C H Lineweaver. 2011. “Cancer Tumors as Metazoa 1.0: Tapping Genes of Ancient Ancestors.” *Physical Biology* 8(1): 015001.

van Dijk, A D J. 2017. “Species-Specific Genome Sequence Databases: A Practical Review.” *Methods in molecular biology (Clifton, N.J.)* 1533: 173–81.

Dvořánková, B et al. 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(5): 679–85.

Gao, Xiao-Lei et al. 2017. “Cytokeratin-14 Contributes to Collective Invasion of Salivary Adenoid Cystic Carcinoma.” *PloS one* 12(2): e0171341.

Gascuel, O. 1997. “BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data.” *Molecular biology and evolution* 14(7): 685–95.

de Grassi, Anna, Cecilia Lanave, and Cecilia Saccone. 2008. “Genome Duplication and Gene-Family Evolution: The Case of Three OXPHOS Gene Families.” *Gene* 421(1–2): 1–6.

Guindon, Stéphane, Frédéric Delsuc, Jean-François Dufayard, and Olivier Gascuel. 2009. “Estimating Maximum Likelihood Phylogenies with PhyML.” *Methods in molecular biology (Clifton, N.J.)* 537: 113–37.

Harris, Charles R. et al. 2020. “Array Programming with NumPy.” *Nature* 585(7825): 357–62.

He, Huan et al. 2015. “Krüppel-like Factor 4 Promotes Esophageal Squamous Cell Carcinoma Differentiation by Up-Regulating Keratin 13 Expression.” *The Journal of biological chemistry* 290(21): 13567–77.

Hesse, Michael, Alexander Zimek, Klaus Weber, and Thomas M Magin. 2004. “Comprehensive Analysis of Keratin Gene Clusters in Humans and Rodents.” *European journal of cell biology* 83(1): 19–26.

Hodgman C, French A, Westhead D.R. *Bioinformatics*. 2nd Edition. Taylor & Francis.

Huang, Wei-Chieh et al. 2019. “A Novel MiR-365-3p/EHF/Keratin 16 Axis Promotes Oral Squamous Cell Carcinoma Metastasis, Cancer Stemness and Drug Resistance via Enhancing B5-Integrin/c-Met Signaling Pathway.” *Journal of experimental & clinical cancer research : CR* 38(1): 89.

Huang, Yanqing et al. 2015. “Overexpression of FGF9 in Prostate Epithelial Cells Augments Reactive Stroma Formation and Promotes Prostate Cancer Progression.” *International journal of biological sciences* 11(8): 948–60.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9(3): 90–95.

Iacobuzio-Donahue, Christine A., Victor E. Velculescu, Christopher L. Wolfgang, and Ralph H.

Hruban. 2012. “Genetic Basis of Pancreas Cancer Development and Progression: Insights from Whole-Exome and Whole-Genome Sequencing.” *Clinical Cancer Research* 18(16): 4257–65.

Jacob, Leni S et al. 2015. “Metastatic Competence Can Emerge with Selection of Preexisting Oncogenic Alleles without a Need of New Mutations.” *Cancer research* 75(18): 3713–19.

Jaffray, David A, and Mary K Gospodarowicz. 2015. *Radiation Therapy for Cancer*.

Jäger, Gerhard. 2018. “Global-Scale Phylogenetic Linguistic Inference from Lexical Resources.” *Scientific data* 5: 180189.

Joosse, Simon A et al. 2012. “Changes in Keratin Expression during Metastatic Progression of Breast Cancer: Impact on the Detection of Circulating Tumor Cells.” *Clinical cancer research : an official journal of the American Association for Cancer Research* 18(4): 993–1003.

Karantza, V. 2011. “Keratins in Health and Cancer: More than Mere Epithelial Cell Markers.” *Oncogene* 30(2): 127–38.

Khanom, Rumana et al. 2012a. “Expression of Basal Cell Keratin 15 and Keratin 19 in Oral Squamous Neoplasms Represents Diverse Pathophysiologies.” *Histology and histopathology* 27(7): 949–59.

———. 2012b. “Expression of Basal Cell Keratin 15 and Keratin 19 in Oral Squamous Neoplasms Represents Diverse Pathophysiologies.” *Histology and histopathology* 27(7): 949–59.

Kim, Hyeong Su et al. 2019. “Pathological and Prognostic Impacts of FGFR2 Overexpression in Gastric Cancer: A Meta-Analysis.” *Journal of Cancer* 10(1): 20–27.

Kontou, Panagiota I. et al. 2016. “Network Analysis of Genes and Their Association with Diseases.” *Gene* 590(1): 68–78.

Kumar, Sudhir et al. 2018. “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.” *Molecular biology and evolution* 35(6): 1547–49.

Lineweaver, Charles H, Paul C W Davies, and Mark D Vincent. 2014. “Targeting Cancer’s Weaknesses (Not Its Strengths): Therapeutic Strategies Suggested by the Atavistic Model.” *BioEssays : news and reviews in molecular, cellular and developmental biology* 36(9): 827–35.

Liu, Zhun et al. 2020. “Keratin 17 Activates AKT Signalling and Induces Epithelial- Mesenchymal Transition in Oesophageal Squamous Cell Carcinoma.” *Journal of proteomics* 211: 103557.

Louka, Aikaterini, Işıl Takan, Athanasia Pavlopoulou, and Alexandros G Georgakilas. 2021. “Bioinformatic Approaches to the Investigation of the Atavistic Genes Implicated in Cancer.” *Frontiers in bioscience (Landmark edition)* 26(8): 279–311.

Marquardt, Stephan et al. 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.

Matsuda, Yoko et al. 2011. “Overexpressed Fibroblast Growth Factor Receptor 2 in the Invasive Front of Colorectal Cancer: A Potential Therapeutic Target in Colorectal Cancer.” *Cancer letters* 309(2): 209–19.

Matsuda, Yoko, Masahito Hagio, Tomoko Seya, and Toshiyuki Ishiwata. 2012. “Fibroblast Growth Factor Receptor 2 IIIc as a Therapeutic Target for Colorectal Cancer Cells.” *Molecular cancer therapeutics* 11(9): 2010–20.

McKinney, Wes, and others. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference, ,* 51–56.

Moll, Roland, Markus Divo, and Lutz Langbein. 2008. “The Human Keratins: Biology and Pathology.” *Histochemistry and cell biology* 129(6): 705–33.

Montazeri, Kamaneh, and Joaquim Bellmunt. 2020. “Erdafitinib for the Treatment of Metastatic Bladder Cancer.” *Expert review of clinical pharmacology* 13(1): 1–6.

Morris, John H et al. 2011. “ClusterMaker: A Multi-Algorithm Clustering Plugin for Cytoscape.” *BMC bioinformatics* 12: 436.

O’Leary, Nuala A et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic acids research* 44(D1): D733-45.

Olofsson, Maria Hägg et al. 2007. “Cytokeratin-18 Is a Useful Serum Biomarker for Early Determination of Response of Breast Carcinomas to Chemotherapy.” *Clinical cancer research : an official journal of the American Association for Cancer Research* 13(11): 3198–3206.

Pavlopoulou, Athanasia, Georgios Pampalakis, Ioannis Michalopoulos, and Georgia Sotiropoulou. 2010. “Evolutionary History of Tissue Kallikreins.” *PloS one* 5(11): e13781.

Rodrigues, Paulo et al. 2018. “NF-KB-Dependent Lymphoid Enhancer Co-Option Promotes Renal Carcinoma Metastasis.” *Cancer discovery* 8(7): 850–65.

van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Salipante, Stephen J, and Marshall S Horwitz. 2006. “Phylogenetic Fate Mapping.” *Proceedings of the National Academy of Sciences of the United States of America* 103(14): 5448–53.

Saville, Jennifer T, Belinda K McDermott, Janice M Fletcher, and Maria Fuller. 2019. “Disease and Subtype Specific Signatures Enable Precise Diagnosis of the Mucopolysaccharidoses.” *Genetics in medicine : official journal of the American College of Medical Genetics* 21(3): 753–57.

Sethuraman, G., and M. Sujasree. 2020. “Decomposition of Certain Complete Graphs and Complete Multipartite Graphs into Almost-Bipartite Graphs and Bipartite Graphs.” *Theory and Applications of Graphs* 7(2): 1–15.

Shannon, Paul et al. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome research* 13(11): 2498–2504.

Sievers, Fabian et al. 2011. “Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7(1): 539.

Szklarczyk, Damian et al. 2017. “The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible.” *Nucleic acids research* 45(D1): D362–68.

Tang, Zefang et al. 2019. “GEPIA2: An Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis.” *Nucleic acids research* 47(W1): W556–60.

Teishima, Jun et al. 2014. “Accumulation of FGF9 in Prostate Cancer Correlates with Epithelial-to-Mesenchymal Transition and Induction of VEGF-A Expression.” *Anticancer research* 34(2): 695–700.

Theofylaktou, Dionysia et al. 2021. “Mining Natural Products with Anticancer Biological Activity through a Systems Biology Approach.” *Oxidative medicine and cellular longevity* 2021: 9993518.

Trigos, Anna S, Richard B Pearson, Anthony T Papenfuss, and David L Goode. 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(24): 6406–11.

UniProt Consortium, The. 2018. “UniProt: The Universal Protein Knowledgebase.” *Nucleic acids research* 46(5): 2699.

Wang, Jizhong et al. 2020. “FGF9 Inhibition by a Novel Binding Peptide Has Efficacy in Gastric and Bladder Cancer per Se and Reverses Resistance to Cisplatin.” *Pharmacological research* 152: 104575.

Wang, Ju et al. 2011. “Antitumor Activity of a Recombinant Soluble Ectodomain of Mutant Human Fibroblast Growth Factor Receptor-2 IIIc.” *Molecular cancer therapeutics* 10(9): 1656–66.

Wang, Juan et al. 2011. “Decline of Serum CYFRA21-1 during Chemoradiotherapy of NSCLC: A Probable Predictive Factor for Tumor Response.” *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 32(4): 689–95.

Wang, L, and T Jiang. 1994. “On the Complexity of Multiple Sequence Alignment.” *Journal of computational biology : a journal of computational molecular cell biology* 1(4): 337–48.

Wang, Peng-Bin et al. 2020. “Keratin 18 Induces Proliferation, Migration, and Invasion in Gastric Cancer via the MAPK Signalling Pathway.” *Clinical and experimental pharmacology & physiology*.

Waskom, Michael et al. 2017. “Mwaskom/Seaborn: V0.8.1 (September 2017).” https://doi.org/10.5281/zenodo.883859.

Wein, L et al. 2017. “FGFR2 Amplification in Metastatic Hormone-Positive Breast Cancer and Response to an MTOR Inhibitor.” *Annals of oncology : official journal of the European Society for Medical Oncology* 28(8): 2025–27.

Wu, Dong-Dong, David M Irwin, and Ya-Ping Zhang. 2008. “Molecular Evolution of the Keratin Associated Protein Gene Family in Mammals, Role in the Evolution of Mammalian Hair.” *BMC evolutionary biology* 8: 241.

Yang, Ziheng, and Bruce Rannala. 2012. “Molecular Phylogenetics: Principles and Practice.” *Nature reviews. Genetics* 13(5): 303–14. Zerbino, Daniel R et al. 2018. “Ensembl 2018.” *Nucleic acids research* 46(D1): D754–61.

Zhou, Lian et al. 2022. “Effects of CD133 Expression on Chemotherapy and Drug Sensitivity of Adenoid Cystic Carcinoma.” *Molecular medicine reports* 25(1).

8. ADDENDA

8.1. Keratin JSON to Tables

ACC

KRT13
ENSG00000171401.14
0
3.97
-2.313
4.7699999999999996e-30

BRCA

KRT14
ENSG00000186847.5
12.009
1.177
2.5789999999999997
0.00337
KRT16
ENSG00000186832.8
32.895
2.235
3.389
0.00921
KRT17
ENSG00000128422.15
1036.137
173.08700000000002
2.575
0.000741

KRT14
ENSG00000186847.5
29.611
411.17199999999997
-3.7510000000000003
4.8900000000000004e-51
KRT15
ENSG00000171346.13
37.969
310.83299999999997
-3
1.64e-34
KRT17
ENSG00000128422.15
38.009
356.46
-3.196
2.8899999999999997e-38

BLCA

KRT14
ENSG00000186847.5
409.32300000000004
1.25
7.511
0.0176

KRT15
ENSG00000171346.13
223.31599999999997
1.59
6.436
0.00000503

KRT16
ENSG00000186832.8
576.888
3.27
7.08
0.000027000000000000002

KRT17
ENSG00000128422.15
2407.632
1.81
9.743
5.429999999999999e-16

KRT19
ENSG00000171345.13
1700.33
254.301
2.736

KRT23
ENSG00000108244.16
8.87
0.43
2.787
0.0326

3.4e-8

KRT13
ENSG00000171401.14
0.07
3.41
-2.043
8.6e-49
KRT17
ENSG00000128422.15
7.82
0.74
2.342
3.9899999999999995e-51
KRT19
ENSG00000171345.13
723.5239999999999
56.61
3.653
1.7600000000000001e-29
KRT20
ENSG00000171431.3
155.912
17.31
3.0989999999999998

CESC

9.5e-7

COAD

KRT23
ENSG00000108244.16
13.96
0.14
3.714
1.6499999999999998e-87

KRT13
ENSG00000171401.14
20.752
40081.53
-10.847999999999999
2.86e-56
KRT15
ENSG00000171346.13
55.736000000000004
1741.8429999999998
-4.941
6.02e-31
KRT17
ENSG00000128422.15
768.365
81.574
3.22
5.300000000000001e-25
KRT24
ENSG00000167916.4
0.04
29.529
-4.876
1.5900000000000002e-28
KRT31
ENSG00000094796.4
0.2
4.125

ESCA

-2.095
7.6e-18

DLBC

KRT23
ENSG00000108244.16
0
11.4
-3.6319999999999997
1.9e-38

KRT32
ENSG00000108759.3
0.035
3.43
-2.0980000000000003
2.5800000000000003e-59

KICH

KRT19
ENSG00000171345.13
1.895
150.361
-5.707999999999999
2.9e-15

GBM

KRT222
ENSG00000213424.8
0.16
13.78
-3.6710000000000003
9.909999999999997e-79

KRT10
ENSG00000186395.6
6.11
89.179
-3.665
9.369999999999999e-35

KRT13
ENSG00000171401.14
294.29
25584.811
-6.437
1.79e-12

KRT19
ENSG00000171345.13
0
174.024
-7.4510000000000005
1.7799999999999998e-153

KRT17
ENSG00000128422.15
6652.5869999999995

1024.603	
2.698	
8.960000000000001e-20	
KRT23	
ENSG00000108244.16	
11.42	**HNSC**
48.802	
-2.004	
1.06e-8	

LAML

LGG

235.51	
0.22	
7.599	
5.439999999999999e-172	
KRT17	
ENSG00000128422.15	
1767.261	
4.76	
8.262	
9.329999999999997e-167	
KRT19	
ENSG00000171345.13	
1407.1879999999999	
254.513	

KRT222
ENSG00000213424.8
1.03
13.78
-2.864
1.7399999999999998e-49

2.4619999999999997
3.8599999999999993e-63

KRT13
ENSG00000171401.14
0.37
5.2
-2.178
6.19e-53

KRT13
ENSG00000171401.14
134.554
5.575
4.3660000000000005
1.5299999999999999e-53
KRT14
ENSG00000186847.5
121.111
1.405
5.666
1.1799999999999996e-82
KRT15
ENSG00000171346.13
356.26199999999994
7.615
5.374
9.579999999999997e-112

KRT16
ENSG00000186832.8

KRT13
ENSG00000171401.14
0.23
6.247000000000001
-2.559
2.41e-31

KRT17
ENSG00000128422.15
33.686
0.44
4.59
5.51e-57

LUAD

KRT19
ENSG00000171345.13

LUSC

OV

473.41	1456.8120000000001
65.416	15.73
2.8369999999999997	6.445
2.4699999999999998e-57	5.0499999999999995e-76
KRT23	**KRT23**
ENSG00000108244.16	ENSG00000108244.16
19.9	15.35
0.09	0.63
4.261	3.326
1.0899999999999998e-54	1.4299999999999998e-43

PAAD

PRAD

KRT15
ENSG00000171346.13
12.11
0.61
3.0260000000000002
8.909999999999999e-44

KRT13
ENSG00000171401.14
2.26
27.305999999999997
-3.1180000000000003
1.5600000000000002e-26

KRT16
ENSG00000186832.8
28.999000000000002
0.15
4.705
1.48e-49

KRT15
ENSG00000171346.13
47.965
317.959
-2.7039999999999997
8.36e-15

KRT17
ENSG00000128422.15
295.824
0.67
7.474
3.529999999999999e-68

KRT23
ENSG00000108244.16
4.12
20.664
-2.081
1.58e-19

KRT19
ENSG00000171345.13

READ **SKCM**

KRT13
ENSG00000171401.14
0.06
3.755
-2.165
4.1e-28

KRT17
ENSG00000128422.15
6.68
0.85
2.0540000000000003
9.649999999999999e-33

KRT19
ENSG00000171345.13
563.471
15.975
5.055
2.55e-14

KRT20
ENSG00000171431.3
158.77
7.1979999999999995
4.285
0.00000125

KRT23
ENSG00000108244.16
42.614
0.12
5.2829999999999995

KRT10
ENSG00000186395.6
13.74
10974.365
-9.54
0

KRT13
ENSG00000171401.14
0.09
3.58
-2.0709999999999997
2.7e-21

KRT14
ENSG00000186847.5
0.82
5307.253000000001
-11.51
9.739999999999995e-250

KRT15
ENSG00000171346.13
0.22
780.956
-9.324
0

KRT16
ENSG00000186832.8
0.18
119.444

9.4e-88	-6.672999999999999
	4.5699999999999999e-103

KRT17
ENSG00000128422.15
0.66
490.07099999999997
-8.209
1.8099999999999995e-162
KRT19
ENSG00000171345.13
0.23
99.016
-6.345
3.2599999999999987e-223
KRT23
ENSG00000108244.16
0.17
134.63299999999998
-6.857
0
KRT31
ENSG00000094796.4
0
18.075
-4.254
5.9099999999999985e-151

-2.059
8e-9

KRT17
ENSG00000128422.15
15.16
0.39
3.5389999999999997
5.36e-34

TGCT

KRT13
ENSG00000171401.14
0.12
14.84
-3.822
6.110000000000001e-23

KRT15
ENSG00000171346.13
0.41
25.221
-4.217
1.8300000000000003e-24

KRT17
ENSG00000128422.15
0.38
5.21
-2.17
0.000273

KRT19
ENSG00000171345.13
25.98
4.46
2.305

KRT13
ENSG00000171401.14
0.2
4

STAD

1.1199999999999999e-9
KRT23
ENSG00000108244.16
0.37
30.721
-4.533
2.92e-43
KRT33A
ENSG00000006059.3
0
3.08
-2.029
7.0199999999999995e-50
KRT19
ENSG00000171345.13
259.465
32.561
2.9560000000000004
9.859999999999999e-49
KRT10
ENSG00000186395.6
16.45
3.29
2.024
3.759999999999999e-62

THCA

KRT14
ENSG00000186847.5
75.015
0.51
5.654
7.26e-61
KRT15
ENSG00000171346.13
56.685
0.18
5.611000000000001
1.5399999999999997e-78
KRT17
ENSG00000128422.15
74.32
0.13
6.059
3.2699999999999992e-115
KRT19
ENSG00000171345.13
555.2130000000001
0.17
8.892999999999999
2.0299999999999994e-193
KRT23
ENSG00000108244.16
0.15
11.35
-3.425
2.7799999999999993e-52

KRT13
ENSG00000171401.14
0.53
8.74
-2.6710000000000003
8.46e-8
KRT17
ENSG00000128422.15
34.298
0.89
4.223
1.59e-20
KRT19
ENSG00000171345.13
478.60699999999997
44.919
3.385
1.26e-18
KRT23
ENSG00000108244.16
18.705
0.16
4.086
1.6999999999999997e-23

THYM UCEC

KRT13
ENSG00000171401.14
0.44
10.117
-2.949
0.00000559
KRT17
ENSG00000128422.15
73.971
0.815
5.367999999999999
2.04e-13
KRT19
ENSG00000171345.13
211.27700000000002
40.464
2.356
0.00894

UCS

8.2. The 253 Jaw Indispensable Gene symbols and Accession Numbers

GENES	H. sapiens	M. musculus	G. gallus	X. laevis	D. rerio	C. milli	Petromyzon spp.	Ciona spp.	Brachiostoma spp.	Strongylocentrotus spp.	Hydra spp.
ACVR2A	NP_001607.1	NP_031422.3	NP_990698.1	XP_018089695.1	NP_001103748.2	XP_007894547.1					
ALX1	NP_008913.2	NP_766141.1	XP_025010336.1	NP_001079116.1	NP_001038539.1	XP_007901967.1					
AMBN	NP_057603.1	NP_033794.1		NP_001083489.1	NP_001138709.1						
AMELY	NP_001134.1	NP_001075447.1	XP_015146083.1	NP_001079152.1	XP_005160745.1	XP_007893773.1					
ARID5B	NP_115575.1	NP_076087.2	XP_015152937.2	XP_018080600.1	NP_001073670.1	XP_007910615.1					
ASXL1	NP_056153.2	NP_001035028.1	XP_015152084.1	NP_001086579.1	XP_001341242.2	XP_007886567.1					
BCL11B	NP_612808.1	NP_067374.2	XP_003641457.1	XP_018085985.1	XP_009291236.1	XP_007903122.1					
BMP4	NP_001193.2	NP_031580.2	NP_990568.3	NP_001095263.1	NP_571417.1	XP_007886339.1					
BMP5	NP_066551.1	NP_031581.2	NP_990479.1	XP_018118958.1	NP_957345.1	XP_007892257.1					
BMP7	NP_001710.1	NP_031583.2	XP_417496.5	NP_001080866.1	NP_571396.1	XP_007901604.1					
BRD4	NP_490597.1	NP_065254.3	XP_015128409.1	XP_018106485.1	NP_001104751.1	XP_007885536.1					

CACNA1S	NP_000060.2	NP_055008.2	XP_024999440.1	OCT94594.1	NP_999891.1	XP_007892902.1					
CCND1	NP_444284.1	NP_031657.1	NP_990712.1	NP_001080245.1	NP_571100.1	AFK11583.1					
CDON	NP_001230526.1	NP_067314.2	XP_417853.3	NP_001079158.1	NP_001075097.1	XP_007894240.1					

CHD1	NP_001261.2	NP_031716.2	XP_015155950.1	XP_018119975.1	NP_001121770.2	XP_007891302.1					
CHST11	NP_060883.1	NP_067414.2	XP_025007655.1	XP_018107832.1	NP_997989.2	XP_007893480.1					
CHUK	NP_001269.3	NP_031726.2	NP_001012922.1	NP_001086127.1	NP_956611.1	XP_007896804.1					
COL11A1	NP_001845.3	NP_031755.2	XP_422303.3	XP_018114134.1	XP_005162870.1	XP_007885476.1					
COL1A1	NP_000079.2	NP_031768.2	XP_024999899.1	NP_001080821.1	NP_954684.1	XP_007908905.1					
COL27A1	NP_116277.2	NP_079961.3	XP_015146330.1	XP_018088698.1	NP_001074044.1	XP_007898565.1					
CREBBP	NP_004371.2	NP_001020603.1	XP_015150110.2	NP_001088637.1	XP_021329192.1	XP_007891802.1					
CTNNB1	NP_001895.1	NP_033948.1	NP_990412.2	NP_001084045.1	NP_571134.2	XP_007899861.1					
CTNNBIP1	NP_064633.1	NP_075954.1	NP_001289087.1	NP_001086620.1	NP_571669.1	XP_007900324.1					

CTSK	NP_000387.1	NP_031828.2	NP_990302.2	XP_018088930.1	NP_001017778.1	XP_007899280.1					
DISP1	NP_001364158.1	NP_081142.3	XP_004935360.1	XP_018120452.1	NP_997965.1	XP_007896461.1					
DKK1	NP_036374.1	NP_034181.2	XP_015134195.1	NP_001079061.1	NP_571078.1	XP_007895204.1					
DLG1	NP_001353136.1	NP_031888.2	XP_015147091.1	XP_018119117.1	NP_955820.1	XP_007896055.1					
DLX2	NP_004396.1	NP_034184.1	XP_025008373.1	NP_001084032.1	NP_571386.2	XP_007888184.1					
DLX5	NP_005212.1	NP_034186.2	NP_989490.1	XP_018124683.1	NP_571381.2	XP_007902942.1					
DLX6	NP_005213.3	NP_034187.1	XP_015137078.1	NP_001090284.1	NP_571398.1	XP_007902943.1					
DNAH11	NP_001264044.1	NP_034190.3	XP_415585.5	XP_018122854.1	XP_009290539.1	XP_007901515.1					
ECE1	NP_001388.1	NP_955011.1	XP_015152737.1	NP_001080378.1	NP_001071260.1	XP_007896015.1					
EGFR	NP_005219.2	NP_997538	NP_990828.2	XP_018123025.1	NP_919405.1	XP_007895818.1					
EGR2	NP_000390.2	NP_034248.2	XP_025007817.1	NP_001079248.1	NP_571072.2	XP_007895410.1					
EYA1	NP_000494.2	NP_034294.2	XP_015138368.1	XP_018097003.1	NP_571268.1	XP_007885157.1					
FBLN1	NP_006477.3	NP_034310.2	NP_989496.1	NP_001089430.1	NP_571117.2	XP_007892884.1					
FBN1	NP_000129.3	NP_032019.2	XP_015147420.1	XP_018108582.1	XP_017207479.2	XP_007909428.1					
FGF18	NP_003853.1	NP_032031.1	NP_990045.1	NP_001008163.1	NP_001013282.1	XP_007889620.1					

FGF8	NP_149355.1	NP_034335.1	NP_001012785.1	NP_001083904.1	NP_571356.2	XP_007904873.1					
FGF9	NP_002001.1	NP_038546.2	NP_989730.1	NP_001079276.1		XP_007898849.1					
FGFR1	NP_075598.2	NP_034336.2	NP_990841.1	NP_001081157.1	NP_694494.2	XP_007909759.1					
FGFR2	NP_000132.3	NP_034337.2	NP_990650.1	NP_001084132.1	NP_001229933.1	XP_007902087.1					
FGFR3	NP_000133.1	NP_032036.2	NP_990840.2	NP_001084170.1	NP_571681.2	XP_007896238.1					
FIGN	NP_060556.2	NP_068362.1	XP_004942874.1	XP_018091343.1	NP_001018411.1	XP_007888032.1					
FLVCR1	NP_054772.1	NP_932142.2	XP_419425.2	XP_018120489.1	NP_001020693.1	XP_007908741.1					
FOXC1	NP_001444.2	NP_032618.2	NP_990337.1	NP_001089846.1		XP_007887802.1					

FOXC2	NP_005242.1	NP_038547.2	NP_990469.1	NP_001083920.1		XP_007887427.1					
FOXF2	NP_001443.1	NP_034355.2	XP_015137672.2	XP_018123225.1	XP_003200754.1	XP_007887803.1					
FOXI3	NP_001129121.1	NP_001094934.1	XP_425185.1	XP_018111695.1	NP_944599.2	XP_007889609.1					
GJA1	NP_000156.1	NP_034418	NP_989917.1	NP_001079129.1	NP_571113.1	XP_007897861.1					
GLI2	NP_005261.2	NP_001074594.1	NP_001258830.1	NP_001081894.1	NP_571042.1	XP_007894561.1					
GLI3	NP_000159.3	NP_032156.2	NP_001258832.1	NP_001081440.1	NP_991291.1	XP_007896289.1					
GPC3	NP_001158089.1	NP_057906.2	XP_025006018.1	XP_002932665.2	NP_001292547.1	XP_007890697.1					
GREM2	NP_071914.3	NP_035955.1	XP_419552.3	NP_001083746.1	NP_001017704.1	XP_007892101.1					
HAND1	NP_004812.1	NP_032239.1	NP_990296.1	NP_001079128.1		XP_007906292.1					
HAND2	NP_068808.1	NP_034532.3	NP_990297.2	NP_001079108.1	NP_571701.3	XP_007894139.1					
HAPLN1	NP_001875.1	NP_038528.3	NP_990813.1	XP_018099885.1	NP_001007791.2	XP_007897935.1					
HIPK2	NP_073577.3	NP_034563.2	XP_015143095.1	NP_001088409.1	NP_001093455.3	XP_007896772.1					
HOXA2	NP_006726.1	NP_034581.1	NP_990481.1	NP_001079219.1	NP_571181.1	XP_007898687.1					
HOXA3	NP_705895.1	NP_034582.1	XP_025002790.1	NP_001080293.1	NP_571609.1	XP_007898681.1					
HRAS	NP_005334.1	NP_032310.2	NP_990623	NP_001084278.1	NP_001017623	XP_007907536.1					

IDUA	NP_000194.2	NP_032351.2	NP_001026604.1	NP_001087031.1	XP_001923689.3	XP_007905890.1					

IFITM5	NP_001020466.1	NP_444318.1	NP_001186427.1	XP_018112954.1	NP_001170783.1	XP_007902982.1					
IHH	NP_002172.2	NP_034674.2	NP_990288.1	NP_001079262.1	NP_001030165.2	XP_007909552.1					
IRF6	NP_006138.1	NP_058547.2	XP_024999581.1	NP_001085345.1	NP_956892.1	XP_007897748.1					
ITGB1	NP_002202	NP_034708.1	NP_001034343.2	NP_001081286.1	NP_001030143.1	XP_007907125.1					
JAG2	NP_002217.3	NP_034718.2	XP_001235689.2	XP_018087787.1	NP_571937.1	XP_007909754.1					
KAT6A	NP_006757.2	NP_001074618.1	XP_424402.4	XP_018106960.1	NP_001116784.4	XP_007908442.1					
KAT6B	NP_036462.2	NP_059507.2	XP_004942127.3	XP_018110516.1	XP_697383.4	XP_007896926.1					
KCNJ2	NP_000882.1	NP_032451.1	NP_001310133.1	XP_018096315.1	NP_001318145.1	XP_007886827.1					
KDM4B	NP_055830.1	NP_742144.1	XP_025000603.1	XP_018099298.1	NP_001076274.1	XP_007899303.1					
KIF3A	NP_001287720.1	NP_032469.2	NP_001025793.1	NP_001084268	NP_001017604.2	XP_007900170.1					
KLF2	NP_057354.1	NP_032478.2	NP_001305352.1	NP_001080430.1	NP_571931.2	XP_007899255.1					
LIMK2	NP_005560.1	NP_034848.1	NP_990446.1	NP_001082219.1	NP_001002651.1	XP_007909768.1					
LOXL3	NP_115992.1	NP_038614.2	XP_015156777.1	XP_018098412.1	NP_001139156.1	XP_007908441.1					
LPAR4	NP_001264929.1	NP_780480.2	NP_001264978.1	NP_001090494.1	XP_001334713.1	XP_007890382.1					
LTBP3	NP_001123616.1	NP_032546.2	XP_025004462.1	XP_002935409.2	NP_001244070.1	XP_007884094.1					

MAPK1	NP_002736.3	NP_036079.1	NP_989481.1	NP_001081344.1	NP_878308.2	XP_007906800					

MMP14	NP_004986.1	NP_032634.3	XP_015147798.1	NP_001084478.1	NP_919397.1	XP_007901389.1					
MMP2	NP_004521.1	NP_032636.1	NP_989751.1	NP_001080697.1	NP_932333.1	XP_007887678.1					
MMP21	NP_671724.1	NP_694423.1	XP_015131123.1	NP_001079285.1	NP_001304682.1	XP_007902157.1					
MN1	NP_002421.3	NP_001074704.1	XP_003642254	XP_018099537.1	XP_001919775.2	XP_007906045.1					
MSX1	NP_002439.2	NP_034965.2	NP_990819.1	NP_001084367.1	NP_571348.1	XP_007895507.1					
NABP2	NP_076973.1	NP_081533.1	XP_025001296.1	NP_001084896.1	XP_005172861.1	XP_007888361.1					
NDST1	NP_001534.1	NP_032332.2	XP_004945002.1	XP_018107482.1	NP_001311384.1	XP_007906348.1					
NFATC1	NP_765975.1	NP_058071.2	XP_015131424.1	XP_018121962.1	NP_001038624.1	XP_007898768.1					
NFIC	NP_001231931.1	NP_032714.1	NP_990602.1	NP_001154859.1	XP_005166996.1	XP_007896879.1					
NPR2	NP_003986.2	NP_776149.1	XP_003642967.1	NP_001084176.1	XP_009300100.1	XP_007895384.1					
OTX2	NP_068374.1	NP_659090.1	NP_989851.2	NP_001084955.1	NP_571326.1	XP_007891617.1					
PAPPA2	NP_064714.2	NP_001078845.1	XP_025008932.1	XP_018088392.1	XP_695417.4	XP_007894444.1					
PAX1	NP_001244025.1	NP_032806.2	XP_015138914.2	NP_001090451.2	NP_001074061.1	XP_007896543.1					
PAX3	NP_852123.1	NP_032807.3	NP_989600.1	NP_001088993.1	NP_571352.1	XP_007887991.1					
PAX7	NP_001128726.1	NP_035169.1	XP_015152318.1	XP_018080096.1	XP_009304561.1	XP_007896063.1					

PAX9	NP_001359005.1	NP_035171.1	NP_990243.2	NP_001167485.1	NP_571373.1	XP_007891546.1					

PCSK6	NP_002561.1	NP_035178.1	XP_025009761.1	NP_001086467.1	XP_021333790.1	XP_007905602.1					
PDGFRB	NP_002600.1	NP_032835.2	XP_025010809.1	XP_018110448.1	NP_001177862.1	XP_007906332.1					
PITX1	NP_002644.4	NP_035227.1	XP_015149472.2	NP_001080981.1	NP_001035436.3	XP_007899502.1					
PITX2	NP_000316.2	NP_035228.2	NP_990341.1	NP_001017227.1	NP_571050.1	XP_007895188.1					
POSTN	NP_006466.2	NP_056599.1	XP_015133099.1	XP_018105718.1	NP_981966.1	XP_007889427.1					
POU3F3	NP_006227.1	NP_032926.2	XP_015133414.1	XP_018102574.1	NP_571177.2	XP_007883344.1					
PRDM16	NP_071397.3	NP_081780.3	XP_417551.3	XP_018083496.1	XP_001922927.6	XP_007900361.1					
PRICKLE1	NP_001138353.1	NP_001028389.1	XP_015137914.1	NP_001082157.1	NP_899185.2	XP_007893374.1					
PRKRA	NP_003681.1	NP_036001.1	XP_015155835.1	NP_001079500.1	NP_001082830.1	XP_007888271.1					
PRRX1	NP_073207.1	NP_035257.1	NP_001264653.1	XP_018114269.1	NP_999899.1	XP_007893876.1					
PTCH1	NP_000255.2	NP_032983.1	NP_990291.2	NP_001082082.1	NP_001292471.1	XP_007898170.1					
PTH1R	NP_000307.1	NP_035329.2	NP_001171046.1	XP_018122519.1	NP_571432.1	XP_007908388.1					
PTPN11	NP_002825.3	NP_035332.1	NP_990299.1	XP_018119081.1	NP_956140.1	XP_007900075					
RFNG	NP_002908.1	NP_033079.1	NP_990278.1	XP_018089606.1	NP_001001830.2	XP_007886742.1					
ROBO1	NP_002932.1	NP_062286.2	XP_015153967.1	NP_001084250.1	NP_001296753.1	XP_007901766.1					

ROR2	NP_004551.2	NP_038874.3	NP_001074185.1	XP_018099378.1	XP_689681.6	XP_007898225.1					

RSPO2	NP_848660.3	NP_766403.1	NP_001305953.1	NP_001088999.1	NP_001268919.1	XP_007900878.1					
SATB2	NP_001165980.1	NP_631885.1	NP_001186039.1	NP_001267548.1	NP_001122004.1	XP_007898262.1					
SETD5	NP_001073986.1	NP_082661.1	NP_001264946.2	XP_018114635.1	XP_698834.4	XP_007888440.1					
SFN	NP_006133.1	NP_061224.2	NP_001280105.1								
SH3PXD2B	NP_001017995.1	NP_796338.2	XP_425197.3	XP_018107654.1	XP_003200955.2	XP_007908750.1					
SHH	NP_000184.1	NP_033196.1	NP_990152.1	NP_001081782.1	NP_571138.1	XP_007906520.1					
SIRT1	NP_036370.2	NP_062786.1	gbAVG72379.1	XP_018082983.1	XP_001334440.4	XP_007895509.1					
SKI	NP_003027.1	NP_035515.2	XP_015152425.1	NP_001084112.1	NP_571010.3	XP_007901663.1					
SLC27A4	NP_005085.2	NP_036119.1	XP_015135039.1	XP_018096544.1	NP_001017737.1	XP_007898514.1					
SMAD2	NP_005892.1	NP_034884.2	NP_989892.1	NP_001084964.1	NP_571441.3	XP_007904996.1					
SMARCA4	NP_001122321.1	NP_001344693.1	NP_990390.1	NP_001165424.1	NP_853634.1	XP_007905293.1					
SOSTDC1	NP_056279.1	NP_079588.1	NP_989704.1	NP_001080917.1	NP_001076314.1	XP_007897897.1					
SOX11	NP_003099.1	NP_033260.4	NP_990518.1	NP_001135834.1	NP_571411.1	XP_007902941.1					
SOX9	NP_000337.1	NP_035578.3	NP_989612.1	NP_001084276.1	NP_571718.1	XP_007886826.1					
SPRY2	NP_001305465.1	NP_036027.1	NP_990131.1	NP_001006932.2	NP_001003842.2	XP_007900755.1					

SPRY4	NP_112226.2	NP_036028.2	NP_001073203.1	XP_018107617.1	NP_571901.2	XP_007884890.1					

TBX1	NP_542378.1	NP_035662.1	XP_025011481.1	NP_001083914.1	NP_899182.1	XP_007899995.1					
TBX15	NP_001317606.1	NP_033349.2	XP_416537.3	XP_018101281.1	NP_001315362.1	XP_007894068.1					
TCOF1	NP_001358552.1	NP_035682.1	XP_015149272.1	XP_018109911.1	XP_002664303.1						
TNFSF11	NP_003692.1	NP_035743.2	NP_001076830.1	XP_018105429.1	XP_002663397.2	XP_007892566.1					
TP53	NP_000537	NP_035770	NP_990595	NP_001081567	NP_001258749	AEW46988					
TP63	NP_003713.3	NP_001120731.1	NP_989682.1	XP_018121105.1	NP_694518.1	AEW46989.1					
VEGFA	NP_001020537.2	NP_001020421.2	NP_990373.1	NP_001079058.1	NP_571483.1	XP_007892525.1					
WNT5A	NP_003383.3	NP_033550.2	NP_990218.1	NP_001079345.1	NP_001073303.1	XP_007888715.1					
YBX1	NP_004550.2	NP_035862.2	XP_015152295.1	NP_001079367.1	NP_571695.1	NP_001279438.1					
ZEB1	NP_001121600.1	NP_035676.1	NP_990462.1	NP_001085962.1	NP_571784.1	XP_007903615.1					
ZIC3	NP_003404.1	NP_033601.2	XP_420237.5	NP_001081088.1	NP_001001950.1	XP_007890960.1					
ZIC5	NP_149123.2	NP_075363.1	XP_025009408.1	NP_001079126.1	NP_991290.1	XP_007904496.1					
ACVR1	NP_001104537.1	NP_031420.2	XP_015145613.1	NP_001083727.1	NP_571420.1	XP_007896316.1	XP_032822706.1				
ALKBH1	NP_006011.2	NP_001096035.1	NP_001026723.1	NP_001087992.1	NP_001018527.1	XP_007904240.1	XP_032827708.1				
ANKH	NP_473368.1	NP_065065.3	NP_001012580.1	NP_001083924.1	NP_919351.1	XP_007899922.1	XP_032807388.1		XP_002596000.1		

AP2B1	NP_001025177.1	NP_001030931.1	XP_415772	NP_001080473.1	NP_956213.2	XP_007894750.1	XP_032823633.1	XP_002122416.3	XP_002610773.1	XP_030848114	XP_012563108.1

APAF1	NP_037361.1	NP_001036023.1	XP_416167.3	NP_001085834.1	NP_571683.1	XP_007905933.1	XP_032812741.1		XP_002598365.1	XP_030844937.1	XP_012566478.1
APC	NP_000029.2	NP_031488.2	XP_004949340.1	XP_018099812.1	NP_001137312.1	XP_007895594.1	XP_032829901.1	XP_018668497.1	XP_002596549.1	XP_030829439.1	XP_012554796.1
ARSB	NP_000037.2	NP_033842.3	XP_003642960.1	XP_018120981.1	XP_021324609.1	XP_007890415.1	XP_032812086.1	XP_002127641.1	XP_002592276.1	XP_030842602.1	
ATMIN	NP_056066.2	NP_808368.3	NP_001362358.1	XP_018113369.1	NP_001093512.1	XP_007887411.1	XP_032816122.1	XP_002126586.1			XP_012554135.1
ATR	NP_001175.2	NP_063917.1	XP_015132446.1	XP_018117212.1	XP_696163.5	XP_007891182.1	XP_032809792.1	XP_018673125.1	XP_002594547.1	XP_030832385.1	XP_012559395.1
BAZ1B	NP_115784.1	NP_035844.2	XP_001233717.4	XP_018104542.1	NP_001334601.1	XP_007894624.1	XP_032803856.1	XP_018670952.1	XP_002591894.1	XP_030849368.1	XP_012565366.1
BNC2	NP_060107.3	NP_766458.3	XP_015135686.1	XP_018112777.1	NP_001314681.1	XP_007890780.1	XP_032803547.1	XP_009861392.1			
C5orf42	NP_075561.3	NP_001156378.1	XP_015133013.1	XP_018121837.1	XP_021334904.1	XP_007909187.1	XP_032826472.1	XP_026692817.1	XP_002594230.1	XP_030852004.1	
CASK	NP_001119527.1	NP_033936.2	XP_005301824.1	NP_001087107.1	NP_694420.1	XP_007899830.1	XP_032835117.1	XP_026690102.1	XP_002607558.1	XP_030837409.1	XP_002163427.3
CBFB	NP_074036.1	NP_071704.3	NP_989901.2	NP_001087047.1	NP_954679.1	NP_001279053.1	XP_032816492.1	NP_001071672.1	XP_002602848.1	NP_001035104.1	XP_004207755.1
CCDC39	NP_852091.1	NP_080498.1	XP_422777.4	XP_018119161.1	XP_001344625.2	XP_007905952.1	XP_032830900.1	XP_009860998.2	XP_002593455.1	XP_030837980.1	
CEP290	NP_079390.3	NP_666121.2	XP_004937659.1	NP_001085087.1	NP_001161739.1	XP_007893519.1	XP_032834088.1	XP_018670980.1	XP_002590663.1	XP_030844423.1	XP_012566704.1
CHRD	NP_003732.2	NP_034023.1	NP_990311.2	NP_001081778.1	NP_571048.2	XP_007900570.1	XP_032827777.1	XP_026690585.1	ABG66525.1	XP_030843065.1	XP_002158106.2
CNBP	NP_003409.1	NP_038521.1	NP_990238.1	NP_001084082.1	NP_001313334.1	XP_007901526.1	XP_032828325.1				
COL2A1	NP_001835.3	NP_112440.2	XP_025001042.1	NP_001081260.1	NP_571367.1	XP_007908719.1	XP_032813113.1	XP_026690723.1		NP_999674.1	

CTGF	NP_001892.1	NP_034347.2	NP_989605.1	NP_001081697.1	NP_001015041.1	XP_007897423.1	XP_032829951.1				

CYP26B1	NP_063938.1	NP_780684.1	XP_015141554.1	XP_018082655.1	NP_997831.1	XP_007904867.1	XP_032808017.1				
CYP51A1	NP_000777.1	NP_064394.2	XP_015136878.1	NP_001091285.1		XP_007895734.1	XP_032816590.1		XP_002606163.1	XP_030851345.1	
DAAM2	NP_001188356.1	NP_001008232.2	XP_025004642.1	XP_018118186.1	XP_001338908.5	XP_007909292.1	XP_032803809.1	XP_026692086.1	XP_002604562.1	XP_030855636.1	XP_012558320.1
DCTN5	NP_115875.1	NP_067621.3	NP_001185568.1	NP_001086487.1	NP_001002497.1	XP_007904011.1	XP_032832114.1	XP_002131615.1	XP_002605052.1	XP_030840147.1	XP_002155568.3
DHRS3	NP_004744.2	NP_035433.1	NP_001264839.1	XP_018083410.1	NP_001003477.1	XP_007885177.1	XP_032824210.1		XP_002586689.1		
DNAH5	NP_001360.1	NP_579943.3	XP_025003476.1	XP_018123134.1	XP_021322524.1	XP_007898716.1	XP_032816356.1	XP_018671051.1		XP_030831333.1	
DNAI1	NP_036276.1	NP_780347.2	XP_003643013.1	XP_018100839.1	XP_021324725.1	XP_007895318.1	XP_032809556.1	XP_002129294.1	XP_002609708.1	XP_030833320.1	XP_012559387.1
DPH1	NP_001374.4	NP_652762.2	NP_001025887.2	XP_012812029.1	NP_001017861.1	XP_007904382.1	XP_032823710.1	XP_026691241.1	XP_002605608.1	XP_030848155.1	XP_002161454.2
DRC1	NP_659475.2	NP_001028632.1	XP_420016.3	NP_001079840.1	NP_001120940.1	XP_007898606.1	XP_032817535.1	XP_004226120.4		XP_030830868.1	
DYNC2H1	NP_001368.2	NP_084127.2	XP_417173.2	XP_018101055.1	XP_021322179.1	XP_007902643.1	XP_032832089.1	XP_026693267.1	XP_019625235.1	XP_030836572.1	XP_012558919.1
EDN1	NP_001946.3	NP_034234.1	XP_418943.2	NP_001090567.1	NP_571594.1	XP_007904565.1	XP_032817360.1				
EDNRA	NP_001948.1	NP_034462.1	NP_989450.1	NP_001080650.1	NP_001092915.2	XP_007897347.1	XP_032811286.1				
EDNRB	NP_001116131.1	NP_031930.1	NP_001001127.1	NP_001079707.1	NP_571272.1	XP_007900764.1	XP_032812242.1				
EPS8L3	NP_573444.2	NP_598628.1	XP_024999597.1	XP_018096619.1	NP_001315008.1	XP_007897795.1	XP_032833964.1	XP_026694557.1	XP_002586560.1	XP_030849025.1	XP_012557569.1
FAM20C	NP_064608.2	NP_085042.2	XP_015149804.2	XP_018091900.1	XP_688892.5	XP_007891748.1	XP_032822771.1	XP_002122861.4	XP_002606109.1	XP_030839755.1	

FGFRL1	NP_001004356.1	NP_473412.1	XP_025005238.1	NP_001086420.1	NP_956670.1	XP_007896240.1	XP_032806318.1	XP_026690297.1	CAI61931.1	NP_001165523.1	

FOXD3	NP_036315.1	NP_034555.3	NP_990282.2	NP_001079026.1	NP_571365.2	XP_007907140.1	XP_032822101.1	XP_009859457.3	XP_002602560.1	XP_003724544.1	
FOXH1	NP_003914.1	NP_032015.1	XP_015137671.2	NP_001081820.1	NP_571577.1	XP_007910257.1	XP_032834948.1	NP_001071992.1	ACE79158.1		
FRAS1	NP_079350.5	NP_780682.3	XP_025006059.1	XP_018124333.1	NP_001124312.1	XP_007895210.1	XP_032805666.1	XP_026692906.1	XP_002601902.1	XP_030835000.1	XP_012562678.1
FUZ	NP_079405.2	NP_081652.2	XP_015151679.1	XP_018081840.1	XP_001337543.1	XP_007910217.1	XP_032833903.1	XP_002130176.1	XP_002604807.1	XP_003723986.2	XP_002154880.2
GAD1	NP_000808.2	NP_032103.2	NP_990244.1	XP_018089468.1	NP_919400.1	XP_007888154.1	XP_032821372.1	XP_026696077.1	XP_002592141.1	XP_784856.2	
GAS1	NP_002039.2	NP_032112.1	XP_004949280.1	XP_018099363.1	NP_001071269.1	XP_007906242.1	XP_032824729.1	XP_002123239.1			
GJB2	NP_003995.2	NP_032151	NP_001257745.1	NP_001080478.1	NP_997990.2	XP_007903773.1	XP_032828672.1				
GLG1	NP_001139139.1	NP_033175.1	NP_990827.1	XP_018116137.1	XP_001332484.4	XP_007902868.1	XP_032830405.1	XP_026689671.1	XP_002612434.1	XP_030831332.1	XP_012560055.1
GRB2	NP_002077.1	NP_032189.1	NP_989742.1	NP_001084357.1	NP_998200.1	XP_007886673.1	XP_032826167.1	XP_002125384.4	XP_019642516.1	XP_001193089.2	XP_002166949.2
GSC	NP_776248.1	NP_034481.1	NP_990662.1	NP_001081278.1	NP_571092.1	XP_007910192.1	XP_032820078.1	NP_001071728.1	AAF97935.1	NP_999663.1	NP_001296713.1
HAT1	NP_003633.2	NP_080391.2	NP_989538.1	XP_018094344.1	NP_001004572.1	XP_007908137.1	XP_032821290.1	XP_002124657.1	XP_002602992.1	XP_030842939.1	XP_012564660.1
HHAT	NP_060664.2	NP_659130.2	XP_001234561.1	XP_018118052.1	NP_001108362.2	XP_007890669.1	XP_032818807.1	XP_002130477.1	XP_002605788.1	XP_030854460.1	XP_012559282.1
HSPB11	NP_057210.2	NP_082670.1	NP_001264542.1	NP_001087826.1	NP_001092897.1	XP_007907160.1			XP_002603858.1	XP_030852059.1	
HSPG2	NP_005520.4	XP_030109088.1	XP_015152746	XP_018081990.1	XP_009295216.1	XP_007896099	XP_032824291.1	XP_018666843.1	XP_019637175.1	XP_030842937.1	XP_004211405.2
IFT140	NP_055529.2	NP_598887.3	NP_001012810.2	XP_018091986.1	XP_695732.4	XP_007891874.1	XP_032824144.1	XP_002125448.2	XP_002607214.1	XP_030840366.1	XP_012555320.1

IFT27	NP_001171172.1	NP_080207.1	NP_001264310.1	NP_001090381.1	NP_001008588.1	NP_001279585.1	XP_032817047.1	XP_009858222.1	XP_002588733.1	XP_030847703.1	XP_004212778.2

IFT57	NP_060480.1	NP_082956.2	XP_416624.4	NP_001108268.1	NP_001001832.1	XP_007895758.1	XP_032822969.1	XP_002121351.1	XP_002604500.1	XP_030833864.1	XP_002154458.3
IKBKAP	NP_003631.2	NP_080355.2	XP_001231708.2	XP_018095285.1	XP_689534.5	XP_007904786.1	XP_032824390.1	XP_009859423.1	XP_002597833.1	XP_030832738.1	XP_012557166.1
JAG1	NP_000205.1	NP_038850.1	XP_415035.4	NP_001083776.1	NP_571936.1	XP_007903695.1	XP_032803601.1				
KAT14	NP_065397.3	NP_852082.2	NP_001186370.1	XP_018118382.1	NP_001002699.2	XP_007892627.1	XP_032804461.1	XP_002121916.1	XP_002602499.1	XP_030828197.1	
LMNA	NP_733821.1	NP_062263.1	NP_990618.1	NP_001095210.1	NP_694503.1	XP_007906490.1	XP_032806857.1				
LRP2	NP_004516.2	NP_001074557.1	XP_004942820.1	XP_018091370.1	NP_001181916.1	XP_007888391.1	XP_032801941.1	XP_026691732.1	XP_002598410.1	XP_030830689.1	
LRRK1	NP_078928.3	NP_666303.3	NP_001274122.2	XP_018108760.1	XP_021333791.1	XP_007905552.1	XP_032806825.1	XP_026690738.1	XP_002609807.1	XP_030832414.1	XP_012555867.1
LTBP1	NP_996826.2	NP_064303.2	XP_015139206.1	XP_018118508.1	XP_017207300.1	XP_007896642.1	XP_032803602.1				
MAP3K7	NP_663304.1	NP_033342.1	XP_015140169.1	NP_001084359.1	NP_001018586.1	XP_007891110.1	XP_032829518.1	NP_001071829.1	XP_002589769.1	XP_030850465.1	XP_002162138.2
MAU2	NP_056144.3	NP_083269.4	XP_425908.2	NP_001124425.1	XP_021322113.1	XP_007897013.1	XP_032829637.1		XP_002609698.1	XP_030833209.1	XP_012560519.1
MBTPS1	NP_003782.1	NP_062683.3	XP_003641945.1	NP_001087381.2	NP_954683.2	XP_007902889.1	XP_032813570.1	XP_002122807.1	XP_002594247.1	XP_030831757.1	XP_002164301.2
MECP2	NP_001104262.1	NP_034918.1	XP_025001604.1	NP_001081854.1	NP_997901.1	XP_007909886.1	XP_032821181.1				
MEGF8	NP_001258867.1	NP_001153872.1	XP_015144416.1	XP_018081796.1	XP_005158088.1	XP_007909276.1	XP_032835707.1	XP_026689375.1	XP_002613359.1	XP_030834386.1	XP_012556758.1
MKS1	NP_060247.2	NP_001034773.2	XP_415705.5	NP_001090590.1	NP_001070841.2	XP_007894621.1	XP_032823388.1	XP_018669336.1	XP_002605664.1	XP_011678596.1	XP_002157754.2
MNT	NP_064706.1	NP_034943.3	XP_015151357.1	XP_018104601.1	NP_001096581.1	XP_007898900.1	XP_032823523.1	XP_009861486.1	XP_002595964.1	XP_003728085.2	XP_012561962.1

MYH10	NP_001242941.1	NP_780469.1	XP_015150613.1	XP_018092877.1	XP_683046.6	XP_007886649.1	XP_032817139.1				

NIPBL	NP_597677.2	NP_081983.2	XP_015133005.1	NP_001086437.1	NP_001154919.2	XP_007909505.1	XP_032801038.1	XP_018671410.1	XP_002587265.1	XP_030852099.1	XP_004213193.2
NOG	NP_005441.1	NP_032737.1	NP_989454.1	XP_018096035.1	NP_571057.1	XP_007891794.1	XP_032828422.1	NP_001071777.1	XP_002595406.1	XP_030855638.1	NP_001296696.1
OSTM1	NP_054747.2	NP_766004.1	NP_001026248.1	XP_018118820.1	XP_002665788.2	NP_001279106.1	XP_032832500.1	XP_002128956.1			XP_002169949.2
OTULIN	NP_612357.4	NP_001013814.2	XP_004935161.3	XP_018124896.1	XP_001341618.4	XP_007899898.1	XP_032823311.1	XP_018667571.1	XP_002598365.1		
PAK1IP1	NP_060376.2	NP_080826.2	NP_001026170.1	NP_001087421.1	NP_957239.2	XP_007899046.1	XP_032808637.1	XP_026691645.1	XP_002609652.1	XP_795222.3	XP_012559538.1
PAX6	NP_000271.1	NP_038655.1	NP_990397.1	XP_018111894.1	NP_571379.1	XP_007885968.1	XP_032805179.1	NP_001027641.1	CAA11364.1	XP_030845154.1	
PDPK1	NP_002604.1	NP_035192.2	NP_001012547.1	XP_018094630.1	NP_991262.1	XP_007904001.1	XP_032822965.1	XP_002124476.1	XP_002606093.1	XP_786576.2	NP_001267831.1
PDS5B	NP_055847.1	NP_780519.3	NP_001012845.2	NP_001089658.1	XP_693953.2	XP_007889466.1	XP_032811251.1				
PDSS2	NP_065114.3	NP_082048.2	NP_001305370.1	XP_018120896.1	NP_001002351.1	NP_001279445.1	XP_032815024.1	XP_026695230.1	XP_002589797.1	XP_011670674.2	XP_002156194.1
PFAS	NP_036525.1	NP_001152991.1	XP_024999398.1	XP_018094887.1	NP_001038667.1	XP_007893212.1	XP_032832973.1	XP_018671183.1	XP_002611349.1	XP_030836182.1	XP_004212171.2
PGAP1	NP_079265.2	NP_001156786.1	XP_015144976.1	NP_001087710.1	XP_001922432.2	XP_007888374.1	XP_032820600.1	XP_002131111.1	XP_002600834.1	XP_030840100.1	XP_012555278.1
PHEX	NP_000435.3	NP_035207.1	XP_015157829.1	XP_018101754.1	NP_001082818.2	XP_007908778.1	XP_032820142.1				
PIGV	NP_060307.2	NP_848813.3	XP_024998857.1	XP_018104933.1	NP_001186989.1	XP_007893122.1	XP_032829362.1	XP_002122985.2	XP_002612582.1	XP_030848488.1	XP_012559454.1
PKD1	NP_001009944.3	NP_038658.2	XP_414854.4	XP_018092070.1	XP_021333081.1	XP_007908643.1	XP_032832098.1	XP_018669270.1	XP_002595676.1		
PKDCC	NP_612379.2	NP_598878.2	XP_025004831.1	NP_001091231.2	NP_001155070.1	XP_007890699.1	XP_032811876.1	XP_002128638.1	XP_002590129.1	XP_030854909.1	

PLXND1	NP_055918.3	NP_080652.2	XP_414455.4	XP_018114545.1	NP_991260.2	XP_007888683.1	XP_032822371.1				

PTHLH	NP_945316.1	NP_032996.2	XP_025002654.1	XP_018107394.1	NP_001019798.2	XP_007903315.1	XP_032804916.1				
PTK7	NP_002812.2	NP_780377.1	NP_001026206.1	NP_001083315.1	XP_021336598.1	XP_007896436.1	XP_032811886.1	XP_018669704.1		XP_030855357.1	
RPGRIP1L	NP_056087.2	NP_775607.2	XP_025010081.1	XP_018113254.1	NP_001233589.2	XP_007887654.1	XP_032834838.1	XP_018666898.1	XP_002612308.1	XP_030832105.1	XP_012559558.1
RYK	NP_002949.2	NP_038677.3	XP_422674.6	NP_001089445.1	NP_001077295.1	XP_007908285.1	XP_032836167.1	XP_002129769.1		XP_011665108.1	XP_012558645.1
SC5D	NP_008849.2	NP_766357.1	XP_015153662.1	NP_001087137.1	NP_001004630.1	XP_007909349.1	XP_032820533.1	XP_002131054.1	XP_002590254.1	XP_030831120.1	
SCHIP1	NP_055390.1	NP_038956.2	XP_004943577.2	XP_018119243.1	NP_001017906.1	XP_007886672.1	XP_032830867.1		XP_002599883.1	XP_030838173.1	
SH3BP2	NP_001116153.1	NP_036023.3	XP_015141267.1	XP_018096046.1	XP_005155638.1	XP_007903445.1	XP_032800705.1	XP_009861676.2	XP_002608882.1	XP_030848580.1	
SIX1	NP_005973.1	NP_033215.2	NP_001038150.1	NP_001089925.1	NP_996978.1	XP_007902022.1	XP_032803935.1				
SLC25A21	NP_085134.1	NP_766165.2	XP_015143145.1	NP_001090527.1	NP_001070100.1	XP_007891544.1	XP_032804225.1	XP_002130378.1		XP_030854904.1	
SLC26A2	NP_000103.2	NP_031911.1	XP_025010793.1	XP_018111616.1	XP_685114.1	XP_007910486.1	XP_032806612.1				
SLC39A13	NP_001121697.2	NP_080997.1	NP_001008471.1	XP_018112898.1	NP_001005306.3	NP_001279227.1	XP_032812053.1	XP_002120593.1	XP_002609106.1	XP_030841333.1	XP_012554328.1
SMG1	NP_055907.3	NP_001026984.1	XP_414907.4	XP_018094605.1	NP_001073513.2	XP_007892012.1	XP_032817005.1	XP_018672013.2	XP_002607176.1	XP_030839843.1	XP_012553895.1
SMO	NP_005622.1	NP_795970.3	XP_015143674.1	NP_001128704.1	NP_571102.2	XP_007889054.1	XP_032809666.1	NP_001071819.1	XP_019641686.1	XP_030830155.1	
SMOC1	NP_001030024.1	NP_071711.2	XP_001231761.2	XP_018084047.1	NP_001188322.1	XP_007891616.1	XP_032811759.1				
SRF	NP_003122.1	NP_065239.1	NP_001239070.1	XP_018120351.1	NP_001103996.1	XP_007896558.1	XP_032820927.1	NP_001071826.1	XP_002610547.1	XP_030854625.1	XP_002156545.1

SUFU	NP_057253.2	NP_056567.2	NP_989595.1	XP_018082933.1	NP_958466.1	XP_007909051.1	XP_032808399.1	XP_002129739.1	XP_019625825.1	XP_796446.2	XP_012560454.1

TFAP2A	NP_001358995.1	NP_035677.3	NP_990425.1	NP_001089041.1	NP_789829.2	XP_007899085.1	XP_032810894.1	NP_001071986.1	XP_002593633.1	XP_030839379.1	
TGFB2	NP_001129071.1	NP_033393.2	NP_001026216.2	NP_001079195.1	NP_919366.1	XP_007902706.1	XP_032818531.1				
TGFBR2	NP_003233.4	NP_033397.3	NP_990759.1	XP_018122960.1	NP_878275.2	XP_007887981.1	XP_032812992.1	NP_001071836.1	XP_002612985.1	XP_030828751.1	
TIPARP	NP_056323.2	NP_849223.2	XP_015147299.1	XP_018119254.1	NP_001038484.1	XP_007886976.1	XP_032824333.1				
TMEM67	NP_714915.3	NP_808529.3	XP_418334.3	XP_018079349.1	XP_700974.4	XP_007900910.1	XP_032813422.1	XP_009862218.1	XP_002593941.1	XP_030831148.1	XP_012554486.1
TNFRSF11B	NP_002537.3	NP_032790.3	NP_001028813.1	XP_018123714.1	;NP_001036153.2	XP_007905722.1	XP_032810850.1		XP_002610826.1		
TRPS1	NP_054831.2	NP_114389.2	XP_025003908.1	XP_018124131.1	NP_001171405.1	XP_007895696.1	XP_032817392.1				
TTC28	NP_001138890.1	NP_001254551.1	XP_015130973.1	XP_018099538.1	XP_017213483.1	XP_007908725.1	XP_032803281.1	XP_026690437.1	XP_002596602.1	XP_030847701.1	XP_012559564.1
TWSG1	NP_065699.1	NP_075540.1	XP_025002764.1	NP_001079870.1	NP_571872.2	XP_007895440.1	XP_032813556.1	XP_018672191.1	XP_002593913.1		
VPS25	NP_115729.1	NP_081052.2	XP_003642862.1	NP_001087424.1	NP_001082831.2	XP_007895107.1	XP_032831290.1	XP_002126561.1	XP_002598439.1	XP_030830088.1	XP_002156482.1
WDR19	NP_079408.3	NP_700440.2	XP_015141146.2	XP_018084648.1	XP_021331990.1	XP_007887304.1	XP_032811146.1	XP_002124215.1	XP_002597830.1	XP_030836953.1	XP_012557446.1
WNT3A	NP_149122.1	NP_033548.1	NP_001165072.1	NP_001079343.1	NP_001007186.1	XP_007905574.1	XP_032802062.1	XP_009859675.1	XP_019618001.1	XP_030830597.1	
XYLT1	NP_071449.1	NP_783576.2	XP_015150019.1	NP_001085934.1	NP_001165868.1	XP_007892011.1	XP_032816793.1				
ZMPSTE24	NP_005848.2	NP_766288.1	XP_417720.1	NP_001087923.1	NP_956186.1	XP_007892978.1	XP_032800885.1	XP_004226161.1	XP_002599069.1	XP_030850470.1	XP_002154940.2
ZNF366	NP_689838.1	NP_001004149.1	XP_429153.4	XP_018120574.1	XP_698186.3	XP_007897984.1	XP_032824509.1		XP_002592273.1		

8.3. References (TABLE 2)

1. Zeng S, Zhang Y, Ma J, Deng G, Qu Y, Guo C, Han Y, Yin L, Cai C, Li Y: BMP4 promotes metastasis of hepatocellular carcinoma by an induction of epithelial–mesenchymal transition via upregulating ID2. *Cancer letters* 2017, 390:67-76.

2. Guo D, Huang J, Gong J: Bone morphogenetic protein 4 (BMP4) is required for migration and invasion of breast cancer. *Molecular and cellular biochemistry* 2012, 363:179-190.

3. Ampuja M, Alarmo E, Owens P, Havunen R, Gorska A, Moses H, Kallioniemi A: The impact of bone morphogenetic protein 4 (BMP4) on breast cancer metastasis in a mouse xenograft model. *Cancer letters* 2016, 375:238-244.

4. Rothhammer T, Poser I, Soncin F, Bataille F, Moser M, Bosserhoff A-K: Bone morphogenic proteins are overexpressed in malignant melanoma and promote cell invasion and migration. *Cancer research* 2005, 65:448-456.

5. Deng H, Makizumi R, Ravikumar T, Dong H, Yang W, Yang W-L: Bone morphogenetic protein-4 is overexpressed in colonic adenocarcinomas and promotes migration and invasion of HCT116 cells. *Experimental cell research* 2007, 313:1033-1044.

6. Zhou Y, Liu Y, Zhang J, Yu D, Li A, Song H, Zhang W, Davis D, Gilbert MR, Liu F: Autocrine BMP4 Signaling Enhances Tumor Aggressiveness via Promoting Wnt/β-Catenin Signaling in IDH1-mutant Gliomas. *Translational Oncology* 2020, 13:125-134.

7. Sachdeva R, Wu M, Johnson K, Kim H, Celebre A, Shahzad U, Graham MS, Kessler JA, Chuang JH, Karamchandani J: BMP signaling mediates glioma stem cell quiescence and confers treatment resistance in glioblastoma. *Scientific reports* 2019, 9:1-14.

8. Eckhardt BL, Cao Y, Redfern AD, Chi LH, Burrows AD, Roslan S, Sloan EK, Parker BS, Loi S, Ueno NT: Activation of canonical BMP4-SMAD7 signaling suppresses breast cancer metastasis. *Cancer Research* 2020.

9. Hayashi T, Oue N, Sakamoto N, Anami K, Oo HZ, Sentani K, Ohara S, Teishima J, Matsubara A, Yasui W: Identification of transmembrane protein in prostate cancer by the Escherichia coli ampicillin secretion trap: expression of CDON is involved in tumor cell growth and invasion. *Pathobiology* 2011, 78:277-284.

10. He X, Lei S, Zhang Q, Ma L, Li N, Wang J: Deregulation of cell adhesion molecules is associated with progression and poor outcomes in endometrial cancer: Analysis of The Cancer Genome Atlas data. *Oncology letters* 2020, 19:1906-1914.

11. Delloye-Bourgeois C, Gibert B, Rama N, Delcros J-G, Gadot N, Scoazec J-Y, Krauss R, Bernet A, Mehlen P: Sonic Hedgehog promotes tumor cell survival by inhibiting CDON pro-apoptotic activity. *PLoS biology* 2013, 11.

12. Brand TM, Iida M, Stein AP, Corrigan KL, Braverman CM, Luthar N, Toulany M, Gill PS, Salgia R, Kimple RJ: AXL mediates resistance to cetuximab therapy. *Cancer research* 2014, 74:5152-5164.

13. Ciardiello F, De Vita F, Orditura M, Tortora G: The role of EGFR inhibitors in nonsmall cell lung cancer. *Current opinion in oncology* 2004, 16:130-135.

14. Hu S, Fu W, Li T, Yuan Q, Wang F, Lv G, Lv Y, Fan X, Shen Y, Lin F: Antagonism of EGFR and Notch limits resistance to EGFR inhibitors and radiation by decreasing tumor- initiating cell frequency. *Science translational medicine* 2017, 9:eaag0339.

15. Lim S-O, Li C-W, Xia W, Lee H-H, Chang S-S, Shen J, Hsu JL, Raftery D, Djukovic D, Gu H: EGFR signaling enhances aerobic glycolysis in triple-negative breast cancer cells to promote tumor growth and immune escape. *Cancer research* 2016, 76:1284-1296.

16. Boone B, Jacobs K, Ferdinande L, Taildeman J, Lambert J, Peeters M, Bracke M, Pauwels P, Brochez L: EGFR in melanoma: clinical significance and potential therapeutic target. *Journal of cutaneous pathology* 2011, 38:492-502.

17. Ueno Y, Sakurai H, Tsunoda S, Choo MK, Matsuo M, Koizumi K, Saiki I: Heregulininduced activation of ErbB3 by EGFR tyrosine kinase activity promotes tumor growth and metastasis in melanoma cells. *International journal of cancer* 2008, 123:340-347.

18. Gan Y, Shi C, Inge L, Hibner M, Balducci J, Huang Y: Differential roles of ERK and Akt pathways in regulation of EGFR-mediated signaling and motility in prostate cancer cells. *Oncogene* 2010, 29:4947-4958.

19. Kelly MP, Lee ST, Lee FT, Smyth FE, Davis ID, Brechbiel MW, Scott AM: Therapeutic efficacy of 177Lu-CHX-A ″-DTPA-hu3S193 radioimmunotherapy in prostate cancer is enhanced by EGFR inhibition or docetaxel chemotherapy. *The Prostate* 2009, 69:92-104.

20. The drugs.com. https://www.drugs.com/drug-class/egfr-inhibitors.html. Accessed 1 May 2020

21. Wei W, Mok SC, Oliva E, Kim S-h, Mohapatra G, Birrer MJ: FGF18 as a prognostic and therapeutic biomarker in ovarian cancer. *The Journal of clinical investigation* 2013, 123:4435-4448.

22. Wei W, Mohapatra G, Birrer MJ: The FGF18/FGFR4 amplicon: Novel therapeutic biomarkers for ovarian cancer. AACR; 2012.

23. Yu Z, Lou L, Zhao Y: Fibroblast growth factor 18 promotes the growth, migration and invasion of MDA-MB-231 cells. *Oncology reports* 2018, 40:704-714.

24. Chen T, Gong W, Tian H, Wang H, Chu S, Ma J, Yang H, Cheng J, Liu M, Li X: Fibroblast growth factor 18 promotes proliferation and migration of H460 cells via the ERK and p38 signaling pathways. *Oncology reports* 2017, 37:1235-1242.

25. Zhang J, Zhou Y, Huang T, Wu F, Pan Y, Dong Y, Wang Y, Chan AK, Liu L, Kwan JS: FGF18, a prominent player in FGF signaling, promotes gastric tumorigenesis through autocrine manner and is negatively regulated by miR-590-5p. *Oncogene* 2019, 38:33-46.

26. Teishima J, Yano S, Shoji K, Hayashi T, Goto K, Kitano H, Oka K, Nagamatsu H, Matsubara A: Accumulation of FGF9 in prostate cancer correlates with epithelial-to- mesenchymal transition and induction of VEGF-A expression. *Anticancer research* 2014, 34:695-700.

27. Cui G, Shao M, Gu X, Guo H, Zhang S, Lu J, Ma H: The value of FGF9 as a novel biomarker in the diagnosis of prostate cancer. *Artificial cells, nanomedicine, and biotechnology* 2019, 47:2241-2245.

28. Sun C, Fukui H, Hara K, Zhang X, Kitayama Y, Eda H, Tomita T, Oshima T, Kikuchi S, Watari J: FGF9 from cancer-associated fibroblasts is a possible mediator of invasion and anti-apoptosis of gastric cancer cells. *BMC cancer* 2015, 15:333.

29. Wang J, Tan X, Guo Q, Lin X, Huang Y, Chen L, Zeng X, Li R, Wang H, Wu X: FGF9 inhibition by a novel binding peptide has efficacy in gastric and bladder cancer per se and reverses resistance to cisplatin. *Pharmacological Research* 2020, 152:104575.

30. Bhattacharya R, Ray Chaudhuri S, Roy SS: FGF9-induced ovarian cancer cell invasion involves VEGF-A/VEGFR2 augmentation by virtue of ETS1 upregulation and metabolic reprogramming. *Journal of cellular biochemistry* 2018, 119:8174-8189.

31. Huang Y, Jin C, Hamana T, Liu J, Wang C, An L, McKeehan WL, Wang F: Overexpression of FGF9 in prostate epithelial cells augments reactive stroma formation and promotes prostate cancer progression. *International journal of biological sciences* 2015, 11:948.

32. Matsuda Y, Ishiwata T, Yamahatsu K, Kawahara K, Hagio M, Peng W-X, Yamamoto T, Nakazawa N, Seya T, Ohaki Y: Overexpressed fibroblast growth factor receptor 2 in the invasive front of colorectal cancer: a potential therapeutic target in colorectal cancer. *Cancer letters* 2011, 309:209-219.

33. Matsuda Y, Hagio M, Seya T, Ishiwata T: Fibroblast growth factor receptor 2 IIIc as a therapeutic target for colorectal cancer cells. *Molecular cancer therapeutics* 2012, 11:2010-2020.

34. Wang J, Liu X-t, Huang H, Xiao G, Zhou Z-y, Chen Y, Yu Z-h, He S-l, Chen A-a, Wang D-d: Antitumor activity of a recombinant soluble ectodomain of mutant human fibroblast growth factor receptor-2 IIIc. *Molecular cancer therapeutics* 2011, 10:1656-1666.

35. Kim HS, Kim JH, Jang HJ, Han B, Zang DY: Pathological and Prognostic Impacts of FGFR2 Overexpression in Gastric Cancer: A Meta-Analysis. *Journal of Cancer* 2019, 10:20.

36. Wein L, Savas P, Van Geelen C, Caramia F, Moodie K, Joshi S, Loi S: FGFR2 amplification in metastatic hormone-positive breast cancer and response to an mTOR inhibitor. *Annals of Oncology* 2017, 28:2025-2027.

37. The U.S. Food and Drug Administration. https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/213736s000lbl.pdf. Accessed 1 May 2020.

38. Montazeri K, Bellmunt J: Erdafitinib for the treatment of metastatic bladder cancer. *Expert Review of Clinical Pharmacology* 2020.

39. Qing J, Du X, Chen Y, Chan P, Li H, Wu P, Marsters S, Stawicki S, Tien J, Totpal K: Antibody-based targeting of FGFR3 in bladder carcinoma and t (4; 14)-positive multiple myeloma in mice. *The Journal of clinical investigation* 2009, 119:1216-1229.

40. Liu X, Jing X, Cheng X, Ma D, Jin Z, Yang W, Qiu W: FGFR3 promotes angiogenesis-dependent metastasis of hepatocellular carcinoma via facilitating MCP-1-mediated vascular formation. *Medical Oncology* 2016, 33:46.

41. Thiyagarajan S, Bhatia N, Reagan-Shaw S, Cozma D, Thomas-Tikhonenko A, Ahmad N, Spiegelman VS: Role of GLI2 transcription factor in growth and tumorigenicity of prostate cells. *Cancer research* 2007, 67:10642-10646.

42. Huang D, Wang Y, Xu L, Chen L, Cheng M, Shi W, Xiong H, Zalli D, Luo S: GLI2 promotes cell proliferation and migration through transcriptional activation of ARHGEF16 in human glioma cells. *Journal of Experimental & Clinical Cancer Research* 2018, 37:1-17.

43. GLI2-mediated melanoma invasion and metastasis

44. Faiao-Flores F, Alves-Fernandes D, Pennacchi PC, Sandri S, Vicente ALSA, Scapulatempo-Neto C, Vazquez VdL, Reis RM, Chauhan J, Goding C: Targeting the hedgehog transcription factors GLI1 and GLI2 restores sensitivity to vemurafenib-resistant human melanoma cells. *Oncogene* 2017, 36:1849-1861.

45. Zhang D, Cao L, Li Y, Lu H, Yang X, Xue P: Expression of glioma-associated oncogene 2 (Gli 2) is correlated with poor prognosis in patients with hepatocellular carcinoma undergoing hepatectomy. *World journal of surgical oncology* 2013, 11:25.

46. Ruan Q, Fang Z-y, Cui S-z, Zhang X-l, Wu Y-b, Tang H-s, Tu Y-n, Ding Y: Thermo-chemotherapy Induced miR-218 upregulation inhibits the invasion of gastric cancer via targeting Gli2 and E-cadherin. *Tumor Biology* 2015, 36:5807-5814.

47. Raven PA, Lysakowski S, Tan Z, D'Costa NM, Moskalev I, Frees S, Struss W, Matsui Y, Narita S, Buttyan R: Inhibition of GLI2 with antisense-oligonucleotides: A potential therapy for the treatment of bladder cancer. *Journal of cellular physiology* 2019, 234:20634-20647.

48. Nagao-Kitamoto H, Nagata M, Nagano S, Kitamoto S, Ishidou Y, Yamamoto T, Nakamura S, Tsuru A, Abematsu M, Fujimoto Y: GLI2 is a novel therapeutic target for metastasis of osteosarcoma. *International journal of cancer* 2015, 136:1276-1284.

49. Rodrigues MFSD, Miguita L, De Andrade NP, Heguedusch D, Rodini CO, Moyses RA, Toporcov TN, Gama RR, Tajara EE, Nunes FD: GLI3 knockdown decreases stemness, cell proliferation and invasion in oral squamous cell carcinoma. *International journal of oncology* 2018, 53:2458-2472.

50. Fu L, Wu H, Cheng SY, Gao D, Zhang L, Zhao Y: Set7 mediated Gli3 methylation plays a positive role in the activation of Sonic Hedgehog pathway in mammals. *Elife* 2016, 5:e15690.

51. Trnski D, Sabol M, Gojević A, Martinić M, Ozretić P, Musani V, Ramić S, Levanat S: GSK3β and Gli3 play a role in activation of Hedgehog-Gli pathway in human colon cancer—Targeting GSK3β downregulates the signaling pathway and reduces cell proliferation. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2015, 1852:2574-2584.

52. Capurro MI, Xiang Y-Y, Lobe C, Filmus J: Glypican-3 promotes the growth of hepatocellular carcinoma by stimulating canonical Wnt signaling. *Cancer research* 2005, 65:6245-6254.

53. Peters M, Farias E, Colombo L, Filmus J, Puricelli L, de Kier Joffé EB: Inhibition of invasion and metastasis by glypican-3 in a syngeneic breast cancer model. *Breast cancer research and treatment* 2003, 80:221-232.

54. Castillo LF, Tascón R, Huvelle MAL, Novack G, Llorens MC, dos Santos AF, Shortrede J, Cabanillas AM, de Kier Joffé EB, Labriola L: Glypican-3 induces a mesenchymal to epithelial transition in human breast cancer cells. *Oncotarget* 2016, 7:60133.

55. Liu Y, Zheng D, Liu M, Bai J, Zhou X, Gong B, Lü J, Zhang Y, Huang H, Luo W: Downregulation of glypican-3 expression increases migration, invasion, and tumorigenicity of human ovarian cancer cells. *Tumor Biology* 2015, 36:7997-8006.

56. Han S, Ma X, Zhao Y, Zhao H, Batista A, Zhou S, Zhou X, Yang Y, Wang T, Bi J: Identification of Glypican-3 as a potential metastasis suppressor gene in gastric cancer. *Oncotarget* 2016, 7:44406.

57. Guo M, Zhang H, Zheng J, Liu Y: Glypican-3: A New Target for Diagnosis and Treatment of Hepatocellular Carcinoma. *Journal of Cancer* 2020, 11:2008.

58. Zhang L, Zou W: Inhibition of integrin β1 decreases the malignancy of ovarian cancer cells and potentiates anticancer therapy via the FAK/STAT1 signaling pathway. *Molecular medicine reports* 2015, 12:7869-7876.

59. Mori R, Ishiguro H, Kuwabara Y, Kimura M, Mitsui A, Tomoda K, Mori Y, Ogawa R, Katada T, Harata K: Targeting β1 integrin restores sensitivity to docetaxel of esophageal squamous cell carcinoma. *Oncology reports* 2008, 20:1345-1351.

60. Xu Z, Zou L, Ma G, Wu X, Huang F, Feng T, Li S, Lin Q, He X, Liu Z: Integrin β1 is a critical effector in promoting metastasis and chemo-resistance of esophageal squamous cell carcinoma. *American journal of cancer research* 2017, 7:531.

61. Song J, Zhang J, Wang J, Wang J, Guo X, Dong W: β1 integrin mediates colorectal cancer cell proliferation and migration through regulation of the Hedgehog pathway. *Tumor Biology* 2015, 36:2013-2021.

62. Barnawi R, Al-Khaldi S, Colak D, Tulbah A, Al-Tweigeri T, Fallatah M, Monies D, Ghebeh H, Al-Alwan M: β1 Integrin is essential for fascin-mediated breast cancer stem cell function and disease progression. *International journal of cancer* 2019, 145:830-841.

63. Das K, Prasad R, Ansari SA, Roy A, Mukherjee A, Sen P: Matrix metalloproteinase-2: A key regulator in coagulation proteases mediated human breast cancer progression through autocrine signaling. *Biomedicine & Pharmacotherapy* 2018, 105:395-406.

64. Mendes O, Kim H-T, Lungu G, Stoica G: MMP2 role in breast cancer brain metastasis development and its regulation by TIMP2 and ERK1/2. *Clinical & experimental metastasis* 2007, 24:341-351.

65. Dong W, Li H, Zhang Y, Yang H, Guo M, Li L, Liu T: Matrix metalloproteinase 2 promotes cell growth and invasion in colorectal cancer. *Acta Biochim Biophys Sin* 2011, 43:840- 848.

66. Bates AL, Pickup MW, Hallett MA, Dozier EA, Thomas S, Fingleton B: Stromal matrix metalloproteinase 2 regulates collagen expression and promotes the outgrowth of experimental metastases. *The Journal of pathology* 2015, 235:773-783.

67. Kim Y, Shiba-Ishii A, Nakagawa T, Iemura S-i, Natsume T, Nakano N, Matsuoka R, Sakashita S, Lee S, Kawaguchi A: Stratifin regulates stabilization of receptor tyrosine kinases via interaction with ubiquitin-specific protease 8 in lung adenocarcinoma. *Oncogene* 2018, 37:5387-5402.

68. Shiba-Ishii A, Kim Y, Shiozawa T, Iyama S, Satomi K, Kano J, Sakashita S, Morishita Y, Noguchi M: Stratifin accelerates progression of lung adenocarcinoma at an early stage. *Molecular cancer* 2015, 14:142.

69. Cetintas VB, Tetik A, Cok G, Kucukaslan AS, Kosova B, Gunduz C, Veral A, Eroglu Z:

Role of 14-3-3σ in resistance to cisplatin in non-small cell lung cancer cells. *Cell biology international* 2013, 37:78-86.

70. Guweidhi A, Kleeff J, Giese N, Fitori JE, Ketterer K, Giese T, Büchler MW, Korc M, Friess H: Enhanced expression of 14-3-3sigma in pancreatic cancer and its role in cell cycle regulation and apoptosis. *Carcinogenesis* 2004, 25:1575-1585.

71. Li Z, Dong Z, Myer D, Yip-Schneider M, Liu J, Cui P, Schmidt CM, Zhang J-T: Role of 14-3-3σ in poor prognosis and in radiation and drug resistance of human pancreatic cancers. *BMC cancer* 2010, 10:598.

72. Shiba-Ishii A, Kano J, Morishita Y, Sato Y, Minami Y, Noguchi M: High expression of stratifin is a universal abnormality during the course of malignant progression of early-stage lung adenocarcinoma. *International journal of cancer* 2011, 129:2445-2453.

73. Nishimaki H, Kasai K, Kozaki K-i, Takeo T, Ikeda H, Saga S, Nitta M, Itoh G: A role of activated Sonic hedgehog signaling for the cellular proliferation of oral squamous cell carcinoma cell line. *Biochemical and biophysical research communications* 2004, 314:313-320.

74. Yan R, Peng X, Yuan X, Huang D, Chen J, Lu Q, Lv N, Luo S: Suppression of growth and migration by blocking the Hedgehog signaling pathway in gastric cancer cells. *Cellular oncology* 2013, 36:421-435.

75. Islam S, Mokhtari R, Noman A, Uddin M, Rahman M, Azadi M, Zlotta A, van der Kwast T, Yeger H, Farhat W: Sonic hedgehog (Shh) signaling promotes tumorigenicity and stemness via activation of epithelial-to-mesenchymal transition (EMT) in bladder cancer. *Molecular carcinogenesis* 2016, 55:537-551.

76. Jiang WG, Ye L, Ruge F, Sun P-H, Sanders AJ, Ji K, Lane J, Zhang L, Satherley L, Weeks HP: Expression of Sonic Hedgehog (SHH) in human lung cancer and the impact of YangZheng XiaoJi on SHH-mediated biological function of lung cancer cells and tumor growth. *Anticancer research* 2015, 35:1321-1331.

77. Szczepny A, Rogers S, Jayasekara WSN, Park K, McCloy RA, Cochrane CR, Ganju V, Cooper WA, Sage J, Peacock CD: The role of canonical and non-canonical Hedgehog signaling in tumor progression in a mouse model of small cell lung cancer. *Oncogene* 2017, 36:5544-5550.

78. Statkiewicz M, Maryan N, Lipiec A, Grecka E, Grygorowicz MA, Omiotek M, Gorska A, Mikula M, Malecki M: The role of the SHH gene in prostate cancer cell resistance to paclitaxel. *The Prostate* 2014, 74:1142-1152.

79. Li X, Wang Z, Ma Q, Xu Q, Liu H, Duan W, Lei J, Ma J, Wang X, Lv S: Sonic hedgehog paracrine signaling activates stromal cells to promote perineural invasion in pancreatic cancer. *Clinical Cancer Research* 2014, 20:4326-4338.

80. Yao Z, Han L, Chen Y, He F, Sun B, Zhang Y, Yang Y, Wang C, Yang Z: Hedgehog signalling in the tumourigenesis and metastasis of osteosarcoma, and its potential value in the clinical therapy of osteosarcoma. *Cell death & disease* 2018, 9:1-12.

81. Riaz SK, Ke Y, Wang F, Kayani MA, Malik MFA: Influence of SHH/GLI1 axis on EMT mediated migration and invasion of breast cancer cells. *Scientific reports* 2019, 9:1-13.

82. Song L, Liu D, Zhao Y, He J, Kang H, Dai Z, Wang X, Zhang S, Zan Y, Xue X: Sinomenine reduces growth and metastasis of breast cancer cells and improves the survival of tumor-bearing mice through suppressing the SHh pathway. *Biomedicine & Pharmacotherapy* 2018, 98:687-693.

83. The U.S. Food and Drug Administration. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2018/210656Orig1s000TOC.cfm. Accessed 1 May 2020.

84. The U.S. Food and Drug Administration. https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/205266s004lbl.pdf. Accessed 1 May 2020.

85. The U.S. Food and Drug Administration. https://www.accessdata.fda.gov/drugsatfda_docs/label/2012/203388lbl.pdf. Accessed 1 May 2020.

86. Walsh AM, Kapoor GS, Buonato JM, Mathew LK, Bi Y, Davuluri RV, Martinez-Lage M, Simon MC, O'Rourke DM, Lazzara MJ: Sprouty2 drives drug resistance and proliferation in glioblastoma. *Molecular Cancer Research* 2015, 13:1227-1237.

87. Park J-W, Wollmann G, Urbiola C, Fogli B, Florio T, Geley S, Klimaschewski L: Sprouty2 enhances the tumorigenic potential of glioblastoma cells. *Neuro-oncology* 2018, 20:1044-1054.

88. Li P, Tao L, Yang J, Cai H, Ju X, Li J, Shao P, Cao Q, Qin C, Meng X: Sprouty2 is associated with prognosis and suppresses cell proliferation and invasion in renal cell carcinoma. *Urology* 2013, 82:253. e251-253. e257.

89. Sutterlüty H, Mayer C-E, Setinek U, Attems J, Ovtcharov S, Mikula M, Mikulits W, Micksche M, Berger W: Down-Regulation of Sprouty2 in Non–Small Cell Lung Cancer Contributes to Tumor Malignancy via Extracellular Signal-Regulated Kinase Pathway-Dependent and-Independent Mechanisms. *Molecular Cancer Research* 2007, 5:509-520.

90. So W-K, Cheng J-C, Fan Q, Wong AS, Huntsman DG, Gilks CB, Leung PC: Loss of Sprouty2 in human high-grade serous ovarian carcinomas promotes EGF-induced E-cadherin down-regulation and cell invasion. *FEBS letters* 2015, 589:302-309.

91. Cheng J-C, Chang H-M, Xiong S, So W-K, Leung PC: Sprouty2 inhibits amphiregulin-induced down-regulation of E-cadherin and cell invasion in human ovarian cancer cells. *Oncotarget* 2016, 7:81645.

92. Saini M, Verma A, Mathew SJ: SPRY2 is a novel MET interactor that regulates metastatic potential and differentiation in rhabdomyosarcoma. *Cell death & disease* 2018, 9:1-15.

93. Holgren C, Dougherty U, Edwin F, Cerasi D, Taylor I, Fichera A, Joseph L, Bissonnette M, Khare S: Sprouty-2 controls c-Met expression and metastatic potential of colon cancer cells: sprouty/c-Met upregulation in human colonic adenocarcinomas. *Oncogene* 2010, 29:5241-5253.

94. Patel R, Gao M, Ahmad I, Fleming J, Singh LB, Rai TS, McKie AB, Seywright M, Barnetson RJ, Edwards J: Sprouty2, PTEN, and PP2A interact to regulate prostate cancer progression. *The Journal of clinical investigation* 2013, 123:1157-1175.

95. Kim M, Jang K, Miller P, Picon-Ruiz M, Yeasky T, El-Ashry D, Slingerland JM: VEGFA links self-renewal and metastasis by inducing Sox2 to repress miR-452, driving Slug. *Oncogene* 2017, 36:5199-5211.

96. English WR, Lunt SJ, Fisher M, Lefley DV, Dhingra M, Lee Y-C, Bingham K, Hurrell JE, Lyons SK, Kanthou C: Differential expression of VEGFA isoforms regulates metastasis and response to anti-VEGFA therapy in sarcoma. *Cancer research* 2017, 77:2633-2646.

97. Kong D-H, Kim MR, Jang JH, Na H-J, Lee S: A review of anti-angiogenic targets for monoclonal antibody cancer therapy. *International journal of molecular sciences* 2017, 18:1786.

98. The drugs.com. https://www.drugs.com/drug-class/vegf-vegfr-inhibitors.htm. Accessed 1 May 2020.

99. Choi J-H, Kim Y-B, Ahn JM, Kim MJ, Bae WJ, Han S-U, Woo HG, Lee D: Identification of genomic aberrations associated with lymph node metastasis in diffuse-type gastric cancer. *Experimental & molecular medicine* 2018, 50:1-11.

100. Nie S, Zhou J, Bai F, Jiang B, Chen J, Zhou J: Role of endothelin A receptor in colon cancer metastasis: in vitro and in vivo evidence. *Molecular carcinogenesis* 2014, 53:E85-E91. 101. Fukui R, Nishimori H, Hata F, Yasoshima T, Ohno K, Yanai Y, Kamiguchi K, Denno R, Sato N, Hirata K: Inhibitory effect of endothelin A receptor blockade on tumor growth and liver metastasis of a human gastric cancer cell line. *Gastric Cancer* 2007, 10:123-128.

102. Rosano L, Spinella F, Salani D, Di Castro V, Venuti A, Nicotra MR, Natali PG, Bagnato A: Therapeutic targeting of the endothelin a receptor in human ovarian carcinoma. *Cancer Research* 2003, 63:2447-2453.

103. Mai HQ, Zeng ZY, Feng KT, Ye YL, Zhang CQ, Liang WJ, Guo X, Mo HY, Hong MH: Therapeutic targeting of the endothelin a receptor in human nasopharyngeal carcinoma. *Cancer science* 2006, 97:1388-1395.

104. Laurberg JR, Jensen JB, Schepeler T, Borre M, Ørntoft TF, Dyrskjøt L: High expression of GEM and EDNRA is associated with metastasis and poor outcome in patients with advanced bladder cancer. *BMC cancer* 2014, 14:638.

105. Said N, Theodorescu D: Permissive role of endothelin receptors in tumor metastasis. *Life sciences* 2012, 91:522-527.

106. Wang H, Zhou X, Zhang Y, Zhu H, Zhao L, Fan L, Wang Y, Gang Y, Wu K, Liu Z:
Growth arrest-specific gene 1 is downregulated and inhibits tumor growth in gastric cancer. *The FEBS journal* 2012, 279:3652-3664.

107. Zhao L, Pan Y, Gang Y, Wang H, Jin H, Tie J, Xia L, Zhang Y, He L, Yao L:

Identification of GAS1 as an epirubicin resistance-related gene in human gastric cancer cells with a partially randomized small interfering RNA library. *Journal of Biological Chemistry* 2009, 284:26273-26285.

108. Li Q, Qin Y, Wei P, Lian P, Li Y, Xu Y, Li X, Li D, Cai S: Gas1 inhibits metastatic and metabolic phenotypes in colorectal carcinoma. *Molecular Cancer Research* 2016, 14:830-840.

109. Jiang Z, Xu Y, Cai S: Down-regulated GAS1 expression correlates with recurrence in stage II and III colorectal cancer. *Human pathology* 2011, 42:361-368.

110. Gobeil S, Zhu X, Doillon CJ, Green MR: A genome-wide shRNA screen identifies GAS1 as a novel melanoma metastasis suppressor gene. *Genes & development* 2008, 22:2932-2940.

111. Sacilotto N, Castillo J, Riffo-Campos ÁL, Flores JM, Hibbitt O, Wade-Martins R, López C, Rodrigo MI, Franco L, López-Rodas G: Growth arrest specific 1 (Gas1) gene overexpression in liver reduces the in vivo progression of murine hepatocellular carcinoma and partially restores gene expression levels. *PloS one* 2015, 10.

112. Jiménez A, López-Ornelas A, Estudillo E, González-Mariscal L, González RO, Segovia J: A soluble form of GAS1 inhibits tumor growth and angiogenesis in a triple negative breast cancer model. *Experimental cell research* 2014, 327:307-317.

113. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, Gillespie T, Park JY, Nam RK, Sugar L: Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer research* 2014, 74:3228-3237.

114. Soady KJ, Kendrick H, Gao Q, Tutt A, Zvelebil M, Ordonez LD, Quist J, Tan DW-M, Isacke CM, Grigoriadis A: Mouse mammary stem cells express prognostic markers for triple- negative breast cancer. *Breast Cancer Research* 2015, 17:31.

115. Andersen RK, Hammer K, Hager H, Christensen JN, Ludvigsen M, Honoré B, Thomsen MBH, Madsen M: Melanoma tumors frequently acquire LRP 2/megalin expression, which

modulates melanoma cell proliferation and survival rates. *Pigment cell & melanoma research* 2015, 28:267-280.

116. Gonias SL, Karimi-Mostowfi N, Murray SS, Mantuano E, Gilder AS: Expression of LDL receptor-related proteins (LRPs) in common solid malignancies correlates with patient survival. *PloS one* 2017, 12.

117. Holt SK, Karyadi DM, Kwon EM, Stanford JL, Nelson PS, Ostrander EA: Association of megalin genetic polymorphisms with prostate cancer risk and prognosis. *Clinical cancer research* 2008, 14:3823-3831.

118. Wang L, Duan W, Kang L, Mao J, Yu X, Fan S, Li L, Tao Y: Smoothened activates breast cancer stem-like cell and promotes tumorigenesis and metastasis of breast cancer. *Biomedicine & Pharmacotherapy* 2014, 68:1099-1104.

119. Richtig G, Aigelsreiter AM, Asslaber M, Weiland T, Pichler M, Eberhard K, Sygulla S, Schauer S, Hoefler G, Aigelsreiter A: Hedgehog pathway proteins SMO and GLI expression as prognostic markers in head and neck squamous cell carcinoma. *Histopathology* 2019, 75:118- 127.

120. Li C, Song G, Zhang S, Wang E, Cui Z: Wnt3a increases the metastatic potential of non-small cell lung cancer cells in vitro in part via its upregulation of Notch3. *Oncology reports* 2015, 33:1207-1214.

121. Qi L, Sun B, Liu Z, Cheng R, Li Y, Zhao X: Wnt3a expression is associated with epithelial-mesenchymal transition and promotes colon cancer progression. *Journal of Experimental & Clinical Cancer Research* 2014, 33:107.

8.4. Ethics Committee Report

İZMİR BİYOTIP VE GENOM MERKEZİ
GİRİŞİMSEL OLMAYAN ARAŞTIRMALAR ETİK KURULU (İBG-GOEK)
KARARI

Toplantı Tarihi	: 25/12/2020	**Toplantı Günü**	: Cuma
Toplantı Sayısı	: 15	**Toplantı Saati**	: 10:30

Sayın Hani ALOTAIBI,

2020-045 Protokol No'lu; sorumlusu olduğunuz "Farklı taksonomik gruplarda kanserin moleküler mekanizmalarının hesaplamalı biyoloji uygulamalarıyla araştırılması" başlıklı araştırmanın uygulanmasında etik açıdan sakınca olmadığına oy birliği ile karar verilmiştir.

Bilgilerinizi ve gereğini rica ederiz.

Prof. Dr. H. Alper BAĞRIYANIK
Başkan

Pandemi süresince online ortamda gerçekleştirilen toplantımızda alınan kararlar tek imzalı olarak düzenlenmektedir. Pandemi sona erdikten sonra ıslak imzalı karar belgesi teslim edilecektir.

8.4. Curriculum Vitae

Işıl Takan

Date of Birth:

Address: Izmir/Turkey

Phone:

E-mail:

Scientific Interests

- Machine Learning - Evolutionary Analysis
- Bioinformatics - System Biology
- Phylogenomics - Biological Networks
- Cancer Evolution - Data processing
- High Throughput Data Analysis

Education

2018- 2022 Ph.D. Student, Dokuz Eylül University, Izmir Biomedicine and Genome Institute,

İzmir, Turkey,

Doctorate Degree in Molecular Biology and Genetics, Research

Area: Bioinformatics

Thesis Title: Application of Computational Biology Approaches for The Investigation of The Molecular Mechanisms of Cancer Across Taxonomic Groups

Supervisor: Dr. Athanasia Pavlopoulou

GPA – 3.43/4.0

2019- AAS Student in Computer Programming, Dokuz Eylül University, İzmir, Turkey

2011–2014 M.S. Student, Ege University İzmir, Turkey,

Master's Degree in Molecular Biology,

Research Area: Bioinformatics
Thesis Title: Bioinformatics Analysis of Prion Gene (PRNP) in Mammalian Species

Supervisor: Prof. Dr. Cemal ÜN GPA

– 3.91/4.0

2006–2011 B.Sc. Student, Ege University İzmir, Turkey, Bachelor's

Degree in Biology

Research Area: Bioinformatics

Thesis Title: Evolutionary Analysis of Sakiz Sheep Using Mitochondrial DNA D-loop HVR1 Gene

Supervisor: Prof. Dr. Cemal ÜN

GPA – 3.09/4.0

Skills

Language skills: Fluent in English ; French (basic-A2) Computer

skills:

Programming: Python, R, Bash scripting, JavaScript, MATLAB/OCTAVE, SQL

Application Programs: Cytoscape, MEGA (Molecular Evolutionary Genetics Analysis), Adobe Photoshop, Adobe Illustrator

Publications

ISI Journals

1. Mehtap O, Benlier N, Takan I, Vasileiou C, Georgakilas AG, Pavlopoulou A, Cetin Z, Saygili EI (2022) “Ginger for Healthy Ageing: A Systematic Review on Current Evidence of Its Antioxidant, Anti-Inflammatory, and Anticancer Properties” Oxidative Medicine and Cellular Longevity, 2022: 16 Oxidative Medicine and Cellular Longevity 2022:9993518
2. Faustino D, Brinkmeier H, Logotheti S, Jonitz-Heincke A, Yilmaz H, Takan I, Peters K, Bader R, Lang H, Pavlopoulou A, Pützer BM, Spitschak A (2022) “Novel integrated workflow allows production and in-depth quality assessment of multifactorial reprogrammed skeletal muscle cells from human stem cells” Cellular and Molecular Life Sciences 79:229
3. Geronikolou SA, Takan I, Pavlopoulou A, Mantzourani M, Chroussos GP (2022) "Thrombocytopenia in COVID-19 and vaccine-induced thrombotic thrombocytopenia” International Journal of Molecular Medicine 49:1-19
4. Pavlopoulou A, Asfa S, Gioukakis E, Mavragani IV, Nikitaki Z, Takan I, Pouget J-P, Harrison L, Georgakilas AG (2021) “In Silico Investigation of the Biological Implications of Complex DNA Damage with Emphasis in Cancer Radiotherapy through a Systems Biology Approach” Molecules 26:7602
5. Louka A, Takan I, Pavlopoulou A, Georgakilas AG. (2021) “Bioinformatic approaches to the investigation of the atavistic genes implicated in cancer” Frontiers in Bioscience-Landmark 26:279-311
6. Theofylaktou D, Takan I, Karakülah G, Biz GM, Zanni V, Pavlopoulou A, Georgakilas AG. (2021) " Mining Natural Products with Anticancer Biological Activity through a Systems Biology Approach" Oxidative Medicine and Cellular Longevity 2021:9993518
7. Marquardt S, Pavlopoulou A, Takan I, Prabir D, Pützer 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

8. Logotheti S, Marquardt S, Richter C, Hain RS, Murr N, Takan I, Pavlopoulou A, Pützer BM (2020) "Neural Networks Recapitulation by Cancer Cells Promotes Disease Progression: A Novel Role of p73 Isoforms in Cancer-Neuronal Crosstalk" Cancers 12:3789

International Conference Presentations

---

1. Takan I, Marquardt S, Logotheti S, Pavlopoulou A (June, 2021) "A systems biology approach for elucidating the 'bing bang' of cancer metastasis" *Hbioinfo bridge* 2020-21 Nicosia, Cyprus
2. Takan I, Theofylaktou D, Karakülah G, Mehmet Biz G, Kontou P, Georgakilas AG, Pavlopoulou A (December, 2021) "Mining natural products with anticancer biological activity through a systems biology approach" *Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB21)* Athens, Greece

3. Takan I, Alotaibi H, Pavlopoulou A (October, 2019) "Phylogenetic analysis of keratins: functional implications" *12th International Symposium on Health Informatics and Bioinformatics (HIBIT 2019)* Izmir, Turkey
4. Takan I, Un C., (April, 2012) "Codon usage bias of scrapie resistance PRNP alleles in Ovis aries", *7th International Symposium on Health Informatics and Bioinformatics (HIBIT 2012)* Cappadocia, Turkey

5. Takan I, Un C., (August, 2012) "Phylogenetic Analysis by Using Mitochondrial DNA D-loop HVR1 Sequences in Ovis aries Breeds*" International ITU 6th Molecular Biology and Genetics Student Congress* Istanbul, Turkey

6. Takan I, Un C., (May, 2011) "Analysis of the nucleotid composition profiles and codon use frequency of PRNP genes from different mammalian species" *6th International Symposium on Health Informatics and Bioinformatics (HIBIT 2011)* Izmir, Turkey
7. Takan I, Un C., (April, 2010) "Evolutionary Analysis of Keratin Gene (KRT84) in Different Mammalian Species" *5th International Symposium on Health Informatics and Bioinformatics (HIBIT 2010)* Antalya, Turkey

National Journals and Conferences

---

1. Takan I, Kızılşahin S. (May, 2012) "Phylogenetic Analysis and Codon Usage Frequency of SHOX Gene In Different Vertebrate Species" *2th National Student Congress of Evolutionary Biology* – Ankara, Turkey

2. Takan I, Un C., (May, 2011) "Evolutionary Analysis of Sakiz Sheep Using Mitochondrial DNA D-loop HVR1 Gene Sequences" *7th National Student Congress of Zoo techniques* Aydın, Turkey

3. Takan I, Un C (December, 2011) "Evaluating Evolutionary Knowledge Through Social Media" *3th The Evolution, Science and Education Symposium,* Istanbul, Turkey

Awards

- PhD: Council of Higher Education YÖK 100/2000 Scholarship
- Bachelor: TUBITAK Student Graduation Thesis Project-2209-Evolutionary Analysis of Sakiz Sheep Using Mitochondrial DNA D-loop *HVR1* Gene Sequences that's supported by THE SCIENTIFIC AND TECHNOLOGICAL RESEARCH COUNCIL OF TURKEY
- HIBIT'19 * Best poster award

8.5. Publications from the Thesis

1. Marquardt S, Pavlopoulou A, Takan I, Prabir D, Pützer 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
2. Takan I, Marquardt S, Logotheti S, Pavlopoulou A (June, 2021) "A systems biology approach for elucidating the 'bing bang' of cancer metastasis" *Hbioinfo bridge 2020-21* Nicosia, Cyprus
3. Takan I, Alotaibi H, Pavlopoulou A (October, 2019) "Phylogenetic analysis of keratins: functional implications" *12th International Symposium on Health Informatics and Bioinformatics (HIBIT 2019)* Izmir, Turkey

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 02 June 2021
doi: 10.3389/fcell.2021.682619

A Systems-Based Key Innovation-Driven Approach Infers Co-option of Jaw Developmental Programs During Cancer Progression

Stephan Marquardt[†], *Athanasia Pavlopoulou*[2,3†], *Işıl Takan*[2,3], *Prabir Dhar*[1], *Brigitte M. Pützer*[1,4] *and Stella Logotheti*[1*]

[1] Institute of Experimental Gene Therapy and Cancer Research, Rostock University Medical Center, Rostock, Germany, [2] İzmir Biomedicine and Genome Center, İzmir, Turkey, [3] İzmir International Biomedicine and Genome Institute, Dokuz Eylül University, İzmir, Turkey, [4] Department Life, Light & Matter, University of Rostock, Rostock, Germany

OPEN ACCESS

Edited by:
Renjie Chai,
Southeast University, China

Reviewed by:
Amélie E. Coudert,
Université Paris Diderot, France
Miriam Gagliardi,
University of Palermo, Italy

***Correspondence:**
Stella Logotheti
Styliani.Logotheti@med.uni-rostock.de

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: 18 March 2021
Accepted: 11 May 2021
Published: 02 June 2021

Citation:
Marquardt S, Pavlopoulou A, Takan I, Dhar P, Pützer BM and Logotheti S (2021) A Systems-Based Key Innovation-Driven Approach Infers Co-option of Jaw Developmental Programs During Cancer Progression. Front. Cell Dev. Biol. 9:682619. doi: 10.3389/fcell.2021.682619

Cancer acquires metastatic potential and evolves via co-opting gene regulatory networks (GRN) of embryonic development and tissue homeostasis. Such GRNs are encoded in the genome and frequently conserved among species. Considering that all metazoa have evolved from a common ancestor via major macroevolutionary events which shaped those GRNs and increased morphogenetic complexity, we sought to examine whether there are any key innovations that may be consistently and deterministically linked with metastatic potential across the metazoa clades. To address tumor evolution relative to organismal evolution, we revisited and retrospectively juxtaposed seminal laboratory and field cancer studies across taxa that lie on the evolutionary lineage from cnidaria to humans. We subsequently applied bioinformatics to integrate species-specific cancer phenotypes, multiomics data from up to 42 human cancer types, developmental phenotypes of knockout mice, and molecular phylogenetics. We found that the phenotypic manifestations of metastasis appear to coincide with agnatha-to-gnathostome transition. Genes indispensable for jaw development, a key innovation of gnathostomes, undergo mutations or methylation alterations, are aberrantly transcribed during tumor progression and are causatively associated with invasion and metastasis. There is a preference for deregulation of gnathostome-specific versus pre-gnathostome genes occupying hubs of the jaw development network. According to these data, we propose our systems-based model as an *in silico* tool the prediction of likely tumor evolutionary trajectories and therapeutic targets for metastasis prevention, on the rationale that the same genes which are essential for key innovations that catalyzed vertebrate evolution, such as jaws, are also important for tumor evolution.

Keywords: cancer evolution, metastasis, jaw development, epithelial-mesenchymal transition, gnathostomes, cyclostomes, tumor evolutionary trajectories, prediction model

INTRODUCTION

Despite advances in cancer management, manifestation of lesions with metastatic potential signals the terminal stage of disease. The term "metastatic potential" may include any combination of cancer phenotypes that enable metastatic dissemination including motility, immune evasion, and ability to survive in circulation and proliferate at distant sites (Birkbak and McGranahan, 2020). Cancer progression is governed by mechanisms distinct from those of initiation (Logotheti et al., 2020). While at early stages cancer cells accumulate driver mutations, at advanced stages, they do not acquire additional, metastasis-specific mutations (Rodrigues et al., 2018), but rather hijack programs of tissue-homeostasis and normal embryonic development and reactivate them in an unusual place, at the wrong time (Logotheti et al., 2020). Metastasis is promoted by aberrant gene regulation, and metastatic transcriptional programs arise from *de novo* combinatorial activation of multiple distinct and developmentally distant transcriptional modules (Rodrigues et al., 2018). We and others have shown that cancer progression is facilitated by ectopic activation of genes that have tissue-restricted profiles (Rousseaux et al., 2013; Richter et al., 2019), or are involved in placenta (Costanzo et al., 2018) and embryonic development (Billaud and Santoro, 2011) including, but not limited to, neuronal development and function (Logotheti et al., 2020). For instance, we have recently provided compelling evidence that genes involved in neuronal development and neurological function are reactivated in tumors and predict poor patient outcomes across various cancers. Tumors co-opt genes essential for the development of distinct anatomical parts of the nervous system, with a frequent preference for cerebral cortex and the neural crest-derived enteric nerves function (Logotheti et al., 2020). In this respect, co-option, the evolutionary process through which a biological function within a specific context may be alternatively used in another context to support a novel function, emerges as a recurrent and prevailing pattern during tumor progression (Billaud and Santoro, 2011).

The gene regulatory networks (GRN) of embryonic development and tissue homeostasis are encoded in the genomes of animal species and define their attributes and morphogenetic complexity (Levine and Davidson, 2005). Considering that tumors can progress to metastatic stages by co-opting such gene programs, it is reasonable to conjure that the metastatic potential largely depends on the gene reservoir of the species on which tumors grow. For example, lesions growing on animals as simple as cnidarians will plausibly usurp the GRNs controlling their corresponding attributes, while mammal tumors have access to GRNs underlying more sophisticated body plans. From a phylogenetic point of view, all metazoa have evolved from a common ancestor via major macroevolutionary events, which advanced the animal body plans, and GRNs which are associated with these events are conserved across species (Levine and Davidson, 2005). Given that these GRNs can be inevitably at the disposal of cancerous tumors, we wondered whether the same key innovations through which the species evolved may have been, in parallel, exploited by the primary tumor, in order to evolve toward metastatic stages. The term "key innovation," as used herein, refers to any novel phenotypic trait that facilitates adaptive radiation and evolutionary success of a taxonomic group (Hunter, 1998), as well as the respective genes and/or GRNs that support establishment of this trait. The epithelial-mesenchymal transition (EMT), that is the hierarchical GRN which controls neural crest, a vertebrate-specific multipotent embryonic cell population which generates several body anatomical structures (Martik and Bronner, 2017) is a representative example of co-option of key innovations toward enhancing metastatic potential. Indeed, the same EMT factors mediating differentiation and migration of neural crest are also ectopically reactivated during tumor progression (Kerosuo and Bronner-Fraser, 2012; Fürst et al., 2019).

All metazoa can develop tumors (Domazet-Lošo et al., 2014), but major differences in their prevalence and metastatic potential are observed across phyla. The "big-bang" of tumor formation is traced to Cnidaria and correlates with the emergence of multicellularity. All cancer-associated genes are conserved in Cnidaria, and *Hydra* tumor cells have an invasive capacity (Domazet-Lošo et al., 2014). However, the aforementioned phylogenetic origin of tumor formation does not coincide with the phenotypic manifestations of aggressiveness, since Cnidaria neither form true metastases nor die of cancer (Domazet-Lošo et al., 2014). Thus, it remains enigmatic how species with lethal cancers have non-metastatic common ancestors, as well as if there are any key innovations that may be linked with increased prevalence of metastasis across the metazoa clades. We hypothesized that if certain key innovations increase organismal fitness of a given species population, they will likely undergo positive selection, despite the risk of being co-opted by the tumors later on in the life of the individuals of the respective population. To explore whether organisms that inherited key innovations from a common ancestor consistently manifest metastatic potential, in contrast to the ones which lack them, we applied phylogeny, that is, the evolutionary history of species in relation to oncogeny (Dawe, 1969). Herein, we revisited and retrospectively juxtaposed cancer reports across taxa on the same evolutionary lineage with mammals, from cnidarians to humans, and then integrated cancer phenotypes of these species with high-throughput data from up to 42 human cancer types, data on developmental phenotypes of knockout mice, and phylogenetic comparative methods. This multidisciplinary meta-analysis allowed us to infer that phenotypic manifestation of metastasis coincides with agnatha-to-gnathostome transition. Genes essential for jaw development, which is a key innovation of gnathostomes, are deregulated in tumor cells and are causatively associated with tumor progression.

MATERIALS AND METHODS

Identification of Jaw-Indispensable Genes

To identify those genes that are indispensable for the development of cartilaginous jaws (JIGs), we screened the Mouse

Genome Informatics (MGI) database (Bult et al., 2019) for genes the knockout of which leads to mouse phenotypes with jaw-related defects, as recently described (Logotheti et al., 2020), using the terms "cartilage," "jaws," "mandible/mandibular," "maxilla," and "micrognathia." Their human orthologs were identified and the official HUGO gene nomenclature committee (HGNC) (Gray et al., 2015) gene symbols were used (**Supplementary Table 2**).

Orthologs Search

The HGNC symbols of the 305 jaw-indispensable genes (JIG; **Supplementary Table 2**) were used initially to retrieve the corresponding, well-annotated, gnathostome (*Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio*, and *Callorhinchus milli*) protein sequences from the publicly available non-redundant sequence database NCBI RefSeq (O'Leary et al., 2016). The canonical or longest known transcripts per protein were selected. For obtaining orthologous sequences from the agnatha genera under study (Petromyzon, Branchiostoma, Ciona, Strongylocentrotus, and Hydra), the retrieved gnathostome sequences were used as probes to iteratively search the agnathan genomes available in NCBI RefSeq and GenBank (Sayers et al., 2019), by applying reciprocal BLASTp and BLASTn (Altschul et al., 1990) with default parameters; an in-house Python script was employed (available on request). The protein domain organization of the novel sequences was explored through SMART v.8.0 (Letunic and Bork, 2018).

To identify the "true orthologs" of each JIG/protein, phylogenetic trees (a total of 305) of the homologous, gnathostome and agnathan, protein sequences were constructed. To this end, the amino acid sequences of the homologous proteins were aligned using Clustal Omega, version 1.2.4 (Sievers and Higgins, 2014a,b) and the resulting multiple sequence alignment was provided as input to the software package MEGA version 10. 1 (Kumar et al., 2018) in order to perform phylogenetic analyses, by employing a neighbor-joining (NJ) and a maximum likelihood (ML) method. The expected number of amino acid substitutions per position was estimated with the JTT model (Jones et al., 1992). The robustness of the inferred phylogenetic trees was evaluated by bootstrapping (100 pseudo-replicates). Only those agnathan sequences that clustered with the known gnathostome sequences under study were considered as "true orthologs." A characteristic example is shown in **Supplementary Figure 1**.

Functional Interaction Networks and Gene Set Enrichment Analysis

The associations among the jaw-indispensable human genes/protein products were investigated in the STRING v11.0 database (Szklarczyk et al., 2019), by selecting a high confidence interaction score (≥0.9). Moreover, Cytoscape v3.8.0 (Shannon et al., 2003), was employed for network processing, visualization and statistical analysis. For the Gene Set Enrichment Analysis, the GSEA-P 2.0 software (Broad Institute, Cambridge, MA, United States) (Subramanian et al., 2005) was used. Enriched hallmark and Gene Ontology terms were plotted against the negative log10 of their individual FDR value (<0.05).

Meta-Analyses of Mutation and Methylation Data From Human Tumors

We juxtaposed the datasets of PanCancer and GENCODEv32 using Ensembl gene IDs and filtered for protein coding genes with a transcript support level TSL < 3, to generate a comprehensive dataset of 19,617 transcripts. Recently identified cancer driver genes and frequently mutated genes were also used (Bamford et al., 2004; Kandoth et al., 2013; Bailey et al., 2018). As control gene lists, we generated 100 lists of 305 random genes by sampling the 19,313 transcripts (without JIGs) without replacement (available on request). The Cancer Gene Census (CGC) list of 723 genes (including two non-coding) was downloaded from the COSMIC (Catalogue of Somatic Mutations in Cancer) website (Bamford et al., 2004) and juxtaposed with JIGs or random lists. For evaluating stochastic events, we calculated the percentage of events by chance ($x\% = 100 \times k/19{,}313$ coding genes without JIGs, $k = 721$ CGC, 299 cancer drivers, 127 frequently mutated) and performed Chi-squared test (stochastic events) or z-test (random gene lists). The number of patients affected by gene mutations and the number of mutations per gene were retrieved from the GDC data portal (Jensen et al., 2017) after uploading the respective gene list. DiseaseMeth database (Xiong et al., 2017) was used to detect the differentially methylated JIGs in several cancer types. All results were downloaded and data for the different cancer types were pooled.

Transcriptome Analysis in CCLE, TCGA, and GEO Databases

Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Atlas (TCGA) transcriptomic data were analyzed as recently described (Logotheti et al., 2020). For identifying differentially regulated transcripts in metastatic versus primary lesions or normal tissue, transcriptome data from the Gene Expression Omnibus (GEO) database (Clough and Barrett, 2016) (study numbers: GSE21510, GSE2509, GSE25976, GSE43837, GSE468, GSE6919, GSE7929, GSE7930, GSE8401) were analyzed by GEO2R. Cox regression analysis was performed by R software using the *coxph* function in the *survival* package.

Statistical Analysis

Unless otherwise stated, statistics were performed by Student's t-test; p values less than 0.05 were considered as significant ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). All statistical tests were two-sided.

RESULTS

Key Innovations in Relation With Tumor Characteristics Across Phylogenetic Taxa: Available Resources and Considerations

As a framework for comparing tumors among metazoa, we used the one proposed by Dawe, one of the pioneers to address oncogeny in relation to phylogeny. The framework is

the phylogenetic tree *per se* and the comparisons are made by ascending the bifurcating tree, in a direction from the last common ancestor to the more recent taxon, by recording changes which occurred proximal to the divergent branches at each node (Dawe, 1969). In line with this, we selected that taxon in each bifurcation of the phylogenetic tree in which an indispensable hallmark macroevolutionary trait occurred first, and was conserved throughout the descendant lineage, up to mammals and humans (Hickman et al., 2001). Porifera was the baseline taxonomic group of our study panel, since sponges do not develop apparent tumors (Domazet-Lošo et al., 2014). In the case of Protochordates, due to their nodal position between invertebrates and vertebrates (Delsuc et al., 2006), both children taxa (Cephalochordates and Urochordates) were included in the analysis.

We extensively searched publicly available databases, for peer-reviewed cancer reports on representatives of these taxa. A degree of inherent heterogeneity due to diverse experimental methods across these reports should be assumed. Notwithstanding, these records are, to date, the best available source of information on tumor characteristics of these taxa. They remain fundamental, especially given that in some species, experimental carcinogenesis protocols cannot be applied, due to animal ethics restrictions and/or because they do not represent conventional laboratory animal models. To eliminate potential bias by comparing pathology reports on isolated cases, we particularly emphasized on studies which (a) followed-up large populations of animals over ample periods of time, and (b) provided a clear number of subjects, histological characterizations, and wherever possible, experimental validations. Another comprehensive source of information is the "Registry of Tumors in Lower Animals" (RTLA), i.e., an official repository of validated tumor reports from a large number of invertebrates and cold blooded vertebrates (Harshbarger, 1969), and the "Overall five year progress report for the registry of tumors in lower animals from September 30, 2002 through September 30, 2007" (personal communication, Dr. P.J. Daschner). We also considered parameters that are particularly important for metastasis, mainly (a) the circulatory system, which offers cancer cells a means for energy supply and migration to secondary sites (Hanahan and Weinberg, 2011) and (b) the immune system, which reflects the innate ability of an organism to detect and eliminate malignant cells and represents a major evolutionary pressure in the tumor microenvironment (Angelova et al., 2018). All chordates have a closed circulatory system (Hickman et al., 2001). Adaptive immunity first occurred in cyclostomes (Cooper and Alder, 2006), whereas all their ancestors possess only innate immunity (Langlet and Bierne, 1982; Cooper and Alder, 2006; Maciel and Oviedo, 2018). Overall, tumors across the selected taxa, in association with aggressiveness, immunity and the circulatory system can be overviewed in **Figure 1** and **Supplementary Table 1**.

Tumor Landscape in Pre-vertebrates

According to available reports (**Supplementary Table 1**), Cnidaria are the most ancient organisms known to develop naturally occurring tumors. Although tumors in Hydra polyps reduce capacity for egg production and rate of population growth, they are non-lethal for the affected individuals (Domazet-Lošo et al., 2014). Platyhelminthes develop spontaneous, non-lethal tumors (Harshbarger, 1969; Tascedda and Ottaviani, 2014), while their exposure to carcinogen type 1A cadmium leads to benign tumors and impairment of their regenerative ability, especially in combination with inactivation of tumor-suppressor genes (Van Roten et al., 2018). No tumors have been reported in Nemertea. Nematodes develop germline cell-derived tumors (Kirienko et al., 2010). Cancer in Molluscs is manifested as a leukemia-like, disseminated neoplasia (DN), and as germinal cell-derived gonadal neoplasia (Barber, 2004; Carballal et al., 2015). Importantly, DN in *Mya arenaria* populating the coast of North America, is a horizontally transmissible form of cancer, whereby the cancer clone, which likely arose in a single individual, is spread to host clams, and bears a genotype distinct from the host genotypes (Metzger et al., 2015). Nevertheless, such deadly tumors are restricted to bivalvia, and, to date, have not been described in other pre-vertebrate taxa on the same evolutionary lineage. Indeed, Echinoderms are resistant to chemical-induced oncogenesis (Wellings, 1969), and either lack spontaneous tumor lesions (Wellings, 1969) or develop non-invasive/non-lethal, pigmented lesions (Fontaine, 1969). Similarly, the protochordates either appear to be cancer-free (Urochordates) (Dawe, 1969; Wellings, 1969) or form benign tumors (Cephalochordates) (Wellings, 1969). Overall, with the exception of Molluscs, tumors in pre-vertebrates are not associated with lethal outcomes (**Figure 1**).

Emergence of Metastasis Coincides With Agnatha-Gnathostomes Split Within the Vertebrate Clade

Cyclostomes, the only living jawless vertebrates (agnatha) (Gess et al., 2006), comprise a monophyletic group (Heimberg et al., 2010), including Petromyzontia (lampreys) and Myxinoidei (hagfishes). Until 70's, only one case of cyclostome cancer had been reported in RTLA. Following this singleton report, Falkmer and colleagues addressed cyclostome tumor pathology in an extensive, thorough and well-controlled manner (Falkmer et al., 1976, 1978; Falkmer, 1998) and, up to this day, this seminal work remains the most comprehensive source of information for carcinogenesis on this enigmatic, though basal, vertebrate superclass. In particular, Falkmer screened, for tumor incidence, two large populations of lampreys caught in Ume/Ricklean rivers, and hagfishes caught inside and outside the Gullmar fjord. In a population of 6,000 lampreys, only one individual (0.017%) presented highly differentiated primary hepatocellular carcinoma. In contrast, tumors were detected in the hagfish population inside the fjord, and this percentage was significantly enhanced versus the tumor-bearing individuals in the open-sea control group (**Figure 2**). Of the 27,300 hagfishes captured within a 5-year period (1972–1976), up to 5.8% exhibited liver neoplasms (adenomas and carcinomas). This percentage was significantly higher compared to the tumor-bearing individuals (2.8%) in the control, open-sea population of 1,183 hagfishes caught outside the fjord. Although the affected hagfishes developed high- or low-differentiation tumors, they showed

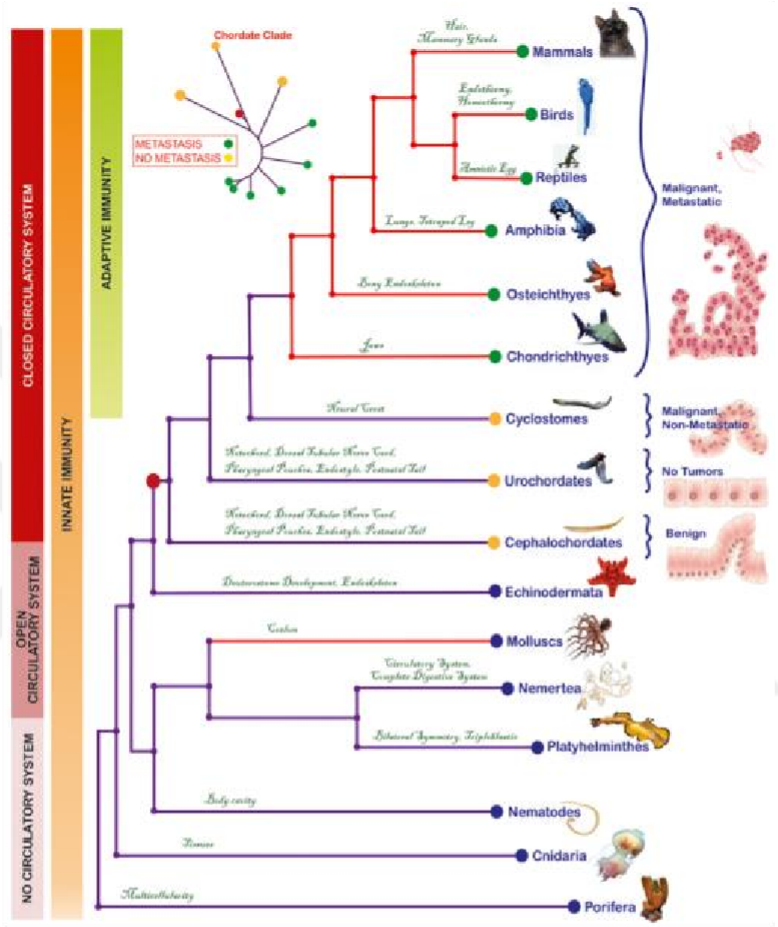


FIGURE 1 | Phylogenetic tree depicting the evolutionary relationships of metazoa and the emergence of metastasis in the cyclostome-to-chondrichthyes transition. Each phylogenetic group has manifested, for the first time, a key innovation that was conserved throughout the lineage. Red-colored clades indicate groups with lethal cancer types. Based on information derived from histological analyses, tumors were classified into benign (corresponding to initiation), malignant/non-metastatic (corresponding to promotion) or malignant, metastatic (corresponding to progression). Invasive and metastatic tumors have been reported from Chondrichthyes to mammals. The mollusks present contagious invasive cancer, while the urochordates are cancer-resistant. The red dot on the tree represents the common chordate ancestor, which gave rise to both, metastatic (green dots) and non-metastatic species (yellow dots). Model of the progression from a normal cell to metastatic cancer was modified from a previous illustration (Iacobuzio-Donahue et al., 2012). The origins of a closed circulatory system and of acquired immunity are also presented.

no macroscopic signs of metastasis (Falkmer et al., 1976, 1978; Falkmer, 1998). This increased cancer incidence was attributed to a combination of polychlorinated biphenyls (PCBs) and dichlorodiphenyltrichloromethylmethane (DDT), two anthropogenic organochloride contaminants which entered the fjord via washout. Due to parasitism of hagfishes on PCB/DDT-contaminated fishes, the organochlorides bioaccumulated in their liver or pancreatic islets, eventually triggering oncogenesis at these sites (Falkmer et al., 1978). The fact that despite their exposure to confirmed carcinogens (Lauby-Secretan et al., 2013; Loomis et al., 2015; Abu-Helil and van der Weyden, 2019), a percentage developed malignant tumors, but none of the 28,483 study individuals developed metastasis, leads to the suggestion that these animals may be metastasis-refractory or metastasis-incapable. Nevertheless, metastatic capability is evident in all descendant lineages. In particular, in Chondrichthyes, at least

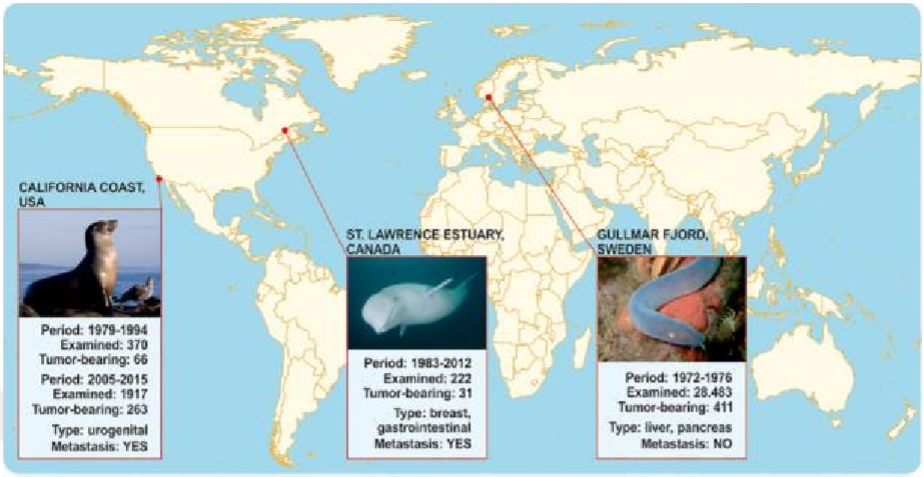


FIGURE 2 | Comparative epidemiology data on metastases in agnathan versus gnathostome representatives that inhabit comparable carcinogen-contaminated aquatic environments. No metastases have been reported in hagfishes inside and outside Gullmar fjord. On the contrary, toothed belugas in Quebec as well as California sea lions that inhabit environments polluted with similar organochlorides are susceptible to metastatic tumors.

50 cases of spontaneous cancer, including invasive (Wellings, 1969) and metastatic (Schlumberger and Lucke, 1948; Ostrander et al., 2004) tumors, have been recorded; in Osteichthyes, the tumor incidence increases and metastatic cases become more frequent (Schlumberger and Lucke, 1948; Wellings, 1969; Couch and Harshbarger, 1985; Groff, 2004); in amphibia (Stacy and Parker, 2004), reptiles (Abu-Helil and van der Weyden, 2019), birds (Abu-Helil and van der Weyden, 2019), and mammals (Albuquerque et al., 2018), there is an increased prevalence of invasive and metastatic cancers (Albuquerque et al., 2018). Collectively, a retrospective overview of tumor reports suggests the consistent occurrence of metastases in chondrichthyes and their descendants, providing hints that the establishment of metastatic potential coincides with the agnatha-to-gnathostome transition within the vertebrate clade (**Figure 1**).

The complicated life-cycle of cyclostomes, as well as ethical considerations regarding their research (Shimeld and Donoghue, 2012), challenge the application of experimental carcinogenesis protocols on adult individuals to simulate Falkmer's field study on a laboratory scale (Shimeld and Donoghue, 2012). Hence, as a surrogate test for corroborating associations between agnatha-to-gnathostome transition and metastatic potential, we sought to juxtapose Falkmer's nodal tumor pathology reports on adult cyclostomes to those of gnathostome representatives living in comparable PCB/DDT-contaminated environments (**Figure 2**). In particular, Beluga whales and California sea lions inhabit aquatic environments with persistent organic pollutants similar to those reported for the Gullmar fjord, and show unusually high cancer prevalence among marine mammals (Deming et al., 2018), which has been associated with exposure to carcinogens (Ylitalo et al., 2005; Randhawa et al., 2015; Lair et al., 2016). As shown in **Figure 2**, in an isolated population of about 900 toothed belugas (*Delphinapterus leucas*) living in the heavily industrialized St Lawrence Estuary of Quebec, Canada, cancer was reported as one of the most frequent causes of death (14%, 31/222 animals) in 222 carcasses found stranded or drifting, from 1983 to 2012. Tumors were often metastatic and fatal, and associated with the gastrointestinal tract (adenocarcinoma of the gastrointestinal mucosa, salivary gland and cholangiocellular carcinoma) and mammary glands. Exposure to carcinogens has been associated to increased cancer incidence (Lair et al., 2016), since living and dead beluga tissues were heavily contaminated by agricultural and industrial contaminants, including PCB/DDTs and their metabolites (Martineau et al., 2002). Similarly, between 1979 and 2015, necropsies of 2,287 sea lions beached-off central California coast in United States revealed high cancer incidence, where the predominant neoplasms were poorly differentiated urogenital carcinomas, with frequent local invasions and widespread metastases (Lipscomb et al., 2000; Deming et al., 2018). Metastasis was diagnosed in 18% (66/370) of necropsied animals from 1979 to 1994. From 2005 to 2015, 14% (263/1917) of cases had cancers, the vast majority of which were metastatic (Lipscomb et al., 2000; Deming et al., 2018), localized in the urogenital tissues and associated with organochloride bioaccumulation (Ylitalo et al., 2005; Randhawa et al., 2015). Hence, in similar carcinogenic environments, gnathostome species appear susceptible to aggressive cancers in contrast to the metastasis-refractory Gullmar fjord hagfishes. Similarly, organochloride pollutants have been correlated with risk of metastasis in human breast cancer patients (Koual et al., 2019). The aforementioned observations suggest that gnathostomes might be more prone to metastasis than agnatha

upon exposure to carcinogenic pollutants. If this is indeed the case, then macroevolutionary gains of gnathostomes, such as the cartilaginous jaws, emerge as a key innovation possibly associated with metastasis.

JIGs Undergo Frequent Mutations and DNA Methylation Alterations in Human Cancers

Agnatha-to-gnathostomes transition was promoted by the evolution of a cartilaginous skull, along with the establishment of jaws. These novelties facilitated the emergence of a complex brain and senses, that, together with the pharyngeal cartilage, allowed gnathostomes to shift to active predation, intermittent feeding and behavioral diversification (Kaucka and Adameyko, 2019). They highlight vertebrates' evolutionary success and, thus, are conserved from chondrichthyes to human (Kaucka and Adameyko, 2019). Based on the observation that the phenotypic manifestations of metastasis coincide with the evolutionary time point of occurrence of gnathostome key innovations, we used computational methods to unravel links between jaw formation and metastatic potential. We postulated that if our hypothesis is valid, then genes supporting formation of cartilaginous jaws would tend to be deregulated during cancer progression. To identify genes essential for the development of cartilaginous jaws, we screened the Mouse Genome Informatics (MGI) database for knockout-mice phenotypes that encompass jaw-related defects. In this way, 305 JIG were identified, all of which exhibit highly conserved human orthologs (**Supplementary Table 2**). The term "JIG," as used herein, refers to any gene which, if impaired, leads to abnormal jaw embryonic development. Notably, as indicated by the calculated ratios of jaws phenotypes to all affected phenotypes (**Supplementary Table 2**), JIGs are not functionally restricted only to jaw development, however mutation in even one of them leads to jaw malformations. GSEA revealed JIGS' involvement in skeletal system and cartilage development, appendage morphogenesis, and pattern specification (**Figure 3A**); molecular functions such as DNA binding, transactivation activity and signaling receptor binding (**Figure 3B**); and hallmark processes like Wnt-beta, TGF-beta and NOTCH signaling, and EMT (**Figure 3C**). STRING analysis indicated that 173 of these factors form a highly interconnected network (**Figure 3D**). Altogether, these data suggest that JIGs interact either physically or functionally within the context of jaw formation, to create a network, to which we, hereafter, refer to as jaw-developmental network (JDN).

Next, we examined whether JIGs undergo frequent genetic and/or epigenetic alterations in tumors. First, we juxtaposed three distinct lists of identified cancer gene drivers and mutations (Bamford et al., 2004; Kandoth et al., 2013; Bailey et al., 2018) with the list of JIGs to determine the number of JIGs that are mutated across human cancer types. To ensure that the association is non-random, we compared to 100 control lists each encompassing an equal number of 305 random, unrelated genes (available on request). A significant enrichment of CGC factors was observed among the JIGs 16.7%, or 51 genes) relative to random gene lists of the same size (3.5 ± 1.2%) and relative to the expected value (3.7%, or 721 of 19,313 coding genes from GENCODEv32, $p < 0.001$) (**Figure 4A**, left). The same tendency was observed when JIGs were juxtaposed to 299 driver genes and mutations identified from a comprehensive PanCancer and PanSoftware analysis spanning 9,423 tumor exomes by employing 26 computational tools (Bailey et al., 2018). JIGs represent 8.9% (27 genes) of the identified cancer drivers versus 1.0 ± 0.6% of random genes ($p < 0.001$) (**Figure 4A**, middle). Similar results were obtained when the same analysis was performed versus a group of 127 significantly mutated genes which have been identified as oncogenesis drivers across 12 major cancer types, whereby most tumors bear two to six of these mutations (Kandoth et al., 2013). A significant enrichment of JIGs for 18 of these genes (5.9% compared to expected 0.6%, which is 127 in 19,313 coding genes) was observed, as opposed to the random genes (0.5 ± 0.4%, $p < 0.001$) (**Figure 4A**, right). Overall, compared to the control lists, JIGs are highly enriched in cancer driver genes and mutations (**Figure 4A** and **Supplementary Table 3**). Then, using gene mutation data from PanCan cohort, we calculated the mutation frequency for all JIGs across human cancer types, and found a significant increase compared to control lists (**Figure 4B**). To further assess whether JIGs tend to be epigenetically altered in cancer, we meta-analyzed data of the DiseaseMeth 2.0 database (**Figure 4C**), and found frequent alterations of DNA methylation in JIGs, while hypomethylation was the most prevalent type of aberration in tumors versus normal tissue controls (131 JIGs hypomethylated, 47 hypermethylated, 47 hypo-/hyper-methylated, $\chi^2 = 62.72$, $p < 0.001$). Collectively, JIGs appear to undergo mutations and/or perturbations of DNA methylation patterns in cancer.

JIGs Are Transcriptionally Deregulated in Aggressive Stages and Predict Cancer Patient Outcomes

To further explore potential links between jaw development and cancer progression, we checked invasive cancer cell lines and patient tumors for changes in JIG transcriptional activity. In this regard, based on a recently described approach (Logotheti et al., 2020), we classified all cell lines of the Cancer Cell Line Encyclopedia (CCLE; Barretina et al., 2012), which includes gene expression data of 962 cell lines, into highly-invasive and less-invasive types, according to the levels of E-cadherin, N-cadherin, Vimentin, ZEB1, and SNAI1, which constitute reliable markers for EMT and tumor progression (Khan et al., 2017). Then, we examined whether JIGs are differentially expressed in highly- versus less-invasive cells across 24 common cancer types. We found that, compared to the control lists, a significantly higher number of JIGs is differentially expressed in high- versus low-aggressive cells (1.8-fold higher, $z = 8.89$, $p < 0.001$), where more JIGs are upregulated (JIGs vs. control: 100 vs. 39.1 ± 6.6 genes) than downregulated (39 vs. 40.4 ± 6.5 genes, $p < 0.003$, **Figure 5A** and **Supplementary Table 3**). These results indicate a non-stochastic tendency for enhanced transcription of JIGs in highly invasive states.

Additionally, to evaluate the clinical relevance of these findings, we meta-analyzed gene expression data in 35 different

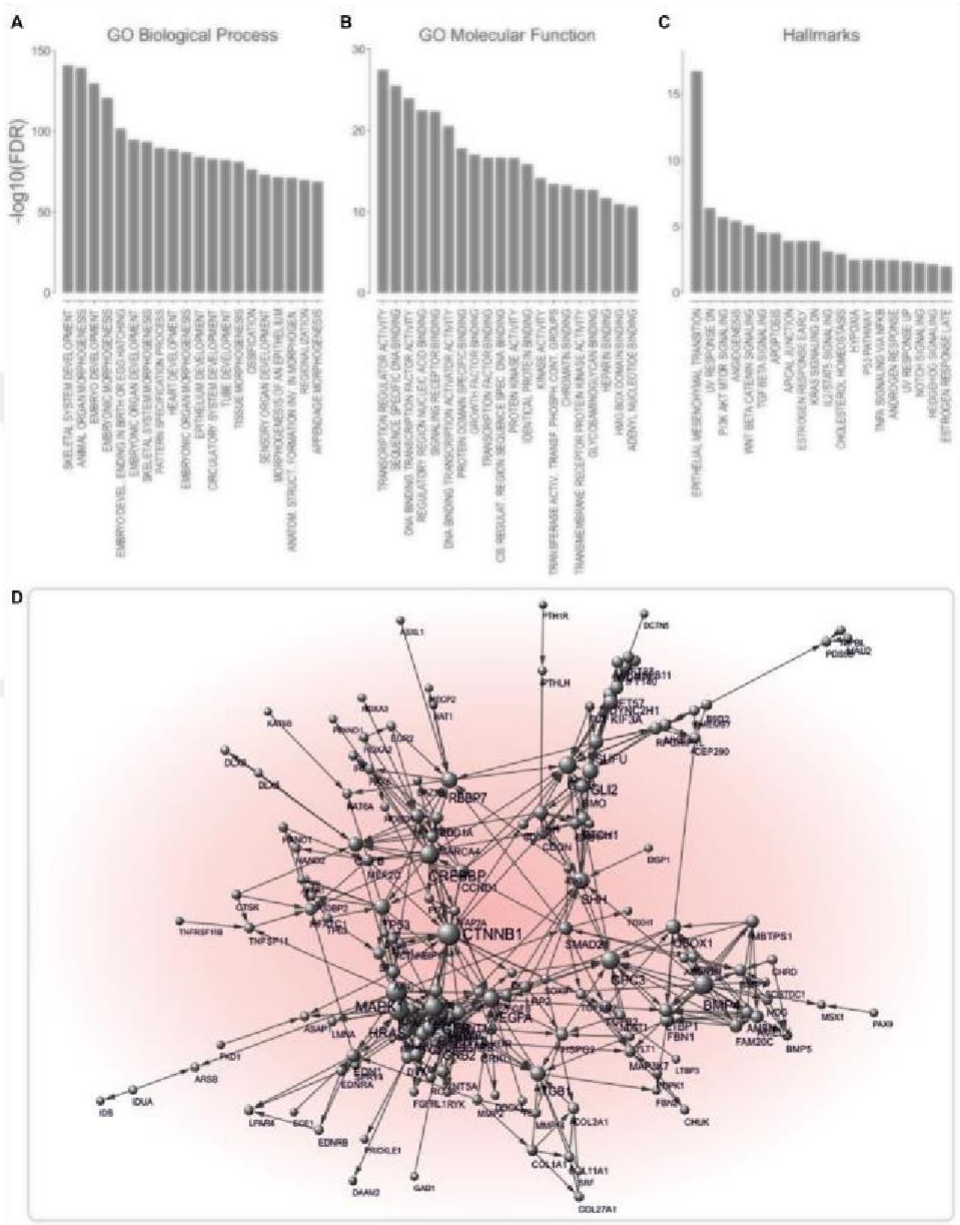


FIGURE 3 | JIGs exhibit a high degree of interconnectedness. **(A–C)** Analysis of enriched GO biological processes **(A)** GO molecular functions, **(B)** and hallmark processes, **(C)** indicates a role of JIGs in the development and EMT, mainly via affecting transcriptional regulation. **(D)** Interaction network of JIGs (interaction confidence score $\geq$ 0.9) shows that 173 of 305 genes/proteins are highly interconnected. The nodes represent genes/proteins and the connecting edges functional associations. Node sizes reflect node degree.

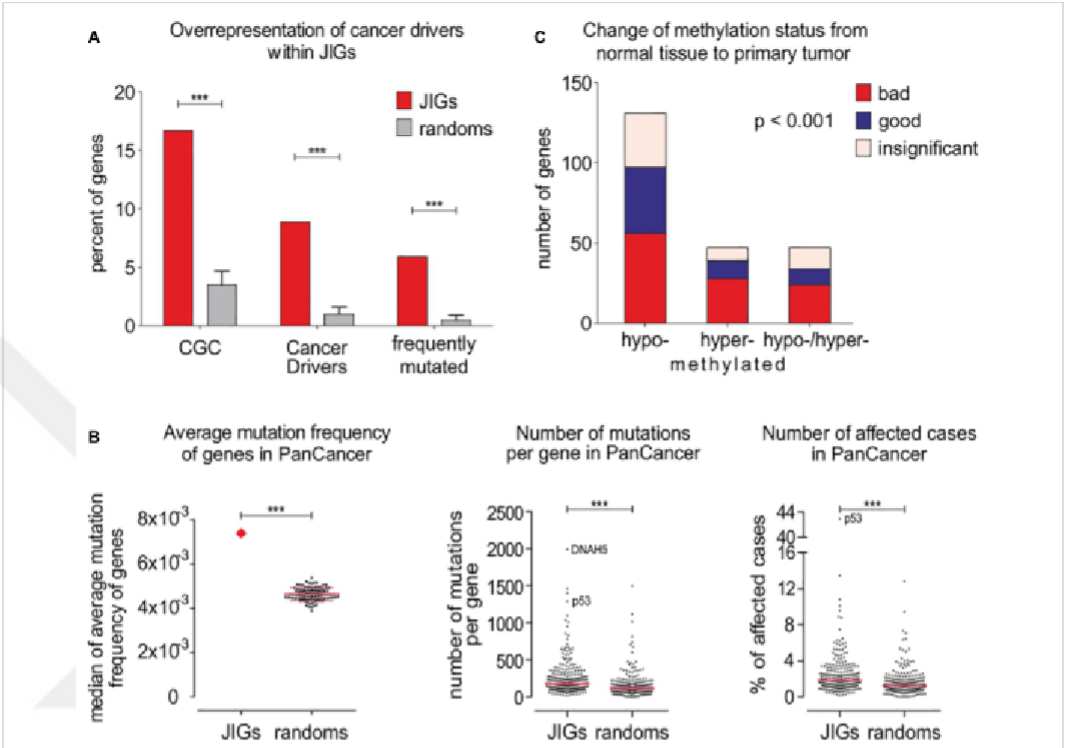


FIGURE 4 | JIGs undergo frequent gene mutations and epigenetic alterations across cancer types. **(A)** JIG enrichments in known and frequently mutated cancer drivers as indicated versus controls were 16.7% vs. 3.5 ± 1.2% ($p < 0.001$), **(left)**; 8.9% vs. 1.0 ± 0.6% ($p < 0.001$), **(middle)**; and 5.9% vs. 0.5 ± 0.4% ($p < 0.001$), **(right)**. **(B)** The average mutation frequency of JIGs is higher in PanCan cohort versus genes of the random lists **(left)**. This entails a higher number of mutations per individual JIG **(center)** and a higher number of patients exhibiting mutations in JIGs **(right)**. Pink lines: medians. **(C)** Almost 74% (225/305) of JIGs undergo aberrant DNA methylation in 42 cancer types versus normal control tissues, whereby 131 are significantly hypomethylated, 47 are hypermethylated, and 47 are either hypomethylated or hypermethylated in a cancer type-dependent manner. Cox regression analysis on PanCan TCGA cohort data showed that more hypo- than hyper- methylated JIGs are correlated with either poor ("bad," red color) or favorable ("good," blue color) patient outcomes. Bars represent SEM.

cancer types from PanCan TCGA cohort (Liu et al., 2018). Interestingly, correlation analyses implied that a large fraction of JIGs are co-expressed in patient tumors, with a tendency to preserve their crosstalks. In particular, relative to the control lists JIGs demonstrated increased value of (i) median expression correlation (MEC) per gene (JIG vs. control: 0.039 ± 0.045 vs. 0.008 ± 0.003, $z = 6.99$, $p < 0.001$, **Figure 5B**, red-boxed diagram), and (ii) more and greater positive correlation areas per gene (13.72 ± 13.05 vs. 3.04 ± 5.39, $z = 33.6$, $p < 0.001$, **Figure 5B**, green-boxed diagram). Using Cox regression, we deciphered all genes associated with patient survival, and subsequently determined the percentage of significant prognostic factors among JIGs. Overall, 227 JIGs correlated with patient outcomes, and showed a higher ratio of poor versus favorable prognostic factors (1.64:1, $\chi^2 = 13.33$, $p < 0.001$) compared to the corresponding ratio of the control lists (1.25:1, $\chi^2 = 2.79$, $p = 0.095$, correspondingly). We also found that 118 JIGs are both, deregulated in highly aggressive states in CCLE and correlated with patient outcomes in PanCan, with a ratio of 1.68:1 for poor versus favorable prognostic factors ($\chi^2 = 7.63$, $p < 0.006$) as opposed to the corresponding 0.93:1 ratio of the control lists ($\chi^2 = 0.1$, $p = 0.755$, **Figure 5C** and **Supplementary Table 3**). Collectively, JIGs appear to become deterministically deregulated in highly invasive cells, with frequent transactivation events, and predict poor patient outcomes. Similar alterations in JIGs also occur in metastatic lesions. In particular, we additionally compared these data with transcriptomes of primary versus metastatic lesions from breast, colon, hepatocellular, medulloblastoma, melanoma, and prostate cancers, that were retrieved from the GEO database (Clough and Barrett, 2016). We found that more JIGs that were poor prognostic factors in PanCan cohort are deregulated in metastases, as opposed to the corresponding favorable factors (1.85:1, $\chi^2 = 15.21$, $p < 0.001$, **Figure 5D** and **Supplementary Table 3**).

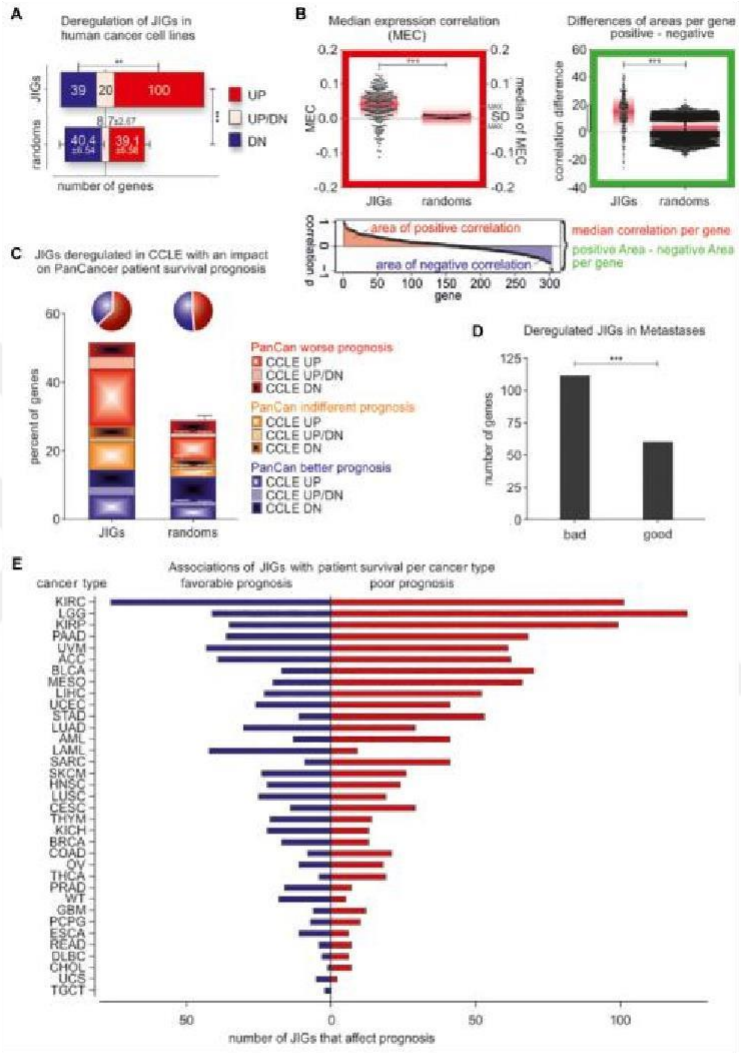


FIGURE 5 | JIGs are transcriptionally deregulated in invasive/metastatic stages and predict clinical outcomes in a cancer type-dependent manner. **(A)** JIG transcripts are deregulated in the highly invasive versus the less-invasive CCLE cell lines, and the total percentage is 1.8-fold higher than the corresponding percentage of control lists. Red: upregulated, blue: downregulated, pink: cancer type-dependent deregulation. Error bars represent mean with SD. **(B)** *Red-boxed diagram:* median expression correlation (MEC) in PanCan and SD of MEC (pink boxes) are higher among JIGs compared to random lists (shown are the 100 medians of MEC with maximal SD). *Green-boxed diagram:* the differences of positive versus negative area of correlation per gene show (a) a significantly larger span and (b) over all a higher positive value compared to 100 random gene lists. Pink lines represent the median, pink boxes represent SD. **(C)** Cox regression analysis of JIGs on the survival of the TCGA PanCan cohort. 227 JIGs correlated with patient outcomes, of which 141 predicted worse and 86 better outcomes. Among 159 deregulated JIGs in highly-aggressive states across several cancer types of CCLE, 118 JIGs (74.2%) influence patient prognosis. Of those, 74 (62.7%) (pie chart, blue) had an adverse effect, 51 were exclusively upregulated in aggressive states of CCLE (CCLE UP, plain red), while additional 10 JIGs were upregulated in some as well as downregulated in other cancer types (CCLE UP/DN, plain pink). Only 44 JIGs differentially expressed in CCLE (dashed blue, plain blue, and dashed pink) had a beneficial effect on survival prognosis (pie chart, green). **(D)** Transcriptional deregulation of JIGs in metastases versus primary tumor or normal tissue, and correlation with JIGs that affect patient prognosis in PanCan. JIG deregulations in 6 metastatic cancer types are more correlated with poor outcomes (1.85:1). **(E)** Survival analysis for JIGs across different TCGA cancers. The prognostic potential of JIGs depends on the cancer type. In the majority of cancer types, more JIGs are associated with poor prognosis and less with favorable prognosis. No cancer tissue was found unaffected by JIGs.

TABLE 1 | Effects of JIGs on cancer patient prognosis (in descending order of total number of JIGs that affect each cancer type).

Abbreviation	Cancer type	Number of JIGs associated with poor prognosis	Number of JIGs associated with favorable prognosis
KIRC	Kidney renal clear cell carcinoma	115	87
LGG	Brain Lower Grade Glioma	130	49
KIRP	Kidney renal papillary cell carcinoma	112	46
PAAD	Pancreatic adenocarcinoma	89	47
UVM	Uveal Melanoma	75	52
ACC	Adrenocortical carcinoma	77	49
MESO	Mesothelioma	88	25
BLCA	Bladder Urothelial Carcinoma	89	21
LIHC	Liver hepatocellular carcinoma	75	29
STAD	Stomach adenocarcinoma	72	21
UCEC	Uterine Corpus Endometrial Carcinoma	54	33
LUAD	Lung adenocarcinoma	46	40
AML	Acute Myeloid Leukemia	60	23
SKCM	Skin Cutaneous Melanoma	42	39
LAML	Acute Myeloid Leukemia	18	58
LUSC	Lung squamous cell carcinoma	40	34
SARC	Sarcoma	55	17
KICH	Kidney Chromophobe	34	35
HNSC	Head and Neck squamous cell carcinoma	34	33
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	39	25
BRCA	Breast invasive carcinoma	26	32
THYM	Thymoma	19	33
OV	Ovarian serous cystadenocarcinoma	31	19
THCA	Thyroid carcinoma	41	9
COAD	Colon adenocarcinoma	33	16
ESCA	Esophageal carcinoma	15	25
GBM	Glioblastoma multiforme	24	15
PCPG	Pheochromocytoma and Paraganglioma	21	16
PRAD	Prostate adenocarcinoma	10	26
WT	Wilms Tumor	7	29
READ	Rectum adenocarcinoma	24	8
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	14	11
CHOL	Cholangiocarcinoma	17	3
UCS	Uterine Carcinosarcoma	7	11
TGCT	Testicular Germ Cell Tumors	1	5

Next, we estimated the effect of each JIG on prognosis for individual cancer types, using Cox regression (**Figure 5E**, **Table 1**, and **Supplementary Table 4**). We found that all JIGs correlate with patient outcomes and that a single JIG can affect prognosis of at least three cancer types. Although there were JIGs implicated in as many as 16 cancer types, no single JIG was universally associated with prognosis in all 35 cancers. Rather, their effect is cancer type-dependent. The top-ten most impacted cancer types, ranked by the total number of JIGs affecting, either beneficially (left) or detrimentally (right), the disease outcome are KIRC, LGG, KIRP, PAAD, UVM, ACC, MESO, BLCA, LIHC, and STAD.

Gnathostome-Specific JIGs Are Preferably Deregulated During Cancer Progression

The developmental origin of the cartilaginous jaw consists a turning point in vertebrate evolutionary history and is attributed to the neural crest (Sauka-Spengler and Bronner-Fraser, 2008), and especially to the cranial neural crest cells, from which cartilage is exclusively formed (Martik and Bronner, 2017). Evolutionary Developmental (Evo-Devo) biology studies comparing embryonic programs in jawless versus early jawed chordates indicate that instead of appearing *de novo*, jaws have rather arisen through the co-option of an ancient developmental pre-pattern (Cerny et al., 2010), in association with corresponding changes in the underlying GRNs. Jaw evolution was driven by incorporation of new genes into a pre-existing dorso-ventral patterning program, which altered the identity of jaw-forming chondrocytes (Cerny et al., 2010). For instance, some transcription factors are components of neural crest GRNs in both, jawed and jawless vertebrates, while other transcription factors are cranial neural crest-specific and included in gnathostomes, but missing from lampreys' GRNs (Martik and Bronner, 2017).

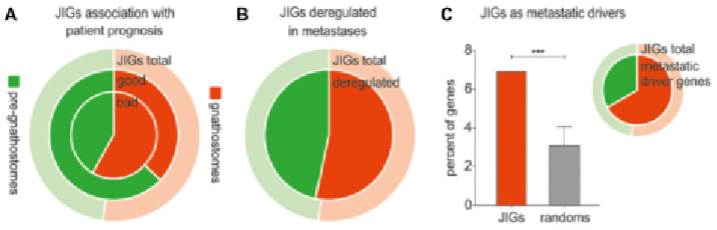


FIGURE 6 | Cancer progression and metastasis are associated with deregulation of evolutionarily younger JIGs. Red indicates gnathostome-specific, green indicates pre-gnathostome orthologs. In all pie charts, the outer, ring represents the composition of the total 305 JIGs in gnathostome-specific (light red) and pre-gnathostome (light green) JIGs. **(A)** Cox regression analysis of JIGs in PanCan revealed that 51.6% of gnathostome-specific orthologs predict worse and only 20.1% better prognosis (ratio 2.56:1) versus 40.4 and 37.0% (ratio 1.09:1) in pre-gnathostome genes. **(B)** In metastasis, a slightly higher number of gnathostome-specific JIGs is deregulated compared to the pre-gnathostomes (inner circle). **(C)** Estimation of gnathostome versus pre-gnathostome genes that are deregulated in a setting of xenograft experimental evolution. Diagram: among the 700 genes which were identified as drivers of tumor evolution, significant enrichment of JIGs was found (6.89%, or 21 genes) as compared to random lists (3.08 ± 0.98%). Pie: of the 21 deregulated JIGs, 14 (66.7%) were of gnathostome and 7 (33.4%) of pre-gnathostome origin.

According to the abovementioned Evo-Devo concepts, the JDN components do not have a similar evolutionary age. It is rather plausible that gene homologs which arose in gnathostomes might have got interconnected with a network that pre-existed in agnatha, perhaps to support morphological novelty. Motivated by this, we questioned whether invasive cancer cells show preference to "usurp" the pre-existing genes or the ones that were incorporated to the JDN after the divergence of jawed vertebrates from cyclostomes. In this respect, we approximated the evolutionary age of all JIGs (**Supplementary Table 5**) and assessed the effects of pre-gnathostome versus gnathostome-specific orthologs on cancer progression. Importantly, to eliminate bias in the estimation of the JIGs' evolutionary age from partially sequenced animal genomes, we included only model organisms with well-annotated genomes. We found that of 305 JIGs, 159 JIGs probably originated in jawed vertebrates, since no "true orthologs" of their corresponding proteins could be detected in jawless species, whereas the other 146 JIGs are of pre-gnathostome origin. Cox regression analysis in PanCan revealed that a higher number of gnathostome-specific JIGs is associated with poor patient outcomes, as opposed to the pre-gnathostome JIGs with an almost equal distribution (82:32 vs. 59:54, $\chi^2 = 8.56$, $p < 0.002$, **Figure 6A**). Consistently, a tendency toward transcriptional deregulation of gnathostome-specific versus pre-gnathostome JIGs was observed in metastases compared to primary tumors ($\chi^2 = 1.09$, $p < 0.296$, **Figure 6B**). We also performed a similar meta-analysis on data produced in a mouse experimental setting, where a tumor's evolution from initiation to metastasis has been simulated *in vivo* via sequential xenografting (Chen et al., 2015). We estimated gnathostome versus pre-gnathostome genes that are deregulated in a setting of xenograft experimental evolution, where a tumor's full-life history from initiation to metastasis was simulated by transforming the immortalized human breast epithelial cell line with *HRAS* and performing sequential xenografting in mice until metastases were observed (Chen et al., 2015). Among the 700 genes which underwent driver expression changes, on the basis that their expression was exclusively increasing or decreasing (Chen et al., 2015), we found that JIGs are significantly enriched ($z = 3.9$, $p < 0.001$), while their majority tends to be of gnathostome rather than pre-gnathostome origin (14:7, $\chi^2 = 1.71$, $p = 0.19$, **Figure 6C** and **Supplementary Table 3**). Collectively, the strong correlations of gnathostome-specific genes with poor and metastatic outcomes, as well as with drivers of experimental tumor evolution underscore a preference for deregulation of evolutionarily younger JIGs toward tumor progression.

Deregulated Expression of JDN Components During Cancer Progression Is More Pronounced in the Gnathostome-Specific Hubs

JIGs appear to form a highly interconnected network (**Figure 3D**). In biological networks, the most highly connected nodes, the so-called "hubs," are considered biologically significant and more relevant to the overall function of the network (Barabási et al., 2011; Kontou et al., 2016; Pavlopoulos et al., 2018). The intra-modular hubs are central to a given network module, with the highest number of connections to the neighboring nodes, whereas inter-modular hubs are intermediate between two or more modules. Taking this into account, we sought to investigate whether gnathostome-specific genes that are deregulated in cancer also occupy hub positions in the JDN. First we identified, through STRING network analysis, 60 nodes representing intra- and inter-modular hubs (**Figure 7A**). By approximating their evolutionary age, we found that these hubs correspond largely to gnathostome-specific versus pre-gnathostome orthologs (**Figure 7B** and **Supplementary Table 5**). Upon comparison of these 60 hubs to CCLE-derived data, we found that 33 (55.0%) of them are deregulated in highly invasive cancer cells, 25 of which are of gnathostome-, and 8 of pre-gnathostome origin ($\chi^2 = 7.76$, $p = 0.005$).

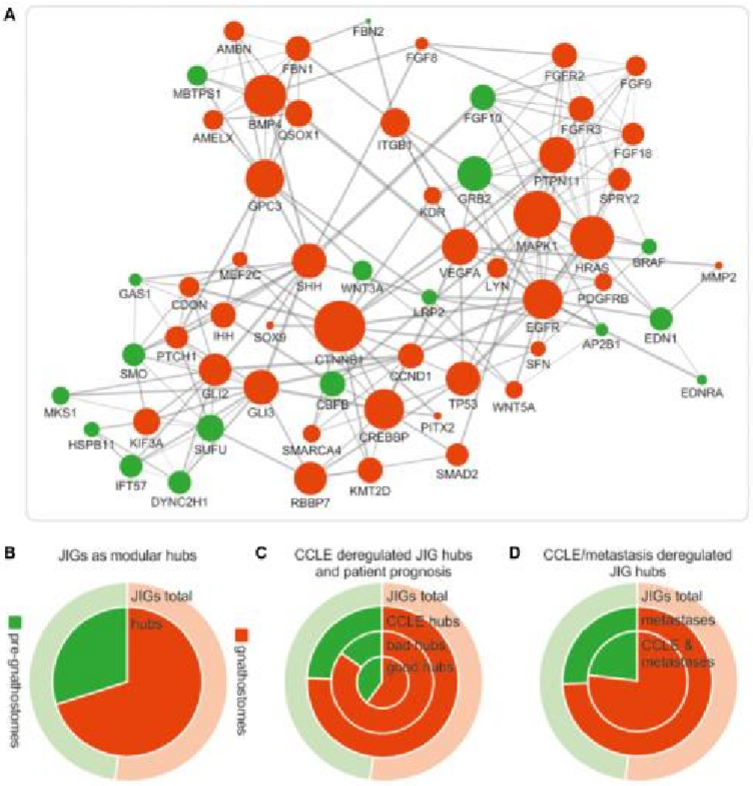


FIGURE 7 | Deregulations of gnathostome-specific hubs of the JDN are more frequent associated with invasive and metastatic stages, as opposed to pre-gnathostome hubs. **(A)** Functional interaction network (confidence interaction score ≥ 0.9) of the gnathostome-specific and pre-gnathostome JIG hubs. The size of the nodes is proportional to the node degree in the "original" JDN network in **Figure 3d**. The 60 nodes that represent intra- and inter-modular hubs in the original network appear to be interconnected and form a rather dense network. **(B)** The gnathostome-specific genes make up 52% (159) of all 305 JIGs (outer ring "JIG total," red), whilst the proportion of gnathostome hubs is significantly enriched in relation to the pre-gnathostome (inner circle "hubs," red). **(C)** More gnathostome-specific hubs that are deregulated in CCLE cells are correlated with worse prognosis in PanCan cohort (ring "bad hubs," red). In contrast, more pre-gnathostome hubs tend to correlate with favorable than poor prognosis (compare ring "bad hubs" with inner circle "good hubs," green). **(D)** More gnathostome-specific hubs are found deregulated, compared to the pre-gnathostome ones, in metastatic lesions (inner ring). This preference is more pronounced for hubs commonly deregulated in metastatic samples and CCLE cells (inner circle).

Notably, all 33 JDN hubs predicted by this approach are indeed causatively linked with invasiveness and/or metastasis in a cancer type-dependent manner, according to experimental evidence in the preclinical setting (**Supplementary Table 6**). In particular, studies using *in vitro* cell lines, mouse xenografts and/or patient samples underscore that most of the hubs promote tumor progression and metastasis. Further data curation in the clinical trials database showed that for several of these hubs, molecular modifiers have been developed and entered clinical trials, either as monotherapies or as combination regimens. In addition, screenings in the drugs.com online pharmaceutical encyclopedia and the publicly available database of the United States Food and Drug Administration (FDA) indicated that small molecules targeting six of these hubs are already approved and marketed drugs for the treatment of aggressive and metastatic cancers. Indeed, as shown in **Supplementary Table 6**, our model was able to identify EGFR which is targeted by approved drugs such as cetuximab, dacomitinib and erlotinib; members of the VEGF (VEGFA/KDR) pathway which are targeted by, for example, bevacizumab, pazopanib or sorafenib; members of the FGFR family of receptors, which are targeted by erdafitinib or pemigatinib; and effectors of the hedgehog signaling pathway (SMO, SHH) which are targeted by glasdegib, sonidegib or vismodegib. Other factors, such as GLI2, GLI3, CDON, MEF2C, PITX, SFN, SPRY2, WNT3A and WNT5A (**Supplementary Table 6**), show

a consistent metastasis-promoting role across several cancer types in preclinical studies, which provides a foundation for the development of molecular modifiers against them and their introduction to clinical trials.

Regarding the age of the hubs of JDN in relation with tumor invasiveness, our studies indicate that among the 60 JDN hubs, the evolutionarily younger ones tend to be more frequently deregulated in aggressive states. The tendency for deregulation of gnathostome-specific hubs as compared to the pre-gnathostome hubs is clinically relevant for cancer patients. We found that more gnathostome-specific hubs are correlated with worse than favorable prognosis (17:10 hubs, correspondingly), whereas the opposite applies for the pre-gnathostome hubs (4:9, $\chi^2 = 2.47$, $p = 0.116$, **Figure 7C**). In addition, aberrant transcription of more gnathostome-specific than pre-gnathostome hub JIGs was observed in metastatic versus primary tumors (35:12 = 2.92, $\chi^2 = 11.26$, $p < 0.001$, **Figure 7D**). This effect is particularly observed among those hubs that are deregulated in highly invasive cancer cell lines (25:8 = 3.13, $\chi^2 = 8.76$, $p = 0.003$, **Figure 7D**). Overall, there is a preference for transcriptional deregulation of gnathostome-specific hubs in the invasive and metastatic stages. Taken together, these results indicate a prevalence of gnathostome-specific genes in occupying hub positions in JDN. However, at the same time, these are preferentially exploited by cancer cells over pre-gnathostome hubs to promote cancer aggressiveness.

DISCUSSION

Cancer is considered an evolutionary and ecological process (Merlo et al., 2006). A neoplasm consists of genetically and epigenetically heterogeneous cell populations that compete for space and resources, evade immune surveillance and cooperate to disseminate to secondary organs. The fitness of cell subpopulations is further shaped by their interactions with cellular and molecular components of the tumor microenvironment. The fittest, or "evolutionarily successful," cell variants are those acquiring capabilities which increase the probability to obtain metastatic potential (Merlo et al., 2006). Darwinian laws apply to both, the tumors and the organisms on which tumors grow. Somatic selection occurs along with organismal selection, following "a mirror within a mirror"-like pattern: heterogeneous cell subpopulations of a growing tumor undergo selection of the fittest within a population of organisms which is under constant evolutionary pressure. Although the timeframe for clonal selection of tumor cells (in months or years) is significantly more narrow than for species selection (millions of years), it is reasonable to envisage that, at a given evolutionary time point, the attributes encoded in the genome of a specific species can be accessed by its cancerous tumor.

Herein, we provide compelling evidence that metastasis is phenotypically manifested within the gnathostome clade, and that genes which are essential for jaw development, a hallmark macroevolutionary trait of gnathostomes, are co-opted during cancer progression. Genes supporting jaw developmental programs tend to undergo mutational and epigenetic changes, with frequent transactivations in invasive and metastatic stages, and a preference for enhanced transcription of the gnathostome-specific versus the ancient ones. These data strongly suggest that the same genes/gene interactions underlying key innovations are also preferentially co-opted within the tumor context toward aggressive outcomes. Certain structures which provide selective advantages at an organismal level, such as placenta development (Costanzo et al., 2018), neural crest formation (Kerosuo and Bronner-Fraser, 2012), and jaw development (this study) may be "hacked" by cancer cells to improve their own fitness, manifested as metastatic potential. The preference of invasive cancer cells to usurp evolutionarily newer hubs of the JDN further reinforces this notion. This observation also implies that there might be a trade-off between the vulnerability to metastasis and the conservation of key innovations that are indispensable for vertebrate fitness and, thus, cannot undergo secondary losses.

Successful prediction of the likely paths of tumor progression is valuable for diagnostic, prognostic, and treatment purposes, but effective models are still not in place (Diaz-Uriarte and Vasallo, 2019). For establishing such prediction models and designing drugs that target events of tumor evolution (Amirouchene-Angelozzi et al., 2017), it is essential to unveil parallels between organismal and clonal selection. Based on our findings, we introduce a systems-based, key innovation-driven model, as an *in silico* tool for prediction of putative prometastatic drivers, on the rationale that genes that are crucial for evolution of a species might be important for tumor evolution. The fact that the results derived using this model are in agreement with the bulk of preclinical studies and with clinical interventions (see **Supplementary Table 6** for details) underscores both, the prediction accuracy of this approach and the translational value of these candidates. In particular, our computational model predicted 33 hub JIGs that are deregulated in highly invasive cell lines of CCLE, consistently with experimental *in vitro*, *in vivo* and/or in patient evidence that these genes mainly support cancer progression. Several of these gene targets have already been translated to marketed drugs, while other factors predicted by this model might represent promising novel targets, since they show consistent metastasis-inducing effects across several cancers. New molecular entities able to inhibit these candidates can be developed and be further assessed in the clinical setting for their potential to prevent metastatic progression or disease recurrence. Out of these 33 hubs, 25 are of gathostome- versus 8 of pregnathostome-origin, indicating a pronounced tendency of cancer cells to usurp the evolutionarily-younger hubs of the JDN network in order to evolve to aggressive stages. Hence, taking into account that a tumor may progress by usurping genes specifically related with vertebrate key-innovations, this model may hold a potential to facilitate the prediction of tumor evolutionary trajectories.

A main future challenge is to design comprehensive experimental settings, where associations between key innovations and metastasis could be investigated at the mechanistic and molecular level, in parallel with corresponding cancer phenotypes. Studies in primitive metastasis-competent taxa versus the incompetent ones could facilitate reconstruction

of the evolutionary history of metastasis by identifying which molecular events that catalyzed gnathostome evolution have also consistently benefited the evolution of primary tumors toward more aggressive stages. Advantageously, a well-established Evo-Devo study system of gnathostomes versus pre-gnathostomes, amenable to experimental and genetic manipulations at embryonic stages (Kuratani et al., 2002; Cerny et al., 2010; Jandzik et al., 2015), could be repurposed for cancer research. This *in vivo* system is comprised of three chordata species of the same evolutionary lineage, presenting progressively aggressive cancer phenotypes from the most ancient to the most recent: amphioxus, the most basal extant chordate, with benign tumors; the metastasis-incompetent jawless lampreys; and the metastasis-competent jawed zebrafish. By comparing lamprey and amphioxus development with that of zebrafish, and other vertebrates like frog and salamander, Evo-Devo specialists attempt to reconstruct genetic and developmental changes underlying the major events in vertebrate evolution (Kuratani et al., 2002; Cerny et al., 2010; Jandzik et al., 2015). Subjecting this system into carcinogenic treatments and comparatively examining the developing lesions in conjunction with genetic and functional changes might provide a glimpse into conserved pathways which are consistently recapitulated in metastases. For instance, it could be investigated whether absence of certain gnathostome-specific genes/gene interactions in the jawless organisms hinders activation of prometastatic cascades in their respective tumors, as opposed to their jawed counterparts. Moreover, studies on cancer-resistant pre-vertebrates, such as echinoderms, urochordates and cephalochordates, could also be considered to dissect their cancer resilience in relation with their hallmark attributes, for instance their increased regeneration ability (Somorjai et al., 2012).

Our key innovation-driven model underscores a preference of invasive tumors for usurping evolutionarily newer features. Otherwise, a recent concept, known as "atavistic model," asserts that neoplasms rely on re-expression of ancestral traits and reverse evolution from multicellularity (MC) to unicellularity (UC; Bussey et al., 2017), which is likely promoted by upregulation of UC genes, disruption of interconnectedness between UC and MC genes (Trigos et al., 2017) and loss-of-function mutations on MC genes (Chen et al., 2015). Furthermore, recent studies show that metastatic competence arises from heterogeneous cancer cell populations without the need for acquisition of additional mutations, and is benefited from further selection of tumor-initiating mutations that seed primary tumorigenesis (Jacob et al., 2015). By unifying these notions, we propose that tumor initiation may be triggered by mutations in evolutionarily old genes governing processes at the "dawn" of multicellularity, such as cell proliferation and genomic stability (Lineweaver et al., 2014). Such alterations in ancient genes may confer genetic heterogeneity (Chen et al., 2015), which is the driving force of evolution. Subsequently, tumor progression may be enabled via selection of clones that entail crosstalks between tumor-initiating mutations already acquired at primary tumor cells and evolutionarily-young genes that support gnathostome key-innovations. Future studies could uncover if such crosstalks guide recurrent routes to metastasis, thereby providing a foundation for the rational design of strategies that prevent cancer evolution.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SL, SM, and BP conceived the project and designed experiments. SM developed bioinformatics pipelines and analyzed high-throughput data. AP and IT designed and performed phylogenomics and network analyses. PD and SL performed literature research. SL, SM, and AP analyzed and interpreted data. IT and SM prepared the figures. SL drafted the manuscript. SL, AP, and BP reviewed and edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by funding from the German Federal Ministry of Education and Research e:Med-MelAutim (BMBF grant 01ZX1905D), German Cancer Aid (Deutsche Krebshilfe grant 70112353), and Deutsche Forschungsgemeinschaft (DFG grant PU 188/17-1). IT acknowledges the YÖK (Yükseköğretim Kurulu) scholarship.

ACKNOWLEDGMENTS

All contributors to this study are included in the list of authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.682619/full#supplementary-material

Supplementary Figure 1 | Maximum likelihood-based tree of protein sequences of the p53 family. A characteristic example of the well-studied *p53* gene family, where distinct orthologs of the three members of this family (i.e., TP53, TP63 and TP73) were found exclusively in gnathostome species. The agnathan p53 homologs occupy a basal position in this phylogenetic tree, suggesting that a primordial *TP53/63/73* gene might have existed in jawless species which after a series of duplications gave rise to the functionally divergent *TP53*, *TP63* and *TP73* genes in jaw-bearing vertebrates.

Supplementary Table 1 | Taxa under study in correlation with their tumor characteristics, their circulatory and immune systems, the level of body organization and their ability to develop metastatic or lethal cancers.

Supplementary Table 2 | Mouse gene knockouts which present jaws abnormalities. The human orthologs are included. The ratio of the jaw-related

phenotypes to all developmental phenotypes affected by each gene knockout is also shown.

Supplementary Table 3 | Associations of each JIG with genetic and epigenetic alterations in tumors, survival of PanCan patients, and expression in highly invasive or metastatic stages. Column annotations: CGC = included in the Cancer Gene Census list, Cancer driver = included among the cancer driver genes in PanCan, freqmut = included among the frequently mutated genes in PanCan, Methylation = differentially methylated JIGs in human tumors versus normal tissue, from DiseaseMeth database (hypo = hypomethylated, hyper = hypermethylated), CCLE = up- or down-regulated in highly-invasive vs. less-invasive human cells from the CCLE, Survival = impact on the survival prognosis of the patients from the PanCancer cohort based on Cox regression analysis, hsa_Metastases = up- or down-regulated in human metastatic vs. primary lesions (GSE21510, GSE2509, GSE25976, GSE43837, GSE488, GSE6919, GSE7929, GSE7930, GSE8401), mmu_metast_driver = included among metastasis driver genes in a mouse model of tumor evolution.

Supplementary Table 4 | JIGs and their effect on patient survival prognosis across different cancer types of TCGA based on Cox regression analysis. Annotations: poor prognosis = 1, favorable prognosis = −1, insignificant = 0.

Supplementary Table 5 | JIG orthologs across pre-gnathostome and gnathostome species. Annotations: green = pregnathostome orthologs, orange: gnathostome-specific orthologs, underlined = network hubs.

Supplementary Table 6 | Comprehensive overview of the preclinical and clinical research studies on the hub JIGs that are deregulated in highly-invasive cell lines of CCLE. The drivers of tumor progression predicted using the key innovation-driven model are in agreement with existing experimental data.

REFERENCES

Abu-Hilil, B., and van der Weyden, L. (2019). Metastasis in the wild: investigating metastasis in non-laboratory animals. *Clin. Exp. Metastasis.* 36, 15–28. doi: 10.1007/s10585-019-09956-3

Albuquerque, T. A. F., Drummond do Val, L., Doherty, A., and de Magalhães, J. P. (2018). From humans to hydra: patterns of cancer across the tree of life. *Biol. Rev. Camb. Philos. Soc.* 93, 1715–1734. doi: 10.1111/brv.12415

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Amirouchene-Angelozzi, N., Swanton, C., and Bardelli, A. (2017). Tumor evolution as a therapeutic target. *Cancer Discov.* 7, 805–817.

Angelova, M., Mlecnik, B., Vasaturo, A., Bindea, G., Fredriksen, T., Lafontaine, L., et al. (2018). Evolution of metastases in space and time under immune selection. *Cell* 175, 751–765.e16.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e18.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* 91, 355–358. doi: 10.1038/sj.bjc.6601894

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Barber, B. J. (2004). Neoplastic diseases of commercially important marine bivalves. *Aquat. Living Resour.* 17, 449–466. doi: 10.1051/alr:2004052

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607.

Billaud, M., and Santoro, M. (2011). Is Co-option a prevailing mechanism during cancer progression? *Cancer Res.* 71, 6572–6575. doi: 10.1158/0008-5472.can-11-2158

Birkbak, N. J., and McGranahan, N. (2020). Cancer genome evolutionary trajectories in metastasis. *Cancer Cell.* 37, 8–19. doi: 10.1016/j.ccell.2019.12.004

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E., and Group, M. G. D. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* 47, D801–D806.

Bussey, K. J., Cisneros, L. H., Lineweaver, C. H., and Davies, P. C. W. (2017). Ancestral gene regulatory networks drive cancer. *Proc. Natl. Acad. Sci. U.S.A.* 114, 6160–6162. doi: 10.1073/pnas.1706990114

Carballal, M. J., Barber, B. J., Iglesias, D., and Villalba, A. (2015). Neoplastic diseases of marine bivalves. *J. Invertebr. Pathol.* 131, 83–106. doi: 10.1016/j.jip.2015.06.004

Cerny, R., Cattell, M., Sauka-Spengler, T., Bronner-Fraser, M., Yu, F., and Medeiros, D. M. (2010). Evidence for the prepattern/cooption model of vertebrate jaw evolution. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17262–17267. doi: 10.1073/pnas.1009304107

Chen, H., Lin, F., Xing, K., and He, X. (2015). The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nat. Commun.* 6:6367.

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Cooper, M. D., and Alder, M. N. (2006). The evolution of adaptive immune systems. *Cell* 124, 815–822. doi: 10.1016/j.cell.2006.02.001

Costanzo, V., Bardelli, A., Siena, S., and Abrignani, S. (2018). Exploring the links between cancer and placenta development. *Open Biol.* 8:180081. doi: 10.1098/rsob.180081

Couch, J. A., and Harshbarger, J. C. (1985). Effects of carcinogenic agents on aquatic animals: an environmental and experimental overview. *J. Environ. Sci. Health Pt. C Environ. Carcinog. Rev.* 3, 63–105. doi: 10.1080/10590508509373329

Dawe, C. J. (1969). Phylogeny and oncogeny. *Natl. Cancer Inst. Monogr.* 31, 1–40.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439, 965–968. doi: 10.1038/nature04336

Deming, A. C., Colegrove, K. M., Duignan, P. J., Hall, A. J., Wellehan, J. F. X., and Gulland, F. M. D. (2018). Prevalence of urogenital carcinoma in stranded California sea lions (*Zalophus californianus*) from 2005-15. *J. Wildl. Dis.* 54, 581–586. doi: 10.7589/2017-08-208

Diaz-Uriarte, R., and Vasallo, C. (2019). Every which way? On predicting tumor evolution using cancer progression models. *PLoS Comput. Biol.* 15:e1007246. doi: 10.1371/journal.pcbi.1007246

Domazet-Lošo, T., Klimovich, A., Anokhin, B., Anton-Erxleben, F., Hamm, M. I., Lange, C., et al. (2014). Naturally occurring tumours in the basal metazoan Hydra. *Nat Commun.* 5:4222.

Falkmer, S. (1998). "The tumour pathology of *Myxine glutinosa*," in *The Biology of Hagfishes*, 1 Edn, eds J. M. Jorgensen, J. P. Lomholt, R. E. Weber, and H. Malte (London: Chapman and Hall), 101–106. doi: 10.1007/978-94-011-5834-3_7

Falkmer, S., Emdin, S. O., Ostberg, Y., Mattisson, A., Sjöbeck, M. L., and Fänge, R. (1976). Tumor pathology of the hagfish, *Myxine glutinosa*, and the river lamprey, *Lampetra fluviatilis*. A light-microscopical study with particular reference to the occurrence of primary liver carcinoma, islet-cell tumors, and epidermoid cysts of the skin. *Prog. Exp. Tumor. Res.* 20, 217–250. doi: 10.1159/000398701

Falkmer, S., Marklund, S., Mattsson, P. E., and Rappe, C. (1978). Hepatomas and other neoplasms in the atlantic hagfish (*Myxine glutinosa*): a histopathologic and chemical study. *Ann. N. Y. Acad. Sci.* 298, 342–355. doi: 10.1111/j.1749-6632.1977.tb19277.x

Fontaine, A. R. (1968). Pigmented tumor-like lesions in an ophluroid echinoderm. *Natl. Cancer Inst. Monogr.* 31, 255–261.

Fürst, K., Steder, M., Logotheti, S., Angerilli, A., Spitschak, A., Marquardt, S., et al. (2019). DNp73-induced degradation of tyrosinase links depigmentation with EMT-driven melanoma progression. *Cancer Lett.* 442, 299–309. doi: 10.1016/j.canlet.2018.11.009

Gess, R. W., Coates, M. I., and Rubidge, B. S. (2006). A lamprey from the Devonian period of South Africa. *Nature* 443, 981–984. doi: 10.1038/nature05150

Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. (2015). Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* 43, D1079–D1085.

Groff, J. M. (2004). Neoplasia in fishes. *Vet. Clin. North Am. Exot. Anim. Pract.* 7, 705–756. doi: 10.1016/j.cvex.2004.04.012

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Harshbarger, J. C. (1969). The registry of tumors in lower animals. *Natl. Cancer Inst. Monogr.* 31, XI–XVI.

Heimberg, A. M., Cowper-Sal-lari, R., Sémon, M., Donoghue, P. C., and Peterson, K. J. (2010). microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19379–19383. doi: 10.1073/pnas.1010350107

Hickman, C. P., Roberts, L. S., and Larson, A. (2001). *Integrated Principles of Zoology*. New York, NY: McGraw-Hill.

Hunter, J. P. (1998). Key innovations and the ecology of macroevolution. *Trends Ecol Evol.* 13, 31–36. doi: 10.1016/s0169-5347(97)01273-1

Iacobuzio-Donahue, C. A., Velculescu, V. E., Wolfgang, C. L., and Hruban, R. H. (2012). Genetic basis of pancreas cancer development and progression: insights from whole-exome and whole-genome sequencing. *Clin. Cancer Res.* 18, 4257–4265. doi: 10.1158/1078-0432.ccr-12-0315

Iacob, L. S., Vanharanta, S., Obenauf, A. C., Pirun, M., Viale, A., Socci, N. D., et al. (2015). Metastatic competence can emerge with selection of preexisting oncogenic alleles without a need of new mutations. *Cancer Res.* 75, 3713–3719. doi: 10.1158/0008-5472.can-15-0562

Jandzik, D., Garnett, A. T., Square, T. A., Cattell, M. V., Yu, J. K., and Medeiros, D. M. (2015). Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue. *Nature* 518, 534–537. doi: 10.1038/nature14000

Jensen, M. A., Ferretti, V., Grossman, R. L., and Staudt, L. M. (2017). The NCI genomic data commons as an engine for precision medicine. *Blood* 130, 453–459. doi: 10.1182/blood-2017-03-735654

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Kaucka, M., and Adameyko, I. (2019). Evolution and development of the cartilaginous skull: From a lancelet towards a human face. *Semin. Cell Dev Biol.* 91, 2–12. doi: 10.1016/j.semcdb.2017.12.007

Kerosuo, L., and Bronner-Fraser, M. (2012). What is bad in cancer is good in the embryo: importance of EMT in neural crest development. *Semin Cell Dev. Biol.* 23, 320–332. doi: 10.1016/j.semcdb.2012.03.010

Khan, F. M., Marquardt, S., Gupta, S. K., Knoll, S., Schmitz, U., Spitschak, A., et al. (2017). Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nat. Commun.* 8:198.

Kirienko, N. V., Mani, K., and Fay, D. S. (2010). Cancer models in *Caenorhabditis elegans*. *Dev. Dyn.* 239, 1413–1448.

Kontou, P. I., Pavlopoulou, A., Dimou, N. L., Pavlopoulos, G. A., and Bagos, P. G. (2016). Network analysis of genes and their association with diseases. *Gene* 590, 68–78. doi: 10.1016/j.gene.2016.05.044

Koual, M., Cano-Sancho, G., Bats, A. S., Tomkiewicz, C., Kaddouch-Amar, Y., Douay-Hauser, N., et al. (2019). Associations between persistent organic pollutants and risk of breast cancer metastasis. *Environ. Int.* 132:105028. doi: 10.1016/j.envint.2019.105028

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kuratani, S., Kuraku, S., and Murakami, Y. (2002). Lamprey as an evo-devo model: lessons from comparative embryology and molecular phylogenetics. *Genesis* 34, 175–183. doi: 10.1002/gene.10142

Lair, S., Measures, L. N., and Martineau, D. (2016). Pathologic findings and trends in mortality in the Beluga (*Delphinapterus leucas*) Population of the St Lawrence Estuary, Quebec, Canada, From 1983 to 2012. *Vet. Pathol.* 53, 22–36. doi: 10.1177/0300985815604726

Langlet, C., and Bierne, J. (1982). Immune characteristics of graft rejection in nemerteans of the genus *Lineus*. *Eur. J. Immunol.* 12, 705–708. doi: 10.1002/eji.1830120902

Lauby-Secretan, B., Loomis, D., Grosse, Y., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., et al. (2013). Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. *Lancet Oncol.* 14, 287–288.

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493–D496.

Levine, M., and Davidson, E. H. (2005). Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4936–4942.

Lineweaver, C. H., Davies, P. C., and Vincent, M. D. (2014). Targeting cancer's weaknesses (not its strengths): therapeutic strategies suggested by the atavistic model. *Bioessays* 36, 827–835. doi: 10.1002/bies.201400070

Lipscomb, T. P., Scott, D. P., Garber, R. L., Krafft, A. E., Tsai, M. M., Lichy, J. H., et al. (2000). Common metastatic carcinoma of California sea lions (*Zalophus californianus*): evidence of genital origin and association with novel gammaherpesvirus. *Vet. Pathol.* 37, 609–617. doi: 10.1354/vp.37-6-609

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.

Logotheti, S., Marquardt, S., Richter, C., Sophie Hain, R., Murr, N., Takan, I., et al. (2020). Neural networks recapitulation by cancer cells promotes disease progression: a novel role of p73 isoforms in cancer-neuronal crosstalk. *Cancers (Basel)* 12:3789. doi: 10.3390/cancers12123789

Loomis, D., Guyton, K., Grosse, Y., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., et al. (2015). Carcinogenicity of lindane, DDT, and 2,4-dichlorophenoxyacetic acid. *Lancet Oncol.* 16, 891–892. doi: 10.1016/s1470-2045(15)00081-9

Maciel, E. I., and Oviedo, N. J. (2018). "Platyhelminthes: molecular dissection of the planarian innate immune system," in *Advances in Comparative Immunology*, ed. E. L. Cooper (Cham: Springer International Publishing), 95–115. doi: 10.1007/978-3-319-76768-0_4

Martik, M. L., and Bronner, M. E. (2017). Regulatory Logic Underlying Diversification of the Neural Crest. *Trends Genet.* 33, 715–727. doi: 10.1016/j.tig.2017.07.015

Martineau, D., Lemberger, K., Dallaire, A., Labelle, P., Lipscomb, T. P., Michel, P., et al. (2002). Cancer in wildlife, a case study: beluga from the St. Lawrence estuary, Québec, Canada. *Environ. Health Perspect.* 110, 285–292. doi: 10.1289/ehp.02110285

Merlo, L. M., Pepper, J. W., Reid, B. J., and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* 6, 924–935. doi: 10.1038/nrc2013

Metzger, M. J., Reinisch, C., Sherry, J., and Goff, S. P. (2015). Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell* 161, 255–263. doi: 10.1016/j.cell.2015.02.042

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745.

Ostrander, G. K., Cheng, K. C., Wolf, J. C., and Wolfe, M. J. (2004). Shark cartilage, cancer and the growing threat of pseudoscience. *Cancer Res.* 64, 8485–8491. doi: 10.1158/0008-5472.can-04-2260

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *Gigascience* 7, 1–31. doi: 10.20429/tag.2020.070202

Randhawa, N., Gulland, F., Ylitalo, G. M., DeLong, R., and Mazet, J. A. K. (2015). Sentinel California sea lions provide insight into legacy organochlorine exposure trends and their association with cancer and infectious disease. *One Health* 1, 37–43. doi: 10.1016/j.onehlt.2015.08.003

Richter, C., Marquardt, S., Li, F., Spitschak, A., Murr, N., Edelhäuser, B. A. H., et al. (2019). Rewiring E2F1 with classical NHEJ via APLF suppression promotes bladder cancer invasiveness. *J. Exp. Clin. Cancer Res.* 38:292.

Rodrigues, P., Patel, S. A., Harewood, L., Olan, I., Vojtasova, E., Syafruddin, S. E., et al. (2018). NF-κB-dependent lymphoid enhancer co-option promotes renal carcinoma metastasis. *Cancer Discov.* 8, 850–865. doi: 10.1158/2159-8290.cd-17-1211

Rousseaux, S., Debernardi, A., Jacquiau, B., Vitte, A. L., Vesin, A., Nagy-Mignotte, H., et al. (2013). Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* 5:186ra66. doi: 10.1126/scitranslmed.3005723

Sauka-Spengler, T., and Bronner-Fraser, M. (2008). A gene regulatory network orchestrates neural crest formation. *Nat. Rev. Mol. Cell Biol.* 9, 557–568. doi: 10.1038/nrm2428

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Res.* 47, D94–D99.

Schlumberger, H. G., and Lucke, B. H. (1948). Tumors of fishes, amphibians, and reptiles. *Cancer Res.* 8, 657–753.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shimeld, S. M., and Donoghue, P. C. (2012). Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). *Development* 139, 2091–2099. doi: 10.1242/dev.074716

Sievers, F., and Higgins, D. G. (2014a). Clustal omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* 1079, 105–116. doi: 10.1007/978-1-62703-646-7_6

Sievers, F., and Higgins, D. G. (2014b). Clustal omega. *Curr. Protoc. Bioinform.* 48, 3.13.1–6.

Somorjai, I. M., Somorjai, R. L., García-Fernàndez, J., and Escrivà, H. (2012). Vertebrate-like regeneration in the invertebrate chordate amphioxus. *Proc. Natl. Acad. Sci. U.S.A.* 109, 517–522. doi: 10.1073/pnas.1100045109

Stacy, B. A., and Parker, J. M. (2004). Amphibian oncology. *Vet. Clin. North Am. Exot. Anim. Pract.* 7, 673–695.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Tascedda, F., and Ottaviani, E. (2014). Tumors in invertebrates. *Inverteb. Surv. J.* 11, 197–203.

Trigos, A. S., Pearson, R. B., Papenfuss, A. T., and Goode, D. L. (2017). Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. *Proc. Natl. Acad. Sci. U.S.A.* 114, 6406–6411. doi: 10.1073/pnas.1617743114

Van Roten, A., Barakat, A. Z. A., Wouters, A., Tran, T. A., Mouton, S., Noben, J. P., et al. (2018). A carcinogenic trigger to study the function of tumor suppressor genes in *Schmidtea mediterranea*. *Dis. Model. Mech.* 11:dmm032573.

Wellings, S. R. (1969). Neoplasia and primitive vertebrate phylogeny: echinoderms, prevertebrates, and fishes–A review. *Natl. Cancer Inst. Monogr.* 31, 59–128.

Xiong, Y., Wei, Y., Gu, Y., Zhang, S., Lyu, J., Zhang, B., et al. (2017). DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res.* 45, D888–D895.

Ylitalo, G. M., Stein, J. E., Hom, T., Johnson, L. L., Tilbury, K. L., Hall, A. J., et al. (2005). The role of organochlorines in cancer-associated mortality in California sea lions (*Zalophus californianus*). *Mar. Pollut. Bull.* 50, 30–39. doi: 10.1016/j.marpolbul.2004.08.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PHYLOGENETIC ANALYSIS OF KERATINS: FUNCTIONAL IMPLICATIONS

Işil Takan[1,2], *Hani Alotaibi*[1,2], *Athanasia Pavlopoulou*[1,2]

1. Izmir Biomedicine and Genome Center, 35340 Balcova, Izmir, Turkey,
2. Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, 35340 Balcova, Izmir, Turkey

Keratins (KRTs) constitute the intermediate filament-forming proteins of epithelial cells. They are also grouped as "soft" and "hard" keratins, on the basis of their physicochemical properties. Hard keratins make up morphological structures such as scales, claws and feathers in birds and hair and nails in mammals. Soft bepithelial keratins are highly involved in epithelial cell protection from mechanical and non-mechanical stressors, and also play a regulatory role in apical–basal plasma membrane polarity, cell size, cell motility, protein synthesis, membrane trafficking and signaling pathways that regulate cell response to wound healing, cell growth and cell death [1, 2]. Because of these properties soft keratins are known to have a prominent 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 [3]. However, there is a lack of a comprehensive phylogenetic analysis of keratins which would enhance our understanding regarding the functional implications of keratins in cancer biology among tumor-bearing species. The human genome contains both functional KRT genes and nonfunctional KRT pseudogenes which are arranged in two uninterrupted clusters on chromosomes 12 and 17 [4]. This exceptional characteristic of KRTs makes them ideal for evolutionary studies aiming to infer the direction and timing of gene duplication events. In order to reconstruct the evolutionary history of the type I KRT gene family, we have performed comprehensive phylogenetic analyses of KRT homologous proteins in multiple genomes. Toward this end, the publicly available genomes, including those that have been completed recently, were searched extensively for keratin homologs. Phylogenetic trees based on the entire length keratin homologous proteins were reconstructed. Furthermore, the chromosomal arrangement of KRT genes was examined, and it was found that their position, transcriptional orientation, and number is preserved in the species under investigation.

Keywords: Keratins; Phylogeny; Gene Duplication

References: [1] Coulombe PA, Omary MB. 'Hard' and 'soft' principles defining the structure, function and regulation of keratin intermediate filaments. Curr Opin Cell Biol. 2002 Feb;14(1):110-22. PubMed PMID: 11792552.
[2] Moll R, Divo M, Langbein L. The human keratins: biology and pathology. Histochem Cell Biol. 2008 Jun;129(6):705-33. PubMed PMID: 18461349.
[3] Karantza V. Keratins in health and cancer: more than mere epithelial cell markers. Oncogene. 2011 Jan 13;30(2):127-38. PubMed PMID: 20890307.
[4] Hesse M, Zimek A, Weber K, Magin TM. Comprehensive analysis