ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps

Deepak Balamurali [ID][†]**, Alessandro Gorohovski** [ID][†]**, Rajesh Detroja, Vikrant Palande, Dorith Raviv-Shay and Milana Frenkel-Morgenstern**[*]

Laboratory of Cancer Genomics and Biocomputing of Complex Diseases, The Azrieli Faculty of Medicine, Bar-Ilan University, Safed 1311502, Israel

ABSTRACT

Chimeric RNA transcripts are formed when exons from two genes fuse together, often due to chromosomal translocations, transcriptional errors or transsplicing effect. While these chimeric RNAs produce functional proteins only in certain cases, they play a significant role in disease phenotyping and progression. ChiTaRS 5.0 (http://chitars.md.biu.ac.il/) is the latest and most comprehensive chimeric transcript repository, with 111 582 annotated entries from eight species, including 23 167 known human cancer breakpoints. The database includes unique information correlating chimeric breakpoints with 3D chromatin contact maps, generated from public datasets of chromosome conformation capture techniques (Hi–C). In this update, we have added curated information on druggable fusion targets matched with chimeric breakpoints, which are applicable to precision medicine in cancers. The introduction of a new section that lists chimeric RNAs in various cell-lines is another salient feature. Finally, using text-mining techniques, novel chimeras in Alzheimer's disease, schizophrenia, dyslexia and other diseases were collected in ChiTaRS. Thus, this improved version is an extensive catalogue of chimeras from multiple species. It extends our understanding of the evolution of chimeric transcripts in eukaryotes and contributes to the analysis of 3D genome conformational changes and the functional role of chimeras in the etiopathogenesis of cancers and other complex diseases.

INTRODUCTION

The transcriptome in eukaryotes is composed of single-stranded sequences of RNAs transcribed from various locations in the total genome. Although most RNA transcripts can be traced back to a single locus, exons from two different genes or from two copies of the same gene sometimes fuse together, leading to the formation of chimeric RNAs (1–22). The various causes of such fusions include transcriptional errors, trans-splicing effects and chromosomal translocations (6,14). Thus, two unrelated genomic loci on different chromosomes may produce a chimeric transcript through a genomic rearrangement event or due to trans-splicing. Similarly, a read-through transcript of two adjacent genomic loci may produce chimeric RNAs (8–9,21). The possibility is high that chimeras are artefacts of template switching by the reverse transcriptase enzyme (considering the difficulty of performing reverse transcriptase-free assays) (9,16). Nonetheless, several studies have identified chimeric transcripts that have also been translated into proteins. This establishes their authenticity and further motivates scientists to curate a list of known chimeras (19–26). A direct correlation has been suggested between the presence of chimeras in the genome and their role in tumorigenesis (18,19). The transcriptome becomes extremely more complex in the context of cancer, due to a high proportion of genomic rearrangements, nucleotide polymorphisms and alterations of the splicing machinery.

One of the most renown gene fusions identified in solid tumors is the TMPRSS2-ERG chimera, a well-documented biomarker of prostate cancer. This chimera has been identified in a high frequency of tested patient samples (28). Novel means, like the delivery of specific siRNAs by targeted liposomal nanovectors and the use of peptidomimetic inhibitors have been attempted, with the aim of mitigating the spread of prostate cancer (29,30). Similarly, FGFR3-TACC3 fusions have been identified as driving factors for cancer progression in multiple tissue types like bladder, lung and brain.

[*]To whom correspondence should be addressed. Tel: +972 72 264 2901; Fax: +972 72 264 2901; Email: milana.morgenstern@biu.ac.il
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

These fusions are currently being targeted using the drug *erdafitinib* (31). Such cases highlight the need to consolidate all known chimeras related to cancer and other disorders, together with information about their breakpoints. This will enable their use as diagnostic tools and as potential drug targets. Gene fusion associations in both solid and liquid tumour development, as well as in other genetic disorders, have been confirmed by the detection of chimeras in cell-lines, using short read sequencing strategies (20). Chimeric transcripts have been shown to be significantly more tissue specific by nature than non-chimeric transcripts (16). Additionally, chimeric RNAs appear not to be a distinct feature of tumour physiology. Recurrent gene fusions have been identified in non-cancerous tissues, as well as in normal cells (27). Chimeras have also been found to lose some of their functional domains, and therefore actively compete with their wild-type genes. This leads to a dominant-negative effect in cancers and other diseases (17). Recent studies have intercepted contact maps of chromosome conformation capture techniques (Hi–C) with known chromosomal translocations. A marked increase has been identified in the correlation between certain tissue types and the Hi–C contact frequencies (32).

As next generation sequencing (NGS) has become the norm for genomic studies in recent years, the identification of chimeric transcripts has increased at an exponential rate (33). Several efforts have been made to catalogue the chimeric transcripts from literature, using a variety of computational methods and text-mining techniques (34). These have resulted in the creation of databases such as the Mitelman database, ChimerDB 3.0, COSMIC, dbCRID, HybridDB, TICdb and FusionGDB. (23,33,35–40). These databases have actively supported research in their respective domains. However, ChiTaRS stands out as a unique server that integrates EST and mRNA sequences, and literature resources, with RNA-sequencing data, expression level and tissue specificity of chimeric transcripts in various tissues and organisms.

The evolution of ChiTaRS

The ChiTaRS database was originally designed to incorporate chimeric transcripts from three organisms (human, mouse and fruit fly) (41). Subsequently, from ChiTaRS-2.1, the database was extended to eight organisms. This increased the evidence of chimeric transcript conservation across these species (42,43). The database was created using sustained bioinformatics analyses of the expressed sequence tags (ESTs) and mRNA sequences from GenBank (44). We also extended our resource of published chimeras, with RNA expression level as verified by RNA sequencing, and whose translated proteins have been shown by means of shotgun and targeted mass-spectrometry analyses (16,42). Interactive links to the protein-protein interaction maps of all the chimeric proteins have also been incorporated into the ChiTaRS database, with the help of the ChiPPI tool, introduced in ChiTaRS-3.1 (43,45). Moreover, ChiTaRS enables visualization of the transcripts and their junction sites by SpliceGrapher, using the genome annotation for the respective organisms from the UCSC genome browser (46). Additionally, an exclusive feature of the ChiTaRS database

is the assessment of junction consistency. This enables ranking chimeras according to the evidence of consistent junction sites. This feature, introduced in ChiTaRS-2.1, has proven beneficial to researchers seeking experimental confirmation of highly ranked chimeric transcripts (42). The ChiTaRS-4.1 version of the database, which was released online in 2018, announced the addition of Hi–C contact maps in nearly 5600 chimeras. This marked an important adjunct to the database. The most recent release, ChiTaRS 5.0, consists of a novel section on druggable fusions, mined from various sources of literature. The database also introduces a distinct section for chimeras identified in 531 human cell-line systems. Moreover, using our ProtFus method (34), we added novel chimeras from neurodegenerative and autoimmune diseases like Alzheimer's disease, schizophrenia and rheumatoid arthritis, thereby empowering research in these other fields. Such features, coupled with many more salient existing database functionalities, make ChiTaRS one of the most up-to-date and sought-after resources for chimeric transcripts.

Improvements in ChiTaRS

The significant updates and improvements to the ChiTaRS database since the last published edition are tabulated in Table 1. The major improvements include: the introduction of a segment of druggable fusion targets matched with chimeric breakpoints, the addition of 3D chromatin contact maps from publicly available Hi–C datasets in four organisms, the creation of a discrete section of chimeric transcripts identified in various human and mice cell-lines, and extension of the annotation of detected chimeras to UniProt, NCBI, GeneCards, PubMed and other public resources.

Database content enhancement

For the current 2019 release of the database, 111 582 chimeric RNAs were collated from 1 393 046 EST and mRNA sequences of eight species (namely, *Homo sapiens, Mus musculus, Drosophila melanogaster, Rattus norvegicus, Bos Taurus, Saccharomyces cerevisiae, Sus scrofa* and *Danio rerio*) retrieved from NCBI/GenBank (44). All chimeric transcripts were aligned to their respective reference genome assemblies using the UCSC BLAT program (47). The details of the reference genomes used for each species are provided in Supplementary Table S1. We included a chimera in the database if its first and second sequence tracts had a minimum identity of 95% and a minimum length of 50nt; and if these two sequences could not be mapped linearly to the reference genome. The number of chimeras detected in each organism is presented in Table 1. Moreover, 23 167 unique cancer breakpoints were collected from the recent publications by Mertens *et al.* and from the Mitelman collection, together comprising ∼11 000 more breakpoints than previously collated (33,35). Thus, we verified, manually, 11 650 breakpoints to confirm their accuracy, with the help of >7700 PubMed articles. Notably, haematological tumour malignancies correlated with a high frequency of reported fusions, particularly, acute lymphoblastic leukaemia (ALL, 710 chimeras), acute myeloid

Table 1. Improvements in ChiTaRS 5.0: significant enhancement over previous database editions

Content	ChiTaRS	ChiTaRS 2.1	ChiTaRS 3.1	ChiTaRS 5.0	Relevance
The collection of chimeric transcripts	16282 (total), 9379 (*Homo sapiens*), 4828 (*Mus musculus*), 2055 (*D. melanogaster*)	29164 (total), 20 753 (*Homo sapiens*), 6226 (*M. musculus*), 2151 (*D. melanogaster*), 4 (*Bos taurus*), 8 (*Rattus norvegicus*), 4 (*Danio rerio*), 5 (*S. cerevisiae*), 13 (*Sus scrofa*)	34 922 (total), 24 608 (*Homo sapiens*), 7457 (*M. musculus*), 2740 (*D. melanogaster*), 6 (*Bos taurus*), 10 (*Rattus norvegicus*), 7 (*Danio rerio*), 5 (*S. cerevisiae*), 89 (*Sus scrofa*)	111 582 (total), 66 243 (*Homo sapiens*), 41 584 (*M. musculus*), 3052 (*D. melanogaster*), 19 (*Bos taurus*), 20 (*Rattus norvegicus*), 67 (*Danio rerio*), 305 (*S. cerevisiae*), 292 (*Sus scrofa*)	Extension of the collection for all eight organisms by ~76 600 new entries
Disease Breakpoints	1280	1428	11 714	23 167	Extension by >11 000 new entries to include unique fusion transcripts with breakpoint information associated with cancer and other complex diseases
Manually verified breakpoints	456	1428	10 285	11 650	Additional breakpoints verified manually
Druggable fusions	No	No	No	680	A novel feature providing information on the use of the chimeric breakpoint as a drug target
Hi–C (3D Chromatin Contact Map) chimeras	No	No	No	5597	The addition of ~5600 chimeras, now matched with 3D chromatin contact maps from Hi–C data
Chimeric Protein-Protein interaction (ChiPPI) networks	No	No	2081 (validated), 22527 (predicted)	9973 (validated), 42 058 (predicted)	Additional protein-protein network analysis maps for ~27 000 chimeras
RNA-seq verified chimeras	175	337	435	937	Extension of the number of chimeras verified in RNA-seq experiments by ~500
Sense-antisense chimeras	No	6044	6485	7521	The incorporation of ~1000 additional sense-antisense chimeras
Chimeras identified in cell-lines	No	No	No	2411	The identification of ~2400 chimeras across 531 human cell lines
Links to external data repositories (GeneCards, OMIM, Pubmed, Ensembl, SNOMED CT & RefSeq)	No	No	33 124	136 458	The addition of ~100 000 links to multiple data repositories for ease of correlation

leukaemia (AML, 639 chimeras) and acute myelomonocytic leukaemia (AMMoL, 599 chimeras) topped the list. Further to the focus on chimeras in cancer, we added novel chimeras that have been linked to the progression of other neurodegenerative and autoimmune disorders, such as dyslexia, schizophrenia, rheumatoid arthritis and systemic lupus erythematosus. To facilitate the understanding of the biological significance of each chimera, we updated the GenBank, RefSeq and Mitelman cross-references for all the genes (total of 202 460 links) (44,48). UniProt, Ensembl, FlyBase (for *D. melanogaster*), SGD (for *S. cerevisiae*), GeneCard and OMIM links (~256 000 links) for the parental genes are also supported in the current version of the ChiTaRS database. This yields a valuable resource for conducting in-depth analyses of chimeric transcripts (49–54). Detailed information regarding available resources for each organism is provided in Supplementary Table S2. We

have also added a new section specifically for 2411 gene fusions from 531 cell lines of human and mouse origin. The cell lines with the highest frequency of chimeric transcripts expressed include HCC1263 (human colorectal adenocarcinoma), VCaP (prostate cancer) and MDA-MB-361 (breast cancer derived from metastatic site – brain), containing 25, 22 and 21 chimeras, respectively.

Interface upgrades in ChiTaRS 5.0

The user interface of ChiTaRS 5.0 has also been upgraded to provide the end-user with a seamless user experience, combining the latest database functionalities with the database existing treasure trove of options. We have added a check-box option in the 'Full Collection and Search' tab for the exclusive search of chimeras with Hi–C contacts (Figure 1). In ChiTaRS-3.1, we introduced protein-protein network

Figure 1. The 'Full Collection & Search' tab on the updated ChiTaRS 5.0 database (http://chitars.md.biu.ac.il/bin/search.pl). The new interface provides options to perform custom searches for chimeras with matched Hi–C contact maps, RNAseq evidences, etc. A sample search is shown below, with each chimera result containing links to external data repositories like NCBI, UniProt, SpliceGrapher, PubMed, GeneCards and ChiPPI. This provides additional information at the user's disposal.

analysis charts for each chimeric protein from chimeric Protein-Protein interactions (ChiPPI) (45). This enabled the end-user to find 'missing' and 'known' interactors for the fusion protein, based on the preserved domains from the parental genes (45). We have now added additional pop-up windows for these ChiPPI networks, and extended annotations of genes and links to PubMed references. A separate tab now facilitates searching for druggable fusions and breakpoints (Supplementary Figure S1). A tab exclusively for chimeras identified in cell lines has also been introduced (Supplementary Figure S2). These features simplify and accelerate analyses, with reduced turnaround time for the end-users. A new user submission form offers users the option to submit chimeric transcripts they have identified. These will be added to the database after due verification. The process for submission is explained in Supplementary Figure S3.

Druggable fusions as tailored therapy targets

Over the years, several papers have documented the role of gene fusions as drivers of cancer and tumour progression. Even rare chimeric transcripts, occurring in low frequencies, can be potential functional drivers (55). In certain cancer types, some fusions recur frequently, generally during tumour initiation (56). Prior to the NGS revolution, identifying gene fusions was extremely difficult. However, currently available NGS technology offers great promise in the field of precision medicine, particularly for using chimeric proteins as druggable fusion targets. Gene fusions can be highly specified. The best examples in support of this theory are the BCR-ABL1 fusion in chronic myeloid leukaemia (CML) and the EML4-ALK gene fusion in non-small cell lung carcinoma (NSCLC). The BCR-ABL1 chimera is used as a target for tyrosine kinase inhibitors like *imatinib* and *nilotinib* (57). Drugs like *crizotinib* and *alectinib*. are used to treat NSCLC by targeting the EML4-ALK fusion event (58). The ALK inhibitor *crizotinib* is approved for treatment of patients with NSCLC in Europe, Japan, the United States and other countries. On the other hand, *alectinib* is a more selective oral ALK inhibitor, and several research groups have demonstrated its effectiveness in systemic disease and in the central nervous system (59). In this version of ChiTaRS, we present a new manually-verified collection of druggable chimeras for various malignancies, linked with their breakpoint information, Medical Subject Headings (MeSH), the Diseases Database (DiseasesDB) and PubMed links to the published manuscripts. We identified 37 drugs/drug targets, including *imatinib*, *crizotinib* and *sorafenib*. These chimeric protein-based therapeutics were curated by us from systematic data mining of published literature and open source repositories like DEPO (56,60). A list of commonly used drugs that target fusions is provided in Table 2. We have provided as separate files the FASTA sequences, the chimeric junction sequence and the drugs that are currently used to target these druggable chimeras. These can be accessed from the '*Downloads*' tab on the ChiTaRS website (http://chitars.md.biu.ac.il/downloads.html). Users may consider this as a **'gold-standard'** druggable fusion dataset to standardize their methods for fusion prediction and testing. The complete list of such drugs can be retrieved from the ChiTaRS database online. We have pro-

vided a searchable list of drugs/chimeras and other features, together with the relevant scientific information (Supplementary Figure S1).

Hi–C contact map correlations with chimeric junctions

The chromosomal translocations frequently occurring in cancer and other disorders often contribute substantially to the progression of these disorder as expressed drivers. Translocations from multiple tissue types have been shown to display high rates of Hi–C contact frequencies. This shapes the 3D genome architecture of the species and can be used in detecting/controlling these abnormalities (32,61–62). Thus, the increased frequency of these specific translocations, together with the existence of rearrangement hotspots, attest to the role of certain internal cellular mechanisms that promulgate specific regions to crosslink. The development of Hi–C, an NGS-based protocol enables describing contact probabilities across the human genome (63). Originally designed to detect known and novel gene fusions in cell-lines, Hi–C has transformed into a robust protocol that can be applied to tissue samples from humans and other organisms, and even detect copy-number variations (64). The ChiTaRS 5.0 database includes Hi–C chromatin contact maps from public datasets for four organisms, namely*, human, mouse, fruit fly* and *yeast*. We have included nearly 5600 chimeras, spanning across 14 cell lines and tissue samples. In humans, we included chimeras with 3D chromatin contacts in K562, HMEC, imr90, gm12878, NHEK, KBM-7, HUVEC and HeLa celllines. We also added Hi–C contacts for chimeras detected in Patski, EKLF$^{-/-}$, F123 and CH12-LX cell-lines from mice. Detailed information about the public datasets used can be retrieved from the '*Methods*' tab of the ChiTaRS website. To date, the mechanisms by which chromatin changes are produced in diseased conditions and their epigenetics are poorly understood. Hi–C contributes to this knowledge as it is a cost- effective method, with high efficiency of identifying these translocations, even when signals are low (64). This is because the contacts generally tend to be on either side of the breakpoint and, therefore, fall into the *cis* and not the *trans* region (65) Such chimeric transcripts, with matched 3D chromatin maps, may also serve as suitable diagnostic / prognostic biomarkers of disease progression (66). During our process of curating chimeras with Hi–C contacts, we noticed a marked correlation in the ratio of translocations to sense-antisense chimeras that we had described previously (42), in relation to 3D chromatin contact maps. Therefore, we propose that changes in chromatin topology influences the formation of sense-antisense chimeras. We also noticed the presence of such chimeras with the conserved Hi–C contacts in certain genes like CDH2, APP and TTN. These are conserved across species, among humans, mice and fruit flies.

We examined the presence of chimeras in evolutionarily conserved genes across three species (*H. sapiens, M. musculus* and *D. melanogaster*) and identified 70 chimeras, of which 29 were overrepresented for three genes only, namely APP, CTBP1 and EEF2 (human) (Figure 2). The 29 chimeras from these three genes are listed in Table 3. The complete list of 70 chimeras is provided in Supplementary
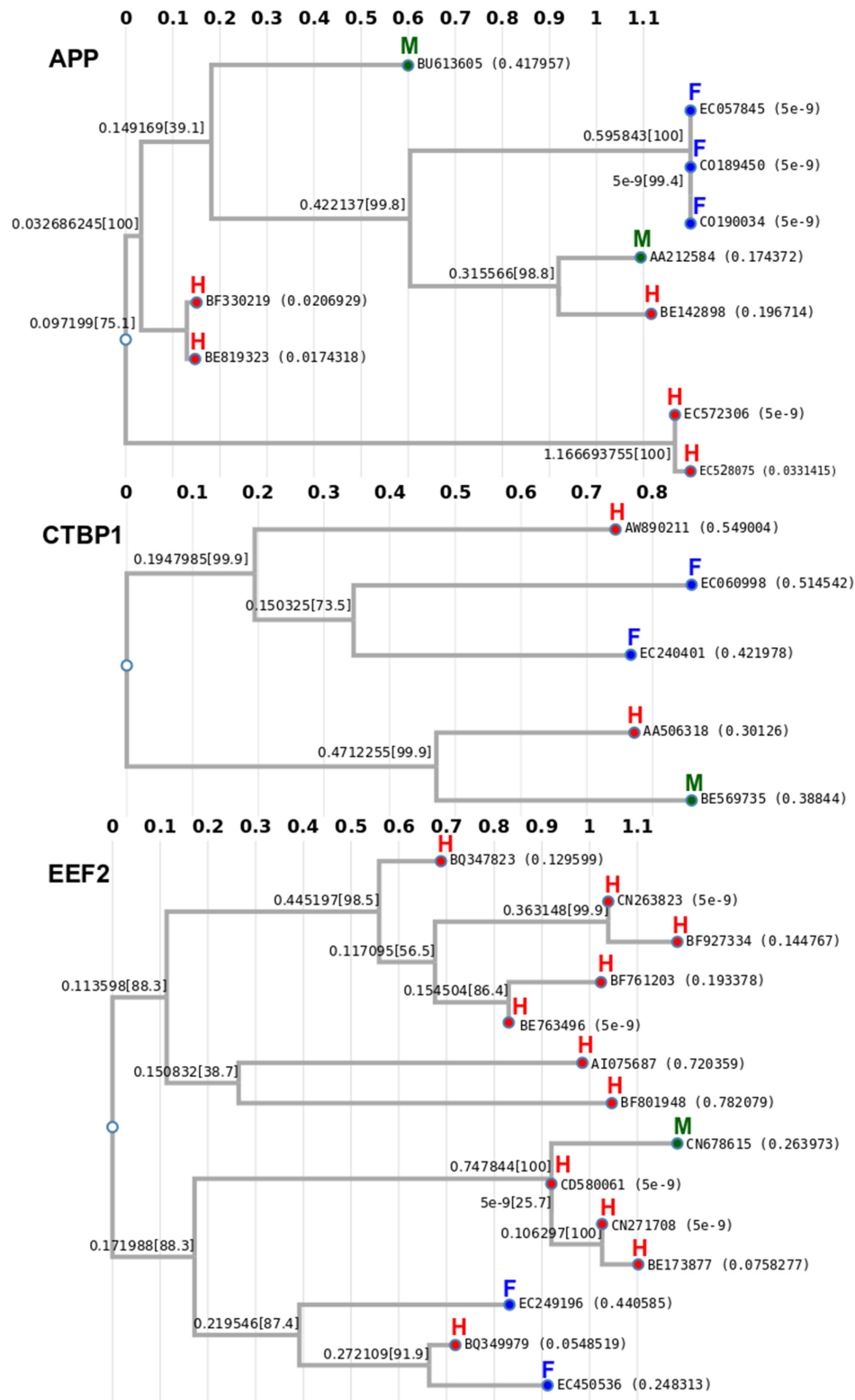
Figure 2. A phylogenetic tree representing the evolution between human (H), mouse (M) and fruit-fly (F) chimeras (represented by NCBI id), with matched Hi–C contacts in APP, CTBP1 and EEF2 genes. We found evolutionary conservation in the Hi–C contacts between orthologs for different organisms and paralogs for the same organisms.

Table 2. List of chimeric RNAs frequently being used as drug targets

S.No.	Drug	Gene fusion	References (PubMed ID)
1	Alectinib	EML4-ALK	25228534, 28890946
2	ATRA + ATO	PML-RARA	15103387
3	AZD4547	FGFR1-TACC1	28890946
4	Crizotinib	RANBP2-ALK	20979472, 28890946
		EML4-ALK	
		PTPRZ1-MET	27748748, 28890946
5	Dasatinib	BCR-ABL1	19779040, 21562040
		RCSD1-ABL1	21863287, 21562040
6	Entrectinib	TPM3-NTRK1	24705251, 28890946
		ETV6-NTRK3	
		VCL-NTRK2	
		AGBL4-NTRK2	
7	Erdafinitib	FGFR3-TACC3	28890946
8	Foretinib	PTPRZ1-MET	27748748
		CLIP2-MET	
		TFG-MET	
9	Imatinib	FIP1L1-PDGFRA	12660384, 16089297
		BCR-PDGFRA	15034867, 11423618
		EML1-ABL1	15713800, 28890946
		NUP214-ABL1	16213363, 28890946
		COL1A1-PDGFB	20194851, 28890946
		BCR-ABL1	28890946, 11423618
10	Lapatinib	TENM4-NRG1	24727320
11	Larotrectinib	TPM3-NTRK1	28890946, 29606586
12	Nilotinib	BCR-ABL1	21562040, 28890946
13	Ponatinib	FGFR1-TACC1	29617662
14	Sorafenib	SLC45A3-BRAF	20526349, 24727320
15	Trastuzumab	ERBB2-CLDN7	28890946

Table S3. Amyloid precursor protein (APP) gene codes for a protein of the same name and is found in many tissues and organs of the body, especially the brain and the spinal cord. While the exact function of the APP gene is yet to be identified, recent studies have shown its involvement in the migration of nerve cells in the brain during early development. Mutation or dysregulation of the APP gene has also been shown in a high proportion of individuals with Alzheimer's disease, and in hereditary cerebral amyloid angiopathy (67). C-Terminal Binding Protein 1 (CTBP1) is a phosphoprotein and a known transcriptional repressor, and dimerizes with CTBP2, its closely related gene. Alterations in CTBP1 are known to cause developmental delay, Wolf-Hirchhorn syndrome and other conditions (68). Eukaryotic translation elongation factor (EEF2) is an essential component of the protein synthesis pathway and is involved in GTP-dependent translocation of nascent protein chains (69). EEF2 has been found to be dysregulated in tracheal carcinoma, spinocerebellar ataxia and other disorders. We identified 10 chimeras across the three species in the conserved APP gene (App in the mouse and AppI in the fruit fly), five chimeras in the CTBP1 gene (Ctbp2 in the mouse, CtBP in the fruit fly) and 14 chimeras in the EEF2 gene (mouse orthologue Eef2, fruit fly orthologue EF2). We analysed the *FASTA* sequence to identify the presence of any functional domains that have been evolutionarily conserved. Data of the multiple sequence alignment and phylogenetic separation are presented in Figure 2.

CONCLUSIONS AND PERSPECTIVES

The comprehensive ChiTaRS 5.0 database will serve as a critical tool dedicated to the study of chimeric transcripts in eukaryotes at 'multi'-omics levels, spanning genomic, transcriptomic and proteomic domains. Our resources such as ChiTaRS, ProtFus and ChiPPI (34,41–43,45) are repeatedly referenced as valuable resources for fusion discovery and annotation (70). The updated version 5.0 of the ChiTaRS database provides a phenomenal increase in annotated and verified chimeric transcripts, compared to the previous ChiTaRS releases, and includes a significant extension of specific research-oriented features like druggable fusions and 3D chromatin contact maps. ChiTaRS 5.0 provides extensive experimental evidence for chimeras specific to cancers, and auto-immune (rheumatoid arthritis, systemic lupus erythematosus) and neurodegenerative disorders (such as Alzheimer's and Parkinson's diseases). These can be effectively applied in the planning of new experiments or for the analysis of large-scale sequencing experiments. Ongoing international consortia like ICGC, PanCancer, 100 000 Genome Project, ADNI, NIAGADS and gnomAD will benefit from this database and from all incremental additions to it, for improving the process of chimera identification and validation. To conclude, the updated ChiTaRS database is designed to advance the field of cancer research as well as evolutionary biology, and ultimately contribute to the therapeutic stratification of diseases using chimeric transcripts, thus validated as potential biomarkers.

DATA AVAILABILITY

The ChiTaRS 5.0 content will be continuously maintained and updated every six months. The database is now publicly accessible at http://chitars.md.biu.ac.il. The previous versions 3.1 and 2.1 are accessible at http://biodb.md.biu.

Table 3. Evolutionarily conserved chimeras with matched 3D chromatin contact maps in APP, CTBP1 and EEF2 genes

Chimera_ID	Gene_ID	Gene	Chr	Transcript	Version	Type_junction
CO189450	31002	Appl	X	NM_001258523.2	dm6	exon–exon
CO190034	31002	Appl	X	NM_001258523.2	dm6	exon–exon
EC057845	31002	Appl	X	NM_001258523.2	dm6	exon–exon
EC060998	41602	CtBP	3R	NM_169491.2	dm6	exon–exon
EC240401	41602	CtBP	3R	NM_001275593.1	dm6	exon–exon
EC249196	35422	EF2	2L	NM_080366.3	dm6	exon–exon
AA212584	11820	App	16	NM_001198823.1	mm10	exon–exon
BU613605	11820	App	16	NM_001198823.1	mm10	exon–exon
BE569735	13017	Ctbp2	7	XM_006507308.3	mm10	exon–intron
CN678615	13629	Eef2	10	NM_007907.2	mm10	exon–exon
BE142898	351	APP	21	NM_000484.3	hg38	exon–exon
BE819323	351	APP	21	NM_000484.3	hg38	exon–intron
BF330219	351	APP	21	NM_000484.3	hg38	exon–intron
EC528075	351	APP	21	NM_000484.3	hg38	exon–exon
EC572306	351	APP	21	NM_000484.3	hg38	exon–exon
AA506318	1487	CTBP1	4	XM_005272263.5	hg38	exon–exon
AW890211	1487	CTBP1	4	XM_005272263.5	hg38	exon–exon
AI075687	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BE173877	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BE763496	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BF761203	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BF801948	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BF927334	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BQ347823	1938	EEF2	19	NM_001961.3	hg38	exon–exon
BQ349979	1938	EEF2	19	NM_001961.3	hg38	exon–exon
CD580061	1938	EEF2	19	NA	hg19	intron-intron
CN263823	1938	EEF2	19	NM_001961.3	hg38	exon–exon
CN271708	1938	EEF2	19	NM_001961.3	hg38	exon–exon
EC450536	1938	EEF2	19	NM_001961.3	hg38	exon–exon

ac.il/chitars.prv/ and http://chitars.bioinfo.cnio.es/, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
2. Guigó,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B., Birney,E. *et al.* (2006) EGASP: the human ENCODE Genome annotation assessment project. *Genome Biol.*, **7**(Suppl. 1), S2.
3. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
4. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E., Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
5. Griffin,T.J., Gygi,S.P., Ideker,T., Rist,B., Eng,J., Hood,L. and Aebersold,R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol. Cell Proteomics*, **1**, 323–333.
6. Finta,C. and Zaphiropoulos,P.G. (2002) Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.*, **277**, 5882–5890.
7. Romani,A., Guerra,E., Trerotola,M. and Alberti,S. (2003) Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res.*, **31**, e17.
8. Akiva,P., Toporik,A., Edelheit,S., Peretz,Y., Diber,A., Shemesh,R., Novik,A. and Sorek,R. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.
9. Parra,G., Reymond,A., Dabbouseh,N., Dermitzakis,E.T., Castelo,R., Thomson,T.M., Antonarakis,S.E. and Guigó,R. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.
10. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

11. Di Segni,G., Gastaldi,S. and Tocchini-Valentini,G.P. (2008) Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 6864–6869.

12. Ortiz de Mendíbil,I., Vizmanos,J.L. and Novo,F.J. (2009) Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PLoS One*, **4**, e4805.

13. Gingeras,T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature*, **461**, 206–211.

14. Li,H., Wang,J., Ma,X. and Sklar,J. (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle*, **8**, 218–222.

15. Edgren,H., Murumagi,A., Kangaspeska,S., Nicorici,D., Hongisto,V., Kleivi,K., Rye,I.H., Nyberg,S., Wolf,M., Borresen-Dale,A.L. *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.*, **12**, R6.

16. Frenkel-Morgenstern,M., Lacroix,V., Ezkurdia,I., Levin,Y., Gabashvili,A., Prilusky,J., Del Pozo,A., Tress,M., Johnson,R., Guigo,R. *et al.* (2012) Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.*, **22**, 1231–1242.

17. Frenkel-Morgenstern,M. and Valencia,A. (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, **28**, i67–i74.

18. Asmann,Y.W., Necela,B.M., Kalari,K.R., Hossain,A., Baker,T.R., Carr,J.M., Davis,C., Getz,J.E., Hostetter,G., Li,X. *et al.* (2012) Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res.*, **72**, 1921–1928.

19. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrette,T., Palanisamy,N. and Chinnaiyan,A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

20. Djebali,S., Lagarde,J., Kapranov,P., Lacroix,V., Borel,C., Mudge,J.M., Howald,C., Foissac,S., Ucla,C., Chrast,J. *et al.* (2012) Evidence for transcript networks composed of chimeric rnas in human cells. *PLoS One*, **7**, e28213.

21. Denoeud,F., Kapranov,P., Ucla,C., Frankish,A., Castelo,R., Drenkow,J., Lagarde,J., Alioto,T., Manzano,C., Chrast,J. *et al.* (2007) Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.*, **17**, 746–759.

22. Prakash,A., Tomazela,D.M., Frewen,B., MacLean,B., Merrihew,G., Peterman,S. and MacCoss,M.J. (2009) Expediting the development of targeted SRM assays: Using data from shotgun proteomics to automate method development. *J. Proteome Res.*, **8**, 2733–2739.

23. Lee,M., Lee,K., Yu,N., Jang,I., Choi,I., Kim,P., EunJang,Y., Kim,B., Kim,S., Lee,B. *et al.* (2017) ChimerDB 3.0: An enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res.*, **45**, D784–D789.

24. McManus,C.J., Duff,M.O., Eipper-Mains,J. and Graveley,B.R. (2010) Global analysis of trans-splicing in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12975–12979.

25. Yu,C.Y., Liu,H.J., Hung,L.Y., Kuo,H.C. and Chuang,T.J. (2014) Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res.*, **42**, 9410–9423.

26. Wu,C.S., Yu,C.Y., Chuang,C.Y., Hsiao,M., Kao,C.F., Kuo,H.C. and Chuang,T.J. (2014) Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res.*, **24**, 25–36.

27. Babiceanu,M., Qin,F., Xie,Z., Jia,Y., Lopez,K., Janus,N., Facemire,L., Kumar,S., Pang,Y., Qi,Y. *et al.* (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res.*, **44**, 2859–2872.

28. Tomlins,S.A., Laxman,B., Varambally,S., Cao,X., Yu,J., Helgeson,B.E., Cao,Q., Prensner,J.R., Rubin,M.A., Shah,R.B. *et al.* (2008) Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia*, **10**, 177–188.

29. Shao,L., Tekedereli,I., Wang,J., Yuca,E., Tsang,S., Sood,A., Lopez-Berestein,G., Ozpolat,B. and Ittmann,M. (2012) Highly specific targeting of the TMPRSS2/ERG fusion gene using liposomal nanovectors. *Clin. Cancer Res.*, **18**, 6648–6657.

30. Wang,X., Qiao,Y., Asangani,I.A., Ateeq,B., Poliakov,A., Cieślik,M., Pitchiaya,S., Chakravarthi,B.V.S.K., Cao,X., Jing,X. *et al.* (2017) Development of peptidomimetic inhibitors of the ERG gene fusion product in prostate cancer. *Cancer Cell*, **31**, 532–548.

31. Costa,R., Carneiro,B.A., Taxter,T., Tavora,F.A., Kalyan,A., Pai,S.A., Chae,Y.K. and Giles,F.J. (2016) FGFR3-TACC3 fusion in solid tumors: Mini review. *Oncotarget*, **7**, 55924–55938.

32. McCord,R.P. and Balajee,A. (2018) 3D genome organization influences the chromosome translocation pattern. *Adv. Exp. Med. Biol.*, **1044**, 113–133.

33. Mertens,F., Johansson,B., Fioretos,T. and Mitelman,F. (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.

34. Tagore,S, Gorohovski,A., Jensen,L.J. and Frenkel-Morgenstern,M. (2019) ProtFus: a comprehensive method characterizing protein-protein interactions of fusion proteins. *PLOS Comput. Biol.*, **15**, e1007239.

35. Mertens,F., Antonescu,C.R. and Mitelman,F. (2016) Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes. *Genes Chromosom. Cancer*, **55**, 291–310.

36. Novo,F.J., de Mendíbil,I.O. and Vizmanos,J.L. (2007) TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, **8**, 33.

37. Kim,D.S., Huh,J.W. and Kim,H.S. (2007) HYBRIDdb: a database of hybrid genes in the human genome. *BMC Genomics*, **8**, 128.

38. Kong,F., Zhu,J., Wu,J., Peng,J., Wang,Y., Wang,Q., Fu,S., Yuan,L.L. and Li,T. (2011) DbCRID: a database of chromosomal rearrangements in human diseases. *Nucleic Acids Res.*, **39**, D895–D900.

39. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.

40. Kim,P. and Zhou,X. (2019) FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Res.*, **47**, D994–D1004.

41. Frenkel-Morgenstern,M., Gorohovski,A., Lacroix,V., Rogers,M., Ibanez,K., Boullosa,C., Leon,E.A., Ben-Hur,A. and Valencia,A. (2013) ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.*, **41**, D142–D151.

42. Frenkel-Morgenstern,M., Gorohovski,A., Vucenovic,D., Maestre,L. and Valencia,A. (2015) ChiTaRS 2.1-an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts. *Nucleic Acids Res.*, **43**, D68–D75.

43. Gorohovski,A., Tagore,S., Palande,V., Malka,A., Raviv-Shay,D. and Frenkel-Morgenstern,M. (2017) ChiTaRS-3.1-the enhanced chimeric transcripts and RNA-seq database matched with protein-protein interactions. *Nucleic Acids Res.*, **45**, D790–D795.

44. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.

45. Frenkel-Morgenstern,M., Gorohovski,A., Tagore,S., Sekar,V., Vazquez,M. and Valencia,A. (2017) ChiPPI: a novel method for mapping chimeric protein-protein interactions uncovers selection principles of protein fusion events in cancer. *Nucleic Acids Res.*, **45**, 7094–7105.

46. Rogers,M.F., Thomas,J., Reddy,A.S.N. and Ben-Hur,A. (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.*, **13**, R4.

47. James Kent,W., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

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

49. Bateman,A., Martin,M.J., O'Donovan,C., Magrane,M., Alpi,E., Antunes,R., Bely,B., Bingley,M., Bonilla,C., Britto,R. *et al.* (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

50. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

51. Thurmond,J., Goodman,J.L., Strelets,V.B., Attrill,H., Gramates,L.S., Marygold,S.J., Matthews,B.B., Millburn,G., Antonazzo,G., Trovisco,V. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.

52. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.

53. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Iny Stein,T., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.*, **54**, 1.30.1–1.30.33.

54. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.

55. Latysheva,N.S. and Babu,M.M. (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.*, **44**, 4487–4503.

56. Gao,Q., Liang,W.W., Foltz,S.M., Mutharasu,G., Jayasinghe,R.G., Cao,S., Liao,W.W., Reynolds,S.M., Wyczalkowski,M.A., Yao,L. *et al.* (2018) Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.*, **23**, 227–238.

57. Baccarani,M., Rosti,G. and Soverini,S. (2019) Chronic myeloid leukemia: the concepts of resistance and persistence and the relationship with the BCR-ABL1 transcript type. *Leukemia*, **33**, 2358–2364.

58. Sasaki,T., Rodig,S.J., Chirieac,L.R. and Jänne,P.A. (2010) The biology and treatment of EML4-ALK non-small cell lung cancer. *Eur. J. Cancer*, **46**, 1773–1780.

59. Vavalà,T. and Novello,S. (2018) Alectinib in the treatment of ALK-positive non-small cell lung cancer: an update on its properties, efficacy, safety and place in therapy. *Ther. Adv. Med. Oncol.*, doi:10.1177/1758835918789364.

60. Sun,S.Q., Mashl,R.J., Sengupta,S., Scott,A.D., Wang,W., Batra,P., Wang,L.B., Wyczalkowski,M.A. and Ding,L. (2018) Database of evidence for precision oncology portal. *Bioinformatics*, **34**, 4315–4317.

61. Kim,K., Eom,J. and Jung,I. (2019) Characterization of structural variations in the context of 3D chromatin structure. *Mol. Cells.*, **42**, 512–522.

62. Di Stefano,M., Di Giovanni,F., Baù,D., Carey,L.B., Marti-Renom,M.A. and Mendoza,M. (2018) Impact of chromosome fusions on 3D genome organization and gene expression in budding yeast. bioRxiv doi: https://doi.org/10.1101/237263, 09 November 2018, preprint: not peerreviewed.

63. Belaghzal,H., Dekker,J. and Gibcus,J.H. (2017) Hi–C 2.0: An optimized Hi–C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*, **123**, 56–65.

64. Harewood,L., Kishore,K., Eldridge,M.D., Wingett,S., Pearson,D., Schoenfelder,S., Collins,V.P. and Fraser,P. (2017) Hi–C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.*, **18**, 125.

65. Rickman,D.S., Soong,T.D., Moss,B., Mosquera,J.M., Dlabal,J., Terry,S., MacDonald,T.Y., Tripodi,J., Bunting,K., Najfeld,V. *et al.* (2012) Oncogene-mediated alterations in chromatin conformation. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 9083–9088.

66. Kamińska,K., Nalejska,E., Kubiak,M., Wojtysiak,J., Żołna,Ł., Kowalewski,J. and Lewandowska,M.A. (2019) Prognostic and predictive epigenetic biomarkers in oncology. *Mol. Diagnosis Ther.*, **23**, 83–95.

67. Grabowski,T.J., Cho,H.S., Vonsattel,J.P.G., William Rebeck,G. and Greenberg,S.M. (2001) Novel amyloid precursor protein mutation in an Iowa family with dementia and severe cerebral amyloid angiopathy. *Ann. Neurol.*, **49**, 697–705.

68. Battaglia,A., Carey,J.C. and South,S.T. (2015) Wolf-Hirschhorn syndrome: A review and update. *Am. J. Med. Genet. Part C Semin. Med. Genet.*, **169**, 216–223.

69. Kaul,G., Pattan,G. and Rafeequi,T. (2011) Eukaryotic elongation factor-2 (eEF2): Its regulation and peptide chain elongation. *Cell Biochem. Funct.*, **29**, 227–234.

70. Boginya,A., Detroja,R., Matityahu,A., Frenkel-Morgenstern,M. and Onn,I. (2019) The chromatin remodeler Chd1 regulates cohesin in budding yeast and humans. *Sci. Rep.*, **9**, 8929.