Published online 11 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D465–D470***
doi: 10.1093/nar/gkx1085

Minimotif Miner 4: a million peptide minimotifs and counting

Kenneth F. Lyon[1,†], Xingyu Cai[2,†], Richard J. Young[1], Abdullah-Al Mamun[2], Sanguthevar Rajasekaran[2,*] and Martin R. Schiller[1,*]

[1]Nevada Institute of Personalized Medicine and School of Life Sciences, University of Nevada, Las Vegas, 89154 4004 NV, USA and [2]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269 2155, USA

Received September 18, 2017; Revised October 13, 2017; Editorial Decision October 16, 2017; Accepted November 09, 2017

ABSTRACT

Minimotif Miner (MnM) is a database and web system for analyzing short functional peptide motifs, termed minimotifs. We present an update to MnM growing the database from ∼300 000 to >1 000 000 minimotif consensus sequences and instances. This growth comes largely from updating data from existing databases and annotation of articles with high-throughput approaches analyzing different types of post-translational modifications. Another update is mapping human proteins and their minimotifs to know human variants from the dbSNP, build 150. Now MnM 4 can be used to generate mechanistic hypotheses about how human genetic variation affect minimotifs and outcomes. One example of the utility of the combined minimotif/SNP tool identifies a loss of function missense SNP in a ubiquitylation minimotif encoded in the excision repair cross-complementing 2 (ERCC2) nucleotide excision repair gene. This SNP reaches genome wide significance for many types of cancer and the variant identified with MnM 4 reveals a more detailed mechanistic hypothesis concerning the role of ERCC2 in cancer. Other updates to the web system include a new architecture with migration of the web system and database to Docker containers for better performance and management. Weblinks: minimotifminer.org and mnm.engr.uconn.edu

INTRODUCTION

Minimotifs are short peptide sequences that are important in evolution and human disease (1–6). Minimotifs play a critical role in interaction of proteins with other proteins and molecules. We refer to an occurrence of minimotif peptide sequences in a protein or peptide as an instance, but sometimes through analysis of multiple instances a consensus sequence pattern is identified that accounts for observed variation and defines observed degeneracy. In 2006, we released the original Minimotif Miner (MnM), which had mostly consensus sequences ($n = \sim 300$) annotated from the literature (7). In the 2008, release of MnM 2, the database grew to >5000 minimotifs and we started including instances (8). We had been using a minimotif syntax put forth by the Seefeld Convention (9), but recognized several deficiencies which led us to propose a new model for minimotifs with a revised syntax that addressed these shortcomings (10). We later revised this model proposing inclusion of structure in the minimotif definition and other minor modifications for MnM 3 in 2011 (11,12). The MnM 3 database utilizing this model grew to ∼300 000 minimotifs.

At the beginning of the MnM project we recognized that predictions from minimotif consensus sequences were excellent at identifying true positives, but lacked specificity, and, only a small percentage of new minimotif predictions were accurate (7). This led us to start the process of generating filters that improved accuracy. We first used proteome frequency, protein surface prediction and evolutionary conservation filters with modest effect (7). Next, we sequentially tested several filters for protein–protein interactions, related molecular functions, genetic interactions and secondary structure, each yielding modest increases in accuracy (11,13–15). However, when all these filters were combined an overall prediction accuracy of 90% was achieved (16). However, this accuracy is limited by the need of additional knowledge about the protein in other databases. Other groups have investigated other methods for improved accuracy as well (17–20).

In 2007, investigators from the eukaryotic linear motif resource group suggested there may be over a million minimotif instances in the cell (21). At the time we agreed. However, we now suggest that this may be a vast underestimate. This hypothesis is supported by this report that the MnM database has grown to >1 million instances. Furthermore,

*To whom correspondence should be addressed. Tel: +1 702 895 5546; Fax: +1 702 895 5728; Email: martin.schiller@unlv.edu
Correspondence may also be addressed to Sanguthevar Rajasekaran. Email: rajasek@engr.uconn.edu
†These authors contributed equally to the paper as first authors.

Table 1. A decade of growth of the MnM database

Category	MnM	MnM 2	MnM 3	MnM 4
Total minimotifs	462	5089	294 933	1 060 436
Consensus	312	858	880	894
Instance	44	4229	294 053	1 059 542
Activity classes				
Post-translational modifications	116	663	210 949	912 735
Binding	162	4689	4922	147 654
Trafficking	34	195	228	229
Required for cell process			47	47
Unique minimotif				
Sequences	312	2224	185 833	590 589
Proteins	<312	1211	49 671	182 868
Targets	<312	687	2620	4586
C-Terminal minimotifs	ND	ND	ND	12 808

ND = not determined.

if we consider that the number of minimotifs in proteins like Tat and Nef from human immunodeficiency virus (22–24), which seem to completely decorate the surface of the protein, there is likely at least another 1–2 of magnitude of minimotif instances in the cell. Given this prevalence in the cell, their functions are likely pervasive in most, if not all cell processes. Herein, we provide more details updating the MnM database and web system to version 4.

MATERIALS AND METHODS

We used separate parsers to extract data from each of four sources of minimotifs: UniProtKB, PhosphoSitePlus, MEROPS and manually annotated entries (25–27). We also used the proteomics standards initiative–modification (PSI-MOD) database from the protein information resource to accurately describe post-translational modifications (PTMs) with this ontology (28). More than 1.3 million missense single nucleotide polymorphisms (SNPs) from single nucleotide polymorphism database (dbSNP) 150 (2) were added to update MnM 4 (29). The integration of the SNP data is used to identify minimotifs that are variable in the human population since minimotifs play a role in evolution and disease (1,2). To identify molecular functions and cell processes that are enriched with minimotifs a Gene Ontology (GO) analysis was performed on 12 405 human RefSeq proteins using the GORILLA gene ontology enrichment analysis tool (30,31). GORILLA produces *P*- and *q*-values for enrichment, where the *q*-value is a *P*-value with a correction for the false discovery rate due to multiple testing.

Binplots were created using ggplot2 in the R programming environment (http://link.springer.com/10.1007/978-0-387-98141-3; https://www.R-project.org). External databases and new annotations were mined and queried with a combination of Java (Sun Microsystems), MySQL (my structured query language) and Python (Python Software Foundation) custom programs. MnM 4 was deployed on a mirrored server with a Linux container architecture (Supplementary Methods and Supplementary Figure S1).

RESULTS

Two strategies were used to update the MnM database. Parsers were modified or built to extract information into the MnM database model. Source databases were universal protein resource (UniProt), PhosphoSitePlus and MEROPS (25–27). A total of 765 503 minimotifs were added to the database. Since the last MnM release many additional minimotifs have been published. PubMed was searched with relevant search terms to identify high-throughput affinity mass spectrometry papers that contained 100 or 1000 s of minimotifs. This produced 27 017 minimotifs annotated from 15 papers. Collectively, MnM has had continued growth to now over 1 million minimotifs in >180 000 unique proteins (Table 1). The data comes from 15 152 research articles. As expected, most of the minimotifs are instances ($n = 1\,059\,542$), with 894 consensus sequences. However, the rich source of instances presents an opportunity for a more standardized approach for generating consensus sequences and position specific-scoring matrices.

Enrichment analysis was performed for GO functions, processes and components in the 12 405 human reference sequence (RefSeq) proteins that contained minimotifs (30–32). The 10 GO categories with the highest *P*-values are listed in Table 2 and top 50 are in Supplementary Table S1.These tables also contain *p*- and *q*-values for term enrichment. Not surprisingly, the most enriched GO terms were related to receptor activity and signal transduction (Table 2).

Since the size of MnM database grew several fold in MnM 4, the nature of the minimotif activities was assessed and quantified through database querying. Most of the minimotifs in MnM 4 (86%) are for PTMs with most of the remaining minimotifs involved in intermolecular binding interactions (Figure 1A and B). Most of the minimotifs were instance with <1000 consensus sequences. The bias toward PTMs and instances likely reflects the increased application of high-throughput experimental approaches, primary affinity mass spectrometry. The new MnM 4 database contains ∼12 000 new C-terminal minimotifs on the last 10 amino acids of a protein, thus is a source of new information for the C-terminome minimotif database (33).

In MnM 4, approximately half of the PTMs subactivities were for phosphorylation sites with acetylation, glycosylation, methylation, ubiquitylation and proteolysis subactivities each approximately equally contributing to ∼35% of the database (Figure 1A). Other activities were not as commonly observed. Some of the growth in MnM 4 came

Table 2. Enrichment of Gene Ontology terms for human proteins with minimotifs[1]

Description	*P*-value	*q*-value	Type
Olfactory receptor activity	2.89E-265	1.14E-261	function
Detection of chemical stimulus involved in perception of smell	2.89E-265	3.95E-261	process
Detection of stimulus involved in sensory perception	8.09E-232	2.76E-228	process
G-protein coupled receptor activity	2.84E-213	5.6E-210	function
Detection of stimulus	1.82E-207	4.98E-204	process
Transmembrane signaling receptor activity	1.17E-190	1.53E-187	function
Transmembrane receptor activity	2.07E-189	2.05E-186	function
G-protein coupled receptor signaling pathway	8.6E-187	1.96E-183	process
Signaling receptor activity	6.33E-182	5E-179	function
Molecular transducer activity	4.6E-161	2.6E-158	function

[1]Supplementary Table S1 is a list of the top 50 enriched GO functions.

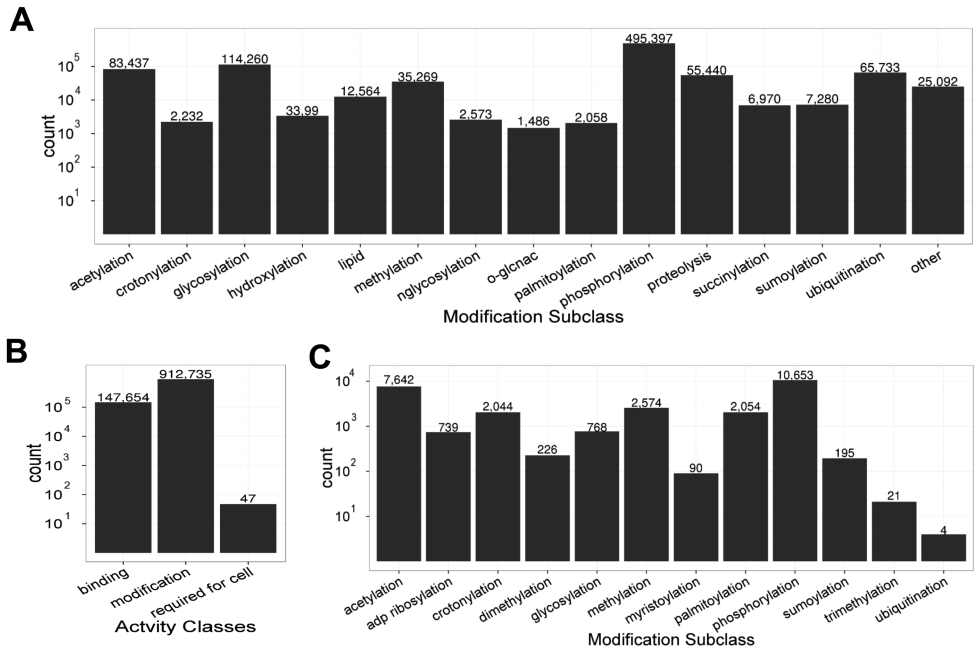

Figure 1. Binplots representing counts of minimotif activity classes. Instances of major minimotif modification activities (**A**) 14 most common modification activity subclasses in MnM 4 (**B**), and the 12 new manually annotated modification activity subclasses for MnM 4 (**C**).

from annotation of published papers using new custom built parsers. This approach and new database imports reduced the bias of MnM 4 toward phosphorylation, with growth in other modification subactivities such as acetylation, methylation and crotonylation (Figure 1C).

There have been several minor modifications in MnM 4. The website architecture was migrated to Docker containers to improve performance and management (Supplementary Methods). MnM 4 maintains flexible navigation among pages. The MnM homepage and results page was updated with a new look, while unnecessary details were removed (Supplementary Figure S2). PSI-MOD and RefSeq were updated. A search progress indicator was added to track progress after submission. For the SNP functions, identifiers for dbSNP (rs number) have been added to the output (Figure 2C).

The 'Show SNPs that change minimotifs' function introduced in a previous version of MnM has become much more interesting given the explosive growth of human whole exome and whole genome sequencing. MnM has a menu selection to show SNPs in the protein sequence window on the query results page. In this window a user can mouse click one or more SNPs to reveal the amino acid encoded by the missense variant. Upon selection of the 'View new minimotifs from SNPs' menu item, a new table displays a list of minimotifs introduced and eliminated by the SNP. Since the original release of this function, dbSNP has grown by several orders of magnitude and the number of missense minimotifs has grown from ∼182 099 in MnM 3 to 1 291 434 in MnM 4 after update with dbSNP, build 150.

To explore the utility of the updated SNP tool, we analyzed *ERCC2*, a gene strongly associated several types of cancer. The *ERCC2* gene encodes a helicase in the TFIIH complex essential for DNA repair through the nucleotide excision repair pathway. The ERCC2 protein (NP000391) was analyzed with MnM 4 producing 100 predicted minimotifs. A total of 37 of these had high combined filter scores suggesting that these are accurate predictions. We next searched for those motifs that had a SNP in a critical position. In MnM 3, ERCC2, typical of other protein

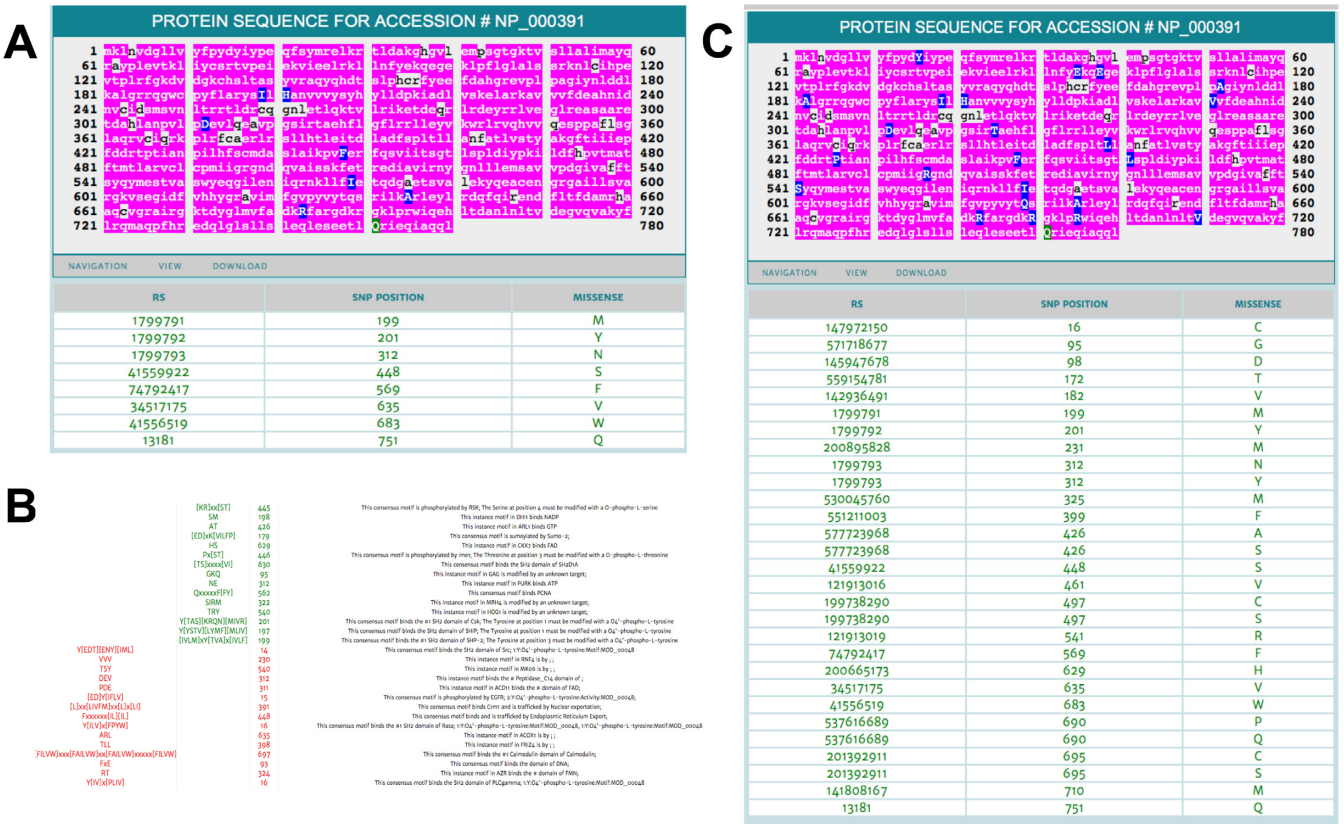

Figure 2. Outputs of SNP functional analysis for excision repair cross-complementing 2 (ERCC2) with MnM. (**A**) MnM 3 output of ERCC2 with minimotifs highlighted magenta and SNPs highlighted dark blue or green. The table shows a list of the eight SNPs that are indicated in the sequence window. (**B**) Example rows of MnM 3 output showing minimotifs introduced (red font) or eliminated (green font) by an SNP that is selected in the sequence window. (**C**) MnM 4 output of ERCC2 with minimotifs highlighted magenta and SNPs highlighted dark blue or green. The table shows a list of the 29 SNPs that are indicated in the sequence window.

queries had <10 SNPs ($n = 8$) that changed 38 minimotifs (Figure 2A). However, the SNP update of ∼1.3 million missense mutations in MnM 4 produces ∼3.5-fold more SNPs in ERCC2 minimotifs (Figure 2B, $n = 29$) with 211 predicted minimotif changes, reflecting the large growth of dbSNP (29). As in MnM 3, the new MnM version codifies minimotifs introduced by a SNP with the font colored green and minimotifs eliminated by a minimotif with the font colored red (Figure 2B).

One of these SNPs of interest was a variant (rs13181) encoding the missense substitution K751Q. This minimotif was a ubiquitylation site (PSI-MOD: MOD:01148) at position K751. MnM4 reveals that the ubiquitylated lysine was changed to glutamine (K751Q) with this SNP. Most certainly this variant creates a loss of function for ubiquitylation at this site. Most often C-terminal ubiquitylation is involved in degradation of the protein, thus it would be expected to increase protein expression. This SNP is of high interest because it reached genome wide significance in several genome wide association studies and meta analyzes for lung, melanoma, breast cancer, glioma, pancreatic, esophageal and ovarian cancers (34–42). Moreover, the variant is observed at relatively high allele frequency (33%) in ∼60 000 sequences from the Exome Aggregation Consortium (EXaC) browser (43). From these published works and our minimotif analysis we are able to generate a new hypothesis that this critical variant eliminates a ubiquitylation site, effecting the degradation of ERCC2. Since this gene is crucial for nucleotide excision repair, this minimotif may provide a clue as to the mechanism of loss of excision repair functions in several cancers.

DISCUSSION

In this paper we report the growth of the MnM database to more than a million minimotifs. However, we think that we are still just in the early stages of minimotif discovery. There are examples of proteins, where the entire protein surface is covered with minimotifs; see examples for tp53, RNAP II, and Histone 3 (44). Furthermore, it is clear that minimotif sites overlap with protein-protein interaction sites, as well as other minimotifs. Their prevalence implies that proteins are much more than simplistic switches for turning on or off an enzymatic function or cell process, but rather a much more complicated functional unit, like an integrated circuit computer chip with many minimotif and protein interaction inputs and outputs. In this role, minimotifs amplify interconnections within the cellular network.

As the number of minimotifs discovered continues to grow, one important question that arises is why are minimotifs not frequently turning up as a vulnerability in disease? There are examples where minimotifs are important

in rare disorders or infectious disease, although this is not commonly observed (2,20,45). Minimotifs are a source of functional genetic variation and important targets of selection and evolution (1). Despite their apparent importance, their general minimal association with disease can be reconciled by the explanation that minimotifs provide a functional redundancy and network robustness, such that loss of function of a single motif only impacts cellular function when it is at a point of network vulnerability. And thus, why minimotifs are only mutated in a few rare disorders. Proteins with multiple minimotifs engaging the same target protein (e.g. many SH3 domain/PxxP interactions) are examples of this encoded robustness, where mutation of one of many PxxP minimotifs in a protein is not likely to significantly influences its interaction with and SH3 protein target (46). This scenario would also explain the minimal influence of minimotif mimetic drugs as disease therapeutics. One exciting possibility is that minimotifs may contribute to genetic risk in common disorders, however, this will require future study.

To help facilitate the study of minimotifs in these types of roles we have enhanced the functionality of MnM 4 to identify minimotifs that are variable in the human population. By identifying those minimotifs residues that are both covalently modified and changed to an amino acid with a chemistry not consistent with that PTM, loss of function minimotifs in the human population can be confidentially inferred as in the ERCC2 ubiquitylation example we highlight herein. We hope the tool to investigate SNPs in minimotifs in MnM 4 will help facilitate study of the roles of minimotifs in selection, evolution, network function, disease, and potentially identify targets for novel therapeutics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Nevada Institute of Personalized Medicine for use of their computers.

FUNDING

National Institutes of Health [GM107983, LM010101, GM079689]; National Science Foundation [1447711]; Nevada Governor's Office of Economic Development Knowledge Fund. Funding for open access charge: National Institutes of Health [GM107983]; National Science Foundation [1447711].
Conflict of interest statement. None declared.

REFERENCES

1. Lyon,K.F., Strong,C.L., Schooler,S.G., Young,R.J., Roy,N., Ozar,B., Bachmeier,M., Rajasekaran,S. and Schiller,M.R. (2015) Natural variability of minimotifs in 1092 people indicates that minimotifs are targets of evolution. *Nucleic Acids Res.*, **43**, 6399–6412.
2. Kadaveru,K., Vyas,J. and Schiller,M.R. (2008) Viral infection and human disease–insights from minimotifs. *Front. Biosci.*, **13**, 6455–6471.
3. Kaneko,T., Huang,H., Zhao,B., Li,L., Liu,H., Voss,C.K., Wu,C., Schiller,M.R. and Li,S.S.-C. (2010) Loops govern SH2 domain specificity by controlling access to binding pockets. *Sci. Signal*, **3**, ra34.
4. Uyar,B., Weatheritt,R.J., Dinkel,H., Davey,N.E. and Gibson,T.J. (2014) Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Mol. Biosyst.*, **10**, 2626–2642.
5. Schiller,M.R. (2016) The minimotif synthesis hypothesis for the origin of life. *J. Transl. Sci.*, **2**, 289–296.
6. Reimand,J., Wagih,O. and Bader,G.D. (2015) Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genet.*, **11**, e1004919.
7. Balla,S., Thapar,V., Verma,S., Luong,T., Faghri,T., Huang,C.H., Rajasekaran,S., del Campo,J.J., Shinn,J.H., Mohler,W.A. *et al.* (2006) Minimotif Miner: a tool for investigating protein function. *Nat. Methods*, **3**, 175–177.
8. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al.* (2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
9. Aasland,R., Abrams,C., Ampe,C., Ball,L.J., Bedford,M.T., Cesareni,G., Gimona,M., Hurley,J.H., Jarchau,T., Lehto,V.P. *et al.* (2002) Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.*, **513**, 141–144.
10. Vyas,J., Nowling,R.J., Maciejewski,M.W., Rajasekaran,S., Gryk,M.R. and Schiller,M.R. (2009) A proposed syntax for Minimotif Semantics, version 1. *BMC Genomics*, **10**, 360–372.
11. Sargeant,D.P., Gryk,M.R., Maciejewski,M.W., Thapar,V., Kundeti,V., Rajasekaran,S., Romero,P., Dunker,K., Li,S.-C., Kaneko,T. *et al.* (2012) Secondary structure, a missing component of sequence-based minimotif definitions. *PLoS One*, **7**, e49957.
12. Mi,T., Merlin,J.C., Deverasetty,S., Gryk,M.R., Bill,T.J., Brooks,A.W., Lee,L.Y., Rathnayake,V., Ross,C.A., Sargeant,D.P. *et al.* (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res.*, **40**, D252–D260.
13. Merlin,J.C., Rajasekaran,S., Mi,T. and Schiller,M.R. (2012) Reducing false-positive prediction of minimotifs with a genetic interaction filter. *PLoS One*, **7**, e32630.
14. Rajasekaran,S., Merlin,J.C., Kundeti,V., Mi,T., Oommen,A., Vyas,J., Alaniz,I., Chung,K., Chowdhury,F., Deverasatty,S. *et al.* (2010) A computational tool for identifying minimotifs in protein-protein interactions and improving the accuracy of minimotif predictions. *Proteins*, **79**, 153–164.
15. Rajasekaran,S., Mi,T., Merlin,J.C., Oommen,A., Gradie,P. and Schiller,M.R. (2010) Partitioning of minimotifs based on function with improved prediction accuracy. *PLoS One*, **5**, e12276.
16. Mi,T., Rajasekaran,S., Merlin,J.C., Gryk,M. and Schiller,M.R. (2012) Achieving High Accuracy Prediction of Minimotifs. *PLoS One*, **7**, e45589.
17. Davey,N.E., Shields,D.C. and Edwards,R.J. (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, **25**, 443–450.
18. Mooney,C., Davey,N., Martin,A.J.M., Walsh,I., Shields,D.C. and Pollastri,G. (2011) In silico protein motif discovery and structural analysis. *Methods Mol. Biol.*, **760**, 341–353.
19. Mooney,C., Pollastri,G., Shields,D.C. and Haslam,N.J. (2011) Prediction of short linear protein binding regions. *J. Mol. Biol.*, **415**, 193–204.
20. Dinkel,H., Michael,S., Weatheritt,R.J., Davey,N.E., Van Roey,K., Altenberg,B., Toedt,G., Uyar,B., Seiler,M., Budd,A. *et al.* (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.*, **40**, D242–D251.
21. Tompa,P., Davey,N.E., Gibson,T.J. and Babu,M.M. (2014) A million peptide motifs for the molecular biologist. *Mol. Cell*, **55**, 161–169.
22. Sarmady,M., Dampier,W. and Tozeren,A. (2011) Sequence- and interactome-based prediction of viral protein hotspots targeting host proteins: a case study for HIV Nef. *PLoS One*, **6**, e20735.
23. Sargeant,D.P., Deverasetty,S., Strong,C.L., Alaniz,I.J., Bartlett,A., Brandon,N.R., Brooks,S.B., Brown,F.A., Bufi,F., Chakarova,M. *et al.* (2014) The HIVToolbox 2 web system integrates sequence, structure, function and mutation analysis. *PLoS One*, **9**, e98810.

24. Evans,P., Dampier,W., Ungar,L. and Tozeren,A. (2009) Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics*, **2**, 27–39.
25. Rawlings,N.D., Barrett,A.J. and Finn,R. (2016) Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, **44**, D343–D350.
26. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
27. Hornbeck,P.V., Chabra,I., Kornhauser,J.M., Skrzypek,E. and Zhang,B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
28. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.
29. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
30. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48–54.
31. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.
32. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
33. Sharma,S., Toledo,O., Hedden,M., Lyon,K.F., Brooks,S.B., David,R.P., Limtong,J., Newsome,J.M., Novakovic,N., Rajasekaran,S. *et al.* (2016) The Functional Human C-Terminome. *PLoS One*, **11**, e0152731.
34. Bernard-Gallon,D., Bosviel,R., Delort,L., Fontana,L., Chamoux,A., Rabiau,N., Kwiatkowski,F., Chalabi,N., Satih,S. and Bignon,Y.-J. (2008) DNA repair gene ERCC2 polymorphisms and associations with breast and ovarian cancer risk. *Mol. Cancer*, **7**, 36–42.
35. Liu,J., Song,J., Wang,M.-Y., He,L., Cai,L. and Chou,K.-C. (2015) Association of EGF rs4444903 and XPD rs13181 polymorphisms with cutaneous melanoma in Caucasians. *Med. Chem.*, **11**, 551–559.
36. Qian,T., Zhang,B., Qian,C., He,Y. and Li,Y. (2017) Association between common polymorphisms in ERCC gene and glioma risk. *Medicine (Baltimore)*, **96**, 1118–1126.
37. Qin,Q., Zhang,C., Yang,X., Zhu,H., Yang,B., Cai,J., Cheng,H., Ma,J., Lu,J., Zhan,L. *et al.* (2013) Polymorphisms in XPD gene could predict clinical outcome of platinum-based chemotherapy for non-small cell lung cancer patients: a meta-analysis of 24 studies. *PLoS One*, **8**, e79864.
38. Smolarz,B., Makowska,M., Samulak,D., Michalska,M.M., Mojs,E., Wilczak,M. and Romanowicz,H. (2014) Single nucleotide polymorphisms (SNPs) of ERCC2, hOGG1, and XRCC1 DNA repair genes and the risk of triple-negative breast cancer in Polish women. *Tumour Biol.*, **35**, 3495–3502.
39. Wu,Y., Lu,Z.-P., Zhang,J.-J., Liu,D.-F., Shi,G.-D., Zhang,C., Qin,Z.-Q., Zhang,J.-Z., He,Y., Wu,P.-F. *et al.* (2017) Association between ERCC2 Lys751Gln polymorphism and the risk of pancreatic cancer, especially among Asians: evidence from a meta-analysis. *Oncotarget*, **8**, 50124–50132.
40. Zhu,M.-L., He,J., Wang,M., Sun,M.-H., Jin,L., Wang,X., Yang,Y.-J., Wang,J.-C., Zheng,L., Xiang,J.-Q. *et al.* (2014) Potentially functional polymorphisms in the ERCC2 gene and risk of Esophageal Squamous Cell Carcinoma in Chinese populations. *Sci. Rep.*, **4**, 6281.
41. Wang,L.-E., Gorlova,O.Y., Ying,J., Qiao,Y., Weng,S.-F., Lee,A.T., Gregersen,P.K., Spitz,M.R., Amos,C.I. and Wei,Q. (2013) Genome-wide association study reveals novel genetic determinants of DNA repair capacity in lung cancer. *Cancer Res.*, **73**, 256–264.
42. Sun,Y., Zhang,H., Ying,H., Jiang,W. and Chen,Q. (2015) A meta-analysis of XPD/ERCC2 Lys751Gln polymorphism and melanoma susceptibility. *Int. J. Clin. Exp. Med.*, **8**, 13874–13878.
43. Karczewski,K.J., Weisburd,B., Thomas,B., Solomonson,M., Ruderfer,D.M., Kavanagh,D., Hamamsy,T., Lek,M., Samocha,K.E., Cummings,B.B. *et al.* (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.*, **45**, D840–D845.
44. Hsu,W.-L., Oldfield,C.J., Xue,B., Meng,J., Huang,F., Romero,P., Uversky,V.N. and Dunker,A.K. (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci.*, **22**, 258–273.
45. Sobhy,H. (2016) A review of functional motifs utilized by viruses. *Proteomes*, **4**, 3–23.
46. Mayer,B.J. (2001) SH3 domains: complexity in moderation. *J. Cell. Sci.*, **114**, 1253–1263.