MethylSight: Taking a wider view of lysine methylation through computer-aided discovery to provide insight into the human methyl-lysine proteome

Biggar, Kyle K.[1], Ruiz-Blanco, Y.B.[2], Charih, F.[3], Fang, Q.[4], Connolly, J.[1], Frensemier, K.[1], Adhikary, H.[1], Li, S.S.C.[+4], Green, J.R.[+3]

Author Affiliations:

1. Institute of Biochemistry, Carleton University, 1125 Colonel By Dr., Ottawa ON Canada
2. Computational Biochemistry, Zentrum fur Medizinische Biotechnologie, Fakultat fur Biologie, Universitat Duisburg-Essen, Universitatsstr. 2, 45117 Essen Germany
3. Department of Computer Engineering, Carleton University, 1125 Colonel By Dr., Ottawa ON Canada
4. Department of Biochemistry, Western University, 1151 Richmond Rd., London ON Canada

[+]Correspondence:

jrgreen@sce.carleton.ca & sli@uwo.ca

SUMMARY

Post-translational lysine methylation has been found to play a fundamental role in the regulation of protein function and the transmission of biological signals. We present the development of a machine learning model for predicting lysine methylation sites among human proteins. The model uses fully-alignment-free features encoding sequence-based information. A total of 57 novel predicted histone methylation sites were selected for evaluation by targeted mass spectrometry, with 51 sites positively reassigned as true methylated sites, while one site was also found to be dynamically responsive to DNA damage. To gain insight into the cellular function of the lysine methylation system, we reveal links between cellular metabolic and GTPase signal transduction, demonstrating a dynamic hypoxia-responsive methylation of the inducible nitric oxide synthase (NOS2). With the growing implication of lysine methylation in human health and disease, the development of methods that help to target its discovery will become of critical importance to understanding its biological implications.

Keywords:

lysine methylation; non-histone methylation; histone methylation;

INTRODUCTION

Post-translational modifications (PTMs) are reversible chemical modifications that play a crucial role in the regulation of protein function and the transmission of biological signals (Mann and Jensen, 2003). This diversity in available chemical protein modifications greatly expands the information potential within the PTM code, allowing cells to exert much greater control over crucial cellular processes. For example, histone proteins and their diverse array of PTMs have been subject to exquisite evolutionary conservation in eukaryotes, and one of the main types of PTMs occurring on histones is the reversible methylation of lysine residues (Martin and Zhang, 2005). Although lysine methylation is commonly known as a PTM of histone proteins, the prevalence of the methylation of non-histone proteins has received considerable attention in recent years, and has been found to play crucial roles in a number of human diseases, including cancer (Zhang et al., 2012; Arrowsmith et al., 2012; Biggar and Li, 2015; Hamamoto et al., 2015). Given the importance of PTMs in protein regulation and cellular function, and the prevalence of its dysregulation in human health and disease, the development of identification technologies have received considerable attention. As a result, there has been a significant effort placed on the development of both *in silico* and mass spectrometry-based enrichment methods to aid in the discovery and exploration of the methyl-lysine proteome (Liu et al., 2013; Carlson et al., 2014; Shi et al., 2015; Wen et al., 2016; Audagnotto and Peraro, 2017).

The number of known methylated proteins and modification sites has grown tremendously in recent years. Indeed, recent advances in identification technologies (i.e., affinity enrichment methods and high-resolution mass spectrometry) have provided insight into a large number of non-histone proteins that undergo lysine methylation, with many of these methylation events shown to have important regulatory functions for the respective proteins (Liu et al., 2013; Carlson et al., 2014). Furthermore, it is now known that the methylation of proteins is extremely dynamic and is involved in a growing number of cellular processes (Wu et al., 2017). These studies suggest a broad role for lysine methylation in regulating protein function, well beyond controlling chromatin dynamics *via* histone methylation. For example, the tumor suppressor p53 is methylated on multiple lysine residues and individual modifications have the capaci-

ty to regulate p53 function through a surprisingly diverse array of mechanisms (West and Gozani, 2011). Further, the catalytic subunit of DNA-dependent protein kinase (DNA-PK), an important regulator of DNA damage repair, is methylated on multiple lysine residues and methylation status dictates its ability to effectively repair damaged DNA (Liu et al., 2013).

Given the extensive regulatory importance that is beginning to be realized for lysine methylation, the successful identification of modification sites has become increasingly important. One of the largest challenges placed on the discovery of lysine-methylated proteins has been limitations in identification technology. It has proven to be difficult to develop specific affinity strategies that are able to enrich for the lysine methylation modification (Liu et al., 2013; Carlson et al., 2014). As a result, the identification of lysine methylation sites has not experienced the same growth in discovery as other PTMs, such as serine/threonine and tyrosine phosphorylation, lysine acetylation, or arginine methylation. However, the development of both new *in silico* prediction resources combined with targeted enrichment strategies will help to aid in the initial annotation of the methyllysine proteome on a proteome scale. Although several affinity strategies that utilize natural methyl-binding domains have been remarkably successful in the identification of new lysine methylation events when coupled with mass spectrometry (Liu et al., 2013; Carlson et al., 2014), these approaches are inherently biased towards the biologically-relevant binding specificity of the domain used for the initial enrichment. *In silico* prediction methods help to overcome this issue by predicting methylation events based on general underlying characteristics of all known modification sites. During the past decade, there have been several attempts to develop methyllysine and methylarginine computational predictors (Table S1) (Chen et al., 2006; Hu et al., 2011; Qiu et al., 2014; Shao et al., 2009; Shi et al., 2012; Shi, et al., 2015; Shien et al., 2009). These studies built their models from the available information of methylated sites extracted from UniProtKB, PhosphoSite-Plus, and PubMed, gathering only a few hundred methylation sites. Therefore, these predictors are limited to approximately 200 non-redundant methyllysine sites for building and assessing their models. Critically, the expected diversity of methyllysine sites can undoubtedly not be represented with such a few number of examples given the impres-

sive growth of validated methylated sites in recent years (Cao and Garcia, 2016). Most notably, there has been a stark lack of experimental validation highlighting the prospective use of such *in silico* methods to aid *in vivo* discovery.

We address these limitations by conducting a model learning approach based, first, on alignment-free features to directly capture the physical and chemical properties of the peptides, rather than relying on domain-specific features that often fail due to the limited amount of available data. At the same time, we enlarged the size of our training dataset to approximately two thousand sites that have been gathered from years of experimental studies on the lysine methylation and deposited in the PhosphoSite database (www.phosphosite.com). Secondly, we treat imbalance by using cost-sensitive learning, thus the datasets are kept in their intrinsic imbalanced ratio during cross-validation and hold-out tests rather than introducing synthetic training data nor losing valuable exemplars through under-sampling. In summary, our method of methyllysine prediction has resulted in a number of promising methylation sites based on comparisons with other existing methods using common independent tests. Moreover, our proteome-wide predictions provide a valuable resource to gain functional insight into the methyllysine proteome, and for the experimental validation of new methylation sites and for the generation of useful hypotheses. The MethylSight user interface, source code, datasets and support vector machine (SVM) models can be freely found at http://methylsight.com.

MATERIALS AND METHODS

Preparing the data sets

The generation of training, calibration, and test datasets are described in the supplemental methods and summarized in Table S2. Feature extraction was accomplished using ProtDCal properties, groups, modifiers and aggregators, as follows: amino acid *properties* are first computed over different *grouping* subsets of amino acids within each input sequence window (Ruiz-Blanco et al., 2015). For example, the hydrophobicity of all charged amino acids within a sequence window could form the basis for a ProtDCal descriptor. In this case, 12 amino acid *properties* were used to

numerically encode the physical-chemical characteristics of the residues. These properties are found in the AAindex database (Kawashima and Kanehisa, 2000) and are also described in the ProtDCal documentation. Fourteen residue *groups* are used based on either side chain structure or using specific residue positions within the input sequence window. The properties can then be *modified* by the computed properties of neighboring amino acids, before applying an *aggregation operation* to reduce the vector down to a scalar quantity, known as a descriptor or feature. Two *modification operators* for capturing vicinity information and twelve *aggregation operators* ultimately transform the property vector of each amino acid group into the final scalar features. The project files with the lists of indices, groups, modification and aggregation operators as well as other parameters for the calculations are provided on the http://methylsight.com website. The above configuration leads to an initial set of 3720 descriptors, which is subsequently filtered to identify those features most useful for methyllysine prediction using a pipeline of supervised and unsupervised feature selection processes.

Feature selection begins with information gain (IG) analysis, which retains only those features whose distribution across all sites in the training data correlates with class label. All the attributes with a non-zero IG value were extracted in this step. Subsequently, an unsupervised redundancy filter is applied, using a single-linkage clustering algorithm with the Spearman correlation coefficient as the similarity measure. Features exhibiting pair-wise correlation above 0.9 are clustered together and only one representative feature from each cluster is kept. Ultimately, the supervised WrapperSubsetEval method, implemented in Weka 3.7.11 (Hall et al., 2009), is used to extract an optimum subset of features for modelling. This method was configured using a Genetic Search for exploring the feature space and potential feature sets are evaluated using the classification F-measure of the positive class in 5-fold cross-validation tests using SVM classifiers with a linear kernel. The cost-sensitive sequential minimal optimization (SMO) algorithm (Cai and Cherkassky, 2012) was used to train all SVM classifiers in this work. The cost matrix reflected the relative class imbalance in the data, such that the false negative error cost is equal to the number of

negative instances and the cost of false positive errors was fixed at the number of positive instances in the data (Table S2).

Training the support vector machine predictor

Following feature selection, a grid-based optimization of SVM hyperparameters was conducted using the training and calibration data. The final model is selected according to the prediction accuracy (in terms of F-measure, precision, and recall). The optimal model is selected based on strong performance in both cross-validation testing and in hold-out calibration testing.

Multiple reaction monitoring mass spectrometry (MRM-MS)

To validate the status of predicted methylation sites, isolated proteins were digested with trypsin and the digest was analyzed by positive ESI LC-MS/MS on a triple quadrupole mass spectrometer (4000 QTRAP, Applied Biosystems Inc.) utilizing Q3 as a linear ion trap. A nanoAcquity UPLC system (Waters) equipped with a C18 analytical column (1.7 µm, BEH130, 75 µm×250 mm) was used to separate the peptides at the flow rate of 300 nl/min and operating pressure of 8000 psi. Peptides were eluted using a 62 min gradient from 95% solvent A (H_2O, 0.1% formic acid) and 5% B (acetonitrile, 0.1% formic acid) to 50% B in 41 min, 6 min at 90% B, and back to 5% for 10 min. Eluted peptides were directly electrosprayed (Nanosource, ESI voltage +2000V) into the mass spectrometer. The instrument was set to monitor up to 200 transitions in each sample with a dwelling time of at least 25 msec/transition.

The *in silico* protease digest patterns (i.e. to generate precursor ions) and the corresponding MRM transitions were compiled using the Skyline™ software made freely available to us by the McCoss Lab, Department of Genome Sciences University of Washington School of Medicine (MacLean et al., 2010). Transitions that are larger than the precursor ion was selected based on the Skyline predictions and the specific b/y ions that allow unambiguous identification of the methylated lysine site were included. Positive identification of a new methylation site required the successful detection of at least three transitions. All transitions used to identify methylation sites are listed in Table S3. An internal NOS2-specific peptide (NH2-QQNESPQPLVETGK-

COOH) was used as a standard to normalize relative NOS2 methylation data to protein abundance.

Functional analysis of prediction methyllysine proteome

To functionally annotate the biological functions enriched in the dataset of known and predicted human lysine methylation sites, we initially used Gene Ontology enrichments to identify biological processes enriched in lysine-methylated proteins. To functionally annotate clusters of interacting proteins within the predicted methyllysine interactome, we used the spatial analysis of functional enrichment (SAFE) component of Cytoscape (v.3.5.1) (Baryshnikova, 2016) using STRING interactions (v.10.5). Functional enrichments based on known protein interactions were carried out at recommended settings.

RESULTS

Demonstrating effectiveness of prediction framework

The achievable prediction recall, precision, and specificity are presented in Figure 2A as a function of decision threshold. As with previous studies, those lysine residues appearing on proteins that have been investigated for methylation, but which have not been reported to be methylated, are here assumed to be negative when training and evaluating predictors. Considering that the number of methylation sites continues to grow significantly, this assumption is known to be flawed (i.e. many of the assumed-negative instances are expected to actually be undocumented positive sites). This leads to a pessimistic estimation of the precision of the obtained model. Therefore, we also computed the precision using a high confidence negative test subset (see Supplemental Methods). Shown as yellow in Figure 2A, this can be considered an optimistic estimator of prediction precision, with the true precision expected to lay between the yellow and grey curves.

The model is subsequently evaluated in the hold-out test set, and the performance is contrasted with other available methylation prediction servers (Figure 2B). In general, the performance of all the methods is very low, which could be a reflection of

the limited training data used to create most of the other servers and the erroneous information of assumed-negative instances that are supplied to the training algorithms. Our method achieved significantly better performance in identifying methylated sites as is shown by our much higher sensitivity. The precision is slightly higher than other predictors which mean that overall, we are able to predict more positive sites than the other methods without sacrificing the false positive prediction rate.

Validation of histone lysine methylation sites

Given effective enrichment methods for the isolation and purification of histone proteins, we chose to validate the methylation status of positively predicted lysine methylation sites in histone proteins. MRM-MS was carried out on purified histone proteins using transitions that were designed for the detection of specific methylation sites. It should be noted that given the high lysine content within histones, it was not possible to validate all predicted methylation sites from trypsin-digested peptides as some sites exist on peptides that are simply too short for proper detection and site-specific identification. Within histone proteins, a total of 74 lysine methylation sites were predicted (Table S4). Given that histone proteins are rich in lysine residues susceptible to trypsin cleavage, from these peptides, only 57 methylation sites were identified to exist on trypsin-digested peptides that we deemed suitable for detection on the QTRAP 4000 MS as determined by the Skyline software. Of these peptides, transitions were selected and optimized for the detection of either the unmethylated or the Kme1, 2, or 3 methyl-modified lysine residues. A total of 51 new histone methylation sites containing 81 different methyl-modifications were successfully validated by MRM-MS and are listed in Table 1. Remarkably, 89% of the sites were found to be actually positive cases of lysine methylation, which outperforms the expected precision and is, therefore, a corroboration of the bias introduced in the model by mislabeled instances assumed to be negative.

DNA damage response of histone H2B(K43) methylation

Given the proximity of the histone H2B(K43) methylation to bound DNA, and a known role of H2B during repair of DNA damage (Hung et al., 2017), we explored the

dynamics of H2B(K43) methylation in response to doxorubicin-induced DNA damage (Figure 3). Histone methylation sites with a known response to periods of DNA damage, specifically histone H3(K4me3) and H3(K9me3), were also included in the analysis to provide a broader scope of analysis (Sun et al., 2009; Faucher et al., 2010; Ayrapetov et al., 2014). Relative methylation status of histone H2B(K43me2 and 3) were found to decrease in response to increasing doxorubicin concentrations following 24 hr treatment (Figure 3C). In contrast, the methylation status of histone H3(K4me3) and H3(K9me3) both dynamically increased in response to increasing concentration of doxorubicin treatment, corroborating with previous studies (Figure 3C). These findings suggest a dynamic response of a previously undocumented H2B methylation site in response to DNA damage.

Prediction of the human methyllysine proteome

To provide insight into the potential scale of the predicted methyllysine proteome, we used our framework to identify proteins harbouring high confidence lysine methylation sites throughout the whole human proteome. A prediction score of 0.7 was chosen for threshold used for the predicted methyllysine sites as this score corresponds to a 95% specificity of the MethylSight algorithm (Figure 2A). A total of 35,973 lysine residues were predicted to be methylated at this threshold; all predicted lysine methylation sites identified within the human proteome are listed in Table S5.

To provide deeper information into the potential biological functions of lysine methylation, the STRING database was used to identify and build networked clusters of interacting proteins that contain predicted methylation sites (Figure 4C). The cellular function of clusters was identified based on the GO enrichment analysis. Results indicated that predicted lysine methylation events were significantly enriched in the regulation of complement activation, positive regulation of the immune response, endonucleolytic cleavage of tricistronic rRNA transcripts, amino acid metabolism, nuclear-transcribed mRNA catabolism, calcium-independent cell-cell adhesion, nuclear protein export, intracellular protein transport, regulation of GTPase mediated signal transduction, among other histone-related biological processes. The VEGF signalling was used as a well-studied GTPase mediated signal transduction example to map

both known (black) and predicted (red) lysine methylation events that may play a role in its regulation (Figure 5A). Predicted NOS2 methylation events were chosen for MRM-MS validation, given the role of NOS2 in nitric oxide production in angiogenesis and hypoxia adaptation.

Validation of NOS2 lysine methylation and hypoxia response

Next, we validated the predicted NOS2 lysine methylation events from NOS2 IP samples obtained from MCF7 cells. The site-specific methylation status of NOS2 at lysine residues K12, K520, and K531 sites was determined in a manner similar as described above using MRM-MS. Although methylation at K12 and K531 could not be detected, the monomethylation of K520 was positively identified as a validated methylation site from MCF7 cells (Figure 5B). Neither the dimethylated or trimethylated state of NOS2(K520) were detected by MRM-MS. As the K520 methylation site is within the calmodulin binding region of NOS2, we then examined the effect of hypoxia on NOS2(K520me1) methylation status. In response to 24hr of 1% oxygen, relative NOS2(K520me1) levels decreased to only 47% of normoxic (i.e., 20% oxygen) levels (Figure 5C).

DISCUSSION

Traditionally, the disease context of lysine methylation has mostly been viewed via its roles in epigenetics, where the aetiology invariably stems from dysregulated histone-dependent transcription programs. Apart from the contribution of histone methylation events, a growing number of lysine-methylated non-histone proteins are being found to directly contribute to cellular dysfunction. For example, the discovery of MAP3K2 methylation at K260 by SMYD3 was shown to be instrumental in the activation of oncogenic Ras/Raf/MEK/ERK signalling and the progression of Ras-driven cancers (Mazur et al., 2014). This example highlights the importance of developing tools that are able to successfully identify new lysine methylation sites for their functional annotation in human health and disease. Indeed, a remarkable amount of atten-

tion has been drawn to the analysis and discovery of non-histone lysine methylation events.

Though many efforts have been devoted to the investigation of protein methylation, the analysis of non-histone methylation at proteome level is still a great challenge. The discovery and mechanistic insight into new lysine modifications will undoubtedly pave the way for the future development and therapeutic application of "epidrugs" in cancer. However, the alteration of protein/peptide physicochemical properties caused by methylation is very small and it is difficult to develop highly efficient enrichment approaches to separate the methylated peptides from the pool of diverse background peptides (Wu et al., 2017). The MethylSight program was developed to help in the efficient discovery of new lysine methylation sites that can then be validated through targeted mass spectrometry.

Currently, state-of-the-art methods for the prediction of post-translational lysine methylation do not provide adequate specificity for the efficient discovery of new *in vivo* methylation events. The AutoMotif server was the first prediction tool for methylation (Plewczynski et al., 2005). Methylated sites with 9 flanking residues were used as a positive dataset, while negative datasets were created using the unmodified corresponding sequences. These data were utilized to train an SVM classifier for the prediction of novel methylation sites. An improvement to this method was published later that year by Daily et al., who proposed that methylated events occur in disordered structures and incorporated this feature into their predictions thereby increasing accuracy (Daily et al., 2005). In the years following, several other prediction algorithms have been developed using an increasing number of features characteristic of known methyllysine sites (such as solvent-accessible surface area and secondary structure). However, these *in silico* approaches require high quality, large methylation site databases using experimentally validated modification sites as positive datasets, a resource which remains elusive. Given the exceptional growth and availability of newly validated lysine methylation sites, we used fully-alignment-free features, which are able to encode structural information from the lysine sites, to train the MethylSight algorithm, a highly accurate SVM-based prediction tool.

A total of 51 new histone methylation sites containing 81 different methyl-modifications were successfully validated by MethylSight (Table 1). To demonstrate the applicability of MethylSight to uncover methylation sites with possible functional implications. Interestingly, analysis of the histone H2B crystal structure (PDB 1AOI) identified the K43 methylation site within 5 angstroms to bound DNA (Figure 3A). Using antibodies designed specifically for the methylated form of histone H2B(K43me2/3), we monitored the response of this methylation to periods of doxorubicin-induced DNA damage (Hung et al., 2017). Indeed, the relative methylation of histone H2B(K43me3) was found to be a response to DNA damage in a doxorubicin concentration-dependent manner (Figure 3C). Previous studies have shown that the H2B(K43) site is also ubiquitinated and has also been shown to have an acetylated variant (Vlaming et al., 2014). The contribution of H2B(K43) methylation to the DNA damage response not directly known at this point, however, H2B is known to be globally ubiquitinated at multiple sites in response to DNA damage (Hung et al., 2017). Specifically, H2B(K123Ub) by Bre1/Rad6 helps to direct DOT1 methylation on H3(K4) methylation. The crosstalk between H2B ubiquitination and H3(K4) and H3(K79) methylation is evolutionarily conserved from yeast to metazoans. Since many other chromatin proteins are also subject to ubiquitination, an important question is which molecular features of ubiquitinated H2B are important for this trans-histone crosstalk *in vivo*. It is possible that H2B(K43me3) could also represent a modification helping to direct site-specific PTM competition between lysine modifications such as Ub, acetylation, and methylation during periods of DNA damage.

To facilitate the high-throughput *in silico* prediction of methylation sites on a proteomic scale, we used MethylSight to screen the complete human proteome from the UniProtKB/Swiss-Prot database (version 2017_07). Our analysis predicted 35,973 methyllysine sites (Table S5). To gain functional insight into the predicted human methyllysine protein network, we used a spatial analysis of functional enrichment (SAFE) (Baryshnikova, 2016) (Figure 4). SAFE was developed as a systematic method for annotating biological networks and examining their functional organization. Our analysis identified our methyllysine network to be enriched in the regulation of complement activation, positive regulation of the immune response, endonucleolytic

cleavage of tricistronic rRNA transcripts, amino acid metabolism, nuclear-transcribed mRNA catabolism, calcium-independent cell-cell adhesion, nuclear protein export, intracellular protein transport, regulation of GTPase mediated signal transduction, among other well-studied histone-related biological processes. This analysis agrees with reports demonstrating a role for lysine methylation in the nuclear localization of heat shock proteins (Cho et al., 2012), calcium signalling events mediated by calmodulin methylation (Haziza et al., 2015), and the regulation of Ras/Raf/MEK/ERK signaling through the methylation of MAP3K2 (Mazur et al., 2014). Indeed such proteome-wide analyses represent a valuable resource for the experimental validation of novel methylation substrates and generation of useful hypotheses.

Given the recent implication of lysine methylation on several examples of GTPase mediated signal transduction, including MAP3K2(K260) and VEGFR1(K831) methylation, we mapped the known and predicted lysine-methylated sites to the VEGFR signal transduction pathway to provide new insight into potential regulation by post-translational lysine methylation (Figure 5A). Indeed MethylSight identified potential methylation site on a number of proteins with direct regulation influence on signalling, including several additional sites on proteins previously known to be lysine-methylated such as VEGFR1 and the guanidine exchange factor, SOS1. To demonstrate the ability of MethylSight to identify methylated sites on non-histone proteins, several predicted sites on NOS2 were selected for MRM-MS based validation. The NOS2 protein was selected for validation given its biologically relevant role in angiogenesis and hypoxia adaptation (Heinecke et al., 2014). Monomethylation at the MethylSight predicted NOS2(K520) site was detected from NOS2 IP samples obtained from MCF7 cells by MRM-MS (Figure 5B). Given that this new methyllysine modified residue is within the calmodulin binding region of NOS2, a region critical for NOS2 function and nitric oxide production, we explore a possible hypoxia-responsive regulation of this methylation site. Indeed, in response to 24hr hypoxia relative monomethylation levels decreased to 47% of normoxic control levels (Figure 5C). These results indicate a possible role of NOS2 methylation in the regulation of its hypoxia-responsive activity, likely dictated by calmodulin binding.

The advances in analyses of lysine methylation at proteome level have been slow compared with other well studied PTMs, such as serine and threonine phosphorylation. Fortunately, progress in this field has been achieved along with advances in its identification technology. Exploiting the recent expansion of publicly available methyllysine datasets, and our combination of *in silico* and wet-lab experiments, we were able to develop and use the MethylSight pipeline to evaluate several new methylation sites. With the further development of novel analytical methods, in-depth exploration of protein lysine methylation can be achieved more easily using *in silico* prediction tools (e.g., MethylSight) that contribute to the deeper understanding of how protein methylation regulates diverse cellular processes.

ACKNOWLEDGEMENTS

This work was supported by a National Science and Engineering Research Council (NSERC) Canada Discovery grants to K.K. Biggar and J.R. Green, and a Canadian Institutes of Health Research (CIHR) grant to S.S.C. Li.

AUTHOR CONTRIBUTIONS

KKB, SSCL and JRG conceived the study. KKB, YRB, FC and JC carried out all validation experiments. KKB, YBRB, and JRG prepared the sequence data. KF and HA carried out all hypoxia experiments, while QF carried out all DNA damage experiments. KKB, YBRB JRG wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Arrowsmith, C.H., Bountra, C., Fish, P.V., Lee, K., Schapira, M. (2012) Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.* 11, 384–400.

Audagnotto, M., Dal Peraro, M. (2017). Protein post-translational modifications: In silico prediction tools and molecular modelling. *Comput. Struct. Biotech. J.* 15, 307-319.

Ayrapetov, M.K., Gursoy-Yuzugullu, O., Xu, C., Xu, Y., Price, B.D. (2014). DNA double-strand breaks promote methylation of histone H3 on lysine 9 and transient formation of repressive chromatin. *Proc. Natl. Acad. Sci, USA* 111(25), 9169-9174.

Baryshnikova, A. (2016). Systematic functional annotation and visualization of biological networks. *Cell Syst.* 2(6), 412-421.

Biggar, K.K., Li, S.S.C. (2015) Non-histone protein methylation as a regulator of cellular signalling and function. *Nat. Rev. Mol. Cell Biol.* 16, 5–17.

Can, F., Cherkassky, V. (2012). Generalized SMO algorithm for SVM-based multitask learning. IEEE Trans. Neural. Netw. Learn. Syst. 23(6), 997-1003.

Cao, X.J., Garcia, B. (2016). Global proteomics analysis of protein lysine methylation. *Curr. Protoc. Protein Sci*. 86, 24.8.1-24.8.19.

Carlson, S.M., Moore, K.E., Green, E.M., Martín, G.M., Gozani, O. (2014) Proteome-wide enrichment of proteins modified by lysine methylation. *Nat. Protoc.* 9, 37–50.

Chen, H., Xue, Y., Huang, N., Yao, X., & Sun, Z. (2006). MeMo: a web tool for prediction of protein methylation modifications. *Nucl. Acids Res. 34*(Web Server), W249–W253.

Cho, HS., Shimazu, T., Toyokawa, G., Daigo, Y., Maehara, Y., Hayami, S., Ito, A., Masuda, K., Ikawa, N., Field, H.I., Tsuchiya, E., Ohnuma, S., Ponder, B.A., Yoshida, M., Nakamura, Y., Hamamoto, R. Enhanced HSP70 lysine methylation promotes proliferation of cancer cells through activation of Aurora kinase B. *Nat. Comm.* 3, 1072 (2012).

Daily, K.M., Radivojac, P., Dunker, A.K. (2005). Intrinsic disorder and protein modifications: building an SVM predictor for methylation. IEEE Symposium on CIBCB, San Diego, California, 475–481.

Faucher, D., Wellinger, R.J. (2010). Methylated H3K4, a transcription-associated histone modification, is involved in the DNA damage response pathway. *PLoS Genet*. 6(8), e1001082.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations* 11(1), 10-18.

Hamamoto, R., Saloura, V., Nakamura, Y. (2015) Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nat. Rev. Cancer* 15, 110–124.

Haziza, S., Magnani, R., Lan, D., Keinan, O., Saada, A., Hershkovitz, E., Yanay, N., Cohen, Y., Nevo, Y., Houtz, R.L., Sheffield, V.C., Golan, H., Parvari, R. (2015). Calmodulin methyltransferase is required for growth, muscle strength, somatosensory development and brain function. *PLoS Genet*. 11(8), e1005388.

Heinecke, J.L., Ridnour, L.A., Cheng, R.Y.S., Switzer, C.H., Lizardo, M.M., Kanna, C., Glynn, S.A., Hussain, S.P., Young, H.A., Ambs, S., Wink, D.A. (2014). Tumor microenvironment-based feed-forward regulation of NOS2 in breast cancer progression. *Proc. Natl. Acad. Sci. USA* 111(17), 6323-6328.

Hu, L.L., Li, Z., Wang, K., Niu, S., Shi, X.H., Li, H.P., Cai, Y.D. (2011). Prediction and analysis of protein methylarginine and methyllysine based on multi sequence features. *Biopolymers* 95(11), 763-771.

Hung, S.H., Wong, R.P., Ulrich, H.D., Kao, C.F. (2017). *Proc. Natl. Acad. Sci. USA* 114(11), E2205-E2214.

Kawashima, S., Kanehisa, M. (2000). AAindex: amino acid index database. *Nucl. Acids Res*. 28(1), 374.

Liu, H., Galka, M., Mori, E., Liu, X., Lin, Y., Wei, R., Pittock, P., Voss, C., Dhami, G., Li, X., Miyaji, M., Lajoie, G., Chen, B., Li, S.S. (2013) A Method for systematic

mapping of protein lysine methylation identifies functions for HP1β in DNA damage response. *Mol. Cell* 50, 723–735.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tab, D.L., Liebler, D.C., MacCoss, M.J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26(7), 966-968.

Mann, M., Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 21(3), 255-261.

Martin, C., Zhang, Y. (2005). The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell. Biol.* 6(11), 838-849.

Mazur, P.K., Reynoird, N., Khatri, P., Jansen, P.W.T.C., Wilkinson, A.W., Liu, S., Barbash, O., Van Aller, G.S., Huddleston, M., Dhanak, D., Tummino, P.J., Kruger, R.G., Garcia, B.A., Butte, A.J., Vermeulen, M., Sage, J., Gozani, O. (2014) SMYD3 links lysine methylation of MAP3K2 to Ras-driven cancer. *Nature* 510, 283–287.

Plewczynski, D., Tkacz, A., Wyrwicz, L.S., Rychlewski, L. (2005). AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics* 21(10), 2525-2527.

Qiu, W.R., Xiao, X., Lin, W.Z., Chou, K.C. (2014). iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *BioMed Research International*, *2014*, 947416.

Ruiz-Blanco, Y.B., Paz, W., Green, J., Marrero-Ponce, Y. (2015). ProtDCal: a program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics*. 16, 162.

Shao, J., Xu, D., Tsai, S.N., Wang, Y., Ngai, S.M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PloS One*, *4*(3), e4920.

Shi, S.P., Qiu, J.D., Sun, X.Y., Suo, S.B., Huang, S.Y., Liang, R.P. (2012). PMeS: Prediction of Methylation Sites Based on Enhanced Feature Encoding Scheme. *PLoS ONE*, *7*(6), e38772.

Shi, Y., Guo, Y., Hu, Y., Li, M. (2015). Position-specific prediction of methylation sites from sequence conservation based on information theory. *Scientific Reports*, *5*(1), 12403.

Shien, D.M., Lee, T.Y., Chang, W.C., Hsu, J. B.K., Horng, J.T., Hsu, P.C., Huang, H.D. (2009). Incorporating structural characteristics for identification of protein methylation sites. *J. Comp. Chem.*, *30*(9), 1532–1543.

Sun, Y., Jiang, X., Xu, Y., Ayrapetov, M.K., Moreau, L.A., Whetstine, J.R., Price, B.D. (2009). Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60. *Nat. Cell. Biol*. 11(11), 1376-1382.

Vlaming, H., van Welsen, T., de Graaf, E.L., Ontoso, D., Maarten Altelaar, A.F., San-Segundo, P.A., Heck, A.J.R., van Leeuwen, F. (2014). Flexibility in crosstalk between H2B ubiquitination and H3 methylation in vivo. *EMBO Rep*. 15(10), 1077-1084.

Wen, P.P., Shi, S.P., Xu, H.D., Wang, L.N., Qiu, J.D. (2016). Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics* 32(20), 3107-3115.

We, Z., Connolly, J., Biggar, K.K. (2017). Beyond histones - the expanding roles of protein lysine methylation. *FEBS J*. 284(17), 2732-2744.

West, L.E., Gozani, O. (2011) Regulation of p53 function by lysine methylation. *Epigenomics* 3, 361–369.

Zhang, X., Wen, H., Shi, X. (2012) Lysine methylation: beyond histones. *Acta Biochim Biophys Sin* 44, 14–27.

FIGURES LEGENDS

Figure 1. Overview of the MethylSight program.

Figure 2. Prediction framework validation. (A) Performance measures of the MethylSight algorithm. The grey and yellow lines are pessimistic and optimistic estimators for true precision (see text), such that the true precision falls between these lines. (B) Comparison of precision and sensitivity of publicly available methyllysine predictors to the MethylSight algorithm.

Figure 3. Characterization of the histone H2B(K43) methylation site. (A) Crystal structure of histone H2B in complex with bound DNA (PDB 1AOI) demonstrating the proximity of the K43 methylation site to bound DNA. Interactions modeled with PyMol. (B) Representative transitions used for the methylation state-specific detection of H2B(K43) methylation patterns for validation by MRM-MS. All mass spectrometry was carried out using an AB SCIEX QTRAP 4000 linear ion trap mass spectrometer. (C) Immunoblot images showing the dynamic H3(K9me3), H3(K4me3), total H2B protein, H2B(K43me2), and H2B(K43me3) methylation in response to periods of doxorubicin-induced DNA damage.

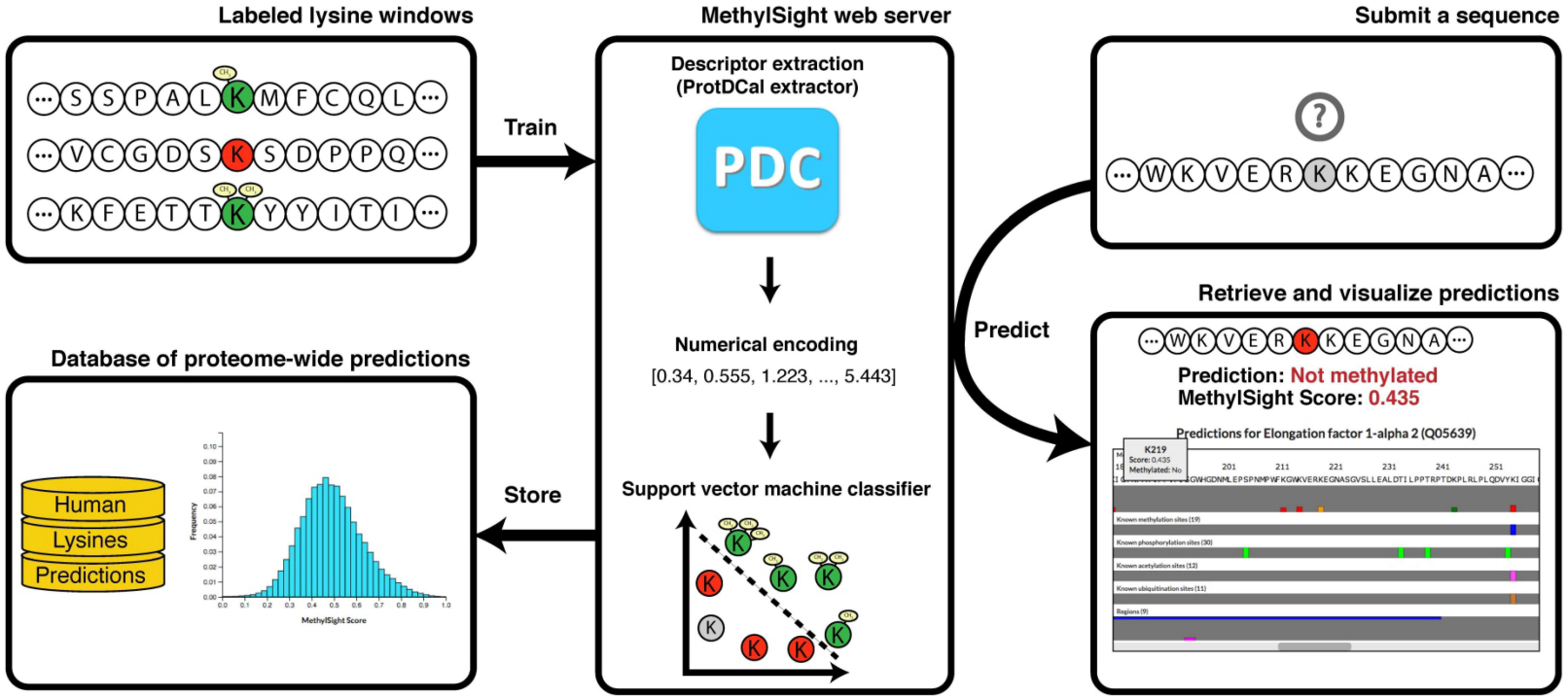
Figure 4. The predicted functional Human methyllysine proteome network. Human Gene Ontology (GO) enrichments of biological processes for both the (A) currently annotated, and (B) predicted methyllysine proteome. The currently annotated methyllysine-modified proteins were downloaded from the PhosphoSitePlus database. (C) To obtain a functional annotation of interacting proteins to reveal functionally relevant groups of lysine methylated proteins, a spatial analysis of functional enrichment was used and visualized using Cytoscape (v.3.5.1). All clusters are differentially coloured and annotated used human GO terms.

Figure 5. Role of lysine methylation on non-histone proteins. (A) known (black) and predicted (red) methylation sites within proteins involved in VEGFR1/2 signal transduction. (B) Representative transitions used for the methylation state-specific detection of NOS2(K520) monomethylation patterns for validation by MRM-MS. (C) Dynamic methylated status of NOS2(K520) in response to a 24hr hypoxic (1% oxygen) environment as monitored by relative peak area of selected transition ions specific for the NOS2(K520me1) peptide. Data normalized to the relative intensity of an internal NOS2 peptide to control for changes in protein expression. All mass spectrometry was carried out using an AB SCIEX QTRAP 4000 linear ion trap mass spectrometer.

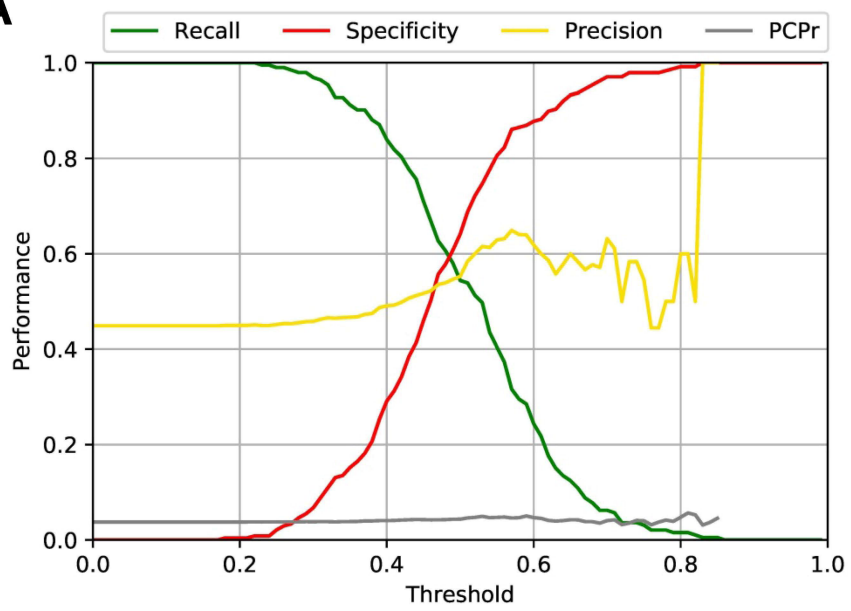
Table 1. MethylSight predicted lysine methylation sites validated by multiple reaction monitoring mass spectrometry. All mass spectrometry was carried out using an AB SCIEX QTRAP 4000 linear ion trap mass spectrometer.

				Methylation state		
Uniprot ID	**Protein**	**Score**	**Site**	**Mono-**	**Di-**	**Tri-**
P07305	H1F0	0.948	171		✓	✓
		0.94	135	✓		
		0.937	158	✓	✓	
		0.937	138		✓	✓
		0.937	136		✓	✓
		0.928	143		✓	✓
		0.909	121			✓
Q8IZA3	H1FOO	0.932	188	✓	✓	✓
		0.921	163	✓		✓
		0.92	171	✓	✓	
		0.91	173	✓	✓	
		0.9	215		✓	
		0.894	263	✓		
		0.88	138	✓	✓	✓
		0.872	152		✓	
		0.847	297		✓	
Q92522	H1FX	0.927	194	✓	✓	
		0.919	184			✓
		0.912	165	✓		✓
		0.91	173		✓	✓
		0.903	178		✓	✓
		0.875	145	✓		
O75367	H2AFY	0.863	7	✓		
P16401	HIST1H1B	0.96	160	✓	✓	✓
		0.956	132	✓	✓	✓
		0.955	206	✓	✓	✓
		0.951	193	✓	✓	

		0.949	167	✓		
		0.947	140	✓		
		0.931	121		✓	
		0.901	111	✓		
		0.881	99		✓	✓
		0.841	77		✓	
		0.826	66	✓		
P16402	HIST1H1D	0.942	206		✓	
		0.931	149			✓
		0.927	160	✓	✓	✓
		0.924	140			✓
		0.903	136			✓
		0.899	119			✓
P10412	HIST1H1E	0.952	129	✓		
		0.95	148			✓
		0.945	196	✓		
		0.942	139			✓
		0.909	116	✓	✓	✓
		0.871	89		✓	✓
		0.85	33	✓		
Q96A08	HIST1H2BA	0.845	13	✓		
O60814	HIST1H2BK	0.724	43		✓	✓
Q6FI13	HIST2H2AA4	0.859	5	✓		✓
Q5QNW6	HIST2H2BF	0.884	34	✓		


Labeled lysine windows
Train
MethylSight web server
Descriptor extraction
(ProtDCal extractor)
PDC
Numerical encoding
[0.34, 0.555, 1.223, ..., 5.443]
Support vector machine classifier
Store
Database of proteome-wide predictions
Human
Lysines
Predictions
Frequency
MethylSight Score
Predict
Submit a sequence
Retrieve and visualize predictions
Prediction: Not methylated
MethylSight Score: 0.435
Predictions for Elongation factor 1-alpha 2 (Q05639)
K219
Score: 0.435
Methylated: No
Known methylation sites (19)
Known phosphorylation sites (30)
Known acetylation sites (12)
Known ubiquitination sites (11)
Regions (9)

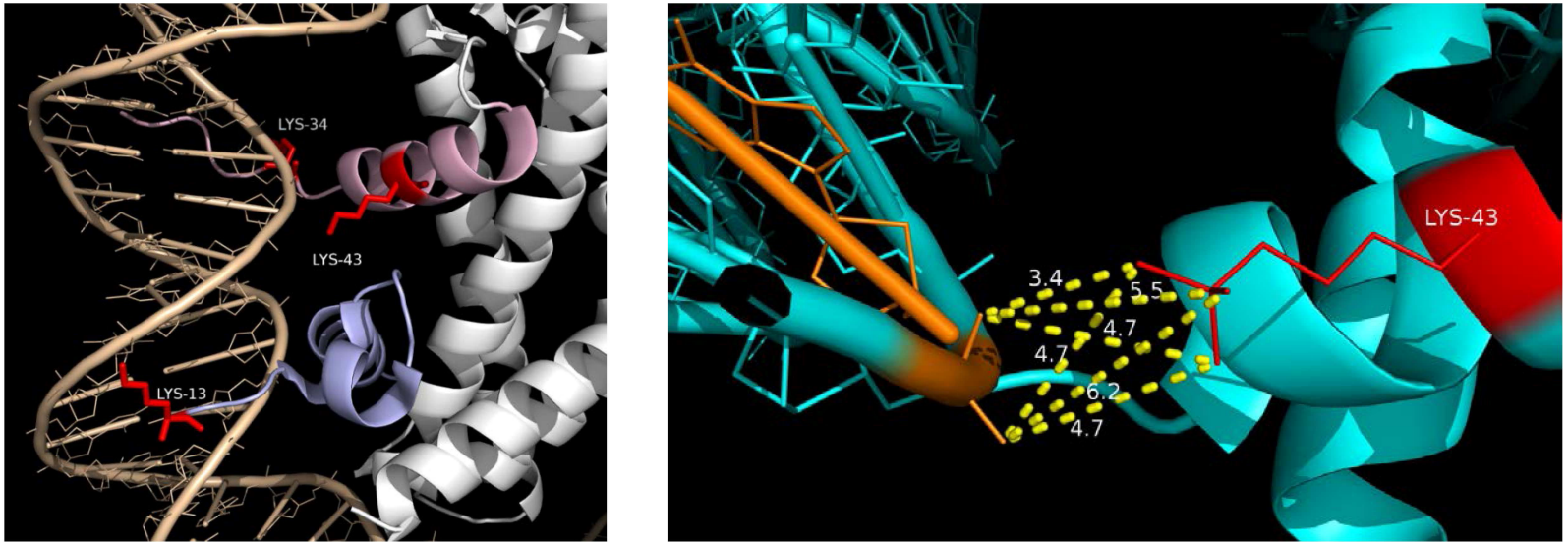
A



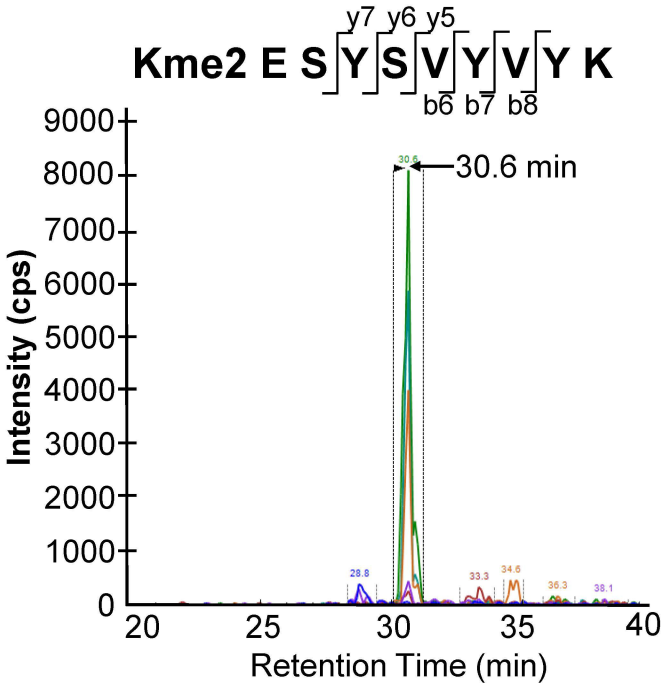
B

METHODS	POSITIVE MATCHES
BPB-PPMS	23.2%
MEMO	16.2%
IMETHYL-PSEAAC	32.6%
MethylSight	86.0%

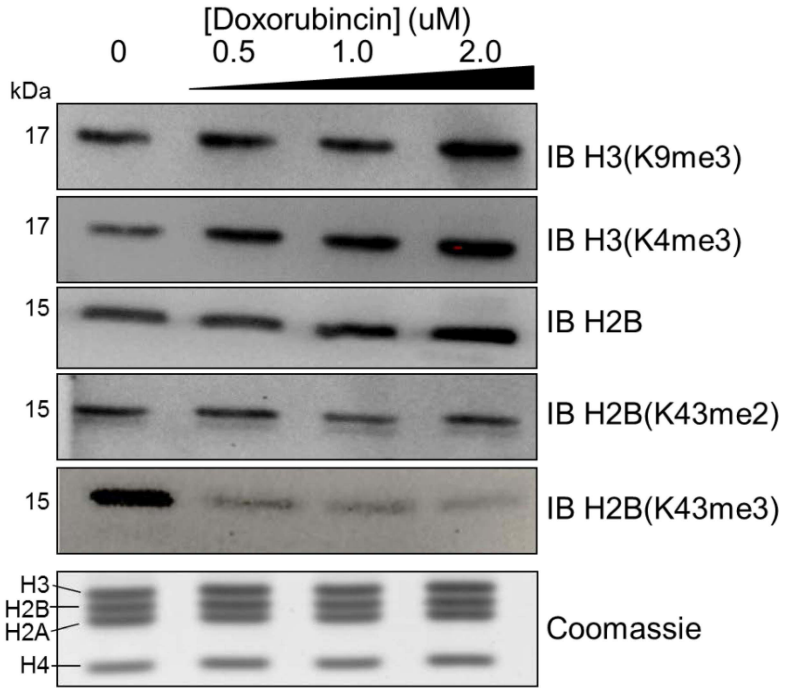
A

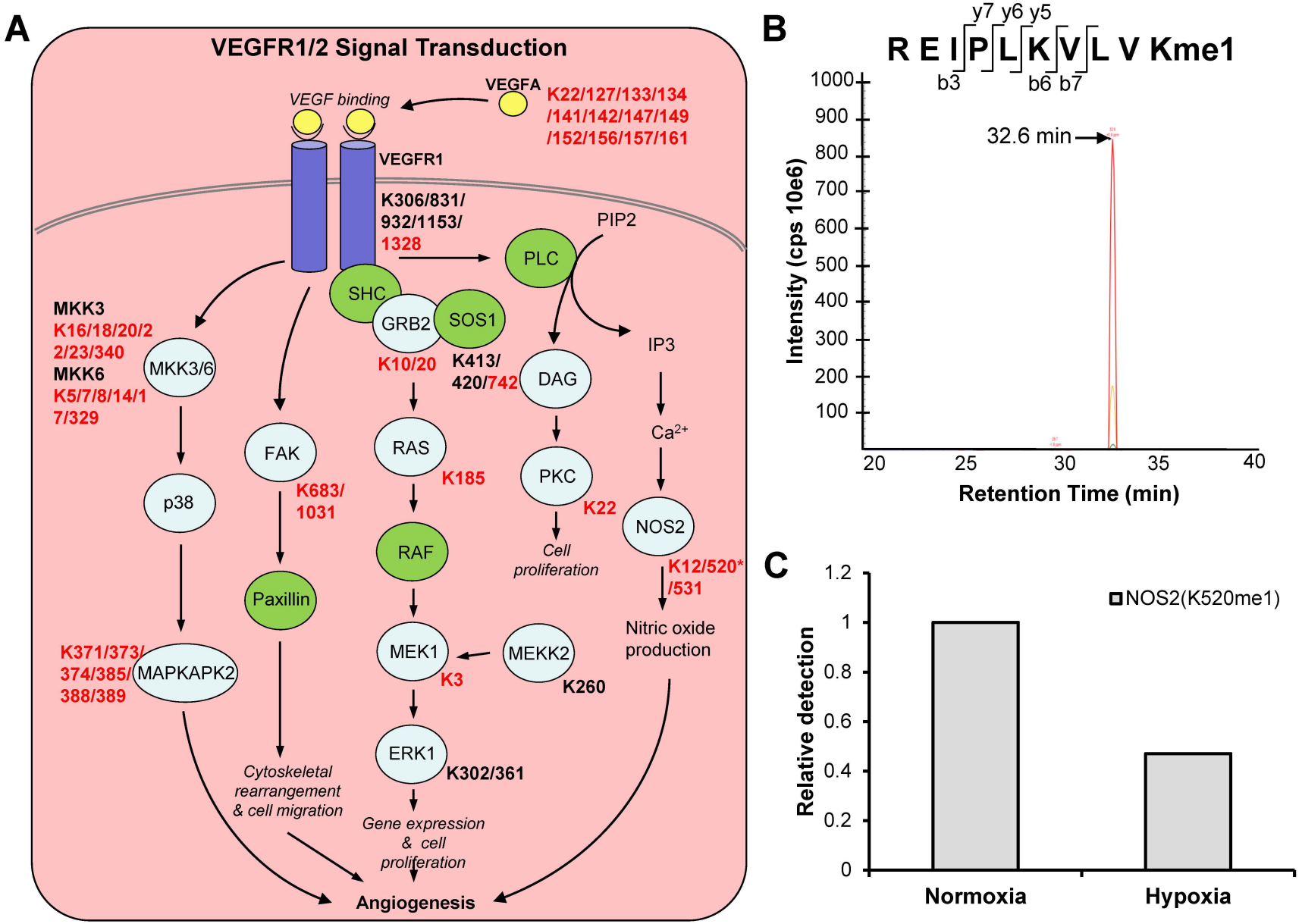

LYS-34
LYS-43
LYS-13
3.4
5.5
4.7
4.7
6.2
4.7
LYS-43

B

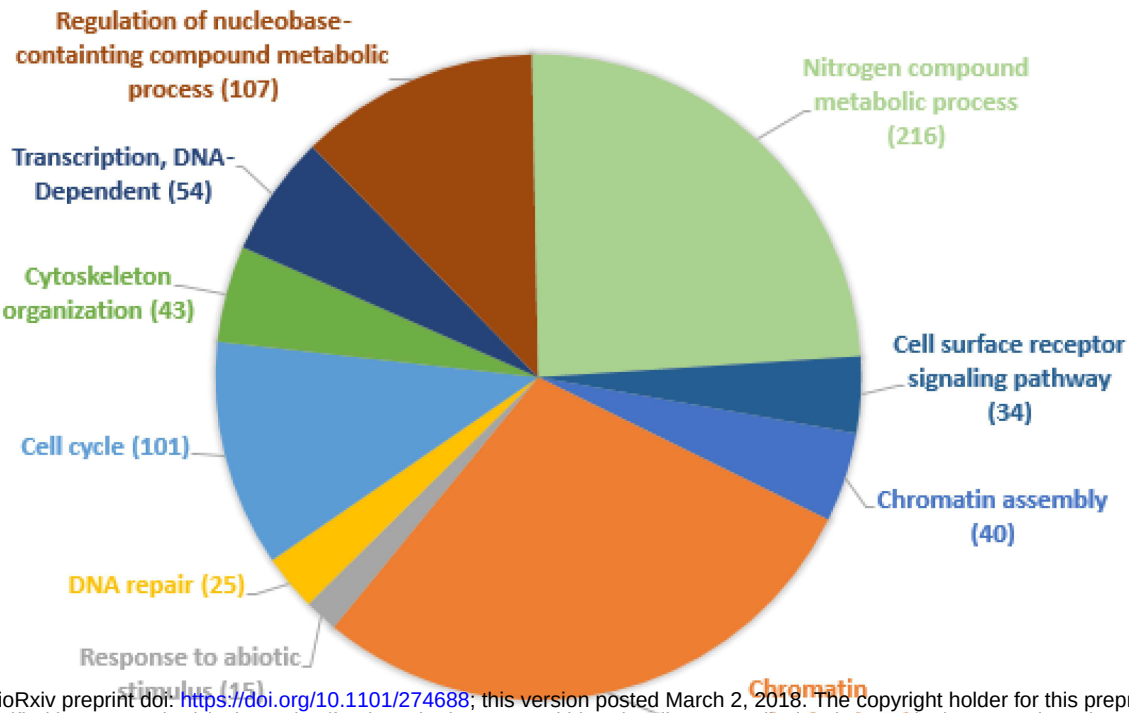

y7 y6 y5
Kme2 E S Y S V Y V Y K
b6 b7 b8
30.6 min
Intensity (cps)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
20
25
30
35
40
Retention Time (min)

C

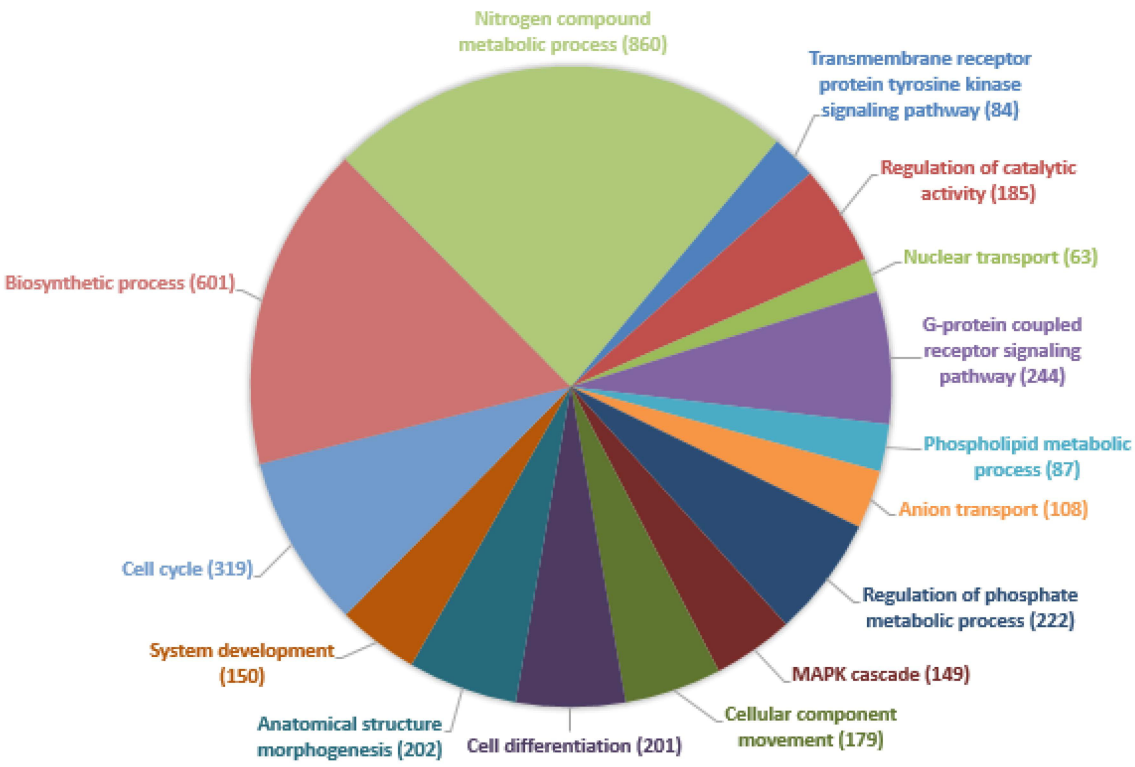

[Doxorubincin] (uM)
0
0.5
1.0
2.0
kDa
17
IB H3(K9me3)
17
IB H3(K4me3)
15
IB H2B
15
IB H2B(K43me2)
15
IB H2B(K43me3)
H3
H2B
H2A
H4
Coomassie


A
VEGFR1/2 Signal Transduction
VEGF binding
VEGFA
K22/127/133/134
/141/142/147/149
/152/156/157/161
VEGFR1
K306/831/
932/1153/
1328
PIP2
PLC
SHC
GRB2
SOS1
K10/20
K413/
420/742
IP3
DAG
MKK3
K16/18/20/2
2/23/340
MKK6
K5/7/8/14/1
7/329
MKK3/6
FAK
K683/
1031
RAS
K185
PKC
K22
Ca2+
NOS2
p38
RAF
Cell
proliferation
K12/520*
/531
Paxillin
Nitric oxide
production
K371/373/
374/385/
388/389
MAPKAPK2
MEK1
K3
MEKK2
K260
ERK1
K302/361
Cytoskeletal
rearrangement
& cell migration
Gene expression
& cell
proliferation
Angiogenesis
B
y7 y6 y5
R E I P L K V L V Kme1
b3
b6 b7
1000
900
800
700
600
500
400
300
200
100
Intensity (cps 10e6)
32.6 min
20
25
30
35
40
Retention Time (min)
C
1.2
1
0.8
0.6
0.4
0.2
0
Relative detection
NOS2(K520me1)
Normoxia
Hypoxia

A



B



C