Phospho-islands and the evolution of phosphorylated amino acids in mammals

Mikhail Moldovan Corresp., 1 , **Mikhail S Gelfand** 1, 2

[1] Skolkovo Institute of Science and Technology, Moscow, Russia

[2] A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia

Corresponding Author: Mikhail Moldovan
Email address: mika.moldovan@gmail.com

Background

Protein phosphorylation is the best studied post-translational modification strongly influencing protein function. Phosphorylated amino acids not only differ in physico-chemical properties from non-phosphorylated counterparts, but also exhibit different evolutionary patterns, tending to mutate to and originate from negatively charged amino acids. The distribution of phosphosites along protein sequences is non-uniform, as phosphosites tend to cluster, forming so-called phospho-islands.

Methods

Here, we have developed an HMM-based procedure for the identification of phospho-islands and studied the properties of the obtained phosphorylation clusters. To check robustness of evolutionary analysis, we consider different models for the reconstructions of ancestral phosphorylation states.

Results

Clustered phosphosites differ from individual phosphosites in several functional and evolutionary aspects including underrepresentation of phosphotyrosines, higher conservation, more frequent mutations to negatively charged amino acids. The spectrum of tissues, frequencies of specific phosphorylation contexts, and mutational patterns observed near clustered sites also are different.

Phospho-islands and the evolution of phosphorylated amino acids in mammals

Mikhail A. Moldovan[1], **Mikhail S. Gelfand**[1,2]

[1] Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, Moscow, Russia 121205

[2] A.A.Kharkevich Institute for Information Transmission Problems (RAS), Bolshoy Karetny Per. 19, bld.1, Moscow, Russia 127051

Corresponding Author:
Mikhail A. Moldovan[1]
Bolshoy Boulevard 30, bld. 1, Moscow, 121205, Russia
Email address: mika.moldovan@gmail.com

Abstract

Background

Protein phosphorylation is the best studied post-translational modification strongly influencing protein function. Phosphorylated amino acids not only differ in physico-chemical properties from non-phosphorylated counterparts, but also exhibit different evolutionary patterns, tending to mutate to and originate from negatively charged amino acids. The distribution of phosphosites along protein sequences is non-uniform, as phosphosites tend to cluster, forming so-called phospho-islands.

Methods

Here, we have developed an HMM-based procedure for the identification of phospho-islands and studied the properties of the obtained phosphorylation clusters. To check robustness of evolutionary analysis, we consider different models for the reconstructions of ancestral phosphorylation states.

Results

Clustered phosphosites differ from individual phosphosites in several functional and evolutionary aspects including underrepresentation of phosphotyrosines, higher conservation, more frequent mutations to negatively charged amino acids. The spectrum of tissues, frequencies of specific phosphorylation contexts, and mutational patterns observed near clustered sites also are different.

Introduction

Protein post-translational modifications (PTMs) are important for a living cell (Schweiger and Linial 2010; Kurmangaliyev et al. 2011; Studer et al. 2016; Huang et al. 2017). By changing physico-chemical properties of proteins, PTMs affect their function, often introducing novel biological features (Pearlman et al. 2011). To date, hundreds of thousands of PTMs in various organisms have been identified and various databases containing information about PTMs have been compiled (Ptacek and Snyder, 2006; Huang et al. 2017).

Protein phosphorylation is likely both the most common and the best studied PTM (Ptacek and Snyder, 2006; Schweiger and Linial, 2010; Huang et al. 2017). Phosphorylation - - introduces a negative charge and a large chemical group to the local protein structure, hence strongly affecting the protein conformation (Pearlman et al. 2011; Nishi et al. 2014). As a result, diverse cellular signaling pathways are based on sequential phosphorylation events (Moses and Landry 2010; Pearlman et al. 2011; Ardito et al. 2017). In eukaryotes, phosphorylation sites (phosphosites) are mainly represented by serines, threonines, and tyrosines (which we here refer to as STY amonoacids), with only a minor fraction involving other amino acids, such as histidine (Fuhs and Hunter 2017; Huang et al. 2017).

Phosphosites are overrepresented in disordered regions (DRs) of proteins, i.e. in regions devoid of secondary or tertiary structure, usually located on the surface of a protein globule (Iakoucheva 2004). Hence, studies of the evolution of phosphosites have mainly concentrated on sites located in DRs (Kurmangaliyev et al. 2011; Miao et al. 2018). In particular, it has been shown, that phosphosites tend to arise from negatively charged amino acids (NCAs) more frequently than their non-phosphorylated counterparts, and, in a number of cases, retain structural features initially maintained by NCAs (Kurmangaliyev et al. 2011; Miao et al. 2018). As phosphorylation is often highly conserved (Macek et al. 2007), experimental limitations on the number of model species with established phosphosites may be overcome in evolutionary studies by formally assigning phosphorylation labels to homologous sites (Kurmangaliyev et al. 2011; Huang et al. 2017). However, this approach requires a degree of caution when dealing with evolutionary trees of substantial depths, *e.g.* only a small fraction of yeast phosphosites are conserved between species separated by ~1400 My (million years), while about a half of phosphosites are conserved at a shorter time (~360 My) (Studer et al. 2016). At smaller distances, this method may be applied to infer some evolutionary properties of phosphosites, *e. g.* in *Drosophila* or in vertebrate species, phosphosites tend to mutate to NCA (Kurmangaliyev et al. 2011; Miao et al. 2018).

Phosphorylation can be both a constitutive modification and a way to transiently modify the protein function (Landry et al. 2014). In the former case, the change of a phosphosite to NCA should not cause a significant fitness reduction, as physico-chemical properties are not strongly affected, whereas in the latter case a mutation would have dire consequences (Moses and Landry 2010; Landry et al. 2014).

In proteins, phosphosites often form co-localized groups called phosphorylation islands or phosphorylation clusters, and about a half of phosphorylated serines and threonines are located in such clusters (Schweiger and Linial 2010). While individual phosphosites function as simple switches, phospho-islands are phosphorylated in a cooperative manner, so that the probability of a phosphorylation event at a focal site strongly depends on the phosphorylation of adjacent sites, and when the number of phosphorylated amino acids exceeds a threshold, the cumulative negative charge of the phosphate groups introduces functionally significant changes to the protein structure (Landry et al. 2014).

Accurate procedures for the identification of phosphosites and next-generation sequencing technologies yielded large numbers of well-annotated phosphosites (Altenhoff et al. 2017; Huang et al. 2017; UniProt 2018) enabling us to develop an accurate automatic procedure for the identification of phospho-site clusters we call phospho-islands. We show that clustered phosphosites exhibit evolutionary properties distinct from those of individual phosphosites, in particular, an enhanced mutation rate to NCA and altered mutational patterns of amino acids in the phosphosite vicinity. Our study complements earlier observations on the general evolutionary patterns in phosphosites with the analysis of mutations in non-serine phosphosites and the demonstration of differences in the evolution of clustered and individual phosphorylated residues.

Materials & Methods

Data

The phosphosite data for human, mouse and rat proteomes were downloaded from the iPTMnet database (Huang et al. 2017). The phosphorylation breadth values for the mouse dataset were obtained from (Huttlin et al. 2010). Human, mouse and rat proteomes were obtained from the UniProt database (UniProt 2018). Vertebrate orthologous gene groups

(OGGs) for human and mouse proteomes were downloaded from the OMA database (Altenhoff et al. 2017). Then, all paralogous sequences and all non-mammalian sequences were excluded from the obtained OGGs.

Alignments and Trees

We searched for homologous proteins in three proteomes with pairwise BLASTp alignments (Altschul et al. 1990). Pairs of proteins with highest scores were considered closest homologs. The information about closest homologs was subsequently used to predict HMR phosphosites. OGG were aligned by the ClustalO multiple protein alignment (Sievers et al. 2014) and, while the HMR phosphosites were identified based on Muscle pairwise protein alignments (Edgar 2004). The mammalian phylogenetic tree was obtained from Timetree (Kumar et al. 2017).

Phosphorylation retention upon mutations

After the identification of homologous protein pairs in human/mouse and mouse/rat proteomes and the proteome alignment construction, we identified homologous phosphosites as homologous STY resudues which were shown to be phosphorylated in both species. We have shown that phosphorylation is retained on S-T and T-S mutation by comparing two pairs of ratention probabilities (Fig. 1C): *p(pS-pS)* with *p(pS-pT | S)* and *p(pT-pT)* with *p(pS-pT | T)* (analogously for the phosphorylation of tyrosines), *p(pX-pX)* being defined as the fraction of X amino acids phosphorylated in both considered species:

$$p(pX-pX)=\frac{\#(pX-pX)}{\#(pX-pX)+\#(pX-X)}$$

and *p(pX*$_1$*-pX*$_2$*)*, as the fraction of phosphorylated X$_1$ residues in one species given that in another species another amino acid residue (X$_2$) is also phosphorylated:

$$p(pX_1-\ pX_2|X_1)=\frac{\#(pX_1-\ pX_2)}{\#(pX_1-\ pX_2)+\#(pX_1-X_2)}$$

$$p(pX_1-\ pX_2|X_2)=\frac{\#(pX_1-\ pX_2)}{\#(pX_1-\ pX_2)+\#(X_1-pX_2)}$$

Homologous phosphosite lists from the human/mouse and human/rat pairs were merged to produce HMR phosphosite list of human phosphosites.

False-positive rates of phosphorylation identification by homologous propagation

We assessed the quality of the phosphorylation prediction via homologous propagation approaches by counting false-positive rates of phosphosite predictions in species with large phosphosite lists. As the numbers of predicted phosphosites drastically differed between species (Huang et al. 2017), we considered multiway predictions in each case as characteristics of the procedure performance. Hence, considering mouse phosphosite predicted by homology with known human phosphosites, we also considered human phosphosites predicted based on known mouse phosphosites. The false-positive rate was assessed as the proportion of correctly predicted phosphosites among the STY aminoacids in one species homologous to phosphosite positions in other considered species.

When assessing the quality of phosphosite predictions based on phosphosites experimentally identified in at least two species, we considered human, mouse. and rat and the lists of phosphosites homologous between human and mouse and between human and rat. In these cases, predictions were made for rat and mouse, respectively with the false-positive rate assessed by the same approach as in the previous case.

Mutation matrices

To obtain single-aminoacid mutation matrices, we first reconstructed ancestral states with the PAML software (Yang 2007). For the reconstruction, we used OGG alignments which did not contain paralogs and pruned mammalian trees retaining only organisms contributing to corresponding OGG alignments. The alignment of both extant and reconstructed ancestral sequences and the corresponding trees were then used to construct mutation matrices, where we distinguished the phosphorylated and non-phosphorylated states of STY amino acids. Here, the phosphorylation state was assigned to STY amino acids using the phosphorylation propagation approach described above. When calculating the mutation matrix, we did not count mutations predicted to happen on branches leading from the root to first-order nodes, as PAML did not reconstruct them well without an outgroup (Koshi and Goldshtein 1996; Yang 2007). Tree pruning and calculating the mutation matrix count were implemented in *ad hoc* python scripts using functions from the ete3 python module.

Disordered regions and phospho-island prediction

Disordered regions were predicted with the PONDR VSL2 software with default parameters (Xue et al. 2010). Phosho-islands were predicted with the Viterbi algorithm (Viterbi 1967). Emission probabilities for the algorithm were obtained as the ratio of density values in the *S* distribution decomposition (likelihood ratio normalized to 1) (Fig. 2a). Transitional probabilities were set to 0.2 to maximize the likeness of obtained distribution of *S* within phospho-islands and the one predicted by the decomposition procedure (Fig. 2b).

Phosphosite contexts

We employed the list of phosphosite contexts as well as the binary decision-tree procedure to define the context of a given phosphosite from Villen et al. 2007. The procedure is as follows. (i) Proline context is assigned if there is a proline at position +1 relative to the phosphosite. (ii) Acidic context is assigned if there are five or six E/D amino acids at positions +1 to +6 relative to the phosphosite. (iii) Basic context is assigned if there is a R/K amino acid at position –3. (iv) Acidic context is assigned if there are D/E amino acids at any of positions +1, +2 or +3. (v) Basic context is assigned if there are at least two R/K amino acids at positions –6 to –1. Otherwise, no context is assigned and we denote this as the “O” (other) context. We consider tyrosine phosphosites separately and formally assign the with the “Y” (tyrosine) context.

Local mutation matrices

We computed local substitution matrices (LSM) as the substitution matrices for amino acids located within a frame with the radius *k* centered at a phosphorylated serine or threonine. When computing LSMs, we did not count mutations of or resulting in STY amino acids to exclude the effects introduced by the presence and abundance of phospho-islands. We have set *k* to 1, 3, 5, and 7 and selected 5 as for this value we observed the strongest effect, that is, obtained the largest number of mutations with frequencies statistically different from those for non-phosphorylated serines and threonines.

Statistics

When comparing frequencies, we used the χ^2 test if all values in the contingency matrix exceeded 20 and Fisher's exact test otherwise. To correct for multiple testing, we used the Bonferroni correction with the scaling factor set to 17 for the substitution vector comparison and to 17×17 for the comparison of substitution matrices with excluded STY amino acids. 95% two-tailed confidence intervals shown in figures were computed by the χ^2 or Fisher's exact test. The significance of obtained Pearson's correlation coefficients was assessed with the F-statistic.

Code availability

Ad hoc scripts were written in Python. Graphs were built using R. All scripts and data analysis protocols are available online at https://github.com/mikemoldovan/phosphosites.

Results

Conserved phosphosites

As protein phosphorylation in a vast majority of organisms has not been studied or has been studied rather poorly (Huang et al. 2017), the evolutionary analyses of phosphosites typically rely on the assumption of absolute conservation of the phosphorylation label assigned to STY amino acids on a considered tree (Kurmangaliyev et al. 2011; Miao et al. 2018). Thus, if, for instance, a serine is phosphorylated in human, we, following this approach, would consider any mutation in the homologous position of the type S-to-X to be a mutation of a phosphorylated serine to amino acid X (Fig. 1b). However, the comprehensive analysis of yeast phosphosites has shown low conservation of the phosphorylation label at the timescales of the order 100 My and more (Studer et al. 2016). Thus, we have considered only orthologous groups of mammalian proteins, present in the OMA database (Altenhoff et al. 2017). The mammalian phylogenetic tree is about 177 My deep (Kumar et al. 2017), which corresponds to about 50% of the phosphorylation loss in the 182 My-deep yeast *Saccharomyces-Lachancea* evolutionary path (Studer et al. 2016). The tree contains three organisms with well-studied phosphoproteomes: human (227834 sites), mouse (92943 sites), and rat (24466 sites) (Huang et al. 2017) (Fig. 1a).

Still, the expected 50% of mispredicted phosphosites could render an accurate evolutionary analysis impossible. This could be partially offset by considering phosphosites conserved in well-studied lineages. Thus, we compiled a set of human phosphosites homologous to residues phosphorylated also in mouse and/or rat, which we will further refer to as human-mouse/rat (HMR) phosphosites. The HMR set consists of 53437 sites covering 54.6% and 61.2% of known mouse and rat phosphosites, respectively, which is consistent with the above-mentioned observation about 50% phosphorylation loss in yeast on evolutionary distances similar to the ones between the human and rodent lineages (Fig. 1ab).

We consider the HMR set to be enriched in accurately predicted phosphosites. Indeed, by considering conserved phosphosites, we substantially reduce the number of mispredictions. If we simply propagated human phosphorylation labels to mouse and *vice versa* we would get about 77.6% and 42.3% of false positive labels, respectively. However, sites conserved between human and rat or sites conserved between rat and mouse would yield about twofold lesser percentages of 41.9% and 19.9% of false positives in mouse and human, respectively. The obtained percentages can be considered as upper estimates of false positive rates, as current experimental phosphosite coverage in mammals cannot guarantee the identification of all conserved phosphosites (Huang et al. 2017). Thus, the HMR dataset is sufficiently robust for the prediction of phosphorylation labels in less-studied mammalian lineages.

Treatment of STY amino acids homologous to phosphorylated ones as phosphorylated yields another possible caveat, stemming from the possible loss of phosphorylation upon STY-to-STY mutations. To assess this effect, we compared the probabilities of phosphosite retention upon pSTY-to-STY mutation, pSTY indicating the phosphorylated state, and the respective probabilities in the situation when a mutation has not occurred for a pair of species with well-established phosphosite lists, i.e. human and mouse (Fig. 1c). We have observed only a minor, insignificant decrease of the probabilities of the phosphorylation retention in the cases of pS-pT and pS-pY mismatches relative to the pT-pT states in mouse and human, indicating the general conservation of the phosphorylation label upon amino acid substitution. An interesting observation here is that the pS-pS states appear to be the most conserved ones (Fig. 1c). Taken together, these results indicate the evolutionary stability of phosphorylation states upon mutation.

The increased evolutionary robustness of the pS state relative to the pT and pY states should manifest as overrepresentation of phosphoserines among phosphosites with respect to non-phosphorylated amino acid positions. Thus, we assessed the relative abundancies of pSTY amino acids in the HMR dataset relative to the established human phosphosite set and to the set of non-phosphorylated STY amino acids. Serines and threonines, comprising the vast majority of the pSTY amino acids, are, respectively, over- and underrepresented in the phosphosite sets (Fig. 1cd). This effect is significantly more pronounced in the HMR dataset relative to the total human phosphosite dataset, further supporting the observation about lower conservation of pT relative to pS, as the HMR dataset is enriched in conserved phosphosites by design.

Phosphorylation islands

The distribution of distances between phosphosites is different from that of randomly chosen serines and threonines even accounting for the tendency of phosphosites to occur in disordered regions (DRs) (Schweiger and Linial 2010) (Fig. 2a). However, this observation depends on an arbitrary definition of phosphorylation islands as groups of phosphosites separated by at most four amino acids (Schweiger and Linial 2010). We have developed an approach that reduces the degree of arbitrariness in the definition of phospho-islands.

Let *S* be the distribution of amino acid distances between adjacent phosphosites in DRs. The logarithm of *S* is not unimodal (Fig. 2a), and we suggest that it is a superposition of two distributions: one generated by phosphosites in phospho-islands and the other reflecting phosphosites outside phospho-islands (left and right peaks, respectively). The latter distribution can be obtained from random sampling from DRs of non-phosphorylated STY amino acids while preserving the amino acid composition and the sample size, as we expect individual phosphosites to emerge independently while maintaining the preference towards DRs (Fig. 2c). Gamma distribution has a good continuous fit to log(*S*+1) for randomly sampled STY amino acids located in DRs. Given its universality and low number of parameters (Friedman et al. 2006; Reiss et al. 2007; Mendoza-Parra et al. 2013), we have selected gamma distribution as a reasonable model for log(*S*+1) (Fig. 2c). Assuming that the distribution of log(*S*+1) values for phosphosites located in phospho-islands should belong to the same family and fixing the parameters of the previously obtained distribution, we decomposed the distribution of log(*S*+1) values into the weighted sum of two gamma distributions, one of which corresponds to STYs located in phospho-islands and the other one, to remaining STYs in DRs (Fig 2a, red and grey curves, respectively). From these two gamma distributions we obtained parameters for a hidden Markov model, which, in turn, was used to map phosphorylation islands. The distributions of *S* values for phosphosites in identified islands and the distribution for other phosphosites yielded a good match to the expected ones (Fig. 2bd).

Both for the HMR and mouse datasets, more than half of phosphosites are located in phospho-islands (61% and 56%, respectively) (Fig. 2e, Suppl. Fig. S8AB). For human phosphosites, however, we see a larger proportion of sites (53%) located outside phospho-islands. In the latter case the distributions in the decomposition differ less, compared to the former two cases (Fig. 2a, Suppl. Fig. S8). It could be caused by a larger density of phosphosites in DRs of the human proteome, resulting from higher experimental coverage; that would lead to generally lower *S* values, which, in turn, could cause the right peak in the log(*S*+1) distribution to merge with the left peak, rendering the underlying gamma-distributions less distinguishable. To validate this explanation, we randomly sampled 40% of human phosphosites, so that the sample size matched the one for mouse phosphosites; however, the results on this rarefied dataset did not change (Fig. 2e, Suppl. Fig. S8C) indicating that our procedure is robust with respect to phosphosite sample sizes. Hence, phospho-islands for the human dataset are identified with a lower accuracy than those for the HMR and mouse datasets. This could be caused by different experimental technique applied to the human phosphosites, compared to the one used for mouse and rat phosphosites, and by a possibly large number of false-positive phosphosites in the former case (Huttlin et al. 2010; Bekker-Jensen et al. 2017; Xu et al. 2017) (see Discussion).

In phospho-islands, the overall pSTY-amino acid composition differs from that of individual phosphosites, mainly because the fraction of threonines is significantly higher in phospho-islands at the expense of the lower fraction of tyrosines (Fig. 2f). Also, the conservation of residues in phospho-islands is larger than that of the individual sites (Fig. 2g). Overall, the general properties of clustered phosphosites seem to differ from those of individual phosphosites.

We do not observe phospho-islands in ordered regions, as the distribution of log(*S*+1) values in this case seemingly cannot be decomposed into a weighted sum of two unimodal distributions and is largely skewed to the left even relative to the distribution of log(*S*+1) values in phospho-islands. Hence, either virtually all pairs of adjacent sites there comprise dense phospho-islands, or a more complex model possibly incorporating tertiary protein structure features is required to infer phospho-islands in this case (Suppl. Fig. S8E).

Mutational patterns of phosphorylated amino acids

Next, we have reconstructed the ancestral states for all mammalian orthologous protein groups not containing paralogs and calculated the proportions of mutations $P(X_1 \rightarrow X_2)$, where X_1 and X_2 are different amino acids. We treated phosphorylated and non-phosphorylated states of STY amino acids as distinct states. We then introduced a measure of difference in mutation rates for phosphorylated STY and their non-phosphorylated counterparts. For a mutation of an STY amino acid X to a non-STY amino acid Z we define $R(X, Z) = P(pX \rightarrow Z) / P(X \rightarrow Z)$. If X^* is another STY amino acid, $R(X, X^*) = P(pX \rightarrow pX^*) / P(X \rightarrow X^*)$. Thus, the R value for a given type of mutations is the proportion of the considered mutation of a phosphorylated STY amino acid among other mutations normalized by the fraction of respective mutations of the non-phosphorylated STY counterpart. The R values are thus not affected by differences in the mutation rates between phosphorylated and non-phosphorylated amino acids, as all probabilities are implicitly normalized by the mutation rates of pX and X.

We firstly consider phosphosites located in DRs. For phosphorserines from the HMR dataset we confirm earlier observations: phosphoserines mutate to NCA more frequently than non-phosphorylated serines (Fig. 3a). The R values for serine mutation to aspartate, R(S,D), and glutamate, R(S,E), are both significantly larger than 1 (1.2, $p<0.01$ and 1.7, $p<0.001$, respectively; χ^2 test) and, interestingly, they differ substantially ($p<0.001$, multiple random Poisson sampling test). Similarly, asparagine and glutamine R values differ, with R(S,N)=0.9 ($p<0.001$, χ^2 test), significantly lower than 1, and R(S,Q)=1.4 ($p<0.001$, χ^2 test), significantly higher than 1. The rate of mutation to lysine significantly differs for phosphorylated and non-phosphorylated serines ($p<0.001$, χ^2 test). Interestingly, the mutation rate to another positively charged amino acid, arginine, is significantly lower than expected ($p<0.01$, χ^2 test). For non-polar amino acids generally no significant differences in the R values between phosphorylated and non-phosporylated serines are observed, but for methionine and proline, the calculated values are significant: R(S,M)>1 ($p<0.001$, χ^2 test) and R(S,P)<1 ($p<0.01$, χ^2 test).

In earlier studies, only mutations of serines or to serines had been considered, as the available data did not allow for statistically significant results for threonine and tyrosine (Kurmangaliyev et al. 2011; Miao et al. 2018). Here, we see that phosphorylated threonines from the HMR dataset tend to mutate to serines (Fig. 3b). At that, phosphorylated serines mutate to threonines more frequently than their non-phosphorylated counterparts for all considered samples, i.e. for the human, mouse and HMR sets (Fig. 3b, Suppl. Figs. S2-S7). Phosphorylated tyrosines tend to avoid mutations to isoleucine ($p<0.05$, χ^2 test) and, for human samples, to arginine ($p<0.05$, χ^2 test) and glycine ($p<0.001$, χ^2 test) (Fig. 3b, Suppl. Figs. S2-S7). Phospho-tyrosines in the mouse dataset show a weaker tendency for the avoidance of the mutations to aspartate than the non-phosphorylated ones ($p<0.05$, χ^2 test) while the rate of pY-to-I mutations is higher (Fig. 3b).

Separate analysis of mutations in phospho-islands and in individual phosphosites yields three observations. Firstly, alterations of mutation patterns of phosphoserines and phosphothreonines (pST) in DRs relative to non-phosphorylated ST in DRs are similar to the patterns observed for the clustered pST and, to a lesser extent, to those observed for individual pSTs (Fig. 3b). This is mostly due to the fact that the mutational patterns of clustered pSTs generally differ from those of their non-phosphorylated counterparts to a greater extent than the mutational patterns of individual phosphoserines do (Fig. 3b, Suppl. Fig. S1). Secondly, for phosphotyrosines, alterations in their mutational patterns brought about by phosphorylation are mostly explained by individual phosphotyrosines. The mutational patterns of individual sites deviate from the ones observed for non-phosphorylated tyrosines more than those of clustered phosphotyrosines (Fig. 3b, Suppl. Figs. S2-S7). Also, if we compare the *R* values calculated for all possible mutations in clustered *vs.* individual phosphosites, the *R* value corresponding to the S-to-E mutation will be significantly higher for the set of clustered phosphosites (p=0.009, χ^2 test, Suppl. Fig. S1). Hence, we posit that the general phosphosite mutational pattern alterations can be explained mostly by mutations in clustered phosphosites for phosphoserines and phosphothreonines and by individual sites when phosphotyrosines are considered.

We also studied mutation patterns in ordered regions (ORs), and observed that phosphothreonines located in ORs demonstrate higher T-to-S mutation rates (Fig. 3b) relative to those of non-phosphorylated threonines located in ORs. Also, sites located in ORs demonstrate enhanced S-to-T and Y-to-T mutation rates relative to non-phosphorylated serines and threonines in ORs, respectively (Fig. 3b).

Phosphosite contexts

Sequence contexts of phosphosites generally fall into three categories: acidic (A), basic (B), and proline (P) motifs, with tyrosine phosphosites comprising a special class (Y) (Villen et al. 2007; Huttlin et al. 2010). For each phosphosite from each dataset we have identified its context. As in previous studies (Villen et al. 2007; Huttlin et al. 2010), phosphosites not assigned with any of these context classes were considered as having “other” (O) motif. We studied the distribution of these motifs for all classes of phosphosites.

In DRs, relative to ORs, we observed a higher percentage of phosphosites with assigned contexts (Fig. 4a). P-phosphosites demonstrate the highest overrepresentation in DRs, with 25% of DR phosphosites having the proline motif. Phospho-islands, compared to individual phosphosites, contain more phosphosites with assigned motifs relative to individual phosphosites. In DRs, there are more B- and P-phosphosites and fewer A-phosphosites and Y-phosphosites among clustered sites than among individual ones.

Phosphorylation breadth

An important feature of a phosphosite is the “phosphorylation breadth”, that is, the number of tissues where it is phosphorylated. In this study, the maximal phosphorylation breadth is nine, as the phosphorylation data for nine mouse tissues are available (Huttlin et al. 2010). Among broadly expressed phosphosites (present in all nine tissues), compared to tissue-specific ones (present in only one tissue), very few sites have unassigned contexts (O) and almost none are tyrosine phosphosites. The fraction of acidic phosphosites (24%) is substantially lower among tissue-specific sites relative to broadly phosphorylated ones (37%) (p<0.001, χ^2 test) (Fig. 4a).

As mentioned above, the pS-to-E mutation yields the highest value, *R*(S,E) (Fig. 3a) and represents the only mutation with significantly different *R* values in phospho-islands and individual sites (*p*=0.009, χ^2 test, Suppl. Fig. S1). At that, *R*(S,E) significantly increase with increasing breadth of expression (Fig. 4b), from *R*(S,E)=1.14 for tissue-specific phosphosites to *R*(S,E)=6.64 for broadly expressed phosphosites (*p*=0.016, t-test).

Finally, we compared percentages of phosphosites with different breadths in ORs *vs.* DRs and in phospho-islands *vs.* individual phosphosites (Fig. 4cd). As the phosphorylation breadth increases, so does the fraction of clustered phosphosites, reaching 85% for sites phosphorylated in nine tissues; the fraction of phosphosites in DRs also increases, reaching 95.4%.

Hence, broadly expressed phosphosites have well-defined motifs, tend towards disordered regions and to phospho-islands, have mostly acidic context, and mutate to NCA more frequently than tissue-specific phosphosites.

Mutation patterns in the proximity of phosphosites

We now show that not only phoshosites require special motifs (Huttlin et al. 2010), but the mutational context of clustered phosphosites differs from that of individual sites. To assess evolutionary dynamics associated with phosphosite motifs, we analyzed mutational patterns in ±3 amino acid windows of HMR ST phosphosites located in DRs and compared them with those of non-phosphorylated ST amino acids. The ±3 window was selected, as it yielded the strongest effect in terms of the number of mutations with rates statistically distinct from the expected ones (Suppl. Fig. S9AB). We did not consider phosphotyrosines, as they have not been shown to possess any discernible general motif apart from the phosphorylated tyrosine itself (Huttlin et al. 2010).

We introduce the measure Q defined as $Q\left(X_1^p \rightarrow X_2^p\right) = P(X_1^p \rightarrow X_2^p)/P(X_1^n \rightarrow X_2^n)$, where X_1^p and X_2^p are amino acids near phosphorylated serines and threonines and X_1^n and X_2^n are amino acids near non-phosphorylated serines and threonines. *Q* measures overrepresentation of a given mutation in the proximity of pST amino acids relative to ST amino acids. We also considered sites located in phospho-islands and individual phosphosites separately (Fig. 5, Suppl. Fig. S9CD).

In the whole HMR dataset, 22 types of non-phosphorylated amino acid substitutions out of the total of 289 have *Q* values statistically different from the expected value 1 (*p*<0.05, χ^2 test with the Bonferroni correction), among them three pairs of mutually reverse mutations (Fig. 5). As expected from the conservation of the phosphosite contexts, mutations between positively and negatively charged amino acids, potentially changing acidic to basic contexts and *vice versa*, are underrepresented, whereas E-to-D, D-to-E and K-to-R, not changing the context type, are overrepresented. The P-to-A substitution is overrepresented, thus indicating the instability of proline contexts. Interestingly, all three mutations with *Q* values exceeding 2.5 involve lysine, two of them being reverse mutations F-to-K and K-to-F. The fourth most overrepresented mutation, Y-to-G with *Q*(Y→G)=2.5, could explain the lack of tyrosine phosphosites in DRs, as a large fraction of DR phosphosites are clustered with the distances between sites not exceeding three amino acids. Thus, a large *Q*(Y→G) value would lead to general underrepresentation of tyrosines in DRs.

Types of mutations with significant *Q* values generally differ near clustered and individual phosphosites (Suppl. Fig. S9CD). E-to-D, not changing the local acidic context type (Huttlin et al. 2010), is overrepresented and E-to-K, disrupting the acidic context (Huttlin et al. 2010), is underrepresented in both cases. On the other hand, around individual phosphosites, Q(F→K)=3.4 and Q(P→A)=1.12, indicating an enhanced birth rate of the basic context and disruption of the proline context, respectively. The R-to-D mutation, disrupting the local basic context, also is overrepresented near individual phosphosites. In general, among seven overrepresented mutations near clustered phosphosites, only the K-to-P mutation disrupts the local basic context in favor of the proline context and among seven overrepresented mutations near individual phosphosites, three mutations (E-to-F, R-to-D, and P-to-A) could be regarded as context-disrupting. Hence, the individual phosphosite contexts are somewhat less evolutionary stable and thus the lower percentage of individual phosphosites with identifiable contexts might be due to specific local context-disrupting mutation patterns for these phosphosites.

Discussion

Clustered vs. individual phosphosites

We have demonstrated that clustered phosphosites differ from non-clustered ones in a number of aspects: (i) overrepresentation of phosphothreonines and underrepresentation of phosphotyrosines in phospho-islands (Fig. 2f); (ii) stronger conservation of clustered phosphoserines and phosphothreonines (Fig. 2g); (iii) larger proportion of sites phosphorylated in many tissues (Fig. 4C); (iv) significantly larger probability of mutations to glutamate for clustered relative to the individual phosphoserines; (v) larger fraction of sites with specific motifs in phospho-islands (Fig. 4A); (vi) mutational patterns in the proximity of phosphosites consistent with the context-retention hypothesis (Fig. 5). What are possible explanations for the observed effects?

Underrepresentation of phosphotyrosines in phospho-islands could be explained by phosphorylation of clustered phosphosites being co-operative. As serines and threonines are more similar to each other in their tendency to being phosphorylated by similar enzymes than they are to tyrosine (Villen et al. 2007; Huttlin et al. 2010; Landry et al. 2014; Studer et al. 2016), one would expect phospho-tyrosines to disrupt co-operative phosphorylation of adjacent ST amino acids by being phosphorylated independently, thus introducing a negative charge which would affect phosphorylation probabilities of the neighbouring amino acids (Landry et al. 2014). Hence phospho-tyrosines could have been purged by selection from pST clusters.

Secondly, phosphosites located in phospho-islands are more conserved than individual ones (Fig. 2g), as opposed to an earlier hypothesis that individual phosphosites are more conserved than their clustered counterparts (Landry et al. 2014). Our result seems to contradict to the notion that the cellular function of phosphosites in an island depends on the number of phosphorylated residues rather than specific phosphorylated sites, whereas individual phosphosites operate as single-site switches and hence should be more conserved (Landry et al. 2014). However, this argument implies that phosphorylation of most individual phosphosites is important for the organism’s fitness, which may be not true (Landry et al. 2014; Miao et al. 2018) and hence our results do not contradict the model of evolution of functionally important phosphosites.

Overrepresentation of phosphosites with defined motifs among the clustered ones (Fig. 4A) and reduced numbers of mutations disrupting the local contexts of the clustered sites

(Suppl. Fig. S9CD) may indicate enhanced selective pressure on clustered phosphosites and their contexts. An indirect support of this claim comes from the overrepresentation of ubiquitously phosphorylated sites among the clustered ones (Fig. 4c). Indeed, broad phosphorylation requires a stronger local context and indicates the reduced probability of a phosphosite being detected simply due to the noise inherent to the phosphorylation machinery (Landry et al. 2014).

Mutations of phosphoserines located in DRs to NCA are generally overrepresented among all mutations of the type pS-to-X relative to the corresponding mutations of non-phosphorylated serines (Fig. 3b). This effect is stronger for clustered phosphosites and for ubiquitously phosphorylated sites. Together with the observation about clustered phosphosites being on average more broadly phosphorylated than the individual ones, this suggests that a large fraction of phosphosite clusters might be phosphorylated (nearly) constitutively, and thus changes of individual phospho-serines to NCAs could experience lesser degrees of negative selection acting upon the corresponding mutations, as these mutations introduce smaller degrees of local electric charge shifts on the protein globule than the mutations of non-phosphorylated serines to NCAs do.

Two types of mutations

In all considered phosphosite datasets, we have observed two types of pSTY-to-X mutations overrepresented relative to STY-to-X mutations (Fig. 3b): (i) pSTY-to-pSTY, especially pT-to-pS mutation and (ii) pSTY-to-NCA, especially pS-to-E mutations. The former effect could be explained by the relaxed selection against pST-to-pST mutations due to the phosphorylation machinery often not distinguishing between serines and threonines (Huttlin et al. 2010; Miao et al. 2018). The overrepresentation of pT-to-pS mutation for all datasets, including sites located in ORs, could stem from the higher probability of phosphosite retention following a pT-to-pS mutation relative to the probability of phosphorylated threonine retention when no mutations have occurred (Fig. 1c). Thus, the observed enhanced pT-to-pS mutation rate could be due to the enhanced evolutionary stability of serine phosphorylation relative to the threonine phosphorylation.

The enhanced serine-to-NCA mutation rates could stem from the physico-chemical similarity of phosphorylated serines and negatively charged amino acids: both types of residues introduce negatively charged groups of similar size to the protein globule. Thus, if phosphorylation is (almost) constitutive, i.e. happens very frequently in a large number of tissues, we would expect the serine-to-NCA mutation rate to be enhanced. Indeed, ubiquitous phosphorylated serines have the pS-to-E mutation rate more than six-fold larger than the S-to-E mutation rate (Fig. 4b). However, the same pattern does not hold for phospho-threonines (Fig. 3B).

Human phosphosites

The results obtained for the human set of phosphosites differ somewhat from those for the mouse and HMR sets, like in cases with different STY amino acids representation among phosphorylated amino acids (Fig. 1D), proportion of phosphosites located in phospho-islands (Fig. 2E) or some mutational patterns of phosphorylated STY amino acids (Fig. 3B). This could be explained by differences in experimental procedures used to obtain phosphosite lists for human and for mouse and rat. Whereas for classic laboratory organisms, phosphosites are obtained directly from the analysis of an organism or an analysis of its live organ (Huttlin et al. 2010), for human phosphosite inference immortalized cell lines, such as HeLa, are used (Bekker-Jensen et al. 2017; Xu et al. 2017), with conditions differing from those *in vivo*, and hence one could expect different patterns of phosphorylation. In particular, the lower rate of mutations to NCA could be explained by overrepresentation of sites with noisy phosphorylation manifesting only in cell lines under the conditions of experiments. The mutation of such a residue to NCA would most likely result in the deleterious effect of an average non-phosphorylated serine mutation to NCA (Jin and Pawson 2012). Thus, we propose that phosphosites conserved between human and rodent lineages, called here HMR sites, are more robust with respect to experimental techniques, and hence are better suited for phosphosite evolutionary studies.

Evolution of non-studied phosphosite groups

Previous studies dedicated to the evolution of phosphosites have focused on phosphoserines located in DRs. The large datasets employed in the present study enabled us to assess the patterns of phosphothreonines, phosphotyrosines and sites located in ORs. Apart from the largely enhanced pT-to-pS mutation proportions relative to T-to-S ones (Fig. 3b) no patterns with straightforward biological explanation were observed in these cases. However, an interesting observation here is the consistent, significantly enhanced rate of pY-to-I mutations relative to the Y-to-I mutations in the mouse and HMR datasets (Fig. 3b).

Perspectives

We propose a simple yet accurate homology-based approach for the ancestral phosphosite inference yielding in our case the set of HMR phosphosites. As the predicted fractions of phosphorylation labels falsely assigned to internal tree nodes are much smaller than the ones for other phosphosite datasets, HMR set poses a valuable source of data for evolutionary studies. A practical extension of our homology-based approach could be a phosphosite prediction procedure incorporating additional pieces of information such as the tendency of phosphosites to cluster, the local phosphosite contexts, and the tree structure into the probabilistic model, which would predict phosphosites with a high degree of accuracy. On the other hand, it would be interesting to infer the interplay between phosphorylation and selection using population-genetics data.

References

2018. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research [Internet] 47:D506–D515. Available from: http://dx.doi.org/10.1093/nar/gky1049

Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, de Farias TM, Zile K, Stevenson C, Long J, et al. 2017. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Research [Internet] 46:D477–D485. Available from: http://dx.doi.org/10.1093/nar/gkx1019

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology [Internet] 215:403–410. Available from: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Ardito F, Giuliani M, Perrone D, Troiano G, Muzio LL. 2017. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). International Journal of Molecular Medicine [Internet] 40:271–280. Available from: http://dx.doi.org/10.3892/ijmm.2017.3036

Bekker-Jensen DB, Kelstrup CD, Batth TS, Larsen SC, Haldrup C, Bramsen JB, Sørensen KD, Høyer S, Ørntoft TF, Andersen CL, et al. 2017. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Systems [Internet] 4:587–599.e4. Available from: http://dx.doi.org/10.1016/j.cels.2017.05.009

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research [Internet] 32:1792–1797. Available from: http://dx.doi.org/10.1093/nar/gkh340

Friedman, N., Cai, L., & Xie, X. S. (2006). Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. Physical Review Letters, 97(16). https://doi.org/10.1103/physrevlett.97.168302

Fuhs, S. R., & Hunter, T. (2017). pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. Current Opinion in Cell Biology, 45, 8–16. https://doi.org/10.1016/j.ceb.2016.12.010

http://etetoolkit.org/

Huang H, Arighi CN, Ross KE, Ren J, Li G, Chen S-C, Wang Q, Cowart J, Vijay-Shanker K, Wu CH. 2017. iPTMnet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Research [Internet] 46:D542–D550. Available from: http://dx.doi.org/10.1093/nar/gkx1104

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villén J, Haas W, Sowa ME, Gygi SP. 2010. A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. Cell [Internet] 143:1174–1189. Available from: http://dx.doi.org/10.1016/j.cell.2010.12.001

Iakoucheva LM. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research [Internet] 32:1037–1049. Available from: http://dx.doi.org/10.1093/nar/gkh253

Jin J, Pawson T. 2012. Modular evolution of phosphorylation-based signalling systems. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet] 367:2540–2555. Available from: http://dx.doi.org/10.1098/rstb.2012.0106

Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of ancestral protein sequences. Journal of Molecular Evolution [Internet] 42:313–320. Available from: http://dx.doi.org/10.1007/BF02198858

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution [Internet] 34:1812–1819. Available from: http://dx.doi.org/10.1093/molbev/msx116

Kurmangaliyev YZ, Goland A, Gelfand MS. 2011. Evolutionary patterns of phosphorylated serines. Biology Direct [Internet] 6:8. Available from: http://dx.doi.org/10.1186/1745-6150-6-8

Landry CR, Freschi L, Zarin T, Moses AM. 2014. Turnover of protein phosphorylation evolving under stabilizing selection. Frontiers in Genetics [Internet] 5. Available from: http://dx.doi.org/10.3389/fgene.2014.00245

Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, Mann M. 2007. Phosphoproteome Analysis ofE. coliReveals Evolutionary Conservation of Bacterial Ser/Thr/Tyr Phosphorylation. Molecular & Cellular Proteomics [Internet] 7:299–307. Available from: http://dx.doi.org/10.1074/mcp.M700311-MCP200

Mendoza-Parra, M.-A., Nowicka, M., Van Gool, W., & Gronemeyer, H. (2013). Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. BMC Genomics, 14(1), 834. https://doi.org/10.1186/1471-2164-14-834

Miao B, Xiao Q, Chen W, Li Y, Wang Z. 2018. Evaluation of functionality for serine and threonine phosphorylation with different evolutionary ages in human and mouse. BMC Genomics [Internet] 19. Available from: http://dx.doi.org/10.1186/s12864-018-4661-6

Moses AM, Landry CR. 2010. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? Trends in Genetics [Internet] 26:462–467. Available from: http://dx.doi.org/10.1016/J.Tig.2010.08.002

Nishi, H., Shaytan, A., & Panchenko, A. R. (2014). Physicochemical mechanisms of protein regulation by phosphorylation. Frontiers in Genetics, 5. https://doi.org/10.3389/fgene.2014.00270

Pearlman SM, Serber Z, Ferrell JE Jr. 2011. A Mechanism for the Evolution of Phosphorylation Sites. Cell [Internet] 147:934–946. Available from: http://dx.doi.org/10.1016/j.cell.2011.08.052

PTACEK J, SNYDER M. 2006. Charging it up: global analysis of protein phosphorylation. Trends in Genetics [Internet] 22:545–554. Available from: http://dx.doi.org/10.1016/j.tig.2006.08.005

Reiss, D. J., Facciotti, M. T., & Baliga, N. S. (2007). Model-based deconvolution of genome-wide DNA binding. Bioinformatics, 24(3), 396–403. https://doi.org/10.1093/bioinformatics/btm592

Schweiger R, Linial M. 2010. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biology Direct [Internet] 5:6. Available from: http://dx.doi.org/10.1186/1745-6150-5-6

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al. 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology [Internet] 7:539–539. Available from: http://dx.doi.org/10.1038/msb.2011.75

Studer RA, Rodriguez-Mias RA, Haas KM, Hsu JI, Vieitez C, Sole C, Swaney DL, Stanford LB, Liachko I, Bottcher R, et al. 2016. Evolution of protein phosphorylation across 18 fungal species. Science [Internet] 354:229–232. Available from: http://dx.doi.org/10.1126/science.aaf2144

Villen J, Beausoleil SA, Gerber SA, Gygi SP. 2007. Large-scale phosphorylation analysis of mouse liver. Proceedings of the National Academy of Sciences [Internet] 104:1488–1493. Available from: http://dx.doi.org/10.1073/pnas.0609836104

Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory [Internet] 13:260–269. Available from: http://dx.doi.org/10.1109/TIT.1967.1054010

Xu H, Chen X, Ying N, Wang M, Xu X, Shi R, Hua Y. 2017. Mass spectrometry-based quantification of the cellular response to ultraviolet radiation in HeLa cells.Huen MS-Y, editor. PLOS ONE [Internet] 12:e0186806. Available from: http://dx.doi.org/10.1371/journal.pone.0186806

Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. 2010. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics [Internet] 1804:996–1010. Available from: http://dx.doi.org/10.1016/j.bbapap.2010.01.011

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution [Internet] 24:1586–1591. Available from: http://dx.doi.org/10.1093/molbev/msm088

Yu Y, Zhou H, Kong Y, Pan B, Chen L, Wang H, Hao P, Li X. 2016. The Landscape of A-to-I RNA Editome Is Shaped by Both Positive and Purifying Selection.Schierup MH, editor. PLOS Genetics [Internet] 12:e1006191. Available from: http://dx.doi.org/10.1371/journal.pgen.1006191

Figure captions

Figure 1 | Phosphosites considered in the study. (A). Venn diagram of iPTMnet human, mouse and rat phosphosites. Intersections correspond to conserved phosphosites. The HMR phosphosite dataset is shown in pink. **(B)**. Phosphosite assignment procedures. Given a tree of a mammalian orthologous gene group and a column in the respective alignment, we assign phosphorylation labels to ancestral and extant amino acids, firstly, by propagating labels from one species to all other species in the tree (shown as separate red and blue arrows) and, secondly, by propagating labels predicted both in the selected species (e.g. human, as shown) and in one of the remaining species (mouse and rat); this corresponds to blue and red arrows entering a given node in the tree. Phosphosites obtained by the latter procedure are referred to as the HMR phosphosite dataset. In both procedures, phosphorylation is considered to be retained both for direct and indirect STY-to-STY mutations. **(C)**. Retention of phosphorylation upon mutation. Bars represent the probability of a conserved modification for the human dataset in the case of mutation and if mutation has not occurred. The letter after the vertical bar is an amino acid over which the probability was normalized. Three asterisks represent *p*<0.001 (χ^2 test). **(D)**. STY amino acid content of three groups of phosphosite datasets.

Figure 2 | Phospho-islands for the HMR phosphosite dataset. (A). The distribution of $log_{10}(S + 1)$ values (pink histogram) and its decomposition in two gamma distributions: the one for phospho-islands (red curve) and for individual phosphosites (red curve). **(B)**. The distribution of $log_{10}(S + 1)$ values for phosphosites predicted to be in phospho-islands. **(C)**. $log_{10}(S + 1)$ values for non-phosphorylated STY amino acids randomly sampled from DRs with the same sample size and amino acid content as in the HMR dataset. **(D)**. $log_{10}(S + 1)$ values for predicted individual phosphosites. **(E)**. Numbers of individual phosphosites and sites in phospho-islands for four datasets. **(F)**. Amino acid content of phospho-islands and individual phosphosites. **(G)**. Frequency of mutations for phosphosites and individual amino acids. Asterisks depict significantly different values ($p < 0.001$, χ^2 test).

Figure 3 | pX_0 -> X_1 substitution vectors. (A). *R* values of the pS→*X* substitutions for serines from the HMR dataset located in DRs. **(B)**. Substitution probabilities for phosphorylated STY amino acids significantly different from those for non-phosphorylated STY amino acids for several datasets. The significance levels are shown with the colors introduced in the panel in (A). Abbreviations on the horizontal axis: ISL – phosphosites located in phospho-islands, IND – individual phosphosites. DR – phosphosites from disordered regions, OR – phosphosites from ordered regions.

Figure 4 | Phosphosite contexts and phosphorylation breadth. (A). Overrepresentation of phosphosite contexts in ordered *vs*. disoredred regions, in phospho-islands *vs*. individual phosphosites and for broadly *vs*. narrowly distributed phosphorylated amino acids. One asterisk and three asterisks indicate statistical significance at the levels of 0.05 and 0.001 respectively (χ^2 test). **(B)**. The dependence of *R*(pS, E) on the phosphosite breadth. Pearson's r^2 is equal to 0.53 with the t-test p=0.016. **(C)**. The dependence of phosphosite fraction in phospho-islands on the phosphorylation breadth ($p=9*10^{-41}$, χ^2 test). **(D)**. Percent of phosphosites in disordered regions *vs*. phosphosite breadth ($p=4.1*10^{-10}$, χ^2 test).

Figure 5 | *Q* values of mutations near ST phosphosites with probabilities significantly different from the expected ones. Solid red lines connect mutually reverse mutations. Dashed lines indicate quazy-reverse mutations of amino acids with common chemical properties.

Figure 1

Figure 1 | Phosphosites considered in the study.

(A). Venn diagram of iPTMnet human, mouse and rat phosphosites. Intersections correspond to conserved phosphosites. The HMR phosphosite dataset is shown in pink. **(B)**. Phosphosite assignment procedures. Given a tree of a mammalian orthologous gene group and a column in the respective alignment, we assign phosphorylation labels to ancestral and extant amino acids, firstly, by propagating labels from one species to all other species in the tree (shown as separate red and blue arrows) and, secondly, by propagating labels predicted both in the selected species (e.g. human, as shown) and in one of the remaining species (mouse and rat); this corresponds to blue and red arrows entering a given node in the tree. Phosphosites obtained by the latter procedure are referred to as the HMR phosphosite dataset. In both procedures, phosphorylation is considered to be retained both for direct and indirect STY-to-STY mutations. **(C)**. Retention of phosphorylation upon mutation. Bars represent the probability of a conserved modification for the human dataset in the case of mutation and if mutation has not occurred. The letter after the vertical bar is an amino acid over which the probability was normalized. Three asterisks represent $p<0.001$ (χ^2 test). **(D)**. STY amino acid content of three groups of phosphosite datasets.

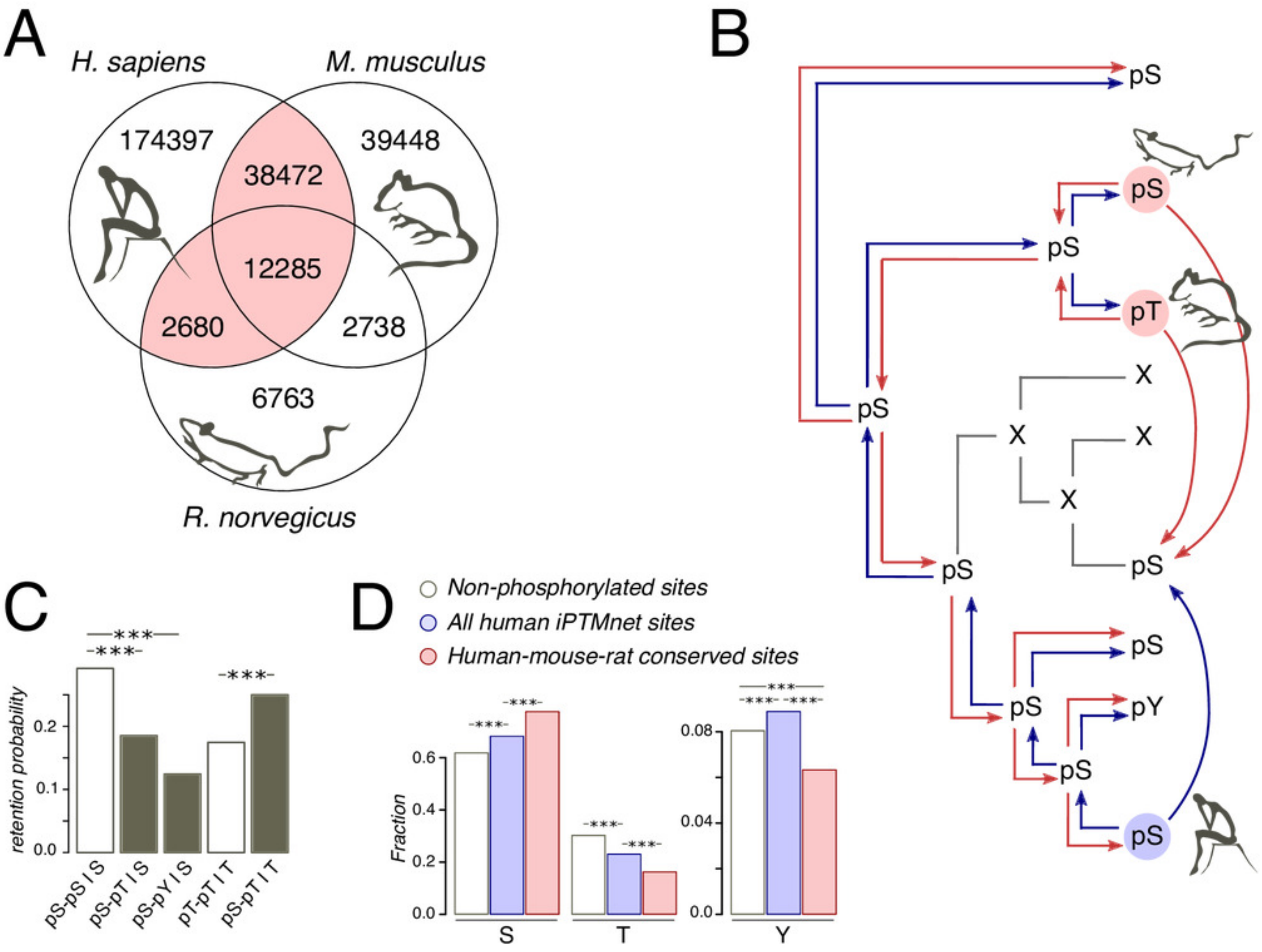

A
H. sapiens
M. musculus
R. norvegicus
174397
38472
39448
12285
2680
2738
6763
B
pS
pT
pY
X
C
—***—
-***-
-***-
retention probability
0.0
0.1
0.2
pS-pS | S
pS-pT | S
pS-pY | S
pT-pT | T
pS-pT | T
D
○ Non-phosphorylated sites
● All human iPTMnet sites
● Human-mouse-rat conserved sites
Fraction
0.0
0.2
0.4
0.6
0.00
0.04
0.08
-***-
S
T
Y

Figure 2

Figure 2 | Phospho-islands for the HMR phosphosite dataset.

(A). The distribution of *$log_{10}(S + 1)$* values (pink histogram) and its decomposition in two gamma distributions: the one for phospho-islands (red curve) and for individual phosphosites (red curve). **(B)**. The distribution of *$log_{10}(S + 1)$* values for phosphosites predicted to be in phospho-islands. **(C)**. *$log_{10}(S + 1)$* values for non-phosphorylated STY amino acids randomly sampled from DRs with the same sample size and amino acid content as in the HMR dataset. **(D)**. *$log_{10}(S + 1)$* values for predicted individual phosphosites. **(E)**. Numbers of individual phosphosites and sites in phospho-islands for four datasets. **(F)**. Amino acid content of phospho-islands and individual phosphosites. **(G)**. Frequency of mutations for phosphosites and individual amino acids. Asterisks depict significantly different values ($p < 0.001$, χ^2 test).

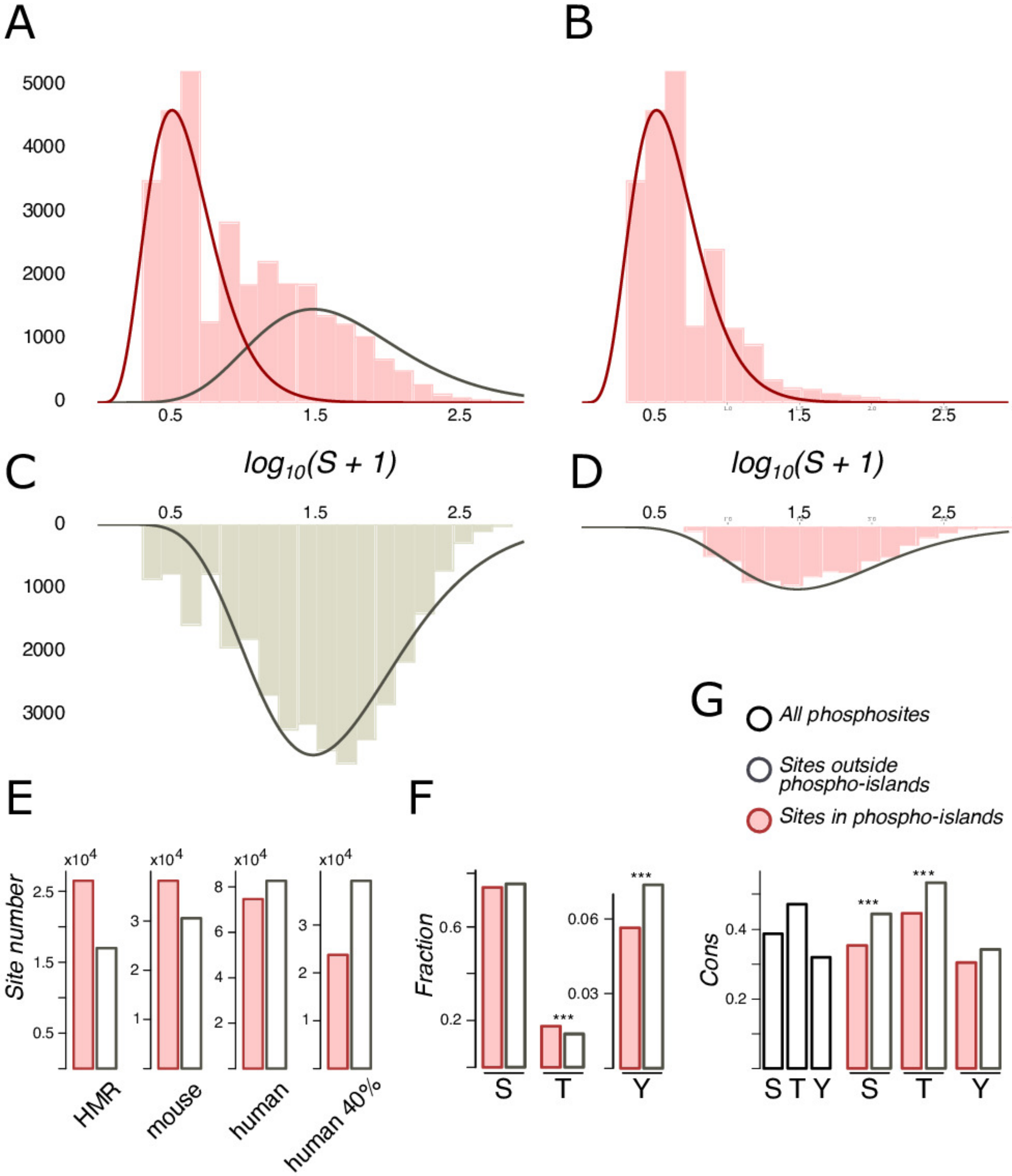

A
B
5000
4000
3000
2000
1000
0
0.5
1.5
2.5
C
$log_{10}(S + 1)$
D
$log_{10}(S + 1)$
0
1000
2000
3000
G
All phosphosites
Sites outside phospho-islands
Sites in phospho-islands
E
F
x10^4
Site number
2.5
1.5
0.5
3
2
1
8
6
4
2
HMR
mouse
human
human 40%
Fraction
0.6
0.2
0.06
0.03
***
S
T
Y
Cons
0.4
0.2
S T Y
S
T
Y

Figure 3

Figure 3 | pX_0 -> X_1 substitution vectors.

(A). *R* values of the pS*X* substitutions for serines from the HMR dataset located in DRs. **(B)**. Substitution probabilities for phosphorylated STY amino acids significantly different from those for non-phosphorylated STY amino acids for several datasets. The significance levels are shown with the colors introduced in the panel in (A). Abbreviations on the horizontal axis: ISL – phosphosites located in phospho-islands, IND – individual phosphosites. DR – phosphosites from disordered regions, OR – phosphosites from ordered regions.

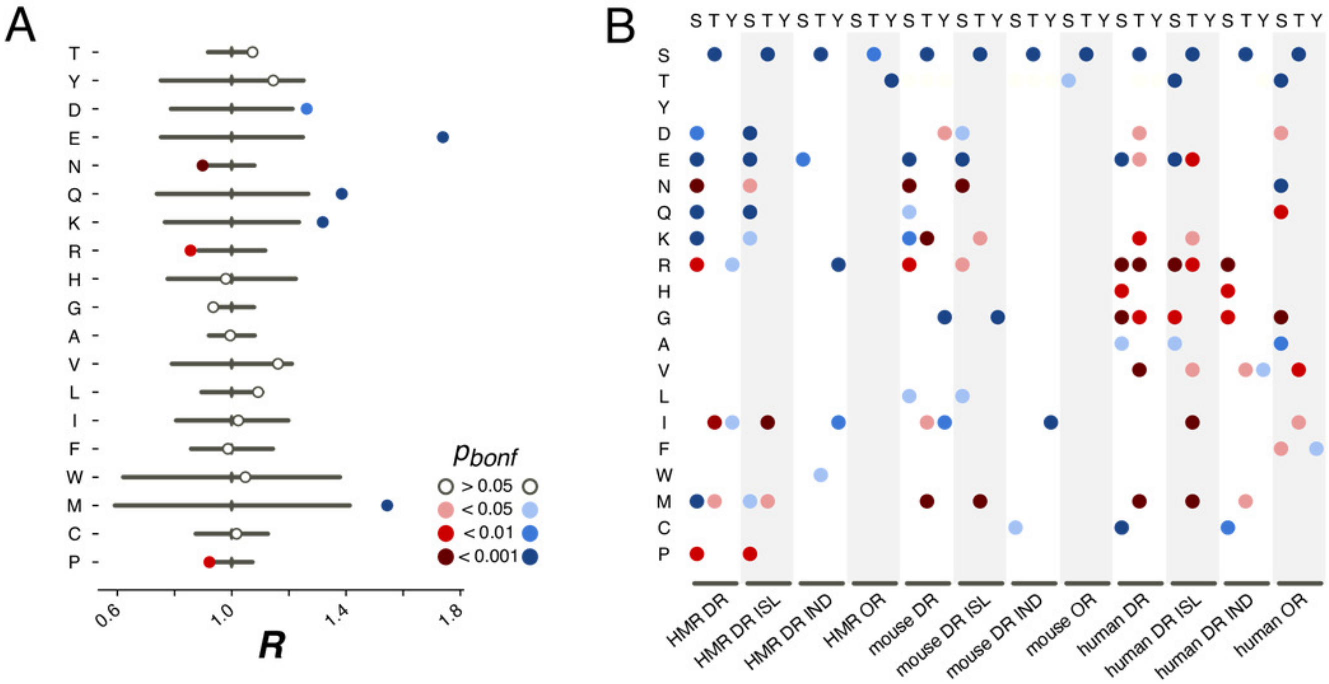


Figure 4

Figure 4 | Phosphosite contexts and phosphorylation breadth.

(A). Overrepresentation of phosphosite contexts in ordered *vs*. disoredred regions, in phospho-islands *vs*. individual phosphosites and for broadly *vs*. narrowly distributed phosphorylated amino acids. One asterisk and three asterisks indicate statistical significance at the levels of 0.05 and 0.001 respectively (χ^2 test). **(B)**. The dependence of *R*(pS,E) on the phosphosite breadth. Pearson's r^2 is equal to 0.53 with the t-test p=0.016. **(C)**. The dependence of phosphosite fraction in phospho-islands on the phosphorylation breadth (p=9*10^{-41}, χ^2 test). **(D)**. Percent of phosphosites in disordered regions *vs*. phosphosite breadth (p=4.1*10^{-10}, χ^2 test).

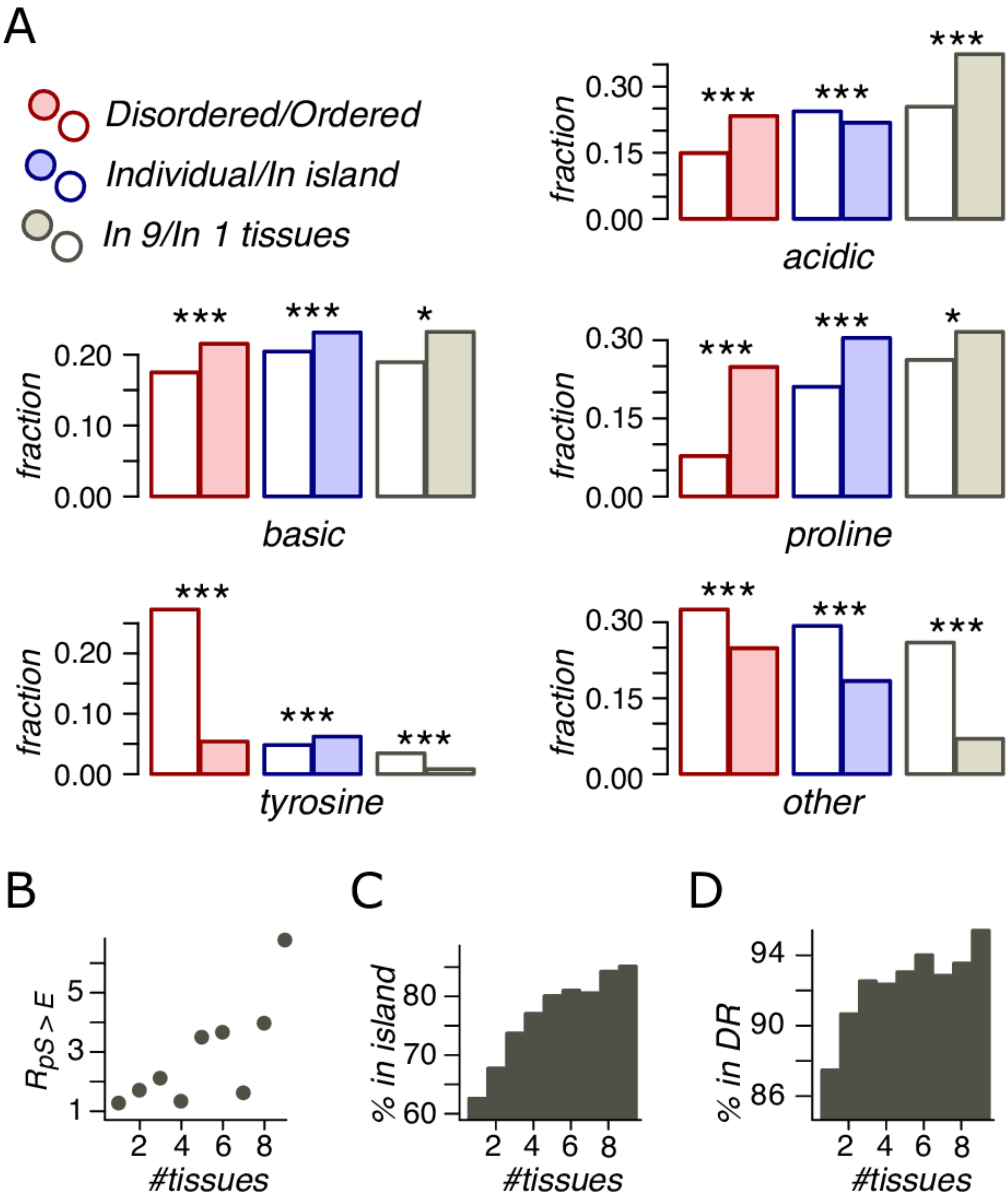
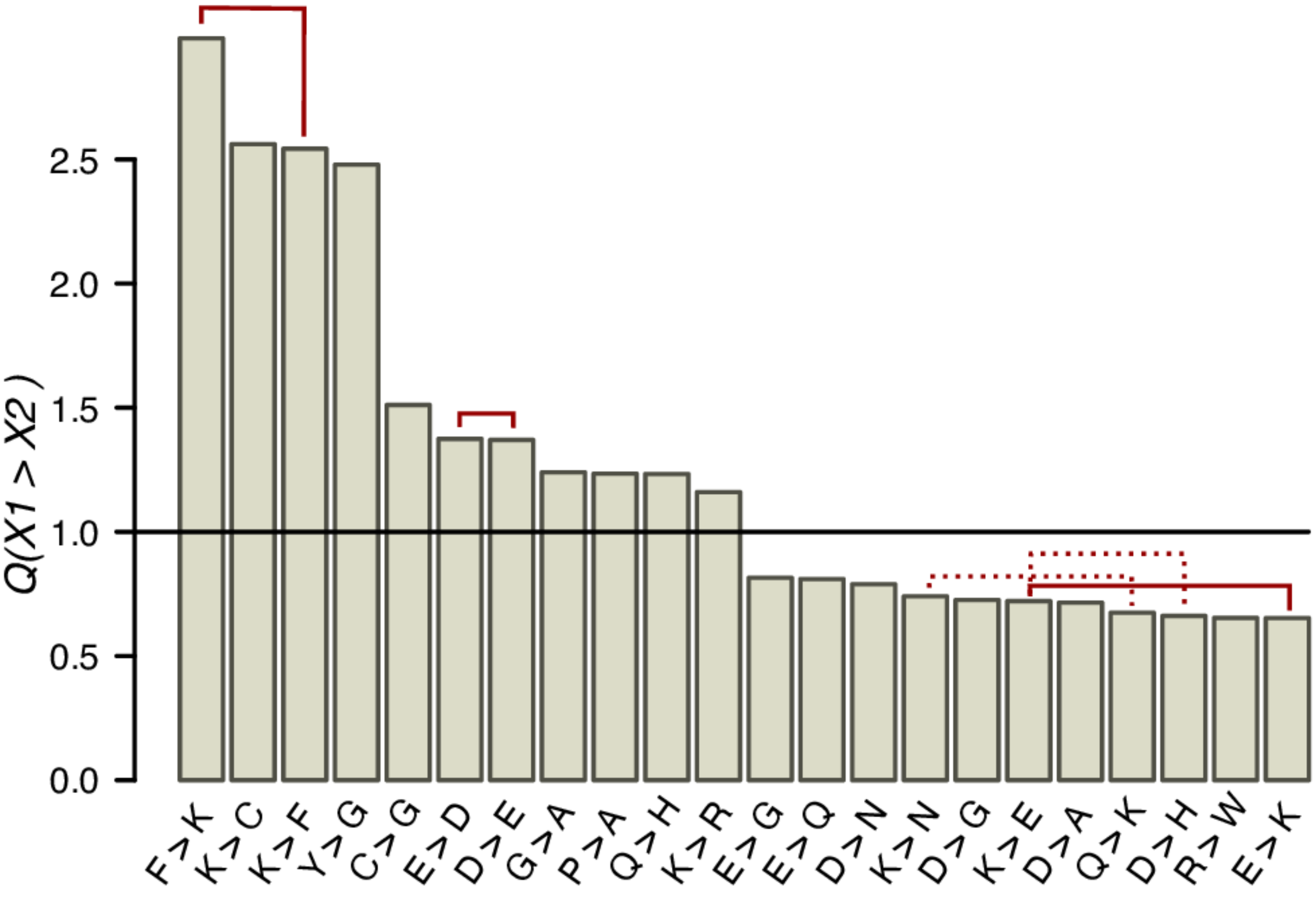

A
Disordered/Ordered
Individual/In island
In 9/In 1 tissues
fraction
0.30
0.15
0.00
***
***
***
acidic
fraction
0.20
0.10
0.00
***
***
*
basic
fraction
0.30
0.15
0.00
***
***
*
proline
fraction
0.20
0.10
0.00
***
***
***
tyrosine
fraction
0.30
0.15
0.00
***
***
***
other
B
$R_{pS>E}$
1
3
5
2
4
6
8
#tissues
C
% in island
60
70
80
2
4
6
8
#tissues
D
% in DR
86
90
94
2
4
6
8
#tissues

Figure 5

Figure 5 | *Q* values of mutations near ST phosphosites with probabilities significantly different from the expected ones.

Solid red lines connect mutually reverse mutations. Dashed lines indicate quazy-reverse mutations of amino acids with common chemical properties.


Q(X1 > X2)
0.0
0.5
1.0
1.5
2.0
2.5
F>K
K>C
K>F
Y>G
C>G
E>D
D>E
G>A
P>A
Q>H
K>R
E>G
E>Q
D>N
K>N
D>G
K>E
D>A
Q>K
D>H
R>W
E>K