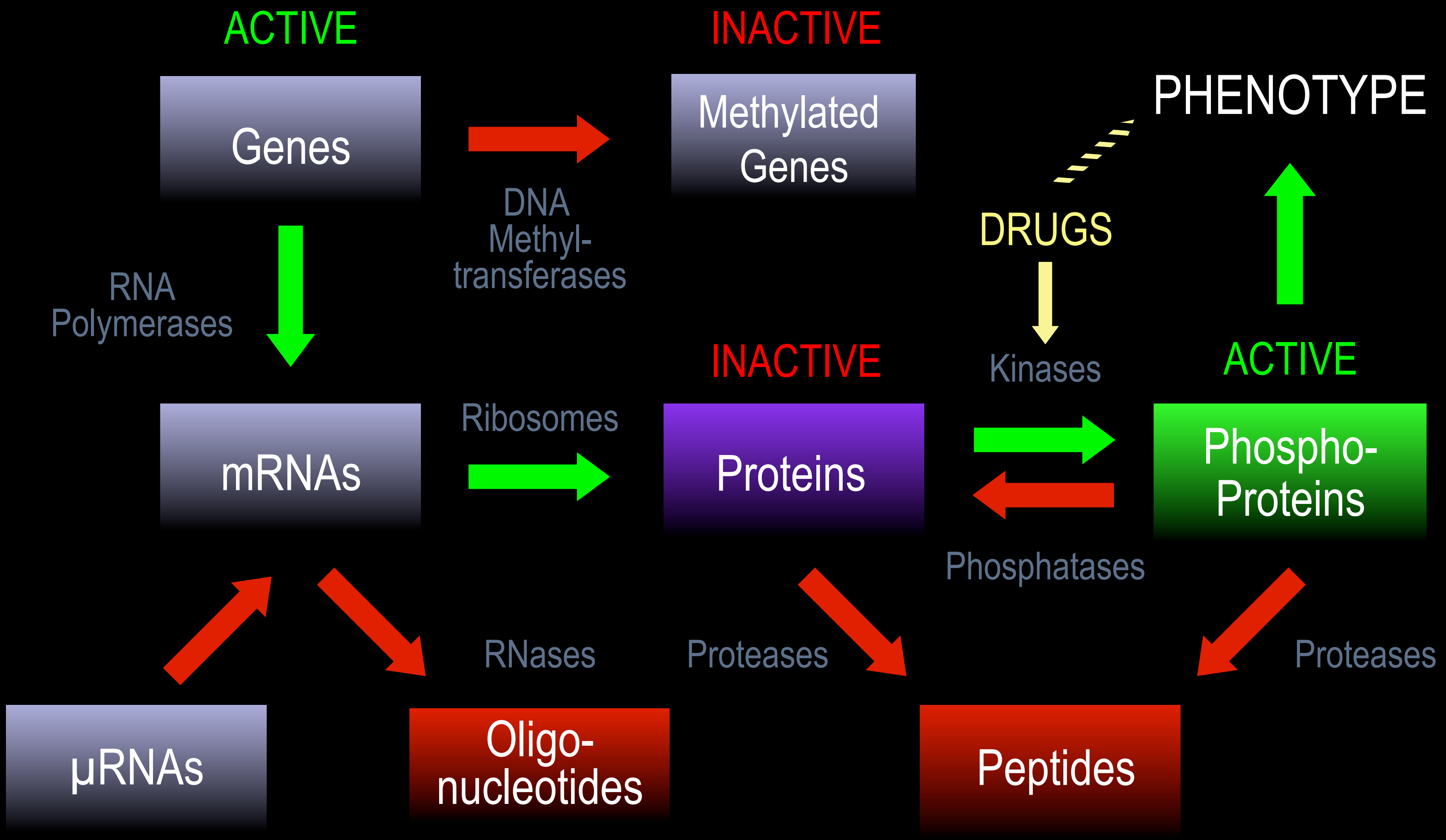
Phosphoproteomics and the Origin and Operations of the Kineome

Presented by Steven Pelech, Ph.D.
Professor, Department of Medicine,
University of British Columbia
President & CSO, Kinexus Bioinformatics Corp.

Presented at the OMICS Group 4th International Conference on Proteomics & Bioinformatics in Northbrook, IL, USA, 2014 Aug. 5

KINEXUS

Phenotype Reflects Functional Protein



ACTIVE

Genes

RNA Polymerases

mRNAs

INACTIVE

Methylated Genes

DNA Methyl-transferases

Ribosomes

PHENOTYPE

DRUGS

Kinases

INACTIVE

Proteins

ACTIVE

Phospho-Proteins

Phosphatases

μRNAs

RNases

Oligo-nucleotides

Proteases

Peptides

Proteases

Protein Kinase Gene Numbers

Common Name	Scientific Name	#Genes	#Protein Kinases
Baker's yeast	*Sacchar. cerevisiae*	6,300	122
Cat	*Felis silvestris*	20,285	
Chicken	*Gallus gallus*	23,000	546
Chimpanzee	*Pan troglodytes*	30,000	587
Cow	*Bos taurus*	22,000	
Dog	*Lupus familis*	40,000	656
Fruit fly	*Dros. melanogaster*	14,000	319 (232)
Honey bee	*Apis mellifera*	10,157	
Human	*Homo sapiens*	21,500	>536 (505)
Maize	*Zea mays*	32,000	
Mouse	*Mus musculus*	23,800	510 (527)
Nematode worm	*Caenorhab. elegans*	22,900	437 (434)
Poplar tree	*Populus trichocarpa*	45,555	
Rat	*Rattus norvegicus*	30,000	521
Red bread mold	*Neurospora crassa*	10,000	103
Rhesus Macaque	*Macaca mulatta*	18,296	821
Rice	*Oryza sativa*	56,000	1429
Sea urchin	*Strongylocentrotus purpuratus*	23,300	353
Thale Cress	*Arabidopsis thaliana*	27,000	1049
Tiger blowfish	*Takifugu rubripes*	31,000	519
Zebra Danio	*Brachydanio rerio*	24,200	

KINEXUS

Human Protein Phosphorylation

PROTEOME

~21,500 total protein genes

KINEOME

>536 protein kinase genes

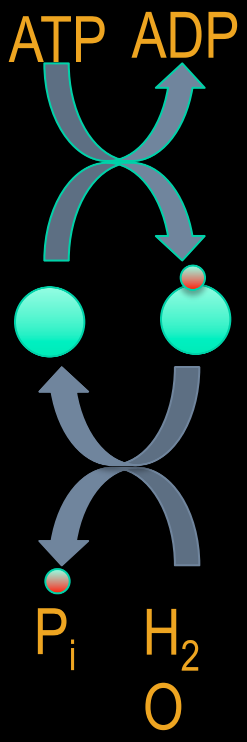
>1000 catalytically active protein kinase isoforms

PHOSPHO-PROTEOME

~1,000,000 predicted phosphorylation sites (>180,000 experimentally confirmed)

PHOSPHATOME

>156 protein phosphatase genes



Statistics on the Human Phosphoproteome

- Number of human phosphosites known: >180,000
- Number of human phosphoproteins identified: >19,800
- Percent serine phosphorylation sites: 53.5%*
- Percent threonine phosphorylation sites: 25.2%*
- Percent tyrosine phosphorylation sites: 21.3%*

*Based on 159,540 human phosphosites identified by mass spectrometry

Number of Human Phosphosites

Based on 60,000 human phosphosites identified by mass spectrometry without specific enrichment of tyrosine phosphorylation

- Percent serine phosphorylation sites: 76%
- Percent threonine phosphorylation sites: 20%
- Percent tyrosine phosphorylation sites: 4%

- Number of tyrosine phosphosites known: >36,194

Minimum number of human phosphosites: (36,000 x 25 =) 900,000

Probable number of human phosphosites: 1,000,000

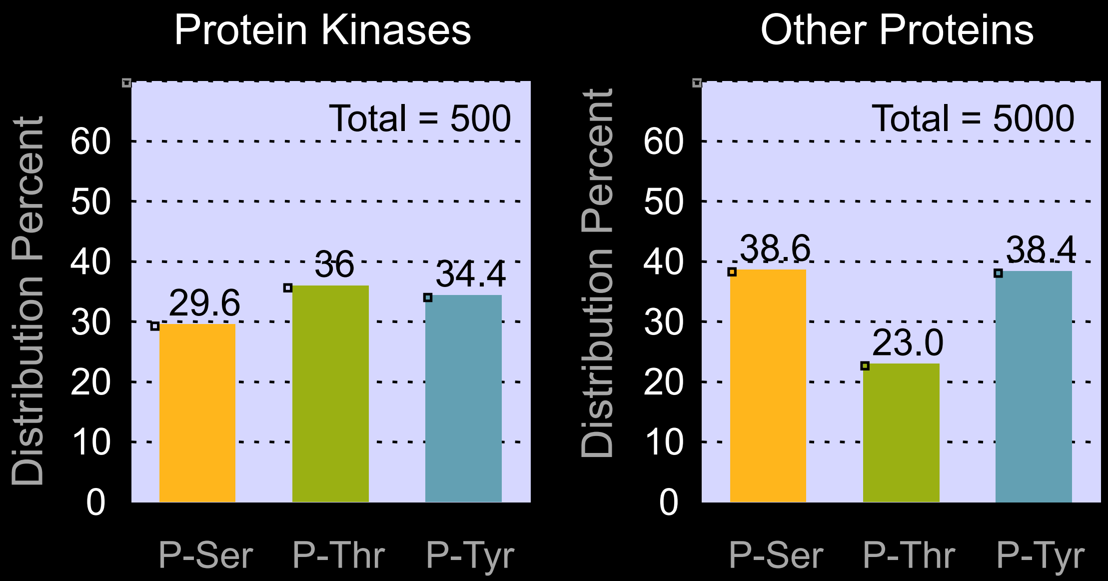
KINEXUS

Number of Phosphosites per Protein

	Selected Amino Acid		
	Serine	Threonine	Tyrosine
Percentage of total phosphosites	76%	20%	4%
Estimated number of total phosphosites (assume 1,000,000 total sites)	760,000	200,000	40,000
Estimated number of phosphosites per average protein (assume 21,500 proteins encoded by human genome)	35.3	9.30	1.86
Phosphosite frequency - Average length of peptide in amino acids with a phosphosite	19.9	75.6	378
Frequency of selected amino acid in general in human proteins	7.1%	6.0%	3.2%
Number of selected amino acids per average protein (average protein is 703)	49.91	42.18	22.50
Percent phosphosites per total available selected amino acids	70.7%	22%	8.3%
Ratio phosphosites per total available selected amino acids	2/3	1/5	1/12

Phosphoamino Acid Distribution

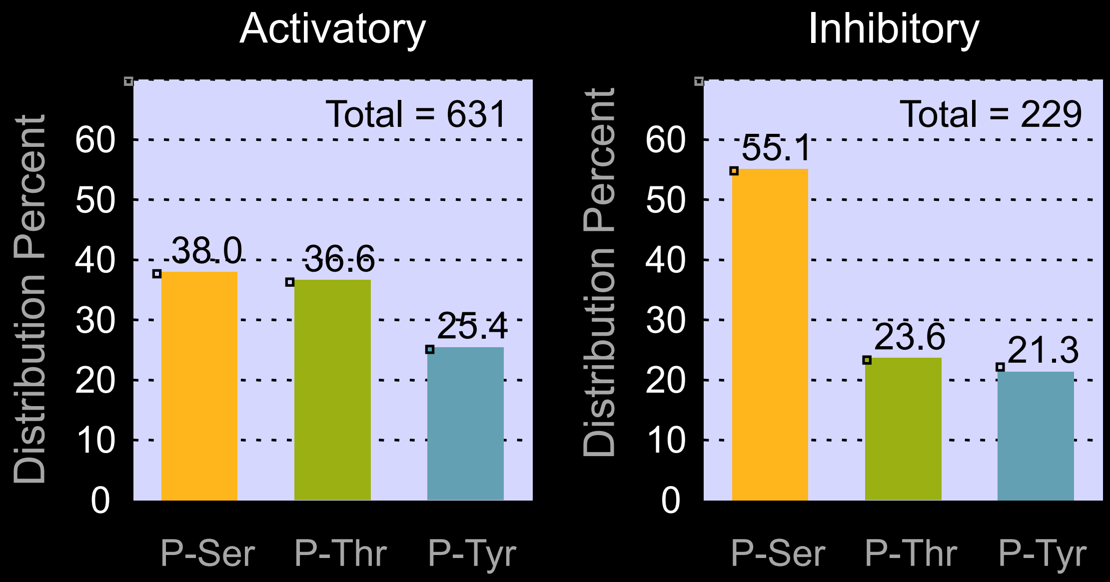
Most Evolutionary Conserved Human Phosphosites



- About 9% of the most conserved phosphosites are found in protein kinases, which represent about 2.5% of all of the human protein genes
- While about 76% of phosphosites in general are P-Ser, in the most conserved phosphosites in protein kinases, P-Thr is the most prevalent

Kinase Phosphoamino Acid Distribution

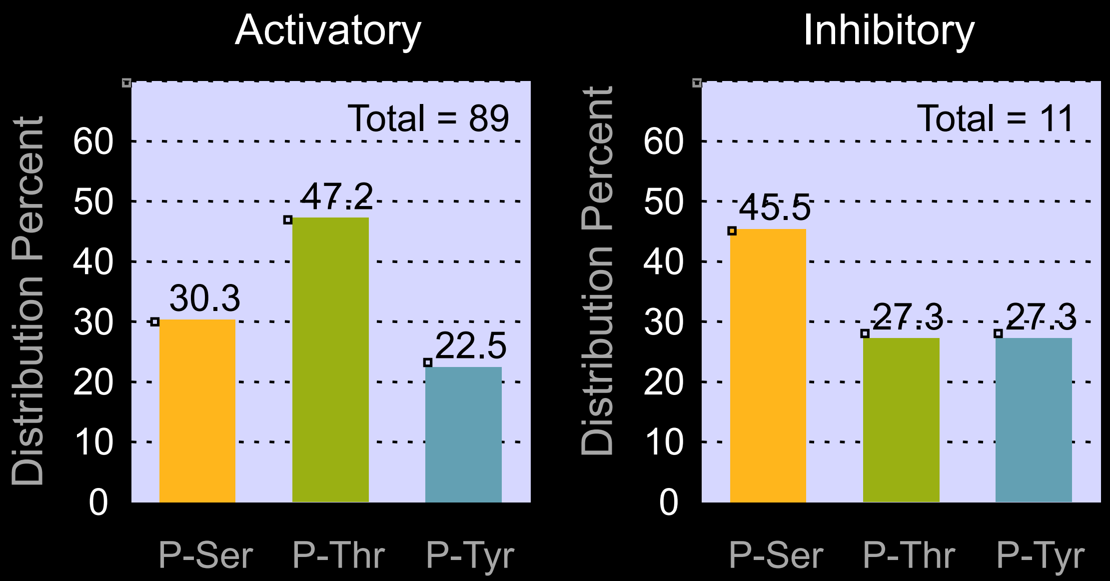
Functional Human Kinase Phosphosites



- In general, functional phosphosites are more likely to be activatory than inhibitory – 2.8:1 ratio

Kinase Phospho-amino Acid Distribution

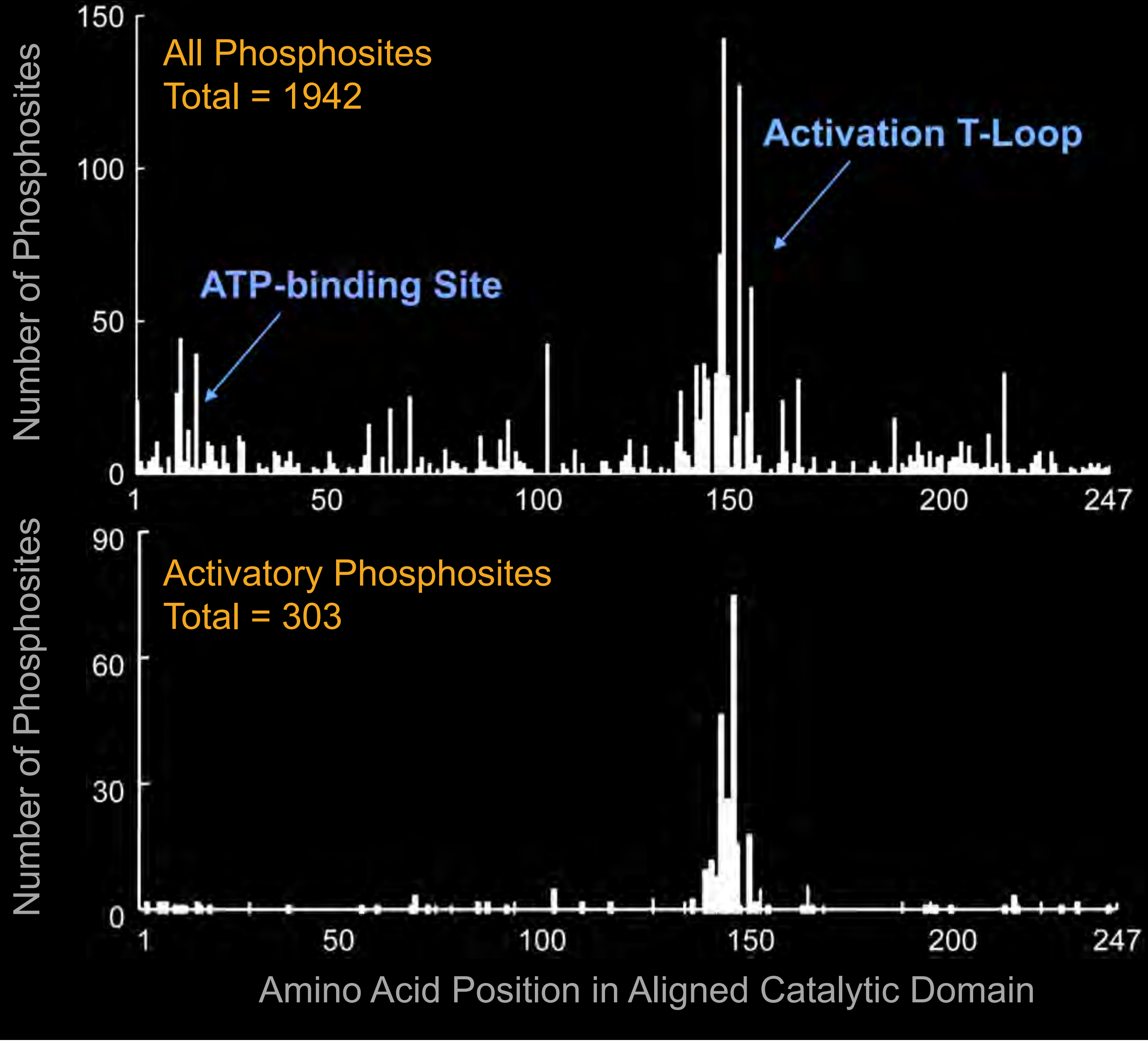
Most Conserved Functional Human Kinase Phosphosites



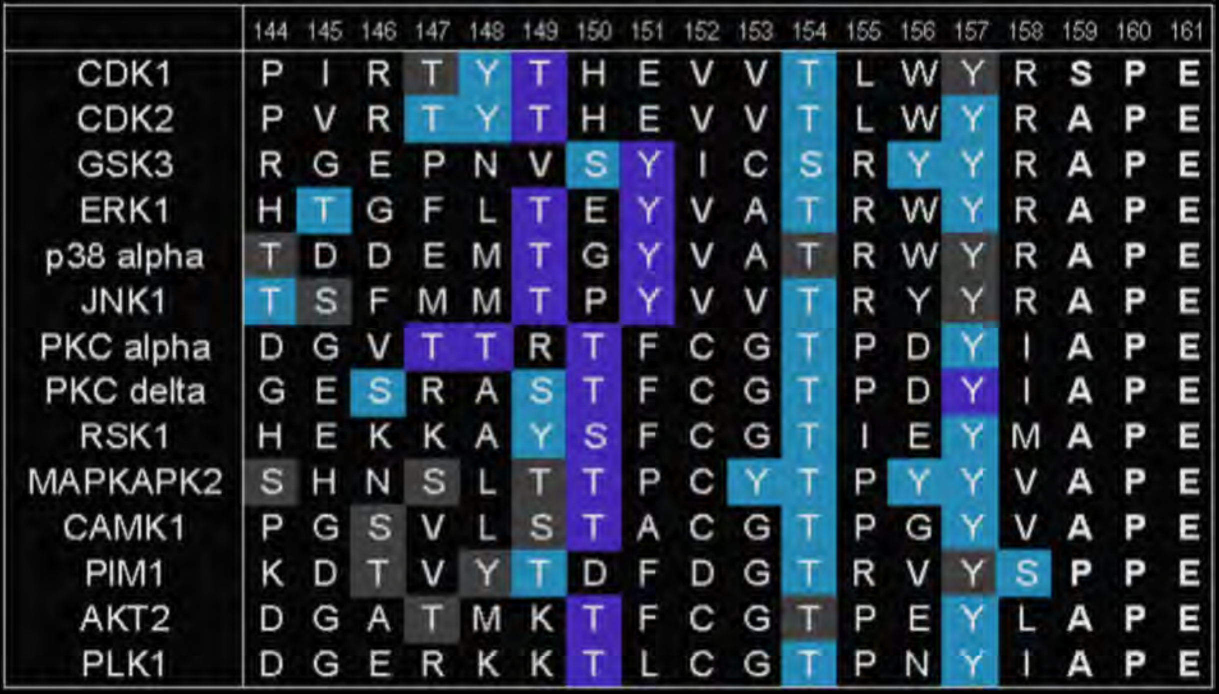
- Functional phosphosites that are conserved are more likely to be activatory than inhibitory in kinases – 8:1 ratio

Human Kinase Catalytic Domain

Distribution of Phosphosites



Phosphosites Between Kinase Subdomains VII and VIII



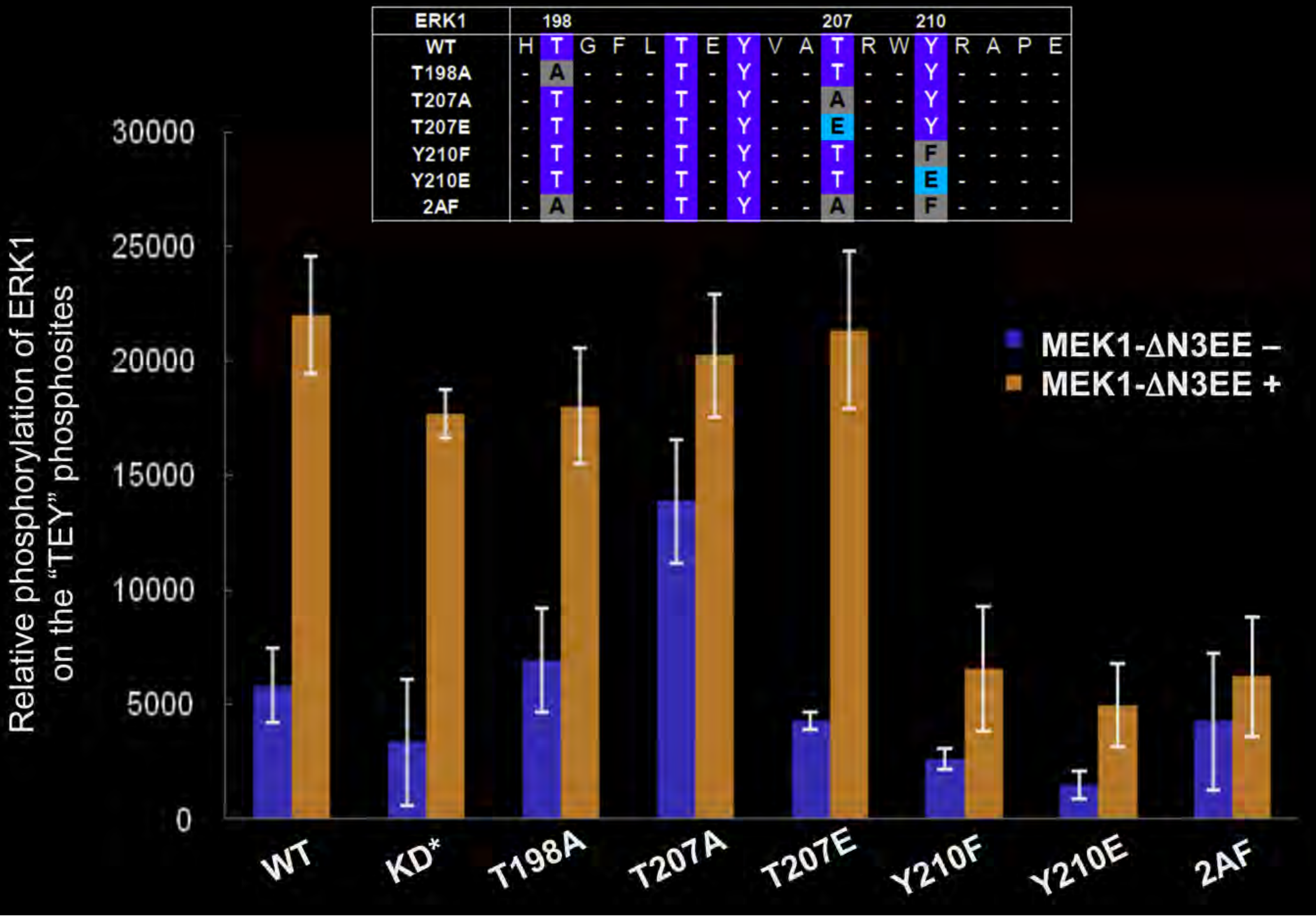
	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161
CDK1	P	I	R	T	Y	T	H	E	V	V	T	L	W	Y	R	S	P	E
CDK2	P	V	R	T	Y	T	H	E	V	V	T	L	W	Y	R	A	P	E
GSK3	R	G	E	P	N	V	S	Y	I	C	S	R	Y	Y	R	A	P	E
ERK1	H	T	G	F	L	T	E	Y	V	A	T	R	W	Y	R	A	P	E
p38 alpha	T	D	D	E	M	T	G	Y	V	A	T	R	W	Y	R	A	P	E
JNK1	T	S	F	M	M	T	P	Y	V	V	T	R	Y	Y	R	A	P	E
PKC alpha	D	G	V	T	T	R	T	F	C	G	T	P	D	Y	I	A	P	E
PKC delta	G	E	S	R	A	S	T	F	C	G	T	P	D	Y	I	A	P	E
RSK1	H	E	K	K	A	Y	S	F	C	G	T	I	E	Y	M	A	P	E
MAPKAPK2	S	H	N	S	L	T	T	P	C	Y	T	P	Y	Y	V	A	P	E
CAMK1	P	G	S	V	L	S	T	A	C	G	T	P	G	Y	V	A	P	E
PIM1	K	D	T	V	Y	T	D	F	D	G	T	R	V	Y	S	P	P	E
AKT2	D	G	A	T	M	K	T	F	C	G	T	P	E	Y	L	A	P	E
PLK1	D	G	E	R	K	K	T	L	C	G	T	P	N	Y	I	A	P	E

Purple, confirmed activation sites; Blue, confirmed phosphosites with unknown function; Grey, potential phosphosites based on similarities

- Kinase Catalytic Domain residues 149-151 most commonly feature activatory phosphosites

ERK1 As a Paradigm

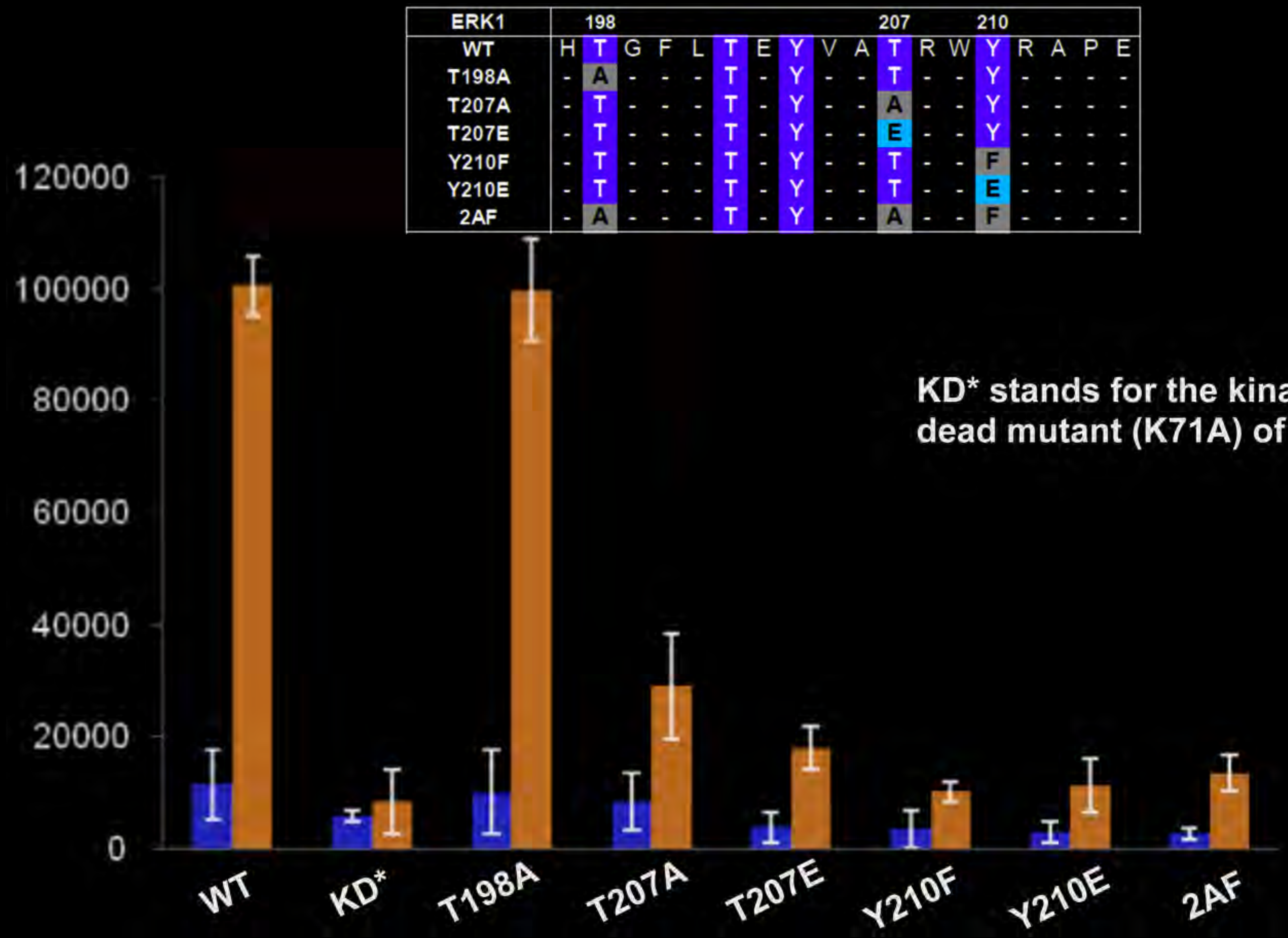
Role Phosphosites Between Kinase Subdomains VII and VIII



ERK1		198									207			210				
WT	H	T	G	F	L	T	E	Y	V	A	T	R	W	Y	R	A	P	E
T198A	-	A	-	-	-	T	-	Y	-	-	T	-	-	Y	-	-	-	-
T207A	-	T	-	-	-	T	-	Y	-	-	A	-	-	Y	-	-	-	-
T207E	-	T	-	-	-	T	-	Y	-	-	E	-	-	Y	-	-	-	-
Y210F	-	T	-	-	-	T	-	Y	-	-	T	-	-	F	-	-	-	-
Y210E	-	T	-	-	-	T	-	Y	-	-	T	-	-	E	-	-	-	-
2AF	-	A	-	-	-	T	-	Y	-	-	A	-	-	F	-	-	-	-

Role of Flanking Phosphosites

Phosphorylation of Myelin Basic Protein by ERK1

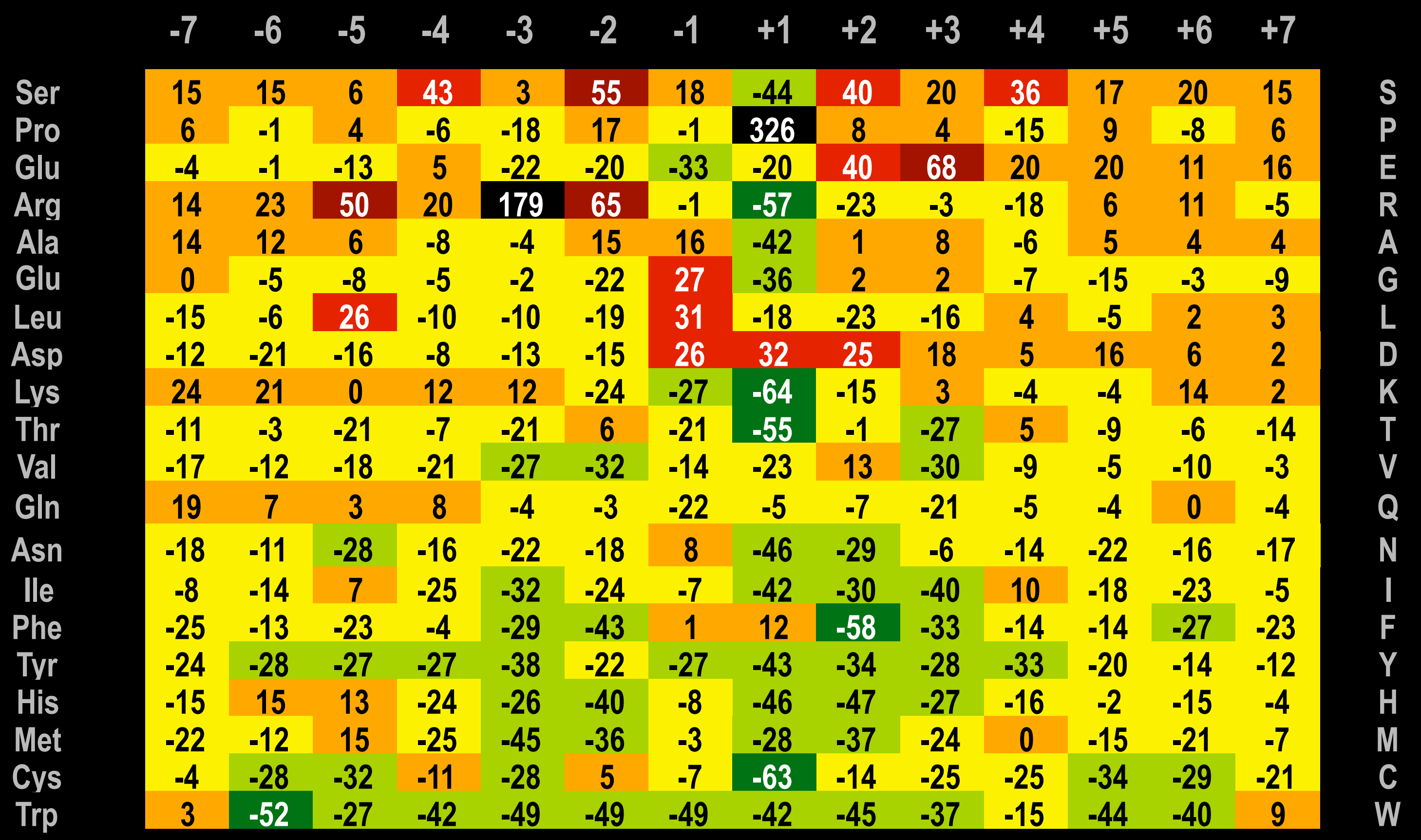


ERK1		198									207			210				
WT	H	T	G	F	L	T	E	Y	V	A	T	R	W	Y	R	A	P	E
T198A	-	A	-	-	-	T	-	Y	-	-	T	-	-	Y	-	-	-	-
T207A	-	T	-	-	-	T	-	Y	-	-	A	-	-	Y	-	-	-	-
T207E	-	T	-	-	-	T	-	Y	-	-	E	-	-	Y	-	-	-	-
Y210F	-	T	-	-	-	T	-	Y	-	-	T	-	-	F	-	-	-	-
Y210E	-	T	-	-	-	T	-	Y	-	-	T	-	-	E	-	-	-	-
2AF	-	A	-	-	-	T	-	Y	-	-	A	-	-	F	-	-	-	-

KD* stands for the kinase-dead mutant (K71A) of ERK1

Human Serine Phosphorylation Sites

Amino Acids Commonly Surrounding Serine Phosphosites



	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	
Ser	15	15	6	43	3	55	18	-44	40	20	36	17	20	15	S
Pro	6	-1	4	-6	-18	17	-1	326	8	4	-15	9	-8	6	P
Glu	-4	-1	-13	5	-22	-20	-33	-20	40	68	20	20	11	16	E
Arg	14	23	50	20	179	65	-1	-57	-23	-3	-18	6	11	-5	R
Ala	14	12	6	-8	-4	15	16	-42	1	8	-6	5	4	4	A
Glu	0	-5	-8	-5	-2	-22	27	-36	2	2	-7	-15	-3	-9	G
Leu	-15	-6	26	-10	-10	-19	31	-18	-23	-16	4	-5	2	3	L
Asp	-12	-21	-16	-8	-13	-15	26	32	25	18	5	16	6	2	D
Lys	24	21	0	12	12	-24	-27	-64	-15	3	-4	-4	14	2	K
Thr	-11	-3	-21	-7	-21	6	-21	-55	-1	-27	5	-9	-6	-14	T
Val	-17	-12	-18	-21	-27	-32	-14	-23	13	-30	-9	-5	-10	-3	V
Gln	19	7	3	8	-4	-3	-22	-5	-7	-21	-5	-4	0	-4	Q
Asn	-18	-11	-28	-16	-22	-18	8	-46	-29	-6	-14	-22	-16	-17	N
Ile	-8	-14	7	-25	-32	-24	-7	-42	-30	-40	10	-18	-23	-5	I
Phe	-25	-13	-23	-4	-29	-43	1	12	-58	-33	-14	-14	-27	-23	F
Tyr	-24	-28	-27	-27	-38	-22	-27	-43	-34	-28	-33	-20	-14	-12	Y
His	-15	15	13	-24	-26	-40	-8	-46	-47	-27	-16	-2	-15	-4	H
Met	-22	-12	15	-25	-45	-36	-3	-28	-37	-24	0	-15	-21	-7	M
Cys	-4	-28	-32	-11	-28	5	-7	-63	-14	-25	-25	-34	-29	-21	C
Trp	3	-52	-27	-42	-49	-49	-49	-42	-45	-37	-15	-44	-40	9	W

Percentage Changes from Expected Random Frequency

Phospho-serine
at 0 position

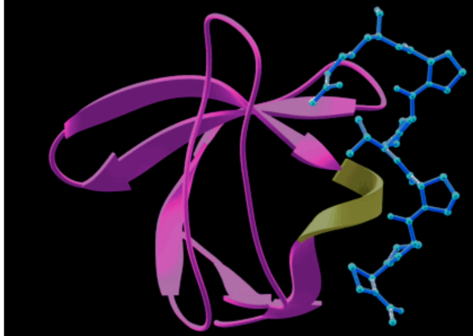
Based on 4035 human sites

% Change from Expected

>100
50 to 100
25 to 49
0 to 24
-25 to -1
-50 to -26
-100 to -51

- Proline appears at the +1 position in ~26% of P-Ser and P-Thr sites

Role of Proline-directed Phosphorylation

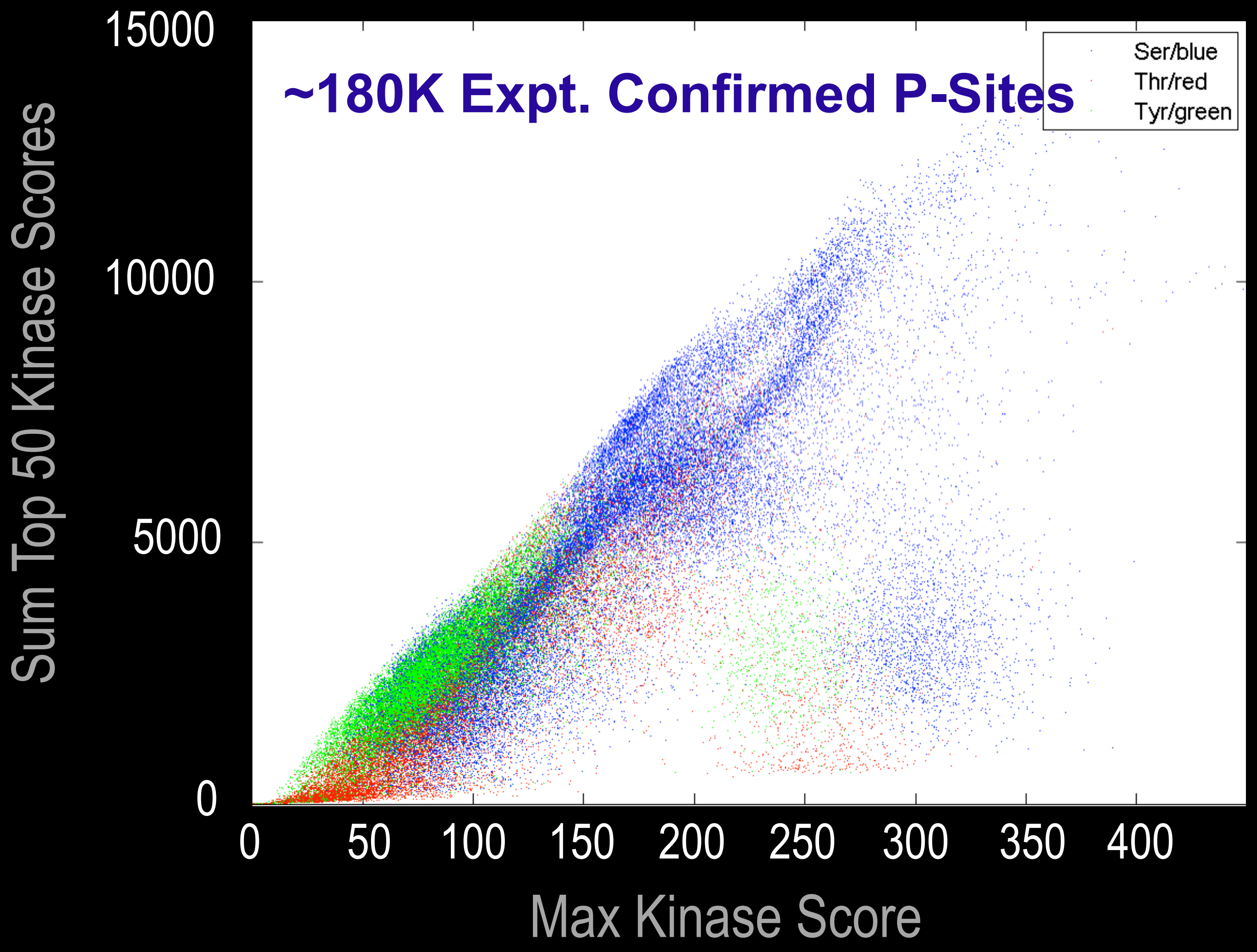



SH3 Domain	WW Domain	EVH1 Domain	GYF Domain
e.g. SEM5	Pin1	EVL	CDC2BP2
Left-hand poly-Pro type II helix	**Pro-Rich**	**Pro-Rich**	**Pro-Rich**
P-X-X-P	Pho-S/Pho-T P-P-X	E/D-F-P-P-P-P-X-D/E	P-P-P-P-G-H-R

Poly-proline-rich binding domains mediate protein-protein interactions. Phosphorylation may regulate these interactions.

KiNEXUS

Human Phosphorylation Sites



Very few phosphorylation sites appear to be targeted with high specificity by only a few kinases. Most phosphorylations are not functionally important.

Role of Hyper-phosphorylation

- PEST sequences are rich in Pro, Glu, Ser and Thr
- PEST sequences are sites of known protein degradation

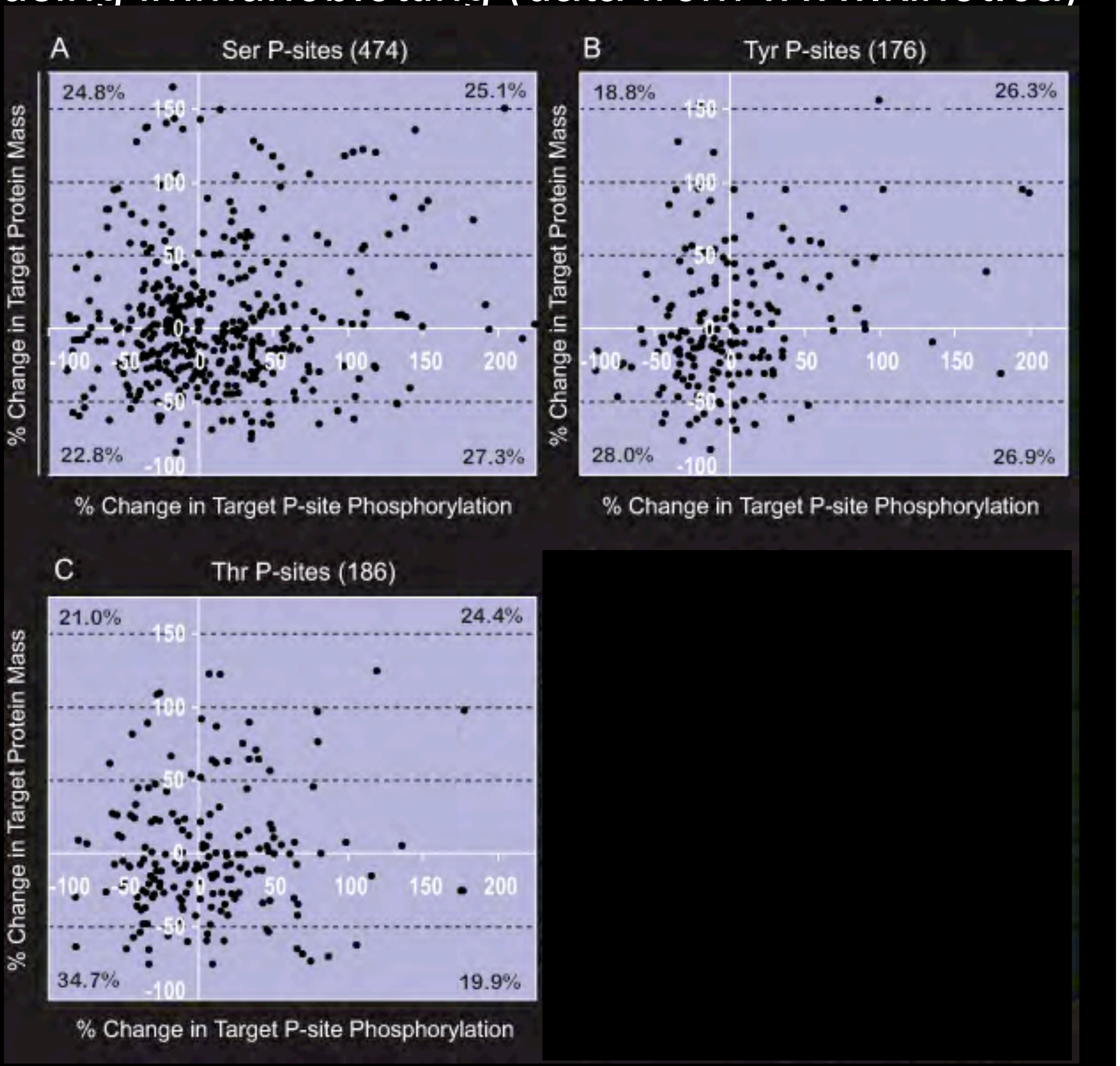
  e.g. cyclins C, D and E; PKC isoforms
- ~ 2,732 PEST sequences of 15 aa length with 3 or more prolines identified in human proteome (~18,619 sequences of 15 aa length with 2 or less prolines also identified - known as ST-rich sites)
- ~42% of PEST sequences appear to be phosphorylatable, compared to ~26% of ST-rich sites, despite more S and T in ST-rich sites
- Phosphorylation may serve to activate PEST sequences

  e.g. plant FRQ and CK1 phosphorylation

Relationship between Protein Phosphorylation and Mass

Based on 80 phospho-sites using immunoblotting (data from www.kinet.ca)

- Serine Phosphosites
  - Positive correlation = 47.9 %
  - Negative correlation = 52.1 %
- Threonine Phosphosites
  - Positive correlation = 59.1 %
  - Negative correlation = 40.9 %
- Tyrosine Phosphosites
  - Positive correlation = 45.7 %
  - Negative correlation = 54.3 %



Role of Hyper-phosphorylation

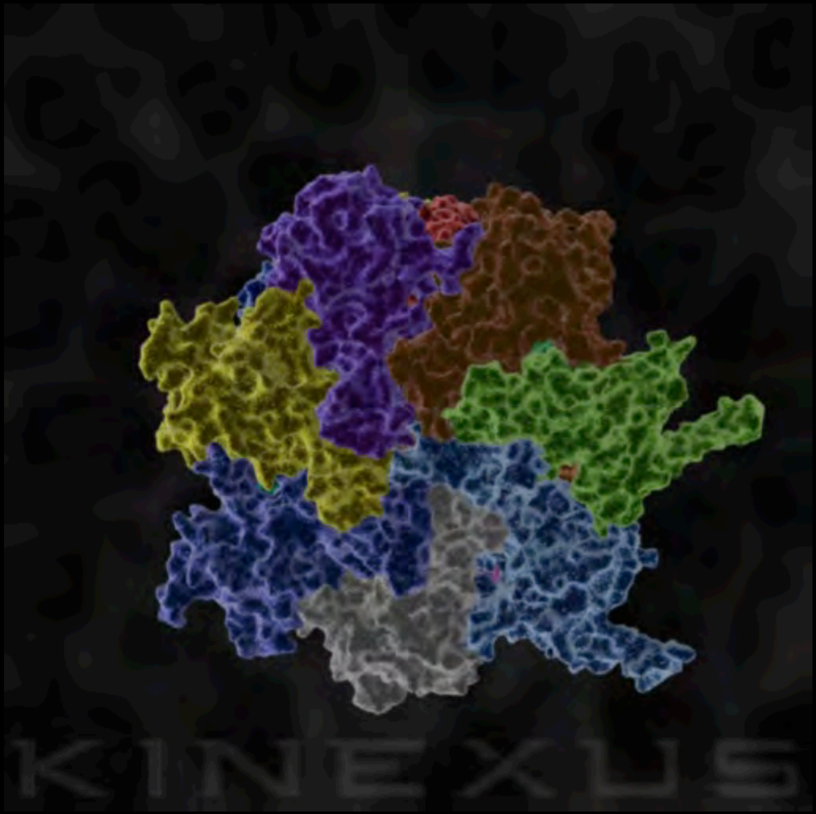
- Hyper-phosphorylation on serine and threonine may mediate the dissociation of protein complexes and protein degradation.
- Conversion to highly charged amino acids may drive protein unfolding
  - Globular proteins are converted to rods when they are coated with detergents such as sodium dodecylsulphate (SDS)
  - This is achieved because of the high negative charge on SDS; the hydrophobic part of the detergent binds to the protein and the negative charge of the sulphate permits solubilization
  - Similarly, the incorporation of highly charged clusters of phosphosites in proteins may also facilitate disruptions of their 3D structures
  - Unfolded proteins should be more susceptible to proteases

Role of Hyper-phosphorylation

20S Proteosome

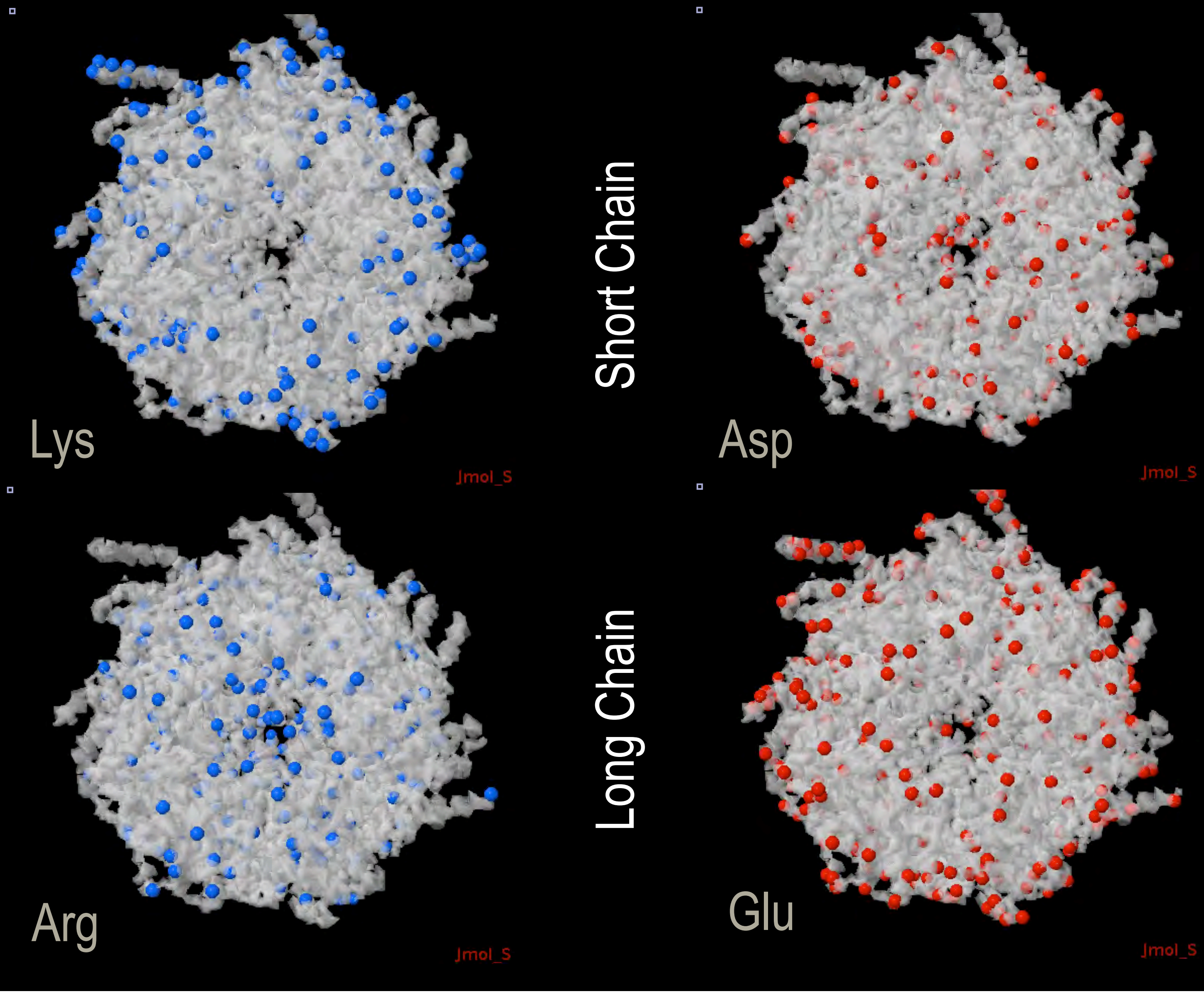
Hyperphosphorylation may facilitate protein degradation

- 20S proteosome mediates protein degradation into amino acids



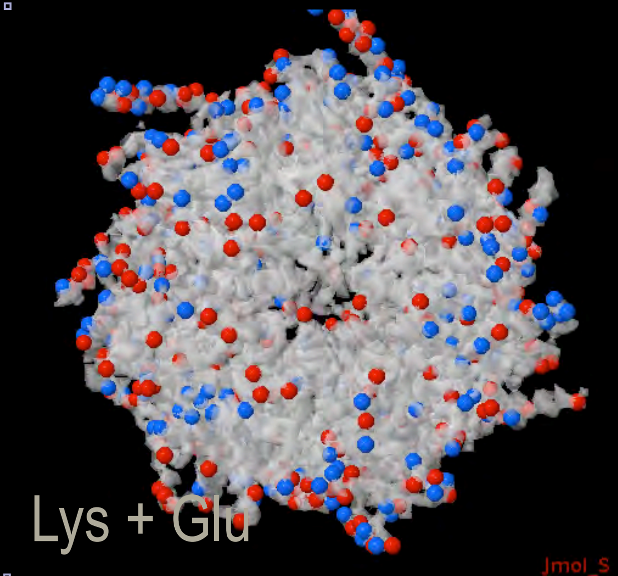
Role of Hyper-phosphorylation

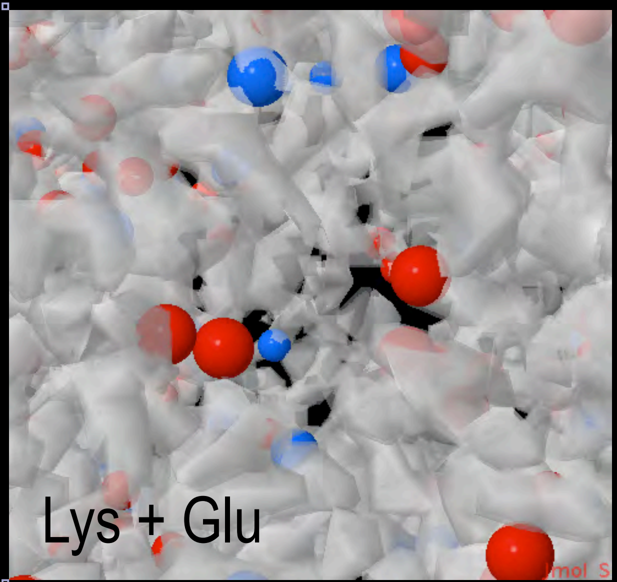
20S Proteosome Active Site

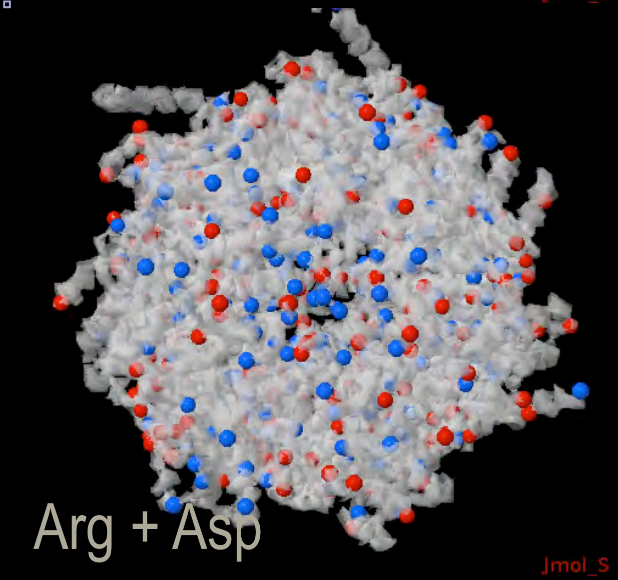


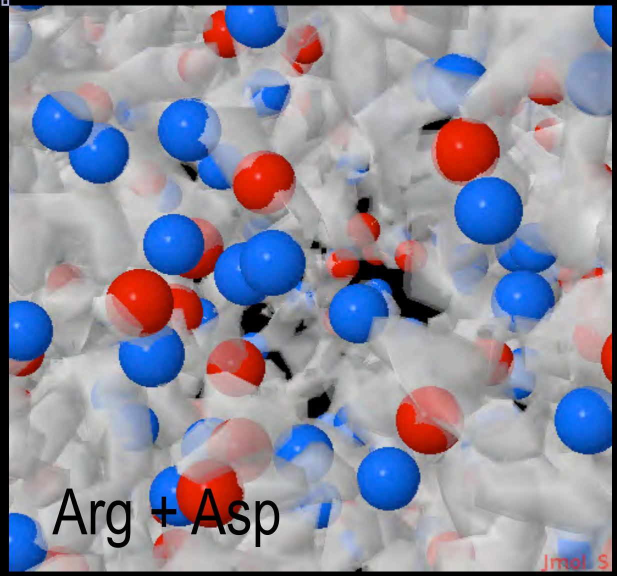
Role of Hyper-phosphorylation

20S Proteosome Active Site



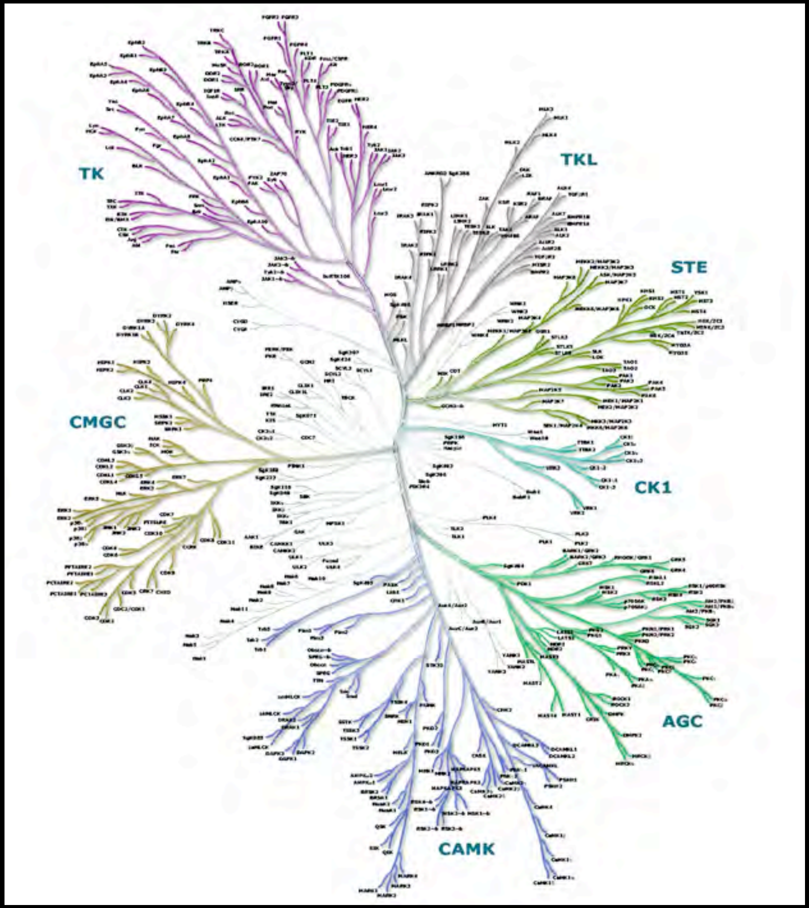






Origin of the Protein Kineome

The Source of the First Protein Kinase Gene



From Manning *et al.* Science 6 December 2002

- Aligned catalytic domains of protein-serine/threonine kinases to derived consensus sequence with 247 aa
- Blast other species and ignore protein kinase sequences
- Got matches with glutamine tRNA ligase (GlnRS; for protein synthesis) and choline kinase (ChK; for phosphatidylcholine biosynthesis)
- GlnRS appears to be the source of the primordial protein kinase

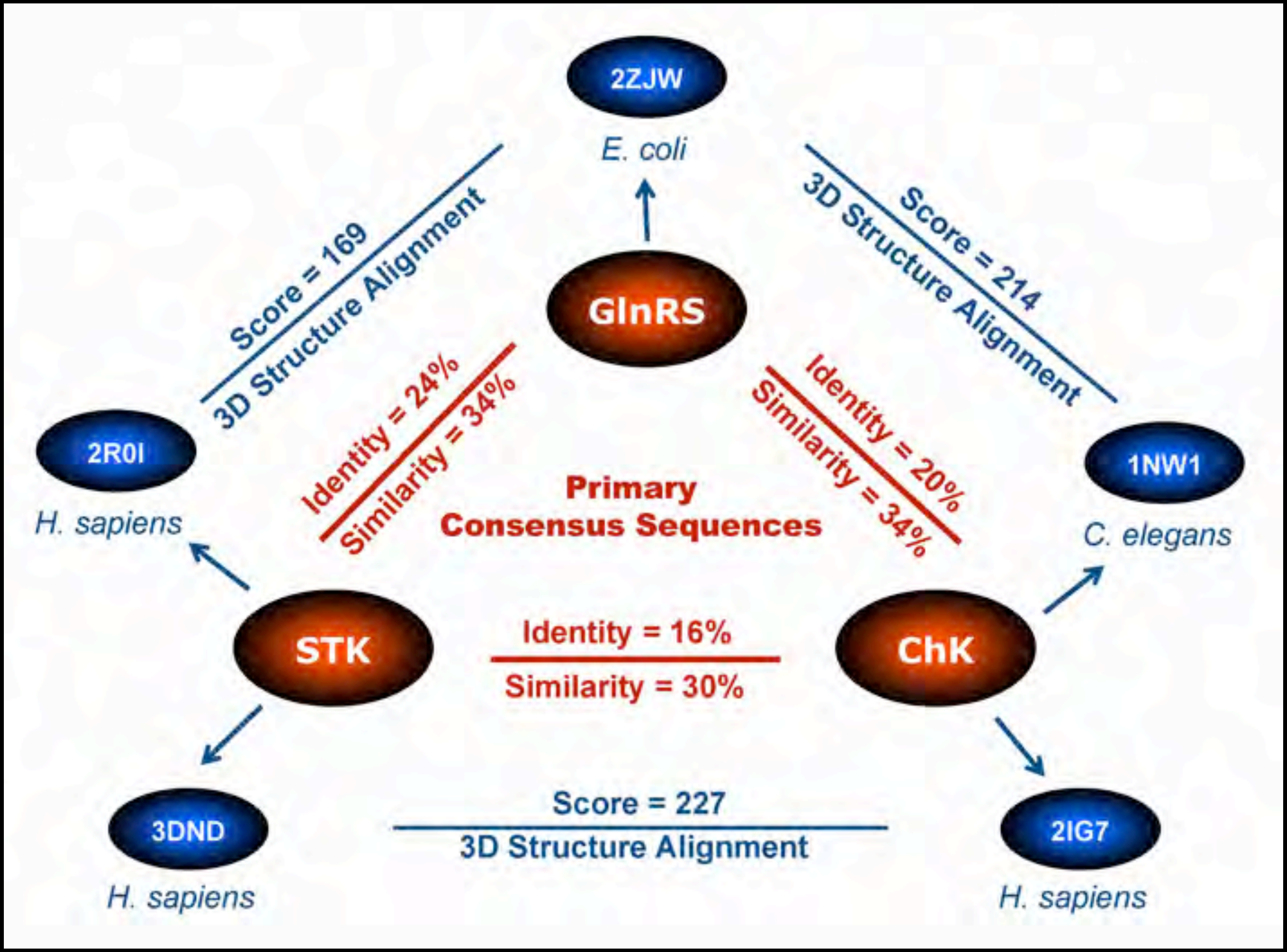
Alignment of Protein Consensus Sequences

```
                      I   I     I   I               G1        II  II      G2                III           G3                    IV G4
GlnRS | F E Q - E C - G V G V V V T P E Q I E E - - - - - - - A V E A V I N R N R E Q L L V E R Y R F N M G L L M G E V R K X L K 57 E
      | + E     E     G   G             +     +               A + +   +     + + R E Q L L   E               L         +   K   L      +
STK   | Y E L L E V L G K G S F G K V Y K A R D K X T G X L V A I K I I K X K D R E Q L L R E I K I L K X : L K H P N I V K - L Y X  K
      |         + +   G         G + +   +           T           + +     +           +   +   +       L K     L                         K
ChK   | G L G P K L Y G V F P N G R L E E F I P S R T L S T E D L R D P E I S X I A E K M A Q F H X L K M P L S K E P K W L F X I E  K

              V   V                                 G5                                                        G6  VI    VI VI VI    VI
GlnRS | T L S L M E Q L R G E A L K F - - H K P G E N Y K - - - T D G Y V V T P N T M E - - L L K Q H L E I T G G V H T R F P P E P N
      |       + M E   L   G     L           K   G +               +                 +       L       H   +         +     R           N
STK   | L Y L V M E Y L S G G D L F D L L K K R G K X L P - - - E E E A R F Y L R Q I L X G L E Y L H S K G X - - I I H R D L K P E N
      |       + + +   L                 + K K     K   L         + E E         L R + + L     L                         H   D L +     N
ChK   | Y L E L V K E L X F X R E A - K I K K L L K I L S Y D L K E E I D N - L R K L L D X L X S P V V F - - - - - C H N D L Q E G N

            G7        VII VII VII VII VII     G8                          VIII VIII VIII  G9            IX  IX    IX
GlnRS | G I L H I G H A K A I N L N F G L A K A N - - - - - G G I C Y L R F D D T N P E K E E E E Y I T S I K N D V E W L G F E P Y K
      |   + L                     + F G L A K               L               +       P E         +       S   K   D +     L G       Y +
STK   | I L L D E D X G H I K L A D F G L A K Q L S X X X X L T T F C G T P E Y M A P E V L L G K X G Y S K K V D I W S L G V I L Y E
      | I L L + E D       +   L   D F     +                 +                 Y                                   +     +       L   +
ChK   | I L L E E D - - K L M L I D F E Y S S Y N P R G F D I A N H F C E W M Y D P F F R A D Y S K Y P T R E X Q L H F I R S Y L G Q

                                        X     G10                     G11         XI        G12   XII   XII
GlnRS | V T Y A S D Y F D Q L Y E W A V E L I R K G L A Y V C H Q T P E E L K 36 E G K A T L R M K L 18 F A P H H R
      | +             F         E     + E L +   K   L           +   P     L      E   K     L R     L         P     R
STK   | L L T G K P P F X G - - E S E L E L L X K I L X X P X L E I P X X L S P  E A K D L L R K L L X  K D P X K R
      |                         E S     E +         +           +       L          + +
ChK   | E - - - - - - - - - - - E S R E E V E D L Y I E V N R F A L A S H L F W  G L W S I V Q
```

- GlnRS = Glutamine tRNA synthetase concensus sequence
- STK = Protein-serine/threonine kinase concensus sequence
- ChK = Choline kinase consensus sequence

Origin of the Protein Kineome

Alignment of Protein Consensus Sequences



Origin of the Protein Kineome

The Emergence of Eukaryotic Cells

- Protein kinases are critical for the development of Molecular Intelligence systems in eukaryotic cells - with larger cells, communication becomes increasingly critical
- Choline kinase catalyzes the phosphorylation of choline, which is the first step in the synthesis of phosphatidylcholine and the main phospholipid found in cell eukaryotic cell membranes
- The duplication of mutation of the glutamine tRNA ligase gene on two separate occasions may have been major steps for the development of successful eukaryotic cells

Credits

- Shenshen Lai
  Ph.D. Student
  UBC

- Javad Safaei
  Ph.D. Student
  UBC

- Dr. Steven Pelech
  President & CSO
  Phone: 604-323-2547 Ext. 10
  E-mail: spelech@kinexus.ca

- Dr. Dirk Winkler
  Lab Manager,
  Kinexus