Cancer signaling networks and their implications for personalized medicine

Creixell, Pau; Linding, Rune

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Cancer signaling networks and their implications for personalized medicine

Pau Creixell

July 19, 2013

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

"Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand."

Albert Einstein (1879-1955)

Contents

Preface

This thesis is submitted as a requirement for obtaining the Ph.D. degree at the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, at the Technical University of Denmark (DTU) and was funded by a DTU scholarship.

All the work was carried out at the Center for Biological Sequence Analysis under the supervision of Professor Rune Linding.

Lyngby, June 2013
Pau Creixell

Abstract

Amongst the unique features of cancer cells perhaps the most crucial one is the change in the cellular decision-making process. While both non-cancer and cancer cells are constantly integrating different external cues that reach them and computing cellular decisions (e.g. proliferation or apoptosis) based on the integration of these cues; this integration and consequently the cellular decisions taken by cancer cells are arguably very distinct from the decisions that would be expected from non-cancer cells. Since cellular signaling networks and its different states are the computational circuits that determine cellular outcome, it is clear to many that these networks will be highly dysregulated in cancer cells. Thus, developing and applying methods that will be capable of mapping and predicting how cancer mutations translate into signaling network perturbations, which could explain cancer development as well as cancer resistance to treatment, represent not only a huge challenge, but also one with potentially extreme benefit for our understanding of the disease and for patients. This thesis summarizes my efforts during the last years in contributing positively to overcome this challenge.

This thesis is divided into six parts. Starting with a brief introduction to the history and some basic concepts of cancer, signaling networks and human protein kinases (part I), we quickly move on to describing existing methods to analyze cancer signaling networks, including methods proposed by us, as well as three of the articles that are part of this PhD thesis (part II). In part III, we illustrate with an article that has been submitted recently, how next-generation sequencing data and mass spectrometry data can be combined to uncover genome-specific signaling networks. In part IV, I describe the two computational methods that I have developed and how they can be integrated with the aim of predicting how signaling networks will be dysregulated in cancer. As a matter of fact, the following part (part V) proves the usefulness of the method by identifying a functional mutation in a group of ovarian clear cell carcinoma cell lines that could cause their resistance to cisplatin treatment. Part VI closes the thesis by summarizing its main points and proposing some future perspectives for the work presented here.

All in all, this work establishes a new framework for the prediction of mechanisms underlying cancer development and evolution which, one would hope, should help close the gap between cancer genotype and phenotype.

Dansk resumé

Cancerceller udviser mange unikke træk, hvoraf ændringer i de cellulære beslutningsprocesser formodentlig hører til blandt de vigtigste. Cancerceller såvel som normale celler integrerer løbende forskellige udefrakommende indtryk til at træffe beslutninger, f.eks. om proliferation eller apoptose. Denne integration hos cancerceller og de deraf afledte beslutninger er distinkt fra den, der observeres hos tilsvarende normale celler. Da cellulære signaleringsnetværk og deres tilhørende tilstande udgør de kredsløb, som styrer cellers adfærd, er det derfor oplagt, at disse netværk ofte vil være fejlregulerede i cancerceller. At udvikle og anvende metoder, som vil være i stand til at forudsige, hvordan mutationer i cancer fører til forandringer i signaleringsnetværk, er en stor udfordring. Men genvinsten er at sådanne metoder vil føre til en bedre forståelse af cancercellers udvikling og lægemiddelresistens og dermed på længere sigt give mulighed for at hjælpe personer, der rammes af cancer. Denne afhandling opsummerer mine bidrag over de seneste år til at løse denne udfordring.

Afhandlingen består af seks dele og starter med en kortfattet introduktion til grundlæggende begreber inden for cancer, signaleringsnetværk og humane proteinkinaser (del I). Herefter følger en beskrivelse af metoder til at analysere cancersignaleringsnetværk – herunder metoder udviklet af os – samt et resumé af tre af de artikler, der udgør en del af denne ph.d.-afhandling (del II). I del III illustrerer vi (med en nyligt indsendt artikel), hvordan data fra next-generation sequencing og massespektrometri kan kombineres til at afdække genomspecifikke signaleringsnetværk. I del IV beskriver jeg de to beregningsmetoder, som jeg har udviklet, samt hvordan de kan kombineres til at forudsige hvordan signaleringsnetværk fejlreguleres i cancer. Den følgende del (V) viser anvendeligheden af denne kombinationsmetode ved at identificere en funktionel mutation, som vil kunne føre til resistens mod behandling med cisplatin i en gruppe af ovarian clear cell carcinoma cellelinjer. Del VI er en opsamling af afhandlingens hovedpointer og afsluttes med en diskussion af perspektiverne samt forslag til videre arbejde.

Denne afhandling etablerer således en ny fremgangsmåde til at beskrive og forudsige de begivenheder og mekanismer, der ligger til grund for udvikling af cancer, og som – med lidt held – kan hjælpe med til at bygge bro mellem de genotypiske og fænotypiske aspekter af denne udvikling.

Acknowledgements

This thesis would not have been possible without the encouragement and support from an immense list of incredible people. My gratitude goes to:

My supervisor Rune Linding. I received from you a perfect combination of guidance and freedom as well as the right research environment to become a more independent and better scientist and person.

My lab mates. In particular, soon-to-be Dr Erwin Schoof, who was in the same s*** at the same time, deserves a very, very special thanks. I don't think either of us knew what was ahead of us when we met back in 2009, but I do not think I could have ever found a better PhD-bro. There are too many situations (going through >400 tissue culture plates, working together until 03:00 in the morning, surfing in Hawaii...) and inappropriate jokes that could be written here. I really hope we stay in touch and keep being friends as well as scientific collaborators for a long time. Always remember to hang lose, mate!
It has been a pleasure to see great people joining the lab during my time in the Linding lab. Cristina Santini has been a constant source of inspiration on what interdisciplinary means. If Erwin was my PhD-bro, Cristina must be my PhD-sister. I have not had enough time to properly meet the most recent additions to the lab (Jinho Kim, Craig Simpson -thanks for checking my English writing for the whole thesis!-, James Longden, Xavier Robin, Oxana Radetskaya, Lene Holberg Blicher -a million thanks for your help translating my abstract to Danish!-, Jesper Ferkinghoff-Borg -thank you for helping with the Danish abstract too!-), but the time we have spent together has been incredibly fruitful and fun, so I would like to thank you for that too. Last but definitely not least, I should thank Antonio Palmeri, as our time in the lab together has been brief but extremely productive, as should be clear from the cover of this PhD thesis ;).

Professor Søren Brunak for welcoming our lab to CBS. It has been a great pleasure to be a PhD student at Center for Biological Sequence analysis, where I have been surrounded by many helpful scientists and collaborators (including Chris Workman, Fred De Masi, Tejal Joshi, Greg Slodkowicz and Juliet Frederiksen). Very special thanks go to Morten Nielsen, Agata Wesolowska (for help with many questions regarding this PhD thesis as well as being a great scientific collaborator) and Ramneek Gupta.

The co-authors on my publications as well as past and current collaborators. It has been a pleasure to work with all of you and I hope we will have an opportunity to work together in the future. A very special thanks goes to members of the Erler, Bodenmiller, Turk and Pawson Labs. Perhaps special thanks to Lara, Tom, Alejandro and Janine for the nice time working together on the different collaborative projects we are running together. Adrian Pasculescu and Chris Tan also deserve to be mentioned here for the great projects we were able to do together.

A massive thanks to the CBS systems admin team (John, Kristoffer, Peter, Olga, Hans Henrik) for being truly research-oriented and always available. I have to pay my deepest respects to John Damm Sørensen for handling my computational jobs with so much care, extending running time of jobs (if necessary, late in the evenings) and, in essence, making possible the many genetic algorithm runs that constitute KINspect, one crucial part of this PhD thesis.

The CBS office administration deserves a massive thanks too. Dorthe, Lone, Annette, Karina and Marlene, I have been blessed to have you on my side when receipts, travel plans or issues with airport lounge access have become more difficult to handle than any other scientific challenges ;).

I should also thank many of my friends who have made this not only possible but also enjoyable. I should here include life-long friends (Quim Gomez, Guillem Marine and the rest of my childhood friends, including those who I have play football with at Junior F.C.), as well as people I have met along the way in Oxford (especially, Sebastien Grifnee, Josh and Kate Woodward and Beata Margitay), London and Copenhagen (Albert, Lidia, Toni, Eli, Marc, Octavi...).
Sebastien, you will never be able to read these words, but I would like to write a few anyway. I was obviously devastated to hear that treatment was not going well a few days before writing this thesis and even more when I heard about your passing away a week before finishing it. This PhD thesis is dedicated to your memory and I really hope one day we will understand enough to cure people like you.
My huge friend, Carles Yurss, deserves an enormous thanks for having been there when I have needed him and visited me everywhere I have lived in. I really hope we can count on each other for many years to come yet.

My highest gratitude to my family. I could obviously never be here without my dad, Joan, and my mom, Yoya, but I should also thank them for having given me the freedom to make mistakes and learn from them - a very scientific raising, one could say. My brother, Jan, and my sister, Maria, who have been siblings and friends at the same time. I have learnt a lot from both of you and I am deeply proud of being able to count you as my brother and sister. And to the rest of my family, perhaps especially to my granddad, Jordi, for the multiple pleasant and fun days we could spend together and my granddad, Josep, for being the spark that initiated my passion for science when he showed me a cell under a microscope when I was not even ten years old. My cousin, Nene, should also be mentioned here as she has always been a source of good scientific advise.

And lastly I will always be grateful to my partner, Marta, for a million things including believing in me, being understanding when I have had to be in the lab for far too long or abroad in conferences, for those evenings and weekends when I am there physically but my mind is on the science, for making it all even fun, and ultimately for accepting the challenge of a different life.

Papers included in the thesis

- **Pau Creixell**, Erwin M. Schoof, Janine T. Erler, Rune Linding. *Navigating cancer network attractors for tumor-specific therapy*. Nature Biotechnology, 30:842–848, 2012.

- **Pau Creixell**, Erwin M. Schoof, Chris S.H. Tan, Rune Linding. *Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues*. Philosophical Transactions of the Royal Society B, 367:2584–2593, 2012.

- Abel Gonzalez-Perez*, Ville Mustonen*, Boris Reva*, Graham R.S. Ritchie*, **Pau Creixell**, Rachel Karchin, Miguel Vazquez, J. Lynn Fink, Karin S. Kassahn, John V. Pearson, Gary Bader, Paul C. Boutros, Lakshmi Muthuswamy, B.F. Francis Ouellette, Jüri Reimand, Rune Linding, Tatsuhiro Shibata, Alfonso Valencia, Adam Butler, Serge Dronov, Paul Flicek, Nick B. Shannon, Hannah Carter, Li Ding, Chris Sander, Josh M. Stuart, Lincoln D. Stein, Nuria Lopez-Bigas and the ICGC Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. *Computational approaches to identify functional and driver variants in cancer genomes*. Manuscript accepted for publication in Nature Methods.

- Erwin M. Schoof*, **Pau Creixell***, Adrian Pasculescu*, Agata Wesolowska-Andersen, Ramneek Gupta, Rune Linding. *Uncovering hidden signaling networks by genome-specific mass spectrometry*. Manuscript submitted to Nature Biotechnology.

* These authors contributed equally.

Papers not included in the thesis

- Jonathan So, Adrian Pasculescu, Anna Y. Dai, Kelly Williton, Andrew James, Vivian Nguyen, **Pau Creixell**, Erwin M. Schoof, John Sinclair, Miriam Barrios-Rodiles, Jun Gu, Aldis Krizus, Ryan Williams, Marina Olhovsky, James W. Dennis, Jeffrey L. Wrana, Rune Linding, Claus Jorgensen, Tony Pawson, Karen Colwill. *Systematic analysis of kinases in TRAIL-induced apoptosis identifies targets for combination therapy*. Manuscript submitted to Molecular Systems Biology.

- **Pau Creixell**, Rune Linding. *Cells, shared memory and breaking the PTM code*. Molecular Systems Biology, 8:598, 2012.

- Chris S.H. Tan, Erwin M. Schoof*, **Pau Creixell***, Adrian Pasculescu, Wendell A. Lim, Tony Pawson, Gary Bader, Rune Linding. *Response to Comment on "Positive selection of tyrosine loss in metazoan evolution"*. Science, 332:917, 2011.

- Adrian Pasculescu, **Pau Creixell***, Erwin M. Schoof*, Marina Olhovsky, Ruijun Tian, Rachel Vanderlaan, Tony Pawson, Rune Linding, Karen Colwill. *CoreFlow: A flexible playground for integration, analysis and modelling of complex biological data*. Manuscript submitted to Journal of Proteomics.

- **Pau Creixell**, Erwin M. Schoof, Lara Perryman, Cristina Costa Santini, Antonio Palmeri, Morten Nielsen, James Longden, Agata Wesolowska, Ramneek Gupta, Marchel Stuiver, Honor M. Rose, Philipp Selenko, Janine Erler, Ben Turk, Bernd Bodenmiller, Rune Linding. *Kinome-wide discovery of network-attacking mutations in cancer*. Manuscript in preparation.

- **Pau Creixell***, Erwin M. Schoof*, Lara Perryman*, Agata Wesolowska-Andersen, Ramneek Gupta, Hiroaki Itamochi, Bernd Bodenmiller, Janine Erler, Rune Linding. *Network drivers of drug resistance in ovarian cancer*. Manuscript in preparation.

* These authors contributed equally.

Abbreviations

MAP2K3	Mitogen-activated protein kinase kinase 3
MS	Mass spectrometry
NGS	Next-generation sequencing
PSSM	Position-specific scoring matrix

Part I

Introduction

Chapter 1

Cancer: Past, present and future.

Despite increasing rates of incidence and our relative inability to fundamentally understand it or treat it, cancer is not a modern disease. The oldest descriptions in humans dates back to ancient Egypt, circa 3000 BC (American Cancer Society, 2012; Mukherjee, 2010), and has recently been described in Neanderthal fossil records more than 120,000 years old (Monge et al., 2013). In this chapter, an extremely concise history of the disease is provided, with a special focus on the most recent developments following the genomic revolution of this century. Next, current and alternative paradigms in the interpretation of cancer mutations are discussed, and we continue by addressing the relevance of signaling networks in cancer disease. A section on human protein kinases, key components of signaling networks, completes this first chapter.

1.1 Brief history of cancer before and after the genomic revolution of 2000

Cancer history in modern humans dates back to early human history, with its oldest description being 5000 years old (American Cancer Society, 2012; Mukherjee, 2010). This description of eight breast tumors was found as part of an Egyptian textbook on trauma surgery called the Edwin Smith Papyrus and it is finished with the conclusive and prevailing statement "There is no treatment" (American Cancer Society, 2012). Recent discoveries from fossil records dating back 120,000 years indicate that cancer also inflicted Neanderthals (Monge et al., 2013). Despite its old history, it would not be until around 400 BC when the disease would receive its current name *cancer*, from one of the fathers of medicine Hippocrates, who used the greek word for crab *carcinos* to describe the finger-like projections that many tumors present (American Cancer Society, 2012; Mukherjee, 2010).

In coming centuries, the advance of general medicine, which has resulted in better treatment and resolution of other diseases, has lead to a significant raise in human life expectancy. This combined with an increasing exposure to mutagenic sources and other cellular insults such as tobacco, chemical agents or UV radiation, has transformed cancer from an almost anecdotal into the epidemic disease that is today. Nonetheless, in more recent decades, important discoveries regarding fundamental cancer processes (such as the mutational basis of cancer, its evolution or the development of metastasis) and its treatment (including radiotherapy and chemotherapy or more recently combination targeted therapy or immunotherapy) have been made by scientists of the magnitude of Marie Curie or Sydney Farber and more recently scientific teams lead by researchers like Harold Varmus, J. Michael Bishop or Robert Weinberg (Mukherjee, 2010).

The turn of the 21st century marked a genomic revolution, epitomized by the publication of the human genome project in 2001 (the International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). This revolution has had clear consequences for cancer research and today there are hundreds of tumors that are being sequenced in the hunt for molecular aberrations that ultimately might lead to a more effective treatment of the disease (Wong et al., 2011). While cancers had been described with some unifying features or hallmarks that to some extend helped our understanding of the disease (Hanahan and Weinberg, 2000, 2011), cancer sequencing revealed a high inter-patient disparity in their tumor genetic mutations (Vogelstein et al., 2013), which led to the hard realization that the interpretation of cancer sequencing data would be the real bottleneck separating generation of new data and generation of new knowledge that could translate into better therapies (Yaffe, 2013). This interpretation gap can be illustrated as the comparison between the number of cancer genome mutations reported and the number of mutations regarded as playing an important role in cancer (Figure 1.1).

1.2 Current and alternatives paradigms in the interpretation of cancer mutations

Since the beginning of this cancer genome era, several paradigms have been proposed to facilitate our understanding of this sequencing data tsunami, with different degrees of success and popularity. Perhaps two of the most established paradigms are the cancer driver/passenger and the oncogene/tumor suppressor classification systems. The cancer-driver paradigm differentiates between mutations that are causally implicated in oncogenesis and confer growth advantage (driver mutations) and mutations that are simply bystanders that occur as a result of the high genomic instability of tumors but do not confer growth advantage to cancer cells (passenger mutations) (Stratton et al., 2009) as shown in Figure 1.2. On the other hand, the oncogene/tumor suppressor paradigm defines as proto-oncogenes and oncogenes those genes that would, under physiological conditions, promote cell growth (proto-oncogenes), but when activated in cancer (oncogenes) lead to an increased cellular proliferation. With the

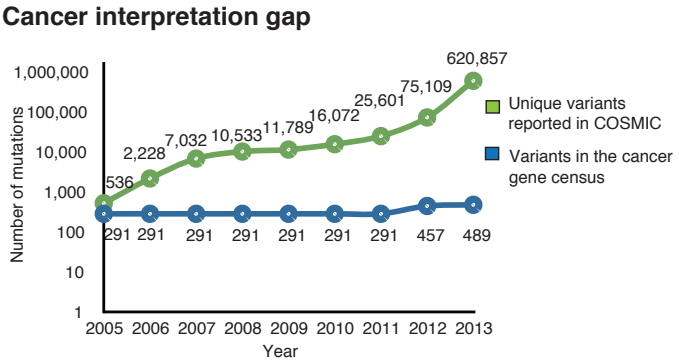


Figure 1.1. Interpretation Gap Evolution. The upper line shows the number of somatic cancer mutations reported in the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes et al., 2001) every year compared to the number of mutations regarded as playing a driving role in cancer (Futreal et al., 2004), in the lower line. Note the Y axis in log scale.

opposite effect, tumor suppressors would inhibit cellular growth under physiological conditions, and when inactivated in cancer cells would promote uncontrolled tumor proliferation (Stehelin et al., 1976). For simplicity, an automotive analogy is often used, where proto-oncogenes can be considered as the accelerator on a car and tumor suppressors are considered its brakes.

While both paradigms have been proven helpful in the identification of some important cancer mutations and genes (such as *BRAF V600E*, present in more than 50% of all malignant melanomas (Davies et al., 2002), or the first cancer gene ever described, the oncogene *src* (Stehelin et al., 1976)), the complexity and non-linearity of the circuits that drive decision-making processes in cells have challenged their limits, as evident from the increasing gap in Figure 1.1. More precisely, at least three fundamental principles of cellular systems that have now been clearly established represent clear oppositions to these relatively simple descriptions of cancer cells. Firstly, given the fact that a protein can be (in)activated by different genetic perturbations (or as we name them, analogous mutations), it cannot be assumed that strong signs of positive selection will be evident and point to specific mutations as being drivers (Creixell et al., 2012). Secondly, it has now been shown that two different mutations that do not drive cancer development by themselves when appearing separately, can lead to cancer when they appear together in the same cell or even in neighboring cells (Wu et al., 2010), in what could be described as two passengers becoming drivers or, as we call them, synthetic oncogenes (Creixell et al., 2012). Finally, a groundbreaking study published in Science in 2005 (Janes et al., 2005) demonstrated that JNK's function in relation to apoptosis (i.e. whether it promotes or inhibits apoptosis) can only be predicted if the different cues and other parts of its network are taken into account, as

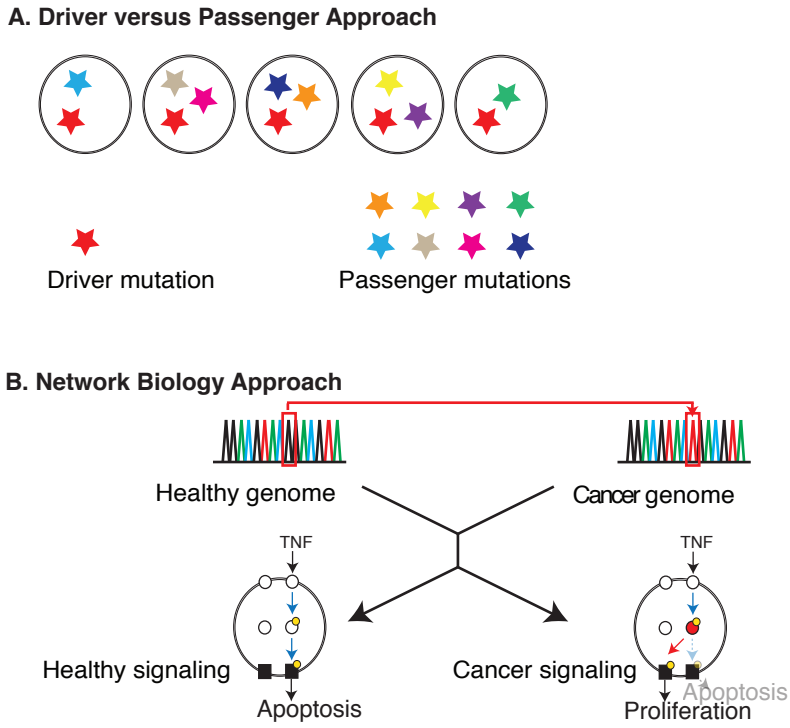


Figure 1.2. Current and alternative cancer paradigms. A. The driver/passenger paradigm largely relies on finding positive selection for the same individual mutation in different tumors. B. In our network biology approach, we predict the impact of mutations at the signaling network level, regardless of whether a given mutation has been seen in different tumors or not, thus aligning with the current trend of personalized medicine.

it can play both a pro-apoptotic and anti-apoptotic cellular role. By extension, binary classifications of protein and mutation function that consider genes and proteins in isolation are unlikely to provide a systematic precise prediction of cellular outcome, including and (one could argue) especially in cancer. Several of these challenges are further discussed in Chapter 3.

1.3 Cancer and signaling networks

While the power of genomics has only been used more recently to support it, the link between cancer and signaling networks was established much longer ago. Once it was clear that cancer cells behave and respond to cues differently from normal cells and

this difference is fundamental in their pathogenesis, it became obvious that perturbations in signaling networks were at the core of the disease, as reviewed by Vogelstein and colleagues (Vogelstein and Kinzler, 2004). The uprise of cancer genomics has not only confirmed but also reinforced the importance of signaling networks and kinases in cancer, as the kinase domain is the domain most often encoded by cancer genes (Futreal et al., 2004). Not surprisingly, many of the proteins that have been given crucial roles in cancer (e.g. Src, BRAF, Abl or the many tyrosine kinase receptors) are human protein kinases themselves.

With the intention of providing an alternative predictive approach to cancer mutation interpretation that would not suffer from the same defects as previous methods and that would focus on signaling networks, we hypothesized that one could try to accurately predict the cellular effect of mutations that affected signaling networks, as these networks crucially regulate cellular decision-making processes perturbed in cancer (Figure 1.2).

1.4 The human kinome - kinase regulation, activation and substrate specificity

Shortly after the publication of the human genome, Gerard Manning and his colleagues published, what is considered a reference article for researchers working on signaling and kinases in particular (Manning et al., 2002). In this article, they defined the human protein kinome superfamily, with 518 active kinases and 106 pseudo-kinases (human proteins that while containing a kinase domain, this domain has lost its catalytic activity), which represents one of the biggest protein superfamilies as well as a key component of signaling networks, thanks to the unique catalytic properties of the kinase domain, which allows signal transduction by phosphorylation (Hanks et al., 1988) (Figure 1.3).

Precisely these 106 pseudo-kinases provided good evidence for amino acid residues that play a critical in Magnesium coordination, ATP binding and phospho-transfer, all of which essential processes for a kinase domain to perform its catalytic activity (i.e. phosphorylation of substrates). Some of these critical residues appeared to be mutated in these different kinases, which provided an explanation for their lack of activity (Zeqiraj and van Aalten, 2010) (Figure 1.4).

In addition to coordination of Magnesium, ATP binding and phospho-transfer, most kinase domains are inactive until they are activated by phosphorylation of one or more residues in their activation segment. This segment consists of a loop that connects the N-lobe and the C-lobe of the kinase domain between two conserved linear motifs (DFG and APE motif). Phosphorylation of the activation segment favors the transition towards a closed active conformation of the domain that will allow phosphorylation of substrates (Hanks and Hunter, 1995; Johnson et al., 1996; Huse and Kuriyan, 2002; Nolen et al., 2004) (Figure 1.4).

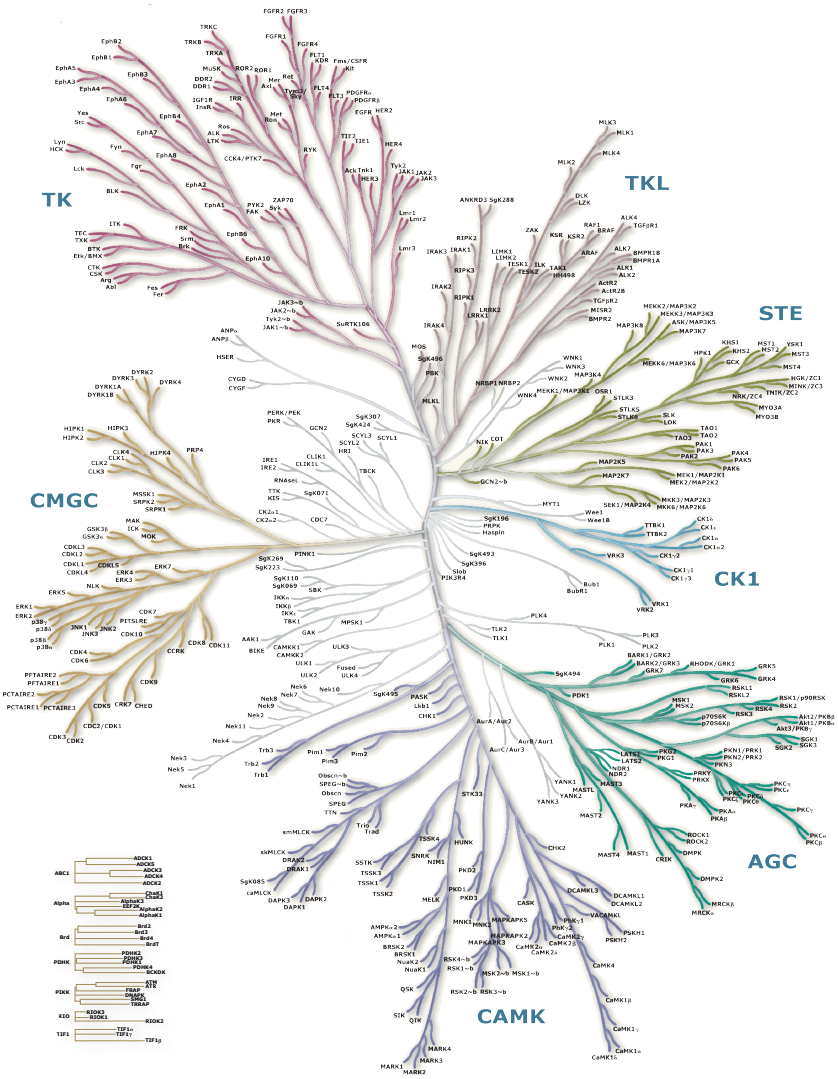


Figure 1.3. The human kinome. Phylogenetic tree illustrating the sequence relationship between the different families of human protein kinases.

Finally, each kinase domain needs to recognize and phosphorylate specific substrates. This rather complex process of selecting substrate encapsulates at least two types of selectivity; at a higher or cellular level, substrate protein specificity, which is driven by different factors such as cellular localization, co-expression of kinase protein and substrate or scaffolding adaptor proteins (Linding et al., 2007); and at a lower or molecular level, peptide specificity, where specific residues within the kinase

domain (also known as determinants of specificity) determine specific preferences at different positions of the substrate peptide (Turk, 2008) (Figure 1.4 and Figure 1.5).

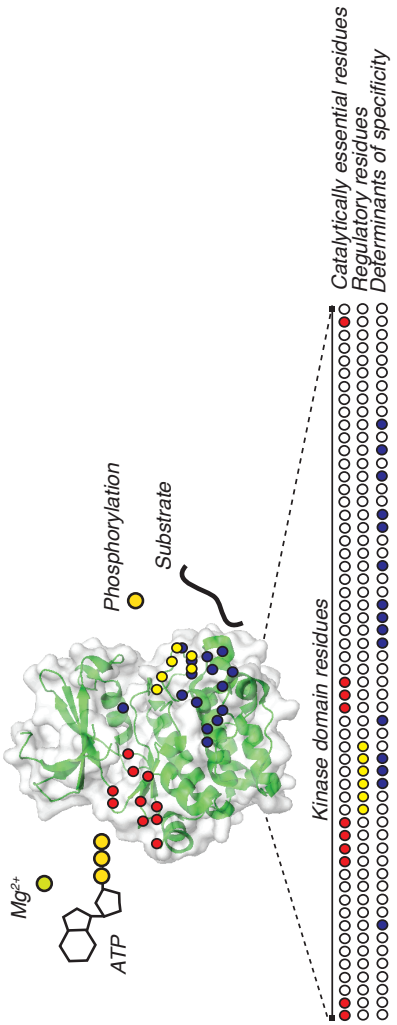


Figure 1.4. Functional residues in the kinase domain. Several specific residues of the kinase domain play crucial roles in activation (catalytically essential residues in red), regulation (regulatory residues in yellow) and substrate specificity (determinants of specificity in blue). An important challenge that is part of this PhD project aims at determining which residues belong to these different categories.

As will be obvious from Chapters 6 and 7, kinase peptide specificity is a major topic of this PhD thesis, especially because we have developed a framework, KINspect,

to predict peptide substrate specificity from kinase domain sequence, crucial for our assessment of cancer mutations in signaling networks (Figure 1.6). Peptide specificity for a given kinase domain can be investigated experimentally by deploying Oriented Peptide Libraries (OPL), where purified kinase domains are exposed to an immense library of peptides that contain random combinations of residues in every position except one fixed residue in a given position, thus making it possible to extrapolate substrate peptide preferences (Yaffe et al., 2001; Hutti et al., 2004; Turk et al., 2006). Matrices resulting from these experiments are subsequently quantified and Position-Specific Scoring Matrices (PSSM) generated. For simplicity and more visual appeal, these matrices are often presented as sequence logos (Figure 1.5).

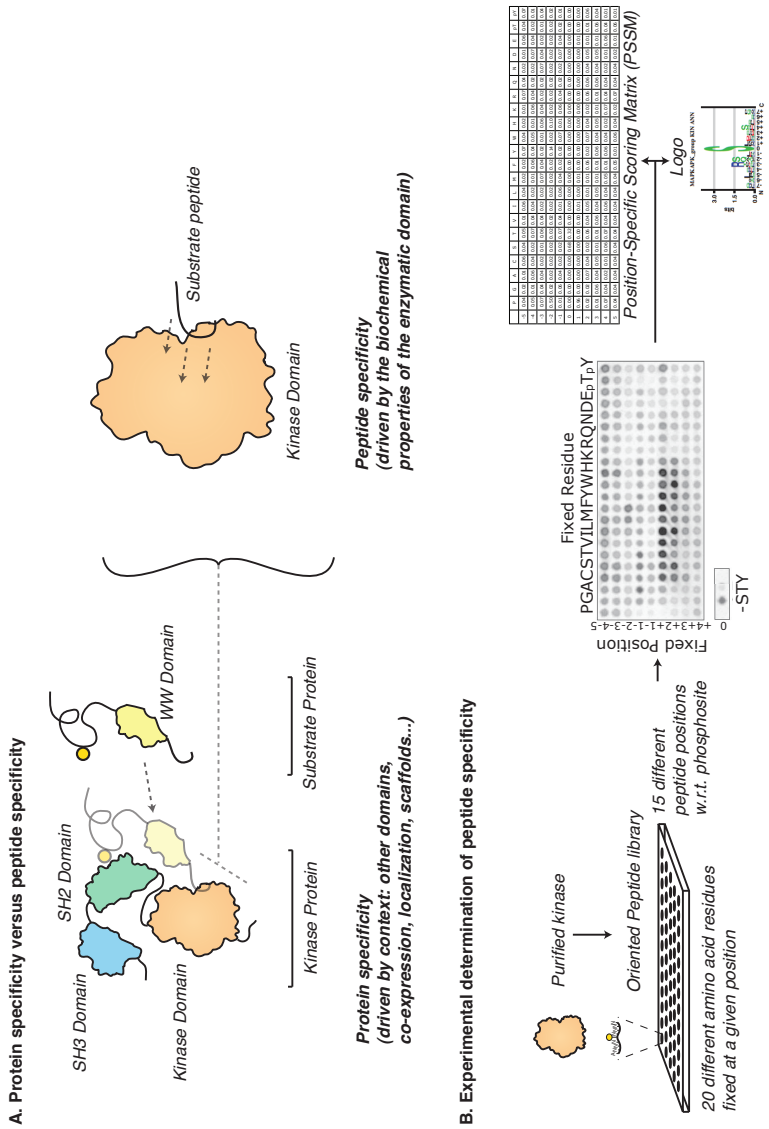


Figure 1.5. Kinase specificity. A. Cellular enzyme specificity, such as kinase specificity, encapsulates two distinct mechanisms of specificity: protein specificity and peptide specificity. Protein specificity determines the interaction between the whole kinase protein and its protein substrate and it is driven by processes such as interactions between other domains and motifs (e.g. SH2 & phospho-tyrosine in this figure), co-expression of the two proteins, cellular localization, scaffold proteins, etc. Peptide specificity is solely driven by the sequence and structure of the kinase domain and drives the phosphorylation of specific peptides within the substrate protein. B. Peptide specificity is determined experimentally by Oriented Peptide Library (OPL) screening, where purified kinases are exposed to random peptides where only particular positions are fixed to one specific residue, thus determining substrate molecular preferences for every given kinase. The experimental results can be turned into Position-Specific Scoring Matrices (PSSM) or logos.

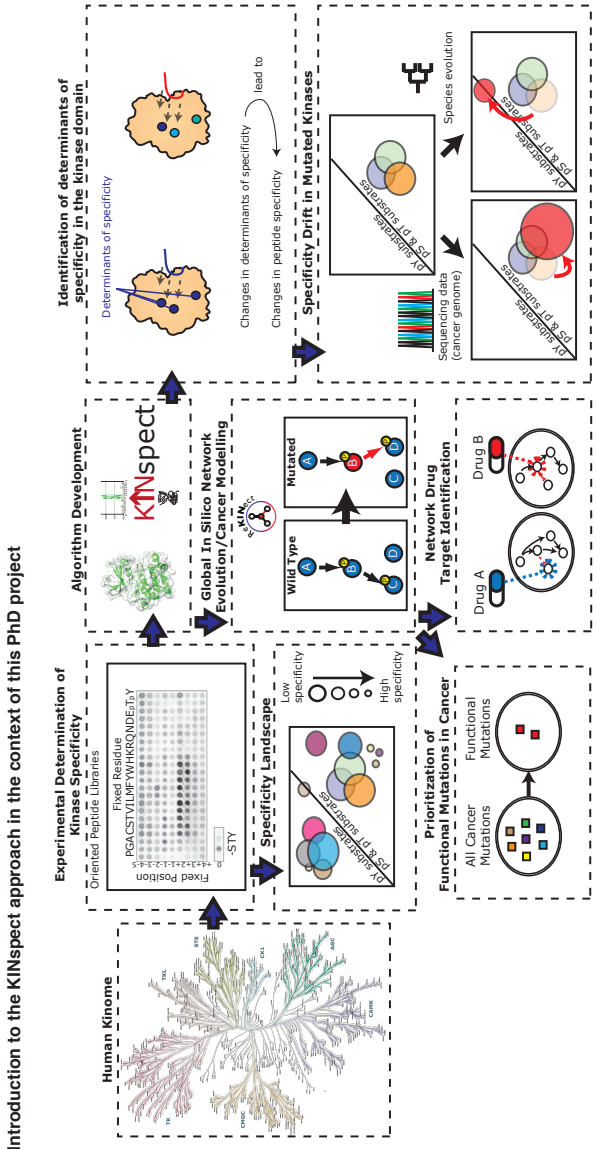


Figure 1.6. Kinase specificity within the scope of this PhD thesis. After determining the specificity for several human protein kinases, all this data can be used for the training of computational algorithms (KINspect) which try to identify the determinants of specificity in the kinase domain. This opens up new research avenues and opportunities in the modeling of specificity in cancer and evolutions or the identification of functional mutations or network drugs for personalized medicine, especially upon integration with other techniques that model the other functional residues within a single framework (ReKINect).

Part II

Existing and new approaches to study cancer signaling networks

Chapter 2

Current methods to identify functional cancer variants

As briefly introduced in Chapter 1, there are multiple computational tools that are being used for the annotation of tumor somatic variants and prediction of their functional impact in cancer. While in my PhD, we became part of the International Cancer Genome Consortium (ICGC), a network of international research institutions involving several specific cancer research projects with the ultimate aim of generating comprehensive catalogues of the genomic abnormalities underlying tumors of 50 different cancer types and subtypes. In this perspective, we provide recommendations for specific computational tools and discuss the immense challenge that represents the identification of mutations that contribute functionally to oncogenesis, tumor maintenance or response to therapy.

Computational approaches to identify functional genetic variants in cancer genomes

Abel Gonzalez-Perez[1,*], Ville Mustonen[2,*], Boris Reva[3,*], Graham R.S. Ritchie[2,4,*], Pau Creixell[5], Rachel Karchin[6], Miguel Vazquez[7], J. Lynn Fink[8], Karin S. Kassahn[8], John V. Pearson[8], Gary Bader[13], Paul C. Boutros[9,10,11], Lakshmi Muthuswamy[9,10], B.F. Francis Ouellette[9,12], Jüri Reimand[13], Rune Linding[5], Tatsuhiro Shibata[14], Alfonso Valencia[7,15], Adam Butler[2], Serge Dronov[2], Paul Flicek[4], Nick B. Shannon[16], Hannah Carter[6], Li Ding[17,18], Chris Sander[3], Josh M. Stuart[19,20], Lincoln D. Stein[9,21], Nuria Lopez-Bigas[1,22] and the ICGC Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group.

[1] Research Unit on Biomedical Informatics, University Pompeu Fabra, Barcelona, 08003, Spain.
[2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
[3] Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA.
[4] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[5] Cellular Signal Integration Group, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark.
[6] Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA.
[7] Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain.
[8] Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane, Queensland 4072, Australia.
[9] Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada.
[10] Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada.
[11] Department of Pharmacology & Toxicology, University of Toronto, Toronto, ON M5S 1A8, Canada.
[12] Department of Cell & Systems Biology, University of Toronto, Toronto, ON M5S 3G4, Canada.
[13] The Donnelly Centre, University of Toronto, Toronto, Canada
[14] Division of Cancer Genomics, National Cancer Center, Chuo-ku, Tokyo, 104-0045, Japan.
[15] Spanish National Bioinformatics Institute, Madrid 28029, Spain.
[16] Cambridge Research Institute, Cambridge CB2 0RE, UK.
[17] The Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA.
[18] Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA.
[19] Biomolecular Engineering Department, University of California, Santa Cruz, CA 95064, USA.
[20] Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA.
[21] Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada.
[22] Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain.

To whom correspondence should be addressed. Email: Nuria Lopez-Bigas <nuria.lopez@upf.edu>, Lincoln Stein <lincoln.stein@oicr.on.ca>

* these authors contributed equally.

Abstract

The International Cancer Genome Consortium (ICGC) aims to catalog genomic abnormalities in tumors from 50 different cancer types. Genome sequencing reveals hundreds to thousands of somatic mutations in each tumor, but only a minority drive tumor progression. We present the result of discussions within the ICGC on how to address the challenge of identifying mutations that contribute to oncogenesis, tumor maintenance or response to therapy, and recommend computational techniques to annotate somatic variants and predict their impact on cancer phenotype.

Introduction

Large-scale sequencing of cancer genomes often reveals many thousands of somatic missense (amino-acid changing) mutations in proteins. However, not all cancer mutations provide a selective ("driving") advantage to cancer cells[1,2]. Many mutations are so-called "passengers" because their impact on protein function is either insignificant or the affected protein is not important for tumor progression. The important practical problem is to determine which mutations are likely drivers. Although the carcinogenicity of a particular mutation depends on concurrent genomic alterations in the cell, one can significantly reduce the number of potential driver candidates by determining the functional impact of each mutation. Thus, a key challenge is to distinguish between functional and non-functional mutations, and by extension between those that contribute to tumorigenesis (drivers) and those that do not (passengers) (see **Box 1** for definitions).

Cancer has been likened to an evolutionary process by which tumor cells gain a fitness advantage over their neighboring cells[2]. The process creates cells with altered abilities such as the circumvention of apoptosis and senescence, deregulated cell division, and failed responses to external cues such as contact-contact inhibition and ligand-mediated cell signaling[3,4]. Normal cells are reprogrammed by changes in the genome that are subsequently selected and clonally expanded. In a similar manner to the way germline mutations can leave behind patterns indicative of negative or positive selection over millions of years, somatic mutations that engender increases in tumor fitness also can leave telltale signs in the protein sequence. The analysis of a given protein can thus reveal a pattern of alterations that recurrently result in its loss of function, as in classic tumor suppressors, like *TP53*, *RB1* or *PTEN*[5].

Mutation events collected across several patient samples can also reveal signs of clustering in the peptide sequence or the three-dimensional protein structure that indicates a critical domain has been modulated. In the extreme case, the presence of the same amino acid change in the same position in different individuals can be a strong indicator of such gain of function or oncogenomic events, as is the case with the *KRAS*[6] or *BRAF*[7] oncogenes. Such patterns can be leveraged by informatics tools to predict if a particular mutational event induces a selectable phenotype.

We review the computational analyses that are commonly carried out after the detection of somatic mutations across a cohort of cancer samples to identify likely functional and likely driver mutations (**Fig. 1**). Our focus will be on single nucleotide variants (SNVs) and small indels (operationally defined here as variants shorter than 50 bp) that change the amino acid sequence or affect regulatory regions. The output of these analyses consists of prioritized lists of mutations, genes and pathways that may undergo follow-up experiments to demonstrate their actual role in cancer.

We divide the process of identifying functional and driver variants into three independent, but related, approaches (**Fig. 1**). The first consists of mapping mutations to annotated functional genomic features, identifying their consequences and determining if they have been previously reported. The second uses computational methods to predict the nature and magnitude of the functional impact of mutation in particular elements (*e.g.,* proteins or regulatory regions). The third employs statistical methods to find signs of positive selection across the cohort. **Figure 1** lists a subset of the computational tools employed in each of the approaches. In the sections that follow, we review the rationale and tools of each approach and conclude by presenting some of the unsolved challenges and future perspectives in the field.

Approach 1: Mutation mapping, annotation and comparison to known variants

The first step in determining the possible functional consequences of somatic mutations is to identify annotated genomic features that may be affected by them. Features that are more likely to encode genomic functions include protein-coding and non-coding transcripts, transcription factor binding sites and other potential regulatory regions. Less well-characterized features, such as highly conserved regions or regions of open chromatin, may also be of interest. There are a variety of software tools that infer the consequences of mutations, but frequently these use different terms and different definitions for the effect itself[8–10] (Supplementary **Table 1**).

A large project such as the ICGC requires a common set of terms describing mutation consequences to facilitate the comparison of results among different groups. We have developed a standard set of 'consequence terms' drawn from the Sequence Ontology[11] (see **Supplementary Table 2**). This list will be extended and updated as the project unfolds. Along with the Sequence Ontology term used to describe the effect of a mutation, we also identify a minimal set of ancillary information that annotation tools should provide for each relevant consequence term, such as coding DNA sequence (CDS), protein relative coordinates, and predicted amino acid substitutions. Several of these annotations will depend on the specific transcript the mutation falls within, and so we recommend that a transcript identifier always be included. Note that this caveat means that a single mutation can, and frequently will, be assigned multiple consequences on multiple transcripts.

We recommend using tools that can output mutation descriptions in the format defined by Human Genome Variation Society (HGVS) at all relevant levels (e.g. DNA-level for all

mutations, and RNA and protein level descriptions where applicable). HGVS nomenclature provides a succinct and feature-centric format for variant descriptions, and some of the tools in **Supplementary Table 1** (e.g. the Ensembl VEP) have options to produce output in this format. We propose a common ranking scheme for the term set that summarizes the effects of a mutation that falls in multiple genomic features, such as multiple transcripts (see **Supplementary Table 2**). In addition, the ranking may be used for prioritizing mutations for follow-up analysis.

When assigning consequence terms to variants, the source of all underlying annotations, such as gene models and regulatory elements, must be noted to clearly document the event. In the context of ICGC, we recommend using the GENCODE[12] comprehensive set of gene models for all gene-associated annotations and identifying the specific release that was used. We advocate the use of GENCODE because of the detailed and frequently updated annotation of splice variants, pseudogenes and non-coding RNA loci, and the ready accessibility of all data for automated annotation via Ensembl and UCSC. Using the same gene models as the ENCODE project[13] will also allow further integration of somatic mutation data and the wider set of ENCODE annotations.

Comparing the list of mutations to catalogues of known variants

An obvious step in determining the implication of detected variants is to identify those that have been observed previously in other cancers, that are involved in other diseases, or that exist as germline polymorphisms. The growing collection of somatic variants detected within the different ICGC projects is a useful source of information, as are databases such as dbSNP[14], 1000 Genomes[15], Catalogue of Somatic Mutations in Cancer (COSMIC)[16] and databases of variants associated with hereditary diseases[17,18]. Several of the tools listed in **Table 1** automatically report if the variant is already known. Since none of these sources are definitive, the ICGC recommends that, at a minimum, projects report matches to variants known in dbSNP, OMIM, 1000 Genomes and COSMIC along with the version number of the database. Although dbSNP has sometimes been used to filter for somatic mutations, historically it contained primarily germline variants. However, in newer releases, many somatic mutations including mutational hotspots are also present, for example in *JAK2*, *KRAS* and *BRAF*. Thus, although we recommend reporting matches in dbSNP we do not recommend using it to filter out somatic mutations.

Approach 2: Assessing the functional impact of mutations

For many variants, no further assessment can be made about their potential impact on cell operation. Nevertheless, for the specific subset of mutations that affect either protein coding sequences or known regulatory sites, one can make computational predictions about their potential effects. In this section we describe computational analyses that may shed light on the possible functions of these variants**.**

Mutations affecting protein coding sequence

A number of computational methods have been developed to differentiate "functional" or "disease-associated" non-synonymous mutations from "non-functional" or polymorphic variants[19–24] (**Supplementary Table 3**). Some of these are specifically designed for cancer variants[25–28]. As a general rule, these approaches use evolutionary information (multiple sequence alignments), secondary and tertiary structure features, physico-chemical properties of amino acids, as well as information about the role of amino acid side chains in the 3D structure of proteins, such as protein surface placement in interaction sites.

Methods aimed at assessing the functional effect of non-synonymous mutations can be classified as "machine learning" and "direct". Machine learning methods use relevant properties of the original and mutant residues (*e.g.,* size, polarity), structural information (*e.g.,* surface accessibility, hydrogen bonding), and/or evolutionary conservation and other features. These methods are then trained to distinguish between positive sets of disease-associated variants and negative control sets of presumably non-functional or passenger variants. In contrast, direct methods assess the effect of a mutation through a computed phenomenological score based on a particular theoretical model that does not require training sets.

Most of these computational approaches have been benchmarked on variants with pronounced phenotypic effects[29] (*e.g.,* functionally deleterious and Mendelian disease-associated variants) and appropriate negative control sets, reporting accuracies close to ~80%. Although not originally designed for this purpose, some of them have been widely employed to rank cancer somatic mutations for their likelihood to be drivers, without previously benchmarking their performance on this problem.

One of the main challenges to produce such benchmarking is the difficulty of collecting well-curated sets of driver and passenger mutations. A recent effort to circumvent this problem employed various datasets of likely driver and likely passenger mutations[25]. Under the assumption that each proxy dataset is incomplete in non-overlapping ways, this study compared the performance of three well-known methods and their impact scores transformed to account for the baseline tolerance across several datasets rather than on individual datasets[25]. In the future, when many more cancer genomes have been sequenced and we understand better the implication of genetic variants on cancer phenotype, it may be possible to collect gold standard datasets to perform more accurate validation.

Given the high-throughput nature of cancer genome projects, one important aspect to consider for tool selection is their computational efficiency when thousands of variants are analyzed. Pre-computation of functional impact scores for all possible mutations in the human proteome is a useful remedy (as done by some tools presented in **Supplementary Table 3**). There is also at least one database (dbNSFP[30]) devoted to collecting and integrating such precomputed functional impact scores from different tools. In some cases it may be useful to visualize the location of mutations in protein 3D structure, if available, to further assess their potential role

with respect to protein stability and/or function, for instance using MuPIT Interactive[31] or the MutationAssessor web server[22].

The output of any computational method should be interpreted as a ranked list of candidate driver variants based on the user-submitted mutations, with the vast majority not likely to be true positives. The purpose of this ranking is to prioritize mutations for further experimental testing. Using a combination of methods based on different theoretical principles (and hence independent error models) may help mitigate false positive and negative rates suffered by any one method alone, thus resulting in a cleaner list of candidates for experimental validation.

Mutations affecting regulatory sites

Only very recently has it become feasible to identify and characterize somatic noncoding mutations that affect putative regulatory sites. Predicting the functional effects of regulatory variants typically starts either by purely statistical approaches, such as the application of machine learning methods to learn motif models from the regulatory sequences, or by modeling the transcription factor (TF) to DNA binding biophysics aided by experimental data such as those obtained from micro-fluidics or protein binding experiments[32,33]. Both approaches result in predictions of binding sites for different TFs within regulatory sequences. There are several tools for making such predictions, such as The Meme Suite[34], and the ENCODE project catalogues a number of relevant experimental data sets[13]. Furthermore, RegulomeDB provides an integrated approach to analyze regulatory variants[35]. It uses datasets from ENCODE[13] and other sources and also uses motif models (eg. from JASPAR[36]).

When a somatic mutation falls within a TF binding site, it is possible to score its effect in multiple ways. Perhaps the simplest is to take the relevant binding site motif model[36] and evaluate the score difference that the variant causes in that binding site's match to the model. This is close in spirit to scores that are derived from multiple alignments, such as PFAM log E value[37]. However, the interpretation of this particular score is not straightforward because the actual binding probability of TF to DNA depends strongly on the factor concentration within the cell and the presence of other protein binding factors and may thus vary across cell types. Furthermore, it is not clear in general whether stronger or weaker predicted binding is better or worse for TF function, and clarifying this will require studying the particular promoter and gene in more detail.

Pleasance *et al.* (ref. 38) used a specific tool[39] to address the functionality of mutations within promoters in a lung cancer cell line. Although somatic mutations did not differ significantly from the null expectation as a set, individual variants were predicted to have significant disruptive effects on potential binding motifs. More recently, systematic analyses integrating TF binding, histone marks, and other epigenomic data were used to identify pathways disrupted by Genome Wide Association Study (GWAS) at the regulatory level[40].

In addition to promoters and enhancers, it is also important to consider possible effects of mutations in splicing, especially now that the connection between splicing and cancer is

becoming increasingly clear (e.g., ref 41). Consequences of mutations in splicing regulatory elements are still difficult to predict but including additional experimental data, such as RNA-Seq, may lead to improvements in this area.

Given that the majority of somatic mutations reside in non-coding sequence, the need to computationally prioritize them for follow-up functional validation is clear. The recent discovery of melanoma driver mutations in the promoter sequence of telomerase reverse transcriptase (*TERT*) gene highlights the potential of regulatory variation to drive tumorigenesis[43,44]. As cancer genome projects are moving toward sequencing whole genomes, more non-coding driving mutations will likely be discovered. To facilitate such discoveries more computational method development to score regulatory variants is needed.

Approach 3: Finding signs of positive selection across a cohort

Independent of whether or not a functional consequence can be predicted for a given mutation, one can assess to what extent a given mutation has been observed at a higher frequency than expected. The rationale for assessing mutation frequency is that driver mutations provide an adaptive advantage to cancer cells (**Box 1**, *e.g., BRAF* V600E mutation found in melanoma[7]) and should thus be positively selected during the clonal evolution of tumors. Provided that similar selective pressures act on different patient tumors and that the same mutation is positively selected, one should be able to trace driver mutations by noting their higher frequency, a common trace of positive selection.

In principle, exploiting this fact to find driver genes is straightforward: it is simply a statistical comparison between the mutation rate observed in a gene versus what is expected under a neutral model. However, in practice this approach involves difficult choices with respect to the selection of appropriate models for neutral evolution. For example, germline variation should not be used to calibrate a null model for somatic mutation analysis[26] because this reflects evolutionary pressures and mutation processes during species evolution rather than during the development of cancer. In addition, many cancers have defects in DNA repair processes that change the neutral mutation rate, which have different regional impacts[38,45,46], and local mutation rate is variable depending on other factors such as replication timing[47].

To accurately identify significantly mutated genes, gene-specific mutation rates should thus be computed. This can be done using synonymous mutations[48] and/or mutations in introns and UTR sequences (eg. InVex)[49]; however, these approaches can only be effectively used in tumors with very high mutation rates. In other cases gene-specific mutation rates must be estimated taking into account factors known to affect mutation rate such as mutation context, replication timing and expression levels (eg. MuSiC[50] and MutSig[51]).

Given the difficulties that are intrinsic to recurrence-based methods, new methods have been developed that try to infer signs of positive selection using alternative means. One such

approach, OncodriveFM[52], consists of detecting genes that exhibit a significant bias towards the accumulation of somatic mutations with high functional impact. This method employs well-known metrics of the functional impact of individual mutations (those in **Supplementary Table 3**) to detect genes and pathways with this functional impact bias[52]. Another novel approach, ActiveDriver[53], involves the discovery of genes significantly enriched for somatic mutations that alter 'active sites' in proteins, such as signaling sites, regulatory domains or linear motifs, assuming that such active mutations are more likely to have a wide-spread downstream effect and lead to a phenotypic advantage for tumor cells[53].

Supplementary Table 4 lists several statistical approaches recently developed to identify candidate driver genes with signs of positive selection in a cohort of tumors[46,48–50,52–54]. As some of these methods are based on different theoretical principles, we recommend applying multiple complementary methods and comparing their results.

Despite these recent advances, future methods will need to capture the high degree of inter-tumor heterogeneity, as different tumors may acquire the same hallmark of cancer by different means (known as analogous mutations[55]). This heterogeneity is clearly underestimated in the current driver/passenger model.

Challenges and future perspectives

Cancer genome sequencing is a rapidly expanding field, and consequently computational methods used to interpret these data are evolving. We have presented a review of classes of practical tools currently available for analysis of a subset of genetic variation data. Because of the rapid evolution of the field, we have purposely avoided recommending particular tools or methods. Instead we present general guidelines to assist in making educated choices of methods that can address particular research problems. A number of pipelines facilitate the user-friendly application of various tools presented here. For instance, CRAVAT[56] maps mutations to their consequences on protein coding genes and it predicts their implication in cancer and disease using CHASM[26] and VEST[57]. IntOGen-mutations[58] provides a way to apply tools of the three approaches, including mapping mutations using Ensembl VEP[8], reporting their functional impact on proteins (using MutationAssessor[22], SIFT[20], PolyPhen2[59] and TransFIC[25]) and identifying genes with signs of positive selection across a cohort using OncodriveFM[52].

It is important to emphasize the limited capacity of these approaches to directly identify the causative mutations of tumor development. Rather, they are intended to prioritize candidates for follow-up experiments that may demonstrate their actual implication in the cancer phenotype. Reporting back the results of these rounds of validation experiments to the method's authors could in principle help them improve their approaches. The current relative scarcity of established spaces for this information exchange should be specifically addressed as part of the development of this field. Furthermore, these validation experiments will contribute to expand the catalogs of well characterized driver and passenger mutations, thus creating appropriate datasets for the development of computational prediction tools.

There are three key challenges in the field of cancer mutation analysis (**Box 2**). The first is to improve the accuracy of prediction of the functional impact of a mutation. Because mutations do not occur in isolation, but coexist with other somatic alterations that work together to alter cellular processes, separate gene-by-gene analyses are error-prone. A promising direction is the integration of multiple sources of biological information[60], and the use of pathway and network analyses in the interpretation of cancer genomes[22,61,62].

The second challenge is to develop reliable computational methods for the classification of mutations by functional impact type: loss of function, gain of function or switch of function[22,61,62]. The computational classification of mutations by type as well as strength of impact will contribute to the more complete elucidation of functional alterations in a cancer genome. The rich information encoded in the 3D structure of proteins, which is not yet well utilized by current approaches, can be particularly useful for deducing both the functional type and cellular consequences of mutations.

Lastly, there is the practical challenge of identifying mutations that confer resistance or sensitivity to a particular form of therapy (see for example[63,64]). We look forward to the day when functional prediction methods support personalized therapeutics, in which the patient's therapy is informed by analysis of the specific genetic alteration profile in an individual tumor. The development of better approaches for analysis of functional and driver mutations will help to facilitate this process and in so doing will support the future development of personalized cancer medicine.

Figure legends

Figure 1. Scheme depicting the three main approaches routinely employed in the analysis of cancer somatic mutations, as reviewed in this perspective. Although there are important relationships of precedence between elements from different approaches, they do not necessarily correspond to sequential steps. Tools employed in each of the approaches are shown in the middle. Integrative pipelines refer to tools that facilitate the use of methods across all approaches (*e.g.,* IntOGen-mutations pipeline).

Box 1: Definitions

We define a functional variant as a genomic variant that affects the molecular function of a protein (as a gain, loss or switch of function). A non-functional variant does not significantly affect the molecular function of a protein. A driver variant confers a selective advantage to a particular tumor cell, while a passenger variant does not. It is important to distinguish between functional versus non-functional and driver versus passenger as they describe different concepts. For example, a mutation might dramatically affect the function of a protein without providing any selective advantage to the tumor (it is a functional passenger variant). Non-synonymous mutations are those that alter the amino acid sequence of a protein.

Box 2: Current Challenges

1. Assess the functional impact of sets of mutations.
Most current methods cannot accurately predict changes in protein and cellular function because changes in tumor phenotype typically result from multiple genetic alterations.

2. Complement the identification of functional and driver mutations by the prediction of how mutations affect protein and cellular function.
There is a need for methods that not only identify functional or driver mutations but also predict the likely cellular outcome resulting from mutations such as gain, loss or switch of function, and how mutations might affect cellular networks.

3. Apply predictive tools to biologically relevant questions such as drug resistance.
The ideal method should not only predict the effect of multiple mutations in an integrative manner and how they affect protein and cellular outcome, but also tackle translational clinical challenges such as drug resistance.

References

1. ICGC *et al*. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
2. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
3. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
4. Hanahan, D. & Weinberg, R. a Hallmarks of cancer: the next generation. *Cell* **144**, 646–74 (2011).
5. Futreal, P. A. *et al*. A census of human cancer genes. *Nature Reviews. Cancer* **4**, 177–183 (2004).
6. Malumbres, M. & Barbacid, M. RAS oncogenes: the first 30 years. *Nature reviews. Cancer* **3**, 459–65 (2003).
7. Davies, H. *et al*. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
8. McLaren, W. *et al*. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26**, 2069–70 (2010).
9. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w (1118) ; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
10. Medina, I. *et al*. VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic acids research* **40**, W54–8 (2012).
11. Hoehndorf, R., Kelso, J. & Herre, H. The ontology of biological sequences. *BMC Bioinformatics* **10**, 377 (2009).
12. Harrow, J. *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760–74 (2012).
13. Dunham, I. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
14. Sherry, S. T. *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308–11 (2001).
15. Project, G., Asia, E., Africa, S., Figs, S. & Tables, S. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135**, 0–9 (2012).
16. Forbes, S. A. *et al*. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–950 (2010).
17. Stenson, P. D. *et al*. The Human Gene Mutation Database: 2008 update. *Genome Medicine* **1**, 13 (2009).
18. NHLBI Exome Sequencing Project (ESP) Exome Variant Server.
19. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**, 1073–1081 (2009).

20. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812–3814 (2003).
21. González-Pérez, A. & López-Bigas, N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *The American Journal of Human Genetics* **88**, 440–449 (2011).
22. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118 (2011).
23. Ryan, M., Diekhans, M., Lien, S., Liu, Y. & Karchin, R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics (Oxford, England)* **25**, 1431–2 (2009).
24. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* **15**, 978–986 (2005).
25. Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome medicine* **4**, 89 (2012).
26. Carter, H. *et al*. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research* **69**, 6660–7 (2009).
27. Kaminker, J. S., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* **35**, W595–598 (2007).
28. Capriotti, E. & Altman, R. B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**, 310–7 (2011).
29. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation* **32**, 358–68 (2011).
30. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. *Human mutation* **32**, 894-9 (2011).
31. Niknafs, N. *et al*. MuPIT Interactive: Webserver for mapping variant positions to annotated, interactive 3D structures. *Human Genetics* In press (2013).
32. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, N.Y.)* **315**, 233–7 (2007).
33. Badis, G. *et al*. Diversity and complexity in DNA recognition by transcription factors. *Science (New York, N.Y.)* **324**, 1720–3 (2009).
34. Bailey, T. L. *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202–8 (2009).
35. Boyle, A. P. *et al*. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790–7 (2012).
36. Bryne, J. C. *et al*. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research* **36**, D102–6 (2008).
37. Clifford, R. J., Edmonson, M. N., Nguyen, C. & Buetow, K. H. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics (Oxford, England)* **20**, 1006–1014 (2004).
38. Pleasance, E. D. *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
39. Hoffman, M. M. & Birney, E. An effective model for natural selection in promoters. *Genome research* **20**, 685–92 (2010).
40. Cowper-Sal Lari, R. *et al*. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature genetics* **44**, 1191–8 (2012).
41. Quesada, V. *et al*. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics* **44**, 47-52 (2011).
42. Desmet, F.-O. *et al*. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* **37**, e67 (2009).
43. Horn, S. *et al*. TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science (New York, N.Y.)* **339**:959-61 (2013).
44. Huang, F. W. *et al*. Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science (New York, N.Y.)* **339**:957-9 (2013).
45. Pleasance, E. D. *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).

46. Lohr, J. G. *et al*. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3879–84 (2012).
47. Stamatoyannopoulos, J. A. *et al*. Human mutation rate associated with DNA replication timing. *Nature genetics* **41**, 393–395 (2009).
48. Greenman, C. *et al*. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
49. Hodis, E. *et al*. A Landscape of Driver Mutations in Melanoma. *Cell* **150**, 251–263 (2012).
50. Dees, N. D. *et al*. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research* **22**:1589-98 (2012).
51. Lawrence M. S. *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* doi:10.1038/nature12213 (2013)
52. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic acids research* **40**:e169 (2012).
53. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology* **9**, 637 (2013).
54. Sjöblom, T. *et al*. The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)* **314**, 268–274 (2006).
55. Creixell, P., Schoof, E. M., Erler, J. T. & Linding, R. Navigating cancer network attractors for tumor-specific therapy. *Nature Biotechnology* **30**, 842–848 (2012).
56. Douville, C. *et al*. CRAVAT: Cancer-Related Analysis of VAriants Toolit. *Bioinformatics* (2013).
57. Carter, H. *et al*. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14(Supl 3):S3**, (2013).
58. Gonzalez-Perez, A., Perez-Llamas, C., Santos, A., Deu-Pons, J. & Lopez-Bigas, N. IntOGen-mutations pipeline: To interpret catalogs of cancer somatic mutations. at <http://www.intogen.org/mutations/analysis> (2013).
59. Adzhubei, I. A. *et al*. A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
60. Masica, D. L. & Karchin, R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer research* **71**, 4550–61 (2011).
61. Lee, W., Zhang, Y., Mukhyala, K., Lazarus, R. A. & Zhang, Z. Bi-Directional SIFT Predicts a Subset of Activating Mutations. *PLOS one* **4**, e8311 (2009).
62. Ng, S. *et al*. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics (Oxford, England)* **28**, i640–i646 (2012).
63. Iyer, G. *et al*. Genome sequencing identifies a basis for everolimus sensitivity. *Science (New York, N.Y.)* **338**, 221 (2012).
64. Valencia, A. & Hidalgo, M. Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome medicine* **4**, 61 (2012).
65. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).
66. Makarov, V. *et al*. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics (Oxford, England)* **28**, 724–5 (2012).
67. Habegger, L. *et al*. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics (Oxford, England)* **28**, 2267–9 (2012).
68. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology* **8**, R232 (2007).
69. Wong, W. C. *et al*. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics (Oxford, England)* **27**, 2147–8 (2011).
70. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics, Fourth Edition*. 545 (Sinauer Associates, Inc.: 2006).

Supplementary Table 2. Tools to annotate variants to genomic features that may be affected using Sequence Ontology Terms

Name	Description	Run by
Ensembl Variant Effect Predictor (VEP)[8] (http://www.ensembl.org)	Predicts the effect of variants with respect to the Ensembl gene set and regulatory build. Given a variant mapped to the reference genome, the system identifies any transcripts or regulatory regions that overlap the variant and uses a rule-based approach to predict the effect that each allele of the variant may have on the feature. Various useful ancillary annotations such as mRNA coordinates, amino acid changes, SIFT & PolyPhen predictions etc. are also reported. It also provides information about variants already known from various sources.	• Using web interface • Using standalone Perl script • Using Ensembl's Perl API • Using Ensembl's REST API
VAGrENT (http://www.sanger.ac.uk/resources/software/vagrent/)	Suite of PERL modules that make use of the ENSEMBL-variation and ENSEMBL-core APIs to retrieve the information on individual variants in a structured format. For each query variant, the user receives all the data on its effect and location at the level of DNA, mRNA, CDS, and protein.	• Using suite of PERL modules in command line
ASOoViR (https://sourceforge.net/p/asoovir/)	Annotates consequence terms of variants using Ensembl gene sets. Ensembl gene models and reference coding sequences are loaded into memory prior to annotation of variants allowing rapid annotation of whole genome scale calls. Annotation is performed on a transcript level basis, identifying associated sequence ontology terms for affected and nearby transcripts. Default output can be obtained on a gene basis, or on a transcript level basis.	• Querying via intermediate scripts in a command-line interface, user-generated scripts using the ruby modules, an interactive Ruby shell (IRB), or via a web browser by running the software in server mode
VARIANT[10] http://variant.bioinfo.cipf.es/	VARIANT (VARIant ANalysis Tool) reports the consequences of variants affecting coding transcripts, as well as noncoding SNVs situated both within the gene and in the neighborhood that could affect different regulatory motifs, splicing signals, and other structural elements, including Jaspar regulatory motifs, miRNA targets, splice sites, exonic splicing silencers, calculations of selective pressures on the particular polymorphic positions. It also provides	• Using a remote database cluster and operates through efficient RESTful Web Services that optimize search and transaction operations • Using command Line interface, REST Web service and Web interface have been

	information about variants already known from various sources.	implemented
Oncotator http://www.broadinstitute.org/cancer/cga/oncotator	Maps SNV and indels to genes, transcripts, functional consequences, and other relevant features. This mapping uses transcript information derived from the UCSC KnownGenes Track. Additionally, Oncotator annotates variants with information from other public data sources, including dbSNP, COSMIC, UniProt, DrugBank, ORegAnno, Cancer Gene Census, Tumorscape, TCGA Copy Number Portal, previously published MutSig Analyses, and the Gene Ontology.	• Using web service or web interface. Results are cached across submissions to improve performance. • Command-line and Python API versions will be publicly available soon.
snpEff[9] http://snpeff.sourceforge.net	Predicts the effect of variants with respect to the Ensembl, NCBI and UCSC gene sets and regulatory build. Identifies transcripts or regulatory regions that overlap the variant and predicts the effect that each allele of the variant may have on the feature. Provides additional annotations such as mRNA coordinates and amino acid changes. Putative effect impact classes are provided for easy categorization. Predictions about loss of function and non-mediated decay can be provided. Accompanying tools SnpSift provides other annotations, such as dbSNP, SIFT, GWAS Catalogue, PolyPhen, Gerp and conservation scores.	• Command line: Java program (platform independent) • Web interface: Integrated to Galaxy project • API: Java API provided.

Other tools for annotating variants to genomic features also exists, however, at the time of writing this perspective they do support the use of sequence ontology terms. Examples include ANNOVAR[65], AnnTools[66], CRAVAT[56] and VAT[67].

Supplementary Table 3. Methods to assess the functional effect of nsSNVs that can be used in a high-throughput manner

Type	Name	Description	Run by
D	**SIFT**[19,20] http://sift.jcvi.org	Given an input protein sequence, SIFT searches for similar sequences against a database defined by the user, builds a multiple sequence alignment of similar proteins and calculates normalized probabilities for all possible substitutions at all positions of the alignment. Based on these probabilities, SIFT classifies substitutions as likely neutral or deleterious.	• Downloading source code and binaries • Using web interface • Other tools, such as Ensembl VEP, also provide precomputed SIFT scores.
D	**PolyPhen2**[59] http://genetics.bwh.harvard.edu/pph2	Naïve Bayes classifier trained from two data sets that contain both deleterious and neutral amino acid changes. Eight sequence-based and three structure-based predictive features, most of them involving comparison of a given property of the wild-type amino acid and its mutated counterpart are the properties used to build the classifier.	• Downloading source code • Using web interface • Other tools, such as Ensembl VEP, also provide precomputed PolyPhen2 scores
D/C	**MutationAssessor**[22] mutationassessor.org	A prediction of the functional impact of protein missense mutations is based on the assessment of evolutionary conservation of amino acid residues in a protein family multiple sequence alignment. The novelty of the approach is in exploiting the evolutionary conservation in protein subfamilies, which are determined by clustering multiple sequence alignments of homologous sequences on the background of conservation of overall function[68].	• Using web interface, which also provides numerous biological annotations and the possibility to inspect mutations in multiple sequence alignment and in 3D structures • Downloading functional impact scores precomputed for all missense SNVs in the reference genome
C	**CHASM**[26] http://www.cravat.us	A random forest classifier is trained on a curated set of driver mutations derived from COSMIC and randomly simulated passenger mutations. It uses eighty-six diverse features (available at SNVBox database[69]), including physio-chemical properties of amino acid residues, scores derived from multiple sequence alignments of protein or DNA, region-based amino acid sequence composition, predicted properties of local protein structure and annotations from	• Downloading source code • Using web interface (results returned via email in spreadsheet and/or tab-delimited format)

		the UniProtKB feature tables.	
D	**VEST**[57] http://www.cravat.us	Random forest classifier trained on mutations from Human Gene Mutation Database and high-frequency nsSNPs from ESP6500. VEST uses features from the SNVBox database[69] (same as CHASM above).	• Downloading source code • Using web interface (results returned via email in spreadsheet and/or tab-delimited format)
C	**transFIC**[25] http://bg.upf.edu/transfic	transFIC (for transformed functional impact scores for cancer) takes the Functional Impact Score (FIS) produced by any method aimed at evaluating the impact of a mutation on the functionality of a protein and transforms it, taking into account the baseline tolerance of similar proteins to functional impacting variants. The transformation can be interpreted as an adjustment for the impact of the somatic variant on cell operation. transFIC has been shown to outperform the original scores in nine proxy datasets of driver and passenger mutations.	• Obtaining transFIC of SIFT, PolyPhen2 and MutationAssessor from a web service and using IntOGen-mutations pipeline • Downloading PERL program to transform any functional impact score
A	**LS-SNP/PDB**[23] (http://ls-snp.icm.jhu.edu/ls-snp-pdb)	LS-SNP/PDB annotates all human SNPs that produce an amino acid change in a protein structure in PDB. Features of each nsSNP's local structural environment, putative binding interactions and evolutionary conservation are displayed. nsSNPs can be filtered by evolutionary conservation, proximity to ligand or domain interface, secondary structure, solvent accessiblity, and severity of amino acid substitution. These annotations allow users to quickly scan a large number of nsSNPs of interest and prioritize those with higher likelihood of impacting normal protein activities.	• Using web interface
D	**CONDEL**[21] http://bg.upf.edu/condel	Condel (Consensus deleteriousness score) is an approach to combine the functional impact scores of non-synonymous single nucleotide variants (nsSNVs). It uses values extracted from the complementary cumulative distributions of the scores produced by individual tools on a dataset of deleterious and neutral nsSNVs as weights to combine them. Tested with	• Obtaining Condel scores for the combination of SIFT, PolyPhen2 and MutationAssessor from a webservice. • Downloading PERL program to combine any functional impact score.

		five well-known tools on a proven dataset of deleterious and neutral nsSNVs, the integrated score outperforms the individual methods.	
D	**Logre (LogR Pfam E-value)**[27] http://www.rbvi.ucsf.edu/Outreach/genentech.html	Wild-type and mutant sequence are aligned to the Pfam HMM that represents the domain where the mutation is located. Then, the logarithm of the ratio of the e-values (LogR E-value) of these two alignments is calculated. Positive ratios correspond to mutations that decrease the fit of the protein sequence to the HMM.	• Using Canpredict web server
D	**MAPP**[24] http://mendel.stanford.edu/SidowLab/downloads/MAPP	A multiple alignment of closely related protein sequences, and a phylogenetic tree of the relationships between the sequences are used to derive six matrices that reflect the constraints faced by the 20 amino acids to occupy each position of the alignment. Each matrix is built on the basis of a single physico-chemical property of the aminoacids. The results for the six physico-chemical properties are then de-correlated to compute a single score that measures the violation of constraints across all properties.	• Dowloading source code

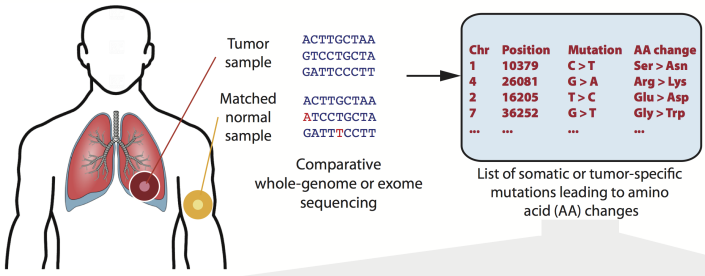
Note that PolyPhen2, CHASM and VEST are “machine learning algorithms” while the rest are “direct” methods. See main text for definition of these terms.

Legend Type
D: Tools designed to discriminate disease-associated variants from polymorphisms.
C: Tools specifically designed to rank likely cancer driver mutations
A: Tools that annotate information on known variants

Supplementary Table 4. Available tools to identify driver genes from a cohort of cancer patients

Name	Description	Run by
MutSig https://confluence.broadinstitute.org/display/CGATools/MutSig	MutSig (for "Mutation Significance") is a package of tools for analyzing mutation data. It operates on a cohort of patients and identifies mutations, genes, and other genomic elements predicted to be driver candidates.	• Requesting MutSig program from the authors *Input required is list of mutations found in each patient and the list of regions sequenced to sufficient depth ("covered") for mutation calling.*
MuSiC[50] http://gmt.genome.wustl.edu/genome-music/	MuSiC (for "Mutational Significance In Cancer") contains a set of tools to analyse cancer genomes, including a package to identify Mutated Genes (SMG package). Various statistical tests (*e.g.,* convolution test [CT], a Fisher's combined P-value test [FCPT] and the likelihood ratio test [LRT]) can be applied for this purpose).	• Downloading MuSiC software tools from their web page *Input required is list of mutations found in each patient (MAF) and alignment files (BAM).* *List of regions of interest also provided*
OncodriveFM[52] http://bg.upf.edu/oncodrive	OncodriveFM uncovers driver genes or gene sets (such as pathways) based on the accumulation of functional mutations across tumor samples as a signal of positive selection. It computes a metric of functional impact using three well-known methods (SIFT, PolyPhen2 and MutationAssessor) and assesses how the functional impact of variants found in a gene across several tumor samples deviates from a null distribution.	• Downloading OncodriveFM as a Perl program • Using IntOGen-mutations pipeline *Input required is the list of mutations found in each patient*
dN/dS	Nonsynonymous /synonymous dN/dS ratio is an approach from evolutionary genetics (see *e.g.,* Hartl and Clark 2007[70]). Ratio dN/dS>1 is consistent with positive selection acting on nonsynonymous mutations, dN/dS=1 with neutral evolution and dN/dS<1 with purifying (negative) selection acting on the variants.	• Downloading one of several (non-cancer specific) implementations of the tool *Input required is coding (nucleotide) sequences of the cancers and the reference sequence*
InVEx[49] http://www.broadinstitute.org/software/invex/	Permutation-based method for ascertaining genes with a somatic mutation distribution showing evidence of positive selection for non-silent mutations. Mutations are permuted on a per-patient, per-	• Downloading InVEx as a python program *Input required is list of mutations found in each patient (MAF) and wiggle file.*

	trinucleotide-context basis across the covered exon, intron and UTR base pairs of a gene, generating a null model of the distribution of mutations to which the observed distribution can be compared to determine statistical significance. The method can operate on whole exome as well as whole genome sequencing data in high mutation rate cancers.	
ActiveDriver[53] http://www.baderlab.org/Software/ActiveDriver/	ActiveDriver is a method for predicting cancer driver genes that show specific mutations in functional sites in protein sequences, such as phosphorylation sites. The gene-centric regression model estimates mutation significance by integrating mutation frequency, protein disorder, number of signaling sites and their proximity to mutations. ActiveDriver is complementary to standard frequency-based methods of mutation significance and helps interpret rare, but site-specific mutations.	• Downloading the ActiveDriver R package and associated data files (protein sequences, disorder predictions, phosphorylation signaling sites) from its web page. *Input required is list of mutations in each patient*



Goal

1 Map mutations

Map mutations to functional genomic features, identify their consequences and compare to catalogs of previously reported mutations

2 Assess effect on the function of a protein or regulatory element

Apply bioinformatic approaches to assess the functional impact of mutations at the level of gene products or regulatory regions

3 Find signs of positive selection across a cohort of tumors

Apply statistical tests to detect mutations, genes and pathways recurrently mutated or with overrepresentation of functional mutations

Tools

Ensembl VEP
VARIANT
Oncotator
VAGrENT
snpEff
ASOoVIR

SIFT
TransFIC
Mutation Assessor
CHASM/ VEST
PolyPhen2

dN/dS
OncodriveFM
ActiveDriver
MutSig
InVEx
MuSiC

Integrative pipelines

Output

List of consequence types for each mutation and information on whether mutations have been previously reported

Ranked list of mutations according to their functional impact

Ranked lists of candidate driver genes and pathways

Chapter 3

New strategies to personalize network medicine

On a more forward-looking and signaling-centric view to the cancer problem than the review included in Chapter 2, in this perspective, we discuss how cancer cells, by hijacking signaling networks, transition from a game governed by Nash equilibria of cooperation between all cells into a new scenario where cancer cells become masters of their own destinies. In reviewing the strategies cancer cells use to become "selfish", we discuss how genetic lesions can lead to altered protein function, changes to the structure and dynamics of signaling networks and, ultimately, cellular phenotype. We also describe general properties of cancer signaling networks and challenges in cancer network biology, and finish by suggesting how the future of personalized medicine could be revolutionized by a combination of relatively new technologies that could allow the discovery of network drugs.

Navigating cancer network attractors for tumor-specific therapy

Pau Creixell[1], Erwin M Schoof[1], Janine T Erler[2] & Rune Linding[1]

Cells employ highly dynamic signaling networks to drive biological decision processes. Perturbations to these signaling networks may attract cells to new malignant signaling and phenotypic states, termed cancer network attractors, that result in cancer development. As different cancer cells reach these malignant states by accumulating different molecular alterations, uncovering these mechanisms represents a grand challenge in cancer biology. Addressing this challenge will require new systems-based strategies that capture the intrinsic properties of cancer signaling networks and provide deeper understanding of the processes by which genetic lesions perturb these networks and lead to disease phenotypes. Network biology will help circumvent fundamental obstacles in cancer treatment, such as drug resistance and metastasis, empowering personalized and tumor-specific cancer therapies.

Cells are constantly computing decisions based on the integration of different cues that reach them at various times. In contrast to single-cell organisms, in multicellular organisms, cellular decisions should, ultimately, benefit the organism as a whole, even if that implies that an individual cell will have to decide to commit suicide. In line with this unique feature, signaling networks have evolved during multicellular evolution to allow cells to integrate cues and make decisions that ensure cooperative behavior between them. By hijacking these mechanisms, cancer cells escape cooperative rules and transition from a game governed by Nash equilibria[1,2] between all cells into a new scenario where cancer cells decide their behavior purely based on their own benefit, or as phrased by Hanahan and Weinberg[3], "become masters of their own destinies." Given the central role played by signaling networks in the integration of cues to compute any cellular responses, we argue that cancer is not simply a disease with a genetic basis, but is one ultimately driven by perturbations at the signaling network level, and that both the 'cue-signal-response' rules of cellular decision-making and the switch in strategy from cooperative to selfish are major, hitherto understudied, hallmarks of cancer[3,4].

In this article, we dissect the strategies cancer cells use to become 'selfish' and drive disease. We first review how genetic lesions can lead to altered protein function, which can result in changes to the structure and dynamics of signaling networks and ultimately cellular phenotype. Next, we describe five general properties of cancer signaling networks (**Fig. 1**) and define five challenges in cancer network biology and propose strategies to overcome them (**Fig. 2**). By meeting these challenges, network biology may fundamentally advance not only basic biology but also patient treatment. Finally, we describe how a combination of relatively new technologies could become a potent cocktail for the discovery of network drugs, and we discuss the practical implementation of personalized and tumor-specific cancer therapy.

From genomic lesions to functional network perturbations

Tumor cells often harbor hundreds to thousands of genetic lesions. But based on the observation that some of these genetic lesions are repeatedly observed in several cancers (e.g., *BRAF V600E*, present in >50% of all malignant melanomas[5]), it has been hypothesized that only a few genetic lesions are causally implicated in cancer development ('drivers'), whereas the majority have no functional consequences ('passengers')[6].

Although this classification has had some use in identifying mutations that are highly prevalent, it is now apparent that a tumor is not, under any circumstances, a static and uniform population of malignant cells. Rather, it is a dynamic ensemble of subpopulations with different abnormalities undergoing molecular evolution[7–9]. Two fundamental principles of cancer signaling networks can explain why a binary driver/passenger classification may be too simplistic to accommodate the complex dynamic nature of tumors. First, different tumors can develop similar phenotypes by acquiring mutations in different proteins[10], in what we term analogous mutations (**Fig. 1a**). Second, it has been shown that two different mutations not capable of causally driving cancer by themselves are able to do so when they appear in combination within the same cells or even within two neighboring cells[11], in what could be described as two passengers becoming drivers or, as we refer to them, synthetic oncogenes (**Fig. 1b**). Thus, patient-to-patient heterogeneity can be driven by the presence of different mutations in the same or in different proteins that lead to a similar signaling state and phenotypic outcome.

Altogether, the intrinsic heterogeneity of tumors makes it a pressing challenge for cancer network biologists to develop tools to identify the extent to which combinations of cancer mutations affect protein function and cellular and phenotypic states (**Fig. 2a**,**b**). Even though several such tools have been developed (reviewed in ref. 12), existing methods are mainly based on protein structure and/or sequence conservation. This is at odds with recent findings that show that cancer mutations tend not to cluster on the most conserved protein regions. In kinases, for example, mutations typically hit the kinase activation segment, a functional, yet largely nonconserved protein region[13].

[1]Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), Lyngby, Denmark. [2]Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. Correspondence should be addressed to J.T.E. (janine.erler@bric.ku.dk) or R.L. (linding@cbs.dtu.dk).

Published online 10 September 2012; doi:10.1038/nbt.2345

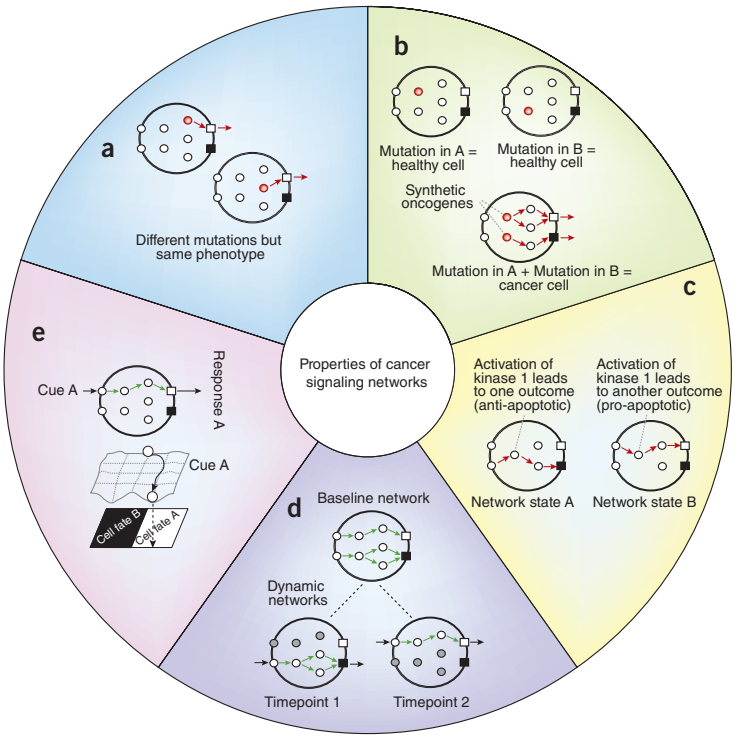


Figure 1 Properties of cancer signaling networks. (**a**) Analogous mutations. Two different tumors may achieve the same signaling and phenotypic outcome with two different mutations (**b**) Synthetic oncogenes. Mutations that are not oncogenic on their own can cooperate when appearing together to drive tumor formation[11]; by analogy to synthetic lethality, we call the genes harboring cooperative mutations, synthetic oncogenes. (**c**) Multivariate nature of signaling networks. The response of a cell to a specific cue depends on, and can only be predicted by taking into account, the state of the cellular signaling networks[25]. This dependency, known as the multivariate nature of signaling networks, is often neglected when classifying mutations and genes as oncogenes or tumor suppressors and cancer drivers or passengers. (**d**) Dynamic networks. Although signaling networks are often represented as static, it is clear that they are highly dynamic entities. Given that the role of signaling networks in computing cellular responses is highly dependent on it, and that cancer mutations will perturb it, this dynamic nature is a critical property of cancer signaling networks. (**e**) Signaling network landscapes. The different states that a signaling network occupies can be represented as a landscape (with stable steady states or attractors represented as valleys and unstable steady states represented as hills), where the cell constantly gets pushed by signaling cues[31,32,39,40]. These states drive cellular and disease phenotypes and represent network drug targets.

Because cancer cells would obtain the greatest fitness advantage from mutations that target the most-functional residues, we reason that a better understanding of the functionality of protein residues would allow more accurate predictions of the consequences of cancer mutations. Functional residues have been defined as those residues required for a protein to perform its molecular function(s), in the sense that they cannot be freely changed without directly affecting the role(s) of the protein[14]. Here we extend this definition to include a more fine-grained and precise definition of protein function as an ensemble of protein features that together describe the different functional capabilities of proteins (e.g., ATP binding, substrate specificity, protein activation or phospho-tyrosine binding). This new definition would not only adapt well to current studies of sequence-function associations[15,16], but also lead to a better description of the effects of a mutation affecting such residues (**Fig. 2a**,**b**).

An insightful example of how to explore this sequence-function relationship in protein domains was carried out by researchers in the Ranganathan and Yaffe laboratories who, using methods from statistical mechanics, generated synthetic WW domains *de novo* that maintained fold and function[17,18]. Further supporting a complex sequence-function relationship, additional studies from the Ranganathan laboratory demonstrated that, in addition to protein architecture described as combinations of modules such as globular domains and linear motifs[19–21], protein domains themselves often have well-defined sectors formed by sparse networks of residues often linking spatially distant regions that contribute cooperatively but unequally to its function[22,23]. Although some targeted studies analyzing several cancer mutations in a single kinase have been conducted[24], similar approaches to those used for WW domains should be pursued to generate high-throughput experimental studies of cancer mutations in the context of signaling networks. These would help gain a better understanding of which amino acid residues can be changed freely without affecting the protein and network function and, most importantly, which cannot.

From network perturbations to cellular phenotypes

The characterization of cellular signaling processes has largely focused on identifying the function of individual genes and proteins. A notable exception is a landmark study[25] on the context dependence of the Jun-activated kinase (JNK) in apoptosis. Before this work, paradoxical results suggested that JNK had a pro-apoptotic function[26], an anti-apoptotic function[27] or even a lack of involvement in apoptosis[28]. The systematic approach undertaken by Janes *et al.*[25] revealed that the phosphorylation status of JNK (and thus its catalytic activity) was not sufficient to determine apoptotic commitment; instead, activation of JNK could lead to both apoptosis and proliferation depending on the cellular signaling network state at the time of activation. Thus, this work demonstrated that a protein's cellular role is not a static property but rather can only be defined dynamically—that is, its role depends on the context of the network it is operating within. Similar context dependencies have been confirmed for other kinases, such as Erk and MK2. Because of this, which is referred to as the multivariate property of signaling networks (**Fig. 1c**), we suggest that it is essential to study cellular context at the systems level.

Although these multivariate molecular networks seem to have evolved a complex structure that makes them robust against deletion of a few proteins[29], they are highly dynamic. Thus, a more accurate description of signaling networks should take into account the fact that a single static network does not exist unchanged over time. Instead, a cell contains a dynamic ensemble of networks whose different permutations are manifested in the cell depending on the different cues the cell is presented

with (**Fig. 1d**). This dynamic nature of signaling networks could, at least in part, explain why all mutant proteins do not seem to be expressed at a given point in time[30], if a substantial part of the proteome is so dynamic that it is expressed only when the cell senses a specific cue.

Moreover, according to a general principle of complex systems introduced in the 1980s[31,32], dynamic cellular networks can only exist in a finite number of states, owing to the constraints that interactions between nodes impose on one another. These network states can be represented as landscapes, where most-probable and least-probable states are represented as valleys and mountains, respectively (**Fig. 1e**). Cells are continuously exploring this landscape and are pushed from one state to another by different environmental or intracellular cues.

Implications for cancer research

The multivariate nature of signaling networks has profound implications for cancer research. Just as it is inaccurate to assign a static function (e.g., apoptotic or anti-apoptotic) to a single protein, it is clear that static interpretations of mutations, that is, driver or passenger mutations, are also misleading. For example, given that the phenotypic role of JNK strongly depends on network state, it is clear that a mutation in JNK (and thus probably any other mutation) should not be statically labeled as a driver or passenger or as an oncogene or tumor suppressor, as such classifications are context dependent (e.g., disease or cell-type specific). Several examples, such as Myc[33] or WT1 (ref. 34) gene products that act as both tumor suppressors and oncogenes, support this idea. These results underscore the importance of assessing mutations based on their effects on signaling networks and of developing novel classification methods to do so. Along these lines, MAP2K4 (one of the protein kinases that can phosphorylate and activate JNK) has been shown to be recurrently lost or mutated in several cancers[35–38]. These represent prime examples of mutations that may display ambivalent phenotypic impact similar to JNK.

Motivated by the example of MAP2K4 and many other mutated kinases[38], we maintain that mutations capable of affecting signaling networks—which we call network-attacking mutations (**Fig. 2c**)—are more likely to affect phenotype than other mutations. Thus, we discuss a general strategy in which mutations in individual cancers are assessed based on, first, the likelihood they will affect protein function, and second, the cellular role of the signaling network that they are operating within (**Fig. 3**). Our strategy extends the concepts introduced by Waddington and elaborated by Kauffman and Huang *et al.*[31,32,39,40], where cancer mutations are turned into perturbations capable of reshaping these landscapes. We represent the cellular response or phenotype as another dimension where each network state (every point in the landscape) is constantly projected to and translated into a cellular decision or phenotypic outcome.

We postulate that network-attacking mutations affect the cell not by perturbing how the signaling landscape is projected to the phenotypic dimension, but by changing the ensemble of dynamic networks that can be manifested in a cell and, in consequence, the number and stability of steady states in the signaling landscape, thus creating new attractor states that only cancer cells can occupy, also known as cancer network attractors (**Fig. 3**). This has additional implications for other mechanisms, such as oncogene and non-oncogene addiction[41], where cancer cells would be trapped in cancer attractor states and could escape from them by reverting the genomic aberration that initially

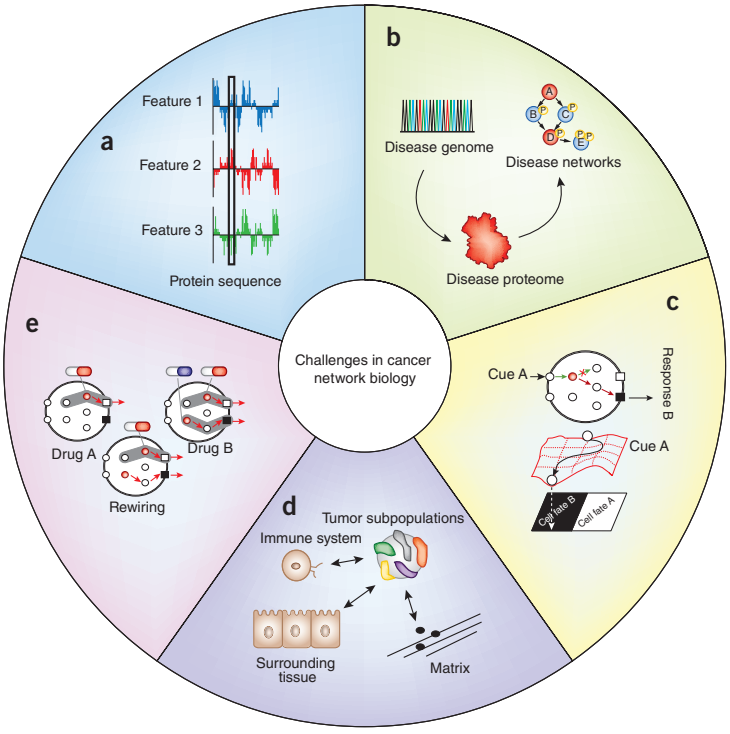


Figure 2 Challenges in cancer network biology. (**a**) Functional consequences of cancer mutations. Using an ensemble of protein-function features (e.g., ATP binding, substrate specificity, activation of the protein kinase or phospho-tyrosine binding), which together represent a comprehensive description of a protein's molecular functions, will enable more accurate and predictive evaluation of cancer mutations. (**b**) Modeling of disease networks. Although experimental and computational tools for modeling molecular networks exist, creating more comprehensive, sensitive and accurate new tools especially designed to model disease-associated networks still represents a big challenge in network biology. (**c**) Network-attacking mutations and cancer network attractors. Network-attacking mutations are mutations that lead to a new cellular phenotype by perturbing signaling networks either at the network structure or the network dynamics level. Network-attacking mutations transform signaling networks, generating new possible network states by changing the number and/or stability of steady states in the signaling landscape[31,32,39,40]. These acquired signaling capabilities lead to alterations in the cell's normal 'cue-signal-output' flow and thereby drive disease phenotypes (see **Fig. 3** for further details). (**d**) Tumor subpopulations and micro-environment. The field is only beginning to comprehend the complex interactions that exist between different co-evolving tumor cell subpopulations and between those cells and the tumor microenvironment, both of which strongly influence tumor progression. (**e**) Network-aware and temporal drugs. As predicted by R.L. and Pawson[66] several years ago, new pharmaceutical strategies that target networks instead of single proteins are becoming available[47,48]. We predict this trend will not only continue, but also include recent advances that highlight the possibility to 'cure' networks using time- and order-dependent therapies[68]. In coming years, the discovery of resistant, metastatic, tissue or cell-specific networks could lead to an even greater advance in the field of network medicine (**Fig. 5**).

caused the perturbed landscape. Given the high degree of determinism that exists between signaling networks, landscapes and phenotypes, we argue that network-attacking mutations are at the heart of all new decision-making capabilities acquired by cancer cells. Consequently, in our view, the study of both network-attacking mutations and new attractor states acquired by cancer cells, that is, cancer network attractors, deserves the highest priority from the field. Such studies should be performed through systematic and quantitative sampling of cell dynamics at multiple levels (e.g., genomic or epigenetic, proteomic and phenotypic), followed by nonlinear interpolation and integrative computational modeling (**Fig. 4**).

The first network-attacking cancer mutation, described more than 15 years ago[42], was a point mutation in the kinase domain of *RET* (M918T), which leads to a switch in peptide specificity. In line with their importance, network-attacking mutations have attracted more attention in recent years[43–48]. Moreover, information has been accumulating steadily about how specificity in signaling networks and modular protein domains emerges[49–51], leading to the definition of determinants of specificity in protein domains[52,53]. These determinants, sometimes referred to as specificity-determining residues, are residues that can lead to substrate specificity changes after mutation. Notably, direct mutagenesis of these determinants of specificity has been used to rewire the entire histidine kinase signaling system in bacteria in a predictive manner[54]. Recent follow-up work indicates that mutations in determinants of specificity prevent cross-talk and allow protein family expansions[55], in a process similar to the one powered by negative selection over Src homology 3 (SH3) protein domains that show similar specificity[56]. We propose that similar studies in human signaling networks, coupled with mapping of cancer mutations on these determinants of specificity, would shed new light on whether signaling rewiring is a general principle of oncogenesis and tumor progression, knowledge of which would in turn be critical as molecular therapies target proteins and their networks and not genes.

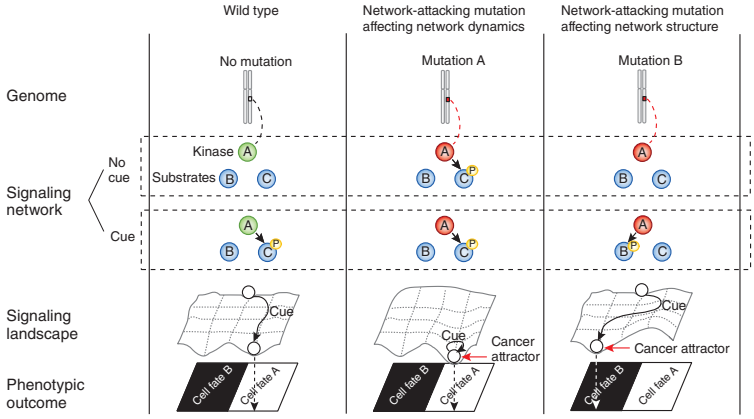


Figure 3 Network-attacking cancer mutations. Proteins are the key elements of signaling networks as a result of their ability to integrate external cues and direct the information flow toward a specific cellular outcome (e.g., epidermal growth factor (EGF) leading to proliferation or tumor necrosis factor alpha (TNF-α) leading to apoptosis). Network-attacking mutations affect the 'cue-signal-output' cellular information flow by affecting either the dynamics (middle), for example, by keeping proteins constitutively active, or the structure (right), by affecting protein specificity, of the signaling networks. Signaling networks can be represented as a landscape with the most likely network states represented as valleys (stable steady states or attractors) and the least likely network states as mountains (unstable steady states). Network-attacking mutations dysregulate signaling networks by perturbing the number and/or stability of steady states in the landscape, effectively creating new cancer-specific attractors that only cancer cells will be able to reach.

Despite the fact that the number of known cancer network-attacking mutations is still relatively low, recent findings suggest that in-frame mutations are enriched on interaction interfaces[57], which implies they are also likely to affect determinants of specificity. Moreover, many fusion proteins have been discovered that likely directly rewire or create new network states[58]. Given the rate at which cancer mutations are being reported and the development of new computational methods for systematically identifying these mutations (**Fig. 2b**), we predict a steep increase in the number of network-attacking mutations that will be uncovered in the coming years.

Personalized cancer network biology

Led by recent advances in sequencing technologies, the amount of data on cancer genome mutations is growing exponentially[59]. Current efforts

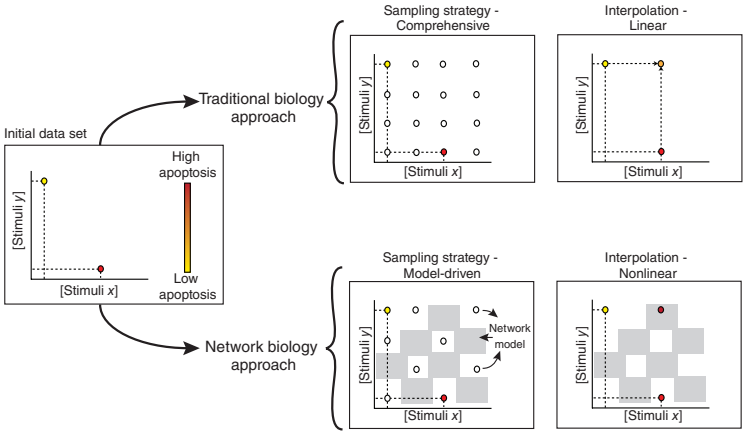


Figure 4 Traditional versus network biology approaches. In more traditional biological approaches, where only one or a few genes or proteins are sampled across a limited set of conditions, there has been limited success in deriving predictive models across conditions or cell types that would require comprehensive sampling. In contrast, network biology relies on systematic sampling across combinations of states that result in increased performance of a network model. Unlike classic approaches, in which the system is stimulated with single specific cellular cues (e.g., growth factor), in the network biology approach, the multivariate nature of signaling networks and the nonlinear relationship between signaling input and output can be successfully elucidated by interrogating the system with multiple orthogonal cues.

from the Cancer Genome Atlas and Cancer Genome Project, now under the umbrella of the International Cancer Genome Consortium[60], will facilitate the annotation and collection of cancer genome data. We foresee similar waves of technological progress and the generation of new consortiums in the cancer proteomics fields in the near future. The establishment of the Clinical Proteomic Tumor Analysis Consortium (http://proteomics.cancer.gov/programs/cptacnetwork), and the implementation of new approaches[61] and labeling techniques[62] optimized for patient samples are encouraging advances in this direction.

These advances, however, will need to be coordinated with new algorithmic and experimental high-throughput methods (e.g., high-content screening) capable of interpreting this flood of information because the functional interpretation of the data is currently the main bottleneck in the field of personalized cancer network biology. Computational integration of large quantitative data sets is also becoming increasingly important, and thus there is a growing requirement for supercomputing infrastructure with large algorithmic dynamic range (e.g., next-generation large shared memory systems). Benchmarking and validation of systematic workflows and algorithms is already receiving increasing attention through initiatives, such as the DREAM challenge[63] and IMPROVER[64].

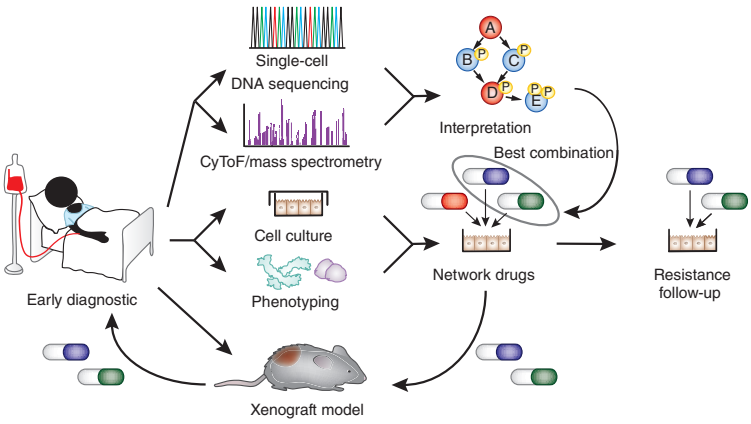


Figure 5 Personalized cancer network biology. The goal of personalized cancer network biology is to be able to treat each tumor with the best combination of drugs tailored to that tumor. Ideally, early diagnosis should be followed by the development of tumor-specific cell lines and xenograft models, cancer genome sequencing, and proteomic and phenotypic analysis. Combinations of network drugs should then be tried in the tumor-specific cell line and xenograft model and eventually transferred back to the patient. Continuing to treat the tumor-specific cell culture with the same network drug combination as is used in the patient may be useful for understanding potential resistance and/or metastasis.

Two emerging areas in network biology that are likely to contribute to the future of cancer research are the study of cell-cell interactions (**Fig. 2d**) and drugs specifically designed to interfere with diseased network dynamics (that is, network drugs; **Fig. 2e**).

R.L. and collaborators[65] studied cell-cell interactions by isotopically labeling two distinct subpopulations of cells, one expressing ephrin-B1$^+$ and the other Eph-B2$^+$, and carrying out a comprehensive phosphoproteomic analysis. This strategy facilitated the first measurements of phosphorylation events during the interaction of two cell subpopulations. The proliferative behavior of cancer cells is still poorly understood in part because it is difficult to experimentally study the transmission of proliferative factors from one cell to its neighbors[3]. Therefore, we argue that a similar isotopic labeling strategy could be used to investigate the cooperation between cells with different oncogenic lesions that together (that is, synthetic oncogenes; **Figs. 1b** and **2d**) lead to tumor formation[11].

Combination drugs that interfere with disease networks (so-called network medicine[66]) have been shown to lead to a better response than single-hit therapies by causing secondary perturbations to signaling networks[47,48,67]. Recent work by the Yaffe laboratory represents a clear leap forward within the field of network medicine[68,69]. Following network modeling, Yaffe and colleagues[68] managed to decode the signaling network dynamics that drive resistance to DNA-damaging chemotherapy. This information was used to sensitize otherwise resistant triple-negative breast cancer cells to conventional DNA-damaging chemotherapy by administering doxorubicin (Adriamycin, Doxil) and erlotinib (Tarceva) in an order- and time-dependent fashion. This could be considered the first example of temporal network drugs (**Figs. 2e** and **5**).

We predict that personalized or even tumor-specific cancer therapy will become a reality in the foreseeable future, starting from early diagnosis of the disease, followed by next-generation sequencing, proteomic analysis, high-throughput profiling of phenotypic cell states in the tumor and design of patient-specific combinations of network drugs with resistance follow-up (**Fig. 5**). Relatively new techniques, such as single-cell and high-depth sequencing[70,71], imaging[72] and cytometry time-of-flight[73], could prove especially valuable for monitoring the number, properties and behavior of different tumor subclones (**Fig. 2d**). Ideally, network drugs, such as the aforementioned order- and time-dependent combination[68], should then be chosen based on the interpretation of sequencing as well as the proteomic and phenotypic analysis of tumor cells and tested on the tumor-specific cell lines and xenograft model. The best-performing combination should ultimately be transferred back to the patient (**Fig. 5**). This whole process should take the shortest time possible to avoid the evolution of the tumor in the patient and the consequent loss of relationship between the primary tumor and the cell line. Tumor-specific cell lines would be kept and treated with the same drugs used in the patient to monitor tumor evolution and treat for resistance and/or metastasis as soon as there is enough evidence of it (**Fig. 5**). Ideally, every patient and paired xenograft or cell line should have a complete electronic record showing the treatment history to facilitate retrospective and cross-disease studies[74,75].

Conclusions

Although we have highlighted some of the challenges that still exist in cancer network biology, substantial progress is also being made. For example, the usage of patient-derived tumor tissue in animal xenograft models to test the response to particular drugs aimed at developing new personalized cancer therapy is rapidly becoming an established technology[76]. Surgical orthotopic implantation to transplant tumors taken directly from the patient to the corresponding organ of immunodeficient mice[77] is currently one of the most promising methods to enable drug screening in patients. In addition, new clinical trials, such as the MD Anderson T9 project[78], are under way in which patients are given therapy that targets tumor-specific aberrations. Nevertheless, the implementation of the strategy depicted in **Figure 5** would benefit from further developments in technology, funding and legislation. For

example, generating models for cancer research that represent human patient diversity[79] and mimicking the complexity of tumor microenvironments (J.T.E. and collaborators)[80] remain extraordinary challenges (**Fig. 2**), and further research efforts and investments are required. As cancer biology becomes a 'big data' science, similar to physics, we expect to see more systematic, data-driven research efforts that will uncover and confront many of the tumor complexities that have remained elusive so far.

Despite recent predictions of >13 million cancer deaths in 2030 (ref. 81), as discussed in this Perspective, we foresee that within this timeframe tumor-specific medicine will become a reality, thanks to a new generation of cancer network biologists who will hopefully overcome these challenges, positively contributing to the battle against this devastating disease and the significant reduction of patient suffering.

ACKNOWLEDGMENTS

We apologize to our colleagues whose work could not be cited due to space limitations. We thank all members of the C-SIG (DTU), the ErlerLab (BRIC), M. Yaffe (MIT) and N. Brunner (KU) for critical input on this manuscript. R.L. is a Lundbeck Foundation Fellow and is supported by a Sapere Aude Starting Grant from The Danish Council for Independent Research and a Career Development Award from Human Frontier Science Program. J.T.E. is supported by a Hallas Møller Stipend from the Novo Nordisk Foundation. Visit http://www.networkbio.org/, http://www.lindinglab.org/ and http://www.erlerlab.org/ for more information on cancer-related network biology.

AUTHOR CONTRIBUTIONS

All correspondence should be addressed to both J.T.E. and R.L.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2345.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Nash, J.F. Equilibrium points in N-person games. *Proc. Natl. Acad. Sci. USA* **36**, 48–49 (1950).
2. Nash, J.F. Non-cooperative games. *Ann. Math.* **54**, 286–295 (1951).
3. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
4. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
5. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
6. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
7. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
8. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
9. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
10. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
11. Wu, M., Pastor-Pareja, J.C. & Xu, T. Interaction between RasV12 and scribbled clones induces tumour growth and invasion. *Nature* **463**, 545–548 (2010).
12. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
13. Dixit, A. *et al.* Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS ONE* **4**, e7485 (2009).
14. Pazos, F. & Bang, J.-W. Computational prediction of functionally important regions in proteins. *Curr. Bioinform.* **1**, 15–23 (2006).
15. Fowler, D.M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
16. Jensen, L.J. *et al. Ab initio* prediction of human orphan protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265 (2002).
17. Socolich, M. *et al.* Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
18. Russ, W., Lowery, D., Mishra, P., Yaffe, M. & Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **437**, 579–583 (2005).
19. Puntervoll, P. *et al.* ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31**, 3625–3630 (2003).
20. Lim, W.A. & Pawson, T. Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* **142**, 661–667 (2010).
21. Seet, B.T., Dikic, I., Zhou, M.M. & Pawson, T. Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.* **7**, 473–483 (2006).
22. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
23. Reynolds, K.A., McLaughlin, R. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).
24. Wan, P.T. *et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855–867 (2004).
25. Janes, K.A. *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
26. Lei, K. & Davis, R.J. JNK phosphorylation of Bim-related members of the Bcl2 family induces Bax-dependent apoptosis. *Proc. Natl. Acad. Sci. USA* **100**, 2432–2437 (2003).
27. Lamb, J.A. *et al.* JunD mediates survival signaling by the JNK signal transduction pathway. *Mol. Cell* **11**, 1479–1489 (2003).
28. Abreu-Martin, M.T. *et al.* Fas activates the JNK pathway in human colonic epithelial cells: lack of a direct role in apoptosis. *Am. J. Physiol.* **276**, G599 (1999).
29. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
30. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* advance online publication, doi:10.1038/nature10933 (4 April 2012).
31. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
32. Kauffman, S.A. & Weinberger, E.D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.* **141**, 211–245 (1989).
33. Uribesalgo, I., Benitah, S.A. & Di Croce, L. From oncogene to tumor suppressor: The dual role of Myc in leukemia. *Cell Cycle* **11**, 1757–1764 (2012).
34. Yang, L., Han, Y., Saurez Saiz, F. & Minden, M.D. A tumor suppressor and oncogene: the WT1 story. *Leukemia* **21**, 868–876 (2007).
35. Ellis, M.J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).
36. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
37. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
38. Greenman, C. *et al.* Pattern of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
39. Waddington, C.H. *The Strategy of the Genes: a Discussion of Some Aspects of Theoretical Biology* (Allen & Unwin, 1957).
40. Huang, S. & Ingber, D.E. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261**, 91–103 (2000).
41. Luo, J., Solimini, N.L. & Elledge, S.J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* **136**, 823–837 (2009).
42. Songyang, Z. *et al.* Catalytic specificity of protein-tyrosine kinases is critical for selective signaling. *Nature* **373**, 536–539 (1995).
43. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
44. Dreze, M. *et al.* 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog. *Nat. Methods* **6**, 843–849 (2009).
45. Pe'er, D. & Hacohen, N. Principles and strategies for developing network models in cancer. *Cell* **144**, 864–873 (2011).
46. Vidal, M., Cusick, M.E. & Barabási, A.-L.L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
47. Schoeberl, B. *et al.* Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. *Sci. Signal.* **2**, ra31 (2009).
48. Huang, P.H. *et al.* Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc. Natl. Acad. Sci. USA* **104**, 12867–12872 (2007).
49. Miller, M.L.L. *et al.* Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**, ra2+ (2008).
50. Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415–1426 (2007).
51. Mok, J. *et al.* Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci. Signal.* **3**, ra12 (2010).
52. Brinkworth, R.I., Breinl, R.A. & Kobe, B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. USA* **100**, 74–79 (2003).
53. Turk, B.E. Understanding and exploiting substrate recognition by protein kinases. *Curr. Opin. Chem. Biol.* **12**, 4–10 (2008).
54. Skerker, J.M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).
55. Capra, E.J., Perchuk, B.S., Skerker, J.M. & Laub, M.T. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222–232 (2012).
56. Zarrinpar, A., Park, S.H. & Lim, W.A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676–680 (2003).
57. Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
58. Brehme, M. *et al.* Charting the molecular network of the drug target Bcr-Abl. *Proc. Natl. Acad. Sci. USA* **106**, 7414–7419 (2009).
59. Wong, K.M.M., Hudson, T.J. & McPherson, J.D. Unraveling the genetics of cancer: genome sequencing and beyond. *Annu. Rev. Genomics Hum. Genet.* **12**, 407–430 (2011).

60. Ledford, H. Big science: the cancer genome challenge. *Nature* **464**, 972–974 (2010).
61. Bensimon, A., Heck, A.J.R. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **81**, 379–405 (2012).
62. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**, 383–385 (2010).
63. Prill, R.J. *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* **5**, e9202 (2010).
64. Meyer, P. *et al.* Verification of systems biology research in the age of collaborative competition. *Nat. Biotechnol.* **29**, 811–815 (2011).
65. Jørgensen, C. *et al.* Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. *Science* **326**, 1502–1509 (2009).
66. Pawson, T. & Linding, R. Network medicine. *FEBS Lett.* **582**, 1266–1270 (2008).
67. Chandarlapaty, S. *et al.* AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity. *Cancer Cell* **19**, 58–71 (2011).
68. Lee, M.J. *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* **149**, 780–794 (2012).
69. Erler, J.T. & Linding, R. Network medicine strikes a blow against breast cancer. *Cell* **149**, 731–733 (2012).
70. Navin, N. *et al.* Tumor evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
71. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
72. Pedersen, M.W. *et al.* Sym004: a novel synergistic anti-epidermal growth factor receptor antibody mixture with superior anticancer efficacy. *Cancer Res.* **70**, 588–597 (2010).
73. Bendall, S.C. & Nolan, G.P. From single cells to deep phenotypes in cancer. *Nat. Biotechnol.* **30**, 639–647 (2012).
74. Roque, F.S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLOS Comput. Biol.* **7**, e1002141 (2011).
75. Jensen, P.B., Jensen, L.J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
76. Blumenthal, R.D. & Goldenberg, D.M. Methods and goals for the use of in vitro and in vivo chemosensitivity testing. *Mol. Biotechnol.* **35**, 185–197 (2007).
77. Hoffman, R.M. Orthotopic mouse models expressing fluorescent proteins for cancer drug discovery. *Expert Opin. Drug Discov.* **5**, 851–866 (2010).
78. Gonzalez-Angulo, A.M., Hennessy, B.T. & Mills, G.B. Future of personalized medicine in oncology: a systems biology approach. *J. Clin. Oncol.* **28**, 2777–2783 (2010).
79. Hunter, K.W. Mouse models of cancer: does the strain matter? *Nat. Rev. Cancer* **12**, 144–149 (2012).
80. Cox, T.R. & Erler, J.T. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. *Dis. Model. Mech.* **4**, 165–178 (2011).
81. WHO. World health organization fact sheet 297 (2012). http://www.who.int/mediacentre/factsheets/fs297/en/

Chapter 4

The genetic code and its consequences for short-term evolution and cancer. A story about serine.

When studying protein evolution, both in short and long timescales, the impact that the genetic code has on the probability of seeing different amino acid residue substitutions is not always taken into account. In this article, we demonstrate how, especially in short-time evolution, such as the one present in complex diseases like cancer, or between phylogenetically close species, the genetic code has a significant impact in determining the evolution of amino acid residues, with mutations between amino acid residues that are close in mutational space (e.g. one nucleotide apart from one another) occurring at a much higher rate than mutations further away in mutational distance (e.g. two or three mutations away from one another). As a result from this, we show that serine, thanks to its unique occupancy within the codon table (six codons distributed across different parts of the table), is the amino acid residue with highest mutability and targetability or, in other words, a mutational hub. Finally, we demonstrate that the cell can fine-tune the mutational activity of different residues when these residues encode functionality, as phosphorylatable residues, such as serine, show a much lower mutability when they are the regulated residue in phosphorylation sites.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues

Pau Creixell, Erwin M. Schoof, Chris Soon Heng Tan and Rune Linding

Phil. Trans. R. Soc. B 2012 **367**, doi: 10.1098/rstb.2012.0076, published 13 August 2012

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

CrossMark click for updates

Phil. Trans. R. Soc. B (2012) **367**, 2584–2593
doi:10.1098/rstb.2012.0076

Research

Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues

Pau Creixell[1], Erwin M. Schoof[1], Chris Soon Heng Tan[2] and Rune Linding[1,*]

[1]*Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), DK-2800 Lyngby, Denmark*
[2]*Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), Vienna, Austria*

As François Jacob pointed out over 30 years ago, evolution is a tinkering process, and, as such, relies on the genetic diversity produced by mutation subsequently shaped by Darwinian selection. However, there is one implicit assumption that is made when studying this tinkering process; it is typically assumed that all amino acid residues are equally likely to mutate or to result from a mutation. Here, by reconstructing ancestral sequences and computing mutational probabilities for all the amino acid residues, we refute this assumption and show extensive inequalities between different residues in terms of their mutational activity. Moreover, we highlight the importance of the genetic code and physico-chemical properties of the amino acid residues as likely causes of these inequalities and uncover serine as a mutational hot spot. Finally, we explore the consequences that these different mutational properties have on phosphorylation site evolution, showing that a higher degree of evolvability exists for phosphorylated threonine and, to a lesser extent, serine in comparison with tyrosine residues. As exemplified by the suppression of serine's mutational activity in phosphorylation sites, our results suggest that the cell can fine-tune the mutational activities of amino acid residues when they reside in functional protein regions.

Keywords: amino acid evolvability; mutation; phosphorylation site evolution

1. INTRODUCTION

Cells are constantly evolving in a race for adaptation to dynamic environmental challenges. As described by François Jacob over three decades ago [1], this process is more analogous to tinkering than to free design, in the sense that nature does not create a new protein function from a blank canvas nor with unlimited resources, but instead evolves through innovation with existing proteins (figure 1*a*,*b*). In line with this principle of functionalization by tinkering, most general models of protein evolution (e.g. duplication–divergence [2], neofunctionalization or subfunctionalization [3]) are based on gene duplication being the main source of new genes, proteins and consequently new cellular function.

In this study, we aim to extend the principle of tinkering in evolution, initially developed by Jacob [1], to include the effect the genetic code has on protein evolution. Our hypothesis is that evolution is not only constrained because it needs to tinker with existing proteins; it is also affected by the genetic code in the sense that genetic variation is not generated by substituting amino acid residues from the evolving protein at random, but instead the genetic code dictates that some amino acid substitutions will be more frequent than others (figure 1*c*).

2. THE INFLUENCE OF THE GENETIC CODE ON MUTATIONAL PATHS

In essence, substitutions between amino acid residues that are far away from each other in mutational space are less likely than between residues that are close to each other (figure 2). For instance, if we had to compute the probability of every amino acid residue to be the target of a mutation from methionine, we would have to consider the mutational distance and the physico-chemical similarity between the two residues. Isoleucine, leucine, phenylalanine, valine, threonine, lysine and arginine are, in terms of mutational distance, the closest residues to methionine, because they are all just one nucleotide mutation away from it (figure 2*a*). Alanine, valine, isoleucine and leucine are the closest residues in physico-chemical distance, because they are small hydrophobic residues similar to methionine (figure 2*b*). Combining these two distances (mutational and physico-chemical) that determine the genetic diversity generated and selection of protein variants, one can rationalize the amino acid substitution frequencies observed along evolution (figure 2*c*).

*Author for correspondence (linding@cbs.dtu.dk; www.lindinglab.org).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2012.0076 or via http://rstb.royalsocietypublishing.org.

One contribution of 13 to a Theme Issue 'The evolution of protein phosphorylation'.

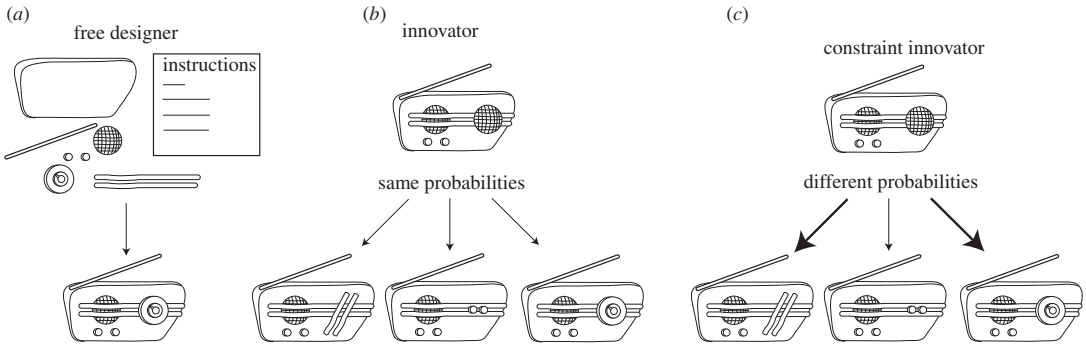


Figure 1. Creative methodologies and evolution. As an analogy to protein evolution in the hunt for new protein function, we have illustrated different strategies to design a radio. (*a*) As Jacob described several years ago, nature does not evolve by creating de novo protein function from a blank canvas resembling a free designer who can build a radio using some predefined instructions and any imaginable radio parts. (*b*) Instead, nature is more of an innovator who tinkers with existing proteins before finding new protein function by a process of mutation and selection. Following with our analogy, the tinkerer does not generate a radio from scratch, but it tinkers with existing devices by combining and substituting pieces, and the best design is selected for. (*c*) In this study, we extend this concept by highlighting the fact that the sources and targets of mutations cannot be chosen arbitrarily, but instead some amino acid substitutions will be more likely than others (different probabilities). Unlike in (*b*), where different substitution probabilities are not considered, tinkering with the loudspeaker in the radio is more likely to lead to some radio parts than others.

Next, we tested the validity and generality of this influence the genetic code has on mutational paths. In principle, one would expect the effect of the genetic code to decrease with time, because longer evolutionary distances would allow several mutations in the same amino acid residues to become more likely (figure 3*a*). As briefly suggested earlier (figure 2*c*), regardless of what amino acid substitution is more probable, purifying selection will act subsequently to disfavour substitutions that would lead to radical changes in the physico-chemical properties of the protein residue. Thus, unlike the effect of the genetic code, we expect the effect of the physico-chemical properties of the different amino acids to remain constant over time. To test the influence of the genetic code and physico-chemical properties on protein evolution, we reconstructed ancestral sequences at different evolutionary distances between humans and other vertebrates (figure 3*b* and see §7 for further details). Supporting our hypothesis, we indeed observed different targets of mutation at different evolutionary distances (figure 3*c*), with mutational targets closer in mutational space for shorter evolutionary distances (L1: human–orangutan) and less influenced by mutational distance for longer evolutionary distances (L7: human–frog).

3. MUTATIONAL PROPERTIES OF AMINO ACID RESIDUES

By expanding our analysis, we computed matrices to reflect the probability of every amino acid residue to mutate and become every other amino acid residue at different evolutionary distances (see the electronic supplementary material, table S1). To better describe the different mutational properties of amino acid residues represented in these matrices, we introduce two new terms, mutability and targetability. We define mutability as the probability of an amino acid residue to mutate, and targetability as the probability of an amino acid to be the result of a mutation. By extension, we have termed our matrices (which effectively contain mutability for each residue on their rows, targetability on their columns and conservation on their diagonal) mutability targetability (MUTA) matrices. The rationale behind developing our MUTA matrices is similar to the rationale behind matrices such as point accepted mutation (PAM) [4] or blocks of amino acid substitution matrix (BLOSUM) [5] but they differ fundamentally in their goal and, in consequence, also in the information they contain (figure 4). While matrices such as PAM or BLOSUM, default matrices used by popular tools such as BLAST (Basic Local Alignment Search Tool) [6], reflect the tendency of some amino acid residues to appear in a multiple sequence alignment of homologue proteins, MUTA matrices describe the probability of the different amino acid residues to mutate (mutability) and be targets of mutation (targetability). Given that MUTA matrices are derived not from conserved blocks but instead from a large range of sequences with different degrees of evolvability, they are likely to be more useful than previous matrices for evolutionary analysis (e.g. the characterization of phosphorylation sites or other protein sequences that do not necessarily reside in conserved protein regions).

To better visualize every amino acid residue's mutational properties, one can represent each amino acid residue as a data point on an x–y scatter plot, i.e. mutability–targetability plot (figure 5*a*,*b*). Following this strategy, we show mutability and targetability for every amino acid residue at different evolutionary distance (figure 5*c*); it is apparent that, contrary to common assumption, different amino acid residues have different mutational properties (i.e. mutability and targetability). Moreover, it is evident that there is a correlation between mutability and targetability whereby amino acids that tend to mutate more are

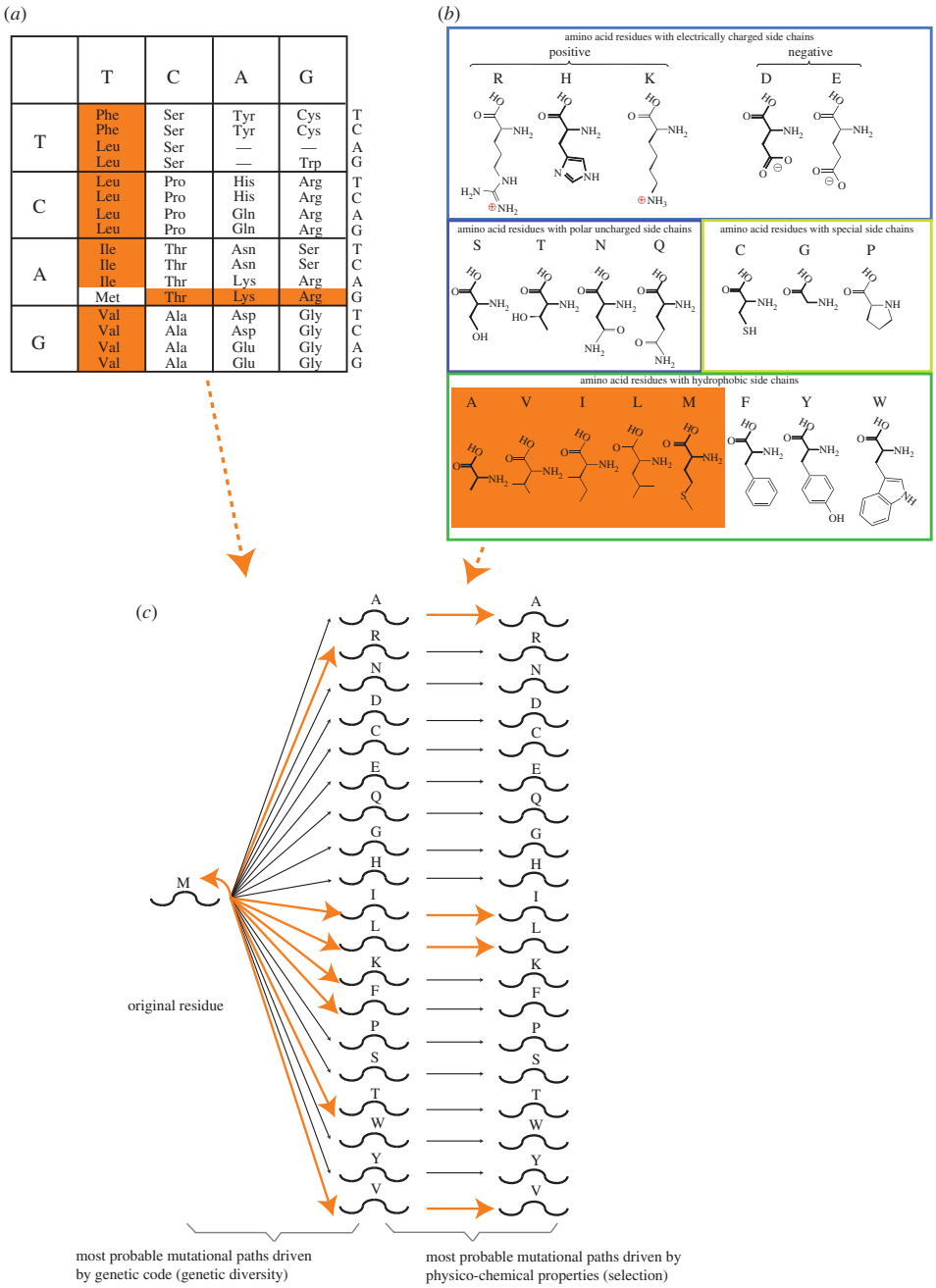


Figure 2. Exploring evolutionary mutational targets. (*a*) In this codon table, we have highlighted amino acid residues that are close to (one nucleotide mutation away from) methionine in mutational space. (*b*) In this table of physico-chemical properties of different amino acid residues, we have highlighted amino acid residues that are close (similar) to methionine in physico-chemical space (adapted from www.wikipedia.org). (*c*) Combining mutational and physico-chemical space allows rationalization of why some mutational paths (amino acid substitutions) are more frequent than others. Here, we have highlighted in orange the most preferred mutational paths owing to short mutational distance (first arrow) and short physico-chemical distance (second arrow). Residue conservation has been illustrated as a loop, and it should be considered as another possible mutational path with very short mutational and physico-chemical distance.

also more likely targets of mutations. This correlation indicates that the dynamic system of amino acid residue substitutions and frequencies lies in equilibrium in a stable steady state, where all the residues balance out residue loss and gain after mutation, which results in only small frequency fluctuations over time.

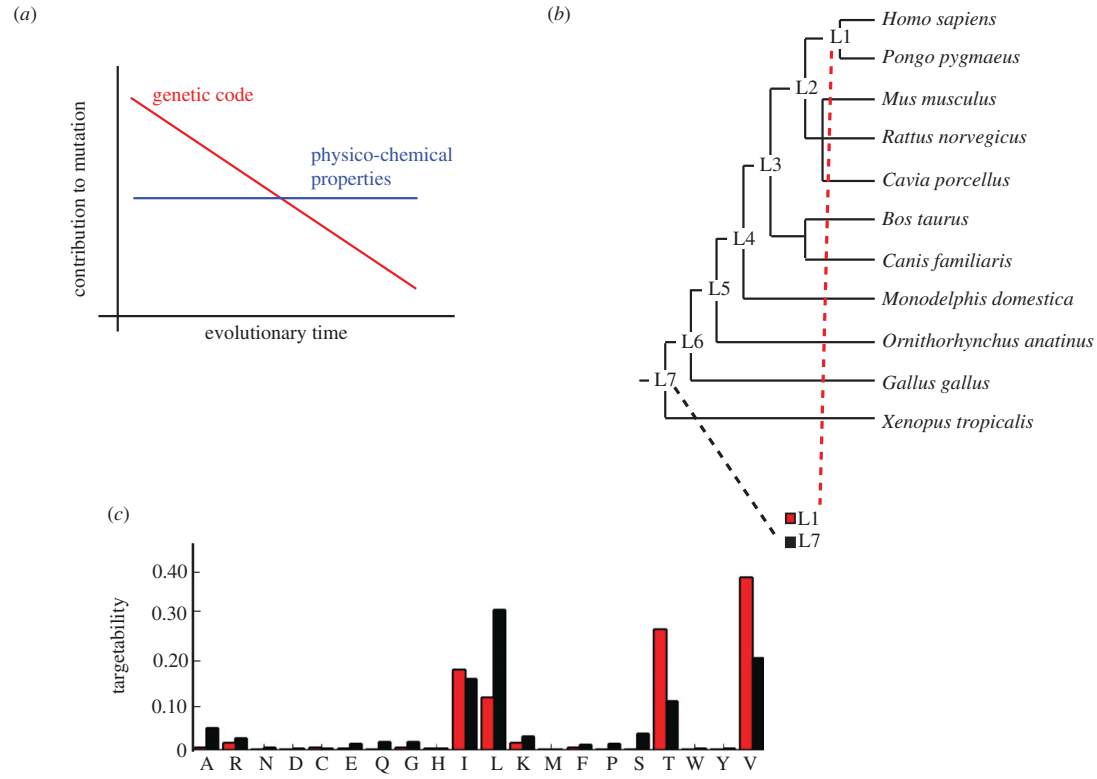


Figure 3. Exploring evolutionary mutational targets. (*a*) The relative contribution of the genetic code (by disfavouring amino acid residue substitutions that require several nucleotide mutations) and the physico-chemical properties (by disfavouring amino acid residue substitutions between dissimilar residues) to mutation will vary over evolutionary time. The restrictions imposed by the genetic code will have higher influence when comparing short-evolutionary distances, whereas the physico-chemical properties of amino acid residues will have a constant influence, because selection against radical changes in physico-chemical space will always be applied before a mutation becomes fixed. (*b*) Graphical representation of the phylogenetic tree whose ancestral sequences (L1, L2, L3, L4, L5, L6 and L7) we have reconstructed as described in §7. (*c*) Here, we confirm the principle described in (*a*), by comparing mutational targets of methionine between L1 and human and between L7 and human and showing that in shorter evolutionary distances (L1: red), methionine tends to mutate only to residues that are one nucleotide mutation away, while for longer times (L7: black), more targets are possible.

In contrast, large discrepancies between mutability and targetability would lead to large fluctuations in frequency and, with time, to extinction or perpetuation (figure 5*b*). This correlation between mutability and targetability is therefore the only path to prevent amino acid residue extinction or perpetuation.

It is also apparent from our mutability–targetability plots that different residues use different evolutionary paths to hold their frequency stable. In one extreme, serine evolves very fast by mutating very often, while also being a more likely target of mutations, i.e. high mutability and high targetability. At the opposite extreme, tryptophan does not mutate frequently, but at the same time it is not a frequent target of mutations either, i.e. low mutability and targetability (figure 5*c*). Analogous to how different nucleotide or protein sequences can evolve at different speed, here we have uncovered that even individual amino acid residues can be fast- or slow-evolving (e.g. serine and tryptophan, respectively). Next, we will investigate the causes and consequences of the mutational properties of the different residues.

4. POSSIBLE CAUSES FOR DIFFERENT MUTATIONAL PROPERTIES OF AMINO ACID RESIDUES

The fact that serine is the fastest-evolving amino acid residue can perhaps give us some insights into why different amino acid residues would present different mutational properties. First, considering mutational space (figure 2*a*), it is apparent that serine is a unique residue in that it is the only amino acid whose six codons are distributed in two different groups, AGY and TCN, that are so far apart from each other (at least two nucleotide mutations away). As a consequence, serine will be more easily reached from another amino acid after mutation, i.e. it is very close in mutational space to most other amino acid residues (in most cases, only one nucleotide mutation away). In addition, from the perspective of physico-chemical distance (figure 2*b*), serine's moderate physico-chemical properties, without a bulky or charged side chain, make it less likely that the amino acid substitution will be rejected by selection (figure 2*c*), because it is close in physico-chemical space to most other amino acid residues.

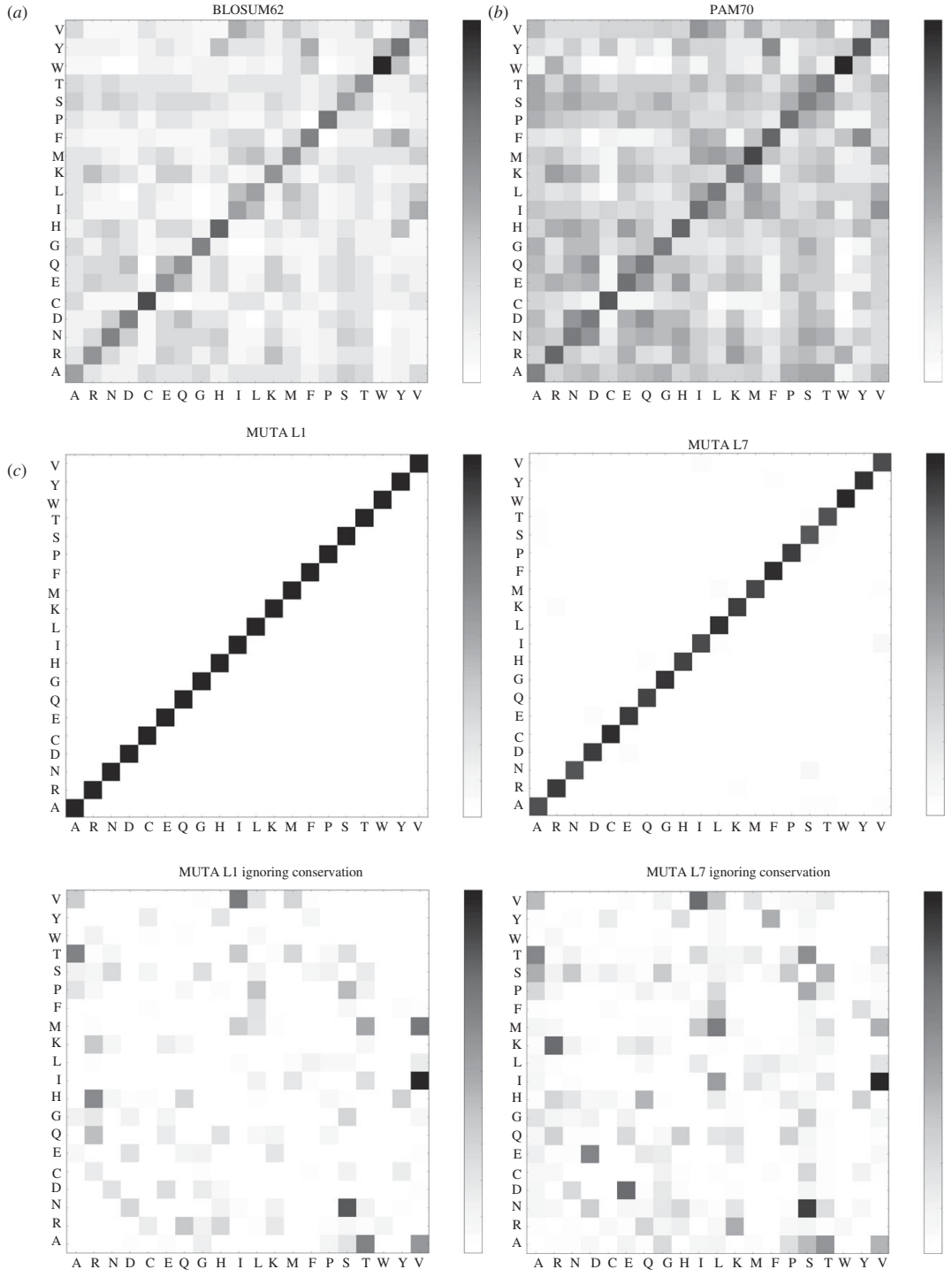


Figure 4. A comparison of amino acid substitution matrices. (*a*) Representation of a normalized version of the BLOSUM62 matrix. (*b*) Representation of a normalized version of the PAM70 matrix. (*c*) Representation of our L1 and L7 MUTA matrices, including versions without conservation (bottom) in order to better visualize the non-conservative amino acid substitutions.

In comparison with serine, the other two amino acid residues coded by six codons (leucine and arginine) do not combine such mutational and physico-chemical proximity with other amino acid residues and, in consequence, are not as fast-evolving as serine. Despite the fact that leucine's physico-chemical properties are

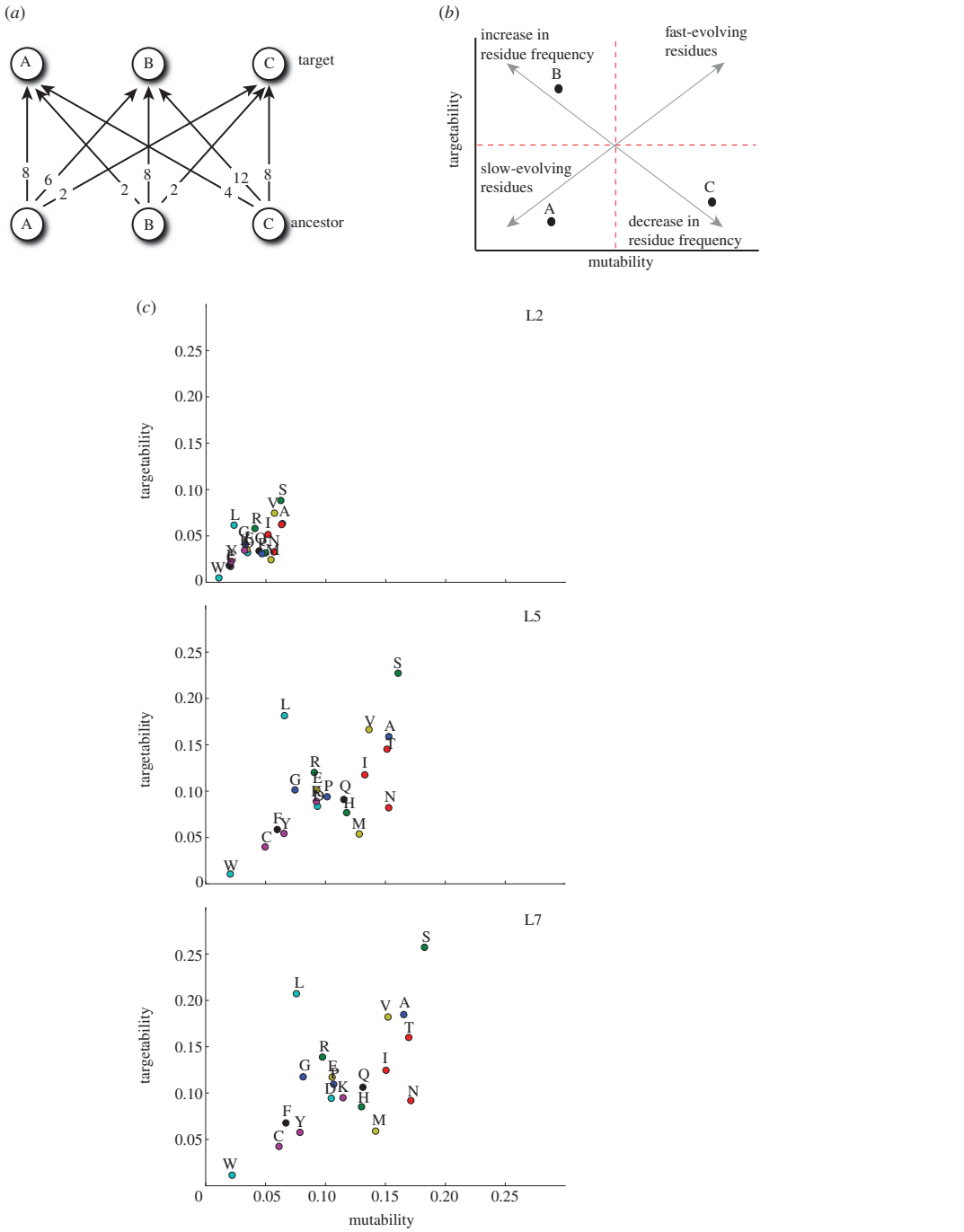


Figure 5. Mutability–targetability plots. (*a*) Toy model to represent three mutable objects and how they can evolve, with each letter representing one element at the ancestral (bottom) or target (top) sequence and each arrow representing the frequency of every possible mutational path. (*b*) Any mutable system, such as the one represented in (*a*), can be represented in a mutability–targetability plot, an *x*–*y* scatter plot where each element (e.g. A–C) is located in a precise coordinate depending on their mutational properties, i.e. how often it mutates (mutability) and how often it is the result of a mutation from another residue (targetability). Depending on their location, we can consider the mutable elements fast or slow evolving (high mutability and high targetability or low mutability and low targetability, respectively) or likely to increase or decrease in frequency (low mutability and high targetability or high mutability and low targetability, respectively). (*c*) Mutability–targetability plots computed for all the amino acid residues at different evolutionary distance (L2, L5 and L7). In order to avoid frequency-related biases, we normalized all mutation frequencies before computing mutabilities and targetabilities for each amino acid (for more information refer to §7).

(like in the case of serine) relatively moderate from a mutational perspective (figure 2*a*), unlike serine's, the two groups of codons that code for leucine, CTN and TTR, are relatively close to each another. As a direct consequence of this close mutational distance between the two groups of codons, the two extra codons that leucine is coded by (TTR) only provide leucine with direct mutational access to six extra codons, compared with eight extra codons that can be directly accessed from serine's two extra codons (AGY). In addition, half of the new codons that can be accessed from leucine's two extra codons are stop codons (TAA, TAG and TGA). Therefore, it can be concluded that, given their mutational proximity to themselves and to stop codons, the six leucine codons cannot contribute to making leucine a more mutable and targetable residue.

On the other hand, arginine which is coded, similar to leucine, by two groups of codons that are relatively close to each other in mutational space (CGN and AGR), would have the potential to have higher mutability and targetability but is probably affected by its extreme physico-chemical properties (charged and large residue), preventing many amino acid substitutions due to natural selection acting against them.

Overall, no other amino acid residue is encoded by as many codons so far apart from each other in mutational space which, combined with its weaker physico-chemical properties, make serine a fast-evolving, mutational hub.

In conclusion, despite the fact that other causes such as bioenergetic costs or tendency to reside in fast-evolving protein regions are also plausible explanations for the mutational properties of the differences residues, we argue that these differences are founded on the mutational and physico-chemical distance from each amino acid residue to every other one of them.

5. IMPLICATIONS FOR PHOSPHORYLATION SITE EVOLUTION

Having described the general mutational properties of the different amino acid residues, we wanted to investigate to what extent cells can modulate these general properties for specific residues. Given the large mutational differences between serine (the fastest-evolving amino acid residue), threonine (a relatively high-evolving residue) and tyrosine (a rather slow-evolving residue), we investigated the consequences that different mutability and targetability may have for protein phosphorylation and evolution of phosphorylation sites.

If these general mutational properties were maintained in phosphorylation sites, one would expect to see fast removal of non-functional phosphorylation sites and fast introduction of a high number of new phosphorylation sites for fast-evolving residues (with high mutability and targetability) like serine or threonine. On the contrary, one would expect to see higher conservation for slow-evolving residues such as tyrosine. We have illustrated these different scenarios for serine, threonine and tyrosine (figure 6*a*).

To test this hypothesis, we computed the sequence conservation of human phosphorylated serines, threonines and tyrosines, and used the sequence conservation of these residues regardless of phosphorylated state as baseline for comparison (figure 6*b*). In addition, to discard the possibility that our results are driven by difference in the likelihood of residues to reside on disordered regions of proteins, we included disorder predictions in our results. In general, our results show the expected trend with phosphorylated residues being more conserved than non-phosphorylated residues, highlighting the likelihood that these are functional sites [7]. Moreover, in line with the general trend for the three residues observed earlier (figure 5*c*), phosphorylated serines and threonines are also much more conserved than phosphorylated tyrosines. Nevertheless, our results (figure 6*b*) also highlight some important subtleties that differ from our previous observations that serine is the most mutationally active residue (figure 5*c*); for instance, we observed a higher degree of conservation for phosphorylated serines than for phosphorylated threonines. Since the overall trend is maintained when taking into account disorder predictions, we can conclude that these observations are not driven by different disorder propensity of the different amino acid residues.

Moreover, we computed the fraction of amino acid residues that were excluded from our analysis because they reside in alignment gaps (see the electronic supplementary material, figure S1), which allowed us to refute the possibility that our observations could be explained by major differences in the propensity of different residues to reside in alignment gaps. In theory, a higher gap propensity of tyrosine (and to a lesser extend threonine) with respect to serine could be a trivial explanation for the different degrees of conservation we observe, because we would have excluded them from our analysis, but the gap propensities we computed do not support this hypothesis.

These results suggest that the cell is indeed capable of modulating the general mutational properties of amino acid residues under special circumstances. Moreover, the higher conservation of phosphorylated serines in comparison with phosphorylated threonines (observation which is in agreement with previous published work [8]) suggests that serine phosphorylation sites have more ancient functional properties, whereas threonine phosphorylation sites have more recent ones. Finally, it is perhaps surprising that the phosphotyrosine system, the signalling system that has appeared most recently in evolution [9], also presents the highest conservation. However, this apparent contradiction can be resolved if this system did not evolve gradually, but instead it evolved by a sudden burst (as supported in the literature [9–11]) and has subsequently remained more conserved (at least at the sequence level) than the phosphoserine or phosphothreonine systems.

6. DISCUSSION AND CONCLUSIONS

In this study, we have uncovered natural forces driving different mutability (probability that a given amino acid residue will be mutated) for different amino acid residues as well as different targetability (probability that a given amino acid residue will be the result of a mutation). These inequalities have made apparent different evolutionary paths for different amino acid residues, with some being slow-evolving and relying for their existence on high conservation (low

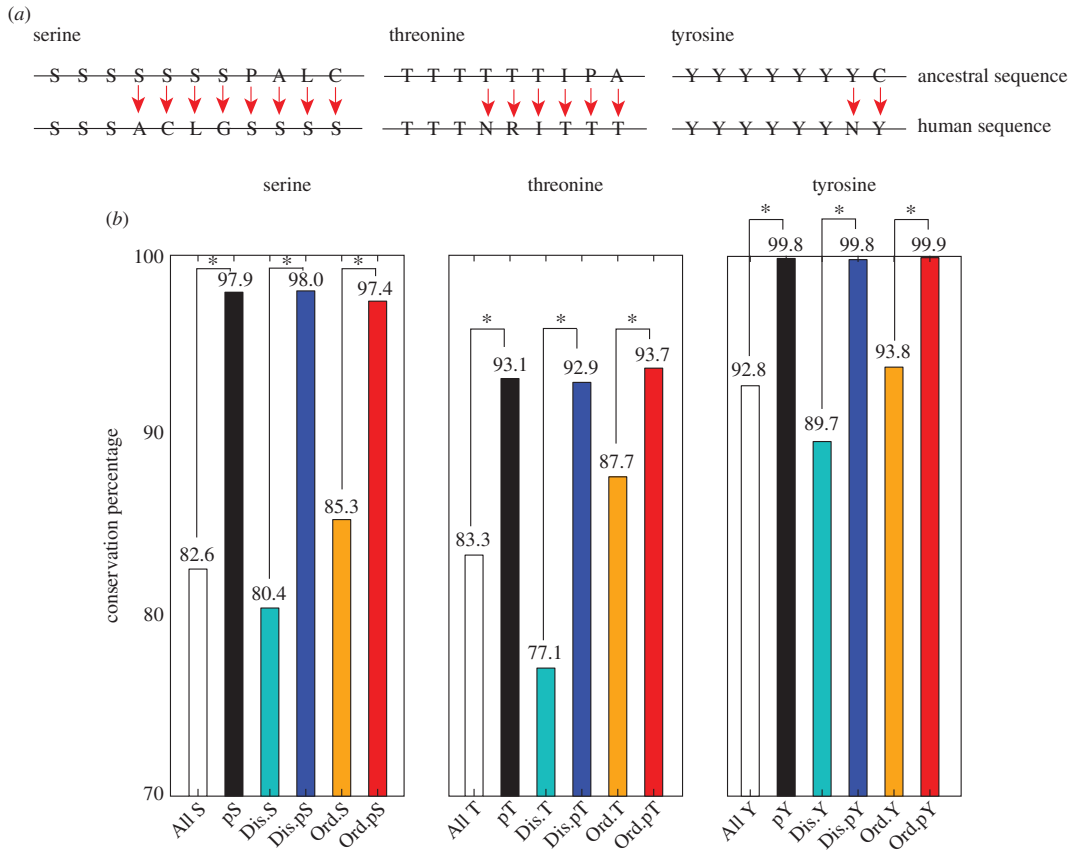


Figure 6. Phosphorylation site evolution. (*a*) As a direct consequence of our previous results, our hypothesis was that phosphorylation at serine, with the highest mutability and targetability rates, would evolve faster than phosphorylation at threonine, which has slightly lower but still relatively high mutability and targetability rates, and much faster than tyrosine, with very low mutability and targetability rates. (*b*) For each phosphorylatable amino acid residue, we have computed the fraction of phosphorylated residue (in black) that is conserved (same amino acid) in the ancestor L7. For comparison, we have computed conservation fraction for the residue, regardless of phosphorylated state (white) and conditional on whether the residue resides in a disordered (cyan both phosphorylated and unphosphorylated residues and blue only phosphorylated residues) or ordered protein region (orange both phosphorylated and unphosphorylated residues and red phosphorylated). Statistical significance ($*p < 0.05$) was assessed by Fisher's exact test.

mutability and targetability), such as tryptophan, and others being fast-evolving and relying for their existence on a high number of mutations leading to it (high mutability and targetability), such as the most mutationally active residue, serine.

In addition, we have computed matrices at different evolutionary distances (and therefore with different degrees of contribution from the genetic code), which may be important for assessing mutations for diseases such as cancer, where somatic mutations are accumulated through a fast evolutionary process. In essence, using our MUTA matrices, computed on short evolutionary distances and highly constrained by the genetic code, one should be able to compute the likelihood of different amino acid residue substitutions occurring in diseases associated with alterations to the genome.

Given the influence that both the genetic code and the physico-chemical properties of amino acid residues have on mutability and targetability, it would perhaps be natural to explore whether some causal relationships exist between them. While several hypotheses for the evolution of the genetic code exist [12], perhaps the most accepted view is that the organization of the genetic code can be explained by a combination of the occupation of codon space by the new amino acid residues as soon as they appeared from their predecessors, and optimization, to some extent, driven by physico-chemical properties of amino acid residues. We therefore argue that the observed mutability and targetability of amino acid residues is, at a higher degree, a consequence rather than a cause of the organization of the genetic code.

Finally, by contrasting the general trends in conservation of the phosphorylatable residues in human (S, T and Y) to the conservation of phosphorylation sites (pS, pT and pY), we have uncovered a higher degree of conservation for phosphorylation sites on serine than expected. This would suggest that the cell can modulate the mutational properties of amino acid residues in special circumstances as, in the case of serine,

given their ancient functional importance, it prevents these residues from evolving as fast as they would under normal circumstances. In line with this perception, it has been reported recently that aspartic and glutamic acid tend to become phosphorylatable residues (serine and threonine) during evolution, in a mechanism that has been suggested as a transition from a static to a dynamic regulation of protein folding [13]. Because these two groups of residues are far from each other in mutational space (two mutations away), we can conclude that, similarly as we found for the phosphotyrosine signalling system in our previous work [8,14], this observation is likely to be driven by positive selection.

It will be important to unravel the plausible mechanisms that have led to different mutability and targetability rates (e.g. amino acid preference to residue in fast-evolving or disordered regions), whether different species with different genetic codes or codon preferences have different residues' mutational properties, and to what extent these properties determine the frequency of every amino acid residue. Moreover, it will also be important to implement tools that can use these metrics to assess the importance of mutations in cell signalling systems associated with cancer progression. We argue this will eventually lead to a better foundation for network-based medicine.

7. MATERIAL AND METHODS

(a) *Alignments and computation of ancestral sequences*

Sequences of known and inferred proteins of 11 vertebrate species, including *Homo sapiens*, with at least 6X genome coverage were retrieved from the Ensembl online database (release 55) at http://jul2009.archive.ensembl.org/info/data/ftp/. These 11 metazoan species are *H. sapiens* (human), *Pongo pygmaeus* (orangutan), *Cavia porcellus* (guinea pig), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Monodelphis domestica* (opossum), *Canis familiaris* (dog), *Bos taurus* (cow), *Ornithorhynchus anatinus* (platypus), *Gallus gallus* (chicken) and *Xenopus tropicalis* (frog). The InParanoid algorithm (v. 2.0) [15] was used to infer orthologous sequences of human proteins across the ten other vertebrate species using the retrieved proteomes. The BLOSUM80 scoring matrix is used with other default parameters in InParanoid. In all cases, only the longest translation of each known/inferred genes was fed into InParanoid for orthologue prediction. The sequence of each known human phosphoprotein was then grouped with its inferred orthologous protein sequences for multiple sequence alignment using the mafft algorithm (v. 6.240, E-INS-i option with default parameters) [16]. Ancestral sequences were inferred from each multiple sequence alignment using the Codeml program in PAML phylogenetic software suite [17]. The phylogenetic relationship depicted in figure 2*b* [18] was input to Codeml with CodonFreq = 2 and using WAG substitution matrix [19].

(b) *From coevolution matrices to mutability and targetability rates*

For each pair of ancestral-human sequences, we computed a 20×20 coevolution matrix describing the evolution tendency of each amino acid, with the ancestral amino acid in the row position and human-aligned residue in the column position. In order to avoid inaccuracies caused by alignment positions with lower quality, we filtered out alignment positions in or next to gaps (see the electronic supplementary material, figure S1 for more information on the fraction of residues excluded). We produced mutability and targetability rates by normalizing the coevolution matrices by row, i.e. effectively balancing out differences in amino acid residue frequencies. The mutability rate for each residue is then measured as the sum of all mutation frequencies, i.e. row sum minus conservation. On the other hand, the targetability rate is measured as the sum of mutation frequencies of all amino acid residues leading to a given amino acid residue, i.e. column sum minus conservation.

(c) *Phosphorylation site evolution*

We have traced the evolution of human phosphorylation sites on serine, threonine and tyrosine by measuring the fraction of each that is conserved versus the fraction that has appeared recently in evolution. More specifically, we compiled a list of human phosphorylation sites obtained from the PhosphoSitePlus [20] and phosphoELM databases [21] and computed what fraction of those are conserved in our inferred ancestral sequences and thus compared how the three signalling systems have evolved. In order to predict disorder propensity for all the proteins analysed, we ran Disopred v. 2.0 [22].

We thank the editor Tony Hunter for critical input on this manuscript. R.L. is a Lundbeck Foundation Fellow. R.L. is further supported by a Sapere Aude Starting Grant from The Danish Council for Independent Research and a Career Development Award from the Human Frontier Science Program.

REFERENCES

1 Jacob, F. 1977 Evolution and tinkering. *Science* **196**, 1161–1166. (doi:10.1126/science.860134)
2 Pastor-Satorras, R., Smith, E. & Solé, R. V. 2003 Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210. (doi:10.1016/S0022-5193(03)00028-6)
3 Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
4 Dayhoff, M. O., Schwartz, R. & Orcutt, B. C. 1978 A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, vol. 5, suppl. 3 (ed. M. O. Dayhoff), pp. 345–352. Washington, DC: National Biomedical Research Foundation.
5 Henikoff, S. & Henikoff, J. G. 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 109 15–109 19. (doi:10.1073/pnas.89.22.10915)
6 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
7 Linding, R. 2010 (R)evolution of complex regulatory systems. *Sci Signal.* **3**, eg4. (doi:10.1126/scisignal.3127eg4)
8 Tan, C. S. H., Pasculescu, A., Lim, W. A., Pawson, T., Bader, G. D. & Linding, R. 2009 Positive selection of

tyrosine loss in metazoan evolution. *Science* **325**, 1686–1688. (doi:10.1126/science.1174301)
9 Lim, W. A. & Pawson, T. 2010 Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* **142**, 661–667. (doi:10.1016/j.cell.2010.08.023)
10 Pincus, D., Letunic, I., Bork, P. & Lim, W. A. 2008 Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc. Natl Acad. Sci. USA* **105**, 9680–9684. (doi:10.1073/pnas.0803161105)
11 Manning, G., Young, S. L., Miller, W. T. & Zhai, Y. 2008 The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc. Natl Acad. Sci. USA* **105**, 9674–9679. (doi:10.1073/pnas.0801314105)
12 Koonin, E. V. & Novozhilov, A. S. 2009 Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111. (doi:10.1002/iub.146)
13 Pearlman, S., Serber Jr, Z. & Ferrell, J. E. 2011 A mechanism for the evolution of phosphorylation sites. *Cell* **147**, 934–946. (doi:10.1016/j.cell.2011.08.052)
14 Tan, C. S. H., Schoof, E. M., Creixell, P., Pasculescu, A., Lim, W. A., Pawson, T., Bader, G. D. & Linding, R. 2011 Response to comment on positive selection of tyrosine loss in metazoan evolution. *Science* **332**, 917. (doi:10.1126/science.1188535)
15 Remm, M., Storm, C. E. & Sonnhammer, E. L. 2001 Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052. (doi:10.1006/jmbi.2000.5197)
16 Katoh, K. & Toh, H. 2008 Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinformatics* **9**, 286–298. (doi:10.1093/bib/bbn013)
17 Yang, Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)
18 Hedges, S. B. 2002 The origin and evolution of model organisms. *Nat. Rev. Genet.* **3**, 838–849. (doi:10.1038/nrg929)
19 Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699. (doi:10.1093/oxfordjournals.molbev.a003851)
20 Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. & Sullivan, M. 2012 PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–D270. (doi:10.1093/nar/gkr1122)
21 Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J. & Diella, F. 2010 Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res.* **39**, D261–D267. (doi:10.1093/nar/gkq1104)
22 Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. 2004 The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139. (doi:10.1093/bioinformatics/bth195)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 3058
doi:10.1098/rstb.2012.0384

Correction

Phil. Trans. R. Soc. B **367**, 2584–2593 (19 September 2012) (doi:10.1098/rstb.2012.0076)

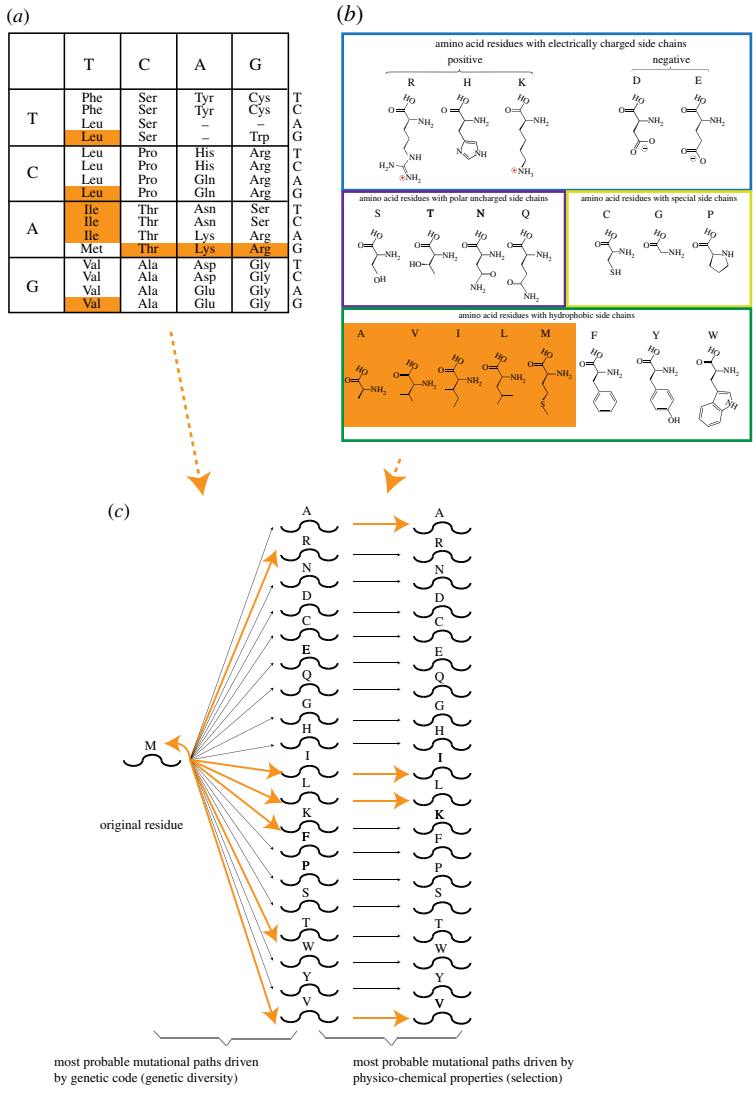
Research article: Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues

Pau Creixell, Erwin M. Schoof, Chris Soon Heng Tan and Rune Linding

Part (*a*) of figure 2 incorrectly highlighted some amino acid residues that are more than one mutation away from methionine. In line with this, part (*c*) erroneously portrayed phenylalanine as being close to methionine in mutational space, and in the main text (§2, THE INFLUENCE OF THE GENETIC CODE ON MUTATIONAL PATHS) it was incorrectly stated that phenylalanine is one of the closest residues to methionine. The corrected figure can be found below. This error does not affect any of our results or conclusions.

(*a*)

	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	–	–	A
	Leu	Ser	–	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G



Part III

Combining NGS and MS data to close the genotype-to-phenotype gap

Chapter 5

Genome-specific MS uncovering hidden networks

Cancer cells and their signaling networks need to be monitored with technologies that allow global and quantitative views of the system upon stimulation with different cues. Mass Spectrometry (MS) is clearly one of the best techniques for this purpose, however, precise global (phospho-)proteomic analysis and monitoring of cancer cells has, for a long time, been hampered by the use of standardized reference databases that mask mutation-associated signaling events. In this article, by combining Next-Generation Sequencing (NGS) and MS data, we observe a direct correlation between MS observability and the fraction of sequencing reads reporting a variant allele, a down-regulation of mutant protein expression and demonstrate how otherwise-hidden cancer-associated signaling networks can be revealed using this strategy.

Uncovering hidden signaling networks by genome-specific proteomics

Erwin M. Schoof*[1], Pau Creixell*[1], Adrian Pasculescu*[2], Agata Wesolowska-Andersen[3], Ramneek Gupta[3] and Rune Linding[1@]

[1]Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), Building 301, DK-2800, Lyngby, Denmark. [2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [3]Functional Human Variation Group, Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), Building 301, DK-2800, Lyngby, Denmark. *These authors contributed equally to this work. [@]Correspondence should be addressed to R.L. (linding@cbs.dtu.dk).

Network biology aims to predict phenotype from multi-scale models of cellular information processing, by integration of quantitative, genome-scale data. While Mass Spectrometry (MS) enables comprehensive sampling of cellular (phospho-)proteomes, the use of wild-type reference sequences results in masking of mutation-associated signaling events. Here we present an integrative strategy combining MS with exome sequencing to perform genome-specific proteomic analysis. Deploying the approach on a colorectal cancer cell line, we uncovered an otherwise-hidden signaling network spanning 177 mutant proteins and 30 mutant phospho-peptides. We observed a direct correlation between the fraction of sequencing reads reporting a variant allele and the likelihood of a mutation to be observed by MS. Additionally, we found a significant decrease in the number of mutant peptides detected by MS compared to wild-type, suggesting a down-regulation of mutant protein expression in cancer cells. We show that genome-specific proteomics experiments enable orthogonal cross-validation of DNA mutations and monitoring of dysregulated signaling networks.

The fields of proteomics and genomics provide complementary views that are essential to integrate in order to predict and understand cellular phenotypes. By combining the two, information linking mutations at the DNA level to amino acid sequences at the protein level can be assessed. This provides a starting point to make inferences about the functional effects of such lesions, thus shaping a far more complete picture of the functional impact of mutations than genomic or proteomic studies alone. Before conclusions can be drawn about the expression of mutations at the protein level, and their potential role in altering cellular information processing, it is imperative they are directly observed experimentally. Here, we present a new strategy for including prior knowledge about the genome of a biological system being probed when conducting mass spectrometry (MS) based proteomics studies. Next Generation Sequencing (NGS) can provide information about mutations occurring throughout the entire genome, and recent advances in the MS field have led to a significant portion of the expressed proteome to be readily observable [Wiśniewski et al., JPR 2009, Beck et al., MSB 2011, Geiger et al., MCP 2012, Munoz et al., MSB 2011] Combined, these technologies pave the way to genome-wide investigations at both the DNA and protein level. Additionally, through optimized enrichment procedures, a large set of PTMs (such as phosphorylation, acetylation and ubiquitination) can be analyzed through MS based studies. Thus, by identification and quantitation of thousands of modified peptides in a single experiment, it is possible to globally monitor altered signaling dynamics in any given biological system.

Traditionally, the raw data produced by an MS experiment, representing the peptides present within a sample, is matched to a database of reference protein sequences, in order to identify the observed peptide spectra using hypothetical spectra derived from *in silico* digestions. If a mutation occurs in these proteins however, the experimental spectra will not match with their theoretical spectra, leading to either misidentification or lack of identification of these proteins (Figure 1A-D). As many mutations have been attributed to play a role in disease [Wong et al. Annu.Rev. Genomics Hum. Genet. 2011, Greenman et al., Nature 2007], it is imperative that the protein dynamics associated with these mutations can be studied. We demonstrate that through the use of a genome-specific spectra database, obtained from accompanying NGS experiments, we can improve the number of identified proteins and phosphorylation sites due to identification of the mutated proteins and peptides, allowing for more accurate signaling network reconstruction (Figure 1E-F). As a proof of principle, we here deployed the HT-29 colon cancer cell line. This cell line has previously been described in terms of its copy number variation, mRNA expression, phospho-proteomics and morphology [Yasui et al., Cancer Research 2004, Schlag et al., Gut 2000, Reichelt et al., Anticancer Res, Kim et al., JPR 2005, Le Bivic et al., PNAS 1988]. However, only limited protein code impacting mutations

have been identified [Ikediobi et al., Mol Can Ther 2006]. We thus analyzed the cell line through deep (phospho-) proteomic and genomic analysis. Using the Q-Exactive Orbitrap platform (Thermo Fisher Scientific) we identified 8,122 unique proteins and 27,872 unique phosphorylation sites in the HT-29 colon cancer cell line including 1,338 phosphorylated tyrosine (pTyr), 22,322 phosphorylated serine (pSer) and 4,212 phosphorylated threonine (pThr) residues. Using the HiSeq platform (Ilumina), we performed exome sequencing of HT-29 cells with an average depth of 80X and >95% of all reads at 10X or more. This resulted in the identification of 7,234 missense variants, equating to 4,561 altered mutated protein sequences with respect to the human reference genome.

The MS spectra generated by the proteomics experiments were searched using several variations of the sequence database. When using the reference Ensembl database, we identified 7,560 proteins and 26,072 phosphorylation sites (of which 14,848 were confidently localized sites); in comparison, by including all possible single mutant proteins in the search database, these numbers increased to 8,122 proteins, 27,872 phosphorylation sites (of which 15,808 were confidently localized sites). In total, we identified 562 additional proteins and 960 confidently localized additional phosphorylation sites by utilizing the genome-specific information for analyzing the proteomics data instead of the reference Ensembl database (Figure 2A). This is a 6-7% increase in identifications compared to using the reference database alone and opens the possibility for looking at the dynamics of the mutant proteins and phosphorylation sites, in addition to being able to distinguish technical artifacts from real variants that are present in the biological sample being studied.

Next, we attempted to investigate a long-standing question of how many mutations are actually expressed at the proteome level. We analyzed whether the fraction of reads reporting a variant allele originating from NGS data, could be correlated to the likelihood of observing the mutation in the proteomics data. From the distribution of peptides (Figure 2B) containing a mutation with a given fraction of reads reporting a variant allele, it is clear that this measure is directly related to the expression rate of the peptides bearing this mutation. In other words, the higher the number of NGS reads of a given mutant allele, the higher the likelihood of identifying the mutant peptide using mass spectrometry, providing orthogonal validation whether a given mutation is present at the genomic level and also expressed at the proteome level.

In order to investigate how many of the variant sites were actually expressed in the cells, we generated all possible tryptic peptides *in silico*, in order to see what percentage was observed. In total, we observed 193 mutated tryptic peptides (of which 30 were phosphorylated peptides), correlating to a 1.5-3.7% of total mutant peptides possibly observable. Compared to the peptide coverage of the non-mutated proteome (7.6% as can be seen from Fig. 2C), this percentage is significantly lower than expected by chance (p-value <10e-6, Wilcoxon test), especially for mutant peptides where the fraction of NGS reads reporting the mutant allele is 1.0 (meaning it is a homozygous mutation and thus the protein should only be present in its mutant form). We propose two explanations behind this significant difference in MS observability between wild-type and mutant peptides: Firstly, it is a possibility that the mutant peptides are being matched to the wild-type sequence through incomplete b- and y-ion series coverage. This could lead to the mutated amino acid not having been observed in the mass spectrum, causing the search algorithm to identify it as wild-type. Secondly, it is possible that mutant proteins show a lower degree of expression than wild-type proteins, rendering them undetectable in the MS experiment performed here. This second hypothesis is well supported by recent findings in another study [Shah et al., Nature, 2012].

By combining the two technologies as described, our method provides a platform to conduct orthogonal cross-validation of mutations using NGS and MS data. More specifically, we investigated whether we could identify cases where there was disagreement between NGS and MS results. Focusing our attention on mutations with the highest NGS evidence of a homozygous mutation (fraction of reads reporting a mutation being 1), we identified three cases for which MS only identified the wild-type peptide. While we cannot conclude that this mutation is not present (the mutant peptide may have simply not been detected by the MS), we can conclude that these positions were either incorrectly identified as mutations by the NGS data or that, at the very least, the reported mutation is heterozygous. Out of all the possible mutated peptides with a reported mutant allele frequency of 1, we detected 82 peptides in our MS data (Online Supplementary Table 1). As three of these peptides were found only as wild-type variants, our data suggests an NGS error rate of around 3.7%. Despite the small sample size of our observation, if we extend this rate to our whole data set, we would estimate that of the 7,234 missense variants reported by NGS in this study, 267 could be false positives.

Of the total number of identified proteins, 9 were identified solely based on their mutated peptides (Online Supplementary Table 2). Given that no wild-type peptides were observed, these proteins would have never

been identified if the reference genome database had been used. For the other 168 proteins, for which at least 1 mutated peptide was observed (Online Supplementary Table 3), the traditional MS approach would have only identified the wild-type variant, while a genome-specific approach is able to identify the mutant variant as well. According to our sequencing results, the number of proteins containing at least one missense mutation, and therefore potential number of additional protein variants identifiable by MS, is 4,561. It is well established that a single amino acid variant can have a significant impact on protein function [e.g. Davies et al., Nature 2002; Songyang et al., Nature 1995], hence underlining the importance of being able to observe these mutant variants.

Additionally, we sought to investigate mutations hitting the region surrounding observed phosphorylation sites. In total, we identified 30 mutated peptides with confidently localized phosphorylation sites using our genome-specific database (Online Supplementary Table 4). In order to assess the 'systems effect' of identifying these genome-specific phosphorylation events, we reconstructed a signaling network model containing all HT-29 phosphorylation sites that would have been missed had we not used a genome-specific proteomics approach. By computational modeling of the upstream kinases using NetworKIN [Linding et al., Cell 2007, Linding et al., NAR 2008] (Figure 3), it seems that several PKC-family members interact with a number of proteins harboring a mutation, indicating a potential involvement in transcriptional regulation and cell migration [Masur et al, Mol Bio Cell 2001]. Additionally, cell cycle related kinases such as PLK1, PLK4 and several CDK-family members seem to interact with a subset of mutated proteins, suggesting these mutations may affect cell cycle or mitosis related signaling in this cell line. These results seem to confirm previous observations of HT-29 being sensitive to Polo Box domain expression levels and responsive to PLK1 inhibition [Fink et al., Mol. Canc. Ther., 2007, Rödel et al., Am J. Pathol, 2010].

In this study, we have provided proof-of-concept of the importance of integrating different types of 'omics data, in order to obtain an accurate foundation for the reconstruction of cellular signaling networks. Due to recent advances in MS technology, it is now possible to obtain deep coverage of the proteome and phospho-proteome, which can be complemented by exome-wide deep sequencing data. While custom MS databases for specific applications have been used in the past [Cheung et al., NBT 2012], we have here generalized and extended the concept of taking into account genome-specific protein sequence information, allowing the identity and dynamics of the mutant proteome to be investigated. In order to assess the functional impact of mutations at a systems level, MS is a key technology, as it allows the analysis of tens of thousands of proteins and phosphorylation sites from a single sample. Considering the ever improving dynamic range in mass spectrometry, the number of observed mutant peptides, while currently relatively modest though significant, will increase in future studies. We demonstrate that conducting genome-specific proteomics experiments is now feasible, even in an un-targeted, global MS setting. It is likely that additional benefits could be gained by deploying a targeted MS approach. Targeted proteomics such as SRM [Picotti et al., Cell 2009, Wolf-Yadlin et al., PNAS 2007] may be the best proteomics strategy to monitor mutant peptides and proteins, as global approaches can currently not guarantee that this specific part of the proteome will be represented in the MS results due to the inherent dynamic range limitation. This is also likely to explain why a large proportion of the mutations reported by the sequencing data could not be observed in the global MS results.

Based on our comparison of experimentally observed wild-type versus mutant peptides, the total number of possibly observable mutant peptides can be up to 30-fold higher than reported in this study. Given the rapid progression in MS and NGS technology, the need for, and benefit of this method is likely to increase significantly in future personalized network medicine studies [Pawson & Linding, FEBS Lett 2008, Creixell et al., Nat Biotech. 2012, Vogelstein et al., Science 2013], where patient samples can undergo NGS and MS experiments to study a disease from the genomic and proteomic perspective, in order to guide the best possible therapeutic strategies. Additionally, it will most likely prove useful in distinguishing between key mutations driving a given disease state, or mutations arising sporadically. In conclusion, through the method described here, we can use the knowledge gained from NGS experiments in order to improve the sensitivity and accuracy of MS experiments, rendering the two technologies a very powerful combination for investigating complex diseases such as cancer, diabetes and neurological illnesses.

FIGURE LEGENDS

FIG.1

A & B) Limitations of unspecific MS

A) Conceptual overview of how a mutated protein is identified as a wild-type protein in an unspecific database search. Due to the lack of the mutated peptide in the reference database, only wild-type peptides are used for matching to the parent protein. B) Only when a genome-specific database is used, can the mutated peptide be matched to its parent sequence and is the correct variant of the protein identified.

C & D) Example of genome-specific mutant peptide identified with our approach.

MS spectra of wild-type (E) and mutant (F) versions of the same peptide. The mutant peptide becomes identifiable due to using an HT-29-specific database for conducting the MS data search.

E & F) Unspecific versus Genome-specific MS approach.

As opposed to previous unspecific MS approaches (C), our genome-specific approach (D) allows for a sample-specific search of MS data by exome sequencing the sample and generating a specific database. This approach allows the identification of mutant proteins that would otherwise be hidden and avoids the mismatching of spectra caused by the absence of a given mutant gene in standard reference databases.

FIG.2

A) Quantification of newly identified proteins and phosphorylation sites.

Pie charts showing the number of proteins and phosphorylation sites identified by searching our MS data using the standard approach or our genome-specific approach.

B) Comparison of reads reporting a mutant allele and MS observability.

As can be observed in this graph, the higher the fraction of reads that report a mutation, the higher the likelihood that this mutant peptide is observed in the MS data.

C) Density plot of overall MS observability.

Almost 8% of all wild-type tryptic peptides are experimentally observed in the Mass Spectrometer.

FIG. 3

Hidden (phospho-) proteome

Signaling network that became apparent only when using the genome-specific approach. Genome specific phosphorylated proteins are represented in red, phosphorylating kinases and binding SH2 domains predicted by NetworKIN are represented in blue and green, respectively.

ACKNOWLEDGMENTS

We would like to thank members of the Linding Lab and the Erler Lab (BRIC, Denmark) for useful input on the manuscript. This work was supported by the Lundbeck Foundation, the Human Frontier Science Program, the Danish Council for Independent Research.

AUTHOR CONTRIBUTIONS

R.L. conceived the project. E.M.S., P.C. and R.L. designed the experiments. E.M.S., P.C. and A.P. performed the experiments. E.M.S., P.C., A.P., and A.W.A analyzed the data, and E.M.S., P.C., A.P. and R.L. wrote the paper. R.G. supervised the genomic analysis. R.L. oversaw the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

Beck et al., The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549 (2011).

Cheung et al., A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature Biotechnology* **30**, 447-452 (2012)

Creixell, P. *et al.* Navigating cancer network attractors for tumor-specific therapy. *Nature Biotechnology* **30**, 842-848 (2012).

Davies, H. *et al*. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002)

Fink et al., Cell type–dependent effects of Polo-like kinase 1 inhibition compared with targeted polo box interference in cancer cell lines. *Mol Cancer Ther* **6**;3189 (2007).

Geiger et al., Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **11**(3) (2012)

Greenman, C. *et al*. Pattern of somatic mutation in human cancer genomes. *Nature* **446**,153–158 (2007).

Ikediobi et al., Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Molecular Cancer Therapeutics* **5**, 2606 (2006).

Kim et al., Global Phosphoproteome of HT-29 Human Colon Adenocarcinoma Cells. *Journal of Proteome Research* **4**, 1339-1346 (2005).

Le Bivic et al., HT-29 cells are an in vitro model for the generation of cell polarity in epithelia during embryonic differentiation. *Proc. Nati. Acad. Sci.* **85**, 136-140 (1988).

Linding, R et al., NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* **36**, D695-9 (2008)

Linding, R. *et al*. Systematic discovery of in vivo phosphorylation networks. *Cell* **129**,1415–1426 (2007).

Masur et al., High PKC a and Low E-Cadherin Expression Contribute to High Migratory Activity of Colon Carcinoma Cells. *Molecular Biology of the Cell* **12**, 1973–1982 (2001).

Munoz et al., The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol Syst Biol.* **7**:550 (2011)

Pawson, T. & Linding, R. Network medicine. *FEBS Lett.* **582**, 1266–1270 (2008)

Picotti et al., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138**, 795-806 (2009)

Reichelt et al., Early oncogene mRNA expression in HT-29 cells treated with the endogenous colon mitosis inhibitor pyroglutamyl-histidyl-glycine. *Anticancer Res.* **22**(2A):991-6 (2002).

Rödel et al., Polo-Like Kinase 1 as Predictive Marker and Therapeutic Target for Radiotherapy in Rectal Cancer. *Am J Pathol.* **177**(2): 918–929 (2010).

Schlag et al., Altered mRNA expression of glycosyltransferases in human colorectal carcinomas and liver metastases. *Gut* **46**:359-366 (2000).

Shah, S. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399(2012).

Songyang, Z. *et al*. Catalytic specificity of protein-tyrosine kinases is critical for selective signaling. *Nature* **373**, 536–539 (1995).

Vogelstein et al., Cancer Genome Landscapes. *Science* **339**(6127), 1546-1558 (2013)

Wiśniewski et al., Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res* **8**(12), 5674-8 (2009).

Wolf-Yadlin et al., Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *PNAS* **104**, 5860-5865 (2007)

Wong, K.M.M., Hudson, T.J. & McPherson, J.D. Unraveling the genetics of cancer: genome sequencing and beyond. *Annu. Rev. Genomics Hum. Genet.* **12**, 407–430 (2011)

Yasui et al., Alteration in Copy Numbers of Genes as a Mechanism for Acquired Drug Resistance. *Cancer Research* **64**, 1403–1410 (2004).

ONLINE METHODS

Sample preparation for sequencing and data analysis. HT-29 cells were grown to 80% confluency in a T-75 flask, and DNA extraction was performed using reagents and instructions provided with the Qiagen QIAamp DNA Mini kit. 5 ug of purified DNA were sent to Roche Nimblegen for full exome sequencing using the SeqCap EZ Human Exome Library v3.0 capture kit. High-quality reads, with > 80x mean coverage and > 95% of exome bases at 10x coverage, were obtained from sequencing and aligned to the NCBI37 reference human genome (version GRCh37) using the Burrows–Wheeler Alignment Tool. The alignment was refined by means of quality score recalibration and around indel realignment using Genome Analysis ToolKit package. SNP calling was performed with SAMtools package using default settings. Next, results were further filtered with VCFtools using standard default settings as well as a minimum 10x sequencing depth threshold set for SNP calling. The data was further analyzed with the help of SAMtools and BEDtools packages and custom-written Perl and Python scripts. Finally, fasta files for both wild-type and mutant protein sequences were generated using the Variant Effector Predictor (VEP) package from Ensembl.

Sample preparation for (phospho-)proteomics. HT-29 cells (obtained from ATCC and regularly checked for mycoplasma contamination) were grown to ~80% confluency in 15cm dishes to provide enough starting material for the phospho peptide enrichment in duplicate (24mg per repeat). Synchronized cells were lysed with ice-cold modified RIPA buffer supplemented with Roche complete protease inhibitor cocktail tablets and ß-glycerophosphate (5mM), NaF (5mM), Na-orthovanadate (1mM, activated). Lysates were sonicated on ice and spun down at 4,400xg for 20mins at 4°C. Proteins were precipitated over-night in ice cold Acetone at -20°C, and dissolved in 6M Urea, 2M Thiourea, 10mM HEPES pH 8.0. Proteins were reduced with 1mM DTT for 1hr, and alkylated with 5.5mM Chloroacetamide for 1hr, after which they were pre-digested with Lysyl Endopeptidase (Wako) at a 1:200 enzyme-to-protein ratio for 4hrs at room temperature (RT). Lysates were diluted 1:4 with 50mM Ammonium Bicarbonate, after which Trypsin (MS grade, Sigma) was added at a 1:200 enzyme-to-protein ratio and left rotating over-night at RT. Enzymatic activity was quenched by adding TFA to a final concentration of 2%, after which the samples were clarified by spinning down at 2,000xg for 5 minutes and desalted using 360mg SepPak columns (Waters WAT020515). Peptides were eluted using 2x 2mL of 40% AcN, 0.1% TFA, and 1x 2ml of 60% Acetonitrile, 0.1% TFA. For the global, Titanium Dioxide (TiO2) based phospho peptide enrichment, the eluent was directly subjected to SCX fractionation, where peptides were separated over a 0-30% Buffer B gradient in 60 minutes at a 1ml/min flowrate (Buffer A: 5mM potassium dihydrogen phosphate, 30% Acetonitrile, pH2.7; Buffer B: 5mM potassium dihydrogen phosphate, 30% Acetonitrile, 350mM potassium chloride, pH2.7). The resulting fractions were pooled according to their chromatography into 11 final samples, which were enriched for phosphorylated peptides. Six aliquots were taken at this point for the global proteome analysis. The TiO2 enrichment was conducted similarly to [Olsen et al., MSPP 2009], with several adjustments. For the TiO2 loading solution, 0.02g/ml dihydrobenzoic acid was dissolved in 30% Acetonitrile and 4% TFA, and the TiO2 beads were incubated in this solution for 15 minutes prior to peptide enrichment. Each pooled SCX fraction was enriched with 1.5mg of TiO2 beads suspended in 6ul of TiO2 loading solution, and left to rotate end-over-end for 30 minutes at RT. The flow-through (early eluting fractions) was enriched three times consecutively, whereas the single SCX chromatography peak peptide samples were enriched twice. Samples were spun at 2000xg for 5 minutes (RT), and pelleted beads were washed with 100ul SCX Buffer B. Subsequently, beads were pelleted again (2000xg, 5minutes, RT) and washed with 100ul 40% Acetonitrile, 0.25% acetic acid, 0.5% TFA. Finally, pelleted beads were re-suspended in 50ul 80% Acetonitrile, 0.5% acetic acid, and transferred to separate in-house packed C8 StageTips [Rappsilber et al., Nat Protoc 2007]. Liquid was spun through at 3000 rpm for 1 minute, after which the phosphorylated peptides were eluted with 1x 20ul 5% Ammonia and 1x 20ul 10% Ammonia, 25% Acetonitrile into a 96-well PCR plate, containing 20ul of 1% TFA, 5% Acetonitrile solution. Peptides were lyophilized to a total volume of 10ul, and acidified with 40ul of 1% TFA, 5% Acetonitrile, after which they were desalted on in-house packed C18 StageTips prior to LC-MS analysis.

For LC-MS analysis, peptides were eluted from the StageTip with 2x 20ul 80% Acetonitrile, 0.1% Formic acid, and lyophilized to 5ul final volume. The eluent was acidified with 1% TFA, 2% Acetonitrile and loaded onto a 50cm C18 EasySpray column (Thermo, ES803), using the Thermo EasyLC 1000 UHPLC system and the column oven operating at 45°C. Peptides were eluted over a 250 minute gradient, ranging from 6-60% of 80% Acetonitrile, 0.1% Formic acid, and the Q Exactive (Thermo) was run in a DD-MS2 top10 method. Full MS spectra were collected at a resolution of 70,000, with an AGC target of 3e6 or maximum injection time of 20ms and a scan range of 300-1750 m/z. The MS2 spectra were obtained at a resolution of 17,500, with an AGC target value of 1e6 or maximum injection time of 80ms. Dynamic exclusion was set to 20s, and ions with a charge state < 2 or unknown were excluded. For the proteome samples, the settings were the same, except for a gradient time of 230mins, maximum MS2 injection time of 60ms and dynamic exclusion of 45s.

The phospho-tyrosine samples were analyzed over a 360 and 480-minute gradient to maximize sample coverage.

For the pTyr specific phospho peptide enrichment, the SepPak eluent (equating to 24mg of peptides) was dried down overnight in a Thermo Express 250 concentrator and stored at -80°C. pTyr specific enrichment was conducted with the novel pTyr-1000 antibody from CST, using the protocols provided by the manufacturer, and the samples were run as technical duplicates on the LC-MS.

Computational analysis of MS data. In order to investigate the effect of using the HT-29-specific FASTA file as the MaxQuant (Version 1.2.7.4) search engine database, we performed the raw data searches in three ways: 1) standard Ensembl v.68 human FASTA, 2) standard Ensembl v.68 human FASTA + all possible single mutant proteins, and 3) all single possible mutant proteins only. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [Vizcaino et al., NAR 2013] with the dataset identifier PXD000267. Variable modifications were set as Methionine oxidation, Protein N-term acetylation and Serine/Threonine/Tyrosine phosphorylation, and Cysteine carbamidomethylation was set as a fixed modification. FDR rates were set to 1%, and the 'match between runs' functionality was activated.
Results from the three independent searches were stored in a MySQL database, and all further analysis was done using scripts written in-house on our "CoreFlow" platform, based on the R statistical package, MySQL and Python. All code and data will be released to the public upon request. Search results filtering was based on phosphorylation localization probability >= 0.75 and a minimum MaxQuant peptide ID score of 50, in order to only use high confidence identifications.
Peptide observability was calculated based on all possibly observable tryptic peptides originating from an *in silico* digest (minimum peptide length of 5 amino acids); the percentage of peptides observed was calculated using the following: peptides observed / total # of peptides observable x 100. For the percentage of Peptides Observed, the data size per bin of Fraction of Reads Reporting a Variant Allele was between 8 and 97 with an average of 21, and the average of total peptides with possible mutation per bin was 800. We considered the data size to be sufficient for the estimation of the percentage of observed peptides. For the MS observability of the non-mutated peptides, we used sampling of the appropriate size from the set of all `in-silico` digested peptides. The sample size was equal to the size of the data set of MS observability of the mutated peptides. To test for statistical significance of the difference in MS observability between the mutant and wild-type peptides, we applied a Wilcoxon statistical test, which does not rely on the assumption of normality or independence between data sets.
The NetworKIN modeling was based using an in-house up-to-date version of NetworKIN v3.0, and the 30 high confidence phosphorylation sites with a surrounding mutation were analyzed. Subsequently, kinase and SH2 domain predictions were filtered to only include predictions with a score of 0.1 and higher in order to reduce false positives. The results were plotted in Cytoscape (http://www.cytoscape.org) [Shannon et al., Gen. Res., 2003] for visual representation.

SUPPLEMENTARY REFERENCES:

Olsen, J et al., High accuracy mass spectrometry in large-scale analysis of protein phosphorylation. *Mass Spectrometry of Proteins and Peptides*, Volume **492**, Chapter 7 (2009)
Rappsilber et al., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols* **2** (8), 1896-906 (2007)
Shannon et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11):2498-504
Vizcaino JA, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res*. **41**(D1):D1063-9 (2013)

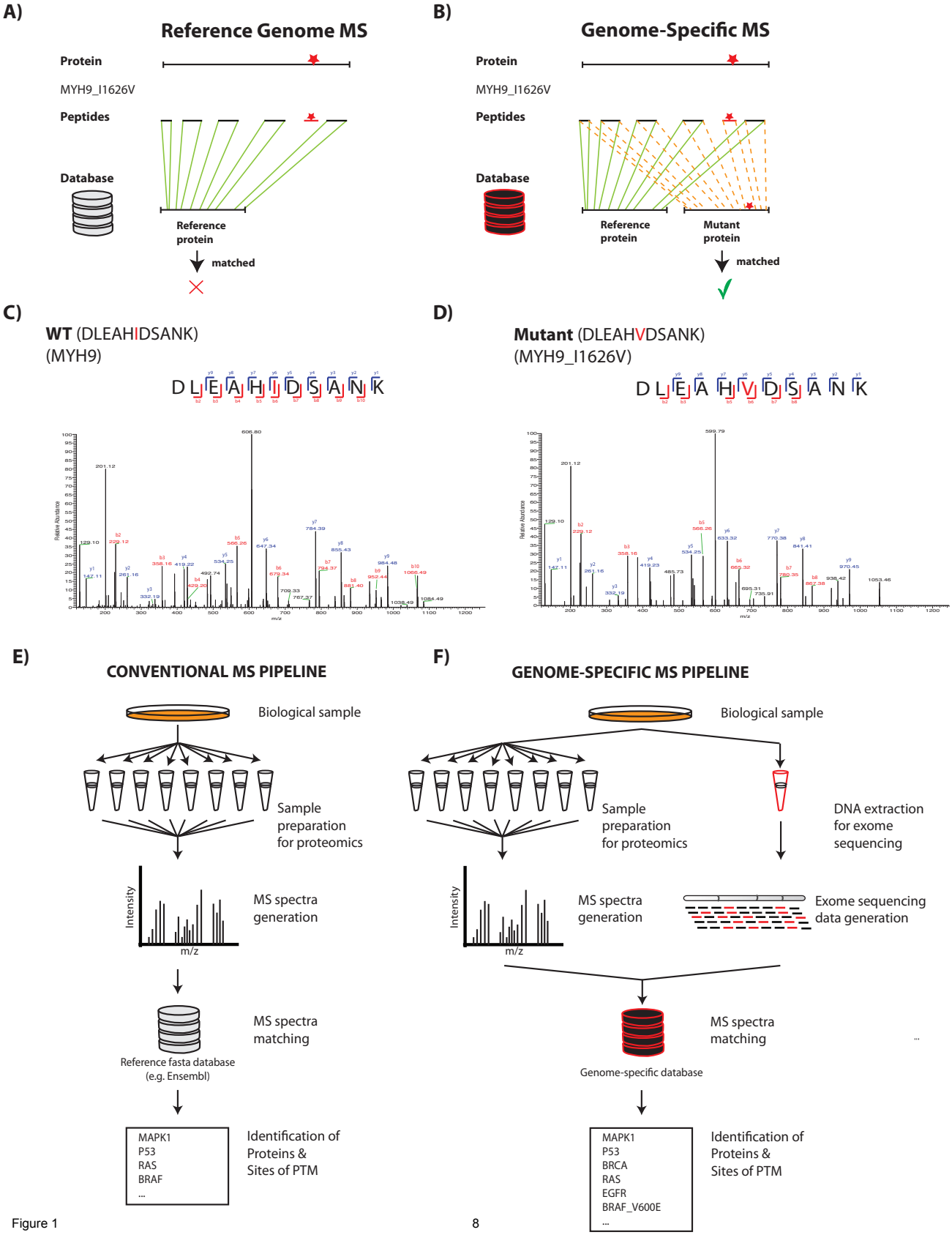
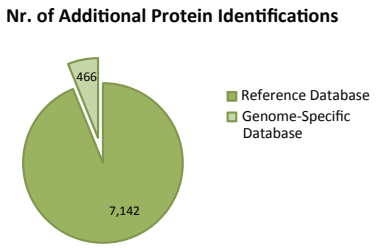
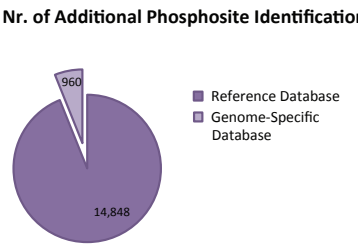

A)
Reference Genome MS
Protein
MYH9_I1626V
Peptides
Database
Reference protein
matched
B)
Genome-Specific MS
Protein
MYH9_I1626V
Peptides
Database
Reference protein
Mutant protein
matched
C)
WT (DLEAHIDSANK)
(MYH9)
D)
Mutant (DLEAHVDSANK)
(MYH9_I1626V)
E)
CONVENTIONAL MS PIPELINE
Biological sample
Sample preparation for proteomics
Intensity
m/z
MS spectra generation
MS spectra matching
Reference fasta database (e.g. Ensembl)
MAPK1
P53
RAS
BRAF
...
Identification of Proteins & Sites of PTM
F)
GENOME-SPECIFIC MS PIPELINE
Biological sample
Sample preparation for proteomics
DNA extraction for exome sequencing
Intensity
m/z
MS spectra generation
Exome sequencing data generation
MS spectra matching
Genome-specific database
MAPK1
P53
BRCA
RAS
EGFR
BRAF_V600E
...
Identification of Proteins & Sites of PTM

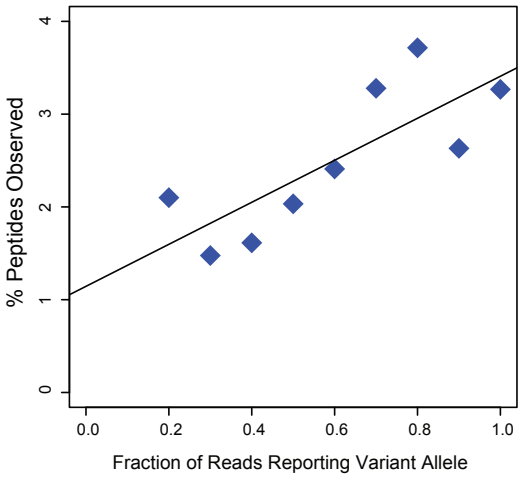
Figure 1

A)


Nr. of Additional Protein Identifications
466
7,142
Reference Database
Genome-Specific Database


Nr. of Additional Phosphosite Identifications
960
14,848
Reference Database
Genome-Specific Database

B)


% Peptides Observed
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Reads Reporting Variant Allele

C)

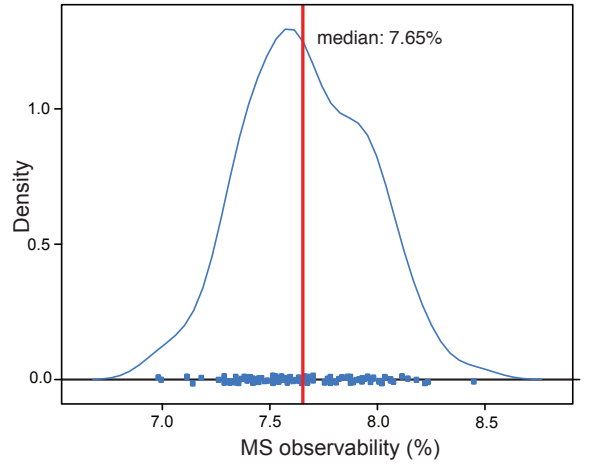

median: 7.65%
Density
0.0
0.5
1.0
7.0
7.5
8.0
8.5
MS observability (%)

Figure 2

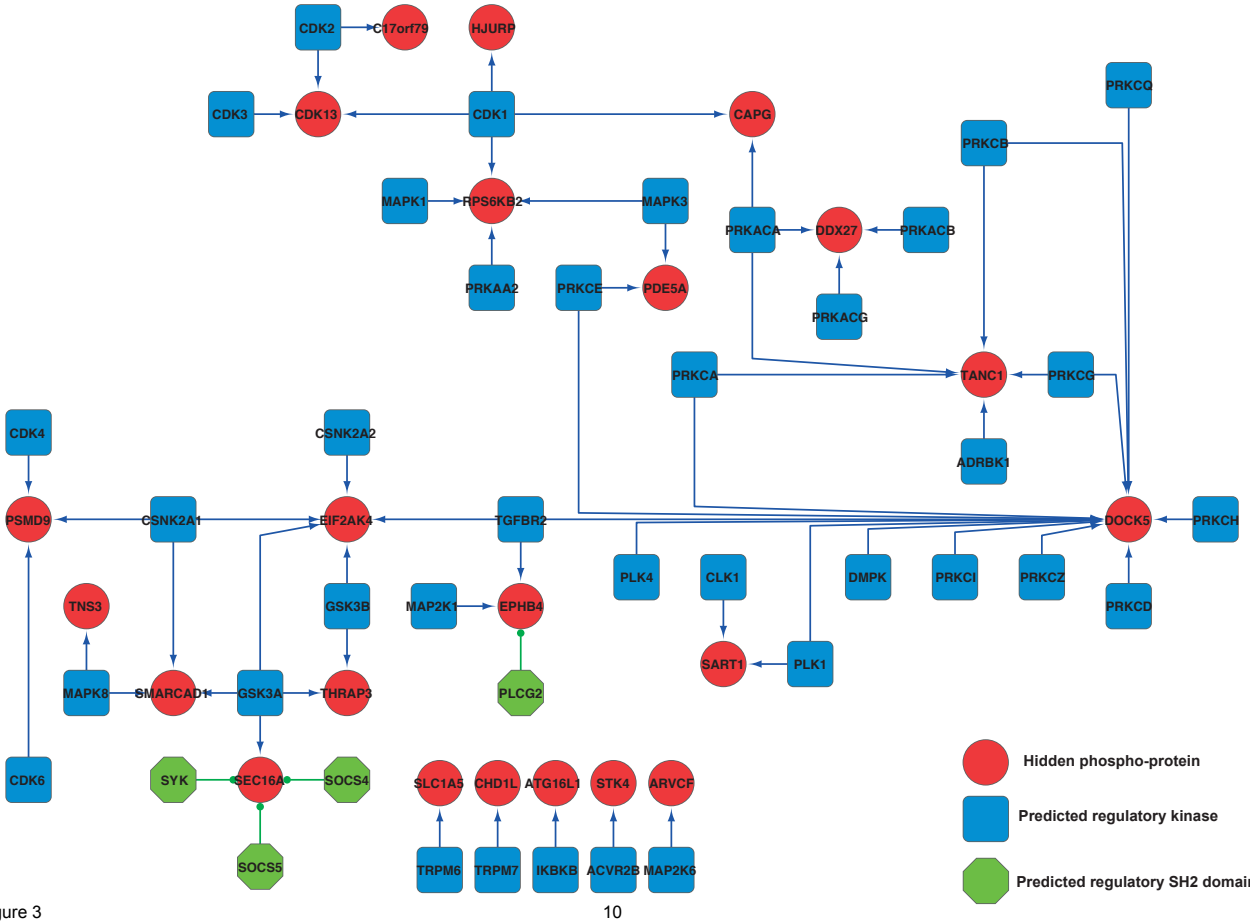

CDK2
C17orf79
HJURP
CDK3
CDK13
CDK1
CAPG
PRKCQ
PRKCB
MAPK1
RPS6KB2
MAPK3
PRKACA
DDX27
PRKACB
PRKAA2
PRKCE
PDE5A
PRKACG
PRKCA
TANC1
PRKCG
CDK4
CSNK2A2
ADRBK1
PSMD9
CSNK2A1
EIF2AK4
TGFBR2
DOCK5
PRKCH
PLK4
CLK1
DMPK
PRKCI
PRKCZ
PRKCD
TNS3
GSK3B
MAP2K1
EPHB4
SART1
PLK1
MAPK8
SMARCAD1
GSK3A
THRAP3
PLCG2
CDK6
SYK
SEC16A
SOCS4
SLC1A5
CHD1L
ATG16L1
STK4
ARVCF
SOCS5
TRPM6
TRPM7
IKBKB
ACVR2B
MAP2K6
Hidden phospho-protein
Predicted regulatory kinase
Predicted regulatory SH2 domain

Figure 3

Supplementary Table 1

Ensembl_Prot_Id	Ensembl_Gene_Id	Gene Name	Aminoacid Mutation	Mutation Ratio Bin	Tryptic pept original	Original Found in WT	Position in pept mut	Triptic as start mut	Tryptic pept mut	Position in pept mut alt	Tryptic pept mut alt	Razor Proteins	max of Scores
ENSP00000170168	ENSG00000079313	REXO1	S759P		1 TLAASGSQSSNGPFPGGQQLK		9 R		TLAASGSQPSNGPEPGGQQLK	9		ENSP00000170168	30.88
ENSP00000216780	ENSG00000100889	PCK2	Q121P		1 DTVQLPYPGGAR		4 R		DTVPLPYPGGAR	4		ENSP00000216780	30.87
ENSP00000221233	ENSG00000077348	EXOSC5	TSM		1 MEEETHTDAK		5		MEEEMHTDAK	5		ENSP00000221233	84.31
ENSP00000221957	ENSG00000105355	PLIN3	V275A		1 AQEALQLSQALSLMETVK		11 R		AQEALQLSQALSLMETVK	11		ENSP00000221957	102.37
ENSP00000225740	ENSG00000108602	ALDH3A1	P329A		1 YIAPTILTDVDPQSAVMQEEIFGPVLPIVCVR		15 R		YIAPTILTDVDPQSAVMQEEIFGPVLPIVCVR	15		ENSP00000225740	63.08
ENSP00000229854	ENSG00000112118	MCM3	E777K		1 DSEEPSSHEQAALSK		3 R		DSKEPPSSHEQAALSK	3		ENSP00000229854	119.23
ENSP00000242108	ENSG00000122547	EEPD1	S343N		1 AVVAEKPSSQLQK		9 K		AVVAEKPSNQLQK	9		ENSP00000242108	117.79
ENSP00000252137	ENSG00000100056	DGCR14	A423V		1 ASYTPSPAR		8 R		ASYTPSPVR	8		ENSP00000252137	129.54
ENSP00000254508	ENSG00000132182	NUP210	A755V		1 LTLAPVTPSPQLDMSCPLQQNK		4 R		LTLVPVTPSPQLDMSCPLQQNK	4		ENSP00000254508	89.03
ENSP00000254508	ENSG00000132182	NUP210	R786L		1 NPRLDLAAYDQEGR		3 R		NPLLDLAAYDQEGR	3		ENSP00000254508	57.8
ENSP00000255194	ENSG00000132842	AFAP1	V585E		1 QLAVPNKK		7 R		QLAVPNEK	7		ENSP00000255194	60.91
ENSP00000258654	ENSG00000136152	CDK3	N747S		1 VNDLAATAYK		2 K		VSDLAATAYK	2		ENSP00000258654	104.52
ENSP00000259008	ENSG00000136492	BRIP1	S919P		1 YSTSPVLLEAASHLSPENFVEDEAK		4 K		YSTPPVLLEAASHLSPENFVEDEAK	4		ENSP00000259008	33.54
ENSP00000259951	ENSG00000206462	HLA-F	P272S		1 WAAVVVPSGEEQR		8 K		WAAVVVPSQEEQR	8		ENSP00000366002	88.55
ENSP00000263635	ENSG00000115183	TANC1	N2515		1 DGNLRLGVQK		3 K		DGSLR	3	DGSLRLGVQK	ENSP00000263635	64.55
ENSP00000263657	ENSG00000115946	PND1	R11G		1 MESEMETQSARAEFGFTQVTR		11		MESEMETQSAGAEFGFTQVTR	11		ENSP00000263657	125.88
ENSP00000263791	ENSG00000128829	EIF2AK4	E556G		1 MPLVEQSPEDSEGQDYVETVIPSNR		12 K		MPLVEQSPEDSGGQDYVETVIPSNR	12		ENSP00000263791	151.34
ENSP00000263867	ENSG00000042493	CAPG	H335R		1 MQYAPNTQVELPQGHESPIFK		16 R		MQYAPNTQVELPQGRESPIFK	16		ENSP00000263867	116.38
ENSP00000264848	ENSG00000114520	C3orf52	G144S		1 LTDVSTSTSPLGR		12 R		LTDVSTSTSPLSR	12		ENSP00000264848	59.19
ENSP00000267884	ENSG00000140319	SRP14	P124A		1 AAAAAAAAAPAAAATAPTTAATTAATAAQ		17 K		AAAAAAAAAPAAATAATTAATTAATAAQ	17		ENSP00000267884	95.83
ENSP00000268097	ENSG00000213614	HEXA	I436V		1 DFYVVEPLAYEQTPQK		4 K		DFYVVEPLAYECTPQK	4		ENSP00000268097	114.56
ENSP00000268150	ENSG00000140545	MFGE8	L76M		1 CVEPLGLMENGNAANSQAAASSVR		7 K		CVEPLGMENGNAANSQAAASSVR	7		ENSP00000268150	33.02
ENSP00000280333	ENSG00000150760	DOCK1	A1857T		1 QASVDSGIVQ		2 K		QTSVDSGIVQ	2		ENSP00000280333	55.57
ENSP00000285039	ENSG00000167105	MYO5B	K781N		1 GWLQKVK		5 R		GWLQNVK	5		ENSP00000285039	60.15
ENSP00000293217	ENSG00000161533	ACOX1	I312M		1 HQSEIKPGEPEPQILDFQTQQYK		5 R		HQSEMKPGEPEPQILDFQTQQYK	5		ENSP00000293217	39.65
ENSP00000294309	ENSG00000162741	TPCN2	G734E		1 DILEEPDEQEILTER		7 R		DILEEPEEQEILTER	7		ENSP00000294309	42.21
ENSP00000295971	ENSG00000163694	RBM47	M565V		1 NAAAAAAMYGGYAGYIPQAAFPAAAIQVPIPDVYQTY		8 K		NAAAAAAVYGGYAGYIPQAAFPAAAIQVPIPDVYQTY	8		ENSP00000295971	73.92
ENSP00000297135	ENSG00000164597	CDC5	F330L		1 ASFWTNMEK		3 R		ASLWTNMEK	3		ENSP00000297135	55.65
ENSP00000300648	ENSG00000089154	GCN1L1	Y2155D		1 IIETYLEATR		5 R		IIEDLLEATR	5		ENSP00000300648	75.69
ENSP00000301286	ENSG00000167676	PLIN4	V124A		1 EVVSSGVTGAMDMAK		2 K		EAVSSGVTGAMDMAK	2		ENSP00000301286	56.48
ENSP00000315013	ENSG00000095321	CHAT	L372M		1 SPLVPLPMPK		3 R		SPMVPLPMPK	3		ENSP00000315013	77.68
ENSP00000315775	ENSG00000103051	CDG4	T162I		1 SEDYEQAAAHTHR		11 R		SEDYEQAAAHIHR	11		ENSP00000315775	51.84
ENSP00000321746	ENSG00000163110	PDLIM5	A345T		1 DSRPSQAAGDNQGDEVK		21 R		SMPESLDSPTSGRPGVTSLTTAAAFKPVGSTGVIK	21		ENSP00000321746	44.96
ENSP00000321746	ENSG00000163110	PDLIM5	S492N		1 IGDEVSALK		7 K		IGDEVPALK	7		ENSP00000321746	79.97
ENSP00000323063	ENSG00000180573	EIF2S3L	S313P		1 SIFSK		5 K		LMCKPFSK	5		ENSP00000251039,ENSP00000323063	95.5
ENSP00000326603	ENSG00000101596	SMCHD1	V708I		1 LSVTWPEGDELLPNEVRPAGTPIGALR		16 R		LSVTWPEGDELLPNEIRPAGTPIGALR	16		ENSP00000326603	57.58
ENSP00000328062	ENSG00000136051	KIAA1033	V901I		1 LGVTPEGQSYLDQFR		3 K		LGITPEGQSYLDQFR	3		ENSP00000328062	57.8
ENSP00000330930	ENSG00000182319	PPAGMIN	R402Q		1 CLGLTGEFQFPAHPEATQPEPYIAESTK		15 R		CLGLTGEFQFPAHPCEATQPEPYIAESTK	15		ENSP00000330930	62.24
ENSP00000331283	ENSG00000185504	C17orf70	T817A		1 MQTNVTFTQAAQDSSAFQDLR		10 R		MQTNATFTQAAQDSSAFQDLR	10		ENSP00000331283	54.29
ENSP00000341597	ENSG00000055147	FAM114A2	G122S		1 AETSLGIPSPSEISTEVK		9 K		AETSLGIPSPSEISTEVK	9		ENSP00000341597	98.51
ENSP00000344818	ENSG00000150991	UBC	S190P		1 EGIPPDQQR		5 K		EGPPPDQQR	5		ENSP00000272317,ENSP00000344818	128.38
ENSP00000345445	ENSG00000100347	SAMM50	I345V		1 FYLGGPTSIR		9 R		FYLGGPTSVR	9		ENSP00000345445	72.01
ENSP00000346634	ENSG00000054118	THRAP3	A201V		1 DSRPSQAAGDNQGDEVK		16 K		DSRPSQAAGDNQGDEVK	16		ENSP00000346634	165.16
ENSP00000347046	ENSG00000138735	PDE5A	A93V		1 ADNSAPQTPTR		5 R		ADNSVPQTPTR	5		ENSP00000347046	33.84
ENSP00000352584	ENSG00000198074	EIF253L	S313P		1 SIFSK		5 K		LMCKPFSK	5		ENSP00000251039,ENSP00000323063	95.5
ENSP00000352995	ENSG00000104880	ARHGEF18	Q701R		1 QLGSANNGQAEDGGSSTGPPR		16 R		RLGSANNGQAEDGGSSTGPPR	16		ENSP00000352995	36.86
ENSP00000358262	ENSG00000113778	CHD1L	A88SS		1 SAVLHAQSSSSSSR		6 K		SAVLHSQSSSSSSR	6		ENSP00000358262	192.24
ENSP00000358832	ENSG00000147403	RPL10	N2025		1 YIPNRGPLDK		4 K		YIPSR	4	YIPSRGPLDK	ENSP00000358832	72.36
ENSP00000360828	ENSG00000124238	DDX27	G766S		1 AGPSFEER		2 R		AGPSFEER	2		ENSP00000360828	87.57
ENSP00000361883	ENSG00000131236	CAP1	S256A		1 SSLFAQNQGESITHALK		2 R		SALFAQINQGESITHALK	2		ENSP00000361883	143.02
ENSP00000366453	ENSG00000119139	TJP2	D482E		1 YQEPPAPQPK		4 R		YQEFPAPQPK	4		ENSP00000366453	90.05
ENSP00000366934	ENSG00000162408	NOL9	S58A		1 LLQAQASGVDWR		7 R		LLQAQAAGVDWR	7		ENSP00000366934	63.73
ENSP00000369814	ENSG00000198610	AKR1C4	Q250R		1 QTPALAILR		1 K		RTPALAILR	1		ENSP00000369927	53.68
ENSP00000371419	ENSG00000102699	PABPC4	M593T		1 HITSNTTAAEFIMSATPTMGNTDFWK		7 K		HITSNTTAAEFIMSATPTMGNTDFWK	7		ENSP00000371419	35.66
ENSP00000372221	ENSG00000206075	SERPINB5	I319V		1 GVALSNVVHK		8 K		GVALSNVVHK	8		ENSP00000372221	72.93
ENSP00000374447	ENSG00000118141	LRCH1	S234P		1 VLPQELVDLPLSLVK		10 K		VLPQELVDLPLSLVK	10		ENSP00000374447	70.62
ENSP00000375681	ENSG00000129451	KLK10	L149P		1 LARPVVLGPR		7 K		LARPVVPGPR	7		ENSP00000375681	73.99
ENSP00000375872	ENSG00000085978	ATG16L1	T300A		1 SVSSPPVPQDWVDTVPSGSK		14 R		SVSSPPVPQDWVDAHPSGSK	14		ENSP00000375872	232.59
ENSP00000376965	ENSG00000066427	ATXN1	V212M		1 VLEANDGSGMLDEDEEDLQR		1 R		MLEANDGSGMLDEDEEDLQR	1		ENSP00000376965	108.71
ENSP00000378438	ENSG00000167173	C15orf39	G491D		1 EGARPPSSPPMPVIDNVFSLAPYR		2 K		EDARPPSSPPMPVIDNVFSLAPYR	2		ENSP00000378438	45.01
ENSP00000378907	ENSG00000108599	AKAP10	I588V		1 MVSDVMQQAQYDQPLEK		6 K		MVSDVMQQAQYDQPLEX	6		ENSP00000378907	95.35
ENSP00000379133	ENSG00000149089	APIP	C76Y		1 IQPEDMFVCDINEK		9 R		IQPEDMFVYDINEK	9		ENSP00000379133	76.26
ENSP00000383365	ENSG00000184470	TNNI3K2	I370T		1 EATSVPHYAIGDVVEGRPELTPIAIMAGR		24 R		EATSVPHYAIGDVVEGRPELTPTAIMAGR	24		ENSP00000383365	146.57
ENSP00000385444	ENSG00000099889	ARVCF	R000Q		1 DVIPMDALGPDDYSTVDQR		18 R		DVIPMDALGPDDYSTVDQR	18		ENSP00000385444	87.85
ENSP00000395249	ENSG00000144567	FAM134A	P419Q		1 LSSPLHFVNTHFNGAGSPFDGVK		19 R		LSSPLHFVNTHFNGAGSPQDGVK	19		ENSP00000395249	180.42
ENSP00000395772	ENSG00000163939	GNL3	V367M		1 NSLEFFTVLAQR		8 R		NSLEFFTMLAQR	8		ENSP00000395772	76.52
ENSP00000396937	ENSG00000185627	PSMD13	N13S		1 DVPGFLQQSQSSGPGQPAVWHR		11 K		DVPGFLQQSQSSGPGQPAVWHR	11		ENSP00000396937	131.18
ENSP00000398131	ENSG00000103342	GSPT1	V100A		1 GPAAPPPVGQAAANNHGAGSGAGGR		9 R		GPAAPPPAGQAANNHGAGSGAGGR	9		ENSP00000398131	80.3
ENSP00000400410	ENSG00000204525	HLA-C	S48A		1 FISVGYVDDTQFVR		3 R		FIAVGYVDDTQFVR	3		ENSP00000366002	86.29
ENSP00000402060	ENSG00000073008	PVR	I340M		1 EGPPSEHSGSMSR		10 K		EGPPSEHSGMSR	10		ENSP00000402060	36.3
ENSP00000404217	ENSG00000105963	ADAP1	G175S		1 FHYLQVAFPGAGDAADLVPK		12 R		FHYLQVAFPGAGDALVPK	12		ENSP00000404217	49.05
ENSP00000404251	ENSG00000144560	VGLL4	I38M		1 IQTLPVASALSSHR		1 R		MQTLPVASALSSHR	1		ENSP00000404251	55.21
ENSP00000405879	ENSG00000201638	TMEM159	E178D		1 SAEFEGLYQI		3 K		SADFEGLYQI	3		ENSP00000405879	46
ENSP00000414051	ENSG00000123485	HJURP	R156G		1 ISRKSPGDPAKPASSPR		3 R		ISGK	3	ISGKSPGDPAKPASSPR	ENSP00000414051	117.38
ENSP00000427867	ENSG00000174093	RP11-1407O15.2	I238L		1 FAIEVAAK		3 K		FALEVAAK	3		ENSP00000320324	73.25
ENSP00000429829	ENSG00000153317	ASAP1	I142V		1 QEEIDESDDDLDDKPSPVK		18 R		QEEIDESDDDLDDKPSPVK	18		ENSP00000429829	112.58
ENSP00000432556	ENSG00000109861	CTSC	I110T		1 VGTASENVYVNAHLK		12 K		VGTASENVYVNTAHLK	12		ENSP00000432556	102.38
ENSP00000443772	ENSG00000256950	RP11-87C12.2	V17A		1 QSGGSSSQAGVVTVSDVQELMR		10 R		QSGGSSSQAGAVTVSDVQELMR	10		ENSP00000443772	56.06
ENSP00000445789	ENSG00000129347	KRI1	G135R		1 YVDEENSDDETSNHR		9 K		YVDEENSDRETSNHR	9		ENSP00000445789	153.86
ENSP00000253108	ENSG00000130811	EIF3G	F282L	ENSP00000253108	2 K		GLAYSFHR			ENSP00000253108	92.062		
ENSP00000353114	ENSG00000185567	AHNAK2	E2503A	ENSP00000353114	3 K		SIAASVDVSAPK			ENSP00000353114	28.806		
ENSP00000374370	ENSG00000070182	SPTB	S439N	ENSP00000349259	5 R		ETWLNENQR			ENSP00000349259	158.87		

Supplementary Table 2

Leading Razor Protein Short	Ensembl_Gene_Id	HGNC Symbol	Aminoacid Mutations	Mutation Position	Mutation Ratios	SIFT predictions	PolyPhen Scores	Tryptic pepts mut	Tryptic pepts mut alt	MUT MS Sequences	MUT MS Sequence Flags
ENSP00000269194	ENSG00000264886		T44A	44	1	tolerated	0.003	SNPATPASK		SFVSTLKSNPATPASK	Mutated
ENSP00000378438	ENSG00000167173	C15orf39	G491D	491	1	tolerated	0	EDARPPSSPPMPVIDNVFSLAPYR		EDARPPSSPPMPVIDNVFSLAPYR	Mutated
ENSP00000263791	ENSG00000128829	EIF2AK4	E556G	556	1	tolerated	0	MPLVEQSPEDSGGQDYVETVIPSNR		MPLVEQSPEDSGGQDYVETVIPSNR	Mutated
ENSP00000239165	ENSG00000120087	HOXB7	T9A	9	0.692308	tolerated	0	MSSLYYANALFSK		MSSLYYANALFSKYPASSSVFATGAFPEQTSCAFASNPQR,SSLYYANALFSK	Mutated
ENSP00000445789	ENSG00000129347	KRI1	G135R	135	1	tolerated	0	YVDEENSDRETSNHR		YVDEENSDRETSNHR	Mutated
ENSP00000366365	ENSG00000204695	OR14I1	M7T	7	0.611111	tolerated	0.002	MVNLTSTSGFLLMGFSDER		MVNLTSTSGFLLMGFSDER,MVNLTSTSGFLLMGFSDERK,VNLTSTSGFLLMGFSDER	Mutated
ENSP00000350447	ENSG00000163535	SGDL2	G9D	9	0.235294	tolerated	0.027	MECPVMETDSLFTSGIK		ECPVMETDSLFTSGIK,ECPVMETDSLFTSGIKR,MECPVMETDSLFTSGIK	Mutated
ENSP00000362285	ENSG00000198246	SLC29A3	R18G	18	0.736842	tolerated	0.003	MAVVSEDDFQHSSNSTYGTTSSSLR		AVVSEDDFQHSSNSTYGTTSSSLR	Mutated
ENSP00000368215	ENSG00000157625	TAB3	W394R	394	1	tolerated	0	SPSPISNQPSPRNQHSLYTATTPSSSPSR		SPSPISNQPSPR	Mutated

Supplementary Table 3

Ensembl_Prot_Id	Ensembl_Gene_Id	Gene Name	Aminoacid Mutation	Position in pept mut	Triptic aa start mut	Tryptic pept	Tryptic pept mut	Position in pept mut alt	Tryptic pept mut alt	MS Sequences	Razor Proteins	MS Modified Sequences Count	MS evidence Count	Razor Proteins Count	min Score	max Score
ENSP00000005226	ENSG00000006611	USH1C	E819D	11	K	IVTDYTLAEAEAALQK	IVTDYTLAEADAALQK	11		IVTDYTLAEADAALQK	ENSP00000005226	1	3	1	53.403	93.51
ENSP00000181839	ENSG00000065883	CDK13	T500A	3	K	GNTETSASASQTNHVK	GNAETSASASQTNHVK	3		GNAETSASASQTNHVK	ENSP00000181839	1	1	1	0	34.094
ENSP00000199764	ENSG00000086548	CEACAM6	G239V	14	R	SDPVTLNVLYGPDGPTISPSK	SDPVTLNVLYGPDVPTISPSK	14		SDPVTLNVLYGPDVPTISPSK	ENSP00000199764	1	3	1	0	164.98
ENSP00000215587	ENSG00000099817	POLR2E	S44F	3	K	AQSGDKPSEGRPR	AQFGDKPSEGRPR	3		AQFGDKPSEGRPR	ENSP00000215587	1	4	1	0	120.14
ENSP00000216129	ENSG00000100304	TTLL12	V297M	8	K	LPLDINPVVHPHGHIFK	LPLDINPMVHPHGHIFK	8		LPLDINPMVHPHGHIFK	ENSP00000216129	1	1	1	0	22.803
ENSP00000216181	ENSG00000100345	MYH9	I1626V	6	K	DLEAHIDSANK	DLEAHVDSANK	6		DLEAHVDSANK	ENSP00000216181	1	10	1	0	213.16
ENSP00000216780	ENSG00000100889	PCK2	Q121P	4	R	DTVQLPPGGAR	DTVPLPPGGAR	4		DTVPLPPGGAR	ENSP00000216780	1	1	1	30.872	30.872
ENSP00000220325	ENSG00000103966	EHD4	V154I	4	K	SISVIDSPGILSGEK	SISIIDSPGILSGEK	4		SISIIDSPGILSGEK	ENSP00000220325	1	4	1	56.482	110.77
ENSP00000221233	ENSG00000077348	EXOSC5	T5M	5		MEEETHTDAK	MEEEMHTDAK	5		MEEEMHTDAK	ENSP00000221233	1	2	1	70.977	84.31
ENSP00000221418	ENSG00000104823	ECH1	E41A	9	R	LTGSSAQEEASGVALGEAPDHSYESLR	LTGSSAQEAASGVALGEAPDHSYESLR	9		LTGSSAQEAASGVALGEAPDHSYESLR	ENSP00000221418	1	2	1	0	184.29
ENSP00000221957	ENSG00000105355	PLIN3	V275A	11	R	AQEALLQLSQVLSLMETVK	AQEALLQLSQALSLMETVK	11		AQEALLQLSQALSLMETVK	ENSP00000221957	1	3	1	0	102.37
ENSP00000225740	ENSG00000108602	ALDH3A1	P329A	15	R	YIAPTILTDVDPQSPVMQEEIFGPVLPIVCVR	YIAPTILTDVDPQSAVMQEEIFGPVLPIVCVR	15		YIAPTILTDVDPQSAVMQEEIFGPVLPIVCVR	ENSP00000225740	1	1	1	63.085	63.085
ENSP00000227868	ENSG00000110435	PDHX	V326L	1	K	VSVNDFIIK	LSVNDFIIK	1		LSVNDFIIK	ENSP00000227868	1	4	1	0	113.22
ENSP00000229854	ENSG00000112118	MCM3	E777K	3	R	DSEEPFSSVEIQAALSK	DSKEPFSSVEIQAALSK	3		DSKEPFSSVEIQAALSK	ENSP00000229854	1	4	1	0	119.23
ENSP00000232607	ENSG00000114491	UMPS	I446V	4	R	GSDIIIVGR	GSDVIIVGR	4		GSDVIIVGR	ENSP00000232607	1	2	1	0	55.549
ENSP00000236959	ENSG00000138363	ATIC	T116S	8	K	TVASPGVTVEEAVEQIDIGGVTLLR	TVASPGVSVEEAVEQIDIGGVTLLR	8		TVASPGVSVEEAVEQIDIGGVTLLR	ENSP00000236959	1	7	1	0	81.922
ENSP00000242108	ENSG00000122547	EEPD1	S343N	9	K	AVVAEKPSSQLQK	AVVAEKPSNQLQK	9		AVVAEKPSNQLQK	ENSP00000242108	1	3	1	0	117.79
ENSP00000252137	ENSG00000100056	DGCR14	A423V	8	R	ASYTPSPAR	ASYTPSPVR	8		ASYTPSPVR	ENSP00000252137	1	4	1	76.588	129.54
ENSP00000252999	ENSG00000130702	LAMA5	V889M	14	R	LELEEAATPEGHAVR	LELEEAATPEGHAMR	14		LELEEAATPEGHAMR	ENSP00000252999	1	2	1	0	48.704
ENSP00000254508	ENSG00000132182	NUP210	A755V	4	R	LTLAPVYTSPQLDMSCPLLQQNK	LTLVPVYTSPQLDMSCPLLQQNK	4		LTLVPVYTSPQLDMSCPLLQQNK	ENSP00000254508	1	5	1	0	89.028
ENSP00000254508	ENSG00000132182	NUP210	R786L	3	R	NPRLDLAAYDQEGR	NPLLDLAAYDQEGR	3		NPLLDLAAYDQEGR	ENSP00000254508	1	2	1	40.025	57.804
ENSP00000255194	ENSG00000132842	AP3B1	V585E	7	R	QLIVPNVK	QLIVPNEK	7		QLIVPNEK	ENSP00000255194	1	1	1	0	60.91
ENSP00000256707	ENSG00000134313	KIDINS220	Q1608H	19	R	SSESSPNHSLHNEVADDSQLEK	SSESSPNHSLHNEVADDSHLEK	19		SSESSPNHSLHNEVADDSHLEK	ENSP00000256707	4	6	1	10.399	61.112
ENSP00000257829	ENSG00000135372	NAT10	R705H	8	K	LNERPAERLDYLGVSYGLTPR	LNERPAEHLDYLGVSYGLTPR	8		LNERPAEHLDYLGVSYGLTPR	ENSP00000257829	1	1	1	0	49.089
ENSP00000258654	ENSG00000136152	COG3	N747S	2	K	VNDLAATAYK	VSDLAATAYK	2		VSDLAATAYK	ENSP00000258654	1	2	1	39.917	96.015
ENSP00000259008	ENSG00000136492	BRIP1	S919P	4	K	YSTSPYLLEAASHLSPENFVEDEAK	YSTPPYLLEAASHLSPENFVEDEAK	4		YSTPPYLLEAASHLSPENFVEDEAK	ENSP00000259008	1	3	1	28.443	33.544
ENSP00000259951	ENSG00000204642	HLA-F	P272S	8	K	WAAVVVPPGEEQR	WAAVVVPSGEEQR	8		WAAVVVPSGEEQR	ENSP00000366002	1	4	1	0	88.552
ENSP00000260197	ENSG00000137642	SORL1	A528T	11	K	TNVYISSSAGAR	TNVYISSSAGTR	11		TNVYISSSAGTR	ENSP00000260197	1	3	1	0	42.789
ENSP00000260227	ENSG00000137673	MMP7	R77H	13	K	FFGLPITGMLNSRVIEIMQKPR	FFGLPITGMLNSHVIEIMQKPR	13		FFGLPITGMLNSHVIEIMQKPR	ENSP00000260227	1	1	1	0	23.726
ENSP00000262265	ENSG00000104872	PIH1D1	P287L	10	K	QLMVAMPLLPVPS	QLMVAMPLLLVPS	10		QLMVAMPLLLVPS	ENSP00000262265	1	1	1	0	72.652
ENSP00000263026	ENSG00000103319	EEF2K	H23R	3	R	AGHDGDSDGDSDDEEGYFICPITDDPSSNQNVNSK	AGRDGDSDGDSDDEEGYFICPITDDPSSNQNVNSK	3		AGRDGDSDGDSDDEEGYFICPITDDPSSNQNVNSK	ENSP00000263026	1	3	1	18.837	43.728
ENSP00000263657	ENSG00000115946	PNO1	R11G	11		MESEMETQSARAEEGFTQVTR	MESEMETQSAGAEEGFTQVTR	11		MESEMETQSAGAEEGFTQVTR	ENSP00000263657	3	5	1	0	125.88
ENSP00000263791	ENSG00000128829	EIF2AK4	E556G	12	K	MPLVEQSPEDSEGQDYVETVIPSNR	MPLVEQSPEDSGGQDYVETVIPSNR	12		MPLVEQSPEDSGGQDYVETVIPSNR	ENSP00000263791	3	11	1	0	151.34
ENSP00000263867	ENSG00000042493	CAPG	H335R	16	R	MQYAPNTQVEILPQGHESPIFK	MQYAPNTQVEILPQGRESPIFK	16		MQYAPNTQVEILPQGRESPIFK	ENSP00000263867	2	4	1	38.746	116.38
ENSP00000264848	ENSG00000114529	C3orf52	G144S	12	R	LTDVYSTSPSLGR	LTDVYSTSPSLSR	12		LTDVYSTSPSLSR	ENSP00000264848	1	1	1	0	59.185
ENSP00000265562	ENSG00000076201	PTPN23	F161L	3	R	EHFPQAYSVDMSR	EHLPQAYSVDMSR	3		EHLPQAYSVDMSR	ENSP00000265562	1	1	1	0	24.964
ENSP00000267884	ENSG00000140319	SRP14	P124A	17	K	AAAAAAAAAPAAAATAPTTAATTAATAAQ	AAAAAAAAAPAAAATAATTAATTAATAAQ	17		AAAAAAAAAPAAAATAATTAATTAATAAQ	ENSP00000267884	1	11	1	0	95.834
ENSP00000268097	ENSG00000213614	HEXA	I436V	4	K	DFYIVEPLAFEGTPEQK	DFYVVEPLAFEGTPEQK	4		DFYVVEPLAFEGTPEQK	ENSP00000268097	1	3	1	0	114.56
ENSP00000268150	ENSG00000140545	MFGE8	L76M	7	K	CVEPLGLENGNIANSQIAASSVR	CVEPLGMENGNIANSQIAASSVR	7		CVEPLGMENGNIANSQIAASSVR	ENSP00000268150	1	1	1	0	33.024
ENSP00000276689	ENSG00000147684	NDUFB9	P146S	8	K	QLQEETPPGGPLTEALPPAR	QLQEETPSGGPLTEALPPAR	8		QLQEETPSGGPLTEALPPAR	ENSP00000276689	1	4	1	126.84	236.16
ENSP00000279907	ENSG00000111647	UHRF1BP1L	T283I	11	K	SMAPEPTQSSTVVASAQQVK	SMAPEPTQSSIVVASAQQVK	11		SMAPEPTQSSIVVASAQQVK	ENSP00000279907	1	1	1	0	24.594
ENSP00000280333	ENSG00000150760	DOCK1	A1857T	2	K	QASVDSGIVQ	QTSVDSGIVQ	2		QTSVDSGIVQ	ENSP00000280333	1	1	1	55.567	55.567
ENSP00000281828	ENSG00000116120	FARSB	V585I	14	K	FELTMPCSSLEINVGPFL	FELTMPCSSLEINIGPFL	14		FELTMPCSSLEINIGPFL	ENSP00000281828	1	1	1	0	26.365
ENSP00000285039	ENSG00000167306	MYO5B	K781N	5	R	GWLQKVK	GWLQNVK	5		GWLQNVK	ENSP00000285039	1	1	1	0	60.151
ENSP00000290541	ENSG00000159377	PSMB4	I234T	3	R	FQIATVTEK	FQTATVTEK	3		FQTATVTEK	ENSP00000290541	1	3	1	0	103.88
ENSP00000291634	ENSG00000160256	FAM207A	V212L	5	R	ASPLVAIGQTLAR	ASPLLAIGQTLAR	5		ASPLLAIGQTLAR	ENSP00000291634	1	1	1	26.124	26.124
ENSP00000293217	ENSG00000161533	ACOX1	I312M	5	R	HQSEIKPGEPEPQILDFQTQQYK	HQSEMKPGEPEPQILDFQTQQYK	5		HQSEMKPGEPEPQILDFQTQQYK	ENSP00000293217	1	1	1	0	39.647
ENSP00000293350	ENSG00000161618	ALDH16A1	T127S	10	R	LLWTLESLVTGR	LLWTLESLVSGR	10		LLWTLESLVSGR	ENSP00000293350	1	3	1	0	52.555
ENSP00000294309	ENSG00000162341	TPCN2	G734E	7	R	DILEEPGEDELTER	DILEEPEEDELTER	7		DILEEPEEDELTER	ENSP00000294309	1	1	1	42.209	42.209
ENSP00000295373	ENSG00000163214	DHX57	N587S	14	K	TTQIPQFILDDSLNGPPEK	TTQIPQFILDDSLSGPPEK	14		TTQIPQFILDDSLSGPPEK	ENSP00000295373	1	3	1	27.924	77.39
ENSP00000295971	ENSG00000163694	RBM47	M565V	8	K	NAAAAAAMYGGYAGYIPQAFPAAAIQVPIPDVYQTY	NAAAAAAVYGGYAGYIPQAFPAAAIQVPIPDVYQTY	8		NAAAAAAVYGGYAGYIPQAFPAAAIQVPIPDVYQTY	ENSP00000295971	1	3	1	0	73.919
ENSP00000297135	ENSG00000164597	COG5	F330L	3	R	ASFWTNMEK	ASLWTNMEK	3		ASLWTNMEK	ENSP00000297135	1	1	1	0	55.649
ENSP00000297185	ENSG00000113013	HSPA9	S639F	5	R	QAASSLQQASLK	QAASFLQQASLK	5		QAASFLQQASLK	ENSP00000297185	1	5	1	0	222.25
ENSP00000300107	ENSG00000166855	CLPX	I488T	3	R	DLIEFGMIPEFVGR	DLTEFGMIPEFVGR	3		DLTEFGMIPEFVGR	ENSP00000300107	1	1	1	0	68.83
ENSP00000300648	ENSG00000089154	GCN1L1	Y2155D	5	R	IIIEYLLEATR	IIIEDLLEATR	5		IIIEDLLEATR	ENSP00000300648	1	2	1	0	75.695
ENSP00000301286	ENSG00000167676	PLIN4	A668V	3	K	GAAQTGVDTAK	GAVQTGVDTAK	3		GAVQTGVDTAK	ENSP00000301286	1	1	1	61.344	61.344
ENSP00000301286	ENSG00000167676	PLIN4	M802T	5	K	GAVQMGVDTAK	GAVQTGVDTAK	5		GAVQTGVDTAK	ENSP00000301286	1	1	1	61.344	61.344
ENSP00000301286	ENSG00000167676	PLIN4	V124A	2	K	EVVSSGVTGAMDMAK	EAVSSGVTGAMDMAK	2		EAVSSGVTGAMDMAK	ENSP00000301286	1	2	1	0	56.482
ENSP00000303511	ENSG00000124587	PEX6	P939Q	10	R	VHDLEEGLEPGSSALMLTMEDLLQAAAR	VHDLEEGLEQGSSALMLTMEDLLQAAAR	10		VHDLEEGLEQGSSALMLTMEDLLQAAAR	ENSP00000303511	1	2	1	0	27.89
ENSP00000304327	ENSG00000172301	C17orf79	S43G	17	R	GAPPSPEAGFATADHSSQER	GAPPSPEAGFATADHSGQER	17		GAPPSPEAGFATADHSGQER	ENSP00000304327	1	3	1	0	98.121
ENSP00000305193	ENSG00000118960	HS1BP3	V260M	13	K	LSPQDPSEDVSSVDPLK	LSPQDPSEDVSSMDPLK	13		LSPQDPSEDVSSMDPLK	ENSP00000305193	1	1	1	27.496	27.496
ENSP00000305918	ENSG00000169925	BRD3	P149A	13	K	VAQMPQEEVELLPPAPK	VAQMPQEEVELLAPAPK	13		VAQMPQEEVELLAPAPK	ENSP00000305918	1	2	1	21.841	28.956
ENSP00000306105	ENSG00000121289	CEP89	R194W	32	R	NQVPLLHEVNSEDDENISHQDGFPGSPPAPQRTQQK	NQVPLLHEVNSEDDENISHQDGFPGSPPAPQWTQQK	32		NQVPLLHEVNSEDDENISHQDGFPGSPPAPQWTQQK	ENSP00000306105	1	2	1	21.275	24.869
ENSP00000307288	ENSG00000166508	MCM7	T415S	8	R	GSSGVGLTAAVLR	GSSGVGLSAAVLR	8		GSSGVGLSAAVLR	ENSP00000307288	1	1	1	0	82.069
ENSP00000308413	ENSG00000175634	RPS6KB2	A420V	1	R	APVSPLK	VPVSPLK	1		VPVSPLK	ENSP00000308413	1	1	1	0	90.629
ENSP00000310448	ENSG00000175467	SART1	G485A	18	R	VENMDISDEEEGGAPPPGSPQVLEEDEAELELQK	VENMDISDEEEGGAPPPASPQVLEEDEAELELQK	18		VENMDISDEEEGGAPPPASPQVLEEDEAELELQK	ENSP00000310448	1	8	1	28.068	187.96
ENSP00000311572	ENSG00000106853	PTGR1	A27S	2	K	TAELPPLK	TSELPPLK	2		TSELPPLK	ENSP00000311572	1	3	1	0	71.379
ENSP00000311648	ENSG00000174607	UGT8	I368M	15	K	AFLSHGGLNSIFETIYHGVPVVGIPLFGDHYDTMTR	AFLSHGGLNSIFETMYHGVPVVGIPLFGDHYDTMTR	15		AFLSHGGLNSIFETMYHGVPVVGIPLFGDHYDTMTR	ENSP00000311648	1	2	1	0	26.268
ENSP00000312143	ENSG00000136205	TNS3	I1132F	23	R	SLGSVSPSSSGFSSPHSGSTISIPFPNVLPDFSK	SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	23		SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	ENSP00000312143	10	16	1	0	120.47
ENSP00000315013	ENSG00000095321	CRAT	L372M	3	R	SPLVPLPMPK	SPMVPLPMPK	3		SPMVPLPMPK	ENSP00000315013	1	2	1	73.985	73.985
ENSP00000315775	ENSG00000103051	COG4	T162I	11	R	SEDYEQAAAHTHR	SEDYEQAAAHIHR	11		SEDYEQAAAHIHR	ENSP00000315775	1	1	1	0	51.841
ENSP00000321746	ENSG00000163110	PDLIM5	A345T	21	R	SMPESLDSPTSGRPGVTSLTAAAAFKPVGSTGVIK	SMPESLDSPTSGRPGVTSLTTAAAFKPVGSTGVIK	21		SMPESLDSPTSGRPGVTSLTTAAAFKPVGSTGVIK	ENSP00000321746	1	1	1	0	44.317
ENSP00000321746	ENSG00000163110	PDLIM5	S136F	9	K	APRPFGSVSSPK	APRPFGSVFSPK	9		APRPFGSVFSPK	ENSP00000321746	1	4	1	0	84.718
ENSP00000321746	ENSG00000163110	PDLIM5	S492N	7	K	ILGEVISALK	ILGEVINALK	7		ILGEVINALK	ENSP00000321746	1	3	1	65.347	79.974
ENSP00000323063	ENSG00000180574	EIF2S3L	S313P	5	K	SIFSK	LMCKPIFSK	5		LMCKPIFSK	ENSP00000253039	1	3	1	0	85.676
ENSP00000323687	ENSG00000175931	UBE2O	A1150T	9	R	IESWLETHALLEK	IESWLETHTLLEK	9		IESWLETHTLLEK	ENSP00000323687	1	1	1	27.409	27.409
ENSP00000323837	ENSG00000163798	SLC4A1AP	P139T	10	R	SLQEEQSRPPTAVSSPGGPAR	SLQEEQSRPTTAVSSPGGPAR	10		SLQEEQSRPTTAVSSPGGPAR	ENSP00000323837	3	4	1	0	121.56
ENSP00000323889	ENSG00000121060	TRIM25	P358L	4	R	LQEPTPSSGDPGEHDPASTHK	LQELTPSSGDPGEHDPASTHK	4		LQELTPSSGDPGEHDPASTHK	ENSP00000323889	1	2	1	0	48.824
ENSP00000326603	ENSG00000101596	SMCHD1	V708I	16	R	LSVTWPEGDELLPNEVRPAGTPIGALR	LSVTWPEGDELLPNEIRPAGTPIGALR	16		LSVTWPEGDELLPNEIRPAGTPIGALR	ENSP00000326603	1	2	1	0	57.578
ENSP00000328062	ENSG00000136051	KIAA1033	V901I	3	K	LGVTPEGQSYLDQFR	LGITPEGQSYLDQFR	3		LGITPEGQSYLDQFR	ENSP00000328062	1	1	1	0	57.804
ENSP00000328251	ENSG00000170854	MINA	A386T	17	K	DHIVLTVLPDQDQSDEAQEK	DHIVLTVLPDQDQSDETQEK	17		DHIVLTVLPDQDQSDETQEK	ENSP00000328251	1	2	1	0	80.638
ENSP00000330836	ENSG00000108010	GLRX3	P123S	9	R	HASSGSFLPSANEHLK	HASSGSFLSSANEHLK	9		HASSGSFLSSANEHLK	ENSP00000330836	2	5	1	0	113.35
ENSP00000330930	ENSG00000182319	PRAGMIN	R402Q	15	R	CLGLTGEPQPPAHPREATQPEPIYAESTK	CLGLTGEPQPPAHPQEATQPEPIYAESTK	15		CLGLTGEPQPPAHPQEATQPEPIYAESTK	ENSP00000330930	1	1	1	62.237	62.237
ENSP00000331983	ENSG00000079691	LRRC16A	G1202S	17	K	LGNDAVSQDSSSPALSGVER	LGNDAVSQDSSSPALSSVER	17		LGNDAVSQDSSSPALSSVER	ENSP00000331983	1	1	1	0	50.883
ENSP00000333283	ENSG00000185504	C17orf70	T817A	10	R	MQTMVTEQATQGSSAPDLR	MQTMVTEQAAQGSSAPDLR	10		MQTMVTEQAAQGSSAPDLR	ENSP00000333283	1	2	1	30.428	54.288
ENSP00000338093	ENSG00000185418	TARSL2	A96G	15	R	QATLESAELEAAQEAGAQPPPSQSQDK	QATLESAELEAAQEGGAQPPPSQSQDK	15		QATLESAELEAAQEGGAQPPPSQSQDK	ENSP00000338093	1	1	1	28.665	28.665
ENSP00000341597	ENSG00000055147	FAM114A2	G122S	9	K	AETSLGIPGPSEISTEVK	AETSLGIPSPSEISTEVK	9		AETSLGIPSPSEISTEVK	ENSP00000341597	1	2	1	96.666	98.507
ENSP00000344818	ENSG00000150991	UBC	S190P	5	K	EGIPSDQQR	EGIPPDQQR	5		EGIPPDQQR	ENSP00000272317	1	3	1	49.418	128.38
ENSP00000345445	ENSG00000100347	SAMM50	I345V	9	R	FYLGGPTSIR	FYLGGPTSVR	9		FYLGGPTSVR	ENSP00000345445	1	1	1	0	72.006
ENSP00000346634	ENSG00000054118	THRAP3	A201V	16	K	DSRPSQAAGDNQGDEAK	DSRPSQAAGDNQGDEVK	16		DSRPSQAAGDNQGDEVK	ENSP00000346634	1	4	1	0	165.16
ENSP00000347046	ENSG00000138735	PDE5A	A93V	5	R	ADNSAPGTPTR	ADNSVPGTPTR	5		ADNSVPGTPTR	ENSP00000347046	1	1	1	0	33.842
ENSP00000347719	ENSG00000141556	TBCD	M617T	12	R	LLSMTLSPDLHMR	LLSMTLSPDLHTR	12		LLSMTLSPDLHTR	ENSP00000347719	1	2	1	0	49.929
ENSP00000348828	ENSG00000114790	ARHGEF26	V29L	8	R	SIPQPHQVLGR	SIPQPHQLLGR	8		SIPQPHQLLGR	ENSP00000348828	1	1	1	0	24.343
ENSP00000350896	ENSG00000196411	EPHB4	T775M	12	R	FLEENSSDPTYTSSLGGK	FLEENSSDPTYMSSLGGK	12		FLEENSSDPTYMSSLGGK	ENSP00000350896	7	12	1	0	268.21
ENSP00000351740	ENSG00000197712	FAM114A1	L116P	11	R	SEIPLQEQNYLAVDSPPSGGGWAGWGSWGK	SEIPLQEQNYPAVDSPPSGGGWAGWGSWGK	11		SEIPLQEQNYPAVDSPPSGGGWAGWGSWGK	ENSP00000351740	1	1	1	0	30.352
ENSP00000351947	ENSG00000163104	SMARCAD1	S247N	15	R	LDHGEESNESAESSSNWEK	LDHGEESNESAESSNNWEK	15		LDHGEESNESAESSNNWEK	ENSP00000351947	3	7	1	0	119.97
ENSP00000352248	ENSG00000109511	ANXA10	M71L	6	R	DLIGDMR	DLIGDLR	6		DLIGDLR	ENSP00000352248	1	1	1	0	45.312
ENSP00000352447	ENSG00000182795	C1orf116	R94W	10	R	ALPITQPTPRGGPEETITQQGR	ALPITQPTPWGGPEETITQQGR	10		ALPITQPTPWGGPEETITQQGR	ENSP00000352447	1	3	1	27.61	95.195
ENSP00000352584	ENSG00000198074	AKR1B10	N313D	16	R	ACNVLQSSHLEDYPFNAEY	ACNVLQSSHLEDYPFDAEY	16		ACNVLQSSHLEDYPFDAEY	ENSP00000352584	1	2	1	0	90.714
ENSP00000352995	ENSG00000104880	ARHGEF18	Q701R	1	R	QLGSANGQAEDGGSSTGPPR	RLGSANGQAEDGGSSTGPPR	1		RLGSANGQAEDGGSSTGPPR	ENSP00000352995	1	1	1	0	36.856
ENSP00000353114	ENSG00000185567	AHNAK2	P2014S	9	R	SIEASVDVPAPK	SIEASVDVSAPK	9		SIEASVDVSAPK	ENSP00000353114	1	1	1	0	28.806
ENSP00000354045	ENSG00000125844	RRBP1	S766L	4	R	EHTSHLEAELEK	EHTLHLEAELEK	4		EHTLHLEAELEK	ENSP00000354045	1	3	1	0	84.718
ENSP00000354086	ENSG00000116221	MRPL37	C366S	16	R	VFHFLVFQLNTTDLDCNEGVK	VFHFLVFQLNTTDLDSNEGVK	16		VFHFLVFQLNTTDLDSNEGVK	ENSP00000354086	1	2	1	0	60.564
ENSP00000354511	ENSG00000093010	COMT	V158M	7	R	MVDFAGVK	MVDFAGMK	7		MVDFAGMK	ENSP00000354511	2	6	1	0	107.9
ENSP00000354532	ENSG00000198805	PNP	G51S	10	K	LTQAQIFDYGEIPNFPR	LTQAQIFDYSEIPNFPR	10		LTQAQIFDYSEIPNFPR	ENSP00000354532	1	4	1	0	84.874
ENSP00000355541	ENSG00000119285	HEATR1	N1694S	15	K	NFGAENPDPFVPVLNTAVK	NFGAENPDPFVPVLSTAVK	15		NFGAENPDPFVPVLSTAVK	ENSP00000355541	1	1	1	0	91.475
ENSP00000355583	ENSG00000143674	RP5-862P8.2	C784G	12	R	SASPPTSLPSTCGEASSPPSLPLSSALGILSTPSFSTK	SASPPTSLPSTGGEASSPPSLPLSSALGILSTPSFSTK	12		SASPPTSLPSTGGEASSPPSLPLSSALGILSTPSFSTK	ENSP00000355583	2	7	1	14.816	50.545
ENSP00000355672	ENSG00000181873	IBA57	G211S	14	R	LLTQDEGPALVPGGR	LLTQDEGPALVPGSR	14		LLTQDEGPALVPGSR	ENSP00000355672	1	4	1	0	56.482
ENSP00000357535	ENSG00000127884	ECHS1	T75I	1	K	TFEEDPAVGAIVLTGGDK	IFEEDPAVGAIVLTGGDK	1		IFEEDPAVGAIVLTGGDK	ENSP00000357535	1	6	1	0	186.23
ENSP00000358262	ENSG00000131778	CHD1L	A885S	6	K	SAVLHAQSSSSSSR	SAVLHSQSSSSSSR	6		SAVLHSQSSSSSSR	ENSP00000358262	3	8	1	0	192.24
ENSP00000359024	ENSG00000166197	NOLC1	S456P	4	K	AALSLPAK	AALPLPAK	4		AALPLPAK	ENSP00000359024	1	2	1	38.962	58.917
ENSP00000359295	ENSG00000117505	DR1	E171D	41	R	QQQAELAQQEWLQMQQAAQQAQLAAASASASNQAGSSQDEEDDDDI	QQQAELAQQEWLQMQQAAQQAQLAAASASASNQAGSSQDEDDDDDI	41		QQQAELAQQEWLQMQQAAQQAQLAAASASASNQAGSSQDEDDDDDI	ENSP00000359295	2	4	1	19.611	69.143
ENSP00000360540	ENSG00000138180	CEP55	T99A	5	R	YSTTTLLEQLEETTR	YSTTALLEQLEETTR	5		YSTTALLEQLEETTR	ENSP00000360540	1	1	1	22.803	22.803
ENSP00000360771	ENSG00000148396	SEC16A	T819A	17	K	ANHSSHQEDTYGALDFTLSR	ANHSSHQEDTYGALDFALSR	17		ANHSSHQEDTYGALDFALSR	ENSP00000360771	1	3	1	90.316	147.53
ENSP00000360771	ENSG00000148396	SEC16A	V884A	6	R	AQQELVPPQQQASPPQLPK	AQQELAPPQQQASPPQLPK	6		AQQELAPPQQQASPPQLPK	ENSP00000360771	2	12	1	0	170.5
ENSP00000360828	ENSG00000124228	DDX27	G766S	2	R	AGPSFEER	ASPSFEER	2		ASPSFEER	ENSP00000360828	1	6	1	45.879	87.568
ENSP00000361883	ENSG00000131236	CAP1	S256A	2	R	SSLFAQINQGESITHALK	SALFAQINQGESITHALK	2		SALFAQINQGESITHALK	ENSP00000361883	1	19	1	0	143.02
ENSP00000361892	ENSG00000101109	STK4	V312M	2	R	EVDQDDEENSEEDEMDSGTMVR	EMDQDDEENSEEDEMDSGTMVR	2		EMDQDDEENSEEDEMDSGTMVR	ENSP00000361892	4	15	1	0	148.87
ENSP00000364986	ENSG00000116771	AGMAT	R140Q	1	R	RIQEAYEK	QIQEAYEK	1		QIQEAYEK	ENSP00000364986	1	1	1	0	89.355
ENSP00000366002	ENSG00000206503	HLA-A	F33S	3	R	YFFTSVSRPGR	YFSTSVSRPGR	3		YFSTSVSRPGR	ENSP00000366002	1	1	1	0	73.067
ENSP00000366002	ENSG00000206503	HLA-A	N151K	6	K	DYIALNEDLR	DYIALKEDLR	6		DYIALKEDLR	ENSP00000366002	1	2	1	0	52.79
ENSP00000366453	ENSG00000119139	TJP2	D482E	4	R	YQEDPPAPQPK	YQEEPPAPQPK	4		YQEEPPAPQPK	ENSP00000366453	2	4	1	0	90.05
ENSP00000366453	ENSG00000119139	TJP2	M668I	4	R	AEQMASVQNAQR	AEQIASVQNAQR	4		AEQIASVQNAQR	ENSP00000366453	1	2	1	72.958	93.011
ENSP00000366620	ENSG00000049239	H6PD	R453Q	16	R	LFGSPLSDYYAYSPVRER	LFGSPLSDYYAYSPVQER	16		LFGSPLSDYYAYSPVQER	ENSP00000366620	1	1	1	0	47.568
ENSP00000366927	ENSG00000137124	ALDH1B1	R107L	4	R	LLNRLADLVER	LLNLLADLVER	4		LLNLLADLVER	ENSP00000366927	1	4	1	0	57.836
ENSP00000366934	ENSG00000162408	NOL9	S58A	7	R	LLQAQASGVDWR	LLQAQAAGVDWR	7		LLQAQAAGVDWR	ENSP00000366934	1	1	1	63.727	63.727
ENSP00000367263	ENSG00000124942	AHNAK	M1572V	4	K	MPSMNIQTHK	MPSVNIQTHK	4		MPSVNIQTHK	ENSP00000367263	1	2	1	0	50.284
ENSP00000369814	ENSG00000198610	AKR1C4	Q250R	1	K	QTPALIALR	RTPALIALR	1		RTPALIALR	ENSP00000369927	1	3	1	36.424	53.683
ENSP00000371419	ENSG00000102699	PARP4	M936T	7	K	HITSNTMAAEFIMSATPTMGNTDFWK	HITSNTTAAEFIMSATPTMGNTDFWK	7		HITSNTTAAEFIMSATPTMGNTDFWK	ENSP00000371419	1	1	1	0	35.66
ENSP00000371419	ENSG00000102699	PARP4	S1459Y	10	R	GFGSYHPSASSPFHFQPSAASLTANLR	GFGSYHPSAYSPFHFQPSAASLTANLR	10		GFGSYHPSAYSPFHFQPSAASLTANLR	ENSP00000371419	1	1	1	22.533	22.533
ENSP00000372221	ENSG00000206075	SERPINB5	I319V	8	K	GVALSNVIHK	GVALSNVVHK	8		GVALSNVVHK	ENSP00000372221	1	4	1	0	63.283
ENSP00000374447	ENSG00000136141	LRCH1	S234P	10	K	VLPQELVDLSLVK	VLPQELVDLPLVK	10		VLPQELVDLPLVK	ENSP00000374447	1	2	1	0	70.622
ENSP00000375681	ENSG00000129451	KLK10	L149P	7	K	LARPVVLGPR	LARPVVPGPR	7		LARPVVPGPR	ENSP00000375681	1	3	1	0	73.985
ENSP00000375829	ENSG00000159166	LAD1	K323E	7	K	NLPSLAKQGASDPPTVASR	NLPSLAEQGASDPPTVASR	7		NLPSLAEQGASDPPTVASR	ENSP00000375829	1	8	1	0	162.65
ENSP00000375872	ENSG00000085978	ATG16L1	T300A	14	R	SVSSFPVPQDNVDTHPGSGK	SVSSFPVPQDNVDAHPGSGK	14		SVSSFPVPQDNVDAHPGSGK	ENSP00000375872	2	9	1	0	232.59
ENSP00000376802	ENSG00000197249	SERPINA1	V237A	12	K	DTEEEDFHVDQVTTVK	DTEEEDFHVDQATTVK	12		DTEEEDFHVDQATTVK	ENSP00000376802	1	1	1	0	29.341
ENSP00000376965	ENSG00000066427	ATXN3	V212M	1	R	VLEANDGSGMLDEDEEDLQR	MLEANDGSGMLDEDEEDLQR	1		MLEANDGSGMLDEDEEDLQR	ENSP00000376965	2	2	1	0	108.71
ENSP00000377845	ENSG00000070669	ASNS	V210E	2	R	DVPLHALYDNVEK	DEPLHALYDNVEK	2		DEPLHALYDNVEK	ENSP00000377845	1	5	1	0	96.591
ENSP00000378438	ENSG00000167173	C15orf39	G491D	2	K	EGARPPSSPPMPVIDNVFSLAPYR	EDARPPSSPPMPVIDNVFSLAPYR	2		EDARPPSSPPMPVIDNVFSLAPYR	ENSP00000378438	1	4	1	0	45.011
ENSP00000378907	ENSG00000108599	AKAP10	I588V	6	K	MIVSDIMQQAQYDQPLEK	MIVSDVMQQAQYDQPLEK	6		MIVSDVMQQAQYDQPLEK	ENSP00000378907	1	1	1	0	95.347
ENSP00000379133	ENSG00000149089	APIP	C76Y	9	R	IQPEDMFVCDINEK	IQPEDMFVYDINEK	9		IQPEDMFVYDINEK	ENSP00000379133	1	3	1	0	76.262
ENSP00000379392	ENSG00000066654	THUMPD1	E311D	7	K	NNQQVPENTEELGQTKPTSNPQVVNEGGAKPELASQATEGSK	NNQQVPDNTEELGQTKPTSNPQVVNEGGAKPELASQATEGSK	7		NNQQVPDNTEELGQTKPTSNPQVVNEGGAKPELASQATEGSK	ENSP00000379392	1	8	1	50.645	146.78
ENSP00000380466	ENSG00000214026	MRPL23	R11Q	8	R	NVVYPLYRLGGPQLR	NVVYPLYQLGGPQLR	8		NVVYPLYQLGGPQLR	ENSP00000380466	1	3	1	0	59.185
ENSP00000381607	ENSG00000084207	GSTP1	I105V	2	K	YISLIYTNYEAGK	YVSLIYTNYEAGK	2		YVSLIYTNYEAGK	ENSP00000381607	1	2	1	37.998	47.621
ENSP00000382241	ENSG00000060749	QSER1	N1018S	6	R	SISGENATSESEFTLGGDDSGVSMNPAR	SISGESATSESEFTLGGDDSGVSMNPAR	6		SISGESATSESEFTLGGDDSGVSMNPAR	ENSP00000382241	1	1	1	0	22.096
ENSP00000383365	ENSG00000184470	TXNRD2	A66S	2	K	VAVVDYVEPSPQGTR	VSVVDYVEPSPQGTR	2		VSVVDYVEPSPQGTR	ENSP00000383365	1	5	1	0	71.692
ENSP00000383365	ENSG00000184470	TXNRD2	G384S	3	R	LFGGSSDLMDYDNVPTTVFTPLEYGCVGLSEEEAVAR	LFSGSSDLMDYDNVPTTVFTPLEYGCVGLSEEEAVAR	3		LFSGSSDLMDYDNVPTTVFTPLEYGCVGLSEEEAVAR	ENSP00000383365	1	1	1	0	36.467
ENSP00000383365	ENSG00000184470	TXNRD2	I370T	24	R	EATSVPHIYAIGDVVEGRPELTPIAIMAGR	EATSVPHIYAIGDVVEGRPELTPTAIMAGR	24		EATSVPHIYAIGDVVEGRPELTPTAIMAGR	ENSP00000383365	1	3	1	0	146.57
ENSP00000383851	ENSG00000162522	KIAA1522	M291V	19	R	VSLQALEAEAEAGAETEAMLQR	VSLQALEAEAEAGAETEAVLQR	19		VSLQALEAEAEAGAETEAVLQR	ENSP00000383851	1	2	1	0	74.748
ENSP00000385444	ENSG00000099889	ARVCF	R900Q	18	R	DVIPMDALGPDGYSTVDRR	DVIPMDALGPDGYSTVDQR	18		DVIPMDALGPDGYSTVDQR	ENSP00000385444	2	4	1	0	87.85
ENSP00000395249	ENSG00000144567	FAM134A	P419Q	19	R	LSSPLHFVNTHFNGAGSPPDGVK	LSSPLHFVNTHFNGAGSPQDGVK	19		LSSPLHFVNTHFNGAGSPQDGVK	ENSP00000395249	4	6	1	24.207	180.42
ENSP00000395772	ENSG00000163938	GNL3	V367M	8	R	NSLEFFTVLAQR	NSLEFFTMLAQR	8		NSLEFFTMLAQR	ENSP00000395772	1	3	1	0	76.522
ENSP00000396937	ENSG00000185627	PSMD13	N13S	11	K	DVPGFLQQSQNSGPGQPAVWHR	DVPGFLQQSQSSGPGQPAVWHR	11		DVPGFLQQSQSSGPGQPAVWHR	ENSP00000396937	1	3	1	32.797	131.18
ENSP00000397925	ENSG00000064270	ATP2C2	M466L	17	K	NAVMGQPTEGALMALAMK	NAVMGQPTEGALMALALK	17		NAVMGQPTEGALMALALK	ENSP00000397925	1	1	1	26.146	26.146
ENSP00000398131	ENSG00000103342	GSPT1	V100A	9	R	GPAAPPPPVGGAANNHGAGSGAGGR	GPAAPPPPAGGAANNHGAGSGAGGR	9		GPAAPPPPAGGAANNHGAGSGAGGR	ENSP00000398131	1	13	1	0	80.303
ENSP00000399168	ENSG00000234745	HLA-B	I218V	8	K	THVTHHPISDHEATLR	THVTHHPVSDHEATLR	8		THVTHHPVSDHEATLR	ENSP00000397867	1	5	1	0	160.81
ENSP00000400410	ENSG00000204525	HLA-C	S48A	3	R	FISVGYVDDTQFVR	FIAVGYVDDTQFVR	3		FIAVGYVDDTQFVR	ENSP00000366002	1	6	1	0	86.291
ENSP00000402060	ENSG00000073008	PVR	I340M	10	K	EGPPSEHSGISR	EGPPSEHSGMSR	10		EGPPSEHSGMSR	ENSP00000402060	1	2	1	0	36.303
ENSP00000402935	ENSG00000131374	TBC1D5	I718V	12	R	GQGQSVQMSGAIK	GQGQSVQMSGAVK	12		GQGQSVQMSGAVK	ENSP00000402935	1	1	1	79.633	79.633
ENSP00000404217	ENSG00000105963	ADAP1	G175S	12	R	FHYLQVAFPGAGDADLVPK	FHYLQVAFPGASDADLVPK	12		FHYLQVAFPGASDADLVPK	ENSP00000404217	1	4	1	0	49.049
ENSP00000404251	ENSG00000144560	VGLL4	I38M	1	R	IQTLPVASALSSHR	MQTLPVASALSSHR	1		MQTLPVASALSSHR	ENSP00000404251	1	1	1	55.213	55.213
ENSP00000409879	ENSG00000011638	TMEM159	E178D	3	K	SAEFEGLYQE	SADFEGLYQE	3		SADFEGLYQE	ENSP00000409879	1	1	1	0	46.001
ENSP00000415430	ENSG00000075218	GTSE1	W525R	13	R	SSGPAPQSLLSAWR	SSGPAPQSLLSARR	13		SSGPAPQSLLSARR	ENSP00000415430	1	3	1	0	19.819
ENSP00000417229	ENSG00000144895	EIF2A	T97S	11	K	NTVLATWQPYTTSK	NTVLATWQPYSTSK	11		NTVLATWQPYSTSK	ENSP00000417229	1	5	1	0	65.805
ENSP00000423463	ENSG00000248672	LY75-CD302	D807E	8	K	TPDWYNPDR	TPDWYNPER	8		TPDWYNPER	ENSP00000423463	1	2	1	42.001	53.683
ENSP00000425809	ENSG00000013375	PGM3	D494N	6	R	VISTTDAER	VISTTNAER	6		VISTTNAER	ENSP00000425809	1	2	1	53.756	55.461
ENSP00000427687	ENSG00000071127	WDR1	I185V	3	K	FTIGDHSR	FTVGDHSR	3		FTVGDHSR	ENSP00000427687	1	5	1	0	152.63
ENSP00000427867	ENSG00000174093	RP11-1407O15.2	I238L	3	K	FAIEVAAK	FALEVAAK	3		FALEVAAK	ENSP00000320324	1	2	1	51.995	73.248
ENSP00000429829	ENSG00000153317	ASAP1	I142V	18	R	QEEIDESDDDLDDKPSPIK	QEEIDESDDDLDDKPSPVK	18		QEEIDESDDDLDDKPSPVK	ENSP00000429829	2	6	1	0	112.58
ENSP00000431794	ENSG00000160710	ADAR	K379R	17	R	NTNSVPETAPAAIPETKR	NTNSVPETAPAAIPETRR	17		NTNSVPETAPAAIPETRR	ENSP00000431794	1	2	1	31.818	50.653
ENSP00000432556	ENSG00000109861	CTSC	I110T	12	K	VGTASENVYVNIAHLK	VGTASENVYVNTAHLK	12		VGTASENVYVNTAHLK	ENSP00000432556	1	3	1	0	102.38
ENSP00000432692	ENSG00000204922	C11orf83	G89S	11	K	NWMVGGEGGAGGR	NWMVGGEGGASGR	11		NWMVGGEGGASGR	ENSP00000432692	2	5	1	0	113.24
ENSP00000433999	ENSG00000023191	RNH1	R188H	15	K	ELTVSNNDINEAGVRVLCQGLK	ELTVSNNDINEAGVHVLCQGLK	15		ELTVSNNDINEAGVHVLCQGLK	ENSP00000433999	1	1	1	0	109.89
ENSP00000436337	ENSG00000227184	EPPK1	T1507A	8	R	QVVSAVTTLVEAAER	QVVSAVTALVEAAER	8		QVVSAVTALVEAAER	ENSP00000436337	1	1	1	27.208	27.208
ENSP00000439093	ENSG00000130779	CLIP1	S941P	6	K	MSGDNSSQLTK	MSGDNPSQLTK	6		MSGDNPSQLTK	ENSP00000439093	1	2	1	58.172	113.61
ENSP00000443772	ENSG00000256950	RP11-87C12.2	V17A	10	R	QSGGSSQAGVVTVSDVQELMR	QSGGSSQAGAVTVSDVQELMR	10		QSGGSSQAGAVTVSDVQELMR	ENSP00000443772	1	5	1	0	56.055
ENSP00000444408	ENSG00000105281	SLC1A5	V512L	10	K	SELPLDPLPVPTEEGNPLLK	SELPLDPLPLPTEEGNPLLK	10		SELPLDPLPLPTEEGNPLLK	ENSP00000444408	1	7	1	0	123.55
ENSP00000445789	ENSG00000129347	KRI1	G135R	9	K	YVDEENSDGETSNHR	YVDEENSDRETSNHR	9		YVDEENSDRETSNHR	ENSP00000445789	2	6	1	0	153.86
ENSP00000454639	ENSG00000172137	CALB2	A110T	7	R	QHVGSSAEFMEAWR	QHVGSSTEFMEAWR	7		QHVGSSTEFMEAWR	ENSP00000454639	1	3	1	0	67.275
ENSP00000459354	ENSG00000126602	TRAP1	D186E	7	R	GVVDSEDIPLNLSR	GVVDSEEIPLNLSR	7		GVVDSEEIPLNLSR	ENSP00000459354	1	5	1	0	151.54
ENSP00000459354	ENSG00000126602	TRAP1	R98G	3	K	DVREWQHEEFYR	DVGEWQHEEFYR	3		DVGEWQHEEFYR	ENSP00000459354	1	7	1	0	131.13
ENSP00000263635	ENSG00000115183	TANC1	N251S	3	K	DGNLR	DGSLR	3	DGSLRLGVQK	DGSLRLGVQK	ENSP00000263635	1	1	1	64.55	64.55
ENSP00000358832	ENSG00000147403	RPL10	N202S	4	K	YIPNR	YIPSR	4	YIPSRGPLDK	YIPSRGPLDK	ENSP00000358832	1	3	1	0	66.962
ENSP00000414051	ENSG00000123485	HJURP	R158G	3	R	ISRK	ISGK	3	ISGKSPGDPAKPASSPR	ISGKSPGDPAKPASSPR	ENSP00000414051	2	5	1	0	117.38
ENSP00000170168	ENSG00000079313	REXO1	S759P	9	R	TLAASGSQSSNGPEPGGQQLK	TLAASGSQPSNGPEPGGQQLK	9		TLAASGSQPSNGPEPGGQQLK	ENSP00000170168	1	1	1	30.882	30.882
ENSP00000216181	ENSG00000100345	MYH9	I1626V	6	K	DLEAHIDSANK	DLEAHVDSANK	6		DLEAHVDSANK	ENSP00000216181	1	7	1	0	186.46
ENSP00000220325	ENSG00000103966	EHD4	V154I	4	K	SISVIDSPGILSGEK	SISIIDSPGILSGEK	4		SISIIDSPGILSGEK	ENSP00000220325	1	3	1	65.092	110.77
ENSP00000221418	ENSG00000104823	ECH1	E41A	9	R	LTGSSAQEEASGVALGEAPDHSYESLR	LTGSSAQEAASGVALGEAPDHSYESLR	9		LTGSSAQEAASGVALGEAPDHSYESLR	ENSP00000221418	1	1	1	0	184.29
ENSP00000227868	ENSG00000110435	PDHX	V326L	1	K	VSVNDFIIK	LSVNDFIIK	1		LSVNDFIIK	ENSP00000227868	1	1	1	0	113.22
ENSP00000236959	ENSG00000138363	ATIC	T116S	8	K	TVASPGVTVEEAVEQIDIGGVTLLR	TVASPGVSVEEAVEQIDIGGVTLLR	8		TVASPGVSVEEAVEQIDIGGVTLLR	ENSP00000236959	1	5	1	0	81.922
ENSP00000252137	ENSG00000100056	DGCR14	A423V	8	R	ASYTPSPAR	ASYTPSPVR	8		ASYTPSPVR	ENSP00000252137	1	3	1	0	129.54
ENSP00000252999	ENSG00000130702	LAMA5	V889M	14	R	LELEEAATPEGHAVR	LELEEAATPEGHAMR	14		LELEEAATPEGHAMR	ENSP00000252999	1	1	1	0	48.704
ENSP00000254508	ENSG00000132182	NUP210	A755V	4	R	LTLAPVYTSPQLDMSCPLLQQNK	LTLVPVYTSPQLDMSCPLLQQNK	4		LTLVPVYTSPQLDMSCPLLQQNK	ENSP00000254508	1	4	1	0	89.028
ENSP00000256707	ENSG00000134313	KIDINS220	Q1608H	19	R	SSESSPNHSLHNEVADDSQLEK	SSESSPNHSLHNEVADDSHLEK	19		SSESSPNHSLHNEVADDSHLEK	ENSP00000256707	3	5	1	0	61.112
ENSP00000258654	ENSG00000136152	COG3	N747S	2	K	VNDLAATAYK	VSDLAATAYK	2		VSDLAATAYK	ENSP00000258654	1	2	1	0	104.52
ENSP00000259008	ENSG00000136492	BRIP1	S919P	4	K	YSTSPYLLEAASHLSPENFVEDEAK	YSTPPYLLEAASHLSPENFVEDEAK	4		YSTPPYLLEAASHLSPENFVEDEAK	ENSP00000259008	1	2	1	33.044	33.544
ENSP00000259951	ENSG00000204642	HLA-F	P272S	8	K	WAAVVVPPGEEQR	WAAVVVPSGEEQR	8		WAAVVVPSGEEQR	ENSP00000366002	1	3	1	0	88.552
ENSP00000262265	ENSG00000104872	PIH1D1	P287L	10	K	QLMVAMPLLPVPS	QLMVAMPLLLVPS	10		QLMVAMPLLLVPS	ENSP00000262265	1	1	1	72.652	72.652
ENSP00000263791	ENSG00000128829	EIF2AK4	E556G	12	K	MPLVEQSPEDSEGQDYVETVIPSNR	MPLVEQSPEDSGGQDYVETVIPSNR	12		MPLVEQSPEDSGGQDYVETVIPSNR	ENSP00000263791	3	12	1	0	151.34
ENSP00000267884	ENSG00000140319	SRP14	P124A	17	K	AAAAAAAAAPAAAATAPTTAATTAATAAQ	AAAAAAAAAPAAAATAATTAATTAATAAQ	17		AAAAAAAAAPAAAATAATTAATTAATAAQ	ENSP00000267884	1	9	1	0	47.331
ENSP00000268097	ENSG00000213614	HEXA	I436V	4	K	DFYIVEPLAFEGTPEQK	DFYVVEPLAFEGTPEQK	4		DFYVVEPLAFEGTPEQK	ENSP00000268097	1	2	1	0	114.56
ENSP00000276689	ENSG00000147684	NDUFB9	P146S	8	K	QLQEETPPGGPLTEALPPAR	QLQEETPSGGPLTEALPPAR	8		QLQEETPSGGPLTEALPPAR	ENSP00000276689	1	3	1	0	151.57
ENSP00000293217	ENSG00000161533	ACOX1	I312M	5	R	HQSEIKPGEPEPQILDFQTQQYK	HQSEMKPGEPEPQILDFQTQQYK	5		HQSEMKPGEPEPQILDFQTQQYK	ENSP00000293217	1	1	1	39.647	39.647
ENSP00000293350	ENSG00000161618	ALDH16A1	T127S	10	R	LLWTLESLVTGR	LLWTLESLVSGR	10		LLWTLESLVSGR	ENSP00000293350	1	2	1	45.915	52.555
ENSP00000295373	ENSG00000163214	DHX57	N587S	14	K	TTQIPQFILDDSLNGPPEK	TTQIPQFILDDSLSGPPEK	14		TTQIPQFILDDSLSGPPEK	ENSP00000295373	1	2	1	53.779	77.39
ENSP00000295971	ENSG00000163694	RBM47	M565V	8	K	NAAAAAAMYGGYAGYIPQAFPAAAIQVPIPDVYQTY	NAAAAAAVYGGYAGYIPQAFPAAAIQVPIPDVYQTY	8		NAAAAAAVYGGYAGYIPQAFPAAAIQVPIPDVYQTY	ENSP00000295971	1	2	1	0	73.919
ENSP00000297185	ENSG00000113013	HSPA9	S639F	5	R	QAASSLQQASLK	QAASFLQQASLK	5		QAASFLQQASLK	ENSP00000297185	1	4	1	0	222.25
ENSP00000300648	ENSG00000089154	GCN1L1	Y2155D	5	R	IIIEYLLEATR	IIIEDLLEATR	5		IIIEDLLEATR	ENSP00000300648	1	1	1	75.695	75.695
ENSP00000301286	ENSG00000167676	PLIN4	A668V	3	K	GAAQTGVDTAK	GAVQTGVDTAK	3		GAVQTGVDTAK	ENSP00000301286	1	1	1	0	61.344
ENSP00000301286	ENSG00000167676	PLIN4	M802T	5	K	GAVQMGVDTAK	GAVQTGVDTAK	5		GAVQTGVDTAK	ENSP00000301286	1	1	1	0	61.344
ENSP00000306105	ENSG00000121289	CEP89	R194W	32	R	NQVPLLHEVNSEDDENISHQDGFPGSPPAPQRTQQK	NQVPLLHEVNSEDDENISHQDGFPGSPPAPQWTQQK	32		NQVPLLHEVNSEDDENISHQDGFPGSPPAPQWTQQK	ENSP00000306105	1	1	1	0	24.869
ENSP00000311648	ENSG00000174607	UGT8	I368M	15	K	AFLSHGGLNSIFETIYHGVPVVGIPLFGDHYDTMTR	AFLSHGGLNSIFETMYHGVPVVGIPLFGDHYDTMTR	15		AFLSHGGLNSIFETMYHGVPVVGIPLFGDHYDTMTR	ENSP00000311648	1	1	1	0	26.268
ENSP00000312143	ENSG00000136205	TNS3	I1132F	23	R	SLGSVSPSSSGFSSPHSGSTISIPFPNVLPDFSK	SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	23		SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	ENSP00000312143	9	17	1	0	120.47
ENSP00000315013	ENSG00000095321	CRAT	L372M	3	R	SPLVPLPMPK	SPMVPLPMPK	3		SPMVPLPMPK	ENSP00000315013	1	3	1	0	77.677
ENSP00000321746	ENSG00000163110	PDLIM5	A345T	21	R	SMPESLDSPTSGRPGVTSLTAAAAFKPVGSTGVIK	SMPESLDSPTSGRPGVTSLTTAAAFKPVGSTGVIK	21		SMPESLDSPTSGRPGVTSLTTAAAFKPVGSTGVIK	ENSP00000321746	1	2	1	0	44.955
ENSP00000321746	ENSG00000163110	PDLIM5	S136F	9	K	APRPFGSVSSPK	APRPFGSVFSPK	9		APRPFGSVFSPK	ENSP00000321746	1	3	1	0	84.718
ENSP00000323063	ENSG00000180574	EIF2S3L	S313P	5	K	SIFSK	LMCKPIFSK	5		LMCKPIFSK	ENSP00000323063	1	3	1	0	95.502
ENSP00000326603	ENSG00000101596	SMCHD1	V708I	16	R	LSVTWPEGDELLPNEVRPAGTPIGALR	LSVTWPEGDELLPNEIRPAGTPIGALR	16		LSVTWPEGDELLPNEIRPAGTPIGALR	ENSP00000326603	1	1	1	0	57.578
ENSP00000330930	ENSG00000182319	PRAGMIN	R402Q	15	R	CLGLTGEPQPPAHPREATQPEPIYAESTK	CLGLTGEPQPPAHPQEATQPEPIYAESTK	15		CLGLTGEPQPPAHPQEATQPEPIYAESTK	ENSP00000330930	1	2	1	27.89	62.237
ENSP00000344818	ENSG00000150991	UBC	S190P	5	K	EGIPSDQQR	EGIPPDQQR	5		EGIPPDQQR	ENSP00000344818	1	2	1	121.64	128.38
ENSP00000346634	ENSG00000054118	THRAP3	A201V	16	K	DSRPSQAAGDNQGDEAK	DSRPSQAAGDNQGDEVK	16		DSRPSQAAGDNQGDEVK	ENSP00000346634	1	3	1	0	165.16
ENSP00000350896	ENSG00000196411	EPHB4	T775M	12	R	FLEENSSDPTYTSSLGGK	FLEENSSDPTYMSSLGGK	12		FLEENSSDPTYMSSLGGK	ENSP00000350896	7	11	1	0	268.21
ENSP00000351947	ENSG00000163104	SMARCAD1	S247N	15	R	LDHGEESNESAESSSNWEK	LDHGEESNESAESSNNWEK	15		LDHGEESNESAESSNNWEK	ENSP00000351947	3	8	1	0	119.97
ENSP00000352447	ENSG00000182795	C1orf116	R94W	10	R	ALPITQPTPRGGPEETITQQGR	ALPITQPTPWGGPEETITQQGR	10		ALPITQPTPWGGPEETITQQGR	ENSP00000352447	1	2	1	53.779	95.195
ENSP00000353114	ENSG00000185567	AHNAK2	D3793N	14	K	LDSAQLEGDLSLADK	LDSAQLEGDLSLANK	14		LDSAQLEGDLSLANK	ENSP00000353114	1	1	1	0	41.092
ENSP00000353114	ENSG00000185567	AHNAK2	P3336L	11	K	SIQASVDVSAPK	SIQASVDVSALK	11		SIQASVDVSALK	ENSP00000353114	1	1	1	0	42.976
ENSP00000354511	ENSG00000093010	COMT	V158M	7	R	MVDFAGVK	MVDFAGMK	7		MVDFAGMK	ENSP00000354511	[illegible]	[illegible]	1	0	107.9
ENSP00000354532	ENSG00000198805	PNP	G51S	10	K	LTQAQIFDYGEIPNFPR	LTQAQIFDYSEIPNFPR	10		LTQAQIFDYSEIPNFPR	ENSP00000354532	[illegible]	[illegible]	1	0	84.874
ENSP00000355672	ENSG00000181873	IBA57	G211S	14	R	LLTQDEGPALVPGGR	LLTQDEGPALVPGSR	14		LLTQDEGPALVPGSR	ENSP00000355672	1	1	1	0	56.482

Supplementary Table 3

ENSP00000358262	ENSG00000131778	CHD1L	A885S	6	K	SAVLHAQSSSSSSR	SAVLHSQSSSSSSR	6		SAVLHSQSSSSSSR	ENSP00000358262	3	9	1	0	192.24
ENSP00000360771	ENSG00000148396	SEC16A	T819A	17	K	ANHSSHQEDTYGALDFTLSR	ANHSSHQEDTYGALDFALSR	17		ANHSSHQEDTYGALDFALSR	ENSP00000360771	1	2	1	90.316	147.53
ENSP00000360771	ENSG00000148396	SEC16A	V884A	6	R	AQQELVPPQQQASPPQLPK	AQQELAPPQQQASPPQLPK	6		AQQELAPPQQQASPPQLPK	ENSP00000360771	2	13	1	0	170.5
ENSP00000360828	ENSG00000124228	DDX27	G766S	2	R	AGPSFEER	ASPSFEER	2		ASPSFEER	ENSP00000360828	1	2	1	81.017	87.568
ENSP00000361883	ENSG00000131236	CAP1	S256A	2	R	SSLFAQINQGESITHALK	SALFAQINQGESITHALK	2		SALFAQINQGESITHALK	ENSP00000361883	1	17	1	0	143.02
ENSP00000361892	ENSG00000101109	STK4	V312M	2	R	EVDQDDEENSEEDEMDSGTMVR	EMDQDDEENSEEDEMDSGTMVR	2		EMDQDDEENSEEDEMDSGTMVR	ENSP00000361892	3	12	1	0	148.87
ENSP00000366002	ENSG00000206503	HLA-A	F33S	3	R	YFFTSVSRPGR	YFSTSVSRPGR	3		YFSTSVSRPGR	ENSP00000366002	1	1	1	73.067	73.067
ENSP00000366365	ENSG00000204695	OR14J1	M7T	7		MVNLTSMSGFLLMGFSDER	MVNLTSTSGFLLMGFSDER	7		MVNLTSTSGFLLMGFSDER	ENSP00000366365	2	2	1	0	4.9784
ENSP00000366453	ENSG00000119139	TJP2	D482E	4	R	YQEDPPAPQPK	YQEEPPAPQPK	4		YQEEPPAPQPK	ENSP00000366453	1	3	1	0	90.05
ENSP00000367263	ENSG00000124942	AHNAK	M1572V	4	K	MPSMNIQTHK	MPSVNIQTHK	4		MPSVNIQTHK	ENSP00000367263	1	1	1	50.284	50.284
ENSP00000367727	ENSG00000157881	PANK4	A6G	6		MAECGASGSGSSGDSLDK	MAECGGSGSGSSGDSLDK	6		MAECGGSGSGSSGDSLDK	ENSP00000367727	1	1	1	10.838	10.838
ENSP00000372221	ENSG00000206075	SERPINB5	I319V	8	K	GVALSNVIHK	GVALSNVVHK	8		GVALSNVVHK	ENSP00000372221	1	4	1	0	72.928
ENSP00000376965	ENSG00000066427	ATXN3	V212M	1	R	VLEANDGSGMLDEDEEDLQR	MLEANDGSGMLDEDEEDLQR	1		MLEANDGSGMLDEDEEDLQR	ENSP00000376965	1	1	1	0	108.71
ENSP00000377845	ENSG00000070669	ASNS	V210E	2	R	DVPLHALYDNVEK	DEPLHALYDNVEK	2		DEPLHALYDNVEK	ENSP00000377845	1	3	1	0	96.591
ENSP00000378438	ENSG00000167173	C15orf39	G491D	2	K	EGARPPSSPPMPVIDNVFSLAPYR	EDARPPSSPPMPVIDNVFSLAPYR	2		EDARPPSSPPMPVIDNVFSLAPYR	ENSP00000378438	1	3	1	0	45.011
ENSP00000379133	ENSG00000149089	APIP	C76Y	9	R	IQPEDMFVCDINEK	IQPEDMFVYDINEK	9		IQPEDMFVYDINEK	ENSP00000379133	1	1	1	0	76.262
ENSP00000383365	ENSG00000184470	TXNRD2	A66S	2	K	VAVVDYVEPSPQGTR	VSVVDYVEPSPQGTR	2		VSVVDYVEPSPQGTR	ENSP00000383365	1	4	1	46.964	71.692
ENSP00000385444	ENSG00000099889	ARVCF	R900Q	18	R	DVIPMDALGPDGYSTVDRR	DVIPMDALGPDGYSTVDQR	18		DVIPMDALGPDGYSTVDQR	ENSP00000385444	1	1	1	0	87.85
ENSP00000395249	ENSG00000144567	FAM134A	P419Q	19	R	LSSPLHFVNTHFNGAGSPPDGVK	LSSPLHFVNTHFNGAGSPQDGVK	19		LSSPLHFVNTHFNGAGSPQDGVK	ENSP00000395249	3	5	1	0	180.42
ENSP00000395772	ENSG00000163938	GNL3	V367M	8	R	NSLEFFTVLAQR	NSLEFFTMLAQR	8		NSLEFFTMLAQR	ENSP00000395772	1	2	1	0	76.522
ENSP00000398131	ENSG00000103342	GSPT1	V100A	9	R	GPAAPPPPVGGAANNHGAGSGAGGR	GPAAPPPPAGGAANNHGAGSGAGGR	9		GPAAPPPPAGGAANNHGAGSGAGGR	ENSP00000398131	1	10	1	0	80.303
ENSP00000399168	ENSG00000234745	HLA-B	I218V	8	K	THVTHHPISDHEATLR	THVTHHPVSDHEATLR	8		THVTHHPVSDHEATLR	ENSP00000399168	1	5	1	0	160.81
ENSP00000400410	ENSG00000204525	HLA-C	S48A	3	R	FISVGYVDDTQFVR	FIAVGYVDDTQFVR	3		FIAVGYVDDTQFVR	ENSP00000366002	1	5	1	0	86.291
ENSP00000404217	ENSG00000105963	ADAP1	G175S	12	R	FHYLQVAFPGAGDADLVPK	FHYLQVAFPGASDADLVPK	12		FHYLQVAFPGASDADLVPK	ENSP00000404217	1	3	1	0	49.049
ENSP00000417229	ENSG00000144895	EIF2A	T97S	11	K	NTVLATWQPYTTSK	NTVLATWQPYSTSK	11		NTVLATWQPYSTSK	ENSP00000417229	1	3	1	0	65.805
ENSP00000417498	ENSG00000145113	MUC4	P4263S	4	K	METPGMTTPSLK	METSGMTTPSLK	4		METSGMTTPSLK	ENSP00000417498	1	1	1	14.561	14.561
ENSP00000423463	ENSG00000248672	LY75-CD302	D807E	8	K	TPDWYNPDR	TPDWYNPER	8		TPDWYNPER	ENSP00000423463	1	1	1	53.683	53.683
ENSP00000427687	ENSG00000071127	WDR1	I185V	3	K	FTIGDHSR	FTVGDHSR	3		FTVGDHSR	ENSP00000427687	1	1	1	0	152.63
ENSP00000431794	ENSG00000160710	ADAR	K379R	17	R	NTNSVPETAPAAIPETKR	NTNSVPETAPAAIPETRR	17		NTNSVPETAPAAIPETRR	ENSP00000431794	1	1	1	0	50.653
ENSP00000432556	ENSG00000109861	CTSC	I110T	12	K	VGTASENVYVNIAHLK	VGTASENVYVNTAHLK	12		VGTASENVYVNTAHLK	ENSP00000432556	1	1	1	0	102.38
ENSP00000432692	ENSG00000204922	C11orf83	G89S	11	K	NWMVGGEGGAGGR	NWMVGGEGGASGR	11		NWMVGGEGGASGR	ENSP00000432692	1	4	1	0	113.24
ENSP00000433999	ENSG00000023191	RNH1	R188H	15	K	ELTVSNNDINEAGVRVLCQGLK	ELTVSNNDINEAGVHVLCQGLK	15		ELTVSNNDINEAGVHVLCQGLK	ENSP00000433999	1	1	1	109.89	109.89
ENSP00000443772	ENSG00000256950	RP11-87C12.2	V17A	10	R	QSGGSSQAGVVTVSDVQELMR	QSGGSSQAGAVTVSDVQELMR	10		QSGGSSQAGAVTVSDVQELMR	ENSP00000443772	1	4	1	0	56.055
ENSP00000444408	ENSG00000105281	SLC1A5	V512L	10	K	SELPLDPLPVPTEEGNPLLK	SELPLDPLPLPTEEGNPLLK	10		SELPLDPLPLPTEEGNPLLK	ENSP00000444408	1	6	1	0	92.307
ENSP00000445789	ENSG00000129347	KRI1	G135R	9	K	YVDEENSDGETSNHR	YVDEENSDRETSNHR	9		YVDEENSDRETSNHR	ENSP00000445789	1	5	1	0	153.86
ENSP00000454639	ENSG00000172137	CALB2	A110T	7	R	QHVGSSAEFMEAWR	QHVGSSTEFMEAWR	7		QHVGSSTEFMEAWR	ENSP00000454639	1	2	1	0	67.275
ENSP00000459354	ENSG00000126602	TRAP1	D186E	7	R	GVVDSEDIPLNLSR	GVVDSEEIPLNLSR	7		GVVDSEEIPLNLSR	ENSP00000459354	1	5	1	49.31	151.54
ENSP00000459354	ENSG00000126602	TRAP1	R98G	3	K	DVREWQHEEFYR	DVGEWQHEEFYR	3		DVGEWQHEEFYR	ENSP00000459354	1	6	1	0	131.13
ENSP00000358832	ENSG00000147403	RPL10	N202S	4	K	YIPNR	YIPSR	4	YIPSRGPLDK	YIPSRGPLDK	ENSP00000358832	1	2	1	0	72.365
ENSP00000414051	ENSG00000123485	HJURP	R158G	3	R	ISRK	ISGK	3	ISGKSPGDPAKPASSPR	ISGKSPGDPAKPASSPR	ENSP00000414051	2	4	1	0	117.38

Supplementary Table 4

Mutated Peptide	Original Peptide	Ensembl_Prot_Id	Ensembl_Gene_Id	Gene Name	Aminoacid Mutation	Modified Sequences	Phospho site Positions	Leading Proteins
ADNSVPGTPTR	ADNSAPGTPTR	ENSP00000347046	ENSG00000138735	PDE5A	A93V	_ADNS(ph)VPGT(ph)PTRK_	92,96	ENSP00000347046:map.4/120523227(A:p.A93V:rs.C598T:c.gCc/gTc:SIFTprediction.tolerated:PolyPhenScore.0.001
ANHSSHQESDTYGALDFTLSR	ANHSSHQESDTYGALDFTLSR	ENSP00000360771	ENSG00000148396	SEC16A	T819A	_ANHSSHQESDTYGALDFALSR_	813	ENSP00000360771:map.9/133969079/C:p.T819A:n.A2489C:c.Acc/Gcc:SIFTprediction.tolerated:PolyPhenScore.0.793
ADQELAPPQQQASPPQLPK	ADQELVPPQQQASPPQLPK	ENSP00000360771	ENSG00000148396	SEC16A	V884A	_ADQELAPPQQQASp(ph)PPQLPK_	891	ENSP00000360771:map.9/133968883/G:p.V884A:n.T2685C:c.gTg/gCg:SIFTprediction.tolerated:PolyPhenScore.0.002
AGPSFEER	AGPSFEER	ENSP00000360828	ENSG00000134228	DDX27	G766S	_AGp(ph)PSFEER_	766	ENSP00000360828:map.20/47859217/A:p.G766S:n.G2305A:c.Ggc/Agc:SIFTprediction.deleterious:PolyPhenScore.0.667
ASYTPSPVR	ASYTPSPAR	ENSP00000252137	ENSG00000100056	DGCR14	A423V	_ASYT(ph)PS(ph)PVR_	421,419	ENSP00000252137:map.22/19121872/A:p.A423V:n.C1312T:c.gCa/gTa:SIFTprediction.tolerated:PolyPhenScore.0.002
DGSURLGVQK	DGNLR	ENSP00000263635	ENSG00000115183	TANC1	N2515	_DGS(ph)URLGVQK_	251	ENSP00000263635:map.2/160019863/G:p.N2515:n.A9895C:c.aAc/aCc:SIFTprediction.tolerated:PolyPhenScore.0
DSRPSQAAGDNQGDEVK	DSRPSQAAGDNQGDEAK	ENSP00000346634	ENSG00000054118	THRAP3	A201V	_SSSp(ph)KDSRPSQAAGDNQGDEVK_ _DSRPSQAAGDNQGDEVKEQTFS(ph)GGTSQDTK_	207,184	ENSP00000346634:map.17/67524317/T:p.A201V:n.C820T:c.gCc/gTc:SIFTprediction.tolerated:PolyPhenScore.0.066
DVIPMDALGPDGYSTVDQR	DVIPMDALGPDGYSTVDRR	ENSP00000385444	ENSG00000099889	ARVCF	R900Q	_SLEGEKT(ph)GS(ph)RDVIPMDALGPDGYSTVDQR_	881,879	ENSP00000385444:map.22/19959473/T:p.R900Q:n.G2699A:c.cGg/cAg:SIFTprediction.tolerated:PolyPhenScore.0.001
EMDQDDEENSEEDEMDSGTMVR	EVDQDDEENSEEDEMDSGTMVR	ENSP00000361892	ENSG00000101109	STK4	V312M	_EMDQDDEENS(ph)EEDEMDSGTMVR_	320	ENSP00000361892:map.20/43629135/A:p.V312M:n.G1029A:c.Gtg/Atg:SIFTprediction.tolerated:PolyPhenScore.0.003
FLEENSSDPTYNSSLGGK	FLEENSSDPTYSSLGGK	ENSP00000350896	ENSG00000196411	EPHB4	T775M	_FLEENSSDPT(ph)YMSSLGGK_ _FLEENS(ph)SDPTY(ph)MSSLGGK_	769,773,774	ENSP00000350896:map.7/100404997/A:p.T775M:n.C2793T:c.aCg/aTg:SIFTprediction.deleterious:PolyPhenScore.0.999
GAPPSPEAGFATADHSSQER	GAPPSPEAGFATADHSSQER	ENSP00000304327	ENSG00000172301	C17orf79	S43G	_GAPPS(ph)PEAGFATADHSGQERETEK_	31	ENSP00000304327:map.17/30183857/C:p.S43G:n.A265G:c.Agt/Ggt:SIFTprediction.tolerated:PolyPhenScore.0
GNAETSASASQTNHVK	GNTETSASASQTNHVK	ENSP00000181839	ENSG00000065883	CDK13	T500A	_ASNTST(ph)PTKGNAETSASASQTNHVK_	494	ENSP00000181839:map.7/40027484/G:p.T500A:n.A2103G:c.Acg/Gcg:SIFTprediction.tolerated:PolyPhenScore.0.073
ISGKSPKDPAKPASSPR	ISRK	ENSP00000414051	ENSG00000123485	HJURP	R158G	_ISGKS(ph)PKDPAKPASSPR_	160	ENSP00000414051:map.2/234750831/C:p.R158C:n.A472C:c.Aga/Iga:SIFTprediction.tolerated:PolyPhenScore.0
LDHGEESNESAESSNNWEK	LDHGEESNESAESSSNWEK	ENSP00000351947	ENSG00000163104	SMARCAD1	S247N	_TRLDHGEES(ph)NES(ph)AESSNNWEK_ _TRLDHGEES(ph)NES(ph)AES(ph)SNNWEK_	239,242,245	ENSP00000351947:map.4/95170839/A:p.S247N:n.G914A:c.aGt/aAt:SIFTprediction.tolerated:PolyPhenScore.0
LSPQDPSEDVSSMDPLK	LSPQDPSEDVSGVDPLK	ENSP00000305193	ENSG00000118960	HS1BP3	V260M	_KLSGs(ph)PQDPSEDVSSMDPLK_	249	ENSP00000305193:map.2/208244982/T:p.V260M:n.G804A:c.Gtg/Atg:SIFTprediction.tolerated:PolyPhenScore.0.003
LSSPLHFVNTHFNGAGSPQDGVK	LSSPLHFVNTHFNGAGSPPDGVK	ENSP00000395249	ENSG00000144567	FAM134A	P419Q	_LSS(ph)PLHFVNTHFNGAGSPQDGVK_	403	ENSP00000395249:map.2/220046975/A:p.P419Q:n.C1392A:c.cCa/cAa:SIFTprediction.tolerated:PolyPhenScore.0.002
MPLVEQSPEDSGGQDYVETVIPSNR	MPLVEQSPEDSEGQDYVETVIPSNR	ENSP00000263791	ENSG00000128829	EIF2AK4	E556G	_MPLVEQS(ph)PEDS(ph)GGQDYVETVIPSNR_	551,555	ENSP00000263791:map.15/40265799/G:p.E556G:n.A1710G:c.gAa/gGa:SIFTprediction.tolerated:PolyPhenScore.0
MQYAPNTQVELPQGREFSPYK	MQYAPNTQVELPQGHESPYK	ENSP00000263867	ENSG00000042493	CAPG	H335R	_MQYAPNTQVELPQGRES(ph)PYK_	337	ENSP00000263867:map.2/85622059/C:p.H335R:n.A1254G:c.cAt/cGt:SIFTprediction.tolerated:PolyPhenScore.0
QEEIDESDDDLDDKPSPVK	QEEIDESDDDLDDKPSPIK	ENSP00000429829	ENSG00000153317	ASAP1	I142V	_QEEIDES(ph)DDDLDDKPS(ph)PVK_	131	ENSP00000429829:map.8/131124559/C:p.I142V:n.A422G:c.Atc/Gtc:SIFTprediction.tolerated:PolyPhenScore.0.068
QSGGSSQAGAVTVSDVQELMR	QSGGSSQAGVVTVSDVQELMR	ENSP00000443772	ENSG00000256950	RP11-87C12.2	V17A	_(ac)S(ph)DEEARQSGGSSQAGAVTVSDVQELMR_	2	ENSP00000443772:map.12/122326812/C:p.V17A:n.T121C:c.gTc/gCc:SIFTprediction.tolerated:PolyPhenScore.0.001
QTSVDSGIVQ	QASVDSGIVQ	ENSP00000280333	ENSG00000150760	DOCK1	A1857T	_KQTS(ph)VDSGIVQ_	1858	ENSP00000280333:map.10/129249662/A:p.A1857T:n.G5678A:c.Gca/Aca:SIFTprediction.tolerated:PolyPhenScore.0
SAVLHSQSSSSSSR	SAVLHAQSSSSSSR	ENSP00000358262	ENSG00000131778	CHD1L	A885S	_SAVLHSQSS(ph)SSSSR_	888	ENSP00000358262:map.1/146767149/T:p.A885S:n.G2673T:c.Gca/Tca:SIFTprediction.tolerated:PolyPhenScore.0
SEUPLDPLPLPTEEGNPLLK	SELPLDPLPVPTEEGNPLLK	ENSP00000444408	ENSG00000105281	SLC1A5	V512L	_STEPELQDKS(ph)KELPLDPLPLPTEEGNPLLK_	503	ENSP00000444408:map.19/47729859/G:p.V512L:n.G2154C:c.Gtc/Ctc:SIFTprediction.tolerated:PolyPhenScore.0.001
SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	SLGSVSPSSSGFSSPHSGSTISFPFPNVLPDFSK	ENSP00000312143	ENSG00000136205	TN53	I1132F	_SLGSVS(ph)PSSSGFSS(ph)PHSGSTISFPFPNVLPDFSK_	1123	ENSP00000312143:map.7/47342611/A:p.I1132F:n.A3752T:c.Atc/Ttc:SIFTprediction.deleterious:PolyPhenScore.0.983
SSESSPNHSLHNEVADDSHLEK	SSESSPNHSLHNEVADDSQLEK	ENSP00000256707	ENSG00000134313	KIDINS220	Q1608H	_SSES(ph)S(ph)SPNHSLHNEVADDSHLEK_	15,931,594	ENSP00000256707:map.2/87134214/A:p.Q1608H:n.G5006T:c.caG/caT:SIFTprediction.tolerated:PolyPhenScore.0.003
SVSSFPVPQDNVDAHPGSGK	SVSSFPVPQDNVDTHPGSGK	ENSP00000375872	ENSG00000085978	ATG16L1	T300A	_S(ph)VSSFPVPQDNVDAHPGSGK_	287	ENSP00000375872:map.2/234183368/G:p.T300A:n.A1155G:c.Act/Gct:SIFTprediction.tolerated:PolyPhenScore.0
VENMDISDEEEGGAPPPASPQVLEEDEAELELQK	VENMDISDEEEGGAPPPGSPQVLEEDEAELELQK	ENSP00000310448	ENSG00000175467	SART1	G485A	_VENMDIS(ph)DEEEGGAPPPAS(ph)PQVLEEDEAELELQK_	474,486	ENSP00000310448:map.11/65735174/C:p.G485A:n.G1546C:c.gGg/gCg:SIFTprediction.tolerated:PolyPhenScore.0.132
VPVSPLK	APVSPLK	ENSP00000308413	ENSG00000175634	RPS6KB2	A420V	_VPVS(ph)PLK_	423	ENSP00000308413:map.11/67202156/T:p.A420V:n.C1304T:c.gCc/gTc:SIFTprediction.tolerated:PolyPhenScore.0.01
VSLQALEAEAEAGAETEAVLQR	VSLQALEAEAEAGAETTAMLQR	ENSP00000383851	ENSG00000162522	KIAA1522	M291V	_VSLp(ph)QALEAEAEAGAETEAVLQR_	274	ENSP00000383851:map.1/93235651/G:p.M291V:n.A941G:c.Agg/Gtg:SIFTprediction.tolerated:PolyPhenScore.0
YVDEENSDRETSNHR	YVDEENSDGETSNHR	ENSP00000445789	ENSG00000129347	KRI1	G135R	_YVDEENS(ph)DRETSNHR_	133	ENSP00000445789:map.19/10672493/T:p.G135R:n.G403A:c.Ggg/Agg:SIFTprediction.tolerated:PolyPhenScore.0

Part IV

Computational methods to predict network-attacking mutations

Chapter 6

Kinome-wide discovery of network-attacking mutations

Cancer cells acquire their oncogenic phenotype by perturbing signaling networks. To date, more than one hundred and fifty thousand distinct cancer somatic mutations have been identified by sequencing. However, functional interpretation of these mutations is still rare due to our limited knowledge of the sequence-function intersection on a systematic scale. We have developed an approach, ReKINect, that successfully predicts the functional impact of mutations on protein kinases and the oncogenic networks that result from them. ReKINect identifies network-attacking mutations that lead to constitutively active or constitutively inactive kinases as well as two different types of network rewiring.

6.1 Introduction

The decision-making processes of cancer cells differ significantly from those of healthy cells (Creixell et al., 2012; Pawson and Linding, 2008; Vogelstein and Kinzler, 2004; Hanahan and Weinberg, 2000). Since these processes are encoded in signaling networks, it should be expected that mutations accumulate in key components of these networks, such as protein kinases, to perturb them (Futreal et al., 2004; Greenman et al., 2007).

In an ever-increasing age of sequencing technologies, genomes of several cancers have been elucidated (Sjöblom et al., 2006; Wood et al., 2007; Greenman et al., 2007; Ding et al., 2008; Ley et al., 2008) and the number of different cancer somatic mutations is continuously increasing and currently in the order of hundreds of thousands (Forbes et al., 2011; Wong et al., 2011). While the field is starting to monitor how mutations accumulate during the evolution of cancer subclones (Nik-Zainal et al., 2012b,a; Ding

et al., 2012), very few of these mutations are deemed functional (the so-called drivers, (Stratton et al., 2009)) and for the vast majority, knowledge is lacking about how they may affect signaling networks and ultimately decision-processes. In other words, it is still largely unknown how many of these mutations are what we call network-attacking mutations (Creixell et al., 2012).
This missing piece of the puzzle may prove critical at, just to name some examples, monitoring the response to treatment (Ding et al., 2012), clustering patients that while presented with distinct mutations at the nucleotide level may harbour the same disease signaling networks (Cancer Genome Atlas Research Network, 2011) or inform combinatorial therapeutic strategies (Huang et al., 2007; Schoeberl et al., 2009; Lee et al., 2012).
While interpretation of how cancer mutations affect protein kinases has been sparse (Greenman et al., 2007) or limited to specific examples (Wan et al., 2004), evolutionary studies have provided insightful evidence of how perturbed kinase sequences affect protein function. Pseudokinases, for instance, are living examples of ancestral protein kinases that have lost their catalytic activity along evolution (Zeqiraj and van Aalten, 2010). Similar examples, now in cancer, exist of how other mutations may lead to constitutionally active kinases by hypothetically mimicking their active phosphorylated state (Davies et al., 2002) or affect peptide specificity (Songyang et al., 1995). We have integrated these different pieces of information in a single framework that allows knowledge-based predictions of network-attacking mutations.
Systematically predicting network-attacking mutations that cause changes in kinase activity and network rewiring represents one of the biggest challenges in our field. With the integrative approach that we present and validate in this article we contribute in closing this gap.

6.2 Results

6.2.1 Network-attacking mutations

As briefly introduced earlier, given that the aim is to predict which mutations attack signaling networks, we first define one schematic classification of the different strategies with which a mutations could affect these networks, i.e. the different types of network-attacking mutations (Fig. 6.1). In essence, mutations can perturb the nodes of signaling networks by: a) changing their dynamic activity, i.e. the conditions under which they become active by either keeping them constitutively active or inactive; b) affecting the structure of signaling networks by, for instance, changing the kinase that phosphorylates and activates a mutated kinase, or by this mutated kinase recognizing and phosphorylating new substrates; and, from the substrate perspective, c) generating new phosphorylation sites or destroying pre-existing phosphorylation sites. As further detailed in the figure legend, these different types of network perturbations also have clear relationship with more traditional classifications of mutations (i.e. hypermorphic or neomorphic gain-of-function and loss-of-function mutations).
Having defined conceptually these different types of network-attacking mutations, our aim can be now focused on identifying whether a given cancer mutation is likely

to belong to any of these categories.

6.2.2 The ReKINect approach

Our approach, ReKINect, combines information about catalytically essential residues (Zeqiraj and van Aalten, 2010), phosphorylation sites on protein kinases (Diella et al., 2004; Hornbeck et al., 2004) and kinase specificity (for further details please refer to Chapter 7) to predict which and how mutations affect signaling networks by perturbing protein kinases.
As shown in Figure 6.2, after exome or genome-wide sequencing of tumors, the algorithm will first map every mutation to determine the ones hitting kinase domains on protein kinases. Subsequently, cancer mutations hitting essential residues will be predicted to constitutively inactivate the mutant kinase. We apply strict knowledge-based rules before labeling a residue as catalytically essential as it should play a critical role in maintaining a kinase catalytically active (i.e. residues involved in ATP binding, Mg2+ coordination or phospho-transfer) (Zeqiraj and van Aalten, 2010) and mutations hitting these same residues will have lead to kinase inactivation (pseudokinases) throughout evolution. Next, ReKINect evaluates whether phosphorylation sites around the mutated residues may be affected by mutations. Comparing which kinase and with which probability is predicted to phosphorylate the wild type and mutant version of the phosphorylation site using NetworKIN (Linding et al., 2007) and NetPhorest (Miller et al., 2008), ReKINect can elucidate increased or decreased phosphorylation propensity or new kinases phosphorylating the site (i.e. upstream rewiring). Moreover, acidic substitutions in close proximity to activating phosphorylation events have been suggested to mimic the active phosphorylated state of the wild type kinase and therefore keep the kinase constitutively active (Davies et al., 2002). Thus, ReKINect also includes this possibility and predicts such mutations as causing constitutive activation of the mutant kinase. Finally, using our new purpose-made algorithm, KINspect (for further details please refer to Chapter 7), our approach assesses whether the mutated kinase is likely to have its peptide specificity affected, which could lead to the phosphorylation of new substrates (i.e. downstream rewiring). Finally, the destruction of phosphorylation sites by integrating data from known and annotated phosphorylation sites (Diella et al., 2004; Hornbeck et al., 2004) or possible generation of phosphorylation site is also reported.
Altogether, then, ReKINect is capable of predicting the different types of network-attacking mutations described in Figure 6.1.

6.2.3 Predictions on publicly-available cancer genome data and four cancer cell lines

In order to test ReKINect's predictive capabilities, we accessed publicly-available cancer genome data from COSMIC (Forbes et al., 2011) and evaluated how many of these mutations were predicted by ReKINect to be network-attacking mutations (as defined in Fig. 6.1). Moreover, we collected a panel of four cancer cell lines (3 ovarian

cancer cell lines -ES2, KOC7C and OVAS- and 1 colorectal cancer cell line -HT29-), on which we performed exome sequencing and global phospho-proteomic analysis, so that ReKINect predictions could be made and validated for the same samples. These predictions are shown in Figure 6.3.

6.2.4 Kinase inactivation predictions

As introduced in Chapter 1, every phosphorylation event is a rather precise and complex process where one kinase domain will 1) become active by being phosphorylated and changing its structural conformation, 2) bind two critical co-factors of the phosphorylation reaction, namely ATP and Mg^{2+}, and 3) recognize and phosphorylate a substrate by catalyzing the phospho-transfer reaction. As also mentioned above, in a process similar to the one which results in the appearance of pseudokinases in evolution (Zeqiraj and van Aalten, 2010), mutations that hit catalytically essential residues, such as those responsible for binding ATP and Mg^{2+}, will turn the affected kinase into a constitutively inactive state. In Figure 6.3 we show how ReKINect has uncovered that a large fraction of mutations that hit the kinase domain ($\sim$4.5%, much larger than expected by chance alone) hit these residues and would, therefore, affect the dynamic behavior of the signaling networks these kinases are embedded within, by constitutively shutting them down. Predictions of kinase inactivation represent a more precise definition of what would be more generally referred to as loss-of-function mutations.

As also evident from Figure 6.3, CamK1d E69V represents a typical example of this type of mutations, as it hits the catalytically essential glutamic acid (E69) that is critical for stabilizing ATP. Thus, this mutation is predicted to lead to kinase inactivation.

6.2.5 Kinase hyper-activation predictions

In contrast to inactivating mutations, other mutations will keep the kinase domain in a constitutive active conformation also affecting signaling networks' dynamic but this time by maintaining the signal "on" all the time.

Since kinase domains typically acquire their active conformation by being phosphorylated on their active segment (Johnson et al., 1996; Nolen et al., 2004), the most popular hypothesis of activating mutations in kinase domains are the so-called phospho-mimicking mutations. In this case, mutations to acidic residues (aspartic acid -D-, or glutamic acid -E-) on or near the activating phosphorylation site would, by adding a negative charge, "mimic" the effect of phosphorylation and "convince" the domain into the active conformation normally only achieved after phosphorylation. This hypothesis is largely based on findings made in BRAF (Davies et al., 2002; Wan et al., 2004), where the commonly found activating mutation V600E is acidic and adjacent to an activating phosphorylation site (Ser599).

With ReKINect we have systematically explored the validity of this hypothesis and identified in several cases the well-established acidic mutation BRAF V600E, in addition to another mutation that could work under the same mechanism, namely CHK2 K373E, as the mutation also leads to an acidic substitution and falls adjacent to a phosphorylation site in the activation segment of the kinase domain (Ser372) (Fig. 6.3).

Other predictions of acidic mutations could also fall within this category, but they either fall further away from known phosphorylation sites or are adjacent to tyrosine phosphorylation (e.g. KIT N822D) which, due to its larger size, one could argue it is less likely to be mimicked by acidic substitutions.

6.2.6 Upstream rewiring predictions

Mutations around phosphorylation sites may, in addition to constitutively activate the kinase domain, also promote a different type of network perturbation, namely upstream rewiring (Fig. 6.2). In this case, mutations to any residue (unlike activating mutations that should be restricted to acidic mutations) would change the sequence of the phosphorylation motif to such extent that it would stop being recognized by the same kinase and would be recognised by a new upstream kinase. Thus, this would lead to an activation of the mutated kinase under different conditions and by a different kinase than in the wild type scenario (Fig. 6.1).

ReKINect redirects mutations hitting phosphorylation motifs to NetworKIN (Linding et al., 2007) and NetPhorest (Miller et al., 2008) which will predict changes in phosphorylation propensity and the possibility that a new kinase phosphorylates the mutated sequence.

As shown in Figure 6.3, the best example of this type of network-attacking mutation is illustrated by EPHB4 T775M, since it hits next to a phosphorylation site in the activation segment of EPHB4 and it leads to a significant change in the predicted upstream kinase for the wild type (Src) and mutant (InsR) peptides.

6.2.7 Downstream rewiring predictions

While there is one example of downstream rewiring following kinase mutation reported in the literature (Songyang et al., 1995), we expect to find several others once the KINspect (Chapter 7) approach is completely finalized.

6.2.8 Destruction of phosphorylation sites

In a similar fashion as we did for activating mutation and upstream rewiring mutations, by integrating the mutational data with data about known phosphorylation

sites from PhosphositePlus (Hornbeck et al., 2004) and PhosphoELM (Diella et al., 2004), we could model the destruction of phosphorylation sites in cancer.

As illustrated in Figure 6.3, we identified 91 mutations hitting and destroying phosphorylation sites on kinase domains which, given the importance of phosphorylation for kinase regulation, in many cases are likely to affect its activity potential. MAP2K3 and its mutation T222M is one of the most clear examples found in our cell lines, as it affects an activating phosphorylation site in the activation segment of MAP2K3, which is known to be required for its activity (Raingeaud et al., 1996) (Fig. 6.3). Thus, while its mechanism is different from standard inactivating mutations that hit catalytically essential residues, by destroying an activating phosphorylation site, this mutation is in effect acting as an inactivating mutation (Fig. 6.1).

6.2.9 Generation of phosphorylation sites

While this category of network-attacking mutations is difficult to predict precisely, except if one considers all the mutations into phosphorylatable residues as potential candidates, after performing global phospho-proteomic analysis of all our cell lines followed by a genome-specific MS search as explained in Chapter 5, we had a unique opportunity to identify new true cancer-specific phosphorylation sites. As shown in Figure 6.4, six distinct high-confident new mutation-generated phosphorylation sites were found in two of the cancer cell lines analyzed, indeed demonstrating that this mechanism occurs in cancer. To the extend of our knowledge, this is the first time that the generation of new phosphorylation sites has been reported in cancer cells.

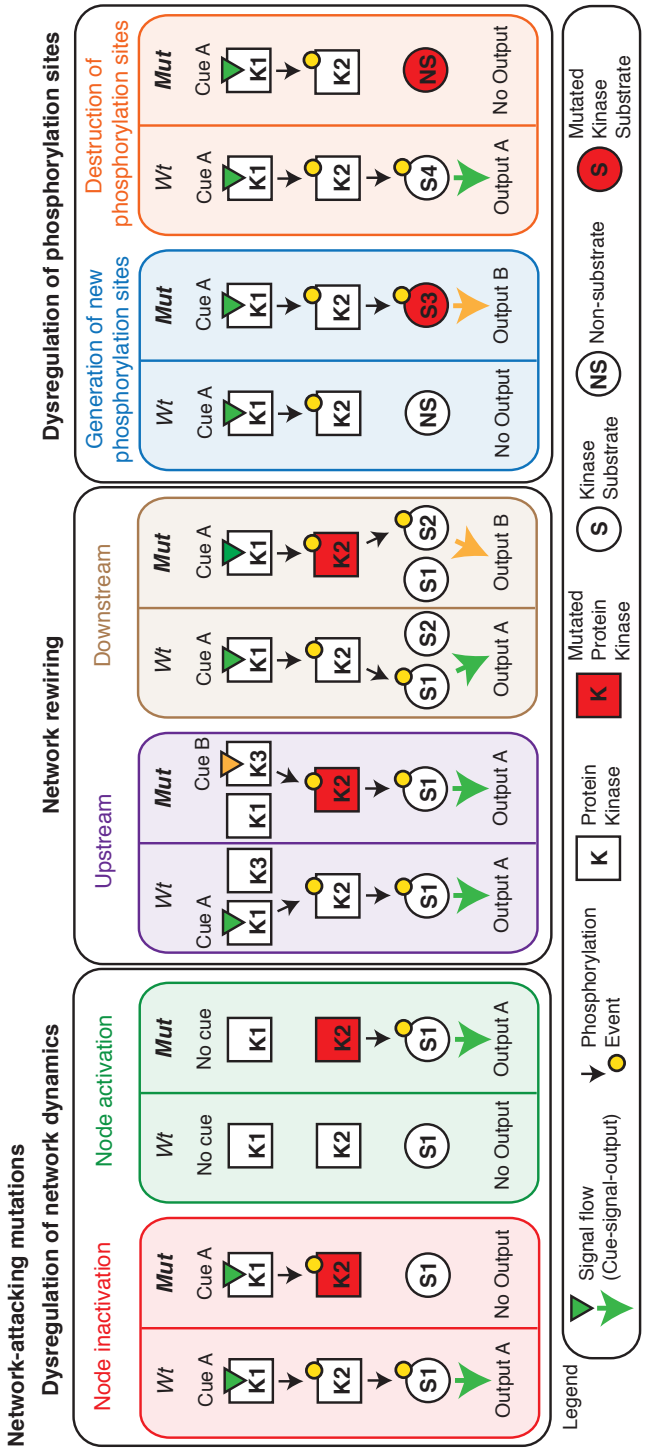


Figure 6.1. (Continued on the following page.)

Figure 6.1. Network-attacking mutations. Mutations can affect signaling networks by perturbing network dynamics (left), network structure (center) or by dysregulating phosphorylation sites (right), and within every category we have two complimentary perturbation possibilities. **Left.** Mutations can affect network dynamics by constitutively inactivating or constitutively activating a kinase. These type of kinase activation and inactivation would fall within the traditional classification of gain-of-function hypermorphic and loss-of-function mutations. **Center.** Mutations can affect network structure by causing upstream rewiring (where mutations in the phosphorylation motif of a kinase switches which upstream kinase phosphorylates the mutant kinase) or downstream rewiring (where kinase specificity is affected and leads to new substrates being phosphorylated). Both of these mutations would have traditionally been considered as neomorphic gain-of-function mutations. **Right.** Finally, mutations on phosphorylation site into non-phosphorylatable residues will effectively destruct that site and, reversely, mutations into phosphorylatable residues could constitute new phosphorylation sites. These mutations could be classified as loss-of-function and neomorphic gain-of-function mutations, respectively.

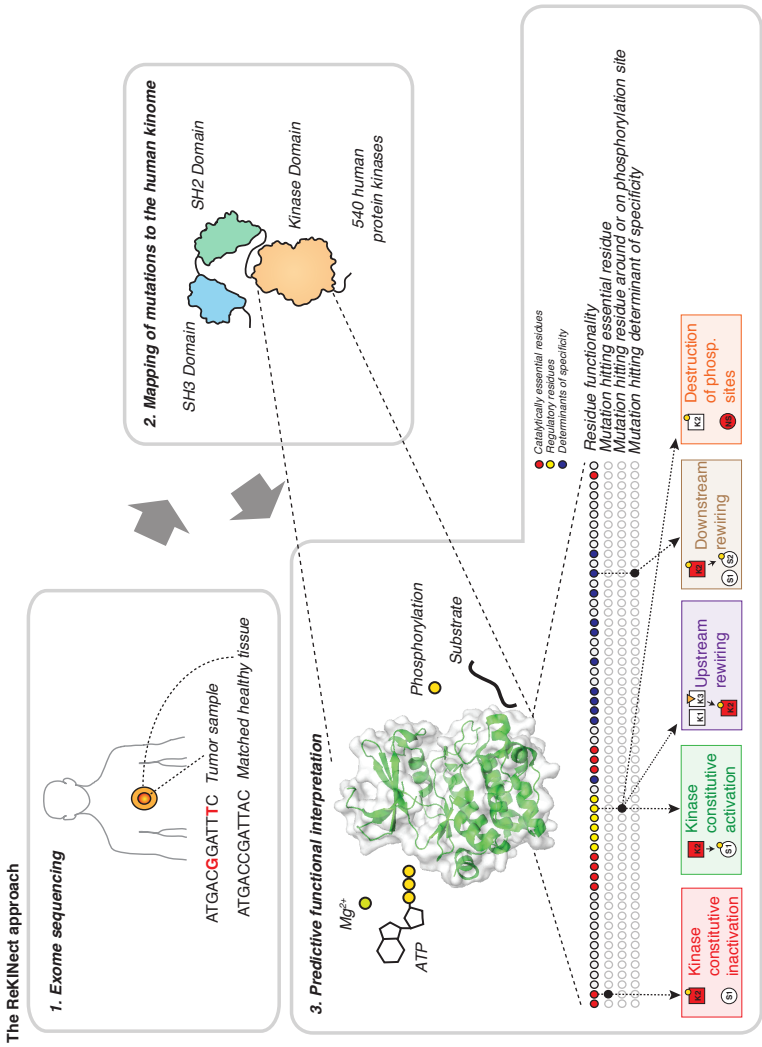


Figure 6.2. The ReKINect approach. Mutations coming from cancer sequencing efforts are first mapped to the human kinome and subsequently compared to known or predicted functional residues (essential residues, regulatory residues and determinants of specificity) in human kinases. As shown in the third panel of the figure, depending on the type of mutation and functional residue it hits, a different signaling network perturbation can be predicted.

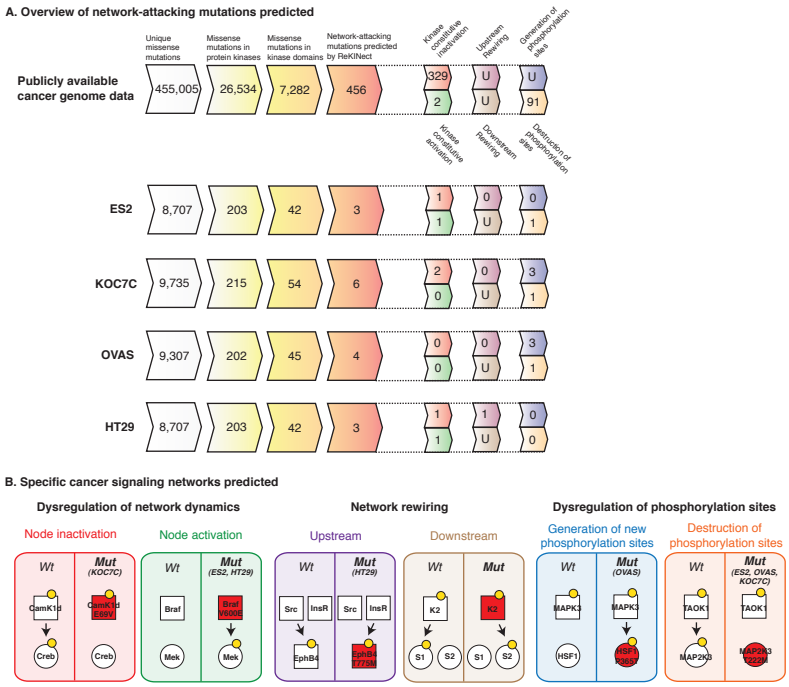


Figure 6.3. ReKINect predictions. A. After filtering for unique missense mutations that hit the kinase domain, the number of different mutations that could be predicted by ReKINect are shown for both publicly-available cancer genome data and cancer cell lines that were part of our panel. Some of the numbers are shown as unknown (U) due to either the lack of information at the point where this PhD needed to be submitted (e.g. KINspect predictions - downstream rewiring) or impossibility to determine the number with some certainty (e.g. how many mutations that become phosphorylatable residues do become phosphorylation sites). B. Specific examples of each type of network-attacking mutations are represented, with the wild type and mutant signaling networks shown side-by-side, with the specific mutant kinase and cell lines harboring the mutation labelled where appropriate. As in part A, due to the fact that KINspect is still being fine-tuned at the point of submission, no example could be given for downstream rewiring.

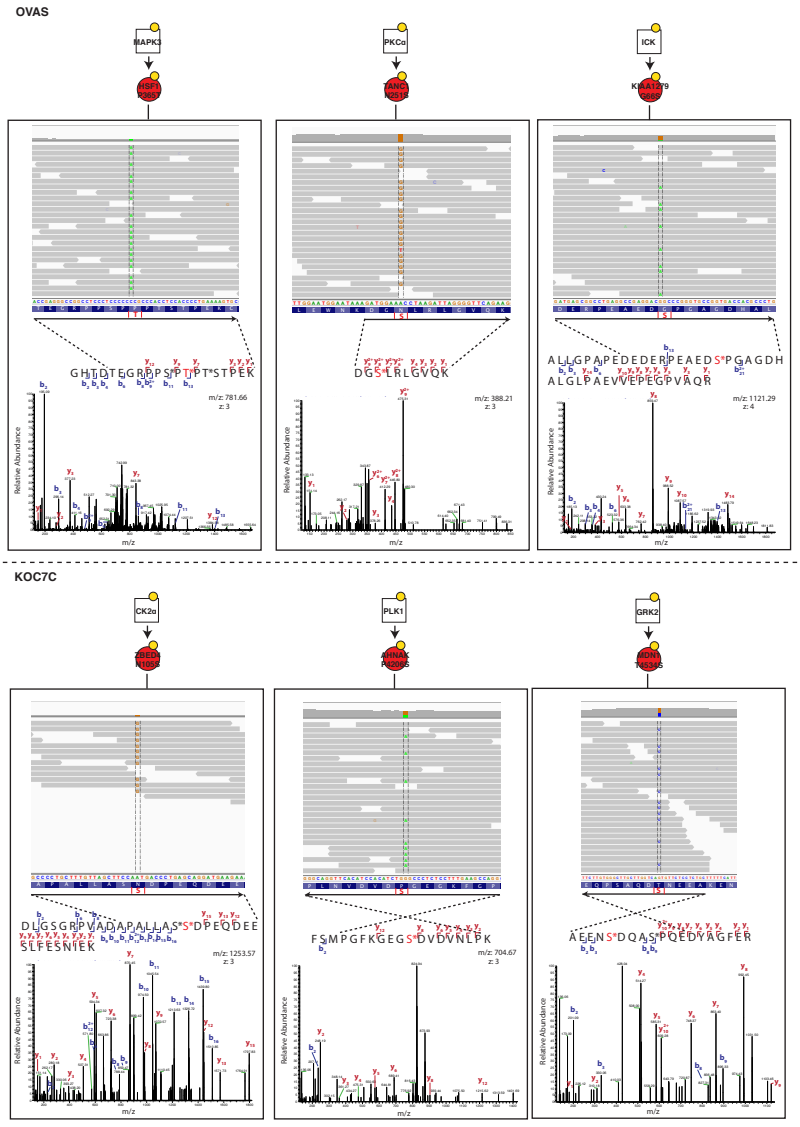


Figure 6.4. Generation of phosphorylation sites. For each of the two cell lines (OVAS at the top and KOC7C at the bottom), three cancer-specific phosphorylation sites were identified (three columns), where the new phosphorylatable residue identified by exome sequence (middle panel) is observed as being phosphorylated by Mass Spectrometry (bottom panel). As illustrated for each case (top part of each of the six columns), NetworKIN (Linding et al., 2007) predictions indicate the most likely kinase responsible for the phosphorylation event observed.

Chapter 7

Uncovering determinants of specificity in the kinase domain

As introduced in Chapter 1 and further detailed in the context of ReKINect in Chapter 6, one of the biggest challenges that we face when modeling how human protein kinases are perturbed by cancer mutations is the identification of residues that drive peptide specificity in the kinase domain, also known as determinants of specificity.

7.1 Introduction

Given this challenge, and after trying different less-successful approaches including mutual information or structural-based methods, here we describe a learning classifier system composed of a genetic algorithm engine combined with a sequence-to-specificity predictive framework, generated as a "free-style" extension to the more traditional structural-based pickpocket method. In this chapter, we demonstrate how this method is capable of finding a quantitative measure of the importance of each residue to specificity (which will naturally lead to the identification of determinants of specificity as those residues that score the highest) and prove the validity of our final specificity mask by showing it is enriched in known determinants of specificity and outperforms the predictive power of previous specificity prediction frameworks.

7.2 Results

7.2.1 The KINspect approach

As shown in Figure 7.1, the KINspect approach consist of different iterative steps that ensure an optimization of specificity masks (vector with as many positions as the

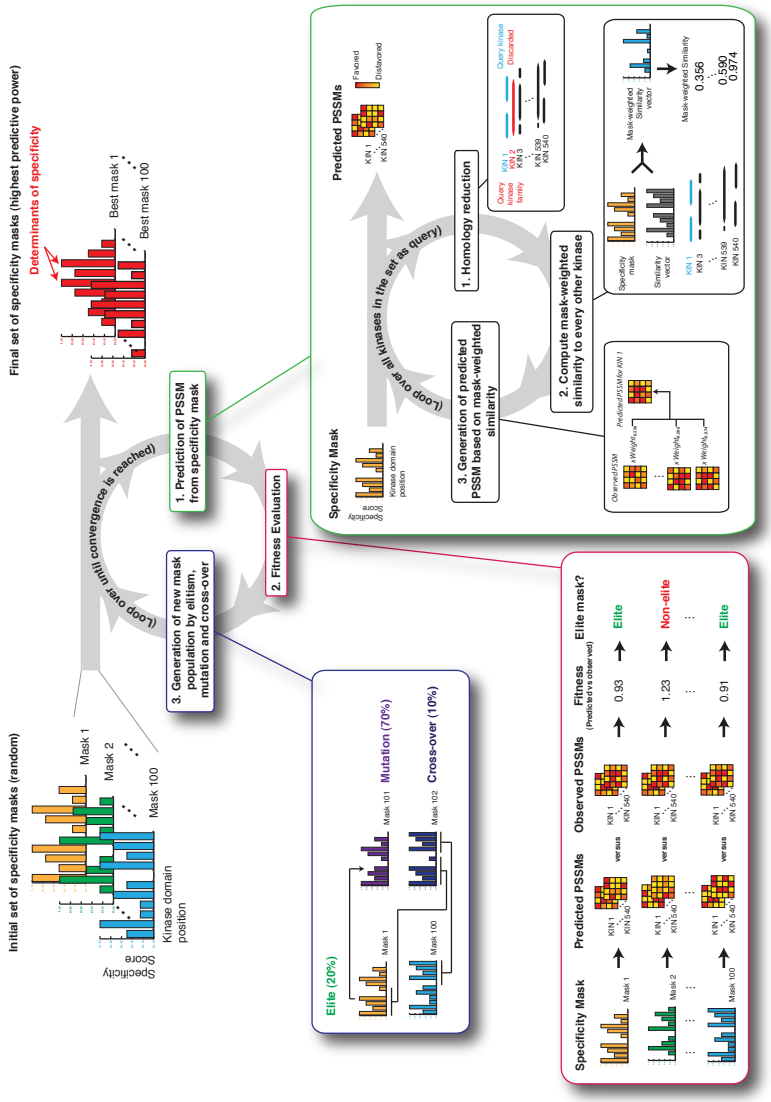


Figure 7.1. (Continued on the following page.)

kinase domain alignment used and values between 0.0 -residues with lowest importance for specificity- and 1.0 -residues with highest importance for specificity-) until those that present highest predictive power and better represent the contribution of each amino acid to the peptide specificity of the kinase domain are obtained.

The workflow is initialized with 100 randomized specificity masks, the predictive power of which will be tested by trying to predict the PSSM for every kinase with it

Figure 7.1. KINspect workflow. The KINspect workflow is designed to identify the specificity mask (vector with as many positions as the kinase domain alignment used and values between 0.0 -residues with lowest importance for specificity- and 1.0 -residues with highest importance for specificity-) that best describes the importance of the different residue for specificity. For the first round of KINspect, the 100 specificity masks are initialized with random values, and the predictive power of each masks is evaluated as follows. **In step 1**, for each specificity mask the system loops over all kinases as query and, using a kinase domain alignment, compare the query kinase to all other kinases (except those belonging to the same kinase family -homology reduction- to avoid over-fitting) at the sequence level, generating a similarity vector. This similarity vector is combined with the specificity mask, so that similarity in high-scoring positions of the mask are reinforced and similarity in low-scoring position of the mask are silenced, effectively producing a mask-weighted similarity vector and sum score for each kinase. These values are subsequently used to integrate the different observed PSSMs into a combined predicted PSSM for the query kinase, as further explained by equations 7.1, 7.2 & 7.3. **In step 2**, after a predicted kinase has been generated for all the kinases in our set, fitness is computed as the median of all the differences between the predicted and the observed PSSM for all the kinases. **In step 3**, the best-performing specificity masks are kept (elite) and new ones are generated by mutation (changing the value of a given position in the mask) and cross-over of the elite sequences (combining two segments of two other masks). Once a new set of masks has been generated, the optimization continues with prediction, fitness evaluation and generation of new masks, until the fitness can not be improved any further (convergence).

(step 1). In this first step, the system loops over all the different kinases, ignoring at every loop both the kinase selected as query and all the kinases belonging to the same kinase family (homology reduction) so that over-fitting is avoided. Next, the similarity between the query kinase and all the other kinases is computed and weighted by the specificity score that each residue has according to the selected specificity mask, following this equation:

$$Sim(KIN1, KIN2) = \sum_{x=1}^{N} \frac{S(KIN1_x, KIN2_x)}{\sqrt{S(KIN1_x, KIN1_x) \cdot S(KIN2_x, KIN2_x)}} SSx \tag{7.1}$$

where x is one position in the kinase domain alignment, $S(KIN1_x, KIN2_x)$ would be the similarity score between the residues of KIN1 and KIN2 in this position x, as determined by a substitution matrix of choice (e.g. BLOSUM (Henikoff and Henikoff, 1992)) and SSx is the specificity score of position x as determined by the specificity mask. By incorporating the specificity score, we achieve a reinforcement of residues deemed important for specificity and dilution of residues deemed as less being less important for specificity. This part of the method is largely inspired and, in essence, represents a generalization of the structure-based method Pickpocket (Zhang et al., 2009).

Subsequently, a final mask-weighted similarity is produced by the following equation:

$$W_{KIN1} = \frac{(Sim(KIN1, KIN2))^{\alpha}}{\sum_{KIN1=A}^{L}(Sim(KIN1, KIN2))^{\alpha}} \tag{7.2}$$

where L represents all the kinases that will be used to assess KIN2, i.e. all of them except those belonging to the same kinase family that KIN2 belongs to, and α represents the parameter that establishes the importance of scoring high similarity for the contribution towards PSSM prediction, with low values of α meaning a more democratic contribution of every kinase, regardless of its similarity, and higher values of α leading to predictions driven by the most similar kinase. The parameter space of α will typically need to be sampled thoroughly to determine the best value, as it is difficult to determine the best performing value until several have ben compared to one another.

Once the final mask-weighted similarity have been determined, the predicted PSSM is generated simply as described in this equation:

$$PSSMPred_{KIN2} = \sum_{KIN1=A}^{L} W_{KIN1} \cdot PSSMObs_{KIN1} \tag{7.3}$$

where $PSSMPred_{KIN2}$ would be the new predicted Position-Specific Scoring Matrix (PSSM) and $PSSMObs_{KIN1}$ represents the observed (i.e. experimentally determined) PSSM. Over 160 PSSMs experimentally determined that are part of the NetPhorest repository (Miller et al., 2008) are used in our training.

Upon prediction of PSSMs for all kinases using the 100 masks, the predictive power of each mask is assessed by comparing the predicted PSSMs to the observed PSSMs using the Frobenius distance (square root of the difference between every value in the two matrices squared) as a measure of performance and the median of all the Frobenius distances as a single fitness value for each mask (Fig. 7.1). Masks showing best performance (i.e. lowest Frobenius distance) will be kept (elite population) and new masks will be generated by mutation and cross-over of these elite masks as shown in Figure 7.1. This whole procedure of prediction, fitness assessment and generation of new mask population is repeated until the fitness can not be improved further, point at which the masks can be considered deeply optimized and containing true determinants of specificity (residues important for specificity scoring high in the mask).

7.2.2 Determinants of specificity identified by KINspect

Since, as described above, the KINspect approach is stochastic in nature and is initialized from random specificity masks, in order to assess the robustness of our results, we

ran the whole KINspect pipeline ten times and compared the results obtained for the different runs. As shown in Figure 7.2, the ten different runs followed a similar path towards convergence and the specificity masks obtained at the end of the ten runs were much more similar to one another than the best-performing masks obtained at the beginning of the process, indicating convergence of the ten runs towards the same solution. Given the high similarity between different runs, we combined the ten final specificity runs into a single consensus specificity mask and observed an enriched in a set of known determinants of specificity that we curated from the literature (Brinkworth et al., 2003; Mok et al., 2010) in this mask, which can be considered an independent benchmarking assessment that demonstrates the selection of true determinants of specificity in our mask. Importantly, in addition to most of the known determinants of specificity, KINspect identified many new residues that had not been identified as determinants until today and gives a quantitative approximation to the contribution of each residue to specificity.

Next, we projected the consensus specificity mask onto the three-dimensional structure of a kinase domain, so that the structural position and possible long-range allosteric interactions between the determinants of specificity identified by KINspect could be appreciated. As shown in Figure 7.3, this representation provides several pieces of interesting insight into how substrate specificity is encoded in the kinase domain. First, it is interesting to observe how the small N-lobe of the domain is largely depleted from determinants of specificity. While, it is to some extend known that most specificity comes from residues that reside in the large C-lobe of the domain, it is still surprising to see that none of the residues that are spatially close to the substrate and reside in the small N-lobe were selected as being important for specificity. Secondly, while some of the high-scoring residues (likely determinants of specificity) are positioned rather far from the active site and substrate, the structure shows a clear paths of coupled residues that connects these far-distant residues with the substrate, thus making it likely that allosteric interactions are involved in specificity. Finally, while some of the residues don't seem to be so well-connected to others or the substrate (especially at the back-end of the small N-lobe) in the conformation shown in Figure 7.3, given the high mobility that kinases show transitioning between "open" and "closed" conformations, we hypothesize that these residues could still form allosteric interactions in a different conformational state which might be important for substrate binding.

7.2.3 Predictive power of KINspect

In order to assess the predictive power of KINspect, as it is standard in machine learning, we used an independent set of PSSMs that was not part of our training set. In this case, we collected a set of PSSMs that were published as part of the DREAM 4 challenge (Ellis and Kobe, 2011). Using this independent set of PSSMs, we could compare our predictive power (assessed as the Frobenius distance between predicted and observed PSSM - lower distance meaning higher predictive power of the method) and concluded that KINspect outperforms previous methods (Brinkworth et al., 2003;

Ellis and Kobe, 2011) in its ability to predict substrate specificity from kinase domain sequence (Fig. 7.4).

7.3 Fine-tuning

Despite the encouraging results presented in this chapter, we are still finalizing the fine-tuning of KINspect, so that the best parameters, including normalization and the α parameter, are found before we integrate KINspect with ReKINect and move on to analyze all our cancer genome data.

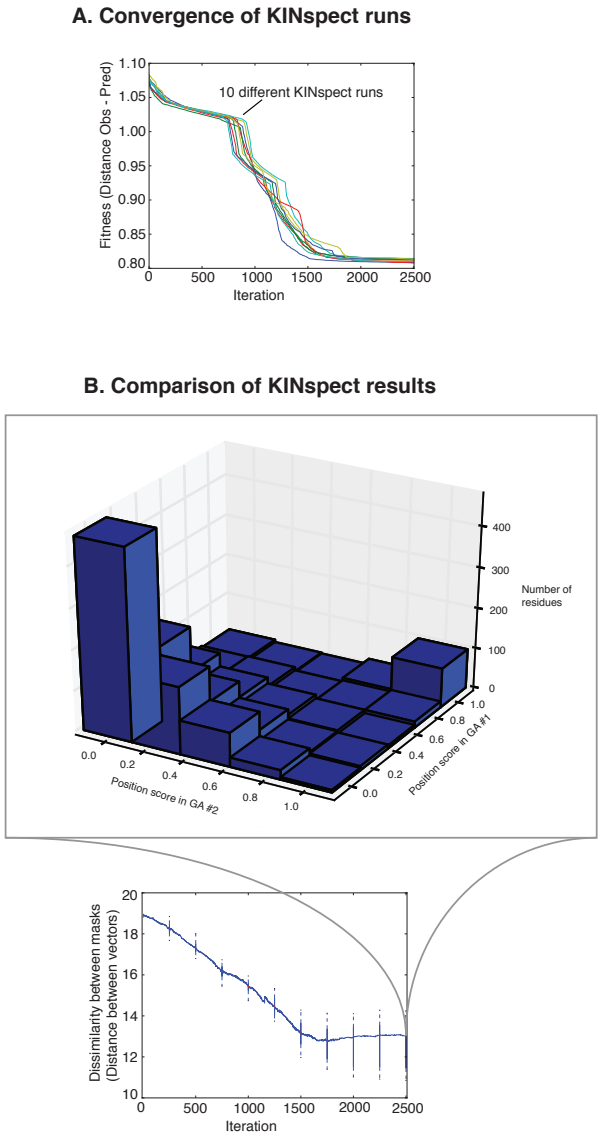


Figure 7.2. (Continued on the following page.)

Figure 7.2. KINspect results. A. Ten different runs of KINspect are compared in terms of their fitness evolution until convergence (2500 generations) and a similar fitness path can be observed for the different runs which, given the parameters are kept homogeneous throughout the runs, is likely to reflect the fitness landscape that is being explored until finding the most predictive solutions. **B.** In addition to converging at a similar generation time, by assessing the dissimilarity between the best performing mask at every iteration for the ten different runs, it can also be demonstrated that the masks tend to be more similar to one another as generation time progresses (lower dissimilarity). For clarity, a comparison between the values of two of the best-performing masks (from two different runs) is shown in the inset. This demonstrates that the different runs converge to the same solution. **C.** By comparing the distribution of all the residues or a set of residues curated from the literature to represent known determinants of specificity, we observe an enrichment (higher score) of these known determinants of specificity in our final masks.

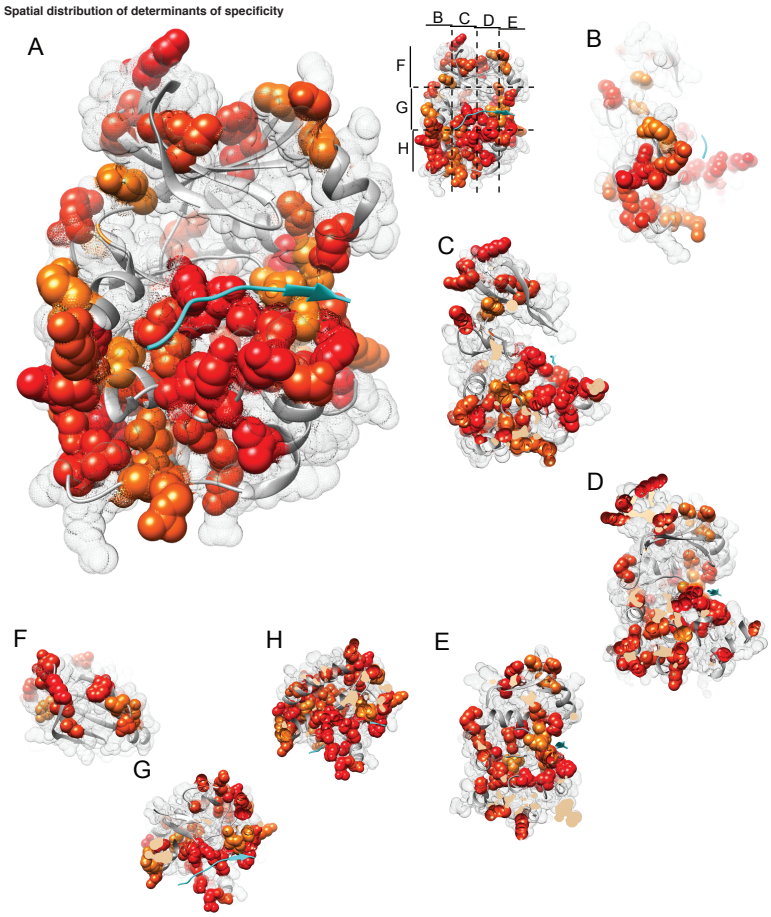


Figure 7.3. Structural representation of the determinants of specificity. A. The three-dimensional structure of the kinase domain of Akt/PKB [PDB code:1O6K] (in surface representation) bound to a substrate (in cartoon representation and cyan) is shown with residues scoring above 0.75 colored in a range between orange and red, red being the highest scoring residues (score of 1.0). For easier perception of the different interactions between the determinants and the substrate, different slices of the representation are shown in the smaller panels as shown in the inset (**B, C, D, E, F, G, H**).

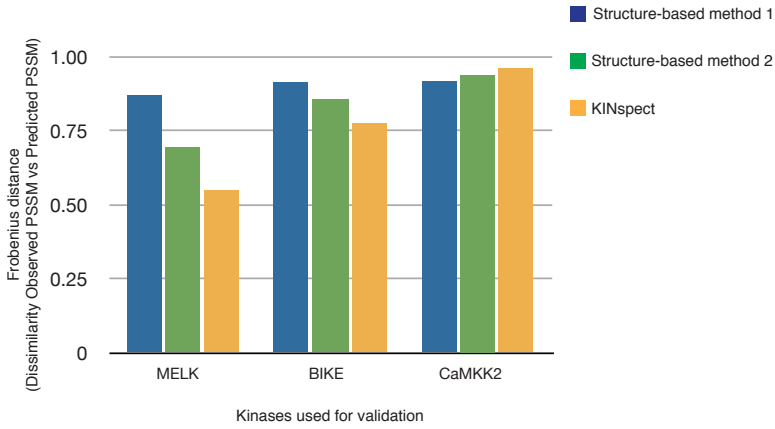


Figure 7.4. Predictive power comparison between methods. KINspect outperforms previous structure-based methods in its ability to predict PSSMs that are close to the experimentally observed ones (smaller Frobenius distance).

Part V

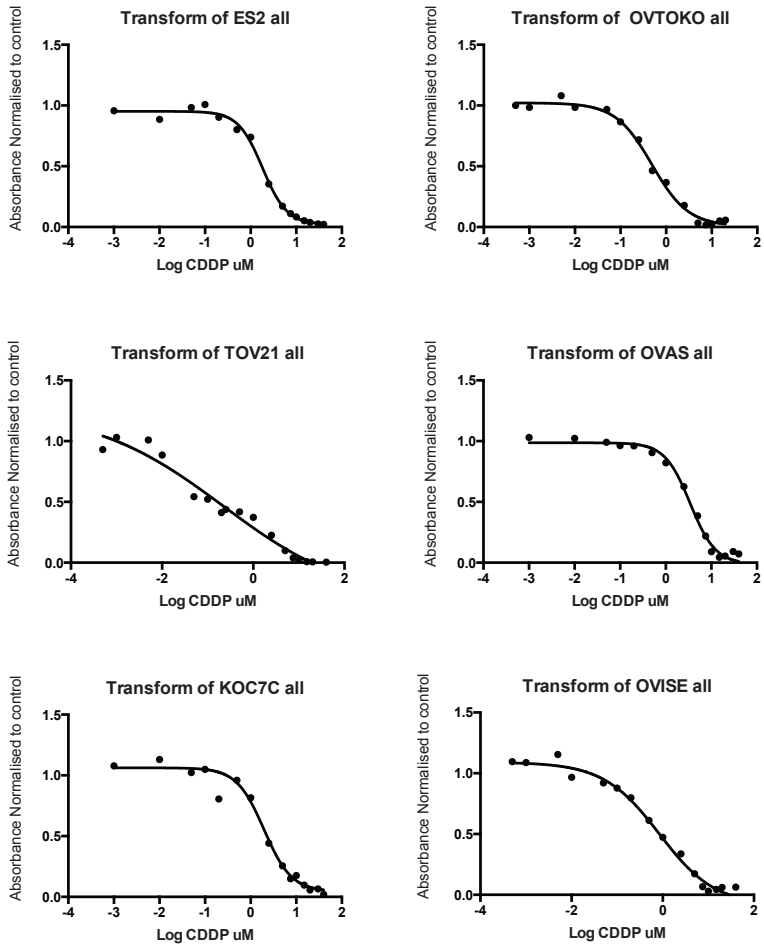
Network-attacking mutations driving resistance to cisplatin in ovarian cancer

Chapter 8

Global sequencing and phospho-proteomic analysis identifies network drivers of drug resistance in ovarian cancer

While previous chapters have been more focused on "proof-of-principle" type of predictions and experimental validations, in this chapter we briefly introduce a more recent study where we applied ReKINect to a biological question that is of high relevance to cancer treatment, namely cancer resistance. To this end, we extended our panel of ovarian cancer cell lines (originally composed of three cell lines and extended to a total of six cell lines for this project), so that we would have the same number of cell lines that would be resistant and sensitive to cisplatin – the standard chemotherapeutic treatment in ovarian clear cell carcinoma. As shown in Figure 8.1, three of our cell lines were resistant to cisplatin (Es-2, Koc7c and Ovas) and three were sensitive (Ovise, Ovtoko and Tov-21). In order to study their differences and suggest a network therapeutic strategy that might prevent the development of cisplatin-resistance, we have performed global exome sequencing combined with ReKINect predictions, global phospho-proteomic analysis, cell viability assays in response to cisplatin developed xenograft models in mice for all the cell lines (Fig. 8.2).

In order to generate some initial hypothesis for potential functional mutations that lead to cisplatin resistance, mutations extracted from exome sequencing of the different cell lines were analyzed with ReKINect and one functional mutation in particular, MAP2K3 T222M, was shown to correlate with resistance to cisplatin (Fig. 8.3). Importantly, as explained in Chapter 6, this mutation destroys an activating phosphorylation site that resides in the activation segment of this kinase (Raingeaud et al., 1996) and therefore leads to its constitutive inactivation. MAP2K3 is directly



	IC50	Range	Fold change	N
ES-2	1.82	1.57-2.10	8	4
Koc7c	1.97	1.67-2.33	9	4
Ovtoko	0.49	0.42-0.57	2	3
Tov-21	0.22	0.10-0.50	1	3
Ovise	0.84	0.61-1.15	4	3
Ovas	3.43	3.08-3.83	16	4

Figure 8.1. Ovarian cell lines cisplatin sensitivity. Actual measurements and summarizing table with IC50 values, where resistant and sensitive cell lines can be derived from (IC50 values above and below 1.00).

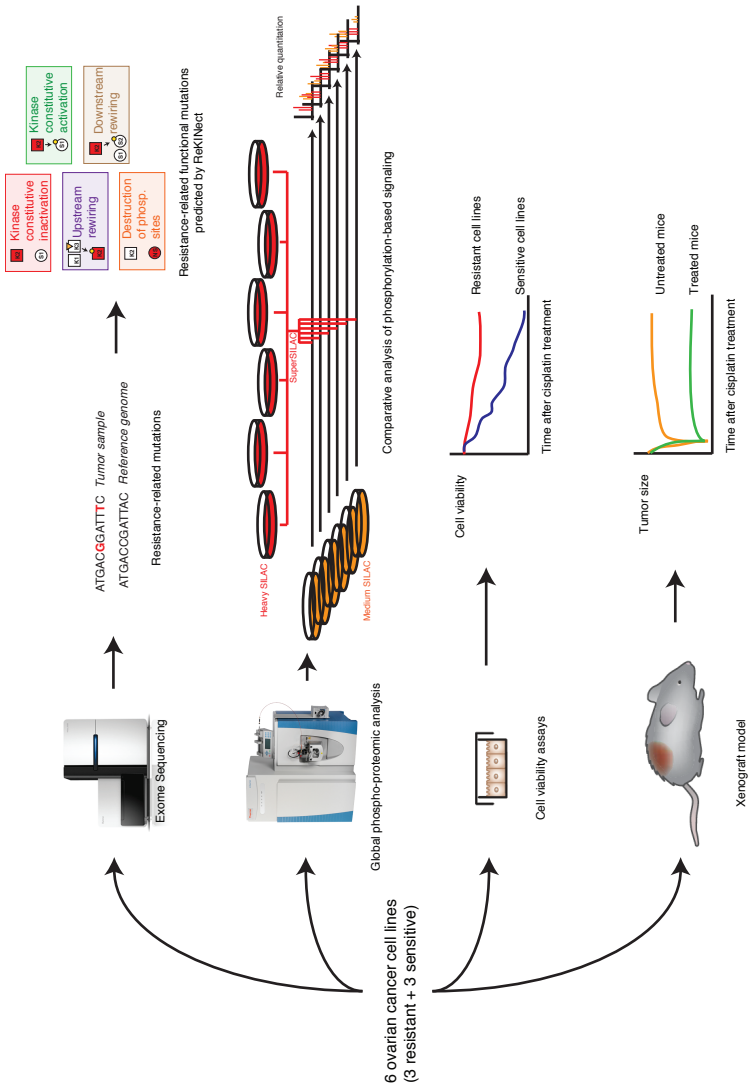


Figure 8.2. Experimental pipeline. 4 different experimental approaches are followed in the hunt of functional mutations that drive resistance to cisplatin, including exome sequencing, global phospho-proteomics, cell viability assays and xenograft models for each ovarian cancer cell line.

upstream p38, one of its known substrates, the dysregulation of which has already been suggested as a mechanism that may lead to cisplatin resistance (Galan-Moya et al., 2011). Despite the fact that p38 (as many other proteins) has already been

suggested as a mechanism for cisplatin resistance, if we can demonstrate the causal relationship between the functional relationship between this specific mutation and cisplatin resistance, this would represent a novel genotype-to-phenotype link in support of this specific mechanism. Thus, in addition to the data analysis that is currently being done to integrate the phospho-proteomic data set generated for all the cell lines, we are currently validating this hypothesis by generating cell lines that either over-express MAP2K3 or where MAP2K3 has been knocked down, so that we can transition from finding a correlation onto demonstrating a causal relationship. If this is successful, we would complete the study by suggesting a specific network medicine therapeutic strategy that would be investigated in our xenograft mouse models.

	Cisplating	MAP2K3
ES-2	Resistant	Mutant
Koc7c	Resistant	Mutant
Ovtoko	Sensitive	Wild type
Tov-21	Sensitive	Wild type
Ovise	Sensitive	Wild type
Ovas	Resistant	Mutant

Figure 8.3. MAP2K3 correlation with resistance. All cell lines presenting the MAP2K3 mutation that leads to a destruction of a phosphorylation site and subsequent inactivation of this kinase are resistant to cisplatin.

Part VI

Epilogue

Chapter 9

Concluding remarks

In this thesis, I have focused on the kinase domain, a protein modular domain that plays critical roles in cellular decision-making processes in healthy cells and is very often perturbed in disease. By conceptually describing network perturbation that may arise in disease, I was able to identify many potential network-attacking mutations. I was also able to propose six specific mechanisms with which these mutations contribute to oncogenesis, thus effectively closing the genotype to phenotype gap.

As network dynamics, network structure and the presence of post-translational modifications are not unique to kinases, we are convinced that network-attacking mutations will be identified for other protein domains and in other diseases. Before an increase in coverage and in overall computational performance (i.e. predictive power) is achieved, it will be crucial to extend our knowledge and data about domain activity, regulation and specificity. For example, it is expected that kinome-wide specificity data should benefit future versions of KINspect and ReKINect.

Another area where we foresee important advances in the near future will be in the development of more advanced computational approaches to predict specificity from sequence, as our current understanding of specificity is not only limited by the amount of data but also by the computational methods used to date, which are not entirely capable of decoding information encapsulated in higher-order relationships between residues (e.g. coupling, allostery or epistasis).

Finally, it is clear that, despite the known enrichment of mutations in protein kinases, a large fraction of cancer mutations hit other parts of the proteome and/or lead to more complex types of mutations than non-synonymous substitutions, which has been the main focus of this thesis. Consequently, fusion proteins, genomic rearrangements, large deletions or insertions or truncated proteins should also become part of ReKINect in the future. Likewise, similarly as has happened with newer versions of NetworKIN (Linding et al., 2007) and NetPhorest (Miller et al., 2008), new versions of ReKINect and KINspect should naturally include more data and predictive capabilities for other protein domains.

Despite this room for improvement, we still believe the frameworks presented in this thesis represent a step forward in our ability to predict disease signaling networks and could hopefully contribute to the development of personalized cancer diagnosis and combinatorial therapeutic treatments in the near future.

Bibliography

American Cancer Society (2012). *The History of Cancer*. American Cancer Society. 3

Brinkworth R.I., Breinl R.A., and Kobe B. (2003). From the Cover: Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *PNAS*, 100(1):74–79. 95

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615. 80

Creixell P., Schoof E.M., Erler J.T., and Linding R. (2012). Navigating cancer network attractors for tumor-specific therapy. *Nature biotechnology*, 30(9):842–848. 5, 79, 80

Davies H., Bignell G.R., Cox C., Stephens P., Edkins S., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954. 5, 80, 81, 82

Diella F., Cameron S., Gemund C., Linding R., Via A., et al. (2004). Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5(1):79+. 81, 84

Ding L., Getz G., Wheeler D.A., Mardis E.R., McLellan M.D., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075. 79

Ding L., Ley T.J., Larson D.E., Miller C.A., Koboldt D.C., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510. 79, 80

Ellis J.J. and Kobe B. (2011). Predicting Protein Kinase Specificity: Predikin Update and Performance in the DREAM4 Challenge. *PLoS ONE*, 6(7):e21169+. 95, 96

Forbes S.A., Bhamra G., Bamford S., Dawson E., Kok C., et al. (2001). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, Chapter 10. 5

Forbes S.A., Bindal N., Bamford S., Cole C., Kok C.Y.Y., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, 39(Database issue):D945–D950. 79, 81

Futreal P.A., Coin L., Marshall M., Down T., Hubbard T., et al. (2004). A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183. 5, 7, 79

Galan-Moya E.M., de la Cruz-Morcillo M.A., Valero M.L., Callejas-Valera J.L., Melgar-Rojas P., et al. (2011). Balance between MKK6 and MKK3 Mediates p38 MAPK Associated Resistance to Cisplatin in NSCLC. *PLoS ONE*, 6(12):e28406. 105

Greenman C., Stephens P., Smith R., Dalgliesh G.L., Hunter C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158. 79, 80

Hanahan D. and Weinberg R.A. (2000). The Hallmarks of Cancer. *Cell*, 100(1):57–70. 4, 79

Hanahan D. and Weinberg R.A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674. 4

Hanks S.K. and Hunter T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB Journal*, 9(8):576–596. 7

Hanks S.K., Quinn A.M., and Hunter T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, 241(4861):42–52. 7

Henikoff S. and Henikoff J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919. 93

Hornbeck P.V., Chabra I., Kornhauser J.M., Skrzypek E., and Zhang B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561. 81, 84

Huang P.H., Mukasa A., Bonavia R., Flynn R.A., Brewer Z.E., et al. (2007). Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12867–12872. 80

Huse M. and Kuriyan J. (2002). The conformational plasticity of protein kinases. *Cell*, 109(3):275–282. 7

Hutti J., Jarrell E., Chang J., Abbott D., Storz P., et al. (2004). A rapid method for determining protein kinase phosphorylation specificity. *Nat Meth*, 1(1):27–29. 10

Janes K.A., Albeck J.G., Gaudet S., Sorger P.K., Lauffenburger D.A., et al. (2005). A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis. *Science*, 310(5754):1646–1653. 5

Johnson L.N., Noble M.E., and Owen D.J. (1996). Active and inactive protein kinases: structural basis for regulation. *Cell*, 85(2):149–158. 7, 82

Lee M.J., Ye A.S., Gardino A.K., Heijink A.M.M., Sorger P.K., et al. (2012). Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell*, 149(4):780–794. 80

Ley T.J., Mardis E.R., Ding L., Fulton B., Mclellan M.D., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72. 79

Linding R., Jensen L.J., Ostheimer G.J., van Vugt M.A.T.M., Jørgensen C., et al. (2007). Systematic Discovery of In Vivo Phosphorylation Networks. *Cell*, 129(7):1415–1426. 8, 81, 83, 89, 109

Manning G., Whyte D.B., Martinez R., Hunter T., and Sudarsanam S. (2002). The Protein Kinase Complement of the Human Genome. *Science*, 298(5600):1912–1934. 7

Miller M.L., Jensen L.J., Diella F., Jorgensen C., Tinti M., et al. (2008). Linear Motif Atlas for Phosphorylation-Dependent Signaling. *Sci. Signal.*, 1(35):ra2+. 81, 83, 94, 109

Mok J., Kim P.M., Lam H.Y.K., Piccirillo S., Zhou X., et al. (2010). Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. *Sci. Signal.*, 3(109):ra12+. 95

Monge J., Kricun M., Radovøiø J., Radovøiø D., Mann A., et al. (2013). Fibrous Dysplasia in a 120,000+ Year Old Neandertal from Krapina, Croatia. *PLoS ONE*, 8(6):e64539. 3

Mukherjee S. (2010). *The Emperor of All Maladies: A Biography of Cancer*. Scribner, 1 edition. 3, 4

Nik-Zainal S., Alexandrov L.B., Wedge D.C., Van Loo P., Greenman C.D., et al. (2012a). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993. 79

Nik-Zainal S., Van Loo P., Wedge D.C., Alexandrov L.B., Greenman C.D., et al. (2012b). The life history of 21 breast cancers. *Cell*, 149(5):994–1007. 79

Nolen B., Taylor S., and Ghosh G. (2004). Regulation of Protein Kinases: Controlling Activity through Activation Segment Conformation. *Molecular Cell*, 15(5):661–675. 7, 82

Pawson T. and Linding R. (2008). Network medicine. *FEBS Letters*, 582(8):1266–1270. 79

Raingeaud J., Whitmarsh A.J., Barrett T., Dérijard B., and Davis R.J. (1996). MKK3- and MKK6-regulated gene expression is mediated by the p38 mitogen-activated protein kinase signal transduction pathway. *Molecular and cellular biology*, 16(3):1247–1255. 84, 103

Schoeberl B., Pace E.A., Fitzgerald J.B., Harms B.D., Xu L., et al. (2009). Therapeutically Targeting ErbB3: A Key Node in Ligand-Induced Activation of the ErbB Receptor-PI3K Axis. *Sci. Signal.*, 2(77):ra31. 80

Sjöblom T., Jones S., Wood L.D., Parsons D.W., Lin J., et al. (2006). The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science*, 314(5797):268–274. 79

Songyang Z., Carraway K.L., Eck M.J., Harrison S.C., Feldman R.A., et al. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature*, 373(6514):536–539. 80, 83

Stehelin D., Varmus H.E., Bishop J.M., and Vogt P.K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173. 5

Stratton M.R., Campbell P.J., and Futreal P.A. (2009). The cancer genome. *Nature*, 458(7239):719–724. 4, 80

the International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 4

Turk B.E. (2008). Understanding and exploiting substrate recognition by protein kinases. *Current opinion in chemical biology*, 12(1):4–10. 9

Turk B.E., Hutti J.E., and Cantley L.C. (2006). Determining protein kinase substrate specificity by parallel solution-phase assay of large numbers of peptide substrates. *Nature protocols*, 1(1):375–379. 10

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., et al. (2001). The Sequence of the Human Genome. *Science*, 291(5507):1304–1351. 4

Vogelstein B. and Kinzler K.W. (2004). Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799. 7, 79

Vogelstein B., Papadopoulos N., Velculescu V.E., Zhou S., Diaz L.A., et al. (2013). Cancer Genome Landscapes. *Science*, 339(6127):1546–1558. 4

Wan P.T., Garnett M.J., Roe S.M., Lee S., Niculescu-Duvaz D., et al. (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, 116(6):855–867. 80, 82

Wong K.M., Hudson T.J., and McPherson J.D. (2011). Unraveling the Genetics of Cancer: Genome Sequencing and Beyond. *Annual Review of Genomics and Human Genetics*, 12(1):407–430. 4, 79

Wood L.D., Parsons D.W., Jones S., Lin J., Sjöblom T., et al. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, 318(5853):1108–1113. 79

Wu M., Pastor-Pareja J.C., and Xu T. (2010). Interaction between RasV12 and scribbled clones induces tumour growth and invasion. *Nature*, 463(7280):545–548. 5

Yaffe M.B. (2013). The Scientific Drunk and the Lamppost: Massive Sequencing Efforts in Cancer Discovery and Treatment. *Sci. Signal.*, 6(269):pe13+. 4

Yaffe M.B., Leparc G.G., Lai J., Obata T., Volinia S., et al. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 19(4):348–353. 10

Zeqiraj E. and van Aalten D.M. (2010). Pseudokinases-remnants of evolution or key allosteric regulators? *Current opinion in structural biology*, 20(6):772–781. 7, 80, 81, 82

Zhang H., Lund O., and Nielsen M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*, 25(10):1293–1299. 93