NOVEL IN-SILICO APPROACHES TO IDENTIFY DRIVER MUTATIONS IN CANCER GENOMICS DATA

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2016

Andrew Michael Hudson

Cancer Research UK Manchester Institute

List of Contents

Word Count: 53120

List of Tables and Figures

Chapter 2 (Publication 1)

Chapter 3

Chapter 4

Chapter 5 (Publication 2)

Chapter 6 (Publication 3)

Chapter 7 (Publication 4)

Appendix 2

List of Abbreviations

AML – Acute Myeloid Leukaemia

ATCC – American Type Culture Collection

ATP – Adenosine Triphosphate

CCLE – Cancer Cell Line Encyclopaedia

COSMIC – Catalogue Of Somatic Mutations in Cancer

CRUK_MI – Cancer Research UK Manchester Institute

dbSNP – Database of Single Nucleotide Polymorphisms

DIPNECH – Diffuse Idiopathic Pulmonary Neuro Endocrine Cell Hyperplasia

DMEM – Dulbecco's Modified Eagle Medium

DMSO – Dimethyl Sulfoxide

DNA – Deoxyribonucleic Acid

DNTP – Deoxynucleotide Triphosphate

DSMZ – Deutsche Sammlung von Mikroorganismen und Zellkulturen

EGF – Epidermal Growth Factor

EHS – Engelbreth-Holm Swarm

EGFR – Epidermal Growth Factor Receptor

FAP – Familial Adenomatous Polyposis

FFPE – Formalin-Fixed Paraffin Embedded

GATK – Genome Analysis Tool Kit

GIST – Gastro Intestinal Stromal Tumour

GOF – Gain of Function

ICGC – International Cancer Genomics Consortium

IGV – Integrative Genomics Viewer

LOF – Loss of Function

MD – Molecular Dynamics

MEN – Multiple Endocrine Neoplasia

NCBI – National Center for Biotechnology Information

NICE – National Institute for Clinical Excellence

NGS – Next Generation Sequencing

NSCLC – Non-small Cell Lung Cancer

PAH – Polycyclic Aromatic Hydrocarbons

PFS – Progression Free Survival

PI3K – Phosphoinositide-3 kinase

PKA – Protein Kinase A

PCR – Polymerase Chain Reaction

RMSD – Root Mean Squared Deviation

RMSF – Root Mean Squared Fluctuation

RNA – Ribonucleic Acid

RPMI – Roswell Park Memorial Institute Medium

RTK – Receptor Tyrosine Kinase

SCLC – Small Cell Lung Cancer

SDM – Site-directed Mutagenesis

SDS – Sodium Dodecyl Sulphate

Ser - Serine

shRNA – Small Hairpin Ribonucleic Acid

siRNA – Small Interfering Ribonucleic Acid

SNV – Single Nucleotide Variant

TCGA – The Cancer Genome Atlas

Thr – Threonine

TS – Tumour Suppressor

Tyr - Tyrosine

VEGF – Vascular Endothelial Growth Factor

WT – Wild Type

Abstract

Novel In-Silico Approaches to Identify Driver Mutations in Genomics Data submitted by Andrew Michael Hudson for the degree of Doctor of Philosophy**,** The University of Manchester, September 2016

The cancer genomics era has now witnessed multiple examples of the clinical benefits of targeting mutated oncogenes with selective inhibitors. Many patients who would have been too frail to withstand chemotherapy, and therefore have no treatment options, have been given additional months or even years of good quality life with these therapies. These successes have driven the search for oncogenes in all tumour subtypes and next generation sequencing (NGS) has been the main tool used for this task. However despite the sequencing of hundreds of thousands of cancer samples with NGS, knowledge is still lacking regarding driver genes for a significant proportion of cancer subtypes. The overarching theme of this thesis is the development of novel in silico approaches to enhance the detection of novel driver mutations.

The first study uses discrepancies in NGS mutation calling between different institutes to identify consistent areas of the exome that are poorly sequenced. These areas are predominantly GC-rich causing a problem for cancers, such as lung cancer, with a mutational signature preferentially affecting GC-rich regions. As proof of principle, a PAK4 mutation in a GC-rich region previously missed by two large sequencing studies is shown to be activating upon the ERK pathway. Analysis of NGS read coverage demonstrates that inadequate NGS of GC-rich regions is prevalent in older projects, potentially obscuring driver mutation detection.

The remainder of the work focuses on using novel filtering algorithms to cut through the mutational noise caused by the hundreds of inconsequential passenger mutations present in each NGS dataset. Linking structural and functional data of critical kinase motifs I developed a specific loss-of-function (LOF) kinase mutation caller. Analysing pan-cancer somatic mutational data with this algorithm revealed MKK7 as a novel tumour suppressor in gastric cancer that was subsequently validated biochemically. The final piece of work applies the LOF algorithm to the germline data of a rare neuroendocrine proliferative condition called DIPNECH. This identified a previously unreported SNV of CDK8 in one patient. Reanalysing pan cancer somatic mutational data reveals similar predicted LOF mutations of CDK8 in small cell lung cancer, also a neuroendocrine malignancy. Additional algorithms were used to filter the DIPNECH data to provide the first genomic insight into this rare disease, highlighting potential genetic links with other neuroendocrine malignancies.

Overall the approaches presented in this thesis offer additional solutions to enhance the detection of driver genes from NGS datasets. Most importantly this work demonstrates the power of using functional considerations to filter driver mutations from the mutational noise. In the future I aim to develop this approach to assist gain-of-function (GOF) kinase mutation identification and write similar algorithms for other protein families.

Declaration

Publication 5 in the Appendix of this thesis has been previously submitted by Corina Antal in support of an application for Doctor of Philosophy at the University of California.

No other portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses

Author Contributions

This thesis is submitted in alternative format as it allows the most appropriate structure to present the three publications that have already been published (2 in main body of thesis and 1 in appendix) in addition to the two publications that are currently being submitted for publication.

The structure of this thesis is as follows:

Chapter 1: Introduction - Cancer Genomics
Chapter 2: **Publication 1** – Review of Bioinformatic Challenges
Chapter 3: Introduction - Functional Data to Filter Datasets
Chapter 4: Materials and Methods
Chapter 5: **Publication 2** – Discrepancies in Cancer Genomic Sequencing
Chapter 6: **Publication 3** – Kinase Motif Filtering of Datasets
Chapter 7: **Publication 4** – DIPNECH germline variants
Chapter 8: Discussion
Chapter 9: References
Chapter 10: Appendices – Including **Publication 5**, Primer List, and R-Script

The authorship contributions for the publications are as follows:

Publication 1: Hudson AM, Wirth C, Stephenson NL, Brognard J, Miller C. Using large-scale genomics data to identify driver mutations in lung cancer: methods and challenges. *Pharmacogenomics* 2015: 16(10); 1149-60

90% of the writing of this review was carried out by the first author with the remaining 10% contributed by the other authors. Crispin Miller and John Brognard helped edit the manuscript.

Publication 2: Hudson AM, Yates T, Li Y, Trotter EW, Fawdar S, Chapman P, Lorigan P, Biankin AV, Miller C, Brognard J. Discrepancies in Cancer Genomic Sequencing Highlight Opportunities for Driver Mutation Discovery. *Cancer Research* 2014: 74; 6390-6

The project was conceived by the first author with supervision from John Brognard and Crispin Miller. Bioinformatic code was written by Tim Yates and Yaoyong Li. The

biochemical validation was performed by Eleanor Trotter. The manuscript was written by the first author and edited by John Brognard and Crispin Miller.

Publication 3: Hudson AM*, Stephenson NL*, Li C, Trotter EW, Fawdar S, Katona G, Bieniasz-Krzywiec P, Howell M, Wirth C, Miller C, Brognard J. Functional Screening of Cancer Datasets Identifies Novel Targets and Mutational Hotspots in the Tumour-Suppressing Kinome. *Indicates equal contribution by authors.

The project was conceived by the first author with supervision from John Brognard and input from Natalie Stephenson. The bioinformatic code was written entirely by the first author with advice from Chris Wirth and Crispin Miller. Natalie Stephenson performed the Molecular Dynamics simulations. Natalie Stephenson and Eleanor Trotter performed the majority of the biochemical experiments with assistance from the first author. Cynthia Li, Gitta Katona and Patricja Bieniasz-Krzywiec were students supervised by the first author and Natalie Stephenson. 90% of the manuscript was written by the first author with the remaining contribution from Natalie Stephenson and editing by John Brognard.

Publication 4: Hudson AM, Leong HS, Miller C, Mansoor W, Brognard J, Germline sequencing of DIPNECH patients reveals a heterogeneous disease with novel SNPs in gene that are somatically mutated in other neuroendocrine malignancies.

The project to sequence the DIPNECH patients was conceived by the first author, Was Mansoor and John Brognard. Patients were identified by Was Mansoor. The experimental strategy and SNV filtering approach was devised and performed entirely by the first author using bioinformatic code written by the first author. Bioinformatic support to prepare SNV data for analysis was performed by Hui Sun Leong and Crispin Miller. 95% of the manuscript was written by the first author.

Publication 5: Antal CE, **Hudson AM**, Kang E, Wirth C, Stephenson NL, Trotter EW, Zanca C, Gallegos LL, Furnari FB, Miller CW, Hunter T, Brognard J, Newton AC. Protein Kinase C Loss-of-Function Mutations in Cancer Reveal Role as Tumor Suppressor, *Cell* 2015: 160(3); 489-502

The second author provided bioinformatic analysis of the mutational data and assisted in the preparation of the manuscript.

Acknowledgements

Firstly I would like to thank my supervisors John Brognard and Paul Lorigan and my advisor Fiona Blackhall for the invaluable support, guidance and mentorship they have given me over the course of the last 4 years. They continue to give me great advice about projects and career plans and I look forward to working with them all again in the future. In addition Tim Illidge has been a great mentor to me and is responsible for introducing me to the world of academic medicine.

I would like to acknowledge the help, encouragement and friendship of everyone in the Signalling Networks in Cancer group both past and present. In particular I have worked closely with Eleanor Trotter and Natalie Stephenson who provided me with essential guidance through the various pitfalls and challenges of lab work. Thank you. I have also had the pleasure of working with Crispin Miller and members of his group including Tim Yates, Chris Wirth, Yaoyong Li and Hui Sun Leong. I thank them all for their advice, technical expertise and patience. For the DIPNECH study I would like to thank the patients who kindly donated their blood, Was Mansoor for proposing the idea and giving me access to his patients and the MCRC Biobank for collecting and processing the samples.

Finally I would like to thank Cancer Research UK and all the people who donate to the charity for funding this research, without whom none of this work would have been possible.

Chapter One

Introduction

Cancer Genomics

1.1 Evolution of Cancer Genomics

Genomics, the study of genes, began with Gregor Mendel's pea plants in 1866 although it took another 50 years to confirm that hereditary material is made from DNA [1]. One of the earliest contributions to this realisation came from Theodor Boveri in 1902 when he discovered that different chromosome imbalances led to phenotypic changes in offspring depending on the different chromosomes that were present [2]. Prior to the discovery of the double helix structure of DNA by Watson, Crick, and Franklin in 1953 [3, 4], it was demonstrated how DNA (not protein) was required to transform cells with a virus [5]. Following the double helix discovery further studies expanded knowledge of DNA function including the discovery of DNA replication via polymerases (1955) [6], the existence of mRNA required for protein translation (1961) [7], and the cracking of the nucleotide code (1962-66) [8]. Frederick Sanger invented a method for sequencing the nucleotide code in 1977 giving researchers an invaluable tool to directly observe how specific genetic aberrations altered cell phenotype [9]. Sanger sequencing allowed a rapid growth in the search for genetic aberrations in many different diseases including cancer and the method is still used today as a validation tool for the newer sequencing technologies.

It had been known since 1911 that an infective agent could induce the formation of sarcoma tumours in chickens. However it required over half a century and the aforementioned advances in genomic knowledge before the causative agent was revealed as a retroviral gene and the world's first oncogene v-src was discovered [10]. Prior to this discovery Boveri had concluded that tumour growth resulted from abnormal chromosomal combinations conferring a growth advantage that is passed on to daughter cells. Substituting the word 'chromosome' for the word 'gene' (which was not used until 1909) in Boveri's conclusion highlights the magnitude of his discovery [2]. In addition to this theory, Boveri is also credited with predicting oncogenes, tumour suppressor genes, germline susceptibility to cancer and even cell cycle checkpoints. Sadly his genius remains relatively unknown compared to the other leading figures in the history of genomics [2]. By the early 1980's the protein product of v-src was shown to be a tyrosine kinase [11] and shortly afterwards the genetic sequences of other viral tumour transforming agent were identified as ras oncogenes [12]. This method of discovering viral transforming agents and linking the sequences to human homologs was used repeatedly in the 1980's to discover other oncogenic human protein kinases such as EGFR [13]. In 1982 the first example of an

activating mutation in an oncogene derived from tumour tissue was discovered in codon 12 of the hras gene in the EJ/T24 bladder cancer cell line. The following year the kras oncogene was found mutated in lung and colon cancer tissues. We now know that ras oncogene mutations occur in approximately 35%, 50%, and 15% of lung adenocarcinoma, bowel adenocarcinoma, and bladder urothelial cancers respectively [14]. In addition many other cancers such as pancreatic adenocarcinoma have very high frequency of ras mutations (95% of cases in the case of pancreatic ductal adenocarcinoma) [15].

In contrast to activating mutations in oncogenes, the 1970's and 80's also saw the identification of genes in humans such as rb1 and tp53 in which loss of normal function led to tumour development [16, 17]. These have since been named tumour suppressor genes and were supported by earlier work on the multi-step nature of carcinogenesis. In 1953 Carl Nordling stated that cancer rates in males increase with the sixth power of the man's age and therefore six mutations are required to form a cancer [18]. With current sequencing technology this estimate has been revised downwards to 3 mutations for certain cancers [19]. Nordling's contribution was important because it introduced the concept of the multi-step acquisition of somatic mutations in carcinogenesis. Knudson used this idea in 1971, before the retinoblastoma gene, rb1, had been identified, to hypothesize why certain retinoblastoma patients developed the disease in early childhood whilst others did not get the disease to later in life [20]. His 'two-hit hypothesis' postulated that children with the disease were born with an inherited abnormal allele and acquired a single somatic mutation early in life. However an adult took longer to present with the disease because they were born with two normal alleles and took time to acquire somatic mutations in both alleles [20]. Sequencing retinoblastoma cases later proved Knudson's hypothesis and the mechanism of requiring a second allele of a tumour suppressor gene to be altered by an additional mutation, copy number loss or gene methylation is frequently observed [21].

The continuing search for oncogenes and tumour suppressor genes was initially limited by the time-consuming nature of genomic sequencing dependent on the process devised by Sanger [17]. In the 1980's prior to the automation of the Sanger technique the sequencing of just one gene could take many years [22]. By the 1990's advances in Sanger sequencing technology allowed an international consortium (Human Genome Project) to contemplate attempting to sequence the entire human genome. In 2001, after astronomical expense and large-scale international

collaboration, the first attempt at sequencing the whole human genome was published [23]. This knowledge enabled the development of next generation sequencing (NGS) platforms with the first commercially available platforms coming online around 2005 [24]. Over the last decade this technology has become sufficiently advanced to allow the whole exome sequencing of a single patient tumour (and recently even a single cell of a tumour) in less than 24 hours. The cost of sequencing a whole genome has dropped from the 3.8 billion dollars it cost the Human Genome Project to sequence the first genome to around 1000 dollars today [25]. The improvement in technology and cost effectiveness has led to Illumina, the world's largest manufacturer of next-generation sequencing machines, to estimate that 1.6 million whole genomes will be sequenced in 2017 [26].

1.2 Clinically actionable genetic drivers from the pre-NGS era

The genetic targets for the first targeted therapies to show clinical efficacy and become the standard of care for solid tumour subtypes in the United Kingdom (BRAF, EGFR, KIT, VEGF-A and HER2) were all discovered in the pre-NGS era. This is partly due to the lead-time between identifying driver mutations and developing efficacious drugs with a proven clinical benefit. Another reason is that these aberrations were discovered in the pre-NGS era because they occur at a relatively high frequency in their respective tumours, a factor that also enhances the commercial value of developing an inhibitor.

1.2.1 Epidermal Growth Factor

Stanley Cohen discovered Epidermal Growth Factor (EGF) in 1962 from the extracts of mouse salivary glands. Injecting the extracts into newborn mice induced multiple physiological effects including tooth eruption and epidermal proliferation. Later Cohen identified the receptor (EGFR) on cell membranes and in 1980 it was found that EGFR transmits EGF signalling via tyrosine kinase activity [27]. By 1983 monoclonal antibodies had been successfully used to block EGFR both in cancer cell lines and mouse models resulting in inhibited growth [28, 29]. This led to the development of cetuximab, a clinical EGFR blocking monoclonal antibody, which demonstrated significant improvement in response rates and survival when combined with irinotecan in the management of colorectal cancer [30]. This benefit is

not seen in colorectal tumours with mutant kras as the MEK-ERK pathway is constitutively activated regardless of signalling through EGFR [31]. Therefore the National Institute for Clinical Excellence (NICE) did not approve cetuximab for KRAS mutation positive tumours, meaning that KRAS mutation testing became the first large-scale genetic test used in clinical practice. Clinical trials have also indicated that cetuximab in combination with radiotherapy is as effective in the treatment of locally advanced head and neck cancers as radiotherapy and cisplatin [32]. Given that cisplatin can be toxic to certain patients, cetuximab is preferred and approved by NICE in certain clinical situations. The actual EGFR gene was located in 1984 due to its close resemblance to the transforming viral protein v-erbB [13]. Other structurally related receptors were subsequently discovered leaving 4 family members; ErbB-1 (EGFR), ErbB-2 (HER2), ErbB-3 (HER3), and ErbB-4 (HER4). Activating epidermal growth factor receptor (EGFR) mutations were found in 2004 by two independent research groups that performed Sanger sequencing on lung cancer samples that were responsive to EGFR inhibitors [33, 34]. EGFR inhibition has subsequently been shown in a randomised clinical trial to give superior progression free survival and lower treatment related toxicity compared to standard doublet chemotherapy in patients with lung adenocarcinoma [35]. In this particular study of East Asian never or light smokers the EGFR mutation rate was 59.7%, which is, much higher than encountered in the UK (only 11%). EGFR inhibition is currently NICE approved only for tumours with EGFR mutation positive or in cases where the mutation status is not known but there is a strong suspicion of being EGFR mutation positive (such as a history of never smoking).

1.2.2 Other receptor tyrosine kinases

Following the discovery of the oncogenic properties of EGFR, another receptor tyrosine family member ErbB-2 (HER2) was found to be overexpressed in breast tumours [36]. Clinical studies reveal that approximately 25% of breast tumours have amplified HER2 and that these patients have worse survival [37]. Analogous to EGFR blockade, HER2 blocking monoclonal antibodies showed in vivo and in vitro activity leading to the development of a HER2 blocking monoclonal antibody for clinical use [38]. Randomised phase 3 trials demonstrated improved survival in patients with amplified HER2 receptor receiving trastuzumab (a HER2 blocking monoclonal antibody) with chemotherapy compared to chemotherapy alone [38]. This led to NICE approval and trastuzumab becoming standard of care for patients

with amplified HER2 status. Lapatinib, a small molecule tyrosine kinase inhibitor of HER2 and EGFR, was also developed to treat HER2 amplified patients. Lapatinib demonstrated increased progression free survival in combination with chemotherapy compared to chemotherapy alone in patients with HER2 amplification who had relapsed on trastuzumab leading to it becoming standard of care in this setting [39].

The discovery of the ErbB family of receptor tyrosine kinases (RTKs) led to the discovery of other families of RTKs such as c-kit that transduces the signal from stem-cell factor (SCF) to the phosphatidylinositol 3-kinase (PI3K) pathway. Activating mutations in c-kit were first discovered in 1993 in a mast cell leukaemia cell line [40]. It was observed in this cell line that c-kit was constitutively phosphorylated and activated in the absence of SCF [40]. Sanger sequencing of the receptor in this cell line revealed a mutation that was confirmed to produce a constitutively activate variant of the kinase. This observation was considered 5 years later by a group investigating the drivers of gastrointestinal stromal tumours (GIST) who found activating mutations in c-kit in 5 out of 6 GISTs via Sanger sequencing [41]. In 2008 a phase 3 randomised control trial confirmed the clinical efficacy of Imatinib, a small molecule tyrosine kinase inhibitor that targets KIT amongst other kinases, against GISTs leading to its adoption into standard clinical practice [42]. Interestingly the kinase domain c-kit mutation originally discovered in the mast cell leukaemia is resistant to the actions of Imatinib [43]. However other high expressing KIT tumours can have dramatic and sustained responses to Imatinib whilst retaining a good quality of life for years.

The vascular endothelial growth factor (VEGF) axis has also been targeted with variable clinical success in a number of solid tumour subtypes. Like many of the aforementioned targets, the story started in the 1980's with the discovery of a growth factor. In this case the VEGF was noted to increase vascular permeability and prevent endothelial cell apoptosis [44]. Unlike EGFR, HER2 and c-kit, the VEGF axis was first targeted with a monoclonal antibody (bevacizumab) that bound the growth factor rather than the receptor [44]. Bevacizumab has demonstrated clinical efficacy in multiple tumour subtypes including colorectal, lung and renal carcinoma [44]. However it is more toxic than many of the other targeted therapies causing hypertension, oedema and renal problems in a significant proportion of patients [45]. Inhibitors against the VEGF receptor tyrosine kinase have also seen variable successes in clinical practice.

1.2.3 BRAF

Human RAF kinases were discovered via the oncogenic murine homolog v-raf in the 1980's [46]. Three human homologs were subsequently identified as araf, braf and craf although activating mutations in these kinases were not discovered until 2002 [47]. In a seminal paper and impressive pre-NGS era sequencing effort 923 cell lines and primary human tumours were sequenced for all coding regions of BRAF and the three RAS isoforms [48]. This was the first example of the sequencing of large number of cancer samples to discover high frequency driver mutations. It revealed that BRAF is mutated in 66% of malignant melanomas with 80% of these mutations being p.V600E. The paper also reported the constitutive activation of the p.V600E BRAF mutation on the MEK-ERK pathway. As a result small tyrosine kinase inhibitors were constructed and shown to be highly effective in the treatment of metastatic melanoma. A randomised phase 3 trial of Vemurafenib (an inhibitor of BRAF) compared to Dacarbazine chemotherapy (the standard of care at the time) showed superior overall and progression free survival in BRAF V600E positive patients [49]. In a cancer that had not seen any progress in efficacious treatments in two decades this result paved the way to the introduction of Vemurafenib into the clinic as the standard of care for melanoma patients. Due to the dire situation for melanoma patients pre-Vemurafenib, the efficacy of the drug in terms of survival and quality of life improvement, and the high frequency of cancers with the mutation, the BRAF V600E story is arguably the best example of the beneficial effects of translational cancer genomics and the benchmark that all new driver mutation discoveries aspire to.

1.2.4 Gene translocations

In addition to missense mutations, the products of chromosomal translocations have also been targeted successfully in the oncology clinic. Fusion proteins result from DNA breaks and subsequent abnormal joining (translocation) of a portion of one gene onto another so that the transcribed mRNA is a hybrid of two genes. The resultant protein may contain functional elements, such as an intact kinase domain, that is either released from the inhibitory control of other lost parts of the protein or enhanced by structural features of the adjoining protein. The most well-known fusion protein in oncology is BCR-ABL in Chronic Myelogenous Leukaemia (CML) caused by the t(9;22)(q34;q11) translocation that was discovered in 1985. The BCR-ABL

fusion protein retains the catalytic activity of ABL but with the addition of part of BCR it more readily dimerises resulting in enhanced autophosphorylation and activation leading to increased activation of downstream proliferative signals [50]. Imatinib, mentioned in section 1.2.2 because of its ability to inhibit oncogenic KIT, also inhibits the catalytic activity of ABL leading to great clinical benefit in patients with CML [51]. More recently in 2007 the EML4-ALK translocation was discovered in a small proportion of lung cancer samples [52]. Whilst some populations such as Europeans have much lower rates, it is estimated that EML4-ALK translocations occur in over 60,000 patients diagnosed with NSCLC annually worldwide [53]. The fusion points in different EML4-ALK translocations are variable but always result in the translation of an intact ALK kinase domain in addition to a dimerization domain of EML4 [54]. Like, BCR-ABL, the EML4-ALK fusion protein is thought to be oncogenic through enhanced dimerization, autophosphorylation and increased activation of downstream signalling [54]. The ALK kinase activity of the fusion protein is inhibited by small molecule inhibitors such as Crizotinib, which (like Imatinib in CML) has revolutionized the treatment of patients with EML4-ALK translocations [55]. In a phase 3 study of patients with ALK rearrangement positive lung cancer, median progression-free survival was more than doubled with Crizotinib compared to standard chemotherapy (7.7 months versus 3.0 months) [55]. This success has resulted in routine genetic testing for ALK rearrangements becoming standard of care in many oncology centres.

The therapies summarised in this section have revolutionised the management of their respective tumour subtypes improving response rates and survival times for patients with specific genetic aberrations often with much fewer side effects than chemotherapy. These successes have driven researchers in the cancer genomics field to find more genetic targets with the ultimate aim being to find genetic drivers for every cancer to develop equally efficacious treatments.

1.3 NGS Era

The completion of the Human Genome Project provided the reference genome and technologies from which to develop NGS sequencing platforms. As NGS technology improved through the mid-2000s it allowed whole genomes and exomes to be sequenced faster and more cheaply, opening up the technology to an increasing number of researchers. The first commercially available NGS machines came online

around 2005 allowing a new approach to target discovery. During this time Bert Vogelstein was completing a final pre-NGS project of epic proportions. His team had embarked on sequencing the whole exomes of multiple breast and colon cancers using Sanger technology [39]. In the end this great effort sequenced the whole exomes of 11 breast and 11 colorectal tumours. For the first time the scale of mutations per cancer sample became apparent. Previously, limited by the resources required to sequence multiple genes by Sanger sequencing, researchers were only able to identify a handful of gene mutations per cancer sample. Vogelstein's study demonstrated that each tumour consisted of approximately 80 mutations with some genes such as TP53 and PIK3CA mutated to a high frequency and many other genes at lower frequencies [56]. With more comprehensive NGS it has been demonstrated that some tumours have thousands of mutations. However Vogelstein's study was the first to indicate that to identify significant functional driver mutations occurring at low frequency requires hundreds of samples from the same cancer subtype to be sequenced.

To co-ordinate the large scale sequencing efforts and correlate the data of different tumour subtypes The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (IGCC), were established. TCGA published the first large scale NGS tumour subtype analysis with mutational analysis of 91 glioblastoma tumours in 2008 [57]. Data has been steadily released over the last 8 years for all major subtypes and all data is publically available via resources such as the cBio portal [14]. The sequencing phase of the TCGA project has now been completed. On completion the glioblastoma dataset consisted of mutational data for 291 cases and gene expression data for over 500 cases [58]. Data from these studies have already identified important driver mutations. For example analysis of brain tumour subtypes reveal approximately 50% possess an IDH1 driver mutation [59]. Whilst there is currently no targeted therapy to exploit IDH1 mutations, the observation that IDH1 mutation positive tumours carry better prognosis allows meaningful stratification of cases to aid clinical decisions in individual patients [59]. NGS has aided the segregation of seemingly homogenous tumour groups into distinct genetic subtypes defined by signalling pathway aberration. Using the genomic profile of 2000 breast cancers, 10 distinct genetic subtypes have been identified that can be used to predict sensitivity to chemotherapy [60, 61]. Currently the most challenging aspect of the practice of clinical oncology is deciding which patients should receive certain toxic treatments and which should receive milder treatments. Enhanced classification of tumours using the genetic principles discussed above greatly improves this decision

making to the benefit of cancer patients. In addition to patient stratification many novel clinically targetable driver genes have been identified and it is expected that over the next few years these discoveries will appear in the clinic. For example in our own group we have discovered using the TCGA dataset that a proportion of non-small cell lung cancer tumours carry ABL1 mutations that may be targeted with existing ABL inhibitors [62].

NGS has also facilitated conceptual shifts in our understanding of tumour biology that would not have been possible without it. The most notable example in this regard is that of tumour heterogeneity. In a landmark study Charles Swanton's group used NGS to sequence multiple regions of the same tumour and show distinct variability in the mutational profile [63]. Some mutations were present in all samples (truncal mutations) and some only present in a subset of often spatially related regions (branch mutations). It is now accepted that, by virtue of their appearance in all regions of the cancer, the truncal mutations occur early in the development of the tumour offering a temporal element to our understanding of tumour biology. Darwinian evolutionary theories of cancer development have been proposed that are analogous to the evolution of species, with distinct clones of cancer cells competing for survival within challenging tissue ecosystems [64]. Sequencing of multiple metastatic deposits from the same patient has allowed complex mapping of these lesions providing insight into the progression of disease and development of therapeutic resistance [65]. Another benefit of NGS is that it is now possible to perform non-invasive monitoring of therapeutic resistance mechanisms. The technology currently allows whole exome sequencing of a single cell or circulating tumour DNA. Combining this approach with advances in circulating tumour cell or circulating tumour DNA extraction allows rapid evaluation of the mutational and expression profile of the tumour at different stages of clinical management using a simple blood test [66].

NGS technology has also assisted in the identification of mechanisms responsible for carcinogenesis in addition to nucleotide sequence variation. Technological advances in NGS platforms and algorithms now allow quantitative analysis for each DNA fragment to provide accurate assessment of copy number [67]. Copy number variations (CNV) are defined as areas of the genome 1 kb or greater that are present at a variable copy number in comparison with a reference genome[68]. CNVs causing the duplication or deletion of whole genes are an important mechanism in carcinogenesis [69]. Most often the mechanism involves deletion of one copy

causing loss of heterozygosity and reliance on the second allele, which may have been inactivated via somatic mutation or inherited variant. Copy number variation can occur for shorter DNA regions when only a portion of the gene is duplicated or lost. This can result in fusion or truncated proteins that may have an oncogenic role. The detection of cancer CNVs by NGS is dependent on a number of factors including heterogeneity of the sample (including normal tissue contamination), quantitative algorithm used, and the efficiency of NGS in sequencing GC-rich regions (as difficult to sequence regions will have a lower read depth which will obscure attempts to determine copy number from read density) [70]. The repetition of smaller segments of DNA less than 1kb also causes challenges for NGS. It is estimated that approximately half of the genome is comprised of repetitive sequence [71]. Repetitive regions longer than the sequencing read length cause problems for NGS because sequence similarity between different regions of the genome mean that the read data cannot be accurately aligned. Small variations or mutations within repetitive regions may not be called if incorrectly mapped to another similar area of the genome. Conversely false-positive variants arise due to incorrect mapping of a normal sequence to a region with high but not complete similarity. The major solution to these problems is to increase read length so that enough unique sequence is provided in each read to allow adequate alignment. As NGS has improved, read length has increased leading to enhanced mapping and sequence analysis but this remains a limitation of the technology [71].

Finally it would be remiss in any discussion about the current status of cancer research to fail to mention immunotherapy, which has enabled further improvements in cancer survival, often providing more durable responses than targeted therapies [72]. Whilst the main targets of current cancer immunotherapy, CTLA4, PD-1, and PD-L1 were discovered in the pre-NGS era [73, 74], NGS is increasingly being used to understand the mechanisms involved. For example NGS was essential for the discovery that tumours with higher somatic mutation burdens respond better to immunotherapy drugs, presumably because a higher mutation load allows greater differentiation between cancer cells and self by the immune system [75]. NGS is also being used to unpick the mechanisms by which cell surface markers such as CTLA4 and PD1 are over-expressed on specific tumours by allowing the discovery of co-expressed genes or aberrant pathways [76].

1.4 Summary

The field of cancer genomics has revolutionised they way we think about cancer and greatly improved the therapeutic repertoire available to offer patients improved survival for relatively good toxicity profiles. Vemurafenib is the poster child for this approach because of the dramatic improvement it caused in a previously untreatable cancer. The development of NGS technologies has exponentially increased the discovery of drivers of carcinogenesis but also changed the way researchers are tackling the cancer biology puzzle. Sequencing data is now used to select targets for further biochemical characterisation and validation whereas previously it was used to explain biochemical observations (such as why EGFR inhibitors are effective in only a proportion of patients). This rearrangement of the pipeline means there is a large emphasis placed on in silico analysis prior to biochemical studies. Challenges persist because of the large number of mutations seen in each tumour and the fact that most of these mutations do not cause a biochemical effect. These 'passenger' mutations create noise that impedes the ability to identify the actual 'driver' mutations. Some believe that the solution is to sequence more tumours to uncover lower frequency drivers and that 600 – 5000 samples per tumour type are required [77]. Others suggest that we should stop sequencing more samples and research money could be better spent on functional studies [78].

Regardless of whether we should continue to invest in obtaining more sequencing data it is certainly true that there are already copious amounts of valuable data ripe for extracting clinical targets if solutions can be provided to aid interpretation and target identification. The overarching theme of this thesis is the use of novel in silico approaches to better analyse existing NGS data to discover novel driver genes. Chapter 2 is a review article published in the journal Pharmacogenomics that describes the passenger mutation problem in more depth along with other challenges of interpreting older NGS data. Chapter 3 is a final introductory chapter that introduces some biochemical and clinical considerations that we have applied to NGS datasets to identify novel driver genes. Chapter 5 is a research paper published in the journal Cancer Research in which we analysed whole exome sequencing data from the Cancer Cell Line Encyclopaedia (CCLE) and Catalogue of Somatic Mutations in Cancer (COSMIC) to identify further opportunities for driver mutation discovery in publically available datasets. Chapter 6 and 7 are research papers in preparation for submission in which we have used the functional and clinical considerations discussed in Chapter 3 to extract novel driver genes from NGS

datasets. Finally, Chapter 8 summarises the approaches used, the novel driver genes identified and looks at how modifications of our approaches can be used to make the most of this precious repository of cancer data.

Chapter Two

Introduction: Paper 1

Using large-scale genomics data to identify driver mutations in lung cancer: methods and challenges

Pharmacogenomics 2015: 16(10); 1149-60

Review

For reprint orders, please contact: reprints@futuremedicine.com

Pharmacogenomics

Using large-scale genomics data to identify driver mutations in lung cancer: methods and challenges

Andrew M Hudson[1], Christopher Wirth[2], Natalie L Stephenson[1], Shameem Fawdar[3], John Brognard*[,1] & Crispin J Miller*[,2]

[1]Signalling Networks in Cancer Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, M20 4BX, UK

[2]RNA Biology Group & Computational Biology Support Team, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, UK

[3]ANDI Centre of Excellence for Biomedical & Biomaterial Research, University of Mauritius, Reduit, Mauritius

*Authors for correspondence:
Tel.: +44 161 3065301;
John.Brognard@cruk.manchester.ac.uk;
Crispin.Miller@cruk.manchester.ac.uk

Lung cancer is the commonest cause of cancer death in the world and carries a poor prognosis for most patients. While precision targeting of mutated proteins has given some successes for never- and light-smoking patients, there are no proven targeted therapies for the majority of smokers with the disease. Despite sequencing hundreds of lung cancers, known driver mutations are lacking for a majority of tumors. Distinguishing driver mutations from inconsequential passenger mutations in a given lung tumor is extremely challenging due to the high mutational burden of smoking-related cancers. Here we discuss the methods employed to identify driver mutations from these large datasets. We examine different approaches based on bioinformatics, *in silico* structural modeling and biological dependency screens and discuss the limitations of these approaches.

Keywords: cancer genomics • challenges • driver mutation • genetic dependency screen • *in silico* analysis • lung cancer

Lung cancer is the most common cause of cancer death in the world; only 16.8% of patients survive to 5 years following a diagnosis of lung cancer [1]. This is in stark contrast to prostate cancer (98.9% surviving to 5 years) and breast cancer (89.2% surviving to 5 years). A major reason for this disparity is that metastatic disease is diagnosed at presentation in the majority of lung cancer cases. In addition, the median age of lung cancer diagnosis is around 70 years, and patients have often smoked for a large period of their life, making successful treatment of lung cancer patients extremely challenging. As a consequence of smoking, many patients possess severe co-existing medical conditions that preclude them from receiving potentially toxic chemotherapeutic regimens. These patients cannot receive an active anticancer treatment and are only eligible for symptomatic palliation. Further, while some patients will benefit from palliative chemotherapy to extend survival, this is often short-lived and accompanied by toxic side effects. Therefore, the promise offered by targeted therapies, with their better-tolerated side effects, is of particular significance for lung cancer patients.

Most clinically effective targeted therapies rely on disruption of 'oncogene addiction' that occurs through genetic mutation or overexpression of genes conferring tumorigenic properties in line with the hallmarks of cancer [2,3]. The success of targeted precision therapies lies in identifying mutated genes that confer a growth or survival advantage (driver mutations) that can be subsequently targeted therapeutically. There have been some notable successes with this approach. EGF receptor (EGFR) inhibitors were first introduced into the clinic for the treatment of non small-cell lung cancer (NSCLC). The IPASS study compared the EGFR inhibitor gefitinib with a standard doublet chemotherapy regimen in patients from East Asia with first-line advanced lung adenocarcinoma [4]. It showed superior progression-free survival (PFS) in the gefitinib arm as well as lower rates of severe toxicity. While the study did not stratify treatment based on

Future Medicine part of fsg

EGFR mutation status, subgroup analysis showed that EGFR mutation positive patients had longer PFS with gefitinib while EGFR mutation negative patients had longer PFS with standard chemotherapy. The study population, consisting of East Asian never or light-smokers, was enriched for EGFR mutations (59.7%) compared with the heavy smoking population that forms the majority of Western lung adenocarcinoma cases (11% with EGFR mutation) [5]. Subsequent trials of gefitinib, erlotinib and afatinib have demonstrated superior PFS in EGFR mutation positive patients when compared with standard chemotherapy leading to these agents being routinely used for treatment in EGFR mutation positive patients [6–9]. More recently, *ALK* rearrangements have been identified in approximately 5% of NSCLC [10]. Again patients with *ALK* rearrangements are more likely to be never/light smokers [11]. Crizotinib, a small-molecule inhibitor of ALK (as well as MET and ROS1 kinases), has been shown to offer improved PFS, lower toxicity and better quality of life compared with standard chemotherapy [12].

Other known oncogenes have been discovered to be mutated in a proportion of NSCLC cases and are currently being assessed in early phase clinical trials. The most commonly mutated gene is *KRAS* (approximately 25% depending on histology and more frequent in heavy smokers) [13]. KRAS itself is not easily targetable and clinical trials have been developed using MEK inhibitors in combination with chemotherapy to block the downstream effects of oncogenic KRAS [14]. This approach has seen some encouraging responses in KRAS mutation positive patients but phase 3 data are awaited and MEK inhibition on its own may not be sufficient in these patients given the multiple downstream effectors of KRAS. An additional downstream target of GOF mutant KRAS is PI3K (phosphoinositide 3-kinase), where activation of PI3K leads to PIP3 (phosphoinositide (3,4,5)-trisphosphate) mediated AKT activation to promote cancer cell survival [15]. It would be expected that combination PI3K/MEK inhibitor therapy would promote tumor regression, however KRAS mutation positive colon cancer PDX models failed to demonstrate tumor regression highlighting the challenges in treating KRAS mutation positive cancers [16]. BRAF is mutated in approximately 3% of patients (with half of cases being the V600E mutation that have been targeted to much success in melanoma) and early phase trials are taking place with BRAF inhibitors [17]. However, the poor clinical response to V600E BRAF inhibition due to EGFR activation in colorectal tumors adds caution to any predictions of efficacy in lung cancer [18]. *HER2* amplification and activating mutations are seen in a proportion of NSCLC but clinical targeting with trastuzumab and lapatnib have not shown the efficacy seen in *HER2* amplified breast cancer patients [19,20]. Other genetic alterations such as *MET* amplifications (8–10%), *RET* rearrangements (1–2%) and gain-of-function mutations in PIK3CA/AKT (2–5%) are being targeted in early phase clinical trials [21].

A biomarker-based precision medicine trial (known as the BATTLE trial) used tumor biopsies to stratify lung cancer patients into different treatment arms based on up-to-date mutational profiling, demonstrating early disease control for certain biomarker–drug combinations and highlighting that this approach is feasible [22]. However, response rates were poor due to a heavily pretreated patient population combined with a lack of identifiable driver mutations and inhibitors for treatment. Overwhelmingly the most important factor in preventing this personalized approach for lung cancer is the lack of identifiable driver mutations. It was recently estimated that at least three driver mutations are required for the development of lung cancers [23]. However, despite the whole exome sequencing of hundreds of lung cancer samples, approximately 50% of NSCLC have no identifiable activating mutations [24]. In this review we shall discuss the different approaches and challenges to mining cancer genomics data to discover druggable driver mutations in lung cancer.

Online aggregated cancer genomics data

To enhance driver mutation discovery, large repositories of cancer genomics data have been published online. Aggregating the data from large numbers of sequenced cancers will aid in the discovery of commonly mutated genes for specific cancer subtypes. cBio is one of the most widely used databases, combining data from The Cancer Genome Atlas (TCGA) samples with other large studies including The Cancer Cell Line Encyclopedia (CCLE) [25,26]. Users can search by gene name to retrieve the frequency of mutation in different cancer subtypes and identify novel targets to evaluate further. The most straightforward approach is to seek genes that are commonly mutated in a large proportion of cancers so that effective and financially viable drug development can be undertaken for that target. However, this approach requires a strategy to distinguish somatic mutations that drive the oncogenic process (driver mutations) from somatic mutations that do not have a functional effect on the cell (passenger mutations) [27]. In lung cancer, this is particularly challenging. Tobacco smoke contains a multitude of powerful carcinogens that form DNA adducts resulting in much higher mutational rates than a majority of other cancers [28,29]. The high mutational burden that yields large numbers of passenger mutations make it difficult to identify the driver mutations amongst

this background of numerous inconsequential mutations [30]. At the time of writing, the average number of protein coding mutations identified in lung squamous and lung adenocarcinoma TCGA samples was 319 and 280, respectively. This is 11–13-times greater than acute myeloid leukemia (AML), for which only 24 protein-coding mutations were reported per sample. In AML, the relatively low mutation burden has aided the discovery of more drivers. A study of 200 AML samples, for example, found that 99.5% possessed a nonsynonymous somatic mutation in a gene of biological significance [31,32]. This is in marked contrast to NSCLC in which only approximately 50% of cases have a known activating mutational driver [24].

Given the relative failure to identify many common mutations in lung cancer it is likely that unknown cases are characterized by small groups of drivers, each accounting for 1–2% of the total. These drivers, affecting only a small proportion of patients can still be beneficial to pursue. This was evident in an expanded Phase 1 study of crizotinib, in which some patients with ROS1 rearrangements (only 1–2% of NSCLC patients) had dramatic responses to the drug with an overall longer median progression free survival than EML4-ALK patients receiving the same drug [33,34]. Given the high incidence of lung cancer, these targets that involve 1–2% of cases, represent a large global patient cohort.

Playing the numbers game

While cancer genomics data can be interrogated for genes with frequent mutations that segment according to histological type, it is not sufficient to assess statistical significance on the assumption that mutation rates for all genes are the same. Instead, more sophisticated models consider additional factors: first, gene-level mutation rates may be normalized according to length because when assuming a uniform background mutation rate, longer genes are more likely to acquire a mutation than shorter ones. Therefore, extremely long genes such as *TTN* have a high mutational frequency (52% in lung squamous TCGA data). Table 1A lists the top 20 most frequently mutated genes in squamous lung cancer. Seven of these encode proteins in the top 20 longest proteins. The median protein length for the top 20 mutated is 4612 amino acids (mean for all proteins screened = 699 amino acids). While this effect is noted, it does not mean that very large proteins do not play a role in oncogenesis and indeed some propose that genes such as TTN may be an important driver of tumor progression [35]. The length of these proteins obviously makes subsequent biological work more challenging and researchers may shy away from the associated technical challenges in search of lower hanging fruit.

Second, another issue that is particularly apparent in samples with a high mutation burden is that many mutations occur in genes that are not expressed as proteins in the given tumor. This is caused, at least in part by the differential effect of transcription-coupled repair; genes that are not expressed are less likely to be repaired and therefore mutations at these loci accumulate [36,37]. Following length correction the top 20 mutated genes (Table 1B) now contain a number of genes, such as olfactory receptors, that are not expressed in squamous lung cancer, and are unlikely to have a functional effect. The length corrected top 20 also contain very small proteins with a small number of incidental mutations as well as larger proteins with biological evidence of significance in lung cancer [38,39].

Third, replication timing is also important, since genes that replicate late will have a depleted pool of nucleotides available, and are therefore more likely to acquire mutations [36]. This phenomenon has been used to explain increased germline variability and somatic mutations in late-replicating regions [36,40,41]. This knowledge may help to explain the high density of mutations in a specific locus, but, as with issues of gene length, it does not rule out the possibility that a late-replication gene might play a significant role in cancer. Replication timing and expression have been used to develop the MutSigCV platform to better identify driver mutations using estimations of the background mutation rates in different cancer types using silent and noncoding mutations in a genetic region [36]. However, it is acknowledged that larger amounts of next generation sequencing is required to get a better picture of local mutation rates and improve the method.

Fourth, intratumoral mutation heterogeneity in NSCLC primary tumors has been demonstrated in two studies [42,43]. Zhang *et al.* showed that 76% of mutations were identified in all regions of individual tumors (including 20 out of 21 known cancer gene mutations) suggesting that analysis of primary tumor genomes will capture most driver mutations. However, it is not known if metastatic lesions that make up a large proportion of clinical presentations and have the most to gain from targeted systemic treatments, share this degree of homogeneity. While the TCGA dataset is comprised of primary tumors, cell lines are frequently derived from metastatic tissue; these differences need to be considered when pursuing candidate mutations in experimental systems based on tumor-derived cell lines. It is also important to consider whether previous lines of therapy have led to the selection of specific genetic clones. The fact that the primary tumors in Zhang *et al.* had especially good concordance for known cancer causing mutations suggests that select-

10.2217/PGS.15.60

Table 1A. Top 20 frequently mutated genes in squamous lung cancer with number of mutated cases and longest corresponding protein length.

Gene	Number mutated cases (out of 178 cases)	Longest protein length (amino acids)
TTN	92	35991
CSMD3	55	3707
RYR2	53	4967
ZFHX4	53	3616
MUC16	51	14507
LRP1B	48	4599
USH2A	43	5202
SYNE1	35	8797
RYR3	34	4873
FLG	31	4061
DNAH5	29	4624
PKHD1	29	4074
MUC17	29	4493
MUC5B	28	5762
AHNAK2	28	5795
SI	28	1827
FAM135B	28	1406
KMT2D	27	5537
HCN1	27	890
CSMD1	27	3565

Seven of these top 20 mutated genes encode proteins in the top 20 longest proteins in the TCGA dataset.

ing mutations with high allele frequency (reported in TCGA data) may be a beneficial strategy. However, uncertainties about the sampling and proportion of normal tissue contamination make this difficult.

Fifth, discrepancies between NGS datasets of the same samples highlight further challenges and opportunities with large-scale genomics data. Two prominent cancer genomics institutes (The Broad Institute [CCLE] and The Sanger Institute [COSMIC]) have published NGS data of commercially available cancer cell lines [26,44]. This work has been extremely beneficial to researchers around the world who have been able to use the data to select relevant cell lines in which to test their hypotheses. We had observed some inconsistencies between the two datasets, leading us to perform a formal comparison of the missense mutations reported by the two different institutes [45]. We demonstrated marked discrepancies between the datasets with only 57% of the mutations reported across 568 cell lines being concordant. We found that one of the major reasons for this discordance was that GC-rich areas of the exome are still proving difficult to sequence by NGS, leading to over 400 significant areas of poor sequencing (cold-spots) in known cancer causing genes and kinases. A conservative estimate suggested that approximately three missense mutations occurring within known cancer census and kinase genes were being missed in each cell line (with many other mutations being missed at other loci). TCGA data are generally of a similar age and obtained with similar technologies, suggesting that these data may also suffer from cold-spots. GC-rich cold spots are more relevant in cancers, such as lung cancers, where mutations occurring more frequently in guanine nucleotides. It is likely, therefore, that mutations in genes with GC rich regions are under-reported in lung cancer, and suggests that these loci harbor additional common mutations that have yet to be identified.

Sixth, another source of disparity between the datasets we studied was poor consensus in the labeling of variants as either germline or somatic. The majority of cell lines do not have paired normal tissue for comparison, and therefore the somatic status of an observed variant was performed by matching to databases of known germline variants. The most common method employed

Table 1B. Top 20 frequently mutated genes in squamous lung cancer normalized for length of longest protein and ranked by length corrected score.

Gene	Length corrected score	Longest protein length (amino acids)	Number of mutations	Comments
CDKN2A	0.152941176	170	26	Evidence of role in lung cancer [38]
REG3A	0.057142857	175	10	
REG1B	0.054216867	166	9	
REG1A	0.054216867	166	9	
KRTAP19–3	0.049382716	81	4	Hair cortex – keratin-associated protein
COX7B2	0.049382716	81	4	
OR5D18	0.044728435	313	14	Olfactory receptor
SST	0.043103448	116	5	
SPANXN1	0.041666667	72	3	Sperm protein associated with the nucleus
OR2T4	0.040229885	348	14	Olfactory receptor
REG3G	0.04	175	7	
OR6F1	0.035714286	308	11	Olfactory receptor
OR5L2	0.035369775	311	11	Olfactory receptor
MANBAL	0.035294118	85	3	
PRAC1	0.035087719	57	2	
NFE2L2	0.033057851	605	20	Evidence of role in lung cancer [39]
LENEP	0.032786885	61	2	
TPTE	0.032667877	551	18	
STATH	0.032258065	62	2	
SLN	0.032258065	31	1	

Proteins previously demonstrated to play a role in lung cancer feature in the list (CDKN2A, NFE2L2) as well as proteins such as olfactory receptors that are unlikely to be expressed in lung cancer and are therefore less likely to undergo transcription coupled repair of somatic mutations. The length correction means that very small proteins with a small number of incidental mutations are also represented.

simply removes all variants with an 'rs' oasis: entry in the dbSNP database [46]. Unfortunately, dbSNP database is rapidly evolving and expanding, with the side effect that the date at which the filtering was performed can greatly affect the final output. Further, database submission is unrestricted leading to the occasional inclusion of a somatic mutation in error. Finally, high GC content can also lead to underreporting of germline variants in hard-to-sequence loci as well as bona fide mutations. As NGS technology improves, common germline SNPs in these regions can be uncovered, but since earlier germline sequencing approaches failed to identify them, they are not filtered against dbSNP leading them to be erroneously reported as rare or somatic.

From numbers to predictions: *in silico* analysis of mutations

Another method used to distinguish between driver and passenger mutations is to consider the structural impact of the resultant amino acid substitution. Tools such as mutationassessor.org, Polyphen2, Provean and SIFT are freely available online, and use information such as protein structure and sequence homology to predict whether a mutation might have a functional impact [47–50]. They can be used to quickly analyze large batches of genomics data to allow researchers to assess whole exome data to select those mutations most likely to alter the function of the protein. A recent evaluation of these different tools demonstrated that they worked well to distinguish known pathogenic mutations from neutral ones, and that predictive power could be further enhanced by combining the outputs from multiple tools [51]. From our experience, loss-of-function mutations (with their presumed greater structural disruption) are more likely to be identified than some more subtle activating oncogenes using these methods [52]. Additional complexity arises because the majority of human protein-coding genes express more than one isoform, with the result that a missense mutation can have different effects according to the

isoform it occurs in, and may not be present at all if it occurs in a spliced exon. Furthermore, proper interpretation of the data is difficult without access to protein expression data, since highly functional mutations will clearly not have an effect if that protein or mutant allele is not expressed. One useful source for these data is the Human Protein Atlas, which provides a valuable online resource in which immunohistochemistry data are used to catalogue the expression of proteins in different cancer types [53].

The challenge of identifying driver mutations is well summarized by Tamborero and colleagues, who state that '*The elucidation of cancer drivers relies on identifying the marks of positive selection that occur during the clonal evolution of tumors*' [54]. These positive marks present themselves in various ways, from the clustering of mutations in a specific protein domain to the correlation of a mutated gene with a specific sub-phenotype present within the patient. Therefore state-of-the-art attempts at driver mutation identification combine a wide breadth of different data to identify the marks of positive selection. One such package is 'MuSiC' which combines statistical tests of mutational frequency and co-occurrence, clinical data, and information pertaining to the frequency of mutations in specific protein domains [55]. Identifying mutations clustered at specific loci (whether using a focused approach based on known protein domain function or just identifying mutations in close proximity to another) can highlight potential mechanisms of positive selection. Another analysis focused on mutations only occurring in phosphosites of proteins to extract novel targets [56]. Increasing understanding of protein domain function, from wet-lab studies, will provide further opportunities to create functionally relevant screens.

Correlating mutational data with copy number deletions, immunohistochemistry, and considering the frequency of truncating mutations may assist in the prediction of loss-of-function mutations. However these cannot be solely relied upon given the effects of co-existing mutations and expression causing varying redundancy and unknowns regarding the presence or absence of a dominant negative effect [57].

If greater computational resources are available, molecular dynamic (MD) simulations can provide more in depth *in silico* approaches with which to assess the effect of a given mutation. MD simulations model the movement of a protein over very short times scales (generally in the nanosecond range), making it possible to predict the structural variations that occur as a consequence of a mutation.

Initial MD simulations are kept relatively short, simulating up to 50 ns of time, due to the large computational burden required for such simulations. These typically focus on biochemically significant mutations that have either been published previously [58,59] or are analyzed *in vitro/in vivo* within the study itself [60,61]. These short simulations are generally utilized to confirm biochemical data and gain further understanding of the structural consequences of the identified mutation. As these simulations are only short, the information gained can range from as little as the position of the mutations in relation of other regions of the protein [60] and movement of key regions of the structure [61] up to changes in binding affinities of key substrates [59]. All of this information can help to understand the potential impact the mutation is having on the protein in question.

More information can be gleaned from longer MD simulation studies. Many longer simulations focus on the alteration of key structural features (e.g., salt bridges, domain or feature orientation and drug binding) and how these affect the free energy landscape of the protein [62–65]. These types of studies have provided information on the progression of these proteins into more active conformations following disease causing mutations [64,66] as well as critical information on the effect of mutations on drug resistance [65,66]. Such information is critical to understanding the structural effects occurring following mutation and providing the research community with this type of analysis could aid in drug development. However, in order to perform these MD simulations, a crystal structure of the protein is required. Furthermore, these more complex *in silico* methods require both a high level of computational knowledge and a large computational resource.

Genetic dependency screens

'Oncogene addiction' describes a phenomenon whereby cancer cells develop a dependency on a specific oncogene that has become either overexpressed or activated by mutation during the development of the cancer. This dependency leaves the cancer vulnerable should the activated oncogene be inhibited or suppressed. A number of mechanisms have been proposed to explain how dependency on a single oncogene occurs in tumors with a high burden of genetic mutations [67]. This dependency creates a desirable differential between the normal and cancer cell that can be exploited and targeted for therapeutic intervention. Most clinically valuable targeted treatments owe their beneficial effect to disrupting the 'oncogene addiction' of a cancer cell to a mutated gene that drives cellular growth or survival. In fact it has been postulated that '*most, if not all, dramatic responses of 'tumor shrinkage' following molecularly targeted therapy result from the acute inactivation of an activated oncoprotein upon which the tumor cells became dependent*" [68]. Substantial effort is now

being channeled into discovering more of these oncogenes against which small molecule inhibitors can be developed for cancer treatment.

Genetic dependency screens aim to exploit the phenomenon of oncogene addiction in a high-throughput manner using small interfering RNA against multiple targets and assessing the functional outcome in the cancer cell [69]. Commercially available si/shRNA libraries mean the method is now widely used by groups investigating novel drivers of oncogenesis leading to novel target discovery [70–75]. Project Achilles is a huge project from The Broad Institute initially undertaking the silencing of thousands of genes with shRNA in hundreds of cell lines [76]. This has yielded some novel targets in different cancer subtypes [77–79].

Rather than performing a genome-wide study, we developed a targeted approach using siRNA to knockdown only those genes harboring somatic mutations in specific lung cancer cell lines and assessed the effects on proliferation and cell survival [61]. In three out of six of these cell lines we discovered novel gain of function mutations that were subsequently validated. One benefit of this approach is a clear endpoint, where the genes identified to harbor potential gain-of-function mutations, can then be generated in the laboratory and tested to determine if the mutation does in fact increase the catalytic activity of the protein. Alternatively the cancer mutant can be expressed in cells to determine if the mutant allele may promote increases in survival or proliferation by more subtle mechanisms such as altered cellular localization or differential substrate specificity. The mutant genes we identified (*PAK5*, *FGFR4* and *MAP3K9*) all activated the MEK-ERK pathway and the respective cell lines had increased sensitivity to MEK inhibitors. Analyzing lung adenocarcinoma TCGA data reveals that the frequencies of cases with mutations in these genes are: *PAK5* (11%), *FGFR4* (5.2%) and *MAP3K9* (4.7%). Inputting these three mutations into different online mutation assessors often predicts the *FGFR4* and *MAP3K9* mutations as unlikely to be pathogenic (Table 2). This highlights the benefit of using a targeted siRNA screen to identify novel drivers that would otherwise not be predicted to be pathogenic.

While the *MAP3K9* mutation in the H2009 cell line was reported by CCLE, it was not reported by an earlier version of COSMIC that also sequenced the cell line. Therefore, using just COSMIC data would have missed this gain-of-function mutation. This demonstrates how a targeted genetic dependency screen relies on high quality genomics data to ensure all genes mutated in a sample are silenced. The issue is also important for nontargeted genetic dependency screens such as Project Achilles as well as pharmacogenomics screens. If the mutational data from these cell lines are not complete it hinders attempts to interpret the phenotypic response of the cell to the knockdown or inhibition of a specific gene.

Interestingly, siRNA knockdown screens of the remaining three cell lines in our study did not identify a stand out driver mutation in terms of cellular proliferation. It remains to be seen whether this is due to no mutational drivers present in the cell lines or that a mutation is present but being missed by inadequate sequencing of GC rich regions. Another consideration is that the driver mutations in these cell lines may exert their effect through loss-of-function mechanisms and a targeted screen like ours will not identify these. Synthetic lethality describes a mechanism by which tumor cells often become more dependent on a gene than a normal cell due to gain or loss of function of a different gene during the development of the cancer [67,80,81]. Therefore, genome-wide knockdown, such as that used in the Achilles project, can be utilized to identify synthetically lethal genes that may be druggable. In addition, it is now possible to perform high-throughput screens for tumor suppressor genes using CRISPR/CAS technology [82].

Conclusion & future perspective

We highlight the challenges of using cancer genomics data aggregated from a large number of samples to identify driver mutations. Given the urgent need for effective targeted therapies against lung cancer, it is important to develop solutions to tackle this problem. The first step is to identify commonly mutated signaling pathways against which to develop targeted therapies. The three clinically successful targets mentioned in this review

Table 2. Mutation predictor data for the three pathogenic mutations discovered with a targeted siRNA screen [61].

Target	Cell line	Mutation	Provean (cut-off = -2.5)	Sift (cut-off = 0.05)	Mutationassessor.org	Polyphen2
FGFR4	H2122	P712T	Neutral (-0.09)	Tolerated (0.242)	Low functional impact	Possibly damaging
MAP3K9	H2009	E179K	Neutral (-2.21)	Tolerated (0.085)	Low functional impact	Possibly damaging
PAK5	H2087	T538N	Deleterious (-2.92)	Damaging (0.045)	Neutral	Probably damaging

The FGFR4 and MAP3K9 mutations would be classified as nonpathological by three out of four of these assessors.

10.2217/PGS.15.60

(EGFR mutation, ALK rearrangements and ROS1 rearrangements) all occur predominantly in nonsmokers. Therefore smokers, often with multiple co-morbidities, lack targeted therapy options. Unfortunately, the mechanisms by which cigarette smoke causes cancer mean that smoking related tumors have a high mutational burden with many passenger mutations. Figure 1 illustrates the sources of potential bias in the analysis of aggregated genomics data. These issues become more problematic when mutational noise is increased. Gene length, expression level and replication timing all have the potential to distort mutation frequencies making it harder to identify driver mutations. We have shown how the ability to sequence difficult regions is improving with technological advances and that older data may be susceptible to bias due to sequencing cold-spots. It is obviously preferable to obtain matched normal tissue samples for comparison and, when this is not possible, inconsistencies in dbSNP reporting will impair the ability to identify driver mutations. Since germline data are obtained from a wide range of sources and situations, it can never be considered as reliable as matched normal tissue. A recent study has demonstrated that false-positive calling of actionable mutations is significantly increased without normal tissue control [83]. Noncoding RNA and intronic mutations are not discussed here but present additional challenges.

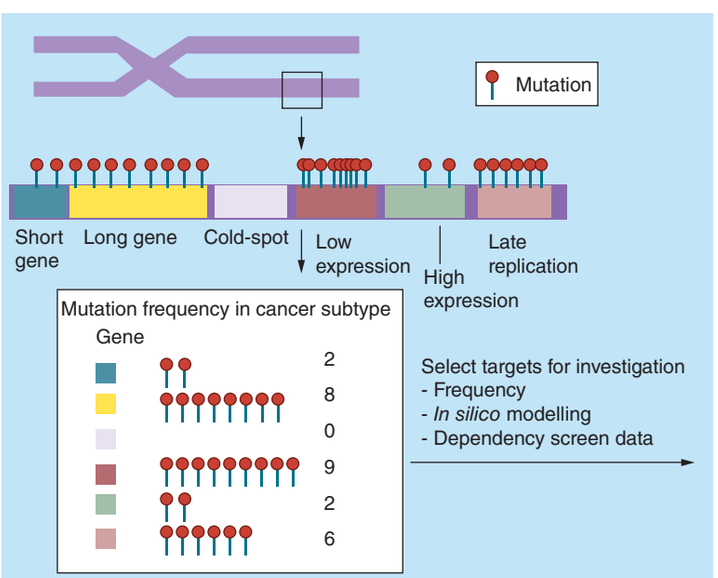


Figure 1. Summary schematic highlighting the factors leading to biases and heterogeneity of mutational data. In areas of uniform mutational rates longer coding genes will demonstrate a higher mutational frequency if the data are not length corrected. Genes that are not expressed and those that replicate late in the cell cycle will have higher mutational rates. Genes with large GC-rich regions will have inadequate sequencing coverage and potential mutations will be missed leading to an underreporting of mutations in these genes.

Structural analysis has the potential to predict functional outcomes on a protein with a high degree of accuracy, but is currently unable to model how complex interactions between proteins within a cancer cell are disrupted as a consequence of co-occurring mutations, variability in gene expression and additional regulatory pathways involving, for example, miRNAs and other noncoding loci. The limitations of mutation assessors are illustrated by our siRNA screen data in which two of three activating mutations would be predicted to be neutral in most of the online mutation assessors. By contrast, potentially detrimental mutations supported by structural studies, are only of relevance if the protein is expressed in the tissue of interest.

Although the state of the art has advanced rapidly, biological confirmation of *in silico* results is still critical, and usually involves knockdown of the gene of interest with small interfering RNA (si/shRNA), combined with functional read-outs for proliferation, cell viability or apoptosis. It is possible to use these approaches to perform high-throughput studies, but limitations such as inadequate sequencing can make it hard in practice to associate observed sensitivity with mutation status, causing targets to be missed. Similarly large-scale inhibitor studies suffer the same limitations [84]. *In silico* modeling is enhanced by the greater understanding of protein structure and function provided by wet-lab studies. Better characterization of protein domain function allows genomic data to be filtered for mutations by areas of functional importance. The biochemical validation of the effects of a mutation also provides valuable information with which to improve the training datasets used to develop these prediction tools. Driver mutation discovery is, therefore, enhanced if there is a virtuous circle in which existing genomics data are reanalyzed in the light of recent functional studies in order to identify further targets for evaluation at the bench – which then support additional rounds of progressive refinement and analysis.

NGS technology continues to progress rapidly, improving the coverage of hard to sequence regions and, as costs decrease, allowing genomes to be sequenced at higher depths. Together, these are contributing to substantial increases in the number of mutations that can be reliably detected. Unfortunately, since the technology itself is unable to distinguish between driver and passenger mutations, these advances come with the challenge of increased levels of mutational noise. The advances in genome profiling are, therefore, increasingly dependent on concomitant improvements in the techniques used to identify actionable mutations. While sequencing of cancer genomes will continue to

accumulate, research efforts should shift to functional genomics to aid in the elucidation of novel drivers so that lung cancer patients can benefit from targeted therapies, likely in combination with immunotherapies. Pinpointing these drivers and targeting them with precision medicines will portend a future where lung cancer patients will be treated with therapies that extend survival while preserving quality of life.

Financial & competing interests disclosure

This work was solely funded by Cancer Research, UK. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Online aggregated cancer genomics data

- Large-scale cancer genomics programs such as TCGA, COSMIC and CCLE have been instrumental in discovering novel mutational drivers of cancer in many cancer subtypes.
- In lung cancer there have been some notable clinical successes in targeting these drivers (EGFR mutations, ALK translocations, ROS translocations). However, the majority of lung cancer patients who benefit from these treatments are never/light smokers.
- Lung cancer associated with smoking is characterized by a high mutational burden and the main hindrance to target discovery is identifying the few driver mutations from hundreds of inconsequential passenger mutations in each sample.

Playing the numbers game

- Using mutational frequency in hundreds of cancer samples to select targets for further investigation is a powerful way to discover common mutations but certain considerations should be made.
- Gene length will influence frequency results to over-represent longer proteins if a length correction is not made.
- Genes with low expression and/or late replication timing will have a higher mutational rate.
- Poor sequencing of GC-rich regions (sequencing cold-spots) will lead to under-reporting of mutations and these cold-spots may be hiding potential high frequency mutations.
- Germline filtering without normal tissue comparison can introduce error.

From numbers to predictions: *in silico* analysis of mutations

- Online mutation prediction programs can be used in a high-throughput manner to analyze whole exome data to select functional mutations.
- Identifying marks of positive selection are fundamental to extracting the driver mutations. These marks are observed in a broad spectrum of data and combining different analyses will likely yield the most success.
- Molecular dynamics simulations provide a more detailed analysis of the structural ramifications of a given mutation but require a known crystal structure and a large computational resource.

Genetic-dependency screens

- si/shRNA screens exploit the oncogene addiction of cancer cells on a high-throughput scale to compare the functional effects of gene knockdown.
- We used a targeted screen to knockdown all mutated genes in specific lung cancer cell lines and discovered three novel mutational drivers of lung cancer (PAK5, MAP3K9, FGFR4).
- Incomplete genomics data (including sequencing cold-spots) and potential loss-of-function mutational drivers may explain why three of the cell lines tested did not have an identifiable driver mutations using the targeted siRNA screen.

Conclusion & future perspective

- Identifying driver mutations in lung cancer genomics data remains a large challenge and there is much opportunity to identify targetable mutations for the benefit of patients.
- The different methods detailed in this review have specific strengths and weaknesses and a combination of approaches is required to capture all driver mutations.
- As sequencing technology improves and becomes cheaper, the scale of mutational data will increase but this will also increase the amount of mutational noise.
- An increased focus on functional genomics is required to develop clinically effective precision medicines from the large-scale data.

References

1 Seer. Surveillance, epidemiology, and end results (seer) program: Seer 18 regs research data Nov 13 sub (2000–2011).

2 Weinstein IB. Addiction to oncogenes – the Achilles heel of cancer. *Science* 297(5578), 63–64 (2002).

3 Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 100(1), 57–70 (2000).

4 Mok TS, Wu YL, Thongprasert S *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* 361(10), 947–957 (2009).

5 Dogan S, Shen R, Ang DC *et al.* Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin. Cancer Res.* 18(22), 6169–6177 (2012).

6 Sequist LV, Yang JC, Yamamoto N *et al.* Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. *J. Clin. Oncol.* 31(27), 3327–3334 (2013).

7 Maemondo M, Inoue A, Kobayashi K *et al.* Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N. Engl. J. Med.* 362(25), 2380–2388 (2010).

8 Rosell R, Carcereny E, Gervais R *et al.* Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced egfr mutation-positive non-small-cell lung cancer (eurtac): a multicentre, open-label, randomised Phase 3 trial. *The Lancet Oncology* 13(3), 239–246 (2012).

9 Mitsudomi T, Morita S, Yatabe Y *et al.* Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (wjtog3405): an open label, randomised Phase 3 trial. *The Lancet Oncology* 11(2), 121–128 (2010).

10 Soda M, Choi YL, Enomoto M *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448(7153), 561–566 (2007).

11 Shaw AT, Yeap BY, Mino-Kenudson M *et al.* Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. *J. Clin. Oncol.* 27(26), 4247–4253 (2009).

12 Solomon BJ, Mok T, Kim DW *et al.* First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N. Engl. J. Med.* 371(23), 2167–2177 (2014).

13 Dearden S, Stevens J, Wu YL, Blowers D. Mutation incidence and coincidence in non small-cell lung cancer: Meta-analyses by ethnicity and histology (mutmap). *Ann. Oncol.* 24(9), 2371–2376 (2013).

14 Janne PA, Shaw AT, Pereira JR *et al.* Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, Phase 2 study. *The Lancet Oncology* 14(1), 38–47 (2013).

15 Luo J, Manning BD, Cantley LC. Targeting the PI3K-AKT pathway in human cancer: rationale and promise. *Cancer Cell* 4(4), 257–262 (2003).

16 Migliardi G, Sassi F, Torti D *et al.* Inhibition of MEK and PI3K/MTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clin. Cancer Res.* 18(9), 2515–2525 (2012).

17 Paik PK, Arcila ME, Fara M *et al.* Clinical characteristics of patients with lung adenocarcinomas harboring BRAF mutations. *J. Clin. Oncol.* 29(15), 2046–2051 (2011).

18 Prahallad A, Sun C, Huang S *et al.* Unresponsiveness of colon cancer to BRAF(v600e) inhibition through feedback activation of EGFR. *Nature* 483(7387), 100–103 (2012).

19 Gatzemeier U, Groth G, Butts C *et al.* Randomized Phase II trial of gemcitabine-cisplatin with or without trastuzumab in HER2-positive non-small-cell lung cancer. *Ann. Oncol.* 15(1), 19–27 (2004).

20 Ross HJ, Blumenschein GR Jr, Aisner J *et al.* Randomized Phase ii multicenter trial of two schedules of lapatinib as first- or second-line monotherapy in patients with advanced or metastatic non-small cell lung cancer. *Clin. Cancer Res.* 16(6), 1938–1949 (2010).

21 Califano R, Abidin A, Tariq NU, Economopoulou P, Metro G, Mountzios G. Beyond EGFR and ALK inhibition: unravelling and exploiting novel genetic alterations in advanced non small-cell lung cancer. *Cancer Treat Rev.* (2015).

22 Kim ES, Herbst RS, Wistuba Ii *et al.* The battle trial: personalizing therapy for lung cancer. *Cancer Discov.* 1(1), 44–53 (2011).

23 Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl Acad. Sci. USA* 112(1), 118–123 (2015).

24 Pao W, Girard N. New driver mutations in non-small-cell lung cancer. *The Lancet. Oncology* 12(2), 175–180 (2011).

25 Cerami E, Gao J, Dogrusoz U *et al.* The CBIO cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2(5), 401–404 (2012).

26 Barretina J, Caponigro G, Stransky N *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391), 603–607 (2012).

27 Haber DA, Settleman J. Cancer: drivers and passengers. *Nature* 446(7132), 145–146 (2007).

28 Hecht SS. Tobacco smoke carcinogens and lung cancer. *J. Natl Cancer Inst.* 91(14), 1194–1210 (1999).

29 Pleasance ED, Stephens PJ, O'meara S *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463(7278), 184–190 (2010).

30 Alexandrov LB, Nik-Zainal S, Wedge DC *et al.* Signatures of mutational processes in human cancer. *Nature* 500(7463), 415–421 (2013).

31 Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* 368(22), 2059–2074 (2013).

32 Chen SJ, Shen Y, Chen Z. A panoramic view of acute myeloid leukemia. *Nat. Genet.* 45(6), 586–587 (2013).

33 Shaw AT, Ou SH, Bang YJ *et al.* Crizotinib in ros1-rearranged non-small-cell lung cancer. *N. Engl. J. Med.* 371(21), 1963–1971 (2014).

34 Gold KA. ROS1– targeting the one percent in lung cancer. *N. Engl. J. Med.* 371(21), 2030–2031 (2014).

35 Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat. Methods* 10(11), 1108–1115 (2013).

36 Lawrence MS, Stojanov P, Polak P *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457), 214–218 (2013).

37 Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* 9(12), 958–970 (2008).

38 Andujar P, Wang J, Descatha A *et al.* P16ink4a inactivation mechanisms in non-small-cell lung cancer patients occupationally exposed to asbestos. *Lung Cancer* 67(1), 23–30 (2010).

39 Singh A, Boldin-Adamsky S, Thimmulappa RK *et al.* RNAI-mediated silencing of nuclear factor erythroid-2-related factor 2 gene expression in non-small cell lung cancer inhibits tumor growth and increases efficacy of chemotherapy. *Cancer Res.* 68(19), 7975–7984 (2008).

40 Koren A, Polak P, Nemesh J *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* 91(6), 1033–1040 (2012).

41 Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. *Nat. Genet.* 41(4), 393–395 (2009).

42 Zhang J, Fujimoto J, Wedge DC *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 346(6206), 256–259 (2014).

43 De Bruin EC, Mcgranahan N, Mitter R *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 346(6206), 251–256 (2014).

44 Forbes SA, Bindal N, Bamford S *et al.* COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 39(Database issue), D945–950 (2011).

45 Hudson AM, Yates T, Li Y *et al.* Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. *Cancer Res.* 74(22), 6390–6396 (2014).

46 Sherry ST, Ward MH, Kholodov M *et al.* DBSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1), 308–311 (2001).

47 Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39(17), e118 (2011).

48 Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* 7(4), 248–249 (2010).

49 Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7(10), e46688 (2012).

50 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat. Protoc.* 4(7), 1073–1081 (2009).

51 Dong C, Wei P, Jian X *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Hum Mol. Genet.* (2014).

52 Brognard J, Zhang YW, Puto LA, Hunter T. Cancer-associated loss-of-function mutations implicate dapk3 as a tumor-suppressing kinase. *Cancer Res.* 71(8), 3152–3161 (2011).

53 Uhlen M, Fagerberg L, Hallstrom BM *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347(6220), 1260419 (2015).

54 Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29(18), 2238–2244 (2013).

55 Dees ND, Zhang Q, Kandoth C *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22(8), 1589–1598 (2012).

56 Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9, 637 (2013).

57 Herskowitz I. Functional inactivation of genes by dominant negative mutations. *Nature* 329(6136), 219–222 (1987).

58 D'ursi P, Orro A, Morra G *et al.* Molecular dynamics and docking simulation of a natural variant of activated protein C with impaired protease activity: implications for integrin-mediated antiseptic function. *J. Biomol. Struct. Dyn.* 33(1), 85–92 (2015).

59 Kumar A, Rajendran V, Sethumadhavan R, Purohit R. Relationship between a point mutation s97c in ck1delta protein and its affect on ATP-binding affinity. *J. Biomol. Struct. Dyn.* 32(3), 394–405 (2014).

60 Liu HC, Lin TM, Eng HL, Lin YT, Shen MC. Functional characterization of a novel missense mutation, HIS147ARG, in A1 domain of FV protein causing type ii deficiency. *Thromb. Res.* 134(1), 153–159 (2014).

61 Fawdar S, Trotter EW, Li Y *et al.* Targeted genetic dependency screen facilitates identification of actionable mutations in FGFR4, MAP3K9, and PAK5 in lung cancer. *Proc. Natl Acad. Sci. USA* 110(30), 12426–12431 (2013).

62 Zhu Y, Wu Y, Luo Y, Zou Y, Ma B, Zhang Q. R102q mutation shifts the salt-bridge network and reduces the structural flexibility of human neuronal calcium sensor-1 protein. *J. Phys. Chem. B* 118(46), 13112–13122 (2014).

63 Corbi-Verge C, Marinelli F, Zafra-Ruano A, Ruiz-Sanz J, Luque I, Faraldo-Gomez JD. Two-state dynamics of the SH3-SH2 tandem of ABL kinase and the allosteric role of the n-cap. *Proc. Natl Acad. Sci. USA* 110(36), E3372–3380 (2013).

64 Sutto L, Gervasio FL. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proc. Natl Acad. Sci. USA* 110(26), 10616–10621 (2013).

65 Sun H, Li Y, Tian S, Wang J, Hou T. P-loop conformation governed crizotinib resistance in g2032r-mutated ROS1 tyrosine kinase: clues from free energy landscape. *PLoS Comput. Biol.* 10(7), e1003729 (2014).

66 Doss GP, Rajith B, Chakraborty C, Nagasundaram N, Ali SK, Zhu H. Structural signature of the g719s-t790m double mutation in the EGFR kinase domain and its response to inhibitors. *Sci. Rep.* 4, 5868 (2014).

67 Torti D, Trusolino L. Oncogene addiction as a foundational rationale for targeted anti-cancer therapy: promises and perils. *EMBO Mol. Med.* 3(11), 623–636 (2011).

68 Sharma SV, Settleman J. Oncogene addiction: setting the stage for molecularly targeted cancer therapy. *Genes Dev.* 21(24), 3214–3231 (2007).

69 Sachse C, Echeverri CJ. Oncology studies using sirna libraries: the dawn of RNAI-based genomics. *Oncogene* 23(51), 8384–8391 (2004).

70 Swanton C, Marani M, Pardo O *et al.* Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer Cell* 11(6), 498–512 (2007).

71 Henderson MC, Gonzales IM, Arora S *et al.* High-throughput rnai screening identifies a role for TNK1 in growth and survival of pancreatic cancer cells. *Mol. Cancer Res.* 9(6), 724–732 (2011).

72 Hu K, Lee C, Qiu D *et al.* Small interfering RNA library screen of human kinases and phosphatases identifies polo-like kinase 1 as a promising new target for the treatment of pediatric rhabdomyosarcomas. *Mol. Cancer Ther.* 8(11), 3024–3035 (2009).

73 Tiedemann RE, Zhu YX, Schmidt J *et al.* Identification of molecular vulnerabilities in human multiple myeloma cells by rna interference lethality screening of the druggable genome. *Cancer Res.* 72(3), 757–768 (2012).

74 Morgan-Lappe SE, Tucker LA, Huang X *et al.* Identification of RAS-related nuclear protein, targeting protein for xenopus kinesin-like protein 2, and stearoyl-coa desaturase 1 as promising cancer targets from an RNAI-based screen. *Cancer Res.* 67(9), 4390–4398 (2007).

75 Thaker NG, Zhang F, Mcdonald PR *et al.* Identification of survival genes in human glioblastoma cells by small interfering RNA screening. *Mol. Pharmacol.* 76(6), 1246–1255 (2009).

76 Cowley G, Weir B, Vazquez F *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data* 1, Article number: 140035 (2014).

77 Cheung HW, Cowley GS, Weir BA *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl Acad. Sci. USA* 108(30), 12372–12377 (2011).

78 Helming KC, Wang X, Wilson BG *et al.* Arid1b is a specific vulnerability in ARID1A-mutant cancers. *Nat. Med.* 20(3), 251–254 (2014).

79 Rosenbluh J, Nijhawan D, Cox AG *et al.* Beta-catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* 151(7), 1457–1473 (2012).

80 Weinstein IB, Joe A. Oncogene addiction. *Cancer Res.* 68(9), 3077–3080; discussion 3080 (2008).

81 Kaelin WG Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nat. Rev. Cancer* 5(9), 689–698 (2005).

82 Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-CAS9 system. *Science* 343(6166), 80–84 (2014).

83 Jones S, Anagnostou V, Lytle K *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* 7(283), 283ra253 (2015).

84 Haibe-Kains B, El-Hachem N, Birkbak NJ *et al.* Inconsistency in large pharmacogenomic studies. *Nature* 504(7480), 389–393 (2013).

Chapter Three

Introduction

Functional Data Used To Filter Datasets

3.1 Filtering NGS data with functional data

The preceding chapter explored the difficulties with extracting driver mutations from the mutational noise caused by inconsequential passenger mutations. Statistical methods based on co-occurrence or deviations from the background mutation rate were discussed. Mutation assessors, which have shown some success in linking the structural consequence of a mutation to a functional prediction, were also highlighted. The drawback of mutation assessors is that they attempt to perform a functional prediction for all proteins, including many of unknown function. As a result these prediction algorithms lack the refinement to determine between gain of function (GOF), loss of function (LOF), and inconsequential aberration of function. One solution is to concentrate on proteins, or regions of proteins, in which the connections between structure and function are already known. Whilst this reduces the pool of genes available to find drivers, it offers results that are easier to interpret and arguably more predictive. Protein kinases are an ideal class of protein to focus on as they have a well-defined shared function linked to structural elements and are numerous (more than 400 classical kinases). The next section describes kinase function and how it links to different structural motifs that were queried in NGS datasets in Chapter 6. The final section in this chapter introduces a rare disease called Diffuse Idiopathic Pulmonary Neuro Endocrine Cell Hyperplasia (DIPNECH). Screening kinase functional motifs in germline variants of DIPNECH was used in Chapter 7 to identify potential drivers of DIPNECH and other neuroendocrine malignancies.

3.2 Protein Kinases

3.2.1 Protein kinase function

Protein kinases are defined as proteins that are capable of attaching phosphate to an amino acid in a protein sequence [79]. The addition of phosphate predominantly occurs on hydroxyl groups of amino acids such as serine, threonine and tyrosine [79]. This biological process is immensely important for the functioning of all cells and is critical for cellular signalling, control of cell cycle, proliferation, gene transcription, cellular motility, response to stress and DNA damage, activation and degradation of proteins to name but a few [80]. The first report of an enzyme that could cause phosphorylation of another protein was made in 1954 when Kennedy demonstrated

that a liver enzyme caused the incorporation of radioactive phosphate into casein [81]. Soon afterwards it was identified that Mg-ATP was required for this activity to take place [82]. A decade later the first protein kinase, Protein Kinase A (PKA), was identified and shown to rely upon a second messenger cAMP for its activation [83]. Furthermore it was demonstrated that activated PKA could directly regulate the activity of different enzymes producing the first example of the function of kinase phosphorylation in cell signalling cascades [82, 83].

Initially phosphorylation by protein kinases was thought to occur exclusively on serine and threonine amino acids. Fortuitously aided by an expired reaction buffer that allowed tyrosine to separate from threonine, Tony Hunter discovered that tyrosine could also be phosphorylated on its hydroxyl group by protein kinases [84]. Histidine, lysine and other amino acids have now been revealed to be capable of phosphorylation despite the absence of hydroxyl groups [82, 85]. However the most prevalent and best studied phosphorylation mechanism is the transfer of the gamma phosphate from magnesium bound adenosine triphosphate (Mg^{2+}-ATP) to the hydroxyl group of an amino acid to yield a phosphorylated amino acid, magnesium bound adenosine diphosphate (Mg^{2+}-ADP) and a hydrogen ion [86]. This phosphorylation of an amino acid leads to a functional effect by changing the conformation of the protein or allowing interaction with other proteins.

3.2.2 Kinase Domain Structure

Shared basic function means that the majority of protein kinases have a shared structure and protein sequence and this has enabled the discovery of additional kinases. Manning et al used mathematical comparisons of protein sequences with known kinase domains to identify all human kinases (The Human Kinome) [87]. Five hundred and eighteen protein kinases were identified accounting for 1.7% of all human genes [87]. Protein kinases can be further subdivided based on typical or atypical sequence similarity (478 kinases versus 40) or substrate specificity for a preference to phosphorylate serine/threonine or tyrosine residues (388 kinases versus 90) [87, 88]. The model used by Manning et al. allowed sequence similarity between kinases to be quantified so that an evolutionary-like network (‘Kinome Tree’) could be constructed. More recently the crystal structures of many kinases have been solved providing a better insight into the conserved structures and phosphorylation mechanism.

All kinase domains are comprised of two lobes, the smaller N-terminal lobe and larger C-terminal lobe [89]. Between these two lobes is a cleft where the adenine ring of ATP orients to facilitate phosphate transfer [89]. Figure 1 taken from Kornev et al. [90] demonstrates the structure of the 'classical' kinase domain with the N-lobe linked to the larger C-lobe via the hinge region, highlighting the interactions between the kinase domain, ATP and the substrate. Protein structures are dominated by two secondary structural arrangements of amino acids, beta sheets and alpha helices. Beta sheets are formed by the hydrogen bonds between strands of amino acids (beta strands) running anti-parallel (or parallel) to another to form a sheet like structure [89]. To run anti-parallel to each other the beta strands depend on loop structures to turn back on themselves. Alpha helices are coiled spring-like structures running clockwise in which the amino acids in sequence form a stable hydrogen-bond backbone [89].

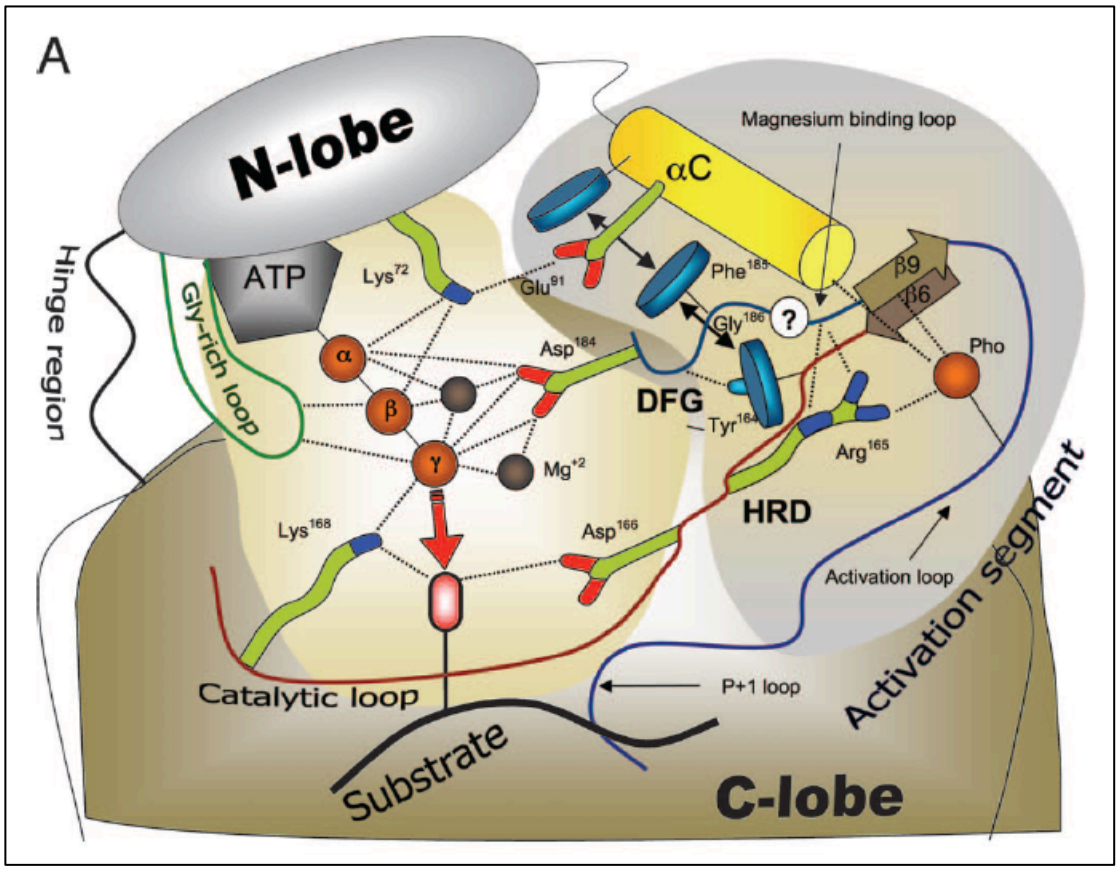


Figure 1: Kinase domain structural elements and their interactions with ATP and the protein substrate. Taken from Kornev at al. [90]. ATP docks into a cleft between the small N-lobe and the larger C-lobe. The red arrow demonstrates the transfer of the gamma phosphate of ATP to a residue on the substrate. Dashed black

lines demonstrate important polar contacts such as within the glycine rich loop that makes connections with the beta and gamma phosphate. Yellow shaded areas mark-out the residues in contact with ATP and/or substrate that catalyse the reaction, such as the DFG motif that anchors ATP via the binding of magnesium ions. The HRD motif is shown positioned in proximity to the substrate binding loop so that it can position the phosphorylation site acceptor group on the peptide substrate. The critical positioning of the activation loop (thick blue line) between the catalytic motifs and the substrate binding P+1 loop means that conformational changes induced by phosphorylation of this region by upstream kinases can greatly affect kinase activity.

3.2.3 N-terminal lobe (N-lobe)

In the following sections the structural locations shall be referenced to Protein Kinase A (PKA) as this is the best-studied kinase in terms of structure [89]. Other kinases may differ slightly in their structure although there is generally a high degree of conservation throughout the human kinome.

The major structure of the N-lobe is a beta sheet comprised of 5 beta strands. The first highly conserved motif occurs in the loop between the first 2 beta strands and is called the 'Glycine-Rich Loop' (G-loop). The typical amino acid sequence for this conserved motif is 'GxGxxG' (where G represents Glycine and x represents any amino acid). The G-loop forms a part of the ATP-binding pocket and its function within this is to position the gamma phosphate of ATP so that it can be catalysed [91]. This function requires a high degree of flexibility of the structure and the small size of the glycine amino acid is critical to achieve this [91]. The first two glycines (G1 and G2) are highly conserved (G1 present approximately 95% of the 411 'typical' active kinases and G2 present in 98%). Biochemical studies have shown a gradual reduction in functional importance from G1 to G3 glycine [91]. Mutagenesis of G3 to another small amino acid had the smallest effect on diminishing kinase activity and this decreased function importance is reflected in its lower rate of conservation in the typical active kinases (78.2%) [91]. The two residues located between G2 and G3 have been shown to play a regulatory mechanism in CDK2 (Tyr-15 and Thr-14) where phosphorylation of either residue results in loss of kinase activity [92]. Molecular dynamics simulations have shown that the phosphorylation of these residues results in misalignment of the ATP with the G-loop [93]. Therefore mutations of these residues to amino acids that cannot be phosphorylated may lead to enhanced ATP alignment and GOF or mutations to phosphomimetic or larger amino acids may lead to LOF.

Following the G-loop of the beta-1 and beta-2 strands another highly conserved region occurs on the beta-3 strand. This conserved motif is designated 'VAxK' (where V= Valine, A = Alanine, K = Lysine and x = any amino acid) although it is the lysine which is the most highly conserved with all typical active kinases possessing a lysine at this position. This lysine has two main functions that are critical for the kinase activity. Firstly it anchors the alpha and beta phosphates of ATP thereby correctly orientating the molecule [94]. Secondly it forms a salt-bridge with a conserved glutamate residue in the C-helix which causes a conformational change in the protein that is critical for kinase activity [94].

The C-helix follows the beta strands of the N-lobe and whilst it is termed a part of the N-lobe it is technically a region that links the N-lobe to the C-lobe [89]. The conserved glutamate residue within the C-helix is critically linked to the activation loop of the kinase and provides the mechanism by which phosphorylation of the activation loop of a kinase (by auto-phosphorylation or an upstream kinase) controls the catalytic activity of the kinase. This is achieved because in an inactive conformation the glutamate interacts with the conserved arginine residue of the HRD motif [94]. When the activation loop is phosphorylated, the phosphorylated residue out-competes the glutamate for the arginine, leaving the glutamate unbound allowing it to move into a position to hydrogen bond with the lysine and form the active conformation of the kinase [94].

3.2.4 C-terminal lobe (C-lobe)

Unlike the N-lobe that is comprised of mainly beta strands forming a beta sheet, the C-lobe is predominantly formed of helices with one beta-sheet [89]. The helices form the structural backbone of the kinase domain whilst the important catalytic machinery of the kinase is contained within the short beta strands of the beta-sheet. This machinery consists of (in order of sequence appearance): the catalytic loop (containing the HRD motif), the Mg-binding loop (containing the DFG motif), the activation loop, and the peptide positioning loop (containing the APE motif) [95].

The catalytic loop is named as such because it is responsible for the catalysis of the transfer of the gamma phosphate from ATP to the protein substrate. It lies between beta strands 6 and 7 and contains the highly conserved HRD motif. The aspartic acid of HRD is the most conserved residue of the motif and acts as a catalytic base to

accept a proton from the phosphorylation site on the substrate protein [94]. It is also responsible for positioning the phosphorylation site acceptor group on the peptide substrate [96]. The arginine of HRD is only present in eukaryotic kinases [96]. It bridges to the activation loop and therefore links the phosphorylation of the activation loop (by an upstream kinase) to the catalytic activity of the kinase. It also bridges to the Mg-binding loop [94]. The histidine of the HRD can often be a tyrosine (YRD) and is a structural element that provides a scaffold to the DFG motif [96]. The residues that flank the HRD in the catalytic loop are also highly conserved and perform critical roles such as the lysine of HRDxK that bridges to the gamma phosphate of ATP and the asparagine of HRDxKxxN that co-ordinates the secondary magnesium ion [94].

The Mg-binding loop contains the highly conserved DFG motif and lies between beta strands 8 and 9 [89]. As its name suggests it is responsible for anchoring ATP via the binding of magnesium ions. The aspartic acid of DFG plays the critical role in this function by binding to the magnesium ion bridging the beta and gamma phosphates of ATP [94]. Via this binding and also other polar contacts to the phosphates of ATP, the aspartic acid is central to the ATP binding. The hydrophobic phenylalanine of DFG makes hydrophobic contacts to HRD and other regions to prevent water from disrupting the phosphate transfer [94]. In many kinases the phenylalanine is replaced by leucine, which is another hydrophobic amino acid. The role of the glycine of DFG is less clear but given its high degree of conservation it must serve an important role [89]. Due to their small size, other conserved glycines are required for flexibility, which may suggest an equivalent role for the glycine of DFG.

The activation loop is less conserved throughout the human kinome than the preceding HRD and DFG motifs suggesting more kinase specific mechanisms of activation by upstream activators. As discussed earlier, phosphorylation of threonine and tyrosine residues in the activation loop cause conformational changes in the kinase through binding the arginine of HRD thus leaving the glutamate of the C-helix to bind the critical lysine of VAxK [94]. The variability of the activation loop sequence is required to allow selectivity of upstream kinases and control of signalling networks. The peptide-positioning loop lies between the activation loop and the highly conserved APE motif [95]. It makes contacts with the portion of the substrate protein that is undergoing phosphorylation and has multiple conserved residues that are present in different subsets of protein kinases [95]. However like the activation loop, there is variability in this region reflecting the structural differences of substrate phosphorylation regions [95]. The APE is a highly conserved motif but its function is

not clearly known. The glutamic acid of APE forms a salt bridge with a highly conserved arginine amino acid towards the furthermost C-terminal end of the kinase domain [97]. The stability of this salt bridge and therefore structural conformation of the kinase is dependent on the phosphorylation of the activation loop suggesting the APE motif is important in regulating the active conformation of the kinase [97].

In addition to conserved motifs identified in the amino acid sequence of kinases, comparison of the three-dimensional structure of kinases can be performed, taking into consideration the bonds between non-adjacent residues of the sequence [98]. This allowed the identification of the F-helix, a critical central structure of the large lobe that is used as an anchor-point for multiple hydrophobic residues important for catalytic activity [90]. Two other important structural features anchored to either end of the F-helix are called the R (regulatory) and C (catalytic) spines [98]. The R spine formation is dependent on activation loop conformation whilst the C-spine relies upon the presence of ATP [98]. The F-helix, R-spine and C-spine all consist of conserved residues, that do not lie in immediate proximity to each other in the amino acid sequence but instead line up upon activation of the kinase. The highly conserved DxxxxG motif seen in the amino acid sequence of kinase domain forms parts of these structures or makes important contacts with them. The D is a component of the R-spine and the G creates a pocket around the F-helix allowing other residues to bind and stabilise it [98]. Residues between the D and G of the DxxxxG motif allow the catalytic loop to anchor to the R and C-spines and are critical to allow formation of the active conformation of the kinase [98]. Overall the DxxxxG motif lies in one of the most critical regions of the kinase, creating contacts that allow the alignment of the R and C spines of the active conformation.

3.2.5 Kinases as opportunities for cancer driver discovery

The preceding sections of this chapter describe how specific amino acid sequences are critical to enable the co-ordination and catalysis of the phosphorylation reaction. Given that approximately one third of all proteins are phosphorylated [99], aberrations in kinase domain structure caused by somatic mutations have the potential to interfere with most cellular processes and all of the ‘hallmarks of cancer’ can be caused by abnormal protein phosphorylation [100]. Therefore it is not surprising that kinase domain mutations feature prominently in some of the most studied cancer mechanisms, such as common EGFR mutations and the BRAF

V600E mutation. Overall kinases are the most commonly mutated class of proteins in cancer [101]. However despite this importance in cancer development there is a disparity in our knowledge across the human kinome. Some kinases such as BRAF, EGFR, KIT, and SRC have been well studied with thousands of papers written on their function in cancer. Conversely the role of many other kinases is not known with some completely uncharacterised, some without known downstream substrates, and even some without a single biochemical study on PubMed (for example CSNK1A1L, which only is mentioned as a top hit of a methylation array study and has not been characterised in that paper or any others) [102]. In 2010 it was estimated that more than 100 kinases have completely unknown function with an additional 200 being largely uncharacterised [101]. Small interfering RNA (siRNA) and short hairpin RNA (shRNA) screens indicate that many of these unknown and uncharacterised kinases are critical for cancer cell survival [103]. Another attractive feature for kinase research is the relative ease of pharmacological inhibition of kinases at clinically achievable dosages compared to other targets such as p53, KRAS and APC [101]. The clinical attrition rate of kinase inhibitors from phase 1 to registration is much lower than other anti-tumour agents (53% versus 82%) [104]. However, like the disparity of publications for different kinases there is also a disparity of available inhibitors with a majority of new inhibitors targeting kinases for which approved drugs are already available [101].

Therefore there are great opportunities to be gained from focussing on the role of the unexplored human kinome in carcinogenesis. The central importance of phosphorylation signalling in cellular processes means that the mechanism is tightly controlled by a conserved structural mechanism. The shared structure and common function of the kinase domain across the kinome mean that biochemical effects such as the mutation of critical residues in a commonly studied kinase can be applied across all members of the kinome. This allows the identification of functional mutations in cancer genomics datasets even if the exact function of the specific kinase, in terms of signalling mechanism, is not known. The method can be exploited to identify driver mutations and prompt further inquiry into the actual mechanisms of the novel kinase. Chapters 6 and 7 report the use of this approach to filter different genomics datasets to report potential driver mutations and kinase targets for further investigation.

3.3 Diffuse Idiopathic Pulmonary Neuroendocrine Hyperplasia (DIPNECH)

3.3.1 Background and Clinical Features

Diffuse Idiopathic Pulmonary Neuroendocrine Cell Hyperplasia (DIPNECH) is a rare condition characterised by hyperplasia of neuroendocrine cells in the distal airways. It was first described in 1992 and approximately only 100 cases have been reported in the literature since [105, 106]. Patients are classically female and middle-aged, presenting with a long-standing chronic cough and dyspnoea [107]. Many patients are non-smokers [108]. Frequently these patients are initially wrongly diagnosed as suffering from bronchial asthma and inhaled bronchodilators appear to benefit some patients confounding the confusion [105, 107-109]. Radiologically there is a spectrum of appearances from bronchiectasis and ill-defined opacity to multiple lung nodules [108, 110]. A proportion of patients present whilst undergoing investigations for other disorders and the radiological appearances can be mistaken for metastatic disease [106, 111, 112]. The symptoms can be mild and many cases do not progress significantly over the course of many years [105, 107]. As a result, some believe that the actual incidence of DIPNECH is much higher than suggested by the small number papers reporting the disease [109, 111].

Given the low number of cases reported in the literature, long-term follow-up data is very scarce. Cytotoxic chemotherapy has been tried in a very small number of cases and been unsuccessful but this is not surprising as the rarity of the condition means that there has been no opportunity to develop a standard chemotherapeutic regimen [105]. Inhaled and systemic corticosteroid treatments have been used with some success to relieve symptoms but again the small numbers mean that evidence is limited [108]. Like carcinoid tumours, that are also neuroendocrine in origin, the use of somatostatin analogues in DIPNECH patients has given symptom improvement [113]. In the most severe cases double lung transplants have been performed with success [114].

3.3.2 Relationship to Other Malignancies

Histopathologically DIPNECH is confined to the respiratory mucosa and does not penetrate through the basement membrane [113, 115]. However a proportion of patients progress to develop carcinoid tumours. In support of this, one small study

observed that 19 out of 25 patients receiving resection of pulmonary carcinoid possessed DIPNECH in the resection specimen [116]. Therefore DIPNECH is thought to be a precursor for pulmonary carcinoid tumours and is included by the World Health Organisation as a preneoplastic condition [117, 118]. The carcinoids that develop in DIPNECH patients are typically low grade and peripheral [116, 117].

It is also noted that a significant number of patients with DIPNECH have a history of non-neuroendocrine cancer. For example, out of 19 patients at our institute 7 have been diagnosed with non-neuroendocrine tumour [M. Howell unpublished data]. In another case series, 5 out of 10 asymptomatic DIPNECH patients had a previous cancer diagnosis including one patient with Multiple Endocrine Neoplasia Type 1 (MEN1) [111]. These observations are skewed by the fact that many patients with asymptomatic DIPNECH only become diagnosed whilst being investigated for another malignant condition. However, a recent report of four cases of DIPNECH appearing in the resection specimens of lung adenocarcinoma may suggest a causal relationship [119]. This is repeated in our own data in which 3 patients were diagnosed following detection of DIPNECH foci in cancer resection specimens [M. Howell unpublished data]. These associations suggest that DIPNECH and other malignancies may share common genetic or environmental aetiologies warranting further investigation. In chapter 7 of this thesis I shall use the germline sequencing of DIPNECH patients alongside filtering of conserved kinase domain variants to identify potential drivers of other neuroendocrine malignancies.

Chapter Four

Materials and Methods

4.1 Materials

4.1.1 Cell Line Media and Cell Lines

For cell culture the following cell line media formulations were used:

RPMI Medium 1640 (1X) + GlutaMAX™ (Life Technologies) supplemented with the following (according to the cell line used):

- Foetal Bovine Serum (labtech.com) concentration depending on cell line
- Additional L-glutamine as GlutaMAX™ (Life Technologies) concentration depending on cell line.

Dulbecco's Modified Eagle's Medium (DMEM) (Sigma) with 4500mg/L glucose, sodium pyruvate, and sodium bicarbonate, without L-glutamine. DMEM was supplemented with the following (according to the cell line used):

- Foetal Bovine Serum (labtech.com) concentration depending on cell line
- Additional L-glutamine as GlutaMAX™ (Life Technologies) concentration depending on cell line.

All media was supplemented with 1% Penicillin Streptomycin (Life Technologies) resulting in 100 Units/ml of Penicillin and 100 µg/ml Streptomycin for general cell line maintenance and for making up freezing media. These antibiotics were not added to media used for plasmid or siRNA transfection.

The following table lists the cell lines used, where they were obtained from, standard cell line media, and modifications of that media as described above:

Cell Line	Source	Media: (FBS = Foetal Bovine Serum)
HEK293T	ATCC (CRL-1573)	RPMI 1640 with 10% FBS
BEAS-2B	ATCC (CRL-9609)	RPMI 1640 with 10% FBS
CAL51	DSMZ (ACC-302)	DMEM with 10% FBS
H1437	ATCC (CRL-5872)	RPMI 1640 with 10% FBS and 4mmol GlutaMAX™
H2009	ATCC	RPMI 1640 with 10% FBS and 4mmol

	(CRL-5911)	GlutaMAX™
H2087	ATCC (CRL-5922)	RPMI 1640 with 10% FBS and 4mmol GlutaMAX™
H2122	ATCC (CRL-5985)	RPMI 1640 with 10% FBS and 4mmol GlutaMAX™
HUG1N	RIKEN (RCB1179)	RPMI 1640 with 10% FBS
IM95	JCRB (JCRB1075.1)	DMEM with 10% FBS
NUGC3	JCRB (JCRB0822)	RPMI 1640 with 10% FBS

Table 1: Cell line sources and information

In the above table, ATCC refers to the American Type Culture Collection (http://lgcstandards-atcc.org), RIKEN is the RIKEN Bioresource Centre (http://cell.brc.riken.jp/), and JCRB is the Japanese Collection of Research Bioresources Cell Bank (http://cellbank.nibiohn.go.jp/).

Cell line stocks were frozen in liquid nitrogen or the -80°C freezer in freezing media consisting of:

- 45% Standard cell line media with modifications above
- 45% Foetal Bovine Serum (labtech.com)
- 10% Dimethyl Sulfoxide (DMSO) (Fisher BioReagents – BP231-100)

4.1.2 Primers

Sanger sequencing, DNA amplification, site-directed mutagenesis and molecular cloning were all carried out using custom-made primers supplied by Eurofin MWG Operon. Primers were re-suspended in H_20 to a stock concentration of 100μM and stored at -20°C. Appendix 2 lists the primer sequences of all primers used.

4.1.3 Plasmids

MAP2K4 and MAP2K7 were obtained in the pCMV6-Entry vector from Origene and PRKCQ was obtained in the pENTR vector from Life Technologies. These were inserted into the pDONR-221 vector using amplification primers containing attb

flanking sites and the BP clonase reaction described below. Wild-type PAK4 plasmid (Addgene Plasmid 23713) was obtained from Addgene [120]. The plasmid was cloned into a Flag-tagged destination vector. The STOP codon and the E119Q mutation were introduced by site-directed mutagenesis using the protocol described below. MAP3K13 and CDK8 were cloned from RNA from HEK293T cells using attb flanking primers and the gene cloning technique described below for insertion into the pDONR-221 vector. DAPK3 in a 3X-Flag vector was provided by Timothy Haystead (Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710).

4.1.4 Reagents

The following reagents and buffers were made up as described below:

Running Buffer A (For Bio-Rad System)

For 1000ml buffer: 100ml of 10 x Tris/Glycine/SDS Buffer (25mM Tris, 192 mM Glycine, 0.1% SDS, pH 8.3) (Bio-Rad) and 900ml double distilled water.

Running Buffer B (For Sigma System)

For 1000ml buffer: 50ml of 20x TruPAGE™ TEA-Tricine SDS Running Buffer (Sigma) in 950ml double distilled water.

Transfer Buffer A (For Bio-Rad System)

For 1000ml buffer: 100ml of 10 x Tris/Glycine Buffer (25mM Tris, 192 mM Glycine, pH 8.3) (Bio-Rad), 200ml Methanol, and 700ml double distilled water.

Transfer Buffer B (For Sigma System)

For 1000ml buffer: 50ml of 20x TruPAGE™ Transfer Buffer (Sigma), 200ml Methanol, and 750ml double distilled water.

4.1.5 Antibodies

The following table lists the primary antibodies used, the concentration in 1% Bovine Serum Albumin in PBS + 0.05% Tween at which they were used and the corresponding secondary antibody (unless specified otherwise the antibody was purchased from Cell Signalling Technology http://cellsignal.com):

Antibody	Concentration	Secondary
Flag: Anti-Flag M2 from Sigma #F3165	1 in 10000	Mouse
Alpha-Tubulin: #T6074 From Sigma	1 in 10000	Mouse
JNK: #9252	1 in 1000	Rabbit
pJNK: (Thr183/Tyr185) #4668	1 in 1000	Rabbit
ERK: #9102	1 in 1000	Rabbit
pERK: (Thr202/Tyr204) #9101	1 in 1000	Rabbit
MKK7: #4172	1 in 1000	Rabbit
pMARCKS: (Ser152/156) #2741	1 in 1000	Rabbit
Phospho-PKC δ/θ (Ser 643/676) #9376	1 in 1000	Rabbit
pThreonine: #9386	1 in 1000	Mouse
Mouse (Secondary): Anti-mouse IgG, HRP-linked #7076	1 in 10000	N/A
Rabbit (Secondary): Anti-rabbit IgG, HRP-linked #7074	1 in 10000	N/A

Table 2: Antibody sources and information

4.2 In vitro methods

4.2.1 Cell Culture

Cell lines were maintained in a humidified incubator at 37 degrees Celsius with a carbon dioxide concentration of 5%. Cell lines were split regularly depending on the growth rate of each individual cell line typically aiming to split at approximately 80% confluency. Likewise splitting density was determined by the individual phenotype of each cell line and the typical split was 1:5 to 1:10 twice per week. Trypsin (0.05%) was used to detach adherent cell lines from tissue culture plates. To ensure that cells were used at a low passage number, cell lines were regularly frozen down for

storage at -80 degrees Celsius in the standard freezing media detailed in the materials section.

4.2.2 Gene Cloning

RNA was extracted from HEK293T cell line using the RNAEasy (QIAGEN) extraction kit using the standard manufacturers protocol. RNA was converted to cDNA using a Reverse Transcriptase reaction. 2.5 µl of the RNA sample was incubated for 5 minutes at 65 degrees Celsius with 2.5 µl oligo dt (Invitrogen), 10 µl deoxynucleoside triphosphate (DNTPs) mix, and 25 µl distilled water. To this mixture 5 µl of RT buffer (New England Biolabs), 2.5 µl RNase inhibitor (New England Biolabs), and 2.5 µl Reverse Transcriptase (New England Biolabs) was added and the sample incubated for 1 hour at 42 degrees Celsius followed by a final step of 90 degrees Celsius for 10 minutes to deactivate the enzyme. This sample was purified using PCR Purification Kit (QIAGEN) using the standard manufacturers protocol. Once the cDNA was acquired the gene of interest was amplified using a PCR protocol with primers complementary to the non-coding regions flanking the gene with the addition of attb binding site sequences. The reaction mix consisted of 12.5 µl Phusion High-Fidelity PCR Master Mix with HF Buffer, 5 µl Betaine, 5 µl cDNA, 1.25 µl Forward Primer, 1.25 µl Reverse Primer. PCR was carried out on a S1000 Thermal Cycler (Bio-Rad). The PCR program below was tailored with different annealing temperatures depending on the melting point of the complementary primers:

1) 98.0 °C for 30 seconds
2) 98.0 °C for 10 seconds
3) 60.0 °C for 30 seconds (annealing temp)
4) 72.0 °C for 2 mins 30 seconds

Steps 2-4 repeated to a total of 40 cycles.

Successful amplification of the target region was verified by running 5 µl of the sample on a 1% Agarose Gel with the addition of 0.02% Nancy 520 (Sigma). The corresponding band on the agarose gel was confirmed using a UV-light source and cut-out of the gel and purified using the Gel Extraction Kit (QIAGEN) following the standard manufacturers protocol. To insert the attb-flanked region of interest the BP Clonase reaction was performed using 7 µl of the above purified attb product, 1 µl of 150ng/µl pDONR221 vector and 2 µl BP Clonase (Thermo Fisher). The sample was

incubated at 25 degrees Celsius for 1 hour before adding 1 µl proteinase K and incubating at 37 degrees Celsius for 10 minutes.

The product of the BP Clonase reaction was transformed into Omnimax E-coli cells (Thermo Fisher) with 1 µl of BP Clonase reaction product added to 25 µl of cells and incubated on ice for 30 minutes. A heat shock was then performed at 42 degrees Celsius for 30 seconds and the sample left on ice for 2 minutes. One millilitre of LB was added and the sample placed into a shaking incubator at 37 degrees Celsius for 1 hour. The sample was then plated on LB plates with 10 µg/ml kanamycin overnight in an incubator at 37 degrees Celsius. Single colonies were selected and grown in liquid broth with 10 µg/ml kanamycin overnight in a shaking incubator at 37 degrees Celsius. DNA was extracted from these samples using the Mini Prep Kit (QIAGEN) and following the standard manufacturers protocol. Sanger sequencing was performed to verify that the correct region had been successfully inserted into the pDONR221 vector.

For stable expression experiments the gene of interest was inserted into the pDEST-Flag destination vector using the LR Clonase Reaction. The reaction mix consisted of 75 ng pDONR221 vector with inserted gene of interest, 75ng empty pDEST-Flag vector, 1 µl LR Clonase reaction mix and was made up to 5 µl with distilled water. The mix was incubated at 25 degrees Celsius for 1 hour before 1 µl of Proteinase K was added and incubated for 10 minutes. The reaction mix was transformed into Omnimax E-coli cells using the same protocol as for the BP Clonase reaction above although the kanamycin was replaced in the LB plates and media with 10 µg/ml Ampicillin to account for the different antibiotic resistance of the pDEST-Flag vector. DNA was extracted from these samples using the Mini Prep Kit (QIAGEN) and following the standard manufacturers protocol. Sanger sequencing was performed to verify that the correct region had been successfully inserted into the pDEST-Flag vector.

4.2.3 Site Directed Mutagenesis

Mutants to compare to wild-type function were created using site-directed mutagenesis using the QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies). The reaction mix consisted of 2.5 µl reaction buffer, 1 µl 50 ng/ µl template plasmid, 0.5 µl forward primer, 0.5 µl reverse primer, 0.5 µl DNTPs, 5 µl Betaine, 15 µl distilled water, 0.5 µl Pfu. PCR was carried out on a S1000 Thermal

Cycler (Bio-Rad). The PCR program below was tailored with different annealing temperatures.

1) 95.0 °C for 30 seconds
2) 95.0 °C for 30 seconds
3) 55.0 °C for 1 minute
4) 68.0 °C for 13 minutes

Steps 2-4 repeated to a total of 18 cycles.

Sequences were confirmed using Sanger sequencing.

4.2.4 Transient Transfection of Plasmids

MKK4 transient transfection experiments were performed in the CAL51 breast cancer cell line. Transfections of other plasmids were performed in the HEK293T cell line. Cells were split into a 12 well format 24 hours before transfection and maintained in the standard cell culture media for that cell line without antibiotics. A transfection mix consisting of 3 µl Attractene Transfection Reagent (QIAGEN), 80 µl Opti-MEM Reduced Serum Medium (Life Technologies) and the plasmid to be transfected (typically 0.8 µg). Each transfection mix is prepared in a sterile Eppendorf and incubated at room temperature for 20 minutes before adding dropwise to the 12 well plate. Cells were lysed on ice after 48 hours using Triton X-100 Cell Lysis Buffer supplemented with protease inhibitor tablet (Roche). Immediately following lysis, samples were frozen at -80 degrees Celsius for 10 minutes before thawing and removing any adherent cells with a cell scraper. Samples were transferred to an Eppendorf before centrifuging at 14,000 rpm for 10 minutes at 4 degrees Celsius. Supernatant was carefully removed and transferred to a clean Eppendorf and 15 µl of Western blot sample buffer added per 100 µl of sample. The sample was boiled at 100 degrees Celsius for 5 minutes and then cooled before Western blotting.

4.2.5 Western Blotting

Two different systems were used for Western Blotting due to changes in lab equipment. For the work in Chapter 4 the Bio-Rad system was used and samples were run on Mini-PROTEAN TGX 12% Precast Gels (Bio-Rad) with Running Buffer

A. Electrophoresis was performed at 200V for 35 minutes on the Bio-Rad PowerPac Basic machine.

For all other chapters the Sigma system was used and samples run on TruPAGE 12% Precast Gels (Sigma) with Running Buffer B. Electrophoresis was performed at 180V for 45 minutes on the Bio-Rad PowerPac Basic machine.

Gels were transferred to Immun-Blot PVDF Membranes (Bio-Rad) membranes in Transfer Buffer A (Bio-Rad system) and Transfer Buffer B (Sigma system) both at 100V for 1 hour. Membranes were blocked in 5% powdered milk in PBS + 0.05% Tween 20 (Fisher BioReagents) for 30 minutes at room temperature. Milk was removed and blots were rinsed with water and the primary antibody added in 1% Bovine Serum Albumin (Sigma) in PBS + 0.05% Tween 20. Blots were incubated at 4 degrees Celsius overnight followed by 3 x 15 minute washes with PBS + 0.05% Tween 20. Blots were incubated for 1 hour with the secondary antibody in 1% Bovine Serum Albumin in PBS + 0.05% Tween. Three further 15 minute washes were performed before incubating each blot in 2ml of Pierce ECL Western Blotting Substrate Reagent (Thermo Scientific) 1 minute at room temperature and then exposure to Amersham Hyperfilm ECL (GE Healthcare) photographic film.

4.2.6 Stable Over-Expression and Colony Formation Assay

Stable over-expression of wild type MKK4 within the CAL51 breast cancer cell line and MKK7 within the IM95 and NUGC3 gastric cancer cell lines was performed by Natalie Stephenson. For anchorage-dependent colony formation assay cells were seeded 100 cells per well of a 6-well plate, treated with tetracycline the following day and left to grow for 3 weeks. For anchorage-independent colony formation assay the cells were seeded 10,000 cells per well of a 6-well plate in 0.35% soft agar +/- tetracycline and left to grow for 3 weeks. Colonies were assessed using 0.05% crystal violet (Sigma) in 25% methanol.

4.3 Clinical methods

4.3.1 DIPNECH Patient Consent and Sample Processing

Ten patients with a pathological diagnosis of DIPNECH were identified for germline whole exome sequencing. All 10 patients were receiving active follow-up at The

Christie NHS Foundation Trust under the same clinical team following a pathology review confirming a diagnosis of DIPNECH. Informed written consent was obtained from each patient for the collection of a whole blood sample for the purposes of germline whole exome sequencing. Where relevant, consent was also requested from patients to analyse the genomics of any archival tissue.

Whole blood was collected in a standard EDTA anti-coagulated blood collection tube and immediately frozen before transfer to The Genomics Diagnostics Laboratory at Manchester Centre For Genomic Medicine, Saint Mary's Hospital, Oxford Road, Manchester, M13 9WL for DNA extraction prior to whole exome sequencing. Archival tissue was obtained via the Manchester Cancer Research Centre Biobank either from their own archive or from archival tissue at collaborating hospitals. Formalin Fixed Paraffin Embedded (FFPE) tissue was prepared for DNA extraction by the Biobank before transfer to St Mary's for DNA extraction.

4.4 Next Generation Sequencing

4.4.1 Cell Line Sequencing

DNA was extracted within 3 passages from delivery from ATCC using DNeasy Blood and Tissue Kit (QIAGEN). The documented passage number for each cell line on delivery from ATCC was; H2009 = passage 23, H2087 = passage 21, H2122 = passage 21, H1437= passage 46. Whole exome sequencing was performed by the CRUK Manchester Institute Core Facility. The samples were enriched with the SureSelect XT Human All Exon V4 library and sequenced on the Illumina HiSeq 2500 performing 2 x 100 bp paired-end sequencing. Variant calling and alignment of the sequencing data to the reference genome (GRCh37/hg19) was performed by Yaoyong Li using the Genome Analysis Tool Kit (GATK: version 3.1.1)[121].

4.4.2 DIPNECH Germline Sequencing

DNA was extracted from whole blood samples by The Genomics Diagnostics Laboratory at Manchester Centre For Genomic Medicine, Saint Mary's Hospital, Oxford Road, Manchester, M13 9WL. Whole exome sequencing was performed by the CRUK Manchester Institute Core Facility. The samples were enriched with the SureSelect XT Human All Exon V6+COSMIC library and sequenced on the Illumina HiSeq 2500 (Rapid Run mode) performing 2 x 100 bp paired-end sequencing.

Variant calling and alignment of the sequencing data to the reference genome (GRCh37/hg19) was performed by Hui Sun Leong using the Genome Analysis Tool Kit (GATK: version 3.1.1)[121]. Post GATK pipeline the variants further filtered to remove any entry with a read depth below 20 or allele frequency below 35%.

4.4.3 Validation of NGS targets

Verification of mutations in respective cell lines and germline samples were made using Sanger sequencing of the same DNA sample as used for the NGS. The list of validation primers are given in Appendix 2. Target region was first amplified using PCR with 12.5 µl Phusion High-Fidelity PCR Master Mix with HF Buffer, 5 µl Betaine, 5 µl cDNA, 1.25 µl Forward Primer, 1.25 µl Reverse Primer. PCR was carried out on a S1000 Thermal Cycler (Bio-Rad). The PCR program below was tailored with different annealing temperatures depending on the melting point of the complementary primers:

1) 98.0 °C for 30 seconds
2) 98.0 °C for 10 seconds
3) 60.0 °C for 30 seconds (annealing temp)
4) 72.0 °C for 2 mins 30 seconds

Steps 2-4 repeated to a total of 40 cycles.

Successful amplification was confirmed by electrophoresis on 1% agarose gel with Nancy 520 (Sigma). Prior to sequencing the PCR product was purified using Illustra ExoProstar Enzymatic PCR Clean-up (GE Healthcare). Sequencing of the amplified target was carried out with a nested primer (Appendix 2) on an ABI13130 16 capillary system (Life Technologies). Sequencing data was analysed using the 4Peaks software (MekenTosj).

4.5 In silico methods

4.5.1 Online Resources

Three main publically accessible genomics datasets were used for analysis:

1) The Cancer Genome Atlas Mutational Data for multiple tumour subtypes accessed via the CGDS-R package and www.cbioportal.org [14, 122].

2) The Cancer Cell Line Encyclopaedia(CCLE) from the Broad Institute for the mutational profiles of over 1000 commercially available cell lines available at www.broadinstitute.org/ccle/home [123]
3) The Catalogue of Somatic Cancer Mutation in Cancer (COSMIC) from the Sanger Institute which aggregates much of the TCGA data as well as performing its own sequencing on commercially available cell lines. Available at www.cancer.sanger.ac.uk/cosmic/ [124].

4.5.2 Computer Packages

The Integrative Genomics Viewer (IGV: Broad Institute) was used to view and analyse NGS Bam file data [125]. The truncating mutation kinase screen and critical motif screen were programmed in the R programming language using RStudio for Macintosh (RStudio Team: Boston, MA www.rstudio.com). For string searches within R the StringR plugin (Version 1.0.0: Hadley Wickham, RStudio) was used. Sanger sequencing data was viewed with 4Peaks for Macintosh (MekenTosj). Protein sequence alignment was carried out using STRAP for Macintosh (Structure Based Sequences Alignment Program: Christophe Gille). Circos plots were constructed using the Circos program downloaded from http://www.circos.ca [126]. Structural modelling of MAP2K4 structures were performed using MODELLER (Version 9.16: Andrewj Sali) and structural images were produced using PyMol (Version 1.5.0.5). Molecular dynamics simulations were performed using GROMACS (GROningen Machine for Chemical Simulation: version 4.5.3) [127].

4.5.3 COSMIC/CCLE comparison and evaluation

Two comparisons were made between the data from CCLE and COSMIC for cell line sequencing. This was because the COSMIC dataset was updated midway through the project to provide mutational data for a much large collection of cell lines. The smaller comparison was carried out using CCLE and COSMIC data downloaded on 14th May 2013. The CCLE data remained the same throughout both analyses occurring in MAF format and filtered by CCLE to remove known common polymorphisms, variants occurring in less than 10% of sequencing reads, putative neutral variants, and any variants occurring outside the coding region (*CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf)*. No information is given by CCLE regarding the algorithms used to call putative neutral variants and neither COSMIC nor CCLE detail their germline filtering

algorithms. COSMIC data was copied from the webpage from each respective cell line as in 2013 there was no function to download the entire dataset. A script was written in Groovy programming language by Tim Yates to compare the mutational profiles of the cell lines reported by COSMIC and CCLE for all genes that were reported as being sequenced by both institutes. Mutations were divided in three categories, those occurring in both data sets, those exclusively in CCLE, and those exclusively in COSMIC. Sequencing bam files were only publically available for CCLE data so the mutations occurring exclusively in COSMIC were viewed in CCLE bamfiles using IGV to categorise the reason why they had been missed by CCLE. The GC content of the mutations exclusively occurring in COSMIC were calculated using the genomic location of the mutation and assessing the region 100bp either side of the mutation.

When COSMIC updated its cell line project to include the NGS of 1015 cell lines the same method was used as above to identify mutations in both datasets and mutations exclusive to either COSMIC or CCLE. The COSMIC data was downloaded on 12th November 2013 (*CosmicCellLineProject_v67_241013.tsv.gv*). The file is filtered for known common polymorphisms although again no specific details are given regarding methodology. This expanded dataset allowed the comparison of 1630 mutually sequenced genes between 568 mutually sequenced cell lines using a similar script to the above analysis written by Tim Yates in Groovy programming language. Mutations were matched by both gene name and genomic location. The 568 cell line mutational profile comparison was also carried out using unfiltered CCLE data. In this dataset common known polymorphisms, putative neutral variants, and non-coding variants were not removed from the analysis. The allele fraction required to register a mutation remained at more than 10% for this dataset. The file (*CCLE_hybrid_capture1650_hg19_allVariants_2012.05.07.maf.gz*) was downloaded from the CCLE website on 22nd November 2013.

Comparing large datasets of different ages result in differences of opinion as to the correct gene transcript. When this occurred the script evaluated the amino acid change for all different transcripts of the same gene and selected the most commonly reported mutation. If there was no majority change CCLE was given precedence followed by COSMIC and then our own (CRUK_MI) sequencing. Cancer Census genes were identified via the COSMIC Cancer Census webpage downloaded on 22nd May 2013 (http://cancer.sanger.ac.uk/cancergenome/projects/census/).

4.5.4 Cold-spot analysis

Genomic regions with consistent poor sequencing across multiple samples were identified using 10 randomly selected lung cancer whole exome sequencing bamfiles. The bam files were downloaded from the Santa Cruz Cancer Genomics Hub (University of California) on 9th January 2014. The sequencing files used for analysis were:

CCLE-NCI-H2286-DNA-08	CCLE-HLF-a-DNA-08
CCLE-NCI-H1944-DNA-08	CCLE-JL-1-DNA-08
CCLE-COR-L95-DNA-08	CCLE-HCC-78-DNA-08
CCLE-NCI-H1373-DNA-08	CCLE-DV-90-DNA-08
CCLE-NCI-H1184-DNA-08	CCLE-DMS153-DNA-08.

All genomic locations in the coding transcripts of 969 protein-coding kinase and Cancer census were queried (Chapter 4, Supplementary Table 6) [128]. A script was written by Yaoyong Li to identify all genomic locations with a mean sequencing read coverage of 4 or fewer sequencing reads across the 10 samples. Cold-spots were defined as inadequate sequencing occurring in exonic regions in 100 consecutive nucleotides. For multiple transcripts of the same gene the longest cold-spot was retained if the genomic start or end position matched other cold-spots. For each sequencing cold-spot the average GC-content (proportion of positions with guanine and cytosine nucleotides compared to all positions in the cold-spot) was calculated. Using the same method the average GC-content of all the coding exons of the 969 genes was made. For genes with multiple transcripts the longest transcript (based on Ensembl Version 70) was used to make the whole exome GC-content calculation.

4.5.5 Critical motif location

Critical kinase domain motifs were located to cross-reference with mutational data and produce filtered lists of mutations with a high likelihood of killing catalytic activity. 411 kinases from Manning et al. were queried [87] (Supplementary Table 1 Chapter 6). All these kinases are described in the Manning paper as being catalytically active by virtue of possessing critical VAIK, HRD, and DFG motifs. A motif location script (Appendix 3) was written in R programming language to identify the location of the VAIK, HRD and DFG motifs within the Genbank [129] sequences for each kinase. Using the DFG location as an anchor point within each sequence and knowing that the APE motif occurs approximately 20-30 amino acids C-terminal to that position the

script searched for variations of the APE motif in that region. The same method was used to identify the glycine-rich loop and salt-bridge E using the lysine of the VAIK as an anchor point and searching for variations of these motifs. Following the critical residue scoring and identification of the APE-6 hotspot this residue was located using the same method and the APE motif as anchor point.

4.5.6 Truncating mutation screen

A truncation mutation screen was performed on the same 411 kinases from Manning et al. that are described in the previous section. Using the E of the APE motif as the C-terminal extent of the kinase domain all truncating mutations occurring N-terminal to this location were recorded. Mutational data from the following TCGA studies (downloaded using CGDS-R package on 11th January 2016) along with the CCLE mutation dataset (downloaded on 11th January 2016) were used to identify the truncation mutations:

genetic_profile_id	**Cancer Subtype**
acc_tcga_mutations	Adrenalcortical
blca_tcga_mutations	Bladder
brca_tcga_mutations	Breast
brca_tcga_pub_mutations	Breast
brca_tcga_pub2015_mutations	Breast
cesc_tcga_mutations	Cervix
coadread_tcga_mutations	Colorectal
coadread_tcga_pub_mutations	Colorectal
gbm_tcga_mutations	Glioblastoma
gbm_tcga_pub_mutations	Glioblastoma
gbm_tcga_pub2013_mutations	Glioblastoma
hnsc_tcga_mutations	Head and Neck
kich_tcga_mutations	Kidney Chromophobe
kich_tcga_pub_mutations	Kidney Chromophobe
kirc_tcga_mutations	Kidney Renal Cell
kirc_tcga_pub_mutations	Kidney Renal Cell
kirp_tcga_mutations	Kidney Renal Papillary Cell
laml_tcga_mutations	Acute Myeloid Leukaemia
lgg_tcga_mutations	Low Grade Glioma
lihc_tcga_mutations	Liver Hepatocellular
luad_tcga_mutations	Lung Adenocarcinoma
lusc_tcga_mutations	Lung Squamous

lusc_tcga_pub_mutations	Lung Squamous
ov_tcga_mutations	Ovarian
ov_tcga_pub_mutations	Ovarian
paad_tcga_mutations	Pancreas
pcpg_tcga_mutations	Phaeochromocytoma and Paraganglioma
prad_tcga_mutations	Prostate
prad_tcga_pub_mutations	Prostate
skcm_tcga_mutations	Melanoma
stad_tcga_mutations	Stomach
thca_tcga_mutations	Thyroid
ucec_tcga_mutations	Uterine Corpus Endometrial
ucec_tcga_pub_mutations	Uterine Corpus Endometrial
ucs_tcga_mutations	Uterine Carcinosarcoma
uvm_tcga_mutations	Ocular Melanoma

Table 3: TCGA studies used for analysis

A length correction was applied to these frequencies to account for inter-kinase differences in the gene length (and therefore mutational frequency bias) N-terminal from the APE position. A top 30 tumour suppressing kinase list was constructed by ranking each kinase by descending length corrected score.

4.5.7 Critical residue scoring and cross-ref with genomics data

The kinase domain amino acid sequences of each of the top 30 tumour suppressing kinases from the truncating mutation screen were aligned from the glycine rich loop to the APE motif using the Strap Alignment Tool. Amino acid conservation and missense mutation frequency were used to identify mutational hotspots of highly conserved residues. The conservation score was calculated as the number of kinases, out of the top 30, that the most frequently observed residue was observed in. The aligned sequence locations were cross-referenced with mutational data from TCGA and CCLE (same data as the preceding section) to identify all missense mutations occurring in each position. The number of top 30 kinases mutated was designated the mutational score. For enhanced stringency to identify critical residues, only residues with a conservation score above 20 were retained for analysis. Multiplying the mutational score by the conservation score produced a combined score for each residue in the kinase domain sequence. Residues were

ranked by descending combined scores to produce a top list of kinase domain residues based on high conservation and mutational frequency.

4.5.8 Pan cancer critical motif screen

All missense mutations (using the same mutational data as the truncation mutation screen above) occurring in the top 13 highest scoring critical residues were recorded and a gene frequency list was constructed to identify the top kinases with high mutational frequencies of these critical regions.

4.5.9 Structural Analysis / Molecular Dynamics

Dr Natalie Stephenson carried out structural analysis and Molecular Dynamics simulations. Structural modelling of wild type MKK4 and mutants of the conserved glycine residue were carried out using an inactive structure of MKK4 (pdb 3ALN). Molecular dynamics simulations were performed using GROMACS and the GROMOS96 53A6 force field parameter set. The simulated conditions included temperature at 37 degrees Celsius and 1.0 bar of pressure. The movement of the molecules were simulated over a 400ns time period. Further information regarding the parameters of the simulations are included in the methods section of Chapter 5.

4.5.10 DIPNECH variant filtering

Various filtering algorithms were applied to the GATK sequencing output . Previously recorded germline variants were removed from the analysis when required by excluding all variants with a dbSNP reference (either build 146 or 147 of the dbSNP database). For INDEL filtering of germline variants the variant was recorded as matching to a dbSNP if there was an INDEL dbSNP listed at any of the genomic locations that was spanned by the INDEL. Somatic missense mutations present in the carcinoid sequencing paper by Fernandez-Cuesta et al. were used to identify DIPNECH variants occurring in genes that are somatically mutated in pulmonary carcinoid tumours [130]. To identify DIPNECH variants occurring in kinases the list of 411 catalytically active kinases (as used in the truncation mutation screen) was cross-referenced with unfiltered DIPNECH SNP data. Variants occurring in critical regions of the kinase domain of the respective kinases were identified using a script written in R programming language that was modified from the motif location screen (Appendix 3). The script identified the location of the glycine rich loop, lysine of VAIK,

saltbridge-E, HRD, DFG, APE and DxxxxG and cross-referenced with the DIPNECH SNP data for variants at those locations or residues occurring +1 and -1 from those locations. The minor allele frequencies for the matching kinase SNPs were manually checked in the National Center for Biotechnology Information (NCBI) Variation Viewer using the genomic location of the matched variant (http://www.ncbi.nlm.nih.gov/variation/view/).

Chapter Five

Paper 2:

Discrepancies in Cancer Genomic Sequencing Highlight Opportunities for Driver Mutation Discovery

Cancer Research 2014: 74; 6390-6

Priority Report

Cancer Research

Discrepancies in Cancer Genomic Sequencing Highlight Opportunities for Driver Mutation Discovery

Andrew M. Hudson[1], Tim Yates[2], Yaoyong Li[3], Eleanor W. Trotter[1], Shameem Fawdar[1], Phil Chapman[3], Paul Lorigan[4], Andrew Biankin[5], Crispin J. Miller[2,3], and John Brognard[1]

Abstract

Cancer genome sequencing is being used at an increasing rate to identify actionable driver mutations that can inform therapeutic intervention strategies. A comparison of two of the most prominent cancer genome sequencing databases from different institutes (Cancer Cell Line Encyclopedia and Catalogue of Somatic Mutations in Cancer) revealed marked discrepancies in the detection of missense mutations in identical cell lines (57.38% conformity). The main reason for this discrepancy is inadequate sequencing of GC-rich areas of the exome. We have therefore mapped over 400 regions of consistent inadequate sequencing (cold-spots) in known cancer-causing genes and kinases, in 368 of which neither institute finds mutations. We demonstrate, using a newly identified PAK4 mutation as proof of principle, that specific targeting and sequencing of these GC-rich cold-spot regions can lead to the identification of novel driver mutations in known tumor suppressors and oncogenes. We highlight that cross-referencing between genomic databases is required to comprehensively assess genomic alterations in commonly used cell lines and that there are still significant opportunities to identify novel drivers of tumorigenesis in poorly sequenced areas of the exome. Finally, we assess other reasons for the observed discrepancy, such as variations in dbSNP filtering and the acquisition/loss of mutations, to give explanations as to why there is a discrepancy in pharmacogenomic studies, given recent concerns with poor reproducibility of data. *Cancer Res; 74(22); 6390–6. ©2014 AACR.*

Introduction

Personalized therapeutic approaches that target genetically activated drivers have significantly improved patient outcome in a number of common and rare cancers. The development of personalized therapeutics relies on affordable, efficient, and accurate cancer genomic sequencing to identify genetic aberrations present in a given tumor, from which actionable mutations can then be obtained (1). To aid novel driver and targeted therapy discovery, the Sanger Institute (Cambridge, United Kingdom) and Broad Institute (Boston, MA) have developed extensive catalogues of mutations found in a large cohort of cell lines. These resources, which are readily accessible to most biomedical researchers via database portals, have greatly facilitated the process of driver gene discovery. Through an initial evaluation of genetic dependencies in non–small cell lung cancer cell lines, we observed inconsistencies in the mutational profiles as reported by the Sanger Institute's Catalogue of Somatic Mutations in Cancer (COSMIC) database and the Broad Institute's Cancer Cell Line Encyclopedia (CCLE; refs. 2–4). We therefore investigated the extent and causes of these discrepancies to identify opportunities to improve the discovery of driver mutations in oncogenes and tumor suppressors.

Materials and Methods

18 cell line comparison between COSMIC and CCLE data

Commercially available cell lines previously sequenced by COSMIC were identified from the Greenman and colleagues paper (5). Eighteen of these cell lines were also sequenced by CCLE using the Hybrid Capture method using the SureSelect Target Enrichment System (Agilent Technologies) and sequencing on Illumina instruments (76bp paired read ends). Mutational data were downloaded from CCLE website on May 14, 2013 (*CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf*). COSMIC data were downloaded for each cell line from their respective webpages on May 14, 2013. Common genes reported as

[1]Signalling Networks in Cancer Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [2]RNA Biology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [3]Computational Biology Support Team, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [4]University of Manchester and The Christie NHS Foundation Trust, Manchester, United Kingdom. [5]Wolfson Wohl Translational Cancer Research Centre, University of Glasgow, United Kingdom.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Y. Li and E.W. Trotter contributed equally to this article.

Corresponding Authors: John Brognard, Signalling Networks in Cancer Group, Cancer Research UK Manchester Institute, Manchester M20 4BX, United Kingdom. Phone: 44-1613065301; Fax: 44-1614463109; E-mail: John.Brognard@cruk.manchester.ac.uk; and Crispin J. Miller, RNA Biology Group and Computational Biology Support Team, Cancer Research UK Manchester Institute, Manchester, M20 4BX, United Kingdom. Phone: 44-1614463176; Fax: 44-1614463109; E-mail: Crispin.Miller@cruk.manchester.ac.uk

doi: 10.1158/0008-5472.CAN-14-1020

sequenced by both institutes were used to compare both datasets. Script A (Supplementary Data) was written in Groovy programming language to compare the genetic location of missense nontruncating mutations recorded by each institute and compare the lists to find conformity. Sequencing bam files for the CCLE hybrid capture sequencing (COSMIC data unavailable) were viewed using the Integrative Genomics Viewer (IGV: Broad Institute; ref. 6) to categorize the mutations only reported in COSMIC. GC content of the missed mutations was calculated with Ensembl Rest API (version 70) reference genome and capturing the sequence 100 bp either side of the mutation.

568 cell line comparison

COSMIC cell line names were compared with the list of cell lines sequenced by CCLE to find 568 mutually sequenced cell lines. CCLE data were downloaded in the filtered MAF file as described above. COSMIC data were downloaded as a complete file from the COSMIC FTP site on November 12, 2013 (*CosmicCellLineProject_v67_241013.tsv.gv*). The comparison of the sequencing of 1,630 mutually sequenced genes by the two datasets was performed using Script B (Supplementary Data). Mutations were matched by genomic location. Given the variability of gene transcripts from which amino acid changes are calculated, the amino acid change reported was derived from the most common resultant amino acid change and where there was no majority change, the CCLE change was reported followed by COSMIC when comparing COSMIC and CRUK MI data only. CCLE data that were unfiltered (data for common polymorphisms, putative neutral variants, and mutations located outside of the CDS not filtered out) and contained all variants with an allelic fraction >10% were obtained from the CCLE website on November 22, 2013 (*CCLE_hybrid_capture1650_hg19_allVariants_2012.05.07.maf.gz*). The COSMIC-only mutations were cross-referenced against the unfiltered CCLE list to identify further mutation matches. Cancer Census genes were identified from the COSMIC Cancer Census webpage (http://cancer.sanger.ac.uk/cancergenome/projects/census/; ref. 7).

Whole-exome sequencing of four cell lines

Cell lines were obtained from ATCC and DNA extracted within three passages of delivery from ATCC corresponding to less than one month from time of receipt. ATCC authenticates cell lines through short tandem repeat profiling, morphology analysis, cytochrome C oxidase I testing, and karyotyping. On arrival from ATCC, the total passage number for each cell line was H2009 = 23, H2087 = 21, H2122 = 21, H1437 = 46. Cells are maintained in RPMI medium-1640 (Invitrogen) with additional 10% FCS (Lonza Group) and 4 mmol/L GlutaMAX (Invitrogen). Cells are split 1:10 at 80% confluency. DNA extraction is performed using DNeasy Blood and Tissue Kit (Qiagen). Whole-exome sequencing was performed using Agilent Sure Select XT Target Enrichment System for Illumina Pair-end Multiplex Sequencing, enriching with the SureSelect XT Human All Exon V4 library and performing 2 × 100 bp paired-end sequencing on the Illumina HiSeq 2500 with TruSeq SBS v3 chemistry (read density: Supplementary Table S5). Average read density for each sample was calculated using the Lander/Waterman equation as detailed in the Illumina Estimating Coverage Technical Note (http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf). Variant calling was made using the Genome Analysis Tool Kit (GATK: Broad Institute; ref. 8). Comparison of conformity with the COSMIC and CCLE mutation calls was made using Script B with data filtered and unfiltered for mutations with dbSNP ids.

Cold-spot analysis

Bam files from hybrid capture used to create the CCLE database are not available for download so ten independent CCLE whole-exome bam files (performed on Illumina HiSeq 2000) were downloaded on January 9, 2014, via the Cancer Genomics Hub (bam files and metadata with experimental info available from https://browser.cghub.ucsc.edu). These files were analyzed for 986 kinase and Cancer Census genes (Supplementary Table S6), among which 969 genes are protein-coding genes as annotated in ENSEMBL human gene database version 70. The lung cancer sequencing files used were: CCLE-NCI-H2286-DNA-08, CCLE-NCI-H1944-DNA-08, CCLE-COR-L95-DNA-08, CCLE-NCI-H1373-DNA-08, CCLE-NCI-H1184-DNA-08, CCLE-HLF-a-DNA-08, CCLE-JL-1-DNA-08, CCLE-HCC-78-DNA-08, CCLE-DV-90-DNA-08, CCLE-DMS153-DNA-08. The reads in the bam files were mapped onto the reference genome hg19. From each bam file, the read coverage at each base of the protein-coding exonic regions of the 969 selected genes was obtained using Samtools Mpileup (9). Sequencing read cold-spots were defined as protein-coding exonic regions spanning 100 nucleotide bps or more and with the averaged read coverage ≤ 4 at each base. Read cold-spots were identified in the sequencing data and the GC content calculated using the bases corresponding to the read cold-spot. Multiple transcripts of the same gene were removed if the genetic location of the identified cold-spot was identical or the start or end genomic location was the same between same gene transcripts (retaining the transcript with the longest read cold-spot). Top 20 cold-spots are defined as gene transcripts (that were sequenced by CCLE and COSMIC) with the largest cold-spot regions. The average GC content for all coding exons was calculated using the longest transcript (Ensemble Version 70) for each of the 969 genes screened for cold-spots. Circos plots were constructed using the Circos software (from http://www.circos.ca; ref. 10).

Verification of PAK4 mutation

Amplification PCR of region of interest was performed using Phusion High Fidelity PCR Master Mix with H.F. Buffer (New England Biolabs; 12.5 μL) with Betaine 5M (Sigma; 5 μL), 250 ng DNA, forward and reverse primers (Eurofin MWG Operon; 1.25 μL each), and water to make reaction volume of 25 μL. PCR was carried out on S1000 Thermal Cycler (Bio-Rad) with the following PCR steps for a total of 40 cycles; (i) 98.0° 30 seconds; (ii) 98.0° 10 seconds; (iii) 62.0° 30 seconds; and (iv) 72.0° for 150 seconds. PCR product purification was carried out with Illustra ExoProstar Enzymatic PCR and Sequencing Clean-up (GE Healthcare). Sequencing was carried out using an ABI13130 16 capillary system (Life Technologies) and sequencing data were analyzed using 4Peaks software (MekenTosj).

PAK4 transient overexpression

Wild-type PAK4 plasmid (Addgene 23713) was obtained from Addgene (deposited by Hahn and Root; ref. 11). The plasmid was cloned into a Flag-tagged destination vector. STOP codon and the E119Q mutation were introduced by site-directed mutagenesis (Quick Change II Kit, Agilent Technologies). Plasmid was transfected into HEK293T cells in a 12-well format using Attractene according to the manufacturer's protocol. Cells were lysed on ice after 48 hours using Triton X-100 Cell Lysis Buffer supplemented with protease inhibitor tablet (Roche). Lysates were resolved on SDS-PAGE gels followed by Western blotting. Primary antibodies used were: Flag M2 and α- tubulin (Sigma); pERK1/2 (T202/Y204) and pJNK (T183/Y185; Cell Signaling Technology). Mouse or rabbit horseradish peroxidase-conjugated antibodies were used as secondary (Cell Signaling Technology). All Western blot analyses are representative of three independent experiments.

Results and Discussion

We compared missense mutations found in 568 cancer cell lines sequenced by CCLE and COSMIC (v67) across 1,630 mutually sequenced genes (3). A total of 45,377 mutations were reported, of which 26,038 were consistent between institutes (57.38%). A total of 4,496 (9.91%) and 14,843 (32.71%) mutations were found solely by CCLE or COSMIC, respectively (Fig. 1). The ISHIKAWAHERAKLIO02ER cell line, sequenced by both institutes using their standard protocols, showed a total of 263 mutations (52 in COSMIC and 213 in CCLE) but no matches, suggesting different cell lines may have been sequenced. Cross-referencing to Cancer Census genes (7) found that 4,058 mutations reported in one, but not both, of the databases were in known cancer-causing genes (Supplementary Table S1). These included mutations in *EGFR*, *TP53*, *BRAF*, *MAP2K1*, and *PIK3CA* (Table 1), highlighting the difficulties faced when using NGS to identify driver mutations even in well-known cancer-causing genes. Our data reveal a marked discrepancy in mutation reporting between the two most prominent resources and that cross-referencing between the databases is imperative.

We had previously performed a pilot comparison of mutational profiles in 18 cancer cell lines sequenced by the Broad Institute's CCLE using Hybrid Capture sequencing (3), and an earlier release of Sanger Institute's COSMIC database (5, 12). Similar to our larger-scale comparison, we observed low consensus between missense mutation detection in mutually sequenced genes (mean 41.33%; Supplementary Fig. S1). Analyzing the raw read data (6) from CCLE suggested that the most common source of discrepancy was poor sequencing read coverage (41%; Fig. 2). We therefore analyzed 10 randomly selected CCLE whole-exome sequencing files to identify regions of poor coverage (cold-spots). We discovered over 400 cold-spots (100 bp or larger) in Cancer Census and kinase genes that we have mapped as a resource for the research community (Fig. 3 and Supplementary Table S2; ref. 10). These cold-spots are rich in GC nucleotides (63.49% compared with 51.74% average GC-content of all exons in target genes) indicating that high GC-content is a major cause of inadequate sequencing coverage. Importantly, we found for CCLE and COSMIC data combined, an 18-fold reduction in mutation

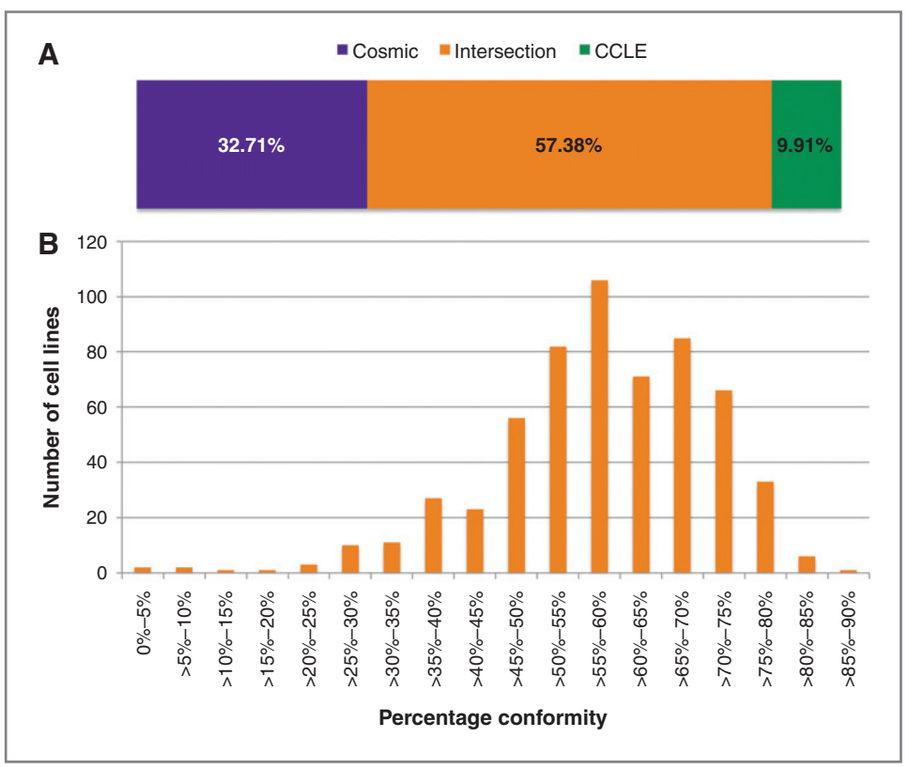


Figure 1. Marked discrepancy is seen in mutation calling between CCLE and COSMIC. A, overall percentage conformity of 46,409 mutations detected by COSMIC and/or CCLE. The intersection between datasets (mutations found by both institutes) accounted for 57.38%. COSMIC-only mutations comprised 32.71% of the dataset and CCLE-only mutations 9.91%. B, the percentage agreement between mutations reported in the 568 cell lines sequenced by both institutes.

Table 1. Mutations in well-known oncogenes and tumor suppressor genes that were detected by only one institute (COSMIC or CCLE)

BRAF	P74A, S76P, V120I, E296K, I326T, **I326V(5)**, R506G, S727G
EGFR	Q71L, R98Q, E282K, S306, V323, K327E, Q408R, L469W, G614S, V654M, G659R, P672R, R677C, T678M, G682V, Q701R, A702D, V738D, A750E, A755D, L815F, L861Q, R973Q, A1076T, T1085N, A1118T, D1127N
FGFR2	R6C, C9S, G89V, E163K, P187S, A315T, Y328S, T341M, A355S, G364E, K401R, L451I, P559H, I643T, C809Y, P814T
HRAS	A11T, G12D, L171P
IDH2	Q95R, H358R, **S408R(2)**
JAK2	V80M, Y96H, T108A, V563I, V617F, L905P, N1129S
KRAS	**G12D(2)**, Q61H, I171M, M188V
MAP2K1	**Q56P(3)**, V85I, A158T, R160K, K185T, V211A
NRAS	**Q61K(6), Q61R(2)**
PIK3CA	K111E, C420R, E542K, E545K, F666L, R770Q
STK11	I46T, Y49D, G56V, K62N, K78N, L105S, D196R, **S216F(2)**, G242W, M392I
TP53	V31I, P47R, D48N, D49H, W53L, A74P, A74S, Y103F, **R110L(2)**, R110P, F113C, F113V, K120N, V122L, C124R, Y126D, M133K, C135R, C176W, E180G, **R181C(2)**, I195T, R213L, **V216L(2)**, V218L, Y220C, N239S, S241F, C242F, M246V, R249S, **R273H(14)**, R283C, R290C, P309S, D324A, R337L, F341C, A347P, G360V, G389W

NOTE: Mutations in bold occurred multiple times (number of occurrences is in parentheses). Supplementary Tables S1a and S1b list the mutations, stratified according to the reporting institute.

density at these loci relative to the remaining exonic regions in the dataset. Extrapolating these data suggests that an additional 1,871 mutations would have been detected in Cancer Census and kinase genes across the 568 cell lines (corresponding to a mean of over three new mutations in Cancer Census or kinase genes per cell line) had the read coverage in the cold-spots been adequate. The *TET2* cold-spot (Fig. 3) is one of the largest of such loci identified, and is not associated with high GC-content. Mutations were reported for this locus in COSMIC, suggesting a sequencing issue specific to the CCLE protocol. This demonstrates that factors, other than inadequate sequencing of GC-content, such as library preparation, reagents, and amplification efficiency can also affect mutation detection at certain loci.

We performed whole-exome sequencing on four of the sequenced lung cancer cell lines (H2009, H1437, H2122, H2087) using an Illumina HiSeq 2500 (achieving over 98% uniquely mapped reads) and a GATK pipeline for mutation detection (8). Our own sequencing identified 27 novel mutations in these four cell lines that were undocumented by COSMIC or CCLE (Supplementary Table S3). Two thirds of these were located in areas of poor read coverage as defined by the CCLE hybrid capture sequencing (less than four reads) but reasonable coverage in our data (mean read depth = 63). The average GC-content 100 bp either side of these newly identified mutations was significantly higher than those where all three institutes were in agreement (60.85% vs. 47.13%, $P = < 10^{-4}$). These findings suggest that the new mutations were previously missed because of being located in GC-rich cold-spots. Although the contribution of factors such as different library preparation and reagents may play a role, our data indicate that NGS efficiency of high GC-rich regions is improving, but earlier datasets are more likely to have missed mutations in GC-rich regions. The majority of The Cancer Genome Atlas and International Cancer Genome Consortium data are of a similar age to CCLE and COSMIC, and therefore subject to similar limitations. Our own more recent sequencing faired better in these regions but still had many GC-rich cold-spots in cancer-associated genes. This is a significant problem, particularly in cancers, including lung cancers, which have a mutational signature predominantly favoring GC-rich trinucleotides (13).

One of the novel mutations identified by our group was in PAK4 (E119Q) in H2009. This mutation lies in a GC-rich (> 76%) area of poor read coverage in CCLE (2 reads; neither reporting the mutation). In contrast, the locus was covered by 39 reads in our data, of which 51% identified the mutation (Supplementary Fig. S2). Given the importance of the PAK kinases in the cancer proliferation and survival pathways (2, 14), we further characterized this mutation. Overexpression of the PAK4 E119Q mutant in 293T cells showed enhanced activation of the ERK pathway compared with the wild-type kinase, suggesting that this is a gain-of-function mutation (Supplementary Fig. S3). These data indicate that additional cancer driver mutations in GC-rich regions will be consistently missed by next-generation cancer genomic sequencing studies, and highlight the potential of developing sequencing platforms to target cold-spot regions for novel cancer gene discovery.

Difference in computational protocols represent another important cause of discrepancy, and includes differences in dbSNP filtering as well as the threshold allelic fraction required to call a mutation. We investigated the effects of dbSNP filtering by comparing the COSMIC-only mutations with unfiltered data from CCLE (the equivalent COSMIC data were

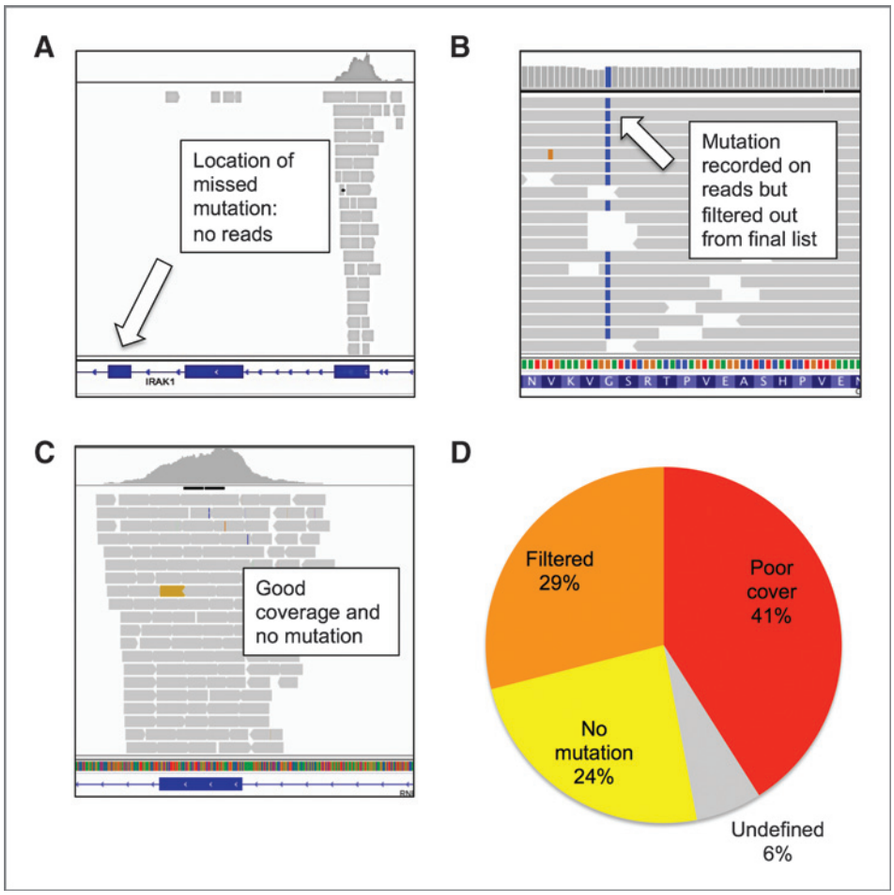


Figure 2. In the original 18 cell line comparison, mutations detected by COSMIC but not CCLE were categorized into: poor coverage with 5 or less reads (A); good read coverage (over 20 reads) and mutation detected on reads but annotated as a dbSNP, neutral variant, outside coding region in all transcripts, or detected on less than 10% of reads, and removed (B); and good coverage, no mutation (C). D, reveals that the most common cause for mutations being missed by CCLE was poor read coverage (41%). Images of read coverage were taken using the Integrative Genomics Viewer.

unavailable). Conformity increased to 67.85% although 10,091 COSMIC-only mutations remained unmatched to CCLE (Supplementary Fig. S4). Therefore, one third of mutations detected only by COSMIC were present on CCLE sequencing reads but were discarded because they were thought to be germline variants. This observation recapitulated the original 18-cell line comparison and our own sequencing also confirmed this with a similar percentage of mutations unreported as a consequence of dbSNP filtering (Supplementary Fig. S5).

By comparing the COSMIC and CCLE data with the four cell lines that we sequenced, we found that 86.34% of the mutations reported by only one database were actually present in our data, suggesting that a minority (approximately 15%–20% based on our two comparisons) of the discrepancy between cell lines is due to acquisition/loss of mutations (Supplementary Table S4). Although a relatively minor factor in our comparisons, the effect of gaining a mutation in a cell line has the potential to greatly affect pharmacogenomic studies. This is highlighted by eight cell lines in the larger comparison that contained activating codon 61 *NRAS* mutations that were reported in only one of the databases (seven reported by COSMIC alone; one by CCLE alone). Analysis of the sequencing data covering the seven *NRAS* mutations not detected by CCLE confirmed good read coverage (mean 220 reads) without evidence of mutation in all seven cases, suggesting loss or gain of the mutation by cell passaging. Passage number is not generally reported in online databases but would greatly assist researchers characterizing the role of specific mutations by indicating whether a mutation has been lost or acquired during passaging.

Although the retrospective nature of our study is unable to control for many sequencing variables such as reagents, polymerases, and platform parameters, we have identified important factors for the discrepancies between the two main cancer genomics databases. These are important findings in the context of a recent study that identified inconsistencies in large pharmacogenomics studies (15). Comparing only 64 genes, this study found some acceptable discrepancies in mutational profiles of cells reported by CCLE and COSMIC but concluded that they were due to differences in the sequencing platforms and variant filtering. Our analysis of a larger panel of genes shows that there is marked discrepancy in sequencing results caused by inadequate sequencing and acquisition of new mutations in addition to variances in dbSNP calling. The authors also concluded that mutational profile was not a major cause of discrepancy in pharmacogenomics data based on the finding that mutational status was not significantly associated with drug response. However, our data show that mutations of

Figure 3. The 20 largest cold-spots detected in cancer census or kinase genes transcripts (of those that were sequenced by both COSMIC and CCLE hybrid capture) using CCLE whole-exome sequencing data. All but one of these cold-spots was located in a high GC-content area and resulted in no mutations being detected by either institute. The *TET2* cold-spot was not located in high-GC content areas and contained mutations detected by COSMIC, indicating that this cold-spot was not present in the COSMIC data. The outer shaded gray plot shows the GC content at each base (calculated as 50 bp either side) with GC content over 70% shaded in red. The middle light green plot shows sequencing read coverage with white troughs representing poor read coverage. The inner three rings record the position of mutations found by both institutes (orange), COSMIC-only (violet), and CCLE (green). Light blue shards show cold-spots over 100 bp in length with the top 20 shaded darker. Data were plotted using a combination of Circos and custom scripts.

cancer-causing genes in sequencing read cold-spots will be frequently undetected, and therefore greatly weaken any analysis attempting to correlate mutation status with drug response. These unsequenced regions of the exome will undoubtedly contain driver mutations, thus mapping cold-spot regions will facilitate novel therapeutic target discovery.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: A.M. Hudson, S. Fawdar, P. Lorigan, A. Biankin, C.J. Miller, J. Brognard
Development of methodology: A.M. Hudson, S. Fawdar, A. Biankin, C.J. Miller, J. Brognard
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A.M. Hudson, E.W. Trotter, J. Brognard
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A.M. Hudson, T. Yates, Y. Li, P. Chapman, C.J. Miller, J. Brognard
Writing, review, and/or revision of the manuscript: A.M. Hudson, Y. Li, P. Lorigan, A. Biankin, C.J. Miller, J. Brognard
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A.M. Hudson, T. Yates, P. Chapman
Study supervision: P. Lorigan, C.J. Miller, J. Brognard

Acknowledgments

The authors thank members of the Signalling Networks in Cancer Group and RNA Biology Groups for helpful discussions, the Core Facility for their advice and support, and Drs. William Newman and Ged Brady for helpful comments and suggestions

Grant Support

This work was fully supported by Cancer Research UK.

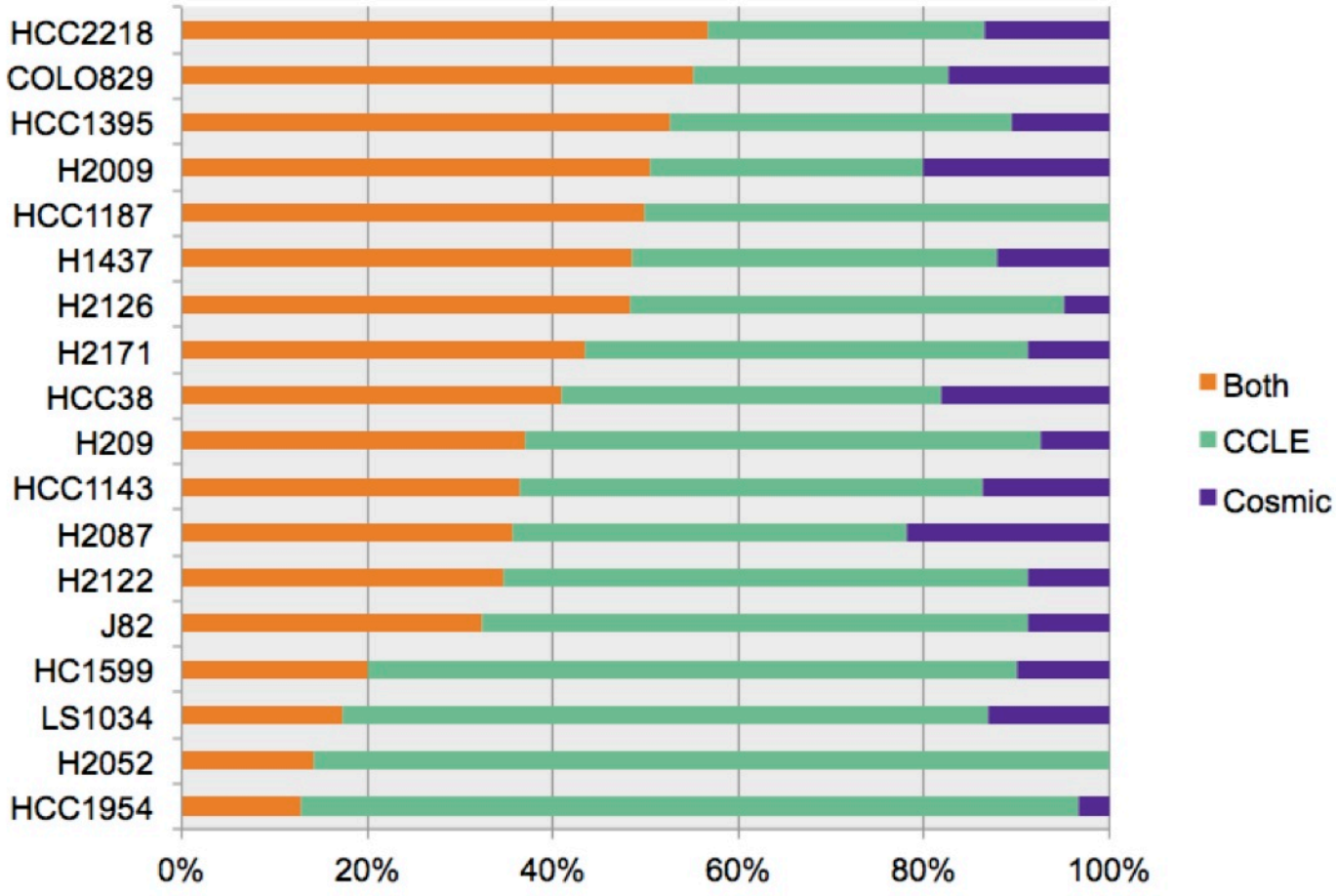
Received April 3, 2014; revised August 14, 2014; accepted September 15, 2014; published OnlineFirst September 25, 2014.

References

1. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, et al. The battle trial: personalizing therapy for lung cancer. Cancer Discov 2011;1:44–53.
2. Fawdar S, Trotter EW, Li Y, Stephenson NL, Hanke F, Marusiak AA, et al. Targeted genetic dependency screen facilitates identification of actionable mutations in FGFR4, MAP3K9, and PAK5 in lung cancer. Proc Natl Acad Sci U S A 2013;110:12426–31.
3. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
4. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 2011;39:945–50.
5. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–8.
6. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
7. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
8. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.
10. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–45.
11. Johannessen CM, Boehm JS, Kim SY, Thomas SR, Wardwell L, Johnson LA, et al. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature 2010;468:968–72.
12. van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, et al. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 2009;41:521–3.
13. Feng Z, Hu W, Hu Y, Tang MS. Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. Proc Natl Acad Sci U S A 2006;103:15404–9.
14. Radu M, Semenova G, Kosoff R, Chernoff J. PAK signalling during the development and progression of cancer. Nat Rev Cancer 2013;14:13–25.
15. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, et al. Inconsistency in large pharmacogenomic studies. Nature 2013;504:389–93.

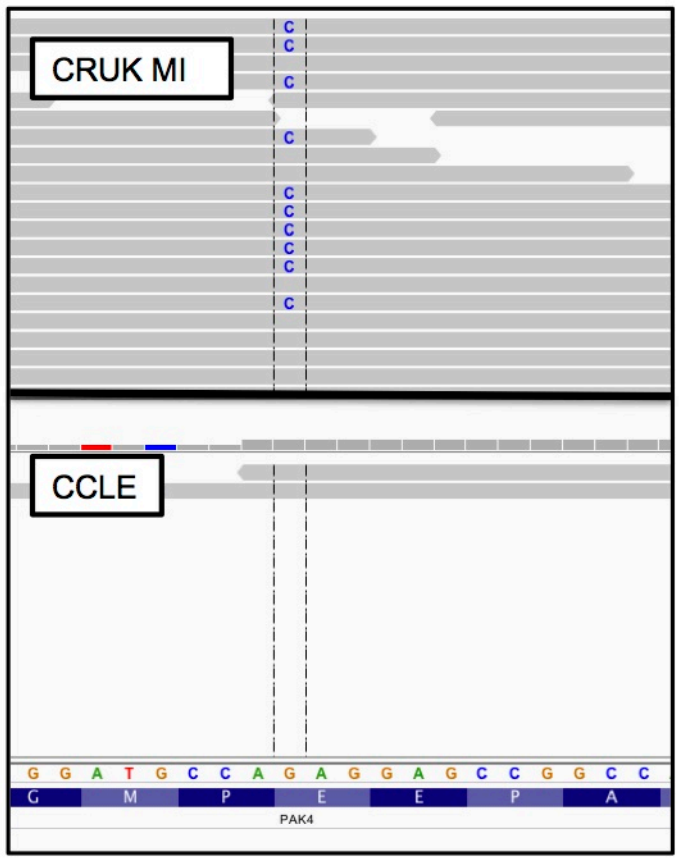
Supp. Figure 1

Original comparison of 18 cell lines sequenced by COSMIC and CCLE show that the conformity of missense mutation detection ranges from 56.75% in HCC2218 cell line to 12.90% in HCC1954.



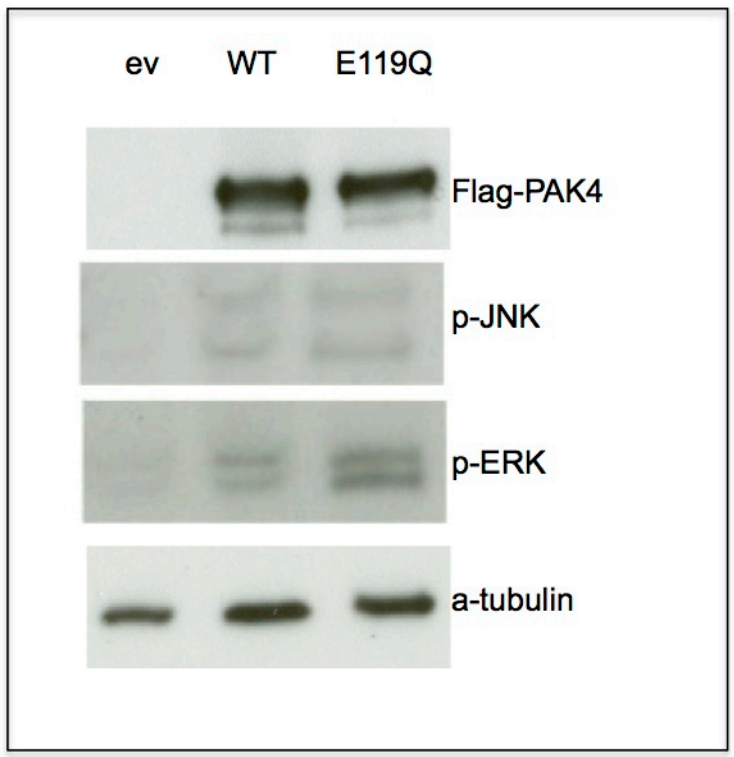
Supp. Figure 2

Images from the Intergrative Genomics Viewer illustrating the read coverage in the CRUK MI sequencing bamfiles and the CCLE hybrid capture bamfiles for the PAK4 p.E119Q mutation in H2009 cell line. The mutation was only reported in the CRUK MI sequencing with mutation seen in 51% of reads in a good coverage area but not reported by CCLE as the area is poorly covered by hybrid capture (only 2 reads and neither showing mutation).



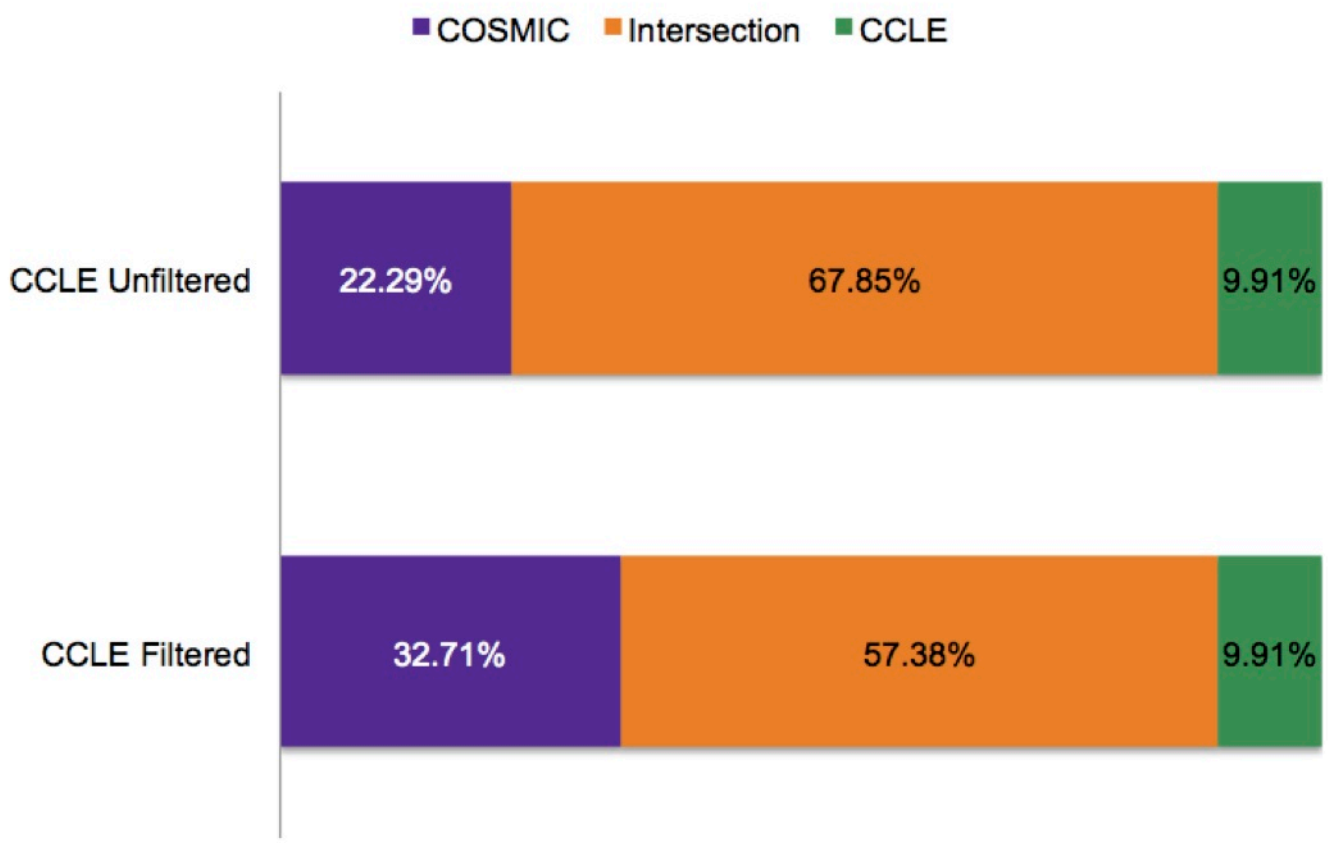
Supp. Figure 3

Western Blot showing PAK4 overexpression in 293T cells. Over-expression of the p.E119Q mutant causes enhanced phosphorylation of ERK compared to over-expression of the wild type plasmid. JNK phosphorylation activity is not affected.



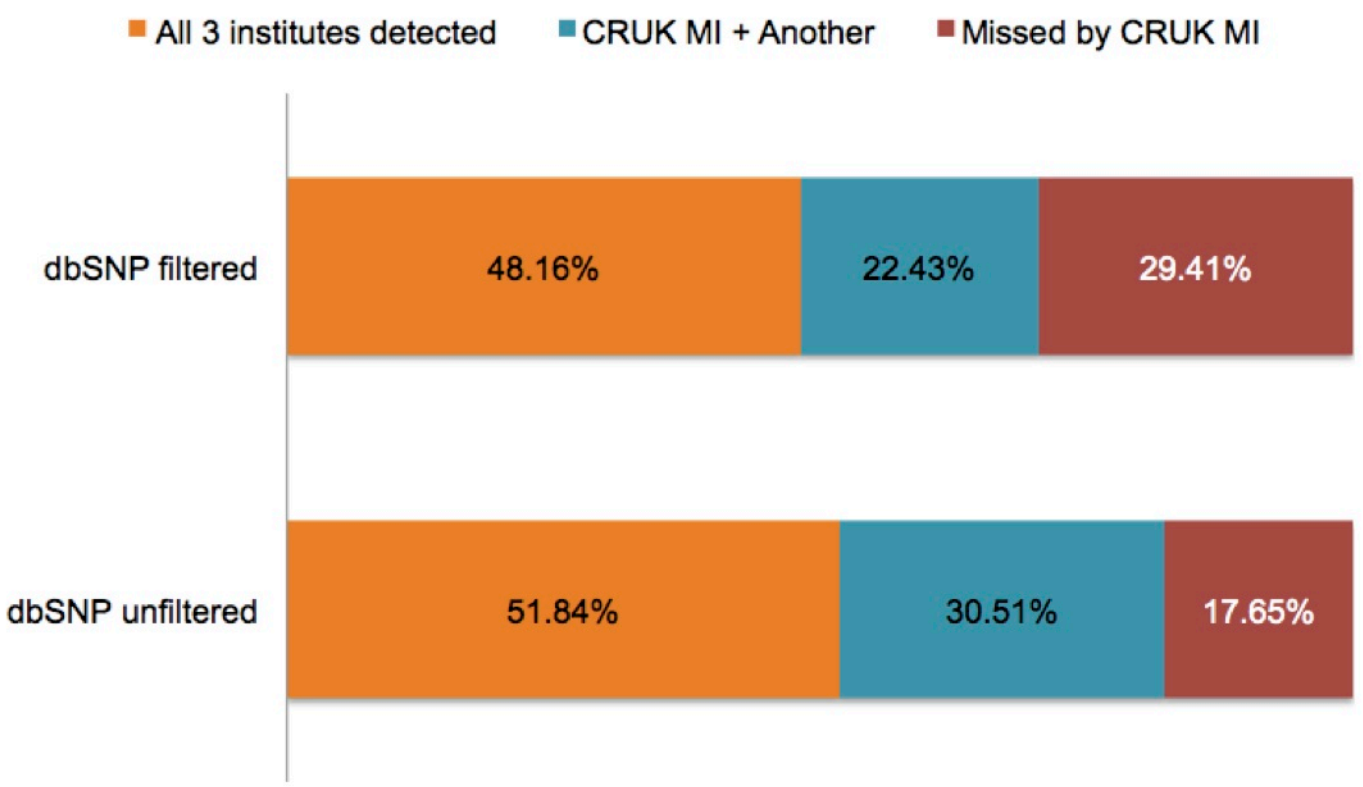
Supp. Figure 4

The comparison of COSMIC and CCLE mutational data for 568 cell lines revealed an overall conformity of 57.38%. Taking the COSMIC only mutations from this analysis and comparing with the dataset from CCLE reported as being unfiltered for germ line SNPs increased the conformity to 67.85% indicating that differences in germ line SNP variant calling is not the sole cause of discrepancy between this data with 10091 unique COSMIC only mutations still remaining. The absence of published germ line SNP unfiltered COSMIC data prevented the opposite analysis being performed with CCLE only mutations.



Supp. Figure 5

Comparison of mutation detection in four cell lines (H2009, H1437, H2122, H2087) between COSMIC and CCLE along with sequencing from our own institute (CRUK MI) reveals 29.41% of mutations missed by CRUK MI. The conformity increases when the comparison is repeated with CRUK MI data that has not been filtered for germ line SNPs with 17.65% now being missed.



Chapter Six

Paper 3:

Functional Screening of Cancer Datasets Identifies Novel Targets and Mutational Hotspots in the Tumour-Suppressing Kinome

Functional Screening of Cancer Datasets Identifies Novel Targets and Mutational Hotspots in the Tumour-Suppressing Kinome.

Andrew M. Hudson[a*], Natalie L. Stephenson[a*], Cynthia Li[a], Eleanor Trotter[a], Gitta Katona[a], Patrycja Bieniasz-Krzywiec[a], Matthew Howell[a], Chris Wirth[b], Crispin J. Miller[b,c], John Brognard[a†]

[a] Signalling Networks in Cancer Group, [b] Computational Biology Support, [c]RNA Biology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester M20 4BX, UK.

* Indicates equal contribution by authors.

[†] To whom correspondence should be addressed. E-mail: John.Brognard@cruk.manchester.ac.uk; Tel: +44(0)161 306 5301

Abstract

A major challenge in cancer genomics, particularly for tumours with a high mutational burden, is identifying genetic drivers from the large numbers of neutral passenger mutations. We present two novel methods of filtering genomic data with functional considerations to identify novel driver mutations that would otherwise be lost in mutational noise. Assessing pan-cancer mutational data of 411 kinases from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopaedia (CCLE) we screened for all truncation mutations occurring N-terminal to the end of the kinase domain to produce a tumour suppressing kinome list. Aligning the kinase domain sequence of these kinases and cross-referencing with TCGA and CCLE allowed the identification of mutational hotspots of conserved residues that we predict are critical for kinase function. We characterise a mutational hotspot in a conserved glycine situated between the activation loop and P+1 loop to show that mutations of this residue frequently cause loss of kinase activity. Next we located this novel critical glycine, in addition to the top 12 critical residues from the mutational hotspot analysis, in 411 kinases. Performing a pan-cancer screen for missense mutations in these 13 critical locations highlighted mutations in *MAP2K7*/MKK7 in gastric adenocarcinoma with a high likelihood of killing catalytic activity. These *MAP2K7* mutations were all validated biochemically as causing profound loss of function of the resultant kinase. Reconstituting wild-type *MAP2K7* in a gastric adenocarcinoma cell line with a *MAP2K7* frameshift mutation confirms the role of MKK7 as a tumour suppressor in this tumour subtype. Filtering next generation sequencing (NGS) data using functional considerations such as these allows the identification of cancer drivers that are not apparent using mutation frequency alone.

It is estimated that within the next 12 months a million cancer samples will have been sequenced by next generation sequencing (NGS) [1]. The greatest challenge now lies in interpreting this data to dissect tumourigenic mechanisms and identify therapeutic targets. A major problem is that the data is often noisy with many inconsequential ‘passenger’ mutations obscuring the detection of driver mutations [2]. This is particularly challenging for tumours with high mutational burdens, such as those associated with strong carcinogens including ultraviolet light and cigarette smoke, where passenger mutations can be a hundred times more numerous than bona fide driver mutations [3]. High mutational burden causes mutational noise both at the level of aggregated tumour subtype studies as well as individual patient data. Aggregating

tumour data from hundreds of cancer samples of the same subtype without a filtering method will only allow the identification of common drivers if they are prevalent enough to be detected above this background mutational noise. Now that most cancer subtypes have been characterised by large-scale sequencing studies we know what many of those common drivers are [4]. However in tumour subtypes such as lung cancer many cases do not posses an identifiable common driver suggesting there is a multitude of lower frequency drivers that we struggle to detect above the noise [5, 6]. The best method to discover these lower frequency cancer drivers is currently under debate [7-9]. Do we continue sequencing more and more samples, or do we shift focus to more functional studies?

In silico methods are already widely used to attempt functional analysis of large genomics data sets [2, 10]. Mutational assessors are tools existing online and within sequencing data pipelines that give a predicted functional score for each mutation based on the amino acid change and the conservation of the locus based on an inter-species comparison [11-14]. Whilst extremely useful these assessors suffer from the limitation that the function of many genes, especially in relation to the changes of their structure, are not very well characterised. As a result there is potential for a significant number of false positive functional mutations that impede driver mutation discovery. Therefore, there is a need to improve genomic analysis and unlock the potential of these huge public datasets. By better linking existing knowledge of a protein's function to the associated structural features offers opportunities for novel functional driver discovery. Protein kinases are a well-characterised class of protein with well-documented mechanisms linking structural motifs to protein function [15-17]. This makes them ideal candidates to develop functional screening methods.

We initially produced a list of candidate tumour suppressing kinases using the frequency of truncating mutations in TCGA and CCLE that were guaranteed to kill catalytic activity. Sequence alignment of this tumour suppressing kinome list allowed the identification of mutational hotspots in conserved regions. The top 12 mutational hotspots were all within motifs already known to be critical for kinase function. We biochemically tested a novel mutational hotspot identified by the analysis in a frequently mutated glycine that is critical for kinase activity in a number of different kinases. We then developed a bioinformatic screen to locate all loss of function (LOF) mutational hotspots in 411 kinases in TCGA and CCLE datasets. Kinases were ranked by the frequency of mutations occurring at LOF motifs. Alongside known tumour suppressors such as BRAF and STK11, we identified and validated a high

incidence of *MAP2K7*/MKK7 LOF mutations in gastric cancer, as an example of the power of this approach. We demonstrate how aligning the sequences of protein families, such as kinases, allows the detection of mutation hotspots in conserved regions and how this approach can be used to identify functionally active mutations above the background passenger mutational noise.

Materials and Methods

Truncation mutation screen

A script was written in R using the StringR package to identify the APE motif in the Genbank sequences of 411 catalytically active kinases with conventional VAIK, HRD and DFG motifs identified in Manning *et al.* (Supplemental. Table 1) [16]. The location of the E of the APE motif was defined as the C-terminal limit of a functional kinase domain. Mutational data from TCGA and CCLE were cross referenced with each kinase APE location to record all truncating mutations occurring N-terminal to this location. Mutational frequencies were length corrected using a mean transcript length (between shortest and longest transcripts) of each kinase to account for inter-kinase differences. A top 30 tumour suppressing kinase list was constructed by ranking kinases by descending length corrected score.

Sequence alignment and residue conservation and mutation scoring

The kinase domain sequence from the first glycine of the GxGxxG loop to APE motif were sequence aligned for each top 30 kinase using the Strap Alignment Tool. A conservation score was calculated as the number of kinases out of the top 30 that the most common amino acid at that position occurs. If a location had more than 5 kinases without a corresponding amino acid (in other words this region is missing from more than 5 or more kinases) the conservation score was calculated as zero. All loci with a conservation score below 20 were removed from the final analysis to leave highly conserved locations. The aligned sequence locations were cross-referenced with the mutational data from TCGA and CCLE to identify mutations at each position and the mutational score was calculated as the number of kinases with a mutation at that position. Multiplying the mutational score by the conservation score produced a combined score to rank each residue in the kinase domain sequence.

Kinase motif based screen

A script was written in the R programming language using the StringR package, which located the top 13 loci (defined by combined score above) within the Genbank sequences for each of the 411 kinases. Mutational data from the TCGA studies obtained via the CGDS-R package (Supplemental. Table 2) and CCLE were cross-referenced with each kinase to identify any point mutations occurring within critical kinase motifs. The genomic mutational data was downloaded on 11th January 2016. A ranked list was constructed from the number of mutations observed per kinase.

Structural Modelling and Molecular Dynamics Simulation

Homology models of WT and mutant MKK4 were created using Modeller 9.16 from PDB ID: 3ALN structure. Molecular dynamics simulations were performed using GROMACS version 5.0 with the GROMOS96 53A6 force field parameter set. All titratable amino acids were assigned their canonical state at physiological pH, short-range interactions were cut off at 1.4 nm and long-range electrostatics were calculated using the particle mesh Ewald summation [18]. Dispersion correction was applied to energy and pressure terms accounting for truncation of van der Waals forces and periodic boundary conditions were applied in all directions. Protein constructs were placed in a cubic box of 100 nM NaCl in simple point charge water with at least 1 nm distance between the protein construct and box edge in all directions. Neutralizing counter ions were added and steepest decent energy minimization was performed, followed by a two-step NVT/NPT equilibration. Both equilibration steps maintained a constant number of particles and temperature, NVT equilibration was performed for 100 ps maintaining a constant volume, followed by 10 ns of NPT equilibration maintaining a constant pressure. Temperature was maintained at 37 ^{o}C by coupling protein and non-protein atoms to separate temperature coupling baths [19], pressure was maintained at 1.0 bar (weak coupling). All position restraints were then removed and simulations were performed for 400 ns using the Nose-Hoover thermostat [20] and the Parrinello-Rahman barostat [21]. Root mean squared fluctuation (RMSF) analysis compared the standard deviation of the atomic position of each a-carbon in the trajectory, fitting to the starting structure as a reference frame. Root mean squared deviation (RMSD) analysis compared the structure of specified groups of residues at each time point of a trajectory with the reference starting structure. Images were created using PyMol version 1.5.0.5.

Plasmids and Transfections

MAP3K13 cDNA was prepared from RNA extracted from HEK293T cells. *PRKCQ*

was bought in the pENTR vector (Ultimate Human ORF Library - Life Technologies), *MAP2K4* and *MAP2K7* were bought in pCMV6-Entry vectors (Origene). Primers containing attB flanking sites were used to amplify up the above constructs before they were inserted into pDONR-221 vector using the BP clonase reaction (Thermo Fisher Scientific). The Gateway LR Clonase system (Thermo Fisher Scientific) was used for cloning from pDONR-221 into a pDEST-FLAG vector created by Dr. Eleanor Trotter from the pReceiver-M12 plasmid (GeneCopoeia). 3X-FLAG *DAPK3* vector was provided by Dr. Timothy Haystead (Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710)

Mutants were created using the Quikchange Site-directed Mutagenesis II Kit (Agilent Technologies) using the manufacturers protocol. Kinase dead mutants were DAPK3 (K42M), MKK4 (K131M), MKK7 (K148M), LZK (K195M) and PRKCQ (K409M). All sequences were confirmed using Sanger sequencing. HEK293T cells or CAL51 cells (for *MAP2K4* transfections) were seeded into 12-well plates (standard transfections) or 6-well plates (immunoprecipitations) and transiently transfected the following day using either Attractene (QIAGEN) for the HEK293T cells or Lipofectamine2000 (Thermo Fisher Scientific) for CAL51 cells according to the manufacturer's protocol.

Protein lysate preparation and immunoblots

24 hours after transfection cells were lysed on ice using Triton X-100 Cell Lysis Buffer (Cell Signalling) supplemented with a protease inhibitor tablet (Roche). Lysates were either resolved on SDS-PAGE gels followed by western blotting or used in an *in vitro* kinase assay (details below). Primary antibodies used were as follows: Flag M2 and a-tubulin (Sigma); MKK7, pJNK (T183/Y185), pMARCKS (S152/S156), pPKC (S676), pThr, pJun (S73) (Cell Signalling Technology); and pSer (Millipore). Mouse or rabbit HRP-conjugated secondary antibodies were used (Cell Signaling). All western blots are representative of at least three independent experiments.

In vitro kinase assays

Cell lysates were incubated with anti-Flag M2 affinity gel (Sigma) for at least 2 hours. Beads were washed with lysis buffer and kinase buffer (Cell Signaling Technology) and a kinase assay was performed in the presence of 200 μM ATP and inactive kinase substrates at 30˚C for 30 min. Autophosphorylation of DAPK3 was assessed using a phospho-Serine antibody [Millipore]. Following addition of 4x reduced SDS

sample buffer, proteins were resolved by SDS-PAGE and analyzed by western blotting.

Generation of *MAP2K4*, *MAP2K7* and *JNK1* tetracycline-inducible cell lines.

Parental CAL51 (*MAP2K4*), CAPAN1 (*MAP2K4*) and IM95 (*MAP2K7*) were used to generate cells with tetracycline-inducible expression. WT plasmids (cloned into pLenti/TO/V5-DEST vector) and pLenti3.3/TR (for tetracycline repressor expression) were transfected into 293FT cells using Lipofectamine2000 to generate a lentiviral stock. Cells were transduced with lentiviral stocks and cell lines generated by antibiotic selection (Blasticidin (Invitrogen) and Geneticin (Gibco)). Tetracycline (Invitrogen) was used to induce expression of WT *MAP2K4* (CAL51, CAPAN1) and WT *MAP2K7* (IM95).

Anchorage-dependent colony formation assay

Cells were seeded at approximately 100 cells/well in a 6-well plate format. The following day, tetracycline was added and cells were left to grow for 3 weeks. Colonies formed were fixed with ice-cold methanol, stained with 0.5% crystal violet (Sigma) solution made up in 25% methanol. Wells were thoroughly washed and air dried. For quantification, 2 ml of 10% acetic acid was added to each well, incubated for 20 mins with shaking and absorbance values read at 595 nm.

Matrigel 3D embedded growth assay

6 well plates were pre-coated with a thin layer of Engelbreth-Holm-Swarm (EHS) (Corning) before cells were seeded at approximately 50,000 cells/well in EHS. 2 ml of media with or without tetracycline was added and fresh media was every 2/3 days for 3 weeks. After 3 weeks cells were stained using 0.05% crystal violet (Sigma) solution made up in 25% methanol. For quantification, colonies formed were counted.

Results

Alignment of predicted tumour suppressing kinome reveals mutational hotspots in conserved regions.

A list of candidate tumour suppressing kinases was identified using truncation mutation frequency in the region within or N-terminal to the kinase domain. The highly

conserved APE motif or its equivalent sequence was located in 411 'catalytically active' human kinases possessing the classical kinase domain motifs from Manning *et al*. This APE motif was used as a conservative C-terminal boundary of functional kinase domains so that all truncating mutations occurring N-terminal to this boundary could be regarded as possessing a high confidence of killing catalytic function. The frequency of truncating mutations found within the TCGA and CCLE datasets occurring N-terminal to this cut-off for each kinase were length corrected (Figure 1) to produce the top 30 kinases by truncating mutation density (Figure 2, Supplemental Table 3). This list comprised of known tumour suppressors such as *MAP2K4*/MKK4 and STK11 along with kinases without a previously published role in tumour suppression [22, 23]. As a proof-of-concept we used the CAPAN1 cell line with a truncation mutation in MKK4, the top ranked tumour suppressor, to show that when WT signalling is restored a significant decrease in anchorage-dependent and anchorage-independent colony forming potential is observed (Supplemental. Figure 1). The kinase domains of these 30 tumour suppressing kinases were then sequence aligned to identify areas of high conservation (Supplemental Table 4). The TCGA and CCLE datasets were queried to capture all missense mutations occurring at each position of the aligned sequences and a combined score based on conservation and mutational frequency was produced for each position (Figure 2). The top 12 residues identified by this combined score were all located within motifs known to be critical for kinase activity (GxGxxG, DFG, HRD, HRD+5, APE, VAIK, and E that salt bridges to critical lysine of VAIK; Table 1). The 13th ranked residue is not located in a known critical motif and corresponds to a glycine that is located in a hinge region between the activation and P+1 loops (at APE motif minus 6 (APE-6)) [17]. This glycine or corresponding residue was located in all 411 kinases and cross-referenced with the TCGA and CCLE genomics data to detect all missense mutations occurring at that position (Supplemental Table 5). Other novel residues occur in the top list at HRD minus 6 residues, HRD minus 7 residues, HRD plus 7 residues and APE minus 3 residues.

Mutations of the APE-6 kill catalytic function in multiple kinases.

Structural modelling demonstrates that APE-6 is located at a critical region that allows the activation loop to move into an active conformation (Figure 3A). The small size of the glycine amino acid that occupies this position in a large number of kinases is essential for the flexibility of this hinge region. Molecular dynamics (MD) simulations of a cancer somatic mutation at this conserved glycine in *MAP2K4*/MKK4 (G265D)

showed a decreased movement of the activation loop compared to that of the WT kinase (Figure 3B). Transient overexpression of *MAP2K4* mutations, G265D and G265C, both seen in cancer samples, demonstrates reduced phosphorylation of the canonical JNK pathway equivalent to a kinase-dead construct (Figure 3C). When WT *MAP2K4* was stably re-expressed in CAL51 cells, harbouring the G265D mutation, a reduction in colony forming potential was observed in both two-dimensional and three-dimensional assays (Figure 3D and E). These data indicate that the G265D MKK4 mutation is a significant LOF driver mutation in the CAL51 cell line. Corresponding glycine mutations observed in other cancer samples in LZK/*MAP3K13* (G315D) and PRKCQ (G541V) were transiently overexpressed and likewise demonstrated reduced activation equivalent to a kinase-dead construct, when phosphorylation of known downstream substrates was assessed (Figure 3F and G). Investigation into the HRD minus 6 (HRD-6) position was also performed. Structural modelling of the HRD-6 residue highlights its close proximity to the R-spine anchoring residue within the aF-helix (Supplemental Figure 2A). The R-spine is formed as part of kinase activation and is critical for catalytic activity [24]. MD simulations of a mutation observed in cancer at HRD-6 in DAPK3 (H131R) showed increased movement of 3 out of the 4 R-spine residues (RS1, RS2 and RS4; Supplemental Figure 2B) *In vitro* kinase assay assessment of DAPK3 H131R mutation demonstrated a loss of kinase activity similar to that observed for kinase dead (Supplemental Figure 2C).

Pan cancer analysis of critical codons highlights a prevalence of LOF MKK7/*MAP2K7* mutations in gastric cancer

Having shown that mutations of APE-6 consistently kills catalytic activity in different kinases, this region was added to the top 12 mutational hotspots of known critical motifs for a pan-cancer analysis of cancer mutation datasets. The TCGA and CCLE datasets were queried for mutations at all 13 residues in 411 kinases. The kinases were ranked by the mutational frequency of the thirteen regions combined (Figure 4A). BRAF was identified as the top hit with 16 mutations throughout both datasets, followed by STK11 and EPHB1 with 11 mutations each and then MKK7 with 10 mutations. BRAF, STK11, EPHB1 and other top hits such as CHEK2 have been previously demonstrated to play a tumour suppressive role in different cancer sub-types [23, 25-27]. MKK7 was selected for further investigation, as 5 out of the 10 detected mutations occurred in a single cancer subtype, gastric adenocarcinoma (Figure 4B). In addition, *MAP2K7* mutations and deletions occur in 7% of cases in the

TCGA gastric adenocarcinoma case series with 40% of the mutated cases possessing more than one mutation, suggesting both alleles are affected. Transient overexpression of the 5 gastric *MAP2K7* mutants in conserved motifs showed LOF on the JNK pathway compared to wild type (WT) (Figure 4C). Furthermore, in the IM95 gastric cell line that harbours an inactivating MKK7 mutation (D290fs) (Supplemental Figure 3), re-expression of WT *MAP2K7* results in a significant decrease in anchorage-dependent colony forming potential (Figure 4D). Considering other genes in the JNK pathway, approximately 22% of gastric cancers harbour alterations in *MAP2K7*, *MAP2K4*, *MAPK8*, *MAPK9*, *MAPK10*, *JUN*, or *ATF2* with a high degree of mutual exclusivity suggesting a significant role for loss of function on the JNK pathway in gastric carcinogenesis (Figure 4E).

Discussion

Cancer genomics has provided large cohorts of data, which has proved extremely valuable for identifying genetic drivers occurring at a high frequency. However, the challenge now lies in dissecting this data to identify those driver mutations that are at a lower frequency, and thus masked by mutational noise. The challenge is even greater for those cancers with a high mutational burden and consequentially higher mutational noise. We present an approach to screen genomic data utilising functional knowledge of the kinase domain structure to identify novel driver mutations that would otherwise be missed.

Truncating mutation frequency was initially used to identify candidate tumour suppressing kinases. Some truncation mutations have been shown to increase catalytic activity by removing regulatory regions occurring C-terminal to the kinase domain [28]. In addition, it is possible that a truncation mutation could cause a kinase to lose their substrate specificity, becoming neomorphic and activating upon alternative pathways [29]. Therefore in our method only truncating mutations occurring N-terminal to the APE motif that we could confidently state would abolish kinase activity by virtue of disrupting the expression of an intact kinase domain. were included. The length-corrected truncation mutation frequency was used to rank all the kinases as potential tumour suppressors. The appearance of two well-known tumour suppressing kinases (STK11 and MKK4) occurring as the top two hits for our screen validated this approach [22, 23].

To identify novel hotspot tumour suppressive residues we aligned the kinase domains of the top 30 tumour suppressing kinases and each residue was assigned a mutational and conservational score. The top 12 residues identified using the product of these two scores occur in well-known motifs critical for kinase function, validating this approach as a method for identifying highly functional mutations. However we accept that by virtue of their conservation and prior recognition there is potential bias in the sequence alignment, which would lead to a higher score for these well-known motifs. We investigated the top novel region identified by this screen, a conserved glycine occurring between the activation and p+1 loops, that is both highly conserved and frequently mutated in different kinases in cancer samples. We demonstrated that mutations seen in cancer samples in this glycine in multiple kinases are loss of function, highlighting this residue as being critical for kinases function. Structural modelling suggests that loss of flexibility at this hinge point results in an inability of the activation loop to move into an active conformation.

Recognising that mutations in the novel glycine, alongside the other top 12 known critical motifs, are highly likely to reduce or abolish kinase catalytic activity we performed a pan-cancer analysis of TCGA and CCLE data to detect all mutations in 13 critical codons in 411 kinases. Like the truncation mutation screen this approach identified known tumour suppressors, many of which were common to both screens (such as STK11 and CHEK2)[23, 30]. Given that the critical codon screen was derived from the truncation screen data some overlap could be expected. However there were also marked differences between the results of the two screens with the archetypal GOF oncogene *BRAF* featuring as the top kinase in the critical motif screen whilst only ranking 88th out of 411 kinases in the truncation screen. Kinase-dead BRAF is thought to act as a scaffold to promote NRAS activation so it is likely that most truncation mutations would not allow this oncogenic mechanism resulting in fewer truncation mutations being observed in the datasets. With this consideration in mind, comparing the results of our two screens may help to identify mechanisms by which kinase dead mutations cause activation upon alternative pathways rather than loss of function of the canonical pathway. Another potential explanation for the differential results of both screens relies on differences of kinase activation. Kinases that dimerise and trans-phosphorylate are potentially more susceptible to the effects of a kinase dead mutant of one allele compared to a truncation mutation of one allele. Many tumour suppressor genes require both alleles to become inactivated as the expression of one functional allele often allows enough protein to maintain function. However in kinases that form dimers to trans-phosphorylate, one kinase dead allele

causes a reduction of 75% activity in kinase function and for tetramers this rises to a 93.75% reduction. Truncation mutants may produce a protein that is insufficient to dimerise and therefore have no effect in the dimer or tetramer trans-phosphorylation activity as long as there is a functional second allele. Therefore there may be a differential in mutation type (truncation or critical codon missense) between tumour suppressors that dimerise or are kinase function dependent and those that act as a scaffold independent of kinase function.

The prevalence of functionally damaging *MAP2K7* mutations in gastric cancer were highlighted using the critical codon screen and these mutations were biochemically validated as playing a tumour suppressing role. Querying mutational data for the whole gene found that *MAP2K7* alterations are observed in 7% of gastric cancers in the TCGA dataset [9]. Even at this moderately high percentage of cases, when ranked by mutational frequency *MAP2K7* is only the 145th most commonly mutated gene in gastric adenocarcinoma, emphasising the difficulty of identifying driver mutations above mutational noise. Gastric cancers carry a high number of mutations per sample although few specific driver mutations are known [31]. By focussing on mutations with a high probability of disrupting kinase function our screen helps remove mutation noise caused by passenger mutations. Using the *MAP2K7* observation as a prompt to query other known pathway members suggests a prominent role of JNK pathway inactivation in almost a quarter of gastric cancers. It is proposed that sequencing more cancer samples will eventually cause lower frequency drivers to become more apparent [32]. Whilst this argument may be true it is clear from our experience with analysis of the TCGA gastric cancer dataset, where 289 cancers have been sequenced and 332 genes have a mutational frequency over 5%, that this process is greatly facilitated when precise functionally derived algorithms are integrated into the pipeline.

Using the human kinome as an example we have shown how the linkage of functional data to NGS data can be beneficial for driver mutation discovery and highlight driver genes that are not apparent using overall mutation frequency of the current number of sequenced cancers available. Kinases are an ideal class of proteins to demonstrate this principle given their well-characterised function and sequence homology. The same approach could be applied to other proteins or functional domains. The key to effectively implement these techniques requires advancing functional and structural research to link to the NGS data. We developed two different screens of the mutation data, which take into consideration different mechanisms of loss of signalling in

cancer networks. Using the truncation screen to produce a tumour suppressing kinome we identified novel residues that are critical for kinase function and are conserved across the kinome. Identifying these additional regions not only enhances the functional screening approaches for NGS analysis but also provides direction for studies into kinase function. In many biomedical fields there has been a rapid growth in sequencing projects with funding diverted away from function studies as a result. Our experience demonstrates the mutual benefits that can be achieved linking functional data to NGS data and highlights the importance of ongoing functional studies.

References

1. Schatz MC. Biological data sciences in genome research. *Genome Res* 25(10), 1417-1422 (2015).
2. Hudson AM, Wirth C, Stephenson NL, Fawdar S, Brognard J, Miller CJ. Using large-scale genomics data to identify driver mutations in lung cancer: Methods and challenges. *Pharmacogenomics* 16(10), 1149-1160 (2015).
3. Alexandrov LB, Nik-Zainal S, Wedge DC *et al*. Signatures of mutational processes in human cancer. *Nature* 500(7463), 415-421 (2013).
4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. *Science* 339(6127), 1546-1558 (2013).
5. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. *The Lancet. Oncology* 12(2), 175-180 (2011).
6. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinformatics* 15, 308 (2014).
7. Ledford H. End of cancer-genome project prompts rethink. *Nature* 517(7533), 128-129 (2015).
8. Marx V. Cancer genomes: Discerning drivers from passengers. *Nat Methods* 11(4), 375-379 (2014).
9. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513(7517), 202-209 (2014).
10. Lawrence MS, Stojanov P, Polak P *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457), 214-218 (2013).
11. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39(17), e118 (2011).
12. Adzhubei IA, Schmidt S, Peshkin L *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 7(4), 248-249 (2010).
13. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7(10), e46688 (2012).
14. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat Protoc* 4(7), 1073-1081 (2009).
15. Taylor SS, Kornev AP. Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem Sci* 36(2), 65-77 (2011).
16. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 298(5600), 1912-1934 (2002).
17. Nolen B, Taylor S, Ghosh G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* 15(5), 661-675 (2004).
18. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald method. *J Chem Phys* 103(19), 8577-8593 (1995).
19. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR. Molecular-dynamics with coupling to an external bath. *J Chem Phys* 81(8), 3684-3690 (1984).
20. Nose S. A unified formulation of the constant temperature molecular-dynamics methods. *J Chem Phys* 81(1), 511-519 (1984).
21. Parrinello M, Rahman A. Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J Appl Phys* 52(12), 7182-7190 (1981).
22. Su GH, Hilgers W, Shekher MC *et al*. Alterations in pancreatic, biliary, and breast carcinomas support mkk4 as a genetically targeted tumor suppressor gene. *Cancer Res* 58(11), 2339-2342 (1998).

23. Marignani PA. Lkb1, the multitasking tumour suppressor kinase. *J Clin Pathol* 58(1), 15-19 (2005).
24. Kornev AP, Haste NM, Taylor SS, Ten Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci U S A* 103(47), 17783-17788 (2006).
25. Heidorn SJ, Milagre C, Whittaker S *et al*. Kinase-dead braf and oncogenic ras cooperate to drive tumor progression through craf. *Cell* 140(2), 209-221 (2010).
26. Kampen KR, Scherpen FJ, Garcia-Manero G *et al*. Ephb1 suppression in acute myelogenous leukemia: Regulating the DNA damage control system. *Mol Cancer Res* 13(6), 982-992 (2015).
27. Stolz A, Ertych N, Bastians H. Tumor suppressor chk2: Regulator of DNA damage response and mediator of chromosomal stability. *Clin Cancer Res* 17(3), 401-405 (2011).
28. Brognard J, Zhang YW, Puto LA, Hunter T. Cancer-associated loss-of-function mutations implicate dapk3 as a tumor-suppressing kinase. *Cancer Res* 71(8), 3152-3161 (2011).
29. Cheung LW, Yu S, Zhang D *et al*. Naturally occurring neomorphic pik3r1 mutations activate the mapk pathway, dictating therapeutic response to mapk pathway inhibitors. *Cancer Cell* 26(4), 479-494 (2014).
30. Hirao A, Kong YY, Matsuoka S *et al*. DNA damage-induced activation of p53 by the checkpoint kinase chk2. *Science* 287(5459), 1824-1827 (2000).
31. Wang K, Yuen ST, Xu JC *et al*. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet* 46(6), 573-582 (2014).
32. Lawrence MS, Stojanov P, Mermel CH *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484), 495-501 (2014).

Figure and Table Legends

Figure 1. Schematic illustrating the truncation mutation screen of 411 kinases to produce top 30 candidate tumour suppressors. A pan cancer analysis across TCGA and CCLE datasets was performed to identify truncating mutations occurring N-terminal to the APE motif. The truncation mutation frequency was length corrected to account for intra-kinase variability in the position of the kinase domain within the overall protein. The kinase domains (GxGxxG to APE motif) of the top 30 kinases determined by length corrected truncation mutation frequency were sequence aligned to identify conserved codons (between the 30 top kinases) and allow re-querying of TCGA and CCLE datasets for mutational frequency at the corresponding location for all 411 'active kinases'.

Figure 2. Output of the truncation mutation screen with top 30 kinases ranked by length corrected truncation mutation frequency. Descending length corrected truncation mutation score ranks the top 30 kinases. A portion of the sequence alignment between DFG and APE motif is displayed with the vertical grey bar representing a portion of poor sequence alignment not displayed. The full alignment is presented in Supplemental Table 4. Shaded aligned positions indicate high conservation. The bar chart above the sequence alignment displays the number of kinases out of the top 30 with mutations at each corresponding position. The line graph below the sequence alignment displays the conservation score for each corresponding position calculated as the number of kinases with the most common amino acid at that position.

Table 1. Top 20 critical codons based on combined score of conservation and mutation frequency in top 30 candidate tumour suppressors. The kinase domains of the top 30 candidate tumour suppressors were aligned to determine a conservation score (number of top 30 kinases with most common amino acid) and mutational score (number of kinases with TCGA or CCLE pan-cancer mutation at that codon). The top 12 critical codons based on combined score are in motifs known to be critical for kinase activity. Novel codons include 13^{th} position (APE -6), 14^{th} position (HRD-6), 15^{th} position (HRD+7), 17^{th} position (HRD-7) and 18^{th} position (APE-3).

Figure 3. Mutations identified in a glycine at position APE-6 are inactivating. A) Position of the conserved APE-6 glycine at a hinge point of the activation loop (shown

in INSR kinase domain). Active kinase conformation is shown in light blue (PDB ID: 1IR3), inactive kinase conformation is shown in bark blue (PDB ID: 1IRK). DFG and APE motifs are shown as sticks, glycine is shown in sticks and spheres. B) MD simulations show a dramatic decrease in the level of movement observed for the activation loop in MKK4 (PDB ID: 3ALN). (i) Root-mean-squared fluctuations (RMSF) of each residue is shown graphically and (ii) structurally, with width and colour of the ribbon showing corresponding level of movement. C) Western blot showing biochemically that mutations in the conserved glycine of MKK4 are LOF towards to the JNK pathway. D) A significant decrease in anchorage-dependent colony forming potential is observed in CAL51 cell line harbouring MKK4 G265D following tetracycline-inducible expression of WT *MAP2K4*. E) A significant decrease in 3D growth is observed following tetracycline-induced expression of WT *MAP2K4* within the CAL51 cell line. Mutations of the conserved APE-6 glycine within LZK/*MAP3K13* (F) and PRKCQ (G) are also LOF.

Figure 4. Critical codon screen identifies MKK7/*MAP2K7* as a target in gastric cancer. A) Table ranking kinases based on the frequency of missense mutations observed within the top 13 critical codons in Table 1 including the novel glycine at APE-6. The top hits include kinases with known tumour suppressive functions (STK11, CHEK2 and TGFBR1) or LOF mutations in carcinogenesis (BRAF). MKK7 was selected for further investigation based on the prevalence of the mutations in a single cancer subtype (5 out of 10 mutations) and no previously documented role of LOF mutations in cancer. B) Schematic highlighting gastric adenocarcinoma mutations identified within key motifs of MKK7. Mutations are indicated by yellow spheres. The E287K mutation (grey sphere) is an additional mutation found in the gastric adenocarcinoma cell line SCH that is annotated in the COSMIC database, as it was not sequenced by CCLE. C) Western blot showing MKK7/*MAP2K7* mutations displayed in (B) are LOF towards downstream substrate JNK. D) 2D colony formation assay with IM95 gastric cell line harbouring LOF MKK7. Parental cell line shows an increase in number of colonies formed following addition of tetracycline. IM95 cell line with tetracycline-inducible expression of *MAP2K7* shows a significant decrease in the number of colonies formed following the addition of tetracycline. E) Query of gastric adenocarcinoma mutation data on www.cbioportal.com reveals 22% of cases possess genetic alterations in *MAP2K7* and JNK pathway genes with a high degree of mutual exclusivity (only alteration positive cases shown in figure).

Figure 1

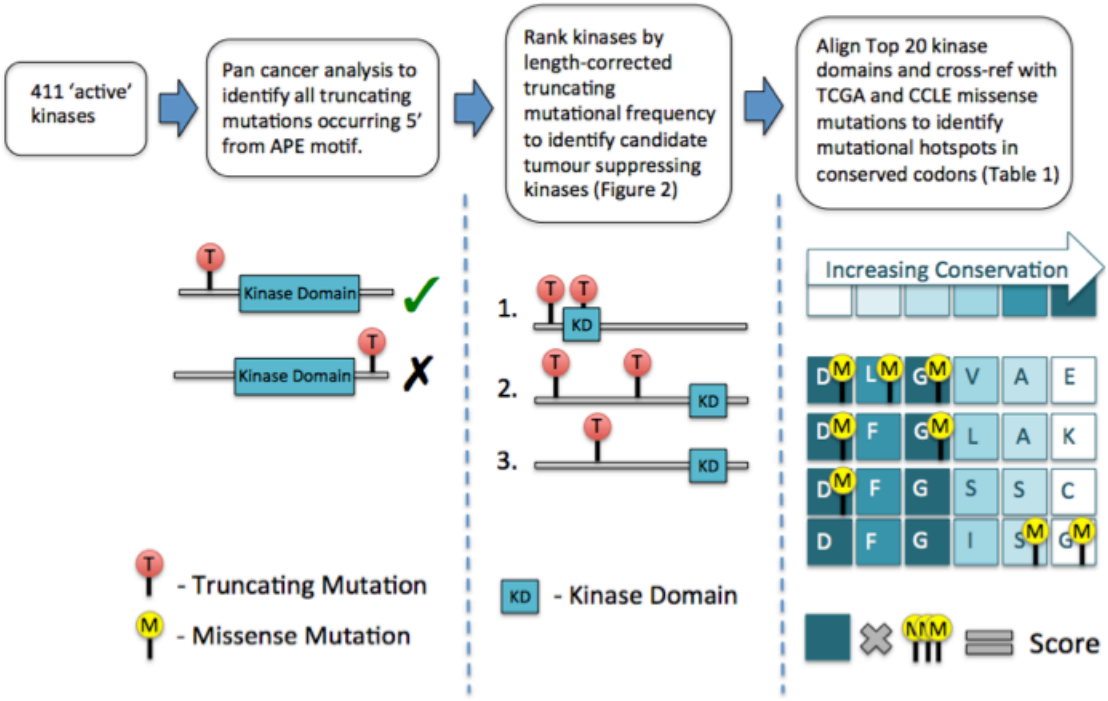

411 'active' kinases
Pan cancer analysis to identify all truncating mutations occurring 5' from APE motif.
Rank kinases by length-corrected truncating mutational frequency to identify candidate tumour suppressing kinases (Figure 2)
Align Top 20 kinase domains and cross-ref with TCGA and CCLE missense mutations to identify mutational hotspots in conserved codons (Table 1)
Kinase Domain
Kinase Domain
1.
2.
3.
KD
Increasing Conservation
- Truncating Mutation
- Missense Mutation
- Kinase Domain
Score

Figure 2

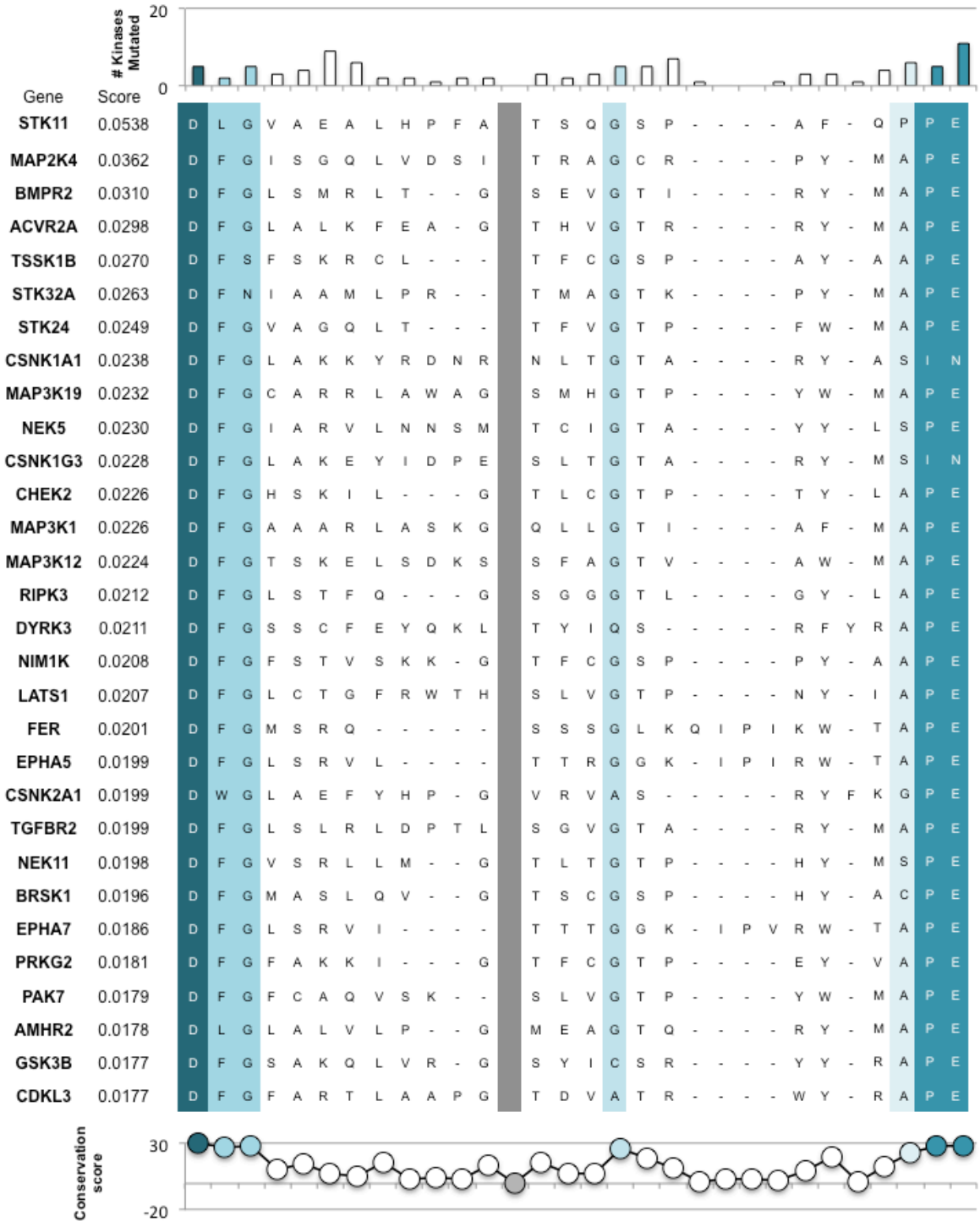

Kinases Mutated
20
0
Gene
Score
STK11 0.0538 D L G V A E A L H P F A T S Q G S P - - - - A F - Q P P E
MAP2K4 0.0362 D F G I S G Q L V D S I T R A G C R - - - - P Y - M A P E
BMPR2 0.0310 D F G L S M R L T - - G S E V G T I - - - - R Y - M A P E
ACVR2A 0.0298 D F G L A L K F E A - G T H V G T R - - - - R Y - M A P E
TSSK1B 0.0270 D F S F S K R C L - - - T F C G S P - - - - A Y - A A P E
STK32A 0.0263 D F N I A A M L P R - - T M A G T K - - - - P Y - M A P E
STK24 0.0249 D F G V A G Q L T - - - T F V G T P - - - - F W - M A P E
CSNK1A1 0.0238 D F G L A K K Y R D N R N L T G T A - - - - R Y - A S I N
MAP3K19 0.0232 D F G C A R R L A W A G S M H G T P - - - - Y W - M A P E
NEK5 0.0230 D F G I A R V L N N S M T C I G T A - - - - Y Y - L S P E
CSNK1G3 0.0228 D F G L A K E Y I D P E S L T G T A - - - - R Y - M S I N
CHEK2 0.0226 D F G H S K I L - - - G T L C G T P - - - - T Y - L A P E
MAP3K1 0.0226 D F G A A A R L A S K G Q L L G T I - - - - A F - M A P E
MAP3K12 0.0224 D F G T S K E L S D K S S F A G T V - - - - A W - M A P E
RIPK3 0.0212 D F G L S T F Q - - - G S G G G T L - - - - G Y - L A P E
DYRK3 0.0211 D F G S S C F E Y Q K L T Y I Q S - - - - - R F Y R A P E
NIM1K 0.0208 D F G F S T V S K K - G T F C G S P - - - - P Y - A A P E
LATS1 0.0207 D F G L C T G F R W T H S L V G T P - - - - N Y - I A P E
FER 0.0201 D F G M S R Q - - - - - S S S G L K Q I P I K W - T A P E
EPHA5 0.0199 D F G L S R V L - - - - T T R G G K - I P I R W - T A P E
CSNK2A1 0.0199 D W G L A E F Y H P - G V R V A S - - - - - R Y F K G P E
TGFBR2 0.0199 D F G L S L R L D P T L S G V G T A - - - - R Y - M A P E
NEK11 0.0198 D F G V S R L L M - - G T L T G T P - - - - H Y - M S P E
BRSK1 0.0196 D F G M A S L Q V - - G T S C G S P - - - - H Y - A C P E
EPHA7 0.0186 D F G L S R V I - - - - T T T G G K - I P V R W - T A P E
PRKG2 0.0181 D F G F A K K I - - - G T F C G T P - - - - E Y - V A P E
PAK7 0.0179 D F G F C A Q V S K - - S L V G T P - - - - Y W - M A P E
AMHR2 0.0178 D L G L A L V L P - - G M E A G T Q - - - - R Y - M A P E
GSK3B 0.0177 D F G S A K Q L V R - G S Y I C S R - - - - Y Y - R A P E
CDKL3 0.0177 D F G F A R T L A A P G T D V A T R - - - - W Y - R A P E
Conservation score
30
-20

Table 1

Rank	Description	Mutation Frequency	Conservation Score	Combined Score
1	AP[E]	11	28	308
2	Salt Bridge E	8	30	240
3	V[A]IK	8	27	216
4=	Gx[G]xxG	7	30	210
4=	HR[D]	7	30	210
6	H[R]D	7	28	196
7	[D]FG	5	30	150
8	HRD plus 5	5	29	145
9	[H]RD	5	28	140
9=	DF[G]	5	28	140
9=	A[P]E	5	28	140
12	[A]PE	6	23	138
13	APE minus 6 - typically G	5	26	130
14	HRD minus 6 - typically H	5	25	125
15	GxGxx[G]	4	24	96
15=	HRD plus 7	4	24	96
17	HRD minus 7	4	20	80
18=	VAI[K]	2	30	60
18=	APE minus 3	3	20	60
20	[G]xGxxG	2	28	56

Figure 3

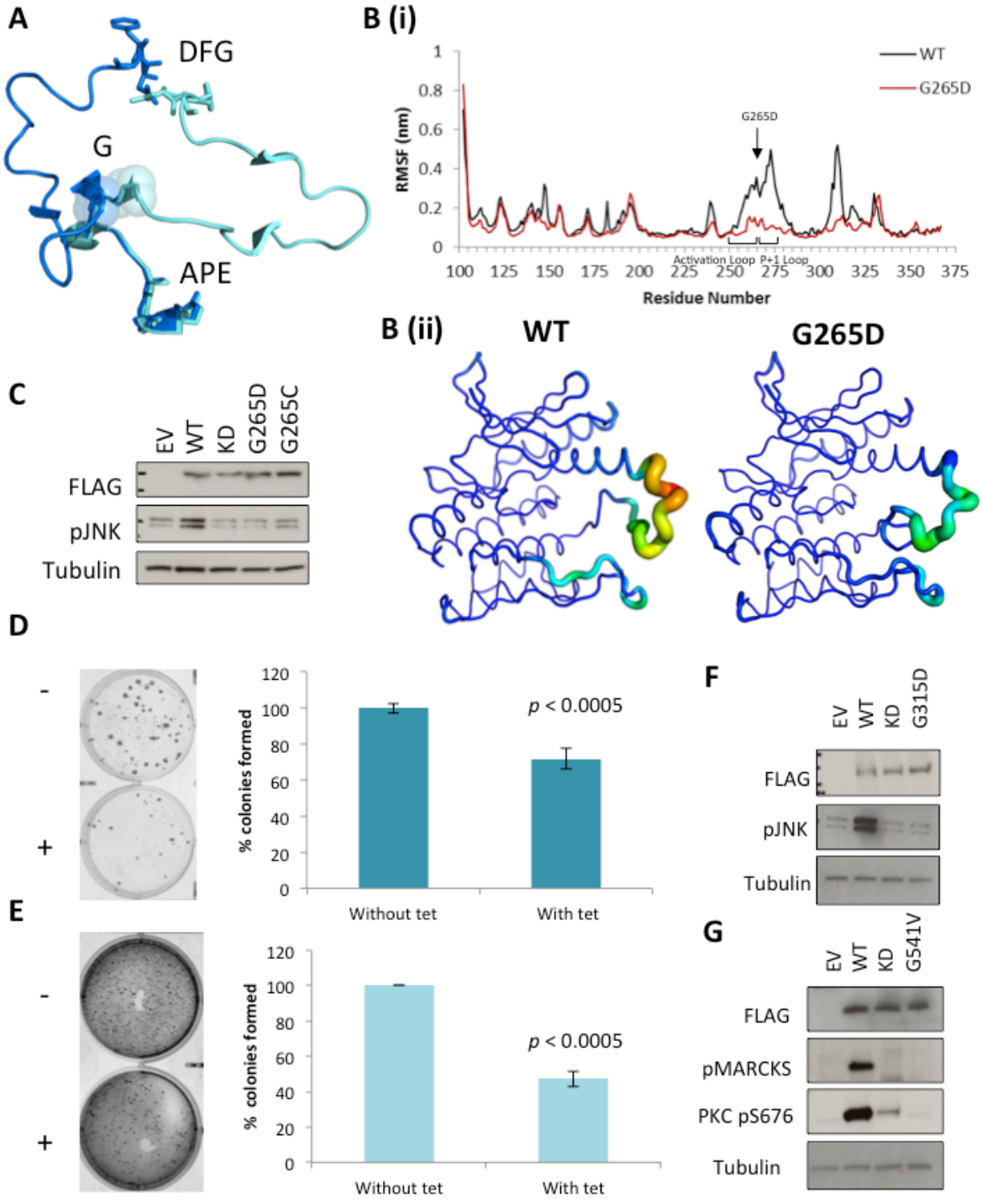
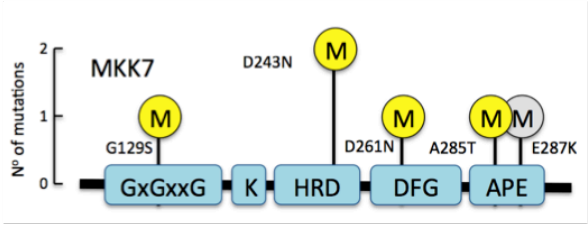

A
DFG
G
APE
B (i)
WT
G265D
RMSF (nm)
G265D
Activation Loop
P+1 Loop
Residue Number
B (ii)
WT
G265D
C
EV
WT
KD
G265D
G265C
FLAG
pJNK
Tubulin
D
-
+
% colonies formed
p < 0.0005
Without tet
With tet
E
-
+
% colonies formed
p < 0.0005
Without tet
With tet
F
EV
WT
KD
G315D
FLAG
pJNK
Tubulin
G
EV
WT
KD
G541V
FLAG
pMARCKS
PKC pS676
Tubulin

Figure 4

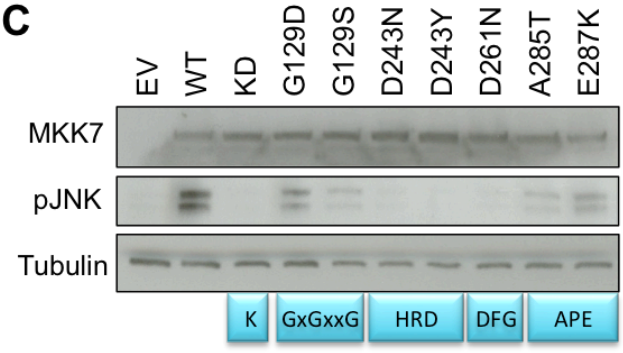
A

Rank	Gene	Mutations
1	*BRAF*	16
2=	*EPHB1*	11
2=	*STK11*	11
4	*MAP2K7*	10
5=	*CHEK2*	8
5=	*MYO3A*	8
5=	*TGFBR1*	8
8=	*DYRK3*	7
8=	*PRKG1*	7

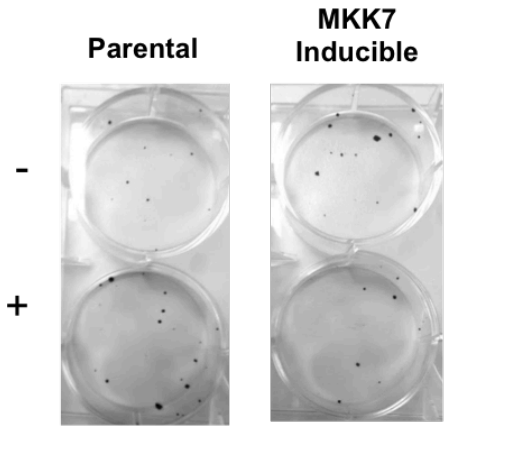
B



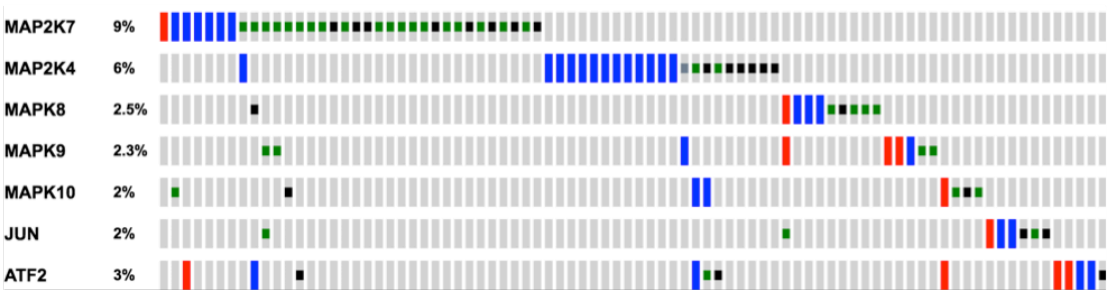
C



D



E



Note: Figure displays the 79 cases with alterations in the 7 pathway genes. The remaining 291 cases without alteration are not displayed.

Supplemental Table and Figure Legends

Supplemental Table 1. 411 kinases screened

Supplemental Table 2. TCGA studies screened

Supplemental Table 3. Top 30 candidate tumour suppressive kinases identified by truncation mutation screen.

Supplemental Table 4. Alignment of top 30 candidate tumour suppressive kinases from GxGxxG motif to APE motif. On enclosed USB stick.

Supplemental Table 5. All missense mutations of APE-7 codon present in TCGA and CCLE datasets.

Supplemental Figure 1. Tetracycline inducible expression of wild-type *MAP2K4* in the CAPAN1 cell line results in significantly decreased anchorage dependent colony formation.

Supplemental Figure 2. A) HRD-6 residue lies in close proximity to the R-spine anchoring residue within aF helix (orange sticks). R-spine shown as yellow spheres, HRD-6 residue shown as red sticks (PKA - PDB ID: 1ATP). B) Root-mean-squared fluctuations (RMSF) from MD simulations highlight increased movement around R spine residues RS1, RS2 and RS4 in DAPK3 H131R (PDB ID: 3BHY). C) *In vitro* kinase assay shows decreased kinase activity within DAPK3 H131R as determined by decreased autophosphorylation.

Supplemental Figure 3. A) Sequence alignment of WT and D290fs *MAP2K7* shows amino acid sequence change caused by the single nucleotide frameshift deletion. Asterix (*) highlight sequence conservation and loss of normal sequence 4 amino acids C-terminal to APE motif and premature termination of sequence 9 codons upstream of the normal termination site. B) Transient expression of D290fs in 293T cell line for 48 hours demonstrates reduced activation of JNK pathway versus wild type evaluated using JNK phosphorylation.

Supplemental Table 1

AAK1	CAMK4	CSF1R	EPHB2	IRAK3	MAP3K5
AATK	CAMKK1	CSK	EPHB3	IRAK4	MAP3K6
ABL1	CAMKK2	CSNK1A1	EPHB4	ITK	MAP3K7
ABL2	CDC42BPA	CSNK1A1L	ERBB2	JAK1	MAP3K8
ACVR1	CDC42BPB	CSNK1D	ERBB4	JAK2	MAP3K9
ACVR1B	CDC42BPG	CSNK1E	ERN1	JAK3	MAP4K1
ACVR1C	CDC7	CSNK1G1	ERN2	KALRN	MAP4K2
ACVR2A	CDK1	CSNK1G2	FER	KDR	MAP4K3
ACVR2B	CDK10	CSNK1G3	FES	KIAA1804	MAP4K4
ACVRL1	CDK11A	CSNK2A1	FGFR1	KIT	MAP4K5
ADRBK1	CDK12	CSNK2A2	FGFR2	LATS1	MAPK1
ADRBK2	CDK13	DAPK1	FGFR3	LATS2	MAPK10
AKT1	CDK14	DAPK2	FGFR4	LCK	MAPK11
AKT2	CDK15	DAPK3	FGR	LIMK1	MAPK12
AKT3	CDK16	DCLK1	FLT1	LIMK2	MAPK13
ALK	CDK17	DCLK2	FLT3	LMTK2	MAPK14
AMHR2	CDK18	DCLK3	FLT4	LMTK3	MAPK15
ANKK1	CDK19	DDR1	FRK	LRRK1	MAPK3
ARAF	CDK2	DDR2	FYN	LRRK2	MAPK4
AURKA	CDK20	DMPK	GAK	LTK	MAPK6
AURKB	CDK3	DSTYK	GRK1	LYN	MAPK7
AURKC	CDK4	DYRK1A	GRK4	MAK	MAPK8
AXL	CDK5	DYRK1B	GRK5	MAP2K1	MAPK9
BLK	CDK6	DYRK2	GRK6	MAP2K2	MAPKAPK2
BMP2K	CDK7	DYRK3	GRK7	MAP2K3	MAPKAPK3
BMPR1A	CDK8	DYRK4	GSK3A	MAP2K4	MAPKAPK5
BMPR1B	CDK9	EGFR	GSK3B	MAP2K5	MARK1
BMPR2	CDKL1	EIF2AK1	HCK	MAP2K6	MARK2
BMX	CDKL2	EIF2AK2	HIPK1	MAP2K7	MARK3
BRAF	CDKL3	EIF2AK3	HIPK2	MAP3K1	MARK4
BRSK1	CDKL4	EIF2AK4	HIPK3	MAP3K10	MAST1
BRSK2	CDKL5	EPHA1	HIPK4	MAP3K11	MAST2
BTK	CHEK1	EPHA2	HUNK	MAP3K12	MAST3
CAMK1	CHEK2	EPHA3	ICK	MAP3K13	MAST4
CAMK1D	CHUK	EPHA4	IGF1R	MAP3K14	MATK
CAMK1G	CIT	EPHA5	IKBKB	MAP3K15	MELK
CAMK2A	CLK1	EPHA6	IKBKE	MAP3K19	MERTK
CAMK2B	CLK2	EPHA7	INSR	MAP3K2	MET
CAMK2D	CLK3	EPHA8	INSRR	MAP3K3	MINK1
CAMK2G	CLK4	EPHB1	IRAK1	MAP3K4	MKNK1

MKNK2	PIK3R4	ROCK1	TEC
MOK	PIM1	ROCK2	TEK
MOS	PIM2	ROR1	TESK1
MST1R	PIM3	ROR2	TESK2
MUSK	PINK1	ROS1	TGFBR1
MYLK	PKMYT1	RPS6KB1	TGFBR2
MYLK2	PKN1	RPS6KB2	TIE1
MYLK3	PKN2	RYK	TLK1
MYLK4	PKN3	SBK1	TLK2
MYO3A	PLK1	SBK2	TNIK
MYO3B	PLK2	SGK1	TNK1
NEK1	PLK3	SGK2	TNK2
NEK10	PLK4	SGK3	TNNI3K
NEK11	PNCK	SGK494	TRIO
NEK2	PRKAA1	SIK1	TSSK1B
NEK3	PRKAA2	SIK2	TSSK2
NEK4	PRKACA	SIK3	TSSK3
NEK5	PRKACB	SLK	TSSK4
NEK6	PRKACG	SNRK	TSSK6
NEK7	PRKCA	SRC	TTBK1
NEK8	PRKCB	SRMS	TTBK2
NEK9	PRKCD	STK10	TTK
NIM1K	PRKCE	STK11	TXK
NLK	PRKCG	STK16	TYK2
NRK	PRKCH	STK17A	TYRO3
NTRK1	PRKCI	STK17B	UHMK1
NTRK2	PRKCQ	STK24	ULK1
NTRK3	PRKCZ	STK25	ULK2
NUAK1	PRKD1	STK26	ULK3
NUAK2	PRKD2	STK3	VRK1
OXSR1	PRKD3	STK32A	VRK2
PAK1	PRKG1	STK32B	WEE1
PAK2	PRKG2	STK32C	WEE2
PAK3	PRKX	STK33	YES1
PAK4	PRPF4B	STK35	ZAK
PAK6	PSKH1	STK36	ZAP70
PAK7	PTK2	STK38	
PASK	PTK2B	STK38L	
PBK	PTK6	STK39	
PDGFRA	RAF1	STK4	
PDGFRB	RET	SYK	
PDIK1L	RIPK1	TAOK1	
PDPK1	RIPK2	TAOK2	
PHKG1	RIPK3	TAOK3	
PHKG2	RIPK4	TBK1	

Supplemental Table 2

genetic_profile_id	Cancer Subtype
acc_tcga_mutations	Adrenalcortical
blca_tcga_mutations	Bladder
brca_tcga_mutations	Breast
brca_tcga_pub_mutations	Breast
brca_tcga_pub2015_mutations	Breast
cesc_tcga_mutations	Cervix
coadread_tcga_mutations	Colorectal
coadread_tcga_pub_mutations	Colorectal
gbm_tcga_mutations	Glioblastoma
gbm_tcga_pub_mutations	Glioblastoma
gbm_tcga_pub2013_mutations	Glioblastoma
hnsc_tcga_mutations	Head and Neck
kich_tcga_mutations	Kidney Chromophobe
kich_tcga_pub_mutations	Kidney Chromophobe
kirc_tcga_mutations	Kidney Renal Cell
kirc_tcga_pub_mutations	Kidney Renal Cell
kirp_tcga_mutations	Kidney Renal Papillary Cell
laml_tcga_mutations	Acute Myeloid Leukaemia
lgg_tcga_mutations	Low Grade Glioma
lihc_tcga_mutations	Liver Hepatocellular
luad_tcga_mutations	Lung Adenocarcinoma
lusc_tcga_mutations	Lung Squamous
lusc_tcga_pub_mutations	Lung Squamous
ov_tcga_mutations	Ovarian
ov_tcga_pub_mutations	Ovarian
paad_tcga_mutations	Pancreas
pcpg_tcga_mutations	Phaeochromocytoma and Paraganglioma
prad_tcga_mutations	Prostate
prad_tcga_pub_mutations	Prostate
skcm_tcga_mutations	Melanoma
stad_tcga_mutations	Stomach
thca_tcga_mutations	Thyroid
ucec_tcga_mutations	Uterine Corpus Endometrial
ucec_tcga_pub_mutations	Uterine Corpus Endometrial
ucs_tcga_mutations	Uterine Carcinosarcoma
uvm_tcga_mutations	Ocular Melanoma

Supplemental Table 3

Gene	Truncation mutations	endCAT	shortestSCORE	longestSCORE	meanSCORE
STK11	12	446	0.05381	0.05381	0.05381
MAP2K4	10	557	0.03620	0.03620	0.03620
BMPR2	12	772	0.03109	0.03109	0.03109
ACVR2A	11	738	0.02981	0.02981	0.02981
TSSK1B	5	370	0.02703	0.02703	0.02703
STK32A	5	380	0.02632	0.02632	0.02632
STK24	5	402	0.02488	0.02488	0.02488
CSNK1A1	5	820	0.02484	0.02283	0.02384
MAP3K19	23	4755	0.02786	0.01859	0.02323
NEK5	4	348	0.02299	0.02299	0.02299
CSNK1G3	5	219	0.02283	0.02283	0.02283
CHEK2	9	1134	0.02311	0.02209	0.02260
MAP3K1	32	2838	0.02255	0.02255	0.02255
MAP3K12	7	626	0.02236	0.02236	0.02236
RIPK3	4	378	0.02116	0.02116	0.02116
DYRK3	8	758	0.02111	0.02111	0.02111
NIM1K	5	240	0.02083	0.02083	0.02083
LATS1	19	1840	0.02065	0.02065	0.02065
FER	14	1779	0.02094	0.01918	0.02006
EPHA5	16	3870	0.02096	0.01885	0.01991
CSNK2A1	4	201	0.01990	0.01990	0.01990
TGFBR2	9	906	0.01987	0.01987	0.01987
NEK11	4	404	0.01980	0.01980	0.01980
BRSK1	4	416	0.01963	0.01963	0.01963
EPHA7	15	1612	0.01861	0.01861	0.01861
PRKG2	11	2422	0.01845	0.01774	0.01810
PAK7	11	1226	0.01794	0.01794	0.01794
AMHR2	7	780	0.01795	0.01759	0.01777
GSK3B	4	226	0.01770	0.01770	0.01770
CDKL3	3	170	0.01765	0.01765	0.01765

Supplemental Table 4

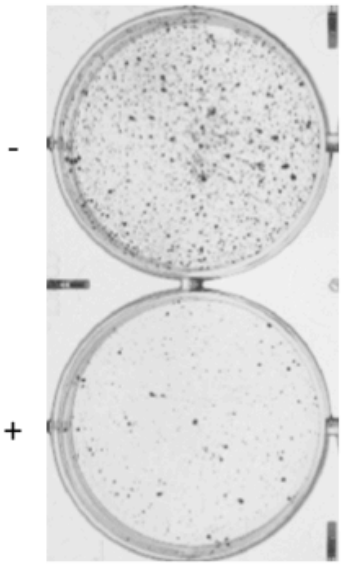
On enclosed USB stick

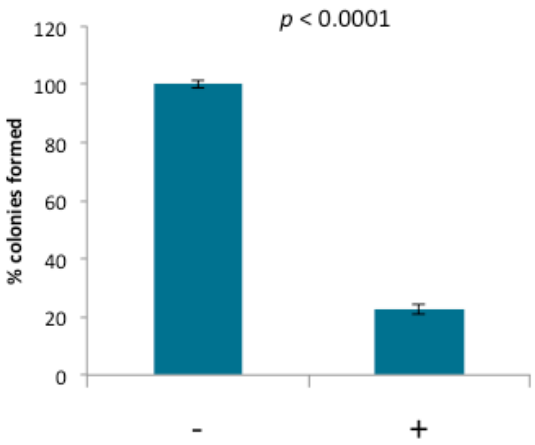
Supplemental Table 5

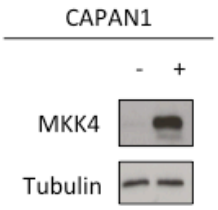
Gene	Study/Cell Line	Chr	Position	Mutation
ADRBK2	skcm_tcga	22	26091075	G352S
AKT1	CCLE_MFE319_ENDOMETRIUM	14	105239613	G311D
AKT1	skcm_tcga	14	105239613	G311D
AURKB	stad_tcga	17	8108691	G235D
BRSK1	lusc_tcga	19	55805501	G192A
CDK15	lusc_tcga	2	202700416	V261L
CDK15	ov_tcga	2	202700416	V261M
CDK20	blca_tcga	9	90584799	G200V
CDK8	CCLE_HCT116_COLORECTAL	13	26959417	V195A
CDKL2	CCLE_HEC1A_ENDOMETRIUM	4	76532422	A163T
CSNK1E	luad_tcga	22	38696770	G175V
CSNK1E	skcm_tcga	22	38696771	G175S
CSNK1G3	CCLE_SNU81_COLORECTAL	5	122911616	G211E
DCLK2	CCLE_EN_ENDOMETRIUM	4	151160985	G570D
DCLK3	lusc_tcga	3	36763059	G515A
DMPK	kirp_tcga	19	46281110	G243S
EIF2AK3	hnsc_tcga	2	88870420	G986V
EPHA2	CCLE_SNU349_KIDNEY	1	16458364	G776D
EPHA7	luad_tcga	6	93964513	G795V
EPHA7	skcm_tcga	6	93964514	G795S
FLT4	lusc_tcga	5	180043372	G1072S
GRK7	ov_tcga	3	141526491	G352A
JAK1	hnsc_tcga	1	65303626	S1043I
JAK3	lihc_tcga	19	17942056	G987S
KALRN	CCLE_LS411N_COLORECTAL	3	124437880	G2842W
KDR	skcm_tcga	4	55956127	G1063E
MAP2K4	CCLE_CAL51_BREAST	17	12016658	G276D
MAP2K4	luad_tcga	17	12016657	G265C
MAP3K13	CCLE_SKNDZ_AUTONOMIC	3	185165669	G315D
MAP3K7	CCLE_KM12_COLORECTAL	6	91266255	G191R
MAP4K2	lusc_tcga	11	64568595	G173V
MAP4K5	CCLE_SNU1040_COLORECTAL	14	50941807	G177D
MAPK7	stad_tcga	17	19284190	A223V
MAPKAPK2	CCLE_EN_ENDOMETRIUM	1	206903425	Y225H
MOS	stad_tcga	8	57025820	G241D
MYLK4	luad_tcga	6	2679608	G265R
MYO3A	skcm_tcga	10	26305800	G187E
NEK8	CCLE_UACC257_SKIN	17	27062264	G165S
PAK3	ucec_tcga	23	110439732	G460E
PHKG1	lusc_tcga	7	56149938	G218W
PHKG2	CCLE_HUH28_BILIARY_TRACT	16	30767511	G189R
PIK3R4	ucec_tcga	3	130463488	R192M

PINK1	lusc_tcga	1	20975099	G409R
PRKAA2	lusc_tcga	1	57159485	G175R
PRKCA	skcm_tcga	17	64738852	G500K
PRKCG	skcm_tcga	19	54403977	G517R
PRKCQ	luad_tcga	10	6498661	G541V
PRKG1	skcm_tcga	10	54041970	G535R
PRKG2	CCLE_NCIH2172_LUNG	4	82096001	G192R
RIPK4	CCLE_JHUEM2_ENDOMETRIUM	21	43171332	G183D
ROR2	luad_tcga	9	94486827	G650V
ROR2	skcm_tcga	9	94486827	G650E
RPS6KB2	CCLE_CGTHW1_THYROID	11	67200497	G231S
RPS6KB2	CCLE_SW579_THYROID	11	67200497	G231S
SBK1	stad_tcga	16	28331595	G210S
STK26	skcm_tcga	23	131202542	G181E
TESK1	skcm_tcga	9	35607612	G219S
TNIK	skcm_tcga	3	170906561	G190E
VRK1	ucec_tcga	14	97321652	G223D
WEE1	lihc_tcga	11	9597784	G264S
WEE2	skcm_tcga	7	141424038	G395E
YES1	ucs_tcga	18	736808	G431S

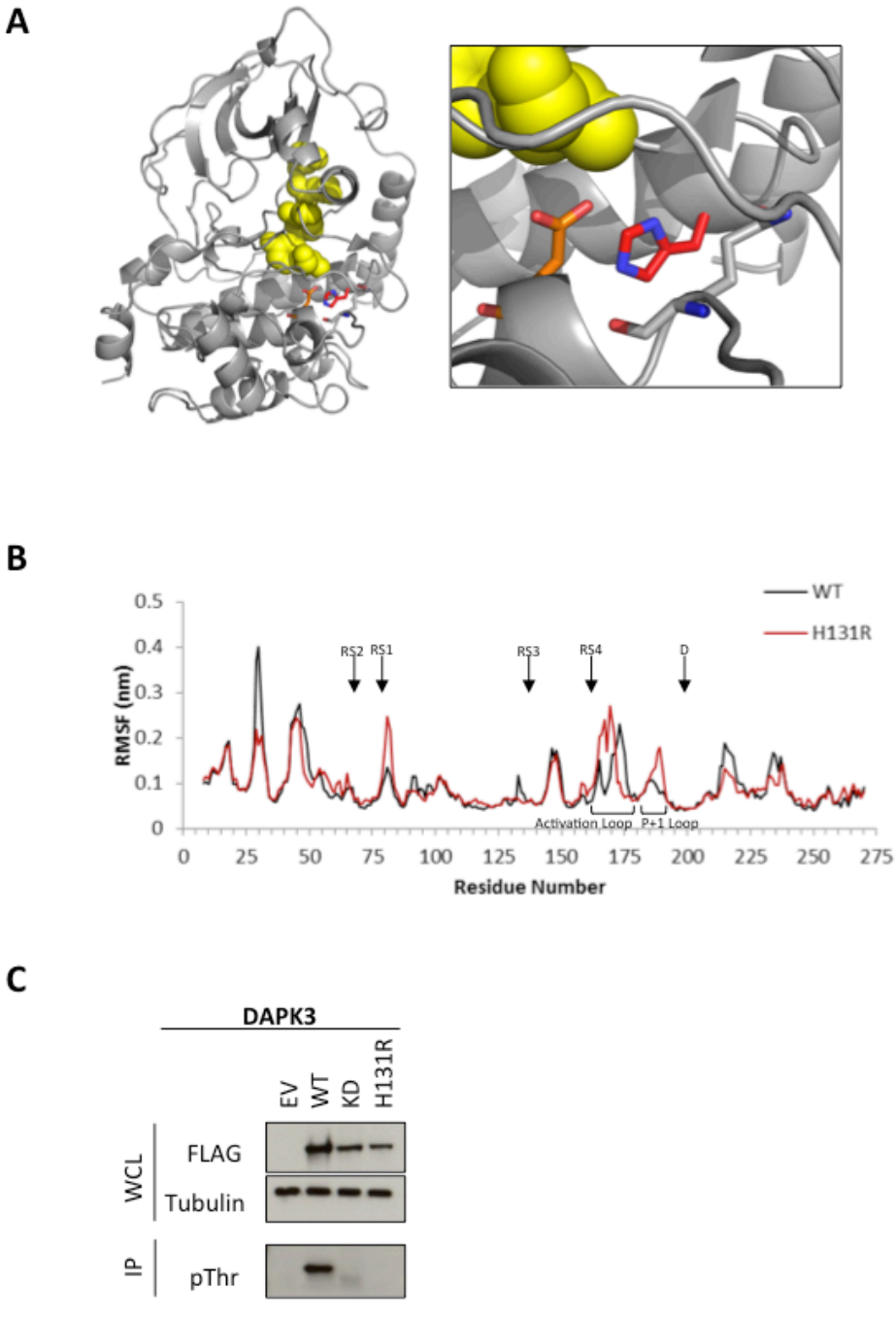
Supplemental Figure 1







Supplemental Figure 2

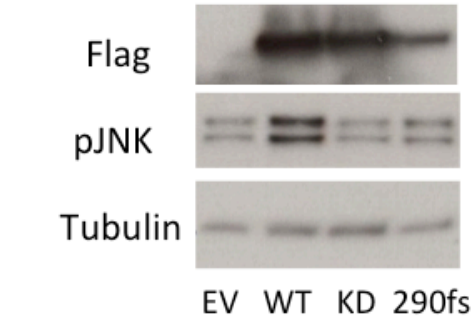


Supplemental Figure 3

A

```
WT          MAASSLEQKLSRLEAKLKQENREARRRIDLNLDISPQRPRPTLQLPLANDGGSRSPSSES
D290fs      MAASSLEQKLSRLEAKLKQENREARRRIDLNLDISPQRPRPTLQLPLANDGGSRSPSSES
            ************************************************************

WT          SPQHPTPPARPRHMLGLPSTLFTPRSMESIEIDQKLQEIMKQTGYLTIGGQRYQAEINDL
D290fs      SPQHPTPPARPRHMLGLPSTLFTPRSMESIEIDQKLQEIMKQTGYLTIGGQRYQAEINDL
            ************************************************************

WT          ENLGEMGSGTCGQVWKMRFRKTGHVIAVKQMRRSGNKEENKRILMDLDVVLKSHDCPYIV
D290fs      ENLGEMGSGTCGQVWKMRFRKTGHVIAVKQMRRSGNKEENKRILMDLDVVLKSHDCPYIV
            ************************************************************

WT          QCFGTFITNTDVFIAMELMGTCAEKLKKRMQGPIPERILGKMTVAIVKALYYLKEKHGVI
D290fs      QCFGTFITNTDVFIAMELMGTCAEKLKKRMQGPIPERILGKMTVAIVKALYYLKEKHGVI
            ************************************************************

WT          HRDVKPSNILLDERGQIKLCDFGISGRLVDSKAKTRSAGCAAYMAPERIDPPDPTKPDYD
D290fs      HRDVKPSNILLDERGQIKLCDFGISGRLVDSKAKTRSAGCAAYMAPERIDPQTPPSRTMT
            ********************************************************  * .

WT          IRADVWSLGISLVELATGQFPYKNCKTDFEVLTKVLQEEPPLLPGHMGFSGDFQSFVKDC
D290fs      SGPTYGAWASRWWSWQQDSFPTRTARRTLRSSPKSYRKSPRFCPDTWASRGTSSPSSKTA
                  : .    .    .** :..:  :.   *  ::.* : *   .  *  .   * .

WT          LTKDHRKRPKYNKLLEHSFIKRYETLEVDVASWFKDVMAKTESPRTSGVLSQPHLPFFR
D290fs      LLKITGRDQSIISYLNTASSSATRRWRWTWRPGSRMSWRRLSHRGLAAS-ASPTCPSSG
            * *   :  .  . *: :  .  .  .      :    : .    :.  :.*  *
```

B



Chapter Seven

Paper 4:

Germline sequencing of DIPNECH patients reveals a heterogeneous disease with rare variants in genes that are somatically mutated in other neuroendocrine malignancies.

Germline sequencing of DIPNECH patients reveals a heterogeneous disease with rare variants in genes that are somatically mutated in other neuroendocrine malignancies

Andrew Hudson[a], Hui Sun Leong [b], Crispin Miller [b c], Was Mansoor [d†], John Brognard[a†]

[a] Signalling Networks in Cancer Group, [b] Computational Biology Support, [c] RNA Biology Group, Cancer Research UK Manchester Institute, The University of Manchester, Manchester M20 4BX, UK, [d] The Christie NHS Foundation Trust, Wilmslow Road, Manchester M20 4BX, UK.

[†] To whom correspondence should be addressed. E-mail: John.Brognard@cruk.manchester.ac.uk and Was.Mansoor@christie.nhs.uk

Abstract

Diffuse Idiopathic Pulmonary Neuroendocrine Cell Hyperplasia (DIPNECH) is a rare disease characterised by neuroendocrine nodules in the lungs and frequently the development of pulmonary carcinoid tumours. The aetiology is unknown with no previous genomic studies. We performed whole exome germline sequencing on 10 patients with DIPNECH, filtering out any SNVs previously annotated in the dbSNP database, and did not discover any common novel SNVs between the cases. Given the clinical association with pulmonary carcinoid tumours we filtered our data to retain only genes previously reported as mutated in pulmonary carcinoids. This highlighted two novel protein-coding SNVs in different patients in MEN1 and PSIP1, two of the most commonly mutated genes in pulmonary carcinoid. MEN1 germline variants are well characterised in Multiple Endocrine Neoplasia (MEN) syndrome and whilst our patient had no other features of this condition DIPNECH has been previously observed in a patient with MEN. The PSIP1 p.R404W SNV we detected is located within the highly conserved MEN1 binding domain and is the most commonly mutated residue of PSIP1 in solid tumours. We also applied a functional screen to the whole dataset to extract SNVs occurring within or immediately adjacent to residues critical for kinase function in 411 kinases. This identified a CDK8 variant in one patient causing a significant amino acid change adjacent to the conserved DxxxxG structural backbone that is highly likely to affect kinase activity. Applying the same functional screen to pan-cancer genomics data reveals a proportion of small cell lung cancers (SCLC) with similar CDK8 mutations at residues critical for kinase function, suggesting a genetic link between these two conditions. Biochemical studies are required to investigate the mechanism by which loss of function CDK8 may drive neuroendocrine tumourigenesis. The CDK8 and PSIP1 examples highlight the benefits of using the analysis of rare diseases to assist in the genomic characterisation of more common conditions such as SCLC where high mutational burdens impede driver mutation discovery.

Introduction

DIPNECH is a rare disease characterised by hyperplasia of neuroendocrine cells in the distal airways [1]. Only a few hundred cases have been reported in the literature worldwide and there is a paucity of data regarding the pathogenesis and aetiology of the disease [2]. It is not known whether there is genetic predisposition to the disease

or if environmental factors are more important. Given the rarity of the disease there are no published genomics studies. The classical presentation is dyspnoea and long-standing cough in middle-aged females who have never smoked [3]. Cross sectional imaging often reveals multiple small proliferations of neuroendocrine tissue (also known as tumourlets). Patients are often misdiagnosed as suffering from bronchial asthma and treated as such, possibly leading to an underestimation of the prevalence of disease [3, 4]. The excess of pulmonary neuroendocrine cells results in increased production of peptides that cause peribronchiolar fibrosis and a progressive obstructive ventilator defect that can be fatal [3]. Somatostatin analogues, used in other neuroendocrine disorders to reduce neuroendocrine cell hormonal production, are tried in DIPNECH and may reduce some symptoms but appear to have little or no effect on pulmonary function [3].

Associations with other neuroendocrine tumours are varied with approximately half of DIPNECH patients presenting with a carcinoid tumour, suggesting a common aetiology [2, 5]. On the other hand there are no case reports of DIPNECH patients developing the more aggressive neuroendocrine derived small cell lung cancer (SCLC) [2]. However the non-surgical management of the vast majority of SCLC may obscure this link. One study identified 4 patients with lung adenocarcinoma and concomitant DIPNECH and suggested a causative link [6]. There are also single case reports of DIPNECH occurring with other malignancies including malignant melanoma where the DIPNECH lesions could have been incorrectly assumed to be pulmonary metastatic disease [7]. It has not been distinguished whether these associations with different malignancies represent a common aetiology or may simply reflect a bias in diagnosis in patients already undergoing lung imaging or resections for their primary cancer.

With monozygotic twins at our institute both suffering with the disease and no environmental link ascertained between our patients locally, we hypothesised that there is a genetic predisposition to developing DIPNECH. Post-diagnosis patients progress very slowly and are rarely managed surgically. As a result, given the rarity of the disease, it was not possible to obtain genetic material from DIPNECH tumourlets in a prospective study. Therefore whole exome sequencing of the germline was performed on 10 patients from our institution. Analysis of germline single nucleotide variants (SNV) discovered no common variants between the samples once all SNVs with a dbSNP reference were filtered out. However using different filtering approaches it was possible to identify novel SNVs occurring in genes that are somatically mutated

in other neuroendocrine malignancies such as carcinoid tumours and small cell lung cancer. These data may provide a genetic link between DIPNECH and other neuroendocrine malignancies. Using a rare proliferative disease provides opportunities to discover driver mutations that would be obscured by the carcinogen-induced high mutational burden of cancers of the same cell type such as SCLC.

Methods

Patient selection

All patients were being actively monitored at The Christie NHS Foundation Trust for a histologically confirmed diagnosis of DIPNECH. Ethical approval was obtained from the Manchester Cancer Research Centre (MCRC) Biobank (a generically approved research tissue bank approved by South Manchester Research Ethics Committee: 07/N1003/161+5). Informed written consent was taken from 10 patients with a diagnosis of DIPNECH to prospectively collect a whole blood sample for germline sequencing. Informed written consent to use any archival tissue for sequencing was also obtained where relevant.

Germline sequencing and validation.

Germline whole exome sequencing was performed on each sample following DNA extraction from a whole blood sample carried out by the Central Manchester Hospitals NHS Trust. Library preparation was carried out using Agilent SureSelect All Exon V6+COSMIC kit. Whole exome sequencing was carried out on an Illumina HiSeq 2500 in Rapid Run mode with paired-end 2 x 100 cycles. MEN1, PSIP1 and CDK8 novel SNVs were validated using Sanger sequencing on an ABI13130 16 capillary system (Life Technologies) and sequence data analysed using 4Peaks software (MekenTosj). Sanger sequencing primers are listed in Supplementary Table 1.

Sequence reads mapping and variant calling

Paired-end sequence reads were aligned to the human reference genome GRCh37/hg19 using BWA (version 0.7.7) [8]. The aligned reads in SAM format were converted to BAM format and sorted using Picard tools (version 1.96, http://picard.sourceforge.net). Local realignment around known INDELs and base quality recalibration were performed using the Genome Analysis Tool Kit (GATK) software (version 3.1.1) on the sorted and marked duplicates BAM files [9]. SNVs and INDELs were identified using GATK's HaplotypeCaller module. In addition to the

parameters described in the GATK best practices, variants were further filtered by requiring a minimum read depth of 20 and minimum allele frequency of 35%.

Variant analysis

Novel SNVs were identified as those occurring at a genomic location without a corresponding dbSNP reference in either build 146 or 147 of the Database of Single Nucleotide Polymorphisms [10]. A carcinoid associated gene list was created by cross-referencing with a list of all genes mutated in pulmonary carcinoids (missense, nonsense, frameshift) from Fernandez-Cuesta et al. [11]. The kinase motif associated list was created by cross-referencing with a list of critical kinase motif locations for each kinase. The kinase motif locations were identified using a list of 411 classical active kinases (Supplementary Table 2) with their respective amino acid sequences for the VAIK, HRD and DFG motifs from Manning et al. [12]. A custom R-script was written to identify each kinase motif in the Genbank sequences of each kinase. The GxGxxG (glycine-rich loop), the E that salt-bridges to the lysine of VAIK, APE (substrate binding motif) and DxxxxG (critical R and C spine contacts) locations for each kinase were identified from the Genbank sequences by string searching for different variants of the motifs using the stringr R package. The critical kinase motifs were defined as GxG of the GxGxxG motif, K of the VAIK motif and the E that salt-bridges to the K, HRD, DFG, APE and DxxxxG. All SNVs that corresponded with a location within these motifs or 1 amino acid either side were recorded as positive.

Results

Ten patients with histologically confirmed DIPNECH underwent germline whole exome sequencing. The patient characteristics are displayed in Table 1 and show 3 distinct clinical patterns; 1) nodular DIPNECH in the absence of other tumours, 2) carcinoid tumour with a background of DIPNECH, 3) other malignancies with a background of DIPNECH. The majority of patients in groups 1 and 2 were never smokers whereas all 3 patients within group 3 had a heavy smoking history. The median age at diagnosis was 60.5 years and the majority of patients in groups 1 and 2 were female.

Across all 10 samples the pipeline detected a total of 119911 SNVs. Removing SNVs with a dbSNP reference left 987 novel SNVs (Supplementary Table 3). None of these novel SNVs were present in more than one sample. At a gene level there were

multiple genes with a novel SNV in more than one sample. These genes included the tumour suppressors NF1, PTEN, and MAP2K4 (5, 3, and 3 samples respectively) that were all intronic SNVs (Table 2). Only TTN had multiple samples (2 samples) with missense or predicted splice site SNVs. In total across all 10 samples there were 9491 INDELs and filtering to remove all without a previous germline reference left only 4 novel INDELs (Supplementary Table 4). None of these novel INDELs occurred in genes with novel SNVs nor did any occur in protein-coding regions although the DDX11L1 INDEL was observed in 9 out of 10 samples. However DDX11L1 is a poorly characterised pseudogene and it is possible that this region has not been adequately sequenced in previous germline sequencing studies.

Given the clinical connection between DIPNECH and pulmonary carcinoid tumours the 987 SNVs were further filtered to retain only genes in which missense, truncating or frameshift mutations were detected in the large pulmonary carcinoid sequencing study by Fernandez-Cuesta et al. [11]. This revealed germline SNVs in carcinoid-associated genes including genes such as MEN1 and PSIP1 that are frequently mutated in pulmonary carcinoids (Table 3). Patient 5, who presented with a Typical Carcinoid Grade 1 on a background of DIPNECH, was found to have a heterozygous p.C409Y MEN1 germline SNV. The C409 residue of MEN1 is located in close proximity to a conserved region in which germline variants are frequently found in Multiple Endocrine Neoplasia (MEN) patients [13]. Likewise Patient 7, who presented with multiple bilateral lung nodules on CT and a wedge resection showing DIPNECH and carcinoid, was found to have a heterozygous p.R404W PSIP1 SNV. The R404 residue of PSIP1 is highly conserved, located within the MEN1 binding domain and is the most frequently somatically mutated residue of PSIP1 in cBioPortal with lung, head and neck, stomach and pancreatic cancer mutations [14-17].

To discover further novel germline drivers of DIPNECH a filtering algorithm was applied to detect SNVs occurring within the critical regions of kinase domains. The whole dataset of SNVs was queried for those occurring in or immediately adjacent to highly conserved motifs critical for kinase function. The cross-referenced motifs consist of the first 3 residues of the glycine rich loop (GxGxxG), the critical lysine of the VAIK motif, the E that salt-bridges to the critical lysine, HRD, DFG, APE and the DxxxxG structural backbone (Figure 1A). This revealed six SNVs of which five had been previously detected in the 1000 Genome Project to varying frequencies (Figure 1B). CDK8 p.C222Y was the only SNV not previously recorded and interspecies comparison of CDK8 sequences confirms that it results in a significant amino acid

change of a highly conserved region (Figure 1D). Querying the mutational datasets in cBioPortal reveals multiple CDK8 cancer mutations in motifs previously demonstrated to be critical for kinase function (Figure 1E) [18]. These samples include a significant proportion of small cell lung cancer mutations including p.D173V affecting the critical DMG motif (analogous to DFG) that is highly likely to kill catalytic function.

Discussion

Our study gives the first genomic insight into DIPNECH, suggesting that it is a heterogeneous disease associated with different germline genetic drivers. A genetic aetiology was suggested by the presence of monozygotic twins at our institute. This has been reinforced with the detection of novel germline variants that are either predicted to have a significant functional effect on the resultant protein or have been recorded previously as commonly mutated genes in other neuroendocrine malignancies. Genetic material was not available from the relatives of these individuals but would be desirable to support a genetic link. It is noteworthy that 3 out of the 10 DIPNECH patients reported first-degree family members with early onset cancer. The heterogeneity of genetic drivers in our patients is unsurprising given the different spectrum of disease presentation displayed by our 10 patients. Clinical presentations ranged from the classical nodular disease in non-smoking females to DIPNECH found in the cancer resection specimens of older male heavy smokers. Such distinct presentations would already suggest a different aetiology whether it be environment or genetic. We found no common novel SNVs and apart from TTN no genes with novel protein coding SNVs in more than one patient. TTN, the longest protein-coding gene of all, is frequently mutated in most cancer subtypes and it is not known if this is merely a consequence of gene length bias or whether TTN functions as an actual genetic driver [19]. A novel non protein-coding INDEL of DDX11L1 was found in 9 out of 10 cases however it occurs in a poorly characterised and rarely sequenced pseudogene, which may not have been adequately sequenced in previous studies. The fact that it is observed in the whole spectrum of our patients, rather than a specific clinical presentation group, suggests that it may be common in the general population. Using the dbSNP database to filter out all previously recognised germline variants from a dataset is limited by the quality of the previous sequencing. We have previously shown how different sequencing studies are discrepant in the reporting of mutations in the same samples due to inadequate sequencing coverage [20]. Filtering high-coverage modern sequencing with older dbSNP data will result in discrepancies leading to the identification of seemingly novel variants that are actually common but

underreported. Another limitation of using dbSNP filtering to identify novel SNVs is associated with the rarity of the disease and the age that it becomes clinically detectable. DIPNECH is thought to be rare although clinical diagnosis is often delayed many years due to symptoms similar to asthma. Removing all variants from our analysis with a dbSNP reference may miss causative SNVs as the dbSNP population may include patients with occult DIPNECH.

Histopathological associations between DIPNECH and pulmonary carcinoids have been made previously [21] so another strategy to filter the SNV list was to retain only genes that had been previously identified as somatically mutated in pulmonary carcinoids. This highlighted two interesting variants of the most commonly mutated genes in pulmonary carcinoids, PSIP1 and MEN1, which are known to interact in the same signalling pathway. In the pulmonary carcinoid sequencing study by Fernandez-Cuesta et al. 13.3% of cases carried mutually exclusive frameshift or truncating mutations of MEN1 or PSIP1 [11]. Approximately 18% of patients with typical pulmonary carcinoids carry a MEN1 mutation and 36% possess loss of heterozygosity (LOH) of MEN1 [21]. Our MEN1 and PSIP1 variants were found in different patients, both of whom presented with pulmonary carcinoids. MEN1 germline variants are best known for their association with the Multiple Endocrine Neoplasia (MEN) syndromes from which the gene gets it name. These patients develop multiple neuroendocrine and endocrine tumours. DIPNECH has been reported previously in one patient with MEN and a gastrinoma although the specific germline variant is not reported [22]. The p.C409Y variant seen in our DIPNECH patient has not been previously identified in MEN patients and whilst pulmonary carcinoids are associated with MEN our patient does not possess any other clinical features of these syndromes [23]. The C409 residue is located in a highly conserved region of MEN1 and the significant amino acid change (from cysteine to tyrosine) results in the variant being predicted as functional by mutationassessor.org. The PSIP1 p.R404W variant we detected in a DIPNECH patient is located in a conserved region of a domain that binds to MEN1 and would therefore be predicted to affect this interaction. Protein-coding PSIP1 germline variants have not been previously documented to play a role in oncogenic diseases. However querying somatic mutational data via the cBioPortal demonstrates that the PSIP1 R404 is the most commonly somatically mutated residue of PSIP1 occurring in multiple cancer subtypes [18].

To assist discovery of functionally relevant germline variants in the DIPNECH dataset we next applied a similar algorithm to the one we have used to identify tumour

suppressor kinases in pan-cancer mutational datasets (Hudson et al., unpublished – Chapter 6 of this thesis). By considering the functional impact of germline variants on the critical motifs of kinases we identified a novel CDK8 variant adjacent to the highly conserved DxxxxG structural backbone with a high likelihood of affecting kinase function. The role of CDK8 on carcinogenesis has been explored but is not comprehensively defined. In colon cancer CDK8 has been postulated as playing an oncogenic role through its regulation of beta-catenin transcriptional activity [24]. Contrary to this our pan-cancer analysis of CDK8 mutations reveals a significant subset affecting residues critical for kinase function and pointing towards a tumour suppressive role in SCLC and bladder cancer. Whilst we have not biochemically validated these CDK8 variants, our experience with significant amino acid changes within conserved kinase motifs suggest these mutations are highly likely to kill catalytic function. CDK8 is known to phosphorylate the E2F1 transcription factor and alter its transcriptional activation [25]. RB1, a well known tumour suppressor with damaging mutations in over 90% of SCLC, is critical to the tumour suppressive functions of E2F1 [26]. It is postulated that CDK8 may also be critical for E2F1 tumour suppressive function and that the damaging mutations drive carcinogenesis through the same pathway as Rb1 inactivation. An interesting observation regarding CDK8 mutations in SCLC is that they occur at a much higher frequency in the cell lines sequenced by CCLE than a major sequencing study of SCLC patient samples by Roman Thomas' group (which only had 1 out of 110 samples with a missense CDK8 mutation) [27]. A major difference between these sequencing studies is that CCLE do not have access to normal tissue from the cell line patients to sequence and remove germline variants from the analysis whereas the Thomas study did. It is possible that the CDK8 mutations detected by CCLE are in fact germline variants, which like our DIPNECH variant may have predisposed the patient to developing a neuroendocrine malignancy. It is widely believed, due to differing genetic profiles, that high-grade neuroendocrine malignancies such as SCLC are separate biological entities to lower grade malignancies such as typical carcinoids and DIPNECH [21]. Further biochemical investigation into the role of CDK8 in neuroendocrine malignancies may reveal a common genetic link that challenges this paradigm for a subset of cases.

Analogous to analysing cancer mutational data to discover somatic driver mutations, functional germline variant discovery is relatively straightforward for common variants of precise clinical phenotypes but very difficult for lower frequency variants in more ambiguous clinical presentations. The task is further complicated in a disease such as DIPNECH because the relative stability of disease in some patients means there are

limited opportunities to collect tumour tissue for comparison. Following the initial analysis of the DIPNECH SNV data it was clear that there were no common recurrent alterations that could be identified by frequency alone. The challenge was then to identify drivers in single cases. We have presented novel methods used to filter germline data based on clinical and biochemical considerations to identify candidate functional SNVs and provide a first insight into the genomic features of DIPNECH. Further biochemical work is required to validate these targets and elucidate the mechanisms of tumourigenesis. However the CDK8 and PSIP1 examples serve to highlight the great potential in investigating rare diseases to assist the genomic characterisation of more common conditions such as SCLC where high mutational burdens impede driver mutation discovery.

References

1. Aguayo SM, Miller YE, Waldron JA, Jr. *et al*. Brief report: Idiopathic diffuse hyperplasia of pulmonary neuroendocrine cells and airways disease. *N Engl J Med* 327(18), 1285-1288 (1992).
2. Wirtschafter E, Walts AE, Liu ST, Marchevsky AM. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia of the lung (dipnech): Current best evidence. *Lung* 193(5), 659-667 (2015).
3. Carr LL, Chung JH, Duarte Achcar R *et al*. The clinical course of diffuse idiopathic pulmonary neuroendocrine cell hyperplasia. *Chest* 147(2), 415-422 (2015).
4. Stenzinger A, Weichert W, Hensel M, Bruns H, Dietel M, Erbersdobler A. Incidental postmortem diagnosis of dipnech in a patient with previously unexplained 'asthma bronchiale'. *Pathol Res Pract* 206(11), 785-787 (2010).
5. Miller RR, Muller NL. Neuroendocrine cell hyperplasia and obliterative bronchiolitis in patients with peripheral carcinoid tumors. *Am J Surg Pathol* 19(6), 653-658 (1995).
6. Mireskandari M, Abdirad A, Zhang Q, Dietel M, Petersen I. Association of small foci of diffuse idiopathic pulmonary neuroendocrine cell hyperplasia (dipnech) with adenocarcinoma of the lung. *Pathology Research and Practice* 209(9), 578-584 (2013).
7. Killen H. Dipnech presenting on a background of malignant melanoma: New lung nodules are not always what they seem. *BMJ Case Rep* 2014, (2014).
8. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26(5), 589-595 (2010).
9. Mckenna A, Hanna M, Banks E *et al*. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9), 1297-1303 (2010).
10. Database of single nucleotide polymorphisms (dbsnp). Bethesda (md): National center for biotechnology information, national library of medicine. (dbsnp build id: 146 & 147). Available from: Http://www.Ncbi.Nlm.Nih.Gov/snp/
11. Fernandez-Cuesta L, Peifer M, Lu X *et al*. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat Commun* 5, 3518 (2014).
12. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 298(5600), 1912-1934 (2002).
13. Concolino P, Costella A, Capoluongo E. Multiple endocrine neoplasia type 1 (men1): An update of 208 new germline variants reported in the last nine years. *Cancer Genet* 209(1-2), 36-41 (2016).
14. Jiao Y, Yonescu R, Offerhaus GJ *et al*. Whole-exome sequencing of pancreatic neoplasms with acinar differentiation. *J Pathol* 232(4), 428-435 (2014).
15. Lin DC, Meng X, Hazawa M *et al*. The genomic landscape of nasopharyngeal carcinoma. *Nat Genet* 46(8), 866-871 (2014).
16. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513(7517), 202-209 (2014).
17. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417), 519-525 (2012).
18. Cerami E, Gao JJ, Dogrusoz U *et al*. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5), 401-404 (2012).
19. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods* 10(11), 1108-1115 (2013).

20. Hudson AM, Yates T, Li Y *et al*. Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. *Cancer Res* 74(22), 6390-6396 (2014).
21. Swarts DRA, Ramaekers FCS, Speel EJM. Molecular and cellular biology of neuroendocrine lung tumors: Evidence for separate biological entities. *Bba-Rev Cancer* 1826(2), 255-271 (2012).
22. Davies S, Gosney J, Hansell D *et al*. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia. An under-recognised spectrum of disease. *Modern Pathol* 19, 165-165 (2006).
23. Duh QY, Hybarger CP, Geist R *et al*. Carcinoids associated with multiple endocrine neoplasia syndromes. *Am J Surg* 154(1), 142-148 (1987).
24. Firestein R, Hahn WC. Revving the throttle on an oncogene: Cdk8 takes the driver seat. *Cancer Res* 69(20), 7899-7901 (2009).
25. Zhao J, Ramos R, Demma M. Cdk8 regulates e2f1 transcriptional activity through s375 phosphorylation. *Oncogene* 32(30), 3520-3530 (2013).
26. Nevins JR. The rb/e2f pathway and cancer. *Hum Mol Genet* 10(7), 699-703 (2001).
27. George J, Lim JS, Jang SJ *et al*. Comprehensive genomic profiles of small cell lung cancer. *Nature* 524(7563), 47-U73 (2015).
28. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39(17), e118 (2011).

Table and Figure Legends

Table 1: Patient characteristics of the 10 sequenced DIPNECH patients. Four patients had a nodular disease pattern (Group 1), 3 had pulmonary carcinoid (Group 2) and associated DIPNECH on resection specimen, and 3 patients had DIPNECH in the resection specimen of other malignancies (Group 3). (p/y = pack year smoking history where 1 p/y is equivalent to smoking 20 cigarettes per day for a whole year).

Table 2: Genes in which more than one patient possessed a novel SNV in that gene. The frequency is the number of patient samples with a novel SNV in the respective gene. The highest frequency genes include known tumour suppressors NF1, MAP2K4 and PTEN however the SNVs in these three genes were all in non-coding regions. TTN is the only gene with more than one protein coding novel SNV.

Table 3: Novel SNVs occurring in genes previously demonstrated to possess somatic missense, truncation or frameshift mutations in pulmonary carcinoids in the case series by Fernandez-Cuesta et al. [11]. Columns D1 to D10 display the case in which the SNV occurred and correlate to the patient characteristics in Table 1.

Figure 1: A) To identify functional SNVs with a high likelihood of affecting kinase activity the total 119911 SNVs were filtered using critical motif locations for 411 classical active kinases. SNVs occurring within or immediately adjacent to the GxGxxG, K, saltbridge-E, HRD, DFG, APE, and DxxxxG were extracted. B) Six SNVs extracted using the critical kinase motif filtering. The CDK8 p.C222Y is the only SNV not previously annotated as a germline variant. C) Validation of the CDK8 p.C222Y using Sanger sequencing. D) Mutation Assessor analysis of the p.C222Y variant predicts it as highly functional with a significant amino acid change in a residue that is highly conserved throughout inter-species sequence comparison [28]. E) Querying the cBioPortal CDK8 mutational data reveals a high frequency of cancer mutations in the conserved critical motifs in SCLC cell lines (red spheres) and other cancer subtypes (green spheres).

Table 1

Case	D1	D2	D3	D4	D5	D6	D7	D8	D10	D9
Age	60	51	66	59	61	54	59	79	67	74
Sex	Female	Female	Female	Female	Female	Female	Male	Male	Male	Female
Pattern	Nodular	Nodular	Nodular	Nodular	Carcinoid	Carcinoid	Carcinoid	Cancer	Cancer	Cancer
Smoking	Never	Never	Never	8 p/y	Never	Never	25 p/y	30 p/y	50 p/y	50 p/y
Other cancer					Breast, Colon			Rectal with lung metastases	Lung, Renal, Thyroid	Lung
SNV			CDK8		MEN1		PSIP1			

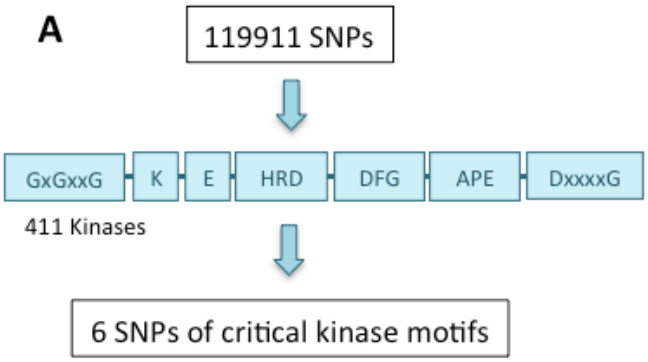
Table 2

Gene	Total Cases		Gene	Total Cases
NF1	5		LRP2	2
PAX5	4		MUC16	2
MAP2K4	3		MUC19	2
PTEN	3		MYO15B	2
SAE1	3		PNPLA7	2
CHD2	2		RP11-296E7.1	2
AMPH	2		RP11-298C3.2	2
ATP13A2	2		SH2B2	2
ATRX	2		SH2B3	2
PHKG2	2		SPTBN4	2
FAM109A	2		TTN	2
IST1	2		UMODL1	2
KDM6A	2		UTP6	2

Table 3

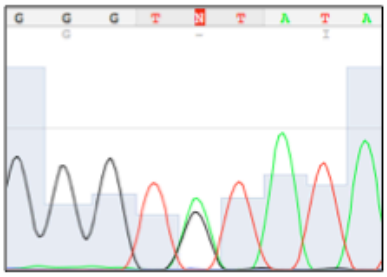
Gene	Location	Mutation	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
TRIM23	5:64913979	R62H	1	0	0	0	0	0	0	0	0	0
MEN1	11:64572630	C409Y	0	0	0	0	1	0	0	0	0	0
GPR68	14:91701127	G100S	0	0	0	0	0	0	0	0	0	1
SLC12A3	16:56947229	W993L	0	0	1	0	0	0	0	0	0	0
ATP8B1	18:55319881	R1032S	0	1	0	0	0	0	0	0	0	0
TTN	2:179425463	E19593K	1	0	0	0	0	0	0	0	0	0
TTN	2:179612506	L4874S	0	0	0	0	1	0	0	0	0	0
PSIP1	9:15468838	R404W	0	0	0	0	0	0	1	0	0	0
PRKCZ	1:1986745	upstream	0	1	0	0	0	0	0	0	0	0
PKD1	16:2154433	upstream	0	0	0	0	0	1	0	0	0	0
TBC1D8	2:101649970	upstream	0	0	0	0	0	1	0	0	0	0
TTN	2:179699212	upstream	0	0	1	0	0	0	0	0	0	0
DOK1	2:74785705	upstream	0	0	0	0	0	0	1	0	0	0
GFPT2	5:179780307	upstream	0	0	0	0	0	0	0	1	0	0
DST	6:56466730	upstream	0	0	0	0	0	0	1	0	0	0
SMARCA2	9:2123802	S1282S	0	1	0	0	0	0	0	0	0	0
PIEZO2	18:10759914	splice_region	0	0	1	0	0	0	0	0	0	0
CYP2C8	10:96827191	noncoding	0	0	0	0	0	0	0	0	0	1
C14orf28	14:45374809	noncoding	0	0	0	1	0	0	0	0	0	0
SETDB1	1:150919608	intron	0	0	0	0	0	0	0	0	1	0
RB1	13:48878925	intron	0	1	0	0	0	0	0	0	0	0
PSMB7	9:127160696	intron	0	0	1	0	0	0	0	0	0	0
RB1	13:49032730	intron	0	0	0	0	0	0	0	0	1	0
CNGB1	16:57921658	intron	0	0	0	0	0	0	0	1	0	0
SMARCB1	22:24169481	intron	0	0	0	0	0	0	1	0	0	0
MAN2A1	5:109027394	intron	0	0	0	0	0	0	1	0	0	0
ANKRD31	5:74502212	intron	0	0	0	0	0	0	1	0	0	0
TNFAIP3	6:138200992	intron	0	0	0	0	1	0	0	0	0	0
PKD1	16:2138744	downstream	1	0	0	0	0	0	0	0	0	0
EVI2A	17:29640474	downstream	0	0	0	0	0	0	0	0	1	0
CRELD1	3:9979330	downstream	0	0	0	1	0	0	0	0	0	0

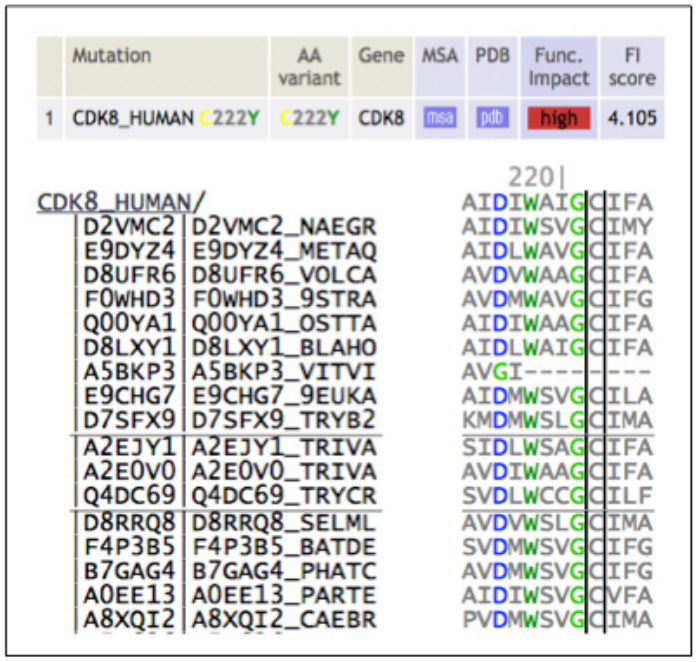
Figure 1

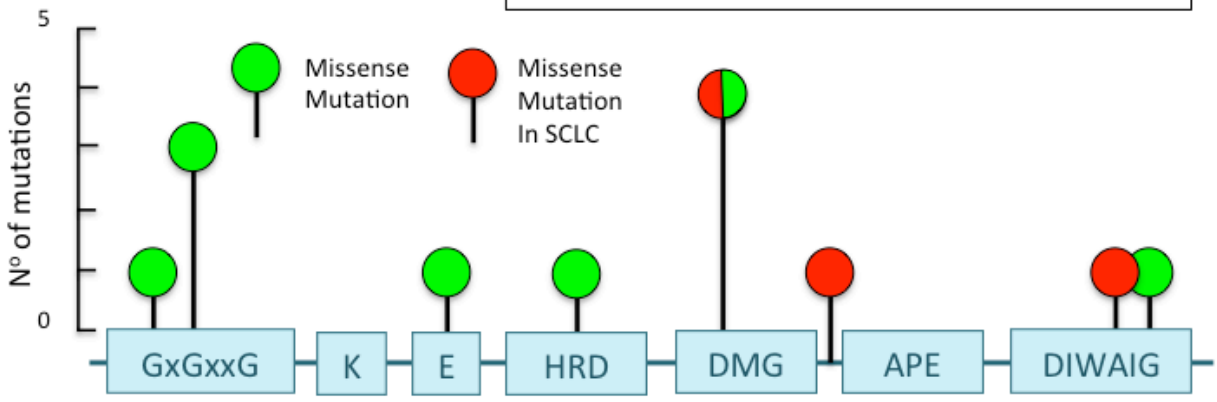


B

Gene	Variant	Cases	1000G MAF	Motif
CDK8	C222Y	1	NA	DxxxxG(x)
MAP2K3	R94L	10	0.4998	K(x)
VRK2	I144V	5	0.3019	HRD(x)
SBK2	C72R	10	0.0821	GxG(x)xG
MAP3K5	I708V	1	0.0018	(x)K
TSSK2	V204M	1	0.0008	DxxxxG(x)







Supplementary Table Legends

Supplementary Table 1: Sanger sequencing validation primers for MEN1, CDK8, PSIP1.

Supplementary Table 2: 411 kinases screened – to avoid duplication in the thesis this table is accessible as Supplemental Table 1 in Chapter 6.

Supplementary Table 3: 987 SNVs identified in the 10 patients without a previous germline reference in the dbSNP database. For brevity, this large list is not printed in the thesis but is present on the enclosed USB stick.

Supplementary Table 4: Novel INDELs (those not previously recorded in the dbSNP germline database) occurring across the 10 samples.

Supplementary Tables

Supplementary Table 1

MEN1_F1	ccaacctatgcttacctttc
MEN1_F2	gtaagagactgatctgtgcc
MEN1_RC1	ctgctctggccatcccatcc
MEN1_RC2	cctgtagtgcccagacctctgt
CDK8_F1	gctgttaattcttaggcgttttg
CDK8_F2	ggtcggacattgtctcttggtg
CDK8_RC1	caggaaacccataaatatttccat
CDK8_RC2	gctgtttctacggatctttga
PSIP1_F1	aggatgtgaacagatgcattgag
PSIP1_F2	gctcagaaacacacagagatg
PSIP1_RC1	tcatctgcctcatgagcaatgg
PSIP1_RC2	ttctgtggcgtatacacagtga

Supplementary Table 2: See Supplemental Table 1 in Chapter 6 p82

Supplementary Table 3: See enclosed USB stick

Supplementary Table 4

GENE	Location	Consequence	REF	ALT
DDX11L1	1:13657-13658	non_coding_variant	CAG	C
BBS1	11:66285926	downstream_variant	CT	C
LPAR6,RB1	13:48983287	downstream_variant	CT	C
MEIS3	19:47916655	intron_variant	CA	C

Chapter Eight

Discussion

8.1 Impediments to driver mutation discovery

The overarching theme of this thesis is based on the observation that despite the sequencing of thousands of cancer samples, knowledge of driver mutations are still lacking for large proportions of different cancer subtypes. The work presented here aims to identify the reasons for this observation and provide novel solutions to better identify cancer drivers. The impediments to driver mutation discovery can be divided into those that are caused by detection limitations and those caused by interpretation limitations. When a tumour sample is sequenced by NGS technology a somatic mutation list is produced containing all somatic mutations detected in the sample. Detection limitations describe situations when driver mutations do not appear on the final mutation lists for the tumour subtype because they have not been detected or removed. Interpretation limitations describe situations when the driver mutations are present on the mutation lists but have not been identified as being driver mutations. The following sections of this discussion summarise these two different types of impediments to driver mutation detection and offer solutions based on the work contained in the thesis, with examples of novel drivers discovered in the process.

8.2 Detection limitations and solutions

8.2.1 Inadequate sequencing

The most common cause of discrepancy between CCLE and COSMIC mutation calling was inadequate sequencing of GC-rich regions and further enquiry demonstrated large sequencing cold-spots where mutational data is missing. This limitation of NGS has the potential to greatly skew the mutation profiles of cancer subtypes so that genes with large GC-rich regions are under-represented. Polycyclic aromatic hydrocarbons (PAH), a major carcinogen in cigarette smoke, preferentially cause mutations of GC nucleotides resulting in a mutational signature favouring GC-rich regions [131]. Therefore, it is logical to assume that smoking-related cancers, such as lung cancers, with GC-rich mutational signatures will be affected most by the inadequacy of sequencing of GC-rich regions and this may explain why a large proportion of lung cancers have no identified mutational driver. Resequencing of lung cancer cell lines with more modern NGS technology detected additional mutations in GC-rich regions that were not detected by either CCLE or COSMIC sequencing using older technologies. This led to the identification of a PAK4 mutation in the H2009 lung

adenocarcinoma cell line. Functional studies of the PAK4 p.E119Q mutation showed hyperactivation of the ERK cellular proliferation pathway compared to wild-type, providing proof-of-principle that driver mutations of GC rich regions have been missed by older NGS technologies previously used to collate mutational data. Whilst a comparison was not done using TCGA data it is expected that the similar age of a proportion of TCGA data is beset with the same limitations. Therefore, caution should be taken when mining older NGS data for driver mutations.

The observation that more modern NGS technology improved the detection of driver mutations suggests that the solution to this particular problem is to use more modern NGS platforms and possibly resequence older samples this way. Using our newer NGS platform sequencing coverage of GC-rich regions was much improved but there were still regions lacking read coverage. This was particularly apparent in the first exons and 5'UTR regions that are well known to have a high GC content. As sequencing technology improves the problem will be resolved but this issue should still be considered when analysing older genomics data. Viewing aggregated mutation data for all cancer types on resources such as COSMIC can often indicate the presence of sequencing cold-spots. When a cold-spot is present the mutational density appears significantly reduced compared to the rest of the gene. Therefore checking genes of interest in this way is a valuable way of assessing the need for alternate sequencing strategies.

8.2.2 Normal variant filtering

The COMSIC/CCLE cell line comparison demonstrated that another cause of discrepancy is the variable removal of presumed germ-line variants. Like GC-rich sequencing this issue is more likely to involve historical data where equivalent normal tissue data is missing. Most commercially available cell lines, which largely originate from the 1980's and 90's, do not have corresponding normal tissue germline data. Therefore when these samples were sequenced known germline variants from the population (such as the 1000 Genome Project) were used to remove potential germline variants and leave the presumed somatic mutations for analysis. The human germline data is constantly being updated so that a proportion of the differences between CCLE and COSMIC can be attributed to the different versions of the germline data used to filter the data. Going forward the obvious solution is to sequence a portion of the normal tissue when sequencing tumour tissue and this is now standard practice for both local studies and large projects such as TCGA. Whilst

this offers the best solution to exclude germline variant from somatic reports it can adversely affect the ability to detect genetic drivers present within the patients germline. This potential limitation is highlighted in the germline sequencing of DIPNECH patients who possessed variants such as the predicted LOF CDK8 variant. Due to its interference with a critical conserved motif we predict the CDK8 p.C222Y variant to have a likelihood of abolishing catalytic activity. If the patient with this variant was to develop a tumour that was subsequently sequenced then filtering out the germline variants could result in this potential driver being missed. This quandary highlights the problem of solely relying on a normal tissue equivalent to filter out germline variants in that predisposing germline drivers will also be excluded from the analysis.

Another potential problem using normal tissue to filter out germline variants occurs if the presumed normal tissue sample is somatically mutated. It is well known that cancers of the upper aerodigestive tract and lung occur within broader regions that have undergone field changes related to a general exposure of the epithelium to the carcinogen (cigarette smoke) [132]. Normal tissue samples are frequently taken at the time of cancer surgery from macroscopically normal tissue. If this macroscopically normal tissue is an area of occult field change or dysplasia then important early somatic drivers of carcinogenesis will also be subtracted from the somatic mutation list. This limits the identification of truncal drivers, the most desirable mutations in cancer genomics to identify. The solution is to ensure that either normal epithelial sampling is taken at a safe distance from the tumour or another source of germline DNA such as whole blood is used.

8.2.3 Case selection

Another question raised by the higher proportion of CDK8 mutations in SCLC cell lines compared to samples from the large genomics studies of SCLC is that of case selection. Genomics data for a specific cancer subtype will be skewed and potentially limit driver detection if the genomics case list does not represent the typical clinical characteristic of that patient group. In George et al. 87% of the 152 fresh frozen SCLC cases sequenced were taken from primary surgery of the tumour [133]. However in everyday clinical practice SCLC is rarely ever treated surgically due to the poor outcomes of this approach with most patients having occult metastatic disease at presentation. Therefore the critical question for a researcher mining this data for therapeutically relevant driver mutations is whether or not the surgical SCLC

genomics cases represent the same disease as the majority of SCLC they treat in the clinic. The fact that it has been possible to operate on the genomic study cases would suggest that they are slowing growing and therefore comprise of different genetic mechanisms to the more typical disease that rapidly metastasizes. The CCLE SCLC samples largely come from metastatic sources and this may explain the higher incidence of CDK8 mutations. It may be argued that the truncal driver mutations that we aim to discover should be present in both the primary and the metastatic samples. However, without a more representative case list from either repository it is difficult to claim that clinically relevant drivers are not being missed. The TCGA NSCLC cases are subject to the same bias with almost all cases being early-stage surgically treated disease.

The barrier to a more representative case list is the lack of availability of genetic material from non-surgical patients. These patients are often more unwell and it is difficult to ethically justify the performance of an invasive procedure for research purposes. In George et al. some samples were taken from higher stage patients at autopsy [133], although this may introduce bias in itself. Other approaches are being made possible because the amount and overall quality of genetic material required for whole exome sequencing is reducing. Solutions now include using circulating tumour cell and circulating free DNA sequencing. Furthermore, improvements in sequencing yield from FFPE samples allow archival biopsy samples. Due to potential links between clinical presentation and tumour biology there will always be bias if only one method is used for sample collection. Therefore, it is important to use a variety of approaches and link as much clinical, demographic and outcome data as possible to aid translation of genomics results to the clinic.

8.3 Interpretation limitations and solutions

8.3.1 Mutational Noise Overview

The greatest challenge in interpreting cancer genomics data in silico is distinguishing the few important driver mutations from the hundreds of inconsequential passenger mutations. These passenger mutations are described as ‘mutational noise’ that obscures the identification of the true driver mutations. Mutational noise is especially problematic in those cancer subtypes, such as lung cancer, with a high mutational burden (high numbers of passenger mutations). In some cancer subtypes the rampant action of a persistent carcinogen on the DNA (whether it be uv-light in melanoma or

cigarette smoke in lung cancer) means there is a high frequency of mutations in many genes that do not have a functional effect on the cancer. For example 10% of TCGA squamous lung cancer samples possess a mutation in the olfactory receptor OR2G6 [134]. In this TCGA dataset alone there are 40 different olfactory receptors with a mutational frequency above 5% of cases. Whilst at this stage we cannot completely rule out an oncogenic role for olfactory receptors in squamous lung cancer there is certainly no evidence to support a role and it seems unlikely. In total, in the TCGA squamous lung cancer dataset [134], there are 319 genes with a mutational frequency above 5% and in the TCGA melanoma dataset it is even higher with 521 genes above 5% (TCGA unpublished). Drivers that occur in a high proportion of cancers, such as BRAF occurring in 50% of melanoma, are easy to discover above a high background passenger mutation rate. However it is easy to see how the identification of lower frequency drivers becomes increasingly difficult as the mutational frequency diminishes. Drivers affecting only 1-2% of cases, which may be pharmacologically tractable and very valuable to find, would be unlikely to be detected by just analysing the frequency data

8.3.2 Statistical Noise Filtering Methods

Statistical methods can be applied to aggregated NGS data to improve the identification of driver mutation. The simplest correction recognises that when analysing mutation frequency at a gene level, longer genes will be reported to have a higher frequency as they contain more DNA to be mutated. Therefore dividing mutation count by nucleotide length of the gene produces a length corrected score that removes gene length bias. Other biases such as replication timing and gene expression level (discussed in chapter 2) can be taken into account with analysis packages such as MutSigCV [135]. MutSigCV also integrates the background mutation rate for different areas of the genome by considering the rates of non-coding mutations. These algorithms are now applied to data directly via the cBio portal [14].

Another important piece of information that can be directly accessed via these portals is the allele frequency of each mutation. This is a useful piece of information given the recent discoveries regarding tumour heterogeneity [63]. Assuming the tumour sample is not contaminated with normal tissue, a low allele frequency suggests the mutation is a branch mutation occurring later on in the tumour development. This suggests that the mutation was not an early driver of the tumourigenic process. This has therapeutic ramifications, as pharmacologically targeting a potential branch driver mutation would

only be beneficial in a small subset of the cancer. For this reason, many would use an allele frequency cut-off for target discovery. Due to the large numbers of tumours now sequenced additional complex statistical methods can be applied to the datasets. We previously applied a Fisher's Exact Test to compare the mutation profiles of 'protein kinase C mutation positive cancers' versus 'protein kinase C mutation negative cancers' [Appendix 1] [136]. This allowed us to detect significant differences in the gene mutation frequency between the two groups compared to what would be expected by chance alone. This approach can be modified to compare the mutation frequencies between one cancer subtype and the rest of the TCGA dataset. Using this method it is expected that many passenger mutation hotspots will be removed because, in the absence of other significant biases, the passenger mutation frequencies should be similar. This type of analysis requires large datasets so grouping cancers together to compare to the rest of the dataset is helpful. For example we have combined head and neck cancer and squamous lung cancers, due to their similar aetiology and pathology, into a 'squamous supergroup'. This supergroup is then compared to all the other TCGA datasets. Applying a Fisher's Exact Test each gene is attributed a q-value, which would be the chance of observing the observed frequency distribution if it was distributed by random chance alone. Using a stringent cut-off q-value (0.01), genes can be ranked by the magnitude of appearance in the supergroup compared to the other TCGA datasets. This approach has produced novel candidates for further investigation in a cancer subtype for which there are few genetic drivers known. There are a few caveats to using this approach. Firstly, the approach may miss pan-cancer drivers that play a significant role in the tumour subtype and grouping cancers by histological subtype itself is questionable in the genomics age. Secondly, biases such as differences in gene expression, influencing transcription coupled repair, and differences in sequencing technology between TCGA studies can be confounding factors. However as a screening tool this approach can highlight genes for further bioinformatic and biochemical consideration.

8.3.3 Using Protein Structural Considerations to Filter Noise

Most NGS pipeline and online genomics resources report some form of mutational assessor score to guide the user to those mutations that are predicted to be the most functionally damaging. The most commonly used assessors include Mutation Taster, Mutationassesor.org, Polyphen2, and Provean Sift [137-140]. All have been designed to predict the functional impact of a mutation based on the type of amino acid change

(e.g. hydrophobic to hydrophilic or acidic to basic) and most importantly the degree of conservation observed in that region throughout different species. The rationale for this method is that amino acids conserved throughout different species must be functionally important. Therefore, when a somatic mutation changes an amino acid conserved throughout the species it is predicted to have higher functional consequences compared to a mutation affecting a non-conserved region. Assessors such as Mutation Taster also use considerations such as splice site interference and loss of protein structural features [137]. Studies have shown that these mutation assessors work well to distinguish between pathogenic and neutral mutations and that combining the results from different assessors improves their predictive power [141]. Filtering mutation lists using a mutation assessor can therefore greatly aid driver discovery. However, the outputs typically consist of a large number of predicted medium-high functional mutations, most of which are unlikely to be significant drivers. For example, the HCT15_LARGE_INTESTINE cell line is reported in cBioPortal to possess 640 missense coding somatic mutations. The Mutation Assessor analysis reports the mutations as unclassified (4 mutations), neutral (125 mutations), low functional effect (238 mutations), medium functional effect (231 mutations), and high functional effect (42 mutations). Functional mutations are expected to be annotated as medium and high functional effect predictions. In the HCT15 cell line this takes the list of 640 missense mutations to 273 mutations (42.7%). Combining outputs with other mutation assessors may further reduce the shortlist but evidently this technique is not powerful enough to solely extract the handful of driver mutations likely to exist in the cell line. Interestingly in the HCT15 data the KRAS G13D mutation and TP53 S241F mutation are both predicted as medium functional effect. Elsewhere in melanoma samples the highly activating BRAF p.V600E mutation is classified as having a low functional effect.

Linking specific protein function to genomics data is essential if in silico methods are to become more accurate in predicting mutational effects. Kinases are a good class of proteins to refine this linkage in because the kinase domain has a specific function that is conserved throughout hundreds of kinases. Using knowledge about the critical structural elements of a kinase domain it was possible to estimate a location within all kinases before which all truncation mutations would cause abolition of kinase activity. This allowed the construction of a loss of function frequency table for all active kinases in the TCGA dataset. The top 30 kinases in terms of frequency of loss of function (LOF) truncating mutations were then used as a list of likely tumour suppressing kinases. The conserved sequence homology between these tumour-

suppressing kinases allowed a comparison of TCGA mutations within the kinase domain that were of a high likelihood of causing LOF. This demonstrated a number of mutational hotspots and biochemical validation showed that these regions are critical for kinase activity.

Identifying LOF hotspot locations has two benefits. Firstly the hotpot locations can be subsequently screened in all kinases in a single patient's genomic data to help identify drivers in that individual tumour. As we discover the functional relevance of different motifs the chances of identifying a driver mutation in a single tumour increases. Secondly the hotspot locations can be used for indicating novel LOF kinases in different cancer subtypes. There will be many mechanisms for causing LOF in a kinase and these will not be immediately apparent in aggregated cancer genomics data. However the occurrence of a mutation in a kinase at a validated LOF hotspot can signal to the researcher that the role of that gene may be tumour suppressive and prompt further analysis or wet-lab studies to discover the effects of the other mutations in that gene. The power of this approach is enhanced when validated LOF hotspots are combined and genes with high frequency of mutations in these regions are identified.

For example in chapter 6 we identified all mutations in CCLE and TCGA in the critical lysine of VAIK, the E that salt-bridges to the lysine, HRD, DFG, APE, the first 2 glycines of GxGxxG, and the conserved G at APE-6 to produce the top kinases mutated in these LOF regions. The approach was validated by the top hits in the table including BRAF and STK11, both kinases with well-documented LOF roles in carcinogenesis [142, 143]. EPHB1, CHEK2 and TGFBR1 feature in the top 10 and have also been documented as playing tumour suppressing roles in a variety of tumour subtypes[144-146]. However, MAP2K7 (MKK7), 4th highest on the list, had not been reported previously to have LOF mutations in cancer. Five out of the 10 mutations were present in gastric cancer and querying the TCGA gastric cancer dataset revealed 7% of samples possessed a MKK7 alteration with many truncating, frameshift and high functional predicted missense mutations [147]. Due to the high mutational burden of gastric cancer, querying the mutational frequency of genes in TCGA gastric cancer samples reveals that based on length-corrected mutation score MAP2K7 is only 145th highest. Based on this modest position in the mutational frequency charts most researchers would overlook MAP2K7 as potential driver of gastric cancer. MKK7 is known to phosphorylate and activate JNK [148]. Expanding the cBio query to include MAPK8, MAPK9, MAPK10, MAP2K4 and ATF2 reveals that

22% of TCGA gastric cancer cases have a mutation in a JNK pathway gene. Furthermore there is a high degree of mutual exclusivity of mutations in these genes suggesting JNK pathway inactivation may be present in a high proportion of gastric cancers. The MKK7 mutations in gastric cancer were all confirmed to be LOF upon the JNK pathway and the stable expression of wild-type MKK7 in mutant cell lines shows decreased colony formation. These data indicate that MKK7 is a tumour suppressor in gastric cancer and we currently exploring mechanisms to exploit the loss of the JNK pathway therapeutically.

Molecular Dynamics simulations are another in silico method that can be used to predict the functional impact of mutations. We used these simulations to assess structural impact of the novel LOF mutation hotspot at APE-6 in MKK4. The simulation indicated that replacing the glycine at 276 with aspartic acid leads to decreased flexibility, preventing movement of the P+1 loop into an active confirmation. Biochemical validation has confirmed that this mutation is LOF. Molecular Dynamics technology will improve as knowledge on protein structure and function advances however currently it is not accurate enough to replace biochemical validation. Moreover these simulations require a large amount of computer processing power and in our institute typically take 2 to 4 weeks to run. Therefore biochemical assessment with site-directed mutagenesis and transient over-expression can often be completed in a shorter time period. As computer-processing power increases then these simulations will be performed much faster and may provide a valuable screening tool for assessing all mutations in a dataset.

8.3.4 Premalignant and Predisposition Condition Analysis

The multistep model of carcinogenesis (Vogelstein 1993) describes how a number of driver mutations are required to transform a normal cell into a tumour cell. Because mutations occur through unpredictable DNA events via the action of a carcinogen the cell must often acquire many hundreds of inconsequential passenger mutations to obtain the required 3-6 driver mutations. Occasionally a DNA damage repair protein is mutated leading to impaired mutational repair and an even greater number of passenger mutations. These factors lead to the mutational noise discussed in the preceding section, requiring sophisticated in silico and biochemical filtering approaches to identify the drivers. However a potentially more straightforward approach to identifying driver mutations is to sequence the cell at an earlier stage in

the carcinogenic process before the accumulation of passenger mutations. This is easier said than done as in many cancers there are no known premalignant lesions. Furthermore, even in cancers with associated premalignant lesions most patients do not present with premalignant disease but later with a fully developed invasive cancer. Finally, the stage at which a premalignant lesion becomes clinically detectable often means that the premalignant cells have acquired a large number of passenger mutations already. For example, a sequencing study of Barrett's oesophagus (a premalignant lesion that develops into oesophageal adenocarcinoma) revealed a high mutational burden in the lesions and lesion heterogeneity [149]. As a result this study did not assist greatly in deciphering the mutational processes occurring in identifying novel drivers of oesophageal adenocarcinoma.

Another approach is to use analysis of germline variants in patients with a genetic predisposition to developing cancer to identify driver events that may occur somatically in similar tumours. An example of this approach is highlighted by the discovery of the genetic basis of familial adenomatous polyposis (FAP), which began as a single case report of a patient with developmental abnormalities, multiple colonic polyps and soft tissue tumours [150]. Analysis of this patient showed a portion of 5q chromosome was missing. It then took 5 years, in the pre-NGS era, for further patients with FAP with similar deletions in 5q to identify the causative tumour suppressor gene APC [151]. Large sequencing studies now reveal APC mutations or loss in a majority of colon cancers. The APC story in colorectal cancer demonstrates that there is a great potential in analysing the germline data of individuals with cancer predisposition syndromes. This approach is much easier in diseases with very characteristic phenotypes where patients often present at an early age with a recognisable pattern of lesions and a strong family history. Things become much more difficult for less conspicuous clinical presentations that may represent more than one syndrome. It also helps, as in the case of APC, if there is a genetic feature such as a large deletion that can focus the attention to the genes in that region.

We performed whole exome sequencing on 10 patients with a rare condition called DIPNECH in which typically middle-aged patients present with areas of neuroendocrine cell proliferations within the lung. Our patient population was heterogeneous with some typical middle-aged non-smoking female patients along with less typical heavy smoking older males. Unsurprisingly, given the different clinical presentations in our case series, we did not find any novel SNVs present in all 10 cases. Furthermore we did not identify any novel SNVs in more than one sample and

at a gene level only TTN (the largest protein coding gene and therefore the most likely to possess multiple protein coding mutations) had missense mutations in more than one case. By honing down on SNVs occurring in critical locations in kinases using the same pipeline we used for somatic driver analysis we identified a SNP in the kinase domain of CDK8 in one patient that was predicted to be highly functional. This variant is a significant amino acid change in a highly conserved critical region of the kinase domain. Given that DIPNECH is a condition of abnormal proliferation of neuroendocrine cells, this discovery led us to enquire about the role of CDK8 in neuroendocrine malignancies. We found similar predicted damaging mutations and homozygous deletions in a significant proportion of small cell lung cancer cell lines. Work is on-going to assess the biochemical consequences of these variants and work out the mechanisms involved to determine if LOF mutations in CDK8 drive the carcinogenic process of neuroendocrine cells.

8.4 Future Work

In addition to validating the biochemical mechanisms of the identified targets such as CDK8, future work derived from this thesis will be based upon further linkage of functional data to genomic datasets to discover novel drivers or therapeutic vulnerabilities in other settings. The kinase motif mutational screen employed in chapters 6 and 7 demonstrates the value of performing specific functional screens of large mutational databases. Kinases were ideal to evaluate this approach with due to the large number of proteins in the human kinome with shared structure and function. There is also the benefit that kinase structure and its relationship with function have been relatively well studied compared to other families of proteins. However as the structural biology of other classes of proteins become known, similar screens could be employed. The next logical step in this regard would be to screen for mutational data in phosphatases that counterbalance the phosphorylation signalling by kinases. Loss of phosphatase activity through mutation of critical residues would result in a reduction of dephosphorylation, which in some situations could cause a pathway to be hyperactive and oncogenic. These activated pathways could be targeted with small molecule inhibitors for potential clinical benefit.

The functional screens used so far have focussed on the mutation of residues critical for kinase function and hence loss of that function. However, it is possible to modify the screen to identify gain of function mutations. Like the loss of function screens

there are two approaches that can be taken. Firstly, direct linkage of predicted gain of function mutations from basic structural knowledge can be applied. For example, there is a subset of kinases with tyrosine immediately prior to the third glycine of the GxGxxG loop. As discussed previously the glycine-rich loop depends on flexibility to position ATP, with phosphorylation of the tyrosine preventing ATP binding. Therefore phosphorylation of the tyrosine can be a control mechanism used to turn off kinase activity. Mutation of the tyrosine to another amino acid that can't be phosphorylated results in increased kinase activity [92]. Screening for these specific mechanisms including phospho-mimetic mutants would allow a gain-of-function analysis of pan cancer genomics data. Secondly, a GOF screen can be developed without specific knowledge of kinase inhibitory mechanisms but using the pan-cancer GOF mutational data itself. By mapping well known GOF mutants to aligned kinase domain sequences, further gain of function mutants in novel kinases can be discovered. For example, the highly activating V600E BRAF mutation occurs at the DFG+4 position. Performing a pan-cancer analysis of the DFG+4 position reveals high frequency known activating mutations in FLT3 (D835), EGFR (L861), and KIT (D816) alongside many other novel kinases with lower frequency mutations. Combining these two approaches for GOF predictions may assist with novel oncogene detection.

Chapters 6 and 7 demonstrated the benefits of the LOF screen in two different scenarios, large-scale pan-cancer data and a smaller SNV study. Applying the screen to the mutational profile of specific cell lines has produced targets for further investigation in my own field of interest, radiation oncology. Yard et al. tested the radiosensitivity of 533 cell lines that had been sequenced by CCLE [152]. Selecting only cell lines of known radioresistant subtypes and retaining the 30 most radiosensitive of these produced a list of radiosensitive outliers. Applying the kinase motif screen to the mutational profiles of these outliers identified damaging missense kinase mutations in uncharacteristically radiosensitive cell lines. Combining this data with an additional screen for truncating kinase mutations in these cell lines produced a frequency table of kinases with LOF mutations in radiosensitive outliers. The top hit from this analysis is NUAK1, a kinase that is thought to inactivate PLK1 and prevent cell cycle progression through the mitotic checkpoint [153]. It is hypothesised that LOF mutations in NUAK1 prevent mitotic arrest post-radiotherapy reducing the time available to repair the DNA damage, leading to mitotic catastrophe and thus enhanced sensitivity to radiotherapy. Interestingly many typically radioresistant tumour subtypes, such as pancreatic and stomach cancer, express higher levels of NUAK1. The advantage of using the CCLE data to discover novel targets for

investigation is that, by virtue of their discovery by the screen, there are proven available radiosensitive cell lines with the desired mutation that can be reverted to wild type using CRISPR technology to observe directly if radiosensitivity is abolished. If proven the mechanism could be exploited clinically with a NUAK1 inhibitor given in combination with radiotherapy to increase the radiosensitivity of tumours.

8.5 Overall Conclusions

The work presented in this thesis resulted from the early failure of an initial project focussed upon a specific kinase target in lung cancer. Chosen by mutational frequency alone the target was not expressed in lung cancer cell lines. The relatively high mutational frequency of this gene was likely secondary to a deficiency of transcription-coupled repair. Searching the mutational datasets for a new target to study I quickly became aware of the discrepancies between the mutational profiling of different sequencing projects. Noticing large areas of genes, often encoded by the first exon, with a clear reduction in mutational frequency compared to the rest of the gene led me to hypothesise that performance in sequencing GC-rich regions was contributing to the discrepancies between datasets. Modern NGS of a limited number of these lung cancer cell lines revealed a functional PAK4 mutation previously missed by COSMIC and CCLE due to impaired sequencing of a GC-rich region. Improved NGS technology will reveal if further high frequency drivers are hidden in these high GC-content sequencing cold spots. It remains to be seen if cancer subtypes with a predisposition to GC-region mutations, such as those associated with cigarette smoke, are hiding high frequency mutations in these regions.

Acknowledging the limitations of NGS but still searching for a novel kinase to study I turned my attention to filtering the available data to find targets for biochemical validation. I was inspired by the different methods my lab colleagues were using to make kinase-dead constructs. Most commonly the lysine of the VAIK motif is mutated to produce a construct with very little kinase activity. This led me to speculate how often this specific mutation occurs in cancer samples in the human kinome. Expanding my knowledge of kinase structure and function I incorporated additional critical residues to create a mutational screen of 411 kinases in the TCGA and CCLE datasets. This identified many novel kinases with mutations predicted to be highly likely to be LOF. MKK7 stood out from this list due to the predilection of a high

frequency of LOF mutations in gastric cancer. These mutations were all validated biochemically to be LOF and colony formation assays demonstrated a tumour suppressive role in gastric cancer. A similar kinase motif screen was used to filter the SNV data of a premalignant neuroendocrine disease and identify a novel CDK8 variant likely to kill kinase activity. This discovery prompted enquiry into CDK8 somatic mutations in TCGA and CCLE, revealing a proportion of SCLC cell lines with similar critical residue CDK8 mutations and suggesting a genetic link between the two conditions. These data highlight the power of using structural and functional considerations to filter genomic data to find driver variants. The DIPNECH data specifically highlights the value of investigating premalignant conditions to aid driver mutation detection in the higher mutational burden cancer samples. Future work will continue to develop additional functional screens in kinases and other families of proteins, applying these to different datasets to assist the discovery of driver mutations.

Chapter Nine

References

References

1. Dahm R. Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Hum Genet* 122(6), 565-581 (2008).
2. Balmain A. Cancer genetics: From boveri and mendel to microarrays. *Nat Rev Cancer* 1(1), 77-82 (2001).
3. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356), 737-738 (1953).
4. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. *Nature* 171(4356), 740-741 (1953).
5. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* 36(1), 39-56 (1952).
6. Bessman MJ, Kornberg A, Lehman IR, Simms ES. Enzymic synthesis of deoxyribonucleic acid. *Biochim Biophys Acta* 21(1), 197-198 (1956).
7. Brenner S, Jacob F, Meselson M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190, 576-581 (1961).
8. Matthaei JH, Jones OW, Martin RG, Nirenberg MW. Characteristics and composition of rna coding units. *Proc Natl Acad Sci U S A* 48, 666-677 (1962).
9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12), 5463-5467 (1977).
10. Martin GS. The road to src. *Oncogene* 23(48), 7910-7917 (2004).
11. Hunter T, Sefton BM. Transforming gene product of rous sarcoma virus phosphorylates tyrosine. *Proc Natl Acad Sci U S A* 77(3), 1311-1315 (1980).
12. Malumbres M, Barbacid M. Ras oncogenes: The first 30 years. *Nat Rev Cancer* 3(6), 459-465 (2003).
13. Downward J, Yarden Y, Mayes E *et al*. Close similarity of epidermal growth factor receptor and v-erb-b oncogene protein sequences. *Nature* 307(5951), 521-527 (1984).
14. Cerami E, Gao J, Dogrusoz U *et al*. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5), 401-404 (2012).
15. Witkiewicz AK, Mcmillan EA, Balaji U *et al*. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* 6, 6744 (2015).
16. Riley DJ, Lee EY, Lee WH. The retinoblastoma protein: More than a tumor suppressor. *Annu Rev Cell Biol* 10, 1-29 (1994).
17. Levine AJ, Oren M. The first 30 years of p53: Growing ever more complex. *Nat Rev Cancer* 9(10), 749-758 (2009).
18. Nordling CO. A new theory on cancer-inducing mechanism. *Br J Cancer* 7(1), 68-72 (1953).
19. Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A* 112(1), 118-123 (2015).
20. Knudson AG, Jr. Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68(4), 820-823 (1971).

21. Cavenee WK, Dryja TP, Phillips RA *et al*. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma*. Nature* 305(5937), 779-784 (1983).
22. Wheeler DA, Wang L. From human genome to cancer genome: The first decade*. Genome Res* 23(7), 1054-1062 (2013).
23. Lander ES, Linton LM, Birren B *et al*. Initial sequencing and analysis of the human genome*. Nature* 409(6822), 860-921 (2001).
24. Shendure J, Ji HL. Next-generation DNA sequencing*. Nat Biotechnol* 26(10), 1135-1145 (2008).
25. Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design*. Nat Rev Genet* 14(5), 321-332 (2013).
26. Schatz MC. Biological data sciences in genome research*. Genome Res* 25(10), 1417-1422 (2015).
27. Cohen S, Carpenter G, King L, Jr. Epidermal growth factor-receptor-protein kinase interactions. Co-purification of receptor and epidermal growth factor-enhanced phosphorylation activity*. J Biol Chem* 255(10), 4834-4842 (1980).
28. Sato JD, Kawamoto T, Le AD, Mendelsohn J, Polikoff J, Sato GH. Biological effects in vitro of monoclonal antibodies to human epidermal growth factor receptors*. Mol Biol Med* 1(5), 511-529 (1983).
29. Aboud-Pirak E, Hurwitz E, Pirak ME, Bellot F, Schlessinger J, Sela M. Efficacy of antibodies to epidermal growth factor receptor against kb carcinoma in vitro and in nude mice*. J Natl Cancer Inst* 80(20), 1605-1611 (1988).
30. Cunningham D, Humblet Y, Siena S *et al*. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer*. N Engl J Med* 351(4), 337-345 (2004).
31. Karapetis CS, Khambata-Ford S, Jonker DJ *et al*. K-ras mutations and benefit from cetuximab in advanced colorectal cancer*. N Engl J Med* 359(17), 1757-1765 (2008).
32. Bonner JA, Harari PM, Giralt J *et al*. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck*. N Engl J Med* 354(6), 567-578 (2006).
33. Lynch TJ, Bell DW, Sordella R *et al*. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib*. N Engl J Med* 350(21), 2129-2139 (2004).
34. Pao W, Miller V, Zakowski M *et al*. Egf receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib*. Proc Natl Acad Sci U S A* 101(36), 13306-13311 (2004).
35. Mok TS, Wu YL, Thongprasert S *et al*. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma*. N Engl J Med* 361(10), 947-957 (2009).
36. Venter DJ, Tuzi NL, Kumar S, Gullick WJ. Overexpression of the c-erbb-2 oncoprotein in human breast carcinomas: Immunohistological assessment correlates with gene amplification*. Lancet* 2(8550), 69-72 (1987).
37. Press MF, Bernstein L, Thomas PA *et al*. Her-2/neu gene amplification characterized by fluorescence in situ hybridization: Poor prognosis in node-negative breast carcinomas*. J Clin Oncol* 15(8), 2894-2904 (1997).

38. Drebin JA, Link VC, Weinberg RA, Greene MI. Inhibition of tumor growth by a monoclonal antibody reactive with an oncogene-encoded tumor antigen. *Proc Natl Acad Sci U S A* 83(23), 9129-9133 (1986).
39. Geyer CE, Forster J, Lindquist D *et al*. Lapatinib plus capecitabine for her2-positive advanced breast cancer. *N Engl J Med* 355(26), 2733-2743 (2006).
40. Furitsu T, Tsujimura T, Tono T *et al*. Identification of mutations in the coding sequence of the protooncogene c-kit in a human mast-cell leukemia-cell line causing ligand-independent activation of c-kit product. *J Clin Invest* 92(4), 1736-1744 (1993).
41. Hirota S, Isozaki K, Moriyama Y *et al*. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science* 279(5350), 577-580 (1998).
42. Demetri GD, Von Mehren M, Blanke CD *et al*. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *New Engl J Med* 347(7), 472-480 (2002).
43. Growney JD, Clark JJ, Adelsperger J *et al*. Activation mutations of human c-kit resistant to imatinib, mesylate are sensitive to the tyrosine kinase inhibitor pkc412. *Blood* 106(2), 721-724 (2005).
44. Ferrara N, Hillan KJ, Gerber HP, Novotny W. Discovery and development of bevacizumab, an anti-vegf antibody for treating cancer. *Nat Rev Drug Discov* 3(5), 391-400 (2004).
45. Shih T, Lindley C. Bevacizumab: An angiogenesis inhibitor for the treatment of solid malignancies. *Clin Ther* 28(11), 1779-1802 (2006).
46. Rapp UR, Goldsborough MD, Mark GE *et al*. Structure and biological activity of v-raf, a unique oncogene transduced by a retrovirus. *Proc Natl Acad Sci U S A* 80(14), 4218-4222 (1983).
47. Solit D, Rosen N. Oncogenic raf: A brief history of time. *Pigment Cell Melanoma Res* 23(6), 760-762 (2010).
48. Davies H, Bignell GR, Cox C *et al*. Mutations of the braf gene in human cancer. *Nature* 417(6892), 949-954 (2002).
49. Chapman PB, Hauschild A, Robert C *et al*. Improved survival with vemurafenib in melanoma with braf v600e mutation. *N Engl J Med* 364(26), 2507-2516 (2011).
50. Goldman JM, Melo JV. Bcr-abl in chronic myelogenous leukemia--how does it work? *Acta Haematol* 119(4), 212-217 (2008).
51. Ren R. Mechanisms of bcr-abl in the pathogenesis of chronic myelogenous leukaemia. *Nat Rev Cancer* 5(3), 172-183 (2005).
52. Soda M, Choi YL, Enomoto M *et al*. Identification of the transforming eml4-alk fusion gene in non-small-cell lung cancer. *Nature* 448(7153), 561-566 (2007).
53. Sasaki T, Rodig SJ, Chirieac LR, Janne PA. The biology and treatment of eml4-alk non-small cell lung cancer. *Eur J Cancer* 46(10), 1773-1780 (2010).
54. Bayliss R, Choi J, Fennell DA, Fry AM, Richards MW. Molecular mechanisms that underpin eml4-alk driven cancers and their response to targeted drugs. *Cell Mol Life Sci* 73(6), 1209-1224 (2016).
55. Shaw AT, Kim DW, Nakagawa K *et al*. Crizotinib versus chemotherapy in advanced alk-positive lung cancer. *N Engl J Med* 368(25), 2385-2394 (2013).

56. Wood LD, Parsons DW, Jones S *et al*. The genomic landscapes of human breast and colorectal cancers. *Science* 318(5853), 1108-1113 (2007).
57. Parsons DW, Jones S, Zhang X *et al*. An integrated genomic analysis of human glioblastoma multiforme. *Science* 321(5897), 1807-1812 (2008).
58. Brennan CW, Verhaak RG, Mckenna A *et al*. The somatic genomic landscape of glioblastoma. *Cell* 155(2), 462-477 (2013).
59. Cohen AL, Holmen SL, Colman H. Idh1 and idh2 mutations in gliomas. *Curr Neurol Neurosci* 13(5), (2013).
60. Curtis C, Shah SP, Chin SF *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486(7403), 346-352 (2012).
61. Ali HR, Rueda OM, Chin SF *et al*. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol* 15(8), 431 (2014).
62. Testoni E, Stephenson NL, Torres-Ayuso P *et al*. Somatically mutated abl1 is an actionable and essential nsclc survival gene. *EMBO Mol Med* 8(2), 105-116 (2016).
63. Gerlinger M, Rowan AJ, Horswell S *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10), 883-892 (2012).
64. Greaves M, Maley CC. Clonal evolution in cancer. *Nature* 481(7381), 306-313 (2012).
65. Gundem G, Van Loo P, Kremeyer B *et al*. The evolutionary history of lethal metastatic prostate cancer. *Nature* 520(7547), 353-357 (2015).
66. Girotti MR, Gremel G, Lee R *et al*. Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma. *Cancer Discov* 6(3), 286-299 (2016).
67. Wang H, Nettleton D, Ying K. Copy number variation detection using next generation sequencing read counts. *BMC Bioinformatics* 15, 109 (2014).
68. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 7(2), 85-97 (2006).
69. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 1(6), 62 (2009).
70. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11(10), 685-696 (2010).
71. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* 13(1), 36-46 (2011).
72. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* 12(4), 252-264 (2012).
73. Linsley PS, Wallace PM, Johnson J *et al*. Immunosuppression in vivo by a soluble form of the ctla-4 t cell activation molecule. *Science* 257(5071), 792-795 (1992).
74. Liu X, Gao JX, Wen J *et al*. B7dc/pdl2 promotes tumor immunity by a pd-1-independent mechanism. *J Exp Med* 197(12), 1721-1730 (2003).
75. Rizvi NA, Hellmann MD, Snyder A *et al*. Cancer immunology. Mutational landscape determines sensitivity to pd-1 blockade in non-small cell lung cancer. *Science* 348(6230), 124-128 (2015).

76. Tumeh PC, Harview CL, Yearley JH *et al*. Pd-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* 515(7528), 568-571 (2014).
77. Lawrence MS, Stojanov P, Mermel CH *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484), 495-501 (2014).
78. Ledford H. End of cancer-genome project prompts rethink. *Nature* 517(7533), 128-129 (2015).
79. Johnson LN. The regulation of protein phosphorylation. *Biochem Soc Trans* 37(Pt 4), 627-641 (2009).
80. Ubersax JA, Ferrell JE, Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol* 8(7), 530-541 (2007).
81. Burnett G, Kennedy EP. The enzymatic phosphorylation of proteins. *J Biol Chem* 211(2), 969-980 (1954).
82. Cohen P. The origins of protein phosphorylation. *Nat Cell Biol* 4(5), E127-130 (2002).
83. Walsh DA, Perkins JP, Krebs EG. An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle. *J Biol Chem* 243(13), 3763-3765 (1968).
84. Hunter T. Tony hunter: Kinase king. Interview by ruth williams. *J Cell Biol* 181(4), 572-573 (2008).
85. Besant PG, Tan E, Attwood PV. Mammalian protein histidine kinases. *Int J Biochem Cell Biol* 35(3), 297-309 (2003).
86. Roskoski R, Jr. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol Res* 100, 1-23 (2015).
87. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 298(5600), 1912-1934 (2002).
88. Endicott JA, Noble ME, Johnson LN. The structural basis for control of eukaryotic protein kinases. *Annu Rev Biochem* 81, 587-613 (2012).
89. Taylor SS, Kornev AP. Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem Sci* 36(2), 65-77 (2011).
90. Kornev AP, Haste NM, Taylor SS, Ten Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci U S A* 103(47), 17783-17788 (2006).
91. Hemmer W, Mcglone M, Tsigelny I, Taylor SS. Role of the glycine triad in the atp-binding site of camp-dependent protein kinase. *J Biol Chem* 272(27), 16946-16954 (1997).
92. Coulonval K, Bockstaele L, Paternot S, Roger PP. Phosphorylations of cyclin-dependent kinase 2 revisited using two-dimensional gel electrophoresis. *J Biol Chem* 278(52), 52052-52060 (2003).
93. Bartova I, Otyepka M, Kriz Z, Koca J. The mechanism of inhibition of the cyclin-dependent kinase-2 as revealed by the molecular dynamics study on the complex cdk2 with the peptide substrate hhasprk. *Protein Sci* 14(2), 445-451 (2005).
94. Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan M, Taylor SS. Dynamics of camp-dependent protein kinase. *Chem Rev* 101(8), 2243-2270 (2001).

95. Nolen B, Taylor S, Ghosh G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* 15(5), 661-675 (2004).
96. Kornev AP, Haste NM, Taylor SS, Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci U S A* 103(47), 17783-17788 (2006).
97. Steichen JM, Iyer GH, Li S *et al*. Global consequences of activation loop phosphorylation on protein kinase a. *J Biol Chem* 285(6), 3825-3832 (2010).
98. Kornev AP, Taylor SS, Ten Eyck LF. A helix scaffold for the assembly of active protein kinases. *Proc Natl Acad Sci U S A* 105(38), 14377-14382 (2008).
99. Sefton BM. Overview of protein phosphorylation. *Curr Protoc Cell Biol* Chapter 14, Unit 14 11 (2001).
100. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 100(1), 57-70 (2000).
101. Fedorov O, Muller S, Knapp S. The (un)targeted cancer kinome. *Nat Chem Biol* 6(3), 166-169 (2010).
102. Garcia-Ibarbia C, Delgado-Calle J, Casafont I *et al*. Contribution of genetic and epigenetic mechanisms to wnt pathway activity in prevalent skeletal disorders. *Gene* 532(2), 165-172 (2013).
103. Manning BD. Challenges and opportunities in defining the essential cancer kinome. *Sci Signal* 2(63), (2009).
104. Walker I, Newell H. Do molecularly targeted agents in oncology have reduced attrition rates? *Nat Rev Drug Discov* 8(1), 15-16 (2009).
105. Aguayo SM, Miller YE, Waldron JA, Jr. *et al*. Brief report: Idiopathic diffuse hyperplasia of pulmonary neuroendocrine cells and airways disease. *N Engl J Med* 327(18), 1285-1288 (1992).
106. Killen H. Dipnech presenting on a background of malignant melanoma: New lung nodules are not always what they seem. *BMJ Case Rep* 2014, (2014).
107. Carr LL, Chung JH, Duarte Achcar R *et al*. The clinical course of diffuse idiopathic pulmonary neuroendocrine cell hyperplasia. *Chest* 147(2), 415-422 (2015).
108. Nassar AA, Jaroszewski DE, Helmers RA, Colby TV, Patel BM, Mookadam F. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia: A systematic overview. *Am J Respir Crit Care Med* 184(1), 8-16 (2011).
109. Stenzinger A, Weichert W, Hensel M, Bruns H, Dietel M, Erbersdobler A. Incidental postmortem diagnosis of dipnech in a patient with previously unexplained 'asthma bronchiale'. *Pathol Res Pract* 206(11), 785-787 (2010).
110. Foran PJ, Hayes SA, Blair DJ, Zakowski MF, Ginsberg MS. Imaging appearances of diffuse idiopathic pulmonary neuroendocrine cell hyperplasia. *Clin Imaging*, (2014).
111. Davies SJ, Gosney JR, Hansell DM *et al*. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia: An under-recognised spectrum of disease. *Thorax* 62(3), 248-252 (2007).
112. Patel C, Tirukonda P, Bishop R, Mulatero C, Scarsbrook A. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia (dipnech)

masquerading as metastatic carcinoma with multiple pulmonary deposits. *Clin Imaging* 36(6), 833-836 (2012).
113. Gorshtein A, Gross DJ, Barak D *et al*. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia and the associated lung neuroendocrine tumors: Clinical experience with a rare entity. *Cancer* 118(3), 612-619 (2012).
114. Zhou H, Ge Y, Janssen B *et al*. Double lung transplantation for diffuse idiopathic pulmonary neuroendocrine cell hyperplasia. *J Bronchology Interv Pulmonol* 21(4), 342-345 (2014).
115. Falkenstern-Ge RF, Kimmich M, Friedel G, Tannapfel A, Neumann V, Kohlhaeufl M. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia: 7-year follow-up of a rare clinicopathologic syndrome. *J Cancer Res Clin Oncol* 137(10), 1495-1498 (2011).
116. Miller RR, Muller NL. Neuroendocrine cell hyperplasia and obliterative bronchiolitis in patients with peripheral carcinoid tumors. *Am J Surg Pathol* 19(6), 653-658 (1995).
117. Gosney JR. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia as a precursor to pulmonary neuroendocrine tumors. *Chest* 125(5 Suppl), 108S (2004).
118. Beasley MB, Brambilla E, Travis WD. The 2004 world health organization classification of lung tumors. *Semin Roentgenol* 40(2), 90-97 (2005).
119. Mireskandari M, Abdirad A, Zhang Q, Dietel M, Petersen I. Association of small foci of diffuse idiopathic pulmonary neuroendocrine cell hyperplasia (dipnech) with adenocarcinoma of the lung. *Pathol Res Pract* 209(9), 578-584 (2013).
120. Johannessen CM, Boehm JS, Kim SY *et al*. Cot drives resistance to raf inhibition through map kinase pathway reactivation. *Nature* 468(7326), 968-972 (2010).
121. Mckenna A, Hanna M, Banks E *et al*. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9), 1297-1303 (2010).
122. Gao J, Aksoy BA, Dogrusoz U *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci Signal* 6(269), pl1 (2013).
123. Barretina J, Caponigro G, Stransky N *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391), 603-607 (2012).
124. Forbes SA, Beare D, Gunasekaran P *et al*. Cosmic: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43(Database issue), D805-811 (2015).
125. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (igv): High-performance genomics data visualization and exploration. *Brief Bioinform* 14(2), 178-192 (2013).
126. Krzywinski M, Schein J, Birol I *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res* 19(9), 1639-1645 (2009).
127. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. Gromacs: Fast, flexible, and free. *J Comput Chem* 26(16), 1701-1718 (2005).

128. Hudson AM, Yates T, Li Y *et al*. Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. *Cancer Res* 74(22), 6390-6396 (2014).
129. Benson DA, Cavanaugh M, Clark K *et al*. Genbank. *Nucleic Acids Res* 41(Database issue), D36-42 (2013).
130. Fernandez-Cuesta L, Peifer M, Lu X *et al*. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat Commun* 5, 3518 (2014).
131. Pleasance ED, Stephens PJ, O'meara S *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463(7278), 184-190 (2010).
132. Kadara H, Wistuba, Ii. Field cancerization in non-small cell lung cancer: Implications in disease pathogenesis. *Proc Am Thorac Soc* 9(2), 38-42 (2012).
133. George J, Lim JS, Jang SJ *et al*. Comprehensive genomic profiles of small cell lung cancer. *Nature* 524(7563), 47-U73 (2015).
134. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417), 519-525 (2012).
135. Lawrence MS, Stojanov P, Polak P *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457), 214-218 (2013).
136. Antal CE, Hudson AM, Kang E *et al*. Cancer-associated protein kinase c mutations reveal kinase's role as tumor suppressor. *Cell* 160(3), 489-502 (2015).
137. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. Mutationtaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 7(8), 575-576 (2010).
138. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat Protoc* 4(7), 1073-1081 (2009).
139. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39(17), e118 (2011).
140. Adzhubei IA, Schmidt S, Peshkin L *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 7(4), 248-249 (2010).
141. Dong C, Wei P, Jian X *et al*. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Hum Mol Genet*, (2014).
142. Hezel AF, Bardeesy N. Lkb1; linking cell structure and tumor suppression. *Oncogene* 27(55), 6908-6919 (2008).
143. Heidorn SJ, Milagre C, Whittaker S *et al*. Kinase-dead braf and oncogenic ras cooperate to drive tumor progression through craf. *Cell* 140(2), 209-221 (2010).
144. Kampen KR, Scherpen FJ, Garcia-Manero G *et al*. Ephb1 suppression in acute myelogenous leukemia: Regulating the DNA damage control system. *Mol Cancer Res* 13(6), 982-992 (2015).
145. Stolz A, Ertych N, Bastians H. Tumor suppressor chk2: Regulator of DNA damage response and mediator of chromosomal stability. *Clin Cancer Res* 17(3), 401-405 (2011).

146. Xu Y, Pasche B. Tgf-beta signaling alterations and susceptibility to colorectal cancer. *Hum Mol Genet* 16 Spec No 1, R14-20 (2007).
147. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513(7517), 202-209 (2014).
148. Tournier C, Whitmarsh AJ, Cavanagh J, Barrett T, Davis RJ. The mkk7 gene encodes a group of c-jun nh2-terminal kinase kinases. *Mol Cell Biol* 19(2), 1569-1581 (1999).
149. Ross-Innes CS, Becq J, Warren A *et al*. Whole-genome sequencing provides new insights into the clonal architecture of barrett's esophagus and esophageal adenocarcinoma. *Nat Genet* 47(9), 1038-1046 (2015).
150. Herrera L, Kakati S, Gibas L, Pietrzak E, Sandberg AA. Gardner syndrome in a man with an interstitial deletion of 5q. *Am J Med Genet* 25(3), 473-476 (1986).
151. Kinzler KW, Nilbert MC, Vogelstein B *et al*. Identification of a gene located at chromosome-5q21 that is mutated in colorectal cancers. *Science* 251(4999), 1366-1370 (1991).
152. Yard BD, Adams DJ, Chie EK *et al*. A genetic basis for the variation in the vulnerability of cancer to DNA damage. *Nat Commun* 7, (2016).
153. Banerjee S, Zagorska A, Deak M, Campbell DG, Prescott AR, Alessi DR. Interplay between polo kinase, lkb1-activated nuak1 kinase, pp1betamypt1 phosphatase complex and the scfbetatrcp e3 ubiquitin ligase. *Biochem J* 461(2), 233-245 (2014).

Chapter Ten

Appendices

10.1 Appendix One:

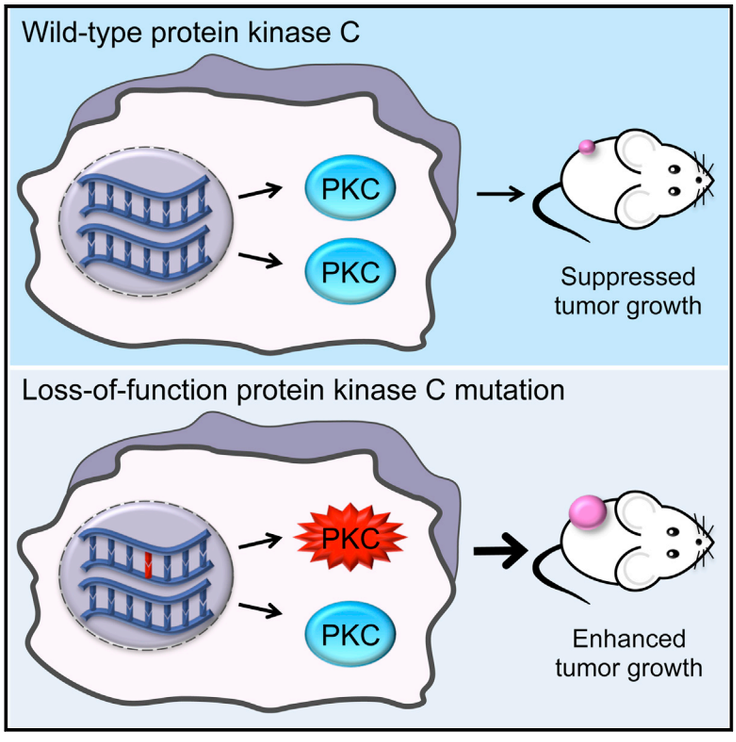
Antal CE, **Hudson AM**, Kang E, Wirth C, Stephenson NL, Trotter EW, Zanca C, Gallegos LL, Furnari FB, Miller CW, Hunter T, Brognard J, Newton AC. Protein Kinase C Loss-of-Function Mutations in Cancer Reveal Role as Tumor Suppressor, ***Cell*** 2015: 160(3); 489- 502

Article

Cell

Cancer-Associated Protein Kinase C Mutations Reveal Kinase's Role as Tumor Suppressor

Graphical Abstract



Authors

Corina E. Antal, Andrew M. Hudson, ..., John Brognard, Alexandra C. Newton

Correspondence

john.brognard@cruk.manchester.ac.uk (J.B.),
anewton@ucsd.edu (A.C.N.)

In Brief

Cancer-associated kinase mutations have generally been characterized as oncogenic, but an analysis of PKC mutations reveals that the majority are loss of function, indicating a tumor-suppressive role for this kinase and a shift in therapeutic strategies targeting PKC.

Highlights

- Cancer-associated PKC mutations are LOF and can act in a dominant-negative manner
- Correcting a heterozygous PKCβ LOF mutation reduces tumor volume
- Hemizygous deletion shows PKC is haploinsufficient for tumor suppression
- Therapeutic strategies should aim to restore PKC activity instead of inhibiting it

Antal et al., 2015, Cell *160*, 489–502
January 29, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.01.001

CellPress

Article

Cell

Cancer-Associated Protein Kinase C Mutations Reveal Kinase's Role as Tumor Suppressor

Corina E. Antal,[1,2] Andrew M. Hudson,[3] Emily Kang,[1] Ciro Zanca,[4] Christopher Wirth,[5] Natalie L. Stephenson,[3] Eleanor W. Trotter,[3] Lisa L. Gallegos,[1,2,7] Crispin J. Miller,[5] Frank B. Furnari,[4] Tony Hunter,[6] John Brognard,[3,*] and Alexandra C. Newton[1,*]

[1]Department of Pharmacology, University of California at San Diego, La Jolla, CA 92093, USA
[2]Biomedical Sciences Graduate Program, University of California at San Diego, La Jolla, CA 92093, USA
[3]Signalling Networks in Cancer Group, Cancer Research UK Manchester Institute, University of Manchester, Manchester M20 4BX, UK
[4]Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA 92093, USA
[5]Applied Computational Biology and Bioinformatics Group, Cancer Research UK Manchester Institute, University of Manchester, Manchester M20 4BX, UK
[6]The Salk Institute, La Jolla, CA 92037, USA
[7]Present address: Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: john.brognard@cruk.manchester.ac.uk (J.B.), anewton@ucsd.edu (A.C.N.)
http://dx.doi.org/10.1016/j.cell.2015.01.001

SUMMARY

Protein kinase C (PKC) isozymes have remained elusive cancer targets despite the unambiguous tumor promoting function of their potent ligands, phorbol esters, and the prevalence of their mutations. We analyzed 8% of PKC mutations identified in human cancers and found that, surprisingly, most were loss of function and none were activating. Loss-of-function mutations occurred in all PKC subgroups and impeded second-messenger binding, phosphorylation, or catalysis. Correction of a loss-of-function PKCβ mutation by CRISPR-mediated genome editing in a patient-derived colon cancer cell line suppressed anchorage-independent growth and reduced tumor growth in a xenograft model. Hemizygous deletion promoted anchorage-independent growth, revealing that PKCβ is haploinsufficient for tumor suppression. Several mutations were dominant negative, suppressing global PKC signaling output, and bioinformatic analysis suggested that PKC mutations cooperate with co-occurring mutations in cancer drivers. These data establish that PKC isozymes generally function as tumor suppressors, indicating that therapies should focus on restoring, not inhibiting, PKC activity.

INTRODUCTION

The protein kinase C (PKC) family has been intensely investigated in the context of cancer since the discovery that it is a receptor for the tumor-promoting phorbol esters (Castagna et al., 1982). This led to the dogma that activation of PKC by phorbol esters promotes carcinogen-induced tumorigenesis (Griner and Kazanietz, 2007), yet targeting PKC in cancer has been unsuccessful.

The PKC family contains nine genes that have many targets and thus diverse cellular functions, including cell survival, proliferation, apoptosis, and migration (Dempsey et al., 2000). PKC isozymes comprise three classes: conventional (cPKC: α, β, γ), novel (nPKC: δ, ε, η, θ), and atypical (aPKC: ζ, ι). cPKC and nPKC isozymes are constitutively phosphorylated at three priming sites (activation loop, turn motif, and hydrophobic motif) to structure PKC for catalysis (Newton, 2003). A pseudosubstrate segment maintains PKC in an autoinhibited conformation that is relieved by second-messenger binding. cPKC isozymes are activated by binding to diacylglycerol (DAG) and Ca^{2+}, whereas nPKC isozymes are activated solely by DAG, events that engage PKC at membranes. Thus, these PKC isozymes have two prerequisites for activation: constitutive processing phosphorylations and second-messenger-dependent relocalization to membranes. Prolonged activation of cPKC and nPKC isozymes with phorbol esters leads to their dephosphorylation and subsequent degradation, a process referred to as downregulation (Hansra et al., 1996; Young et al., 1987). aPKC isozymes bind neither Ca^{2+} nor DAG.

PKC has proved an intractable target in cancer therapeutics (Kang, 2014). PKCι was proposed to be an oncogene in lung and ovarian cancers (Justilien et al., 2014; Regala et al., 2005; Zhang et al., 2006), and PKCε was categorized as an oncogene because of its ability to transform cells (Cacace et al., 1993). However, for most PKC isozymes, there is conflicting evidence as to whether they act as oncogenes or as tumor suppressors. For example, PKCδ is considered a tumor suppressor because of its pro-apoptotic effects (Reyland, 2007). However, it promotes tumor progression of lung and pancreatic cancers in certain contexts (Mauro et al., 2010; Symonds et al., 2011). Similarly, both overexpression and loss of PKCζ in colon cancer cells have been reported to decrease tumorigenicity in nude mice or cell lines, respectively (Luna-Ulloa et al., 2011; Ma et al., 2013). Likewise, PKCα was reported to both induce (Walsh et al., 2004; Wu et al., 2013) and suppress colon cancer cell proliferation (Gwak et al., 2009) and to suppress colon tumor formation in the $APC^{Min/+}$ model (Oster and Leitges, 2006). Based on the dogma that PKC isozymes contribute positively to cancer progression, many PKC inhibitors have entered clinical trials; however, they have been ineffective (Mackay and Twelves, 2007).

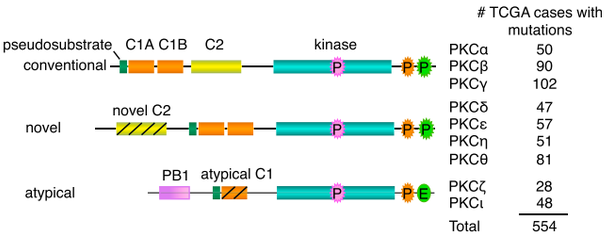


Figure 1. A Multitude of Cancer-Associated Mutations Have Been Identified within the Nine PKC Genes

(Left) Domain structure of conventional (α, β, γ), novel (δ, ε, η, θ), and atypical (ζ, ι) PKC members showing priming phosphorylation sites: activation loop (pink), turn motif (orange), and hydrophobic motif (green). (Right) Number of TCGA cases with cancer-associated mutations (missense, nonsense, insertions, deletions, splice site, or translation start site) identified within each of the PKC genes.

In fact, a recent meta-analysis of controlled trials of PKC inhibitors combined with chemotherapy versus chemotherapy alone revealed that PKC inhibitors significantly decreased response rates and disease control rates in non-small cell lung cancer (Zhang et al., 2014). Why has inhibiting PKC failed in the clinic? It has been well established that prolonged or repetitive treatment with phorbol esters depletes cPKC and nPKC isozymes from cells (Blumberg, 1980; Nelson and Alkon, 2009), bringing into question whether loss of PKC, rather than its activation, promotes tumorigenesis.

PKC is frequently mutated in human cancers. To uncover whether loss or gain of PKC function contributes to cancer progression, we selected mutations throughout the primary sequence and family membership and assessed their functional impact. Specifically, we asked how these cancer-associated mutations alter the signaling output of PKC using our genetically encoded reporter, C kinase activity reporter (CKAR) (Violin et al., 2003). Characterization of 46 of these mutations revealed that most reduced or abolished PKC activity and none were activating. Bioinformatic analysis of all PKC mutations revealed that they may cooperate with co-occurring mutations in oncogenes and tumor suppressors known to be regulated by PKC. Correction of one patient-identified, heterozygous, loss-of-function (LOF) PKCβ mutation in a colon cancer cell line significantly decreased tumor size in mouse xenografts, indicating that loss of PKC function enhances tumor growth. Our data are consistent with PKC isozymes functioning generally as tumor suppressors, reversing the paradigm that their hyperactivation promotes tumor growth.

RESULTS

A Multitude of Cancer-Associated Mutations Have Been Identified within the Nine PKC Genes

554 mutations (as of October 2014), of which most are heterozygous, have been identified in diverse cancers (Cerami et al., 2012; Gao et al., 2013) within cPKC (242), nPKC (236), and aPKC (76) isozymes (Figure 1). These mutations reside throughout the entire coding region, with no apparent mutational hotspots. Therefore, we conducted a comprehensive study of mutations within PKC domains and within interdomain regions to determine how they affect PKC signaling to contribute to cancer pathogenesis. 46 mutations of both conserved and non-conserved residues were selected from all three classes of PKC isozymes (Table 1 and Table S1).

PKC Mutations in the Regulatory C1 and C2 Domains Are LOF

The C1 domains of cPKC and nPKC isozymes are critical for their activation because they mediate PKC translocation to membranes via binding to DAG. Thus, we investigated how C1 domain mutations alter PKC translocation and activation. To measure agonist-dependent PKC activity, COS7 cells co-expressing the FRET-based PKC reporter (CKAR) and equal levels of either wild-type (WT) or mutant mCherry-tagged PKC were stimulated with the cell-permeable DAG, DiC8, or the phorbol ester, phorbol 12,13-dibutyrate (PDBu), and phosphorylation-dependent FRET ratio changes were recorded. Phorbol esters serve as an effective although non-physiological tool to maximally activate PKC because they bind with 100-fold higher affinity to C1 domains compared to DAG (Mosior and Newton, 1998). A mutation identified in a colorectal cancer tumor altered a residue (PKCα H75Q) required for coordination of Zn^{2+} and thus for folding of the C1 domain (Figure 2A). This mutation ablated agonist-stimulated activity, as evidenced by a lower FRET ratio trace compared with that of cells containing only endogenous PKC (Figure 2B). This lower activity suggests that the mutant is dominant negative toward global PKC output. Within a head and neck cancer patient, a mutation altered a critical residue (PKCα W58L) required for controlling the affinity for DAG, but not phorbol ester (Dries et al., 2007) (Figure 2A). This mutation also abolished DiC8-induced and basal activity but retained some PDBu-induced activity, consistent with this residue selectively regulating DAG affinity (Figures 2B and S1A). Because membrane translocation is a prerequisite for activation of cPKC isozymes, we compared the translocation of YFP-tagged WT and mutant PKC to membrane-targeted CFP using FRET (Antal et al., 2014). Mutation of either residue impaired translocation upon stimulation with DiC8, phorbol ester (Figure 2C), or the natural agonist UTP (Figure 2D), accounting for the inability of these agonists to activate the mutants. Lastly, we asked how these mutations affected the processing phosphorylations of PKC. PKCα H75Q, but not W58L, was unphosphorylated, likely because the misfolded C1A domain of the H75Q mutant prevented its processing (Figure 2E). Three additional mutations within the C1A domains of PKCα (G61W), PKCβ (G61W), and PKCγ (Q62H) also exhibited reduced agonist-induced PKC activity (Figures S1B–S1D). Our analysis of nine C1 domain mutations revealed that five reduced or abolished activity while none were hyperactivating (Tables 1 and S1). Inactivation occurred by altering two key inputs required for PKC function: disruption of binding to DAG or processing by phosphorylations.

The C2 domain of cPKC isozymes is also critical for activation, as it mediates Ca^{2+}-dependent pre-targeting to plasma membrane, where these isozymes bind DAG and become activated (Newton, 2003). One mutation identified within the C2 domain of PKCγ (D193N) was present in colorectal and ovarian cancers

Table 1. Loss-of-Function PKC Mutations in Cancer

Mutation[a]	Activity	Domain	Cancer(s)	Residue Importance	Allele Frequency	Other Mutations[b]
γ G23E	none[c]	PS	colorectal	adding negative charge to pseudosubstrate	N/A	γ G23W δ G146R ι G128C
ε R162H	low		head and neck	non-conserved	0.15	
α W58L	none[c]	C1A	head and neck	DAG binding; conserved in all C1a domains	0.22	γ W57splice θ W171*
α G61W	low		lung	conserved in cPKC C1a domains	0.05	β G61W
β G61W	low		lung	conserved in cPKC C1a domains	0.06	α G61W
γ Q62H	none[c]		lung	conserved in all PKC isozymes	0.45	α Q63H ε Q197P
α H75Q	none[d]		colorectal	coordinates Zn^{2+}; conserved in all C1 domains	N/A	η H284Y ι H179Y
γ D193N	none[c]	C2	colorectal/melanoma/ovarian	Ca^{2+} binding site	0.28	
γ T218M	none[c]		stomach	non-conserved	0.42	γ T218R
γ D254N	low		endometrial/ovarian	Ca^{2+} binding site	0.43	
α G257V	none[c]		lung	conserved in cPKC isozymes	0.12	
γ F362L	none[c]	Kinase	endometrial	conserved in cPKC and nPKC isozymes	0.21	γ F362fs β F353L
β Y417H	none[c]		liver	conserved in cPKC isozymes	0.67	γ Y431F
ζ E421K	none[d]		breast	APE motif; conserved in most protein kinases	N/A	α E508K ι E423D
α F435C	none[c]		endometrial	conserved in cPKC and nPKC isozymes	0.31	
α A444V	low		endometrial/breast	conserved in cPKC and nPKC isozymes	0.27	β A447T γ A461T γ A461V δ A454V θ A485T ι S359C
γ G450C	none[c]		endometrial/lung/liver	conserved in cPKC isozymes	0.41	ε R502*
α D481E	low		colorectal	DFG motif; conserved in most protein kinases	N/A	β D484N γ D498N ι D396E
β A509V	none[d]		breast	APE motif; conserved in most protein kinases	N/A	α A506V α A506T β A509T
β A509T	none[c]		colorectal	APE motif; conserved in most protein kinases	0.53	α A506V α A506T β A509V
γ P524R	none[d]		pancreatic	APE motif; conserved in most protein kinases	N/A	γ P524L δ P517S ε P576S θ P548S
δ D530G	none[d]		colorectal	anchors the conserved regulatory spine; conserved in all eukaryotic kinases	N/A	β D523N γ D537G γ D537Y
δ P568A	none[c]		head and neck	conserved in all PKC isozymes	0.16	δ P568S β P561H γ P575H
β G585S	low		lung	conserved in all PKC isozymes	N/A	η G598V
η K591E	low		breast	reversal of conserved charge	N/A	η K591N θ R616Q
η R596H	none[d]		colorectal	conserved in all PKC isozymes	0.50	
η G598V	none[d]		lung	conserved in all PKC isozymes	N/A	β G585S

(*Continued on next page*)

Table 1. *Continued*

Mutation[a]	Activity	Domain	Cancer(s)	Residue Importance	Allele Frequency	Other Mutations[b]
β P619Q	none[d]	C-tail	endometrial	PXXP motif; conserved in AGC kinases	0.48	

PKC mutations showing no activity with any agonist, no activity with physiological stimuli, or reduced activity in response to physiological stimuli. Allele frequencies were obtained from cBioPortal.

[a]Mutations examined in this study.

[b]Other mutations present at the same/corresponding residue in the same/other PKC isozymes.

[c]Kinase-dead.

[d]No response to physiological stimuli.

and in melanoma. Another (D254N) was found in endometrial and ovarian cancers. Because both of these Asp residues (Figure 2F) coordinate Ca^{2+} (Medkova and Cho, 1998), we monitored their activation upon elevation of intracellular Ca^{2+} with thapsigargin, a sarco/endoplasmic reticulum Ca^{2+}-ATPase inhibitor (Rogers et al., 1995). In contrast to WT PKCγ, neither mutant was activated (Figure 2G) nor translocated to the plasma membrane (Figure 2H) following thapsigargin addition, consistent with impaired Ca^{2+} binding. However, both mutants retained full responses to phorbol esters, consistent with unimpaired C1 domains. To further substantiate the inability of the mutants to bind Ca^{2+}, we monitored PKC oscillatory translocation stimulated by histamine-induced oscillatory Ca^{2+} release in HeLa cells (Violin et al., 2003). Whereas WT PKCγ exhibited oscillatory translocation in some cells, the C2 domain mutants were unresponsive to histamine (Figure 2I). Thus, these C2 domain mutations dampen PKCγ activity because they impede Ca^{2+} binding. Mutation of two other C2 domain residues that are not directly involved in Ca^{2+} binding (PKCγ T218M and PKCα G257V) also caused LOF (Figure S1D and S1E); PKCα G257V was LOF because it was not processed by phosphorylation (Figure S1F), whereas the remaining C2 domain mutants were (data not shown). Our analysis of six C2 domain mutations revealed four LOF mutations and no hyperactivating ones (Tables 1 and S1).

PKC Mutations in the Kinase Domain Are LOF

We next evaluated 21 kinase domain mutations, two of which were within PKCδ: D530G in colorectal cancer and P568A in head and neck cancer (Figure 3A). Asp530 functions as an anchor for the kinase regulatory spine, a highly conserved structural element of eukaryotic kinases (Kornev et al., 2006; Kornev et al., 2008); not surprisingly, the D530G mutant was kinase dead and not primed by phosphorylation (Figures 3B and 3C). Mutation of the conserved Pro568 to Ala also prevented a response to natural agonist stimulation but maintained some PDBu-stimulated activity, likely because a small pool of this mutant was phosphorylated (Figures 3B and 3C).

Strikingly, all three PKCη mutations examined (K591E, R596H, and G598V) altered its subcellular localization by pre-localizing it at the plasma membrane prior to stimulation (Figure 3D). However, despite constitutive membrane association, these mutants had reduced basal and stimulated activity as read out by a phospho-(Ser) PKC substrate antibody (Figure 3E) because they were not processed by phosphorylation (Figure 3F). We have previously shown that unprocessed nPKC isozymes have exposed C1 domains that induce constitutive membrane association (Antal et al., 2014).

A number of mutations were present within the highly conserved APE motif that is involved in substrate binding and allosteric activation of kinases (Kornev et al., 2008). PKCγ P524R and PKCβ A509V mutations ablated activity by preventing processing phosphorylations, and both exhibited dominant-negative roles (Figures 3G–3J). PKCβ A509T (colorectal cancer) also showed loss of function in response to UTP but was modestly activated by the potent ligand PDBu (Figure 3I), likely because a small pool of it was phosphorylated (Figure 3J). A LOF mutation that prevented processing of the atypical PKCζ was also found within the APE motif (E421K; Figure S1G).

Further analysis revealed that 16 out of 21 kinase domain mutations that we analyzed (Tables 1 and S1) resulted in full or partial LOF, with the majority preventing processing by phosphorylation. For example, PKCα F435C, PKCα A444V, PKCβII Y417H, PKCβII G585S, and PKCγ G450C had impaired phosphorylation and reduced activity (Figures S1C–S1F and S1H–S1J). However, partial LOF mutations were also observed in cases in which phosphorylation was maintained—PKCα D481E (Figures S1B and S1F) and PKCγ F362L (Figures S1D and S1J), suggesting that these mutations likely decrease PKC's intrinsic catalytic activity.

The Majority of Cancer-Associated PKC Mutations Are LOF

Our analysis of 46 mutations present within eight of the PKC genes revealed that ~61% (28) of them were LOF and none were activating (Figure 4A). A lack of identification of activating mutations is not an artifact of our assays, as activating PKC mutations that increase PKC affinity for DAG or decrease autoinhibition are readily detectable (data not shown). LOF mutations were identified within cPKC (α, β, γ), nPKC (δ, ε, η), and aPKC (ζ) isozymes and occurred within the C1, C2, and kinase domains as well as the pseudosubstrate and C-terminal tail (Figure 4B). For example, the PKCγ G23E pseudosubstrate mutation was not processed by phosphorylation (Figure S1J) and thus lacked any UTP-stimulated activity (Figure S1D), and the PKCε R162H pseudosubstrate mutation showed reduced agonist-stimulated and basal activity (Figures S1K and S1L). The PKCβ P619Q C-terminal tail mutation, residing within a conserved PXXP motif required for processing (Gould et al., 2009), was also LOF as it prevented PKC phosphorylation (Figure S1H). Overall, PKC LOF occurred by diverse mechanisms, most commonly by preventing processing phosphorylations or ligand binding, and as such, there were no mutational hotspots for loss of function. However, we identified seven LOF mutation "warmspots" (Sun et al., 2007) that fell within highly conserved regions of PKC—one within the pseudosubstrate and six within the kinase domain (Figure 4C). Thus,

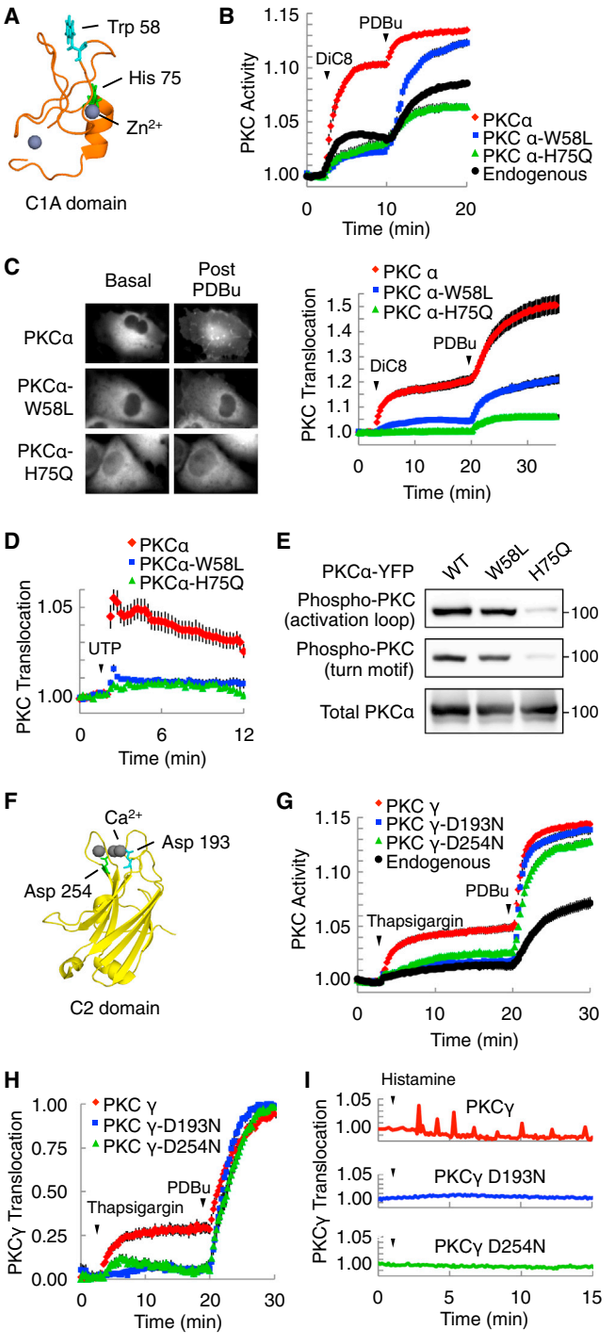


Figure 2. PKC Mutations in the Regulatory C1 and C2 Domains Are LOF

(A) Solution structure of the C1A domain of PKCγ (PDB 2E73) showing the corresponding PKCα His75 residue that coordinates Zn^{2+} and PKCα Trp58.

(B) Normalized FRET ratio changes (mean ±SEM) representing DiC8- (10 μM) followed by PDBu- (200 nM) induced PKC activity as read out by CKAR in COS7 cells co-expressing CKAR and either mCherry-tagged WT, mutant PKCα, or no exogenous PKC (endogenous).

(C) (Left) Representative YFP images of the indicated PKC isozymes under basal and PDBu-treated conditions (200 nM; 15 min) showing relocalization of WT, but not mutant PKCα, to membranes. (Right) Normalized FRET ratio changes (mean ±SEM) quantifying translocation of YFP-tagged PKCα proteins toward a membrane-targeted CFP upon stimulation with 10 μM DiC8, followed by 200 nM PDBu.

(D) Normalized FRET ratio changes (mean ±SEM) showing PKC translocation following UTP (100 μM) stimulation.

(E) Immunoblot showing the phosphorylation state of the indicated YFP-tagged PKCα proteins.

(F) Crystal structure of the C2 domain of PKCγ (PDB 2UZP) highlighting Asp193 and Asp254 residues involved in Ca^{2+} binding.

(G) Normalized FRET ratio changes (mean ±SEM) showing PKC activity as read out by CKAR upon elevation of intracellular Ca^{2+} stimulated by thapsigargin (5 μM), followed by PDBu (200 nM).

(H) Normalized FRET ratio changes (mean ±SEM) showing translocation of YFP-tagged PKCγ constructs toward membrane-localized CFP upon stimulation of COS7 cells with thapsigargin (5 μM) followed by PDBu (200 nM). Data were normalized to the maximal amplitude of translocation for each cell and then scaled from 0 to 1 using the equation: X = (Y − Ymin)/(Ymax − Ymin), where Y = normalized FRET ratio, Ymin = minimum value of Y, and Ymax is maximum value of Y.

(I) Normalized FRET ratio changes displaying oscillatory translocation of YFP-tagged WT PKCγ, but not PKCγ mutants D193N and D254N, in HeLa cells co-expressing membrane-targeted CFP and stimulated with 10 μM histamine. Data are representative traces from individual cells of three independent experiments.

See also Figure S1.

inactivating mutations targeted conserved regulatory elements and frequently hit the same residue, whereas mutations that exhibited no difference from WT occurred more randomly (Table S1).

Analysis of cancer types most frequently harboring PKC mutations revealed that, although PKC isozymes are mutated across many cancers, PKC mutations are enriched in certain cancers (Figure 4D). Namely, PKC isozymes are mutated in 20%–25% of melanomas, colorectal cancers, or lung squamous cell carcinomas but are mutated in <5% of ovarian cancers, glioblastoma, or breast cancers (Cerami et al., 2012; Gao et al., 2013). Additionally, nPKC isozymes are most commonly mutated in gastrointestinal cancers (pancreatic, stomach, and colorectal), which have a lower mutation burden than melanomas and lung cancers, highlighting their importance in this type of cancer (Figure 4D). The majority of PKC mutations are heterozygous, with an allele frequency varying from 0.05 to 0.67 for the mutations characterized (Tables 1 and S1). This indicates that PKC mutations can be truncal events in regards to tumor heterogeneity and exist in a majority of the cells within a tumor or can be branchal events acquired later in tumorigenesis as the tumor progresses to a more aggressive stage. This is consistent with PKC mutations being co-driver events that enhance tumorigenesis mediated by primary drivers.

Dominant-Negative PKCβ Mutation Confers a Tumor Growth Advantage

Because the majority of PKC mutations examined were LOF, we tested whether we could rescue HCT116 colon cancer cells that have a heterozygous LOF frameshift mutation in the C2 domain of PKCβ by overexpressing WT PKCβII. This resulted in a dramatic reduction in anchorage-independent growth (Figure S2A), a hallmark of cellular transformation. Thus, we next used CRISPR/Cas9-mediated genome editing to ask whether

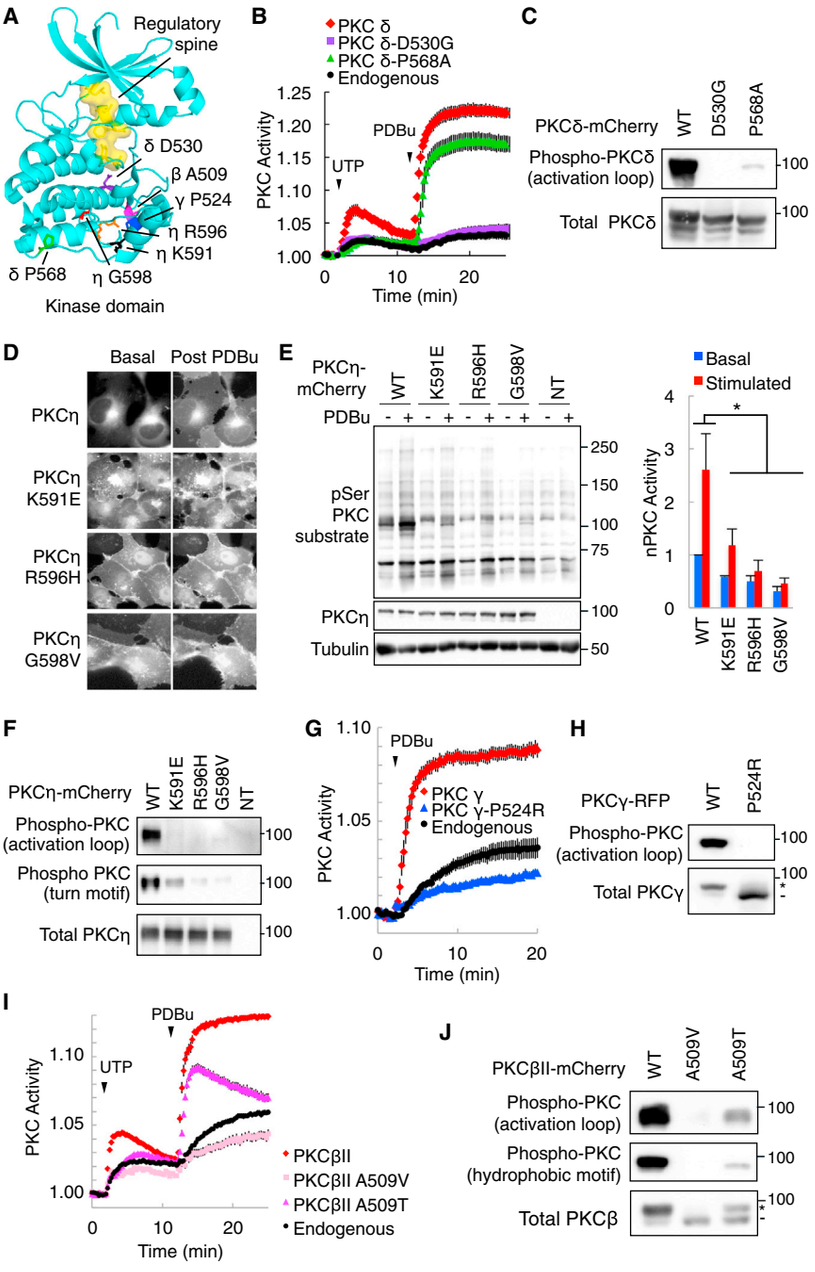


Figure 3. PKC Mutations in the Kinase Domain Are LOF

(A) Crystal structure of the kinase domain of PKCβII (PDB 2I0E) highlighting cancer-associated residues and the regulatory spine (yellow space filling).

(B) Normalized FRET ratio changes (mean ±SEM) showing PKC activity of PKCδ constructs in COS7 cells co-expressing the plasma membrane-targeted, PKCδ-specific reporter PM-δCKAR. Cells were stimulated with UTP (100 μM) followed by PDBu (200 nM).

(C) Immunoblot analysis of the phosphorylation state of PKCδ WT and mutants.

(D) Representative mCherry images of mCherry-tagged PKCη WT or mutants showing localization under basal conditions and 15 min post 200 nM PDBu addition to COS7 cells.

(E) (Left) Immunoblot showing PKC substrate phosphorylation. COS7 cells overexpressing the indicated constructs were pre-treated with 4 μM Gö6976 for 10 min to inhibit cPKC isozymes and were then stimulated or not with 200 nM PDBu to activate nPKC isozymes. (Right) Immunoblots were quantified and normalized to total PKCη levels and tubulin. Data represent averages of three independent experiments ±SEM. Comparisons for basal and stimulated activity were made using a repeated-measures one-way ANOVA followed by post hoc Dunnett's multiple comparison test. *$p < 0.05$ as compared with the WT group.

(F) Immunoblot analysis of the phosphorylation state of mCherry-tagged PKCη WT and mutants.

(G) Normalized FRET ratio changes (mean ± SEM) showing PKC activity from COS7 cells co-expressing CKAR and RFP-tagged PKCγ mutants stimulated with 200 nM PDBu.

(H) Immunoblot depicting PKCγ WT and P524R phosphorylation. The asterisk denotes phosphorylated and the dash unphosphorylated PKCγ.

(I) Normalized FRET ratio changes (mean ±SEM) showing PKC activity of PKCβII constructs in COS7 cells co-expressing CKAR. Cells were stimulated with UTP (100 μM) followed by PDBu (200 nM).

(J) Immunoblot depicting mCherry-tagged PKCβII WT and mutant phosphorylation. The asterisk denotes phosphorylated and the dash unphosphorylated PKCβII.

See also Figure S1.

reverting an endogenous LOF allele to WT would also rescue cell growth. We used DLD1 colon cancer cells because they harbor a PKCβ A509T LOF mutation (Figure 3I) to assess whether a heterozygous LOF PKC mutation could confer a survival advantage, as most cancer-associated PKC mutations are heterozygous. We reverted the mutation to WT in three isogenic clones (Figures S2B and S2C) and confirmed that no sequence alterations existed within the top two most likely predicted off-targets (data not shown). Correction of the A509T mutation in the endogenous PKCβ (*PRKCB*) allele caused a slight but reproducible increase in the PKCβ levels and a >2-fold increase in PKCα levels, although neither reached statistical significance (Figure 5A). Immunoblot analysis with a phospho-(Ser) PKC substrate antibody revealed significantly higher basal PKC activity in the corrected cells (Figure 5B). This is consistent with the DLD1 parental cells having reduced PKC activity because of the LOF PKCβ mutation and the lower PKCα levels. We next tested the ability of these cells to grow in suspension. Consistent with having higher PKC activity and a more tumor-suppressive phenotype, the corrected cells were less viable in suspension (Figure 5C) because they were less capable of forming the compact multicellular aggregates formed by the DLD1 parental cells (Figure 5D). Moreover,

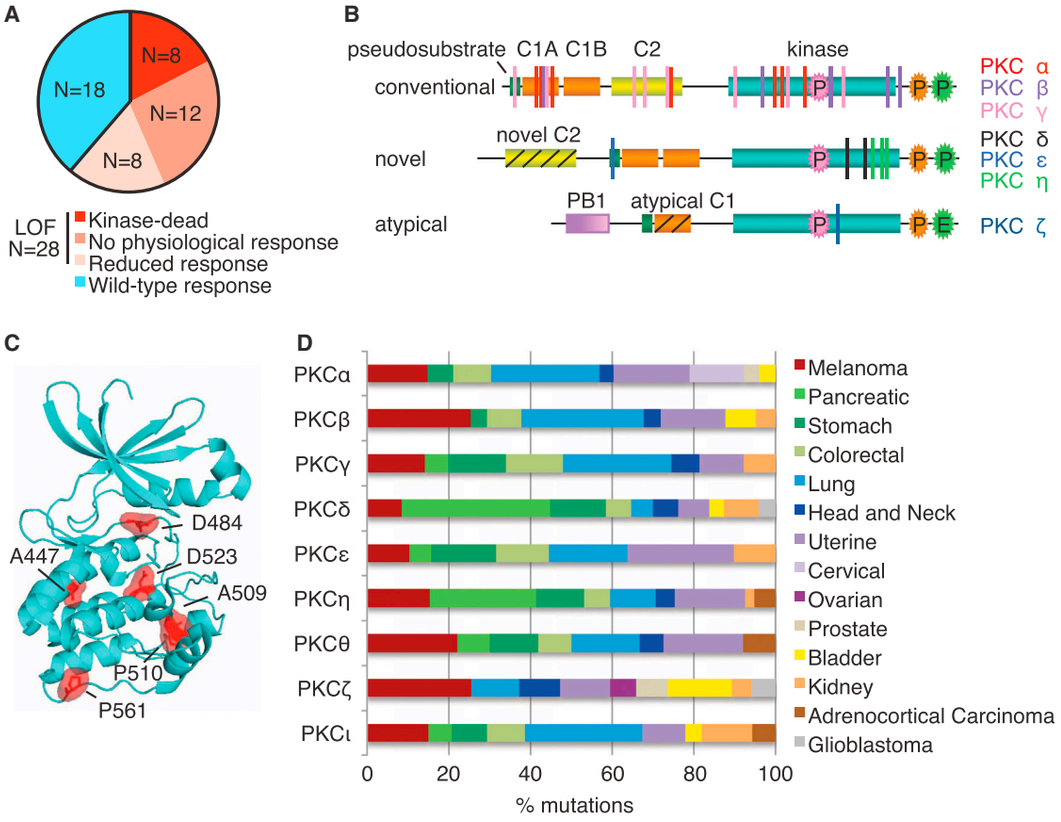


Figure 4. The Majority of PKC Mutations Are LOF

(A) Pie chart of the functional impact of the investigated PKC mutations, with bright red representing mutations that lack any activity, medium red representing mutations that show no response to physiological stimuli (DAG or Ca^{2+} elevation) but some response to non-physiological phorbol esters, light red representing mutations that display reduced activity to physiological stimuli compared to the corresponding WT isozyme, and blue representing no difference from the corresponding WT PKC isozyme.

(B) Domain structure of cPKC, nPKC, and aPKC isozymes, overlaid with the LOF mutations color coded by isozyme.

(C) Crystal structure of the kinase domain of PKCβII (PDB 2I0E) highlighting "warmspot" residues mutated in at least four tumor samples within the various PKC isozymes.

(D) Bar graph depicting the percentage of mutations distributed in the indicated cancers for each PKC isozyme.

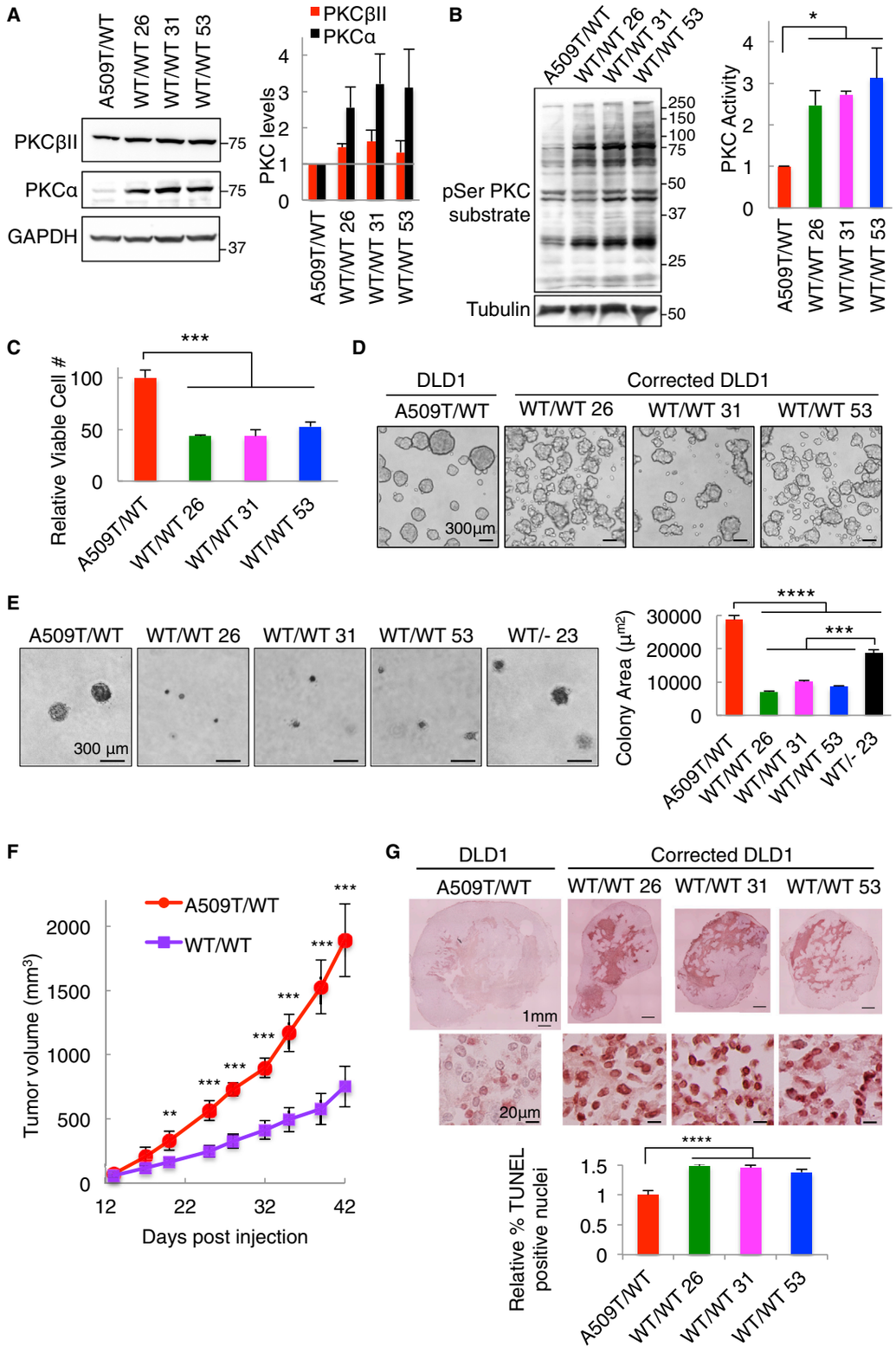
the corrected clones had decreased anchorage-independent growth potential (Figure 5E). These results corroborate those obtained from the HCT116 cells overexpressing PKCβII, demonstrating that partial loss of PKCβ activity is necessary for growth in soft agar. However, in a 2D proliferation assay, the DLD1-corrected cells proliferated at similar rates to the DLD1 parental cells (Figure S2D), indicating that it is not the proliferation rates that differ between these cells but, rather, their ability to grow in the absence of anchorage.

To determine whether PKC displays haploinsufficiency, we knocked out the mutant PKCβ allele in DLD1 cells by creating a frameshift deletion using genome engineering (Figure S2E). This hemizygous clone (WT/- 23), containing only one WT allele and thus expressing lower PKCβII levels (Figure S2F), exhibited significantly increased anchorage-independent growth potential compared to cells containing two WT alleles, indicating that PKCβII is haploinsufficient for tumor suppression (Figure 5E). Additionally, the PKCβ hemizygous cells did not grow as well as the PKCβ A509T mutated cells in soft agar, indicating that this mutation had a dominant-negative effect.

To definitively establish whether a heterozygous LOF PKCβ mutation facilitates tumor growth in vivo, the DLD1 parental or corrected cells were subcutaneously injected into the flanks of nude mice and tumor growth was monitored. Consistent with our cellular data, the tumors derived from the corrected cells were significantly smaller than those from the DLD1 parental cells (Figures 5F and S2G). This reduced growth correlated with increased apoptosis as assessed by TUNEL staining of tumor sections (Figure 5G). These data demonstrate that a heterozygous, dominant-negative PKCβ mutation can significantly increase tumor growth, thus establishing PKCβ as a tumor suppressor.

DISCUSSION

Here we establish that clinical trials targeting PKC have been based on the wrong assumption; it is not inactivation of PKC



(legend on next page)

but, rather, activation that suppresses tumor growth. Thus, we propose that therapies should target mechanisms to restore the PKC signaling output rather than reduce it. Our comprehensive analysis revealed that 61% of the PKC mutations characterized were LOF and none were activating. We did not account for nonsense mutations or deletions, so an even higher proportion of PKC mutations are LOF. Corroborating our data, three other LOF PKC mutations have been previously described. A LOF PKCα mutation (D294G in C2 domain) was identified in three types of cancer (Alvaro et al., 1993; Prévostel et al., 1997; Zhu et al., 2005) and a LOF PKCζ mutation (S514F in the kinase domain) was identified in colorectal cancer (Galvez et al., 2009). A partial LOF mutation in PKCι (R471C), present in three distinct cancers, disrupted substrate binding and induced abnormal epithelial polarity (Linch et al., 2013). To our knowledge, no gain-of-function PKC mutations have been observed in cancer. The identification of LOF mutations throughout the PKC family and in diverse cancers supports a general role for PKC isozymes as tumor suppressors.

Strikingly, several LOF PKC mutations (e.g., PKCβ A509V, PKCγ P524R, and PKCα W58L, H75Q, and G257V) acted in a dominant-negative manner by decreasing global endogenous PKC activity. Moreover, the presence of mutant PKCβ A509T protein in DLD1 cells reduced PKCα levels. One mechanism for this cross-PKC dominant-negative effect is that the LOF PKC impairs the priming phosphorylations of other PKCs, thus reducing their steady-state levels. This is supported by a prior study demonstrating that unprocessed kinase-dead PKC isozymes prevent the phosphorylation of other PKC isozymes, likely because their phosphorylation requires common titratable components (Garcia-Paramio et al., 1998). This dominant-negative role of LOF mutations is corroborated by studies showing that kinase-dead PKC isozymes function in a dominant-negative manner to exhibit tumorigenic effects on cells (Galvez et al., 2009; Hirai et al., 1994; Kim et al., 2013; Lu et al., 1997). Importantly, although some PKC mutations were dominant negative, loss of PKC such as would occur from nonsense mutations or gene deletions also conferred a growth advantage (Figure 5E), indicating that PKC is haploinsufficient for tumor suppression.

A tumor-suppressive role of PKC is supported by PKC gene knockout mouse models and cellular studies. PKCα-deficient (*Prkca*$^{-/-}$) mice developed spontaneous intestinal tumors (Oster and Leitges, 2006). In an APC$^{Min/+}$ background, loss of PKCα induced more aggressive tumors and decreased survival (Oster and Leitges, 2006), and in the context of oncogenic Kras, PKCα deletion increased lung tumor formation (Hill et al., 2014). Deletion of PKCζ in mice that are PTEN haploinsufficient resulted in larger, more invasive prostate tumors and enhanced intestinal tumorigenesis in an APC$^{Min/+}$ background (Ma et al., 2013). Knockdown of PKCδ in colon cancer cells increased tumor growth in nude mice (Hernández-Maqueda et al., 2013). Conversely, overexpression of PKC revealed a protective role. Re-expression of PKCβI in colon cancer cells (Choi et al., 1990) or of PKCδ in keratinocytes (D'Costa et al., 2006) or overexpression of PKCζ in colon cancer cells (Ma et al., 2013) or in Ras-transformed fibroblasts (Galvez et al., 2009) decreased tumorigenicity in nude mice.

Clinical data reveal lower PKC protein levels and activity in tumor tissue compared with cognate normal tissue, also supporting a tumor-suppressive role for PKC. Total PKC activity was significantly lower in human colorectal cancers versus normal mucosa because of decreased PKCβ and PKCδ (Craven and DeRubertis, 1994) or PKCβ and PKCε protein levels (Pongracz et al., 1995). PKCα protein was downregulated in 60% of human colorectal cancers (Suga et al., 1998), and PKCζ was downregulated in renal cell carcinoma (Pu et al., 2012) and non-small cell lung cancer (Galvez et al., 2009). Decreased PKCβ and PKCδ levels correlated with increased tumor grade in bladder cancer (Koren et al., 2000; Langzam et al., 2001; Varga et al., 2004), and decreased PKCδ levels correlated with increased grade in endometrial cancer and glioma (Reno et al., 2008; Mandil et al., 2001). PKCη was downregulated in colon and hepatocellular carcinomas, and lower PKCη expression was associated with poorer long-term survival (Davidson et al., 1994; Lu et al., 2009). However, increased PKCι protein and DNA copy number levels have been observed in certain cancers (Perry et al., 2014; Regala et al., 2005). PKCι is part of the 3q26 amplicon, and its increased DNA copy number levels correlate with increased mRNA expression (Figure S3). However, DNA copy number

Figure 5. Correction of a Heterozygous LOF PKCβ Mutation Reduces Growth in Soft Agar, Suspension, and a Xenograft Model

(A) Immunoblot (left) and quantification (right; mean ±SEM) of PKCβII, PKCα, and GAPDH levels in the DLD1 cells.

(B) Immunoblot (left) and quantification (right; mean ±SEM) of phospho-(Ser) PKC substrates. Comparisons were made using a repeated-measures one-way ANOVA followed by post hoc Dunnett's multiple comparison test. *$p < 0.05$ as compared with the DLD1 parental cells. Data represent the mean of three independent experiments ± SEM.

(C) Relative viable cell number (mean ±SEM) as assessed by a trypan blue exclusion assay after 72 hr in suspension from three independent experiments. Comparisons were made by using a one-way ANOVA followed by post hoc Dunnett's Multiple Comparison test. ***$p < 0.001$ as compared with the DLD1 parental cell group.

(D) Representative phase contrast images of DLD1 cells grown in suspension for 24 hr.

(E) (Left) Colony formation assay in soft agar. (Right) Quantification of colony area (mean ±SEM) for colonies with a diameter ≥50 μm from three to six independent experiments. Comparisons were made using a one-way ANOVA followed by post hoc Tukey's multiple comparison test. ****$p < 0.0001$ and ***$p < 0.001$ as compared with the DLD1 parental cell group.

(F) Tumor growth is presented as the mean tumor volume (mm^3) ±SEM, with the red representing data from mice injected with the DLD1 parental cells (A509T/WT; five mice) and purple representing data of the three corrected clones (17 mice total). Comparisons were made using a two-tailed, unpaired Student's t test for each time point. **$p < 0.005$ and ***$p < 0.0005$.

(G) (Top) Representative fields from TUNEL-stained slides of tumors derived from the DLD1 cells. (Bottom) Quantification of TUNEL-positive nuclei (mean ±SEM). Comparisons were made using a one-way ANOVA followed by post hoc Dunnett's Multiple Comparison test. ****$p < 0.0001$ as compared with the DLD1 parental cell group.

See also Figure S2.

and mRNA levels do not correlate for cPKC genes (Figure S3). In fact, for PKCα, copy number levels inversely correlate with protein levels in breast cancer (Myhre et al., 2013), the cancer in which PKCα is most amplified (Cerami et al., 2012; Gao et al., 2013). A number of studies reported increased mRNA expression of other PKC genes in cancer; however, mRNA expression and protein levels often poorly correlate (Myhre et al., 2013). Thus, clinical data of this sort are consistent with a tumor-suppressive function of PKC isozymes, although there might be context specific exceptions for PKCι.

The recent discovery that germline LOF mutations in PKCδ are causal drivers of autoimmune lymphoproliferative syndrome and systemic lupus erythematosus, disorders associated with the acquisition of cancer-associated phenotypes, supports a bona fide tumor-suppressive role of PKC in humans (Belot et al., 2013; Kuehn et al., 2013; Salzer et al., 2013). Both diseases are characterized by increased proliferation and decreased apoptosis of B cells (Belot et al., 2013; Kuehn et al., 2013), and patients frequently develop lymphomas (Bernatsky et al., 2005; Mellemkjaer et al., 1997). Moreover, we found that siblings homozygous for a LOF PKCδ mutation have reduced levels of PKCζ (data not shown), supporting a dominant-negative role of LOF mutations.

How could decreased PKC activity enhance tumorigenesis? One possibility is that PKC isozymes suppress oncogenic signaling by repressing signaling from oncogenes or stabilizing tumor suppressors. Supporting this, unbiased bioinformatic analysis of tumor samples harboring PKC LOF mutations revealed that *TP53* (p53) is one of most frequently mutated genes in tumors harboring LOF mutations for each PKC isozyme (Table 2). PKC might promote the tumor-suppressive function of p53 by stabilizing the WT protein. Considerable evidence suggests that phosphorylation by PKCδ stabilizes p53, thus promoting apoptosis (Abbas et al., 2004; Yoshida et al., 2006), but the role of other PKC isozymes is less clear. *KRAS* was also among the top ten genes mutated in cancers harboring PKC mutations for seven of the PKC isozymes (Table 2), specifically with mutation at Gly12 (Table S3). This argues that PKC might suppress Kras signaling, such that loss of PKC would be required for Kras to exert its full oncogenic potential. Consistent with this, PKC modulates both the activity and localization of Kras through phosphorylation of Ser181 (Bivona et al., 2006). Although the role of this phosphorylation site in tumors remains controversial (Barceló et al., 2014), our analysis is consistent with loss of PKC enhancing its oncogenic potential. In fact, the DLD1 and HCT116 cells used in our assays contained an oncogenic Kras mutation (G13D) that is necessary for the ability of these cells to grow in soft agar (data not shown). This suggests that LOF PKC mutations are not major cancer drivers but, rather, co-drivers that contribute to cancer progression.

We also analyzed which kinase or cancer census genes (genes implicated in cancer) are significantly more commonly mutated (>15-fold) in tumors harboring PKC mutations versus tumors lacking PKC mutations (Table S4). This allowed us to identify proteins that might be important co-drivers or represent novel genetic dependencies for PKC. The tumor suppressor LATS2, which inhibits the Hippo pathway, and the kinases ROCK1 and ROCK2, which are required for the anchorage independent growth and invasion of non-small cell lung cancer cells, were among the top 20 mutated proteins that were significantly enriched in tumors harboring PKC mutations (Table S4). Our analysis suggests that mutations in these genes provide a greater proliferative advantage upon loss of PKC signaling. We also performed an analysis of cancer-specific genes frequently co-mutated with PKC in lung cancer, colorectal cancer, or melanoma. This revealed very little overlap in co-mutated genes between the three cancers and also between the three classes of PKC isozymes (Table S5), suggesting that the individual PKC isozymes regulate distinct pathways in different cancers. Interestingly, cancers with a high PKC mutation burden, such as melanoma and colorectal cancers, show little PKC amplification. Conversely, cancers that have higher PKC amplification rates, such as breast and ovarian cancers, have few PKC mutations (Cerami et al., 2012; Gao et al., 2013), consistent with PKC mutations having a smaller or different role in breast and ovarian cancers.

The foregoing data provide a mechanism for why inhibiting PKC has proved unsuccessful and, in fact, detrimental in cancer clinical trials: it is not gain of function but, rather, LOF that confers a survival advantage. Therefore, therapeutic strategies should target ways to restore PKC activity. Bryostatin-1, a PKC agonist, also failed as a therapeutic and, in fact, exhibited counter-therapeutic effects in cervical cancer (Nezhat et al., 2004), likely because it downregulates PKC (Szallasi et al., 1994). Therefore, strategies to activate PKC without downregulating it hold significant clinical potential. An important ramification of this study is that drugs that inhibit proteins involved in the processing of PKC cause loss of PKC. Notably, both mTOR and HSP90 inhibitors, currently in use in the clinic (Don and Zheng, 2011; Neckers and Workman, 2012), prevent processing of PKC (Gould et al., 2009; Guertin et al., 2006) and would thus have the detrimental effect of removing its tumor suppressive function. Restoring PKC activity would have to accompany other chemotherapeutics, given that PKC isozymes act as the brakes, not the primary drivers, to oncogenic signaling. Our finding that decreased PKC activity enhances tumor growth challenges the concept of inhibiting PKC isozymes in cancer and underscores the need for therapies that restore or stabilize PKC activity in cells.

EXPERIMENTAL PROCEDURES

FRET Imaging and Analysis

Cells were imaged as described previously (Gallegos et al., 2006). For activity measurements, cells were co-transfected with the indicated mCherry-tagged PKC and CKAR or plasma membrane-targeted CKAR, as indicated. For translocation experiments, cells were co-transfected with the indicated YFP-tagged PKC and membrane-targeted CFP.

Generation of CRISPR Cell Lines

The CRISPR/Cas9 genome-editing system was employed to generate DLD1 cell lines in which the PKCβ A509T mutation was reverted to WT or knocked out. For the nuclease method, DLD1 cells were transiently transfected with the hSpCas9 vector containing the gRNA PKCβ-a, the PAGE-purified 70-mer ssODN (Figure S2B), and pMAX-GFP. For the double nickase method, DLD1 cells were transfected with two hSpCas9n vectors containing either gRNA PKCβ-a or PKCβ-b, the ssODN, and pMAX-GFP. GFP^+ cells were sorted 72 hr later. To reduce off-target mutagenesis, one of the clones (WT/WT 53) was made using a double-nicking approach that requires the

Table 2. Top 20 Genes with Mutations that Co-Occur with PKC Mutations

PKCα (50)	PKCβ (90)	PKCγ (102)	PKCδ (47)	PKCε (57)	PKCη (51)	PKCθ (81)	PKCι (48)	PKCζ (28)
BLID (7)	TP53 (42)	TP53 (52)	**KRAS** (13)	GNG4 (5)	SPINK7 (5)	TP53 (42)	SPRR2G (6)	TNP1 (3)
TP53 (23)	KRTAP6-2 (6)	CDKN2A (17)	TP53 (22)	**KRAS** (11)	RPL39 (3)	CDKN2A (13)	TP53 (26)	TP53 (15)
KRTAP19-5 (4)	PCP4 (4)	**KRAS** (16)	CDKN2A (9)	DEFB114 (4)	**KRAS** (11)	**KRAS** (14)	CDKN2A (10)	CNPY1 (3)
SPRR2E (4)	**KRAS** (12)	HTN1 (4)	CD52 (3)	CNPY1 (5)	DEFB114 (4)	SPANXN5 (5)	BANF1 (5)	SPATA8 (3)
REG3A (8)	OR4A15 (21)	SPRR2G (5)	CNPY1 (4)	SVIP (4)	PLN (3)	DEFB110 (4)	LACRT (7)	SPANXN3 (4)
H3F3C (6)	POM121L12 (18)	DEFB115 (6)	SPINK13 (4)	CXCL10 (5)	DEFB115 (5)	KRTAP15-1 (8)	CXCL9 (6)	KRTAP19-5 (2)
MLLT11 (4)	REG1A (10)	DNAJC5B (12)	ATP5E (2)	KRTAP19-3 (4)	LELP1 (5)	DEFB119 (5)	**KRAS** (9)	VPREB1 (4)
PI3 (5)	NRAS (11)	REG3G (10)	RPL39 (2)	COX7C (3)	DEFB116 (5)	PPIAL4G (9)	RETNLB (5)	GNG4 (2)
SNURF (3)	PLN (3)	SPATA8 (6)	COX7B2 (3)	KRTAP19-8 (3)	KRTAP19-8 (3)	DPPA5 (6)	WFDC10B (4)	ATP6V1G3 (3)
CDKN2A (7)	GNG4 (4)	REG1A (9)	OR4K1 (11)	SPINK7 (4)	IAPP (4)	CRYGB (9)	DEFB110 (3)	CDKN2A (4)
GNG3 (3)	CDKN2A (9)	POM121L12 (16)	FDCSP (3)	TP53 (18)	NPS (4)	SPANXN2 (9)	TMSB15B (2)	DEFB119 (2)
DAOA (6)	DEFA4 (5)	TRAT1 (10)	CARTPT (4)	BANF1 (4)	WFDC10B (4)	KRTAP19-3 (4)	GNG7 (3)	LGALS1 (3)
RPL39 (2)	OR2L13 (16)	HIST1H2AA (7)	DUSP22 (7)	TMSB15B (2)	S100A7L2 (5)	DYNLRB2 (6)	CNPY1 (4)	SCGB1D1 (2)
SVIP (3)	LCE1B (6)	SPINK13 (5)	BANF1 (3)	DEFA4 (4)	CNPY1 (4)	SPATA8 (5)	LSM8 (4)	NANOS2 (3)
PLN (2)	SPANXN3 (7)	CCK (6)	DYNLL2 (3)	POM121L12 (12)	TP53 (17)	KRTAP19-8 (3)	KRTAP19-5 (3)	CCL17 (2)
FAM19A2 (5)	KRTAP19-3 (4)	OR4K1 (16)	LYRM5 (3)	GYPA (6)	DPPA5 (5)	RIPPLY3 (9)	SPANXN5 (3)	NRAS (4)
CPLX4 (6)	TRAT1 (9)	OR4A5 (16)	ATP6V1G3 (4)	DYNLRB2 (5)	DEFB131 (3)	POM121L12 (14)	CSTL1 (6)	CCL1 (2)
SEC22B (8)	IFNB1 (9)	CCL7 (5)	DEFB128 (3)	HIST1H2BB (5)	SPINK13 (4)	OR4N2 (14)	DEFA4 (4)	PATE4 (2)
CTXN3 (3)	KRTAP19-8 (3)	B2M (6)	MAP1LC3B2 (4)	HIST1H2BI (5)	RPL10L (9)	DEFB115 (4)	SPANXD (4)	POM121L12 (6)
KRTAP19-3 (3)	KRTAP8-1 (3)	PCP4 (3)	GPX5 (7)	FGFR1OP2 (10)	SPRR2A (3)	OTOS (4)	EDDM3A (6)	CRIPT (2)

Data were normalized based on gene length, and the number of co-occurring cases is listed in parentheses. Two genes are highlighted: *TP53* is underlined, and *KRAS* is in bold.

cooperation between two nickase Cas9 enzymes (Ran et al., 2013). CRISPR-targeted clones were expanded and gDNA was extracted using a Quick-gDNA MiniPrep Kit (Zymo Research Corporation) and were screened for the presence of two wild-type alleles by PCR using primers spanning the A509 locus, followed by restriction digest with BtgZI. This restriction site was only present in the WT allele, and correction of the A509T mutation introduced this site into the other allele. The presence a WT allele at both loci was confirmed by Sanger sequencing (Eton Bioscience).

Xenograft Model

Athymic Nude-*Foxn1*^{nu} mice (Harlan) were housed in compliance with the University of California San Diego Institutional Animal Core and Use Committee. 3×10^6 DLD1 cells in 100 μl PBS were injected subcutaneously into the right flank of each 4-week-old female mouse. Tumor dimensions were recorded twice weekly and tumor volume was calculated as 1/2 × length × width^2. Mice were euthanized 43 days after injection, and tumors were excised. One tumor was excluded, as it did not engraft well (DLD1p), and another was excluded, as it was not subcutaneous (WT/WT 31).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, three figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.01.001.

AUTHOR CONTRIBUTIONS

L.L.G., J.B., T.H., and A.C.N. initiated the study. C.E.A. and A.C.N. conceived the experiments and wrote the manuscript. C.E.A. performed the experiments with assistance from E.K. for imaging and immunoblots and from C.Z. for the xenograft model. F.B.F. advised on the use of the xenograft model. A.M.H., C.W., C.J.M., and J.B. performed the bioinformatic analysis. N.L.S. and E.W.T. made the tetracycline-inducible PKCβII HCT116 cells.

ACKNOWLEDGMENTS

We thank the lab for helpful comments, the Moores Cancer Center Histology Core for the TUNEL staining, Meghdad Rahdar for cell sorting, and Jack Dixon for equipment use. This work was supported by NIH GM43154 to A.C.N., NIH NS080939 and the James S. McDonnell Foundation to F.B.F., and NIH CA82683 to T.H. C.E.A. was supported by the UCSD Graduate Training Program in Cellular and Molecular Pharmacology (T32 GM007752) and the NSF Graduate Research Fellowship (DGE1144086). A.M.H., C.W., N.L.S., E.W.T., C.J.M., and J.B. were supported by Cancer Research UK. T.H. is a Frank and Else Schilling American Cancer Society Professor and holds the Renato Dulbecco Chair in Cancer Research.

Received: August 21, 2014
Revised: November 12, 2014
Accepted: December 24, 2014
Published: January 22, 2015

REFERENCES

Abbas, T., White, D., Hui, L., Yoshida, K., Foster, D.A., and Bargonetti, J. (2004). Inhibition of human p53 basal transcription by down-regulation of protein kinase Cdelta. J. Biol. Chem. *279*, 9970–9977.

Alvaro, V., Lévy, L., Dubray, C., Roche, A., Peillon, F., Quérat, B., and Joubert, D. (1993). Invasive human pituitary tumors express a point-mutated alpha-protein kinase-C. J. Clin. Endocrinol. Metab. *77*, 1125–1129.

Antal, C.E., Violin, J.D., Kunkel, M.T., Skovsø, S., and Newton, A.C. (2014). Intramolecular conformational changes optimize protein kinase C signaling. Chem. Biol. *21*, 459–469.

Barceló, C., Paco, N., Morell, M., Alvarez-Moya, B., Bota-Rabassedas, N., Jaumot, M., Vilardell, F., Capella, G., and Agell, N. (2014). Phosphorylation at Ser-181 of oncogenic KRAS is required for tumor growth. Cancer Res. *74*, 1190–1199.

Belot, A., Kasher, P.R., Trotter, E.W., Foray, A.P., Debaud, A.L., Rice, G.I., Szynkiewicz, M., Zabot, M.T., Rouvet, I., Bhaskar, S.S., et al. (2013). Protein kinase cδ deficiency causes mendelian systemic lupus erythematosus with B cell-defective apoptosis and hyperproliferation. Arthritis Rheum. *65*, 2161–2171.

Bernatsky, S., Boivin, J.F., Joseph, L., Rajan, R., Zoma, A., Manzi, S., Ginzler, E., Urowitz, M., Gladman, D., Fortin, P.R., et al. (2005). An international cohort study of cancer in systemic lupus erythematosus. Arthritis Rheum. *52*, 1481–1490.

Bivona, T.G., Quatela, S.E., Bodemann, B.O., Ahearn, I.M., Soskis, M.J., Mor, A., Miura, J., Wiener, H.H., Wright, L., Saba, S.G., et al. (2006). PKC regulates a farnesyl-electrostatic switch on K-Ras that promotes its association with Bcl-XL on mitochondria and induces apoptosis. Mol. Cell *21*, 481–493.

Blumberg, P.M. (1980). In vitro studies on the mode of action of the phorbol esters, potent tumor promoters: part 1. Crit. Rev. Toxicol. *8*, 153–197.

Cacace, A.M., Guadagno, S.N., Krauss, R.S., Fabbro, D., and Weinstein, I.B. (1993). The epsilon isoform of protein kinase C is an oncogene when overexpressed in rat fibroblasts. Oncogene *8*, 2095–2104.

Castagna, M., Takai, Y., Kaibuchi, K., Sano, K., Kikkawa, U., and Nishizuka, Y. (1982). Direct activation of calcium-activated, phospholipid-dependent protein kinase by tumor-promoting phorbol esters. J. Biol. Chem. *257*, 7847–7851.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBiol cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Choi, P.M., Tchou-Wong, K.M., and Weinstein, I.B. (1990). Overexpression of protein kinase C in HT29 colon cancer cells causes growth inhibition and tumor suppression. Mol. Cell. Biol. *10*, 4650–4657.

Craven, P.A., and DeRubertis, F.R. (1994). Loss of protein kinase C delta isozyme immunoreactivity in human adenocarcinomas. Dig. Dis. Sci. *39*, 481–489.

D'Costa, A.M., Robinson, J.K., Maududi, T., Chaturvedi, V., Nickoloff, B.J., and Denning, M.F. (2006). The proapoptotic tumor suppressor protein kinase C-delta is lost in human squamous cell carcinomas. Oncogene *25*, 378–386.

Davidson, L.A., Jiang, Y.H., Derr, J.N., Aukema, H.M., Lupton, J.R., and Chapkin, R.S. (1994). Protein kinase C isoforms in human and rat colonic mucosa. Arch. Biochem. Biophys. *312*, 547–553.

Dempsey, E.C., Newton, A.C., Mochly-Rosen, D., Fields, A.P., Reyland, M.E., Insel, P.A., and Messing, R.O. (2000). Protein kinase C isozymes and the regulation of diverse cell responses. Am. J. Physiol. Lung Cell. Mol. Physiol. *279*, L429–L438.

Don, A.S., and Zheng, X.F. (2011). Recent clinical trials of mTOR-targeted cancer therapies. Rev. Recent Clin. Trials *6*, 24–35.

Dries, D.R., Gallegos, L.L., and Newton, A.C. (2007). A single residue in the C1 domain sensitizes novel protein kinase C isoforms to cellular diacylglycerol production. J. Biol. Chem. *282*, 826–830.

Gallegos, L.L., Kunkel, M.T., and Newton, A.C. (2006). Targeting protein kinase C activity reporter to discrete intracellular regions reveals spatiotemporal differences in agonist-dependent signaling. J. Biol. Chem. *281*, 30947–30956.

Galvez, A.S., Duran, A., Linares, J.F., Pathrose, P., Castilla, E.A., Abu-Baker, S., Leitges, M., Diaz-Meco, M.T., and Moscat, J. (2009). Protein kinase Czeta represses the interleukin-6 promoter and impairs tumorigenesis in vivo. Mol. Cell. Biol. *29*, 104–115.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Garcia-Paramio, P., Cabrerizo, Y., Bornancin, F., and Parker, P.J. (1998). The broad specificity of dominant inhibitory protein kinase C mutants infers a common step in phosphorylation. Biochem. J. *333*, 631–636.

Gould, C.M., Kannan, N., Taylor, S.S., and Newton, A.C. (2009). The chaperones Hsp90 and Cdc37 mediate the maturation and stabilization of protein kinase C through a conserved PXXP motif in the C-terminal tail. J. Biol. Chem. *284*, 4921–4935.

Griner, E.M., and Kazanietz, M.G. (2007). Protein kinase C and other diacylglycerol effectors in cancer. Nat. Rev. Cancer *7*, 281–294.

Guertin, D.A., Stevens, D.M., Thoreen, C.C., Burds, A.A., Kalaany, N.Y., Moffat, J., Brown, M., Fitzgerald, K.J., and Sabatini, D.M. (2006). Ablation in mice of the mTORC components raptor, rictor, or mLST8 reveals that mTORC2 is required for signaling to Akt-FOXO and PKCalpha, but not S6K1. Dev. Cell *11*, 859–871.

Gwak, J., Jung, S.J., Kang, D.I., Kim, E.Y., Kim, D.E., Chung, Y.H., Shin, J.G., and Oh, S. (2009). Stimulation of protein kinase C-alpha suppresses colon cancer cell proliferation by down-regulation of beta-catenin. J. Cell. Mol. Med. *13* (8B), 2171–2180.

Hansra, G., Bornancin, F., Whelan, R., Hemmings, B.A., and Parker, P.J. (1996). 12-O-Tetradecanoylphorbol-13-acetate-induced dephosphorylation of protein kinase Calpha correlates with the presence of a membrane-associated protein phosphatase 2A heterotrimer. J. Biol. Chem. *271*, 32785–32788.

Hernández-Maqueda, J.G., Luna-Ulloa, L.B., Santoyo-Ramos, P., Castañeda-Patlán, M.C., and Robles-Flores, M. (2013). Protein kinase C delta negatively modulates canonical Wnt pathway and cell proliferation in colon tumor cell lines. PLoS ONE *8*, e58540.

Hill, K.S., Erdogan, E., Khoor, A., Walsh, M.P., Leitges, M., Murray, N.R., and Fields, A.P. (2014). Protein kinase Cα suppresses Kras-mediated lung tumor formation through activation of a p38 MAPK-TGFβ signaling axis. Oncogene *33*, 2134–2144.

Hirai, S., Izumi, Y., Higa, K., Kaibuchi, K., Mizuno, K., Osada, S., Suzuki, K., and Ohno, S. (1994). Ras-dependent signal transduction is indispensable but not sufficient for the activation of AP1/Jun by PKC delta. EMBO J. *13*, 2331–2340.

Justilien, V., Walsh, M.P., Ali, S.A., Thompson, E.A., Murray, N.R., and Fields, A.P. (2014). The PRKCI and SOX2 oncogenes are coamplified and cooperate to activate Hedgehog signaling in lung squamous cell carcinoma. Cancer Cell *25*, 139–151.

Kang, J.-H. (2014). Protein kinase C (PKC) isozymes and cancer. New J. Sci. *2014*, 231418.

Kim, J.Y., Valencia, T., Abu-Baker, S., Linares, J., Lee, S.J., Yajima, T., Chen, J., Eroshkin, A., Castilla, E.A., Brill, L.M., et al. (2013). c-Myc phosphorylation by PKCζ represses prostate tumorigenesis. Proc. Natl. Acad. Sci. USA *110*, 6418–6423.

Koren, R., Langzam, L., Paz, A., Livne, P.M., Gal, R., and Sampson, S.R. (2000). Protein kinase C (PKC) isoenzymes immunohistochemistry in lymph node revealing solution-fixed, paraffin-embedded bladder tumors. Appl. Immunohistochem. Mol. Morphol. *8*, 166–171.

Kornev, A.P., Haste, N.M., Taylor, S.S., and Eyck, L.F. (2006). Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc. Natl. Acad. Sci. USA *103*, 17783–17788.

Kornev, A.P., Taylor, S.S., and Ten Eyck, L.F. (2008). A helix scaffold for the assembly of active protein kinases. Proc. Natl. Acad. Sci. USA *105*, 14377–14382.

Kuehn, H.S., Niemela, J.E., Rangel-Santos, A., Zhang, M., Pittaluga, S., Stoddard, J.L., Hussey, A.A., Evbuomwan, M.O., Priel, D.A., Kuhns, D.B., et al. (2013). Loss-of-function of the protein kinase C δ (PKCδ) causes a B-cell lymphoproliferative syndrome in humans. Blood *121*, 3117–3125.

Langzam, L., Koren, R., Gal, R., Kugel, V., Paz, A., Farkas, A., and Sampson, S.R. (2001). Patterns of protein kinase C isoenzyme expression in transitional cell carcinoma of bladder. Relation to degree of malignancy. Am. J. Clin. Pathol. *116*, 377–385.

Linch, M., Sanz-Garcia, M., Soriano, E., Zhang, Y., Riou, P., Rosse, C., Cameron, A., Knowles, P., Purkiss, A., Kjaer, S., et al. (2013). A cancer-associated mutation in atypical protein kinase Cι occurs in a substrate-specific recruitment motif. Sci. Signal. *6*, ra82.

Lu, Z., Hornia, A., Jiang, Y.W., Zang, Q., Ohno, S., and Foster, D.A. (1997). Tumor promotion by depleting cells of protein kinase C delta. Mol. Cell. Biol. *17*, 3418–3428.

Lu, H.C., Chou, F.P., Yeh, K.T., Chang, Y.S., Hsu, N.C., and Chang, J.G. (2009). Analysing the expression of protein kinase C eta in human hepatocellular carcinoma. Pathology *41*, 626–629.

Luna-Ulloa, L.B., Hernández-Maqueda, J.G., Santoyo-Ramos, P., Castañeda-Patlán, M.C., and Robles-Flores, M. (2011). Protein kinase C ζ is a positive modulator of canonical Wnt signaling pathway in tumoral colon cell lines. Carcinogenesis *32*, 1615–1624.

Ma, L., Tao, Y., Duran, A., Llado, V., Galvez, A., Barger, J.F., Castilla, E.A., Chen, J., Yajima, T., Porollo, A., et al. (2013). Control of nutrient stress-induced metabolic reprogramming by PKCζ in tumorigenesis. Cell *152*, 599–611.

Mackay, H.J., and Twelves, C.J. (2007). Targeting the protein kinase C family: are we there yet? Nat. Rev. Cancer *7*, 554–562.

Mandil, R., Ashkenazi, E., Blass, M., Kronfeld, I., Kazimirsky, G., Rosenthal, G., Umansky, F., Lorenzo, P.S., Blumberg, P.M., and Brodie, C. (2001). Protein kinase Calpha and protein kinase Cdelta play opposite roles in the proliferation and apoptosis of glioma cells. Cancer Res. *61*, 4612–4619.

Mauro, L.V., Grossoni, V.C., Urtreger, A.J., Yang, C., Colombo, L.L., Morandi, A., Pallotta, M.G., Kazanietz, M.G., Bal de Kier Joffé, E.D., and Puricelli, L.L. (2010). PKC Delta (PKCdelta) promotes tumoral progression of human ductal pancreatic cancer. Pancreas *39*, e31–e41.

Medkova, M., and Cho, W. (1998). Mutagenesis of the C2 domain of protein kinase C-alpha. Differential roles of Ca2+ ligands and membrane binding residues. J. Biol. Chem. *273*, 17544–17552.

Mellemkjaer, L., Andersen, V., Linet, M.S., Gridley, G., Hoover, R., and Olsen, J.H. (1997). Non-Hodgkin's lymphoma and other cancers among a cohort of patients with systemic lupus erythematosus. Arthritis Rheum. *40*, 761–768.

Mosior, M., and Newton, A.C. (1998). Mechanism of the apparent cooperativity in the interaction of protein kinase C with phosphatidylserine. Biochemistry *37*, 17271–17279.

Myhre, S., Lingjærde, O.C., Hennessy, B.T., Aure, M.R., Carey, M.S., Alsner, J., Tramm, T., Overgaard, J., Mills, G.B., Børresen-Dale, A.L., and Sørlie, T. (2013). Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. Mol. Oncol. *7*, 704–718.

Neckers, L., and Workman, P. (2012). Hsp90 molecular chaperone inhibitors: are we there yet? Clin. Cancer Res. *18*, 64–76.

Nelson, T.J., and Alkon, D.L. (2009). Neuroprotective versus tumorigenic protein kinase C activators. Trends Biochem. Sci. *34*, 136–145.

Newton, A.C. (2003). Regulation of the ABC kinases by phosphorylation: protein kinase C as a paradigm. Biochem. J. *370*, 361–371.

Nezhat, F., Wadler, S., Muggia, F., Mandeli, J., Goldberg, G., Rahaman, J., Runowicz, C., Murgo, A.J., and Gardner, G.J. (2004). Phase II trial of the combination of bryostatin-1 and cisplatin in advanced or recurrent carcinoma of the cervix: a New York Gynecologic Oncology Group study. Gynecol. Oncol. *93*, 144–148.

Oster, H., and Leitges, M. (2006). Protein kinase C alpha but not PKCzeta suppresses intestinal tumor formation in ApcMin/+ mice. Cancer Res. *66*, 6955–6963.

Perry, A.S., Furusato, B., Nagle, R.B., and Ghosh, S. (2014). Increased aPKC Expression Correlates with Prostatic Adenocarcinoma Gleason Score and Tumor Stage in the Japanese Population. Prostate Cancer *2014*, 481697.

Pongracz, J., Clark, P., Neoptolemos, J.P., and Lord, J.M. (1995). Expression of protein kinase C isoenzymes in colorectal cancer tissue and their differential activation by different bile acids. Int. J. Cancer *61*, 35–39.

Prévostel, C., Martin, A., Alvaro, V., Jaffiol, C., and Joubert, D. (1997). Protein kinase C alpha and tumorigenesis of the endocrine gland. Horm. Res. *47*, 140–144.

Pu, Y.S., Huang, C.Y., Chen, J.Y., Kang, W.Y., Lin, Y.C., Shiu, Y.S., Chuang, S.J., Yu, H.J., Lai, M.K., Tsai, Y.C., et al. (2012). Down-regulation of PKCζ in renal cell carcinoma and its clinicopathological implications. J. Biomed. Sci. *19*, 39.

Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell *154*, 1380–1389.

Regala, R.P., Weems, C., Jamieson, L., Khoor, A., Edell, E.S., Lohse, C.M., and Fields, A.P. (2005). Atypical protein kinase C iota is an oncogene in human non-small cell lung cancer. Cancer Res. *65*, 8905–8911.

Reno, E.M., Haughian, J.M., Dimitrova, I.K., Jackson, T.A., Shroyer, K.R., and Bradford, A.P. (2008). Analysis of protein kinase C delta (PKC delta) expression in endometrial tumors. Hum. Pathol. *39*, 21–29.

Reyland, M.E. (2007). Protein kinase Cdelta and apoptosis. Biochem. Soc. Trans. *35*, 1001–1004.

Rogers, T.B., Inesi, G., Wade, R., and Lederer, W.J. (1995). Use of thapsigargin to study Ca2+ homeostasis in cardiac cells. Biosci. Rep. *15*, 341–349.

Salzer, E., Santos-Valente, E., Klaver, S., Ban, S.A., Emminger, W., Prengemann, N.K., Garncarz, W., Müllauer, L., Kain, R., Boztug, H., et al. (2013). B-cell deficiency and severe autoimmunity caused by deficiency of protein kinase C δ. Blood *121*, 3112–3116.

Suga, K., Sugimoto, I., Ito, H., and Hashimoto, E. (1998). Down-regulation of protein kinase C-alpha detected in human colorectal cancer. Biochem. Mol. Biol. Int. *44*, 523–528.

Sun, S., Schiller, J.H., and Gazdar, A.F. (2007). Lung cancer in never smokers—a different disease. Nat. Rev. Cancer *7*, 778–790.

Symonds, J.M., Ohm, A.M., Carter, C.J., Heasley, L.E., Boyle, T.A., Franklin, W.A., and Reyland, M.E. (2011). Protein kinase C δ is a downstream effector of oncogenic K-ras in lung tumors. Cancer Res. *71*, 2087–2097.

Szallasi, Z., Smith, C.B., Pettit, G.R., and Blumberg, P.M. (1994). Differential regulation of protein kinase C isozymes by bryostatin 1 and phorbol 12-myristate 13-acetate in NIH 3T3 fibroblasts. J. Biol. Chem. *269*, 2118–2124.

Varga, A., Czifra, G., Tállai, B., Németh, T., Kovács, I., Kovács, L., and Bíró, T. (2004). Tumor grade-dependent alterations in the protein kinase C isoform pattern in urinary bladder carcinomas. Eur. Urol. *46*, 462–465.

Violin, J.D., Zhang, J., Tsien, R.Y., and Newton, A.C. (2003). A genetically encoded fluorescent reporter reveals oscillatory phosphorylation by protein kinase C. J. Cell Biol. *161*, 899–909.

Walsh, M.F., Woo, R.K., Gomez, R., and Basson, M.D. (2004). Extracellular pressure stimulates colon cancer cell proliferation via a mechanism requiring PKC and tyrosine kinase signals. Cell Prolif. *37*, 427–441.

Wu, B., Zhou, H., Hu, L., Mu, Y., and Wu, Y. (2013). Involvement of PKCalpha activation in TF/VIIa/PAR2-induced proliferation, migration, and survival of colon cancer cell SW620. Tumour Biol. *34*, 837–846.

Yoshida, K., Liu, H., and Miki, Y. (2006). Protein kinase C delta regulates Ser46 phosphorylation of p53 tumor suppressor in the apoptotic response to DNA damage. J. Biol. Chem. *281*, 5734–5740.

Young, S., Parker, P.J., Ullrich, A., and Stabel, S. (1987). Down-regulation of protein kinase C is due to an increased rate of degradation. Biochem. J. *244*, 775–779.

Zhang, L., Huang, J., Yang, N., Liang, S., Barchetti, A., Giannakakis, A., Cadungog, M.G., O'Brien-Jenkins, A., Massobrio, M., Roby, K.F., et al. (2006). Integrative genomic analysis of protein kinase C (PKC) family identifies PKCiota as a biomarker and potential oncogene in ovarian carcinoma. Cancer Res. *66*, 4627–4635.

Zhang, L.L., Cao, F.F., Wang, Y., Meng, F.L., Zhang, Y., Zhong, D.S., and Zhou, Q.H. (2014). The protein kinase C (PKC) inhibitors combined with chemotherapy in the treatment of advanced non-small cell lung cancer: meta-analysis of randomized controlled trials. Clin. Trans. Oncol. Published online October 29, 2014. http://dx.doi.org/10.1007/s12094-014-1241-3.

Zhu, Y., Dong, Q., Tan, B.J., Lim, W.G., Zhou, S., and Duan, W. (2005). The PKCalpha-D294G mutant found in pituitary and thyroid tumors fails to transduce extracellular signals. Cancer Res. *65*, 4520–4524.

10.2 Appendix Two: Primer List

SDM = Site Directed Mutagenesis

Gene	Purpose	Primer
PAK4	Mutation validation	TCCATGGCATCTCTTCATTGCGTC
PAK4	Mutation validation	CTGTGTGTGTCGTGCTGTCCTGT
PAK4	Mutation validation	TGAGGACAGAGGCAGGGACCCAG
PAK4	SDM (Stop) F	CAGAACCGCACCAGATGACCAACTTTCTTGTACAA
PAK4	SDM (Stop) RC	TTGTACAAGAAAGTTGGTCATCTGGTGCGGTTCTG
PAK4	SDM (E119Q) F	CCAGGAAAATGGGATGCCACAGGAGCCGGC
PAK4	SDM (E119Q) RC	GCCGGCTCCTGTGGCATCCCATTTTCCTGG
MAP2K4	SDM (KD) F	CAAACCAAGTGGGCAAATAATGGCAGTTATGAGAA TTCGGTCAACA
MAP2K4	SDM (KD) RC	TGTTGACCGAATTCTCATAACTGCCATTATTTGCCC ACTTGGTTTG
MAP2K4	SDM (G265D) F	CCATGTATGGCCTACAGTCAGCATCTCTTGTCTTG
MAP2K4	SDM (G265D) RC	CAAGACAAGAGATGCTGACTGTAGGCCATACATGG
MAP2K4	SDM (G265C) F	GCCAAGACAAGAGATGCTTGCTGTAGGCCATACAT
MAP2K4	SDM (G265C) RC	ATGTATGGCCTACAGCAAGCATCTCTTGTCTTGGC
MAP2K7	SDM (KD) F	CCACGTCATTGCCGTTATGCAAATGCGGCG
MAP2K7	SDM (KD) RC	CGCCGCATTTGCATAACGGCAATGACGTGG
MAP2K7	SDM (G129S) F	GAGATGGGCAGCAGCACCTGCGGCC
MAP2K7	SDM (G129S) RC	GGCCGCAGGTGCTGCTGCCCATCTC
MAP2K7	SDM (G129D) F	GATGGGCAGCGACACCTGCGGCC
MAP2K7	SDM (G129D) RC	GGCCGCAGGTGTCGCTGCCCATC
MAP2K7	SDM (D243N) F	GTGTCATCCACCGCAACGTCAAGCCCTCC
MAP2K7	SDM (D243N) RC	GGAGGGCTTGACGTTGCGGTGGATGACAC
MAP2K7	SDM (D243Y) F	GTGTCATCCACCGCTACGTCAAGCCCTCC
MAP2K7	SDM (D243Y) RC	GGAGGGCTTGACGTAGCGGTGGATGACAC
MAP2K7	SDM (D261N) F	CAGATCAAGCTCTGCAACTTCGGCATCAGCG
MAP2K7	SDM (D261N) RC	CGCTGATGCCGAAGTTGCAGAGCTTGATCTG
MAP2K7	SDM (A285T) F	GCCGCCTACATGACACCCGAGCGCA
MAP2K7	SDM (A285T) RC	TGCGCTCGGGTGTCATGTAGGCGGC
MAP2K7	SDM (E287K) F	CTACATGGCACCCCAGCGCATTGACCC
MAP2K7	SDM (E287K) RC	GGGTCAATGCGCTGGGGTGCCATGTAG
MAP2K7	SDM (290fs) F	CGAGCGCATTGACATGGCGGCGTCC
MAP2K7	SDM (290fs) RC	GGACGCCGCCATGTCAATGCGCTCG
MAP3K13	attb F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCATGG CCAACTTTCAGGA
MAP3K13	attb RC	GGGGACCACTTTGTACAAGAAAGCTGGGTCTTACC AGGTAGCAGAGC
MAP3K13	SDM (G315D) F	CAAGATGTCATTTGCTGACACGGTCGCATGGATGG
MAP3K13	SDM (G315D) RC	CCATCCATGCGACCGTGTCAGCAAATGACATCTTG
PKCq	SDM (G541V) F	GACGAATACCTTCTGTGTGACACCTGACTACATCG
PKCq	SDM (G541V) RC	CGATGTAGTCAGGTGTCACACAGAAGGTATTCGTC
MEN1	SNV validation F1	CCAACCTATGCTTACCTTTTC
MEN1	SNV validation F2	GTAAGAGACTGATCTGTGCC
MEN1	SNV validation RC1	CTGCTCTGGCCATCCCATCC
MEN1	SNV validation RC2	CCTGTAGTGCCCAGACCTCTGT
CDK8	SNV validation F1	GCTGTTAATTCTTAGGCGTTTTG
CDK8	SNV validation F2	GGTCGGACATTGTCTCTTGGTG

CDK8	SNV validation RC1	CAGGAAACCCATAAATATTTCCAT
CDK8	SNV validation RC2	GCTGTTTCTACGGATCTTTGA
PSIP1	SNV validation F1	AGGATGTGAACAGATGCATTGAG
PSIP1	SNV validation F2	GCTCAGAAACACACAGAGATG
PSIP1	SNV validation RC1	TCATCTGCCTCATGAGCAATGG
PSIP1	SNV validation RC2	TTCTGTGGCGTATACACAGTGA

10.3 Appendix Three:

R programming script locate critical motifs

```
#script to locate critical kinase motifs

#load StringR package
library("stringr", lib.loc="~/Library/R/3.1/library")


# match kinases with GENBANK sequences
genbank <- read.csv ( file = 'SOURCE_DEFINITIVE_GENBANK_TRIM.csv')
names   <- read.csv (file = 'SOURCE_DEFINITIVEGenename_TRIM.csv')
genbankkinases = merge(genbank, names, by = 'Gene', nomatch = 0)

# extract motifs as per Manning et al
HRDmotifs  <- as.vector(genbankkinases$HRDmotif)
VAIKmotifs <- as.vector(genbankkinases$VAIKmotif)
DFGmotifs  <- as.vector(genbankkinases$DFGmotif)

# count number of occurences of motif in each seq (using StringR package)
numberofDFG  <- str_count(genbankkinases$Protein_Seq,  DFGmotifs)
numberofVAIK <- str_count(genbankkinases$Protein_Seq, VAIKmotifs)
numberofHRD  <- str_count(genbankkinases$Protein_Seq,  HRDmotifs)
genbankkinases$numberofVAIK <- numberofVAIK
genbankkinases$numberofHRD  <- numberofHRD
genbankkinases$numberofDFG  <- numberofDFG

# remove any entry with number of VAIK or HRD or DFG = 0
genbankkinases <- subset(genbankkinases, (numberofVAIK != 0))
genbankkinases <- subset(genbankkinases, (numberofHRD  != 0))
genbankkinases <- subset(genbankkinases, (numberofDFG  != 0))

# SOURCE_motif2use is a file with the correct motif to use when there are multiple
motifToUse     <- read.csv ('SOURCE_motif2use.csv')
genbankkinases <- merge    (genbankkinases, motifToUse, by = 'f')

#clean up genbankkinases
colnames(genbankkinases)[2] <- "Gene"
colnames(genbankkinases)[3] <- "Length"
colnames(genbankkinases)[4] <- "Protein_Seq"
colnames(genbankkinases)[6] <- "VAIKmotif"
colnames(genbankkinases)[7] <- "HRDmotif"
colnames(genbankkinases)[8] <- "DFGmotif"

#locate all motifs
HRDmotifs  <- as.vector(genbankkinases$HRDmotif)
VAIKmotifs <- as.vector(genbankkinases$VAIKmotif)
DFGmotifs  <- as.vector(genbankkinases$DFGmotif)

# locate VAIK
firstVAIK <- str_locate(genbankkinases$Protein_Seq, VAIKmotifs)
firstVAIK <- as.data.frame(firstVAIK)
genbankkinases$firstVAIK <- as.numeric(firstVAIK$start)
genbankkinases$actual_lysine <- as.numeric(genbankkinases$firstVAIK + 3)

#locate all HRD
firstHRD <- str_locate(genbankkinases$Protein_Seq, HRDmotifs)
```

```
firstHRD <- as.data.frame(firstHRD)
genbankkinases$firstHRD <- firstHRD$start
firstHRD <- as.numeric(firstHRD$start)
endofstring = as.numeric(firstHRD +1000000)
beyondfirstHRD <- substr(genbankkinases$Protein_Seq, firstHRD + 1, endofstring)
secondHRD <- str_locate(beyondfirstHRD, HRDmotifs)
secondHRD <- as.data.frame(secondHRD)
genbankkinases$secondHRD <- secondHRD$start
secondHRD = as.numeric(genbankkinases$secondHRD + firstHRD)
genbankkinases$secondHRD <- secondHRD
beyondsecondHRD <- substr(genbankkinases$Protein_Seq, secondHRD + 1, endofstring)
thirdHRD <- str_locate(beyondsecondHRD, HRDmotifs)
thirdHRD <- as.data.frame(thirdHRD)
genbankkinases$thirdHRD <- thirdHRD$start
thirdHRD = as.numeric(genbankkinases$thirdHRD + secondHRD)
genbankkinases$thirdHRD <- thirdHRD

#locate DFG
firstDFG <- str_locate(genbankkinases$Protein_Seq, DFGmotifs)
firstDFG <- as.data.frame(firstDFG)
genbankkinases$firstDFG <- firstDFG$start
firstDFG <- as.numeric(firstDFG$start)
endofstring = as.numeric(firstDFG +1000000)
beyondfirstDFG <- substr(genbankkinases$Protein_Seq, firstDFG + 1, endofstring)
secondDFG <- str_locate(beyondfirstDFG, DFGmotifs)
secondDFG <- as.data.frame(secondDFG)
genbankkinases$secondDFG <- secondDFG$start
secondDFG = as.numeric(genbankkinases$secondDFG + firstDFG)
genbankkinases$secondDFG <- secondDFG
beyondsecondDFG <- substr(genbankkinases$Protein_Seq, secondDFG + 1, endofstring)
thirdDFG <- str_locate(beyondsecondDFG, DFGmotifs)
thirdDFG <- as.data.frame(thirdDFG)
genbankkinases$thirdDFG <- thirdDFG$start
thirdDFG = as.numeric(genbankkinases$thirdDFG + secondDFG)
genbankkinases$thirdDFG <- thirdDFG

# choose correct HRD
genbankkinases$actual_HRD_H <- ifelse (genbankkinases$HRDtouse == '1',
genbankkinases$firstHRD, ifelse (genbankkinases$HRDtouse == '2',
genbankkinases$secondHRD , genbankkinases$thirdHRD))
genbankkinases$actual_HRD_R <- genbankkinases$actual_HRD_H + 1
genbankkinases$actual_HRD_D <- genbankkinases$actual_HRD_H + 2

# choose correct DFG
genbankkinases$actual_DFG_D <- ifelse (genbankkinases$DFGtouse == '1',
genbankkinases$firstDFG, ifelse (genbankkinases$DFGtouse == '2',
genbankkinases$secondDFG , genbankkinases$thirdDFG))
genbankkinases$actual_DFG_F <- genbankkinases$actual_DFG_D + 1
genbankkinases$actual_DFG_G <- genbankkinases$actual_DFG_D + 2

# find APE motif
APEfragstart <- genbankkinases$actual_DFG_G + 10
genbankkinases$APEfragstart <- as.numeric(APEfragstart)
APEfragend <- genbankkinases$actual_DFG_G + 70
genbankkinases$APEfragend <- as.numeric(APEfragend)
APEfrag <- str_sub(genbankkinases$Protein_Seq, genbankkinases$APEfragstart,
genbankkinases$APEfragend )
genbankkinases$APEfrag <- APEfrag

# find APE
```

```
numAPE <- str_count(APEfrag,  'APE')
genbankkinases$numAPE <- as.numeric(numAPE)
APELoc <- str_locate(APEfrag, 'APE')
APELoc <- as.data.frame(APELoc)
genbankkinases$APELoc <- as.numeric(APELoc$start)
genbankkinases$locAPEprotein <- genbankkinases$APEfragstart + genbankkinases$APELoc
- 1

# find PE
numPE <- str_count(APEfrag,  'PE')
genbankkinases$numPE <- as.numeric(numPE)
PELoc <- str_locate(APEfrag, 'PE')
PELoc <- as.data.frame(PELoc)
genbankkinases$PELoc <- as.numeric(PELoc$start)
genbankkinases$locPEprotein <- genbankkinases$APEfragstart + genbankkinases$PELoc - 2
# minus 2 as PE is 1 further back

# find GTxxxxxxE
numGTx6E <- str_count(APEfrag,  'GT[A-Z]{6}E')
genbankkinases$numGTx6E  <- as.numeric(numGTx6E)
GTx6ELoc <- str_locate(APEfrag, 'GT[A-Z]{6}E')
GTx6ELoc <- as.data.frame(GTx6ELoc)
genbankkinases$GTx6ELoc <- as.numeric(GTx6ELoc$start)
genbankkinases$locGTx6Eprotein <- genbankkinases$APEfragstart +
genbankkinases$GTx6ELoc + 5

# find GTxxxxxNE
numGTx5NE <- str_count(APEfrag,  'GT[A-Z]{5}NE')
genbankkinases$numGTx5NE  <- as.numeric(numGTx5NE)
GTx5NELoc <- str_locate(APEfrag, 'GT[A-Z]{5}NE')
GTx5NELoc <- as.data.frame(GTx5NELoc)
genbankkinases$GTx5NELoc <- as.numeric(GTx5NELoc$start)
genbankkinases$locGTx5NEprotein <- genbankkinases$APEfragstart +
genbankkinases$GTx5NELoc + 5

# find GTxxxxxD
numGTx6D <- str_count(APEfrag,  'GT[A-Z]{6}D')
genbankkinases$numGTx6D  <- as.numeric(numGTx6D)
GTx6DLoc <- str_locate(APEfrag, 'GT[A-Z]{6}D')
GTx6DLoc <- as.data.frame(GTx6DLoc)
genbankkinases$GTx6DLoc <- as.numeric(GTx6DLoc$start)
genbankkinases$locGTx6Dprotein <- genbankkinases$APEfragstart +
genbankkinases$GTx6DLoc + 5

# find AxE
numAxE <- str_count(APEfrag,  'A[A-Z]E')
genbankkinases$numAxE <- as.numeric(numAxE)
AxELoc <- str_locate(APEfrag, 'A[A-Z]E')
AxELoc <- as.data.frame(AxELoc)
genbankkinases$AxELoc <- as.numeric(AxELoc$start)
genbankkinases$locAxEprotein <- genbankkinases$APEfragstart + genbankkinases$AxELoc -
1

# find APD
numAPD <- str_count(APEfrag,  'APD')
genbankkinases$numAPD  <- as.numeric(numAPD)
APDLoc <- str_locate(APEfrag, 'APD')
APDLoc <- as.data.frame(APDLoc)
genbankkinases$APDLoc <- as.numeric(APDLoc$start)
```

```
genbankkinases$locAPDprotein <- genbankkinases$APEfragstart + genbankkinases$APDLoc
- 1

# find PPD
numPPD <- str_count(APEfrag,  'PPD')
genbankkinases$numPPD  <- as.numeric(numPPD)
PPDLoc <- str_locate(APEfrag, 'PPD')
PPDLoc <- as.data.frame(PPDLoc)
genbankkinases$PPDLoc <- as.numeric(PPDLoc$start)
genbankkinases$locPPDprotein <- genbankkinases$APEfragstart + genbankkinases$PPDLoc
- 1

# find GTxxY
numGTxxY <- str_count(APEfrag,  'GT[A-Z]{2}Y')
genbankkinases$numGTxxY  <- as.numeric(numGTxxY)
GTxxYLoc <- str_locate(APEfrag, 'GT[A-Z]{2}Y')
GTxxYLoc <- as.data.frame(GTxxYLoc)
genbankkinases$GTxxYLoc <- as.numeric(GTxxYLoc$start)
genbankkinases$locGTxxYprotein <- genbankkinases$APEfragstart +
genbankkinases$GTxxYLoc + 5

# find PIR
numPIR <- str_count(APEfrag,  'PIR')
genbankkinases$numPIR  <- as.numeric(numPIR)
PIRLoc <- str_locate(APEfrag, 'PIR')
PIRLoc <- as.data.frame(PIRLoc)
genbankkinases$PIRLoc <- as.numeric(PIRLoc$start)
genbankkinases$locPIRprotein <- genbankkinases$APEfragstart + genbankkinases$PIRLoc +
4

# find Gx7E
numGx7E <- str_count(APEfrag,  'G[A-Z]{7}E')
genbankkinases$numGx7E  <- as.numeric(numGx7E)
Gx7ELoc <- str_locate(APEfrag, 'G[A-Z]{7}E')
Gx7ELoc <- as.data.frame(Gx7ELoc)
genbankkinases$Gx7ELoc <- as.numeric(Gx7ELoc$start)
genbankkinases$locGx7Eprotein <- genbankkinases$APEfragstart +
genbankkinases$Gx7ELoc + 5

# find YxAP (captures MAPKAPK5)
numYxAP <- str_count(APEfrag,  'Y[A-Z]{1}AP')
genbankkinases$numYxAP  <- as.numeric(numYxAP)
YxAPLoc <- str_locate(APEfrag, 'Y[A-Z]{1}AP')
YxAPLoc <- as.data.frame(YxAPLoc)
genbankkinases$YxAPLoc <- as.numeric(YxAPLoc$start)
genbankkinases$YxAPLocprotein <- genbankkinases$APEfragstart +
genbankkinases$YxAPLoc + 1

# find WYxxPR (captures MAPK4 and MAPK6)
numWYxxPR <- str_count(APEfrag,  'WY[A-Z]{2}PR')
genbankkinases$numWYxxPR  <- as.numeric(numWYxxPR)
WYxxPRLoc <- str_locate(APEfrag, 'WY[A-Z]{2}PR')
WYxxPRLoc <- as.data.frame(WYxxPRLoc)
genbankkinases$WYxxPRLoc <- as.numeric(WYxxPRLoc$start)
genbankkinases$WYxxPRLocprotein <- genbankkinases$APEfragstart +
genbankkinases$WYxxPRLoc + 2

# choose correct APE
genbankkinases$loc_actual_APE_A <- ifelse (genbankkinases$numAPE <1,
genbankkinases$locGTx5NEprotein, genbankkinases$locAPEprotein)
```

```
genbankkinases$loc_actual_APE_B <- genbankkinases$loc_actual_APE_A
genbankkinases$loc_actual_APE_C   <- ifelse (genbankkinases$loc_actual_APE_A %in%
NA, genbankkinases$locGTx6Eprotein, genbankkinases$loc_actual_APE_B)
genbankkinases$loc_actual_APE_D <- genbankkinases$loc_actual_APE_C
genbankkinases$loc_actual_APE_E   <- ifelse (genbankkinases$loc_actual_APE_C %in%
NA, genbankkinases$locPEprotein, genbankkinases$loc_actual_APE_D)
genbankkinases$loc_actual_APE_F <- genbankkinases$loc_actual_APE_E
genbankkinases$loc_actual_APE_G   <- ifelse (genbankkinases$loc_actual_APE_E %in%
NA, genbankkinases$locGTx6Dprotein, genbankkinases$loc_actual_APE_F)
genbankkinases$loc_actual_APE_H <- genbankkinases$loc_actual_APE_G
genbankkinases$loc_actual_APE_I   <- ifelse (genbankkinases$loc_actual_APE_G %in% NA,
genbankkinases$locGTxxYprotein, genbankkinases$loc_actual_APE_H)
genbankkinases$loc_actual_APE_J <- genbankkinases$loc_actual_APE_I
genbankkinases$loc_actual_APE_K   <- ifelse (genbankkinases$loc_actual_APE_I %in% NA,
genbankkinases$locAPDprotein, genbankkinases$loc_actual_APE_J)
genbankkinases$loc_actual_APE_L  <- genbankkinases$loc_actual_APE_K
genbankkinases$loc_actual_APE_M   <- ifelse (genbankkinases$loc_actual_APE_K %in%
NA, genbankkinases$locPPDprotein, genbankkinases$loc_actual_APE_L)
genbankkinases$loc_actual_APE_N  <- genbankkinases$loc_actual_APE_M
genbankkinases$loc_actual_APE_O   <- ifelse (genbankkinases$loc_actual_APE_M %in%
NA, genbankkinases$WYxxPRLocprotein, genbankkinases$loc_actual_APE_N)
genbankkinases$loc_actual_APE_P <- genbankkinases$loc_actual_APE_O
genbankkinases$loc_actual_APE_Q   <- ifelse (genbankkinases$loc_actual_APE_O %in%
NA, genbankkinases$locAxEprotein, genbankkinases$loc_actual_APE_P)
genbankkinases$loc_actual_APE_R <- genbankkinases$loc_actual_APE_Q
genbankkinases$loc_actual_APE_S   <- ifelse (genbankkinases$loc_actual_APE_Q %in%
NA, genbankkinases$locPIRprotein, genbankkinases$loc_actual_APE_R)
genbankkinases$loc_actual_APE_T <- genbankkinases$loc_actual_APE_S
genbankkinases$loc_actual_APE_U   <- ifelse (genbankkinases$loc_actual_APE_S %in%
NA, genbankkinases$locGx7Eprotein, genbankkinases$loc_actual_APE_T)
genbankkinases$loc_actual_APE_V <- genbankkinases$loc_actual_APE_U
genbankkinases$loc_actual_APE   <- ifelse (genbankkinases$loc_actual_APE_U %in% NA,
genbankkinases$YxAPLocprotein, genbankkinases$loc_actual_APE_V)

genbankkinases$APE_m7loc   <- genbankkinases$loc_actual_APE - 7
genbankkinases$APE_m6loc   <- genbankkinases$loc_actual_APE - 6
genbankkinases$APE_m5loc   <- genbankkinases$loc_actual_APE - 5
genbankkinases$APE_m4loc   <- genbankkinases$loc_actual_APE - 4
genbankkinases$APE_m3loc   <- genbankkinases$loc_actual_APE - 3
genbankkinases$APE_m2loc   <- genbankkinases$loc_actual_APE - 2
genbankkinases$APE_m1loc   <- genbankkinases$loc_actual_APE - 1
genbankkinases$APE_A1loc   <- genbankkinases$loc_actual_APE
genbankkinases$APE_P1loc   <- genbankkinases$loc_actual_APE + 1
genbankkinases$APE_E1loc   <- genbankkinases$loc_actual_APE + 2

genbankkinases$APE_m7      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m7loc , genbankkinases$APE_m7loc)
genbankkinases$APE_m6      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m6loc , genbankkinases$APE_m6loc)
genbankkinases$APE_m5      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m5loc , genbankkinases$APE_m5loc)
genbankkinases$APE_m4      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m4loc , genbankkinases$APE_m4loc)
genbankkinases$APE_m3      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m3loc , genbankkinases$APE_m3loc)
genbankkinases$APE_m2      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m2loc , genbankkinases$APE_m2loc)
genbankkinases$APE_m1      <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_m1loc , genbankkinases$APE_m1loc)
```

```
genbankkinases$APE_A1       <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_A1loc , genbankkinases$APE_A1loc)
genbankkinases$APE_P1       <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_P1loc , genbankkinases$APE_P1loc)
genbankkinases$APE_E1       <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$APE_E1loc , genbankkinases$APE_E1loc)

genbankkinases$APE_MOTIF <- paste(genbankkinases$APE_A1, genbankkinases$APE_P1,
genbankkinases$APE_E1)
genbankkinases$PE_MOTIF <- paste(genbankkinases$APE_P1, genbankkinases$APE_E1)

# if no G in APEm6 then look for G in APEm8 and then APEm9 and then A's in these positions
genbankkinases$correctedG_loc <- ifelse (genbankkinases$APE_m6 != 'G',
genbankkinases$APE_m9loc, genbankkinases$APE_m6loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG != 'G',
genbankkinases$APE_m5loc, genbankkinases$correctedG_loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG != 'G',
genbankkinases$APE_m8loc, genbankkinases$correctedG_loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG != 'G',
genbankkinases$APE_m10loc, genbankkinases$correctedG_loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG != 'G',
genbankkinases$APE_m11loc, genbankkinases$correctedG_loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG != 'G',
genbankkinases$APE_m7loc, genbankkinases$correctedG_loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

genbankkinases$correctedG_loc <- ifelse (genbankkinases$correctedG == 'G',
genbankkinases$correctedG_loc, genbankkinases$APE_m6loc)
genbankkinases$correctedG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$correctedG_loc , genbankkinases$correctedG_loc)

#end of kinase domain = APE (+ defined end)for truncation screen)
genbankkinases$endCAT <- as.numeric(genbankkinases$APE_E1loc)

# find GxGxxG
GxGfragstart <- genbankkinases$actual_lysine - 45
genbankkinases$GxGfragstart <- as.numeric(GxGfragstart)

# NEED TO PUT THIS LINE IN FOR ALL MOTIF SEARCHES - TELLING THE FRAG TO
START AT 1 IF MINUS NUMBER
genbankkinases$GxGfragstart <- ifelse(genbankkinases$GxGfragstart < 1,
1,genbankkinases$GxGfragstart )

GxGfragend <- genbankkinases$actual_lysine - 4
```

```
genbankkinases$GxGfragend <- as.numeric(GxGfragend)
GxGfrag <- str_sub(genbankkinases$Protein_Seq, genbankkinases$GxGfragstart,
genbankkinases$GxGfragend )
genbankkinases$GxGfrag <- GxGfrag

# find GxGxxG
numGxGxxG <- str_count(GxGfrag,  'G[A-Z]G[A-Z]{2}G')
genbankkinases$numGxGxxG <- as.numeric(numGxGxxG)
GxGxxGLoc <- str_locate(GxGfrag, 'G[A-Z]G[A-Z]{2}G')
GxGxxGLoc <- as.data.frame(GxGxxGLoc)
genbankkinases$GxGxxGLoc <- as.numeric(GxGxxGLoc$start)
genbankkinases$locGxGxxGprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxGxxGLoc - 1

# find GxGxF
numGxGxF <- str_count(GxGfrag,  'G[A-Z]G[A-Z]F')
genbankkinases$numGxGxF <- as.numeric(numGxGxF)
GxGxFLoc <- str_locate(GxGfrag, 'G[A-Z]G[A-Z]F')
GxGxFLoc <- as.data.frame(GxGxFLoc)
genbankkinases$GxGxFLoc <- as.numeric(GxGxFLoc$start)
genbankkinases$locGxGxFprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxGxFLoc - 1

# find GxGxY
numGxGxY <- str_count(GxGfrag,  'G[A-Z]G[A-Z]Y')
genbankkinases$numGxGxY <- as.numeric(numGxGxY)
GxGxYLoc <- str_locate(GxGfrag, 'G[A-Z]G[A-Z]Y')
GxGxYLoc <- as.data.frame(GxGxYLoc)
genbankkinases$GxGxYLoc <- as.numeric(GxGxYLoc$start)
genbankkinases$locGxGxYprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxGxYLoc - 1

# find GxFG
numGxFG <- str_count(GxGfrag,  'G[A-Z]FG')
genbankkinases$numGxFG <- as.numeric(numGxFG)
GxFGLoc <- str_locate(GxGfrag, 'G[A-Z]FG')
GxFGLoc <- as.data.frame(GxFGLoc)
genbankkinases$GxFGLoc <- as.numeric(GxFGLoc$start)
genbankkinases$locGxFGprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxFGLoc - 3
#above is minus 3 rather than minus 1 to take account of the GxFG starting 2 on from GxGxxG

# find GxGxxA
numGxGxxA <- str_count(GxGfrag,  'G[A-Z]G[A-Z]{2}A')
genbankkinases$numGxGxxA <- as.numeric(numGxGxxA)
GxGxxALoc <- str_locate(GxGfrag, 'G[A-Z]G[A-Z]{2}A')
GxGxxALoc <- as.data.frame(GxGxxALoc)
genbankkinases$GxGxxALoc <- as.numeric(GxGxxALoc$start)
genbankkinases$locGxGxxAprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxGxxALoc - 1

# find GxGxxS
numGxGxxS <- str_count(GxGfrag,  'G[A-Z]G[A-Z]{2}S')
genbankkinases$numGxGxxS <- as.numeric(numGxGxxS)
GxGxxSLoc <- str_locate(GxGfrag, 'G[A-Z]G[A-Z]{2}S')
GxGxxSLoc <- as.data.frame(GxGxxSLoc)
genbankkinases$GxGxxSLoc <- as.numeric(GxGxxSLoc$start)
genbankkinases$locGxGxxSprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxGxxSLoc - 1
```

```
# find SxGxxG
numSxGxxG <- str_count(GxGfrag,  'S[A-Z]G[A-Z]{2}G')
genbankkinases$numSxGxxG <- as.numeric(numSxGxxG)
SxGxxGLoc <- str_locate(GxGfrag, 'S[A-Z]G[A-Z]{2}G')
SxGxxGLoc <- as.data.frame(SxGxxGLoc)
genbankkinases$SxGxxGLoc <- as.numeric(SxGxxGLoc$start)
genbankkinases$locSxGxxGprotein <- genbankkinases$GxGfragstart +
genbankkinases$SxGxxGLoc - 1

# find GxxG
numGxxG <- str_count(GxGfrag,  'G[A-Z]{2}G')
genbankkinases$numGxxG <- as.numeric(numGxxG)
GxxGLoc <- str_locate(GxGfrag, 'G[A-Z]{2}G')
GxxGLoc <- as.data.frame(GxxGLoc)
genbankkinases$GxxGLoc <- as.numeric(GxxGLoc$start)
genbankkinases$locGxxGprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxxGLoc - 3

# find G4xG
numG4xG <- str_count(GxGfrag,  'G[A-Z]{4}G')
genbankkinases$numG4xG <- as.numeric(numG4xG)
G4xGLoc <- str_locate(GxGfrag, 'G[A-Z]{4}G')
G4xGLoc <- as.data.frame(G4xGLoc)
genbankkinases$G4xGLoc <- as.numeric(G4xGLoc$start)
genbankkinases$locG4xGprotein <- genbankkinases$GxGfragstart +
genbankkinases$G4xGLoc - 1

# find GxG
numGxG <- str_count(GxGfrag,  'G[A-Z]G')
genbankkinases$numGxG <- as.numeric(numGxG)
GxGLoc <- str_locate(GxGfrag, 'G[A-Z]G')
GxGLoc <- as.data.frame(GxGLoc)
genbankkinases$GxGLoc <- as.numeric(GxGLoc$start)
genbankkinases$locGxGprotein <- genbankkinases$GxGfragstart + genbankkinases$GxGLoc
- 1

# find GxTxF
numGxTxF <- str_count(GxGfrag,  'G[A-Z]T[A-Z]F')
genbankkinases$numGxTxF <- as.numeric(numGxTxF)
GxTxFLoc <- str_locate(GxGfrag, 'G[A-Z]T[A-Z]F')
GxTxFLoc <- as.data.frame(GxTxFLoc)
genbankkinases$GxTxFLoc <- as.numeric(GxTxFLoc$start)
genbankkinases$locGxTxFprotein <- genbankkinases$GxGfragstart +
genbankkinases$GxTxFLoc - 1

# find GG
numGG <- str_count(GxGfrag, 'GG')
genbankkinases$numGG <- as.numeric(numGG)
GGLoc <- str_locate(GxGfrag, 'GG')
GGLoc <- as.data.frame(GGLoc)
genbankkinases$GGLoc <- as.numeric(GGLoc$start)
genbankkinases$locGGprotein <- genbankkinases$GxGfragstart + genbankkinases$GGLoc -
3

# choose correct GxG
genbankkinases$loc_actual_gxg_A <- ifelse (genbankkinases$numGxGxxG <1,
genbankkinases$locGxGxFprotein, genbankkinases$locGxGxxGprotein)
genbankkinases$loc_actual_gxg_B <- genbankkinases$loc_actual_gxg_A
genbankkinases$loc_actual_gxg_C   <- ifelse (genbankkinases$loc_actual_gxg_A %in% NA,
genbankkinases$locGxFGprotein, genbankkinases$loc_actual_gxg_B)
```

```
genbankkinases$loc_actual_gxg_D <- genbankkinases$loc_actual_gxg_C
genbankkinases$loc_actual_gxg_E   <- ifelse (genbankkinases$loc_actual_gxg_C %in% NA,
genbankkinases$locGxGxYprotein, genbankkinases$loc_actual_gxg_D)
genbankkinases$loc_actual_gxg_F <- genbankkinases$loc_actual_gxg_E
genbankkinases$loc_actual_gxg_G   <- ifelse (genbankkinases$loc_actual_gxg_E %in% NA,
genbankkinases$locGxGxxAprotein, genbankkinases$loc_actual_gxg_F)
genbankkinases$loc_actual_gxg_H <- genbankkinases$loc_actual_gxg_G
genbankkinases$loc_actual_gxg_I   <- ifelse (genbankkinases$loc_actual_gxg_G %in% NA,
genbankkinases$locGxGxxSprotein, genbankkinases$loc_actual_gxg_H)
genbankkinases$loc_actual_gxg_J <- genbankkinases$loc_actual_gxg_I
genbankkinases$loc_actual_gxg_K   <- ifelse (genbankkinases$loc_actual_gxg_I %in% NA,
genbankkinases$locSxGxxGprotein, genbankkinases$loc_actual_gxg_J)
genbankkinases$loc_actual_gxg_L <- genbankkinases$loc_actual_gxg_K
genbankkinases$loc_actual_gxg_M   <- ifelse (genbankkinases$loc_actual_gxg_K %in% NA,
genbankkinases$locGxxGprotein, genbankkinases$loc_actual_gxg_L)
genbankkinases$loc_actual_gxg_N <- genbankkinases$loc_actual_gxg_M
genbankkinases$loc_actual_gxg_O   <- ifelse (genbankkinases$loc_actual_gxg_M %in% NA,
genbankkinases$locG4xGprotein, genbankkinases$loc_actual_gxg_N)
genbankkinases$loc_actual_gxg_P <- genbankkinases$loc_actual_gxg_O
genbankkinases$loc_actual_gxg_Q   <- ifelse (genbankkinases$loc_actual_gxg_O %in% NA,
genbankkinases$locGxGprotein, genbankkinases$loc_actual_gxg_P)
genbankkinases$loc_actual_gxg_R <- genbankkinases$loc_actual_gxg_Q
genbankkinases$loc_actual_gxg_S   <- ifelse (genbankkinases$loc_actual_gxg_Q %in% NA,
genbankkinases$locGxTxFprotein, genbankkinases$loc_actual_gxg_R)
genbankkinases$loc_actual_gxg_T <- genbankkinases$loc_actual_gxg_S
genbankkinases$loc_actual_gxg   <- ifelse (genbankkinases$loc_actual_gxg_S %in% NA,
genbankkinases$locGGprotein, genbankkinases$loc_actual_gxg_T)

#extract GxGxxG motif and location/aa for each
startGxG <- genbankkinases$loc_actual_gxg
endGxG <- genbankkinases$loc_actual_gxg + 5
genbankkinases$GxGxxGmotif <- str_sub(genbankkinases$Protein_Seq, startGxG, endGxG)

genbankkinases$GxGxxG_G1     <- str_sub(genbankkinases$GxGxxGmotif, 1, 1)
genbankkinases$GxGxxG_X1     <- str_sub(genbankkinases$GxGxxGmotif, 2, 2)
genbankkinases$GxGxxG_G2     <- str_sub(genbankkinases$GxGxxGmotif, 3, 3)
genbankkinases$GxGxxG_X2     <- str_sub(genbankkinases$GxGxxGmotif, 4, 4)
genbankkinases$GxGxxG_X3     <- str_sub(genbankkinases$GxGxxGmotif, 5, 5)
genbankkinases$GxGxxG_G3     <- str_sub(genbankkinases$GxGxxGmotif, 6, 6)


genbankkinases$GxGxxG_G1loc   <- genbankkinases$loc_actual_gxg
genbankkinases$GxGxxG_X1loc   <- genbankkinases$loc_actual_gxg + 1
genbankkinases$GxGxxG_G2loc   <- genbankkinases$loc_actual_gxg + 2
genbankkinases$GxGxxG_X2loc   <- genbankkinases$loc_actual_gxg + 3
genbankkinases$GxGxxG_X3loc   <- genbankkinases$loc_actual_gxg + 4
genbankkinases$GxGxxG_G3loc   <- genbankkinases$loc_actual_gxg + 5

genbankkinases$GXGXXG_MOTIF <- paste(genbankkinases$GxGxxG_G1,
genbankkinases$GxGxxG_G2, genbankkinases$GxGxxG_G3)

genbankkinases$CATALYTIC_FRAG <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$GxGxxG_G1loc, genbankkinases$APE_E1loc )

#find salt bridge e
genbankkinases$saltbridgeEfragSTART <- genbankkinases$actual_lysine + 5
genbankkinases$saltbridgeEfragEND <- genbankkinases$actual_lysine + 30
genbankkinases$saltbridgeEfrag <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$saltbridgeEfragSTART, genbankkinases$saltbridgeEfragEND )
saltbridgeEfrag <- genbankkinases$saltbridgeEfrag
```

```
#find ExxxL
numExxxL <- str_count(saltbridgeEfrag,  'E[A-Z]{3}L')
genbankkinases$numExxxL <- as.numeric(numExxxL)
ExxxLLoc <- str_locate(saltbridgeEfrag,  'E[A-Z]{3}L')
ExxxLLoc <- as.data.frame(ExxxLLoc)
genbankkinases$ExxxLLoc <- as.numeric(ExxxLLoc$start)
genbankkinases$ExxxLLocprotein <- genbankkinases$saltbridgeEfragSTART +
genbankkinases$ExxxLLoc - 1

#find ExxI
numExxI <- str_count(saltbridgeEfrag,  'E[A-Z]{2}I')
genbankkinases$numExxI <- as.numeric(numExxI)
ExxILoc <- str_locate(saltbridgeEfrag,  'E[A-Z]{2}I')
ExxILoc <- as.data.frame(ExxILoc)
genbankkinases$ExxILoc <- as.numeric(ExxILoc$start)
genbankkinases$ExxILocprotein <- genbankkinases$saltbridgeEfragSTART +
genbankkinases$ExxILoc - 1

#find ExxxM
numExxxM <- str_count(saltbridgeEfrag,  'E[A-Z]{3}M')
genbankkinases$numExxxM <- as.numeric(numExxxM)
ExxxMLoc <- str_locate(saltbridgeEfrag,  'E[A-Z]{3}M')
ExxxMLoc <- as.data.frame(ExxxMLoc)
genbankkinases$ExxxMLoc <- as.numeric(ExxxMLoc$start)
genbankkinases$ExxxMLocprotein <- genbankkinases$saltbridgeEfragSTART +
genbankkinases$ExxxMLoc - 1

#find ExxL
numExxL <- str_count(saltbridgeEfrag,  'E[A-Z]{2}L')
genbankkinases$numExxL <- as.numeric(numExxL)
ExxLLoc <- str_locate(saltbridgeEfrag,  'E[A-Z]{2}L')
ExxLLoc <- as.data.frame(ExxLLoc)
genbankkinases$ExxLLoc <- as.numeric(ExxLLoc$start)
genbankkinases$ExxLLocprotein <- genbankkinases$saltbridgeEfragSTART +
genbankkinases$ExxLLoc - 1

#find QxxxE
numQxxxE <- str_count(saltbridgeEfrag,  'Q[A-Z]{3}E')
genbankkinases$numQxxxE <- as.numeric(numQxxxE)
QxxxELoc <- str_locate(saltbridgeEfrag,  'Q[A-Z]{3}E')
QxxxELoc <- as.data.frame(QxxxELoc)
genbankkinases$QxxxELoc <- as.numeric(QxxxELoc$start)
genbankkinases$QxxxELocprotein <- genbankkinases$saltbridgeEfragSTART +
genbankkinases$QxxxELoc + 3

# choose correct saltbridge
genbankkinases$loc_actual_saltbridge_A <- ifelse (genbankkinases$numExxxL <1,
genbankkinases$ExxILocprotein, genbankkinases$ExxxLLocprotein)
genbankkinases$loc_actual_saltbridge_B <- genbankkinases$loc_actual_saltbridge_A
genbankkinases$loc_actual_saltbridge_C   <- ifelse (genbankkinases$loc_actual_saltbridge_A
%in% NA, genbankkinases$ExxxMLocprotein, genbankkinases$loc_actual_saltbridge_B)
genbankkinases$loc_actual_saltbridge_D <- genbankkinases$loc_actual_saltbridge_C
genbankkinases$loc_actual_saltbridge_E   <- ifelse (genbankkinases$loc_actual_saltbridge_C
%in% NA, genbankkinases$ExxLLocprotein, genbankkinases$loc_actual_saltbridge_D)
genbankkinases$loc_actual_saltbridge_F <- genbankkinases$loc_actual_saltbridge_E
genbankkinases$loc_actual_saltbridge   <- ifelse (genbankkinases$loc_actual_saltbridge_E
%in% NA, genbankkinases$QxxxELocprotein, genbankkinases$loc_actual_saltbridge_F)
```

```
genbankkinases$checkE <- str_sub(genbankkinases$Protein_Seq,
genbankkinases$loc_actual_saltbridge, genbankkinases$loc_actual_saltbridge)
```