


5-2016

Accurate mutation annotation and functional prediction enhance the applicability of -omics data in precision medicine

Tenghui Chen

Follow this and additional works at: http://digitalcommons.library.tmc.edu/utgsbs_dissertations

 Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Chen, Tenghui, "Accurate mutation annotation and functional prediction enhance the applicability of -omics data in precision medicine" (2016). *UT GSBS Dissertations and Theses (Open Access)*. Paper 666.

This Dissertation (PhD) is brought to you for free and open access by the Graduate School of Biomedical Sciences at DigitalCommons@The Texas Medical Center. It has been accepted for inclusion in UT GSBS Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@The Texas Medical Center. For more information, please contact laurel.sanders@library.tmc.edu.

**ACCURATE MUTATION ANNOTATION AND FUNCTIONAL
PREDICTION ENHANCE THE APPLICABILITY OF -OMICS DATA IN
PRECISION MEDICINE**

by
Tenghui Chen, M.S.

APPROVED:

Supervisory Professor: Ken Chen, Ph.D.

Keith Baggerly, Ph.D.

Roel Verhaak, Ph.D.

Han Liang, Ph.D.

Marcos Estecio, Ph.D.

APPROVED:

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

**ACCURATE MUTATION ANNOTATION AND FUNCTIONAL
PREDICTION ENHANCE THE APPLICABILITY OF -OMICS DATA IN
PRECISION MEDICINE**

A
THESIS

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment
of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

by
Tenghui Chen, M.S.
Houston, Texas
May, 2016

DEDICATION

To all my family, my mentors and my friends,
who have stood beside me through all this time,
for their help, support and love,
I couldn't have done this without you.

ACKNOWLEDGEMENTS

I would have never been able to complete my dissertation work without the guidance of my advisor, my advisory committee members, the help from my academic collaborators, and the consistent support from my family and friends.

My most sincere thanks go to my Ph.D advisor, Dr. Ken Chen, for his excellent guidance, caring, patience and continuous support throughout my research process. He created an excellent academic atmosphere for me to freely pursue my interest in translational genomics research. I am truly thankful for his continuous instructions and encouragements that helped me shape my theoretical knowledge, computational and technical skills for research. I would like to thank Dr. John Weinstein and Dr. Sanjey Shete, for their patient guidance and suggestions in my Ph.D research projects. I am also very grateful for the great help and advices from my advisory committee members: Dr. Keith Baggerly, Dr. Han Liang, Dr. Roel Verhaak and Dr. Marcos Estecio. They have routinely given me advice on improving my thesis work and tried to help me become a better scientist. I owe my thanks to Dr. Gordon Mills, Dr. Funda Metric-Berstam, Dr. Hao Zhao, Dr. Yuan Qi, Dr. Patrick Ng, Dr. Jia Zeng and other colleagues in the Institute for Personalized Cancer Therapy (IPCT), for their great suggestions on T200 related projects. I also owe my thanks to Dr. Wanding Zhou, Dr. Zechen Chong, Dr. Zixing Wang, Dr. Yong Mao, Xian Fan and Hamim Zafar from Dr. Ken Chen's research group. We have built an exciting collaborative environment together within the group. I would like to thank Dr. Jun Zhang, Mary Rohrdanz and Dr. Chris Wakefield for their information&technology

support.

I thank Dr. William Mattox, Dr. Andrew Bean, Ms. Lourdes Perez and other staff members from UT-GSBS for their great help and support in the past several years.

I would also like to thank my parents and my wife, for their unconditional support that encouraged me to freely pursue my academic research career over the past five years. Finally, I would like to thank all my friends for their continuous support and friendship.

ABSTRACT

ACCURATE MUTATION ANNOTATION AND FUNCTIONAL PREDICTION ENHANCE THE APPLICABILITY OF -OMICS DATA IN PRECISION MEDICINE

Tenghui Chen, M.S.
Advisory Professor: Ken Chen, Ph.D.

Clinical sequencing has been recognized as an effective approach for enhancing the accuracy and efficiency of cancer patient management and therefore achieve the goals of personalized therapy. However, the accuracy of large scale sequencing data in clinics has been constrained by many different aspects, such as clinical detection, annotation and interpretation of the variants that are observed in clinical sequencing data. In my Ph.D thesis work, I mainly investigated how to comprehensively and efficiently apply high dimensional -omics data to enhance the capability of precision cancer medicine. Following this motivation, my dissertation has been focused on two important topics in translational genomics.

1) Developing a computational approach to resolve ambiguities in existing clinical genomic annotations and to facilitate correct diagnostic and treatment decisions. I have developed a multi-level variant annotator, TransVar, to perform precise annotation at genomic, mRNA and protein levels. TransVar implements three main functions: 1) it performs an innovative “reverse annotation” function, which identifies the genomic variants that can be translated into a given protein

variant through alternative splicing. This function significantly improves the accuracy of genomic testing in clinics and functional validation in genomic laboratories; 2) It performs “equivalence annotation”, which identifies the protein variants having identical genomic origins with a given protein variant. This function resolves annotation inconsistencies among variants imported from different data sources, and is crucial for precise mutation biomarker identification and functional prediction; 3) It improves “forward annotation” (i.e., translation of genomic variants to protein variants) over existing annotators by more rigorously implementing the Human Genome Variation Society (HGVS) nomenclature. Our study also tried to illustrate the ambiguities of annotation among different transcript databases and different mutation types. TransVar standardizes mutation annotation and enables precise characterization of genomic variants in the context of functional genomic studies and clinical decision support and will significantly advance genomic medicine.

2) Developing a statistical framework to precisely identify hotspot mutations and investigate their functional impact on tumorigenesis and drug therapeutic response using large-scale -omics data. I have proposed a statistical model, which utilizes characteristics of genomic data to nominate 702 cancer type-specific hotspot mutations in 549 genes. It models background mutation rate variations among different genes, mutation subtypes and di-nucleotide sequence contexts and effectively identifies hotspots that have more than the expected number of recurrent mutations. We then investigate the mutational signatures represented by the hotspot mutations and find they vary from one tumor type to another, suggesting distinct

mutational positive selections during different cancer progressions. In addition, we build an integrative statistical framework by using transcriptomics, proteomics and pharmacogenomics data to investigate the diverse functions of each hotspot mutation under different disease and biological contexts and to associate the effects of mutations on RNA/protein expression, pathway activity, and drug sensitivity. We not only validate diverse functions of well-known hotspot mutations in different contexts, but also identify some novel hotspot mutations such as *MAP3K4* A1199 deletion, *NR1H2* R175 insertion, and *GATA3* P409 insertion with different functional associations. Our study addresses a long-term challenge of explicitly distinguishing driver mutations from passengers, and nominates a set of putative driver mutations that possess diverse functional potentials.

The translational genomics research I conducted in my Ph.D study will benefit the cancer research community. The tools I developed will answer translational genomics questions such as identification of biomarkers for clinical diagnostics and treatment, and promote our understanding of the biological function of driver mutations towards the realization of personalized medicine.

TABLE OF CONTENTS

APPROVAL PAGE	i
TITLE PAGE	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vi
LIST OF FIGURES	xiii
LIST OF TABLES	xvi
CHAPTERS	
1. Introduction	1
1.1. Application of next generation sequencing (NGS) data in personalized medicine	2
1.1.1. Impact of NGS techniques on human health	2
1.1.2. DNA NGS data analysis workflow	3
1.1.3. Challenges of DNA clinical sequencing and analysis	6
1.2. Mutation annotation in NGS data analysis	8
1.3. Mutation annotation ambiguities in translational and functional genomics study	9
1.3.1. Forward annotation	9
1.3.2. Reverse annotation	11
1.3.3. Limitations of current variant annotation algorithms	12
1.4. Cancer gene prediction	13

1.5. Driver mutation prediction	15
1.6. Application of high-dimensional -omics data in translational genomics	18
1.7. Impact of variant annotation on hotspot mutation prediction	19
1.8. Motivation and Rationale of the thesis study	20
2. TransVar: a multi-level variant annotator for precision genomics	23
2.1. Materials and methods	24
2.1.1. COSMIC and TCGA somatic mutation data	24
2.1.2. Transcriptome definitions	24
2.1.3. Reverse annotation	24
2.1.4. Forward annotation	26
2.1.5. Equivalence annotation	27
2.1.6. Tool implementation	28
2.2. Results	30
2.2.1. Overview of the functions of TransVar	30
2.2.2. Forward annotation of COSMIC mutations using TransVar, ANNOVAR, VEP, snpEff and Oncotator	32
2.2.3. Forward annotation ambiguities in TCGA and COSMIC mutation data	37
2.2.4. Forward annotation of RNA-editing sites by TransVar	37
2.2.5. Reverse annotation accuracy for SNV, indels and frame-shift variants	38

2.2.6. Reverse annotating protein phosphorylation sites using TransVar	43
2.2.7. Impact of TransVar on designing experimental validation	44
2.2.8. Impact of TransVar on Pharmacogenomics and hotspot mutation prediction	45
2.2.9. Web interface of TransVar	50
2.2.10. Command line usage of TransVar	52
2.3. Discussion	53
3. Hotspot mutations delineating diverse mutational signature and biological utilities across cancer types	55
3.1. Materials and methods	56
3.1.1. COSMIC somatic mutation data	56
3.1.2. Cancer gene candidates	56
3.1.3. Definition of hotspot mutations	56
3.1.4. TCGA pan-cancer data	59
3.1.5. Cancer Cell Line Encyclopedia (CCLE) mutation and drug sensitivity data	60
3.1.6. Tumor-type prevalence of hotspot mutations	60
3.1.7. Conservation score comparison	60
3.2. Results	62
3.2.1. Variable mutation rates among different tumor types and mutation subtypes	62
3.2.2. Identifying hotspot mutations in COSMIC	64

3.2.3. Evaluate the performance of hotspot mutation identification	66
3.2.4. Sequence context signature of hotspot mutations	69
3.2.5. Exploring the biological utilities of hotspot mutations using TCGA mRNA/protein expression data	71
3.2.6. Exploring the pharmacogenomics properties of hotspot mutations	74
3.2.7. Tumor type-specific hotspot mutations	79
3.2.8. Conservation and protein-domain characteristics of the hotspot mutations	85
3.3. Discussion	87
4. Conclusions and future directions	91
4.1. Conclusions	92
4.2. Future directions	95
APPENDIX	98
BIBLIOGRAPHY	133
VITA	156

LIST OF FIGURES

Figure 1.1 A pipeline overview of DNA next generation sequencing data analysis	5
Figure 2.1 Schematic overview of TransVar	31
Figure 2.2 Comparison of forward annotation consistency among TransVar, VEP, ANNOVAR, snpEff and Oncotator	34
Figure 2.3 Comparison of forward annotation consistency between COSMIC and TCGA mutation data using TransVar	36
Figure 2.4 Inconsistent forward annotation on performed in TCGA and COSMIC	49
Figure 2.5 The web interface of TransVar that shows how to perform the reverse annotation and what type of information could be obtained from the output	50
Figure 2.6 A screenshot of the homepage of TransVar	51
Figure 3.1 A schematic overview of HotDriver	58
Figure 3.2 Statistics of the mutation distribution in different tumor types in COSMIC	63
Figure 3.3 Illustration of hotspot mutations definition and functional utility analysis	65
Figure 3.4 Number of hotspot mutations defined in individual tumor types using COSMIC data	68

Figure 3.5 Mutational signatures of hotspot mutations in 16 tumor types	70
Figure 3.6 Functional implications of hotspot mutations in RNA and protein expression	73
Figure 3.7 Functional implications of hotspot mutations in signaling pathway activity	75
Figure 3.8 Functional implications of hotspot mutations in drug sensitivity	77
Figure 3.9 Prevalence of <i>TP53</i> hotspot mutations in different TCGA cancer types	81
Figure 3.10 Prevalence of hotspot mutations in different TCGA cancer types and their functional implications	84
Figure 3.11 Compare the conservation and proteomic domain localization of the hotspot and the non-hotspot mutations	86
Figure S2.1 Illustration of reporting ambiguity for a 3 bp insertion CTG at (A) genomic/mRNA levels and (B) protein level	98
Figure S3.1 The percentage of different mutational subtypes across all defined hotspot mutations	99
Figure S3.2 The significance of overlap (y-axis, calculated using Fisher exact test) between hotspot-mutation-containing-genes and previously known cancer genes at various adjusted p value cutoffs (x-axis)	99
Figure S3.3 Relationship between the number of hotspot mutations and the total number of mutations (mutation burden) in each tumor type	100

Figure S3.4 Functional implications of hotspot mutations in RNA and protein expression	101
Figure S3.5 Functional implications of hotspot mutations in drug sensitivity	102
Figure S3.6 Functional implications of hotspot mutations in drug sensitivity	103
Figure S3.7 Prevalence of hotspot mutations in different TCGA cancer types	104
Figure S3.8 Numbers of highly prevalent hotspot mutations in different tumor types	104

LIST OF TABLES

Table 1.1 Popular 2 nd generation sequencing technologies in the market	3
Table 2.1 Reverse annotation consistency of COSMIC protein identifiers via different transcript databases	40
Table 2.2 Reverse annotation consistency of COSMIC mRNA identifiers via different transcript databases	42
Table 2.3 Reverse annotation results of FGFR2:p.N549K using TransVar	45
Table 2.4 Clinically actionable cancer mutations with non-unique genomic origins	47
Table 3.1 Hotspot mutations exclusively detected in only one tumor type in TCGA pan-cancer data	83
Table S2.1 Comparing the annotation consistency of different mutation types using TransVar, VEP, ANNOVAR, snpEff and Oncotator	105
Table S3.1 Number of samples in 17 tumor types in COSMIC v71	106
Table S3.2 20 mutation subtypes that were included in the statistical modeling of hotspot mutation definition	107
Table S3.3 Number of samples available in different TCGA cancer types	108
Table S3.4 2×2 table of calculating the prevalence of target mutation B in samples A	109

Table S3.5 List of the predicted hotspot mutations in different tumor types based on COSMIC version 71	110
Table S3.6 A full list of the hotspot mutations that were highly prevalent in specific cancer types in TCGA	130

CHAPTER 1

Introductions

1.1 Application of next generation sequencing (NGS) data in personalized medicine

1.1.1 Impact of NGS techniques on human health

Starting from the 1970's, DNA sequencing techniques have continuously revolutionized our understanding of the human genome and enhanced our capability for learning biological principles from human genetics. The development of sequencing technologies originated from the pioneer works of Walter Gilbert [1] and Frederick Sanger [2]. After that, DNA sequencing technology continuously improved in terms of both instruments and mechanics to advance the generality and accuracy with which we understand human genome. In 2001, the human genome project (HGP) [3] was finished, which enabled us to have a close look at the human genetic codes for the first time and prompted the potential of examining different diseases in a personalized revolution. The achievements of the HGP lie in several aspects: 1) It gave a hint that people could actually utilize the genetic information to improve the understanding of disease susceptibilities and indications of disease prevention; 2) It involved the collaborations of multiple research institutes such as the Sanger Institute and the biotech industries, and increased the possibility of transferring human genetic research into business and in turn helping better monitor human health; 3) Most importantly, it let people believe that the era of personalized medicine is not far away, and comprehensively improved the clinical diagnostics and drug development using human genetic information.

After the HGP, which represented the 1st generation sequencing technology,

multiple 2nd generation sequencing technologies have emerged (Table 1.1), such as 454 sequencing (<http://www.my454.com>), Solexa/Illumina (<http://www.illumina.com>), SOLiD (<http://www.appliedbiosystems.com>), and Polonator (<http://www.polonator.org>). Each technology had its own advantages and weaknesses. Thanks to the development of those technologies, we have been able to continuously understand the human health based on genetics and dramatically reduce the cost of sequencing from more than 100,000 dollars per genome to a few thousand dollars. These are important for making DNA sequencing more applicable and useful, as it becomes affordable to individuals and allows people to infer more health related indications based on genetic information.

Table 1.1 Popular 2nd generation sequencing technologies in the market

Sequencing platform	Sequencing chemistry	Read length	Template preparation	Application
Roche 454	Pyrosequencing	400bp	Emulsion PCR	WGS and WES of microbes
Illumina Hiseq 3000	Reversible terminator chemistry	2*125bp	Solid phase	Human WGS, WES, RNA-seq and methylation
ABI/Life Tech SOLiD	Sequencing by ligation	2*60bp	Emulsion PCR	Human WGS, WES, RNA-seq and methylation
Polonator	Reversible terminator chemistry	25-55bp	Single molecule	Human WGS, WES, RNA-seq and methylation

* WGS represents Whole Genome Sequencing; WES represents whole exome sequencing

1.1.2 DNA NGS data analysis workflow

In the practice of DNA NGS data analysis (**Figure 1.1**), for each analyzed sample, fastq files with sequencing reads are provided. The first step usually

involves prior sequencing data quality control using software such as FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) to perform sequence trimming and make sure the overall sequencing data quality is sufficient to perform downstream analysis. After that, the reads can be aligned to a human genome reference [4] using alignment tools such as BWA [5] and novoAlign (<http://www.novocraft.com/products/novoalign/>), followed by local reads realignment using GATK [6] and PCR duplicate removal using Picard (<http://sourceforge.net/projects/picard/>) or Samtools [7]. After obtaining a bam file of aligned and filtered reads, further data quality assessments should be performed to evaluate the read duplicate rate, coverage, and coverage uniformity. This step is very useful in evaluating the workability of a sequencing platform and the quality of sequenced samples.

After data quality assessment, multiple types of genomic alterations such as single nucleotide variant (SNV), small insertions and deletions (indels), structural variants (SV) and copy number alterations (CNA) could be investigated. There have been multiple algorithms developed for each type of genomic alteration study, for example, VarScan2 [8], GATK [6] and Mutect [9] are the most popular SNV detection tools and achieve high accuracy; GATK [6], Pindel [10] and Scalpel [11] are well known in detecting indels; BreakDancer [12], DELLY [13] and novoBreak are capable of detecting SVs; ExomeCNV [14], EXCAVATOR [15], and CONTRA [16] are well known algorithms for detecting CNAs. One challenge of using these tools is that none of them generally performs best for all sequencing platforms or types of sequencing data, therefore, given the specific data type, a detailed

comparison should be performed to make sure the tool achieves a good performance on the given data.

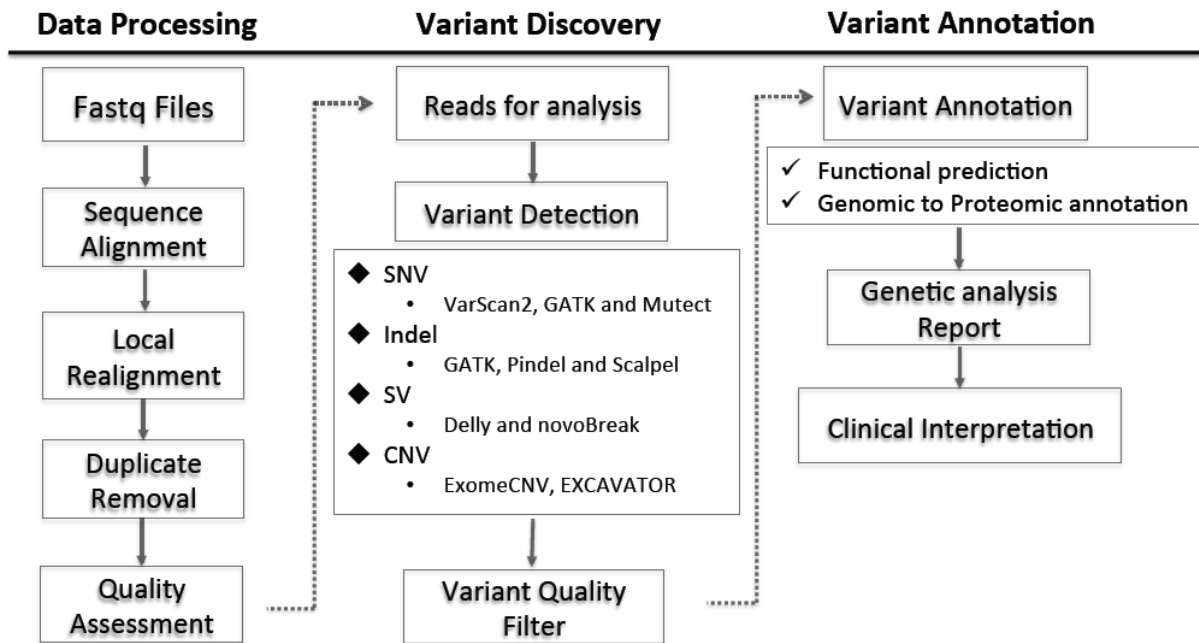


Figure 1.1 A pipeline overview of DNA next generation sequencing data analysis

After detection of genomic alterations, one critical step is to annotate the mutations from the genomic level to the protein level. There have been many tools developed to perform such mutation annotation, such as ANNOVAR [17], Variant effect predictor (VEP) [18], SnpEff [19] and Oncotator [20]. Each tool was implemented with different annotation roles as defined by the developers and preferentially uses different transcript databases among Ensembl [21], Refseq [22], UCSC [23], GENCODE [24], etc. Another important step is to characterize functional mutation events that could potentially drive disease development or represent indications of clinical diagnostics and drug treatment. In this step, several driver mutation prediction algorithms were widely used such as CanDrA [25], CHASM [26],

VEST [26], TransFIC (<http://bg.upf.edu/transfic/home>) and MutationAssessor [27] to predict a genomic SNV is a driver mutation or not. In addition, some databases such as myCancerGenome in Vanderbilt University (<http://www.mycancergenome.org>) and the Institute for Personalized Cancer Therapy in MD Anderson Cancer Center (<http://pct.mdanderson.org>) have actively curated the potentially clinically functional and actionable mutations through literature learning and collected protein variants that could potentially be used to indicate clinical therapy responses.

1.1.3 Challenges of DNA clinical sequencing and analysis

With the development of sequencing technologies over the last decade, there are many different types of sequencing which have become routinely used in scientific research, such as whole genome sequencing (WGS), exome sequencing (WES), RNA sequencing and methylation sequencing. In addition, as it comes to measure the applicability of using sequencing data to enhance health management and clinical treatment, many more cost-effective and practical methods have been developed and applied, such as target of DNA sequencing, which significantly enhances the affordability of utilizing sequencing data in clinical management.

There are already a lot of exciting applications of using whole-genome sequencing and exome sequencing to improve the understanding of inheriting familiar diseases. For instance, a direct relationship between a specific gene locus and disease was revealed by studying a family with four siblings that were affected by Charcot-Marie-Tooth disease (a peripheral polyneuropathy) [28]. Furthermore, analyses focusing on individual genomes have also been published previously [29-

32], including the first complete individual high-throughput method [33].

Besides the familiar inherited diseases, cancer is a class of diseases that consist of multiple disease types and could potentially benefit from the application of personalized therapeutic approaches, particularly given the wide spectrum of mutations that must be analyzed and the complexity of cancer-related genome variation: germline susceptibility, somatic single/multiple nucleotide substitutions and small insertion/deletion mutations, copy number variations, and structural variants.

Although it was increasingly recognized that the personal genome profiles obtained from clinical sequencing data can help inform more accurate clinical decision making [33, 34], the implementation of cancer genomic medicine is critically constrained by a lack of precise understanding of the impact of individual somatic mutations on tumor pathophysiology and response to cancer therapy. Currently, multiple challenges still exist to accomplish the goal of personalized therapy. For example, 1) limitations of current sequencing technologies and computational algorithms in accurately characterizing genomic alterations such as sequencing error, inadequate coverage and uneven coverage uniformity, which may cause loss of information that can be used in clinics; 2) Inconsistent annotations among different data sources and tools, which lead to an ambiguous interpretation of mutation consequences in clinical diagnostics and biomarker indications; 3) The lack of ability to distinguish genomic alterations that confer tumorigenesis (i.e. drivers), from those that provide no selective advantage to tumor growth but occur stochastically in cancer development; 4) Inadequate practice of using -omics data to reveal the specific function of different genomic alterations in different diseases and

biological contexts.

1.2 Mutation annotation in NGS data analysis

Downstream interpretation of genomic alterations that are characterized in mutation detection algorithms is a critical component in NGS data analysis, bridging the identification of mutations and the application in determining functional and disease relevant mutations. One fundamental utility of variant annotation is to categorize each variant based on its relationship to coding sequences in the defined genome and see how it may change the coding sequence and then affect the protein level structure.

The coding sequences of the reference genome refer to, generally speaking, the genes. The “gene” comes to refer to a genomic region that produces poly-adenylated mRNAs through transcription and followed-up translation that encodes an expressed protein [7]. To date, our principal understanding of the protein-coding sequences in the human genome is summarized in the set of transcript isoforms we continuously curate and believe to exist. Thus, in the mutation annotation, the annotation depends on the set of transcripts that were given. There are several widely used annotation databases and browsers such as Ensembl [21], Refseq [22], GENCODE [24], and UCSC [23], which independently contain different sets of transcripts that can be used for variant annotation, as well as a wealth of information of many other kinds as well, such as ENCODE [35] data about the function of non-coding regions of the genome. In this way, a transcript set may also include information about non-coding regions in the genome that mainly function by

regulating the expression of coding transcripts.

Variant annotation could be straightforward and unbiased, in the cases that, only one transcript exists for the gene in which the given genomic variant locates or different transcripts exist for the gene but the given genomic variant locates in a position where identical coding sequences are obtained based on different alternative splicing. However, frequently we encounter more complex situations in the annotations. One situation is that a genomic variant could be transcribed by different transcripts and then further translated into proteins with the variant in different relative positions, then the question is about which one transcript to choose to report as the annotation result for the given variant. Another situation is a genomic variant is annotated by different transcript databases and then further translated into proteins with the variant in different relative positions, then the question will be which database to use for the annotation. When a genomic variant is mapped to multiple potential genes, the situation can be even more complex as prioritizing the genes should be performed in addition to choosing a transcript.

Based on the above mentioned facts, different transcripts existing for one particular gene and even different genes share a similar sequence in the DNA level, therefore the usages of different transcript isoforms or annotation tools would likely produce discordant protein variants given an identical genomic variant.

1.3 Mutation annotation ambiguities in translational and functional genomics study

1.3.1 Forward annotation

In a functional genomics study, the biologists frequently want to validate a set of putative functional mutations with the help of bioinformatics analysis. Frequently, people would prefer to nominate the functional mutation candidates based on the variant frequency. This scenario requires a precise estimation of the variant frequency in a mutation cohort or across different mutation cohorts. However, different annotation practices, especially for data from different cohorts, would produce inconsistent annotation results for a genomic variant based on different transcript usages, and therefore underestimate the variant frequency and lose potentially promising candidates for functionally relevant study. For example, chr7:g.55221821G>A was annotated to *EGFR:A244T* based on transcript ENST00000455089, while was annotated to *EGFR:A289T* in COSMIC based on transcript ENST00000275493. *EGFR:A244T* was considered as a hotspot mutation in TCGA; However, *EGFR:A244T* does not even exist in COSMIC because the identical genomic variant was annotated to a totally different protein variant *EGFR:A289T*, which may be confusing when researchers look at these two annotation results separately and would argue that whether *EGFR:A244T* is a truly functional mutation since it was observed in TCGA but not in COSMIC.

Lack of specification can clearly affect the interpretation of variants and may have serious implications for patient management as well as for understanding of the biology. For example, it has been shown that cancer cells that carry *BRAF:V600E* have higher sensitivity to RAF inhibitors (RAF265 and PLX4720) than those with other *BRAF:V600X* variants [36]. If only the amino acid position and the reference amino acid are given (e.g. *BRAF:V600*), a treatment with little likelihood of

benefit for the patient might be selected [37]. Inconsistent annotations can result in discordance among molecular testing results from different laboratories, generating difficulties for clinical decision-making, particularly in the absence of strong decision support. For example, *ABL*:T315I and *ABL*:M351T, which are frequently seen in chronic myeloid leukemia patients, confer resistance to Imatinib treatment [38], but they can also be reported as T334I and M370T, respectively, based on different transcript isoforms. That difference in annotation may mislead the clinical decision-making.

In genetic diagnosis or counseling for germline variants, the mutation annotation ambiguities could also be misleading. For example, chr10:123258036A>C, which is associated with Crouzon Syndrome [39], can be annotated as N550H in *FGFR2b* or N549H in *FGFR2c*. However, it could also be annotated as N432H, N433H, N460H and N461H based on the usage of other transcript isoforms. Those variants have not been shown to be associated with Crouzon Syndrome in any literature or clinical reports. If a variant annotator chooses to report one of the variants not known to be associated with Crouzon Syndrome, a clinical diagnosis or counseling opportunity might be missed.

1.3.2 Reverse annotation

Conversely, different DNA variants such as chr7:55249076_55249077CT>AG and chr7:55242470T>C can result in the same cDNA and protein variant, *EGFR*:p.L747S, which mediates acquired resistance of non-small cell lung cancer to tyrosine kinase inhibitors [40]. However, the multiple options of genomic variants for

the protein variant could lead to another commonly encountered issue, which is uncertain genomic origin for a given protein variant. This exposes an important gap in the clinical genomic validation process, for example, when a biological researcher or a clinician observes *EGFR*:p.L747S in a genetic report as mentioned above, without the information of concrete transcript ID and genomic origin, it is impossible to precisely determine how to construct the cDNA sequence for experimental functional validation.

1.3.3 Limitations of current variant annotation algorithms

Differences among widely used annotation algorithms such as ANNOVAR [17], SnpEff [19] and VEP [18] and among various transcript databases such as Ensembl [41], RefSeq [42] and GENCODE [24], further increase the extent of ambiguity. *Ad hoc* post-annotation filters such as reporting of the variant on the longest transcript may also be a problem because the longest transcript may be different in different systems and, further, may not represent the transcript actually present. Inconsistent use of conventions in annotating indels, i.e., reporting the left-most (left-aligned) or the right-most (right-aligned) position in the context of repetitive sequences can further increase ambiguity. A recent investigation indicated that variable usage of annotation algorithms and/or transcript databases may cause greater than 50% discrepancy in annotating loss-of-function variants identified from genomic sequencing [43]. Our investigation of the COSMIC database reveals that 92,444 out of 1,010,316 (9.1%) of somatic DNA variants have been reported more than once at different locations on proteins based on differences in isoform.

Despite such significant challenges in ongoing research and clinical practice, we have not had an annotation tool that is capable of quantifying the extent of such ambiguities and resolving them in a systematic way. Existing tools (e.g., ANNOVAR, snpEff and VEP) perform what we call “forward annotation”, which maps a variant characterized at the genomic level to a set of cDNA or protein isoforms. No tool, to the best of our knowledge, has general capability for what we call “reverse annotation,” which reverse-maps a variant at a cDNA or protein level to the genome. A previous algorithm, Mutalyzer [44], can reverse-annotate variants at the cDNA level, but it has limited functionality: it annotates only single nucleotide variants; it does not allow input of a gene name, nor offering analysis of variants at the protein level. Without a fully automated reverse annotation tool, translation from a functional protein site (such as Y308/S473 in *AKT* and Y1068/Y1172 in *EGFR*) to a genomic identifier would involve a tedious manual process that is not scalable.

1.4 Cancer gene prediction

With the increasingly used NGS data in biomedical research, a hot topic has been cancer gene identifications. The main goal was to find gene candidates that could drive the cancer development, promote cancer cell proliferation or enhance the cancer cell viability. To achieve this goal, there have been several methods that were developed in the past few years to predict driver genes. Essentially, different methods have largely diverse assumptions for defining driver genes and differences in the results reflect the differences in the methodologies.

One assumption was to consider all significantly mutated genes across the

cancer genome as driver genes. Motivated by that, a gene is nominated as a driver if it contains significantly more mutations than expected from a null background model [45, 46]. A variety of practical algorithms have been developed in the context of large-scale cancer genome sequencing, differing mainly in how they model background mutations. For example, MuSiC [47] assumes a homogenous background mutation rate across all genes. In each individual gene, a binomial model is used to calculate the significance level (p value) of each mutation subtype, then Fisher's method is applied to combine the p values across different mutation subtypes and come up with an unified p value to indicate the significance level of the investigated gene. With realizations that the background mutation rates were not uniform across all the genes, MutSigCV [48] models the heterogeneous background mutation rate for each gene–patient–category combination based on the observed silent mutations in the specifically investigated gene and non-coding mutations in the surrounding chromosomal regions. Because in most cases these data are too sparse to estimate an accurate background mutation rate, the method tries to increase the accuracy by pooling data from other genes with similar properties (for example, replication time, expression level). Furthermore, the method identifies the significantly mutated genes by incorporating factors such as di-nucleotide sequence context, cancer type and epigenetic elements.

The other assumption was that driver genes tend to have highly recurrent mutations that are enriched in clusters. Some algorithms have also been developed to nominate driver genes using cluster-based methods. For example, OncodriveCLUST [49] estimates a background model from coding-silent mutations

and tests whether the mutations on each individual site are significantly mutated, after that, the method arbitrarily chooses to combine mutational hotspots that are within 5 amino acids of each other, and detects clusters of missense mutations that are significantly mutated and therefore where driver genes lie. Essentially, those clusters would be likely to contain mutations that can alter the protein structure and thus affect the gene function during cancer development. E-Driver [50] exploits the internal distribution of missense mutations between different proteins' functional regions (for example, functional domains or intrinsically disordered regions) to nominate clusters of missense mutations in protein-protein interaction (PPI) interfaces that show a bias in their mutation rate as compared with other regions of the same protein, providing evidence of positive selection and suggesting that these proteins may be actual cancer drivers.

However, increasingly many studies indicate that a mutation may have substantially different functions at different amino acid positions in the same gene [51, 52] and may be associated with different clinical utilities in different disease and biological contexts [53, 54]. Similarly, not every mutation within a cancer gene can be assumed to have equal function in one cancer type or across different cancer types. In addition, the previous studies that focused on identifying significantly mutated genes mostly ignored potential functional mutations in infrequently mutated genes, and in under-investigated mutation types such as insertions and deletions.

1.5 Driver mutation prediction

To characterize the function of individual mutations, multiple computational

tools such as TransFIC (<http://bg.upf.edu/transfic/home>) [55], CHASM [26], Condel [56], MutationAssessor [27] and CanDrA [25] have been developed. Essentially, CHASM uses a random forest classifier and incorporates 49 features to rank the investigated mutations with different probabilities of being driver (the lower p-value indicates higher driver potential). In terms of the training data, CHASM uses manually curated oncogenic functional mutations as true drivers from breast, colorectal, and pancreatic tumor resequencing studies [46, 57] and the COSMIC database [58], and uses synthetic passenger mutations that are generated by sampling from eight multinomial distributions that depend on dinucleotide context and tumor type [26]. CanDrA uses a support vector machine classifier and incorporates 95 different features to classify the investigated mutations to be either driver or passenger in a tumor type-specific manner. In terms of the training data, CanDrA defines a driver mutation as one that is observed in at least two different samples, from either TCGA or COSMIC. Compared to CHASM and CanDrA, other methods such as Condel, MutationAssessor, Polyphen2 [59] and SIFT [60], tend to use a scoring system to predict the functional impact of a mutation. For example, Condel uses a weighted average of the normalized scores that are integrated from five different tools including Logre, MAPP, MutationAssessor, Polyphen2 and SIFT. TransFIC takes as input the Functional Impact Score of a somatic mutation observed in cancer provided by MutationAssessor, Polyphen2 and SIFT. It then compares that score to the distribution of scores of germline SNVs observed in genes with similar functional annotations (for instance genes with the same molecular function as provided by the Gene Ontologies), and eventually reports a Z-

score to indicate the tolerance of the mutation.

Although the algorithms have been actively used in identifying functional mutations from multiple aspects and contributed significantly to the cancer research, several limitations have been frequently observed: 1) Many algorithms are designed to tackle a generic problem about the function of a driver, but do not account for cancer type specificity and ignore the potential functional heterogeneity of a mutation in different cancers; 2) Many machine learning-based algorithms simply select as positive training data mutations in known cancer genes, which may significantly bias the assessment of mutations in previously unknown non-cancer genes, while it is unfair to assume that the driver mutations only occur in known cancer genes; 3) Previous studies generally lack adequate functional assessment, which usually involves interrogation of RNA expression, protein activity and drug response data, to investigate the biological and therapeutic relevance of predicted mutations; 4) All the algorithms predict a large number of driver mutations (most are non-recurrent) and thus significantly sacrifice specificity, which is impractical for thorough functional genomic validations and individual evaluations using currently available functional data.

Since the potential driver mutations are supposed to be defined under a specific disease context, a driver mutation prediction algorithm that does not take into consideration disease-specific factors such as cancer type, disease stage, mutation prevalence, mutation spectrum, and other clinical characteristics, may not be accurate enough. As a result, these functional predictions often disagree with each other and are not accurate enough for practical use. The number of clinically

actionable mutations, which could potentially be used as training data, will likely remain in the hundreds (currently 285 in MyCancerGenome.org and 269 in PersonalizedCancerTherapy.org). Therefore, it is critical to improve predictive approaches that do not completely rely on known actionable mutations, but are capable of accurately predicting driver mutations.

1.6 Application of high-dimensional -omics data in translational genomics

With the accumulating knowledge of cancer genome, many genetic variants have been shown to affect tumor expression architecture through the transcriptional regulation either in -cis or in -trans manner [61, 62]. The effort to map genetic variation to specific expression quantitative trait loci (eQTL) has illustrated the value of transcriptomics data in the functional interpretation of genetic variants [63]. Thus, the transcriptomics data could serve as a unique resource to evaluate the impact of individual genomic mutations on the expression level. In addition to transcriptomics data such as RNA sequencing or microarray data, in the past few years, proteomics data such as Mass Spectrometry [64] or Reverse Phase Protein Array (RPPA) [65] has become one of most promising data types for investigating biological activity at the gene/pathway level, as well as the effects of post-translational modification. Therefore, in terms of investigating the genome-wide effect of genomic alterations, proteomics data will serve as a novel data type.

Cancer pharmacogenomics studies have recently become an important way for discovering the molecular determinants of drug response, thus representing the potential benefits of personalized cancer treatment. Multiple works such as seminal

work on the NCI-60 cancer cell lines [66, 67] and other subsequent research efforts [68, 69] highlighted specific genetic alterations as drug targets or biomarkers of drug response. For example, *BRAF*:V600E and *NRAS* mutations have been shown to increase cancer cell sensitivity to MEK inhibitors (such as AZD6244) [53, 70]. However, it was also elucidated that cells with *BRAF*:V600E mutations are sensitive to RAF inhibitors in melanoma but not in colon cancer [70], which suggested the disease specificity of the biomarker indication. The emerging large-scale pharmacogenomics datasets such as CCLE [71] and GDSC [72] have been designed to facilitate an increased understanding of the molecular features that influence drug response in cancer cells and will enable the design of improved cancer therapies.

1.7 Impact of variant annotation on hotspot mutation prediction

Regarding hotspot mutations, it is common to make predictions using either genomic or protein variants. Many computational algorithms have tried to predict the impact of functional mutations on the genomic level [25, 26, 60]. Compared to using genomic variants, one advantage of using protein variants is to substantially increase the statistical power, since different genomic variants could happen at one identical protein amino acid position and actually represent very similar biological functions [73, 74]. For hotspot mutation prediction that relies on the amino acid residue information, it is important to have a correct representation of the protein variant frequency within a defined cohort or across different cohorts. Fundamentally, the protein variant was inferred from the genomic variant that was detected in each sample within a specific cohort, therefore, it is important to know that whether an

observed protein variant was from one single genomic variant and whether a genomic variant was annotated to a uniform protein variant.

The variant annotation process from the genomic to protein level, is not an absolute one-to-one correspondence. Ambiguities may come from: 1) different annotation rules that are implemented; 2) different transcript databases that are used; 3) different transcript isoforms that are used within one identical transcript database. Due to those concerns, one genomic variant is frequently annotated to various protein variants in different practices. This means if we solely use the protein residue information to identify the functional variants such as hotspot mutations, the frequency of protein residues could be largely underestimated and the function of a specific protein residue may be misinterpreted because of incorrect frequency calculations.

1.8 Motivations and Rationale of the thesis study

Next generation sequencing has been recognized as an effective approach to enhance the accuracy and efficiency of cancer patient management. Based on the massive information that is obtained from analyzing such high dimensional data, we could potentially achieve the goals of personalized therapy. Through the investigations since the emergence of NGS techniques, the accuracy of applying the NGS data in clinics has been largely constrained by many different aspects, such as clinical detection, annotation and interpretation of the variants that were observed in clinical sequencing data.

A large amount of ambiguity exists in current mutation annotations, which could

potentially prevent the precise identification of functional mutations, biomarkers, and target therapies in basic research and clinical practice. For example, one genomic variant can be annotated to different protein variants based on different transcript isoform usages within one dataset or across different datasets, which could lead to significant underestimation of targeted protein variant frequency and misinterpretation of the functional impact of the targeted protein variant. Due to the annotation ambiguities, the genomic variant that corresponds to a protein biomarker for a drug treatment indication could be annotated to another different protein variant and lead to incorrect clinical decision-making. Given these problematic practices in research and in clinics, it is important to systematically investigate the existence of annotation ambiguities in the commonly used mutation datasets such as COSMIC and TCGA. In addition, development of a method that enables cross-level variant annotations, would allow the researchers and clinicians to fully capture the potential genomic origins of an observed protein variant and minimize the ambiguities of using genomic information to make a decision. Motivated by these considerations, we developed a novel variant annotator, TransVar, which performs multi-level variant annotation such as forward annotation from genomic to RNA and to protein level, and reverse annotation from protein to RNA and to genomic level. The novel reverse and equivalence annotation function of TransVar could potentially contribute in: 1) experimental validation design; 2) clinical pharmacogenomics; and 3) hotspot mutation prediction.

Another critical challenge of oncogenomics and pharmacogenomics is to distinguish genomic alterations that confer tumorigenesis (i.e. drivers), from those

that provide no selective advantage to tumor growth but occur stochastically in cancer development. Previously many driver gene algorithms have been proposed to distinguish cancer related genes that could promote cancer development. However, more recently researches found that 1) it is incorrect to assume equal function of different mutations within one cancer gene in all cancer types. For example, one mutation such as *BRAF*:V600E, can have different functional indications in different cancer types. In *BRAF*, quite a lot of mutations do not actually have clear functional indications in many cancer types; 2) it is unfair to simply focus on significantly mutated genes while ignoring infrequently mutated genes. More and more studies indicate now that infrequently mutated genes could also be functional in certain cancer types. After that, some driver mutation algorithms were proposed to distinguish driver mutations from passengers. However, most of the methods still assume the same function of a variant in different cancer types; they did not consider using additional functional data to justify the performance of the driver mutation prediction, and they do not assess the function of a variant in a specific biological context. Therefore, people gradually realized that it is important to obtain accurate biological and therapeutic interpretations of a mutation in a tumor type-specific manner to help improve the efficacy of using genomics information in clinical applications. With those motivations, we systematically identified tumor type-specific hotspot mutations in 17 tumor types, and analyzed the potential impact of hotspot mutations by performing genome-wide and population-based analysis across different tumor types and assessing functionality using transcriptomics, proteomics and pharmacogenomics data.

CHAPTER 2

TransVar: a multi-level variant annotator for precision genomics

(Most of the methods and results in this chapter have been published online in Nature Methods, November 2015: Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A Rohrdanz, James M Melott, Chris Wakefield, Jia Zeng, John N Weinstein, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, “TransVar: a multilevel variant annotator for precision genomics”. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)

2.1 Materials and Methods

2.1.1 COSMIC and TCGA somatic mutation data

We downloaded the COSMIC somatic mutation dataset version 67 for our study. As introduced in the COSMIC data source, this total mutation set includes many sources of curated mutation data. In terms of TCGA mutations, we downloaded TCGA pan-cancer level-3 somatic mutation data from Synapse (<https://www.synapse.org/#!Synapse:syn300013>) as last updated in November 2014.

2.1.2 Transcriptome definitions

For human genome reference, TransVar supports hg18, hg19 and hg38. TransVar supports transcriptome definitions in 1) UCSC knownGene [75], 2) RefGene built from the UCSC table browser [76], 3) Ensembl annotation [41] in General Transfer Format (GTF), 4) RefSeq annotation [42] in General Feature Format version 3 (GFF3), 5) Consensus Coding Sequence (CCDS) [77], 6) GENCODE [24] release 19, and 7) AceView [78]. TransVar not only supports the transcript annotation in human, but also supports which in mouse. When using TransVar, the users can use any one or a combination of different transcript sources or else user-provided definitions (such as ncRNA and miRNA databases, as long as they are based on the same genome reference assembly).

2.1.3 Reverse annotation

TransVar is designed for reverse annotation of four categories of mutations at

the mRNA level (single-nucleotide substitution, insertion, deletion and block substitution) and four categories of aberrations at the protein level (single-amino acid substitution, insertion, deletion and frame-shift). The input format of TransVar was designed to follow the Human Genome Variation Society (HGVS) nomenclature (<http://www.hgvs.org/mutnomen/>) [44]. Small formatting variation from HGVS is allowed to accommodate non-standard identifiers frequently seen in the literature (as illustrated by examples on the web site, www.transvar.net). The output of TransVar strictly follows the HGVS nomenclature. For each input variant, based on the gene name and the correspondingly available transcript information in the defined databases, TransVar iterates through all of the associated transcripts and infers the relative coordinates on each transcript based on the genomic coordinates of the coding sequence defined for each transcript. TransVar builds in multiple checkpoints to restrict the search scope of valid transcripts. That filter takes into account: 1) the length of the transcript, 2) the sequence of the optionally provided reference transcript or isoform, 3) exon boundaries in the transcript, and 4) any transcript identifiers provided. For protein-level variants, TransVar provides parsimonious inference of nucleotide changes that could best explain the observed amino acid change. Taking single-amino acid substitution as an example, TransVar iterates over all possible target codon sequences to identify a set of most likely base changes that minimize the distance between the altered codon and the reference sequence in each transcript. Meanwhile, TransVar outputs all candidate variant identifiers to a report that informs the user of all possible sourcing variant annotations.

For insertions and deletions, TransVar aligns the alternative indel sequences to

the reference sequence when generating identifiers at the mRNA and protein levels, conforming to the HGVS nomenclature. TransVar uses a walk-and-roll strategy for the realignment to ensure accurate positioning of the resulting identifiers despite the presence of repeats, intron splicing, and redundant codon usage. Contrary to the HGVS specification---3'-alignment (or right-alignment or C-terminal-alignment under the protein representation) of indels---a commonly used rule in the genomics community is to 5'-align (or left-align or N-terminal-align) indels. Thus, TransVar also provides a left-aligned identifier in the output to allow users to reference annotations created under other rules. For frame-shift variants, TransVar iterates from the reported location all possible single and double nucleotide insertions and deletions and reports those that result in the corresponding frame-shift at the protein level, as delimited by the amino acid sequence between the first altered amino acid and the first stop-codon.

2.1.4 Forward annotation

Forward annotation starts from a genomic location. The input follows HGVS nomenclature (<http://www.hgvs.org/mutnomen/>) [44] and small formatting variations are allowed (as illustrated in the examples on the web site). The output strictly follows HGVS nomenclature. TransVar hashes transcript definitions based on each transcript's start and end positions and retrieves all of the isoform definitions that overlap with the genomic coordinate of the input identifier. If the input variant falls within the span of an exon, TransVar reports the consequence of the variant and generates variant identifiers at both the mRNA and protein levels. TransVar also reports mRNA identifiers for variants that overlap intronic or UTR regions. Additional

sequence features can optionally be hashed to allow more extended annotation of regulatory elements. Those features are similar in concept to ones in existing annotators such as ANNOVAR, snpEff and VEP.

The input to TransVar can use a specific variant (e.g., chr6:g.4241214_4241218del) or simply a genomic interval (e.g., chr6:g.4241214_4241218) without specifying the reference and alternative alleles. For example, chr6:g.4241214A>T denotes a single-nucleotide substitution, whereas omitting the alternative allele T indicates a single genomic position. Both forms are valid inputs to TransVar. The annotation of genomic position/interval is of great utility for understanding the potential translational consequences of indels and structural variations (SVs) if the corresponding breakpoints on the transcripts can be revealed. For example, for chr3:g.178936091_178936192, a 102-bp interval, one can use TransVar to show that it encodes *PIK3CA*:p.E545_R555, which begins in the coding sequence of *PIK3CA* exon 10 and ends in the intron between exon 10 and 11, covering codons 1633 to 1664 and 70 intron bases. That variation may introduce not just an indel but also a novel splicing, resulting in complex mRNA and protein products. Given a long altered genomic interval, TransVar can also reveal which genes are contained within the interval.

2.1.5 Equivalence annotation

TransVar automates the search for alternative codon identifiers that are potentially of the same genomic origin, a task we call “equivalence annotation.” Two codon identifiers such as *MET*:p.T1010 and *MET*:p.T992 are considered equivalent

because they can be translated from the same genomic variant chr7:g.116411990C>T, based on different isoform definitions. The equivalence annotation functionality can be used to ascertain the functional or clinical interpretation of important variants and also provide more accurate estimation of variant frequency in a disease cohort. For a protein level identifier, TransVar performs reverse annotation and then forward annotation to map the resulting genomic variants back to protein level. Both steps may result in multiple valid candidate identifiers that are equivalent to the original variant. All of the equivalent identifiers are reported in the TransVar output. By providing specific transcript identifiers from different databases, TransVar allows the user to map a protein-level variant annotated using one of the databases (e.g., RefSeq) to protein-level variants annotated using a different database (e.g., Ensembl). That type of analysis can be particularly important for decision support in patient management. For example, *MET*:p.T1010 is generally designated as an activating germline SNP, whereas *MET*:p.T992 is not annotated in most decision-support algorithms [79].

TransVar maximally uses available information such as reference sequence to reduce the number of equivalent identifiers. For example, for *ABL1*:p.255, TransVar finds 6 equivalent codon identifiers (p.236, p.237, p.254, p.256, p.273, and p.274) in the RefGene database. When the reference amino acid (E) is provided, only 2 codons remain (E236 and E274), instead of 6. Including the reference amino acid can substantially reduce the number of equivalent identifiers. For example, for a set of 1821 hotspot mutations in COSMIC, specifying the reference amino acid reduces the number of mutations with non-unique codon identifiers from 1260 (69.19%) to

1021 (56.07%) based on Ensembl v75.

2.1.6 Tool Implementation

TransVar provides reverse, forward and equivalence annotation algorithms that can be accessed via a command line utility, a programmable Python API, or a web interface (www.transvar.net). The web interface uses application programming interface (API) calls in the common gateway interface (CGI) and allows use of TransVar without writing code or issuing commands in a terminal. The user can either type in the variant identifiers or upload a file for batch processing. TransVar can optionally load the coordinates of the entire transcriptome definition into memory but reads in the corresponding reference transcript/protein sequences only when necessary. That procedure limits the memory footprint to the size of the transcriptome definition. For the web interface, the transcript definitions are stored in disk-indexed database tables. Gene names (in the case of reverse annotation) and genomic locations (in the case of forward annotation) are indexed to facilitate quick look-up. TransVar can map directly to different versions of human genome assemblies. Choices of reference assemblies and transcript definitions are provided as options in the web-form submission. Each line characterizes the annotation based on one specified transcript definition. If no valid gene name matches or if no transcript definition matches, a warning message is provided in the last field of the output.

2.2 Results

2.2.1 Overview of the functions of TransVar

To facilitate standardization and reveal inconsistency in existing variant annotations, we have designed a novel variant annotator, TransVar, to perform three main functions supporting diverse reference genomes and transcript databases (Figure 2.1): (i) “forward annotation”, which annotates all potential effects of a genomic variant on mRNAs and proteins; (ii) “reverse annotation”, which traces an mRNA or protein variant to all potential genomic origins; and (iii) “equivalence annotation”, which, for a given protein variant, searches for alternative protein variants that have identical genomic origin but are represented based on different isoforms.

Essentially, uncertainty frequently exists in mutation annotation. One DNA sequence with a mutation could be transcribed into different transcript isoforms based on alternative splicing, and therefore be further translated into proteins with the mutation on different relative positions (Figure 2.1 Upper). For example, chr14:g.105239423C>T (hg19) can be forwardly annotated to coding mutations AKT1: E17K, AKT1: E322K or non-coding mutation AKT1:intronic. Compared to forward annotation, reverse annotation from the protein level to genomic level is a novel concept. One protein with a variant could potentially come from multiple transcripts with the mutations on various locations, and multiple genomic variants are responsible for each of the corresponding transcript (Figure 2.1 Lower). For example, EGFR: p. L747S (hg19) can be reversely annotated to

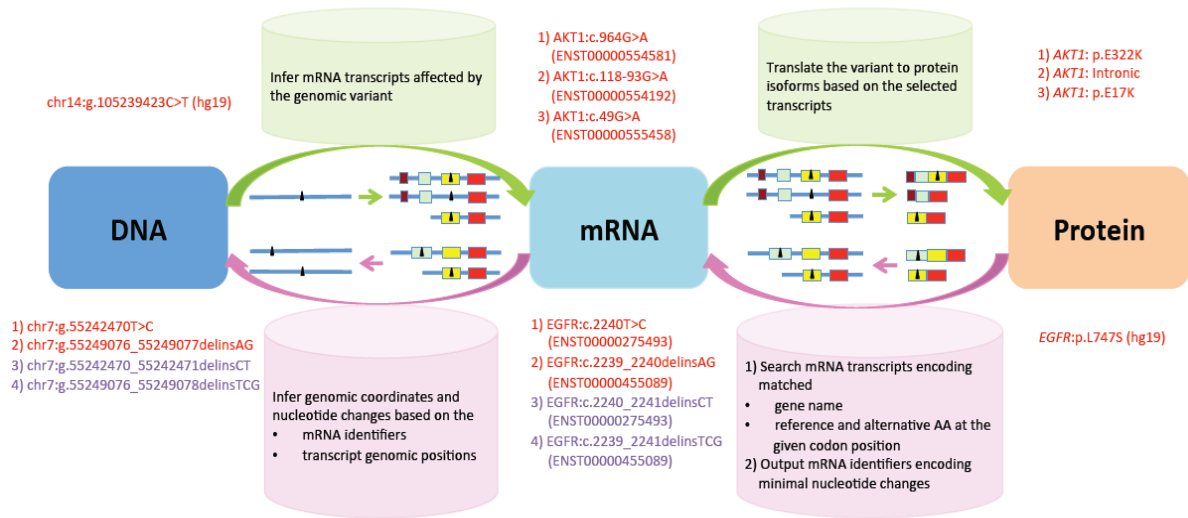


Figure 2.1 Overview of TransVar. TransVar performs forward (green arrows) and reverse annotation (pink arrows) and considers all possible mRNA transcripts or protein isoforms available in user-specified reference genome and transcript databases (colored boxes representing exons in various transcripts or isoforms of a gene). Given a variant (black triangle) at any of the genomic, mRNA or protein levels, TransVar is able to infer the associated variants at the other two levels. In reverse annotation, TransVar searches all potential transcripts and reports one variant on each transcript. When there are multiple variants on the same transcript, TransVar reports the variant with minimal nucleotide changes (red text) instead of other *alternatives* (purple text). (Figure reprinted from *TransVar: a multilevel variant annotator for precision genomics*. Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A Rohrdanz, James M Melott, Chris Wakefield, Jia Zeng, John N Weinstein, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, *Nature Methods*, 2015. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)

chr7:g.55242470T>C and chr7:g.55249076_55249077delinsAG. In addition, multiple genomic alterations on an identical transcript may result in a similar protein variant.

In the reverse annotation practice of TransVar, When there are multiple variants on the same transcript, TransVar reports the variant with minimal nucleotide changes chr7:g.55242470T>C and chr7:g.55249076_55249077delinsAG, instead of other alternatives such as chr7:g.55242470_55242471delinsCT, and chr7:g.55249076_55249078delinsTCG.

2.2.2 Forward annotation of COSMIC mutations using TransVar, ANNOVAR, VEP, snpEff and Oncotator

To illustrate the degree of inconsistency in existing variant data and evaluate TransVar's accuracy in performing comprehensive annotation, we conducted forward annotation on COSMIC mutation data v67, using TransVar and several widely used variant annotators, ANNOVAR [17], VEP [18], snpEff [19], and Oncotator [80], and asked whether the resulting protein identifiers (gene name, protein coordinates, and reference amino acid (AA)) match those in COSMIC.

We downloaded 964,132 unique single-nucleotide substitutions (SNSs) such as chr1:g.87369101C>A, 3,715 multi-nucleotide substitutions (MNSs) such as chr10:g.52595929_52595930delinsAA, 11,761 in-frame and frame-shift insertions such as chr2:g.69741762_69741762insTGC and chr12:g.9021138_9021138insG, 24,595 in-frame and frame-shift deletions such as chr3:g.137843433_137843435del and chr19:g.58863869_58863869del, and 166 block substitutions (BLSs) from the catalogue of somatic mutations in cancer (COSMIC v67). BLS represents in-frame

or frame-shift replacements that potentially incorporate both insertion and deletion in one event, such as chr5:g.112175419_112175425delinsGCA and chr7:g.55249018delinsAACCCCT.

We used the genomic and corresponding protein variants reported in COSMIC as the ground truth for this comparison. The ground truth here was clearly subject to any annotation biases when variants were submitted to COSMIC. Differences from the ground truth largely indicate inconsistency among annotators or rules used (instead of correctness in any absolute sense). Although TransVar can jointly utilize multiple transcript databases in annotation, we tried to use a consistent one for our comparison in Table S2.1. Specifically, TransVar (version 2.1.15.20150827), ANNOVAR (released on 2015Mar22), VEP (version 81) and snpEff (version 4.1) used Ensembl v75, while Oncotator (version 1.5.1.0) used GENCODE v19. Minor differences may exist among the instances of databases used by these tools. We used the default settings of each tool. We considered one genomic variant as being annotated consistently with COSMIC if the results reported by an algorithm contain the corresponding entry in COSMIC, as each algorithm may output multiple annotation entries for a given genomic variant, due to alternative isoform usages. For all types of variants, we required that the gene name, protein coordinate and reference amino acid match exactly with those in COSMIC.

For SNSs, ANNOVAR, VEP, Oncotator and snpEff achieved similar results with around 92% consistencies. TransVar achieved slightly higher consistency at 96% (Figure 2.2 and Table S2.1). Since the forward annotation algorithm in TransVar did not have major differences from others, such difference in consistency may be

attributed to minor differences in the databases used or other implementation details. The small percentage (4%) of SNSs that TransVar failed to annotate consistently was due mostly to invalid or imprecise use of gene names (e.g., *ANXA8* for *ANXA8L2*) or missing matched transcript definitions (e.g., unmatched reference alleles at specified coordinates) in Ensembl v75. Thus, TransVar's forward annotation algorithm might have achieved the best possible result in this experiment. In addition, we found that the consistency of TransVar dropped to 68.8% if only the longest transcripts in Ensembl v75 were selected. This indicated the importance of considering all available transcripts in performing annotation to avoid inaccurate interpretation.

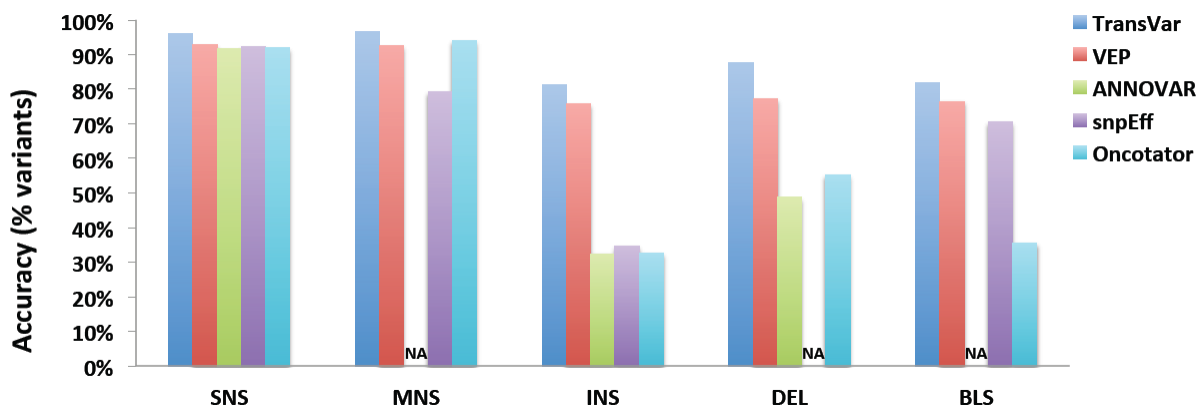


Figure 2.2 Comparison of forward annotation consistency among TransVar, VEP, ANNOVAR, snpEff and Oncotator. Plotted are percentages of variants (Y axis) that had matched protein annotations in COSMIC v67 based on 964,132 unique SNSs, 3,715 MNSs, 11,761 INSs, 24,595 DELs and 166 BLSs (X axis). NA: Protein level annotations not available. (Figure reprinted from *TransVar: a multilevel variant annotator for precision genomics*. Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A Rohrdanz, James M Melott, Chris Wakefield, Jia Zeng, John N Weinstein, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, *Nature Methods*, 2015. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)

TransVar achieved higher consistency than ANNOVAR, snpEff, Oncotator and VEP in annotating insertions, deletions and BLSs, due mainly to its more comprehensive support of different indel reporting rules (Figure 2.2 and Table S2.1). Most indel variants are reported at 3'-aligned amino acid (AA) positions in COSMIC, conforming to the HGVS conventions. However, ANNOVAR, snpEff and Oncotator reported them at only the 5'-aligned AA positions, resulting in relatively large extents of inconsistency. VEP achieved better consistency because it reported indels at the 3'-aligned AA positions. TransVar achieved the highest consistency because it reported not only 3'-aligned AA positions but also 5'-aligned, as well as unshifted AA positions as alternatives. Interestingly, there are some variants for which none of the annotators (including TransVar) could produce annotation consistent with COSMIC. For example, chr11:g.32417943_32417943delinsGGG, which was annotated as *WT1*:p.302 in COSMIC, was consistently annotated as *WT1*:p.121, p.141, p.158, or p.370 by TransVar, snpEff, Oncotator and VEP, suggesting some transcripts that generated the original annotations in COSMIC have become obsolete in Ensembl v75 or GENCODE v19.

The above findings of the annotation inconsistencies between TransVar and other variant annotators can largely be attributed to a lack of standardization among variant annotations (codon or AA positions of variants) submitted to COSMIC and among conventions implemented in various annotators. Inconsistency in annotations blurred the lines of evidence for variant frequency estimation and led to inaccurate determination of variant function. TransVar revealed hidden inconsistency in these variant annotations by comprehensively outputting alternative annotations in all

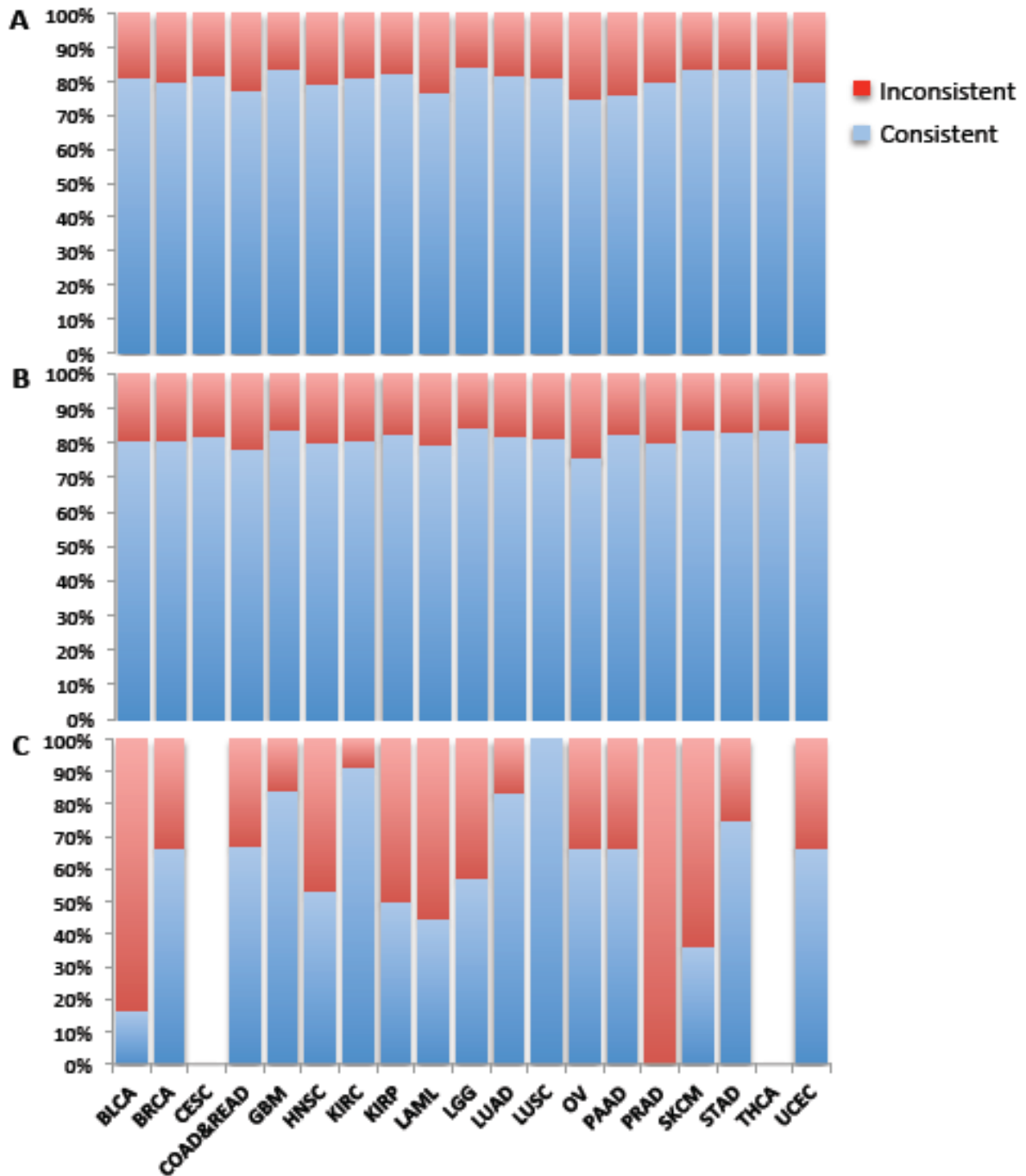


Figure 2.3 Comparison of forward annotation consistency between COSMIC and TCGA mutation data using TransVar. The datasets were investigated in several different ways: (A) using all the available mutations, (B) using all the point mutations (including missense, nonsense and silent mutations), and (C) using all the indel mutations.

available transcripts in standard HGVS nomenclature, and thus resulted in greater consistency in this experiment.

2.2.3 Forward annotation ambiguities in TCGA and COSMIC mutation data

Beside the annotation ambiguities that exist when using different transcript databases or different database versions, we also found a large inconsistency of annotation among various data sources such as TCGA and COSMIC. To specifically investigate the forward annotation inconsistency between TCGA and COSMIC, we thoroughly compared the annotation datasets from TCGA and COSMIC in 19 cancer types. In Figure 2.3A, in the majority of the investigated cancer types, the inconsistency rate was above 10%, which means more than 10% of the genomic variants shared by TCGA and COSMIC were actually annotated to different protein variants. To further investigate the ambiguities from different types of mutations, we dissect the mutation sets into SNV and indels.

In Figure 2.3B, we found most of the inconsistency rates of SNV annotation were a bit lower than 10%, while in Figure 2.3C, we found most of the inconsistency rates of indel annotation were higher than 10%. The results indicated that indels overall have a higher annotation ambiguity as compared to SNVs. As we explained above, this might be due to the more complex rules implemented when annotating indels, such as alignment options.

2.2.4 Forward annotation of RNA-editing sites by TransVar

Recent studies indicate that RNA-editing contributes to human disease,

including cancer [81, 82]. RNA-editing sites are typically discovered by comparing matched RNA-seq and DNA-seq reads based on their alignment to a common genomic reference [83]. However, the functional interpretation of resulting RNA-editing sites may be ambiguous due to lack of clarity about the isoform used. For example, an RNA-editing site in the coding region of one isoform may be in a UTR region of another. To quantify the extent of such ambiguity in current RNA-editing studies, we downloaded a set of 1,379,403 curated A-to-I RNA-editing sites [83]. Using TransVar, based on the human reference assembly GRCh37 and Ensembl v75 transcript database, we found that 401,146 (29.8%) sites could affect different regions on different isoforms. For example, chr12:g.69237552 could affect the coding region of exon 1 or exon 2, the 3'-UTR, or the intronic region between exons 5 and 6 of various isoforms of *MDM2*. TransVar also revealed many sites such as chr3:g.126299981 and chr6:g.43585896 that could be annotated as either in an intronic region or in the UTR, whereas the original annotation reported only one of the possibilities [83].

2.2.5 Reverse annotation accuracy for SNV, indels and frame-shift variants

TransVar's novel reverse annotation can be used to ascertain if two protein variants have an identical genomic origin, thus reducing inconsistency in annotation data. It can also reveal whether or not a protein variant has non-unique genomic origins and requires caution in genetic and clinical interpretation. We evaluated TransVar's reverse annotation accuracies of single amino acid substitutions (SASs), amino acid insertions (INSPs), deletions (DELPs) and frame-shift substitutions (FSPs) in various databases.

For SASs, given a protein variant, the accuracy was calculated based on whether the position of the genomic variants inferred by TransVar contains the one that was reported in COSMIC. Using the Ensembl database v75 (Table 2.1), we found 91.8% of the SASs protein variants in COSMIC can be accurately traced back to the genomic variants by using TransVar's reverse annotation. Among those SASs, only 79.8% of the SASs protein variants can be uniquely traced back to the genomic variants by using TransVar reverse annotation, the remaining 12% have multiple genomic mapping positions because of different transcript isoform usages. We also tried to estimate the reverse annotation accuracies using the RefSeq v105 and CCDS v37.3 databases, the consistencies and uniqueness of reverse annotation were different from what has been observed using Ensembl database v75 (Table 2.1). In addition, we tried to combine the above-mentioned three databases, and found that the consistency was improved to 94.7% while the uniqueness was dropped to 74.7% due to redundant transcript isoforms that are used in different transcript databases.

For INSPs, DELPs and FSPs, We regarded that a reverse annotation reported by TransVar was consistent if the genomic start position of the annotated variant was within 5 bp of the original genomic start position in COSMIC. Uniquely mapping INSP, DELP and FSP to genomic coordinates was much more challenging than SASs and sometime impossible due to repeats, which resulted in lower consistency in these variant types (Table 2.1). For example, an insertion of a glutamine (Q) into any location in a five-glutamine peptide usually results in a notation of the insertion before the first glutamine following the reporting rule of 5'-alignment (Figure S2.1).

Table 2.1 Reverse annotation consistency of COSMIC protein identifiers via different transcript databases

		Ensembl	RefSeq	CCDS	Merged*
SAS	Uniquely	79.8%	81.0%	85.8%	74.2%
	Consistently	91.8%	88.6%	88.0%	94.7%
INSP	Uniquely	74.4%	74.7%	75.6%	73.8%
	Consistently	75.6%	75.7%	76.0%	75.8%
DELP	Uniquely	68.4%	75.2%	64.7%	65.8%
	Consistently	87.6%	87.3%	83.9%	87.9%
FSP	Uniquely	44.0%	45.1%	47.7%	41.7%
	Consistently	50.9%	49.2%	48.8%	55.8%

*The merged database contains all the transcripts in Ensembl v75, RefSeq (release 105) and CCDS (NCBI release 37.3) databases. A variant is called “consistently” annotated if its reverse-annotation result contains the matched original genomic identifier from COSMIC, and “uniquely” annotated if its reverse-annotation result matched uniquely with the genomic identifier in COSMIC (allowing ± 5 bp positioning ambiguity). *(Table reprinted from TransVar: a multilevel variant annotator for precision genomics. Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A Rohrdanz, James M Melott, Chris Wakefield, Jia Zeng, John N Weinstein, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, Nature Methods, 2015. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)*

Using the Ensembl database v75, we found 75.6% of INSPs, 87.6% of DELPs and 50.9% of FSPs in COSMIC can be accurately traced back to the genomic variants by using TransVar reverse annotation. Among those variants, 74.4% of INSPs, 68.4% of DELPs and 44.4% of FSPs can be uniquely traced back to the genomic variants by using TransVar reverse annotation, the remaining variants have multiple genomic mapping positions because of different transcript isoform usages. Similar to what have been investigated in SASs, we also tried to estimated the reverse annotation accuracies on INSPs, DELPs and FSPs using RefSeq v105 and CCDS v37.3 databases, the consistencies and uniqueness of reverse annotation were different from what has been observed using Ensembl database v75 (Table 2.1). When combining the three transcript databases, the consistency was much improved while the uniqueness rate was further dropped due to redundant transcript isoform usages.

Besides the different consistencies that were observed using different transcript databases in an identical mutation types. We consistently found that insertions have a lower fraction of unique mapping compared with deletions. This is because of the lack of a reference allele that can be used to constrain the search scope of transcripts. For example, ABCA5:c.742_743insA does not contain reference bases to constrain the search whereas a deletion ABCA10:c.1328_1331delTGTC, contains “TGTC” which provides the deleted reference bases to constrain the search. Insertions on the protein level are less affected since the first and the last amino acids are usually specified in the identifier (e.g., ACIN1:p.S647_A648insRS), which enables TransVar to include reference

Table 2.2 Reverse annotation consistency of COSMIC mRNA identifiers via different transcript databases

		Ensembl	RefSeq	CCDS	Merged*
SNS	Uniquely	63.7%	73.2%	89.7%	49.1%
	Consistently	95.5%	92.3%	91.8%	97.3%
INSN	Uniquely	24.1%	34.1%	60.5%	11.9%
	Consistently	81.0%	78.3%	77.9%	85.1%
DELN	Uniquely	61.0%	66.2%	80.2%	50.5%
	Consistently	90.5%	87.2%	86.9%	93.9%
BLSN	Uniquely	78.0%	80.2%	86.4%	71.9%
	Consistently	92.0%	89.7%	88.8%	95.3%

*The merged database is a combination of Ensembl v75, RefSeq (release 105) and CCDS (NCBI release 37.3). A variant is called “consistently” annotated if its reverse-annotation result contains the matched genomic identifier in COSMIC, and “uniquely” annotated if its reverse-annotation result exactly matched the genomic identifier in COSMIC (allowing ± 5 bp positioning ambiguity). (Table reprinted from *TransVar: a multilevel variant annotator for precision genomics*. Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A Rohrdanz, James M Melott, Chris Wakefield, Jia Zeng, John N Weinstein, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, *Nature Methods*, 2015. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)

alleles to narrow the search scope and reduce the ambiguity.

Furthermore, we also investigated the reverse annotation performance of TransVar using the cDNA variants. Similar to what has been observed using the protein variants, a large number of inconsistencies were observed using different types of variants and different transcript databases (Table 2.2).

2.2.6 Reverse annotating protein phosphorylation sites using TransVar

TransVar's novel reverse annotation functionality can greatly facilitate identification of the genomic origins of variants that are functionally interesting at protein or mRNA levels, such as those identified from mass spectrometry data. For example, identifying the genomic locations of phosphorylation sites (such as p.Y308/p.S473 in AKT1 and p.Y1068/p.Y1172 in EGFR) can lead to the recognition of DNA variants that affect phosphorylation-mediated signal transduction, a biological process central to current pharmacogenomics research. Translation from a protein (e.g., EIF4ENIF1:p.Y580) to a genomic identifier (chr22:g.31845362_31845364) usually involves a tedious manual process that is not scalable. First, the UniProt ID must be mapped to a transcript ID. Then, a sentinel site with a known protein-level coordinate must be identified to enable one to measure the distance from the desired site to the sentinel site and step the corresponding number of amino acids. That process must be repeated for all of the isoforms until a matching reference amino acid is identified. With TransVar, all of those processes are automated. As a demonstration, we downloaded and analyzed 191,903 sites of protein phosphorylation in human proteins from PhosphoSitePlus

[84]. In just a few minutes of compute time, we were able to map 167,696 sites (87.4%) to genomic coordinates using CCDS transcripts and 187,464 (97.69%) sites using a combined transcript definition from CCDS, Ensembl and RefSeq. Most of the mapping failures were attributable to obsolete UniProt identifiers that were no longer present in the current releases.

2.2.7 Impact of TransVar on designing experimental validation

In clinical genetics and translational genomics investigations, frequently only the protein variant information is provided to the biological researchers. However, it is critical to know the genomic information of the specific protein variant to allow for precise experimental validation to show the exact function of the variant. For example, FGFR2:p.N549K is a protein variant from FGFR2, which is known to be an oncogene that promotes cell proliferation, however, the specific role of FGFR2:p.N549K is still unknown. It is important for the researchers to know the chromosomal and cDNA locations of this protein variant, then the genomic and transcriptomic information can be used to implement a specific mutation on the wild-type FGFR2 transcript and investigate the functional consequences through *in vitro* and *in vivo* experiments. Therefore, a tool that is capable of performing reverse annotation from protein to genomic level would be highly desired. With this motivation, we implemented the reverse annotation function of TransVar. In this case, TransVar could efficiently take FGFR2:p.N549K as an input and output all 3 potential genomic variants along with their corresponding transcript identifiers (Table 2.3). Given the results provided by TransVar, the users will be able to fully capture the potential ambiguities in annotating the specific variant and obtain the necessary

information to design the experiments.

Given the truth that we observed large ambiguities in both forward and reverse annotations due to different transcript isoform usages, we investigated 537 clinically actionable protein variants and found that 78 (14.5%) could be annotated to multiple genomic origins (Table 2.4), indicating the criticalness of using a reverse annotation tool such as TransVar to resolve such ambiguities in clinics.

Table 2.3 Reverse annotation results of FGFR2:p.N549K using TransVar

Input	Transcript	Coordinate	Region
FGFR2:p.N549K	ENST00000357555	chr10:g.123247577A>C/c.1647T>G/p.N549K	Exon 13
FGFR2:p.N549K	ENST00000358487	chr10:g.123258034A>C/c.1647T>G/p.N549K	Exon 12
FGFR2:p.N549K	ENST00000360144	chr10:g.123247580G>C/c.1647C>G/p.N549K	Exon 13

2.2.8 Impact of TransVar on Pharmacogenomics and hotspot mutation prediction

Another application of TransVar lies in clinical pharmacogenomics and functional mutation prediction.

When performing clinical treatment decision-making using genetic biomarker status, frequently the clinicians just use the protein variant information to make a treatment decision. However, due to the fact that a genomic variant that is discovered in a patient could potentially be annotated to different protein variants because of multiple transcript isoform usages, an ambiguity in using the protein variant as the biomarker could come up and potentially lead to incorrect treatment

decision making. For example, EGFR:p.T790M is a well known biomarker that indicates resistance to Tyrosine kinase inhibitor (TKI) treatment [85-87]. In the COSMIC dataset, EGFR:p.T790M (ENST00000275493) is frequently observed. However, this protein variant is absent in the TCGA dataset. When we investigate the genomic origin of this protein variant in COSMIC, it corresponds to chr7:g.55249071C>T, but if we trace this genomic variant chr7:g.55249071C>T in TCGA, it was actually annotated to EGFR:p.T745M (ENST00000455089) because of different transcript isoform usage (Figure 2.4A). Therefore, based on the annotation that was performed in TCGA, the biomarker indication could be missed. With the equivalence annotation function of TransVar, we were able to infer that EGFR:p.T745M is actually an equivalent protein identifier for EGFR:p.T790M. By this manner, even if EGFR:p.T745M is observed in a clinical genetic report, the clinicians could still link it to EGFR:p.T790M and know that the patient may have drug resistance to TKI treatment.

It is common for a variant and its functional consequence to be described based on protein level information, such as hotspot mutations. However, due to the fact that a genomic variant that is discovered in a patient could potentially be annotated to different protein variants because of multiple transcript isoform usages, an ambiguity could be observed when try to define a hotspot mutation and investigate its functional consequence based on the protein variant information. For example, EGFR:p.A244T (ENST00000455089) is a hotspot mutation that was detected in the TCGA dataset. However, this variant is absent in the COSMIC dataset. When we try to figure out the genomic origin of this protein variant in TCGA,

Table 2.4 Clinically actionable cancer mutations with non-unique genomic origins

Protein Identifier	# of Possible Genomic Origins	PMID
CDKN2A:M53_R58del	9	11255261
CDKN2A:R87P	6	12352668
ERBB2:L755_T759del	5	23220880
ERBB2:774_775insAYVM	5	16843263 19122145
CDKN2A:A36P	3	11595726
CDKN2A:L63V	3	11255261
FGFR1:N546K	3	21367659 15509736 16186508
EGFR:L747_S752del	3	18981003
EGFR:E746_A750del	3	16373402 15837736 17318210
AKT1:P42T	3	23134728
FGFR2:G227E	3	19147536
FGFR2:N549H	3	17525745
FGFR2:N549K	3	22383975 17525745
PTCH1:P681L	2	
PTCH1:Q688*	2	
CDKN2A:A57V	2	19260062
CDKN2A:R58*	2	12362978 11255261
CDKN2A:W110*	2	11255261
CDKN2A:E69*	2	8668202
CDKN2A:P114S	2	23190892
CDKN2A:P114L	2	7777061 8755727
CDKN2A:R80*	2	8668202
CDKN2A:R124fs*22	2	11255261
ABL1:V299L	2	23086624 19201023 21509757
ABL1:E255V	2	20038234 15293570 19164531 21505103
ABL1:E255K	2	12663457 12692682 20697894 19768693
ABL1:L248_K274del	2	17008892 21221851 18354488
ABL1:L184_K274del	2	18354488
EGFR:L838V	2	19147750
EGFR:H835L	2	21422421
EGFR:G810S	2	19147750
EGFR:L747S	2	17973572
EGFR:E709G	2	16205628
EGFR:E709A	2	19671738
EGFR:E709K	2	19726454
EGFR:A702S	2	19020901
EGFR:R521K	2	
BRAF:T241P	2	19206169

KIT:D816V	2	7691885 7512180 23777495 19718013
KIT:D816A	2	16188233 22847983
KIT:D816F	2	9990072
KIT:D816I	2	19865100
KIT:D816Y	2	21504297 19698218 9990072
KIT:D816H	2	21504297 14695343 16188233
KIT:D579del	2	12727838 9797363
KIT:V559_V560del	2	9797363 9438854
KIT:V559del	2	9989791
KIT:W557_K558del	2	12727838 12918066 16203282
KIT:Y553_Q556del	2	9438854 12727838
KIT:M552_Y553del	2	12727838
KIT:P551_V555del	2	11719439 12727838 9438854
KIT:K550_K558del	2	15824741 12727838
PIK3CA:G118D	2	23246288 22949682
ERBB2:V842I	2	23220880
ERBB2:L755S	2	16397024 18413839 23220880
ERBB2:R896C	2	23220880
ERBB2:T733I	2	16397024 18413839
AKT1:R25C	2	23246288 8702995 18823366 17138652
AKT1:Q79K	2	23134728
AKT1:D32Y	2	23134728
AKT1:E17K	2	17611497 23134728 23888070
AKT1:L52R	2	23134728
AKT1:T435P	2	23246288
FLT3:I836del	2	
FLT3:I836del>MN	2	12036858 12663439
CDK4:K22Q	2	9426066 9228064 9712735
CDK4:K22R	2	22197931 9228064 9426066 9712735
CDK4:K22M	2	9228064 9426066 9712735
KRAS:F156L	2	17875937
KRAS:R164Q	2	20147967
PTEN:L181P	2	21828076
PTEN:V166I	2	21828076
PTEN:Q171A	2	21828076
PTEN:S170R	2	21828076 10866302
PTEN:C105F	2	19644652 10560660 10866302
PTEN:K13E	2	14711368
FGFR2:K659E	2	22383975 17525745 23527311
FGFR2:K659N	2	23527311 17525745

PMID: PubMed IDs of the publications that indicate the clinical actionability of mutations.
 Ensembl v75 and GRCh37/hg19 assembly were used to obtain this result.

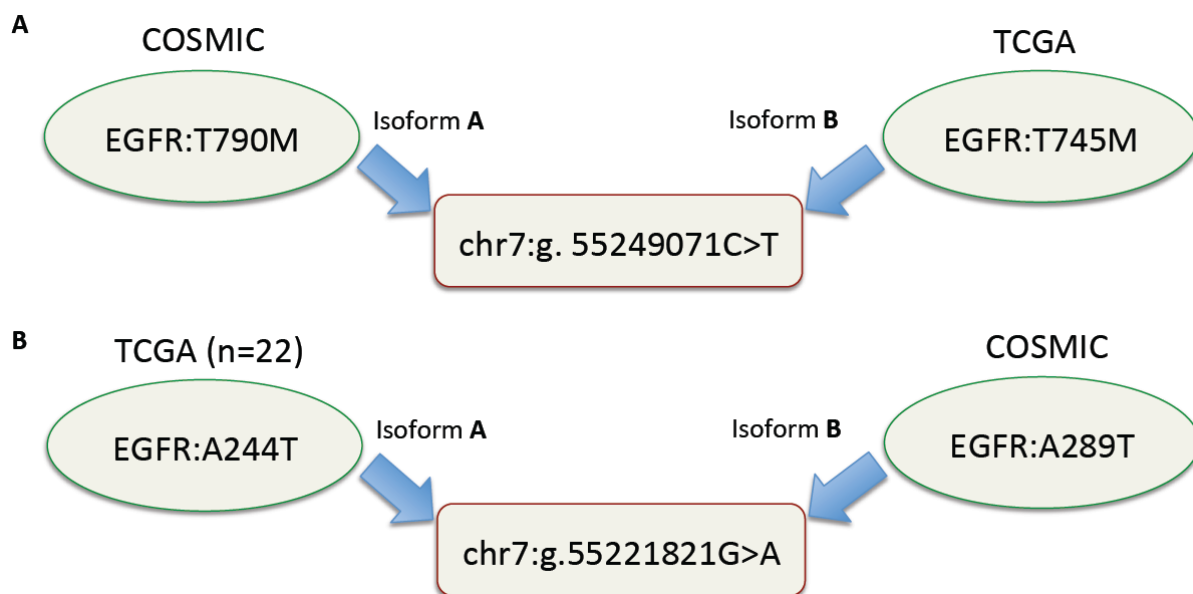


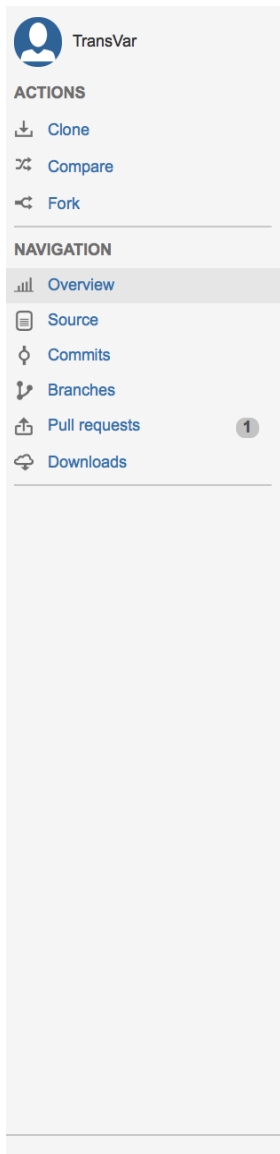
Figure 2.4 Inconsistent forward annotation in TCGA and COSMIC. (A) chr7:g.55249071C>T was annotated to EGFR:p.T790M using transcript isoform ENST00000275493 in COSMIC while it was annotated to EGFR:p.T745M using transcript isoform ENST00000455089 in TCGA. (B) chr7:g.55221821G>A was annotated to EGFR:p.A244T using transcript isoform ENST00000455089 in TCGA while was annotated to EGFR:p.A289T using transcript isoform ENST00000342916 and ENST00000344576 in COSMIC.

it corresponds to chr7:g.55221821G>A, but if we track this genomic variant chr7:g.55221821G>A in COSMIC, it was actually annotated to EGFR:p.A289T (ENST00000342916 and ENST00000344576) because of different transcript isoform usage (Figure 2.4B). As a result, based on the inconsistent annotations that were used in COSMIC and TCGA, the frequency of EGFR:p.A244T/EGFR:p.A289T can be largely underestimated. In addition, some potential hotspot mutations may be missed due to this type of annotation inconsistencies within a single mutation cohort or across different mutation cohorts. With the equivalence annotation function of TransVar, we were able to infer that EGFR:p.A289T is actually an equivalent protein identifier for EGFR:p.A244T. In this way, we would be able to recover the frequency of observed protein variants and correctly classify their functionalities.

2.2.9 Web interface of TransVar

input	transcript	gene	strand	coordinates (gDNA/cDNA/protein)	region	info
FGFR2:p.N549K	ENST00000357555 (protein_coding)	FGFR2	-	chr10:g.123247577A>C/c.1647T>G/p.N549K	cds_in_exon_13	reference_codon=AAT;candidate_codons=
FGFR2:p.N549K	ENST00000358487 (protein_coding)	FGFR2	-	chr10:g.123258034A>C/c.1647T>G/p.N549K	cds_in_exon_12	reference_codon=AAT;candidate_codons=
FGFR2:p.N549K	ENST00000360144 (protein_coding)	FGFR2	-	chr10:g.123247580G>C/c.1647C>G/p.N549K	cds_in_exon_13	reference_codon=AAC;candidate_codons=

Figure 2.5 The web interface of TransVar that shows how to perform reverse annotation and what type of information could be obtained from the output.



- o Download and install
 - dependency
 - download the program
 - install
 - System-wise install (need root)
 - Local install
 - quick start
 - install/specify reference genome assembly
 - transcript annotations
- o Usage
 - specify transcript annotation
 - specify reference assembly
 - view current configuration
 - batch processing
 - reverse annotation of protein sites
 - reverse annotation of protein motif
 - reverse annotation of protein range
 - reverse annotation of single amino acid substitution
 - annotate with additional resources
 - reverse annotation of single nucleotide variation (SNV)
 - reverse annotation of cDNA region
 - reverse annotation of nucleotide insertion
 - reverse annotation of nucleotide deletion
 - reverse annotation of nucleotide block substitution
 - reverse annotation of nucleotide duplication
 - reverse annotation of amino acid insertion
 - reverse annotation of amino acid deletion
 - reverse annotation of amino acid block substitution
 - reverse annotation of amino acid frame-shift
 - search alternative codon identifiers
 - infer potential codon identity
 - annotate SNP from genomic locations
 - annotate a short genomic region
 - annotate a long genomic region
 - annotate a deletion from genomic location
 - annotate an insertion from genomic location
 - annotate block substitution from genomic locations
 - annotate promoter region
 - annotate non-coding RNA
- o FAQ
 - how can TransVar take VCF as input?
 - Can TransVar automatically decompose a haplotype into multiple mutations?
 - Can TransVar use 3-letter code instead of 1-letter code for protein?
 - How can I let TransVar output sequence context?
 - Can TransVar report results in one line for each query?
 - I got 'gene_not_recognized', what's wrong?
 - Does TransVar support alternative format for MNV such as c.508_509CC>TT?
 - Does TransVar support relaxed input without 'g.', 'c.' and 'p.'?
 - When I annotate a variant for protein identifier, why would I end up getting results in another variant type?
- o Future work
- o Bug report and feature request
- o Reference

Figure 2.6 A screenshot of the homepage of TransVar. It provides detailed information of obtaining the source code, installing TransVar and running examples of different annotation functions.

To allow for easy usage of TransVar, we developed a web interface (www.transvar.net), in which the users are free to perform different types of annotation (forward, reverse and equivalence annotation) using different reference genomes and different transcript databases. Currently we support human reference genome hg18, hg19 and hg38 as well as all available transcript databases that correspond to each reference genome. For example, when performing the reverse annotation, the user could choose 'Reverse annotation: Protein' as the annotation option followed up any specific reference genome and any available transcript database, and the results will output all the available genomic variants along with their corresponding transcript identifiers of the given protein variant (Figure 2.5).

2.2.10 Command line usage of TransVar

In addition to the web interface, we also provide a command line tool for batch analysis. On the homepage of TransVar (Figure 2.6), we provide detailed instructions of how to obtain the source code and how to install TransVar on the user's machine, and a bunch of examples to run different types of annotations of TransVar, such as forward, reverse and equivalence annotation.

2.3 Discussion

We developed TransVar, which is a comprehensive variant annotator that performs multi-level variant annotation such as forward annotation from genomic to RNA and to protein level, and reverse annotation from protein to RNA and to genomic level. We developed TransVar as a web-based bioinformatics tool to enable the complete and accurate characterization of the origin and functionality of genomic variants identified by the research community.

Using TransVar, we have identified frequent ambiguities in the current transcript databases (such as Ensembl, RefSeq and CCDS) available for basic and translational research. These include ambiguities in translation among genomic, cDNA and protein levels that involve cancer hotspot mutations, biomarkers that affect clinical treatment decision-making, potentially clinically actionable variants, protein phosphorylation sites and RNA-editing sites. With TransVar, we were able to (i) pinpoint such ambiguities efficiently; (ii) associate variant identifiers accurately across genomic, cDNA and protein levels; and (iii) achieve unambiguous data exchange across different sources.

Our results indicated an urgent need to standardize variant annotation and to improve the quality of information technology that implements the standards. It is not sufficient to specify only the variants themselves; the isoforms that have been used in annotation must also be specified completely in variant databases, and the information must be included in data-sharing processes.

One important application of TransVar lies in revealing all the potential

genomic origins of a given protein variant. In our current study, we have focused on somatic mutations in cancer and have created lists and tables that are useful for disambiguating clinically actionable mutations. Similar lists and tables can be created for other diseases using TransVar to ensure accurate use of information for clinical decision support, genetic diagnosis and counseling.

CHAPTER 3

Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types

(Most of the methods and results in this chapter have been accepted for a publication in BMC Genomics: Tenghui Chen, Zixing Wang, Wanding Zhou, Zechen Chong, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen, “Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types”. The manuscript is currently in final proofreading and not online yet.)

3.1 Materials and Methods

3.1.1 COSMIC somatic mutation data

We downloaded the COSMIC somatic mutation dataset version 71 for our study. As introduced in the COSMIC data source, this total mutation set (12250 samples) includes many sources of curated mutation data. We excluded samples that underwent targeted-sequencing [88], and selected only those that were subjected to either whole genome or whole exome sequencing (Table S3.1). In this manner, we ensured that all the exons of investigated genes were uniformly examined in the selected samples.

3.1.2 Cancer gene candidates

We collected a set of candidate cancer genes from the literatures, which included 546 genes reported in cancer gene census [89], 435 genes in PanCan12 [90], and 221 genes reported in Lawrence *et al* [91]. For OncodriveCLUST [49] and e-Driver [50], we applied their algorithms to predict tumor type-specific driver genes using COSMIC v71 mutation data. We used $q\text{-value} < 0.01$ and $q\text{-value} < 0.05$ to determine driver genes in OncodriveCLUST and in e-Driver, respectively.

3.1.3 Definition of hotspot mutations

Our algorithm identifies hotspots based on amino acid (AA) positions (Figure 3.1). To make sure we have an adequate count of mutations, five major mutation types were included in our statistical modeling: missense, nonsense, coding-silent, insertion and deletion. For missense, nonsense and coding-silent mutations, six

types of di-nucleotide sequence context were considered: A/T transition (ATts), A/T transversion (ATtv), CpG island G/C transition (CpG_CGts), non-CpG island G/C transition (NoCpG_CGts), CpG island G/C transversion (CpG_CGtv), non-CpG island G/C transversion (NoCpG_CGtv), as previously introduced [47]. Altogether, 20 mutation subtypes were considered (Table S3.2). For each mutation subtype in each gene, we counted the number of subtype-specific mutations across all the samples. For each gene, we calculated the mean subtype-specific mutation rate as the total number of subtype-specific mutations in the coding regions (E) divided (normalized) by the protein length. We calculated a p-value based on the number of observed subtype-specific mutations (O) in a given AA, assuming the number of mutations in each mutation subtype follows a Poisson distribution. After obtaining a p-value for each mutation subtype, we computed an integrated p-value for each AA based on Fisher's method [92]

$$x = -2 \sum_{i=1}^k \log(\text{pois}(O_i, E_i)),$$

where i represents a mutation subtype, and pois the Poisson distribution; x follows a chi-square distribution with $2k$ degrees of freedom, where k is the number of mutation subtypes tested. We further applied false discovery rate (FDR) correction [93] and reported hotspot mutations in AA positions with adjusted p-value < 0.001 in COSMIC.

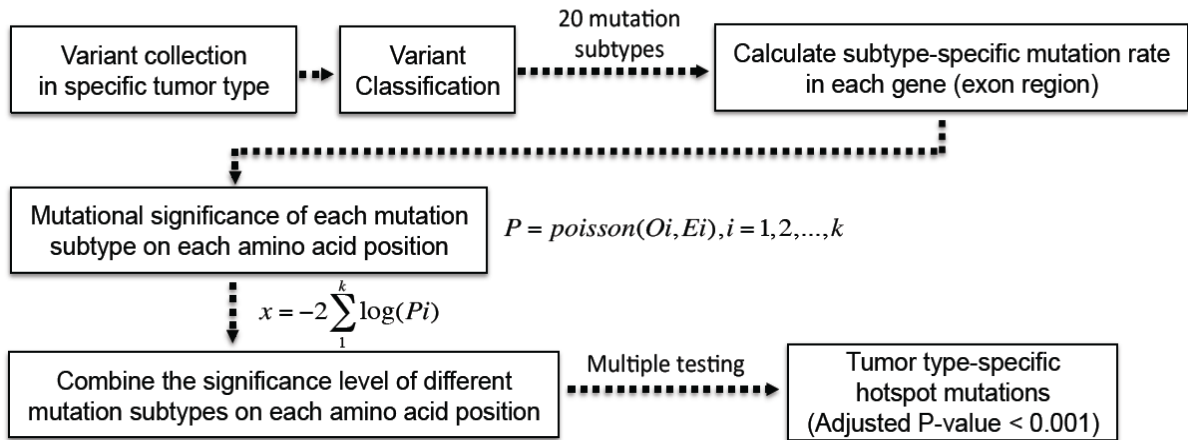


Figure 3.1 A schematic overview of HotDriver. Providing a mutational profile from a specific tumor type, the variants were classified into 20 mutation subtypes, then the mutation subtype-specific mutation rates were computed for each investigated gene and the significance level of each amino acid position in the corresponding gene was calculated. After that, the significance level of each amino acid position was calculated by combining p values from different mutation subtypes using Fisher's method, and an adjusted p value was computed for each amino acid position.

3.1.4 TCGA pan-cancer data

We downloaded TCGA pan-cancer level-3 somatic mutation, copy number alteration and RNA expression data from Synapse (<https://www.synapse.org/#!/Synapse:syn300013>) as last updated in November 2014, and RPPA data from TCPA (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html) [94] for 19 cancer types. More than 4400 tumor samples were assayed by whole exome sequencing, total RNA sequencing [95], or reverse phase protein array (RPPA) technologies. The number of tumor samples available for each cancer type is listed in Table S3.3. In terms of copy number alterations, we called deletions where the normalized copy number value is less than -1 and amplifications where the value is greater than 1. In terms of RNA expression data, we used the normalized TCGA level-3 RNA expression data in our study. To allow for log transformation, the RPKM values of 0 were set to the minimum nonzero RPKM in the given samples. We applied \log_2 transformation to all mRNA RPKM expression values, as described by Jacobsen *et al.* [96]. In terms of protein expression data, we analyzed the expression level of 181 proteins in total using RPPA, which contains 181 high-quality antibodies targeting 128 total proteins and 53 post-translationally modified proteins. We used the normalized level-3 RPPA data (level-4 data for Breast invasive carcinoma) in our study [94].

To test associations between mutations and RNA expression, we used samples that had both somatic mutations and RNA expression data available. To test associations between mutations and protein expression, we used samples that had

both somatic mutation and RPPA data available (Table S3.3).

3.1.5 Cancer Cell Line Encyclopedia (CCLE) mutation and drug sensitivity data

The CCLE [71] contains drug activity data for 24 different compounds in 504 cell lines and somatic mutation data of 906 cell lines. In our analysis, we included cell lines with both drug sensitivity and mutation data. Drug sensitivity data were fit using a logistical-sigmoidal function and described by 4 different variables: the maximal effect level (A_{max}), the drug concentration at half-maximal activity of the compound (EC_{50}), the concentration at which the drug response reached an absolute inhibition of 50% (IC_{50}), and the activity area, which is the area above the dose-response curve [71]. In our analysis, we used the activity area (the under curve area), which captures both efficacy and potency of drug activity according to the CCLE, to measure drug responses.

3.1.6 Tumor-type prevalence of hotspot mutations

To measure the prevalence of a hotspot mutation in tumor type **A**, we calculated the number of **A** samples that contain a target mutation **B**, the number of **A** samples that do not contain **B**, the number of non-**A** samples that contain **B**, and the number of non-**A** samples that do not contain **B**, respectively (Table S3.4). Then we used Fisher's exact test to compute the significance and applied an FDR correction. A hotspot is considered to be highly prevalent in a specific tumor type if the adjusted p-value < 0.01 .

3.1.7 Conservation score comparison

We downloaded the chromosomal base-wise Genomic Evolutionary Rate Profiling (GERP) scores computed by GERP++ [97]. In our study, we extracted the resistant substitution (RS) scores from the nucleotide bases that belong to hotspot mutations and that belong to non-hotspot mutations, and tested if the scores between these two groups were significantly different. A higher RS score represents stronger evolutionary conservation.

3.2 Results

3.2.1 Variable mutation rates among different tumor types and mutation subtypes

As mentioned previously in the methods, we classified all the mutations into 20 subtypes based on both mutation types and di-nucleotide sequence contexts (Table S3.2). In the COSMIC mutation dataset, skin, stomach, bladder and colon tumors have relatively high overall mutational rates, which were consistent with a previous report [48]. Besides, we also observed high mutational rates in bone and endometrium tumors (Figure 3.2). However, we observed highly variable mutational rates across different mutation subtypes (Kruskal-Wallis H-test, $p=2.22e-05$). In one tumor type, different mutation subtypes have largely inconsistent mutational rates. For example, in bone tumors, nonsense non-CpG C/G transversion has a mutation rate of 0.69/Mb while nonsense CpG C/G transition has a mutation rate of 14.2/Mb. In skin tumors, missense non-CpG C/G transition has a mutation rate of 6.18/Mb while silent AT transversion has a mutation rate of 0.53/Mb. Similarly, the mutational rate of one mutation subtype can vary substantially across different tumor types (Kruskal-Wallis H-test, $p=3.49e-40$). For example, missense non-CpG C/G transition has an average rate of 6.18/Mb in skin tumors, much higher than in brain tumors (0.61/Mb). Therefore, to identify potential drivers that are positively selected in cancer, it is important to account for mutation rate variations among mutation subtypes and sequence contexts in different tumor types, instead of assuming a uniform background mutation rate and examining only variant frequencies in the population.

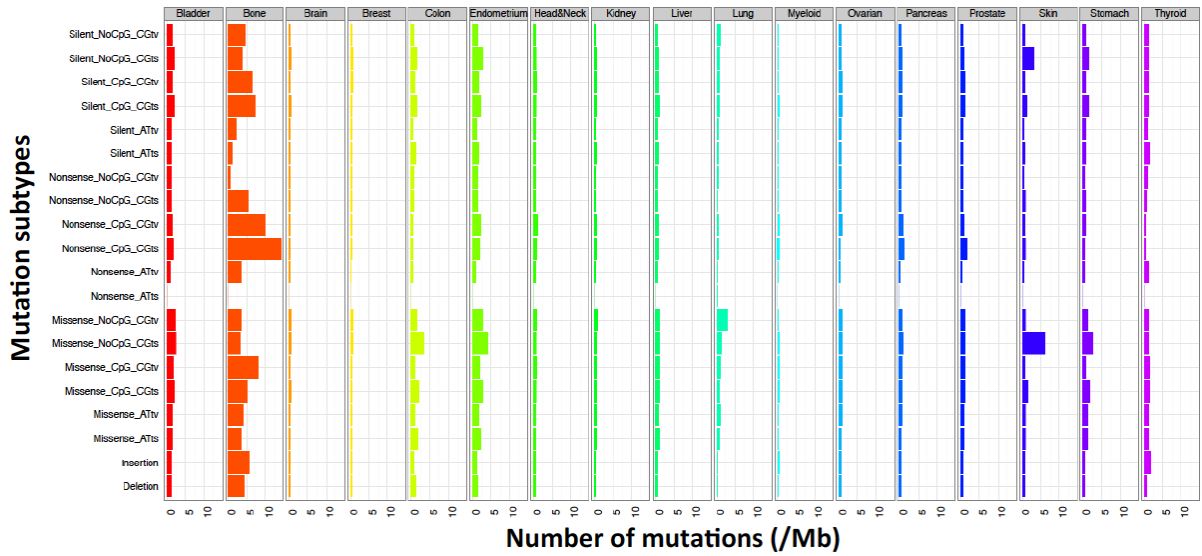


Figure 3.2 Statistics of the mutation distribution in different tumor types in COSMIC.
 The mutation rate of 20 mutation subtypes in 17 main tumor types of COSMIC v71 whole genome and whole exome sequencing data.

3.2.2 Identifying hotspot mutations in COSMIC

We started with all the mutations in 17 tumor types in COSMIC v71 (Figure 3.3). Only data that were obtained via either whole exome or whole genome sequencing were used (Methods, Table S3.1) [88]. Estimation of background mutation rates may be biased by outlier hyper-mutated samples. To avoid such bias, we calculated the mean μ and the standard deviation σ of the number of mutations in each sample, labeled the samples with numbers of mutations greater than $\mu + 2\sigma$ as hyper-mutated, and excluded them from further consideration (Table S3.1).

Our goal was to identify hotspot mutations within individual genes (Methods) and to explore their potentially biological utilities under different biological and disease contexts. The large number of samples in COSMIC made it possible to reliably estimate a background mutation rate for each gene in each tumor type and mutation subtype (Methods). We identified a hotspot mutation as the set of genomic aberrations that affect an amino acid (AA) position and occur significantly more frequently than expected from the background. In total, we identified a set of 702 putative hotspot mutations in 549 genes in 17 tumor types (Figure 3.3, Methods).

We measured the composition of different mutational subtypes in the hotspot mutations (Figure S3.1). As expected, 510 (72.65%) were missense and 17 (2.42%) were nonsense, occupying a high proportion of hotspot mutations. We also identified 31 insertion (4.42%) and 78 deletion (11.11%) hotspots, which were largely ignored in previous studies [49, 50] and potentially offer novel candidates for driver mutation and cancer gene prediction. Besides, we examined the insertion and

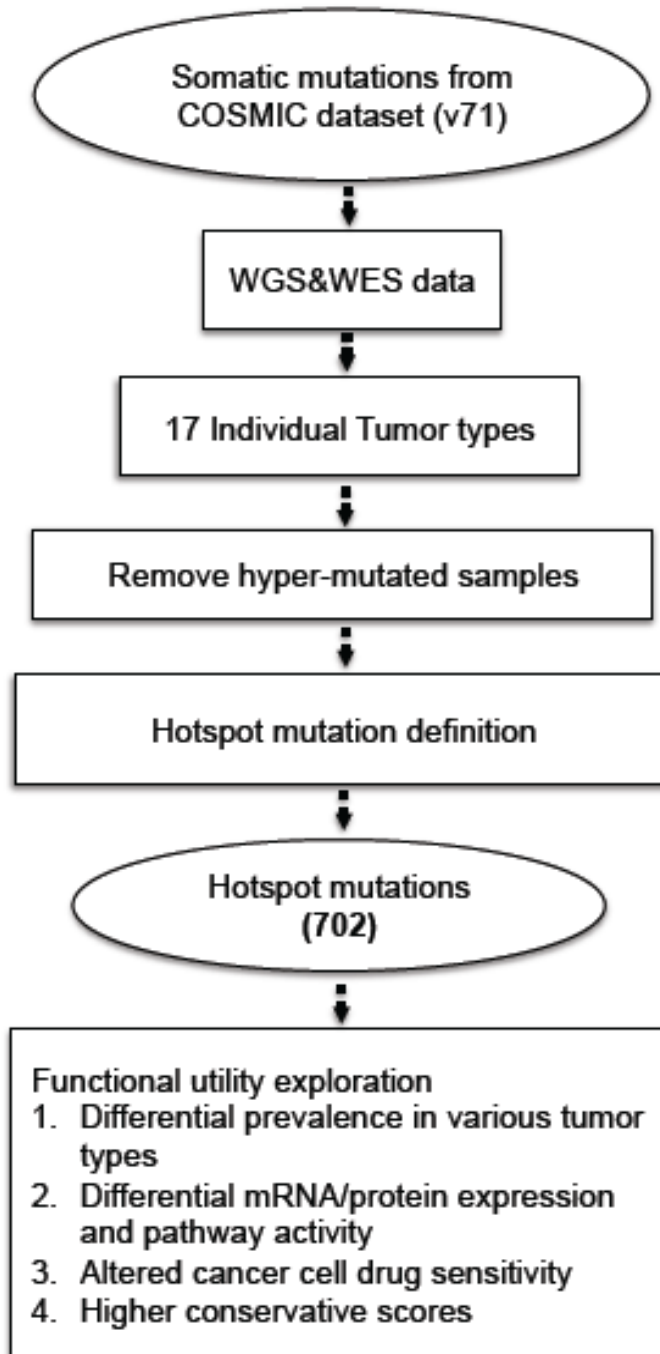


Figure 3.3 Illustration of hotspot mutation definition and functional utility analysis. We used COSMIC v71 data as the input. We first selected the samples that were examined with whole genome or whole exome sequencing, and then removed the hyper-mutated samples in each tumor type. Hotspot mutations were identified in individual tumor types, and the biological utility investigations were performed.

deletion hotspots and found that 17/31 were in-frame insertions and 17/78 were in-frame deletions. Among the remaining frame-shift insertion and deletion hotspots, more than 70% have slightly different start positions and/or sizes. For example, the ESRP1 N512 hotspot deletion has two genomic variants chr8:95686611A/- and chr8:95686611-95686612AA/-.

3.2.3 Evaluate the performance of hotspot mutation identification

We found that the hotspot-mutation-containing-genes (HMCGs) identified in our study overlapped significantly (98/546 vs 451/24405, Fisher exact test, $p=1.28e-53$) with the 546 cancer genes reported in the Cancer Gene Census (CGC). Among 24,951 available genes in COSMIC, 549 genes were identified to contain at least one hotspot, among which 98 were the CGC cancer genes. Similarly, we found that HMCGs overlapped significantly with the significantly mutated genes reported in TCGA PANCAN analysis (101/435 vs 448/24516, Fisher exact test, $p=6.56e-74$) and in Lawrence *et al* (73/221 vs 476/24630, Fisher exact test, $p=2.56e-65$). The non-overlapping genes were detected due likely to that 1) the previous studies had different background mutation rate assumptions than our study; 2) they detected large numbers of tumor suppressors that do not contain clear hotspot mutations; 3) our study was not only able to detect hotspot mutations in known cancer genes, but was also capable of detecting hotspot mutations in infrequently mutated genes, which may have previously unknown biological functionality; 4) our study included mutation types (indels) that previous studies did not. To evaluate the robustness of our statistical modeling, we examined the extent of overlap between HMCGs and the union of the above mentioned cancer gene sets, and found that the overlap

remained highly significant when we chose various adjusted p value cutoffs to identify the hotspot mutations (Figure S3.2), which indicated the statistical robustness of our approach.

Furthermore, we found significantly overlapped genes between our set with those predicted by other cluster-based methods such as e-Driver [50] (151/552 vs 398/24499, Fisher exact test, $p=3.42e-139$) and OncodriveCLUST [49] (106/489 vs 443/24462, Fisher exact test, $p=2.31e-74$). Additionally, regarding the mutational clusters, we found 213 hotspots overlapped with 1125 significant mutational clusters as identified by e-Driver (213/1125 vs 489/92822, Proportional test, $p=2.14e-87$) and 261 hotspots overlapped with 1042 significant mutational clusters as predicted by OncodriveCLUST (261/1042 vs 441/89561, Proportional test, $p=4.98e-121$). Non-overlapping results were found due mainly to: 1) e-Driver and OncodriveCLUST predicted clusters based mainly on missense mutations in a uniform mutational background; 2) our study identified not only missense hotspot mutations but also a substantial proportion of insertion (4.42%) and deletion (11.11%) hotspots (Figure S3.1); 3) our study chose a more stringent statistical significance cutoff to increase the confidence of identified hotspot mutations.

The number of hotspot mutations varied to a great extent from one tumor type to another (Figure 3.4 and Table S3.5). Most tumor types had 5 to 100 hotspot mutations. However, colorectal cancer had 253 hotspot mutations despite its relatively small sample size (684 samples), including a high proportion of insertion (10%) and deletion (23%) hotspot mutations (Figure 3.5). In contrast, only 65 hotspot mutations were found in myeloid cancer (1,344 samples). Such enrichment

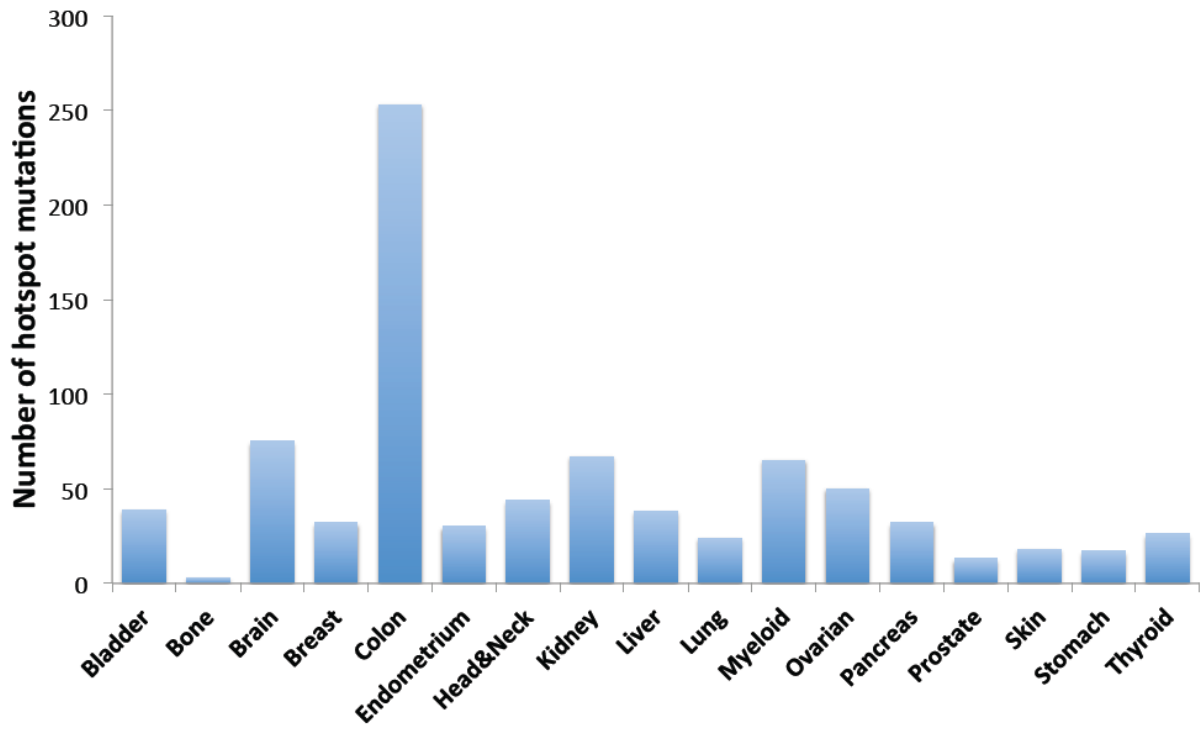


Figure 3.4 Number of hotspot mutations defined in individual tumor types using COSMIC data.

may reflect a higher extent of genetic heterogeneity in the initiation and progression of colorectal cancer, as has been suggested previously [98, 99] and also that colorectal cancer is predominantly driven by mutations rather than by copy number alterations [100]. In addition, we examined the numbers of hotspot mutations and the total numbers of mutations (mutation burden) in each tumor type, but did not find a clear correlation between them (Figure S3.3).

3.2.4 Sequence context signature of hotspot mutations

We investigated the mutational signatures of 702 hotspot mutations under different sequence contexts across different tumor types. As shown in Figure 3.5, in 7 different tumor types (stomach, ovarian, brain, breast, skin, pancreas and kidney cancer), NoCpG_CGts was the most prevalent sequence context compared to other sequence contexts under which the hotspot mutations happened ($p < 0.05$), indicating a higher strength of positive selection on DNA sequences with NoCpG_CGts mutation. In 3 tumor types (head&neck, liver, and myeloid cancer), NoCpG_CGtv appears to be the most prevalent sequence context ($p < 0.05$). In several tumor types such as brain and ovarian cancer, although NoCpG_CGtv did not act as the predominant mutation sequence context, it represented a fairly high percentage (brain: 32% and ovarian: 35%). However, in some tumor types such as bladder cancer, the hotspot mutations are significantly enriched in ATtv sequence context (35%, $p = 1.77e-2$).

In terms of the specific sequence context that hotspot mutations occur across different tumor types, although insertion is not the most prevalent sequence context

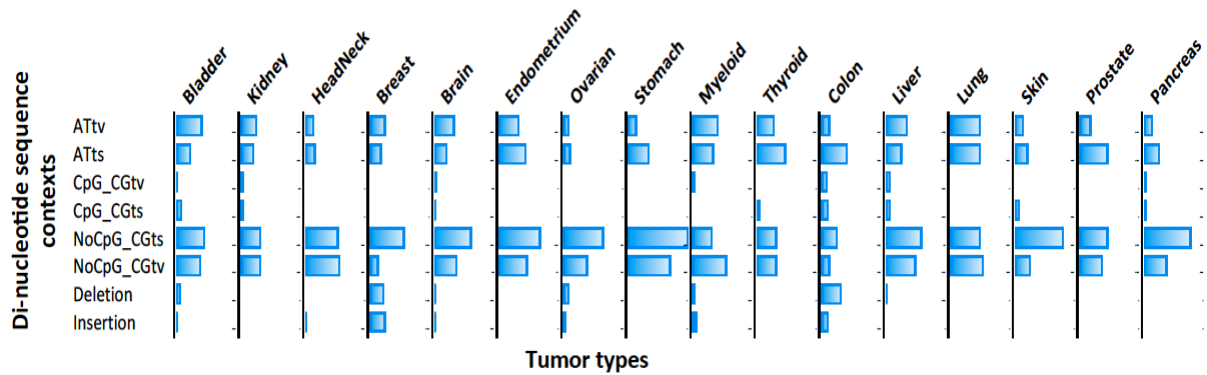


Figure 3.5 Mutational signatures of hotspot mutations in 16 tumor types. The x-axis represents the tumor types and the y-axis represent the 8 types of sequence contexts (concatenating missense, nonsense and silent mutations). Each bar represents the percentage of specific sequence contexts under which the hotspot mutations happen. In each tumor type, the addition of the percentages of different sequence contexts might be larger than 1, because one or more types of mutations may happen on a single hotspot driver mutation site.

within breast cancer, the percentage of insertion in breast cancer (22%) was significantly higher than in any other tumor types ($p= 1.14e-02$), similarly, the percentage of deletion in colorectal cancer (27%) was obviously higher than in other tumor types ($p=1.84e-4$), so as the percentage of ATts (36%, $p=5.84e-3$) in colorectal and ATtv (35%, $p=3.73e-3$) in myeloid cancer.

These observations revealed the common genomic features such as NoCpG_CGts and NoCpG_CGtv sequence context that were positively selected across various tumor types as well as distinct genomic features that occurred in individual tumor types, and highlighted the significance of investigating the hotspot mutations under different sequence contexts separately to better understand their genetic complexities and functional indications.

To gain a novel functional insight with respect to these mutations that were predicted based on statistics of mutation data, we performed a set of additional statistical tests to associate these 702 hotspot mutations with functional evidence.

3.2.5 Exploring the biological utilities of hotspot mutations using TCGA mRNA/protein expression data

The functional consequences of mutations may manifest in two aspects: affecting the expression of the gene containing the mutation or leading to abnormal signaling pathway activity. To address these questions, we divided the mRNA and protein expression values of a set of TCGA samples into multiple groups based on the mutational status of a specific gene in these samples: having a hotspot mutation, no hotspot mutation, or no mutations [95]. Only mutations occurring at least twice were included in the comparison. Mann-Whitney U tests were performed to measure the difference between samples with individual hotspot mutations and samples with non-hotspot mutations, as well as between samples with individual hotspot mutations and samples without mutation [96]. Among 702 hotspot mutations, we found 42 hotspot mutations resulted in significant mRNA or protein expression alterations (Table S3.5).

It is known that TP53 contains gain of function mutations associated with increased expression of TP53 [101, 102] through down-regulation of downstream targets such as *MDM2/MDM4*, which in return attenuate the suppression on the expression of *TP53*. However, it is not well investigated whether different mutations in *TP53* exhibit different functions across different cancer types. Motivated by this,

we examined the association of *TP53* hotspot mutations and RNA and protein expression of *TP53* in different cancer types. To focus on the effect of mutations on *TP53* expression, we excluded samples harboring *TP53* deletions from the analysis (Methods). As shown in Figure 3.6A, in breast invasive carcinoma (BRCA), samples with R175, R248 and R273 missense mutations have obviously higher mRNA or protein expression levels, compared to samples with non-hotspot mutations and with no mutation in *TP53*. In ovarian serous cystadenocarcinoma (OV), similar effects were observed for R248 and R273, which are associated with increases in the *TP53* mRNA and protein expressions (Figure S3.4). However, in rectal adenocarcinoma (READ), although R175 is associated with increases in *TP53* RNA expressions similar to what is observed in BRCA, R248 and R273 missense mutations are not significantly associated with the *TP53* mRNA or protein expression, comparing to samples with non-hotspot or no mutations in *TP53* (Figure 3.6B), implicating distinct functions of R248 and R273 in different disease contexts. In addition, G108 frame-shift deletion, I195 missense and R213 nonsense mutations, which were uniquely detected as hotspot mutations in BRCA, OV and READ respectively, are associated with either reduced or enhanced *TP53* expression in corresponding cancer types, suggesting the functional heterogeneity of hotspot mutations in different cancer types (Figure 3.6 and Figure S3.4).

Instead of altering the RNA/protein level, certain mutations may function via altering downstream protein activity through signaling transduction. For example, activation of *PIK3CA* could lead to activation of downstream targets such as *AKT* phosphorylation [103]. A set of *PIK3CA* mutations have been detected in various

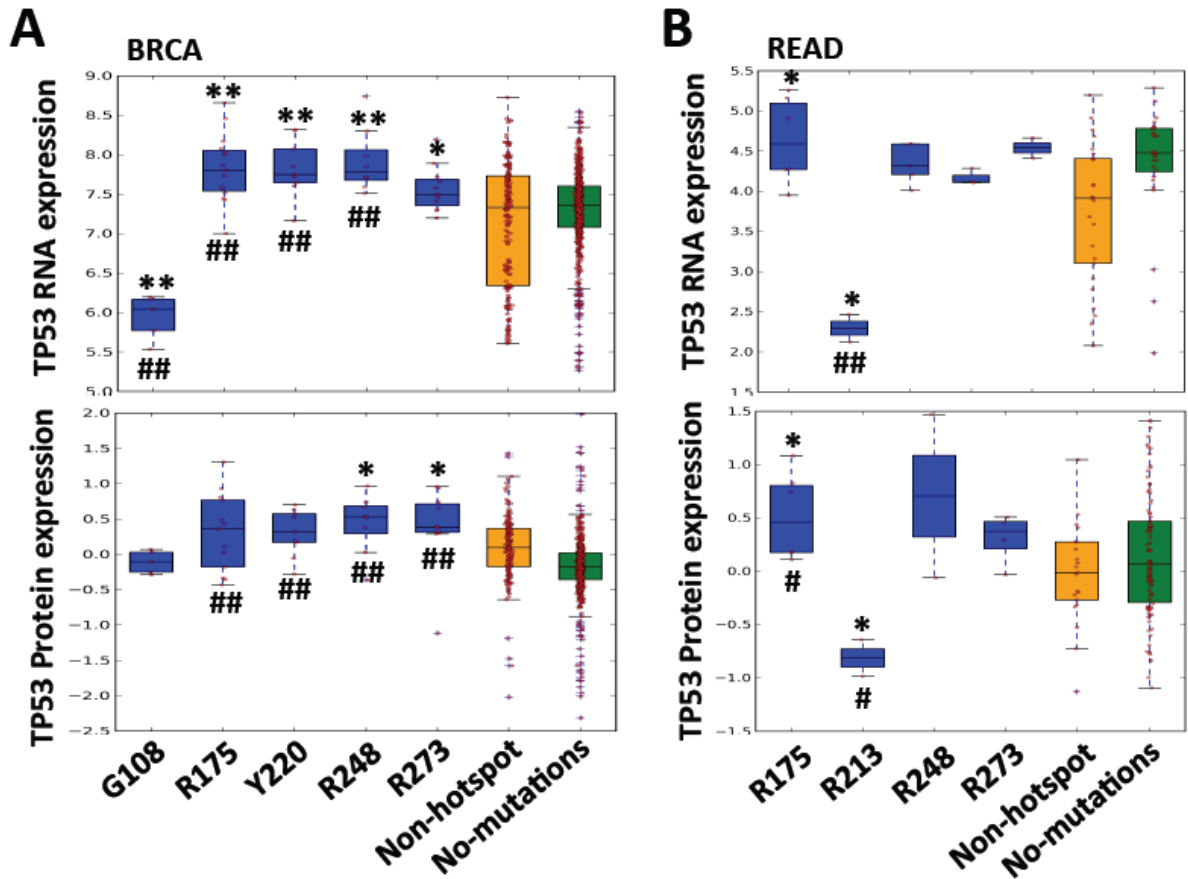


Figure 3.6 Functional implications of hotspot mutations in RNA and protein expression. (A) In BRCA, tumor samples with G108 deletion hotspot mutations in *TP53* exhibit lower *TP53* RNA expression than those with non-hotspot mutations and without *TP53* mutations. In contrast, tumor samples with missense hotspot mutations (R175, Y220, R248 and R273) in *TP53* show higher *TP53* RNA and protein expression. **(B)** In READ, tumor samples with R175 missense mutations show higher *TP53* RNA and protein expression than those with non-hotspot mutations and without *TP53* mutations, while R213 nonsense mutations has the opposite effect. * indicates $p < 0.05$ and ** indicates $p < 0.001$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ and ## indicates $p < 0.001$ between samples with specified hotspot mutations and samples without mutations in examined gene.

cancer types such as BRCA and colon adenocarcinoma (COAD). We examined the association of individual *PIK3CA* mutations and *AKT* activation by comparing the phosphorylated *AKT* levels in samples with various *PIK3CA* mutations to those in samples without *PIK3CA* mutation. Surprisingly, in BRCA, only *PIK3CA* H1047 was associated with dramatically higher *AKT* pT308 and pS473 levels, compared to samples that did not have any *PIK3CA* mutations (Figure 3.7A); in COAD, only *PIK3CA* E542 were associated with significantly higher *AKT* pT308 and pS473 levels, compared to samples that did not have any *PIK3CA* mutations (Figure 3.7B). Notably, in both cases, *PIK3CA* mutations did not affect the total *AKT* level (data not shown), suggesting that different *PIK3CA* mutations in different cancer types may selectively activate *AKT* via signaling transduction, rather than expression regulation.

Therefore, the availability of mRNA and protein expression data enable an opportunity to characterize the biological consequences of different mutations in one cancer type, as well as one mutation under different cancer contexts, reiterating the rationale of distinguishing the function of individual mutations in different disease contexts.

3.2.6 Exploring the pharmacogenomics properties of hotspot mutations

It has been shown that cancer cells respond to specific drugs when they harbor mutations in driver genes such as *BRAF* and *NRAS* [53]. However, it is not entirely clear whether different mutations in a driver gene can trigger different drug responses. Here, we assessed the effects of individual mutations on drug

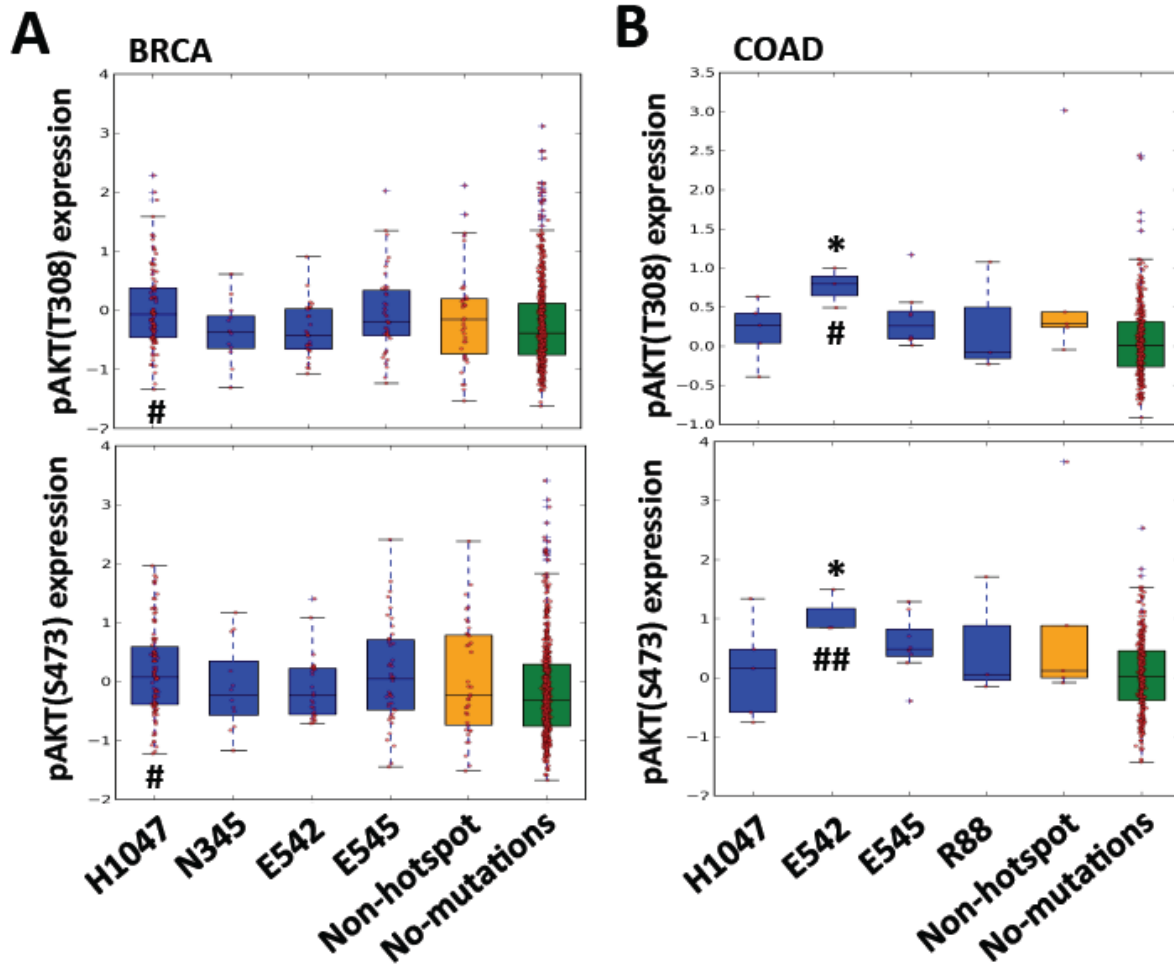


Figure 3.7 Functional implications of hotspot mutations in signaling pathway activity.

(A) In BRCA, tumor samples with H1047 missense hotspot mutations in *PIK3CA* show higher AKT pT308 and pS473 levels than those with no mutations in *PIK3CA*. (B) In COAD, tumor samples with E542 missense hotspot mutations in *PIK3CA* show higher AKT pT308 and pS473 levels than those with no mutations in *PIK3CA*. * indicates $p < 0.05$ and ** indicates $p < 0.001$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ and ## indicates $p < 0.001$ between samples with specified hotspot mutations and samples without mutations in examined gene.

responsiveness using data from the CCLE [71]. We divided cancer cell-line samples into different groups, depending on whether they contain specific hotspot, non-hotspot, or no mutations in investigated gene candidates. Only mutations occurring at least twice were included in the comparison. Mann-Whitney U tests were performed to measure the difference between samples with individual hotspot mutations and samples with non-hotspot mutations, as well as between samples with individual hotspot mutations and samples without mutation [96]. Among 702 hotspot mutations, we found 35 hotspot mutations lead to significantly altered drug sensitivities (Table S3.5).

We first illustrated the effect of individual hotspot mutations in *BRAF*, *KRAS* and *NRAS* on the sensitivity of cancer cells treated by *MEK* inhibitors (PD-0325901 and AZD6244). As expected, cells with *BRAF* V600E mutations demonstrated significantly higher sensitivity to *MEK* inhibitors than those without *BRAF* mutations (data not shown). We also divided cells depending on their mutational status in *NRAS*. Specifically, we found that cells with *NRAS* Q61 hotspot mutations demonstrated significantly higher sensitivity to *MEK* inhibitors than those with non-hotspot mutations and those without mutations in *NRAS* (Figure 3.8A). We further divided cells depending on their mutational status in *KRAS* and found only cells with *KRAS* G12 hotspot mutations demonstrated significantly higher sensitivity to *MEK* inhibitors than those with non-hotspot mutations and those without mutations in *KRAS* (Figure 3.8A).

It has been reported that *TP53* mutations make cancer cells resistant to MDM2 inhibitor (Nutlin-3) [104]. We specifically investigated the effect of different hotspot

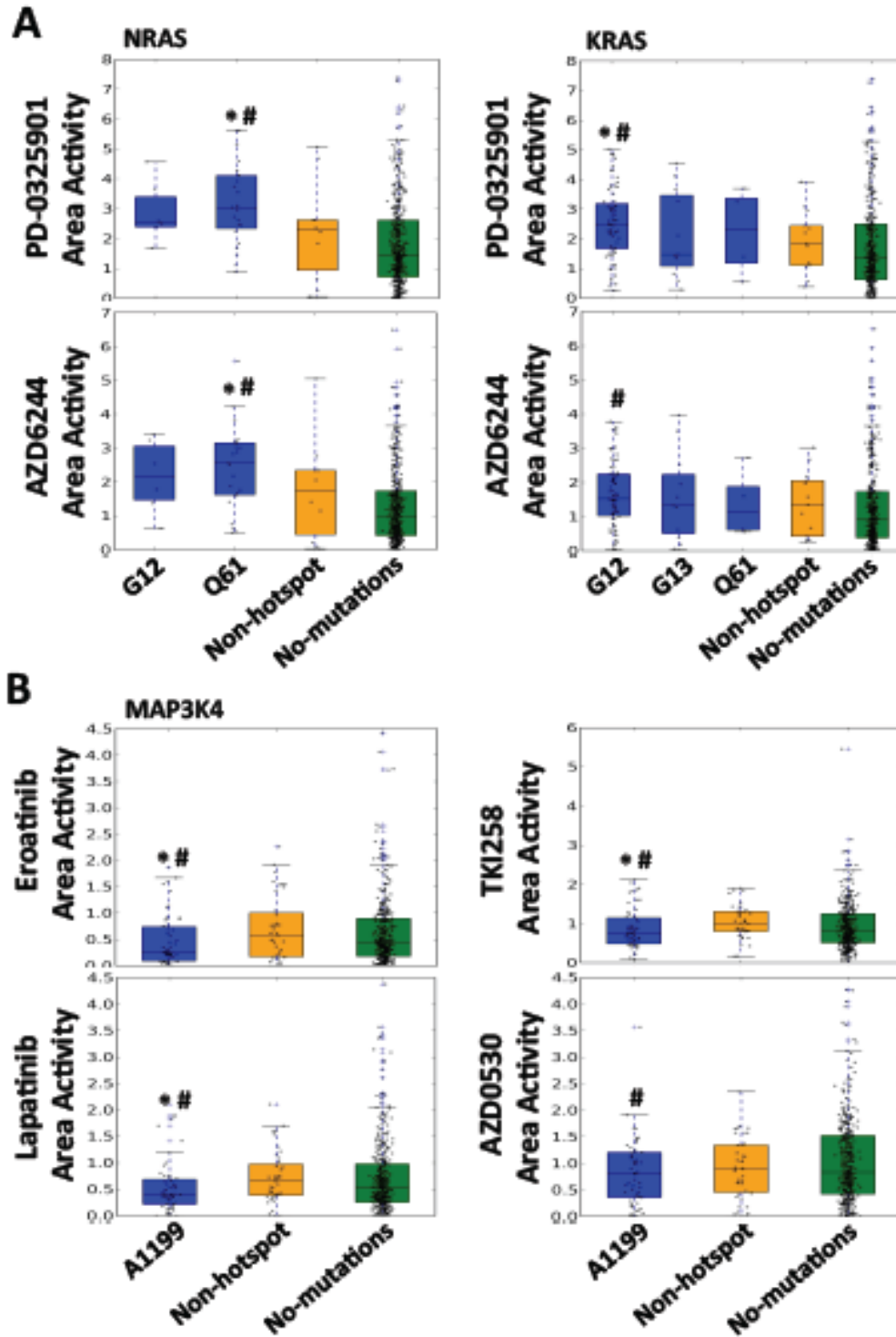


Figure 3.8 Functional implications of hotspot mutations in drug sensitivity. (A) Cancer cells with *NRAS* Q61 or *KRAS* G12 missense hotspot mutations exhibit higher sensitivity to MEK inhibitors (PD-0325901 and AZD6244) than those with non-hotspot mutations or without any mutations in *NRAS* or *KRAS*. **(B)** Cancer cells with *MAP3K4* A1199 deletion

hotspot mutations exhibit lower sensitivity to different EGFR inhibitors (Erlotinib, Lapatinib, TKI258 and AZD0530) than those with non-hotspot mutations or without any mutations in *MAP3K4*. * indicates $p < 0.05$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ between samples with specified hotspot mutations and samples without mutations in examined gene.

mutations and non-hotspot mutations of *TP53*. Surprisingly, we found cancer cells with four *TP53* hotspot mutations (R175, R213, R248 and R273) showed significantly lower sensitivity to Nutlin-3 compared to cells without *TP53* mutations (Figure S3.5 upper panel). Previous study has also suggested that HSP90 inhibitor (17-AAG) exhibits different effects on *TP53* wild-type and mutant cells [105], we specifically measured the effect on cells with different *TP53* mutations and found that only cells with two nonsense hotspot mutations (R213 and R342) are resistant to 17-AAG (Figure S3.5 lower panel) compared to cells without *TP53* mutations, all the other missense hotspot mutations do not show significant effects.

Epidermal growth factor (*EGF*) is one of the high affinity ligands of *EGFR*. The *EGF/EGFR* system induces cell growth, differentiation, migration, adhesion and cell survival through various interacting signaling pathways such as the *MAPK* pathway [106], in which *MAP3K4* is an important component [107]. Clinically, *EGFR* inhibitors such as Erlotinib were used to repress *EGFR* signaling activations and suppress tumor cell growth. However, we found that cancer cell-lines with *MAP3K4* A1199 deletion hotspot mutations were more resistant to all four examined *EGFR* inhibitors (Erlotinib, Lapatinib, TKI258 and AZD0530) in comparison to cancer cell-lines without *MAP3K4* mutations (Figure 3.8B). These *EGFR* hotspot mutant cell-lines are also more resistant to three inhibitors (Erlotinib, Lapatinib and TKI258) in

comparison to cell-lines containing non-hotspot mutations in *MAP3K4* (Figure 3.8B), suggesting the unique function of *MAP3K4* A1199 deletion in disrupting the *MAPK* pathway function and its potential biomarker utility.

It has been well known that *KRAS* occupies a central role in multiple *RTK* signaling pathways, such as the IGF1-R and MET signaling pathway [108]. As expected, we measured the response of cancer cells with *KRAS* mutations and without *KRAS* mutations to IGF1-R inhibitor (AEW541) and c-MET inhibitors (PF2341066 and PHA-665752) and observed that cancer cells with *KRAS* mutations are resistant to IGF1-R inhibitor and c-MET inhibitors compared to cancer cell without *KRAS* mutation (data not shown). To specifically investigate the effect of individual *KRAS* mutations, we grouped the cancer cells by *KRAS* hotspot mutations, and found they were functionally diverse. As shown in Figure S3.6, *KRAS* G13 demonstrated the ability of making cells resistant to IGF1-R compared to cells with non-hotspot mutations and cells without *KRAS* mutations, *KRAS* G12 and *KRAS* Q61 showed minor resistant effect. *KRAS* G12 and G13 made cells resistant to c-MET inhibitors compared to cells with non-hotspot mutations and cells without *KRAS* mutations, while *KRAS* Q61 does not have notable resistant effects.

These observations above support that hotspot mutations we identified may have distinct roles in mediating signaling pathways and are associated with different drug sensitivities. Therefore, it is critical to obtain accurate genomic information and interpret them in context-specific manner in order to achieve desirable outcomes in personalized cancer treatment.

3.2.7 Tumor type-specific hotspot mutations

TCGA pan-cancer data enabled us to investigate the diverse function of a cancer gene in different tumor types. For example, *TP53* was found to be important in most tumor types [109], and *APC* was important mostly in rectal (READ) and colon adenocarcinoma (COAD) [110]. However, related studies have thus focused on characterizing genes and it is largely unclear whether individual mutations demonstrate functionality that is specific to different tumor types. We performed an analysis to assess whether a hotspot mutation in our set is highly prevalent in specific tumor types. Among all the 702 hotspots, we found that 68 were highly prevalent in one tumor type, 11 in two tumor types, 2 (*KRAS* G12 and *PIK3CA* E542) in three tumor types, and 1 (*KRAS* G13) in four tumor types (Figure S3.7 and Table S3.6). Among these, 34 hotspot mutations such as *CD209* R129 missense (4.0%) in bladder cancer, *MAGI1* Q421 insertion (0.8%) and *NR1H2* Q175 insertion (1.8%) in breast cancer were not well investigated based on previous studies and are potentially novel targets.

Of the 21 hotspot mutations detected in *TP53* (Figure 3.9), 2 were found to be prevalent in multiple cancer types (R248 in bladder urothelial carcinoma (BLCA), BRCA and OV, R273 in lower grade glioma (LGG), BRCA and OV), and 9 (G108, R158, R175, I195, R213, Y220, R249, R282, E285) in one tumor type, confirming the functional diversity of *TP53* hotspot mutations in different cancer types (Figure 3.9). Our results indicated that the function of mutations in a gene may be highly heterogeneous in different tumor types. In addition, most hotspots appeared to be highly homogeneous, containing one subtype of mutation at any given amino acid

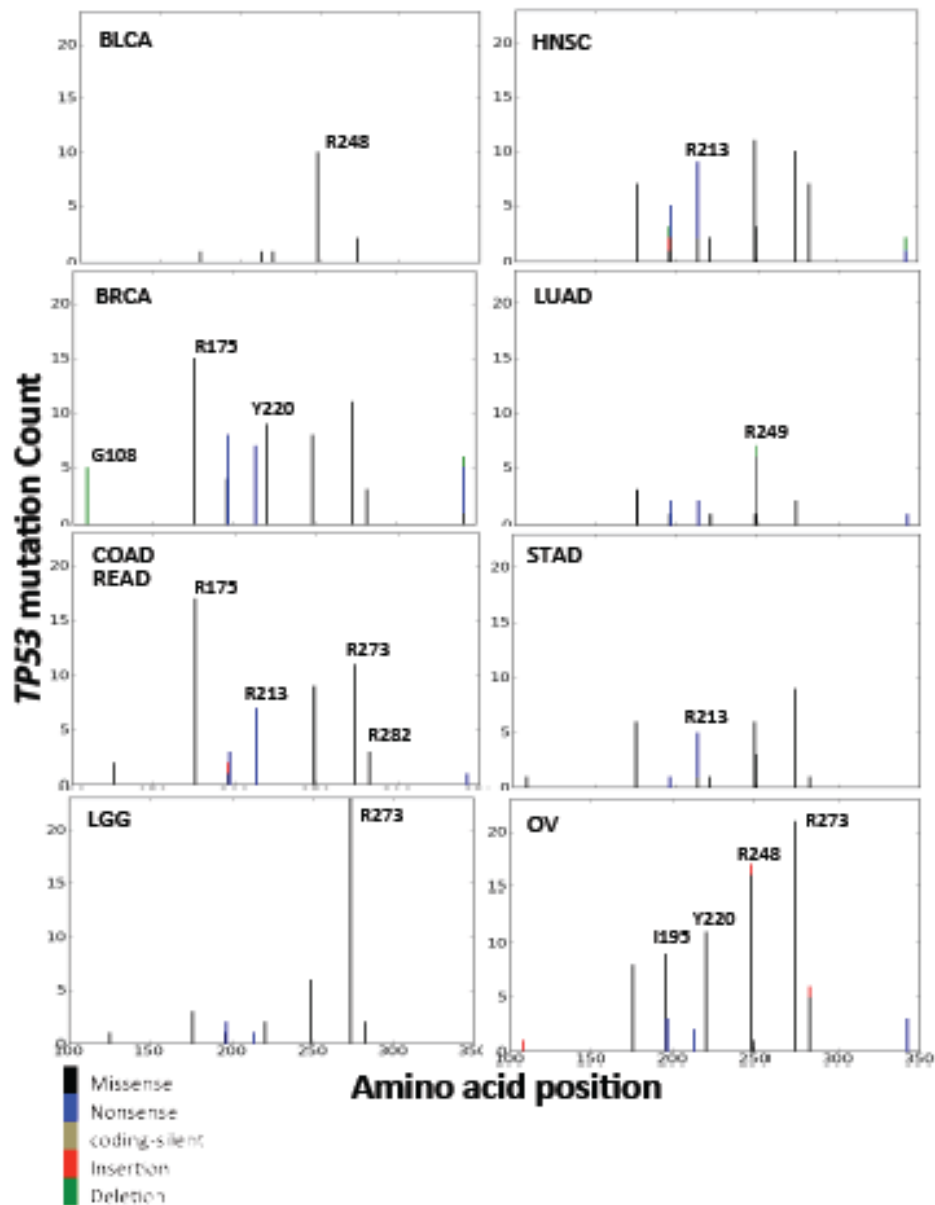


Figure 3.9 Prevalence of *TP53* hotspot mutations in different TCGA cancer types. The hotspot mutations of *TP53* are differentially prevalent in different tumor types, indicating their differential functions.

position.

Among various tumor types, COAD&READ had the highest number (24) of tumor type-enriched hotspot mutations (15 in *APC*, 3 in *KRAS*, 2 in *FBXW7*, and 1 in *TP53*, *SMAD4*, *NRAS*, and *ERBB3* respectively), which again suggested a high degree of genetic complexity in colorectal cancer. We did not identify any tumor type-specific hotspot mutations in kidney renal papillary cell carcinoma (KIRP), which suggests that the development of KIRP may not involve a unique pathway and so may be generally similar to that of other tumor types (Figure S3.8).

We further identified 30 hotspot mutations that were exclusively detected in only one tumor type (Table 3.1). Included were *DNMT3A* R882 and *NPM1* W288, which occur in 14.9% and 25.6% of acute myeloid leukemia (LAML) patients, respectively and have been shown important in LAML oncogenesis [111]. Besides these expected hotspots, we found some potentially novel hotspots. For example, we found an in-frame insertion hotspot mutation, *NR1H2* Q175 in 1.8% of BRCA patients, further investigation using BRCA mRNA expression data showed that *NR1H2* Q175 insertion is associated with reduced mRNA expression of *NR1H2*, comparing to *NR1H2* non-hotspot mutations (Mann-Whitney U test, $p=2.60e-2$, Figure 3.10A). Although having been reported to regulate cholesterol homeostasis and tumorigenesis of liver cancer [112], the role of *NR1H2* Q175 insertion in BRCA has not been well characterized. In addition, *GATA3* P409, a frame-shift insertion hotspot mutation was detected in 1.6% of BRCA patients. BRCA samples with *GATA3* P409 insertions had higher expressions of *GATA3* compared to samples

Table 3.1 Hotspot mutations exclusively detected in only one tumor type in TCGA pan-cancer data

Tumor_Type	Gene	aa_Pos	Frequency	adj_Pvalue
BLCA	RXRA	S427	5.00%	4.26E-08
BRCA	GATA3	P409	1.60%	5.48E-09
BRCA	GOLGA6L2	E537	0.90%	1.27E-04
BRCA	MAGI1	Q421	0.80%	4.95E-04
COAD&READ	APC	R876	5.30%	4.01E-14
KIRC	NEFH	P655	0.70%	2.16E-03
LAML	FLT3	D835	8.20%	1.94E-21
LAML	NPM1	W288	25.60%	1.58E-69
LAML	DNMT3A	R882	14.90%	4.24E-38
SKCM	AGAP10	M293	2.80%	1.62E-08
SKCM	C15orf23	S24	2.80%	1.62E-08
SKCM	PCDHGA1	R293	2.00%	3.48E-06
SKCM	TRRAP	S722	1.60%	4.83E-05
STAD	BMPR2	N583	2.00%	1.49E-04
STAD	CCDC43	R216	2.00%	1.49E-04
STAD	ESRP1	N512	3.30%	3.25E-07
STAD	FAM18A	F140	2.00%	1.49E-04
STAD	GTF2I	N440	2.00%	1.49E-04
STAD	STAMBPL1	K405	2.00%	1.49E-04
STAD	ZNF365	K399	2.00%	1.49E-04
STAD	CNBD1	L396	2.00%	4.68E-04
STAD	DOCK3	P1852	6.60%	2.47E-13
STAD	PGM5	I98	10.50%	5.36E-22
STAD	SLC3A2	K331	2.60%	3.15E-05
STAD	UBR5	E2121	5.30%	1.72E-10
UCEC	FGFR2	S252	3.60%	6.31E-11
UCEC	MAX	H28	1.60%	4.55E-05
UCEC	BCOR	N1459	3.20%	7.14E-09
UCEC	PIK3CA	R93	2.40%	1.25E-06

Note: Frequency was calculated by dividing number of mutations over number of samples in specific tumor type; adjusted p-value was computed based on Fisher's exact test followed by FDR correction.

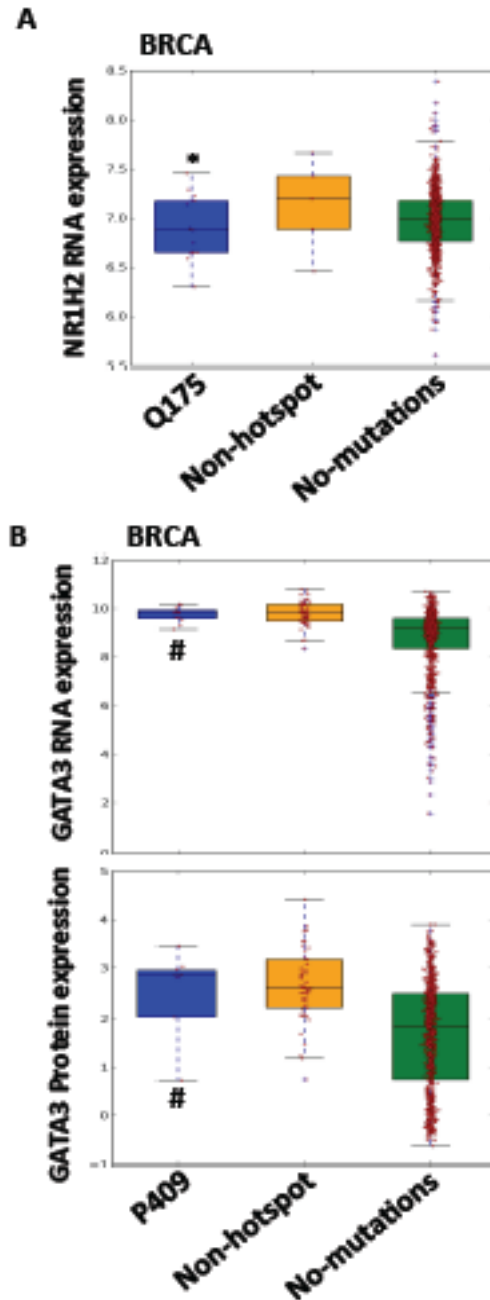


Figure 3.10 Prevalence of hotspot mutations in different TCGA cancer types and their functional implications. (A) In BRCA, samples with *NR1H2* Q175 in-frame insertion hotspot mutations have significantly lower *NR1H2* expression compared to samples with *NR1H2* non-hotspot mutations. **(B)** In BRCA, sample with *GATA3* P409 insertion hotspot mutations have obviously higher *GATA3* compared to samples without *GATA3* mutation. * indicates $p < 0.05$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ between samples with specified hotspot mutations and samples without mutations in examined gene.

without *GATA3* mutations based on both the BRCA mRNA expression (Mann-Whitney U test, $p=2.03e-2$) and RRPAs data (Mann-Whitney U test, $p=5.94e-2$, Figure 3.10B). Because *GATA3* has been proposed as a prognostic biomarker in breast cancer [113], the high frequency of *GATA3* P409 and elevated *GATA3* expression in BRCA make them potential useful therapeutic targets in clinics.

3.2.8 Conservation and protein-domain characteristics of the hotspot mutations

In general, functional and structural important mutations are expected to locate in highly evolutionally conserved region and domain in the protein. To evaluate our hotspot mutation, we used the RS scores computed by GERP++ [97], to measure the evolutionary constraints across different chromosomal sites (Methods). We compared the RS score difference between the sites that belong to hotspot mutations and those belong to non-hotspot mutations. The RS scores of 702 hotspot mutations were significantly higher than those of non-hotspot mutations (Figure 3.11A), suggesting the sites that harbor hotspot mutations were more conserved than those do not. In addition, we also examined the relative location of mutations on the protein. The non-hotspot mutations were evenly distributed across different domains of the protein (lower panel), while the hotspot mutations showed clustering in the middle and the terminals (Figure 3.11B, upper panel), suggesting the functional preference of mutations in different protein domains.

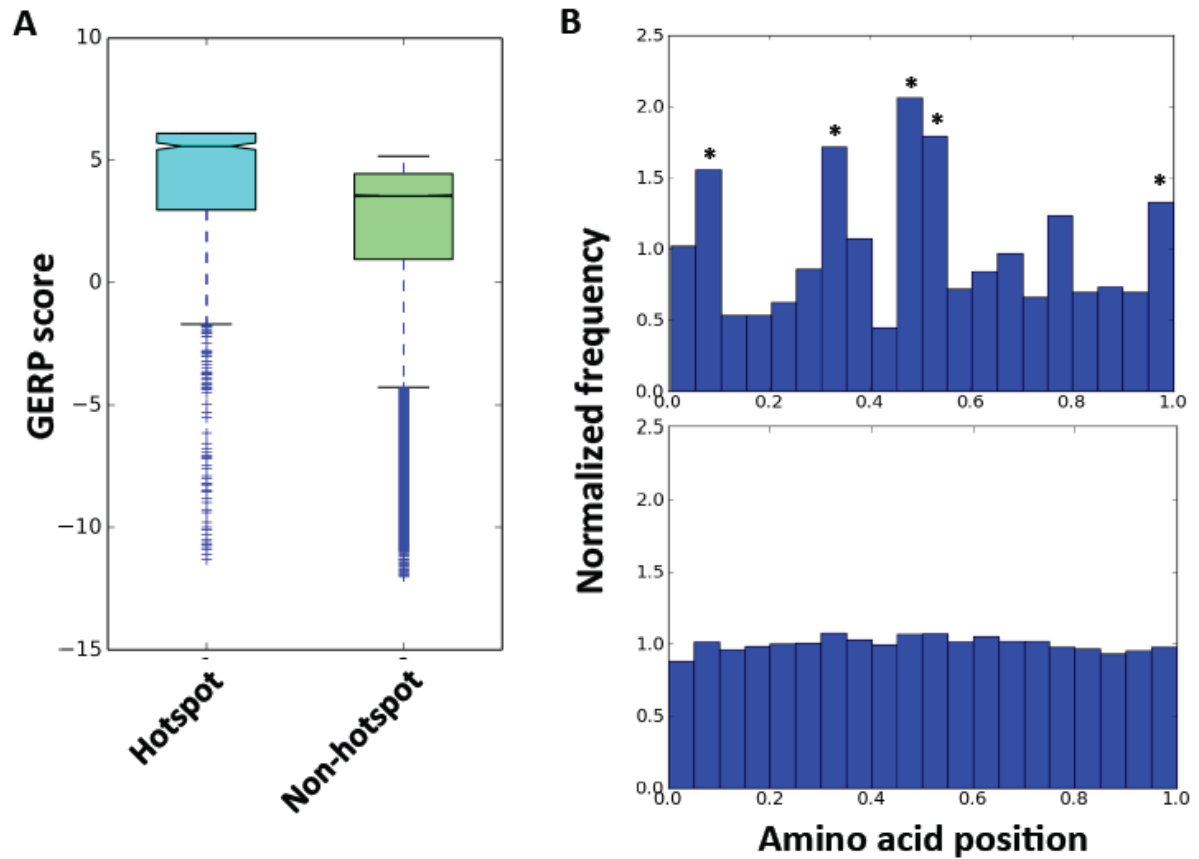


Figure 3.11 Comparing the conservation and proteome domain localization of the hotspot and the non-hotspot mutations. (A) Comparison of GERP score between the hotspot and non-hotspot mutations. (B) Investigation of the proteome domain location of the hotspot (upper) and non-hotspot (lower) mutations.

3.3 Discussion

We nominated 702 hotspot mutations in 549 genes from the COSMIC database, among which 53 were associated with statistically significant functional evidences in currently available TCGA and CCLE data (Table S3.5). The rest of the hotspot mutations could not be associated with additional functional evidence, which may due to sparseness in the data and limitations in the current knowledge bases. For example, only 187 antibodies were measured on the RPPAs, the sample size was relatively small and some observed patterns might change as the sample size increases in the future. Nonetheless, our study revealed differential biological consequences and pharmacogenomics utilities of mutations under different disease contexts and highlighted the significance of allocating the specific function of individual mutations using functional genomics and pharmacogenomics data. These aspects have not been systematically explored in previous studies. Besides investigating previous known hotspot mutations in different contexts, we also nominated a set of novel hotspot mutations such as those in *MAP3K4*, *NR1H2* and *GATA3* with corresponding functional associations, which represent good candidates for developing predictive biomarkers and drug targets.

Investigating the mutational signatures in different cancer types has been a useful tool for understanding the underlying biological processes of cancer development. Alexandrov *et al.* [114] dissected all the mutations in TCGA into 21 distinct mutational signatures with diverse sequence context enrichments and associated them with different phenotypes such as age of the patient at cancer diagnosis, known mutagenic exposures or defects in DNA maintenance. Kandoth *et*

al. [115] investigated 12 cancer types in TCGA and reported that mutations were enriched in C/G transitions such as C->T and C/G transversions such as C->A in different cancer types using all the mutation data. In our study, we focused on predicted hotspot mutations and illustrated the mutational signatures that hotspot mutations represented. We found that hotspot mutations were enriched in NoCpG_CGts and NoCpG_CGtv sequence context in 10 tumor types and some sequence contexts such as ATtv in bladder cancer. In addition, we elucidated that insertion mutations were highly enriched in breast cancer and deletion mutations were enriched in colorectal cancer, which was a novel finding in our study.

Another novel contribution of our current investigation was to highlight the criticalness of distinguishing the biological roles of individual hotspot mutations within one cancer gene under different disease contexts. Different hotspot mutations within one gene can exhibit diverse functional indications. For example, only *PIK3CA* H1047 but not any other hotspot mutations enhances the *AKT* pathway activity in BRCA, while only *PIK3CA* E542 enhances the *AKT* pathway activity in COAD. Previous studies observed that *PIK3CA* H1047R and E545K both result in a constitutively active enzyme with oncogenic capacity but the effect of H1047R is much stronger than E545K [116, 117]. We may not have seen obvious E545K enhancement of the *AKT* pathway activity because: 1) insufficient samples carrying the *PIK3CA* E545K mutation in our current analysis; and 2) highly sparse expression of phospho-AKT in samples without *PIK3CA* mutation. Similarly, one hotspot mutation can represent different functional relevance in different cancer types. For example, with *TP53*, R248 and R273 significantly increase its RNA and protein

expression in BRCA and OV but not in READ. In addition, different *TP53* hotspot mutations were prevalent in various cancer types, and 30 hotspot mutations exclusively occur in only one cancer type. Although it was interesting to observe the differential functional correlation of hotspot mutations in different disease contexts, to further improve the convincingness of the conclusions achieved in our study, power analysis would be an ideal way to evaluate the reliability of functional correlation analysis, especially when measuring the differential functional impact of an identical hotspot mutation across different cancer types.

Along the line of identifying hotspot mutations, it was commonly assumed that mutations close to each other are expected to exhibit similar functions and grouping nearby mutations as a hotspot would improve the power of identifying driver mutations. One important observation of our study was we found that even hotspot mutations close to each other could have distinct biological implications in the same cancer type. For example, *PIK3CA* E542 was significantly associated with enhancement of phospho-AKT activities in COAD, while E545 did not; cell-lines with *KRAS* G13 were resistant to IGF-1R inhibitor (AEW541), while those with G12 did not (data not shown). Nearby hotspot mutations demonstrated distinct functions under different disease context. Simply collapsing mutations based on proximity and assuming that nearby mutations have the same functions may result in errors in functional prediction.

Although available functional genomic data prohibited us from systematically characterizing every hotspot mutation we predicted, our integrative assessment based on mRNA expression, protein activity, drug sensitivity, and tumor specificity

data in TCGA and CCLE, indicated potential utility of each of our predicted hotspot mutations. Such functional characterization can be unequivocally improved in the future by using systematic pathway-aware algorithms such as DriverNet [118] and PARADIGM-SHIFT [119], and by integrating additional functional genomic datasets such as Genomics of Drug Sensitivity in Cancer (GDSC) [72]. In addition, further dissecting the mutation data into different groups would be helpful to distinguish distinct mutation profiles and precisely investigate the specific function of hotspot mutations in different cancer subtypes. For example, different cancer subtype groups (such as MSI and non-MSI in colorectal cancer, ER+, HER2+ and TNBC in breast cancer) or considering the co-founding factors across the population (such as age, sex, ethics). When evaluate the enrichment of hotspot mutations in a specific cancer type, it is also valuable to dive into cancer subtypes and investigate whether any hotspot mutations only occur in specific cancer subtype (for example, TNBC in breast cancer). Importantly, our results demonstrated a high degree of functional heterogeneity at the mutational level, which has not been sufficiently apprehended or investigated in current research and clinical practice. Despite all the caveats, the hotspot mutations we identified provide a step forward in cataloging hotspot driver mutations in different cancer types and biological contexts, which is critical for realizing the promise of personalized cancer medicine.

CHAPTER 4

Conclusions and future directions

4.1 Conclusions

In the dissertation study, I focused on investigating the possibility of applying -omics data to enhance the capability of precision medicine. To achieve this essential goal, we performed the studies through two aspects: 1) Illustrated the potential ambiguities in variant annotations, and developed TransVar to help resolve such ambiguities and greatly increase the accuracy of applying variant annotations in different research and clinical fields; 2) Proposed a population-based statistical model to identify hotspot mutations in amino acid resolution, and elucidated their mutational signatures and diverse biological utilities across different cancer types.

We developed TransVar, which is a comprehensive variant annotator that performs multi-level variant annotation such as forward annotation from genomic to RNA and to protein level, and reverse annotation from protein to RNA and to genomic level. We implemented the command line tool and make TransVar highly flexible of handling different formats of data input. Essentially, TransVar supports not only the standard variant format that was recommended by HGVS nomenclature, but also formats that were frequently used by researchers.

We investigated the various areas that the reverse and equivalence annotation function of TransVar could potentially contribute to: 1) experimental validation design; 2) clinical pharmacogenomics; and 3) hotspot mutation prediction. With the comprehensive functions of TransVar, the ambiguities that were encountered frequently in biological investigation and clinical treatment decision-making could be largely removed. In addition, our investigation revealed the frequent annotation

inconsistencies in current databases. We used the mutation data in TCGA and COSMIC to show large inconsistencies between the annotation practices in these two mutation databases. We used the COSMIC data to compare the forward annotation consistencies that were achieved by TransVar and other existing annotations, and showed that TransVar could provide the most comprehensive information available to allow the users to fully capture the potential protein variants that were annotated from a specific genomic variant. Our study also tried to illustrate the ambiguities of reverse annotation among different transcript databases and different mutation types. With both forward and reverse annotation that was enabled in TransVar, we can reveal hidden inconsistency and significantly improve the precision of translational and clinical genomics. The source code and detailed instructions for TransVar are available at <https://bitbucket.org/wanding/transvar> and a web interface is at <http://www.transvar.net>.

By investigating mutation data of 17 different tumor types in COSMIC, we observed a large discordancy of mutation rates across different mutation subtypes and tumor types. By respecting those mutation variations, we developed a population-based statistical model to nominate 702 hotspot mutations in 549 cancer genes using COSMIC data in a gene, tumor type, mutation subtype and sequence context specific manner. We illustrated the common and distinct mutational signatures of hotspot mutations across different tumor types, we found a high enrichment of Non-CpG island C/G transition and transversion in 10 tumor types, insertion hotspots are highly prevalent in Breast cancer, and deletion hotspots are enriched in colon cancer. We also employed multi-dimensional functional evidence

(RNA sequencing, reverse phase protein array and pharmacogenomics data) to demonstrate the diverse functional relevance of hotspot mutations in different biological and disease contexts and nominate a set of novel hotspot mutations such as *MAP3K4* A1199 deletion, *NR1H2* R175 insertion, and *GATA3* P409 insertion with different functional associations. Our results will promote our understanding of the process of genomic positive selection by investigating the mutational signatures on hotspot mutations and facilitate ongoing efforts in cancer target discovery and development [120]. The source code used for our analysis is available at <https://sourceforge.net/projects/hotdriver/>.

4.2 Future directions

With the quick expansion of high dimensional data in the past few years, more and more attention has been paid to exploring the utility of -omics data in clinical applications. These greatly improved our capability of performing clinical management and developing novel drugs, and therefore allowed us to get close to the goal of personalized medicine.

Through the current functions enabled by TransVar, we are able to perform cross-level variant annotation in genomic, RNA and protein levels. The functions of TransVar could help the researchers and clinicians resolve some ambiguities such as experimental validation design, clinical pharmacogenomics and hotspot mutation prediction. In the future, we will try to further expand the potential applications of TransVar in other biological contexts: 1) annotate the functional impact of a variant such as in any specific protein domain locations, epigenetic regulatory domains, etc; 2) explore the application of TransVar in shRNA design to make sure that the shRNA targeted region globally exists in all transcript isoforms of a specific gene; 3) investigate the effects of genomic editing CRISPR-Cas9 on the protein level and the consequence on the RNA expression.

We developed a population-based statistical model to nominate 702 hotspot mutations in 549 cancer genes using COSMIC data in a gene, tumor type, mutation subtype and sequence context specific manner. We employed multi-dimensional functional evidence to demonstrate the diverse functional relevance of hotspot mutations in different contexts and nominate a set of novel hotspot mutations with

functional indications. In the future, we can apply the statistical model to different in-house or publically available mutation cohorts to identify hotspot mutations.

Specifically, it would be interesting to identify hotspot mutations in different public datasets such as ICGC and TCGA, and investigate whether the hotspot mutations nominated are different and if there are any specific hotspot mutations occur in specific dataset. Beside, our current model focused on identify hotspot mutations of single amino acids, we will further investigate whether grouping close-by mutations will enhance the power of identifying functional mutations. In terms of functional indication of hotspot mutations, we will try to 1) evaluate the pathway activities and systematically evaluate the effect of hotspot mutations on different pathway activities; and 2) utilize power analysis to evaluate the confidence of our statistical analysis.

With the advance of genomics and pharmacogenomics research, people gradually realize that that clinical drug response is partially determined by the genomic alterations and gene expression changes in each particular patient. In the future, we will be interested in utilizing comprehensive -omics data to help predict specific drug response and looking for promising biomarker signatures that could help inform the clinical treatment decision making given a patient's -omic profiles. Through this type of study, we will be able to 1) implement a statistical model that is suitable to identify the drug response in a drug-based manner; and 2) discover biomarkers that can be used to stratify the patient groups when decide which drug should be choose to treat the patient. These observations could greatly help the capability of future patient diagnosis and clinical treatment. In addition, given

different molecular data types for each drug, we tend to investigate which we could achieve the best drug response prediction, such as using a specific data type or a combination of different data types. Furthermore, we tend to evaluate whether additional knowledge can be helpful to further improve the drug response prediction, such as accurate mutation annotation, driver/hotspot mutation information and pathway network knowledge. This will be a good continuation of our efforts in the current thesis and will illustrate whether my thesis can be further helpful in different biological and therapeutic contexts.

Appendix

Supplemental Figures:

(A) DNA Level (genomic/mRNA)

CAG TTG TTG TTG CTG **CTG** CTG GTA
CAG TTG TTG **TTG** CTG **CTG** **CTG** GTA

(B) Protein Level

L **Q** Q Q Q **Q** **Q** Y

Legend:
■ True-position (blue)
■ 5'-aligned (orange)
■ 3'-aligned (green)

Figure S2.1 Illustration of reporting ambiguity for a 3 bp insertion CTG at (A) genomic/mRNA levels and (B) protein level. Note that positional justifications on the protein level are agnostic of the base sequence on the DNA level and may result in variable amount of shifts.

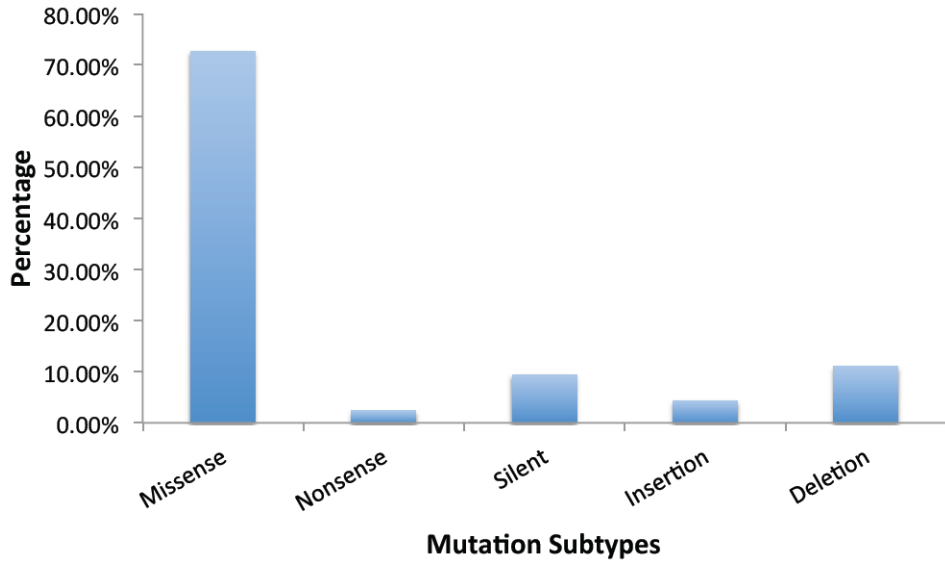


Figure S3.1 The percentage of different mutational subtypes across all defined hotspot mutations. On each hotspot locus, only the mutational subtype that occupies the highest number of mutations was counted.

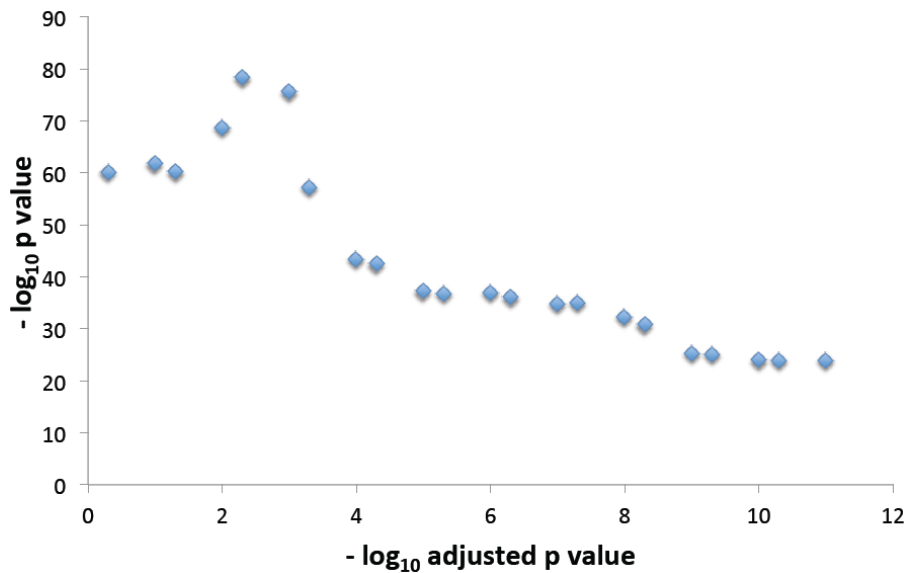


Figure S3.2 The significance of overlap (y-axis, calculated using Fisher exact test) between hotspot-mutation-containing-genes and previously known cancer genes at various adjusted p value cutoffs (x-axis).

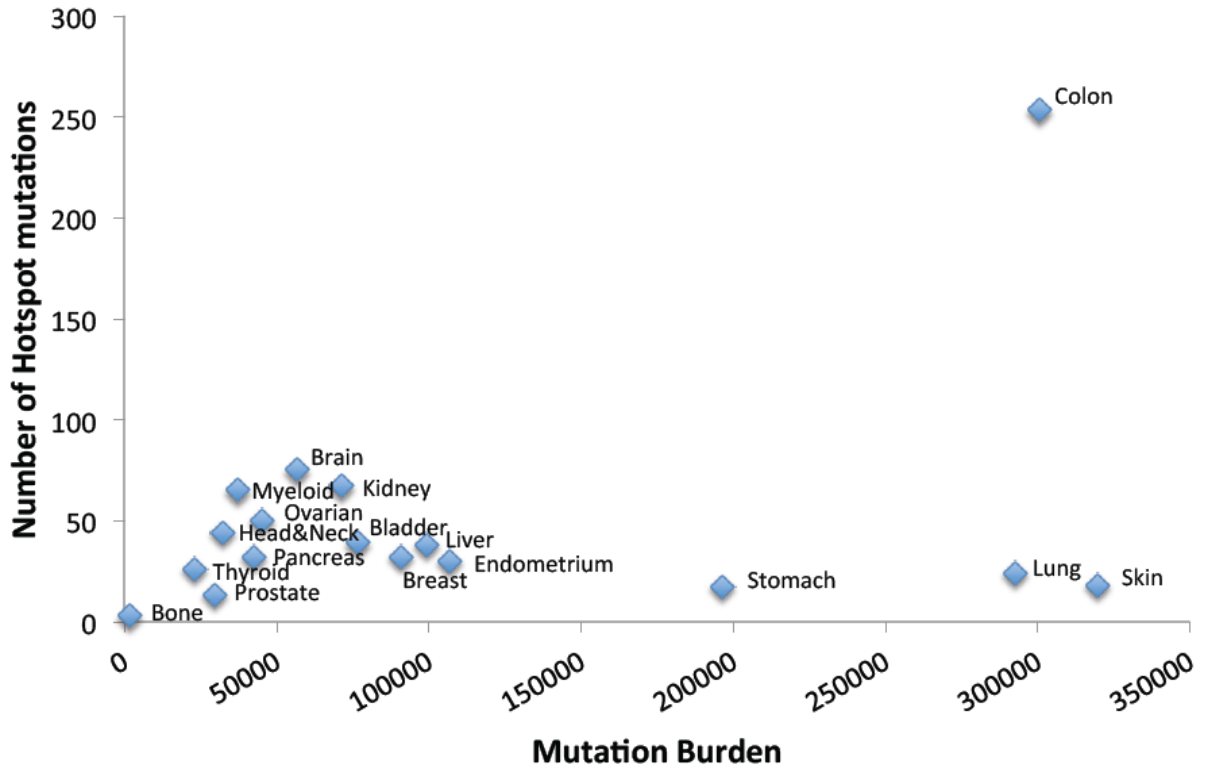


Figure S3.3 Relationship between the number of hotspot mutations and the total number of mutations (mutation burden) in each tumor type.

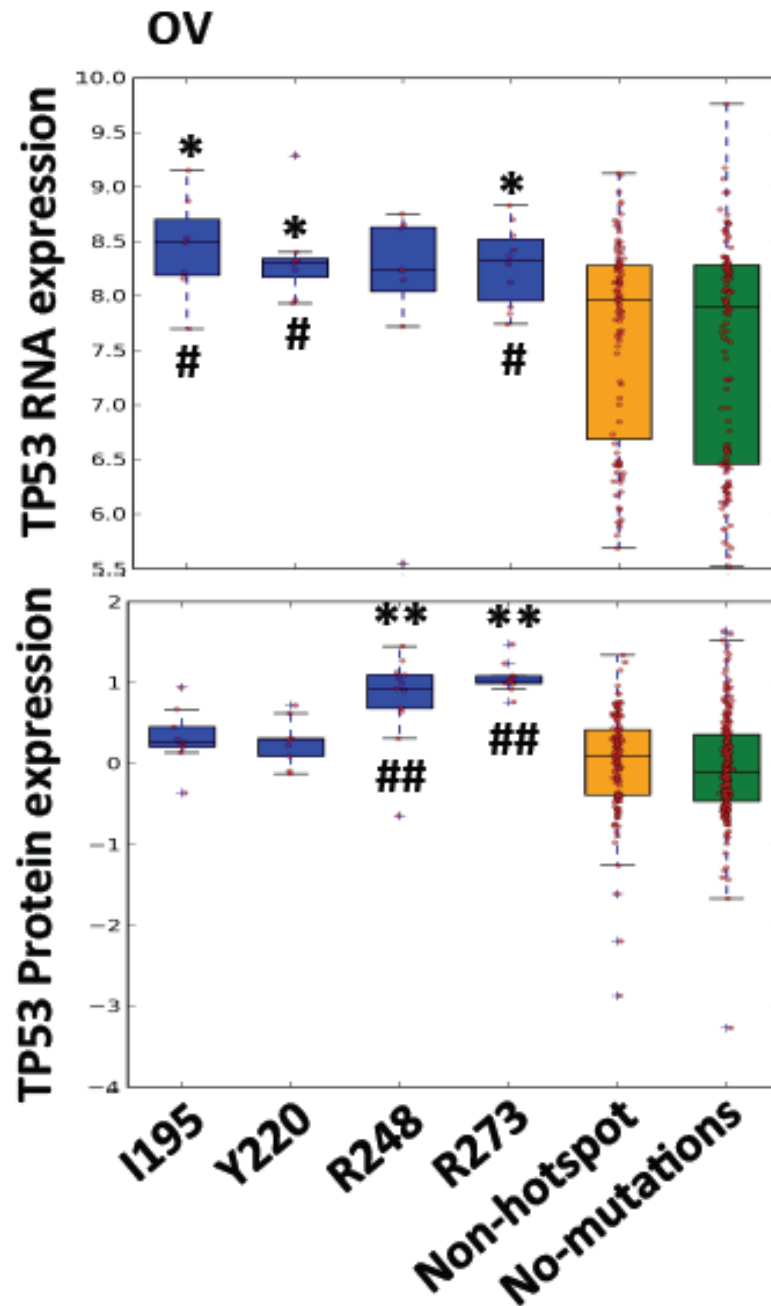


Figure S3.4 Functional implications of hotspot mutations in RNA and protein expression. In OV, tumor samples with missense hotspot mutations (I195, Y220, R248 and R273) in *TP53* show higher *TP53* RNA and protein expression than those with non-hotspot mutations and without *TP53* mutations. * indicates $p < 0.05$ and ** indicates $p < 0.001$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ and ## indicates $p < 0.001$ between samples with specified hotspot mutations and samples without mutations in examined gene.

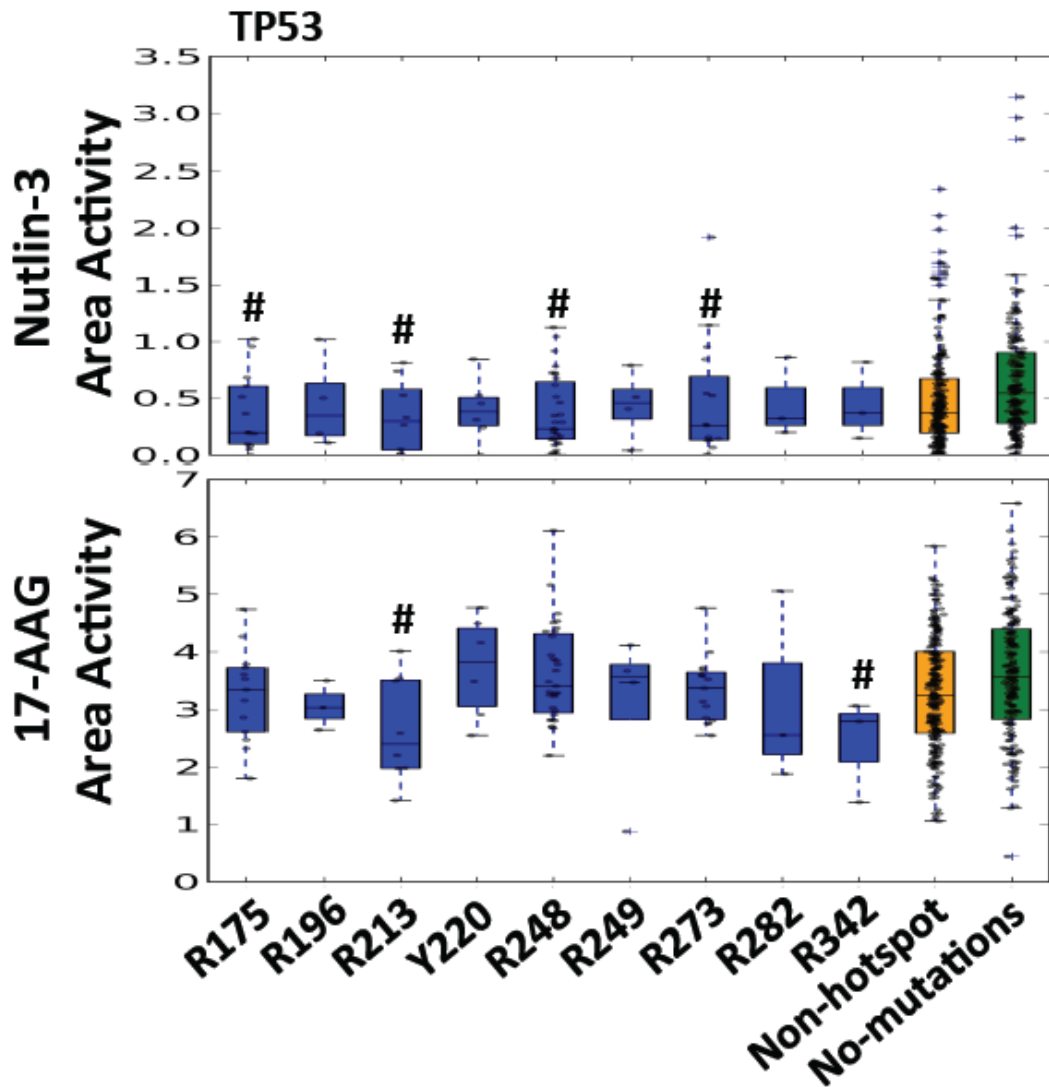


Figure S3.5 Functional implications of hotspot mutations in drug sensitivity. Cancer cells with *TP53* R175, R213, R248 and R273 hotspot mutations show resistant to MDM2 inhibitor (Nutlin-3) compared to those without *TP53* mutations, while cancer cells with *TP53* R213 and R342 nonsense mutations are resistant to HSP90 inhibitor (17-AAG) compared to those without *TP53* mutations. * indicates $p < 0.05$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ between samples with specified hotspot mutations and samples without mutations in examined gene.

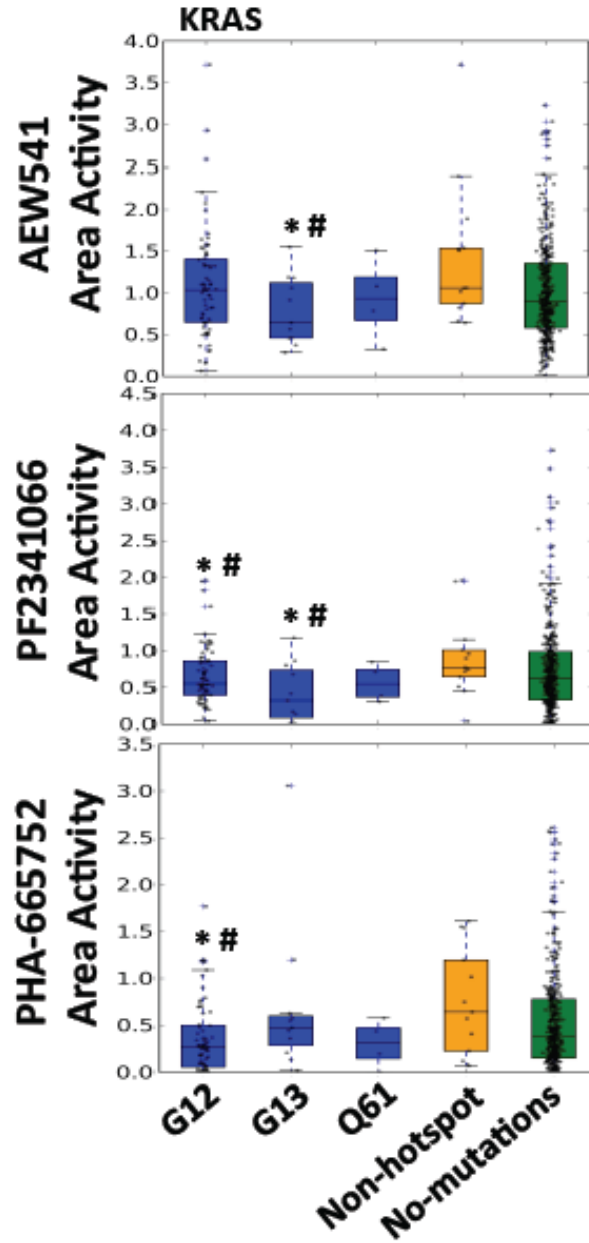


Figure S3.6 Functional implications of hotspot mutations in drug sensitivity. Cancer cells with *KRAS* G13 missense hotspot mutations show resistant to IGF-1R inhibitor (AEW541) compared to those with non-hotspot mutations and without *KRAS* mutations, while cancer cells with *KRAS* G12 and G13 missense mutations are resistant to c-MET inhibitor (PF2341066 and PHA-665752) compared to those with non-hotspot mutations and without *TP53* mutations. * indicates $p < 0.05$ between samples with specified hotspot mutations and samples with non-hotspot mutations in examined gene; # indicates $p < 0.05$ between samples with specified hotspot mutations and samples without mutations in examined gene.

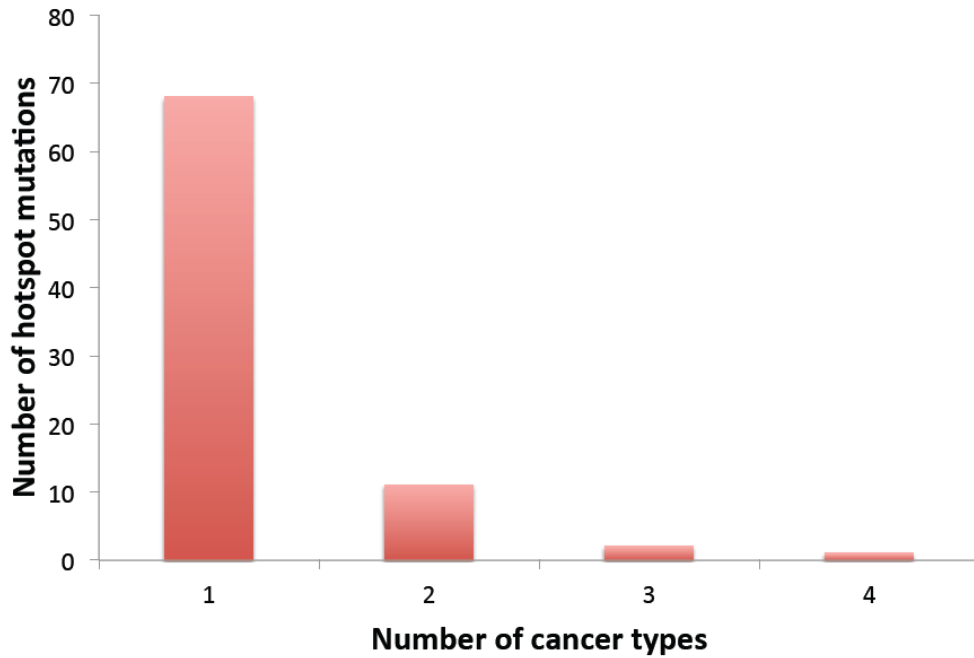


Figure S3.7 Prevalence of hotspot mutations in different TCGA cancer types. 82 hotspot mutations were highly prevalent in one or more cancer types. Most are highly prevalent in only one tumor type, while a few were in two or more tumor types.

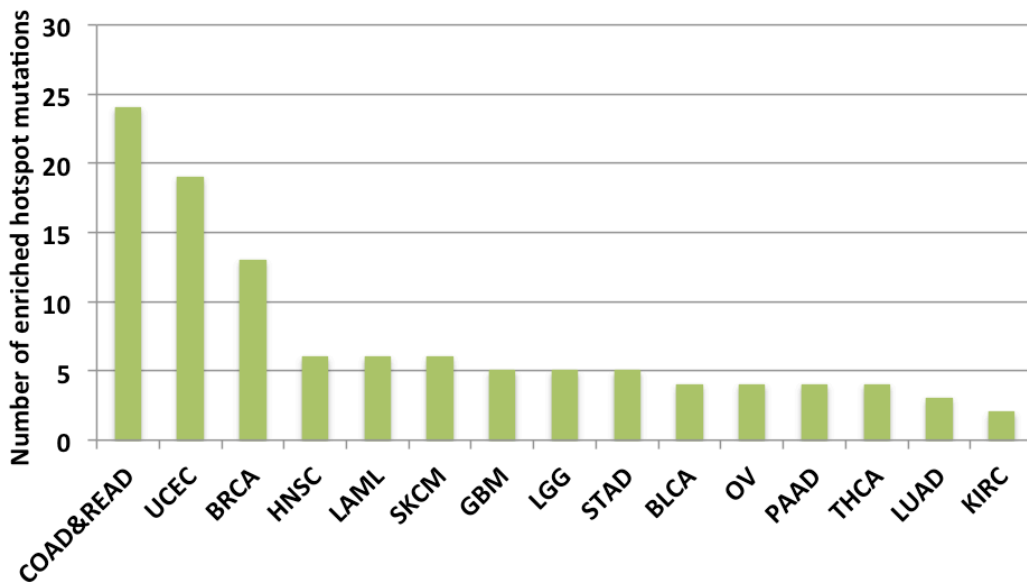


Figure S3.8 Numbers of highly prevalent hotspot mutations in different tumor types

Supplemental Tables:

Table S2.1 Comparing the annotation consistency of different mutation types using TransVar, VEP, ANNOVAR, snpEff and Oncotator. We annotated 964,162 unique SNSs, 3,715 MNSs, 11,761 INSs, 24,595 DELs and 166 BLSs in catalogue of somatic mutations in cancer (COSMIC) and counted if the resulting annotations (gene names, amino acid positions and alterations) match the corresponding protein identifiers in COSMIC.

	TransVar	VEP	ANNOVAR	snpEff	Oncotator
SNS	96.1%	92.9%	91.6%	91.9%	91.8%
MNS	96.8%	92.6%	NA ¹	77.5%	92.6%
INS²	80.6%	75.8%	32.4%	34.6%	38.4%
DEL²	87.8%	77.1%	48.7%	NA ¹	55.6%
BLS	81.9%	75.6%	NA ¹	70.5%	35.5%

¹ Protein level annotations not available. ² TransVar reports both 5'-aligned and 3'-aligned results; VEP only reports the 3'-aligned protein variants with or without --shift_hgvs option, while ANNOVAR, snpEff and Oncotator report only 5'-aligned results.

Table S3.1 Number of samples in 17 tumor types in COSMIC v71

Tumor Type	COSMIC samples *	WGS&WEX ^	Exclude Hyper-mutator #
Bladder	3872	364	358
Bone	704	81	79
Brain	8457	1366	1354
Breast	4994	1152	1140
Colon	29413	694	684
Endometrium	2293	271	260
Head&Neck	3036	710	699
Kidney	3616	879	867
Liver	2448	900	890
Lung	10520	969	951
Myeloid	52500	1344	1336
Ovarian	3378	647	640
Pancreas	5561	800	789
Prostate	953	508	501
Skin	9072	655	650
Stomach	3615	621	613
Thyroid	13967	444	439

* Number of samples that were collected by COSMIC v71; ^ Number of samples that were subjected to either whole genome or whole exome sequencing; # Number of samples after excluding samples that were shown to be hyper-mutated.

Table S3.2 20 mutation subtypes that were included in the statistical modeling of hotspot mutation definition

Mutation_subtype
Missense A/T transition
Missense A/T transversion
Missense non-CpG C/G transition
Missense non-CpG C/G transversion
Missense CpG C/G transition
Missense CpG C/G transversion
Nonsense A/T transition
Nonsense A/T transversion
Nonsense non-CpG C/G transition
Nonsense non-CpG C/G transversion
Nonsense CpG C/G transition
Nonsense CpG C/G transversion
Silent A/T transition
Silent A/T transversion
Silent non-CpG C/G transition
Silent non-CpG C/G transversion
Silent CpG C/G transition
Silent CpG C/G transversion
Insertion
Deletion

Table S3.3 Number of samples available in different TCGA cancer types.

Abbreviation	Tumor_Type	Mutation	RNA	RPPA	Mut_RNA	Mut_RPPA
BRCA	Breast Invasive Carcinoma	772	817	748	752	637
KIRC	Kidney Renal Clear Cell Carcinoma	417	470	455	391	386
THCA	Thyroid Carcinoma	323	426	NA	303	NA
OV	Ovarian Serous Cystadenocarcinoma	316	263	413	163	210
HNSC	Head & Neck Squamous Cell Carcinoma	306	303	213	299	208
GBM	Glioblastoma Multiforme	291	161	216	150	146
SKCM	Skin Cutaneous Melanoma	253	NA	NA	NA	NA
UCEC	Uterine Corpus Endometrioid Carcinoma	248	333	404	239	203
LUAD	Lung Adenocarcinoma	230	353	238	169	135
COAD-READ	Colon & Rectum Adenocarcinoma	224	263	466	217	157
LAML	Acute Myeloid Leukemia	194	173	NA	169	NA
LUSC	Lung Squamous Cell Carcinoma	178	220	196	177	112
LGG	Brain Lower Grade Glioma	170	205	NA	166	NA
STAD	Stomach Adenocarcinoma	151	58	NA	58	NA
KIRP	Kidney Renal Papillary Cell Carcinoma	100	78	NA	77	NA
BLCA	Bladder Urothelial Carcinoma	99	96	128	95	92
PRAD	Prostate Adenocarcinoma	83	142	NA	72	NA
CESC	Cervical Squamous Cell Carcinoma	39	97	NA	38	NA
PAAD	Pancreatic Adenocarcinoma	34	41	NA	19	NA

Note: The number of samples with somatic mutation data, RNA expression data, RPPA data, or both data types (Mut_RNA and Mut_RPPA) in each TCGA cancer type. The data was last updated in November 2014.

Table S3.4 2×2 table of calculating the prevalence of target mutation *B* in samples *A*

Number of A samples with B mutations	Number of A samples without B mutations
Number of non- A samples with B mutations	Number of non- A samples without B mutations

Table S3.5 List of the predicted hotspot mutations in different tumor types based on COSMIC version 71

Gene	aaPos	Tumor_type
ATXN2L	R382	Bladder
CD209	R129	Bladder
CD36	V103	Bladder
COQ10A	L35	Bladder
COQ10A	L50	Bladder
COQ10A	L67	Bladder
ENSG00000196306	D252	Bladder
FGFR3	R248	Bladder
FGFR3	S249	Bladder
FGFRL1	H485	Bladder
FRG2C	A277	Bladder
HRAS	G12	Bladder
HRAS	Q61	Bladder
KDM3A	R157	Bladder
KRAS	G12	Bladder
KRTAP10-1	D159	Bladder
MUC4	D2389	Bladder
MUC6	Y1920	Bladder
NEFH	P655	Bladder
NOTCH2	A21	Bladder
PCDH11X	A89	Bladder
PIK3CA	E542	Bladder
PIK3CA	E545	Bladder
PLXNA1	G36	Bladder
RALGAPA1	Q15	Bladder
RP5-1086D14.3	F374	Bladder
RXRA	S427	Bladder
SENP6	Y599	Bladder
TCF7L2	A87	Bladder
TMPRSS13	A77	Bladder
TMPRSS13	Q78	Bladder
TNKS1BP1	A944	Bladder
TP53	R248	Bladder
TP53	E285	Bladder
ZBTB17	A144	Bladder
ZBTB17	A207	Bladder
ZNF233	L662	Bladder
ZNF83	G267	Bladder

ZNF83	E293	Bladder
IDH1	R132	Bone
OBSCN	L1039	Bone
TTN	R139	Bone
ACTR3B	V136	Brain
AKR1C2	L261	Brain
AKR1C2	Y5	Brain
ANK2	L1097	Brain
ANK2	T3651	Brain
ASPM	K3446	Brain
BRAF	V600	Brain
CDC42BPA	P422	Brain
CDKL5	S603	Brain
CHEK2	Y390	Brain
CHEK2	A392	Brain
CHEK2	R519	Brain
CHEK2	P522	Brain
CHEK2	P536	Brain
COBL	P23	Brain
CTNNB1	S33	Brain
EGFR	A289	Brain
EGFR	G598	Brain
EPHA5	V475	Brain
ERC2	R20	Brain
FGFR1	N546	Brain
FGFR1	K656	Brain
FKBP9	R107	Brain
FKBP9	I274	Brain
FKBP9	I42	Brain
H3F3A	K28	Brain
HIF1A	K213	Brain
HIF1A	D238	Brain
IDH1	R132	Brain
IDH2	R172	Brain
IRS4	A640	Brain
ITGAV	R775	Brain
KLF4	K409	Brain
KPNA2	F17	Brain
KTN1	E687	Brain
MAP3K6	N614	Brain
MAP3K6	N622	Brain
MAX	R60	Brain
NBPF10	K3445	Brain

NBPF10	E3455	Brain
NCOA6	Q269	Brain
PARD3B	G308	Brain
PARG	A584	Brain
PIK3C2B	R41	Brain
PIK3CA	H1047	Brain
PIK3CA	E545	Brain
POTEM	V308	Brain
RGPD8	P1760	Brain
RPSAP58	Q111	Brain
SMAD4	L23	Brain
SOX11	L326	Brain
SYNPO2	R340	Brain
TBK1	E109	Brain
TBK1	T79	Brain
THAP3	E160	Brain
TP53	R158	Brain
TP53	R175	Brain
TP53	H179	Brain
TP53	R196	Brain
TP53	R213	Brain
TP53	Y220	Brain
TP53	G245	Brain
TP53	R248	Brain
TP53	R249	Brain
TP53	R273	Brain
UBBP4	L149	Brain
UBBP4	R73	Brain
WASH3P	G374	Brain
ZNF429	E395	Brain
ZNF429	N426	Brain
ZNF429	K528	Brain
ZNF429	K556	Brain
ZNF429	K568	Brain
ZNF429	R672	Brain
ZNF814	D404	Brain
AKT1	E17	Breast
BCL6B	S244	Breast
BTNL8	V21	Breast
C10orf140	E428	Breast
DDX11	E310	Breast
FLJ42177	V295	Breast
GOLGA6L2	E537	Breast

HSPD1	R24	Breast
KCNN3	L66	Breast
KRAS	G12	Breast
MAGI1	Q421	Breast
MAP3K4	A1199	Breast
NCOA3	Q1255	Breast
NCOR2	Q510	Breast
NR1H2	Q175	Breast
PIK3CA	H1047	Breast
PIK3CA	N345	Breast
PIK3CA	E542	Breast
PIK3CA	E545	Breast
RBMX	P106	Breast
SF3B1	K700	Breast
TBP	Q76	Breast
TP53	G108	Breast
TP53	R175	Breast
TP53	C176	Breast
TP53	H193	Breast
TP53	R196	Breast
TP53	R213	Breast
TP53	Y220	Breast
TP53	R248	Breast
TP53	R273	Breast
USP36	K959	Breast
AATK	G703	Colon
ABCA7	A2045	Colon
ABCF1	K76	Colon
ACTR5	F6	Colon
ADAD2	G44	Colon
ADAM22	P81	Colon
ADAM29	T746	Colon
ALDH2	L189	Colon
ALOX12	P41	Colon
ANKRD30A	P818	Colon
ANTXR2	A357	Colon
ANXA2	C8	Colon
APC	R1114	Colon
APC	R1450	Colon
APC	R876	Colon
ARHGAP5	V474	Colon
ARHGEF33	S582	Colon
ATM	R337	Colon

ATR	I774	Colon
ATXN7	T781	Colon
BAX	E41	Colon
BCAS4	E56	Colon
BCL9L	P1128	Colon
BIRC6	K4503	Colon
BMPR1A	R486	Colon
BMPR2	N583	Colon
BRAF	V600	Colon
BRAT1	R644	Colon
BRCA2	S1682	Colon
C17orf82	L186	Colon
C1orf106	R538	Colon
C8orf55	L63	Colon
C8orf80	N631	Colon
CACNA1H	D2133	Colon
CASP5	R23	Colon
CBWD6	L260	Colon
CCDC43	R216	Colon
CD3G	K71	Colon
CD8B	L9	Colon
CDC42BPA	S344	Colon
CDKN2D	R30	Colon
CEBPB	A147	Colon
CHEK2	S372	Colon
CHEK2	K373	Colon
CHML	S514	Colon
CLK4	R2	Colon
CLOCK	L123	Colon
CSNK1D	S97	Colon
CSNK1E	N172	Colon
CYP21A2	L10	Colon
D2HGDH	R55	Colon
DDHD1	G112	Colon
DDX26B	L120	Colon
DDX4	P48	Colon
DEFB126	P106	Colon
DLC1	K237	Colon
DLC1	R347	Colon
DNAH3	S1608	Colon
DNTTIP1	P13	Colon
DOCK3	P1852	Colon
DOT1L	G1386	Colon

DSC2	K270	Colon
DSG4	A601	Colon
DSPP	S879	Colon
EBPL	F172	Colon
EEF1D	D189	Colon
ENSG00000269210	G98	Colon
ERI1	L16	Colon
ESRP1	N512	Colon
ETHE1	A2	Colon
ETNK2	P10	Colon
FADS3	A18	Colon
FAM18A	F140	Colon
FAM190A	D515	Colon
FAM190A	E611	Colon
FAM194B	E135	Colon
FAM194B	E136	Colon
FAM194B	E138	Colon
FAM194B	Y139	Colon
FBXW7	R465	Colon
FEZ2	P50	Colon
FGFR1	T141	Colon
GLTPD2	D209	Colon
GOT1L1	T415	Colon
GPHB5	R53	Colon
GPRIN2	G237	Colon
GRID2IP	A221	Colon
GRIK2	N849	Colon
GSG2	R82	Colon
GSX1	L78	Colon
GTPBP3	T66	Colon
HAP1	K4	Colon
HCLS1	P368	Colon
HECW1	R1502	Colon
HLA-DQA1	G79	Colon
HSD17B1	G313	Colon
HSPA12B	C627	Colon
IDH3A	S8	Colon
IFI27	A43	Colon
IFITM3	P55	Colon
IGF2R	D1317	Colon
IRF5	P183	Colon
JPH4	A502	Colon
KDM5A	G1200	Colon

KIAA1024	S702	Colon
KIF17	R13	Colon
KIF25	W3	Colon
KLF14	S112	Colon
KLHL30	A213	Colon
KNDC1	A432	Colon
KNDC1	V806	Colon
KRAS	G12	Colon
KRAS	G13	Colon
KRT4	G155	Colon
KRTAP4-3	R26	Colon
KRTAP4-3	Q31	Colon
KRTAP9-1	C153	Colon
KSR1	P291	Colon
LATS2	A324	Colon
LIG1	K152	Colon
LMAN1	E305	Colon
LRRIQ1	R329	Colon
MAGEB2	R23	Colon
MAML2	Q596	Colon
MAPK9	K56	Colon
MEGF11	A102	Colon
MEGF11	A177	Colon
MLL3	Y366	Colon
MLL3	T3698	Colon
MRE11A	F237	Colon
MSH3	K374	Colon
MTX1	T63	Colon
MUC2	T1541	Colon
MUC2	T1542	Colon
MUC4	T113	Colon
MUC6	P1570	Colon
MUC6	P1571	Colon
MUC6	T1911	Colon
MYOM1	R212	Colon
NEFH	E645	Colon
NEFH	A646	Colon
NFATC3	K474	Colon
NFKB2	P423	Colon
NIN	E1559	Colon
NIPBL	K603	Colon
NOTCH3	P1521	Colon
NRAS	G12	Colon

NRAS	Q61	Colon
NTSR2	A54	Colon
OBSCN	A94	Colon
OPRD1	C27	Colon
OR5K4	I301	Colon
OR6C76	K311	Colon
ORAI1	P43	Colon
PANX2	P505	Colon
PANX2	L555	Colon
PCDHA7	L352	Colon
PCDHGC3	N689	Colon
PCSK5	E562	Colon
PDXDC1	A384	Colon
PDXDC1	A407	Colon
PHF2	P988	Colon
PIK3CA	H1047	Colon
PIK3CA	E542	Colon
PIK3CA	E545	Colon
PIK3CA	R88	Colon
PKD1L2	N236	Colon
PKDCC	G17	Colon
PLEC	L1184	Colon
PLEC	A1976	Colon
POLE	V1394	Colon
PPP2R3B	P533	Colon
PRRT4	R391	Colon
PRSS36	S423	Colon
PTPLA	E64	Colon
PTPRD	R584	Colon
PURB	A107	Colon
RAPGEF6	S640	Colon
RASA2	E759	Colon
RASSF5	L16	Colon
RBBP8	K357	Colon
RBP1	P18	Colon
RHPN2	A353	Colon
RIC8A	P210	Colon
RLIM	S501	Colon
RNF145	K23	Colon
RNF145	N27	Colon
SBK2	A298	Colon
SCARF2	R727	Colon
SCRIB	P1450	Colon

SF1	E28	Colon
SLAMF1	S277	Colon
SLC39A3	V236	Colon
SLC3A2	K331	Colon
SMAD4	R361	Colon
SNX13	L652	Colon
SNX18	G204	Colon
SOLH	A840	Colon
SP5	A75	Colon
SPHK1	A34	Colon
SRPR	K170	Colon
STAMBPL1	K405	Colon
STARD3NL	T130	Colon
SVIL	M1863	Colon
SgK069	A298	Colon
SgK223	G350	Colon
SgK493	G160	Colon
TAF1B	N66	Colon
TBC1D8B	K118	Colon
TCERG1	K957	Colon
TCF15	T113	Colon
TCF15	V114	Colon
TEAD2	H295	Colon
TFAM	E148	Colon
TMBIM4	Y174	Colon
TMEM131	G44	Colon
TMEM151B	L332	Colon
TMEM60	A78	Colon
TMPRSS13	A77	Colon
TMPRSS13	Q78	Colon
TNRC6C	N615	Colon
TP53	R175	Colon
TP53	R213	Colon
TP53	R248	Colon
TP53	R273	Colon
TRRAP	K159	Colon
TTK	R854	Colon
TTLL11	A163	Colon
TTN	I2725	Colon
TTN	I2771	Colon
TYSND1	L258	Colon
UBR5	E2121	Colon
UHRF1	P674	Colon

USP35	T411	Colon
USP35	T655	Colon
USP36	K959	Colon
USP6	V148	Colon
VEZT	S74	Colon
VPS13A	R161	Colon
VPS13A	K372	Colon
VSIG10L	M356	Colon
WDTC1	E290	Colon
ZBTB42	R15	Colon
ZDBF2	K1728	Colon
ZNF233	L662	Colon
ZNF365	K399	Colon
ZNF516	V1037	Colon
ZNF518A	T929	Colon
ZNF696	R341	Colon
ZNF717	E818	Colon
ZNF814	G320	Colon
ZNF837	A242	Colon
ZSCAN1	L54	Colon
AGAP10	H228	Endometrium
BCOR	N1425	Endometrium
BCOR	N1459	Endometrium
CTNNB1	S33	Endometrium
CTNNB1	S37	Endometrium
FBXW7	R465	Endometrium
FGFR2	S252	Endometrium
GTF2I	N440	Endometrium
KRAS	G12	Endometrium
KRAS	G13	Endometrium
MAX	H28	Endometrium
OR8I2	A270	Endometrium
PIK3CA	H1047	Endometrium
PIK3CA	G118	Endometrium
PIK3CA	E542	Endometrium
PIK3CA	E545	Endometrium
PIK3CA	Q546	Endometrium
PIK3CA	R88	Endometrium
PIK3CA	R93	Endometrium
PPP2R1A	P179	Endometrium
PTEN	R130	Endometrium
RGPD3	L812	Endometrium
RGPD3	R816	Endometrium

RP11-231C14.2	H27	Endometrium
RSBN1L	L432	Endometrium
SPDYE3	S418	Endometrium
TLK2	R262	Endometrium
TMEM106B	R140	Endometrium
TP53	R248	Endometrium
VIT	G344	Endometrium
ALK	F1174	Head&Neck
ALK	R1275	Head&Neck
ATP2B3	G272	Head&Neck
BRD4	A467	Head&Neck
CACNA1G	S1109	Head&Neck
CHD4	R1353	Head&Neck
CHD4	P799	Head&Neck
DDR1	I141	Head&Neck
DMBT1	G158	Head&Neck
DMBT1	S716	Head&Neck
DNM1	G146	Head&Neck
DOCK7	D185	Head&Neck
ERBB2IP	T1116	Head&Neck
ERBB2IP	S750	Head&Neck
ERBB3	N101	Head&Neck
ESR2	H115	Head&Neck
FAM22D	C197	Head&Neck
FAM83A	H119	Head&Neck
FYB	D285	Head&Neck
HEATR8	N219	Head&Neck
HEATR8	A701	Head&Neck
HRAS	G12	Head&Neck
HRAS	G13	Head&Neck
HRAS	Q61	Head&Neck
MCMDC2	R369	Head&Neck
MLST8	D181	Head&Neck
NACAD	P905	Head&Neck
NFASC	V258	Head&Neck
NUTM2A	C197	Head&Neck
PIK3CA	H1047	Head&Neck
PIK3CA	E542	Head&Neck
PIK3CA	E545	Head&Neck
PTPRD	V225	Head&Neck
QKI	D131	Head&Neck
RP11-368J21.2	A454	Head&Neck
SETDB1	F234	Head&Neck

SLMAP	I195	Head&Neck
TP53	R175	Head&Neck
TP53	R196	Head&Neck
TP53	R213	Head&Neck
TP53	Y220	Head&Neck
TP53	R248	Head&Neck
TP53	R282	Head&Neck
TTN	R62	Head&Neck
ABCA6	V801	Kidney
ADAMTS10	R182	Kidney
AHNAK	P5445	Kidney
APOBEC3H	N15	Kidney
ARHGEF5	E487	Kidney
C22orf31	A261	Kidney
C8orf45	Y111	Kidney
CCKBR	R396	Kidney
CCKBR	R465	Kidney
CDH5	V141	Kidney
CENPH	K138	Kidney
CHEK2	K373	Kidney
COL5A3	P176	Kidney
CUZD1	R355	Kidney
DIEXF	K229	Kidney
F2RL2	C143	Kidney
FCRLA	L196	Kidney
GBP6	K155	Kidney
GCNT2	F107	Kidney
GNLY	L76	Kidney
GOLGA6L10	A469	Kidney
GPR158	K1027	Kidney
IL34	N155	Kidney
KIF6	D714	Kidney
LILRB3	R143	Kidney
LRRK2	I1294	Kidney
MCCC2	A249	Kidney
MED13	P1012	Kidney
MLLT3	S167	Kidney
N4BP2	G560	Kidney
NIPBL	N141	Kidney
NUP155	P990	Kidney
OPRK1	V380	Kidney
OR11H4	P89	Kidney
OR2B11	T244	Kidney

PIP4K2A	R133	Kidney
POTEB	C221	Kidney
PREX2	G233	Kidney
PSKH2	G312	Kidney
RFX3	I519	Kidney
RNF213	S1184	Kidney
RNF213	S3160	Kidney
RP13-996F3.4	V159	Kidney
RPL8	G153	Kidney
RPS12	I94	Kidney
RPS9	L25	Kidney
RYBP	G291	Kidney
SCYL2	Q715	Kidney
SEMA5B	G44	Kidney
SERPINA10	E64	Kidney
SH3GL1	R184	Kidney
SLC11A2	Y357	Kidney
SLC2A5	P496	Kidney
SLC36A2	G317	Kidney
SRBD1	R464	Kidney
SRGAP3	R559	Kidney
TLK2	R262	Kidney
TUBA3E	A126	Kidney
UBA7	L383	Kidney
UBBP4	R73	Kidney
UBE3C	T888	Kidney
UBE4A	N800	Kidney
UPF3A	L91	Kidney
XPNPEP1	S299	Kidney
ZNF462	Q759	Kidney
ZNF605	S82	Kidney
ZNF776	V299	Kidney
ACO1	S174	Liver
AGAP3	G21	Liver
AUTS2	L102	Liver
CACNA1C	T79	Liver
CACNA1G	Q767	Liver
CFLAR	I190	Liver
CNOT4	A188	Liver
COL6A2	G283	Liver
CTNNB1	S45	Liver
CTNNB1	T41	Liver
CTNNB1	S33	Liver

CTNNB1	S37	Liver
CTNNB1	G34	Liver
DACT3	G278	Liver
FGFR1	A21	Liver
FUBP1	Y582	Liver
HSPD1	R24	Liver
HSPD1	G56	Liver
KCNMA1	T254	Liver
KIF4A	R598	Liver
MAGI1	Q410	Liver
MAPK9	G35	Liver
MBD1	G177	Liver
NPM1	G90	Liver
PCDH11X	V38	Liver
PIK3CA	H1047	Liver
SEC16A	R280	Liver
TMEM50B	A139	Liver
TMPRSS13	S70	Liver
TP53	R249	Liver
TP53	Y163	Liver
TP53	G245	Liver
TP53	R213	Liver
TTN	E155	Liver
TTN	S147	Liver
TTN	G156	Liver
YEATS2	L316	Liver
ZNF208	V325	Liver
AGAP10	H228	Lung
ATXN3	K295	Lung
CD6	S52	Lung
CHEK2	K373	Lung
DSPP	D881	Lung
EGFR	L858	Lung
KRAS	G12	Lung
KRAS	Q61	Lung
KRT2	G104	Lung
MUC6	S2085	Lung
PIK3CA	H1047	Lung
PIK3CA	E542	Lung
PIK3CA	E545	Lung
PNKP	P16	Lung
RP11-671M22.1	R443	Lung
RPSAP58	Q111	Lung

TP53	T125	Lung
TP53	R158	Lung
TP53	R248	Lung
TP53	R249	Lung
TP53	R273	Lung
U2AF1	S34	Lung
WASH3P	G374	Lung
ZNF814	S332	Lung
ACSM3	G485	Myeloid
AKIRIN2	Y201	Myeloid
AKT3	K172	Myeloid
ALPP	D64	Myeloid
ANGPTL4	Q143	Myeloid
AP4B1	M100	Myeloid
BRAF	V600	Myeloid
BZRAP1	S1417	Myeloid
CALR	K385	Myeloid
CCND1	Y44	Myeloid
CD79B	Y196	Myeloid
CLCN2	G715	Myeloid
CNDP1	L20	Myeloid
CXCL10	R93	Myeloid
DCTN2	R181	Myeloid
DNMT3A	R882	Myeloid
EIF3D	A43	Myeloid
EIF4G1	K643	Myeloid
EZH2	Y602	Myeloid
EZH2	Y646	Myeloid
FLT3	D835	Myeloid
GRN	L46	Myeloid
HLA-DRB1	H256	Myeloid
IDH1	R132	Myeloid
IDH2	R140	Myeloid
JAK2	V617	Myeloid
KRAS	G12	Myeloid
MICAL2	A388	Myeloid
MPL	W515	Myeloid
MUC6	P1965	Myeloid
MYD88	L265	Myeloid
NEDD9	D178	Myeloid
NEFH	P655	Myeloid
NIT1	R33	Myeloid
NOTCH1	P2514	Myeloid

NPM1	W288	Myeloid
NPR2	K683	Myeloid
NRAS	Q61	Myeloid
NSMAF	R850	Myeloid
NSMAF	R881	Myeloid
ORM1	R86	Myeloid
P4HA1	R118	Myeloid
PARP4	A1096	Myeloid
PEX6	V788	Myeloid
PFKFB3	K147	Myeloid
PKD1L2	N236	Myeloid
PLIN4	A811	Myeloid
PLIN4	A883	Myeloid
PSME2	V17	Myeloid
PSRC1	D7	Myeloid
PYGM	A610	Myeloid
RECQL4	L1132	Myeloid
RPE65	V287	Myeloid
RPS16	V100	Myeloid
SF3B1	K700	Myeloid
SF3B1	G742	Myeloid
SH3BP1	R229	Myeloid
TP53	R248	Myeloid
TUBA1B	D76	Myeloid
U2AF1	Q157	Myeloid
U2AF1	S34	Myeloid
USP3	E443	Myeloid
XPO1	E571	Myeloid
ZNF217	R629	Myeloid
ZNF98	T451	Myeloid
ABL1	S349	Ovarian
ANKRD36C	V306	Ovarian
ARFIP1	Q373	Ovarian
BAHD1	W653	Ovarian
BEND5	S173	Ovarian
BEND5	S342	Ovarian
CDSN	H260	Ovarian
CEP152	G1415	Ovarian
CHEK2	R519	Ovarian
CHRNA5	T148	Ovarian
CLASP1	A1272	Ovarian
CNGA2	I524	Ovarian
CYFIP2	D543	Ovarian

DDX23	R351	Ovarian
DHX8	H768	Ovarian
FCAMR	T88	Ovarian
FUBP1	G358	Ovarian
HHLA2	V145	Ovarian
KIAA0355	G168	Ovarian
KLK2	Q39	Ovarian
KRAS	G12	Ovarian
LRRC34	V374	Ovarian
LRRIQ4	G151	Ovarian
MYCBP2	S3024	Ovarian
MYOF	T92	Ovarian
NFXL1	R815	Ovarian
NOXA1	A300	Ovarian
PEAR1	P62	Ovarian
PKD1L2	N236	Ovarian
PLEKHG6	G68	Ovarian
POLDIP3	A94	Ovarian
POLR1C	V193	Ovarian
PREX1	K872	Ovarian
PTGER3	H314	Ovarian
SCUBE3	P821	Ovarian
SIK2	T878	Ovarian
SI	W1086	Ovarian
TP53	Y103	Ovarian
TP53	R175	Ovarian
TP53	I195	Ovarian
TP53	Y220	Ovarian
TP53	N239	Ovarian
TP53	R248	Ovarian
TP53	R273	Ovarian
WDR37	R303	Ovarian
ZBBX	L299	Ovarian
ZFP112	H827	Ovarian
ZFR2	P441	Ovarian
ZNF510	T356	Ovarian
ZYG11A	N606	Ovarian
ADRA1A	R235	Pancreas
ATP2B3	E151	Pancreas
BMPR1A	D414	Pancreas
CACNA1C	N348	Pancreas
ERBB2IP	R1037	Pancreas
GALR3	L256	Pancreas

GNAS	R201	Pancreas
GNAS	R844	Pancreas
GRIA4	R368	Pancreas
IFT43	L44	Pancreas
KLKB1	I199	Pancreas
KRAS	G12	Pancreas
KRAS	Q61	Pancreas
MSL2	S419	Pancreas
NTM	T166	Pancreas
NTNG1	S289	Pancreas
PCSK5	S487	Pancreas
PRKCB	G88	Pancreas
SF3B1	K700	Pancreas
SHPRH	Q798	Pancreas
SLC26A5	E30	Pancreas
SLC8A3	D72	Pancreas
TP53	R175	Pancreas
TP53	C176	Pancreas
TP53	R196	Pancreas
TP53	R213	Pancreas
TP53	G245	Pancreas
TP53	R248	Pancreas
TP53	R273	Pancreas
TP53	R342	Pancreas
TTN	Q369	Pancreas
USP20	V422	Pancreas
AGAP10	H228	Prostate
ENSG00000103472	W375	Prostate
FAM129C	G603	Prostate
MCMDC2	T314	Prostate
MED12	L1224	Prostate
NBPF10	K3445	Prostate
NCOA6	Q269	Prostate
RGPD8	P1760	Prostate
SPOP	F133	Prostate
SYT16	R131	Prostate
UBBP4	L149	Prostate
ZDHHC11	A303	Prostate
ZNF91	R333	Prostate
AGAP10	H228	Skin
AGAP10	M293	Skin
BRAF	V600	Skin
C15orf23	S24	Skin

CTAGE1	F739	Skin
DDX11	R167	Skin
DTNA	R255	Skin
IDH1	R132	Skin
LEPR	I740	Skin
NLRP1	R698	Skin
NRAS	Q61	Skin
PCDHGA1	R293	Skin
RAC1	P29	Skin
RGPD8	P1760	Skin
RGS7	R44	Skin
TP53	R213	Skin
TRRAP	S722	Skin
WASH3P	G374	Skin
CNBD1	L396	Stomach
ERBB3	V104	Stomach
KRAS	G12	Stomach
KRAS	G13	Stomach
PGM5	I98	Stomach
PIK3CA	H1047	Stomach
PIK3CA	E542	Stomach
PIK3CA	E545	Stomach
RIMS2	S231	Stomach
TP53	R175	Stomach
TP53	C176	Stomach
TP53	R196	Stomach
TP53	R213	Stomach
TP53	G245	Stomach
TP53	R248	Stomach
TP53	R273	Stomach
TP53	R282	Stomach
ADRA1A	A104	Thyroid
BRAF	V600	Thyroid
CACNA1A	H2219	Thyroid
CBWD6	A154	Thyroid
CLIP1	L271	Thyroid
DIDO1	T580	Thyroid
FBXW10	L261	Thyroid
GOLGA8B	A488	Thyroid
HAVCR1	L34	Thyroid
HRAS	Q61	Thyroid
KREMEN1	P4	Thyroid
MAML3	Q491	Thyroid

MAML3	Q493	Thyroid
MLL3	I882	Thyroid
MUC16	P328	Thyroid
MUC6	N1519	Thyroid
NEFH	E645	Thyroid
NOTCH1	T349	Thyroid
NRAS	Q61	Thyroid
OBSCN	A998	Thyroid
PI4KA	P1714	Thyroid
RAB11FIP3	A30	Thyroid
RGPD8	P1121	Thyroid
RP11-578F21.5	Q441	Thyroid
SCN5A	D1978	Thyroid
TMPRSS13	S70	Thyroid

Table S3.6 A full list of the hotspot mutations that were highly prevalent in specific cancer types in TCGA

Cancer_Type	Gene	aa_Pos	Mut_Percent	adj_Pvalue
thca	AGAP10	H228	2.80%	7.35E-04
skcm	AGAP10	M293	2.80%	1.62E-08
brca	AKT1	E17	2.30%	7.97E-09
coadread	APC	R1114	2.70%	2.54E-06
coadread	APC	R1450	8.40%	4.96E-22
coadread	APC	R876	5.30%	4.01E-14
ucec	BCOR	N1459	3.20%	7.14E-09
stad	BMPR2	N583	2.00%	1.49E-04
thca	BRAF	V600	56.20%	3.31E-145
skcm	BRAF	V600	37.80%	5.74E-49
coadread	BTNL8	V21	1.30%	9.06E-03
skcm	C15orf23	S24	2.80%	1.62E-08
stad	CCDC43	R216	2.00%	1.49E-04
paad	CD209	R129	8.60%	2.98E-04
blca	CD209	R129	4.00%	3.27E-04
stad	CNBD1	L396	2.00%	4.68E-04
ucec	CTNNB1	G34	2.00%	1.88E-04
lgg	DDX11	R167	2.90%	1.23E-03
laml	DNMT3A	R882	14.90%	4.24E-38
stad	DOCK3	P1852	6.60%	2.47E-13
coadread	ERBB3	V104	2.20%	5.63E-04
stad	ESRP1	N512	3.30%	3.25E-07
stad	FAM18A	F140	2.00%	1.49E-04
coadread	FBXW7	R465	5.30%	6.20E-09
ucec	FBXW7	R465	2.80%	9.17E-04
ucec	FGFR2	S252	3.60%	6.31E-11
blca	FGFR3	S249	4.00%	7.02E-05
laml	FLT3	D835	8.20%	1.94E-21
brca	GOLGA6L2	E537	0.90%	1.27E-04
stad	GTF2I	N440	2.00%	1.49E-04
hnsk	HRAS	G12	2.00%	1.36E-05
thca	HRAS	Q61	3.70%	9.05E-10
lgg	IDH1	R132	76.60%	1.28E-169
laml	IDH1	R132	9.70%	8.35E-04
coadread	KRAS	G12	29.80%	8.22E-39
luad	KRAS	G12	22.90%	7.20E-24
paad	KRAS	G12	57.10%	1.17E-17
ucec	KRAS	G12	14.50%	1.24E-09

coadread	KRAS	G13	4.90%	4.80E-07
ucec	KRAS	G13	3.60%	7.62E-05
stad	KRAS	G13	3.90%	1.01E-03
paad	KRAS	Q61	11.40%	3.44E-05
brca	MAGI1	Q421	0.80%	4.95E-04
ucec	MAX	H28	1.60%	4.55E-05
gbm	NBPF10	E3455	2.10%	1.37E-03
kirc	NEFH	P655	0.70%	2.16E-03
laml	NPM1	W288	25.60%	1.58E-69
brca	NR1H2	Q175	1.80%	2.44E-10
coadread	NRAS	G12	2.70%	3.80E-04
skcm	NRAS	Q61	23.60%	1.89E-44
thca	NRAS	Q61	8.00%	7.74E-07
skcm	PCDHGA1	R293	2.00%	3.48E-06
stad	PGM5	I98	10.50%	5.36E-22
brca	PIK3CA	E542	4.10%	5.20E-06
ucec	PIK3CA	E542	5.20%	1.04E-03
hnsc	PIK3CA	E542	4.60%	2.27E-03
brca	PIK3CA	E545	6.50%	2.92E-06
hnsc	PIK3CA	E545	6.80%	2.43E-03
ucec	PIK3CA	G118	2.40%	1.97E-04
brca	PIK3CA	H1047	15.40%	6.47E-53
ucec	PIK3CA	H1047	8.00%	2.75E-03
brca	PIK3CA	N345	1.70%	1.63E-04
ucec	PIK3CA	Q546	4.40%	5.53E-07
ucec	PIK3CA	R88	4.40%	4.50E-08
ucec	PIK3CA	R93	2.40%	1.25E-06
ucec	PTEN	R130	23.30%	1.09E-57
skcm	RAC1	P29	3.90%	2.25E-10
skcm	RGS7	R44	2.40%	2.87E-05
hnsc	RPSAP58	Q111	3.90%	5.20E-06
blca	RXRA	S427	5.00%	4.26E-08
brca	SF3B1	K700	1.00%	1.08E-04
blca	SLAMF1	S277	3.00%	6.19E-04
stad	SLC3A2	K331	2.60%	3.15E-05
coadread	SMAD4	R361	3.10%	7.01E-05
stad	SMAD4	R361	3.90%	7.97E-05
stad	STAMBPL1	K405	2.00%	1.49E-04
stad	TAF1B	N66	2.60%	7.88E-05
blca	TP53	E285	3.00%	2.96E-03
brca	TP53	G108	0.60%	5.97E-03
ov	TP53	G245	2.80%	6.70E-03
ov	TP53	I195	2.80%	1.27E-04

lusc	TP53	R158	4.50%	4.48E-06
coadread	TP53	R175	7.60%	3.44E-07
hnsc	TP53	R213	2.90%	8.65E-03
blca	TP53	R248	10.00%	1.66E-04
ov	TP53	R248	5.40%	1.03E-03
luad	TP53	R249	3.00%	1.95E-04
lgg	TP53	R273	17.50%	2.64E-16
ov	TP53	R273	6.60%	2.30E-04
ov	TP53	Y220	3.50%	1.28E-04
skcm	TRRAP	S722	1.60%	4.83E-05
laml	U2AF1	S34	3.60%	7.26E-06
luad	U2AF1	S34	2.20%	1.89E-03
thca	UBBP4	L149	2.20%	6.23E-03
stad	UBR5	E2121	5.30%	1.72E-10
brca	USP36	K959	1.30%	2.17E-05
stad	ZNF365	K399	2.00%	1.49E-04
gbm	ZNF814	D404	3.10%	7.15E-05
blca	ZNF814	D404	4.00%	5.27E-03
paad	ZNF91	R333	14.30%	1.31E-07

BIBLIOGRAPHY

1. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci U S A* 1977, **74**:560-564.
2. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94**:441-448.
3. Weissenbach J: **Human genome project: past, present, future.** *Ernst Schering Res Found Workshop* 2002:1-9.
4. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526**:68-74.
5. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
6. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**:1297-1303.
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.
8. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy**

- number alteration discovery in cancer by exome sequencing.** *Genome Res* 2012, **22**:568-576.
9. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.** *Nat Biotechnol* 2013, **31**:213-219.
 10. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.** *Bioinformatics* 2009, **25**:2865-2871.
 11. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC: **Accurate de novo and transmitted indel detection in exome-capture data using microassembly.** *Nat Methods* 2014, **11**:1033-1036.
 12. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6**:677-681.
 13. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: **DELLY: structural variant discovery by integrated paired-end and split-read analysis.** *Bioinformatics* 2012, **28**:i333-i339.
 14. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF: **Exome sequencing-based copy-number**

- variation and loss of heterozygosity detection: ExomeCNV.**
Bioinformatics 2011, **27**:2648-2654.
15. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, Romeo G, Pippucci T, De Bellis G, Abbate R, Gensini GF: **EXCAVATOR: detecting copy number variants from whole-exome sequencing data.** *Genome Biol* 2013, **14**:R120.
 16. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringer KL: **CONTRA: copy number analysis for targeted resequencing.** *Bioinformatics* 2012, **28**:1307-1313.
 17. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38**:e164.
 18. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26**:2069-2070.
 19. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.** *Fly (Austin)* 2012, **6**:80-92.
 20. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G: **Oncotator: cancer variant annotation tool.** *Hum Mutat* 2015, **36**:E2423-2429.

21. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM: **Ensembl 2012**. *Nucleic Acids Res* 2012, **40**:D84-90.
22. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy**. *Nucleic Acids Res* 2012, **40**:D130-135.
23. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The UCSC Genome Browser database: 2014 update**. *Nucleic Acids Res* 2014, **42**:D764-770.
24. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters

- N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**:1760-1774.
25. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K: **CanDrA: cancer-specific driver missense mutation annotation with optimized features.** *PLoS One* 2013, **8**:e77945.
26. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer Res* 2009, **69**:6660-6667.
27. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39**:e118.
28. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362**:1181-1191.
29. Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, Wu R, Chen C, Li X, Zhou L, He M, Li Z, Sun X, Jia W, Chen J, Yang S, Zhou F, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Liu C, Jiang H, Wang Y, Zheng H, Sun L, Liu X,

- Jiang Z, Feng D, Chen J, Wu S, Zou J, Zhang Z, Yang R, Zhao J, Xu C, Yin W, Guan Z, Ye J, Zhang H, Li J, Kristiansen K, Nickerson ML, Theodorescu D, Li Y, Zhang X, Li S, Wang J, Yang H, Wang J, Cai Z: **Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder.** *Nat Genet* 2011, **43**:875-878.
30. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR: **Initial genome sequencing and analysis of multiple myeloma.** *Nature* 2011, **471**:467-472.
31. Cancer Genome Atlas Research N: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**:609-615.
32. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz

- G, Rubin MA, Garraway LA: **The genomic complexity of primary human prostate cancer.** *Nature* 2011, **470**:214-220.
33. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148**:1293-1307.
34. **Personal Genomics.** [/http://en.wikipedia.org/wiki/Personal_genomics%5D](http://en.wikipedia.org/wiki/Personal_genomics%5D).
35. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57-74.
36. Long GV, Menzies AM, Nagrial AM, Haydu LE, Hamilton AL, Mann GJ, Hughes TM, Thompson JF, Scolyer RA, Kefford RF: **Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma.** *J Clin Oncol* 2011, **29**:1239-1246.
37. Ascierto PA, Kirkwood JM, Grob JJ, Simeone E, Grimaldi AM, Maio M, Palmieri G, Testori A, Marincola FM, Mozzillo N: **The role of BRAF V600 mutation in melanoma.** *J Transl Med* 2012, **10**:85.
38. Elias MH, Baba AA, Azlan H, Rosline H, Sim GA, Padmini M, Fadilah SA, Ankathil R: **BCR-ABL kinase domain mutations, including 2 novel**

- mutations in imatinib resistant Malaysian chronic myeloid leukemia patients-Frequency and clinical outcome.** *Leuk Res* 2014, **38**:454-459.
39. Pollock PM, Gartside MG, Dejeza LC, Powell MA, Mallon MA, Davies H, Mohammadi M, Futreal PA, Stratton MR, Trent JM, Goodfellow PJ: **Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes.** *Oncogene* 2007, **26**:7158-7162.
40. Yamaguchi F, Fukuchi K, Yamazaki Y, Takayasu H, Tazawa S, Tateno H, Kato E, Wakabayashi A, Fujimori M, Iwasaki T, Hayashi M, Tsuchiya Y, Yamashita J, Takeda N, Kokubu F: **Acquired resistance L747S mutation in an epidermal growth factor receptor-tyrosine kinase inhibitor-naive patient: A report of three cases.** *Oncol Lett* 2014, **7**:357-360.
41. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM: **Ensembl 2014.** *Nucleic Acids Res* 2014, **42**:D749-755.
42. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA,

- Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: an update on mammalian reference sequences.** *Nucleic Acids Res* 2014, **42**:D756-763.
43. McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P: **Choice of transcripts and software has a large effect on variant annotation.** *Genome Med* 2014, **6**:26.
44. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE: **Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker.** *Hum Mutat* 2008, **29**:6-13.
45. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**:153-158.

46. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**:1108-1113.
47. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L: **MuSiC: identifying mutational significance in cancer genomes.** *Genome Res* 2012, **22**:1589-1598.
48. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CW, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES,

- Getz G: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499**:214-218.
49. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.** *Bioinformatics* 2013, **29**:2238-2244.
50. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A: **A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces.** *PLoS Comput Biol* 2015, **11**:e1004518.
51. Cully M, You H, Levine AJ, Mak TW: **Beyond PTEN mutations: the PI3K pathway as an integrator of multiple inputs during tumorigenesis.** *Nat Rev Cancer* 2006, **6**:184-192.
52. Cheung LW, Yu S, Zhang D, Li J, Ng PK, Panupinthu N, Mitra S, Ju Z, Yu Q, Liang H, Hawke DH, Lu Y, Broaddus RR, Mills GB: **Naturally occurring neomorphic PIK3R1 mutations activate the MAPK pathway, dictating therapeutic response to MAPK pathway inhibitors.** *Cancer Cell* 2014, **26**:479-494.
53. Solit DB, Garraway LA, Pratilas CA, Sawai A, Getz G, Basso A, Ye Q, Lobo JM, She Y, Osman I, Golub TR, Sebolt-Leopold J, Sellers WR, Rosen N: **BRAF mutation predicts sensitivity to MEK inhibition.** *Nature* 2006, **439**:358-362.
54. Holderfield M, Deuker MM, McCormick F, McMahon M: **Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond.** *Nat Rev Cancer* 2014, **14**:455-467.

55. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N: **Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation.** *Genome Med* 2012, **4**:89.
56. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.** *Am J Hum Genet* 2011, **88**:440-449.
57. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Jr., Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW: **An integrated genomic analysis of human glioblastoma multiforme.** *Science* 2008, **321**:1807-1812.
58. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR: **Cosmic 2005.** *Br J Cancer* 2006, **94**:318-322.
59. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248-249.
60. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4**:1073-1081.

61. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**:346-352.
62. Li Q, Seo J-H, Stranger B, McKenna A, Pe'er I, LaFramboise T, Brown M, Tyekucheva S, Freedman ML: **Integrative eQTL-based analyses reveal the biology of breast cancer risk loci.** *Cell* 2013, **152**:633-641.
63. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M: **Effect of predicted protein-truncating genetic variants on the human transcriptome.** *Science* 2015, **348**:666-669.
64. Wu WW, Wang G, Baek SJ, Shen RF: **Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D gel- or LC-MALDI TOF/TOF.** *J Proteome Res* 2006, **5**:651-658.
65. Spurrier B, Ramalingam S, Nishizuka S: **Reverse-phase protein lysate microarrays for cell signaling analysis.** *Nat Protoc* 2008, **3**:1796-1808.
66. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, Beroukhi R, Milner DA, Granter SR, Du J, Lee C, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson ML, Fisher DE, Sellers WR: **Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma.** *Nature* 2005, **436**:117-122.
67. Staunton JE, Slonim DK, Collier HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR:

- Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci U S A* 2001, **98**:10787-10792.
68. Lin WM, Baker AC, Beroukhim R, Winckler W, Feng W, Marmion JM, Laine E, Greulich H, Tseng H, Gates C, Hodi FS, Dranoff G, Sellers WR, Thomas RK, Meyerson M, Golub TR, Dummer R, Herlyn M, Getz G, Garraway LA: **Modeling genomic diversity and tumor dependency in malignant melanoma.** *Cancer Res* 2008, **68**:664-673.
69. Sos ML, Michel K, Zander T, Weiss J, Frommolt P, Peifer M, Li D, Ullrich R, Koker M, Fischer F, Shimamura T, Rauh D, Mermel C, Fischer S, Stuckrath I, Heynck S, Beroukhim R, Lin W, Winckler W, Shah K, LaFramboise T, Moriarty WF, Hanna M, Tolosi L, Rahnenfuhrer J, Verhaak R, Chiang D, Getz G, Hellmich M, Wolf J, Girard L, Peyton M, Weir BA, Chen TH, Greulich H, Barretina J, Shapiro GI, Garraway LA, Gazdar AF, Minna JD, Meyerson M, Wong KK, Thomas RK: **Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions.** *J Clin Invest* 2009, **119**:1727-1740.
70. Inamdar GS, Madhunapantula SV, Robertson GP: **Targeting the MAPK pathway in melanoma: why some approaches succeed and other fail.** *Biochem Pharmacol* 2010, **80**:624-637.
71. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C,

- Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature* 2012, **483**:603-607.
72. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ: **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.** *Nucleic Acids Res* 2013, **41**:D955-961.
73. Ng PC, Henikoff S: **Predicting the effects of amino acid substitutions on protein function.** *Annu Rev Genomics Hum Genet* 2006, **7**:61-80.
74. Dogruluk T, Tsang YH, Espitia M, Chen F, Chen T, Chong Z, Appadurai V, Dogruluk A, Eterovic AK, Bonnen PE, Creighton CJ, Chen K, Mills GB, Scott KL: **Identification of Variant-Specific Functions of PIK3CA by Rapid Phenotyping of Rare Mutations.** *Cancer Res* 2015, **75**:5341-5354.
75. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22**:1036-1046.
76. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Research* 2004, **32**:D493-D496.

77. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, DiCuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu WY, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes (vol 19, pg 1316, 2009).** *Genome Research* 2009, **19**:1506-1506.
78. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** *Genome Biology* 2006, **7**.
79. Liu S, Meric-Bernstam F, Parinyanitikul N, Wang B, Eterovic AK, Zheng X, Gagea M, Chavez-MacGregor M, Ueno NT, Lei X, Zhou W, Nair L, Tripathy D, Brown PH, Hortobagyi GN, Chen K, Mendelsohn J, Mills GB, Gonzalez-Angulo AM: **Functional consequence of the MET-T1010I polymorphism in breast cancer.** *Oncotarget* 2015, **6**:2604-2614.
80. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G: **Oncotator: Cancer Variant Annotation Tool.** *Hum Mutat* 2015.
81. Slotkin W, Nishikura K: **Adenosine-to-inosine RNA editing and human disease.** *Genome Med* 2013, **5**:105.

82. Chen L, Li Y, Lin CH, Chan TH, Chow RK, Song Y, Liu M, Yuan YF, Fu L, Kong KL, Qi L, Li Y, Zhang N, Tong AH, Kwong DL, Man K, Lo CM, Lok S, Tenen DG, Guan XY: **Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma.** *Nat Med* 2013, **19**:209-216.
83. Ramaswami G, Li JB: **RADAR: a rigorously annotated database of A-to-I RNA editing.** *Nucleic Acids Res* 2014, **42**:D109-113.
84. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.** *Nucleic Acids Res* 2012, **40**:D261-270.
85. Linardou H, Dahabreh IJ, Bafaloukos D, Kosmidis P, Murray S: **Somatic EGFR mutations and efficacy of tyrosine kinase inhibitors in NSCLC.** *Nat Rev Clin Oncol* 2009, **6**:352-366.
86. Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, Johnson BE, Eck MJ, Tenen DG, Halmos B: **EGFR mutation and resistance of non-small-cell lung cancer to gefitinib.** *N Engl J Med* 2005, **352**:786-792.
87. Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MF, Kris MG, Varmus H: **Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain.** *PLoS Med* 2005, **2**:e73.

88. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC)**. *Curr Protoc Hum Genet* 2008, **Chapter 10**:Unit 10 11.
89. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**:177-183.
90. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N: **Comprehensive identification of mutational cancer driver genes across 12 tumor types**. *Sci Rep* 2013, **3**:2650.
91. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: **Discovery and saturation analysis of cancer genes across 21 tumour types**. *Nature* 2014, **505**:495-501.
92. Fisher RA: *Statistical methods for research workers*. Edinburgh, London,: Oliver and Boyd; 1925.
93. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**:289-300.
94. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang JY, Broom BM, Verhaak RG, Kane DW, Wakefield C, Weinstein JN, Mills GB, Liang H: **TCPA: a**

- resource for cancer functional proteomics data.** *Nat Methods* 2013, **10**:1046-1047.
95. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45**:1113-1120.
96. Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C: **Analysis of microRNA-target interactions across diverse cancer types.** *Nat Struct Mol Biol* 2013, **20**:1325-1332.
97. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++.** *PLoS Comput Biol* 2010, **6**:e1001025.
98. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP: **CpG island methylator phenotype in colorectal cancer.** *Proc Natl Acad Sci U S A* 1999, **96**:8681-8686.
99. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL: **Genetic alterations during colorectal-tumor development.** *N Engl J Med* 1988, **319**:525-532.
100. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers.** *Nat Genet* 2013, **45**:1127-1133.
101. Oren M, Rotter V: **Mutant p53 gain-of-function in cancer.** *Cold Spring Harb Perspect Biol* 2010, **2**:a001107.

102. Kang HJ, Chun SM, Kim KR, Sohn I, Sung CO: **Clinical relevance of gain-of-function mutations of p53 in high-grade serous ovarian carcinoma.** *PLoS One* 2013, **8**:e72609.
103. Yuan TL, Cantley LC: **PI3K pathway alterations in cancer: variations on a theme.** *Oncogene* 2008, **27**:5497-5510.
104. Shangary S, Wang S: **Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy.** *Annu Rev Pharmacol Toxicol* 2009, **49**:223-241.
105. Lin K, Rockcliffe N, Johnson GG, Sherrington PD, Pettitt AR: **Hsp90 inhibition has opposing effects on wild-type and mutant p53 and induces p21 expression and cytotoxicity irrespective of p53/ATM status in chronic lymphocytic leukaemia cells.** *Oncogene* 2008, **27**:2445-2455.
106. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H, Subbannayya Y, Goel R, Jacob HK, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R, Ramachandra YL, Rahiman BA, Prasad TS, Lin JX, Houtman JC, Desiderio S, Renauld JC, Constantinescu SN, Ohara O, Hirano T, Kubo M, Singh S, Khatri P, Draghici S, Bader GD, Sander C, Leonard WJ, Pandey A: **NetPath: a public resource of curated signal transduction pathways.** *Genome Biol* 2010, **11**:R3.
107. Whitmarsh AJ, Davis RJ: **Role of mitogen-activated protein kinase kinase 4 in cancer.** *Oncogene* 2007, **26**:3172-3184.

108. Kim WY, Prudkin L, Feng L, Kim ES, Hennessy B, Lee JS, Lee JJ, Glisson B, Lippman SM, Wistuba, II, Hong WK, Lee HY: **Epidermal growth factor receptor and K-Ras mutations and resistance of lung cancer to insulin-like growth factor 1 receptor tyrosine kinase inhibitors.** *Cancer* 2012, **118**:3993-4003.
109. Olivier M, Hollstein M, Hainaut P: **TP53 mutations in human cancers: origins, consequences, and clinical use.** *Cold Spring Harb Perspect Biol* 2010, **2**:a001008.
110. Fodde R, Smits R, Clevers H: **APC, signal transduction and genetic instability in colorectal cancer.** *Nat Rev Cancer* 2001, **1**:55-67.
111. Lin J, Yao DM, Qian J, Chen Q, Qian W, Li Y, Yang J, Wang CZ, Chai HY, Qian Z, Xiao GF, Xu WR: **Recurrent DNMT3A R882 mutations in Chinese patients with acute myeloid leukemia and myelodysplastic syndrome.** *PLoS One* 2011, **6**:e26906.
112. Zhao C, Dahlman-Wright K: **Liver X receptor in cholesterol metabolism.** *J Endocrinol* 2010, **204**:233-240.
113. Gulbahce HE, Sweeney C, Surowiecka M, Knapp D, Varghese L, Blair CK: **Significance of GATA-3 expression in outcomes of patients with breast cancer who received systemic chemotherapy and/or hormonal therapy and clinicopathologic features of GATA-3-positive tumors.** *Hum Pathol* 2013, **44**:2427-2431.
114. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler

- AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500**:415-421.
115. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L: **Mutational landscape and significance across 12 major cancer types.** *Nature* 2013, **502**:333-339.
116. Meyer DS, Koren S, Leroy C, Brinkhaus H, Muller U, Klebba I, Muller M, Cardiff RD, Bentires-Alj M: **Expression of PIK3CA mutant E545K in the mammary gland induces heterogeneous tumors but is less potent than mutant H1047R.** *Oncogenesis* 2013, **2**:e74.
117. Janku F, Wheler JJ, Naing A, Falchook GS, Hong DS, Stepanek VM, Fu S, Piha-Paul SA, Lee JJ, Luthra R, Tsimberidou AM, Kurzrock R: **PIK3CA**

- mutation H1047R is associated with response to PI3K/AKT/mTOR signaling pathway inhibitors in early-phase clinical trials.** *Cancer Res* 2013, **73**:276-284.
118. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP: **DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer.** *Genome Biol* 2012, **13**:R124.
119. Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM: **PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.** *Bioinformatics* 2012, **28**:i640-i646.
120. Cancer Target D, Development N, Schreiber SL, Shamji AF, Clemons PA, Hon C, Koehler AN, Munoz B, Palmer M, Stern AM, Wagner BK, Powers S, Lowe SW, Guo X, Krasnitz A, Sawey ET, Sordella R, Stein L, Trotman LC, Califano A, Dalla-Favera R, Ferrando A, Iavarone A, Pasqualucci L, Silva J, Stockwell BR, Hahn WC, Chin L, DePinho RA, Boehm JS, Gopal S, Huang A, Root DE, Weir BA, Gerhard DS, Zenklusen JC, Roth MG, White MA, Minna JD, MacMillan JB, Posner BA: **Towards patient-based cancer therapeutics.** *Nat Biotechnol* 2010, **28**:904-906.

VITA

Tenghui Chen, the son of Zonghai Chen and Shuzhen Zhang, was born on August 31st, 1985 in Quanzhou, Fujian province, China. He graduated from Quangan No.1 high school in Quanzhou in June 2004. After that, he joined Xiamen University with major in Biotechnology and obtained his bachelor degree in June 2008. He continuously stayed in Xiamen University for the master degree study with a major in Cell Biology and gained his master degree in June 2011. In August 2011, he moved to Houston, Texas to pursue his Ph.D degree in Biostatistics, Bioinformatics and Systems Biology program at the University of Texas Health Science Center at Houston and MD Anderson Cancer Center. He conducts thesis research under the supervision of Dr. Ken Chen in the department of Bioinformatics and Computational Biology and expects to finish his Ph.D study in May 2016.

Permanent address:

Xucuo Village No.26, Houlong Town, Quangan District,
Quanzhou, Fujian, China, 362800