Winter 12-15-2016

Discovery and Functional Implication of Genetic Alterations Associated with Clonal Hematopoietic Expansion

Mingchao Xie
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Li Ding, Chair
Barak A. Cohen
Todd E. Druley
John R. Edwards
Daniel C. Link
Gary D. Stormo

Discovery and Functional Implication of Genetic Alterations Associated with Clonal Hematopoietic Expansion

by

Mingchao Xie

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2016
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

It took me much longer time to write this chapter than what I thought, because there's so many people I want to thank, and so many great memories I want to share.

First of all, I would like to express my sincere gratitude to my advisor, Dr. Li Ding, for the generous support and help during my Ph.D. study. Her perseverance and passion for science inspired me over the years, and led me to realize how to become a great scientist. I always appreciated her guidance and encouragement, and I could not imagine how I could have finished my thesis project so smoothly without her help.

In addition to my advisor, I would like to thank my thesis committee: Dr. Gary Stormo, for the invaluable feedback on my thesis project and wonderful guidance before and after I joined graduate school; Dr. Barak Cohen, for the great encouragement and inspiring discussions on all of the questions I had for my research project and Ph.D. training as a Computational Biology student; Dr. Todd Druley and Dr. John Edwards, for the thought-provoking questions that shaped my thesis work, starting from the qualifying exam and through my thesis proposal and thesis updates, encouraging me to expand my horizons; and last but not least, my thesis committee chair, Dr. Daniel Link, whose incredible knowledge base in oncology and multi-faceted approaches in experimental biology allowed me to appreciate the importance of experiments in biological study.

In addition, to the past and present Ding Lab members and MGI members, thank you for the stimulating discussions and wonderful collaborations in the past four years. I also want to thank Dr. Feng Chen and his lab members for all of the help and support on my experimental work. There is not enough space for me to individually thank everyone, but I have learned many

incredible things from everyone in this amazing group, who I had the honor to share my time with during my graduate school.

Also, I want to thank, Prof. Ting Wang. It is he to lead me into the exciting field of biological research, computational genomics study, and encourage me to pursue a PhD degree in Washington University. Besides, I would also thank every staff member in DBBS whose hard work allows us to focus on graduate studies and research, especially Jeanne Silvestrini.

I would also thank my dear friends outside of the lab, who made my life in graduate school colorful, for the fun we have had in the past few years.

I would express a deep sense of gratitude to my parents and my brother for their dedication and the many years of support even though I was so far away from them.

Above all, I am so grateful for my wife, Meng, for all the love and care she has given to me. Thank you for always being there for me, believing in me and encouraging me to pursue my dreams.

Mingchao Xie

Washington University in St. Louis

December 2016

Dedicated to my family.

ABSTRACT OF THE DISSERTATION

Discovery and Functional Implication of Genetic Alterations Associated with Clonal Hematopoietic Expansion

by

Mingchao Xie

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2016

Associate Professor Li Ding, Chair

Cancers, including hematologic malignancies, arise as a result of the stepwise accumulation of mutations. Some early mutations that potentially initiate clonal expansion might exist in patients many years before they develop obvious disease symptoms. Therefore, identifying and characterizing these early mutations are critical to understanding the genetic basis of tumorigenesis. Here, we analyzed blood-derived DNA sequencing data from 2,728 individuals without apparent hematologic malignancies and identified 77 blood-specific mutations in 31 cancer-associated genes. Importantly, 83% of all mutations occurred in 19 genes that have been previously linked to hematological malignancies, such as *DNMT3A*, *TET2*, *JAK2*, and *ASXL1*. By investigating these mutations in different hematologic diseases, we identified several recurrently mutated genes that may be disease initiators. To obtain more comprehensive profiling of genes and variants associated with clonal hematopoietic expansion, we processed an additional 3,221 normal blood samples from The Cancer Genome Atlas (TCGA) and developed a statistical approach to systematically identify blood-specific mutations in all human genes. 26 genes were significantly mutated in human blood samples, including *PPM1D*. Functional

validation showed that *PPM1D* mutations suppressed the phosphorylation of TP53 at Ser15, suggesting that the blood-specific mutants in *PPM1D* retain its phosphatase activity in regulating TP53. We also characterized rare copy number variations (CNVs) in blood samples and discovered about half of the individuals examined carried rare somatic CNVs in their blood. Some of these CNVs were associated with genes involved in hematological malignancies, such as *JAK2*, *ASXL1*, and *FLT3*. In summary, we systematically identified early genomic alterations in normal blood cells by utilizing the large-scale sequencing data and further determined the functional impact of the mutations in the recurrently mutated gene. Our comprehensive analysis of blood-specific genomic alterations will shed light on understanding the complex mechanisms of hematologic malignancies and also facilitate the development of more efficient strategies for early detection, prevention, and treatment of hematologic cancer.

Chapter 1: Introduction to Cancer Genomics

Cancer is a leading cause of death worldwide. According to the International Agency for Research on Cancer, there were 14.1 million new cancer cases and 8.2 million cancer-related deaths in 2012 [1]. Furthermore, as a result of the growing and aging human population, the number of annual cancer cases is expected to double in the next two decades [1]. The majority of cancer cases are sporadic, caused by exposure to various carcinogens, such as radiation, chemical mutagens, or viral/bacterial infections. However, approximately 5–10% of cancers are inherited [1].

Cancer is "a disease of the genome." Scientists first realized the role of the genome in cancer development in the early twentieth century when Theodor Boveri proposed that cancers are caused by chromosomal abnormalities [2]. This groundbreaking hypothesis was proven in the 1970s by the discovery of the Philadelphia Chromosome and the identification of the *BCR–ABL* fusion gene in chronic myelogenous leukemia [3, 4].

Because cancer is attributed to genomic alterations in cells, it is essential to identify the genes that tend to accumulate mutations during cancer development. Based on their effects on tumorigenesis, cancer-associated genes are classified into two categories: oncogenes and tumor suppressor genes. In 1979, *src* was identified as the first oncogene in a chicken retrovirus, which induces tumors in connective tissues, such as bone and muscle, of infected animals [5]. Five years later, the first tumor suppressor gene, *RB1*, was discovered in the retinoblastoma studies [6]. Unlike oncogenes, where a genetic alteration on a single allele can lead to cellular transformation, the inactivation of tumor suppressor gene requires genetic silencing of both alleles, which is so-called "two-hit hypothesis," proposed by Knudson [7].

After decades of study, many important discoveries in cancer research have been achieved by traditional molecular analysis. However, due to the complex pathophysiology, it is very difficult to comprehensively study cancer progression by relying on traditional approaches alone. The Human Genome Project opened up new avenues for cancer research, leading biologists to realize that the complete sequencing of the cancer genome would be a superb way to systematically study genes and mutations involved in cancer development. Based on this idea, the first cancer genome from a patient with acute myeloid leukemia was decoded in 2008 [8]. Since then, the number of sequenced cancer genomes has grown exponentially. By 2015, more than 8,000 cases across various cancer types were sequenced and analyzed by TCGA, the majority of which are high-coverage whole exome sequencing data [9, 10]. The ubiquity of these data has spurred active research in multiple areas, including identification of novel pathogenic somatic mutations, detection of germline cancer susceptibility variants, and cancer evolution. These studies have provided significant insights into the characterization of the mutational landscape within the cancer genome and the understanding of the mechanisms underlying tumorigenesis. These advancements are expected to facilitate the development of novel strategies for cancer diagnosis, prevention, and treatment. Although cancer genomic studies have made remarkable progress, some key challenges still need to be addressed, as discussed below.

1.1 Mutation landscapes in cancer genomes

1.1.1 Mutational heterogeneity in cancer

Although cancers are caused by the genomic mutations, the patterns of mutations are divergent, both mutation frequency and spectra varying substantially across cancer types, among individuals with the same cancer, and even within a single cancer genome, which indicates cancer heterogeneity [9-12].

First, mutation frequency varies in different cancer types, averaging from 0.1 events per Mb in pediatric cancers or adult leukemia to more than 10 events per Mb in melanoma and lung cancers [9, 10]. This difference can be partly explained by the divergent mutagenic mechanisms underlying each cancer type. For example, pediatric cancers often occur in tissues that have fewer cell divisions. As a result, there is less time for mutations to accumulate. While melanomas and lung cancers are primarily induced by chronic exposure to well-known carcinogens, such as ultraviolet (UV) radiation and tobacco smoke, leading to extremely high mutation rates in skin and lung tissue.

Mutation frequency also varies substantially among individuals with the same cancer type. Remarkable example can be seen in lung cancer and leukemia. In lung cancer, the median frequency of non-synonymous mutations varies by more than 1,000-fold across the entire lung cancer cohort, ranging from 0.1 to 100 mutations per Mb [10]. Similarly, although leukemia exhibits fewer mutations, the mutation frequency in the entire leukemia cohort also spans three orders of magnitude (from 0.01 to 10/Mb per individual), as lung cancer does [10].

Moreover, at the single cancer genome level, the mutation frequency also appears to be divergent across the genome, correlated with transcription level and DNA replication timing [11, 12]. Specifically, late-replicating and low-expression regions, such as intergenic and non-coding regions, tend to have much higher mutation rates than the early replicating and actively transcribed regions due to depletion of the pool of free nucleotides and transcription-coupled repair, respectively [10-12].

In addition to mutation frequency, mutation spectra also vary in different cancer types [9, 10]. Skin cancers show a distinct pattern of C->T mutations, which is caused by misrepair of UV-

induced covalent bonds between adjacent pyrimidines [9, 10]. Lung cancer is the most common cancer caused by smoking. Correspondingly, C->A transversion, a signature of tobacco smoking exposure, is dominant in lung cancer patients' genomes [9, 10]. In addition, bladder, cervical, and head and neck cancers usually show frequent C->T or G->A mutations, possibly due to off-target modification of DNA by the APOBEC family of cytidine deaminases [10, 13]. All of these observations suggest that the study of mutation spectra could facilitate understanding of specific tumor etiology.

Mutational heterogeneity is a hallmark of cancer. To some extent, mutational heterogeneity reflects the diversity and complexity of cancer. Divergent mutation patterns conceal the key mechanisms underlying tumorigenesis and make it difficult to determine the potential targets for directed therapies. Large-scale cancer genome sequencing studies have provided an enormous amount of information on the genomic alterations in a variety of human cancers, which has sharply improved our understanding of the genetic basis for cancer. However, the unexpectedly large number of mutations detected in human cancers also dramatically increases the difficulty in identifying the critical genes or mutations that cause cancer initiation, progression, and metastasis.

1.1.2 Passenger and driver mutations in the cancer development

Cancer is caused by genomic alterations, but not all alterations present in the cancer genome promote tumorigenesis. According to their roles in tumor growth, alterations can be categorized as "drivers" or "passengers." The majority of alterations in cancer genome are incidental "passengers"—they have no effect on tumorigenesis. On the other hand, the "driver" mutations are under positive selection during cancer progression and constantly confer a growth advantage to the cancer cells [9, 10, 14].

In comparison to passenger mutations, the number of driver mutations is extremely small. It is estimated that 2 to 7 driver mutations are sufficient to convert a normal human cell to a malignant cell [9, 15-16]. Given the importance of driver mutations in cancer development, distinguishing them from passenger mutations has been a central goal of cancer research for years. At present, more than 600 genes have been discovered in several large-scale cancer genomic studies, with substantial evidence showing that their mutations are associated with cancer progression, and many of these genes are transcription factors, protein kinases, cell cycle regulators, or DNA/histone modifiers [9, 10, 17-19]. Some of these genes are commonly mutated in several cancer types. For example, *TP53*, an essential gene in the cell cycle regulation pathway, was mutated in about half of all cases in certain studies, especially in serous ovarian (95% of cases) and serous endometrial carcinomas (89% of cases) [9, 17]. Genes in the classical PI3K signaling pathway, including *PIK3CA*, *PTEN*, *PIK3R1*, *TLR4*, *PIK3CG*, and *AKT1*, are also frequently mutated in most cancer types, such as breast cancer, endometrial cancer, head and neck cancer, and glioblastoma multiforme [9]. Also, some drivers show tumor-type specificity. For example, *APC* is a well-known tumor suppressor gene, whose mutations predominate in the colon and rectal carcinoma [9], as well as *VHL* in kidney cancer [9, 17].

The analysis of driver mutations is critical for understanding the molecular mechanisms of tumorigenesis and the identifying of therapeutic targets. However, the role of mutations as passengers or drivers is not immutable. Sometimes, due to cancer treatment or changes of microenvironment, a passenger mutation can become a driver mutation and confer resistance to the selective pressure, expanding at relapse or growing in new sites [20]. Therefore, the identification of driver mutations is still of primary importance for cancer genomic study.

1.2 The heritability of cancer

1.2.1 Hereditary cancer syndromes

The majority of cancer cases are sporadic, however, some common cancers, including breast, ovarian, colorectal, and prostate cancers, exhibit clear family aggregation patterns, showing cancer more frequently occurred in family members at an earlier age than in the general population [21-24]. Based on monozygotic twins studies, the family aggregation of cancer is believed to be primarily caused by inherited genetic factors instead of shared environmental factors [21, 22].

It is estimated that at least 5%-10% of all cancers are inherited. Based on the two-hit hypothesis [7], a person with an inherited mutation obtain one copy of an abnormal gene passed from their parents, where another copy of the functional gene could be inactivated with a high chance during cell division by random processes, such as DNA mismatch, chromosome rearrangement, deletion, or insertion. Therefore, they are much more likely to develop cancer at an early age. However, for individuals without inherited mutations, two mutations, one in each allele of the gene, need to be accumulated in the same cell, and this process could take much longer.

Hereditary cancers often show different features from sporadic cancers. First, several family members tend to have the same or similar cancers at an earlier age. Second, cancer is more likely to develop in more than one place in the body. For example, patients with germline pathogenic variants in *BRCA1* or *BRCA2* have high risk of developing breast and ovarian cancers, as well as the other cancer types, such as prostate, pancreatic, and melanoma at an earlier age [25-28]. Meanwhile, they also have much higher risk for cancer recurrence than the general population. This phenomenon is “hereditary breast and ovarian cancer syndrome (HBOC)”. For HBOC individuals, the lifetime risks for developing breast and ovarian cancers are 55%-85% and 25%-

45%, respectively, while for the general population, the corresponding risks are much lower, only 12.3% and 1.3%, respectively [29, 30]. A similar phenomenon is also found in Lynch syndrome, often called hereditary nonpolyposis colorectal cancer (HNPCC) [31].

1.2.2 Cancer predisposition genes and variants

In the past 30 years, many inherited loci and genes have been identified as high-risk candidates associated with cancer predisposition by family-based studies and genome-wide association studies (GWAS) [32-42]. However, these candidates can only explain a small proportion of the cancer heritability, which leaves a door open for many more candidates to be discovered.

Since GWAS primarily focuses on identifying common disease risk-associated variants, typically with a minor allele frequency (MAF) > 1%, it is reasonable that rare variants could have the potential to explain additional disease risk. On the other hand, based on the theory of evolution, disease-associated variants are likely to be rare as a result of negative selection. Therefore, the hypothesis that many more unknown rare variants may contribute to the missing heritability in cancer has been commonly accepted.

Based on current studies, there are two types of rare variants associated with cancer predisposition. The first type of rare variant has high penetrance to disease, and their frequencies are very low in the population, most often less than 0.1%. Such would include, for example, pathogenic variants in *BRCA1* [34], *BRCA2* [35], *APC* [36], *TP53* [37], and some mismatch repair genes, such as *MSH2* [38], *MLH1* [39]. Individuals with relevant variants in these genes have a much higher risk of developing cancer. Another type of rare variant, with frequencies ranging from 0.1% to 0.5%, have moderate penetrance to disease, but their combined effects can be sufficient to cause cancer progression. For example, genes in a common pathway (e.g., DNA damage checkpoint or repair pathways), such as *CHEK2*, *BRIP1*, and *PALB2*, have been reported

to carry moderate-effect rare variants that are associated with breast cancer predisposition [40-42].

A rapidly growing number of publicly available deep-coverage exome sequencing data is a valuable resource for the study of rare germline cancer susceptibility variants [43-45]. Kanchi *et al.* analyzed 429 whole exome sequencing data from TCGA serous ovarian cancer patients and detected 3,635 high confident, rare (<1% MAF in the population and cohort) truncations and 22,953 missense variants with predicted functional impact [45]. By comparing the frequency of germline variants with healthy control data, several novel candidate germline susceptibility variants in known ovarian genes (such as *BRCA1*, *BRCA2*, *ATM* [46], and *PALB2*), as well as several genes not previously associated with ovarian cancer (such as, *ASXL1*, *RB1*, *NF1*, *CDKN2A* and *EXO1*), were identified [45]. This investigation established a foundation for future susceptibility variants studies in other cancer types.

Large-scale sequencing data hold great promise for rare cancer susceptibility variant discovery. However, to determine the effect of candidate variants and genes prioritized by sophisticated computational filtering strategies, the replication in additional datasets and extensive experimental functional validation are essential.

1.3 The evolution of cancer

1.3.1 Model of cancer evolution

Cancer progression is a complex, dynamic, and cumulative process that involves stages of cancer initiation, invasion, and metastasis. The first landmark perspective on cancer progression took place in 1976 when Peter Nowell proposed that cancer is a clonal disease restricted by Darwinian evolution [47]. Based on this theory, the typical view of cancer progression can be characterized

as successive waves of clonal expansion, in which clones with the best growth advantage (evolutionary fitness) arise from a small subclone to become a dominant clone, giving rise to disease relapse and metastasis.

Intra-tumor heterogeneity is central to cancer evolution, as it provides a pool of genetic variants that can be used for selection, increasing the probability of a subclone with a fitness advantage existing in the cell population. On the other hand, environmental factors, such as tumor microenvironment, carcinogenic exposures, and cancer treatments, provide selective pressures to guide the evolutionary process within a tumor and shape the clonal architecture.

Several studies have already demonstrated clonal evolution in several common cancer types, such as renal carcinomas, acute myeloid leukemia, breast cancers, and glioblastoma [48-52]. Ding *et al.* studied the primary tumor and relapse samples from eight AML patients, and found two main clonal evolution patterns during AML relapse: either a founding clone in the primary tumor gained new driver mutations, or a subclone survived in cancer treatment gained additional mutations [48]. This study confirmed previous research on cancer evolution and enhanced the understanding of genetic changes acquired during AML progression and relapse.

In another study, Nik-Zainal *et al.* analyzed tumor sequence data from 21 breast cancer patients, investigated the content of subclonal variations within the cancer samples, and examined mutation changes over time [51]. By reconstructing the dominant subclonal lineage during breast cancer development, they revealed the appearance of the most-recent common ancestor, which contains the full complement of somatic mutations found in all tumor cells. This study uncovered a new cancer evolution pattern, which is different from that observed in AML, and shed light on

the investigation of the mechanism of breast cancer development from breast organogenesis to carcinogenesis in adults.

1.3.2 Clonal expansions in normal cells

Current cancer genomic studies have already made substantial progress in our understanding of the temporal order of mutations or clonal architectures during cancer progression. However, many knowledge gaps still remain. For example, how normal cells obtain growth advantages progressively, how they initiate the clonal expansion, and how and when they break through the evolutionary bottleneck and eventually result in clonal dominance in the newly formed tumor. All of these questions remain largely unanswered.

In adult epithelial tumors, about 5-7 driver mutations are required for cancer progression [9,14]. Although in hematopoietic malignancies, this number might be lower, at least two lesions are still needed, each belonging to a distinct allele [9, 53]. However, it is not so easy for a normal cell to acquire two spontaneous driver mutations on different alleles. One possible way to achieve that quickly is clonal expansion. When a normal cell acquires one driver mutation due to a random process, it obtains a growth advantage that triggers clonal expansion. Although this clonal expansion is not sufficient to cause cancer directly, it could expand the pool of mutated cell for further selection, thus increasing the chance for one mutated cell to obtain an additional driver mutation, and thereby contributing to cancer development when neither of them is sufficient to cause it alone. Therefore, measuring the genomic changes in normal tissues over time could help us quantify the extent of mutation and understand the dynamics of clonal expansion in the early stages of cancer development.

There are two significant challenges in this study. First, based on the model of cancer evolution, a series of clonal expansions during cancer development is a long-term process, which usually

takes decades. Therefore, collecting sufficient sequencing data from normal tissues to monitor clonal expansion is a big challenge. Second, in the early stages of clonal expansion, the driver mutation is only present in a very tiny fraction of cells, making them very difficult to detect precisely. However, recently, a wide variety of large-scale sequencing projects, including TCGA and ICGC, have generated a number of high coverage exome sequencing data in normal tissues, especially in blood samples, which makes the study of clonal expansion in noncancerous tissue feasible. Investigating clonal expansion in normal tissue will provide great insight into the dynamic change of genomic alterations from organogenesis to tumorigenesis, and allow us to better understand cancer initiation and progression. In addition, this study can also be translated into effective clinical strategies, helping us to promote targeted screening and preventive measures, and to monitor people's health status and disease risk.

In summary

Cancer is a complex disease, and decades of detailed genetic research indicate that it is more complicated than we initially thought. Cancer genome studies so far have laid a foundation for us to better understand the genetic basis for cancer, which demonstrates the power of sequencing technology and genomic information for analyzing complex diseases. However, we need to realize that the cancer study is a long path, and there is still a lot of unknown in cancer, just like what we discussed in this chapter.

1.4 Reference

1. Ferlay, J., I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. GLOBOCAN 2012 v1. 0. *Cancer incidence and mortality worldwide: IARC CancerBase* 11, 2014 (2013).

2. Boveri, Theodor. Zur frage der entstehung maligner tumoren. *Gustav Fischer*, (1914).

3. Nowell, P., and D. Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Landmarks in Medical Genetics: Classic Papers with Commentaries* 132, 103 (2004).

4. Rowley, Janet D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. 290-293 (1973).

5. Oppermann, Hermann, Arthur D. Levinson, Harold E. Varmus, Leon Levintow, and J. Michael Bishop. Uninfected vertebrate cells contain a protein that is closely related to the product of the avian sarcoma virus transforming gene (src). *Proceedings of the National Academy of Sciences* 76, 1804-1808 (1979).

6. Murphree, A. Linn, and William F. Benedict. Retinoblastoma: clues to human oncogenesis. *Science* 223, 1028-1033 (1984).

7. Knudson, Alfred G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* 68, 820-823 (1971).

8. Ley, Timothy J., Elaine R. Mardis, Li Ding, Bob Fulton, Michael D. McLellan, Ken Chen, David Dooling *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66-72 (2008).

9. Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).

10. Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-218 (2013).

11. Fousteri, Maria, and Leon HF Mullenders. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell research* 18, 73-84 (2008).

12. Stamatoyannopoulos, John A., Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev. Human mutation rate associated with DNA replication timing. *Nature genetics* 41, 393-395 (2009).

13. Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell *et al*. Signatures of mutational processes in human cancer. *Nature* 500, 415-421 (2013).

14. Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature* 458, 719-724 (2009).

15. Schinzel, Anna C., and William C. Hahn. Oncogenic transformation and experimental models of human cancer. *Frontiers in bioscience: a journal and virtual library* 13, 71-84 (2007)

16. Beerenwinkel, Niko, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W. Kinzler, Victor E. Velculescu, Bert Vogelstein, and Martin A. Nowak. Genetic progression and the waiting time to cancer. *PLOS Computational Biology* 3, e225 (2007).

17. Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science* 339, 1546-1558 (2013)

18. Pritchard, Colin C., Stephen J. Salipante, Karen Koehler, Christina Smith, Sheena Scroggins, Brent Wood, David Wu *et al*. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of Molecular Diagnostics* 16, 56-67 (2014).

19. Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin *et al*. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology* 31, 1023-1031 (2013).

20. Mullighan, Charles G., Letha A. Phillips, Xiaoping Su, Jing Ma, Christopher B. Miller, Sheila A. Shurtleff, and James R. Downing. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* 322, 1377-1380 (2008)

21. Lichtenstein, Paul, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England journal of medicine* 343, 78-85 (2000).

22. Goldgar, David E., Douglas F. Easton, Lisa A. Cannon-Albright, and Mark H. Skolnick. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *Journal of the National Cancer Institute* 86, 1600-1608 (1994).

23. Edwards, Stephen M., Zsofia Kote-Jarai, Julia Meitz, Rifat Hamoudi, Questa Hope, Peter Osin, Rachel Jackson *et al*. Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *The American Journal of Human Genetics* 72, 1-12 (2003).

24. Pharoah, Paul DP, Antonis C. Antoniou, Douglas F. Easton, and Bruce AJ Ponder. Polygenes, risk prediction, and targeted prevention of breast cancer. *New England Journal of Medicine* 358, 2796-2803 (2008).

25. Parmigiani, Giovanni, Donald A. Berry, and Omar Aguilar. "Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2." *The American Journal of Human Genetics* 62, 145-158 (1998).

26. Antoniou, A. C., P. P. D. Pharoah, P. Smith, and D. F. Easton. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer* 91, 1580-1590 (2004).

27. Thompson, Deborah, Douglas F. Easton, and Breast Cancer Linkage Consortium. Cancer incidence in BRCA1 mutation carriers. *Journal of the National Cancer Institute* 94, 1358-1365 (2002).

28. Lubinski, Jan, Catherine M. Phelan, Parviz Ghadirian, Henry T. Lynch, Judy Garber, Barbara Weber, Nadine Tung *et al*. Cancer variation associated with the position of the mutation in the BRCA2 gene. *Familial cancer* 3, 1-10 (2004).

29. Antoniou, Anthony, P. D. P. Pharoah, Steven Narod, Harvey A. Risch, Jorunn E. Eyfjord, J. L. Hopper, Niklas Loman *et al*. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics* 72, 1117-1130 (2003).

30. Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers. *Journal of the National Cancer Institute* 91, 1310-1316 (1999).

31. Lynch, Henry T., and Albert de la Chapelle. Hereditary colorectal cancer. *New England Journal of Medicine* 348, 919-932 (2003).

32. Rahman, Nazneen. Realizing the promise of cancer predisposition genes. *Nature* 505, 302-308 (2014).

33. Foulkes, William D. Inherited susceptibility to common cancers. *New England Journal of Medicine* 359, 2143-2153 (2008).

34. Skolnick, Mark H. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266, 66-71 (1994).

35. Wooster, Richard, Graham Bignell, Jonathan Lancaster, Sally Swift, Sheila Seal, Jonathan Mangion, Nadine Collins *et al*. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378, 789-792 (1995).

36. Groden, Joanna, Andrew Thliveris, Wade Samowitz, Mary Carlson, Lawrence Gelbert, Hans Albertsen, Geoff Joslyn *et al*. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* 66, 589-600 (1991).

37. Malkin, David, Frederick P. Li, Louise C. Strong, Joseph F. Fraumeni Jr, Camille E. Nelson, David H. Kim, Jayne Kassel *et al*. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250, 1233-1238 (1990).

38. Fishel, Richard, Mary Kay Lescoe, M. R. S. Rao, Neal G. Copeland, Nancy A. Jenkins, Judy Garber, Michael Kane, and Richard Kolodner. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 75, 1027-1038 (1993).

39. Papadopoulos, Nickolas, Nicholas C. Nicolaides, Ying-Fei Wei, Steven M. Ruben, Kenneth C. Carter, Craig A. Rosen, William A. Haseltine, Robert D. Fleischmann, Claire M. Fraser, and Mark D. Adams. Mutation of a mutL homolog in hereditary colon cancer. *Science* 263, 1625-1629 (1994).

40. CHEK2 Breast Cancer Case-Control Consortium. CHEK2* 1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *The American Journal of Human Genetics* 74, 1175-1182 (2004).

41. Seal, Sheila, Deborah Thompson, Anthony Renwick, Anna Elliott, Patrick Kelly, Rita Barfoot, Tasnim Chagtai *et al*. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nature genetics* 38, 1239-1241 (2006).

42. Rahman, Nazneen, Sheila Seal, Deborah Thompson, Patrick Kelly, Anthony Renwick, Anna Elliott, Sarah Reid *et al*. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics* 39, 165-167 (2007).

43. Comino-Méndez, Iñaki, Francisco J. Gracia-Aznárez, Francesca Schiavi, Iñigo Landa, Luis J. Leandro-García, Rocío Letón, Emiliano Honrado *et al*. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nature genetics* 43, 663-667 (2011).

44. Smith, Miriam J., James O'Sullivan, Sanjeev S. Bhaskar, Kristen D. Hadfield, Gemma Poke, John Caird, Saba Sharif *et al*. Loss-of-function mutations in SMARCE1 cause an inherited disorder of multiple spinal meningiomas. *Nature genetics* 45, 295-298 (2013).

45. Kanchi, Krishna L., Kimberly J. Johnson, Charles Lu, Michael D. McLellan, Mark DM Leiserson, Michael C. Wendl, Qunyuan Zhang *et al*. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications* 5 (2014).

46. Athma, Prasanna, Rebecca Rappaport, and Michael Swift. Molecular genotyping shows that ataxia-telangiectasia heterozygotes are predisposed to breast cancer. *Cancer genetics and cytogenetics* 92, 130-134 (1996).

47. Nowell, Peter C. The clonal evolution of tumor cell populations. *Science* 194, 23-28 (1976).

48. Ding, Li, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey *et al*. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481, 506-510 (2012).

49. Walter, Matthew J., Dong Shen, Li Ding, Jin Shao, Daniel C. Koboldt, Ken Chen, David E. Larson *et al*. Clonal architecture of secondary acute myeloid leukemia. *New England Journal of Medicine* 366, 1090-1098 (2012).

50. Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine* 366, 883-892 (2012).

51. Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine *et al*. The life history of 21 breast cancers. *Cell* 149, 994-1007 (2012).

52. Kim, Jinkuk, In-Hee Lee, Hee Jin Cho, Chul-Kee Park, Yang-Soon Jung, Yanghee Kim, So Hee Nam *et al*. Spatiotemporal evolution of the primary glioblastoma genome. *Cancer cell* 28, 318-328 (2015).

53. Gilliland, D. Gary, and James D. Griffin. The roles of FLT3 in hematopoiesis and leukemia. *Blood* 100, 1532-1542 (2002).

Chapter 2: Patterns and Functional Implications of Rare Germline Variants across 12 Cancer Types[†]

2.1 Abstract

Large-scale cancer sequencing data enable discovery of rare germline cancer susceptibility variants. Here we systematically analyze 4,034 TCGA cancer cases representing 12 cancer types. Rare germline truncations in 114 cancer-susceptibility-associated genes vary widely, from 4% (AML) to 19% (ovarian cancer), with a notably high frequency of 11% in stomach cancer. Burden testing identifies 13 cancer genes with significant enrichment of rare truncations, some associated with specific cancers (e.g. *RAD51C*, *PALB2*, and *MSH6* in AML, stomach, and endometrial cancers, respectively). Significant, tumor-specific loss of heterozygosity occurs in 9 genes (*ATM*, *BAP1*, *BRCA1/2*, *BRIP1*, *FANCM*, *PALB2*, and *RAD51C/D*). Moreover, our homology directed repair assay of 68 *BRCA1* rare missense variants supports the utility of allelic enrichment analysis for characterizing variants of unknown significance. The scale of this analysis and the somatic-germline integration enable the detection of rare variants that may affect individual susceptibility to tumor development, a critical step toward precision medicine.

2.2 Introduction

At least 3% of all cancer cases are thought to have a strong hereditary component, with large variation being found across cancer types [1]. For example, it was recently estimated that up to 20-25% of ovarian cancers are due to a germline loss-of-function variant in one of several genes that confer moderate to high risk [2, 3], while other cancer types (e.g. lung) have strong

[†] Charles Lu[*], Mingchao Xie[*], Michael C. Wendl[*], *et al*. "Patterns and functional implications of rare germline variants across 12 cancer types." *Nature communications* 6 (2015) (* contributed equally to this work)

environmental components with little evidence of genetic predisposition [4]. The absence of heritability in some cancers may be due to low or medium penetrance alleles [5]. Genome wide association studies (GWAS) have been instrumental in identifying hundreds of common low-effect risk alleles across multiple cancer types [6]. The availability of large scale normal and tumor sequencing data from cancer cases now allows for discovery of rare variants influencing cancer susceptibility through analysis of both germline and somatic sequencing data.

Tumorigenesis is a complex process that often involves close interactions between germline and somatic variants. Their cooperation is best exemplified by the "two-hit hypothesis" [7], in which a tumor suppressor gene is inactivated by the combination of an initial germline mutation of one allele, followed by the somatic inactivation of the other. Loss of heterozygosity (LOH), whereby the wild-type allele for a two-hit tumor suppressor is eliminated, has been implicated in many cancers [8, 9]. Advancing our understanding of cooperative germline-somatic dynamics and their implications for tumorigenesis requires large cohort studies using sequencing data from both germline and somatic tissues, as well as new tools to reliably detect allelic loss.

We have previously reported that whole exome sequencing data can be successfully employed to identify both known high penetrance cancer genes in ovarian cancer, as well as new candidate predisposition alleles for downstream functional characterization [3]. Here, we extend this work to 12 cancer types with the goal of describing the landscape of germline variants (truncation and missense) and analyzing the effect of germline variants on somatic mutations using >4,000 cancer cases. Our analysis shows a diverse set of genes potentially contributing to predisposition with variable frequencies and levels. Stomach cancer has a relatively high rate of rare germline truncations, in large part due to frequent *PALB2* and *ATM* mutations. Genes and local hotspots of significant allelic enrichment within functional domains were discovered through integrating

germline and somatic data. Germline and somatic integration sheds insights on genes influencing somatic mutation frequencies and genes/pathways involved in the entire life history of individual tumors. Experimental validation of 68 *BRCA1* variants, with 62 having previously unknown functional significance or not reported by the NHGRI Breast Cancer Information Core (BIC) database, identified 9 with complete or partial loss of homology directed repair (HDR) function, further supporting LOH analysis results. Such discovery of new cancer susceptibility genes and functional characterization of variant alleles will be an important step toward generating an actionable catalog for personalized treatment of cancer.

2.3 Results

2.3.1 Cancer types and sample characteristics

We searched for candidate germline cancer predisposition variants in the exome sequence data from 4,034 cancer patients across 12 diverse cancer types: breast adenocarcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (AML), low grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC). The numbers of cases from each tumor type ranged from 178 (PRAD) to 770 (BRCA) and are listed in Table 2.1. Of the 3,548 TCGA cases with available ethnicity information, 88.1% were Caucasian (*n*=3,125), 6.3% were African American (*n*=225), 5.2% were Asian (*n*=183), and 0.4% (*n*=15) were American Indian/Alaska Native. Patients (*n*=3,827) were diagnosed between 10-90 years (mean 59.9 ± 13.2 years) with LUSC and LGG having the highest and lowest mean ages, respectively (Table 2.1). The sex distribution is generally consistent with U.S. general population cancer statistics for these malignancy types.

Age of onset distribution is bimodal for LGG, LUAD, and STAD, with some evidence of a bimodal distribution for OV, KIRC, HNSC, and GBM. Distinct age of onset populations may indicate discrete mutational or disease processes (Figure 2.1a).

Sequencing data for an additional 1,627 TCGA cases were collected for 10 out of the 12 cancer types (AML and STAD not included) for validating findings from the discovery cohort. In the validation cohort, 1,388 TCGA cases had available demographic information, of which 1,173 cases had ethnicity information, where 83.8% were Caucasian (983 out of 1,173), 12.79% were African American (150 out of 1,173), 2.98% were Asian (35 out of 1,173), and 0.4% (5 out of 1,173) were American Indian/Alaska Native/Native Hawaiian or Other Pacific Islander. Patients (n=1,388) were diagnosed between 19-90 years (mean 59.8 ± 13.1 years), with LUAD and LGG having the highest and lowest mean ages, respectively (Figure 2.1a and Table 2.1).

Sequencing data for samples from the National Heart Lung and Blood Institute (NHBLI) Women's Health Initiative Exome Sequencing Project (WHISP) were downloaded, processed, and used for comparison of genetic variants to TCGA cancer cases. After extensive quality checks (see Methods), 1,039 Caucasians with an average age of 63.7 ± 7.9 years (mean ± s.d., range 50-79) were selected as controls for downstream burden test analyses (Figure 2.1b). NHLBI variant calls for 6,503 samples (4,300 Caucasians and 2,203 African American) were also downloaded from the NHLBI Exome Variant Server (ESP6500SI-V2, http://evs.gs.washington.edu/EVS/) for additional comparative analyses.

2.3.2 Landscape of germline truncation and missense variants

Germline variant calling was conducted using VarScan [10], GATK [11], and Pindel [12] for TCGA discovery (4,034) and validation (1,627) samples and WHI (1,039) controls. False-positive filters were applied to the intersected indel calls to ensure high quality for downstream

analyses. Missense variants were further analyzed by comparing to recurrent somatic mutation sites and IARC and ClinVar databases (Supplementary Note). Examination of coverages in the TCGA and WHI samples across the exome showed comparable depths, with averages of 115.3 and 106.2, respectively (see Supplementary Table 2.1). Specifically, there is a high positive correlation (Pearson Correlation R= 0.98) of the percentage of coding regions with at least 30X coverage between WHI (70.8%) and TCGA (71.4%) samples across the 624 cancer genes selected based on several recent studies [13-17] (Supplementary Figure 2.1).

We identified 2,089 truncation variants (splice site, frameshift indels, nonstop, and nonsense) in the TCGA discovery cohort in 624 cancer-associated genes (see Methods and Supplementary Table 2.2-2.4). We limited our analysis to variants whose minor allele frequency between our discovery dataset and that of NHLBI ESP 6,503 was ≤0.05%, based on the distribution of minor allele frequencies across *BRCA1* and *BRCA2* truncations detected (Supplementary Figure 2.2). After manual curation, we retained 838 truncation variants in 249 genes previously implicated in cancer (Supplementary Table 2.2); 69 of them with whole genome sequencing coverage have all been confirmed (Supplementary Table 2.4).

We conducted a more stringent investigation of the distribution of the rare truncation variants (MAF ≤ 0.05%) across cancer types using 2 different gene sets: 114 well-known cancer susceptibility genes reported by Rahman *et al*. [1] and 47 DNA repair genes associated with Fanconi Anemia pathway [3], with 15 overlapping between the two sets (Figure 2.1c, and Supplementary Table 2.5 and 2.6). Examination of the 114 susceptibility genes revealed that ovarian (19%, 95% confidence interval (CI): 16%-23%) and stomach (11%, 95% CI: 8%-14%) cancers have the highest percentage of cases carrying rare truncation variants, while AML (4%, 95% CI: 2%-8%) and GBM (4%, 95% CI: 3%-8%) have the lowest number of such events

(Figure 2.1c). Ovarian (17%, 95% CI: 14%-20%), prostate (8%, 95% CI: 4%-12%) and breast (8%, 95% CI: 6%-10%) cancers exhibit the highest percentage of cases harboring rare truncations when the 47 DNA repair genes associated with Fanconi Anemia pathway are included. Stomach cancer (8%, 95% CI: 5%-11%) in the Fanconi Anemia pathway-related genes also displayed relatively high truncation rates. Interestingly, LGG (2%, 95% CI: 1%-5%) and KIRC (3%, 95% CI: 2%-5%) have the lowest truncation rates in the Fanconi Anemia pathway-related genes, consistent with the small numbers of somatic variants identified in these two cancer types.

2.3.3 Genes significantly associated with cancer predisposition

Out of 4,034 total discovery cases, 3,125 were identified as Caucasians based on reported clinical data. We performed burden analysis in Caucasians (3,125 cases vs. 1039 WHI Caucasian controls, see Supplementary Table 2.7) using well-established methods [18, 19] (see Methods). To obtain the most comprehensive information, we also performed comparisons between the TCGA 4,034 cases and ESP 6,503 (downloaded variant calls, see Supplementary Table 2.8**)**. We searched for genes displaying significantly higher rare truncation variant frequencies than the background rate derived from WHI 1,039 control set (see Methods) and identified 13 significant genes (FDR $\leq$ 5%) using TFT calculations [19], 5 from cross cancer type analysis and an additional 8 from individual cancer type analysis, with *BRCA1*, *BRCA2, ATM, BRIP1* and *PALB2* as the top 5 ranked genes associated in the pan-cancer analysis, and other genes including *CNKSR1*, *EME2*, *MRE11A*, *MSH6*, *PIK3C2G*, *RAD51C*, *RAD51D*, and *XRCC2* associated with specific cancer types (Figure 2.2a and 2.2b, and Supplementary Table 2.7).

We detected 53 *BRCA1* rare truncation variants across 7 cancer types and 50 *BRCA2* rare truncation variants across 6 cancer types (Figure 2.2c). As expected, most variants were detected

in ovarian and breast cancer cases. However, 7 *BRCA1* and 6 *BRCA2* germline truncations (MAF ≤ 0.05%) were detected in other cancer types (3 each in endometrial, stomach, and lung cancers, 2 in kidney cancer, and 1 each in prostate and head and neck cancers). The average age at diagnosis of *BRCA1* and *BRCA2* germline truncation carriers vs. non-carriers was non-significantly younger for endometrial (52.7 vs. 63.1), stomach (59.7 vs. 66.1), and lung (63.0 vs. 66.1) cancers, providing support that these variants may contribute to younger onsets of these cancer types, though additional data is required for confirmation and to reach statistical significance. We also observed 32 truncations in *BRCA1* and *BRCA2* interacting proteins: *PALB2* (n=12, 4 in stomach, 3 in ovarian, 2 in head and neck and each in breast, lung, and prostate cancers), *BRIP1* (n=16, 3 each in breast, ovarian, and lung, 2 in stomach, one each in GBM, HNSC, KIRC, LGG, and UCEC), and *BAP1* (n=2, in kidney), and *BARD1* (n=2, 1 each in PRAD and BRCA). *ATM*, ataxia telangiectasia mutated, was the third most significant gene and the third highest in number of rare truncation variants; a total of 28 were found in *ATM* (23) and its homolog, *ATR* (5) (Supplementary Table 2.7). Our study bolsters evidence for previously claimed *ATM/ATR* associations with breast cancer with observations of 4 *ATM* and 4 *ATR* truncations in breast cancer cases. Notably, 19 *ATM* truncations were also detected in other cancer types, mostly in lung, stomach, and prostate cancers, the respective fractions of cases being 1.1% (5 out of 462 cases), 1.2% (4 out of 321 cases), and 3.4% (6 out of 178 cases). These fractions are all higher than the observed 0.5% in breast cancer. Both *ATM* and *ATR* are serine/threonine protein kinases that act upstream from cell cycle check point proteins *CHEK2* (6) and *CHEK1* (1), respectively. The rest of the significant genes were linked to various DNA repair pathways. For example, *MSH6* (11) is a component of the mismatch repair pathway and *XRCC2* (7), *RAD51C* (6), *NBN* (9) are all part of the DNA double strand repair pathway. *ERCC1*

(3) and *ERCC2* (10) are involved in transcription-coupled nucleotide excision repair. Four rare truncations (2 in LGG) were also found in *MUTYH* (a mutY homolog), involved in oxidative DNA damage repair (Supplementary Table 2.7**)**.

We also sought to identify genes enriched for truncations that were significantly associated with single or a subset of cancer types. *RAD51C* was found to be significant in OV and significant and top ranked in AML, while *PALB2* truncations were associated with STAD and OV (Figure 2.2b). *PMS2*, involved in colorectal [20] and endometrial cancer [21] predisposition, showed suggestive association with HNSC (TFT, FDR = 14%) and LGG (TFT, FDR = 10%) in the discovery set, but did not reach the 5% FDR threshold (Figure 2.2b). Significant enrichment of *MSH6* (6 were close to the C-terminus of the protein) and *MRE11A* truncations were found in UCEC. Other notable genes that were significant in a specific cancer type included *EME1* in KIRC and *FANCM* in BRCA (Figure 2.2b and Supplementary Table 2.7). Notably, we observed several novel associations between specific cancer types and genes, including *RAD51C* in AML, *ATM* in PRAD, *PALB2* and *EME2* in STAD.

To further evaluate these findings, we investigated rare truncations in those 13 significant genes, as well as an additional 21 suggestive genes having FDR ≤ 15% (TFT) using another independent set of 1,627 cancer cases from 10 of the 12 cancer types (see Methods). Our analysis showed that additional rare truncations (MAF ≤ 0.05%) were identified in 29 out of these genes in the validation set (Supplementary Table 2.3). The overall frequencies correlate positively (Pearson coefficient of 0.6167, Supplementary Figure 2.3). Notably, 10 rare *PMS2* truncations were found in the validation set, with 4 from UCEC, 2 each from LUAD and LUSC, and 1 each from BRCA and PRAD; these observations confirm the significance of *PMS2* in susceptibility

and broaden its role in cancer types not previously implicated. Another example is *XPA* detected as significant using the discovery cohort and confirmed by the identification of 2 additional rare truncations (E111* and V244fs) in prostate cancer using the validation cohort. Although 3 additional *ATM* rare truncations were found in BRCA and GBM in the validation cohort, no events were detected in LUAD and PRAD, two cancer types with significant results in the discovery cohort. Overall, our results from the validation cohort strengthen provisional conclusions derived in the discovery phase, but also indicate that larger cohorts are required for accurately assessing frequencies of germline mutations as well as detecting low frequency events in individual cancer types.

2.3.4 LOH analysis of rare truncation and missense variants

While burden analysis can identify genes with significant enrichment of rare truncations, association studies have limitations, specifically with respect to inference about functional implications of specific variants. LOH analysis can uncover heterozygous germline variants that are under potential selection in the tumor, one of the key indications being increased VAF in the tumor sample. With no LOH, it would be expected that the VAF detected in tumor relative to the normal tissue derived DNA would be 1 while with complete LOH the VAF ratio would be 2. Because tumor samples are not completely free of normal tissue and can exhibit clonal heterogeneity, evidence for LOH is increasingly strong for VAF ratios approaching 2. The combined use of burden tests that can narrow the search space for germline variants of functional importance with LOH analysis can solidify support for both putative genes and specific variants involved in cancer susceptibility.

With respect to genes, we first tested the expanded list of 34 significant or nearly significant genes (known and likely oncogenes excluded) in burden analysis (see Methods) for evidence of

somatic loss of the wild-type allele. A total of 7 genes, *BRCA1, BRCA2, RAD51D, PALB2, RAD51C, ATM,* and *BRIP1* were significant (FDR ≤ 5%) along with 2 genes (*BAP1* and *FANCM*) near significance **(**Supplementary Table 2.9 and 2.10, and Figure 2.2 and 2.3a**)**. Consistent with expectations, *BRCA1* and *BRCA2* had the highest percentage of significant variants demonstrating LOH (44 of 48 (92%) and 21 of 30 (70%), respectively). Other genes demonstrating variants with LOH include: *PALB2*, which functions in maintenance and repair and cooperates with *BRCA2* [22] (5 significant truncation mutations of 11, 45%), *ATM*, which is activated by double-strand breaks (8 of 17 significant, 47%), *BAP1*, a transcriptional repressor involved in BRCA1-mediated cell growth suppression [23] (2 of 2, 100%), and *FANCM,* which plays a role in DNA repair [24] (3 of 9, 33%). In all, 99 of 264 (38%) truncation variants showed significant LOH. It is worth noting that although LOH in cases with *BRCA1* and *BRCA2* truncations mutations were largely restricted to OV and BRCA, the majority of LOH truncations in other genes (e.g., *ATM*, *PALB2*, *BAP1*, *FANCM*) were found across cancer types (Figure 2.3a).

We further compared VAFs of missense variants in the 7 significant LOH genes above, finding that 4 in *BRCA1*, *ATM*, *BRCA2*, and *RAD51C* are significant. This underscores both our findings from rare truncation analysis (Supplementary Table 2.11 and 2.12, and Figure 2.3b) and the potential importance of missense events in cancer. The significant missense VAFs in these genes range from 13% to 23% (Figure 2.3b), while other genes average 9%. Of all individual missense events, 173 of 1170 (11%) showed significant LOH (FDR ≤ 1%) (Supplementary Table 2.12). Significant events for *ATM* and *BRCA1* were concentrated in BRCA, HNSC, and OV, while *RAD51C* did not show preference (Figure 2.3b). Of note, our LOH analysis identified G245V in *TP53* as highly significant (FDR = 1.18e-07) although no rare *TP53* truncations were found.

To further investigate the effect of missense events on cancer susceptibility, we sought to determine whether there are any larger informative patterns associated with their LOH, specifically whether the significant instances of LOH spatially cluster in or near specific protein regions/domains. Indeed, analysis shows statistically significant difference in spatial clustering, further supporting the mechanistic roles of these variants in cancer (Figure 2.3c). For example, there is a strong grouping of variants (FDR = 0.34%) that overlaps both a kinase-like and a PIK kinase domain near the end of *ATM*, which participate in chromosome maintenance and repair. We also found clusters overlapping the BRCT (FDR = 5%) and RING domains (FDR = 0.39%), which participate in the DNA repair functionality of *BRCA1*. 2 *BRCA2* clusters (FDRs = 6.5% and 8.9%) in the oligonucleotide/oligosaccharide binding motif (OB fold) domains, important in the DNA damage response, are near significant (Supplementary Table 2.13).

2.3.5 Somatic and germline interactions and clinical associations

We followed stringent filtering strategies for standardizing specificity across the Pan-Cancer somatic variant calls for 3,368 cases in this study [13] (Supplementary Table 2.14). We first used MuSiC [25] to search for genes demonstrating co-occurring or mutually exclusive germline and somatic mutations (Figure 2.4a,b and Supplementary Table 2.15,2.16). Our pan-cancer analysis using 34 burden test genes of interest and 54 cancer-associated genes with recurrently mutated somatic variants (frequency ≥ 5 across cancer types), detected significant mutual exclusivity between *BRCA1*/*BRCA2* germline truncations and *IDH1* somatic mutations, which is likely confounded by cancer-type specificity: *BRCA1*/*BRCA2* germline truncations were most prevalent in BRCA and OV, whereas *IDH1* somatic variants are mostly found in AML, GBM and BLCA. To mitigate the cancer type specific effect, we investigated co-occurrence and mutual exclusivity within each cancer type (requiring recurrently mutated somatic variants with frequency ≥ 2

across cancer types) (Supplementary Table 2.16). Notably, *ATM* germline truncations were found to be mutually exclusive of *TP53* somatic mutations in LUAD (permutation test, P = 0.041), consistent with the paradigm that ATM activates TP53 to trigger apoptosis [26] and the need to disrupt only one gene to confer an anti-apoptotic effect. As expected, we also observed co-occurrence of *BRCA1* germline truncations and *TP53* somatic mutations in BRCA (permutation test, P = 0.012) [27], as well as mutual exclusivity between *BRCA1*/*BRCA2* germline truncations and *PIK3CA* somatic mutations in BRCA (permutation test, P = 0.01 and P = 0.03). *BRCA1* germline truncations have previously been reported to be associated with the basal subtype breast cancer [28], which tends to exhibit a molecular profile similar to ovarian cancer [29]. Our findings are consistent with the association between basal subtype breast cancer and frequent *TP53* and infrequent *PIK3CA* mutations [30]. Additionally, we also observed a co-occurrence of *BRCA2* germline truncations and *TP53* somatic mutations in ovarian cancer, as expected. Our data suggest that the combinational effects of *BRCA1*/*BRCA2* germline mutations, along with the high frequency of LOH events and somatic *TP53* mutations result in aggressive basal subtype breast cancer and ovarian cancer.

Interestingly, the distribution of *BRCA1*, *BRCA2*, and *ATM* rare germline truncations with their somatic mutations across cancer types varies with the high frequency of *ATM* in prostate, lung, and stomach cancers and *BRCA1* and *BRCA2* germline events in ovarian and breast cancers (Figure 2.5a and Supplementary Table 2.17). Collectively, these analyses show distinct combinations of germline and somatic mutations contribute to the development of individual cancer types.

We also examined germline variants having significant impact on carriers' somatic mutation frequencies. Analysis of the expanded 34 burden test genes revealed that patients with germline

BRCA1 and *BRCA2* truncations had significantly higher somatic mutation frequencies than cases without such changes in both breast and ovarian cancers (Figure 2.5b and Supplementary Table 2.18). Since the correlation between BRCA1/2 germline and higher somatic mutation rate may be characteristic of the basal subtype breast cancer, we compared the mutation frequency of basal cases with BRCA1/2 germline truncation to basal cases without BRCA1/2 germline truncation and found the former have significantly higher mutation rate (Supplementary Figure 2.4, Wilcoxon rank-sum test, P = 9e-4).

In addition, *RAD51C* and *RAD51D* germline truncations are positively correlated with increased somatic mutation frequencies in ovarian cancer. *FANCM* and *EME1* germline truncations are positively correlated with increased somatic mutation frequencies in HNSC (Wilcoxon rank-sum test, P = 0.046) and KIRC (Wilcoxon rank-sum test, P = 0.027), respectively. In UCEC, *MSH6* germline truncations are found to be significantly associated with higher mutation frequencies, as expected (Wilcoxon rank-sum test, P = 0.014) (Figure 2.5b and Supplementary Table 2.18). Further, 81 cases carried *MSH2* germline variants (MAF ≤ 0.05%, including 1 truncation variant), and they also showed higher somatic mutation frequency (Wilcoxon rank-sum test, P = 3.63e-03).

The joint analysis of all 12 cancer types including cancer type as a covariate identified *BRCA1*, *BRCA2*, and *PMS2* as having strong correlations with a younger age of onset (P = 5.20e-07, 2.04e-04, and 0.049, respectively; MuSiC GLM analysis, Figure 2.5c and Supplementary Table 2.19). Analysis of individual cancer types revealed significant early onset for germline truncations of *FANCA* in HNSC, *BRIP1* in LUSC, and *ATM* in STAD (Figure 2.5c and Supplementary Table 2.20). Not surprisingly, we found that germline truncation variants in 47 Fanconi Anemia genes and 114 cancer susceptibility reported in Rahman et al. were significantly

enriched in younger patients according to Wilcoxon rank-sum testing (P = 1.08e-03 and 1.38e-04, respectively).

2.3.6 Functional validation of *BRCA1* missense variants

To investigate the effect of missense variants on *BRCA1* function and evaluate LOH analysis for missense variants, 68 variants were selected based on MAF and protein domains for functional validation using the HDR assay [31] (see Methods and Supplementary Table 2.21); 47 of them had previously been assigned as variants of unknown clinical importance in the NHGRI Breast Cancer Information Core (BIC) database and 15 variants were not reported at all in BIC. One known deleterious truncation mutation in the carboxyl-terminus of the *BRCA1* protein Q1779fs and 3 other truncations, E1250*, E1415fs and E23fs discovered in UCEC, were also included in the experiment. We successfully introduced 68 missense variants and 4 truncation variants into full-length *BRCA1* expression plasmid pcDNA-5'HA-BRCA1 for the *in vitro* HDR assay as previously described [31, 32] (Supplementary Table 2.21). All mutant constructs were confirmed by sequencing and protein expression (Supplementary Figure 2.5) and tested in triplicate using the *in vitro* assay. The percentages of cells showing GFP expression were normalized to homologous recombination levels observed in cells depleted of endogenous BRCA1 and rescued by transfection of the wild-type BRCA1 expression vector (see Methods).

Among all tested variants, all 4 truncations (3 from UCEC) and 6 missense variants retained less than 30% of homologous recombination activities relative to wide-type (WT) *BRCA1,* and are therefore considered HDR-defective (Supplementary Table 2.22). These missense variants included C61G (observed in 4 cases), C64G (2 cases), T1685I (1 case), R1699W (2 cases), L1786P (1 case) and G1788V (1 case); all of them showed significant enrichments in the tumor samples based on LOH analysis (Figure 2.6a). Comparative analysis of RNA-seq data from 2

carriers and 4 non-carriers suggests C64G is in fact a variant affecting splicing (Supplementary Figure 2.6), consistent with a previous report [33], and our results suggest that should some of the C64G mRNAs be properly spliced, the protein is not active in DNA repair. Of particular interest, L1786P, identified and validated as HDR-defective in our study, has not been previously designated as pathogenic, despite observations in two previous studies [34, 35]. Our analysis of the crystal structure of the BRCT domain showed that the substitution of leucine with proline in L1786P will likely result in the termination of the alpha helix structure, which may cause the loss of *BRCA1* HDR function. Interestingly, an additional 3 variants, A1708V, M1783T and R1835Q (from one patient each) consistently displayed less than 70% HDR function in comparison to WT *BRCA1* (partial HDR-defective, Figure 2.6a); all three had previously been designated as variants of unknown significance (VUS) in the BIC database. It is worth noting that A1708V and R1835Q were found in male patients with kidney and stomach cancers, respectively; both developed cancers at age of 48. A1708V has previously been characterized as a low to moderate risk variant [36] and R1835Q has been identified in a Malay population of early-onset breast cancer patients with a personal or family breast cancer history [37]. One endometrial cancer patient harboring M1783T was diagnosed at age of 65. The BRCA1 protein harboring this variant was previously shown to possess enhanced protease sensitivity [38]. Further, our analysis shows that all 7 HDR-defective or partial defective missense variants from the BRCT domain are either positioned in the center of the structure or on the surface responsible for protein-protein interactions, while the 5 HDR-WT variants from the BRCT domain tested are mapped to the periphery of the structure (Figure 2.6b). In addition, these 9 HDR-defective (or partial HDR-defective) missense variants are mutually exclusive to *BRCA1* somatic mutations and germline truncation variants (Supplementary Table 2.14 and Supplementary Table 2.2).

Using the systematic *BRCA1* missense variant validation data, we evaluated the prediction power of LOH analysis for identifying candidate variants of functional relevance. Without LOH analysis filtering, we observed a rate of 4.7% (3 of 64 validated), but *BRCA1* validation of candidates filtered through LOH was 38.1% (8 of 21) (Supplementary Table 2.23). The significant difference (P-value = 0.0004, Fisher's test) suggests LOH offers an effective sieve for candidates, which in this case gives an estimated enrichment factor of 8-fold.

2.4 Discussion

This study of over 4000 cancer cases is the largest integrated analysis of germline and somatic variants to date. Our systematic analysis indicated that an estimated 18% of cancer cases from the TCGA cohort had one or more rare truncations in 624 genes associated with cancer. Further, there was significant enrichment of rare truncation variants in 13 genes and suggestive evidence of increases in 21 more, comprising 8.3% (333 out of 4,034) of TCGA cancer cases.

We observed several significant associations in specific cancer types: *RAD51C* in AML, *ATM* in PRAD, *PALB2* in STAD. Across cancer types, a higher percentage of breast and ovarian cancer cases were identified as having rare truncation variants in cancer genes versus other cancer types, due predominantly to high frequencies in *BRCA1/2*. The percentage of breast and ovarian cancer cases carrying *BRCA1/2* germline truncation variants in the TCGA cohort was 4.4% and 11.6%, respectively, consistent with previous reports [39-42]. Interestingly, stomach cancer has the second highest percentage of rare truncations in 114 genes previously reported [1], largely due to the contributions from *ATM*, *BRIP1*, *PALB2*, *XRCC2*, and others. In contrast, for KIRC and GBM, truncation variants in the 34 associated germline genes were uncommon, identified in only less than 6% of cases (Figure 2.2d). These results contribute to our understanding of the

genetic architecture in cancers, complementing the known effect of common and tagged variants from array-based studies as well as the estimate of overall heritability from twin studies in multiple cancer types [43, 44].

Our results indicated that germline truncation and missense variants in several genes were under selection in the tumor, with *ATM, BRCA1, BRCA2,* and *RAD51C* determined as significant from both truncation and missense analyses and *BAP1, BRIP1, FANCM, PALB2,* and *RAD51D* from truncation analysis alone. As a proof of concept, we performed functional validation for 68 *BRCA1* missense variant sites using HDR assay; our experimental efforts identified 9 variants from 14 patients with complete or partial defective HDR function and validated our LOH analysis for effective enrichment of variants under functional selection (an estimated 8-fold enrichment in *BRCA1*).

More importantly, our integrated germline and somatic study identified *BRCA1*, *BRCA2*, *RAD51C*, *RAD51D*, *FANCM*, *EME1* and *MSH6* germline truncations significantly associated with increased somatic mutation frequencies in specific cancer types, suggesting that germline defects in DNA repair expand to the somatic level. Further, our search for co-occurring or mutually exclusive germline truncation/somatic mutations across 12 cancer types revealed a number of important insights in terms of genes and pathways involved including: 1) the association between germline *BRCA1/2* germline truncations and frequent *TP53* and infrequent *PIK3CA* somatic mutations confirm breast cancer clinical subtype classification and 2) *ATM* as a bona-fide (3rd frequently truncated) susceptibility gene demonstrated by both burden and LOH analyses, is the only common gene highly mutated at both germline and somatic levels.

Although our study has been revealing at a genetic level, we are mindful of the limitations of the TCGA dataset, including the lack of detailed family history information that would further inform the potential pathogenicity of germline variants. Despite the large sample size overall, our inferences are limited for specific cancer types because of small case numbers. In addition, the vast majority of TCGA cases in our sample set were of Northern European background, emphasizing the need for the development of a reference source of genomic data on germline cancer predisposition variants from ancestrally diverse population groups. Nonetheless, this study is the largest to date that has integrated somatic and germline alterations to identify important genes across 12 major types contributing to cancer susceptibility and our results provide a promising list of candidate genes for definitive association and functional analyses. The combination of high throughput discovery and experimental validation should identify the most functionally and clinically relevant variants for cancer risk assessment.

2.5 Methods

2.5.1 Access and Inclusion

Approval for access to TCGA case sequence and clinical data was obtained from the database of Genotypes and Phenotypes **(***dbGaP***)** (document #3281 Discover germline cancer predisposition variants). We selected a total of 4,034 discovery cases and 1,627 validation cases with germline and tumor DNA sequenced by exome capture followed by next generation sequencing on Illumina or SOLiD platforms. All cases met our inclusion criteria of 50% coverage of the targeted exome having at least 20X coverage in both germline and tumor samples.

2.5.2 Control cohort

NHLBI variant calls for 6,503 samples (2,203 African-Americans and 4,300 European-Americans unrelated individuals) were downloaded from the NHLBI GO Exome Sequencing

Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/; accessed on August 26, 2013). For comparative analysis, all ESP variants were filtered for < 0.1% total MAF to minimize false positives. For the Women's Health Initiative Sequencing Project (WHISP) sample set (N = 1039) as part of the NHLBI ESP cohort, we performed variant analyses using methods described in the following section. All variants were processed using the same tools as for the TCGA cohort. dbGaP accession id for NHLBI ESP is phs00281.

2.5.3 Germline variant calling and filtering

Sequence data from paired tumor and germline samples were aligned independently to NCBI Build 37 of the human reference using BWA v0.5.9 and de-duplicated using Picard 1.29. Germline SNPs were identified using Varscan (2.2.6 –min-var-freq 0.10 --p-value 0.1 --min-coverage 8 –map-quality 10), and GATK (revision5336) in single-sample mode for normal and tumor BAMs. For breast and endometrial cancer samples, we also used population-based methods, but found differences to be minimal. Germline indels were identified using Varscan 2.2.9 (--min-coverage 3 –min-var-freq 0.2 –p-value 0.10 –strand-filter 1 –map-quality 10) and GATK (revision5336, only for AML, BRCA, OV, and UCEC) in single sample mode. We also applied Pindel (version 0.2.4x, May 8, 2013; --window-size 1) on each pair of tumor and germline sequencing data (for some samples, multiple normal files are used if available) for indel prediction. For the analysis, we preset the insertion size to 500 if this information was not provided in the BAM header.

For each cancer type, all variants were limited to coding regions as defined by ensemble 70. In addition to the coding regions, the two base pairs flanking each exon to cover splice donor/acceptor sites were included. SNVs were based on the union of GATK and VarScan. They were subsequently processed through our in-house false-positive filter (all default

parameters --min-homopolymer 10). We required that indels were called by at least 2 out of 3 callers (GATK, Varscan, Pindel) when all three callers were applied. In addition we also included Pindel unique calls (at least 30X coverage and 20% VAF). All combined indels were then processed through our false-positive-filter (all default --min-homopolymer 10 –min-var-freq 0.2 --min-var-count=6). We then applied additional annotation and minor allele frequency filters as previously reported [45].

The predictions for 4,034 TCGA cases consist of 2,709,906 variants (1,655,391 missense, 947,045 silent, 36,009 nonsense, 18,693 splice site, 2,041 nonstop/readthrough, 30,508 frameshift indels and 20,219 in frame indels) with minor allele frequency ≤1% in 1000 Genomes, ESP 6,503 dataset, Discovery 4,034 cohort, and additional annotation filters as previously reported [3]; of these, 1,842,459 variants were from 3,125 Caucasian TCGA cases. Using the same processing for the 1,039 WHI Caucasian controls, we identified 516,219 variants, consisting of 319,698 missense, 176,862 silent, 6274 nonsense, 3541 splice site, 355 nonstop/readthrough, 6101 frameshift indels, and 3568 in-frame indels.

2.5.4 Cancer Associated Genes

A total of 624 candidate cancer-associated genes were compiled from nine sources, including recently published large-scale cancer studies, publicly available screening panels, and unpublished preliminary analysis of publicly available data sources. We retained 204 genes shared across at least two of the nine sources and a literature search was conducted to identify evidence supporting inclusion of any remaining unique genes. A subset of 518 genes originated from recent publications, including 294 genes from Frampton *et al.* [17], 125 from Kandoth *et al.* [13], 212 from Lawrence *et al.* [15], 194 from Pritchard *et al.* [16], 114 from Rahman [1], and 124 from Vogelstein *et al* [14]. Thirty-nine additional genes were included based on the analysis

of driver mutations in publicly available TCGA data, the published guidelines for return of results of the American College of Genetics and Genomics [46], and 18 novel cancer driver genes identified in recently published large-scale studies.

2.5.5 Germline sites overlapping with recurrent somatic mutations

Recurrent somatic mutations were extracted from the high confidence filtered set of somatic mutations [13] and germline variants overlapping them were further filtered to remove those having a reported global MAF < 0.5% in the NHLBI Exomes (ESP6500SI-V2). Remaining variants were filtered to remove artifacts due to ambiguous alignments, simple repeats, reference sequence errors, putative somatic mutations in adjacent normal tissue, somatic mutations associated with clonal expansion in blood [47], and variants with a VAF<10% in tumor or normal. No germline mutations were found to overlap somatic mutations in the same individual.

In addition to sites described in the main text, several rare germline variants overlapping somatic mutations in genes associated with toxin metabolism were also identified. This included three cases carrying *CYP2D6* (H352R) as well as one carrier of *ABCC2* (E943K; rs3740065). Variants in both genes have been reported to be associated with poor outcome in post-menopausal women treated with tamoxifen but their association with cancer predisposition remains undetermined [48, 49]. Additionally, a germline variant at somatic R423Q site was found in the *CARD11* oncogene [50] and another germline variant S650L in *PDGFRB* was identified. Interestingly, a *FLT3* germline variant (R387Q) was identified to have an overlapping somatic mutation in endometrial cancer.

2.5.6 Identifying significant genes using burden tests

We determined the MAF cutoff for rare variants as 0.05% based on balancing the inclusion of possible false-positives versus the loss of possible true-positives in subsequent burden test and

LOH analysis. For example, if one presumes that p-values <0.01 have a reasonable possibility of being retained as significant in a multiple hypothesis test, the 0.05 threshold only excludes 2 such points out of a total of 47 for *BRCA1* and 1 such point out of a total of 52 for *BRCA2*. Conversely, it excludes 24 points in the MAF range up to 1% that are very unlikely to show significance. Points having MAF > 1% are likewise not likely to be of interest (Supplementary Figure 2.2).

Burden test analysis was performed by comparing the frequency of rare germline truncation mutations in cancer associated genes from the Pan-Cancer 12 germline dataset (from 12 cancer types) (cohort size = 4,034) with WHI 1,039 control samples and those downloaded from the NHLBI Exome Sequencing Project (ESP 6,503 including 2,203 African-Americans and 4,300 European-Americans unrelated individuals). Variant calling on the TCGA and WHI dataset was done as previously described in the methods section. Variants for the ESP 6,503, along with their minor allele frequency were downloaded from http://evs.gs.washington.edu/EVS/). The truncation variants (nonsense, splice_site, and frameshift indels) from both groups were limited to a list of genes previously associated with cancer (See Cancer Associated Genes section). Further filtering includes retaining variants with < 1% minor allele frequency from 1000 Genomes Project, and < 1% cohort frequency in each cancer type. A pooled minor allele frequency (the average minor allele frequency of each variant between the test and control group) was calculated for each variant and only those whose pooled minor allele frequency was less than 0.05% were kept for burden analysis. We excluded events having insufficient numbers of observations, defined here as fewer than 3 in the combined cases and controls for the ESP cohort and fewer than 2 in the WHI cohort. We subjected the data to the Total Frequency Test (TFT), evaluating the one-tailed P-value in each case (observations significantly greater than

controls). For reference, we also evaluated the data using the Cohort Allelic Sum Test (CAST), although these results were not carried forward for analysis, because they correlate with TFT. The TFT probabilities were then ranked by the standard False Discovery rate (FDR). This procedure was performed for each cancer type vs. the control group. In addition an overall burden test was performed for Pan-Cancer 12 germline dataset vs. the control group. A FDR cutoff of 10% for the Pan-Cancer 12 germline dataset was used.

2.5.7 Statistical Methods of Loss of heterozygosity (LOH) Analysis

Next-generation sequencing provides direct read counts of reference and variant alleles and each pair of counts comprises an observational sample of the actual variant allele fraction (VAFs) at its site. We devised several statistical procedures using these counts to test for AI at sites within a subset of genes hypothesized to be relevant across cancer types and, moreover, to test the genes themselves for significant content of such sites. This is one component of a larger method to assess loss of function alleles in these genes.

The evaluation at each tumor variant site (truncation or missense) is based on two complementary aspects related to its VAF: (1) whether it is significantly higher than the VAF at its corresponding site in the matched normal sample and (2) whether it is significantly higher than the characteristic VAF in the general population of genes having somatic mutations. The first aspect was implemented using Fisher's exact test [51] on a 2X2 table of allele type (reference and variant) vs. sample type (tumor and normal). For the second test, we permuted all combinations of reference counts and variant counts of the somatic events for all other genes, thus obtaining a null distribution that can be used for computing tailed p-values.

Each of these 2 calculations uses some component of unique information not available to the other: they are essentially independent tests of the same hypothesis. We used a standard

transformation method from the mathematical statistics literature to combine these values into a single, overall result [52]. The list of p-values for the entire complement of tested sites was then corrected for multiple hypothesis testing bias and ranked using the standard Benjamini-Hochberg False Discovery Rate (FDR) calculation [53].

For the second type of test at the gene level, we took the following approach for truncation events. All mutated sites for the candidate gene were cataloged, as were all sites outside of that gene, the latter representing the mutation "background". The statistical difference between the two sets was then calculated using a standard difference-of-means t-test on the tumor variant allele fractions of the 2 groups, where the number of degrees of freedom is 2 less than the total number of sites in the test. This procedure was repeated for each gene of interest, after which multiple testing correction was again applied in the context of FDR. With respect to missense events, we found this procedure was not sufficiently sensitive, so we used an alternative test based on comparing the fraction of missense sites within each gene that showed significant LOH on the individual level to the corresponding fraction in a background set consisting of the genes from burden testing that did not show significant LOH for truncations. To minimize noise, we adopted somewhat strict criteria for this particular test: to be tallied as LOH, a site must have had a maximum of 1% FDR in the site test and we only tested genes that satisfied the following inclusion criteria: a minimum difference of fractional values of 2 percentage points and at least 3 events showing LOH at the 1% FDR level. We then applied Fisher's exact test on 2X2 tables of missense type (significant LOH and no discernable imbalance) vs. cohort (test gene and background population), after which FDR was once again applied to the result.

An important aspect of the above methods is pre-conditioning of inputs. Previous studies [54] have discarded sites based on their inability to attain a significant P-value under the test being

used, pointing out incidentally that excluding sites directly improves FDR. The latter observation is undoubtedly true, but this view misses the importance of the confidence level associated with a VAF estimate, as determined by the size of the sample used for its computation. Because of depth variations both between samples and within samples, the reliability (confidence) of VAF estimates as calculated from read counts varies from site to site. A specified confidence interval for each VAF furnish is a rigorous metric upon which reliability can be assessed and low-reliability points subsequently excluded from analysis. Because VAFs can approach the extremes of 0 and 1 and are also sometimes based on only 10 or 15 reads, the standard interval from sampling theory is not particularly useful. Instead, we used Wilson's interval [55], which does not suffer appreciably in these circumstances. We chose an interval of 90% confidence (Z-score of approximately 1.65), removing events whose larger distance (above or below the calculated VAF) exceeded 12%. The remaining "high-quality" data were then used in the tests described above. Results having FDR ≤ 20% were prioritized as significant.

2.5.8 LOH analysis of germline truncations and missense variants

We applied our LOH analysis method as a refinement step to the burden analysis. Specifically, we tested all sites (and by extension the genes containing those sites) that burden testing identified as being significant, either in a Pancancer context or as associated with a specific cancer type. Here, we used a FDR of 15% to capture the widest set of genes that could be significant. In a sense, we used our LOH method as a "confidence filter" situated on top of burden analysis to eliminate false positives. With oncogenes removed, the list of candidates at this stage consisted of 32 genes, including *ATM, BRCA1, BRCA2, BRIP1, MSH6,* and *RAD51C* (the "burden test genes"). We also separated missense from nonsense alterations, the latter

typically resulting in truncated, non-functional protein products, and analyzed these sets separately.

The statistical procedure outlined above is straightforward, but can be applied in various ways. For assessing the burden test genes, we selected each one individually and constructed the corresponding null distribution from all remaining non-burden test genes. That is, we excluded from the null all those genes for which there was already some evidence of possible significance. The same principle applied to testing individual sites: no variants from burden test genes were included in the null distributions.

2.5.9 Calculation of hotspots of significant LOH

The calculation method for LOH discussed above identifies instances where the observed VAF in the tumor is higher than what is attributable to chance. Building on this, we now describe a subsequent calculation that identifies groups of such instances that are clustered spatially. These groupings are so-called "hotspots of significant LOH" and signal likely biological relevance. The null hypothesis is that instances of LOH, whether statistically significant or not, are distributed randomly. Since we are primarily interested in discovery, test regions are implemented as unbiased "sliding windows" rather than as specific domains, linkers, etc. A relevant LOH observation must satisfy 2 conditions:

condition *A*: the LOH is statistically significant, as described above

condition *B*: the LOH resides within the current test window

Status and spatial placement are independent of one another, meaning that the Bernoulli probability of a single LOH observation can be calculated as

$$p_b = P(A \cap B) = \frac{D_s \cdot W}{(D_s + D_n) \cdot L} \quad (1)$$

where W and L are the sizes of the test window and protein, respectively (in units of amino acids) and D_s and D_n are the total numbers of significant and non-significant LOHs observed for the protein. This expression indicates observations are of greater weight to the degree that the significant LOHs are more rare (as compared to non-significant LOHs) in the test set and that they cluster within tighter regions. LOHs are independently and identically distributed under the null hypothesis, meaning the mass probability of k observations of significant LOH within the window is then

$$P(k) = \binom{D_s + D_n}{k} \cdot p_b^k \cdot (1 - p_b)^{D_s + D_n - k} \quad (2)$$

and the significance (test) probability of k observations within a test window is

$$P_S = P(k) + P(k+1) + \cdots + P(D_s + D_n) \quad (3)$$

Since p_b is constant over a given protein for a given window size, appreciable caching can be used to economize the calculation. We use a slide-step of 1 amino acid and scan window sizes from 30 to 200, taking regions of significance to be characterized by their smallest P_S. The software automatically merges overlapping significant regions. Standard FDR analysis, as described above, is then performed on the resulting list of hotspots.

2.5.10 Functional validation of BRCA1 variants

Variants were incorporated into a full-length *BRCA1* expression plasmid, pcDNA-5'HA-BRCA1, using Q5 site-directed mutagenesis kit (New England BioLabs). Primer sequencesare

available in the Supplementary table 2.21. All of the desired variants were confirmed by sequencing.

HeLa-DR cells, a stable derivative of HeLa cells containing the genomic integration of the recombination substrate vector, pDR-GFP were used for the homology-directed recombination assay. Co-transfection of HeLa-DR cells with the *BRCA1* expression plasmid containing the test variant and siRNA targeting the 3'-UTR of the *BRCA1* gene to deplete endogenous BRCA1 expression was performed. Two days later, cells were transfected again with the siRNA, BRCA1 expression plasmid, and the I-SceI expression plasmid. After 3 days, cells were harvested by trypsinization and the fraction of GFP-positive cells was determined using a FACScalibur flow cytometer (BD Biosciences model E1202). The plasmids and cell line used in this study have been described previously [31].

All *BRCA1* variants were tested in triplicate and the percentage of cells with GFP expression was normalized to the rescue by wild-type BRCA1 expression plasmid.

2.6 References

1. Rahman, Nazneen. Realizing the promise of cancer predisposition genes. *Nature* 505, 302-308 (2014).

2. Walsh, Tom, Silvia Casadei, Ming K. Lee, Christopher C. Pennil, Alex S. Nord, Anne M. Thornton, Wendy Roeb *et al*. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proceedings of the National Academy of Sciences* 108, 18032-18037 (2011).

3. Kanchi, Krishna L., Kimberly J. Johnson, Charles Lu, Michael D. McLellan, Mark DM Leiserson, Michael C. Wendl, Qunyuan Zhang *et al*. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications* 5 (2014).

4. Schwartz, Ann G. Genetic Epidemiology of Cigarette Smoke–Induced Lung Disease. *Proceedings of the American Thoracic Society* 9, 22-26 (2012).

5. Bodmer, Walter, and Ian Tomlinson. Rare genetic variants and the risk of cancer. *Current opinion in genetics & development* 20, 262-267 (2010).

6. Yu, Wei, Ajay Yesupriya, Anja Wulf, Lucia A. Hindorff, Nicole Dowling, Muin J. Khoury, and Marta Gwinn. GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *European Journal of Human Genetics* 19, 1095-1099 (2011).

7. Knudson, Alfred G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* 68, 820-823 (1971).

8. Nomoto, Shuji, Nobuhiro Haruki, Yoshio Tatematsu, Hiroyuki Konishi, Tetsuya Mitsudomi, Toshitada Takahashi, and Takashi Takahashi. Frequent allelic imbalance suggests involvement of a tumor suppressor gene at 1p36 in the pathogenesis of human lung cancers. *Genes, Chromosomes and Cancer* 28, 342-346 (2000).

9. Saito, Misato, Aikou Okamoto, Takashi Kohno, Satoshi Takakura, Hideo Shinozaki, Seiji Isonishi, Takaomi Yasuhara *et al*. Allelic imbalance and mutations of the PTEN gene in ovarian cancer. *International Journal of Cancer* 85, 160-165 (2000).

10. Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22, 568-576 (2012).

11. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20, 1297-1303 (2010).

12. Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865-2871 (2009).

13. Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).

14. Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science* 339, 1546-1558 (2013).

15. Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature* 505, 495-501 (2014).

16. Pritchard, Colin C., Stephen J. Salipante, Karen Koehler, Christina Smith, Sheena Scroggins, Brent Wood, David Wu *et al*. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of Molecular Diagnostics* 16, 56-67 (2014).

17. Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin *et al*. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology* 31, 1023-1031 (2013).

18. Morgenthaler, Stephan, and William G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615, 28-56 (2007).

19. Basu, Saonli, and Wei Pan. Comparison of statistical tests for disease association with rare variants. *Genetic epidemiology* 35, 606-619 (2011).

20. Haghighi, Mahdi Montazer, Ramin Radpour, Katayoun Aghajani, Narges Zali, Mahsa Molaei, and Mohammad Reza Zali. Four novel germline mutations in the MLH1 and PMS2 mismatch repair genes in patients with hereditary nonpolyposis colorectal cancer. *International journal of colorectal disease* 24, 885-893 (2009).

21. Stewart, Alison. Genetic testing strategies in newly diagnosed endometrial cancer patients aimed at reducing morbidity or mortality from lynch syndrome in the index case or her relatives. *PLoS currents* 5 (2013).

22. Sy, Shirley MH, Michael SY Huen, and Junjie Chen. PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proceedings of the National Academy of Sciences* 106, 7155-7160 (2009).

23. Jensen, David E., Monja Proctor, Sandra T. Marquis, Heather Perry Gardner, Seung I. Ha, Lewis A. Chodosh, Alexander M. Ishov *et al*. BAP1: a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression. *Oncogene* 16, (1998).

24. Gari, Kerstin, Chantal Décaillet, Alicja Z. Stasiak, Andrzej Stasiak, and Angelos Constantinou. The Fanconi anemia protein FANCM can promote branch migration of Holliday junctions and replication forks. *Molecular cell* 29, 141-148 (2008).

25. Dees, Nathan D., Qunyuan Zhang, Cyriac Kandoth, Michael C. Wendl, William Schierding, Daniel C. Koboldt, Thomas B. Mooney *et al*. MuSiC: identifying mutational significance in cancer genomes. *Genome research* 22, 1589-1598 (2012).

26. Cheng, Qian, Lihong Chen, Zhenyu Li, William S. Lane, and Jiandong Chen. ATM activates p53 by regulating MDM2 oligomerization and E3 processivity. *The EMBO journal* 28, 3857-3867 (2009).

27. Buller, Richard E., Thomas A. Lallas, Mark S. Shahin, Anil K. Sood, Melanie Hatterman-Zogg, Barrie Anderson, Joel I. Sorosky, and Patricia A. Kirby. The p53 mutational spectrum associated with BRCA1 mutant ovarian cancer. *Clinical cancer research* 7, 831-838 (2001).

28. Sørlie, Therese, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* 100, 8418-8423 (2003).

29. Hoadley, Katherine A., Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max DM Leiserson *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929-944 (2014).

30. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61-70 (2012).

31. Ransburgh, Derek JR, Natsuko Chiba, Chikashi Ishioka, Amanda Ewart Toland, and Jeffrey D. Parvin. Identification of breast tumor mutations in BRCA1 that abolish its function in homologous DNA recombination. *Cancer research* 70, 988-995 (2010).

32. Pierce, Andrew J., Peng Hu, Mingguang Han, Nathan Ellis, and Maria Jasin. Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. *Genes & development* 15, 3237-3242 (2001).

33. Yang, Yongping, Srividya Swaminathan, Betty K. Martin, and Shyam K. Sharan. Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. *Human molecular genetics* 12, 2121-2131 (2003).

34. Meyer, Peter, Theda Voigtlaender, Claus R. Bartram, and Ruediger Klaes. Twenty-three novel BRCA1 and BRCA2 sequence alterations in breast and/or ovarian cancer families in Southern Germany. *Human mutation* 22, 259-259 (2003).

35. Hayes, Finbarr, Charmagne Cayanan, Daniela Barillà, and Alvaro NA Monteiro. Functional assay for BRCA1: mutagenesis of the COOH-terminal region reveals critical residues for transcription activation. *Cancer research* 60, 2411-2418 (2000).

36. Lovelock, Paul K., Amanda B. Spurdle, Myth TS Mok, Daniel J. Farrugia, Sunil R. Lakhani, Sue Healey, Stephen Arnold *et al*. Identification of BRCA1 missense substitutions that confer partial functional activity: potential moderate risk variants? *Breast Cancer Research* 9, 1 (2007).

37. Thirthagiri, E., S. Y. Lee, P. Kang, D. S. Lee, G. T. Toh, S. Selamat, S. Y. Yoon *et al*. Evaluation of BRCA1 and BRCA2 mutations and risk-prediction models in a typical Asian country (Malaysia) with a relatively low incidence of breast cancer. *Breast Cancer Research* 10, 1 (2008).

38. Williams, R. Scott, Daniel I. Chasman, D. Duong Hau, Benjamin Hui, Albert Y. Lau, and JN Mark Glover. Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations. *Journal of Biological Chemistry* 278, 53007-53016 (2003).

39. Risch, Harvey A., John R. McLaughlin, David EC Cole, Barry Rosen, Linda Bradley, Elaine Kwan, Elaine Jack *et al*. Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. *The American Journal of Human Genetics* 68, 700-710 (2001).

40. Malone, Kathleen E., Janet R. Daling, David R. Doody, Li Hsu, Leslie Bernstein, Ralph J. Coates, Polly A. Marchbanks *et al*. Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years. *Cancer research* 66, 8297-8308 (2006).

41. Rubin, Stephen C., M. Anne Blackwood, Christina Bandera, Kian Behbakht, Ivor Benjamin, Timothy R. Rebbeck, and Jeff Boyd. BRCA1, BRCA2, and hereditary nonpolyposis colorectal cancer gene mutations in an unselected ovarian cancer population: relationship to family history and implications for genetic testing. *American journal of obstetrics and gynecology* 178, 670-677 (1998).

42. Pal, Tuya, Jenny Permuth-Wey, Judith A. Betts, Jeffrey P. Krischer, James Fiorica, Hector Arango, James LaPolla *et al*. BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer* 104, 2807-2816 (2005).

43. Lu, Yi, Weronica E. Ek, David Whiteman, Thomas L. Vaughan, Amanda B. Spurdle, Douglas F. Easton, Paul D. Pharoah *et al*. Most common 'sporadic'cancers have a significant germline genetic component. *Human molecular genetics* 23, 6112–6118 (2014).

44. Lichtenstein, Paul, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England journal of medicine* 343, 78-85 (2000).

45. Kanchi, Krishna L., Kimberly J. Johnson, Charles Lu, Michael D. McLellan, Mark DM Leiserson, Michael C. Wendl, Qunyuan Zhang *et al*. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications* 5 (2014).

46. Rehm, Heidi L., Sherri J. Bale, Pinar Bayrak-Toydemir, Jonathan S. Berg, Kerry K. Brown, Joshua L. Deignan, Michael J. Friez *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine* 15, 733-747 (2013).

47. Xie, Mingchao, Charles Lu, Jiayin Wang, Michael D. McLellan, Kimberly J. Johnson, Michael C. Wendl, Joshua F. McMichael *et al*. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* 20, 1472-1478 (2014).

48. Goetz, Matthew P., Stacey K. Knox, Vera J. Suman, James M. Rae, Stephanie L. Safgren, Matthew M. Ames, Daniel W. Visscher *et al*. The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. *Breast cancer research and treatment* 101, 113-121 (2007).

49. Kiyotani, Kazuma, Taisei Mushiroda, Chiyo K. Imamura, Naoya Hosono, Tatsuhiko Tsunoda, Michiaki Kubo, Yusuke Tanigawara *et al*. Significant effect of polymorphisms in CYP2D6 and ABCC2 on clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients. *Journal of Clinical Oncology* 28, 1287-1293 (2010).

50. Lenz, Georg, R. Eric Davis, Vu N. Ngo, Lloyd Lam, Thaddeus C. George, George W. Wright, Sandeep S. Dave *et al*. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* 319, 1676-1679 (2008).

51. Sokal, R. R. R., F.J. Biometry, 3rd ed. *W H Freeman and Co.* (1995).

52. Fisher, R. A. Statistical Methods for Research Workers. *Oliver and Boyd, London, 7th edition.* (1938).

53. Benjamini, Yoav, and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 289-300 (1995).

54. LaFramboise, Thomas, Ninad Dewal, Katherine Wilkins, Itsik Pe'er, and Matthew L. Freedman. Allelic selection of amplicons in glioblastoma revealed by combining somatic and germline analysis. *PLoS Genetis* 6, e1001086 (2010).

55. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. American Statistical Association* 22, 209-212 (1927).

Acknowledgements

This work was supported by the National Cancer Institute grants R01CA180006 and R01CA178383 to L.D., R01CA141090 to J.D.P. and PO1CA101937 to T.J.L., the National Human Genome Research Institute grants U01HG006517 to L.D., R01HG007069 to B.J.R. and U54HG003079 to R.K.W., the National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK087960 and DoD grant PC130118 to F.C., NHGRI Genome Analysis Training Program (T32 HG000045) to M.X., Ministry of Education in Taiwan Fellowship to K.-L.H. and CMB training grant (GM 007067) to R.J. We want to thank the Analysis Pipeline group at the McDonnell Genome Institute for developing the automated sequence analysis pipelines. We also acknowledge The Cancer Genome Atlas (cancergenome.nih.gov) as the source of primary data and members of the TCGA Research Network for helpful discussions.

Contributions

This work was finished by a team in the lab. I was the major contributor to this project. With the other lab members' help, I developed the analysis pipeline and performed statistical analysis, including variant detection and filtering, variant association study, the interaction of somatic mutation and germline variant. Besides, I worked with Jie Ning, the lab technician, designed the experiment and analyzed the functional validation data.

Table 2.1: Case numbers from individual cancer types and basic clinical features for cancer cases included in this study.

Cancer Type	Cohort	Samples	Age (Mean ± S.D.)	Gender		Ethnicity					Vital Status		
				M (%)	F (%)	Other	Asian	African America	Caucasians	NA	Alive	Deceased	NA
BRCA	Discovery	770	58.2 ± 13.2	1.0	99.0	1	50	53	578	88	682	88	0
	Validation	217	59.5 ± 12.8	0.5	99.5	0	6	64	124	6	189	11	0
GBM	Discovery	267	59.6 ± 14.0	60.9	39.1	0	4	21	237	5	89	176	1
	Validation	124	61.0 ± 12.4	62.7	37.3	0	3	3	107	5	36	81	1
HNSC	Discovery	291	60.9 ± 12.4	71.3	28.7	1	4	26	223	7	153	108	0
	Validation	222	60.3 ± 11.3	73.4	26.6	0	5	9	174	4	155	37	0
KIRC	Discovery	452	60.7 ± 12.1	64.8	35.2	0	7	20	419	6	306	146	0
	Validation	42	58.5 ± 12.2	76.9	23.1	0	1	7	31	0	27	12	0
AML	Discovery	200	55.0 ± 16.1	54.5	45.5	0	2	15	181	2	67	133	0
LGG	Discovery	223	43.0 ± 13.5	57.7	42.3	0	0	9	209	2	169	51	0
	Validation	240	43.5 ± 13.5	51.9	48.1	1	4	6	167	3	167	14	0
LUAD	Discovery	462	65.2 ± 9.9	46.3	53.7	1	5	23	297	61	294	93	0
	Validation	94	66.7 ± 9.5	46.1	53.9	0	1	3	67	5	51	25	0
LUSC	Discovery	193	67.7 ± 9.3	74.8	25.2	0	0	4	113	22	93	46	0
	Validation	183	66.1 ± 8.4	78.4	21.6	0	6	3	79	51	103	36	0
OV	Discovery	429	59.4 ±11.8	0	100	2	15	19	370	23	207	218	4
	Validation	68	61.2 ± 10.3	0	100	1	3	2	58	4	40	27	1
PRAD	Discovery	178	60.4 ± 6.9	100	0	0	2	6	130	10	147	1	0
	Validation	157	60.4 ± 6.7	100	0	0	0	1	9	118	127	1	0
STAD	Discovery	321	66.0 ± 10.7	61.1	38.9	0	81	4	175	46	281	25	0
UCEC	Discovery	248	63.1 ± 11.1	0	100	10	13	25	193	7	231	17	0
	Validation	280	64.7 ± 11.3	0	100	3	6	52	167	19	221	26	0
TOTAL	Discovery	4034	59.9 ± 13.2	39.3	60.7	15	183	225	3125	279	2719	1102	5
	Validation	1627	59.7 ± 13.1	44.1	55.9	5	35	150	983	215	1116	270	2

*Other indicates American Indian, Alaska Native, Hawaiian, Pacific Islander

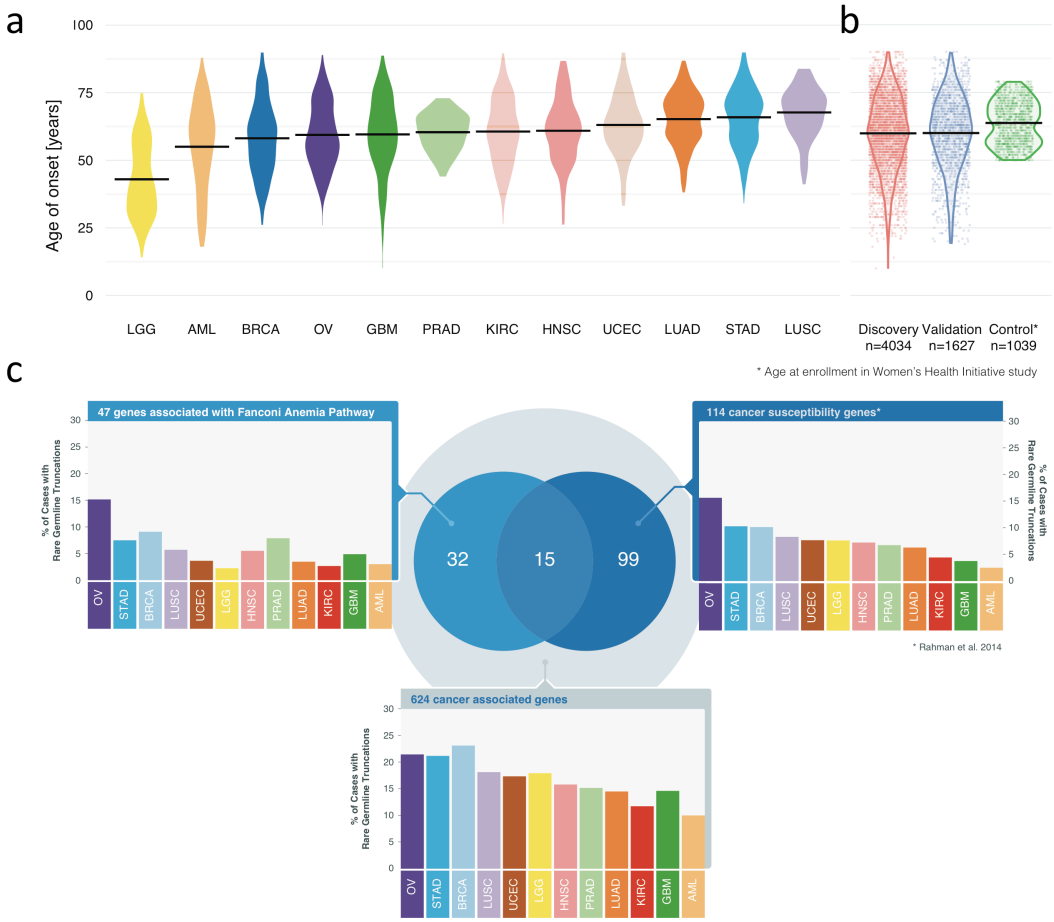


Figure 2.1: Characteristics of the data. Data are distributed by age, cancer, cohort, and carrier frequency. **(a)** Age of onset by cancer type. Age varies on average across cancer types, from 43 yr in LGG to 67.7 yr in LUSC. Note that LGG, LUAD, and STAD show clear bimodal characteristics. **(b)** Age distributions for discovery, validation, and control cohorts. **(c)** Comparison of cancer gene truncation carrier frequencies across 12 cancer types. The distribution of rare germline truncation variants for 12 cancer types (represented as the percent of cases in each cancer type with rare germline truncation mutation) in 2 different groups of cancer associated genes (labeled on top of each bar plot): 114 cancer susceptibility genes from Rahman et al [1] and 47 genes associated with the DNA repair (Fanconi Anemia) pathway. There are 15 genes common to both groups. The total number of unique genes from these 2 groups is 131.

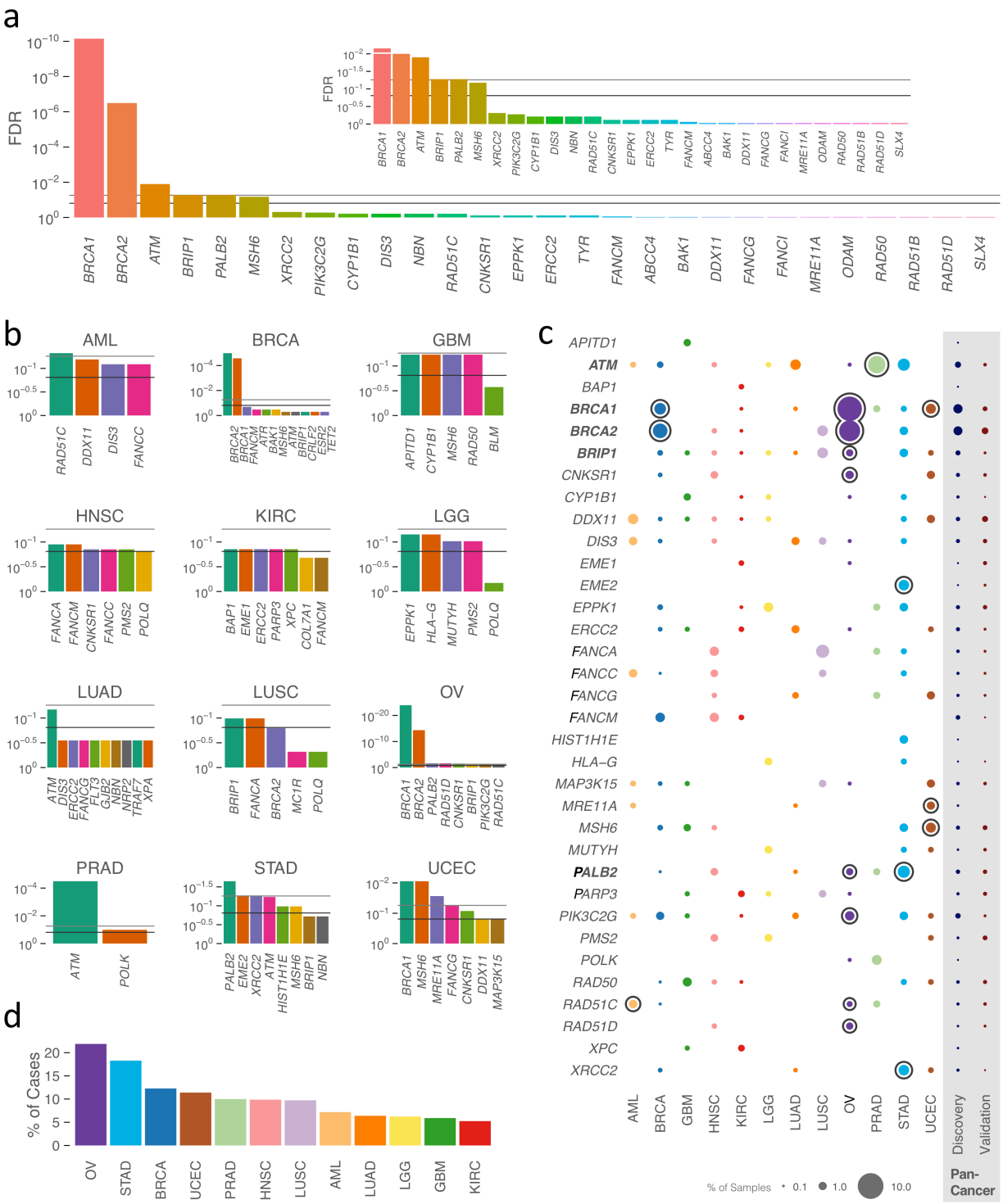


Figure 2.2: Burden analysis reveals distinct set of cancer susceptibility genes across 12 cancer types. A total of 34 genes of interest were identified by burden analysis by comparing the frequencies of rare truncation variants in Caucasian cancer cases (n = 3,125) vs. their frequencies in the WHI control population (n = 1,039). Two oncogenes (*ABL2* and BCR) were omitted. **(a)** Significant genes across Pan-cancer types. Data were analyzed with the Total Frequency Test (TFT) followed by False Discovery Rate (FDR) ranking. Dark horizontal line indicates the 5%

FDR threshold, which is satisfied by 5 genes, including *BRCA1, BRCA2, ATM, BRIP1,* and *PALB2*. Inset shows closer visual resolution. **(b)** Significant genes for specific cancer types. Each plot shows the top tested genes, by FDR, from the same TFT analysis procedure for all 12 individual cancer types. 8 genes in addition to the 5 shown in (a) are significant at the 5% FDR level from cancer type specific analysis. **(c)** Cohort frequencies of genes. Bubble plot shows frequency of rare truncation mutation as a percentage of cases in each cohort (all 4,034 cases included for frequency calculation). The x-axis denotes the test group of a specific cancer type, the Pan-Cancer discovery cohort (4,034), and the validation cohort (1,627). Genes found to be significant at 5% FDR using the Pan-Cancer discovery cohort are labeled in boldface. Grey rings indicate genes that are significant (TFT, FDR ≤ 5%) for a particular cohort on the x-axis. **(d)** Percentage of cases carrying rare truncation in the 34 genes of interest across 12 cancer types in the discovery cohort.

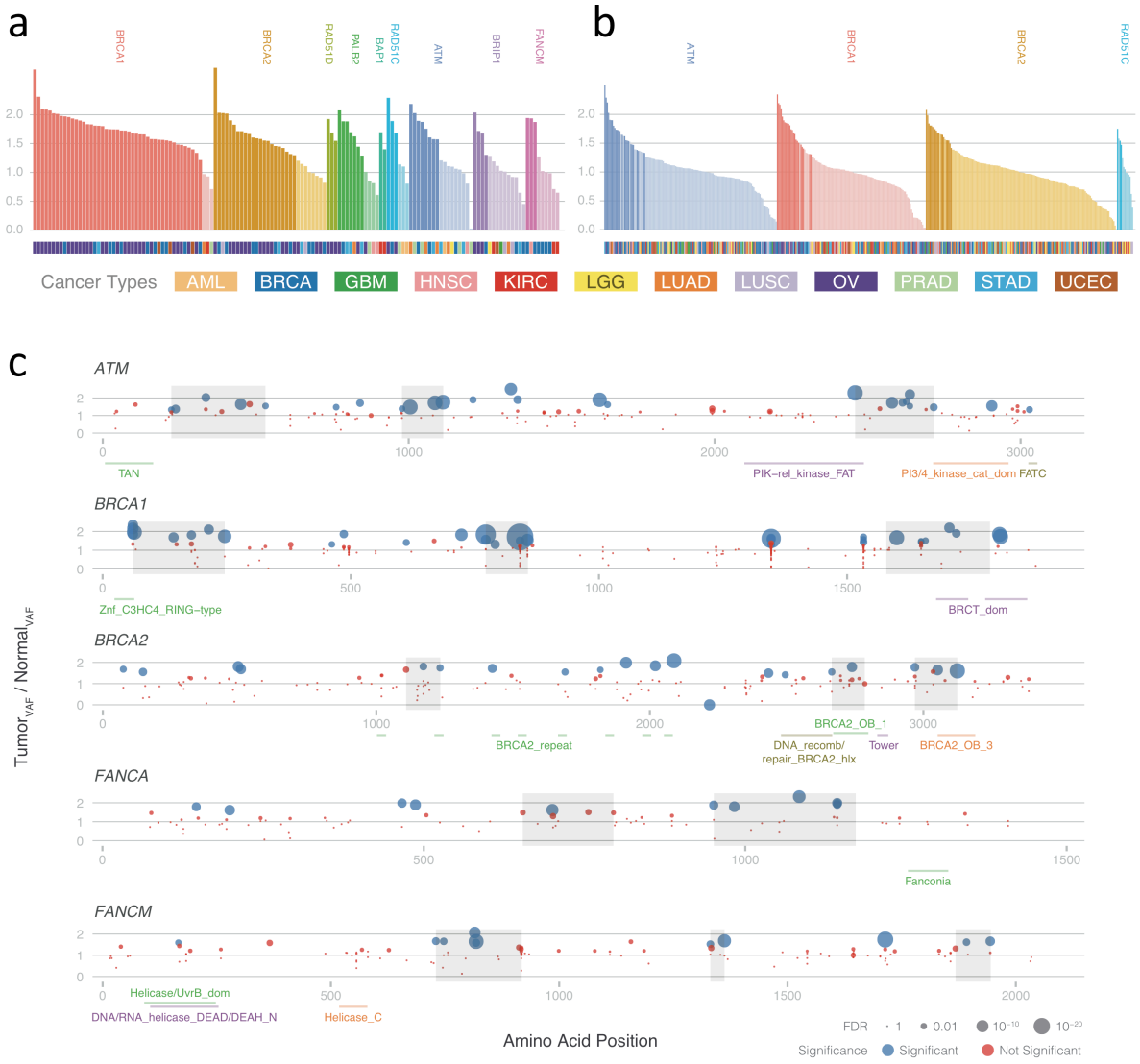


Figure 2.3: Analysis of loss of heterozygosity in rare truncation and missense variants. (a) Bar plot shows individual truncations from 9 genes (FDR shown) with lengths representing ratios of tumor to normal variant allele fractions (i.e. the fraction of reads containing the variant allele). Statistically significant events, defined as FDR ≤ 5%, are shaded boldly, while non-significant events are muted, with colors corresponding to genes. Cancer source of each truncation is shown underneath, for example most *BRCA1* variants occur in ovarian and breast cancers and all *BAP1* variants in KIRC. **(b)** Bar plot for individual missense variants from 4 genes having elevated frequencies of such variants that show very significant LOH, i.e. at the 1% FDR level. **(c)** Dot plot shows individual missense variants where abscissa and ordinate are amino acid position and ratio of tumor to normal variant allele fraction. Blue and red indicate significant (FDR ≤ 5%) and non-significant events, respectively, with size of blue dots proportional to negative log of the FDR. Annotated domains from the PFAM database are aligned with position, while shaded areas indicate "hotspot" regions where variants having significant LOH cluster more than the rate explainable by chance. Plots are shown for *ATM*, *BRCA1*, *BRCA2*, *FANCA*, and *FANCM*.

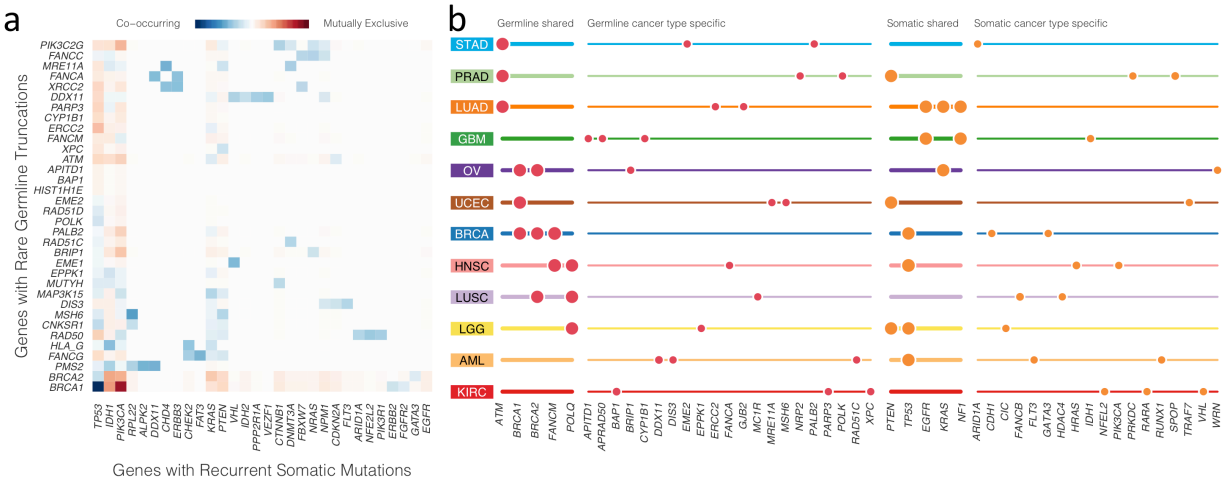


Figure 2.4: Molecular interactions between rare germline variants and somatic mutations within and across cancer types. (a) Heatmap demonstrates the significance of interactions between 34 burden test significant genes and 54 cancer-associated genes (only 30 were shown) with recurrently mutated somatic variants across cancer types. Red-white color scale and blue-white color scale depict the negative log of P-value for mutual exclusivity and co-occurrence, respectively. Both are based on the MuSiC permutation test (n=10,000). **(b)** Abacus plot displays the distribution of significant, mutually-exclusive rare germline variants and somatic mutations across all 12 cancer types. Unique combinations of germline and somatic variants contribute to the development of individual cancer types. Bigger dots indicate recurrent genes across cancer types, while smaller dots indicate cancer type specific genes.

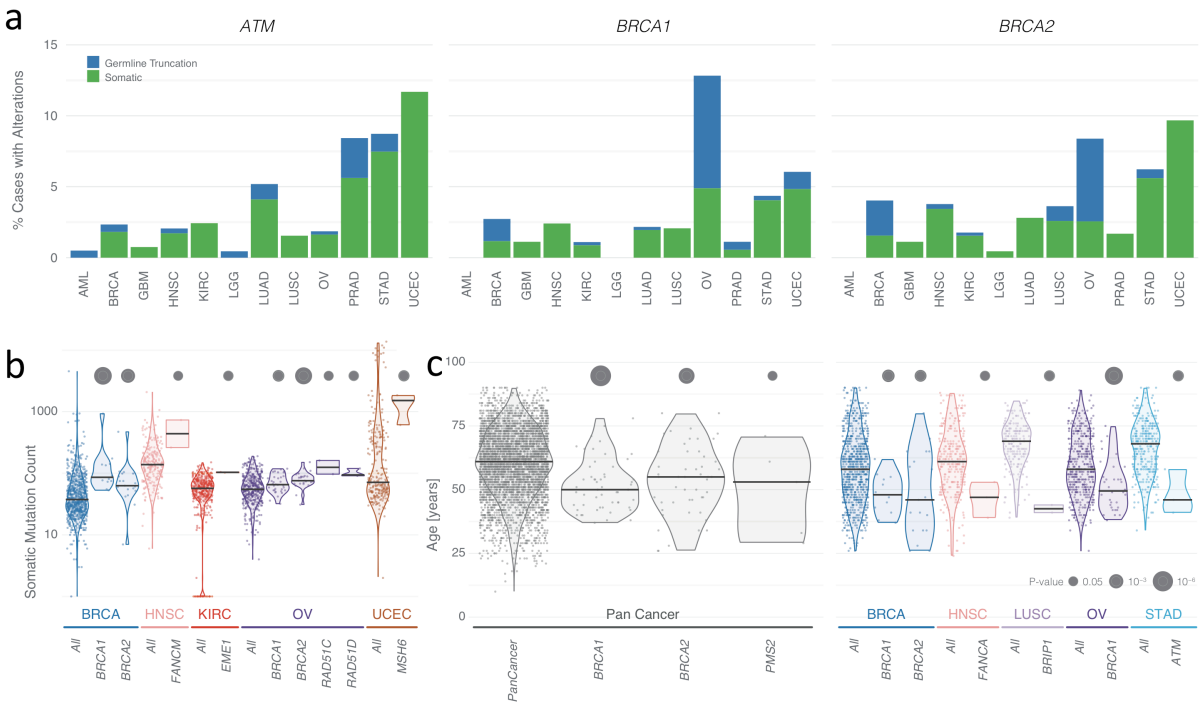


Figure 2.5: Germline variants correlate with somatic mutations and age at diagnosis. Panel **(a)** illustrates the distribution of *BRCA1*, *BRCA2*, and *ATM* somatic and germline mutations across cancer types. Panels **(b)** and **(c)** display genes significantly correlated with somatic mutation frequency and younger age of onset in different cancer types and in Pan-Cancer. The width of the shape indicates the density, and the horizontal line indicates the median. P-value is calculated by permutation test and scales with size of the dots.

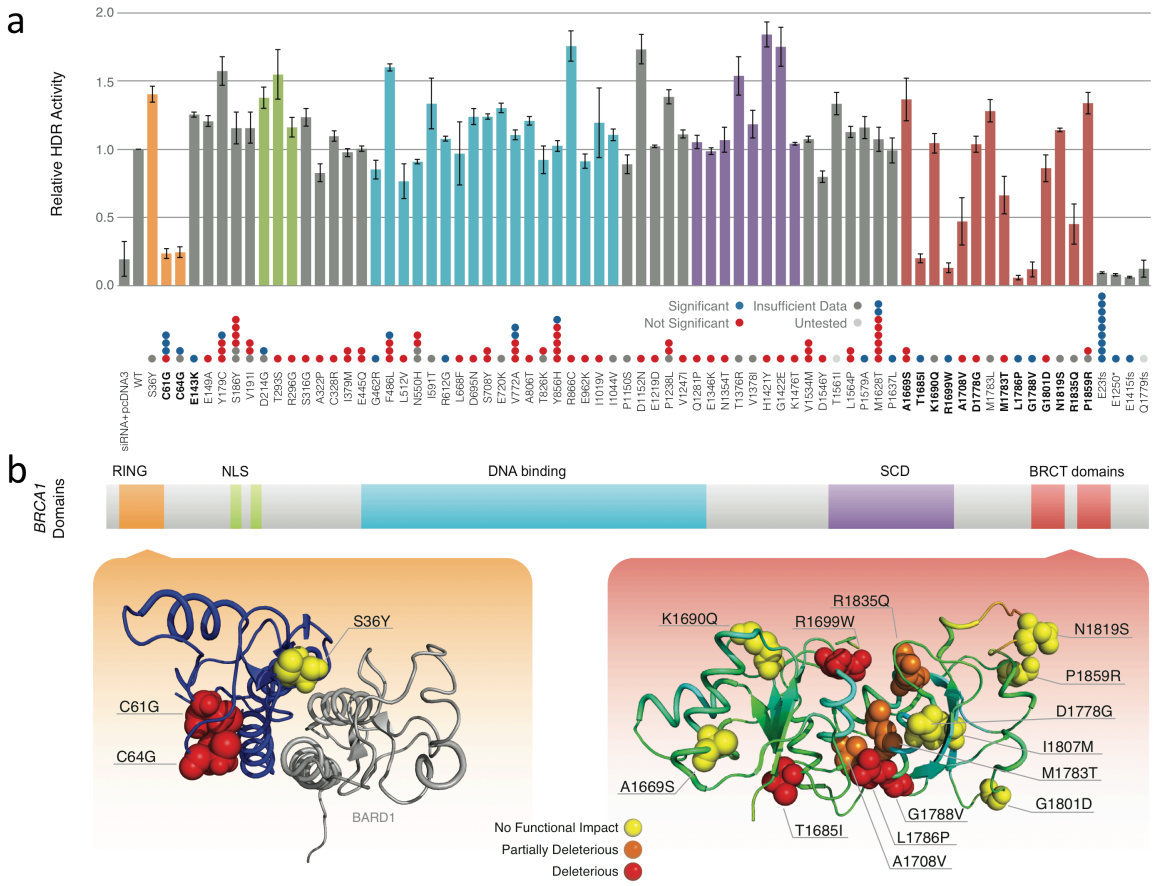


Figure 2.6: Functional validation of *BRCA1* missense and truncation variants. (a) 68 rare missense and 4 truncation variant sites were tested by HDR assay. All samples were depleted of endogenous BRCA1 by transfection of a siRNA targeting the 3'-UTR. Indicated in the legend are the plasmids transfected to test for rescue of BRCA1 activity. "pcDNA3" is empty vector, and "WT" represents wild-type BRCA1 plasmid. The y-axis denotes the relative HDR activity to the wild-type BRCA1 protein. Error bars depict standard deviations from the mean. Dots on the x-axis represent LOH statuses, each dot corresponding to one case. Blue, red, dark gray, and light gray denote statistical significance, non-significance, unknown LOH (due to lack of sufficient coverage), and untested, respectively. Variants in different functional domains are tagged with different colors as follows: orange=RING domain; green=nuclear localization signal (NLS); blue=DNA binding region; purple=a SQ/TQ cluster domain (SCD), and red=*BRCA1* C-Terminal domain (BRCT). All the HDR assays were tested in triplicate. **(b)** Crystal structure of the BRCA1 RING (left) domain in complex with the BARD1 RING domain (labeled in gray) and BRCT domain (right panel) are displayed, with HDR-defective variants labeled in red and partial HDR-defective variants tagged in orange. Variants in yellow are functional in the HDR assay.

Chapter 3: Age-related Cancer Mutations Associated with Clonal Hematopoietic Expansion[‡]

3.1 Abstract

Several genetic alterations characteristic of leukemia and lymphoma have been detected in the blood of individuals without apparent hematological malignancies. TCGA provides a unique resource for comprehensive discovery of mutations and genes in blood that may contribute to the clonal expansion of hematopoietic stem/progenitor cells. Here, we analyzed blood-derived sequence data from 2,728 individuals from TCGA and discovered 77 blood-specific mutations in cancer-associated genes, the majority being associated with advanced age. Remarkably, 83% of these mutations were from 19 leukemia/lymphoma-associated genes, and nine were recurrently mutated (*DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *TP53*, *GNAS*, *PPM1D*, *BCORL1* and *SF3B1).* We identified 14 additional mutations in a very small fraction of blood cells, possibly representing the earliest stages of clonal expansion in hematopoietic stem cells. Comparison of these findings to mutations in hematological malignancies identified several recurrently mutated genes that may be disease initiators. Our analyses show that the blood cells of more than 2% of individuals (5–6% of people older than 70 years) contain mutations that may represent premalignant events that cause clonal hematopoietic expansion.

3.2 Introduction

Blood cells are continuously regenerated by hematopoietic stem/progenitor cells (HSPCs). Human HSPCs divide only rarely (estimated at once a month), but have self-renewal properties

[‡] Mingchao Xie[*], Charles Lu[*], Jiaying Wang[*], *et al*. "Age-related mutations associated with clonal hematopoietic expansion and malignancies." *Nature medicine* 20, 1472-1478 (2014) (* contributed equally to this work)

that sustain survival for decades. As HSPCs divide, they accumulate rare, random mutations that generally do not affect function [1]. However, some mutations confer advantages in self-renewal and/or proliferation, resulting in clonal expansion of the affected cells. Although these "initiating" mutations do not lead directly to disease, they can cooperate with subsequent mutations to cause hematopoietic malignancies. For example, *BCR-ABL* and *BCL2* translocations have been found in blood cells of individuals without overt hematological malignancies [2-4]. The frequency of such events appears to increase with age, with a similar trend being found for somatic structural changes in the nuclear genomes of blood cells [1, 5]. Single nucleotide polymorphism array analysis from large genome-wide association study cohorts showed ~2–3% of normal individuals of advanced age (70s and 80s) harbor leukemia-associated copy number changes that include genes such as *DNMT3A* (encoding DNA (cytosine-5-)-methyltransferase 3α) and *TET2* (encoding tet methylcytosine dioxygenase 2) [6, 7]. More recently, somatic recurrent *TET2* mutations were detected in the blood of elderly women without overt hematological malignancies [8], and *DNMT3A* mutations were reported in nonleukemic cells [9].

These findings have collectively led to the hypothesis that certain genetic mutations may confer advantages to affected HSPCs, resulting in enhanced cell renewal, clonal expansion or both. However, it is unclear whether the effect involves only a small number of genes or many more genes related to leukemia and lymphoma and whether their participation in promoting clonal expansion necessarily leads to clones resembling cancer cells. Here, we address the former question by analyzing variations in 2,728 blood samples in TCGA. We observed many individuals with age-related hematopoietic clonal mosaicism and concurrent presence of over 60 mutations in 19 leukemia- and/or lymphoma-associated genes. Our study identified not only genes but also specific mutations, associated with the clonal expansion process. Additional

statistical analysis identified low-level (2–10% VAFs) recurrent leukemic mutations in a substantial number of cases, possibly in the early stages of clonal expansion. Moreover, our analysis suggests that *DNMT3A*, *TET2*, *JAK2* (encoding Janus kinase 2), *ASXL1* (encoding additional sex combs–like transcriptional regulator 1), *SF3B1* (encoding splicing factor 3b, subunit 1) and *TP53* (encoding tumor protein p53) have distinct and overlapping roles in the development of myeloproliferative neoplasm (MPN), myelodysplastic syndrome (MDS), chronic lymphocytic leukemia (CLL) and/or acute myelogenous leukemia (AML). Finally, these results also incidentally highlight the need for caution when using blood as a reference for a surrogate 'germline' genome, especially in older individuals.

3.3 Results

3.3.1 Cancer types and sample characteristics

We searched for variants present in the blood normal controls across 2,728 cancer patients (Supplementary Table 3.1) from 11 diverse cancer types: breast adenocarcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), brain low grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), and uterine corpus endometrioid carcinoma (UCEC). The numbers of cases from each tumor type range from 57 (KIRC) to 673 (BRCA) and are listed in Supplementary Table 3.2. Patients were diagnosed between 10–90 years (mean 59.5 ± 13.1 years) and 22.1% were deceased at the time of TCGA sample procurement (Supplementary Table 3.2). TCGA collects clinical data regarding diagnosis and prior treatment of neoplasms during the sample submission process. To ensure that our dataset was comprised only of individuals having first-time primary cancers and having had no treatment with radiation

and/or chemotherapy, we excluded those having reported histories of these events as identified at https://tcga-data.nci.nih.gov/annotations/ and all clinical data (July 30th, 2014). However, five patients with synchronous tumors not associated with blood were included, since these synchronous tumors would be unlikely to affect variant analysis in corresponding blood samples.

3.3.2 Variant calling and filtering strategies

Variants in the 2,728 blood normal controls were identified with VarScan (single nucleotide variant and indel), GATK (single nucleotide variant and indel), and Pindel (indel) (see Methods). False-positive filters were subsequently applied prior to downstream analysis and interpretation (see Methods). Out of the 49,317,027 variants (previously reported OV counts [10] were not included here) that passed false positive filters, 1,622,485 with minor allele frequency of <1% in the 1000 Genomes reference and in each cancer cohort were retained for further analysis; this consists of 1,025,632 missense, 529,505 synonymous, 19,663 nonsense, 10,976 splice site, 926 nonstop/readthrough, 20,275 frameshift indels, and 15,508 in frame indels (Supplementary Table 3.3). We used a stringent filtering strategy described previously [11] for standardizing specificity across the Pan-Cancer somatic variant calls for available matched tumor samples (Supplementary Table 3.4).

3.3.3 Variants contributing to hematopoietic clonal expansion

The collection of both tumor and matched blood normal exome data by TCGA provides a unique comparative resource for identifying those somatic variants in blood that contribute to clonal expansion. We set out to identify both rare truncation variants (RTV), that is, those having <1% MAF in both the 1000 Genomes collection and the cohort data, and variants overlapping with recurrent somatic mutations (also called Known Hotspot Variants (KHV)) found in the analysis of 12 TCGA cancer types (see Methods and Supplementary Table 3.5 and 3.6). Subsequently,

1,598 RTVs in 556 cancer-associated genes based on several recent studies [11-15] (see Methods and Supplementary Table 3.7) were manually reviewed to remove false positive calls and 136 KHVs in the same set of cancer genes were identified. The resulting list was further filtered to remove polymorphisms present in 1000 Genomes (see Methods) and having greater than 0.1% MAF as reported in the current Exome Variant Server data release (ESP6500SI-V2, http://evs.gs.washington.edu/EVS/) from 6,503 samples drawn from multiple NHLBI Exome Sequencing Project (ESP) cohorts. We focused on those mutations found in blood normal samples, but not present or present only at very low levels in either the tumor samples or tumor adjacent normal samples, as this pattern is highly suggestive of somatic mutation in HSPCs introduced by the clonal expansion process. Inflammatory lymphocytes/macrophages/neutrophils will infiltrate different tumors to different extents. Therefore, the hematopoietic mutations do not have to be completely absent in the tumor sample.

Our analysis of RTVs and KHVs in 556 selected cancer-associated genes identified 70 blood-specific mutations in 58 individuals. We further performed comparative analysis of blood versus tumor samples for these 58 individuals, with the goal of detecting all blood-specific nonsynonymous mutations in the 556 cancer-associated genes; this analysis identified seven additional events (those that were likely loss of heterozygosity related to copy number alterations in the tumor were not included), yielding a final list of 77 blood-specific mutations in 58 cases (Table 3.1 and Supplementary Table 3.8). For five of those 58 cases that also had adjacent normal tissue analyzed, blood variants were absent in the adjacent normal tissue. Interestingly, among the 31 genes harboring these events, 19 have already been linked to hematological malignancies (Figure 3.1a). More strikingly, 64 out of the total 77 events (83%) were in these 19 genes, examples as follows: *DNMT3A* [16] (18 cases), *TET2* [17] (9 cases), *JAK2* [18] (8 cases),

ASXL1 [19, 20] (6 cases), *TP53* [21] (4 cases), *SF3B1* [22] (2 cases), *BCORL1* [23] (2 case), *ASXL2* [24] (1 case), and *SH2B3* [25] (1 case) (Table 3.1 and Figure 3.1a). The overall frequency of blood-specific mutations increased with age (logistic regression analysis, $P = 2.38e{-}08$), for example, 0.9% of the cases were in their 40s, 1.0% in their 50s, 1.8% in their 60s, 5.3% in their 70s, and 6.1% in their 80s (Figure 3.1b). The blood-specific mutations were found in all 11 cancer types (Figure 3.1b). Frequencies of individual *TET2*, *DNMT3A*, *ASXL1*, and *SF3B1* mutations also show association with age, (all FDR values <0.034, logistic regression). Interestingly, *TET2, ASXL1,* and *SF3B1* mutations were predominantly found in the oldest age groups (70s and 80s) and *DNMT3A* mutations in 60s to 80s. In contrast, *JAK2* mutations, which did not achieve significance, trended in both younger (40s and 50s) and older age groups (70s) (Figure 3.1c).

The average age of the 54 individuals with blood-specific mutations in the 19 leukemia or lymphoma-associated genes was 70.0 ± 9.9 years, significantly higher than that of the larger TCGA cohort used in this study (difference of means test, $P = 3.4e{-}10$) (Figure 3.2a). Notably, the six individuals having two blood-specific mutations in the nine recurrently mutated genes are relatively older, with ages of 64 (*DNMT3A*: R882H; 36% VAF; *TET2*: H863fs; 12% VAF), 72 (*JAK2*: V617F; 73% VAF and *TET2*: T229fs; 19% VAF), 75 (*DNMT3A*: Y584fs; 38% VAF and *TET2*: Q764fs; 33% VAF), 76 (*JAK2*: V617F; 42% VAF and *ASXL1*: R548fs; 35% VAF), 83 years (*TET2*: F381fs; 50% VAF and *TET2*: Q888*; 20% VAF), and unknown age (*BCORL1*: G883E; 17% VAF and *TP53*: Q136*; 18% VAF), respectively (Table 3.1 and Supplementary Table 3.8). We also compared the distribution of variant allele fractions for these 64 events versus inherited sites identified in the same sample set; we observed a clear shift towards lower

VAFs in the blood-specific sites (Figure 3.2b), suggesting the majority are present in only a fraction of the blood cells.

Although *GNAS* mutations have been found in leukemia [26], activating gain-of-function mutations in *GNAS* are best known for their involvement in polyostotic fibrous dysplasia and McCune-Albright syndrome [27]. Interestingly, previous studies showed that activating mosaic *GNAS* mutations could affect various tissue types and the non-mosaic state for activating *GNAS* mutations may be lethal for the embryo [28-30]. This is consistent with our finding of mosaic *GNAS* R202H in transcript ENST00000354359 (also known as R201H in transcript ENSP00000360126) in the three blood samples of TCGA cases (11.5%, 14.4%, 21.4% VAFs, respectively). It is also worth noting that two blood-specific truncation mutations were detected in *PPM1D*, a gene recently found to be associated with breast and ovarian predisposition with mosaic signatures [31], but not with hematological malignancies.

Due to the fact that the blood-specific variants were present in only a fraction of the blood cells, we postulated that certain low level variants associated with clonal expansion were not captured by the variant detection tools. Therefore, we performed read count-based analysis for a set of known hotspot variants, including R882 in *DNMT3A*, R132 in *IDH1*, R172/R140 in *IDH2*, V617 in *JAK2*, K700 in *SF3B1*, and S34 in *U2AF1* for the entire TCGA cohort. We compared the distributions of VAFs between tumor and blood normal samples and found that the strongest differences were at R882 in *DNMT3A* and V617 in *JAK2* (Figure 3.3 and Table 3.1). We devised a statistical procedure (see Methods) to identify additional "low VAF" sites significantly above the background error rate. Our analysis identified 14 blood samples having low-level hot spot variants (12 of them are not part of the 58 cases identified), including eight in *DNMT3A*, four in *JAK2*, one in both *SF3B1* and *IDH2*, but none in *IDH1* or *U2AF1*. The average age for these 14

cases is 64.0 ± 14.9 years, again higher than the entire TCGA cohort studied (Supplementary Table 3.9). We performed deep sequencing for five selected low-level hot spot variants (two R882C and one R882H sample sites in *DNMT3A*, one V617F in *JAK2*, and one K700E in *SF3B1*) to evaluate our detection approach. We achieved more than 450,000x average coverage for each site and all of the five variants were validated as bona fide low-level blood specific mutations (Supplementary Table 3.10 and see Methods). By including low VAF events, the overall frequencies of blood-specific mutations in different age groups are: 1.2%, 1.3%, 2.2%, 6.1%, and 6.8% in their 40s, 50s, 60s, 70s, and 80s, respectively (Supplementary Figure 3.1). Collectively, we estimate that approximately 2% and 3.5% of individuals over the ages of 40 and 60, respectively, and without overt hematological malignancies carry blood-specific mutations that are associated with hematological malignancies.

3.3.4 Known Hotspot Variants in NHLBI exome sequencing project control cohort

To confirm that these mutations are also present in an independent set of normal blood samples, we examined the NHLBI exome sequencing project (ESP) control cohort. We searched for RTVs and KHVs in 6,503 ESP samples, focusing on four AML-associated genes discovered in the TCGA set (*DNMT3A*, *TET2*, *JAK2*, and *ASXL1*). When we applied a 0.1% MAF threshold in ESP (to focus on rare variants and to prevent potential false positives), we identified 8 RTVs and 13 KHVs in *DNMT3A*, 13 RTVs in *TET2,* and 7 RTVs in *ASXL1*, and 3 KHVs in *JAK2* (Supplementary Table 3.11). For a subset of ESP samples (n = 557 WHISP), we re-aligned reads and performed variant analysis using the same process applied for the TCGA cohort. This allowed us to confirm RTVs and KHVs detected by the pipeline, and also to detect low VAF KHVs potentially missed by variant detection tools. Our pipeline detected one RTV each in *DNMT3A*, *TET2,* and *ASXL1*, four *DNMT3A* R882H mutations, and two *JAK2* V617F mutations.

Careful analysis of recurrent hotspot sites (described in the previous section) using the 557 WHISP samples identified an additional three *DNMT3A* R882, one *JAK2* V617, one *GNAS* R202 and two *SF3B1* K700 variants with VAFs ranging from 2% to 10% (Figure 3.3 and Supplementary Table 3.12). Based on these analyses, over 2% of the 557 WHISP cases contain mutations in these five selected genes. However, sequencing of matched non-blood samples from these cases would be required to prove that they are truly blood-specific mutations.

3.3.5 Mutations in TCGA blood samples and patients with hematological malignancies

We next compared the mutation frequencies in 25 genes frequently mutated in at least one of the five following cohorts: TCGA 58 blood samples, 151 myeloproliferative neoplasm (MPN) cases reported by Nangalia *et al.* [32], 150 myelodysplastic syndrome (MDS) cases reported by Walter *et al.* [33], 160 chronic lymphocytic leukemia (CLL) by Landau *et al.* [34], and 200 TCGA AML cases [21]. *DNMT3A*, *TET2*, *ASXL1, TP53,* and *SF3B1* were found to be consistently mutated in at least four groups, while *JAK2* was more specifically mutated in 58 TCGA blood samples and MPN patients (Figure 3.4a). No mutations were found in TCGA blood samples in genes such as *IDH1*, *NRAS*, *RUNX1*, and *PHF6*, which are significantly mutated in AML and frequently mutated in MPN and/or MDS. Several genes showed cohort specificity: *GNAS* was mutated only in TCGA blood samples and *CEBPA*, *WT1*, *PTPN11*, *KIT*, *SMC1A*, and *SMC3* were preferably mutated in the AML cohort. We reasoned that common mutations among the cohorts (e.g., in *DNMT3A*, *TET2*, *ASXL1*, *SF3B1*, *JAK2*, and *TP53*) may be relevant for initiating HSPC clonal expansion, and are also likely to be early, initiation events for hematological malignancies, such as MPN, MDS, CLL, and AML. On the other hand, genes specific for MPN, MDS, and/or AML (e.g., *NRAS*, *RUNX1, NPM1,* and *FLT3*) are more likely to be subsequent, cooperating mutations that are involved in the progression of these diseases. These observations also show both distinct

and common connections among these five cohort groups, suggesting that the TCGA cohort consists of a combination of precursor mutations that may sometimes evolve to MPN, MDS, CLL, and/or AML, although subsequent, collaborating mutations are clearly required (Supplementary Figure 3.2). Finally, we compared the average number of mutations among these 4 cohorts (not including MDS) in 556 selected cancer-associated genes; this showed TCGA blood samples having the fewest mutations, MPN and CLL cohorts having intermediate numbers of mutations, and AML patients harboring the highest number of mutations (Figure 3.4b).

3.4 Discussion

We identified age-related hematopoietic clonal expansion and the concurrent presence of leukemia/lymphoma-associated mutations in about 2% of individuals studied, who did not have reported hematological malignancies; this frequency reaches 5–6% for individuals who are at least 70 years of age. Our investigations of 2,728 TCGA blood samples identified 64 mutations in 19 genes known for their roles in hematological malignancies with VAFs above 10%, and an additional 14 mutations with lower VAFs (2 to 10%) when site-specific analysis was conducted. While many of these genes (e.g., *DNMT3A*, *JAK2*, *ASXL1*, and *TET2*) are established drivers for hematological malignancies, others (e.g., *PPM1D)* have not yet been implicated. The wide range of VAFs is indicative of the different stages of clonal expansion among individuals. Our finding also supports the hypothesis that mutations in genes such as *DNMT3A*, *JAK2*, *ASXL1*, *TET2, GNAS,* and others are likely to be initiating events for MPN, MDS, CLL, and/or AML, while epigenetic changes have also been previously implicated [35, 36]. Importantly, the lack of detectable mutations in *IDH1*, *RUNX1*, *NRAS*, *NPM1*, and *FLT3* in both TCGA blood samples and WHISP cases supports the idea that these mutations are usually cooperating mutations that are important for disease progression.

These data suggest that extra care is required when using blood as a surrogate reference for the germline genome, especially in elderly individuals. First, there is an obvious risk of blood-specific variants in individuals without overt hematological malignancy being mistaken as germline variants. Secondly, germline alleles in cancer samples could be mistaken as tumor-specific variants when comparing to blood samples. Lastly, connections were made between mosaic *PPM1D* mutations in lymphocytes and the predisposition to breast and ovarian cancers [31]. While the influence of the immune system on tumorigenesis is well known, it is also possible that the blood-specific mutations are independent, unrelated events that are simply associated with the clonal expansion of HSPCs.

Our unbiased mutational analysis using sequencing data of relatively high depth (average of 107.5x coverage, Supplementary Table 3.13) allowed us to detect hot-spot mutations down to 2–3% VAFs; we discovered that 5–6% cases with advanced age (over 70 years) carry blood-specific mutations known to be involved in hematological malignancies. Some of these individuals may be undergoing hematopoietic clonal expansion (Figure 3.5), but most probably do not progress to overt disease, since the incidence of myeloid malignancies in the elderly is less than 0.1% [37]. Participants providing specimens to TCGA are de-identified/coded, so it is not feasible to determine whether a participant with a leukemia-associated mutation actually progressed to a malignant hematologic disease (Supplementary Figure 3.2). Using SNP array data from GENEVA study cohorts (melanoma, lung health, prostate cancer etc.), Laurie *et al.* showed roughly 2–3% of elderly individuals (over 70 years) have chromosomal anomalies in blood samples [7]. We therefore suggest that our estimate of frequency may be conservative, since other types of alterations (such as gene fusions and copy number alterations) were not included in our study. Regardless, these data clearly show that the elderly often acquire clonal

"skewing" in their hematopoietic compartments and that this may represent a contributing factor to the development of hematologic malignancies.

3.5 Methods

We analyzed the peripheral blood sequence data from 2,728 individuals having had first-time primary cancer and no radiation or chemotherapy treatment. Exome data were aligned to human reference build 37 using BWA [38] and variants were identified using VarScan [39], GATK [40], and Pindel [41], with stringent downstream filtering to standardize specificity. Variant annotation was based on Ensembl release 70_37_v5. The list of 556 cancer-associated genes was compiled from publicly available screening panels, published studies, and preliminary analysis of publicly available data sources [11-15, 42]. Read count analysis was performed with our bam-readcount tool, available at https://github.com/genome/bam-readcount. Low level blood-specific events were discovered using a two-stage process including pre-filtering candidate non-mosaic samples using Bayes' Rule and then the detection of high probability mosaic sites using Fisher's exact test.

3.5.1 Variant calling and annotation strategies

Exome sequencing data were aligned to NCBI Build 37 of the human reference using BWA v0.5.9 and de-duplicated using Picard 1.29. Single nucleotide variants were identified by Varscan (version 2.2.6: –min-var-freq 0.1 –p-value 0.1 –min-coverage 8 –map-quality 10), and GATK (revision 5336: –T UnifiedGenotyper –R GRCh37-lite –et NO_ET –l INFO –U ALL –validation_strictness SILENT). Indels were identified using Varscan (version 2.2.9: –min-coverage 3 –min-var-freq 0.2 –p-value 0.1 –strand-filter 1 –map-quality 10), GATK (revision5336: –T IndelGenotyperV2 –R GRCh37-lite –window_size 300 –et NO_ET –U ALL), as well as Pindel (version 0.2.4x, May 8, 2013: –window-size 1). For Pindel analysis, we preset

the insertion size to 500 if this information is not provided in the BAM header. SNVs were based on the union of GATK and VarScan. They were subsequently processed through our in-house false-positive filter (–min-homopolymer 10). We required that indels were called by at least 2 out of 3 callers (GATK, Varscan, Pindel). In addition we also included Pindel unique calls (at least 30X coverage and 20% VAF). All combined indels were then processed through our false positive filter (–min-homopolymer 10 –min-var-freq 0.2 –min-var-count 6). We then applied additional annotation and minor allele frequency filters as previously reported [10].

Variant transcript annotation is based on all human transcripts obtained from Ensembl Release 70_37_v5. The reference alleles and positions were derived from the sequence and coordinates of GRCh37. All transcripts were annotated and a single representative was selected for each somatic mutation based on the significance of the predicted functional effect of each mutation, ordered from most significant to least significant as follows: nonsense, frameshift, splice site, in frame, missense, no stop (nonstop/readthrough), synonymous, and RNA. Splice site mutations were restricted to substitutions, deletions, or insertions overlapping the 2bp intronic sequence defined as the canonical splice donor or splice acceptor. RNA mutations were restricted solely to transcripts without an annotated open reading frame. Mutations affecting 3'UTR, 5'UTR, intronic sequence, and intergenic sequence were discarded for the purposes of downstream analysis.

3.5.2 Recurrent somatic mutations in 12 cancer types

We collected extensively filtered somatic variants in 3,355 TCGA samples from 12 cancer types (Supplementary Table 3.5), and selected recurrent mutations appearing more than once at a given genomic position (Supplementary Table 3.6).

3.5.3 Compiling cancer-associated gene list

A total of 556 candidate cancer-associated genes were compiled using nine sources, including recently published large-scale cancer studies, publicly available screening panels, and unpublished preliminary analysis of publicly available data sources. The 204 genes shared across at least two of the nine sources were retained and a literature search was conducted to identify evidence supporting inclusion of any remaining unique genes. A subset of 518 genes originated from recent publications, including 294 genes from Frampton *et al* [15], 125 genes from Kandoth *et al* [11], 212 genes from Lawrence *et al* [13], 194 genes from Pritchard *et al* [14], and 124 genes from Vogelstein *et al* [12]. Thirty-nine additional genes were included based on the analysis of driver mutations in 20 TCGA cancer types (unpublished), recommendations in accordance with the standards and guidelines of the American College of Genetics and Genomics [42], and 18 novel cancer driver genes identified in recently published large-scale studies (Supplementary Table 3.7).

3.5.4 Readcount analysis and statistical approaches for identifying significant low level variants

Read counts for variants were determined using our internally developed tool bam-readcount (https://github.com/genome/bam-readcount). For sites to be included in the downstream statistical test, we require greater than 30X coverage for both blood normal and matched tumor samples. We assessed the false-discovery rate (FDR) using a two-stage process. The first step is a purpose-specific pre-filter to eliminate candidates that can be confidently identified as having originated from non-mosaic samples, as these identically fail the inclusion criterion for significance testing. It targets deeply covered sites whose apparent variant reads actually represent base-calling errors. Take the sample space for blood classification as consisting of two mutually exclusive, collectively exhaustive (MECE) statuses, “normal”, $S=N$, and “mosaic”,

$S=M$, and let the candidate blood site data, D, consist of T spanning reads, of which V and T-V report variant and reference counts, respectively. If we presume that the rate of mosaicism (the fraction of altered cells, assumed as roughly 2%), ρ, is much larger than the Phred-determined likelihood of base-calling error for any single read, ε, i.e. ρ/ε>>1, then the conditional probabilities are binomial, $P(D \mid S) = C_{T,V}\, p^{V} (1-p)^{T-V}$, where $C_{T,V}$ is the number of combinations of T objects chosen V at a time and p=ε for $S=N$ and p=ρ for $S=M$. Given a prior estimate of $P(N)$=0.999, we can formulate $P(N|D)$ directly from Bayes' Rule, from which additional algebraic manipulation and suitable asymptotic approximation show

$$P(N \mid D) = \frac{1}{1 + P(M) \cdot (\rho/\varepsilon)^{V} \cdot \exp\left[-\rho(T-V)\right] / P(N)} \quad .$$

Candidates are culled as non-mosaic if this probability exceeds 95%, though in practice most cases actually removed have $P(N|D)$>99.5%. The remaining set of events is subsequently passed to the second step, which is a traditional Fisher exact (table) test for association between sample type and variant allele fraction followed by standard Benjamini-Hochberg FDR assessment. We report events having ≤20% FDR with at least 3 supporting reads and greater than 2% variant allele fraction.

3.5.5 Analysis of NHLBI ESP data

NHLBI variants calls for 6,503 samples were downloaded from the NHLBI Exome Variant Server http://evs.gs.washington.edu/EVS/. All variants were processed using the same tools as for the TCGA cohort. For comparative analysis, all ESP variants are filtered for < 0.1% total MAF to minimize false positives. The Women's Health Initiative Exome Sequencing Project (WHISP) is part of the National Heart, Lung, and Blood Institute's (NHLBI), Grand Opportunity Exome Sequencing Project (https://www.fhcrc.org/en/labs/phs/projects/cancer-

prevention/projects/whisp.html). WHISP data for 614 samples were downloaded from dbGaP, verified for file integrity, and then imported as BAM files into our data warehouse. Alignment to the reference genome GRCh37-lite was carried out using BWA v0.5.9 with parameters –t 4 –q 5 and marking of duplicates by Picard v1.46. Variant calling was performed as described in the "Variant calling and annotation strategies". For quality control purposes, we included WHISP samples with read mapping rates greater than 80%, duplication rates less than 40%, and at least 10,000 SNVs detected in the target region. The 557 Caucasian WHISP samples selected for this study, on average, had mapping rates of ~95%, duplication rates of ~9%, and ~18,000 SNVs called in the target region.

3.5.6 Deep sequencing validation

Five candidate low-level variants (two R882C and one R882H sample sites in *DNMT3A*, one V617F in *JAK2*, and one K700E in *SF3B1*), 5 positive controls, and 4 negative controls were selected for validation using deep sequencing (Supplementary Table 3.10). Primer pairs tailed with sample-specific indexes were designed for individual target sites and further used for PCR amplifications. Indexed libraries were made for tumor and blood pools respectively. We then generated sequencing data using 1 lane of MiSeq run with read length of 2x250. Custom references were created by including specific primer and index sequences. MiSeq reads were aligned to the custom references using BWA (bwa aln –t 8; bwa sampe). Allelic counts for the variants were obtained using in house tool bam-readcount (bam-readcount –b 0 –q 30).

3.6 References

1. Welch, John S., Timothy J. Ley, Daniel C. Link, Christopher A. Miller, David E. Larson, Daniel C. Koboldt, Lukas D. Wartman *et al*. The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150, 264-278 (2012).

2. Limpens, Jacqueline, R. Stad, C. Vos, C. De Vlaam, D. De Jong, G. J. Van Ommen, E. P. M. K. Schuuring, and Ph M. Kluin. Lymphoma-associated translocation t (14; 18) in blood B cells of normal individuals. *Blood* 85, 2528-2536 (1995).

3. Liu, Y. A. F. E. I., Antonio M. Hernandez, Darryl Shibata, and Gino A. Cortopassi. BCL2 translocation frequency rises with age in humans. *Proceedings of the National Academy of Sciences* 91, 8910-8914 (1994).

4. Biernaux, C., M. Loos, A. Sels, G. Huez, and P. Stryckmans. Detection of major bcr-abl gene expression at a very low level in blood cells of some healthy individuals. *Blood* 86, 3118-3122 (1995).

5. Forsberg, Lars A., Chiara Rasi, Hamid R. Razzaghian, Geeta Pakalapati, Lindsay Waite, Krista Stanton Thilbeault, Anna Ronowicz *et al*. Age-related somatic structural changes in the nuclear genome of human blood cells. *The American Journal of Human Genetics* 90, 217-228 (2012).

6. Jacobs, Kevin B., Meredith Yeager, Weiyin Zhou, Sholom Wacholder, Zhaoming Wang, Benjamin Rodriguez-Santiago, Amy Hutchinson *et al*. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature genetics* 44, 651-658 (2012).

7. Laurie, Cathy C., Cecelia A. Laurie, Kenneth Rice, Kimberly F. Doheny, Leila R. Zelnick, Caitlin P. McHugh, Hua Ling *et al*. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature genetics* 44, 642-650 (2012).

8. Busque, Lambert, Jay P. Patel, Maria E. Figueroa, Aparna Vasanthakumar, Sylvie Provost, Zineb Hamilou, Luigina Mollica *et al*. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature genetics* 44, 1179-1181 (2012).

9. Shlush, Liran I., Sasan Zandi, Amanda Mitchell, Weihsu Claire Chen, Joseph M. Brandwein, Vikas Gupta, James A. Kennedy *et al*. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* 506, 328-333 (2014).

10. Kanchi, Krishna L., Kimberly J. Johnson, Charles Lu, Michael D. McLellan, Mark DM Leiserson, Michael C. Wendl, Qunyuan Zhang *et al*. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications* 5 (2014).

11. Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).

12. Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science* 339, 1546-1558 (2013).

13. Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495-501 (2014).

14. Pritchard, Colin C., Stephen J. Salipante, Karen Koehler, Christina Smith, Sheena Scroggins, Brent Wood, David Wu *et al*. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of Molecular Diagnostics* 16, 56-67 (2014).

15. Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin *et al*. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology* 31, 1023-1031 (2013).

16. Ley, Timothy J., Li Ding, Matthew J. Walter, Michael D. McLellan, Tamara Lamprecht, David E. Larson, Cyriac Kandoth *et al*. DNMT3A mutations in acute myeloid leukemia. *New England Journal of Medicine* 363, 2424-2433 (2010).

17. Delhommeau, François, Sabrina Dupont, Véronique Della Valle, Chloe James, Severine Trannoy, Aline Masse, Olivier Kosmider *et al*. Mutation in TET2 in myeloid cancers. *New England Journal of Medicine* 360, 2289-2301 (2009).

18. Kralovics, Robert, Francesco Passamonti, Andreas S. Buser, Soon-Siong Teo, Ralph Tiedt, Jakob R. Passweg, Andre Tichelli, Mario Cazzola, and Radek C. Skoda. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *New England Journal of Medicine* 352, 1779-1790 (2005).

19. Gelsi-Boyer, Véronique, Virginie Trouplin, José Adélaïde, Julien Bonansea, Nathalie Cervera, Nadine Carbuccia, Arnaud Lagarde *et al*. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *British journal of haematology* 145, 788-800 (2009).

20. Abdel-Wahab, Omar, Jie Gao, Mazhar Adli, Anwesha Dey, Thomas Trimarchi, Young Rock Chung, Cem Kuscu *et al*. Deletion of Asxl1 results in myelodysplasia and severe developmental defects in vivo. *The Journal of experimental medicine* 210, 2641-2659 (2013).

21. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013, 2059-2074 (2013).

22. Yoshida, Kenichi, Masashi Sanada, Yuichi Shiraishi, Daniel Nowak, Yasunobu Nagata, Ryo Yamamoto, Yusuke Sato *et al*. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478, 64-69 (2011).

23. Li, Meng, Roxane Collins, Yuchen Jiao, Peter Ouillette, Dale Bixby, Harry Erba, Bert Vogelstein, Kenneth W. Kinzler, Nickolas Papadopoulos, and Sami N. Malek. Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. *Blood* 118, 5914-5917 (2011).

24. Zhang, Jinghui, Li Ding, Linda Holmfeldt, Gang Wu, Sue L. Heatley, Debbie Payne-Turner, John Easton *et al*. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* 481, 157-163 (2012).

25. Perez-Garcia, Arianne, Alberto Ambesi-Impiombato, Michael Hadler, Isaura Rigo, Charles A. LeDuc, Kara Kelly, Chaim Jalas *et al*. Genetic loss of SH2B3 in acute lymphoblastic leukemia. *Blood* 122, 2425-2432 (2013).

26. Bejar, Rafael, Kristen Stevenson, Omar Abdel-Wahab, Naomi Galili, Björn Nilsson, Guillermo Garcia-Manero, Hagop Kantarjian *et al*. Clinical effect of point mutations in myelodysplastic syndromes. *New England Journal of Medicine* 364, 2496-2506 (2011).

27. Weinstein, Lee S. Gsα Mutations in Fibrous Dysplasia and McCune-Albright Syndrome. *Journal of Bone and Mineral Research* 21, 120-124 (2006).

28. Aldred, Micheala A., and Richard C. Trembath. Activating and inactivating mutations in the human GNAS1 gene. *Human Mutation* 16, 183-189 (2000).

29. Lumbroso, Serge, Françoise Paris, and Charles Sultan. Activating Gsα mutations: analysis of 113 patients with signs of McCune-Albright Syndrome—a European collaborative study. *The Journal of Clinical Endocrinology & Metabolism* 89, 2107-2113 (2004).

30. Ringel, Matthew D., William F. Schwindinger, and Michael A. Levine. Clinical Implications of Genetic Defects in G Proteins: The Molecular Basis of McCune-Albright Syndrome and Albright Hereditary Osteodystrophy. *Medicine* 75, 171-184 (1996).

31. Ruark, Elise, Katie Snape, Peter Humburg, Chey Loveday, Ilirjana Bajrami, Rachel Brough, Daniel Nava Rodrigues *et al*. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* 493, 406-410 (2013).

32. Nangalia, Jyoti, Charles E. Massie, E. Joanna Baxter, Francesca L. Nice, Gunes Gundem, David C. Wedge, Edward Avezov *et al*. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *New England Journal of Medicine* 369, 2391-2405 (2013).

33. Walter, M. J., D. Shen, J. Shao, L. Ding, B. S. White, C. Kandoth, C. A. Miller *et al*. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* 27, 1275-1282 (2013).

34. Landau, Dan A., Scott L. Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S. Lawrence, Carrie Sougnez *et al*. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152, 714-726 (2013).

35. Akalin, Altuna, Francine E. Garrett-Bakelman, Matthias Kormaksson, Jennifer Busuttil, Lu Zhang, Irina Khrebtukova, Thomas A. Milne *et al*. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genetics* 8, e1002781 (2012).

36. Lu, Chao, Sriram Venneti, Altuna Akalin, Fang Fang, Patrick S. Ward, Raymond G. DeMatteo, Andrew M. Intlekofer *et al*. Induction of sarcomas by mutant IDH2. *Genes & development* 27, 1986-1998 (2013).

37. Siegel, Rebecca, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2013. *CA: a cancer journal for clinicians* 63, 11-30 (2013).

38. Li, Heng, and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).

39. Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22, 568-576 (2012).

40. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20, 1297-1303 (2010).

41. Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865-2871 (2009).

42. Rehm, Heidi L., Sherri J. Bale, Pinar Bayrak-Toydemir, Jonathan S. Berg, Kerry K. Brown, Joshua L. Deignan, Michael J. Friez *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine* 15, 733-747 (2013).

Acknowledgements

This work was supported by US National Cancer Institute grants R01CA180006 to L.D. and PO1CA101937 to T.J.L. and US National Human Genome Research Institute grants U54HG003079 to R.K.W. and U01HG006517 to L.D. M.J.W. is supported by a Leukemia and Lymphoma Society Scholar Award (1230-14), and J.S.W. is supported by R00HL103975. We thank M. Wyczalkowski for suggestions on figures and C. Kandoth on somatic mutation analysis. We also acknowledge The Cancer Genome Atlas as the source of primary data.

Contributions

This project was finished by a team in the lab. I was the major contributor to this project. With the help of Charles Lu and Jiayin Wang, I performed data analysis, including identification of blood somatic mutation, the association between blood somatic mutation and age, and comparison between blood somatic mutations and mutations in different disease cohorts. Besides, I worked with Heather Schmidt and performed the deep sequencing validation for the low VAF variants.

Table 3.1: Blood-specific mutations in 9 recurrently mutated genes identified in TCGA cases. Asterisks indicate nonsense mutations. VAF represents variant allele fraction. 11 cancer types were investigated in this study: BRCA (breast adenocarcinoma), GBM (glioblastoma multiforme), HNSC (head and neck squamous cell carcinoma), KIRC (kidney renal clear cell carcinoma), LGG (brain low grade glioma), LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma), OV (ovarian carcinoma), PRAD (prostate adenocarcinoma), STAD (stomach adenocarcinoma), and UCEC (uterine corpus endometrioid carcinoma).

Gene	Mutation	Case			Gene	Mutation	Case		
		Type	Age	VAF			Type	Age	VAF
DNMT3A	p.R882C	GBM	81	15.79%	*JAK2*	p.V617F	GBM	57	21.52%
		STAD	60	18.29%			GBM	72	73.39%
		STAD	69	12.17%			KIRC	59	28.57%
	p.R882H	BRCA	62	21.43%			LGG	45	15.87%
		GBM	64	35.56%			LUAD	72	27.62%
		LUSC	76	31.91%			LUAD	76	41.62%
	e13+1	KIRC	79	15.94%			UCEC	59	35.90%
		LUAD	76	11.11%			UCEC	74	42.92%
	p.E469*	GBM	72	20.60%	*ASXL1*	p.Q575*	LUAD	75	20%
	p.F851fs	BRCA	64	34.88%		p.Q733*	LUAD	72	14.29%
	p.K577fs	HNSC	72	24.14%		p.Q733fs	UCEC	81	27.27%
	p.N516fs	LUSC	71	33.33%		p.R548fs	LUAD	76	35.03%
	p.S770*	STAD	75	16.03%		p.Y591*	STAD	65	17.88%
	p.W314*	UCEC	74	22.06%		p.Y591fs	LUSC	56	29.70%
	p.Y584fs	GBM	75	38%	*TP53*	p.C275Y	OV	52	14.29%
	e12-1	PRAD	60	35.79%		p.Q136*	LUAD	null	18%
	e21-2	GBM	76	11.81%		p.Q144*	STAD	62	15.96%
	e22-1	UCEC	77	33.85%		p.R273L	LUAD	70	34.62%
TET2	p.F381fs	GBM	83	50%	*GNAS*	p.R202H	GBM	76	14.44%
	p.H863fs	GBM	64	11.67%			HNSC	59	11.54%
	p.K889*	OV	85	15.09%			LUAD	69	21.43%
	p.Q531*	KIRC	48	11.90%	*PPM1D*	p.Q520*	BRCA	79	35.42%
	p.Q644*	UCEC	89	16.78%		p.S468*	UCEC	49	21.23%
	p.Q764fs	GBM	75	33.01%	*BCORL1*	p.G883E	LUAD	null	16.67%
	p.Q831fs	LUAD	75	26.42%		p.S264*	PRAD	56	22.45%
	p.Q888*	GBM	83	20.39%	*SF3B1*	p.K700E	GBM	89	13.86%
	p.R550*	LUAD	76	16.25%			KIRC	77	43.04%
	p.T229fs	GBM	72	19.05%					

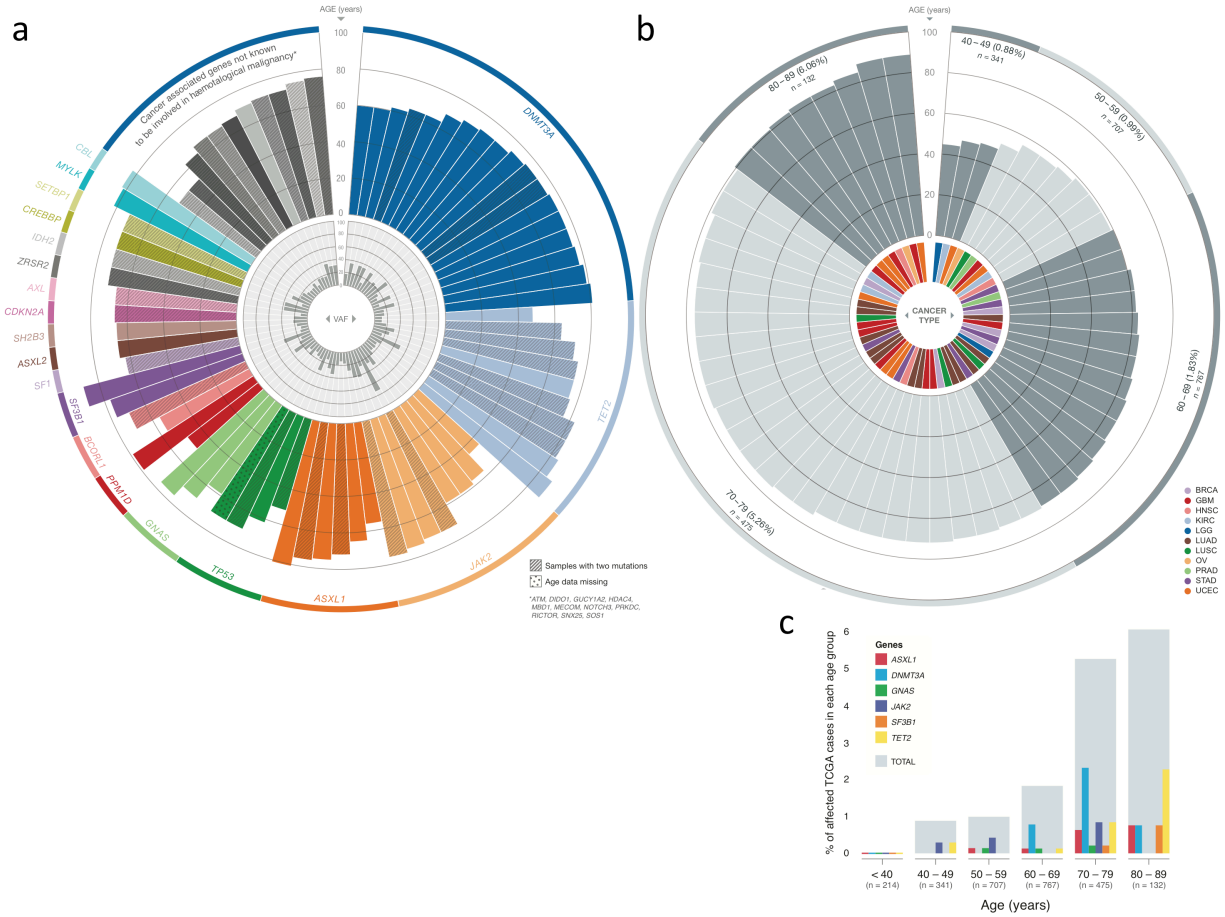


Figure 3.1: Blood-specific mutations identified in 58 out of 2,728 TCGA cases from 11 cancer types. (**a**) Rose chart illustrating the distribution of blood-specific nonsynonymous mutations in 31 genes. The variant allele fractions (VAF) of the 77 mutations are indicated in the center. (**b**) Rose chart illustrating the age distribution of samples with blood-specific mutations. Higher frequencies of blood-specific mutations are found in older age groups (60s, 70s, and 80s) versus younger ones (40s and 50s). The cancer type distribution is shown in the center. (**c**) Distribution of blood-specific mutations in *DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *SF3B1*, and *GNAS* in different age groups. Total includes all blood-specific mutations in 556 cancer associated genes identified in each age group.

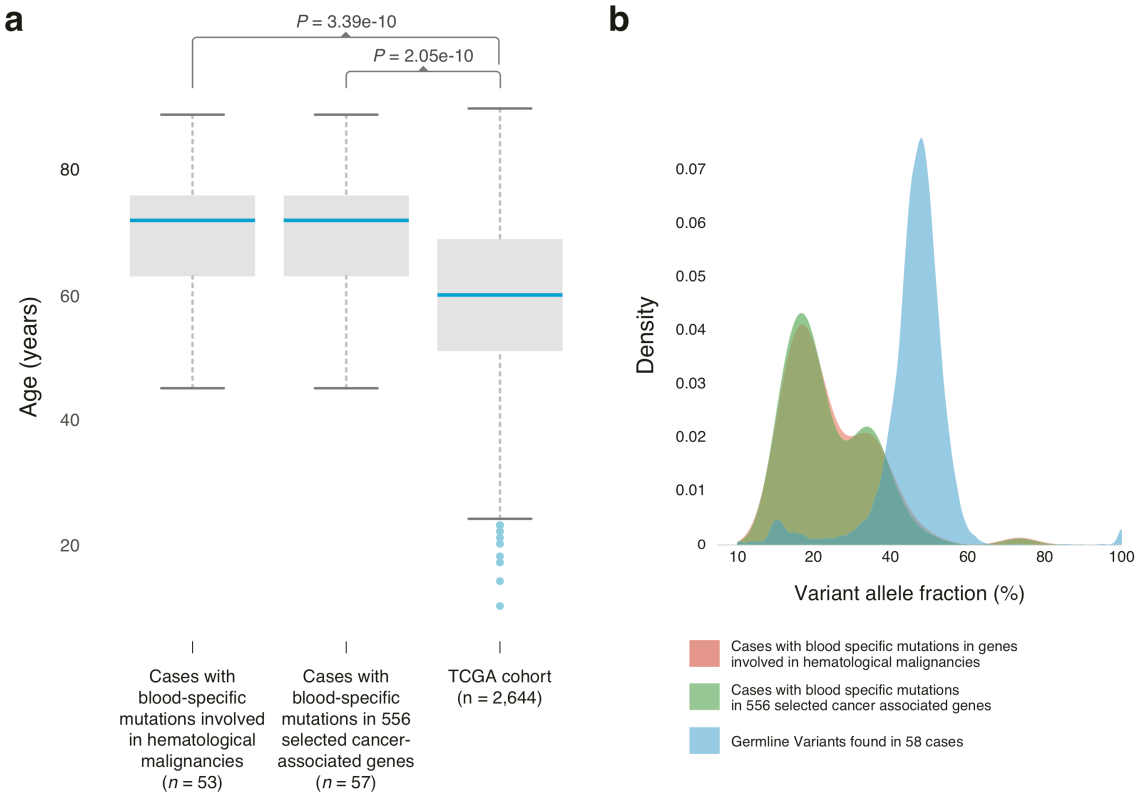


Figure 3.2: Blood-specific mutations and their association with age. (**a**) Box plot showing positive correlation between blood-specific mutations in leukemia/lymphoma genes and age. Age information is not available for one of the 58 cases with blood specific mutations. (**b**) The wide spectrum and lower average variant allele fractions in blood-specific mutations, compared to the ~50% VAF for germline variants identified in the same samples.

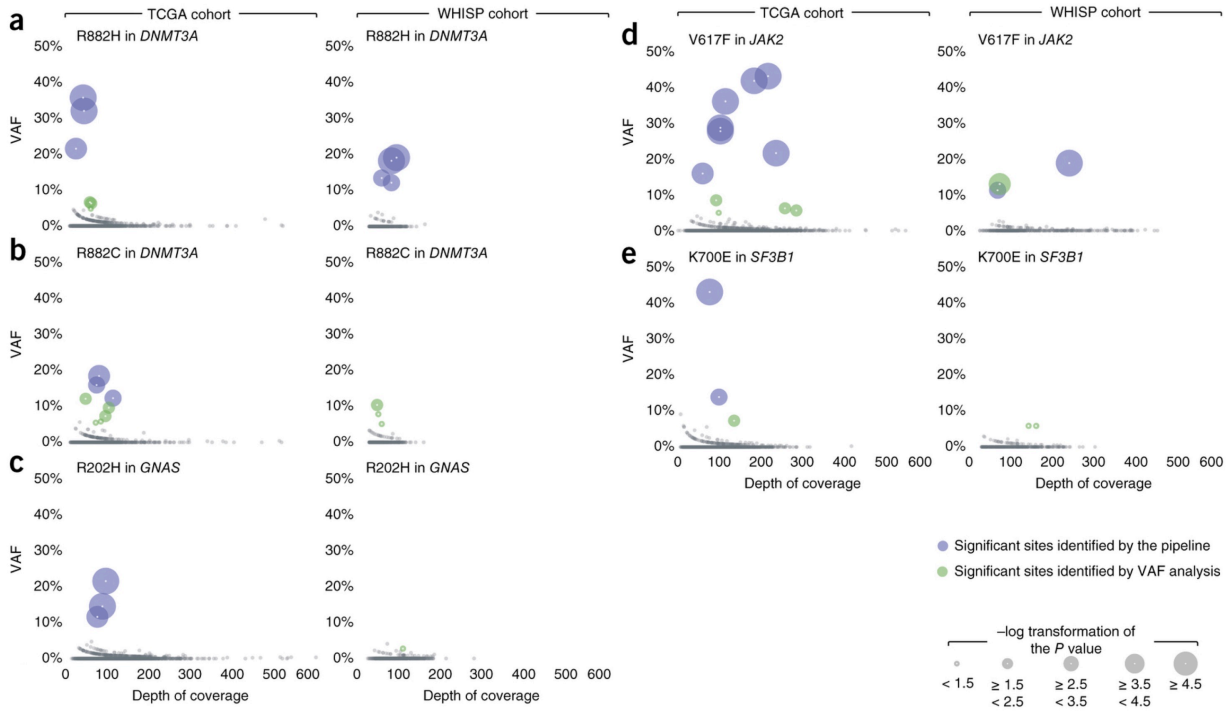


Figure 3.3: Low VAF blood-specific, hotspot mutations identified in the TCGA and WHISP cohorts using a readcount based approach. Blood-specific mutations identified by the variant detection pipeline are in blue. An additional 14 blood-specific events (13 shown in green) with VAFs between 2% and 10% were identified in the TCGA samples and their positive associations with older ages were confirmed. 13 hotspot variants were identified in WHISP samples (n = 557) and seven (in green) have variant allele fractions ranging from 2% to ~10%. One *JAK2* V617F identified in a TCGA sample was not shown due to a VAF higher than 50%. (**a**) *DNMT3A* R882C, (**b**) *DNMT3A* R882H, (**c**) *GNAS* R202H, (**d**) *JAK2* V617F, and (**e**) *SF3B1* K700E.

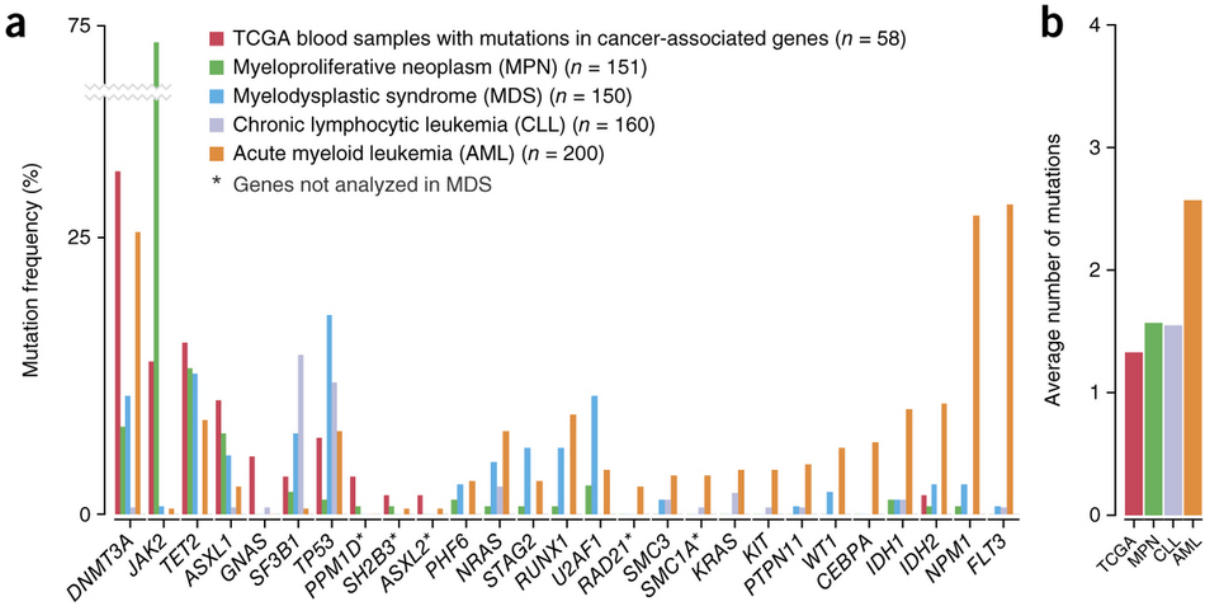


Figure 3.4: Comparison of mutation frequencies in blood samples from 58 TCGA cases with mutations in cancer-associated genes in 151 MPN, 150 MDS, 160 CLL, and 200 AML cases. (**a**) Mutation frequencies of major genes involved in hematological malignancies. (**b**) The average number of non-synonymous mutations found in TCGA blood normal cases, MPN, MDS, and AML patients across 556 cancer associated genes.

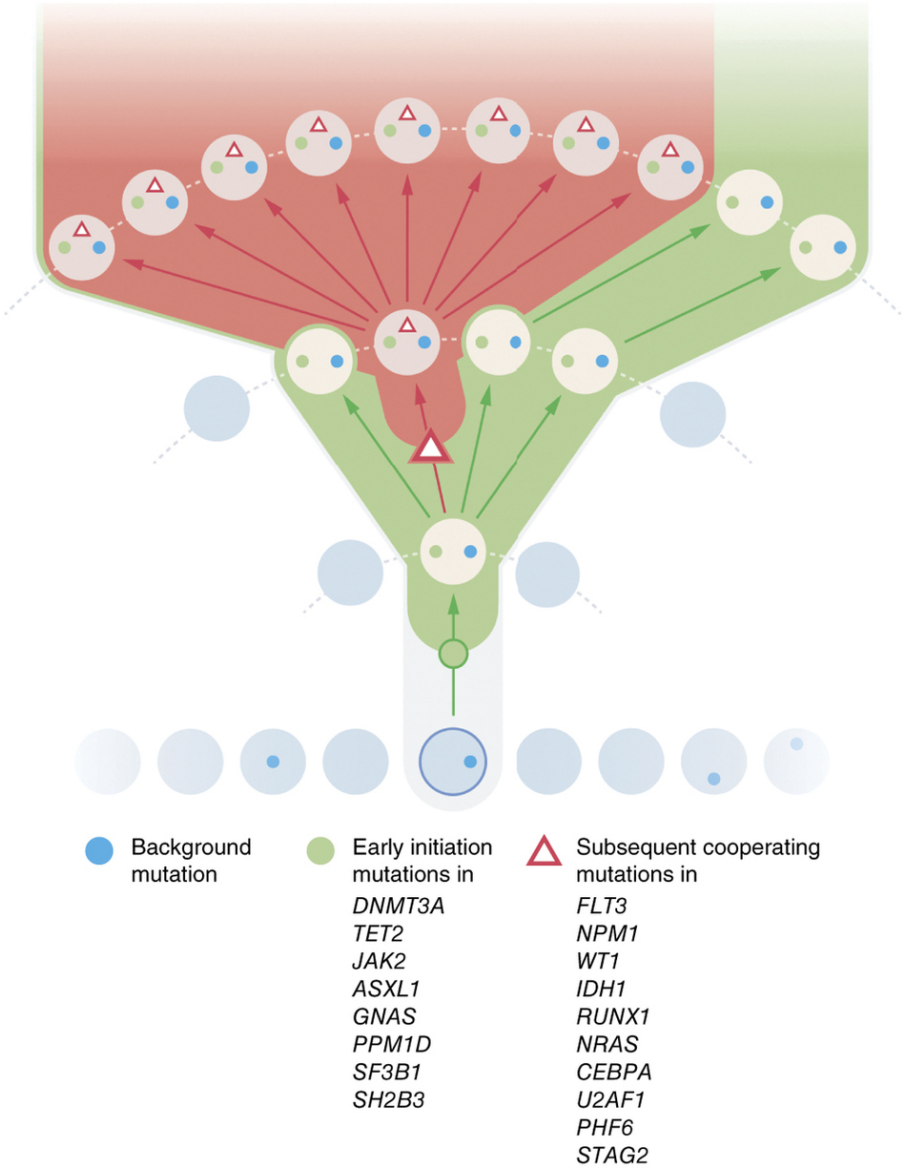


Figure 3.5: **Clonal expansion model.** The distinct roles of a set of genes including *DNMT3A*, *ASXL1*, *TET2*, *GNAS*, *JAK2*, *PPM1D, IDH1, NRAS, NPM1, and FLT3* in the initiation of hematopoietic clonal expansion.

Chapter 4: Genome-wide Pre-existing Genomic Alterations in Noncancerous Blood Associated with Clonal Hematopoietic Expansion

4.1 Abstract

Cancer development is a prolonged process. Mutations that initiate clonal expansion could exist in patients many years before disease symptoms are apparent. It is essential to identify and characterize these early mutations in order to understand the genetic basis of tumorigenesis. Here, I processed 5,949 normal blood samples from participants in TCGA and developed a statistical approach to systematically identify blood-specific mutations across all human genes. 13,345 blood-specific mutations in total were discovered, and they were significantly enriched in cancer and AML-associated genes. Blood-specific mutations were significantly distributed across 26 genes, including well-known hematologic malignancy associated genes (e.g., *DNMT3A*, *ASXL1*, *TET2*, *JAK2*, *IDH2*, and *SF3B1*), as well as previously unreported genes, including *PPM1D*. All blood-specific *PPM1D* truncation mutations identified in our study were located in the C-terminal regulatory domain, and functional validation showed that they suppressed phosphorylation of p53 at Ser15 *in vitro*. Lastly, I also identified blood-specific CNVs associated with genes related to clonal expansion. This comprehensive analysis of genomic alterations in noncancerous blood samples sheds light on the complex origins of hematologic malignancies and could facilitate the development of strategies for early detection and prevention of hematologic cancer.

4.2 Introduction

Blood cells are generated by the proliferation and differentiation of a very small population of pluripotent HSPCs, which have the ability to replenish themselves by self-renewal [1]. To ensure genomic stability and retain their capability, HSPCs rarely divide so as to avoid mutation occurrence and accumulation in the genome [2]. However, mutations are inevitable during cell division, and some of them have the potential to cause clonal hematopoietic expansion in "healthy" individuals before they develop severe disease symptoms. What such mutations are and how they drive clonal hematopoietic expansion are still unclear so far.

Recently, a few research groups have investigated these questions using large-scale sequencing data derived from individuals without overt hematologic disorders [3-5]. Two groups used blood samples alone to identify somatic mutations and found that the frequency of somatic mutations leading to clonal hematopoietic expansion increased with age and that about 10% of individuals older than 65 years carried these high-risk mutations in genes associated with myeloid and lymphoid cancers [3, 4]. These studies preliminarily provided insights into the connection between aging and the initiation of hematologic cancer. However, due to the lack of appropriate control samples from the same individual, the accuracy of somatic variant detection is concerning.

In a previous study, we analyzed blood-derived sequence data from TCGA and assessed the incidence of specific mutations within 556 cancer-associated genes by comparing variant allele fractions between blood and matched tumor tissues [5]. We found that the majority of somatic mutations identified in blood samples occurred in leukemia/lymphoma associated genes, including *DNMT3A*, *TET2*, *JAK2*, *ASXL1* and so on. However, since this study was limited to

candidate genes recurrently mutated in cancer, more comprehensive investigation of all human genes is needed.

Furthermore, structural variants, particularly copy number alterations, associated with hematologic malignancies were also observed in normal blood samples [6-8]. 2-3% of persons older than 70 years of age carried large chromosomal abnormalities involving genes associated with hematologic cancers [7, 8]. However, precisely identifying blood-specific somatic structural variants and characterizing their roles in the clonal hematopoietic expansion remains a challenge.

Here, I studied 5,949 whole exome sequencing data from TCGA, and systematically investigated blood-specific genomic alterations, including SNVs, small indels, and copy number variants (CNVs) across all human genes. To precisely determine somatic events, the matched tumor-adjacent normal tissue or tumor tissue from the same individual has been processed in the identical way such that it effectively can serve as a control. In this study, we showed a comprehensive genomic alteration profile within noncancerous blood cells and investigated the association between these alterations and hematologic cancer, and this study will benefit our understanding of hematologic cancer progression and future investigation for clinical applications.

4.3 Results

4.3.1 Exome-wide discovery of pre-existing mutations in blood

To identify blood-specific somatic mutations, I analyzed whole exome sequencing data generated from peripheral blood cells of 5,949 TCGA participants with no reported overt hematological malignancy. Samples represented 21 diverse cancer cohorts (Table 4.1). The

average age of these participants was 59 years old (ranging from 10 to 90), 3,639 were women (57.4%), and 4,738 were Caucasian (75.2%).

Variant calling in blood samples was conducted using previously described methods [5, 9]. The final set of variants was composed of 1,071,919 missense, 26,423 nonsense, 25,803 frame shift indels, 13,215 splice site mutations, and 1,202 nonstop mutations. By performing read count–based statistical analysis for all of the variants identified in copy number neutral regions (see Methods), 29,232 missense variants (2.7%) and 1,704 truncation variants (2.6%; nonsense, frame shift indel, nonstop, and splice site mutations) were classified as blood-specific somatic mutations. To capture potentially pathogenic missense mutations, further analysis was restricted to the set of missense mutations occurring in highly conserved sites (UCSC conservation score = 1.0; 11,641 missense variants remained).

The average variant allele fraction (VAF) of these mutations is 23.4% in blood samples and 2.6% in control samples (tumor or adjacent normal samples). Compared to the germline mutations identified in the same cohort, whose average VAF should be 50%, blood-specific mutations were only present in a small subset of the cells from which we obtained DNA for analysis, suggesting the presence of an early stage of clonal expansion (Figure 4.1a).

In addition, I detected well-known acute myeloid leukemia (AML) hotspot mutations in these 5,949 individuals using methods described previously [5] and identified a set of deleterious mutations that were significantly enriched in normal blood samples, including *DNMT3A*/R882 (28 cases), *JAK2*/V617F (19 cases), *IDH2*/R140Q (11 cases), *SF3B1*/K700E (6 cases), and *GNAS*/R202H (6 cases) (Supplementary Table 4.1).

4.3.2 Significantly mutated genes associated with clonal hematopoietic expansion

In all of the detected blood-specific mutations, 944 were present in 351 cancer-associated genes, selected based on several recent studies [10-14], and 628 mutations occurred in 176 genes that were recurrently mutated in *de novo* AML (Supplementary Table 4.2) [15]. Considering the small fraction of cancer and AML-associated genes in the human genome, blood-specific somatic mutations were significantly enriched in these two gene groups ($p < 2.2e-16$, respectively, Fisher's exact test) (Figure 4.1b). This result indicates that the mutations observed in blood samples were not randomly distributed, but rather, those associated with hematologic malignancy appear more likely to be retained.

A total of 181 genes were truncated in at least 2 individuals, including 26 cancer-associated or AML-associated genes. By combining the highly conserved missense mutations and AML hotspot variants, I performed the significantly mutated gene (SMG) test using the Mutational Significance in Cancer (MuSiC) suite of tools [16]. This analysis yielded 26 genes with a higher-than-expected mutation prevalence [false discovery rate (FDR) < 0.01], including genes with well-known relevance to hematologic malignancies (e.g., *DNMT3A*, *ASXL1*, *TET2*, *JAK2*, *IDH2*, and *SF3B1*), genes that have been implicated in the pathogenesis of hematologic diseases (such as *EPHB2* [17], *PTN* [18], *TIE1* [19], *GNB1* [20], and *ASH1L* [21]), as well as the genes that have not been reported in leukemia/lymphoma studies, such as *PPM1D*, *SPOP*, *MYH4*, *EMID2*, *FRG1*, *FRG1B*, *PCMTD1*, *TMEM196*, *ZNF318*, *EPPK1*, *GBAS*, *MORC2*, *SLC9A4*, *ZKSCAN4*, and *TMC1* (Figure 4.2, Supplementary Table 4.3).

DNMT3A, *TET2*, *ASXL1*, and *JAK2* were the top four most frequently mutated genes in our study, identified in 149 individuals, accounting for about 2.5% of the entire discovery cohort.

Remarkably, approximate 60% of *DNMT3A* mutations were potentially deleterious (26 truncation mutations and 29 R882 hotspot mutations), while for *ASXL1*, *TET2*, and *JAK2*, the ratio of potentially deleterious mutations was 100%, 86.4%, and 88.2%, respectively. The existence of these putative driver mutations associated with hematological malignancies suggested that a premalignant state was common in asymptomatic persons.

Besides, *TP53* and *NRAS* were also significantly mutated in the normal blood samples (FDR = 0.014 and 0.015, respectively), suggesting that they were likely to be involved in the process of early clonal hematopoietic expansion as well. In addition, a few leukemia/lymphoma-associated genes, such as *KDM6A*, *MLL3*, and *ATM*, did not pass the SMG test, but were recurrently mutated in the normal blood samples. For instance, 4 individuals had *KDM6A* mutations, 11 individuals carried *ATM* mutations, and 13 cases observed with *MLL3* mutations. Although these genes were frequently mutated in normal blood samples, more evidence is needed to determine if they are involved in the initiation of clonal expansion.

4.3.3 Identification of somatic CNV from blood samples using whole exome sequencing data

Chromosomal deletions and amplifications are often found in patients with AML and myelodysplastic syndromes [22, 23]. To investigate if the hematologic malignancy associated CNVs are present in noncancerous blood samples, I characterized the blood-specific CNV events from the same cohort using a statistical toolset, XHMM [24], which is specifically designed for detection of exon-resolution CNVs using whole exome sequence data (see Methods).

After systematically removing the individual, batch, and target effects by principal-component analysis of XHMM, I identified 143,455 CNVs from 5,456 individuals. Since common CNVs usually do not confer much advantage for diseases, I restricted the analysis to rare CNVs with a

minor allele frequency (MAF) of less than 0.5% by filtering out common instances in more than 55 individuals. Due to technical concerns, only CNVs on autosomal chromosomes were retained for the following analysis. To identify the high-confidence somatic CNVs in blood, I compared the genotyping quality scores of each rare CNV called in blood samples to the relevant target region in the control samples, such as tumor sample or adjacent normal sample from the same individual, and only those whose genotype in control were determined to be diploid with high certainty (quality score ≥ 60) were categorized as blood-specific somatic CNVs.

Using this method, 13,522 rare CNVs in autosomal chromosomes (about 22.9% of all well-defined, rare CNVs detected in the blood samples) were categorized as blood-specific. Duplication events were much more common than deletions, with 8,584 duplications (median length of 14.5 kb, ranging from 0.07kb to 3.6 Mb) and 4,938 deletions observed (median length of 18.5 kb, ranging from 0.12 kb to 3.7 Mb) (Figure 4.3a). About half of individuals (2,407, 44.1%) were detected to have rare blood-specific CNVs, with 881 individuals (16.1%) carrying only one rare CNV and 396 individuals (7.2%) carrying more than 10 (Figure 4.3a). On average, each individual carried 5.6 blood-specific rare CNVs (3.58 duplications and 2.05 deletions). The average age of individuals carrying blood-specific CNVs is 60.1 years old, which is significantly older than the average age of the individuals without blood-specific CNV (57.3 years old, p-value = 9.647e-13, student t-test) (Figure 4.3b).

4.3.4 Somatic CNVs associated with clonal hematopoietic expansion

The rare blood-specific CNVs observed in this cohort affected 8,329 unique autosomal genes. Of these genes, 58 were previously well-defined cancer-associated genes [12], including 25 oncogenes observed in copy number duplication regions, and 18 tumor suppressor genes in copy number deletion regions, in 154 individuals (Figure 4.4a, Supplementary Table 4.4). 16 of these

genes occurred in more than one sample, such as *KIT* (2 cases), *MET* (2 cases), *ACVR1B* (2 cases), *RB1* (2 cases), *SMARCA4* (2 cases), *CTNNB1* (3 cases), *ERBB2* (3 cases), *FLT3* (3 cases), *MLL3* (4 cases), *MLL2* (5 cases), *MSH6* (5 cases), *JAK2* (5 cases), *PIK3CA* (5 cases), *EGFR* (7 cases), *CASP8* (7 cases), and *NOTCH2* (9 cases) (Figure 4.4a). Importantly, most of these genes were associated with hematological malignancies.

In addition, several genes that were previously reported as clonal expansion-related genes were also affected by blood-specific CNVs. For example, *ASXL1* was entirely deleted in an 88-year old individual (Figure 4.4b), and *CREBBP* was partially deleted in another 68-year old individual. *GNAS*, *IDH2*, and *SF3B1* were found in the CNV duplication region from older individuals as well (88, 76, 81 years old, respectively) (Supplementary Table 4.4). Although *DNMT3A* and *TET2* were recurrently mutated in normal blood samples, no blood-specific CNVs were found to be associated with them, even though they were amply covered by data across all samples (Supplementary Figure 4.1).

4.3.5 Somatic gain-of-function *PPM1D* mutations in normal blood samples

We identified one novel recurrently mutated gene, *PPM1D*, with significant enrichment of somatic mutations in blood samples. The average VAF of *PPM1D* mutations in the discovery cohort was even higher than that of *DNMT3A*, *TET2*, or *ASXL1*, suggesting that *PPM1D* has the potential to play an important role in clonal hematopoietic expansion. However, besides its association with breast and ovarian cancer predisposition [25], *PPM1D* has not been reported in any hematologic malignancy study.

To confirm that *PPM1D* mutations are also present in an independent set of normal blood samples, I examined the ExAC (The Exome Aggregation Consortium) database and found 17 *PPM1D* rare truncation mutations in 60,706 unrelated individuals (Supplementary Table 4.5).

All of these mutations occurred in the C-terminus, and 16 of them were located in the exon 6, which is consistent with our TCGA discovery cohort (Figure 4.5a).

PPM1D functions as a phosphatase that dephosphorylates and inactivates many DNA damage response mediators such as TP53 [26]. It has been reported that PPM1D truncating alterations in the C-terminus could enhance the ability of PPM1D to hinder activation of DNA damage response proteins [27]. To examine the functional consequences of the *PPM1D* blood-specific mutations identified in our study, I assessed the impact of 3 truncation mutations, C478*, Q520* and K549fs, as well as 2 missense mutations, Q404E and R599K, on TP53 phosphorylation by introducing them into HEK293T cells, which carry wild-type *TP53*. As controls, I also introduced wild-type PPM1D, a phosphatase-dead D314A PPM1D mutant, and vector alone into HEK293T cells, respectively. Expression of the PPM1D mutants decreased the level of phosphorylation of TP53 at Ser15 after 10 Gy of irradiation (Figure 4.5b). This effect was also achieved by expressing wild-type PPM1D but not as strong as PPM1D mutant. These results demonstrated that blood-specific *PPM1D* mutants retained phosphatase activity against TP53, and suggested that PPM1D mutations were potentially involved in the cell proliferation regulation, stopping the cells from apoptosis after DNA damage. However, to fully evaluate the impact of PPM1D mutations in cell proliferation and clonal hematopoietic expansion, more *in vivo* experiments are needed.

4.4 Discussion

We observed that blood-specific somatic mutations were commonly present in healthy individuals, and the frequency of potential driver mutations (truncation mutations in cancer-associated genes) in the general population is consistent with previous studies [3-5]. While taking

highly conserved missense mutations into consideration, the fraction of people carrying mutations in cancer genes increases to 7.9% in their 50s and 16.6% in 70s.

I identified 26 SMGs with blood-specific somatic mutations in this study, including 6 established drivers for hematologic malignancy, such as *DNMT3A*, *ASXL1*, *TET2*, *JAK2*, *IDH2*, and *SF3B1*. Approximate 3.0% of individuals carried blood-specific mutations in these 6 genes. Of the remaining SMGs, 4 genes appear to be involved in leukemia/lymphoma. *EPHB2* (Ephrin type-B receptor 2) was found to play important roles in colon cancer development and glioma cell invasion [28, 29]. It was recently reported that EPHB2 was strongly expressed in $CD133^+$ cells and half of $CD34^+$ cells, suggesting that it might play a role in HPSC function through regulatory effects on cell adhesion, migration, and differentiation [17]. *GNB1* [Guanine nucleotide binding protein (G Protein), beta polypeptide 1] is involved in the PI3K-Akt signaling pathway, and recurrent *GNB1* mutations conferred cytokine-independent growth in IL-3 dependent lymphoid cells and promoted myeloid dendritic cell neoplasms *in vivo* [20]. *PTN* (Pleiotrophin) is an angiogenic factor during tumor growth and promotes invasion and metastasis in different tumor types [30, 31]. PTN was highly expressed in $CD19^+$ B cells from B-cell acute lymphoblastic leukemia (ALL) and B-cell chronic lymphocytic leukemia (CLL) patients, which suggests that it could be involved in the development of lymphocytic leukemia [18]. *Tie1* (tyrosine kinase with immunoglobulin-like and EGF-like domains 1) was expressed in the very early stages of hemopoietic and endothelial cell development, and it might be associated with the initial stage of B-cell CLL [19].

In this study, I found *PPM1D* truncation mutations were enriched in blood samples, and all of these truncation mutations were located in the protein's C-terminal regulatory domain. *In vitro* validation experiments demonstrated that these mutations increased PPM1D function on the

regulation of TP53 phosphorylation after DNA damage, suggesting that PPM1D blood-specific mutations might have regulatory effects on cell proliferation. Considering similar mutations found in ExAC, *PPM1D* has the potential to be involved in the early stages of clonal hematopoietic expansion.

In addition, I found that blood-specific CNVs were commonly present in healthy individuals. At least 58 cancer-associated genes were affected by these CNVs. Some of the somatic CNVs were associated with clonal expansion related genes, such as *JAK2*, *FLT3*, *ASXL1*, *CREBBP*, *IDH2*, *SF3B1*, *MLL3*, and *GNAS*. All of these results suggested that CNVs might be an important contributor to the clonal hematopoietic expansion. Although *DNMT3A* and *TET2* were frequented mutated in normal blood samples, no CNV events were found so far associated with these two genes. It could be due to the small sample size, or limitation of CNV detection using exome-sequencing data, but it also could indicate that point mutations and short indels are two major alteration forms for these two genes, but not CNVs. To fully explain that, additional sequencing data, especially whole genome sequencing data, or SNP-array data might be crucial.

In summary, this study provides a comprehensive profile of genomic alterations in noncancerous blood samples. It reveals that genomic alterations associated with clonal hematopoietic expansion are commonly present in older individuals, suggesting a potential risk of pre-malignancy in asymptomatic persons. This study provides numerous potential targets for novel clinical strategies, and also guides future investigation on clonal expansion.

4.5 Methods

4.5.1 Variant calling and annotation strategies

Whole exome sequencing data from blood samples were aligned to NCBI human reference Build 37 using BWA v0.5.9 and then de-duplicated using Picard v1.29. SNVs were detected by VarScan (v2.2.6) and GATK (revision 5336: −T UnifiedGenotyper). Only the SNVs called by both GATK and VarScan and passing our in-house false-positive filter (−min-homopolymer 10) were retained for the following analysis. Indels were identified using VarScan (v2.2.9), GATK (revision5336: −T IndelGenotyperV2), and Pindel (0.2.4x). We required that indels were detected by at least 2 out of 3 software (GATK, VarScan, Pindel). In addition, Pindel unique calls (minimal 30× coverage and 20% VAF) were also included. Moreover, then, all combined indels were processed through the false-positive filter (–min-homopolymer 10 −min-var-freq 0.2 −min-var-count 6). Additional annotation and minor allele frequency filters were applied as previously reported [5, 9].

Variant transcript annotation is based on all human transcripts obtained from Ensembl Release 70_37_v5. All transcripts were annotated and one single representative was selected for each variant based on the significance of the predicted functional effect, ordered as follows: nonsense, frame shift indels, splice site, in-frame indels, missense, nonstop, and synonymous. Splice site mutations were restricted to substitutions, deletions or insertions overlapping the 2-bp intronic sequence defined as the canonical splice acceptor or splice donor. Mutations affecting 3′ UTR, 5′ UTR, intergenic sequence and intronic sequence were discarded for the downstream analysis.

4.5.2 Determination of the blood-specific mutations

Read counts for variants were determined using our internally developed tool bam-readcount. The variants, which were located on copy number neutral regain, had read coverage greater than 20× in both blood normal and matched control samples (tumor or adjacent normal), and variant supporting reads in blood sample greater than 2, were retained for the downstream statistical test.

Let the data on the candidate variant as D, consist of T spanning reads, of which V and $T–V$ represent variant and reference counts, respectively, and variant allele fraction as p, and the conditional probabilities can be calculated based on binomial distribution as equation (1):

$$P(D \mid Variant\_Status) = C_{T,V} p^{V} (1-p)^{T-V} \quad (1)$$

Subsequently, we formulate $P(Variant\_Status|D)$ directly from Bayes' rule as the following equation (2):

$$P(Variant\_Status \mid D) = \frac{(C_{T,V} p_B^V (1-p_B)^{T-V})(C_{t,v} p_T^v (1-p_T)^{t-v}) P(Variant\_Status)}{P(D_B, D_T)} \quad (2)$$

Here, p_B and p_T indicate variant allele fraction in the blood sample and tumor/adjacent-normal sample, respectively, and t and v represent spanning reads and variant supporting reads for candidate variant in tumor/adjacent-normal sample.

Taking the variants in blood samples classified as two mutually exclusive, collectively exhaustive statuses, 'Germline Variants', $S = G$, and 'Blood-specific Somatic Variants', $S = S$, we can determine the variant status of each candidate by calculating LOD (log odds) score:

$$\begin{aligned} LOD &= \ln \frac{P(S \mid D)}{P(G \mid D)} \\ &= V \ln(\frac{p_{S,B}}{p_{G,B}}) + (T-V) \ln(\frac{1-p_{S,B}}{1-p_{G,B}}) + v \ln(\frac{\varepsilon}{p_{G,T}}) + (t-v) \ln(\frac{1-\varepsilon}{1-p_{G,T}}) + \ln \frac{P(S)}{P(G)} \end{aligned} \quad (3)$$

We conservatively set VAF to 0.5 for a germline heterozygous variants in both blood and tumor/adjacent-normal samples, while for blood-specific variants in tumor/adjacent-normal sample, we assume the VAF in tumor/adjacent-normal is equal to the probability of error of that base call (each base has an associated Phred-like quality score q where $\varepsilon = 10^{-q/10}$). Based on the pan-cancer analysis, the median frequency of somatic mutation, $P(S)$, in leukemia is 0.037 per Mb [10,11], and the frequency of rare germline mutation, $P(G)$, is 50 per Mb. To obtain the high-confident blood-specific mutation, we use very stringent LOD score, 2.94, as a cutoff, which indicates $P(S|D)$ as 95% and $P(G|D)$ as 5%.

4.5.3 Significantly mutated genes analysis

We applied the SMG test in the MuSiC suite to identify significant genes. This test assigns mutations to seven categories: AT transversion, AT transition, CG transition, CpG transition, CpG transversion, CG transversion, and indel, and then uses statistical approaches based on convolution, the hypergeometric distribution (Fisher's exact test), and likelihood to combine the category-specific binomials to calculate overall P values. All P-values were combined using the methods described previously [9].

4.5.4 Copy number alteration detection

Copy number alterations were called by XHMM as described previously [24]. The exome information of human gene transcripts was downloaded from Ensembl Release 70_37_v5, padded by 2 bp on each side. The mean depth of sequencing coverage on each exon was calculated by GATK DepthOfCoverage module (mapping quality greater than 20) across all of the 388,584 targets for the exome sequencing data from 5,949 individuals, including blood samples, tumor-adjacent normal samples or tumor samples. Samples and targets with outlier read-depth values, low complexity, and extreme GC content were filtered out before principal-

component analysis and read-depth normalization [24]. Taking use of PCA-normalized read-depth data and removing any targets with high read-depth variance (standard deviation greater than 30), sample-level z scores of read depths were calculated by centering relative to all target depths in the sample. After that, CNVs were called using default XHMM parameters. In addition, all called CNVs were statistically genotyped by forward–backward HMM algorithm across all samples.

4.5.5 Identification of rare blood specific CNV

To select the rare events, any CNV that overlaps 50% of another CNV that occurs in more than 1% of all individuals (55 in our study) was filtered out [24]. To identify the high confident blood-specific somatic CNV, we require its quality score (SQ) greater than or equal to 60 in the blood sample, and the NQ (Phred-scaled qualities of No CNV event) greater than or equal to 60 in the matched control sample (tumor-adjacent normal or tumor samples) as well. Genes overlapped with the blood specific CNVs were collected for the following analysis.

4.5.6 Cell culture, transfection, and irradiation

HEK293T cells were cultured in DMEM (Invitrogen) media supplemented with 10% FBS (Gibco). For transfection experiments, HEK293T cells were plated in six-well plates and transfected with 2 μg of plasmid using Lipofectamine 3000 (Invitrogen). After 24 h of transfection, the cells were exposed to 10 Gy of gamma irradiation (X-rays) and were harvested at 2 hours after irradiation.

4.5.7 Western blotting

Cells were washed twice with ice-cold 1× PBS and then suspended in RIPA lysis buffer (Santa Cruz Biotechnology) supplemented with protease inhibitor (Santa Cruz Biotechnology). After sonication and centrifugation, the protein concentrations were measured by bicinchoninic acid

(BCA) assay. A total of 30 µg of protein per sample was denatured at 70 °C for 10 min with NuPAGE LDS Sample Buffer (Life Technologies) and resolved by electrophoresis. Proteins were transferred to a nitrocellulose membrane, and membranes were blocked in Odyssey blocking buffer (LI-COR). After blocking, membranes were incubated with primary antibody, including mouse antibody to PPM1D (Santa Cruz Biotechnology, sc-376257); rabbit antibody to GAPDH (Santa Cruz Biotechnology, sc-25778); rabbit antibody to p53 (Cell Signaling Technology, 9282); rabbit antibody to p53 phosphorylated at Ser15 (Cell Signaling Technology, 9284), overnight at 4 °C. Membranes were washed four times using 1x PBS with 0.1% Tween-20 and then incubated with infrared fluorescence, IRDye® secondary antibodies for 45 mins, and the signal was detected using the LI-COR System.

4.6 Reference

1. Abkowitz, Janis L., Sandra N. Catlin, Monica T. McCallie, and Peter Guttorp. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood* 100, 2665-2667 (2002).

2. Catlin, Sandra N., Lambert Busque, Rosemary E. Gale, Peter Guttorp, and Janis L. Abkowitz. The replication rate of human hematopoietic stem cells in vivo. *Blood* 117, 4460-4466 (2011).

3. Jaiswal, Siddhartha, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley *et al*. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* 371, 2488-2498 (2014).

4. Genovese, Giulio, Anna K. Kähler, Robert E. Handsaker, Johan Lindberg, Samuel A. Rose, Samuel F. Bakhoum, Kimberly Chambert *et al*. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine* 371, 2477-2487 (2014).

5. Xie, Mingchao, Charles Lu, Jiayin Wang, Michael D. McLellan, Kimberly J. Johnson, Michael C. Wendl, Joshua F. McMichael *et al*. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* 20, 1472-1478 (2014).

6. Forsberg, Lars A., Chiara Rasi, Hamid R. Razzaghian, Geeta Pakalapati, Lindsay Waite, Krista Stanton Thilbeault, Anna Ronowicz *et al*. Age-related somatic structural changes in the nuclear genome of human blood cells. *The American Journal of Human Genetics* 90, 217-228 (2012).

7. Jacobs, Kevin B., Meredith Yeager, Weiyin Zhou, Sholom Wacholder, Zhaoming Wang, Benjamin Rodriguez-Santiago, Amy Hutchinson *et al*. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature genetics* 44, 651-658 (2012).

8. Laurie, Cathy C., Cecelia A. Laurie, Kenneth Rice, Kimberly F. Doheny, Leila R. Zelnick, Caitlin P. McHugh, Hua Ling *et al*. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature genetics* 44, 642-650 (2012).

9. Lu, Charles, Mingchao Xie, Michael C. Wendl, Jiayin Wang, Michael D. McLellan, Mark DM Leiserson, Kuan-lin Huang *et al*. Patterns and functional implications of rare germline variants across 12 cancer types. *Nature communications* 6 (2015).

10. Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).

11. Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-218 (2013).

12. Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science* 339, 1546-1558 (2013).

13. Pritchard, Colin C., Stephen J. Salipante, Karen Koehler, Christina Smith, Sheena Scroggins, Brent Wood, David Wu *et al*. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of Molecular Diagnostics* 16, 56-67 (2014).

14. Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin *et al*. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology* 31, 1023-1031 (2013).

15. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013, 2059-2074 (2013).

16. Dees, Nathan D., Qunyuan Zhang, Cyriac Kandoth, Michael C. Wendl, William Schierding, Daniel C. Koboldt, Thomas B. Mooney *et al*. MuSiC: identifying mutational significance in cancer genomes. *Genome research* 22, 1589-1598 (2012).

17. Lazarova, P., Q. Wu, G. Kvalheim, Z. Suo, K. W. Haakenstad, K. Metodiev, and J. M. Nesland. Growth factor receptors in hematopoietic stem cells: EPH family expression in CD34+ and CD133+ cell populations from mobilized peripheral blood. *International journal of immunopathology and pharmacology* 19, 49-56 (2005).

18. Du, Chun-xian, Lan Wang, Yan Li, Wei Xiao, Qin-lian Guo, Fei Chen, and Xin-ti Tan. Elevated expression of pleiotrophin in lymphocytic leukemia CD19+ B cells. *APMIS* 122, 905-913 (2014).

19. Aguayo, Alvaro, Taghi Manshouri, Susan O'Brien, Michael Keating, Miloslav Beran, Charles Koller, Hagop Kantarjian, Anna Rogers, and Maher Albitar. Clinical relevance of Flt1 and Tie1 angiogenesis receptors expression in B-cell chronic lymphocytic leukemia (CLL). *Leukemia research* 25, 279-285 (2001).

20. Walter, M. J., D. Shen, J. Shao, L. Ding, B. S. White, C. Kandoth, C. A. Miller *et al*. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* 27, 1275-1282 (2013).

21. Zhu, Li, Qin Li, Stephen HK Wong, Min Huang, Brianna J. Klein, Jinfeng Shen, Larissa Ikenouye *et al*. ASH1L links histone H3 lysine 36 di-methylation to MLL leukemia. *Cancer discovery* CD-16 (2016)

22. Mullighan, Charles G., Salil Goorha, Ina Radtke, Christopher B. Miller, Elaine Coustan-Smith, James D. Dalton, Kevin Girtman *et al*. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 446, 758-764 (2007).

23. Jacoby, Meagan A., and Matthew J. Walter. Detection of copy number alterations in acute myeloid leukemia and myelodysplastic syndromes. *Expert review of molecular diagnostics* 12, 253-264 (2012).

24. Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker *et al*. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics* 91, 597-607 (2012).

25. Ruark, Elise, Katie Snape, Peter Humburg, Chey Loveday, Ilirjana Bajrami, Rachel Brough, Daniel Nava Rodrigues *et al*. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* 493, 406-410 (2013).

26. Lu, Xiongbin, Bonnie Nannenga, and Lawrence A. Donehower. PPM1D dephosphorylates Chk1 and p53 and abrogates cell cycle checkpoints. *Genes & development* 19, 1162-1174 (2005).

27. Kleiblova, Petra, Indra A. Shaltiel, Jan Benada, Jan Ševčík, Soňa Pecháčková, Petr Pohlreich, Emile E. Voest *et al*. Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. *The Journal of cell biology* 201, 511-521 (2013).

28. Batlle, Eduard, Jeffrey T. Henderson, Harry Beghtel, Maaike MW van den Born, Elena Sancho, Gerwin Huls, Jan Meeldijk *et al*. β-Catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* 111, 251-263 (2002).

29. Nakada, Mitsutoshi, Jared A. Niska, Nhan L. Tran, Wendy S. McDonough, and Michael E. Berens. EphB2/R-Ras signaling regulates glioma cell adhesion, growth, and invasion. *The American journal of pathology* 167, 565-576 (2005).

30. Fang, Wenjing, N. Hartmann, D. T. Chow, A. T. Riegel, and A. Wellstein. Pleiotrophin stimulates fibroblasts and endothelial and epithelial cells and is expressed in human cancer. *Journal of Biological Chemistry* 267, 25889-25897 (1992).

31. Yanagisawa, Hiroko, Yukari Komuta, Hitoshi Kawano, Masashi Toyoda, and Kazunori Sango. Pleiotrophin induces neurite outgrowth and up-regulates growth-associated protein (GAP)-43 mRNA through the ALK/GSK3β/β-catenin signaling in developing mouse neurons. *Neuroscience research* 66, 111-116 (2010).

Acknowledgement

This work was supported by the National Cancer Institute grants R01CA180006 and R01CA178383 to L.D., the National Human Genome Research Institute grants U01HG006517 to L.D., and NHGRI Genome Analysis Training Program (T32 HG000045) to M.X. We also thank The Cancer Genome Atlas as the source of primary data.

We acknowledge the following people for their contribution and help during data process and experiment validation: Feng Chen, Piyush Tripathi, Rajees Varghese, Kalyan Manda, Jie Ning, Dinesh K. Thotala and Andrea Collins.

Table 4.1: Samples used in this study.

Cancer Type	Number of Cases	Age (Year +/- S.D.)	Cancer Type	Number of Cases	Age (Year +/- S.D.)
BLCA	215	67.87 +/- 11.17	LUAD	363	65.13 +/- 9.721
BRCA	874	58.12 +/- 12.84	LUSC	220	66.61 +/- 8.947
CESC	175	48.09 +/- 13.07	OV	358	59.46 +/- 11.47
COAD	262	65.53 +/- 13.18	PRAD	283	60.51 +/- 6.948
GBM	340	60.14 +/- 13.34	READ	106	63.92 +/- 12.15
HNSC	459	60.98 +/- 11.89	SARC	122	61.31 +/- 14.16
KICH	8	47.12 +/- 11.48	SKCM	343	56.97 +/- 15.75
KIRC	72	58.23 +/- 11.57	STAD	254	65.63 +/- 10.45
KIRP	103	57.69 +/- 12.33	THCA	392	46.46 +/- 15.63
LGG	449	43.0 +/- 13.49	UCEC	463	63.83 +/- 11.16
LIHC	88	59.94 +/- 14.12			
Total	**5949**	**58.71 +/- 14.21**			

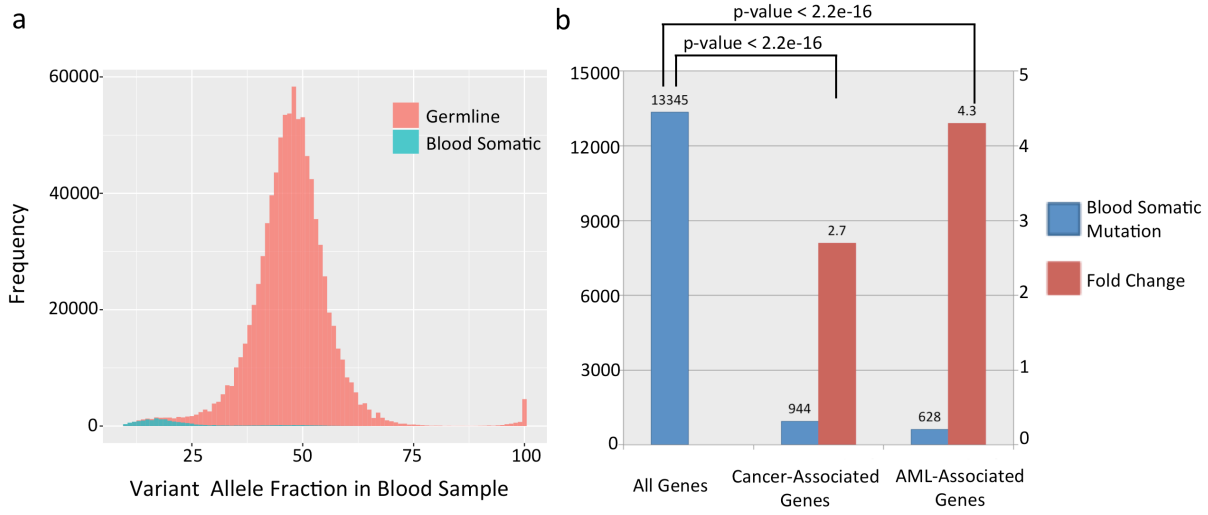


Figure 4.1: Blood-specific mutations identified in 2,363 TCGA cases from 21 cancer types. (**a**) Blood-specific somatic mutations were only present in a small subset of blood cells. Variant allele fraction distributions in blood samples are shown here, red indicating germline variants and green indicating blood-specific somatic mutations. (**b**) Blood specific somatic mutations were enriched in the cancer-associated and AML-associated genes. The red bar represents the number of mutations in each given gene group. Blue bar indicates the folder change of mutations in each given gene group compared to the fraction of each gene group in all human genes. The p-value is calculated by Fisher's exact test.

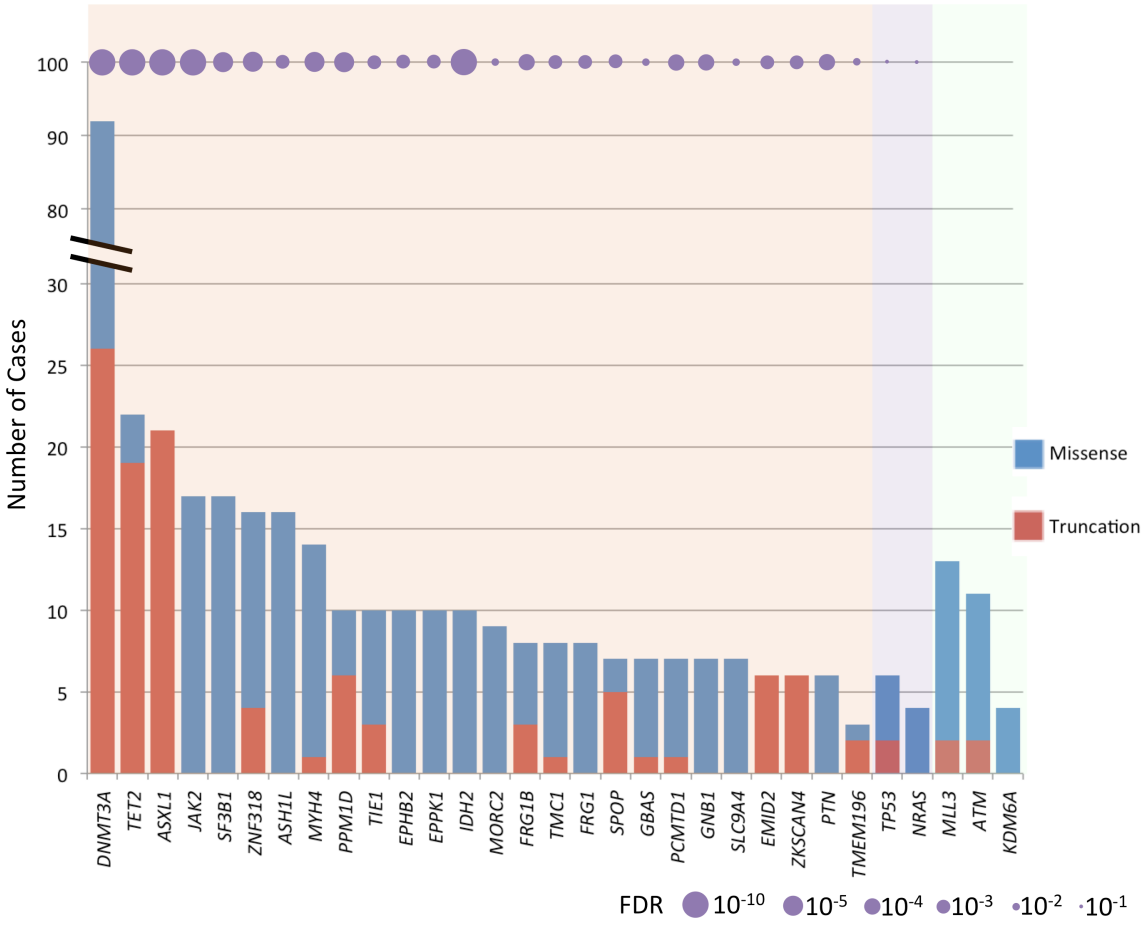


Figure 4.2: Significantly mutated genes identified in blood samples. The plot illustrates mutation frequencies of 26 SMGs and 5 recurrently mutated leukemia/lymphoma-associated genes, red representing truncation mutations and blue indicating highly conserved missense mutations. Purple dots on the top of each bar indicate significant events, with the size of dots proportional to the negative log of the FDR.

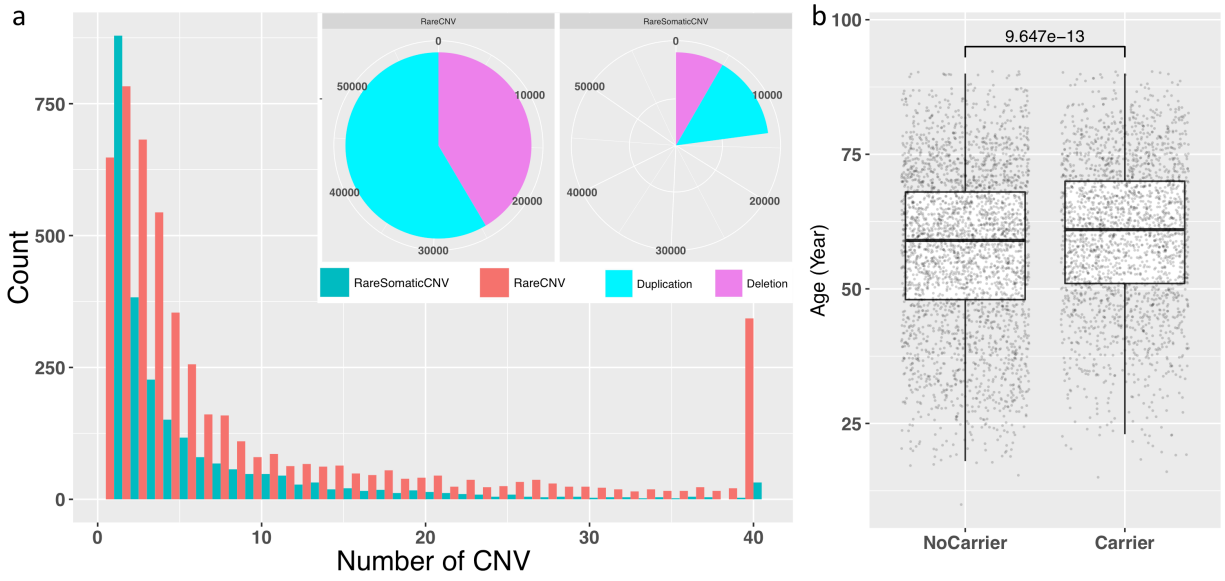


Figure 4.3: Overview of rare CNVs identified by XHMM in blood samples. (**a**) Characterization of rare CNVs across all of the participants. The histogram indicates the distribution of the number of rare CNVs and blood-specific rare CNVs per individual, and pie chart denotes the fraction of CNV deletion and duplication identified in blood samples. (**b**) Positive correlation between blood-specific CNV and carriers' age. Graph shows median central line, 50% confidence interval box, and 95% confidence interval whiskers.

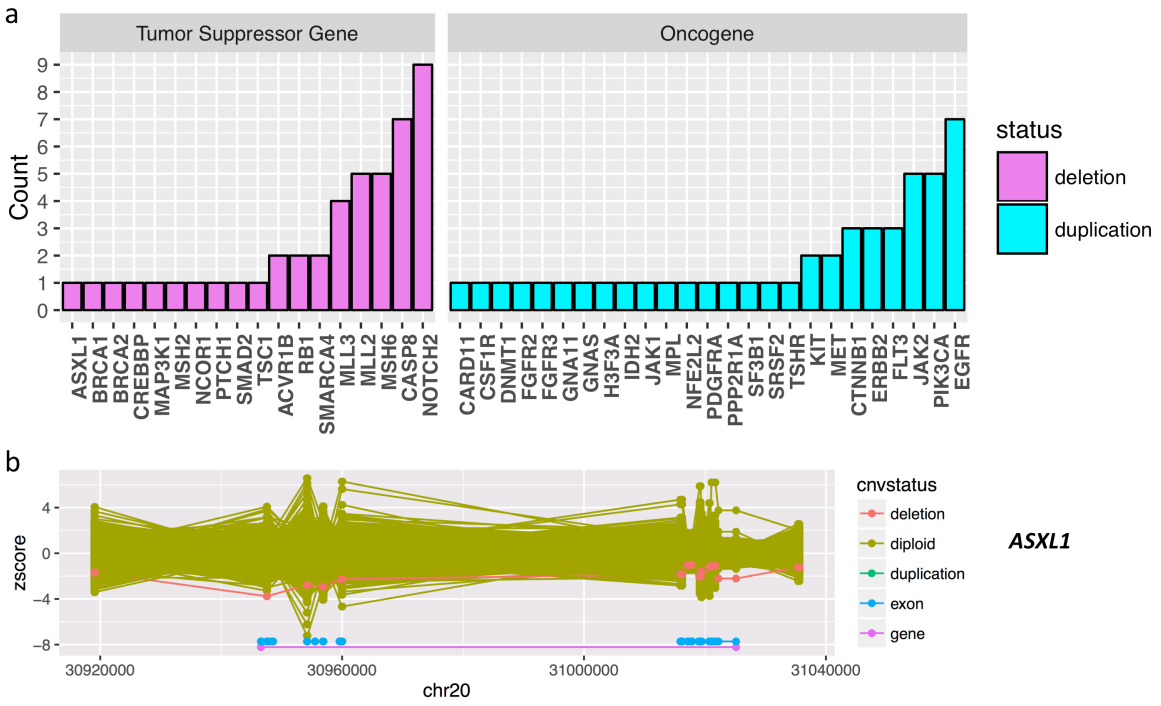


Figure 4.4: Cancer-associated genes affected by rare blood-specific CNVs. (**a**) Tumor suppressor genes and oncogenes were affected by CNV deletion and duplication in blood. (**b**) Copy number variation region plot for *ASXL1*. This plot shows normalized read depths at given targets across all the individuals with blood-specific CNVs. Samples with deletion were colored in red, duplications in green, and diploid in brown. Purple indicates the annotated gene region, and blue dots mark the location of the exome targets.

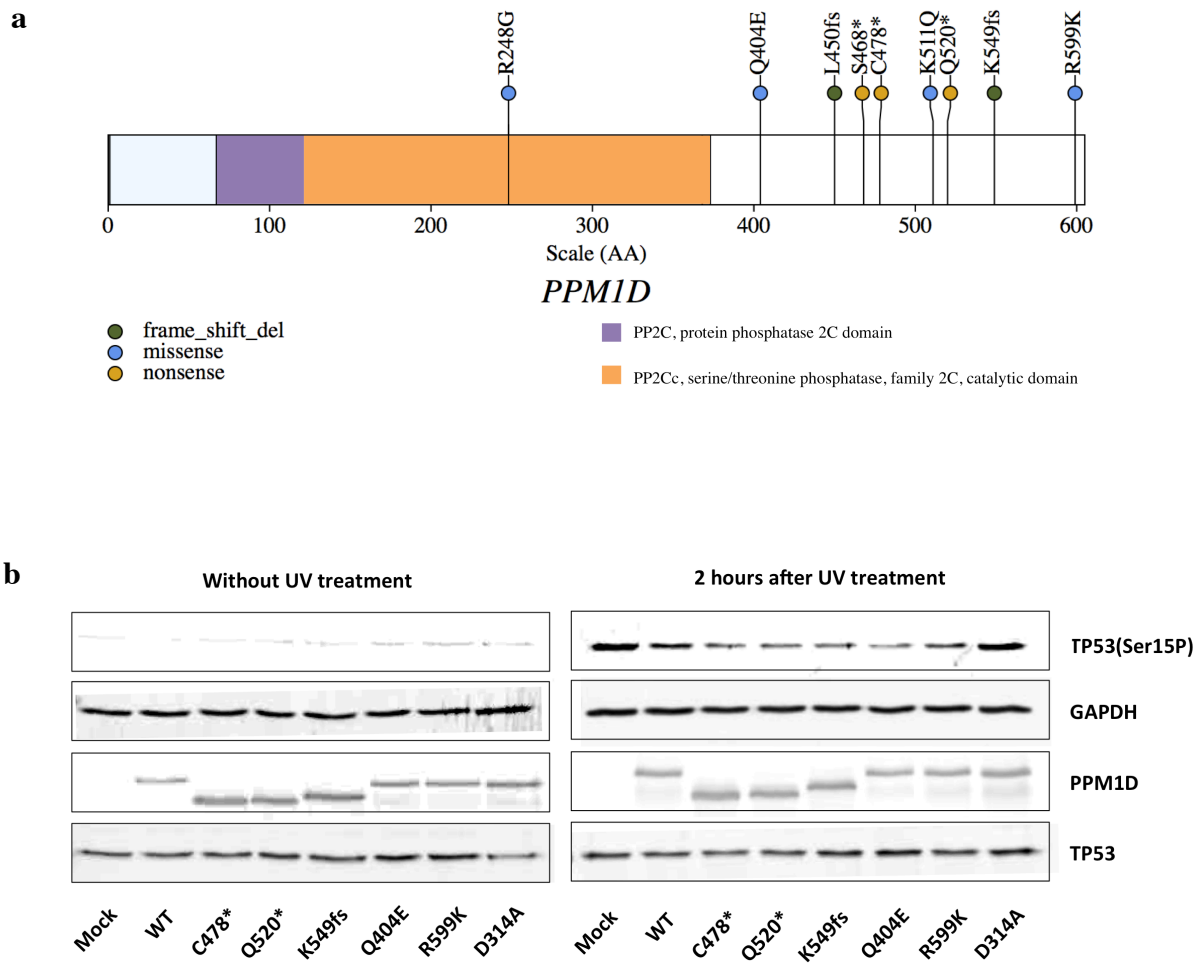


Figure 4.5: Functional validation of *PPM1D* blood-specific somatic mutations. (**a**) Blood-specific mutations were identified in *PPM1D*. Each dot corresponds to one case, orange, blue and dark green denoting nonsense mutations, missense mutations, and frame-shift indels, respectively. Variants in different functional domains are indicated with colors as follows: orange, PP2Cc, serine/threonine phosphatase, family 2C, catalytic domain; and purple, PP2C, protein phosphatase 2C domain. (**b**) Impact of PPM1D mutations on the phosphorylation of TP53 after DNA damage. Western blot shows the expression level of PPM1D, GAPDH, TP53 and TP53 phosphorylation at Ser15 with or without UV treatment. Indicated in the legend are the plasmids transfected to test. 'Mock' is empty vector and 'WT' represents wild-type PPM1D plasmid.

Chapter 5: Discussion and Future Directions

In cancer genome studies, blood samples represent a critical resource for revealing genetic alterations associated with cancer progression. Making use of TCGA whole exome sequencing data, we have identified many germline variants associated with cancer predispositions, as well as blood-specific somatic precursors for clonal hematopoietic expansion, in "normal" blood samples. However, this is just the beginning of a long journey, as many unknown variants, genes and molecular mechanisms underlying hematologic malignancy remain to be discovered. Here, we discuss other types of genomic alterations potentially present in noncancerous blood samples, and consider the collusion between genetic alterations and epigenetic defects that may contribute to clonal hematopoietic expansion.

5.1 Gene fusions in normal blood samples

A fusion gene is a chimera product formed by two separate genes and caused by chromosomal rearrangements, including inter-chromosomal or intra-chromosomal translocation [1, 2]. Gene fusions are extremely powerful mutations, which can dramatically change targeted gene expression through the juxtaposition of new promoter or enhancer regions, or create a chimeric protein with novel function.

Gene fusions are a very common event in the initial step of tumorigenesis, especially for hematologic cancers. The first fusion gene, as known as the Philadelphia chromosome, was discovered in the 1970s in chronic myelogenous leukemia (CML) [3]. It is caused by a reciprocal translocation between chromosome 9 and chromosome 22, leading to two genes, *BCR* and *ABL1*, joining together [4]. ABL1 as a tyrosine kinase plays a role in cell differentiation, cell division, and stress response [5]. The *BCR-ABL1* fusion gene creates a constantly active tyrosine kinase, which causes cell proliferation uncontrollably [6]. *BCR-ABL1* has been found in more than 95%

of CML patients and 25%–40% adults with acute lymphoblastic leukemia (ALL) [7, 8]. It is also worth noting that the frequency of the Philadelphia chromosome increases with age in ALL patients [9].

To date, 284 well-curated fusion genes, derived from 326 different reference genes have been identified in human neoplasia [10]. Most of the involved genes are transcription factors or tyrosine kinases. Interestingly, more than 25% of these fusion genes were found in hematologic disorders, and several genes were frequently involved, forming fusion genes with multiple partners, such as *KMT2A* (*MLL*), *ALK*, and *ETV6*, with more than 50, 10, and 5 partners, respectively [10]. A few fusion genes show strong cancer-type specificity, commonly present in particular leukemia or lymphoma subtypes. For instance, *PML-RARA* in acute promyelocytic leukemia and *IGH-CCND1* in mantle cell lymphoma [11, 12]. Moreover, some fusion genes could occur at very early stages of development, resulting in severe consequences. For example, the expression of an *MLL4-AF4* fusion gene at early stages of the embryo's development can cause infants to be born with a very aggressive form of leukemia [13].

These facts reflect a widely-held opinion that fusion genes play an important role in hematologic malignancies, which leads to some important biological questions: Are fusion genes also commonly present in noncancerous blood samples? And what are their roles in clonal hematopoietic expansion? To answer these questions, we first need to know how to detect fusion genes.

Fluorescence in situ hybridization (FISH) and RT-PCR has been used to identify fusion gene for decades [12]. However, these methods require prior knowledge of fusion partner genes, so it is hard for them to detect novel fusion genes. Despite this limitation, FISH and PCR are still most

valuable and commonly used technologies for fusion gene detection in both research and diagnosis of human neoplasia, because of high sensitivity and specificity, low costs, simplicity, and speed.

Recently, sequencing technologies, particularly RNA-Seq, have become more and more powerful for identifying and characterizing novel transcripts and fusion genes [14-17]. A number of new fusion genes have been detected using RNA-Seq data of tumors and cancer cell lines, such as *SLC45A3–ELK4* and *MSMB-NCOA4* in prostate cancer [14, 15], *VAPB-IKZF3* in breast cancer [16], and *RB1-ITM2B* in melanoma [17]. Aside RNA-Seq, whole genome sequencing (WGS) is also a very powerful resource for fusion gene detection, and it even provides more comprehensive and unbiased characterization of chromosomal translocation and gene fusion. Using WGS technology, a variety of fusion genes have been discovered in different cancers. For example, Welch *et al*. applied WGS to AML patients and identified classic bcr3 *PML-RARA* fusion gene and also found two novel fusion genes, *LOXL1-PML*, and *RARA-LOXL1* [18].

Considering the important roles that fusion genes play in hematologic malignancies and rapidly growing number of available RNA-Seq and WGS data, identifying fusion genes from "normal" blood samples will become feasible. It could enhance our understanding of the origins and progression factors of cancer, and provide potential preventive or therapeutic targets in anti-cancer treatments.

5.2 Epigenomic changes in clonal hematopoietic expansion

Besides genomic alterations, aberrant epigenetic changes are an important characteristic feature in a variety of cancers, including hematologic malignancy [19-22]. By incorporation with genomic DNA, the epigenome regulates gene expression in different cell types by altering

chromatin density and the accessibility of DNA to cellular machinery. Disruption of proper epigenetic maintenance can lead to inappropriate activation or inactivation of different genes or signal pathways, resulting in severe consequences, including cancer [23]. Although the role of epigenetic abnormalities in promoting hematologic malignancies is widely accepted, the detailed molecular mechanism remains unclear, so far.

Recently, genome-wide methylation patterns have been evaluated in blood cancer [22, 24]. A large study involving 344 AML patients showed that AML could be categorized into 16 subclasses according to methylation signatures. Some of the specific signatures were associated with well-known AML genomic alterations, such as *PML-RARA*, *CBFB-MYH11*, and *RUNX1-RUNX1T1* (*AML1-ETO*) [24]. However, *PML-RARA* had very limited impacts on overall DNA methylation by itself. A recent study using a murine model suggested that *DNMT3A* is required for *PML-RARA* to initiate acute promyelocytic leukemia *in vivo* [25]. This result suggests the methylation pattern in AML could be caused by the interaction of multiple genomic alterations.

DNA methylation regulators, such as *DNMT3A*, *TET2*, and *IDH1/2*, were recurrently mutated in leukemia and lymphoma patients with distinct DNA methylation phenotypes [26, 28]. For example, *DNMT3A* mutations occur in about 20% of AML patients [26]. AML patients with *DNMT3A* R882 mutations showed a pattern of global hypomethylation, especially at CpG islands. Some of the hypomethylated genes were previously linked to AML, such as homeobox-containing transcription factors [27]. On the other hand, *TET2* and *IDH1/2* mutations can cause DNA hypermethylation phenotype in AML patients [28]. Perhaps due to their similar biological effects, *TET2* and *IDH1/2* mutations were mutually exclusive with each other in AML [24].

In addition, genes that are involved in histone modification were also frequently mutated in AML. For instance, *ASXL1* mutations are present in 10-25% myelodysplastic syndrome, 10-15% myeloproliferative neoplasia, 5-30% AML and 43%–58% chronic myelomonocytic leukemia [29, 30]. *ASXL1* mutations can lead to loss of PRC2-mediated histone H3 lysine27 tri-methylation (H3K27me3), resulting in myeloid transformation [31].

Genes involved in epigenetic regulation were not only present in patients with hematologic malignancies, but also exist in "healthy" elderly individuals, such as *DNMT3A*, *TET2*, *ASXL1*, and *IDH2* [32]. Although a number of genome-wide studies and functional studies have been performed to explore the role of these genes in directing aberrant epigenetic changes and leukemogenesis, the detailed mechanisms linking these genes to the dynamic epigenetic change in the clonal hematopoietic expansion, and further passing through the bottleneck of the pre-leukemic state have yet to be fully elucidated.

Based on current studies, many known cancer-associated genes are regulated through direct or indirect epigenetic alterations. Therefore, it will be critical to apply epigenomic platforms to comprehensively elucidate how these epigenetic modifier mutations contribute to epigenetic alterations and gene expression in the initiation stage of hematologic malignancies. Studies of this nature have the potential to benefit the development of novel therapeutic strategies that aim to reverse epigenetic alterations in patients with mutations in epigenetic modifiers.

In summary, we have discussed the genomic alterations and mechanisms that are likely involved in the development of hematologic malignancies. As we pursue these goals, it will be critical to realize the necessity of functional validation of variants identified by computational and statistical approaches. Genome sequencing already represents a significant breakthrough for

cancer research, but it alone cannot address all these questions. Ultimately, functional studies are required to confirm analytical prediction in the proposed studies and make definite conclusions.

5.3 Reference

1. Mitelman, Felix, Bertil Johansson, and Fredrik Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer* 7, 233-245 (2007).

2. Edwards, Paul AW. Fusion genes and chromosome translocations in the common epithelial cancers. *The Journal of pathology* 220, 244-254 (2010).

3. Rowley, Janet D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. 290-293 (1973).

4. Lifshitz, B., E. Fainstein, C. Marcelle, E. Shtivelman, R. Amson, R. P. Gale, and E. Canaani. bcr genes and transcripts. *Oncogene* 2, 113-117 (1988).

5. SHAUL, Yosef, and BEN-YEHOYADA Merav. Role of c-Abl in the DNA damage stress response. *Cell research* 15, 33-35 (2005).

6. Wapner, Jessica. The Philadelphia chromosome: a genetic mystery, a lethal cancer, and the improbable invention of a lifesaving treatment. *Workman Publishing*, 2014.

7. Dreazen, Orna, Ivana Klisak, Gary Jones, Winston G. Ho, Robert S. Sparkes, and Robert Peter Gale. Multiple molecular abnormalities in Ph1 chromosome positive acute lymphoblastic leukaemia. *British journal of haematology* 67, 319-324 (1987).

8. Ottmann, Oliver G., and Heike Pfeifer. Management of Philadelphia chromosome–positive acute lymphoblastic leukemia (Ph+ ALL). *ASH Education Program Book* 2009, 371-381 (2009).

9. Pui, Ching-Hon, and William E. Evans. Treatment of acute lymphoblastic leukemia. *New England Journal of Medicine* 354, 166-178 (2006).

10. The Catalogue Of Somatic Mutations In Cancer, http://cancer.sanger.ac.uk/cosmic/fusion

11. Larson, Richard A., Koji Kondo, James W. Vardiman, Ann E. Butler, Harvey M. Golomb, and Janet D. Rowley. Evidence for a 15; 17 translocation in every patient with acute promyelocytic leukemia. *The American journal of medicine* 76, 827-841 (1984).

12. Li, Jian-Yong, Fanny Gaillard, Anne Moreau, Jean-Luc Harousseau, Christian Laboisse, Noël Milpied, Régis Bataille, and Hervé Avet-Loiseau. Detection of translocation t (11; 14)(q13; q32) in mantle cell lymphoma by fluorescence in situ hybridization. *The American journal of pathology* 154, 1449-1452 (1999).

13. Barrett, Neil A., Camille Malouf, Chrysa Kapeni, Wendi A. Bacon, George Giotopoulos, Sten Eirik W. Jacobsen, Brian J. Huntly, and Katrin Ottersbach. Mll-AF4 Confers Enhanced Self-Renewal and Lymphoid Potential during a Restricted Window in Development. *Cell Reports* 16, 1039-1054 (2016).

14. Maher, Christopher A., Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458, 97-101 (2009).

15. Nacu, Serban, Wenlin Yuan, Zhengyan Kan, Deepali Bhatt, Celina Sanchez Rivers, Jeremy Stinson, Brock A. Peters *et al*. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC medical genomics* 4, 1 (2011).

16. Edgren, Henrik, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H. Rye *et al*. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome biology* 12, 1 (2011).

17. Berger, Michael F., Joshua Z. Levin, Krishna Vijayendran, Andrey Sivachenko, Xian Adiconis, Jared Maguire, Laura A. Johnson *et al*. Integrative analysis of the melanoma transcriptome. *Genome research* 20, 413-427 (2010).

18. Welch, John S., Peter Westervelt, Li Ding, David E. Larson, Jeffery M. Klco, Shashikant Kulkarni, John Wallis *et al*. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *Jama* 305, 1577-1584 (2011).

19. Jones, Peter A., and Stephen B. Baylin. The fundamental role of epigenetic events in cancer. *Nature reviews genetics* 3, 415-428 (2002).

20. Sharma, Shikhar, Theresa K. Kelly, and Peter A. Jones. Epigenetics in cancer. *Carcinogenesis* 31, 27-36 (2010).

21. Schoofs, T., W. E. Berdel, and C. Müller-Tidow. Origins of aberrant DNA methylation in acute myeloid leukemia. *Leukemia* 28, 1-14 (2014).

22. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013, 2059-2074 (2013).

23. Feinberg, Andrew P., Rolf Ohlsson, and Steven Henikoff. The epigenetic progenitor origin of human cancer. *Nature reviews genetics* 7, 21-33 (2006).

24. Figueroa, Maria E., Sanne Lugthart, Yushan Li, Claudia Erpelinck-Verschueren, Xutao Deng, Paul J. Christos, Elizabeth Schifano *et al*. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer cell* 17, 13-27 (2010).

25. Cole, Christopher B., Angela M. Verdoni, Shamika Ketkar, Elizabeth R. Leight, David A. Russler-Germain, Tamara L. Lamprecht, Ryan T. Demeter, Vincent Magrini, and Timothy J. Ley. PML-RARA requires DNA methyltransferase 3A to initiate acute promyelocytic leukemia. *The Journal of clinical investigation* 126, 85 (2016).

26. Ley, Timothy J., Li Ding, Matthew J. Walter, Michael D. McLellan, Tamara Lamprecht, David E. Larson, Cyriac Kandoth *et al*. DNMT3A mutations in acute myeloid leukemia. *New England Journal of Medicine* 363, 2424-2433 (2010).

27. Qu, Ying, Andreas Lennartsson, Verena I. Gaidzik, Stefan Deneberg, Mohsen Karimi, Sofia Bengtzén, Martin Höglund, Lars Bullinger, Konstanze Döhner, and Sören Lehmann. Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes. *Epigenetics* 9, 1108-1119 (2014).

28. Figueroa, Maria E., Omar Abdel-Wahab, Chao Lu, Patrick S. Ward, Jay Patel, Alan Shih, Yushan Li *et al*. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer cell* 18, 553-567 (2010).

29. Gelsi-Boyer, Véronique, Mandy Brecqueville, Raynier Devillier, Anne Murati, Marie-Joelle Mozziconacci, and Daniel Birnbaum. Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. *Journal of hematology & oncology* 5, 1 (2012).

30. Thol, Felicitas, Inna Friesen, Frederik Damm, Haiyang Yun, Eva M. Weissinger, Jürgen Krauter, Katharina Wagner *et al*. Prognostic significance of ASXL1 mutations in patients with myelodysplastic syndromes. *Journal of Clinical Oncology* 2499–2506 (2011).

31. Abdel-Wahab, Omar, Mazhar Adli, Lindsay M. LaFave, Jie Gao, Todd Hricik, Alan H. Shih, Suveg Pandey *et al*. ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer cell* 22, 180-193 (2012).

32. Xie, Mingchao, Charles Lu, Jiayin Wang, Michael D. McLellan, Kimberly J. Johnson, Michael C. Wendl, Joshua F. McMichael *et al*. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* 20, 1472-1478 (2014).

Appendix 1 Supplementary Materials for Chapter 2

Supplementary Note

We also examined germline sites overlapping recurrent somatic mutations found in the 12 TCGA cancer types [1] (Supplementary Table 2.24). After stringent filtering, we identified 34 missense germline hotspot variants that overlapped recurrent somatic mutations in cancer associated genes, based on the somatic mutations reported in Kandoth *et al* [1]. Most of these 34 hotspot missense germline variants affected conserved nucleotide/amino acid residues. This list includes six variants in the DNA-binding domain of *TP53*, all occurring at five previously identified hotspots (R110H, R158C, R267Q, R175C, and G245V). G245V has been reported by the IARC as nonfunctional [2], while the four remaining variants were reported to have partial functionality (http://p53.iarc.fr/). One *ATM* (R2691C) variant, involved in CLL [3], is also known to interact with TP53 and can result in the transformation of the ATP binding pocket. Another prominent cluster of rare germline missense mutations appeared in DNA-repair (Fanconi Anemia) pathway. These included two recurrent variants (A625T) in *PARP1*, somatically mutated in bladder and endometrial cancer [1], a variant (E201K) in *DDX11*, proximal to a validated functional missense variant (R263Q) responsible for Warsaw Breakage [4], and one variant (R1084C) in *FANCA*. In addition, a germline variant (K140N) in the *BRCA1*-binding-partner *BARD1* was identified and its three adjacent residues were found to be recurrently mutated in multiple cancer types (COSMIC). A missense variant (E2020K) in *BRCA2*, recurrently mutated in other samples, is currently classified as a variant of unknown significance (BIC) [5].

Using existing clinical significance data from the NCBI ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/), a total of 101 rare germline missense variants were listed as Pathogenic. Of these, 40 were from tumor suppressor genes (TSG) or encoded proteins involved in DNA repair, including *POLH* (DNA repair; 9 variants), *BUB1B* (TSG; 9), *VHL* (TSG; 5), and *APC* (TSG; 5), among others. *BRCA2* was also on the list but had only 2 variants. The low occurrence of *BRCA2* variants and lack of *BRCA1* variants merit further investigation and suggest that this approach may be improved by querying additional clinical databases specific to breast cancer and other types of cancer. Other identified variants were involved in other diseases (e.g. *LRRK2* for Parkinson's disease) and also included an oncogene (*RET* proto-oncogene, 8) and DNA transcription factors (e.g. *AR*, 4). Additionally, only *TYR* and *BRCA2* are common to this list and the list of significant non-oncogenes obtained from truncation variant analysis, suggesting that the combination of both methods could be especially useful for identifying germline variants involved in cancer.

Supplementary References

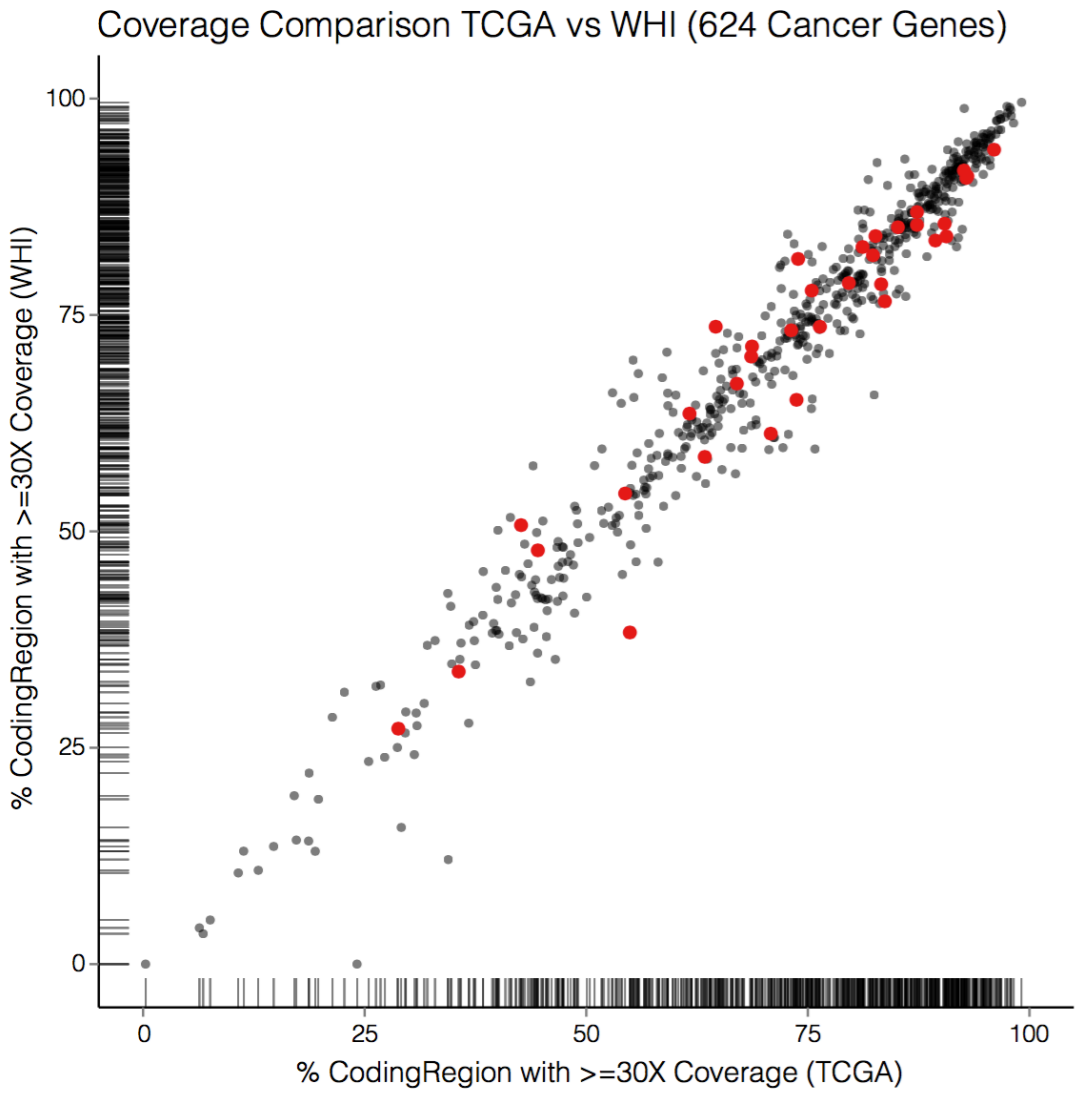
1. Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333-339 (2013).

2. Kato, Shunsuke, Shuang-Yin Han, Wen Liu, Kazunori Otsuka, Hiroyuki Shibata, Ryunosuke Kanamaru, and Chikashi Ishioka. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proceedings of the National Academy of Sciences* 100, 8424-8429 (2003).

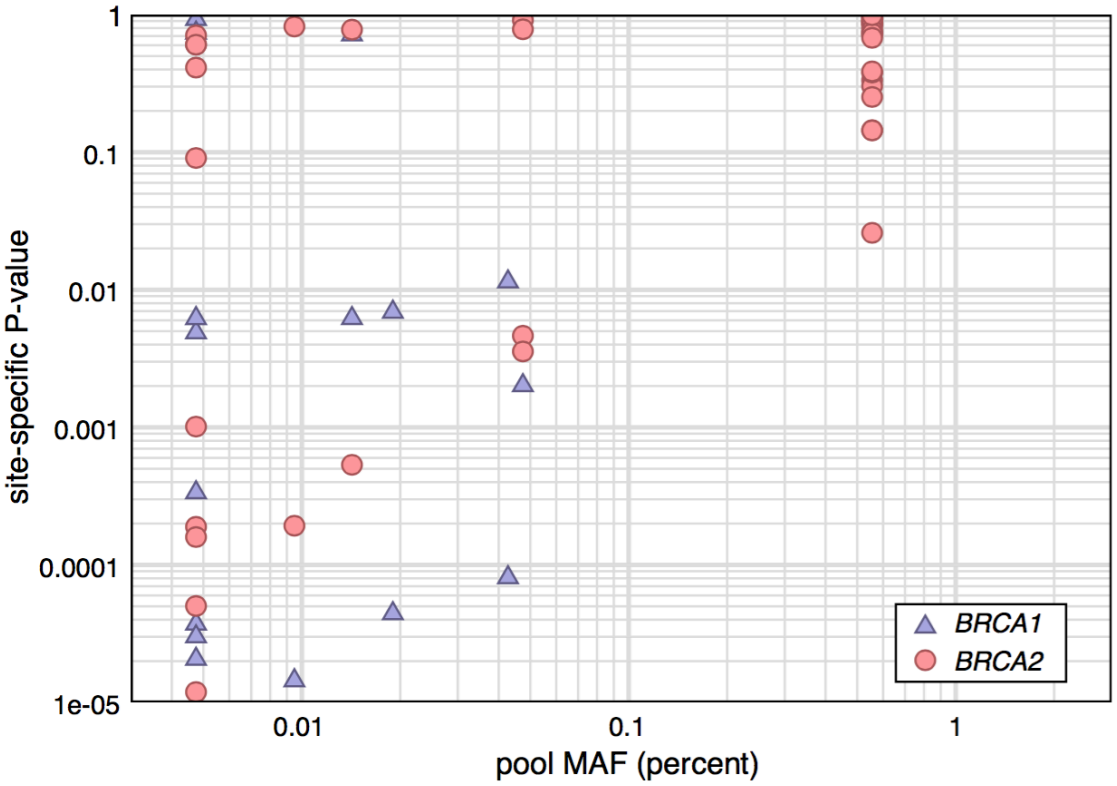
3. Guarini, Anna, Marilisa Marinelli, Simona Tavolaro, Emanuele Bellacchio, Monia Magliozzi, Sabina Chiaretti, Maria Stefania De Propris *et al*. ATM gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression. *Haematologica* haematol-2011 (2011).

4. Capo-Chichi, José-Mario, Sanjay Kumar Bharti, Joshua A. Sommers, Tony Yammine, Eliane Chouery, Lysanne Patry, Guy A. Rouleau *et al*. Identification and biochemical characterization of a novel mutation in DDX11 causing Warsaw breakage syndrome. *Human mutation* 34, 103-107 (2013).

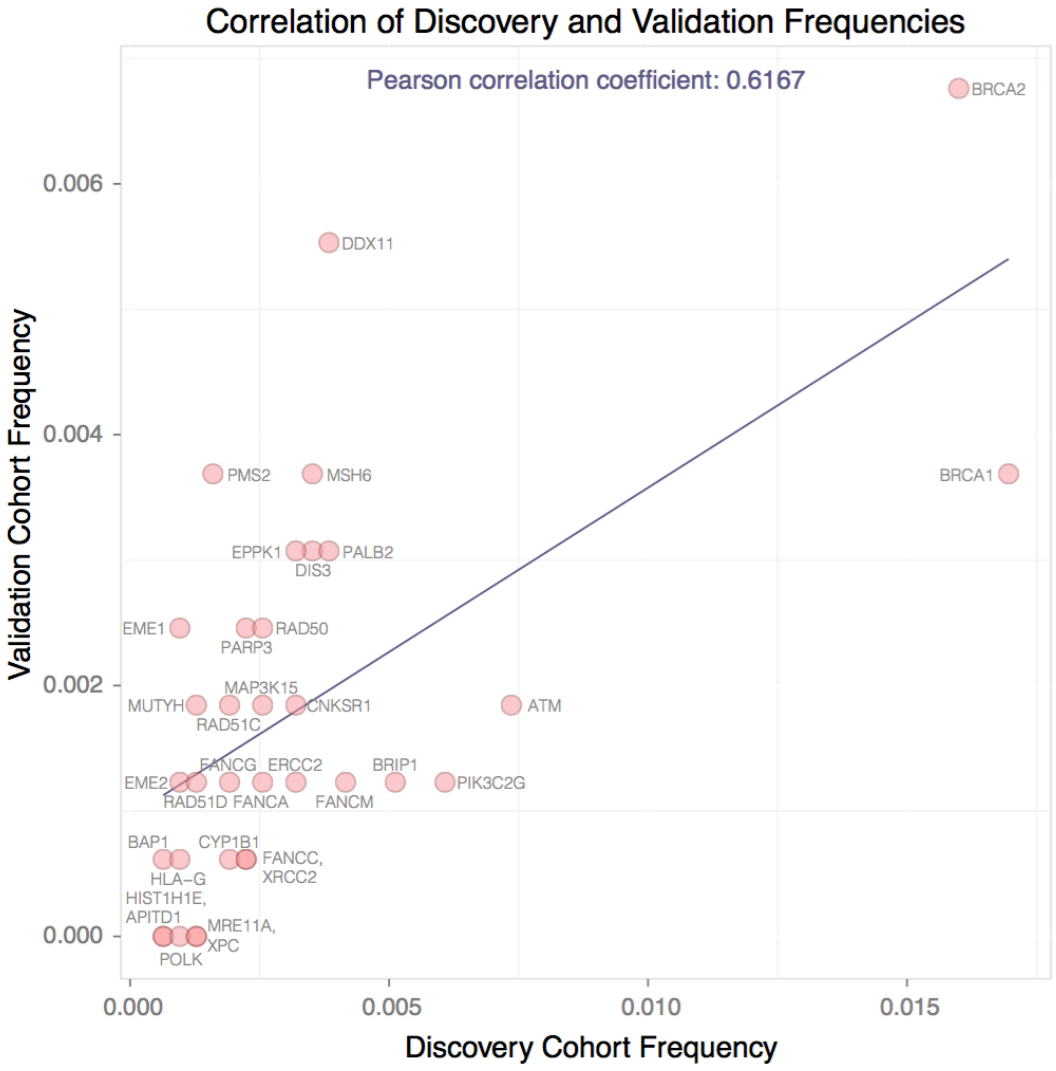
5. Szabo, Csilla, Anthony Masiello, Joseph F. Ryan, and Lawrence C. Brody. The breast cancer information core: database design, structure, and scope. *Human mutation* 16, 123 (2000).



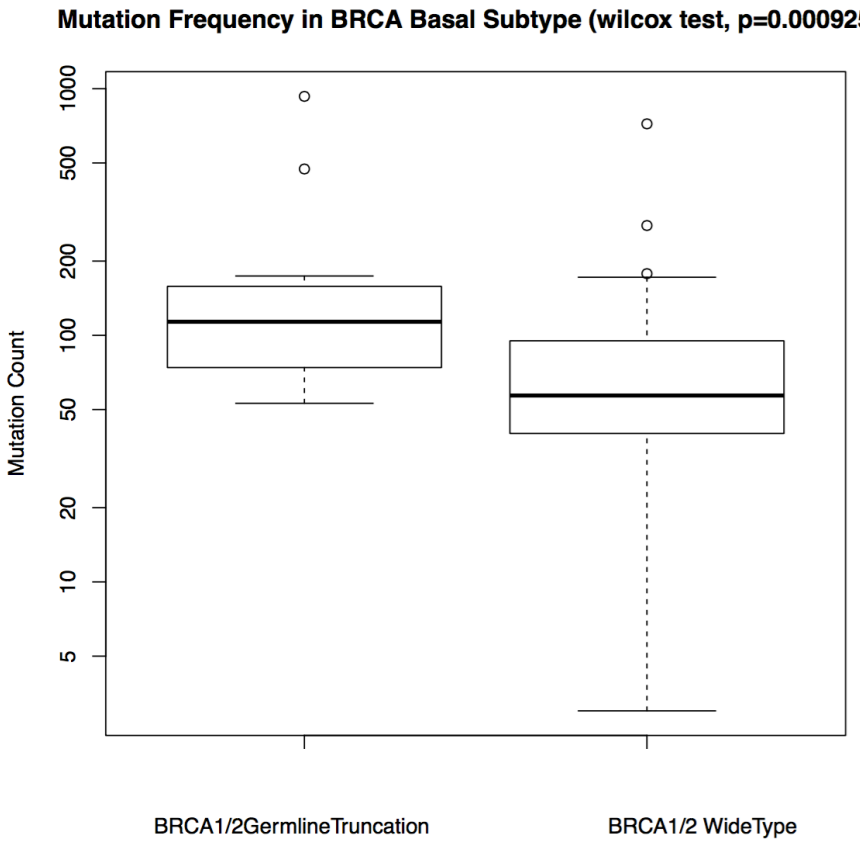
Supplementary Figure 2.1: Comparison of Coverage between the Caucasian TCGA cohort and WHISP cohort. We compared the coverage of the 624 cancer genes between 3,125 TCGA Caucasian and 1,039 WHI cases. The scatter plot shows the mean percent of coding regions with ≥30X coverage for each of these 624 genes in the TCGA cohort (X axis) and WHI (Y axis), with the Pearson's correlation coefficient of 0.98. A subset of the 624 dots (denoted in red) represents genes with significant enrichment of rare truncation variants determined in the TCGA cohort by the burden analysis.



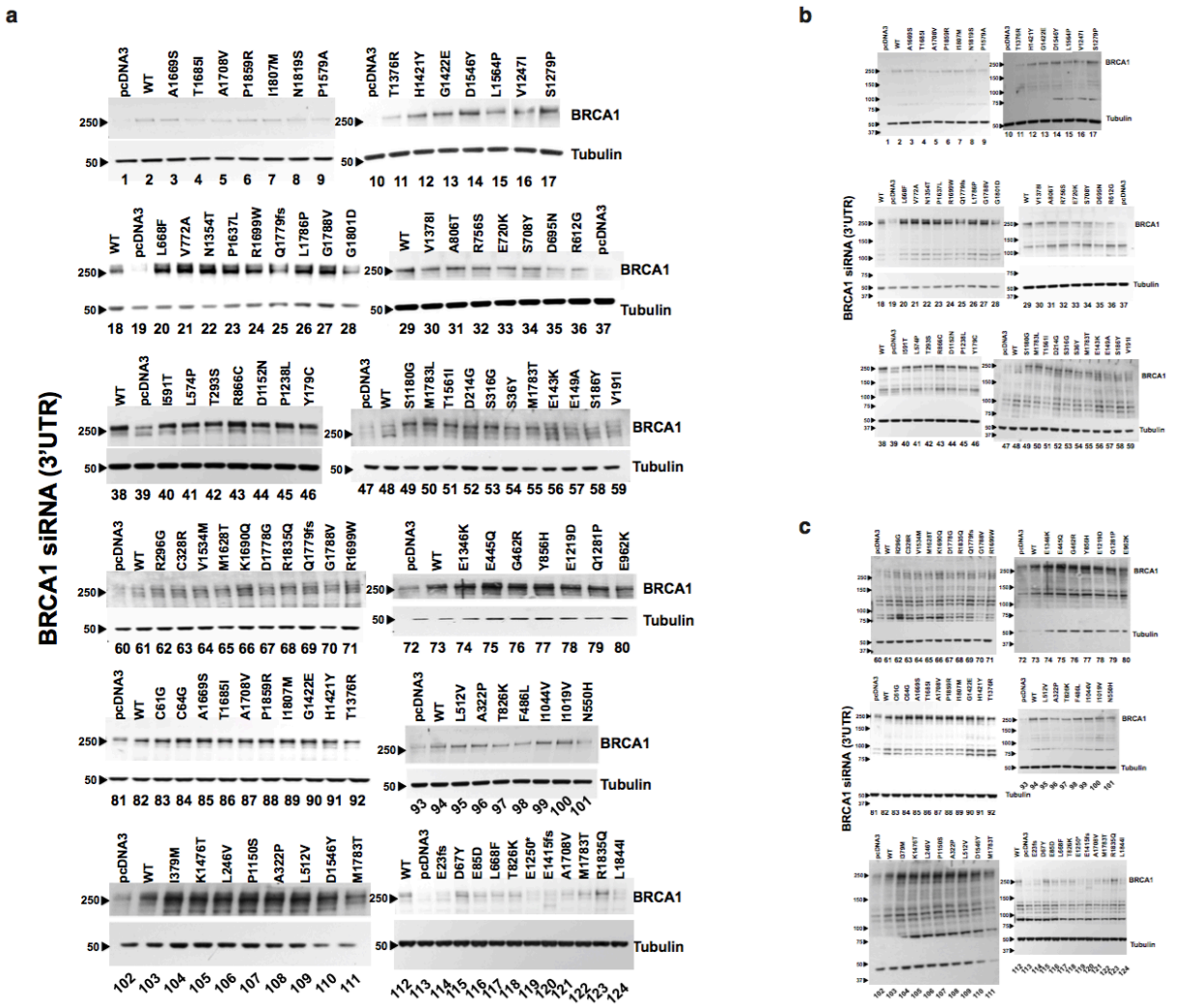
Supplementary Figure 2.2: Minor allele frequency analysis of *BRCA1* and *BRCA2* rare germline truncations. The scatter plot shows the relationship between MAFs of *BRCA1* and *BRCA2* germline truncation variants and site-specific P-values from LOH analysis. For both BRCA1 and BRCA2, the variants with low MAFs tend to display significant LOH. Selecting 0.05% cutoff for rare variants was based on balancing the inclusion of possible false-positives versus the loss of possible true-positives in subsequent burden test and LOH analysis.



Supplementary Figure 2.3: Correlation of truncation frequencies in 32 genes of interest between discovery set and validation set. Using the 3,125 and 1,627 cases in the discovery and validation cohorts, respectively, we used gene-specific tallies to calculate frequencies of truncations in the 32 genes found to be significant by burden testing. Figure shows the discovery and validation frequencies for each gene plotted on the abscissa and ordinate, respectively. These data were regressed using the ordinary least-squares (OLS) calculation. A Pearson's correlation coefficient of 0.6167 was found and the resulting regression line is also plotted (red).

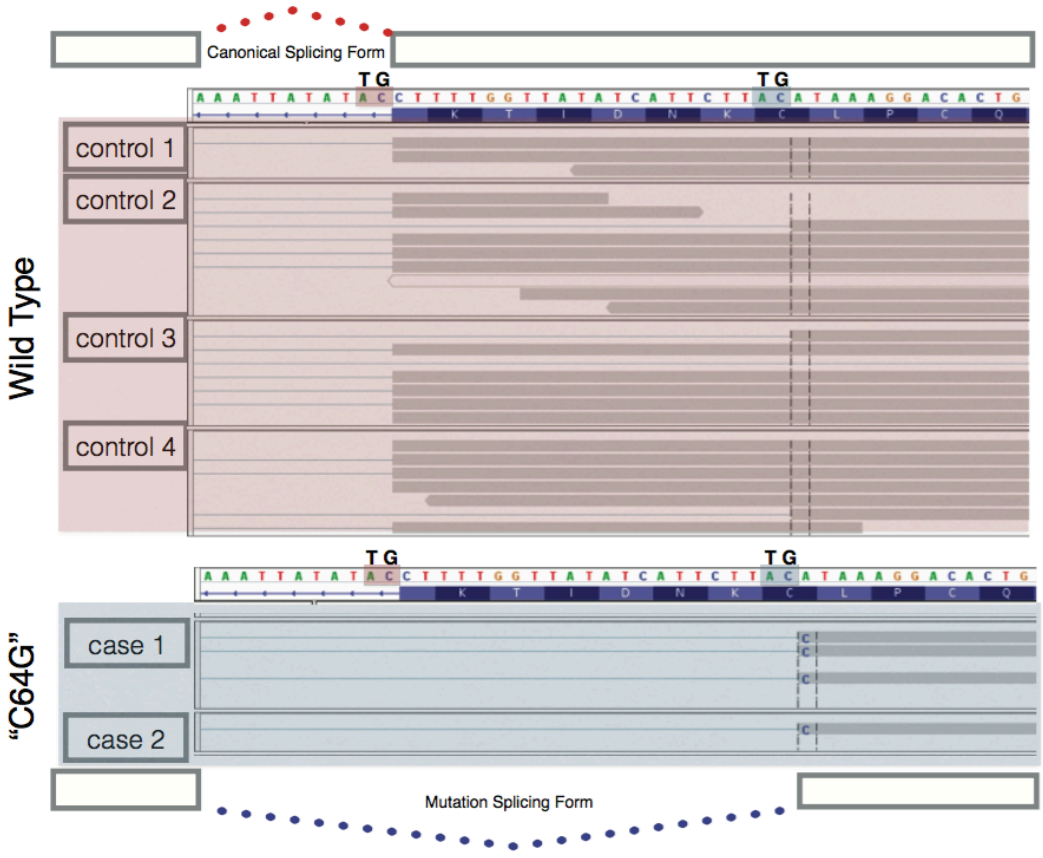


Supplementary Figure 2.4: Mutation rate comparison within BRCA basal subtype. Boxplot shows the mutation rate distribution in basal cases with BRCA1/2 rare germline truncation (n=12) and basal cases without BRCA1/2 rare germline truncation variants (n=101, 2 cases missing somatic mutation information). Graph shows median central line, 50% confidence interval box and 95% confidence interval whiskers. P-value is calculated by Wilcoxon rank-sum test.



Supplementary Figure 2.5: Western blot analysis of expression of BRCA1 mutant constructs. HDR assay results were tested for protein expression of the BRCA1 variant from the transfected plasmid. Cells that remained following flow cytometry analysis were extracted and soluble proteins analyzed by immunoblots. Since multiple experiments were done, each panel should be compared separately from the others. In all samples, the endogenous BRCA1 protein had been depleted by transfection of the siRNA targeting the 3'-UTR of the *BRCA1* mRNA. All blots were probed for BRCA1, and strips show the proteins that migrated at about the 250 kDa marker, and strips were probed for a-tubulin, which migrated near the 50 kDa marker, as a loading control. Samples from the various BRCA1 plasmid transfections were as follows: vector only (lanes 1, 10, 19, 37, 39, 47, 60, 72, 81, 93, 102, 113); WT (lanes 2, 18, 29, 38, 48, 61, 73, 82, 94, 103, 112); E23fs (lane 114); S36Y (lane 54); C61G (lane 83); C64G (lane 84); D67Y (lane 115); E85D (lane 116); E143K (lane 56); E149A (lane 57); Y179C (lane 46); S186Y (lane 58); V191I (lane 59); D214G (lane 52); L246V (lane 106); T293S (lane 42); R296G (lane 62); S316G (lane 53); A322P (lanes 96, 108); C328R (lane 63); I379M (lane 104); E445Q (lane 75); G462R (lane 76); F486L (lane 98); L512V (lanes 95, 109); N550H (lane 101); L574P (lane 41); I591T (lane 40); R612G (lane 36); L668F (lanes 20, 117); D695N (lane 35); S708Y

(lane 34); E720K (lane 33); R756S (lane 32); V772A (lane 21); A806T (lane 31), T826K (lanes 97, 118); Y856H (lane 77); R866C (lane 43); E962K (lane 80), I1019V (lane 100); I1044V (lane 99); P1150S (lane 107); D1152N (lane 44); S1180G (lane 49); E1219D (lane 78); P1238L (lane 45); V1247I (lane 16); E1250* (lane 119); S1279P (lane 17); Q1281P (lane 79); E1346K (lane 74); N1354T (lane 22); T1376R (lanes 11, 92); V1378I (lane 30); E1415fs (lane 120); H1421Y (lanes 12, 91); G1422E (lanes 13, 90); K1476T (lane 105); V1534M (lane 64); D1546Y (lanes 14, 110); T1561I (lane 51); L1564P (lane 15); P1579A (lane 9); M1628T (lane 65); P1637L (lane 23); A1669S (lanes 3, 85); T1685I (lanes 4, 86); K1690Q (lane 66); R1699W (lanes 24, 71); A1708V (lanes 5, 87, 121); D1778G (lane 67); Q1779fs (lanes 25, 69); M1783L (lane 50); M1783T (lanes 55, 111, 122); L1786P (lane 26); G1788V (lanes 27, 70); G1801D (lane 28); I1807M (lanes 7, 89); N1819S (lane 8); R1835Q (lanes 68, 123); L1844I (lane 124); and P1859R (lanes 6, 88).



Supplementary Figure 2.6: Impact of C64G mutation in *BRCA1* on splicing. Integrated genomic viewer screen capture of C64G mutation, leading to the activation of an infrequently used splice site, is shown. Two ovarian cases with C64G mutation and four ovarian cases without this mutation were shown.

Supplementary Table 2.1: Coverage and variant calling stats for discovery and control cohorts.

Coverage Stats

Mini depth	Cancer Type	Target space covered (%)	Mean depth	Stdev depth
10x	BRCA	92	110.8	57.8
	GBM	90.6	143.4	47.5
	HNSC	94.6	92	22.2
	KIRC	91.8	146.6	70.5
	LAML-WXS	90.3	164.9	61
	LGG	94.4	98.3	28.1
	LUAD	93.5	97.1	31
	LUSC	93.6	105.8	41.7
	OV	82.8	146.1	69
	PRAD	94.6	101.1	30.4
	STAD	94.6	102.3	27.7
	UCEC	93.4	105.9	64.9
	WHISP	92.1	106.2	33.3
20x	BRCA	85.7	110	58.1
	GBM	86.1	142.7	47.6
	HNSC	89.5	91.2	22.4
	KIRC	86.6	146	70.7
	LAML-WXS	85.3	164.1	61.2
	LGG	89.4	97.5	28.4
	LUAD	88.1	96.2	31.2
	LUSC	88.3	104.9	41.9
	OV	77.3	145.4	69.2
	PRAD	89.6	100.3	30.6
	STAD	89.6	101.5	27.9
	UCEC	87.3	105.2	65.2
	WHISP	86.3	105.4	33.5
30x	BRCA	78	108.3	59
	GBM	81.8	141.6	47.8
	HNSC	83.9	89.6	22.8
	KIRC	81.5	144.8	71.1
	LAML-WXS	80.8	162.9	61.6
	LGG	84	96.1	28.8
	LUAD	82.2	94.6	31.7
	LUSC	82.6	103.4	42.4
	OV	72.3	144.2	69.6
	PRAD	84.3	98.9	31
	STAD	84.2	100	28.5
	UCEC	78.6	103.1	66.3
	WHISP	80.3	104	34
40x	BRCA	69.6	105.5	60.6
	GBM	77.6	140.1	48.2
	HNSC	77.5	87.2	23.4
	KIRC	76.1	143.1	71.8
	LAML-WXS	76.3	161.2	62.1
	LGG	77.9	93.8	29.5
	LUAD	75.6	92.2	32.4
	LUSC	76.3	101	43.2
	OV	67.6	142.5	70.1
	PRAD	78.3	96.7	31.7
	STAD	78.3	97.8	29.3
	UCEC	68.4	99.7	68.3
	WHISP	73.7	101.8	34.9

Variant calling stats

	WHISP		Pan12		Pan12 Caucasian		Pan12 African American	
Variant type	All Genes	Cancer Genes	All Genes	Cancer Genes	All Genes	Cancer Genes	All Genes	Cancer Genes
Nonsense	6241	219	36009	1251	27462	919	3185	136
Nonstop	353	11	2041	65	1474	50	279	6
Splice site	3529	123	18693	553	14005	442	1878	30
Frame shift indels	6055	266	30508	1242	22034	973	3118	88
In-frame indels	3545	240	20219	965	13687	637	2545	128
Missense	318072	13630	1655391	70919	1157583	50869	207504	7905
Silent	175729	9120	947045	47114	606214	30789	156778	7342
Total Variants	513524	23609	2709906	122109	1842459	84679	375287	15635
Variants/sample	494.25	22.72	671.77	30.27	589.59	27.10	1667.94	69.49

Supplementary Table 2.2: Summary of germline truncation variants identified in 624 cancer associated genes. Rare germline truncation variants (<0.05% MAF in discovery case and control combined) identified in 624 cancer associated genes across 4,034 cancer cases.

The table is too large to display here. So it is hosted by *Nature Communications* website: http://www.nature.com/article-assets/npg/ncomms/2015/151209/ncomms10086/extref/ncomms10086-s3.xlsx

Supplementary Table 2.3: Rare germline truncation variants (<0.05% MAF in case and control combined) identified in 32 genes of interest across 1,627 validation cancer cases.

Chr	Position	Reference	Variant	Gene	Amino acid change	Cancer	Tumor			Normal		
							Ref	Var	VAF	Ref	Var	VAF
7	152357811	A	-	*XRCC2*	p.F32fs	BRCA	51	24	32	39	24	38.1
7	6018315	GA	-	*PMS2*	p.L729fs	BRCA	12	0	0	7	17	70.83
8	144940189	C	-	*EPPK1*	p.L2412fs	GBM	19	15	44.12	22	15	40.54
8	144947107	C	-	*EPPK1*	p.L106fs	LGG	29	37	56.06	29	20	40.82
8	144947370	C	-	*EPPK1*	p.E18fs	LGG	27	23	46	33	19	36.54
8	144946136	TGAG	-	*EPPK1*	p.T428fs	LGG	1	10	90.91	4	18	81.82
8	144941718	C	-	*EPPK1*	p.V1902fs	LGG	29	16	35.56	27	15	35.71
8	144946136	TGAG	-	*EPPK1*	p.T428fs	LUAD	15	5	25	11	10	47.62
7	6018315	GA	-	*PMS2*	p.L729fs	LUAD	25	16	39.02	22	12	35.29
7	6018315	GA	-	*PMS2*	p.L729fs	LUSC	29	56	65.88	16	21	56.76
3	142234279	AA	-	*ATR*	p.F1487fs	LUSC	40	63	61.17	67	38	36.19
7	6026386	TTACC	-	*PMS2*	p.S669fs	PRAD	45	16	26.23	55	29	34.52
7	6018315	GA	-	*PMS2*	p.L729fs	UCEC	44	34	43.59	48	39	44.83
1	26513692	AA	-	*CNKSR1*	p.N190fs	UCEC	16	24	60	21	27	56.25
1	26514778	C	-	*CNKSR1*	p.P247fs	BRCA	39	42	51.85	32	31	49.21
2	48026251	AAGAG	-	*MSH6*	p.R379fs	UCEC	7	6	46.15	20	13	39.39
2	48026890	C	-	*MSH6*	p.P591fs	UCEC	10	21	67.74	16	11	40.74
2	48027686	T	-	*MSH6*	p.I856fs	UCEC	7	14	66.67	28	19	40.43
2	48030640	C	-	*MSH6*	p.F1088fs	UCEC	9	31	77.5	29	14	32.56
5	131915039	GT	-	*RAD50*	p.A134fs	HNSC	34	20	37.04	34	21	38.18
6	33543680	CT	-	*BAK1*	p.E32fs	HNSC	67	34	33.66	61	37	37.76
9	35078692	CAGT	-	*FANCG*	p.T72fs	PRAD	31	11	26.19	34	22	39.29
11	89028498	AG	-	*TYR*	p.E519fs	LUSC	10	12	54.55	20	15	42.86
12	18446857	A	-	*PIK3C2G*	p.I315fs	UCEC	68	50	42.37	55	40	42.11
13	32903605	TG	-	*BRCA2*	p.V220fs	OV	3	13	81.25	17	9	34.62
13	32903605	TG	-	*BRCA2*	p.V220fs	BRCA	20	16	44.44	20	10	33.33
13	32913703	TACT	-	*BRCA2*	p.T1738fs	OV	22	59	72.84	32	24	42.86
13	32914061	GAAAC	-	*BRCA2*	p.E1857fs	OV	18	71	79.78	135	69	33.82
13	32915134	T	-	*BRCA2*	p.Y2215fs	BRCA	6	28	82.35	34	26	43.33
13	32920968	AATA	-	*BRCA2*	p.I2315fs	BRCA	78	57	42.22	169	122	41.92
13	32945138	AG	-	*BRCA2*	p.E2846fs	BRCA	60	45	42.86	23	16	41.03
16	23647268	A	-	*PALB2*	p.L200fs	UCEC	8	35	81.4	22	17	43.59
17	41243480	TTGA	-	*BRCA1*	p.N1355fs	BRCA	36	24	40	49	18	26.87
17	41244778	TAAC	-	*BRCA1*	p.V923fs	PRAD	35	38	52.05	55	37	40.22
17	41276045	CT	-	*BRCA1*	p.E23fs	GBM	70	36	33.96	113	49	30.25
17	48453264	AG	-	*EME1*	p.N233fs	LGG	82	55	40.15	96	52	35.14
17	59761199	A	-	*BRIP1*	p.S1070fs	UCEC	31	24	43.64	25	17	40.48
17	59821794	TT	-	*BRIP1*	p.K752fs	LGG	71	42	37.17	55	53	49.07
13	32912338	TG	-	*BRCA2*	p.V1283fs	LUSC	8	1	11.11	0	0	0
9	100444655	AC	-	*XPA*	p.V244fs	PRAD	75	59	44.03	88	51	36.69
12	18747475	AGTT	-	*PIK3C2G*	p.V1354fs	HNSC	29	11	27.5	19	8	29.63
13	32914438	T	-	*BRCA2*	p.S1982fs	OV	55	82	59.85	42	17	28.81
16	1825947	C	-	*EME2*	p.A375fs	HNSC	16	0	0	29	0	0
12	31247556	-	T	*DDX11*	p.S469fs	UCEC	72	56	43.75	55	38	40.86
13	32954272	-	A	*BRCA2*	p.T3084fs	LUSC	43	42	49.41	31	23	42.59
2	38298106	-	A	*CYP1B1*	p.S464fs	BRCA	40	18	31.03	31	34	52.31
13	32907420	-	A	*BRCA2*	p.I605fs	LUAD	20	15	42.86	15	20	57.14
17	41209079	-	G	*BRCA1*	p.Q247fs	OV	223	602	72.97	125	102	44.93
17	41209079	-	G	*BRCA1*	p.Q247fs	BRCA	24	38	61.29	69	39	36.11

17	41234477	-	T	*BRCA1*	p.S1434fs	LUSC	111	96	46.38	75	52	40.94
17	56774172	-	C	*RAD51C*	p.C176fs	LUAD	61	50	45.05	57	55	49.11
2	48033981	-	TTGA	*MSH6*	p.K1357fs	BRCA	31	12	27.91	28	21	42.86
2	48033981	-	TTGA	*MSH6*	p.K1357fs	LGG	30	12	28.57	44	20	31.25
12	31256907	C	A	*DDX11*	p.C951*	BRCA	24	14	36.84	19	17	47.22
12	31256907	C	A	*DDX11*	p.C951*	BRCA	25	10	28.57	17	9	34.62
12	31256907	C	A	*DDX11*	p.C951*	BRCA	20	10	33.33	14	19	57.58
3	142279156	A	T	*ATR*	p.L497*	HNSC	54	33	37.93	48	55	53.4
12	31256907	C	A	*DDX11*	p.C951*	KIRC	59	68	53.12	66	65	49.62
8	144945558	C	A	*EPPK1*	p.E622*	KIRC	6	5	45.45	14	8	36.36
7	6013096	C	T	*PMS2*	p.W841*	LUAD	10	9	47.37	9	7	43.75
12	31256907	C	A	*DDX11*	p.C951*	PRAD	2	5	71.43	2	10	83.33
7	6026709	G	A	*PMS2*	p.R563*	UCEC	42	58	57.43	49	58	54.21
6	29797323	C	T	*HLA-G*	p.Q255*	UCEC	129	63	32.81	71	56	44.09
7	6026514	G	A	*PMS2*	p.R628*	UCEC	38	45	54.22	34	31	47.69
12	31256907	C	A	*DDX11*	p.C951*	UCEC	24	21	46.67	17	8	32
8	144947235	G	A	*EPPK1*	p.Q63*	UCEC	41	41	49.4	32	16	33.33
12	31256907	C	A	*DDX11*	p.C951*	UCEC	15	7	31.82	10	21	67.74
5	131973889	C	T	*RAD50*	p.R1198*	LGG	40	36	47.37	50	55	52.38
5	131973895	C	T	*RAD50*	p.R1200*	UCEC	14	12	46.15	22	18	45
9	35074486	G	A	*FANCG*	p.R548*	UCEC	44	60	57.69	41	44	51.76
9	100451874	C	A	*XPA*	p.E111*	PRAD	15	15	50	16	13	44.83
13	32913457	C	G	*BRCA2*	p.Y1655*	HNSC	33	34	50.75	21	37	63.79
16	1825377	G	T	*EME2*	p.E255*	BRCA	57	45	44.12	57	40	40.82
16	23614792	G	C	*PALB2*	p.Y1183*	BRCA	33	31	48.44	46	38	45.24
16	23647443	T	A	*PALB2*	p.K142*	BRCA	43	72	62.61	56	54	49.09
16	23649415	G	A	*PALB2*	p.Q23*	LGG	71	55	43.65	63	73	53.68
17	33428320	C	T	*RAD51D*	p.W288*	LGG	79	64	44.76	106	71	39.44
17	33433425	G	A	*RAD51D*	p.R206*	OV	14	160	91.95	5	8	61.54
17	41246494	C	A	*BRCA1*	p.E352*	BRCA	13	52	80	50	32	39.02
17	56772543	C	T	*RAD51C*	p.Q133*	UCEC	4	9	69.23	4	7	63.64
17	56787223	C	T	*RAD51C*	p.R237*	PRAD	94	47	33.33	91	58	38.93
1	26507075	G	A	*CNKSR1*	p.W55*	UCEC	22	65	74.71	9	8	47.06
5	131930642	C	G	*RAD50*	p.Y625*	GBM	151	127	45.68	88	105	54.4
6	33541983	T	A	*BAK1*	p.R127*	BRCA	16	3	15.79	10	14	58.33
6	33541983	T	A	*BAK1*	p.R127*	UCEC	10	6	37.5	15	8	34.78
11	108175528	C	T	*ATM*	p.R1875*	GBM	21	20	48.78	15	21	58.33
11	108183151	G	T	*ATM*	p.E1978*	BRCA	11	41	78.85	30	28	48.28
11	108183151	G	T	*ATM*	p.E1978*	BRCA	5	21	80.77	21	22	51.16
14	45658326	C	T	*FANCM*	p.Q1701*	BRCA	3	1	25	24	19	44.19
12	31255955	G	A	*DDX11*	e23+1	BRCA	48	32	40	44	43	49.43
7	6043425	T	A	*PMS2*	e4-2	LUSC	22	24	52.17	26	17	39.53
7	6022454	C	T	*PMS2*	e12+1	UCEC	0	11	100	21	17	44.74
16	23641791	C	G	*PALB2*	e5-1	OV	0	3	100	2	5	71.43
16	89813299	C	T	*FANCA*	e34-1	BRCA	6	6	50	25	20	44.44
16	89881023	T	A	*FANCA*	e3-2	LUAD	34	25	42.37	33	37	52.86
17	48458123	G	A	*EME1*	e8-1	BRCA	14	49	77.78	23	30	56.6
2	48033791	-	TAAC	*MSH6*	e9+1	LUAD	38	19	33.33	26	14	35

Supplementary Table 2.4: 69 germline truncations validated using whole genome sequencing data.

The file is too large to display here, it is hosted by *Nature Communications* website:
http://www.nature.com/article-assets/npg/ncomms/2015/151209/ncomms10086/extref/ncomms10086-s5.xlsx

Supplementary Table 2.5: Cancer associated gene lists used in this study, including 624 cancer associated genes, 114 cancer susceptibility genes reported in Rahman et al., 47 genes from Fanconi Anemia pathway.

624 Cancer Associated Genes

CEP76	*DPYD*	*FANCI*	*GSK3B*	*KIT*	*MXRA5*	*PHOX2B*	*RBM10*	*SLX4*	*TLR4*
CERS2	*ECSCR*	*FANCL*	*GSTP1*	*KLF4*	*MYB*	*PIK3C2G*	*RBMX*	*SMAD2*	*TMEM127*
CHD4	*EGFR*	*FANCM*	*GUCY1A2*	*KLHL6*	*MYC*	*PIK3C3*	*RECQL4*	*SMAD3*	*TMPRSS2*
CHD8	*EGR3*	*FAS*	*H3F3A*	*KMT2A*	*MYCL*	*PIK3CA*	*REL*	*SMAD4*	*TNF*
CHEK1	*EIF2S2*	*FAT1*	*H3F3C*	*KMT2B*	*MYCN*	*PIK3CG*	*RET*	*SMARCA4*	*TNFAIP3*
CHEK2	*EIF3A*	*FAT3*	*HAUS3*	*KMT2C*	*MYD88*	*PIK3R1*	*REV3L*	*SMARCB1*	*TNFRSF14*
CHUK	*EIF4A2*	*FBXW7*	*HDAC4*	*KMT2D*	*MYLK*	*PIK3R2*	*RHBDF2*	*SMARCD1*	*TOP1*
CIC	*ELANE*	*FCGR1A*	*HES1*	*KRAS*	*NAV3*	*PLCG2*	*RHEB*	*SMARCE1*	*TOP3A*
CNBD1	*ELF3*	*FCGR2A*	*HFE*	*LIFR*	*NBN*	*PML*	*RHOA*	*SMC1A*	*TOP3B*
CNKSR1	*EME1*	*FCGR3A*	*HGF*	*LMO1*	*NBPF1*	*PMS2*	*RICTOR*	*SMC3*	*TP53*
COL7A1	*EME2*	*FGF10*	*HIF1A*	*LRP1B*	*NCOR1*	*PMS2CL*	*RIT1*	*SMO*	*TP53BP1*
COMT	*EML4*	*FGF12*	*HIST1H1C*	*LRP2*	*NEIL1*	*PNRC1*	*RMI1*	*SNX25*	*TPMT*
CRBN	*EP300*	*FGF14*	*HIST1H1E*	*LRRK2*	*NF1*	*POLD1*	*RMI2*	*SOCS1*	*TPX2*
CREBBP	*EPHA2*	*FGF19*	*HIST1H2BD*	*MALAT1*	*NF2*	*POLE*	*RMRP*	*SOD2*	*TRAF3*
CRIPAK	*EPHA3*	*FGF23*	*HIST1H3B*	*MAN1B1*	*NFE2L2*	*POLH*	*RNF43*	*SOS1*	*TRAF7*
CRKL	*EPHA5*	*FGF3*	*HIST1H4E*	*MAP2K1*	*NFKBIA*	*POLI*	*ROS1*	*SOX10*	*TRIM37*
CRLF2	*EPHB1*	*FGF4*	*HLA-A*	*MAP2K2*	*NKX2-1*	*POLK*	*RPA1*	*SOX17*	*TRRAP*
CSF1R	*EPHB2*	*FGF6*	*HLA-B*	*MAP2K4*	*NOTCH1*	*POLQ*	*RPA2*	*SOX2*	*TSC1*
CTCF	*EPHB6*	*FGF7*	*HLA-G*	*MAP3K1*	*NOTCH2*	*PORCN*	*RPA4*	*SOX9*	*TSC2*
CTNNA1	*EPPK1*	*FGFBP1*	*HMBS*	*MAP3K13*	*NOTCH3*	*POU2AF1*	*RPL22*	*SPEN*	*TSHR*
CTNNB1	*ERBB2*	*FGFR1*	*HNF1A*	*MAP3K15*	*NOTCH4*	*POU2F2*	*RPL5*	*SPOP*	*TSHZ2*
CUL4A	*ERBB3*	*FGFR2*	*HRAS*	*MAP4K1*	*NPM1*	*PPM1D*	*RPS14*	*SPRY4*	*TSHZ3*
CUL4B	*ERBB4*	*FGFR3*	*HSP90AB1*	*MAP4K3*	*NQO1*	*PPP2R1A*	*RPS15*	*SRC*	*TYMS*
CUX1	*ERCC1*	*FGFR4*	*IDH1*	*MAPK1*	*NRAS*	*PPP6C*	*RPS2*	*SRSF2*	*TYR*
CYLD	*ERCC2*	*FH*	*IDH2*	*MAPK8IP1*	*NRP2*	*PRDM1*	*RPTOR*	*SRY*	*U2AF1*
CYP17A1	*ERCC3*	*FLCN*	*IGF1*	*MAX*	*NSD1*	*PRKAR1A*	*RUNX1*	*STAG2*	*U2AF2*
CYP1B1	*ERCC4*	*FLT1*	*IGF1R*	*MBD1*	*NTN4*	*PRKDC*	*RUNX1T1*	*STAT3*	*UGT1A1*
CYP2C19	*ERCC5*	*FLT3*	*IGF2*	*MC1R*	*NTRK1*	*PRLR*	*RUNX3*	*STAT4*	*UMPS*
CYP2C8	*ERG*	*FLT4*	*IKBKE*	*MCL1*	*NTRK2*	*PRPF40B*	*RXRA*	*STK11*	*UROD*
CYP2D6	*ESR1*	*FOXA1*	*IKZF1*	*MDM2*	*NTRK3*	*PRSS1*	*SBDS*	*STK19*	*USP1*
CYP3A4	*ESR2*	*FOXA2*	*IL7R*	*MDM4*	*NUP93*	*PRSS8*	*SDHA*	*STK38*	*USP9X*
CYP3A5	*ETV1*	*FOXL2*	*ING1*	*MECOM*	*ODAM*	*PTCH1*	*SDHAF2*	*STX2*	*VANGL2*
DAXX	*ETV4*	*FOXQ1*	*INHA*	*MED12*	*OTUD7A*	*PTEN*	*SDHB*	*SUFU*	*VEZF1*
DCAF6	*ETV5*	*FUBP1*	*INHBA*	*MED23*	*PAK3*	*PTPN11*	*SDHC*	*SULT1A1*	*VHL*
DDB2	*ETV6*	*FZD1*	*INPPL1*	*MEF2A*	*PAK7*	*PTPRC*	*SDHD*	*SUZ12*	*WAC*
DDR1	*EWSR1*	*GAB2*	*IPO7*	*MEF2B*	*PALB2*	*PTPRD*	*SERPINA1*	*SYK*	*WAS*
DDR2	*EXO1*	*GATA1*	*IRF4*	*MEN1*	*PAPD5*	*QKI*	*SERPINB13*	*SZRD1*	*WASF3*
DDX11	*EXT1*	*GATA2*	*IRS2*	*MET*	*PARP1*	*RAB40A*	*SETBP1*	*TAF1*	*WDR48*
DDX3X	*EXT2*	*GATA3*	*ITGAV*	*MGA*	*PARP2*	*RAC1*	*SETD2*	*TBC1D12*	*WISP3*

DDX5	*EZH1*	*GBA*	*ITK*	*MIR142*	*PARP3*	*RAD21*	*SETDB1*	*TBL1XR1*	*WNK1*
DIAPH1	*EZH2*	*GID4*	*ITPA*	*MITF*	*PARP4*	*RAD50*	*SF1*	*TBX3*	*WRN*
DICER1	*EZR*	*GJB2*	*JAK1*	*MLH1*	*PAX5*	*RAD51*	*SF3B1*	*TCEB1*	*WT1*
DIDO1	*FAH*	*GNA11*	*JAK2*	*MNDA*	*PBRM1*	*RAD51B*	*SGK1*	*TCF7L2*	*XPA*
DIS3	*FAM129B*	*GNA13*	*JAK3*	*MORC4*	*PCBP1*	*RAD51C*	*SH2B3*	*TELO2*	*XPC*
DIS3L2	*FAM46C*	*GNAQ*	*JUN*	*MPL*	*PCDH10*	*RAD51D*	*SH2D1A*	*TERT*	*XPO1*
DKC1	*FANCA*	*GNAS*	*KAT6A*	*MRE11A*	*PDAP1*	*RAD52*	*SIN3A*	*TET2*	*XRCC2*
DLC1	*FANCB*	*GNB1*	*KDM5A*	*MSH2*	*PDCD2L*	*RAD54L*	*SIRPA*	*TFG*	*XRCC3*
DNER	*FANCC*	*GPC3*	*KDM5C*	*MSH6*	*PDGFRA*	*RAF1*	*SIRT4*	*TGFBR1*	*ZFHX3*
DNMT1	*FANCD2*	*GPR124*	*KDM6A*	*MTAP*	*PDGFRB*	*RALY*	*SLC19A1*	*TGFBR2*	*ZNF217*
DNMT3A	*FANCE*	*GPS2*	*KDR*	*MTHFR*	*PDK1*	*RARA*	*SLC22A2*	*TIMM17A*	*ZNF703*
DOCK8	*FANCF*	*GRIN2A*	*KEAP1*	*MTOR*	*PDSS2*	*RASA1*	*SLC25A13*	*TIPARP*	*ZRANB3*
DOT1L	*FANCG*	*GRM3*	*KIF5B*	*MUTYH*	*PHF6*	*RB1*	*SLCO1B3*	*TLK2*	*ZRSR2*

47 genes in Fanconi Anemia Pathway

APITD1	*BRCA1*	*ERCC1*	*FANCB*	*FANCG*	*MLH1*	*RAD51*	*RPA1*	*SMC3*	*WDR48*
ATM	*BRCA2*	*ERCC4*	*FANCC*	*FANCI*	*PALB2*	*RAD51C*	*RPA2*	*TELO2*	*XPC*
ATR	*BRIP1*	*ERCC5*	*FANCD2*	*FANCL*	*PMS2CL*	*REV3L*	*RPA4*	*TOP3A*	
ATRIP	*CHEK2*	*EXO1*	*FANCE*	*FANCM*	*POLI*	*RMI1*	*SLX4*	*TOP3B*	
BLM	*EME1*	*FANCA*	*FANCF*	*HES1*	*POLK*	*RMI2*	*SMC1A*	*USP1*	

114 Rahman Genes

ABCB11	*CBL*	*DIS3L2*	*FANCA*	*ITK*	*NF2*	*PTPN11*	*SDHB*	*STAT3*	*VHL*
ALK	*CDC73*	*DKC1*	*FANCC*	*KIT*	*PALB2*	*RAD51C*	*SDHC*	*STK11*	*WAS*
APC	*CDH1*	*DOCK8*	*FANCG*	*MAX*	*PDGFRA*	*RAD51D*	*SDHD*	*SUFU*	*WRN*
ATM	*CDK4*	*EGFR*	*FH*	*MEN1*	*PHOX2B*	*RB1*	*SERPINA1*	*TERT*	*WT1*
AXIN2	*CDKN1B*	*ELANE*	*FLCN*	*MET*	*PMS2*	*RECQL4*	*SH2D1A*	*TGFBR1*	*XPA*
BAP1	*CDKN2A*	*ERCC2*	*GATA2*	*MLH1*	*POLD1*	*RET*	*SLC25A13*	*TMEM127*	*XPC*
BLM	*CEBPA*	*ERCC3*	*GBA*	*MSH2*	*POLE*	*RHBDF2*	*SMAD4*	*TNFRSF6 (FAS)*	
BMPR1A	*CHEK2*	*ERCC4*	*GJB2*	*MSH6*	*POLH*	*RMRP*	*SMARCA4*	*TP53*	
BRCA1	*COL7A1*	*ERCC5*	*GPC3*	*MTAP*	*PRKAR1A*	*RUNX1*	*SMARCB1*	*TRIM37*	
BRCA2	*CYLD*	*EXT1*	*HFE*	*MUTYH*	*PRSS1*	*SBDS*	*SMARCE1*	*TSC1*	
BRIP1	*DDB2*	*EXT2*	*HMBS*	*NBN*	*PTCH1*	*SDHA*	*SOS1*	*TSC2*	
BUB1B	*DICER1*	*FAH*	*HRAS*	*NF1*	*PTEN*	*SDHAF2*	*SRY*	*UROD*	

Supplementary Table 2.6: Frequencies of rare truncation variants in 3 gene lists across 12 cancer types.

Fanconi Anemia Pathway (47 Genes)	Rare truncation variants	Sample Size (n)	Truncation variant/sample
OV	73	429	0.1702
BRCA	65	770	0.0844
PRAD	14	178	0.0787
STAD	25	321	0.0779
LUSC	11	193	0.057
HNSC	16	291	0.055
GBM	13	267	0.0487
LAML	8	200	0.04
LUAD	16	462	0.0346
UCEC	8	248	0.0323
KIRC	13	452	0.0288
LGG	5	223	0.0224
Rahman et al. 2014 (114 Genes)	**Rare truncation variants**	**Sample Size (n)**	**Truncation variant/sample**
OV	81	429	0.1888
STAD	34	321	0.1059
BRCA	69	770	0.0896
PRAD	15	178	0.0843
LUSC	16	193	0.0829
LGG	17	223	0.0762
HNSC	22	291	0.0756
UCEC	18	248	0.0726
LUAD	32	462	0.0693
KIRC	21	452	0.0465
GBM	12	267	0.0449
LAML	8	200	0.04
Cancer Associated Genes (624 genes)	**Rare truncation variants**	**Sample Size (n)**	**Truncation variant/sample**
OV	114	429	0.2657
BRCA	163	770	0.2117
STAD	67	321	0.2087
LUSC	35	193	0.1813
LAML	36	200	0.18
LGG	38	223	0.1704
PRAD	30	178	0.1685
UCEC	39	248	0.1573
HNSC	44	291	0.1512
LUAD	68	462	0.1472
GBM	39	267	0.1461
KIRC	55	452	0.1217

Supplementary Table 2.7: Burden analysis results for Pan-Cancer discovery cohort using rare truncation variants from 624 cancer associated genes in 3,125 Caucasian samples. The control cohort is 1,039 WHI Caucasian cases.

PAN12

# RANK	GENE	PVALCAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*BRCA1*	0.0000	0.0000	0.0000	0	53	2078	6250
2	*BRCA2*	0.0000	0.0000	0.0000	3	50	2078	6250
3	*ATM*	0.0008	0.0003	0.0126	2	23	2078	6250
4	*BRIP1*	0.0016	0.0015	0.0526	1	16	2078	6250
5	*PALB2*	0.0003	0.0016	0.0526	0	12	2078	6250
6	*MSH6*	0.0005	0.0028	0.0666	0	11	2078	6250
7	*XRCC2*	0.0041	0.0237	0.4844	0	7	2078	6250
8	*PIK3C2G*	0.1437	0.0295	0.5278	5	19	2078	6250
9	*NBN*	0.0404	0.0385	0.6113	1	9	2078	6250
10	*RAD51C*	0.0071	0.0405	0.6113	0	6	2078	6250
11	*CYP1B1*	0.0071	0.0405	0.6113	0	6	2078	6250
12	*DIS3*	0.0896	0.0471	0.6113	2	11	2078	6250
13	*TYR*	0.0126	0.0691	0.7597	0	5	2078	6250
14	*CNKSR1*	0.1237	0.0695	0.7597	2	10	2078	6250
15	*ERCC2*	0.1237	0.0695	0.7597	2	10	2078	6250
16	*EPPK1*	0.1236	0.0695	0.7597	2	10	2078	6250
17	*FANCM*	0.2711	0.1034	0.8699	4	13	2078	6250
18	*RAD51D*	0.0227	0.1179	0.9364	0	4	2078	6250
19	*MRE11A*	0.0227	0.1179	0.9364	0	4	2078	6250
20	*ODAM*	0.0227	0.1179	0.9364	0	4	2078	6250
21	*SLX4*	0.0227	0.1179	0.9364	0	4	2078	6250
22	*RAD51B*	0.0227	0.1179	0.9364	0	4	2078	6250
23	*BAK1*	0.0227	0.1179	0.9364	0	4	2078	6250
24	*DDX11*	0.3275	0.1397	0.9364	4	12	2078	6250
25	*FANCI*	0.1409	0.1410	0.9364	1	6	2078	6250
26	*ABCC4*	0.1409	0.1410	0.9364	1	6	2078	6250
27	*FANCG*	0.1409	0.1410	0.9364	1	6	2078	6250
28	*RAD50*	0.2243	0.1457	0.9364	2	8	2078	6250
29	*SLC19A1*	0.0416	0.2012	0.9919	0	3	2078	6250
30	*SGK1*	0.0416	0.2012	0.9919	0	3	2078	6250
31	*POLI*	0.0416	0.2012	0.9919	0	3	2078	6250
32	*HLA-G*	0.0416	0.2012	0.9919	0	3	2078	6250
33	*BCR*	0.0416	0.2012	0.9919	0	3	2078	6250
34	*TRAF7*	0.0416	0.2012	0.9919	0	3	2078	6250
35	*TP53BP1*	0.0416	0.2012	0.9919	0	3	2078	6250
36	*RAD52*	0.0416	0.2012	0.9919	0	3	2078	6250
37	*POLH*	0.0416	0.2012	0.9919	0	3	2078	6250
38	*KRAS*	0.0416	0.2012	0.9919	0	3	2078	6250
39	*IRF4*	0.0416	0.2012	0.9919	0	3	2078	6250
40	*IL7R*	0.0416	0.2012	0.9919	0	3	2078	6250
41	*IDH1*	0.0416	0.2012	0.9919	0	3	2078	6250
42	*GPS2*	0.0416	0.2012	0.9919	0	3	2078	6250
43	*FAH*	0.0416	0.2012	0.9919	0	3	2078	6250
44	*EXT2*	0.0416	0.2012	0.9919	0	3	2078	6250
45	*ERCC1*	0.0416	0.2012	0.9919	0	3	2078	6250
46	*EME2*	0.0416	0.2012	0.9919	0	3	2078	6250
47	*EME1*	0.0416	0.2012	0.9919	0	3	2078	6250
48	*DIS3L2*	0.0416	0.2012	0.9919	0	3	2078	6250
49	*FANCC*	0.2937	0.2063	0.9919	2	7	2078	6250
50	*SLC25A13*	0.2067	0.2120	0.9919	1	5	2078	6250
51	*ESR2*	0.2067	0.2120	0.9919	1	5	2078	6250

52	*ATR*	0.2067	0.2120	0.9919	1	5	2078	6250
53	*ABL2*	0.2067	0.2120	0.9919	1	5	2078	6250
54	*XPA*	0.2067	0.2120	0.9919	1	5	2078	6250
55	*NRP2*	0.2067	0.2120	0.9919	1	5	2078	6250
56	*FLT3*	0.2067	0.2120	0.9919	1	5	2078	6250
57	*FANCA*	0.4235	0.2640	0.9919	3	8	2078	6250
58	*GJB2*	0.4235	0.2640	0.9919	3	8	2078	6250
59	*PARP2*	0.2961	0.3131	0.9919	1	4	2078	6250
60	*TET2*	0.2961	0.3131	0.9919	1	4	2078	6250
61	*BUB1B*	0.2961	0.3131	0.9919	1	4	2078	6250
62	*WDR48*	0.0786	0.3433	0.9919	0	2	2078	6250
63	*USP9X*	0.0786	0.3433	0.9919	0	2	2078	6250
64	*UROD*	0.0786	0.3433	0.9919	0	2	2078	6250
65	*TPX2*	0.0786	0.3433	0.9919	0	2	2078	6250
66	*SUZ12*	0.0786	0.3433	0.9919	0	2	2078	6250
67	*RUNX1*	0.0786	0.3433	0.9919	0	2	2078	6250
68	*PDCD2L*	0.0786	0.3433	0.9919	0	2	2078	6250
69	*NQO1*	0.0786	0.3433	0.9919	0	2	2078	6250
70	*MYB*	0.0786	0.3433	0.9919	0	2	2078	6250
71	*MET*	0.0786	0.3433	0.9919	0	2	2078	6250
72	*KAT6A*	0.0786	0.3433	0.9919	0	2	2078	6250
73	*HIST1H2BD*	0.0786	0.3433	0.9919	0	2	2078	6250
74	*HIST1H1E*	0.0786	0.3433	0.9919	0	2	2078	6250
75	*HGF*	0.0786	0.3433	0.9919	0	2	2078	6250
76	*HFE*	0.0786	0.3433	0.9919	0	2	2078	6250
77	*FLT1*	0.0786	0.3433	0.9919	0	2	2078	6250
78	*FLCN*	0.0786	0.3433	0.9919	0	2	2078	6250
79	*FANCF*	0.0786	0.3433	0.9919	0	2	2078	6250
80	*ETV4*	0.0786	0.3433	0.9919	0	2	2078	6250
81	*EPHB2*	0.0786	0.3433	0.9919	0	2	2078	6250
82	*EPHA5*	0.0786	0.3433	0.9919	0	2	2078	6250
83	*DAXX*	0.0786	0.3433	0.9919	0	2	2078	6250
84	*CYP2C19*	0.0786	0.3433	0.9919	0	2	2078	6250
85	*CDKN2A*	0.0786	0.3433	0.9919	0	2	2078	6250
86	*BARD1*	0.0786	0.3433	0.9919	0	2	2078	6250
87	*BAP1*	0.0786	0.3433	0.9919	0	2	2078	6250
88	*AZGP1*	0.0786	0.3433	0.9919	0	2	2078	6250
89	*APOL2*	0.0786	0.3433	0.9919	0	2	2078	6250
90	*APITD1*	0.0786	0.3433	0.9919	0	2	2078	6250
91	*AKT3*	0.0786	0.3433	0.9919	0	2	2078	6250
92	*ACO1*	0.0786	0.3433	0.9919	0	2	2078	6250
93	*PARP3*	0.5037	0.3477	0.9919	3	7	2078	6250
94	*RAD54L*	0.4698	0.3897	0.9919	2	5	2078	6250
95	*PMS2*	0.4698	0.3897	0.9919	2	5	2078	6250
96	*MAP3K15*	0.6003	0.3988	0.9919	4	8	2078	6250
97	*BLM*	0.7338	0.4376	0.9919	7	12	2078	6250
98	*TELO2*	0.4116	0.4510	0.9919	1	3	2078	6250
99	*MNDA*	0.4116	0.4510	0.9919	1	3	2078	6250
100	*MED23*	0.4116	0.4510	0.9919	1	3	2078	6250
101	*HIST1H4E*	0.4116	0.4510	0.9919	1	3	2078	6250
102	*GPR124*	0.4116	0.4510	0.9919	1	3	2078	6250
103	*ERBB3*	0.4116	0.4510	0.9919	1	3	2078	6250
104	*DNMT3A*	0.4116	0.4510	0.9919	1	3	2078	6250
105	*CRLF2*	0.4116	0.4510	0.9919	1	3	2078	6250
106	*COMT*	0.4116	0.4510	0.9919	1	3	2078	6250
107	*ALK*	0.4116	0.4510	0.9919	1	3	2078	6250
108	*COL7A1*	0.6715	0.4940	0.9919	4	7	2078	6250
109	*ABCC2*	0.6716	0.4940	0.9919	4	7	2078	6250
110	*HNF1A*	0.5713	0.5152	0.9919	2	4	2078	6250

111	*MUTYH*	0.5713	0.5152	0.9919	2	4	2078	6250
112	*POLE*	0.5713	0.5152	0.9919	2	4	2078	6250
113	*XPC*	0.5713	0.5152	0.9919	2	4	2078	6250
114	*WRN*	0.6703	0.5613	0.9919	3	5	2078	6250
115	*CHEK2*	0.7393	0.5974	0.9919	4	6	2078	6250
116	*APC*	0.5506	0.6276	0.9919	1	2	2078	6250
117	*ATRIP*	0.5506	0.6276	0.9919	1	2	2078	6250
118	*DOCK8*	0.5506	0.6276	0.9919	1	2	2078	6250
119	*ERCC4*	0.5506	0.6276	0.9919	1	2	2078	6250
120	*ITGAV*	0.5506	0.6276	0.9919	1	2	2078	6250
121	*NEIL1*	0.5506	0.6276	0.9919	1	2	2078	6250
122	*PRKDC*	0.5506	0.6276	0.9919	1	2	2078	6250
123	*RECQL4*	0.5506	0.6276	0.9919	1	2	2078	6250
124	*SBDS*	0.5506	0.6276	0.9919	1	2	2078	6250
125	*SDHA*	0.6747	0.6580	0.9919	2	3	2078	6250
126	*CASP8*	0.6748	0.6580	0.9919	2	3	2078	6250
127	*TOP3A*	0.6748	0.6580	0.9919	2	3	2078	6250
128	*PARP4*	0.7493	0.6825	0.9919	3	4	2078	6250
129	*ERCC3*	0.8920	0.7524	0.9919	7	8	2078	6250
130	*EXO1*	0.8199	0.8008	0.9919	3	3	2078	6250
131	*AURKB*	0.8200	0.8008	0.9919	3	3	2078	6250
132	*FGFR4*	0.8200	0.8008	0.9919	3	3	2078	6250
133	*FAT1*	0.8200	0.8008	0.9919	3	3	2078	6250
134	*ROS1*	0.8200	0.8008	0.9919	3	3	2078	6250
135	*CBLC*	0.7726	0.8042	0.9919	2	2	2078	6250
136	*HAUS3*	0.7726	0.8042	0.9919	2	2	2078	6250
137	*SERPINA1*	0.7726	0.8042	0.9919	2	2	2078	6250
138	*FANCD2*	0.8814	0.8093	0.9919	5	5	2078	6250
139	*POLQ*	0.9839	0.8555	0.9919	18	19	2078	6250
140	*POLK*	0.8993	0.8895	0.9919	4	3	2078	6250
141	*FANCL*	0.8994	0.8895	0.9919	4	3	2078	6250
142	*ZRANB3*	0.9677	0.9693	0.9919	6	3	2078	6250
143	*MC1R*	0.9874	0.9747	0.9919	10	6	2078	6250

BRCA

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*BRCA2*	0.0000	0.0000	0.0000	3	19	2078	1156
2	*BRCA1*	0.0003	0.0000	0.0000	0	12	2078	1156
3	*FANCM*	0.0231	0.0158	0.2048	4	8	2078	1156
4	*MSH6*	0.0414	0.0341	0.3327	0	3	2078	1156
5	*BAK1*	0.0414	0.0341	0.3327	0	3	2078	1156
6	*ATR*	0.0449	0.0410	0.3327	1	4	2078	1156
7	*ATM*	0.0792	0.0915	0.5098	2	4	2078	1156
8	*ESR2*	0.0832	0.1033	0.5098	1	3	2078	1156
9	*CRLF2*	0.0832	0.1033	0.5098	1	3	2078	1156
10	*TET2*	0.0831	0.1033	0.5098	1	3	2078	1156
11	*BRIP1*	0.0831	0.1033	0.5098	1	3	2078	1156
12	*XRCC2*	0.0785	0.1052	0.5098	0	2	2078	1156
13	*TYR*	0.0785	0.1052	0.5098	0	2	2078	1156
14	*RAD52*	0.0785	0.1052	0.5098	0	2	2078	1156
15	*RAD51B*	0.0785	0.1052	0.5098	0	2	2078	1156
16	*IRF4*	0.0785	0.1052	0.5098	0	2	2078	1156
17	*FLT1*	0.0785	0.1052	0.5098	0	2	2078	1156
18	*AKT3*	0.0785	0.1052	0.5098	0	2	2078	1156
19	*PIK3C2G*	0.0985	0.1090	0.5098	5	6	2078	1156
20	*EPPK1*	0.1437	0.1968	0.5098	2	3	2078	1156
21	*CHEK2*	0.1881	0.2405	0.5098	4	4	2078	1156
22	*ABCC2*	0.1879	0.2405	0.5098	4	4	2078	1156
23	*SBDS*	0.1595	0.2474	0.5098	1	2	2078	1156

24	*MNDA*	0.1595	0.2474	0.5098	1	2	2078	1156
25	*DNMT3A*	0.1595	0.2474	0.5098	1	2	2078	1156
26	*RAD54L*	0.2658	0.3914	0.5872	2	2	2078	1156
27	*ERCC2*	0.2658	0.3914	0.5872	2	2	2078	1156
28	*CNKSR1*	0.2658	0.3914	0.5872	2	2	2078	1156
29	*DIS3*	0.2658	0.3914	0.5872	2	2	2078	1156
30	*CASP8*	0.2658	0.3914	0.5872	2	2	2078	1156
31	*AURKB*	0.3851	0.5212	0.6558	3	2	2078	1156
32	*MAP3K15*	0.5044	0.6309	0.7689	4	2	2078	1156
33	*DDX11*	0.5044	0.6309	0.7689	4	2	2078	1156
34	*FANCD2*	0.6140	0.7197	0.8256	5	2	2078	1156
35	*ZRANB3*	0.7082	0.7898	0.8801	6	2	2078	1156
36	*BLM*	0.7850	0.8439	0.9142	7	2	2078	1156
37	*ERCC3*	0.7850	0.8439	0.9142	7	2	2078	1156
38	*MC1R*	0.9237	0.9388	0.9635	10	2	2078	1156
39	*POLQ*	0.9865	0.9842	0.9842	18	3	2078	1156

GBM

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*RAD50*	0.0584	0.0118	0.0588	2	3	2078	474
2	*MSH6*	0.0782	0.0125	0.0588	0	2	2078	474
3	*CYP1B1*	0.0782	0.0125	0.0588	0	2	2078	474
4	*APITD1*	0.0782	0.0125	0.0588	0	2	2078	474
5	*BLM*	0.2380	0.2665	0.2665	7	2	2078	474

HNSC

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*PALB2*	0.0782	0.0159	0.1116	0	2	2078	446
2	*FANCA*	0.0732	0.0299	0.1116	3	3	2078	446
3	*FANCM*	0.0870	0.0475	0.1116	4	3	2078	446
4	*PMS2*	0.1180	0.0803	0.1405	2	2	2078	446
5	*FANCC*	0.1180	0.0803	0.1405	2	2	2078	446
6	*CNKSR1*	0.1178	0.0803	0.1405	2	2	2078	446
7	*POLQ*	0.1668	0.1562	0.1562	18	5	2078	446

KIRC

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*EME1*	0.0784	0.0174	0.1395	0	2	2078	838
2	*BCR*	0.0784	0.0174	0.1395	0	2	2078	838
3	*BAP1*	0.0784	0.0174	0.1395	0	2	2078	838
4	*XPC*	0.0663	0.0187	0.1395	2	3	2078	838
5	*PARP3*	0.0823	0.0338	0.1395	3	3	2078	838
6	*ERCC2*	0.1284	0.0871	0.1395	2	2	2078	838
7	*FANCM*	0.1943	0.1822	0.2082	4	2	2078	838
8	*COL7A1*	0.1943	0.1822	0.2082	4	2	2078	838

LAML

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*RAD51C*	0.0781	0.0118	0.0474	0	2	2078	362
2	*DDX11*	0.0773	0.0322	0.0644	4	3	2078	362
3	*FANCC*	0.1099	0.0612	0.0816	2	2	2078	362
4	*DIS3*	0.1099	0.0612	0.0816	2	2	2078	362

LGG

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*EPPK1*	0.0581	0.0117	0.0703	2	3	2078	418
2	*HLA-G*	0.0782	0.0125	0.0703	0	2	2078	418
3	*ABL2*	0.0943	0.0347	0.0703	1	2	2078	418
4	*PMS2*	0.1119	0.0642	0.0963	2	2	2078	418

5	*MUTYH*	0.1119	0.0642	0.0963	2	2	2078	418
6	*POLQ*	0.6076	0.6711	0.6711	18	2	2078	418

LUAD

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*ATM*	0.0236	0.0048	0.0673	2	5	2078	594
2	*TRAF7*	0.0783	0.0405	0.2835	0	2	2078	594
3	*ERCC2*	0.0801	0.0589	0.2835	2	3	2078	594
4	*DIS3*	0.0801	0.0589	0.2835	2	3	2078	594
5	*GJB2*	0.1064	0.1004	0.2835	3	3	2078	594
6	*FANCG*	0.1127	0.1052	0.2835	1	2	2078	594
7	*XPA*	0.1125	0.1052	0.2835	1	2	2078	594
8	*NRP2*	0.1125	0.1052	0.2835	1	2	2078	594
9	*NBN*	0.1125	0.1052	0.2835	1	2	2078	594
10	*FLT3*	0.1125	0.1052	0.2835	1	2	2078	594
11	*WRN*	0.2046	0.2653	0.3376	3	2	2078	594
12	*FAT1*	0.2044	0.2653	0.3376	3	2	2078	594
13	*ABCC2*	0.2605	0.3477	0.3745	4	2	2078	594
14	*PIK3C2G*	0.3219	0.4267	0.4267	5	2	2078	594

LUSC

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*FANCA*	0.0642	0.0203	0.1017	3	3	2078	226
2	*BRIP1*	0.0924	0.0333	0.1017	1	2	2078	226
3	*BRCA2*	0.1253	0.0957	0.1596	3	2	2078	226
4	*MC1R*	0.2951	0.3844	0.4805	10	2	2078	226
5	*POLQ*	0.3250	0.4083	0.4805	18	3	2078	226

OV

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*BRCA1*	0.0000	0.0000	0.0000	0	34	2078	740
2	*BRCA2*	0.0000	0.0000	0.0000	3	25	2078	740
3	*RAD51D*	0.0415	0.0062	0.0248	0	3	2078	740
4	*PALB2*	0.0413	0.0062	0.0248	0	3	2078	740
5	*CNKSR1*	0.0421	0.0125	0.0299	2	4	2078	740
6	*BRIP1*	0.0576	0.0214	0.0428	1	3	2078	740
7	*PIK3C2G*	0.0514	0.0232	0.0428	5	5	2078	740
8	*RAD51C*	0.0784	0.0338	0.0507	0	2	2078	740
9	*SDHA*	0.1513	0.1564	0.2085	2	2	2078	740
10	*FANCD2*	0.3097	0.3782	0.4538	5	2	2078	740
11	*BLM*	0.4349	0.5133	0.5600	7	2	2078	740
12	*POLQ*	0.9206	0.9052	0.9052	18	2	2078	740

PRAD

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*ATM*	0.0085	0.0000	0.0000	2	6	2078	260
2	*POLK*	0.1339	0.1021	0.1021	4	2	2078	260

STAD

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*PALB2*	0.0221	0.0019	0.0224	0	4	2078	350
2	*XRCC2*	0.0410	0.0090	0.0538	0	3	2078	350
3	*EME2*	0.0410	0.0090	0.0538	0	3	2078	350
4	*ATM*	0.0413	0.0195	0.0584	2	4	2078	350
5	*MSH6*	0.0781	0.0432	0.1036	0	2	2078	350
6	*HIST1H1E*	0.0781	0.0432	0.1036	0	2	2078	350
7	*NBN*	0.1097	0.1116	0.1913	1	2	2078	350
8	*BRIP1*	0.1097	0.1116	0.1913	1	2	2078	350
9	*POLE*	0.1484	0.1929	0.2572	2	2	2078	350

10	*EPPK1*	0.1484	0.1929	0.2572	2	2	2078	350
11	*BRCA2*	0.1938	0.2788	0.3041	3	2	2078	350
12	*PIK3C2G*	0.3015	0.4447	0.4447	5	2	2078	350

UCEC

# RANK	GENE	PVAL.CAST	PVAL.TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*MSH6*	0.0410	0.0013	0.0088	0	3	2078	386
2	*BRCA1*	0.0410	0.0013	0.0088	0	3	2078	386
3	*MRE11A*	0.0781	0.0116	0.0271	0	2	2078	386
4	*FANCG*	0.0931	0.0323	0.0566	1	2	2078	386
5	*CNKSR1*	0.1103	0.0601	0.0841	2	2	2078	386
6	*MAP3K15*	0.1498	0.1299	0.1515	4	2	2078	386
7	*DDX11*	0.1495	0.1299	0.1515	4	2	2078	386

Supplementary Table 2.8: Burden analysis results for Pan-Cancer discovery cohort using rare truncation variants from 624 cancer associated genes in 4,034 cases. The control cohort is ESP 6503 sample set.

PAN12

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)
1	*BRCA1*	0.0000	0.0000	0.0000	10	53	13006	8068
2	*BRCA2*	0.0000	0.0000	0.0000	44	50	13006	8068
3	*ATM*	0.0001	0.0000	0.0000	13	23	13006	8068
4	*BRIP1*	0.0006	0.0000	0.0001	7	16	13006	8068
5	*MSH6*	0.0040	0.0002	0.0043	5	11	13006	8068
6	*NBN*	0.0079	0.0006	0.0138	4	9	13006	8068
7	*PIK3C2G*	0.0097	0.0008	0.0153	21	19	13006	8068
8	*XRCC2*	0.0104	0.0009	0.0153	2	7	13006	8068
9	*RAD51C*	0.0117	0.0010	0.0153	1	6	13006	8068
10	*ERCC2*	0.0097	0.0010	0.0153	6	10	13006	8068
11	*PALB2*	0.0172	0.0028	0.0333	11	12	13006	8068
12	*DIS3*	0.0362	0.0095	0.1050	12	11	13006	8068
13	*DDX11*	0.0379	0.0096	0.1050	14	12	13006	8068
14	*BLM*	0.0379	0.0096	0.1050	14	12	13006	8068
15	*FLT3*	0.0327	0.0099	0.1050	2	5	13006	8068
16	*RAD51B*	0.0378	0.0125	0.1050	1	4	13006	8068
17	*SLC19A1*	0.0416	0.0131	0.1050	0	3	13006	8068
18	*IRF4*	0.0416	0.0131	0.1050	0	3	13006	8068
19	*ERCC1*	0.0416	0.0131	0.1050	0	3	13006	8068
20	*FANCM*	0.0606	0.0181	0.1202	18	13	13006	8068
21	*SLC25A13*	0.0492	0.0213	0.1347	3	5	13006	8068
22	*MRE11A*	0.0591	0.0306	0.1849	2	4	13006	8068
23	*PARP2*	0.0591	0.0306	0.1849	2	4	13006	8068
24	*ATR*	0.0708	0.0387	0.2143	4	5	13006	8068
25	*TYR*	0.0708	0.0387	0.2143	4	5	13006	8068
26	*NRP2*	0.0708	0.0387	0.2143	4	5	13006	8068
27	*GJB2*	0.0870	0.0416	0.2143	10	8	13006	8068
28	*MC1R*	0.0786	0.0418	0.2143	6	6	13006	8068
29	*SGK1*	0.0699	0.0432	0.2143	1	3	13006	8068
30	*KRAS*	0.0699	0.0432	0.2143	1	3	13006	8068
31	*POLH*	0.0699	0.0432	0.2143	1	3	13006	8068
32	*GPS2*	0.0699	0.0432	0.2143	1	3	13006	8068
33	*DIS3L2*	0.0699	0.0432	0.2143	1	3	13006	8068
34	*RAD51D*	0.0877	0.0582	0.2275	3	4	13006	8068
35	*HNF1A*	0.0877	0.0582	0.2275	3	4	13006	8068
36	*BUB1B*	0.0877	0.0582	0.2275	3	4	13006	8068
37	*CRLF2*	0.1089	0.0891	0.3202	2	3	13006	8068
38	*TP53BP1*	0.1089	0.0891	0.3202	2	3	13006	8068
39	*RAD52*	0.1089	0.0891	0.3202	2	3	13006	8068
40	*CNKSR1*	0.1896	0.1017	0.3383	18	10	13006	8068
41	*CYP1B1*	0.1717	0.1188	0.3855	9	6	13006	8068
42	*ABCC4*	0.1717	0.1188	0.3855	9	6	13006	8068
43	*FANCA*	0.2013	0.1246	0.3855	14	8	13006	8068
44	*FLT1*	0.1347	0.1405	0.4248	1	2	13006	8068
45	*CDKN2A*	0.1347	0.1405	0.4248	1	2	13006	8068
46	*SBDS*	0.1347	0.1405	0.4248	1	2	13006	8068
47	*RUNX1*	0.1347	0.1405	0.4248	1	2	13006	8068
48	*HIST1H1E*	0.1347	0.1405	0.4248	1	2	13006	8068
49	*ETV4*	0.1347	0.1405	0.4248	1	2	13006	8068
50	*AKT3*	0.1347	0.1405	0.4248	1	2	13006	8068
51	*SDHA*	0.1587	0.1476	0.4248	3	3	13006	8068

52	*BCR*	0.1587	0.1476	0.4248	3	3	13006	8068
53	*GPR124*	0.1586	0.1476	0.4248	3	3	13006	8068
54	*FANCL*	0.1586	0.1476	0.4248	3	3	13006	8068
55	*ERBB3*	0.1586	0.1476	0.4248	3	3	13006	8068
56	*CHEK2*	0.2124	0.1539	0.4248	10	6	13006	8068
57	*RAD50*	0.2785	0.1851	0.4320	16	8	13006	8068
58	*XPC*	0.2196	0.1919	0.4400	6	4	13006	8068
59	*FANCI*	0.2572	0.1929	0.4400	11	6	13006	8068
60	*POLQ*	0.4185	0.1950	0.4400	47	19	13006	8068
61	*HIST1H4E*	0.2184	0.2146	0.4680	4	3	13006	8068
62	*MNDA*	0.2184	0.2146	0.4680	4	3	13006	8068
63	*FAT1*	0.2184	0.2146	0.4680	4	3	13006	8068
64	*ALK*	0.2184	0.2146	0.4680	4	3	13006	8068
65	*ABL2*	0.2671	0.2180	0.4680	9	5	13006	8068
66	*PMS2*	0.2671	0.2180	0.4680	9	5	13006	8068
67	*HGF*	0.2071	0.2379	0.4723	2	2	13006	8068
68	*CYP2C19*	0.2071	0.2379	0.4723	2	2	13006	8068
69	*UROD*	0.2071	0.2379	0.4723	2	2	13006	8068
70	*PRKDC*	0.2071	0.2379	0.4723	2	2	13006	8068
71	*ITGAV*	0.2071	0.2379	0.4723	2	2	13006	8068
72	*HIST1H2BD*	0.2071	0.2379	0.4723	2	2	13006	8068
73	*HFE*	0.2071	0.2379	0.4723	2	2	13006	8068
74	*BAP1*	0.2071	0.2379	0.4723	2	2	13006	8068
75	*APITD1*	0.2071	0.2379	0.4723	2	2	13006	8068
76	*PARP3*	0.3376	0.2470	0.4723	15	7	13006	8068
77	*RAD54L*	0.3212	0.2675	0.4723	10	5	13006	8068
78	*ESR2*	0.3212	0.2675	0.4723	10	5	13006	8068
79	*TOP3A*	0.2862	0.2864	0.4821	5	3	13006	8068
80	*MED23*	0.2862	0.2864	0.4821	5	3	13006	8068
81	*MUTYH*	0.3390	0.3076	0.5051	8	4	13006	8068
82	*BAK1*	0.3390	0.3076	0.5051	8	4	13006	8068
83	*ABCC2*	0.4348	0.3302	0.5291	17	7	13006	8068
84	*NQO1*	0.2922	0.3372	0.5339	3	2	13006	8068
85	*EPHB2*	0.2922	0.3372	0.5339	3	2	13006	8068
86	*APC*	0.2922	0.3372	0.5339	3	2	13006	8068
87	*ERCC4*	0.2922	0.3372	0.5339	3	2	13006	8068
88	*EPHA5*	0.2922	0.3372	0.5339	3	2	13006	8068
89	*CBLC*	0.2922	0.3372	0.5339	3	2	13006	8068
90	*AZGP1*	0.2922	0.3372	0.5339	3	2	13006	8068
91	*HLA-G*	0.3596	0.3595	0.5339	6	3	13006	8068
92	*TELO2*	0.3596	0.3595	0.5339	6	3	13006	8068
93	*FGFR4*	0.3596	0.3595	0.5339	6	3	13006	8068
94	*ODAM*	0.4037	0.3681	0.5339	9	4	13006	8068
95	*WRN*	0.4364	0.3718	0.5339	12	5	13006	8068
96	*XPA*	0.4364	0.3718	0.5339	12	5	13006	8068
97	*COL7A1*	0.5329	0.4166	0.5712	19	7	13006	8068
98	*POLK*	0.4354	0.4313	0.5854	7	3	13006	8068
99	*POLI*	0.4354	0.4313	0.5854	7	3	13006	8068
100	*EXT2*	0.4354	0.4313	0.5854	7	3	13006	8068
101	*CASP8*	0.4354	0.4313	0.5854	7	3	13006	8068
102	*SUZ12*	0.3845	0.4320	0.5854	4	2	13006	8068
103	*RECQL4*	0.3845	0.4320	0.5854	4	2	13006	8068
104	*ATRIP*	0.3845	0.4320	0.5854	4	2	13006	8068
105	*APOL2*	0.3845	0.4320	0.5854	4	2	13006	8068
106	*FANCC*	0.5806	0.4597	0.5854	20	7	13006	8068
107	*FANCD2*	0.5522	0.4760	0.5917	14	5	13006	8068
108	*ZRANB3*	0.5110	0.4999	0.6157	8	3	13006	8068
109	*IL7R*	0.5110	0.4999	0.6157	8	3	13006	8068

110	*EME1*	0.5110	0.4999	0.6157	8	3	13006	8068
111	*MYB*	0.4783	0.5189	0.6218	5	2	13006	8068
112	*MAP3K15*	0.6752	0.5311	0.6307	25	8	13006	8068
113	*ERCC3*	0.6752	0.5311	0.6307	25	8	13006	8068
114	*DNMT3A*	0.5836	0.5640	0.6580	9	3	13006	8068
115	*PARP4*	0.6547	0.5948	0.6880	13	4	13006	8068
116	*SLX4*	0.6547	0.5948	0.6880	13	4	13006	8068
117	*FAH*	0.6512	0.6228	0.7080	10	3	13006	8068
118	*IDH1*	0.6512	0.6228	0.7080	10	3	13006	8068
119	*TRAF7*	0.6512	0.6228	0.7080	10	3	13006	8068
120	*POLE*	0.7084	0.6436	0.7133	14	4	13006	8068
121	*FLCN*	0.6509	0.6642	0.7300	7	2	13006	8068
122	*USP9X*	0.7236	0.7224	0.7875	8	2	13006	8068
123	*FANCG*	0.8480	0.7434	0.8039	24	6	13006	8068
124	*COMT*	0.8129	0.7650	0.8205	13	3	13006	8068
125	*TET2*	0.8681	0.7977	0.8488	18	4	13006	8068
126	*EME2*	0.8522	0.8014	0.8488	14	3	13006	8068
127	*DOCK8*	0.8362	0.8134	0.8518	10	2	13006	8068
128	*FANCF*	0.8362	0.8134	0.8518	10	2	13006	8068
129	*EPPK1*	0.9536	0.8507	0.8771	44	10	13006	8068
130	*AURKB*	0.9111	0.8602	0.8800	16	3	13006	8068
131	*ACO1*	0.9090	0.8767	0.8901	12	2	13006	8068
132	*EXO1*	0.9718	0.9337	0.9407	20	3	13006	8068
133	*ROS1*	0.9945	0.9755	0.9755	25	3	13006	8068

BRCA

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*BRCA1*	0.0006	0.0000	0.0000	10	12	13006	1540
2	*BRCA2*	0.0002	0.0000	0.0000	44	19	13006	1540
3	*FANCM*	0.0110	0.0004	0.0050	18	8	13006	1540
4	*ATR*	0.0335	0.0016	0.0152	4	4	13006	1540
5	*CRLF2*	0.0511	0.0035	0.0269	2	3	13006	1540
6	*PIK3C2G*	0.0421	0.0121	0.0766	21	6	13006	1540
7	*FLT1*	0.0878	0.0154	0.0835	1	2	13006	1540
8	*SBDS*	0.0877	0.0154	0.0835	1	2	13006	1540
9	*RAD51B*	0.0877	0.0154	0.0835	1	2	13006	1540
10	*AKT3*	0.0877	0.0154	0.0835	1	2	13006	1540
11	*CHEK2*	0.0574	0.0160	0.0835	10	4	13006	1540
12	*MSH6*	0.0684	0.0167	0.0835	5	3	13006	1540
13	*XRCC2*	0.0977	0.0293	0.0856	2	2	13006	1540
14	*RAD52*	0.0977	0.0293	0.0856	2	2	13006	1540
15	*ATM*	0.0732	0.0319	0.0856	13	4	13006	1540
16	*BRIP1*	0.0822	0.0321	0.0856	7	3	13006	1540
17	*BAK1*	0.0898	0.0418	0.0934	8	3	13006	1540
18	*ABCC2*	0.0993	0.0636	0.1344	17	4	13006	1540
19	*ESR2*	0.1066	0.0649	0.1344	10	3	13006	1540
20	*TYR*	0.1200	0.0663	0.1344	4	2	13006	1540
21	*MNDA*	0.1200	0.0663	0.1344	4	2	13006	1540
22	*MC1R*	0.1456	0.1123	0.1939	6	2	13006	1540
23	*ERCC2*	0.1456	0.1123	0.1939	6	2	13006	1540
24	*CASP8*	0.1595	0.1376	0.2178	7	2	13006	1540
25	*ZRANB3*	0.1742	0.1639	0.2492	8	2	13006	1540
26	*DNMT3A*	0.1897	0.1911	0.2792	9	2	13006	1540
27	*TET2*	0.1943	0.1971	0.2792	18	3	13006	1540
28	*RAD54L*	0.2061	0.2187	0.2968	10	2	13006	1540
29	*DIS3*	0.2409	0.2748	0.3600	12	2	13006	1540
30	*DDX11*	0.2783	0.3306	0.4187	14	2	13006	1540
31	*BLM*	0.2783	0.3306	0.4187	14	2	13006	1540

32	*FANCD2*	0.2783	0.3306	0.4187	14	2	13006	1540
33	*AURKB*	0.3182	0.3851	0.4435	16	2	13006	1540
34	*CNKSR1*	0.3600	0.4376	0.4891	18	2	13006	1540
35	*ERCC3*	0.5144	0.5995	0.6509	25	2	13006	1540
36	*MAP3K15*	0.5144	0.5995	0.6509	25	2	13006	1540
37	*EPPK1*	0.6323	0.6804	0.6988	44	3	13006	1540
38	*POLQ*	0.6816	0.7206	0.7206	47	3	13006	1540

GBM

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*APITD1*	0.0830	0.0027	0.0134	2	2	13006	534
2	*RAD50*	0.0635	0.0074	0.0185	16	3	13006	534
3	*MSH6*	0.0906	0.0090	0.0185	5	2	13006	534
4	*CYP1B1*	0.1017	0.0223	0.0279	9	2	13006	534
5	*BLM*	0.1164	0.0453	0.0453	14	2	13006	534

HNSC

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*FANCA*	0.0630	0.0072	0.0504	14	3	13006	582
2	*POLQ*	0.0435	0.0080	0.0504	47	5	13006	582
3	*FANCM*	0.0706	0.0131	0.0504	18	3	13006	582
4	*PMS2*	0.1048	0.0272	0.0504	9	2	13006	582
5	*PALB2*	0.1114	0.0373	0.0523	11	2	13006	582
6	*CNKSR1*	0.1369	0.0816	0.0952	18	2	13006	582
7	*FANCC*	0.1448	0.0962	0.0962	20	2	13006	582

KIRC

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*XPC*	0.0517	0.0017	0.0132	6	3	13006	904
2	*BAP1*	0.0850	0.0046	0.0184	2	2	13006	904
3	*BCR*	0.0884	0.0075	0.0200	3	2	13006	904
4	*PARP3*	0.0708	0.0133	0.0266	15	3	13006	904
5	*ERCC2*	0.0992	0.0199	0.0318	6	2	13006	904
6	*EME1*	0.1070	0.0308	0.0410	8	2	13006	904
7	*FANCM*	0.1521	0.1081	0.1235	18	2	13006	904
8	*COL7A1*	0.1572	0.1173	0.1235	19	2	13006	904

LAML

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*RAD51C*	0.0803	0.0012	0.0048	1	2	13006	400
2	*DDX11*	0.0589	0.0045	0.0090	14	3	13006	400
3	*DIS3*	0.1079	0.0315	0.0420	12	2	13006	400
4	*FANCC*	0.1312	0.0720	0.0720	20	2	13006	400

LGG

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X(CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*HLA-G*	0.0931	0.0115	0.0688	6	2	13006	446
2	*MUTYH*	0.0981	0.0179	0.0688	8	2	13006	446
3	*ABL2*	0.1011	0.0216	0.0688	9	2	13006	446
4	*PMS2*	0.1008	0.0216	0.0688	9	2	13006	446
5	*EPPK1*	0.1199	0.0769	0.0922	44	3	13006	446
6	*POLQ*	0.2462	0.2769	0.2769	47	2	13006	446

LUAD

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N (CASES)

1	*ATM*	0.0233	0.0005	0.0073	13	5	13006	924
2	*ERCC2*	0.0566	0.0043	0.0299	6	3	13006	924
3	*FLT3*	0.0877	0.0088	0.0410	2	2	13006	924
4	*GJB2*	0.0685	0.0129	0.0452	10	3	13006	924
5	*DIS3*	0.0755	0.0194	0.0543	12	3	13006	924
6	*NRP2*	0.0978	0.0209	0.0543	4	2	13006	924
7	*NBN*	0.0978	0.0209	0.0543	4	2	13006	924
8	*FAT1*	0.0978	0.0209	0.0543	4	2	13006	924
9	*TRAF7*	0.1329	0.0785	0.1221	10	2	13006	924
10	*WRN*	0.1464	0.1029	0.1440	12	2	13006	924
11	*XPA*	0.1463	0.1029	0.1440	12	2	13006	924
12	*ABCC2*	0.1834	0.1704	0.1988	17	2	13006	924
13	*PIK3C2G*	0.2170	0.2282	0.2458	21	2	13006	924
14	*FANCG*	0.2442	0.2724	0.2724	24	2	13006	924

LUSC

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*FANCA*	0.0574	0.0039	0.0196	14	3	13006	386
2	*MC1R*	0.0913	0.0095	0.0238	6	2	13006	386
3	*BRIP1*	0.0939	0.0121	0.0238	7	2	13006	386
4	*POLQ*	0.1160	0.0711	0.0889	47	3	13006	386
5	*BRCA2*	0.2123	0.2206	0.2206	44	2	13006	386

OV

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*BRCA1*	0.0000	0.0000	0.0000	10	34	13006	858
2	*BRCA2*	0.0000	0.0000	0.0000	44	25	13006	858
3	*RAD51D*	0.0482	0.0010	0.0040	3	3	13006	858
4	*PIK3C2G*	0.0321	0.0027	0.0080	21	5	13006	858
5	*RAD51C*	0.0828	0.0042	0.0102	1	2	13006	858
6	*BRIP1*	0.0584	0.0054	0.0108	7	3	13006	858
7	*CNKSR1*	0.0514	0.0088	0.0151	18	4	13006	858
8	*SDHA*	0.0922	0.0134	0.0201	3	2	13006	858
9	*PALB2*	0.0706	0.0146	0.0201	11	3	13006	858
10	*BLM*	0.1568	0.1221	0.1465	14	2	13006	858
11	*FANCD2*	0.1567	0.1221	0.1465	14	2	13006	858
12	*POLQ*	0.4738	0.5606	0.5606	47	2	13006	858

PRAD

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*ATM*	0.0087	0.0000	0.0000	13	6	13006	356
2	*POLK*	0.0914	0.0093	0.0093	7	2	13006	356

STAD

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*XRCC2*	0.0456	0.0005	0.0060	2	3	13006	642
2	*PALB2*	0.0372	0.0020	0.0118	11	4	13006	642
3	*ATM*	0.0409	0.0032	0.0129	13	4	13006	642
4	*HIST1H1E*	0.0826	0.0041	0.0129	1	2	13006	642
5	*NBN*	0.0964	0.0192	0.0461	4	2	13006	642
6	*EME2*	0.0795	0.0244	0.0488	14	3	13006	642
7	*MSH6*	0.1013	0.0262	0.0488	5	2	13006	642
8	*BRIP1*	0.1118	0.0428	0.0642	7	2	13006	642
9	*POLE*	0.1541	0.1200	0.1601	14	2	13006	642
10	*PIK3C2G*	0.2057	0.2140	0.2568	21	2	13006	642
11	*EPPK1*	0.4277	0.5208	0.5681	44	2	13006	642

12	*BRCA2*	0.4277	0.5208	0.5681	44	2	13006	642

UCEC

# RANK	GENE	PVAL CAST	PVAL TFT	FDR.TFT ONLY	X (CONTROLS)	X (CASES)	N (CONTROLS)	N(CASES)
1	*MSH6*	0.0468	0.0004	0.0029	5	3	13006	496
2	*BRCA1*	0.0531	0.0020	0.0069	10	3	13006	496
3	*MRE11A*	0.0827	0.0023	0.0069	2	2	13006	496
4	*DDX11*	0.1133	0.0400	0.0699	14	2	13006	496
5	*CNKSR1*	0.1254	0.0600	0.0841	18	2	13006	496
6	*FANCG*	0.1444	0.0950	0.1108	24	2	13006	496
7	*MAP3K15*	0.1480	0.1013	0.1108	25	2	13006	496

Supplementary Table 2.9: Gene-based LOH analysis using rare truncation variants in significant genes from burden analysis.

Gene	P-value	FDR
BRCA1	5.55E-17	8.88E-16
BRCA2	3.50E-14	2.80E-13
RAD51D	1.21E-05	6.46E-05
PALB2	0.0035285	0.014114
BAP1	0.0050624	0.0161997
RAD51C	0.0098984	0.0263957
ATM	0.021792	0.0498103
BRIP1	0.0329	0.0658
FANCM	0.0824	0.146489
POLK	0.10393	0.166288
APITD1	0.29351	0.426924
HIST1H1E	0.31589	0.421187
FANCA	0.35538	0.437391
MRE11A	0.36367	0.415623
HLA-G	0.40043	0.427125
PMS2	0.41588	0.41588

Supplementary Table 2.10: Site-based LOH analysis for rare truncation variants in 624 cancer associated genes.

Gene	Chr	Position	Ref	Var	Normal			Tumor			Cancer Type	Pool MAF	FDR
					Ref	Var	VAF	Ref	Var	VAF			
BRCA1	17	41209079	-	G	940	735	43.88	777	3639	82.4	OV	0.047	0
BRCA1	17	41209079	-	G	447	300	40.16	121	468	79.46	OV	0.047	6.1E-46
BRCA1	17	41246433	C	T	307	302	49.51	37	377	90.84	OV	0.005	1.0E-45
BRCA1	17	41245159	C	A	531	481	47.53	153	603	79.45	OV	0.005	4.7E-42
BRCA1	17	41245073	G	-	479	383	44.43	150	518	77.54	OV	0.009	1.7E-37
PALB2	16	23647108	-	A	186	133	41.69	50	320	86.49	OV	0.014	4.2E-34
BRCA1	17	41245410	G	C	164	144	46.75	27	255	90.43	OV	0.005	4.1E-30
BRCA1	17	41209079	-	G	313	196	38.51	221	529	70.53	OV	0.047	4.0E-27
BRCA1	17	41244145	G	A	253	224	46.96	96	357	78.81	OV	0.005	5.2E-22
BRCA1	17	41251825	G	-	115	80	41.03	40	227	85.02	OV	0.005	1.8E-21
BRCA1	17	41209079	-	G	410	295	41.84	38	147	79.46	OV	0.047	1.2E-18
POLK	5	74892232	C	T	96	74	43.27	12	126	91.3	OV	0.014	1.3E-18
BRIP1	17	59857686	G	T	112	85	42.71	22	148	87.06	OV	0.005	1.1E-17
BRCA1	17	41276045	CT	-	175	137	43.91	27	144	84.21	OV	0.043	3.2E-17
RAD51C	17	56801399	AG	-	282	121	30.02	58	128	68.82	OV	0.005	1.2E-16
BRCA1	17	41243513	T	-	150	102	40.48	36	142	79.78	BRCA	0.019	8.1E-15
BRCA1	17	41246038	G	-	355	275	43.65	47	153	76.5	OV	0.005	1.0E-14
BRCA1	17	41276045	CT	-	190	105	35.59	43	127	74.71	OV	0.043	1.4E-14
MSH6	2	48032149	C	G	56	27	32.53	9	81	90	GBM	0.005	1.5E-14
BRCA2	13	32906495	G	T	88	90	50.56	3	75	96.15	OV	0.005	3.3E-14
ATM	11	108183151	G	T	155	99	38.98	31	117	79.05	PRAD	0.014	9.6E-14
RAD51C	17	56780562	C	T	77	57	42.54	68	279	80.4	BRCA	0.009	1.1E-13
BRIP1	17	59934505	TTGT	-	307	204	39.92	93	203	68.58	OV	0.005	2.0E-13
BRCA1	17	41276045	CT	-	118	68	36.56	42	136	76.4	OV	0.043	4.3E-13
BRCA2	13	32913778	T	G	47	23	32.86	5	61	92.42	OV	0.005	6.2E-13
BRCA1	17	41245332	-	AG	118	108	47.79	38	170	81.73	OV	0.005	1.9E-12
BRCA1	17	41245091	G	-	226	192	45.93	53	170	76.23	OV	0.005	2.3E-12
RAD51D	17	33434045	G	A	211	218	50.46	61	218	78.14	OV	0.005	3.1E-12
BRIP1	17	59821942	T	GGA	161	135	45.61	56	181	76.37	OV	0.005	1.1E-11
BRCA1	17	41209079	-	G	315	254	44.64	99	225	69.44	BRCA	0.047	2.5E-11
PALB2	16	23647357	TC	-	74	55	42.64	39	160	80.4	BRCA	0.009	5.3E-11
BRCA1	17	41209079	-	G	375	298	44.28	239	413	63.34	OV	0.047	1.5E-10
BRCA2	13	32929170	A	T	207	168	44.56	42	129	75.44	OV	0.005	3.2E-10
BRCA1	17	41242962	-	GA	56	23	29.11	15	64	81.01	OV	0.005	5.0E-10
BRCA1	17	41276045	CT	-	128	103	44.59	41	137	76.97	BRCA	0.043	5.3E-10
BRCA1	17	41243513	T	-	118	126	51.64	99	325	76.65	OV	0.019	9.3E-10
BRCA2	13	32907420	-	A	99	68	40.72	52	151	74.38	OV	0.009	1.2E-09
BRCA1	17	41246652	AC	-	109	78	41.71	12	63	84	OV	0.005	1.7E-09
BRCA1	17	41199682	C	T	38	52	57.78	15	155	91.18	OV	0.005	2.0E-09
BRCA1	17	41209079	-	G	134	83	38.25	62	139	69.15	STAD	0.047	5.7E-09
BRCA1	17	41276045	CT	-	130	62	32.29	83	143	63.27	OV	0.043	8.2E-09
FANCM	14	45658326	C	T	94	65	40.88	20	76	79.17	HNSC	0.038	1.8E-08
BRCA1	17	41243800	C	A	94	111	54.15	22	112	83.58	UCEC	0.005	7.9E-08
PALB2	16	23640545	G	A	77	60	43.8	14	66	82.5	STAD	0.005	1.0E-07
BRCA2	13	32931944	GT	-	100	54	35.06	31	77	71.3	BRCA	0.005	1.3E-07
BRCA1	17	41209079	-	G	93	79	45.93	20	81	80.2	BRCA	0.047	1.5E-07
BRCA2	13	32914438	T	-	138	90	39.47	62	130	67.71	OV	0.047	1.5E-07
BRCA2	13	32913044	G	T	131	111	45.87	47	131	73.6	BRCA	0.005	1.5E-07
BRCA1	17	41243581	G	A	105	99	48.53	46	145	75.92	BRCA	0.005	2.2E-07
ATM	11	108173702	G	-	110	86	43.88	70	167	70.46	BRCA	0.005	3.7E-07
BRCA2	13	32903605	TG	-	131	78	37.32	17	52	75.36	BRCA	0.009	4.3E-07
BRCA2	13	32913381	C	G	88	92	51.11	51	163	76.17	OV	0.005	2.1E-06
BRCA2	13	32914529	A	T	28	39	58.21	6	69	92	OV	0.005	2.7E-06

BRCA1	17	41209079	-	G	138	114	45.24	25	75	75	OV	0.047	3.2E-06
BRCA2	13	32914438	T	-	120	84	41.18	35	84	70.59	OV	0.047	3.8E-06
BRCA1	17	41276045	CT	-	64	50	43.86	17	64	79.01	OV	0.043	5.0E-06
ABL2	1	179079416	C	T	282	157	35.76	199	224	52.83	LGG	0.047	8.6E-06
BRCA1	17	41245073	G	-	137	121	46.9	62	140	69.31	OV	0.009	1.5E-05
BRCA2	13	32914438	T	-	91	87	48.88	30	93	75.61	OV	0.047	2.1E-05
ATM	11	108172487	C	-	38	19	33.33	22	59	72.84	LUAD	0.005	4.3E-05
BRCA2	13	32968951	C	T	40	24	37.5	17	54	76.06	LUSC	0.005	4.6E-05
BRCA1	17	41276045	CT	-	52	34	39.53	9	36	80	OV	0.043	4.6E-05
RAD51D	17	33433425	G	A	32	35	52.24	6	45	88.24	OV	0.009	4.6E-05
RAD51D	17	33433425	G	A	60	45	42.86	7	33	82.5	OV	0.009	5.5E-05
BRCA1	17	41201209	G	-	51	40	43.96	23	72	75.79	OV	0.005	6.2E-05
FANCM	14	45654441	G	-	40	22	35.48	50	111	68.94	BRCA	0.005	6.9E-05
BRCA1	17	41244016	-	A	66	38	36.54	44	86	66.15	BRCA	0.005	7.4E-05
BRIP1	17	59871048	A	C	56	95	62.91	45	205	82	BRCA	0.005	8.5E-05
BRCA1	17	41234535	C	-	110	105	48.84	88	190	68.35	UCEC	0.005	0.00010
BRCA1	17	41243513	T	-	43	31	41.89	30	82	73.21	BRCA	0.019	0.00015
BRCA1	17	41245587	T	-	37	15	28.85	29	58	66.67	OV	0.005	0.00017
BRCA1	17	41276045	CT	-	85	66	43.71	40	87	68.5	UCEC	0.043	0.00027
BAP1	3	52436624	G	A	122	117	48.75	58	127	68.28	KIRC	0.005	0.00032
BRCA2	13	32914438	T	-	81	71	46.71	25	66	72.53	OV	0.047	0.00046
BRCA2	13	32937541	-	C	226	187	45.28	150	222	59.68	BRCA	0.005	0.00052
BRCA2	13	32913837	AA	-	30	39	56.52	19	89	82.41	BRCA	0.009	0.00057
ATM	11	108121753	AG	-	121	69	36.32	72	96	57.14	LUAD	0.024	0.00088
BRCA2	13	32913604	AATA	-	75	56	42.75	19	47	71.21	OV	0.005	0.00090
BRCA1	17	41246625	C	-	79	55	41.04	33	65	66.33	BRCA	0.005	0.00103
RAD51C	17	56780562	C	T	44	35	44.3	16	47	74.6	OV	0.009	0.00126
ATM	11	108216491	G	-	46	25	35.21	24	48	66.67	BRCA	0.005	0.00129
MRE11A	11	94180442	G	A	140	119	45.95	191	291	60.37	UCEC	0.009	0.00133
PALB2	16	23646388	G	-	138	87	38.67	111	141	55.95	OV	0.005	0.00152
BRCA2	13	32930687	C	T	91	85	48.3	33	77	70	STAD	0.014	0.00152
ATM	11	108165719	-	CT	78	40	33.9	36	53	59.55	LUAD	0.005	0.00219
FANCM	14	45628392	-	A	31	20	39.22	13	36	73.47	BRCA	0.005	0.00261
BAP1	3	52441973	C	T	29	24	45.28	13	43	76.79	KIRC	0.005	0.00260
BRCA2	13	32968863	C	G	25	27	51.92	8	38	82.61	BRCA	0.005	0.00260
ATM	11	108121753	AG	-	130	79	37.8	36	53	59.55	STAD	0.024	0.00425
BRCA1	17	41209079	-	G	54	33	37.93	74	110	59.78	BRCA	0.047	0.00554
PALB2	16	23649207	ACAA	-	62	33	34.74	39	56	58.95	STAD	0.009	0.00619
BRCA2	13	32906640	A	-	69	73	51.41	34	79	69.91	OV	0.005	0.01000
BRCA1	17	41243513	T	-	43	37	46.25	22	52	70.27	OV	0.019	0.00997
BRCA1	17	41215906	C	-	96	64	40	53	74	58.27	BRCA	0.005	0.01288
BRCA2	13	32914438	T	-	121	116	48.95	62	108	63.53	OV	0.047	0.01533
BRCA1	17	41256250	-	T	409	278	40.47	271	262	49.16	BRCA	0.005	0.01629
ATM	11	108115640	T	-	36	15	29.41	48	59	55.14	PRAD	0.005	0.01725
DDX11	12	31256740	G	A	121	64	34.59	159	147	48.04	LAML	0.009	0.02419
BRCA2	13	32914438	T	-	98	74	43.02	31	49	61.25	OV	0.047	0.03180
PALB2	16	23649207	ACAA	-	38	23	37.7	19	30	61.22	LUAD	0.009	0.06352
PALB2	16	23647357	TC	-	136	104	43.33	63	80	55.94	STAD	0.009	0.07285
MSH6	2	48033981	-	TTGA	72	50	40.98	53	68	56.2	BRCA	0.028	0.07762
HLA-G	6	29796574	A	T	66	59	46.83	46	73	60.83	LGG	0.005	0.08982
BRCA1	17	41276045	CT	-	80	69	46.31	27	44	61.97	OV	0.043	0.09724
PIK3C2G	12	18691246	G	-	42	32	43.24	53	77	59.23	LUAD	0.014	0.10222
MAP3K15	X	19392724	G	A	83	59	41.55	90	105	53.3	LAML	0.009	0.10491
EPPK1	8	144941508	C	A	42	33	42.86	19	31	59.62	STAD	0.009	0.14932
MSH6	2	48033745	AAGC	-	39	19	32.76	35	37	51.39	UCEC	0.005	0.15187
FANCM	14	45636336	C	T	31	38	55.07	14	33	70.21	BRCA	0.019	0.18008
FANCG	9	35075554	G	C	62	65	51.18	49	80	62.02	UCEC	0.005	0.19452

Supplementary Table 2.11: Gene-based LOH analysis for rare missense variants in 624 cancer associated genes.

Gene	P-Value	FDR
BRCA1	0.000723162	0.004338975
ATM	0.001159173	0.004338975
BRCA2	0.065008945	0.13001789
RAD51C	0.100971489	0.151457234
BRIP1	0.258099528	0.309719433
RAD51D	0.329113579	0.329113579

Supplementary Table 2.12: Site-based LOH analysis for rare missense variants in 624 cancer associated genes.

Gene	Chr	Position	Ref	Var	Normal			Tumor			Cancer Type	Pool MAF	FDR
					Ref	Var	VAF	Ref	Var	VAF			
BRIP1	17	59937223	G	C	248	221	47.12	10	426	97.71	OV	0.028	1.06E-71
BRCA1	17	41245027	G	A	644	613	48.69	146	732	83.28	OV	0.247	2.68E-59
PARP3	3	51978526	T	C	13	65	83.33	179	0	0	LAML	0.005	1.76E-43
XRCC2	7	152346287	T	C	113	139	55.16	14	460	96.64	OV	0.043	1.35E-42
BRCA1	17	41245233	A	G	231	230	49.78	25	267	91.44	OV	0.033	1.09E-32
BRCA1	17	41243509	T	C	262	286	52.19	86	483	84.89	OV	0.489	5.04E-30
POLK	5	74869615	G	T	149	121	44.65	152	0	0	LUAD	0.005	1.22E-25
EME1	17	48453202	C	A	168	162	49.09	38	270	87.66	OV	0.009	5.20E-24
FANCM	14	45658366	C	T	101	104	50.24	51	360	87.59	GBM	0.043	5.75E-20
FANCC	9	98011545	C	T	87	90	50.85	14	182	92.86	BRCA	0.033	8.02E-19
MSH6	2	48032068	T	C	188	180	48.91	107	407	78.88	OV	0.005	3.92E-18
EPPK1	8	144941013	G	T	139	111	44.22	39	205	84.02	KIRC	0.005	3.72E-18
ATM	11	108201008	C	G	128	85	39.91	9	96	91.43	OV	0.009	3.57E-18
BRCA2	13	32968940	A	T	217	246	52.12	41	239	83.57	BRCA	0.005	5.46E-18
PIK3C2G	12	18435620	A	T	477	437	47.81	225	528	69.84	OV	0.024	1.81E-17
BRCA1	17	41201181	C	A	35	46	56.79	4	213	98.16	OV	0.005	1.89E-17
BRCA1	17	41201187	A	G	142	121	45.66	37	197	84.19	BRCA	0.005	3.89E-17
FANCM	14	45644409	A	G	217	204	48.34	69	274	79.88	KIRC	0.014	3.82E-17
BRCA1	17	41258495	A	C	56	54	49.09	5	130	96.3	OV	0.009	3.69E-17
PALB2	16	23641340	G	A	97	63	39.38	137	478	77.6	BRCA	0.047	5.06E-17
MAP3K15	X	19410570	C	T	183	187	50.54	22	159	87.85	OV	0.005	5.53E-17
BRCA1	17	41223196	G	C	273	239	46.5	78	264	77.19	OV	0.009	5.66E-17
ATM	11	108142070	A	G	332	324	49.39	179	494	73.29	OV	0.024	6.09E-17
BRCA2	13	32914755	C	T	92	80	46.51	3	86	96.63	OV	0.005	6.10E-17
BRIP1	17	59821942	T	A	131	137	47.4	24	181	83.41	OV	0.005	9.46E-17
RAD50	5	131977907	C	T	106	106	50	31	216	87.45	OV	0.005	9.28E-17
RAD50	5	131924529	G	C	84	84	49.7	3	95	96.94	OV	0.005	1.24E-16
ATM	11	108150267	C	G	172	189	52.35	10	120	92.31	OV	0.005	1.32E-16
ATM	11	108143552	G	A	102	112	52.34	19	176	90.26	HNSC	0.005	2.83E-16
BAP1	3	52437281	G	C	14	38	73.08	65	0	0	LAML	0.005	5.00E-16
ATM	11	108165748	A	G	41	44	51.76	2	116	98.31	UCEC	0.019	1.01E-15
EME2	16	1825096	C	T	108	88	44.9	21	132	86.27	STAD	0.005	2.51E-14
BRIP1	17	59878645	T	C	417	456	52.23	88	288	76.6	OV	0.005	3.77E-14
EME1	17	48453305	G	A	107	72	40	97	0	0	BRCA	0.005	8.05E-14
BRCA1	17	41246812	A	C	47	57	54.81	5	125	95.42	OV	0.024	1.02E-13
BRIP1	17	59885868	T	C	71	67	48.55	30	201	87.01	UCEC	0.005	1.51E-13
EPPK1	8	144940828	C	G	147	172	53.92	93	378	80.08	HNSC	0.009	4.81E-13
ERCC2	19	45858047	C	T	87	67	43.51	9	85	90.43	LGG	0.014	4.73E-13
FANCM	14	45646041	G	A	63	84	57.14	4	102	96.23	GBM	0.014	9.19E-13
MSH6	2	48025871	T	C	114	87	43.07	14	92	86.79	PRAD	0.009	1.30E-12
MAP3K15	X	19389563	C	T	114	96	45.71	10	85	89.47	BRCA	0.005	1.38E-12
BRIP1	17	59793364	G	A	55	64	53.78	15	164	91.62	OV	0.014	1.50E-12
FANCA	16	89815165	G	A	40	29	40.85	4	75	94.94	BRCA	0.005	5.67E-12
MSH6	2	48026778	T	A	211	234	52.47	274	730	72.64	OV	0.005	2.07E-11
BRCA1	17	41245381	T	C	132	105	44.3	27	118	80.82	UCEC	0.161	2.36E-11
MRE11A	11	94180441	C	T	163	132	44.59	102	273	72.8	UCEC	0.038	2.32E-11
BRCA1	17	41244982	A	G	199	198	49.75	72	236	76.62	STAD	0.033	2.54E-11
ATM	11	108158333	C	G	39	23	37.1	5	67	93.06	STAD	0.005	3.04E-11
FANCA	16	89838139	A	T	131	132	50.19	40	173	81.22	UCEC	0.005	7.15E-11
MUTYH	1	45800164	C	T	79	101	56.11	6	87	93.55	LGG	0.005	1.66E-10
ATM	11	108202716	A	C	97	85	46.7	37	156	80.83	OV	0.005	2.61E-10
FANCG	9	35078332	T	C	26	26	50	1	67	98.53	STAD	0.005	2.57E-10
BRCA2	13	32914229	T	C	91	67	42.41	16	85	84.16	BRCA	0.005	4.06E-10

FANCM	14	45644401	C	T	75	50	40	18	88	83.02	OV	0.009	7.25E-10
MAP3K15	X	19392655	C	T	173	145	45.6	40	133	76.88	OV	0.005	8.99E-10
BRCA1	17	41258504	A	C	64	42	39.62	12	73	85.88	BRCA	0.019	1.13E-09
POLK	5	74892218	T	G	49	47	48.96	10	94	90.38	GBM	0.009	1.27E-09
ATM	11	108121543	C	T	91	91	50	34	158	82.29	BRCA	0.005	1.27E-09
MRE11A	11	94180450	C	T	189	212	52.48	65	224	76.98	BRCA	0.005	1.33E-09
MRE11A	11	94212892	T	C	38	32	45.71	5	70	93.33	LUAD	0.005	1.35E-09
RAD50	5	131976461	G	A	42	48	53.33	3	69	95.83	BRCA	0.005	1.39E-09
ERCC2	19	45854910	C	G	94	92	48.94	15	92	85.98	LGG	0.033	3.68E-09
ATM	11	108224538	T	C	215	179	45.43	85	203	70.49	BRCA	0.005	5.35E-09
BRCA2	13	32915144	G	A	44	24	35.29	105	0	0	LAML	0.005	6.49E-09
BRCA2	13	32907098	G	C	45	48	51.61	5	68	93.15	BRCA	0.005	1.46E-08
FANCA	16	89851273	G	A	45	40	47.06	8	73	89.02	BRCA	0.009	1.83E-08
BRCA2	13	32914550	G	A	50	53	51.46	3	54	94.74	GBM	0.005	1.87E-08
RAD51C	17	56787260	G	A	103	99	49.01	13	78	85.71	BRCA	0.005	1.87E-08
FANCA	16	89825017	A	C	83	78	48.45	11	72	86.75	OV	0.019	3.57E-08
BRCA2	13	32954181	G	A	58	72	55.38	8	82	91.11	PRAD	0.005	3.80E-08
PALB2	16	23614833	G	A	114	117	50.65	50	175	77.78	OV	0.014	6.12E-08
FANCA	16	89874705	T	G	107	98	47.8	45	151	77.04	OV	0.005	6.52E-08
XPC	3	14214383	C	T	42	30	41.67	6	53	89.83	UCEC	0.005	6.97E-08
BRCA1	17	41258504	A	C	103	60	36.81	71	152	68.16	BRCA	0.019	8.19E-08
BRCA1	17	41258504	A	C	67	38	36.19	8	45	84.91	BRCA	0.019	8.20E-08
MAP3K15	X	19398357	C	T	51	50	49.02	5	56	91.8	OV	0.009	8.82E-08
BRCA1	17	41219645	G	A	44	28	38.89	11	64	85.33	OV	0.005	9.74E-08
FANCG	9	35076812	G	A	119	107	47.35	36	122	77.22	KIRC	0.005	1.25E-07
ATM	11	108203612	T	G	48	30	38.46	10	56	84.85	BRCA	0.009	2.14E-07
BRCA2	13	32937554	G	A	114	107	48.42	9	57	86.36	UCEC	0.009	2.32E-07
BRCA1	17	41256153	C	T	99	80	44.44	42	128	74.85	STAD	0.005	2.33E-07
FANCA	16	89813078	G	C	28	27	49.09	1	38	97.44	OV	0.047	2.70E-07
BRCA2	13	32907128	A	G	88	88	50	15	80	84.21	HNSC	0.005	2.76E-07
BRCA1	17	41247892	T	C	63	44	41.12	7	45	86.54	OV	0.014	4.41E-07
PARP3	3	51979205	C	T	55	54	49.54	15	87	85.29	LUSC	0.005	4.52E-07
MUTYH	1	45798309	C	T	138	108	43.72	176	342	65.9	OV	0.005	5.53E-07
PARP3	3	51978552	C	G	155	97	38.34	98	170	63.2	LUAD	0.014	9.77E-07
PARP3	3	51982404	G	A	39	38	49.35	13	86	86.87	OV	0.024	9.87E-07
BRCA1	17	41245233	A	G	130	106	44.92	72	169	70.12	LUAD	0.033	1.28E-06
MAP3K15	X	19380945	T	C	93	99	51.56	19	88	82.24	OV	0.005	1.70E-06
EPPK1	8	144942354	T	C	57	34	37.36	23	75	76.53	HNSC	0.014	1.78E-06
EME1	17	48456028	G	A	31	32	50.79	11	85	88.54	BRCA	0.033	2.13E-06
FANCM	14	45667962	G	T	187	189	50.27	9	51	83.61	OV	0.014	2.82E-06
BRCA2	13	32929291	A	C	100	123	55.16	22	103	82.4	OV	0.005	3.35E-06
DIS3	13	73351586	T	A	33	32	49.23	6	55	90.16	LUAD	0.005	3.57E-06
FANCA	16	89851334	G	C	73	60	45.11	4	34	89.47	BRCA	0.009	4.00E-06
BRCA1	17	41245176	A	G	376	353	48.36	213	366	63.21	LAML	0.005	5.07E-06
BRCA1	17	41251803	T	C	82	56	40.58	30	83	73.45	HNSC	0.038	6.27E-06
PALB2	16	23641001	C	G	36	27	42.86	11	60	84.51	STAD	0.005	6.90E-06
BRCA1	17	41243509	T	C	83	92	52.57	15	75	83.33	OV	0.489	7.70E-06
FANCA	16	89828357	C	T	44	40	47.62	5	42	89.36	BRCA	0.009	7.98E-06
BRCA1	17	41243509	T	C	80	71	47.02	28	94	77.05	HNSC	0.489	1.02E-05
PMS2	7	6045549	C	A	209	159	43.21	192	302	60.89	LGG	0.024	1.08E-05
BRCA2	13	32912763	C	G	59	44	42.72	40	116	73.42	LUSC	0.005	1.15E-05
POLK	5	74865291	A	G	35	26	42.62	7	45	86.54	LUAD	0.014	1.23E-05
EPPK1	8	144942462	C	T	34	28	45.16	10	59	85.51	BRCA	0.014	1.27E-05
FANCA	16	89877200	G	C	49	44	47.31	10	55	84.62	BRCA	0.038	1.61E-05
EME2	16	1826153	C	T	61	45	42.45	18	64	78.05	LUAD	0.005	1.86E-05
ATM	11	108117798	C	T	32	23	41.82	9	50	84.75	BRCA	0.014	2.26E-05
FANCG	9	35076015	C	T	72	65	47.45	48	139	74.33	STAD	0.005	2.39E-05
ATM	11	108115564	A	G	197	175	47.04	161	291	64.38	OV	0.005	2.65E-05
BRCA2	13	32953604	G	A	63	55	46.22	12	54	81.82	LGG	0.038	3.35E-05

MRE11A	11	94197302	T	C	66	85	55.92	9	60	86.96	LUSC	0.014	3.48E-05
BRCA1	17	41215948	G	A	42	30	41.67	17	64	79.01	OV	0.014	4.47E-05
EPPK1	8	144945071	C	T	145	155	51.67	50	137	72.87	UCEC	0.009	4.82E-05
ATM	11	108158399	A	G	28	27	49.09	2	29	93.55	HNSC	0.024	5.03E-05
PIK3C2G	12	18435145	G	A	91	95	51.08	5	36	87.8	OV	0.019	5.32E-05
DIS3	13	73333982	T	C	60	69	53.49	23	98	80.99	OV	0.005	5.29E-05
XRCC2	7	152346287	T	C	58	75	56.39	61	228	78.89	OV	0.043	5.26E-05
DIS3	13	73355036	C	T	157	160	50.47	90	204	69.39	BRCA	0.005	5.67E-05
DIS3	13	73340135	T	C	47	31	39.74	11	44	80	BRCA	0.019	5.80E-05
FANCA	16	89813078	G	C	40	27	40.3	14	54	79.41	GBM	0.047	6.14E-05
PARP3	3	51978529	C	T	210	176	45.48	75	143	65.3	STAD	0.014	7.61E-05
BRCA1	17	41246092	A	G	64	40	38.46	27	68	71.58	HNSC	0.038	8.01E-05
BRCA1	17	41245027	G	A	80	74	48.05	52	138	72.25	UCEC	0.247	8.40E-05
BRCA2	13	32900252	A	G	56	62	52.54	18	79	81.44	OV	0.009	9.92E-05
PARP3	3	51978503	G	T	149	117	43.98	95	170	64.15	LUAD	0.005	0.00011
FANCA	16	89813078	G	C	27	20	42.55	14	63	80.77	STAD	0.047	0.00013
BAP1	3	52438541	T	C	135	107	44.21	36	82	69.49	KIRC	0.005	0.00017
EPPK1	8	144942519	C	T	74	53	41.73	26	67	72.04	KIRC	0.005	0.00019
MUTYH	1	45798837	C	T	113	79	41.15	73	129	63.86	LUAD	0.005	0.00023
RAD51D	17	33434093	C	T	101	71	41.28	44	89	66.92	OV	0.014	0.00026
FANCM	14	45644409	A	G	45	46	50.55	18	74	80.43	BRCA	0.014	0.00026
RAD51D	17	33428027	A	T	78	67	46.21	22	66	74.16	LUSC	0.028	0.00028
RAD51C	17	56772310	C	T	35	45	56.25	23	111	82.84	OV	0.005	0.00032
XRCC2	7	152346287	T	C	173	141	44.9	146	235	61.52	KIRC	0.043	0.00035
ATM	11	108203541	C	T	65	47	41.96	20	56	72.73	BRCA	0.005	0.00034
FANCM	14	45642287	A	T	70	56	44.09	21	61	73.49	OV	0.005	0.00036
RAD51D	17	33428027	A	T	59	68	53.54	10	51	82.26	BRCA	0.028	0.00037
ATM	11	108205832	T	C	59	70	54.26	22	87	79.82	LUAD	0.009	0.00037
EPPK1	8	144942329	C	T	41	43	50.59	17	72	80	HNSC	0.009	0.00040
BRCA2	13	32912007	C	T	65	39	37.5	31	65	67.71	KIRC	0.047	0.00054
ATM	11	108137953	A	C	65	45	40.91	30	70	70	BRCA	0.005	0.00055
BRCA1	17	41243509	T	C	225	228	50.22	103	199	65.89	BRCA	0.489	0.00055
BRCA1	17	41226488	C	A	24	36	60	22	131	85.62	BRCA	0.285	0.00062
PMS2	7	6027143	G	A	50	39	43.33	27	74	73.27	LUAD	0.005	0.00072
FANCM	14	45645949	C	T	96	78	44.83	44	95	68.35	KIRC	0.038	0.00072
FANCG	9	35078184	C	T	111	68	37.57	25	50	66.67	STAD	0.009	0.00082
MAP3K15	X	19380887	C	G	133	75	36.06	62	88	58.67	HNSC	0.014	0.00084
RAD51C	17	56774155	T	C	60	68	53.12	11	50	81.97	UCEC	0.019	0.00084
FANCM	14	45642337	A	G	61	54	46.55	12	44	77.19	STAD	0.009	0.00087
FANCM	14	45665709	A	C	65	48	42.48	40	88	68.75	LUAD	0.005	0.00093
PALB2	16	23632740	A	G	45	32	41.56	49	111	68.94	OV	0.014	0.00108
ATM	11	108203580	A	G	50	36	41.86	14	43	75.44	LUSC	0.005	0.00110
FANCC	9	97879600	G	C	25	36	59.02	4	37	90.24	OV	0.005	0.00148
BRCA2	13	32912190	C	T	33	28	45.9	11	44	80	LUSC	0.005	0.00159
BRCA1	17	41226488	C	A	30	29	49.15	8	40	83.33	HNSC	0.285	0.00164
BRCA2	13	32937333	A	G	73	65	47.1	24	65	73.03	GBM	0.047	0.00166
ATM	11	108153488	A	G	81	35	30.17	44	59	57.28	LUAD	0.009	0.00191
BRCA2	13	32913562	A	C	54	56	50.91	14	52	78.79	BRCA	0.033	0.00205
BRCA2	13	32930613	T	C	321	208	39.32	84	107	55.73	LAML	0.005	0.00225
BRCA1	17	41226488	C	A	82	73	47.1	20	55	73.33	BRCA	0.285	0.00226
BRCA2	13	32893369	G	C	64	48	42.86	15	41	71.93	LUSC	0.028	0.00285
DDX11	12	31255911	C	T	48	31	39.24	25	56	69.14	BRCA	0.019	0.00289
BRCA1	17	41245714	T	C	65	74	53.24	31	93	75	OV	0.005	0.00292
ATM	11	108236150	G	A	164	163	49.85	63	126	66.67	OV	0.028	0.00353
ATM	11	108168053	A	G	116	58	33.33	74	87	54.04	BRCA	0.019	0.00439
MSH6	2	48027683	A	T	57	51	47.22	32	78	70.91	LUSC	0.038	0.00558
ATM	11	108141988	T	C	136	96	41.38	136	182	57.23	BRCA	0.043	0.00609
DDX11	12	31249614	G	A	33	43	56.58	50	171	77.38	BRCA	0.028	0.00638
RAD50	5	131939181	G	C	33	21	38.89	18	45	71.43	STAD	0.028	0.00656

ATM	11	108203612	T	G	83	59	41.55	43	76	63.87	BRCA	0.009	0.00662
ATM	11	108121787	G	A	89	57	39.04	51	79	60.31	HNSC	0.009	0.00724
ATM	11	108115522	A	G	150	173	53.4	37	91	71.09	BRCA	0.019	0.00725
BRCA1	17	41223021	G	A	37	43	53.75	10	43	81.13	OV	0.005	0.00744
BRCA1	17	41223048	A	G	32	40	55.56	18	71	79.78	BRCA	0.057	0.00764
ATM	11	108128246	T	A	46	53	53.54	13	49	79.03	HNSC	0.043	0.00761
BRCA1	17	41246164	C	T	88	82	48.24	199	346	63.25	LUSC	0.009	0.00779
FANCC	9	97873856	C	G	133	111	45.49	55	96	63.16	BRCA	0.005	0.00832
BRCA1	17	41223048	A	G	46	38	45.24	81	163	66.8	BRCA	0.057	0.00879
CNKSR1	1	26515965	C	T	41	68	61.82	14	68	82.93	LGG	0.024	0.00880
BRCA2	13	32913947	C	T	27	26	49.06	8	34	80.95	BRCA	0.014	0.00904
FANCM	14	45605730	G	A	41	31	43.06	30	68	68.69	BRCA	0.009	0.00901
ERCC2	19	45854910	C	G	172	159	47.32	201	304	59.96	STAD	0.033	0.01000
PALB2	16	23647569	G	A	64	60	48.39	16	45	73.77	BRCA	0.028	0.01025
ATM	11	108121632	A	C	32	27	45.76	14	43	75.44	OV	0.009	0.01107
FANCM	14	45623169	G	A	114	90	44.12	16	37	69.81	GBM	0.005	0.01118
ATM	11	108183194	A	C	120	74	38.14	201	228	53.15	BRCA	0.028	0.01237
BRCA2	13	32911818	C	T	60	39	39	29	53	64.63	KIRC	0.005	0.01262
XPC	3	14220032	C	G	50	27	35.06	31	51	62.2	LUAD	0.043	0.01266
DDX11	12	31231421	C	T	42	29	40.85	13	33	71.74	OV	0.028	0.01316
FANCA	16	89836623	C	G	32	37	53.62	9	38	80.85	BRCA	0.014	0.01421
FANCM	14	45645955	A	C	100	91	47.64	70	126	63.96	OV	0.019	0.01446
FANCA	16	89838136	T	C	164	148	47.44	99	157	61.33	BRCA	0.014	0.01520
EME1	17	48453514	C	T	48	47	49.47	53	120	69.36	OV	0.005	0.01670
CNKSR1	1	26515361	C	T	69	62	46.27	53	104	66.24	OV	0.005	0.01681
RAD50	5	131924538	A	G	200	144	41.86	74	99	57.23	GBM	0.019	0.01837
FANCM	14	45665637	A	G	136	126	48.09	68	118	63.44	OV	0.005	0.01838
FANCM	14	45644691	A	G	122	97	44.29	71	109	60.56	LUSC	0.009	0.01982
DDX11	12	31236777	C	T	150	116	43.61	76	110	59.14	LUAD	0.005	0.01974
XPC	3	14212018	T	A	78	61	43.88	16	36	69.23	BRCA	0.005	0.02122
XPC	3	14199801	T	C	39	36	48	32	77	70.64	KIRC	0.005	0.02182
MSH6	2	48027541	G	A	194	201	50.89	148	241	61.79	GBM	0.005	0.02372
MRE11A	11	94169056	C	T	82	58	41.43	53	81	60.45	LGG	0.005	0.02680
EME1	17	48453272	G	A	73	44	37.61	52	72	58.06	LUAD	0.005	0.02729
FANCA	16	89839732	G	T	28	32	53.33	10	38	79.17	BRCA	0.005	0.02940
PIK3C2G	12	18435620	A	T	144	167	53.7	80	157	66.24	LAML	0.024	0.03184
XPC	3	14199969	C	T	68	70	50.36	72	144	66.36	UCEC	0.009	0.03171
DDX11	12	31249614	G	A	35	20	36.36	29	50	63.29	STAD	0.028	0.03361
EPPK1	8	144946503	C	T	93	78	45.61	41	72	63.72	STAD	0.005	0.03524
BRCA1	17	41243509	T	C	69	58	45.67	70	118	62.77	STAD	0.489	0.03633
EPPK1	8	144947076	C	T	36	17	32.08	27	41	60.29	STAD	0.005	0.03625
DIS3	13	73333944	G	A	61	69	53.08	11	35	74.47	UCEC	0.019	0.03755
ATM	11	108192118	G	T	138	142	50.71	125	208	62.28	UCEC	0.033	0.04014
ATM	11	108183194	A	C	92	109	54.23	58	125	68.31	KIRC	0.028	0.04226
BRCA1	17	41246411	A	C	126	119	48.57	60	102	62.96	PRAD	0.147	0.04925
EPPK1	8	144941205	G	A	46	47	50.54	50	106	67.95	LUAD	0.028	0.06117
MRE11A	11	94180388	T	C	114	137	54.58	54	112	67.47	BRCA	0.005	0.07082
MAP3K15	X	19392700	C	T	84	68	44.74	27	48	64	GBM	0.014	0.07055
ERCC2	19	45855574	G	A	70	72	50.7	42	84	66.67	GBM	0.014	0.07494
BRIP1	17	59934481	C	T	157	158	50.16	129	199	60.67	BRCA	0.033	0.07950
BRCA1	17	41251803	T	C	130	84	39.07	125	136	52.11	BRCA	0.038	0.08176
FANCG	9	35075302	C	T	63	65	50.78	21	48	69.57	BRCA	0.009	0.08444
MSH6	2	48033735	G	C	112	113	50.22	111	180	61.64	BRCA	0.005	0.08434
RAD50	5	131940620	C	T	52	27	34.18	39	49	55.68	GBM	0.043	0.08546
MSH6	2	48027037	G	A	136	110	44.72	117	153	56.67	LGG	0.009	0.08513
EPPK1	8	144942246	T	C	57	46	44.66	38	65	63.11	BRCA	0.009	0.08610
BRCA2	13	32944563	G	A	0	362	100	1	444	99.33	BRCA	0.005	0.09000
FANCM	14	45658156	G	A	1	187	99.47	0	117	100	BRCA	0.038	0.09830
FANCM	14	45644706	A	G	110	101	47.87	51	83	61.94	OV	0.047	0.10640

ATM	11	108119760	T	C	182	163	47.25	110	151	57.85	UCEC	0.005	0.10708
RAD50	5	131939072	G	A	92	72	43.9	88	119	57.49	GBM	0.019	0.11089
MAP3K15	X	19413274	G	A	108	60	35.5	152	144	48.65	LUAD	0.005	0.11062
ATM	11	108163377	A	G	281	188	40	264	249	48.44	KIRC	0.005	0.11341
FANCA	16	89836366	T	C	41	36	46.75	15	33	68.75	STAD	0.014	0.12526
FANCG	9	35078282	C	G	51	48	48.48	23	47	67.14	KIRC	0.009	0.12747
BRCA2	13	32972575	G	A	31	42	57.53	20	59	74.68	LUAD	0.024	0.12730
PALB2	16	23646867	A	C	101	123	54.91	26	59	69.41	BRCA	0.009	0.12835
PMS2	7	6026906	C	T	81	80	49.38	37	68	64.76	STAD	0.019	0.12801
EPPK1	8	144946799	C	T	24	22	47.83	17	41	70.69	STAD	0.028	0.12778
RAD50	5	131973868	G	A	75	46	38.02	30	40	57.14	LUAD	0.005	0.13377
PMS2	7	6026840	T	C	0	32	100	0	35	100	BRCA	0.005	0.14086
ATM	11	108138061	G	C	0	65	100	0	97	100	LUAD	0.005	0.14028
FANCC	9	97879669	G	A	0	32	100	0	37	97.37	PRAD	0.009	0.13970
ATM	11	108202273	G	A	49	41	45.56	33	57	63.33	STAD	0.005	0.14933
ERCC2	19	45854910	C	G	128	120	47.62	93	136	58.62	KIRC	0.033	0.14891
BRCA1	17	41245546	G	A	45	35	43.75	17	32	65.31	OV	0.005	0.14940
XRCC2	7	152346389	G	T	70	74	51.39	67	120	63.83	STAD	0.014	0.14899
RAD51C	17	56770018	C	T	48	24	33.33	44	50	52.63	KIRC	0.014	0.14928
BRCA2	13	32913898	A	C	62	69	52.67	86	159	64.9	BRCA	0.014	0.15380
BRIP1	17	59924502	T	C	166	158	48.77	198	267	57.42	BRCA	0.014	0.15401
BRCA2	13	32937554	G	A	152	157	50.81	124	187	60.13	UCEC	0.009	0.16263
DDX11	12	31242979	C	A	47	39	45.35	37	62	61.39	BRCA	0.033	0.16256
BRCA2	13	32954037	A	C	50	31	38.27	20	30	60	BRCA	0.009	0.16614
FANCM	14	45645426	G	A	36	21	36.84	19	29	60.42	BRCA	0.009	0.17201
MUTYH	1	45798269	T	C	64	67	51.15	53	95	63.76	OV	0.024	0.19622
XRCC2	7	152346287	T	C	73	82	52.9	75	135	64.29	STAD	0.043	0.19586
BRCA2	13	32929222	A	C	92	68	42.24	63	80	55.94	OV	0.005	0.19756
EME1	17	48453487	G	A	35	25	41.67	28	45	61.64	HNSC	0.009	0.19740
FANCM	14	45605738	G	C	33	31	47.69	15	33	68.75	LUAD	0.014	0.19847

Supplementary Table 2.13: LOH analysis of rare missense variants for discovering hotspot clusters.

Rank	Region	P-Value	FDR
RANK 1:	(EPPK1: 1635 to 1726)	0.000103934	0.003118018
RANK 2:	(PARP3: 144 to 283)	0.000307675	0.00461512
RANK 3:	(ATM: 2459 to 2717)	0.000341543	0.003415432
RANK 4:	(BRCA1: 61 to 246)	0.000521565	0.003911739
RANK 5:	(ERCC2: 695 to 754)	0.000642717	0.003856304
RANK 6:	(FANCM: 730 to 918)	0.000980384	0.004901922
RANK 7:	(XRCC2: 61 to 124)	0.001459117	0.00625336
RANK 8:	(FANCA: 951 to 1172)	0.002081455	0.007805457
RANK 9:	(MRE11A: 404 to 649)	0.002114544	0.00704848
RANK 10:	(FANCM: 1331 to 1362)	0.002505345	0.007516036
RANK 11:	(EME1: 211 to 326)	0.002879257	0.007852518
RANK 12:	(BRCA1: 612 to 857)	0.006289302	0.015723254
RANK 13:	(FANCA: 654 to 795)	0.008787266	0.020278307
RANK 14:	(RAD50: 1191 to 1264)	0.010473836	0.022443933
RANK 15:	(PMS2: 418 to 519)	0.011760717	0.023521434
RANK 16:	(MAP3K15: 539 to 738)	0.01202091	0.022539206
RANK 17:	(FANCG: 106 to 279)	0.012120043	0.021388311
RANK 18:	(FANCC: 334 to 407)	0.015479323	0.025798871
RANK 19:	(XPC: 13 to 112)	0.015946328	0.025178413
RANK 20:	(EPPK1: 2073 to 2198)	0.025175917	0.037763876
RANK 21:	(FANCM: 1868 to 1945)	0.027610014	0.039442877
RANK 22:	(RAD50: 763 to 884)	0.038359536	0.052308458
RANK 23:	(BRCA1: 1579 to 1788)	0.040442038	0.052750484
RANK 24:	(BRCA2: 2665 to 2786)	0.049978996	0.062473745
RANK 25:	(BRCA2: 1109 to 1234)	0.054021647	0.064825977
RANK 26:	(ATM: 337 to 532)	0.05986289	0.069072566
RANK 27:	(ATM: 978 to 1113)	0.060714125	0.067460139
RANK 28:	(EPPK1: 116 to 307)	0.070785715	0.075841838
RANK 29:	(MAP3K15: 897 to 1048)	0.088932652	0.091999295
RANK 30:	(BRCA2: 2969 to 3124)	0.088986997	0.088986997

Supplementary Table 2.14: Somatic mutations discovered in 3,368 out of 4,034 cancer cases.

The file is too large to display here, it is hosted by *Nature Communications* website:
http://www.nature.com/ncomms/2015/151209/ncomms10086/extref/ncomms10086-s15.xlsx

Supplementary Table 2.15: Somatic and Germline Mutation Relationship (mutual exclusive/co-occuring) across 12 cancer types. The genes used are 34 burden test significant genes and recurrent mutated genes (>= 5 somatic mutations across all cancer types).

Co-occurrence Test

Gene1	Gene2	Count of Gene1	Count of Gene2	Co-occur Count	Mutually exclusivity Count	Permutation co-occur Count	Permutation mutually exclusivity Count	P-value of Co-occur	P-value of exclusivity
TP53	*BRCA1*	708	53	29	703	13.3563	734.2874	0	1
RPL22	*MSH6*	27	11	2	34	0.1102	37.7796	0.004	1
DDX11	*PMS2*	5	5	1	8	0.0089	9.9822	0.0089	1
CHD4	*MRE11A*	6	4	1	8	0.0091	9.9818	0.0091	1
ALPK2	*PMS2*	5	5	1	8	0.0097	9.9806	0.0096	0.9999
FAT3	*FANCG*	5	6	1	9	0.0106	10.9788	0.0106	1
VHL	*EME1*	11	3	1	12	0.0125	13.975	0.0124	0.9999
DDX11	*FANCA*	5	8	1	11	0.0141	12.9718	0.014	0.9999
IDH1	*HLAG*	198	3	2	197	0.2217	200.5566	0.015	0.9994
ERBB3	*XRCC2*	6	7	1	11	0.0162	12.9676	0.0162	1
CHD4	*XRCC2*	6	7	1	11	0.0168	12.9664	0.0166	0.9998
ERBB3	*FANCA*	6	8	1	12	0.0175	13.965	0.0173	0.9998
CHEK2	*HLAG*	18	3	1	19	0.0225	20.955	0.0224	0.9999
VEZF1	*DDX11*	6	12	1	16	0.0266	17.9468	0.0264	0.9998
NFE2L2	*RAD50*	10	8	1	16	0.0299	17.9402	0.0299	1
CHEK2	*FANCG*	18	6	1	22	0.0381	23.9238	0.0376	0.9995
PPP2R1A	*DDX11*	8	12	1	18	0.038	19.924	0.0376	0.9996
PIK3R1	*RAD50*	13	8	1	19	0.0383	20.9234	0.0377	0.9994
DNMT3A	*MRE11A*	28	4	1	30	0.042	31.916	0.0418	0.9998
IDH1	*PMS2*	198	5	2	199	0.3703	202.2594	0.0449	0.9962
VHL	*DDX11*	11	12	1	21	0.0486	22.9028	0.0476	0.999
CTNNB1	*PIK3C2G*	49	19	2	64	0.3582	67.2836	0.0478	0.995
RPL22	*PMS2*	27	5	1	30	0.0498	31.9004	0.0489	0.9991

Mutual Exclusivity Test

Gene1	Gene2	Count of Gene1	Count of Gene2	Co-occur Count	Mutually exclusivity Count	Permutation co-occur Count	Permutation mutually exclusivity Count	P-value of Co-occur	P-value of exclusivity
PIK3CA	*BRCA1*	444	53	1	495	8.4768	480.0464	1	2.00E-04
IDH1	*BRCA1*	198	53	0	251	3.85	243.3	1	0.0152
IDH1	*BRCA2*	198	50	0	248	3.6275	240.745	1	0.0213
PIK3CA	*BRCA2*	444	50	3	488	7.9735	478.053	0.9902	0.0294
PIK3CA	*PIK3C2G*	444	19	0	463	3.0323	456.9354	1	0.0367

*Gene1 indicates gene with recurrent somatic mutations
*Gene2 indicates gene with rare germline truncation variants

Supplementary Table 2.16: Somatic and Germline Mutation Relationship (mutual exclusive/co-occuring) for individual cancer types. The genes used are 34 burden test significant genes and recurrent mutated genes (>=2 somatic mutations in particular cancer type).

Co-occurrence Test

Gene1	Gene2	Count of Gene1	Count of Gene2	Co-occur Count	Mutually exclusivity Count	Permutation co-occur Count	Permutation mutually exclusivity Count	P-value of Co-occur	P-value of exclusivity	Cancer Type
FANCB	*DIS3*	1	2	1	1	0.0024	2.9952	0.0024	1	BRCA
CARD11	*PIK3C2G*	1	6	1	5	0.0076	6.9848	0.0076	1	BRCA
WNK1	*XRCC2*	3	2	1	3	0.0079	4.9842	0.0079	1	BRCA
TP53	*BRCA1*	185	12	7	183	2.9616	191.0768	0.0118	0.9977	BRCA
VEZF1	*DDX11*	5	2	1	5	0.0135	6.973	0.0135	1	BRCA
KRAS	*MAP3K15*	5	2	1	5	0.0143	6.9714	0.0142	0.9999	BRCA
KRAS	*EPPK1*	5	3	1	6	0.0225	7.955	0.0222	0.9997	BRCA
POLD1	*BRCA2*	1	19	1	18	0.0227	19.9546	0.0227	1	BRCA
CYP2D6	*FANCM*	2	8	1	8	0.0228	9.9544	0.0228	1	BRCA
WNK1	*PIK3C2G*	3	6	1	7	0.0241	8.9518	0.024	0.9999	BRCA
FGFR3	*BRCA1*	2	12	1	12	0.0321	13.9358	0.0319	0.9998	BRCA
KDM6A	*BRCA1*	3	12	1	13	0.0485	14.903	0.0476	0.9991	BRCA
KDR	*ERCC2*	1	1	1	0	0.0049	1.9902	0.0049	1	GBM
FAM46C	*MAP3K15*	1	1	1	0	0.0049	1.9902	0.0049	1	GBM
FLT1	*CYP1B1*	1	2	1	1	0.0093	2.9814	0.0093	1	GBM
NF1	*MSH6*	4	2	1	4	0.0366	5.9268	0.0366	1	GBM
APC	*ATM*	1	1	1	0	0.0029	1.9942	0.0029	1	HNSC
FAT3	*FANCG*	2	1	1	1	0.0079	2.9842	0.0079	1	HNSC
INPPL1	*MSH6*	2	1	1	1	0.0084	2.9832	0.0084	1	HNSC
ARID5B	*PALB2*	1	2	1	1	0.009	2.982	0.009	1	HNSC
CRKL	*CNKSR1*	1	2	1	1	0.0092	2.9816	0.0092	1	HNSC
SETDB1	*FANCM*	1	3	1	2	0.0116	3.9768	0.0116	1	HNSC
INHBA	*FANCM*	1	3	1	2	0.013	3.974	0.013	1	HNSC
MAPK1	*DDX11*	3	1	1	2	0.0159	3.9682	0.0159	1	HNSC
DDX11	*PMS2*	2	2	1	2	0.018	3.964	0.018	1	HNSC
DDX11	*FANCA*	2	3	1	3	0.0246	4.9508	0.0246	1	HNSC
NCOR1	*CNKSR1*	4	2	1	4	0.0347	5.9306	0.0344	0.9997	HNSC
NFE2L2	*RAD50*	10	1	1	9	0.0427	10.9146	0.0427	1	HNSC
ACO1	*ERCC2*	1	2	1	1	0.005	2.99	0.005	1	KIRC
CRIPAK	*PARP3*	1	3	1	2	0.0078	3.9844	0.0078	1	KIRC
TP53	*BAP1*	5	2	1	5	0.0242	6.9516	0.0242	1	KIRC
TP53	*FANCM*	5	2	1	5	0.0256	6.9488	0.0254	0.9998	KIRC
PBRM1	*DDX11*	16	1	1	15	0.0347	16.9306	0.0347	1	KIRC
CHD4	*MRE11A*	1	1	1	0	0.0051	1.9898	0.0051	1	LAML
SOS1	*FANCC*	1	2	1	1	0.0118	2.9764	0.0118	1	LAML
CHD4	*MUTYH*	1	2	1	1	0.0075	2.985	0.0075	1	LGG
SMARCA4	*HLAG*	3	2	1	3	0.0266	4.9468	0.0264	0.9998	LGG
PTEN	*BRIP1*	7	1	1	6	0.0275	7.945	0.0275	1	LGG
CCND2	*DIS3*	1	3	1	2	0.0198	3.9604	0.0198	1	LUAD
BRWD3	*DIS3*	2	3	1	3	0.04	4.92	0.0394	0.9994	LUAD
EML4	*DIS3*	1	1	1	0	0.0096	1.9808	0.0096	1	LUSC
FAT3	*BRCA2*	1	2	1	1	0.0225	2.955	0.0225	1	LUSC
IGF1	*BRCA2*	1	2	1	1	0.0247	2.9506	0.0247	1	LUSC
SMAD4	*FANCA*	1	3	1	2	0.0345	3.931	0.0345	1	LUSC
MAPK8IP1	*BRIP1*	1	3	1	2	0.0064	3.9872	0.0064	1	OV
NRAS	*BRIP1*	2	3	1	3	0.0146	4.9708	0.0146	1	OV
TP53	*BRCA2*	315	25	24	292	19.8331	300.3338	0.0147	0.9984	OV
B2M	*RAD51D*	2	3	1	3	0.0155	4.969	0.0155	1	OV
RB1	*PALB2*	3	3	1	4	0.0238	5.9524	0.0237	0.9999	OV

NRAS	*PIK3C2G*	2	5	1	5	0.0243	6.9514	0.0242	0.9999	OV
INHBA	*EPPK1*	1	1	1	0	0.0071	1.9858	0.0071	1	PRAD
GUCY1A2	*EPPK1*	1	1	1	0	0.0096	1.9808	0.0096	1	PRAD
CHEK2	*FANCG*	3	1	1	2	0.0206	3.9588	0.0206	1	PRAD
CRIPAK	*ATM*	1	5	1	4	0.0377	5.9246	0.0377	1	PRAD
PIK3CA	*BRCA1*	5	1	1	4	0.0379	5.9242	0.0379	1	PRAD
ERCC3	*MUTYH*	1	1	1	0	0.0136	1.9728	0.0136	1	STAD
EIF3A	*FANCC*	1	1	1	0	0.0148	1.9704	0.0148	1	STAD
AURKB	*MUTYH*	1	1	1	0	0.0152	1.9696	0.0152	1	STAD
BUB1B	*FANCC*	1	1	1	0	0.0156	1.9688	0.0156	1	STAD
AKT3	*DIS3*	1	1	1	0	0.0164	1.9672	0.0164	1	STAD
TMPRSS2	*MUTYH*	1	1	1	0	0.0184	1.9632	0.0184	1	STAD
GPR124	*FANCC*	2	1	1	1	0.0266	2.9468	0.0266	1	STAD
BCOR	*HLAG*	2	1	1	1	0.027	2.946	0.027	1	STAD
BCL6	*MUTYH*	2	1	1	1	0.028	2.944	0.028	1	STAD
HSP90AB1	*MUTYH*	2	1	1	1	0.028	2.944	0.028	1	STAD
DDR2	*EPPK1*	1	2	1	1	0.0292	2.9416	0.0292	1	STAD
MAP4K1	*DIS3*	2	1	1	1	0.0292	2.9416	0.0292	1	STAD
B2M	*DIS3*	2	1	1	1	0.0302	2.9396	0.0302	1	STAD
TSC2	*MUTYH*	2	1	1	1	0.0312	2.9376	0.0312	1	STAD
ATR	*MUTYH*	2	1	1	1	0.0328	2.9344	0.0328	1	STAD
SMARCB1	*MUTYH*	3	1	1	2	0.0412	3.9176	0.0412	1	STAD
SOS1	*XRCC2*	1	3	1	2	0.0416	3.9168	0.0416	1	STAD
ARID5B	*DIS3*	3	1	1	2	0.0426	3.9148	0.0426	1	STAD
HGF	*FANCA*	3	1	1	2	0.0448	3.9104	0.0448	1	STAD
EWSR1	*DIS3*	3	1	1	2	0.0452	3.9096	0.0452	1	STAD
NOTCH3	*MUTYH*	3	1	1	2	0.0468	3.9064	0.0468	1	STAD
GRM3	*FANCA*	3	1	1	2	0.0476	3.9048	0.0476	1	STAD
EPHB1	*FANCA*	3	1	1	2	0.0492	3.9016	0.0492	1	STAD
BRE	*RAD50*	2	1	1	1	0.0462	2.9076	0.0462	1	UCEC

Mutual Exclusivity Test

Gene1	***Gene2***	**Count of Gene1**	**Count of Gene2**	**Co-occur Count**	**Mutually exclusivity Count**	**Permutation co-occur Count**	**Permutation mutually exclusivity Count**	**P-value of Co-occur**	**P-value of exclusivity**	**Cancer Type**
PIK3CA	*BRCA1*	235	12	0	247	3.7578	239.4844	1	0.0101	BRCA
PIK3CA	*BRCA2*	235	19	2	250	5.9606	242.0788	0.9944	0.0297	BRCA
TP53	*ATM*	96	5	0	101	2.2913	96.4174	1	0.0412	LUAD
TP53	*PALB2*	88	4	0	92	2.4414	87.1172	1	0.0224	STAD
PIK3CA	*BRCA1*	116	3	0	119	2.2264	114.5472	1	0.0126	UCEC
PTEN	*BRCA1*	119	3	0	122	2.2404	117.5192	1	0.0138	UCEC

*Gene1 indicates gene with recurrent somatic mutations

*Gene2 indicates gene with rare germline truncation variants

Supplementary Table 2.17: Distribution of *BRCA1*, *BRCA2*, and *ATM* germline truncation variants and somatic mutations across 12 cancer types.

Type	Gene	BRCA	GBM	HNSC	KIRC	LAML	LGG	LUAD	LUSC	OV	PRAD	STAD	UCEC
Somatic	ATM	1.82%	0.75%	1.72%	2.43%	0.00%	0.00%	4.11%	1.55%	1.63%	5.62%	7.48%	11.69%
	BRCA1	1.17%	1.12%	2.41%	0.88%	0.00%	0.00%	1.95%	2.07%	4.90%	0.56%	4.05%	4.84%
	BRCA2	1.56%	1.12%	3.44%	1.55%	0.00%	0.45%	2.81%	2.59%	2.56%	1.69%	5.61%	9.68%
Germline Truncation	ATM	0.52%	0.00%	0.34%	0.00%	0.50%	0.45%	1.08%	0.00%	0.23%	2.81%	1.25%	0.00%
	BRCA1	1.56%	0.00%	0.00%	0.22%	0.00%	0.00%	0.22%	0.00%	7.93%	0.56%	0.31%	1.21%
	BRCA2	2.47%	0.00%	0.34%	0.22%	0.00%	0.00%	0.00%	1.04%	5.83%	0.00%	0.62%	0.00%

Supplementary Table 2.18: Genes with rare germline truncation variants associated with somatic mutation frequencies.

Cancer Type	Gene	Clinic Feature	Method	Num of Cases in each cancer type	Statistic	P-value
BRCA	*BRCA1*	Somatic Mutation Frequency	wilcox	758	1.3150E+03	2.6613E-05
OV	*BRCA2*	Somatic Mutation Frequency	wilcox	426	2.6150E+03	5.9785E-05
BRCA	*BRCA2*	Somatic Mutation Frequency	wilcox	758	3.9915E+03	1.3088E-03
OV	*BRCA1*	Somatic Mutation Frequency	wilcox	426	4.8970E+03	1.0307E-02
OV	*RAD51D*	Somatic Mutation Frequency	wilcox	426	1.6050E+02	2.5850E-02
OV	*RAD51C*	Somatic Mutation Frequency	wilcox	426	6.1000E+01	3.6887E-02
OV	*ERCC2*	Somatic Mutation Frequency	wilcox	426	0.0000E+00	8.4695E-02
OV	*CYP1B1*	Somatic Mutation Frequency	wilcox	426	1.5500E+01	1.1004E-01
OV	*BRIP1*	Somatic Mutation Frequency	wilcox	426	9.8650E+02	9.8071E-02
UCEC	*MSH6*	Somatic Mutation Frequency	wilcox	248	6.5000E+01	1.4462E-02
BRCA	*PIK3C2G*	Somatic Mutation Frequency	wilcox	758	2.9180E+03	3.3905E-02
OV	*PIK3C2G*	Somatic Mutation Frequency	wilcox	426	5.5600E+02	6.9914E-02
OV	*PALB2*	Somatic Mutation Frequency	wilcox	426	2.6750E+02	8.4550E-02
UCEC	*BRIP1*	Somatic Mutation Frequency	wilcox	248	2.4050E+02	1.0365E-01
KIRC	*EME1*	Somatic Mutation Frequency	wilcox	387	3.5000E+01	2.6730E-02
OV	*DIS3*	Somatic Mutation Frequency	wilcox	426	6.0500E+01	2.1793E-01
OV	*CNKSR1*	Somatic Mutation Frequency	wilcox	426	5.5600E+02	2.4072E-01
UCEC	*RAD50*	Somatic Mutation Frequency	wilcox	248	2.0000E+01	1.5020E-01
UCEC	*DDX11*	Somatic Mutation Frequency	wilcox	248	4.1150E+02	1.0244E-01
OV	*MAP3K15*	Somatic Mutation Frequency	wilcox	426	8.3500E+01	2.9602E-01
STAD	*EME2*	Somatic Mutation Frequency	wilcox	220	3.4550E+02	1.5640E-01
LGG	*MUTYH*	Somatic Mutation Frequency	wilcox	217	1.0850E+02	2.3030E-01
LGG	*DDX11*	Somatic Mutation Frequency	wilcox	217	5.1500E+01	3.7123E-01
LGG	*HLA_G*	Somatic Mutation Frequency	wilcox	217	1.0000E+02	1.9506E-01
UCEC	*CNKSR1*	Somatic Mutation Frequency	wilcox	248	7.7000E+01	9.5363E-02
LGG	*BRIP1*	Somatic Mutation Frequency	wilcox	217	4.8500E+01	3.4616E-01
STAD	*EPPK1*	Somatic Mutation Frequency	wilcox	220	8.8000E+01	1.4840E-01
LGG	*PMS2*	Somatic Mutation Frequency	wilcox	217	3.0600E+02	3.0576E-01
LGG	*CYP1B1*	Somatic Mutation Frequency	wilcox	217	1.8500E+01	1.5529E-01
KIRC	*BRCA2*	Somatic Mutation Frequency	wilcox	387	5.4500E+01	2.1664E-01
UCEC	*MRE11A*	Somatic Mutation Frequency	wilcox	248	4.3150E+02	6.7090E-02
LUAD	*DIS3*	Somatic Mutation Frequency	wilcox	228	3.0000E+00	9.4662E-02
STAD	*ATM*	Somatic Mutation Frequency	wilcox	220	7.1250E+02	2.6437E-02
UCEC	*PMS2*	Somatic Mutation Frequency	wilcox	248	3.9000E+01	2.4063E-01
UCEC	*PIK3C2G*	Somatic Mutation Frequency	wilcox	248	1.9150E+02	3.4573E-01
KIRC	*EPPK1*	Somatic Mutation Frequency	wilcox	387	3.0900E+02	3.0111E-01
OV	*ATM*	Somatic Mutation Frequency	wilcox	426	1.0800E+02	3.9768E-01
KIRC	*ERCC2*	Somatic Mutation Frequency	wilcox	387	2.1050E+02	2.7004E-01
OV	*PARP3*	Somatic Mutation Frequency	wilcox	426	3.0800E+02	4.3977E-01
BRCA	*FANCC*	Somatic Mutation Frequency	wilcox	758	3.0000E+01	1.1171E-01
LAML	*DIS3*	Somatic Mutation Frequency	wilcox	196	7.1500E+01	1.2580E-01
KIRC	*XPC*	Somatic Mutation Frequency	wilcox	387	8.8500E+01	3.5180E-01
UCEC	*ERCC2*	Somatic Mutation Frequency	wilcox	248	5.3500E+01	3.3162E-01
LGG	*PARP3*	Somatic Mutation Frequency	wilcox	217	2.0700E+02	1.1577E-01
BRCA	*RAD51C*	Somatic Mutation Frequency	wilcox	758	6.1000E+01	1.4737E-01
STAD	*DIS3*	Somatic Mutation Frequency	wilcox	220	1.7000E+01	1.4743E-01
KIRC	*BAP1*	Somatic Mutation Frequency	wilcox	387	2.5750E+02	4.2080E-01
KIRC	*FANCM*	Somatic Mutation Frequency	wilcox	387	1.8850E+02	2.1408E-01

STAD	*PALB2*	Somatic Mutation Frequency	wilcox	220	1.5700E+02	1.2495E-01
UCEC	*MUTYH*	Somatic Mutation Frequency	wilcox	248	1.9450E+02	3.2471E-01
HNSC	*FANCA*	Somatic Mutation Frequency	wilcox	238	5.5950E+02	8.1384E-02
KIRC	*CYP1B1*	Somatic Mutation Frequency	wilcox	387	1.2250E+02	5.3085E-01
LGG	*EPPK1*	Somatic Mutation Frequency	wilcox	217	3.8100E+02	5.8159E-01
KIRC	*BRCA1*	Somatic Mutation Frequency	wilcox	387	3.0000E+00	8.9776E-02
KIRC	*RAD50*	Somatic Mutation Frequency	wilcox	387	2.6950E+02	4.9624E-01
HNSC	*FANCM*	Somatic Mutation Frequency	wilcox	238	4.2000E+01	4.5960E-02
LAML	*PIK3C2G*	Somatic Mutation Frequency	wilcox	196	1.4500E+02	4.0546E-01
STAD	*FANCA*	Somatic Mutation Frequency	wilcox	220	7.0000E+00	1.0825E-01
LUSC	*MAP3K15*	Somatic Mutation Frequency	wilcox	82	7.9000E+01	1.0839E-01
LAML	*MAP3K15*	Somatic Mutation Frequency	wilcox	196	6.5500E+01	5.7715E-01
STAD	*HIST1H1E*	Somatic Mutation Frequency	wilcox	220	2.8700E+02	4.4460E-01
LAML	*FANCC*	Somatic Mutation Frequency	wilcox	196	2.4750E+02	5.0602E-01
BRCA	*BRIP1*	Somatic Mutation Frequency	wilcox	758	1.2860E+03	6.8601E-01
KIRC	*DDX11*	Somatic Mutation Frequency	wilcox	387	3.3300E+02	2.1169E-01
LUSC	*DIS3*	Somatic Mutation Frequency	wilcox	82	5.8500E+01	4.5969E-01
BRCA	*MAP3K15*	Somatic Mutation Frequency	wilcox	758	4.8600E+02	3.8344E-01
BRCA	*PALB2*	Somatic Mutation Frequency	wilcox	758	1.1550E+02	2.3022E-01
GBM	*PIK3C2G*	Somatic Mutation Frequency	wilcox	197	1.6500E+01	1.5428E-01
OV	*POLK*	Somatic Mutation Frequency	wilcox	426	1.5700E+02	6.5467E-01
STAD	*XRCC2*	Somatic Mutation Frequency	wilcox	220	2.3800E+02	4.2687E-01
BRCA	*XRCC2*	Somatic Mutation Frequency	wilcox	758	8.9100E+02	6.6358E-01
LAML	*MRE11A*	Somatic Mutation Frequency	wilcox	196	4.4500E+01	3.5276E-01
KIRC	*PIK3C2G*	Somatic Mutation Frequency	wilcox	387	3.7500E+01	1.6523E-01
BRCA	*FANCM*	Somatic Mutation Frequency	wilcox	758	3.5610E+03	3.6284E-01
OV	*EME1*	Somatic Mutation Frequency	wilcox	426	1.7050E+02	7.3574E-01
STAD	*BRCA2*	Somatic Mutation Frequency	wilcox	220	1.3700E+02	3.6899E-01
BRCA	*DDX11*	Somatic Mutation Frequency	wilcox	758	2.8150E+02	6.5917E-01
LUAD	*XRCC2*	Somatic Mutation Frequency	wilcox	228	1.3550E+02	7.4392E-01
PRAD	*POLK*	Somatic Mutation Frequency	wilcox	171	7.6500E+01	1.8614E-01
STAD	*CYP1B1*	Somatic Mutation Frequency	wilcox	220	1.6150E+02	4.1740E-01
STAD	*FANCC*	Somatic Mutation Frequency	wilcox	220	4.8000E+01	3.3679E-01
BRCA	*CNKSR1*	Somatic Mutation Frequency	wilcox	758	9.8100E+02	4.6781E-01
LAML	*RAD51C*	Somatic Mutation Frequency	wilcox	196	3.2700E+02	9.6388E-02
PRAD	*ATM*	Somatic Mutation Frequency	wilcox	171	2.3000E+02	2.9006E-01
BRCA	*EPPK1*	Somatic Mutation Frequency	wilcox	758	9.6450E+02	6.5806E-01
STAD	*BRCA1*	Somatic Mutation Frequency	wilcox	220	1.4600E+02	5.7080E-01
BRCA	*DIS3*	Somatic Mutation Frequency	wilcox	758	1.0425E+03	3.5500E-01
LGG	*ATM*	Somatic Mutation Frequency	wilcox	217	2.1500E+02	8.9035E-02
BRCA	*RAD50*	Somatic Mutation Frequency	wilcox	758	5.1300E+02	5.4023E-01
HNSC	*CNKSR1*	Somatic Mutation Frequency	wilcox	238	1.2000E+02	2.3355E-01
LUSC	*BRCA2*	Somatic Mutation Frequency	wilcox	82	4.6000E+01	8.3269E-01
BRCA	*ERCC2*	Somatic Mutation Frequency	wilcox	758	6.1500E+02	6.4955E-01
LUSC	*PARP3*	Somatic Mutation Frequency	wilcox	82	6.0000E+01	4.2213E-01
LUSC	*FANCA*	Somatic Mutation Frequency	wilcox	82	5.0000E+01	7.0376E-01
BRCA	*MSH6*	Somatic Mutation Frequency	wilcox	758	7.6600E+02	3.3350E-01
LUAD	*ATM*	Somatic Mutation Frequency	wilcox	228	1.8800E+02	6.8638E-01
GBM	*CYP1B1*	Somatic Mutation Frequency	wilcox	197	1.7100E+02	7.6952E-01
LUAD	*ERCC2*	Somatic Mutation Frequency	wilcox	228	1.6600E+02	5.2175E-01
PRAD	*EPPK1*	Somatic Mutation Frequency	wilcox	171	1.2400E+02	4.3532E-01
BRCA	*ATM*	Somatic Mutation Frequency	wilcox	758	1.7170E+03	6.3305E-01
STAD	*MSH6*	Somatic Mutation Frequency	wilcox	220	1.8100E+02	6.8376E-01
LUAD	*PIK3C2G*	Somatic Mutation Frequency	wilcox	228	3.1250E+02	3.5445E-01
GBM	*RAD50*	Somatic Mutation Frequency	wilcox	197	2.2700E+02	6.9451E-01

STAD	*RAD50*	Somatic Mutation Frequency	wilcox	220	1.2550E+02	8.0718E-01
GBM	*BRIP1*	Somatic Mutation Frequency	wilcox	197	9.1000E+01	9.0899E-01
LAML	*DDX11*	Somatic Mutation Frequency	wilcox	196	1.9650E+02	3.4203E-01
STAD	*HLA_G*	Somatic Mutation Frequency	wilcox	220	9.0500E+01	7.7082E-01
STAD	*MUTYH*	Somatic Mutation Frequency	wilcox	220	6.0000E+00	1.0483E-01
UCEC	*MAP3K15*	Somatic Mutation Frequency	wilcox	248	2.1500E+02	7.6274E-01
STAD	*PIK3C2G*	Somatic Mutation Frequency	wilcox	220	2.1450E+02	9.7329E-01
KIRC	*PARP3*	Somatic Mutation Frequency	wilcox	387	3.9000E+02	9.7724E-01
LAML	*ATM*	Somatic Mutation Frequency	wilcox	196	9.9500E+01	9.7882E-01
PRAD	*FANCG*	Somatic Mutation Frequency	wilcox	171	8.3500E+01	9.8383E-01
PRAD	*BRCA1*	Somatic Mutation Frequency	wilcox	171	8.3500E+01	9.8383E-01
UCEC	*FANCG*	Somatic Mutation Frequency	wilcox	248	2.4350E+02	9.8421E-01
UCEC	*XRCC2*	Somatic Mutation Frequency	wilcox	248	1.1550E+02	9.1656E-01
STAD	*BRIP1*	Somatic Mutation Frequency	wilcox	220	1.0400E+02	9.3725E-01
GBM	*MAP3K15*	Somatic Mutation Frequency	wilcox	197	6.8500E+01	6.1002E-01
GBM	*ERCC2*	Somatic Mutation Frequency	wilcox	197	1.2700E+02	6.1620E-01
GBM	*XPC*	Somatic Mutation Frequency	wilcox	197	4.8500E+01	3.8880E-01
GBM	*MSH6*	Somatic Mutation Frequency	wilcox	197	7.5000E+01	6.9231E-01
GBM	*PARP3*	Somatic Mutation Frequency	wilcox	197	8.5000E+00	1.1752E-01
HNSC	*BRCA2*	Somatic Mutation Frequency	wilcox	238	1.0450E+02	8.4422E-01
HNSC	*FANCC*	Somatic Mutation Frequency	wilcox	238	1.5050E+02	6.4660E-01
HNSC	*DIS3*	Somatic Mutation Frequency	wilcox	238	1.6150E+02	5.3618E-01
HNSC	*RAD50*	Somatic Mutation Frequency	wilcox	238	1.2450E+02	9.3619E-01
HNSC	*DDX11*	Somatic Mutation Frequency	wilcox	238	3.3000E+01	2.1601E-01
HNSC	*MSH6*	Somatic Mutation Frequency	wilcox	238	1.0100E+02	8.0457E-01
HNSC	*ATM*	Somatic Mutation Frequency	wilcox	238	8.4000E+01	6.2068E-01
HNSC	*RAD51D*	Somatic Mutation Frequency	wilcox	238	1.7100E+02	4.4912E-01
HNSC	*FANCG*	Somatic Mutation Frequency	wilcox	238	1.2050E+02	9.8258E-01
HNSC	*PALB2*	Somatic Mutation Frequency	wilcox	238	1.1800E+02	1.0000E+00
HNSC	*PMS2*	Somatic Mutation Frequency	wilcox	238	2.9800E+02	5.2588E-01
LUSC	*BRIP1*	Somatic Mutation Frequency	wilcox	82	1.9000E+01	3.7495E-01
PRAD	*FANCA*	Somatic Mutation Frequency	wilcox	171	1.0000E+02	7.6890E-01
PRAD	*RAD51C*	Somatic Mutation Frequency	wilcox	171	1.9000E+01	1.8443E-01
PRAD	*PALB2*	Somatic Mutation Frequency	wilcox	171	9.0500E+01	9.1930E-01
UCEC	*BRCA1*	Somatic Mutation Frequency	wilcox	248	3.4500E+02	8.5860E-01

Supplementary Table 2.19: Genes with rare germline truncation variants associated with younger age of initial diagnosis with cancer type as a covariate.

Gene	Clinic Feature	p-value	Covariants
BRCA1	Age at Diagnosis	5.20E-07	Cancer Type
BRCA2	Age at Diagnosis	2.04E-04	Cancer Type
ERCC2	Age at Diagnosis	3.18E-02	Cancer Type
DIS3	Age at Diagnosis	4.66E-02	Cancer Type
PMS2	Age at Diagnosis	4.87E-02	Cancer Type
ATM	Age at Diagnosis	5.39E-02	Cancer Type
FANCM	Age at Diagnosis	6.69E-02	Cancer Type
PIK3C2G	Age at Diagnosis	1.08E-01	Cancer Type
BRIP1	Age at Diagnosis	1.40E-01	Cancer Type
EPPK1	Age at Diagnosis	1.84E-01	Cancer Type
POLK	Age at Diagnosis	2.48E-01	Cancer Type
RAD50	Age at Diagnosis	3.16E-01	Cancer Type
PALB2	Age at Diagnosis	3.28E-01	Cancer Type
RAD51C	Age at Diagnosis	3.37E-01	Cancer Type
MAP3K15	Age at Diagnosis	3.45E-01	Cancer Type
PARP3	Age at Diagnosis	3.56E-01	Cancer Type
XPC	Age at Diagnosis	3.94E-01	Cancer Type
CNKSR1	Age at Diagnosis	4.15E-01	Cancer Type
BAP1	Age at Diagnosis	4.42E-01	Cancer Type
FANCA	Age at Diagnosis	4.71E-01	Cancer Type
EME1	Age at Diagnosis	5.12E-01	Cancer Type
RAD51D	Age at Diagnosis	5.27E-01	Cancer Type
FANCG	Age at Diagnosis	5.33E-01	Cancer Type
MUTYH	Age at Diagnosis	5.54E-01	Cancer Type
XRCC2	Age at Diagnosis	6.30E-01	Cancer Type
DDX11	Age at Diagnosis	7.03E-01	Cancer Type
HLA_G	Age at Diagnosis	7.12E-01	Cancer Type
MSH6	Age at Diagnosis	7.44E-01	Cancer Type
FANCC	Age at Diagnosis	7.69E-01	Cancer Type
CYP1B1	Age at Diagnosis	7.75E-01	Cancer Type
APITD1	Age at Diagnosis	8.69E-01	Cancer Type
MRE11A	Age at Diagnosis	8.91E-01	Cancer Type
HIST1H1E	Age at Diagnosis	9.15E-01	Cancer Type
EME2	Age at Diagnosis	9.53E-01	Cancer Type

Supplementary Table 2.20: Genes with rare germline truncation variants associated with younger age of initial diagnosis for each cancer type.

Cancer Type	Gene	Clinic Feature	Method	Number of total cases in cancer cohort	P-value
BRCA	*BRCA2*	Age at Diagnosis	wilcox	770	8.71E-03
BRCA	*BRCA1*	Age at Diagnosis	wilcox	770	5.04E-03
BRCA	*PIK3C2G*	Age at Diagnosis	wilcox	770	2.14E-01
BRCA	*RAD50*	Age at Diagnosis	wilcox	770	2.89E-01
BRCA	*ATM*	Age at Diagnosis	wilcox	770	4.85E-01
BRCA	*ERCC2*	Age at Diagnosis	wilcox	770	2.03E-01
BRCA	*MSH6*	Age at Diagnosis	wilcox	770	4.63E-01
BRCA	*CNKSR1*	Age at Diagnosis	wilcox	770	4.20E-01
BRCA	*RAD51C*	Age at Diagnosis	wilcox	770	6.89E-01
BRCA	*DIS3*	Age at Diagnosis	wilcox	770	3.74E-01
BRCA	*FANCC*	Age at Diagnosis	wilcox	770	5.89E-01
BRCA	*PALB2*	Age at Diagnosis	wilcox	770	6.46E-01
BRCA	*DDX11*	Age at Diagnosis	wilcox	770	9.99E-01
BRCA	*BRIP1*	Age at Diagnosis	wilcox	770	8.64E-01
BRCA	*EPPK1*	Age at Diagnosis	wilcox	770	8.46E-01
BRCA	*MAP3K15*	Age at Diagnosis	wilcox	770	9.31E-01
BRCA	*FANCM*	Age at Diagnosis	wilcox	770	1.75E-01
BRCA	*XRCC2*	Age at Diagnosis	wilcox	770	9.95E-01
GBM	*MSH6*	Age at Diagnosis	wilcox	267	1.66E-01
GBM	*PIK3C2G*	Age at Diagnosis	wilcox	267	1.63E-01
GBM	*APITD1*	Age at Diagnosis	wilcox	267	9.38E-01
GBM	*CYP1B1*	Age at Diagnosis	wilcox	267	8.15E-01
GBM	*BRIP1*	Age at Diagnosis	wilcox	267	5.21E-01
GBM	*MAP3K15*	Age at Diagnosis	wilcox	267	6.59E-01
GBM	*RAD50*	Age at Diagnosis	wilcox	267	9.19E-01
GBM	*XPC*	Age at Diagnosis	wilcox	267	1.01E-01
GBM	*ERCC2*	Age at Diagnosis	wilcox	267	7.90E-01
GBM	*DDX11*	Age at Diagnosis	wilcox	267	8.00E-01
GBM	*PARP3*	Age at Diagnosis	wilcox	267	5.86E-01
HNSC	*FANCA*	Age at Diagnosis	wilcox	290	2.95E-02
HNSC	*MSH6*	Age at Diagnosis	wilcox	290	9.32E-02
HNSC	*RAD50*	Age at Diagnosis	wilcox	290	1.25E-01
HNSC	*RAD51D*	Age at Diagnosis	wilcox	290	2.39E-01
HNSC	*BRCA2*	Age at Diagnosis	wilcox	290	2.74E-01
HNSC	*FANCM*	Age at Diagnosis	wilcox	290	2.20E-01
HNSC	*PMS2*	Age at Diagnosis	wilcox	290	7.69E-02
HNSC	*DDX11*	Age at Diagnosis	wilcox	290	2.12E-01
HNSC	*FANCC*	Age at Diagnosis	wilcox	290	8.59E-01
HNSC	*CNKSR1*	Age at Diagnosis	wilcox	290	8.36E-01
HNSC	*BRIP1*	Age at Diagnosis	wilcox	290	8.30E-01
HNSC	*FANCG*	Age at Diagnosis	wilcox	290	8.30E-01
HNSC	*DIS3*	Age at Diagnosis	wilcox	290	5.87E-01
HNSC	*ATM*	Age at Diagnosis	wilcox	290	6.63E-01
HNSC	*PALB2*	Age at Diagnosis	wilcox	290	7.86E-01
KIRC	*BRCA1*	Age at Diagnosis	wilcox	452	1.38E-01
KIRC	*EPPK1*	Age at Diagnosis	wilcox	452	1.36E-01
KIRC	*BRCA2*	Age at Diagnosis	wilcox	452	4.81E-01
KIRC	*BRIP1*	Age at Diagnosis	wilcox	452	4.48E-01
KIRC	*PIK3C2G*	Age at Diagnosis	wilcox	452	2.52E-01
KIRC	*BAP1*	Age at Diagnosis	wilcox	452	3.10E-01
KIRC	*FANCM*	Age at Diagnosis	wilcox	452	1.05E-01
KIRC	*ERCC2*	Age at Diagnosis	wilcox	452	4.44E-01
KIRC	*XPC*	Age at Diagnosis	wilcox	452	8.07E-01

KIRC	*PARP3*	Age at Diagnosis	wilcox	452	7.14E-01
KIRC	*CYP1B1*	Age at Diagnosis	wilcox	452	6.65E-01
KIRC	*DDX11*	Age at Diagnosis	wilcox	452	7.88E-01
KIRC	*RAD50*	Age at Diagnosis	wilcox	452	4.30E-01
KIRC	*EME1*	Age at Diagnosis	wilcox	452	6.62E-02
LAML	*DDX11*	Age at Diagnosis	wilcox	200	3.93E-01
LAML	*FANCC*	Age at Diagnosis	wilcox	200	4.95E-01
LAML	*DIS3*	Age at Diagnosis	wilcox	200	6.19E-01
LAML	*MAP3K15*	Age at Diagnosis	wilcox	200	7.49E-01
LAML	*PIK3C2G*	Age at Diagnosis	wilcox	200	2.35E-01
LAML	*ATM*	Age at Diagnosis	wilcox	200	3.63E-01
LAML	*MRE11A*	Age at Diagnosis	wilcox	200	9.79E-01
LAML	*RAD51C*	Age at Diagnosis	wilcox	200	1.87E-01
LGG	*MUTYH*	Age at Diagnosis	wilcox	223	6.96E-01
LGG	*PMS2*	Age at Diagnosis	wilcox	223	6.64E-01
LGG	*BRIP1*	Age at Diagnosis	wilcox	223	3.01E-01
LGG	*DDX11*	Age at Diagnosis	wilcox	223	9.44E-01
LGG	*EPPK1*	Age at Diagnosis	wilcox	223	6.43E-01
LGG	*HLA_G*	Age at Diagnosis	wilcox	223	2.15E-01
LGG	*ATM*	Age at Diagnosis	wilcox	223	5.39E-01
LGG	*PARP3*	Age at Diagnosis	wilcox	223	1.09E-01
LGG	*CYP1B1*	Age at Diagnosis	wilcox	223	5.09E-01
LUAD	*MRE11A*	Age at Diagnosis	wilcox	381	3.58E-01
LUAD	*BRIP1*	Age at Diagnosis	wilcox	381	3.21E-01
LUAD	*ERCC2*	Age at Diagnosis	wilcox	381	1.38E-01
LUAD	*PALB2*	Age at Diagnosis	wilcox	381	4.95E-01
LUAD	*ATM*	Age at Diagnosis	wilcox	381	2.36E-01
LUAD	*FANCG*	Age at Diagnosis	wilcox	381	3.20E-01
LUAD	*BRCA1*	Age at Diagnosis	wilcox	381	9.53E-01
LUAD	*DIS3*	Age at Diagnosis	wilcox	381	9.00E-01
LUAD	*PIK3C2G*	Age at Diagnosis	wilcox	381	8.59E-01
LUAD	*XRCC2*	Age at Diagnosis	wilcox	381	1.04E-01
LUSC	*BRIP1*	Age at Diagnosis	wilcox	185	1.65E-02
LUSC	*BRCA2*	Age at Diagnosis	wilcox	185	1.44E-01
LUSC	*DIS3*	Age at Diagnosis	wilcox	185	1.07E-01
LUSC	*FANCA*	Age at Diagnosis	wilcox	185	3.75E-01
LUSC	*MAP3K15*	Age at Diagnosis	wilcox	185	4.82E-01
LUSC	*PARP3*	Age at Diagnosis	wilcox	185	7.64E-01
LUSC	*FANCC*	Age at Diagnosis	wilcox	185	7.15E-01
OV	*BRCA1*	Age at Diagnosis	wilcox	428	1.69E-05
OV	*EME1*	Age at Diagnosis	wilcox	428	1.30E-01
OV	*MAP3K15*	Age at Diagnosis	wilcox	428	1.13E-01
OV	*BRCA2*	Age at Diagnosis	wilcox	428	9.74E-02
OV	*PALB2*	Age at Diagnosis	wilcox	428	7.38E-01
OV	*RAD51C*	Age at Diagnosis	wilcox	428	8.25E-01
OV	*RAD51D*	Age at Diagnosis	wilcox	428	9.72E-01
OV	*ATM*	Age at Diagnosis	wilcox	428	7.34E-01
OV	*BRIP1*	Age at Diagnosis	wilcox	428	9.20E-01
OV	*CNKSR1*	Age at Diagnosis	wilcox	428	5.28E-01
OV	*POLK*	Age at Diagnosis	wilcox	428	4.76E-01
OV	*PARP3*	Age at Diagnosis	wilcox	428	6.98E-01
OV	*PIK3C2G*	Age at Diagnosis	wilcox	428	6.47E-01
OV	*CYP1B1*	Age at Diagnosis	wilcox	428	6.07E-01
OV	*ERCC2*	Age at Diagnosis	wilcox	428	4.42E-01
OV	*DIS3*	Age at Diagnosis	wilcox	428	6.07E-01
PRAD	*FANCG*	Age at Diagnosis	wilcox	177	1.58E-01
PRAD	*RAD51C*	Age at Diagnosis	wilcox	177	3.88E-01
PRAD	*BRCA1*	Age at Diagnosis	wilcox	177	3.88E-01
PRAD	*FANCA*	Age at Diagnosis	wilcox	177	1.12E-01

PRAD	*EPPK1*	Age at Diagnosis	wilcox	177	3.08E-01
PRAD	*PALB2*	Age at Diagnosis	wilcox	177	2.56E-01
PRAD	*POLK*	Age at Diagnosis	wilcox	177	8.35E-02
PRAD	*ATM*	Age at Diagnosis	wilcox	177	9.61E-01
STAD	*ATM*	Age at Diagnosis	wilcox	315	1.56E-02
STAD	*HLA_G*	Age at Diagnosis	wilcox	315	1.49E-01
STAD	*BRCA1*	Age at Diagnosis	wilcox	315	1.49E-01
STAD	*HIST1H1E*	Age at Diagnosis	wilcox	315	3.74E-01
STAD	*CYP1B1*	Age at Diagnosis	wilcox	315	2.69E-01
STAD	*XRCC2*	Age at Diagnosis	wilcox	315	2.96E-01
STAD	*EPPK1*	Age at Diagnosis	wilcox	315	2.50E-01
STAD	*PIK3C2G*	Age at Diagnosis	wilcox	315	3.42E-01
STAD	*RAD50*	Age at Diagnosis	wilcox	315	2.28E-01
STAD	*FANCA*	Age at Diagnosis	wilcox	315	3.73E-01
STAD	*PALB2*	Age at Diagnosis	wilcox	315	1.22E-01
STAD	*EME2*	Age at Diagnosis	wilcox	315	2.16E-01
STAD	*MSH6*	Age at Diagnosis	wilcox	315	4.73E-01
STAD	*DDX11*	Age at Diagnosis	wilcox	315	5.13E-01
STAD	*BRIP1*	Age at Diagnosis	wilcox	315	1.03E-01
STAD	*MUTYH*	Age at Diagnosis	wilcox	315	6.21E-01
STAD	*BRCA2*	Age at Diagnosis	wilcox	315	7.52E-01
STAD	*DIS3*	Age at Diagnosis	wilcox	315	8.81E-02
STAD	*FANCC*	Age at Diagnosis	wilcox	315	8.99E-01
UCEC	*BRCA1*	Age at Diagnosis	wilcox	248	1.11E-01
UCEC	*FANCG*	Age at Diagnosis	wilcox	248	5.79E-02
UCEC	*DDX11*	Age at Diagnosis	wilcox	248	7.97E-01
UCEC	*CNKSR1*	Age at Diagnosis	wilcox	248	5.36E-01
UCEC	*MAP3K15*	Age at Diagnosis	wilcox	248	7.40E-01
UCEC	*BRIP1*	Age at Diagnosis	wilcox	248	6.20E-01
UCEC	*MUTYH*	Age at Diagnosis	wilcox	248	9.00E-01
UCEC	*MSH6*	Age at Diagnosis	wilcox	248	3.88E-01
UCEC	*XRCC2*	Age at Diagnosis	wilcox	248	1.00E+00
UCEC	*PMS2*	Age at Diagnosis	wilcox	248	3.60E-01
UCEC	*PIK3C2G*	Age at Diagnosis	wilcox	248	5.25E-01
UCEC	*RAD50*	Age at Diagnosis	wilcox	248	4.50E-01
UCEC	*MRE11A*	Age at Diagnosis	wilcox	248	2.31E-01
UCEC	*ERCC2*	Age at Diagnosis	wilcox	248	7.16E-01

Supplementary Table 2.21: Primers used for creating 72 *BRCA1* expression constructs with 68 rare missense variants introduced and 4 control truncation constructs.

***BRCA1* Variant Annotation**		**Forward Primer Sequence**	**Reverse Primer Sequence**
Missense (68)	S36Y	GAACCTGTCTaCACAAAGTGTG	CTTGATCAACTCCAGACAG
	C61G	GCCTTCACAGgGTCCTTTATG	CCTTTCTTCTGGTTGAGAAG
	C64G	GTGTCCTTTAgGTAAGAATGATATAAC	TGTGAAGGCCCTTTCTTC
	E143K	GAGTGAACCCaAAAATCCTTCC	TGTAGAAGTCTTTTGGCAC
	E149A	TCCTTGCAGGcAACCAGTCTC	AGGATTTTCGGGTTCACTC
	Y179C	ACGTCTGTCTgCATTGAATTG	CTTTTGAGGTTGTATCCG
	S186Y	GGATCTGATTaTTCTGAAGATACC	CAATTCAATGTAGACAGACG
	V191I	TGAAGATACCaTTAATAAGGCAAC	GAAGAATCAGATCCCAATTC
	D214G	GGAACCAGGGgTGAAATCAGTTTG	TTGAGGGGTGATTTGTAACAATTC
	T293S	TTATTACTCAgTAAAGACAGAATGAATG	ACTGCTGTTCTCATGCTG
	R296G	CACTAAAGACgGAATGAATGTAG	AGTAATAAACTGCTGTTCTC
	S316G	CTTAGCAAGGgGCCAACATAAC	CCAGGCTGTTTGCTTTTATTAC
	A322P	TAACAGATGGcCTGGAAGTAAGG	TGTTGGCTCCTTGCTAAG
	C328R	TAAGGAAACAcGTAATGATAGGCG	CTTCCAGCCCATCTGTTATG
	I379M	ATAGCAGCATgCAGAAAGTTAATG	TTAGTGTTATCCAAGGAACATC
	E445Q	ATGTAAAAGTcAAAGAGTTCACTCC	ATTAAAGCCTCATGAGGATC
	G462R	CAAAATATTTaGGAAAACCTATCGGAAG	TCTTCAATATTACTCTCTACTGATTTG
	F486L	TATAGGAGCAcTTGTTACTGAG	ATTAGATTTTCAGTTACATGGC
	L512V	TACATCAGGCgTTCATCCTGAG	GGTCTCCTTTTACGCTTTAATTTATTTG
	N550H	GAATATTACTcATAGTGGTCATGAGAATAAAAC	ATCACTTGACCATTCTGC
	I591T	AGCAGCAGTAcAAGCAATATG	TATAGGTTCAGCTTTCGTTTTG
	R612G	GAATAGGCTGgGGAGGAAGTC	TTTTTAGGTGCTTTTGAATTGTG
	L668F	AAACCTACAAtTCATGGAAGGTAAAGAACC	CTGCTGTGCCTGACTGGC
	D695N	ACATGACAGTaATACTTTCCCAG	CTTTTACTTGTCTGTTCATTTG
	S708Y	GCACCTGGTTaTTTTACTAAG	ATTTGTTAACTTCAGCTCTG
	E720K	TGAACTTAAAaAATTTGTCAATCCTAG	CTGGTATTTGAACACTTAGTAAAAG
	V772A	ATTTCATTGGcACCTGGTACTG	ACTGCTACTCTCTACAGATC
	A806T	GAGTCAGTGTaCAGCATTTGAAAAC	ACACATTTATTTGGTTCTGTTTTTG
	T826K	AGAAATGACAaAGAAGGCTTTAAG	ATTATCTTTGGAACAACCATG
	Y856H	TGATGCTCAGcATTTGCAGAATAC	AGTTCACTTTCTTCCATTTCTATG
	R866C	GGTTTCAAAGtGCCAGTCATTTG	TTGAATGTATTCTGCAAATACTG
	E962K	CAGAGGCAACaAAACTGGACTC	AACTGAGATGATAGACAAAAC
	I1019V	AAATGAGAACgTTCCAAGTACAG	CCCATTTCTCTTTCAGGTG
	I1044V	CTCAAGCAATgTTAATGAAGTAGG	CTGGCTCCTTTAAAAACATTTTC
	P1150S	TTCTGAGACAtCTGATGACCTG	CAAACCTGAGATGCATGAC
	D1152N	GACACCTGATaACCTGTTAGATG	TCAGAACAAACCTGAGATG
	E1219D	TATCTAGTGAcGATGAAGAGCTTCC	AGTTCTCTTCTGAGGACTC
	P1238L	AACAATATACtTTCTCAGTCTACTAG	TACTTTACCAAATAACAAGTGTTG
	V1247I	GCATAGCACCaTTGCTACCGA	CTAGTAGACTGAGAAGGTATATTG
	Q1281P	AAGGCATCTCcGGAACATCAC	TGCCAATATTACCTGGTTAC
	E1346K	AGATGATGAAaAAAGAGGAACG	GAAACCAATTCCTTGTCAC
	N1354T	TTGGAAGAAAcTAATCAAGAAGAGC	GCCCGTTCCTCTTTCTTC

	T1376R	GAGAGTGAAAgAAGCGTCTCTG	ACACCCAGATGCTGCTTC
	V1378I	TGAAACAAGCaTCTCTGAAGACTG	CTCTCACACCCAGATGCTG
	H1421Y	GTTAGAACAGtATGGGAGCCAG	ACAGCTTCTAGTTCAGCC
	G1422E	GAACAGCATGaGAGCCAGCCT	TAACACAGCTTCTAGTTCAGC
	K1476T	TCTGCTGACAcGTTTGAGGTG	AAGGCCTTCTGGATTCTG
	V1534M	GGTTGTTGATaTGGAGGAGCAAC	TTAATGAGCTCCTCTTGAGATG
	D1546Y	TGGGCCACACtATTTGACGGA	GACTCTTCCAGCTGTTGC
	T1561I	CTAGAGGGAAtCCCTTACCTG	ATCTTGCCTTGGCAAGTAAG
	L1564P	ACCCCTTACCcGGAATCTGGAATC	TCCCTCTAGATCTTGCCTTG
	P1579A	TGAATCTGATgCTTCTGAAGAC	GGGTCATCAGAGAAGAGG
	M1628T	TATAATGCAAcGGAAGAAAGTGTG	CCCAGCAGTATCAGTAGTATG
	P1637L	AGGGAGAAGCtAGAATTGACAG	GCTCACACTTTCTTCCATTG
	A1669S	GTACAAGTTTtCCAGAAAACACC	ACGAGCATAAATTCTTCTGG
	T1685I	GAAGAGACTAtTCATGTTGTTATGAAAAC	AGTAATTAGATTAGTTAAAGTGATG
	K1690Q	TGTTGTTATGcAAACAGATGCTG	TGAGTAGTCTCTTCAGTAATTAG
	R1699W	TGTGTGTGAAtGGACACTGAAATATTTTC	AACTCAGCATCTGTTTTCATAAC
	A1708V	CTAGGAATTGtGGGAGGAAAATG	AAAATATTTCAGTGTCCGTTC
	D1778G	ATGCCCACAGgTCAACTGGAA	GTTGGTGAAGGGCCCATA
	M1783L	ACTGGAATGGcTGGTACAGCTG	TGATCTGTGGGCATGTTG
	M1783T	CTGGAATGGAcGGTACAGCTG	TTGATCTGTGGGCATGTTG
	L1786P	ATGGTACAGCcGTGTGGTGCTTC	CCATTCCAGTTGATCTGTGG
	G1788V	CAGCTGTGTGtTGCTTCTGTG	TACCATCCATTCCAGTTG
	G1801D	TTCACCCTTGaCACAGGTGTC	TGATGAAAGCTCCTTCAC
	N1819S	ACAGAGGACAgTGGCTTCCATG	CCAGGCATCTGGCTGCAC
	R1835Q	GTGGTGACCCaAGAGTGGGTG	AGGTGCCTCACACATCTG
	P1859R	CCCCAGATCCgCCACAGCCAC	TATCAGGTAGGTGTCCAGCTC
Positive Control (4)	E23fs	TGTCCCATCTGTCTGGAG	CTAAGATTTTCTGCATAGCATTAATG
	E1250*	CGTTGCTACCTAGTGTCTGTC	GTGCTATGCCTAGTAGAC
	E1415fs	AAGCTGTGTTAGAACAGC	TAGTTCAGCCATTTCCTG
	Q1779fs	AACTGGAATGGATGGTACAG	ATCTGTGGGCATGTTGGTG

Supplementary Table 2.22: Homologous directed recombination assay results for 68 missense constructs, and 4 truncations as positive controls.

Variants		Rep1	Rep2	Rep3	Rep4	Rep5	Rep6	Mean	Stdev	BIC database annotation
Wide Type		1.0000	1.0000	1.0000				1.0000	0.0000	
Negative Control	siRNA+ pcDNA3	0.0820	0.0867	0.0804	0.2487	0.3140	0.3679	0.1966	0.1301	
Missense (68)	S36Y	1.4331	1.3359	1.4377				1.4022	0.0575	Not Reported
	C61G	0.2232	0.2054	0.2798				0.2361	0.0388	Pathogenic/likely pathogenic
	C64G	0.2113	0.2917	0.2321				0.2450	0.0417	Unknown clinical importance
	E143K	1.2487	1.2738	1.2377				1.2534	0.0185	Unknown clinical importance
	E149A	1.1607	1.2126	1.2393				1.2042	0.0399	Not Reported
	Y179C	1.6730	1.5735	1.4668				1.5711	0.1031	Unknown clinical importance
	S186Y	1.1529	1.0398	1.2707				1.1545	0.1155	Unknown clinical importance
	V191I	1.1277	1.0618	1.2817				1.1571	0.1128	Unknown clinical importance
	D214G	1.4574	1.3024	1.3708				1.3769	0.0777	Unknown clinical importance
	T293S	1.6777	1.3412	1.6209				1.5466	0.1801	Not Reported
	R296G	1.2361	1.0979	1.1450				1.1597	0.0703	Not Reported
	S316G	1.1626	1.2492	1.2903				1.2340	0.0652	Unknown clinical importance
	A322P	0.8928	0.8769	0.8848	0.7508	0.7938	0.7554	0.8258	0.0666	Unknown clinical importance
	C328R	1.1262	1.1073	1.0539				1.0958	0.0375	Not Reported
	I379M	0.9384	0.9897	0.9954				0.9745	0.0314	No clinical importance
	E445Q	0.9875	1.0228	0.9977				1.0027	0.0182	Unknown clinical importance
	G462R	0.9327	0.8141	0.8084				0.8518	0.0702	Unknown clinical importance
	F486L	1.6108	1.5692	1.6200				1.6000	0.0270	Unknown clinical importance
	L512V	0.8803	0.8130	0.9293	0.6754	0.6523	0.6385	0.7648	0.1260	Not Reported
	N550H	0.9092	0.8892	0.9262				0.9082	0.0185	Unknown clinical importance
	I591T	1.5450	1.2441	1.2109				1.3333	0.1841	Unknown clinical importance
	R612G	1.0725	1.0613	1.0936				1.0758	0.0164	Unknown clinical importance
	L668F	1.1397	0.5592	0.8360	1.0713	1.1452	1.0536	0.9675	0.2295	Unknown clinical importance
	D695N	1.2263	1.3010	1.1839				1.2371	0.0593	Unknown clinical importance
	S708Y	1.2553	1.2185	1.2397				1.2378	0.0185	Unknown clinical importance
	E720K	1.3255	1.3144	1.2631				1.3010	0.0333	Unknown clinical importance
	V772A	1.0836	1.0820	1.1453				1.1037	0.0361	Unknown clinical importance
	A806T	1.2363	1.2107	1.1750				1.2074	0.0308	Unknown clinical importance
	T826K	0.8554	1.0134	1.0736	0.8809	0.9060	0.8006	0.9216	0.1024	Unknown clinical importance
	Y856H	0.9966	1.0718	1.0068				1.0251	0.0408	Unknown clinical importance
	R866C	1.6611	1.7275	1.8768				1.7551	0.1104	No clinical importance
	E962K	0.9361	0.8506	0.9464				0.9111	0.0526	Not Reported
	I1019V	1.0523	1.4908	1.0385				1.1938	0.2572	Unknown clinical importance
	I1044V	1.1446	1.1138	1.0569				1.1051	0.0445	Unknown clinical importance
	P1150S	0.9396	0.9179	0.8107				0.8894	0.0690	Unknown clinical importance
	D1152N	1.6659	1.8555	1.6706				1.7306	0.1081	Unknown clinical importance
	E1219D	1.0319	1.0148	1.0171				1.0213	0.0093	Unknown clinical importance
	P1238L	1.4431	1.3626	1.3460				1.3839	0.0520	Unknown clinical importance
	V1247I	1.1118	1.0806	1.1391				1.1105	0.0293	Not Reported
	Q1281P	1.0445	1.1049	1.0046				1.0513	0.0505	Unknown clinical importance
	E1346K	0.9920	1.0046	0.9578				0.9848	0.0242	Unknown clinical importance
	N1354T	1.1270	1.1115	0.9612				1.0666	0.0915	Not Reported

	T1376R	1.4040	1.6880	1.5200				1.5373	0.1428	Unknown clinical importance
	V1378I	1.1014	1.2977	1.1483				1.1825	0.1025	Unknown clinical importance
	H1421Y	1.9300	1.7460	1.8440				1.8400	0.0921	Unknown clinical importance
	G1422E	1.6780	1.6580	1.9140				1.7500	0.1424	Not Reported
	K1476T	1.0547	1.0331	1.0342				1.0407	0.0122	Not Reported
	V1534M	1.0806	1.0916	1.0492				1.0738	0.0220	Unknown clinical importance
	D1546Y	0.7393	0.8231	0.8105	0.7512	0.8365	0.8236	0.7974	0.0414	Unknown clinical importance
	T1561I	1.2994	1.4240	1.2736				1.3323	0.0805	Unknown clinical importance
	L1564P	1.1255	1.0855	1.1655				1.1255	0.0400	No clinical importance
	P1579A	1.2503	1.1274	1.0982				1.1586	0.0807	Not Reported
	M1628T	1.0005	1.1717	1.0461				1.0728	0.0887	Unknown clinical importance
	P1637L	0.9277	0.9443	1.0930				0.9883	0.0911	Unknown clinical importance
	A1669S	1.2530	1.2976	1.5446				1.3651	0.1571	Unknown clinical importance
	T1685I	0.1935	0.2381	0.1726				0.2014	0.0335	Unknown clinical importance
	K1690Q	1.0759	0.9613	1.0963				1.0445	0.0728	Not Reported
	R1699W	0.1254	0.1269	0.0659	0.1492	0.1618	0.1555	0.1308	0.0351	Pathogenic/likely pathogenic
	A1708V	0.5938	0.6604	0.5362	0.5030	0.1951	0.3360	0.4707	0.1735	Unknown clinical importance
	D1778G	1.0037	1.0005	1.1042				1.0361	0.0590	Not Reported
	M1783L	1.2264	1.2371	1.3739				1.2791	0.0822	Unknown clinical importance
	M1783T	0.5188	0.5450	0.5268	0.8146	0.8222	0.8708	0.6616	0.1377	Unknown clinical importance
	L1786P	0.0514	0.0774	0.0465				0.0585	0.0166	Unknown clinical importance
	G1788V	0.0740	0.0913	0.0484	0.1712	0.1791	0.1602	0.1207	0.0562	Pathogenic/likely pathogenic
	G1801D	0.9132	0.9241	0.7461				0.8612	0.0998	Not Reported
	N1819S	1.1245	1.1460	1.1479				1.1395	0.0130	Unknown clinical importance
	R1835Q	0.5827	0.5796	0.6016	0.2930	0.3187	0.3358	0.4519	0.1499	Unknown clinical importance
	P1859R	1.3304	1.4196	1.2619				1.3373	0.0791	Unknown clinical importance
Positive control (4)	E23fs	0.0957	0.0888	0.0997				0.0947	0.0055	
	E1250*	0.0908	0.0826	0.0715				0.0816	0.0097	
	E1415fs	0.0566	0.0648	0.0640				0.0618	0.0045	
	Q1779fs	0.0852	0.0867	0.0368	0.1712	0.1712	0.1963	0.1246	0.0635	

Supplementary Table 2.23: Summary of BRCA1 validation status for A.I./non A.I. events and the enrichment factor.

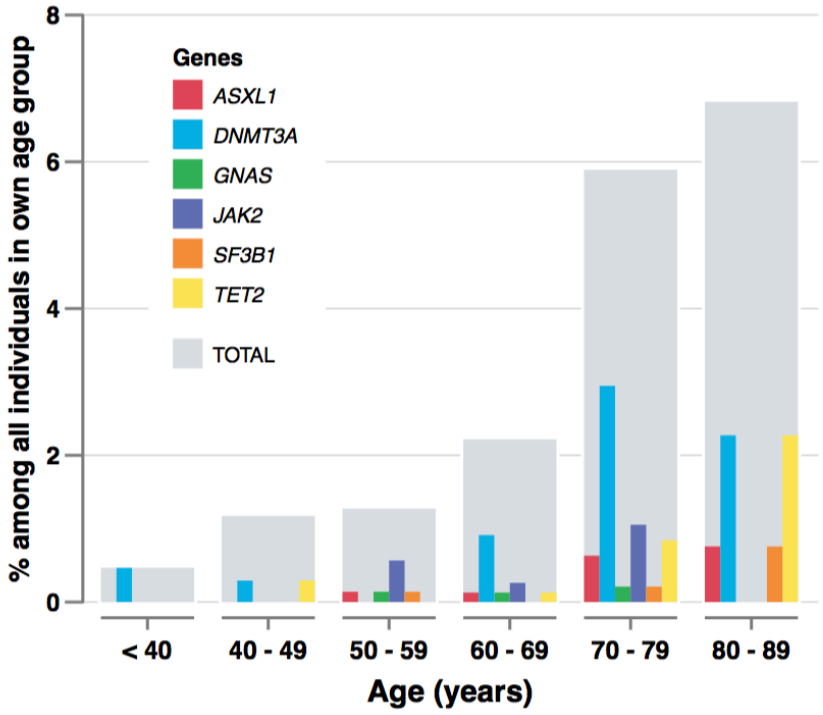
Chr	Position	Ref	Var	AA Change	Normal			Tumor			Cancer Type	MAF	AI FDR	AI Status
					Ref	Var	VAF	Ref	Var	VAF				
17	41245233	A	G	p.V772A	231	230	49.78	25	267	91.44	OV	0.033	1.3E-32	S
17	41201181	C	A	p.G1788V	35	46	56.79	4	213	98.16	OV	0.004	2.1E-17	S
17	41201187	A	G	p.L1786P	142	121	45.66	37	197	84.19	BRCA	0.004	4.1E-17	S
17	41258495	A	C	p.C64G	56	54	49.09	5	130	96.3	OV	0.009	3.9E-17	S
17	41223196	G	C	p.P1579A	273	239	46.5	78	264	77.19	OV	0.009	5.8E-17	S
17	41244982	A	G	p.Y856H	199	198	49.75	72	236	76.62	STAD	0.033	2.5E-11	S
17	41258504	A	C	p.C61G	64	42	39.62	12	73	85.88	BRCA	0.019	1.1E-09	S
17	41258504	A	C	p.C61G	103	60	36.81	71	152	68.16	BRCA	0.019	8.0E-08	S
17	41258504	A	C	p.C61G	67	38	36.19	8	45	84.91	BRCA	0.019	8.0E-08	S
17	41219645	G	A	p.T1685I	44	28	38.89	11	64	85.33	OV	0.004	9.5E-08	S
17	41256153	C	T	p.E143K	99	80	44.44	42	128	74.85	STAD	0.004	2.3E-07	S
17	41247892	T	C	p.D214G	63	44	41.12	7	45	86.54	OV	0.014	4.3E-07	S
17	41245233	A	G	p.V772A	130	106	44.92	72	169	70.12	LUAD	0.033	1.2E-06	S
17	41251803	T	C	p.Y179C	82	56	40.58	30	83	73.45	HNSC	0.038	6.1E-06	S
17	41215948	G	A	p.R1699W	42	30	41.67	17	64	79.01	OV	0.014	4.4E-05	S
17	41246092	A	G	p.F486L	64	40	38.46	27	68	71.58	HNSC	0.038	7.8E-05	S
17	41245714	T	C	p.R612G	65	74	53.24	31	93	75	OV	0.004	0.00290	S
17	41223021	G	A	p.P1637L	37	43	53.75	10	43	81.13	OV	0.004	0.00739	S
17	41223048	A	G	p.M1628T	32	40	55.56	18	71	79.78	BRCA	0.056	0.00758	S
17	41246164	C	T	p.G462R	88	82	48.24	199	346	63.25	LUSC	0.009	0.00774	S
17	41223048	A	G	p.M1628T	46	38	45.24	81	163	66.8	BRCA	0.056	0.00872	S
17	41251803	T	C	p.Y179C	130	84	39.07	125	136	52.11	BRCA	0.038	0.08128	NS
17	41245546	G	A	p.L668F	45	35	43.75	17	32	65.31	OV	0.004	0.14731	NS
17	41251893	T	G	p.E149A	69	49	41.53	75	90	54.55	STAD	0.004	0.27013	NS
17	41223048	A	G	p.M1628T	82	58	41.43	369	388	51.05	LUSC	0.056	0.29012	NS
17	41223048	A	G	p.M1628T	57	36	38.71	36	42	53.85	STAD	0.056	0.37041	NS
17	41244952	G	A	p.R866C	66	54	45	67	86	56.21	HNSC	0.004	0.40519	NS
17	41246092	A	G	p.F486L	172	125	42.09	188	183	49.33	KIRC	0.038	0.41902	NS
17	41258504	A	C	p.C61G	39	32	43.84	36	50	58.14	BRCA	0.019	0.51012	NS
17	41246566	A	G	p.C328R	128	98	43.17	92	98	51.58	LUAD	0.004	0.52505	NS
17	41246662	T	C	p.R296G	98	125	56.05	62	104	62.65	KIRC	0.004	0.59230	NS
17	41223048	A	G	p.M1628T	58	37	38.95	70	67	48.91	LGG	0.056	0.68186	NS
17	41201196	A	G	p.M1783T	129	93	41.7	54	54	50	UCEC	0.042	0.72804	NS
17	41246092	A	G	p.F486L	132	121	47.83	81	95	53.98	BRCA	0.038	0.74465	NS
17	41234517	G	A	p.H1421Y	134	123	47.86	125	132	51.16	KIRC	0.004	1	NS
17	41223048	A	G	p.M1628T	104	98	48.51	122	133	51.95	STAD	0.056	1	NS
17	41243512	C	T	p.E1346K	81	72	47.06	78	82	50.93	KIRC	0.009	1	NS
17	41244982	A	G	p.Y856H	81	77	48.73	61	68	52.71	STAD	0.033	1	NS
17	41246215	C	G	p.E445Q	108	98	47.34	285	287	50.09	BRCA	0.009	1	NS
17	41223240	A	G	p.L1564P	37	31	45.59	45	47	51.09	UCEC	0.038	1	NS
17	41246584	C	G	p.A322P	168	127	43.05	237	199	45.64	KIRC	0.004	1	NS
17	41245233	A	G	p.V772A	98	88	47.31	88	89	50	HNSC	0.033	1	NS
17	41245071	G	T	p.T826K	92	65	41.4	39	33	45.83	UCEC	0.023	1	NS
17	41243706	T	G	p.Q1281P	98	82	45.56	103	96	48	LGG	0.009	1	NS

17	41245233	A	G	p.V772A	87	83	48.82	106	110	50.93	HNSC	0.033	1	NS
17	41226387	C	A	p.D1546Y	38	36	48.65	49	52	51.49	KIRC	0.004	1	NS
17	41246092	A	G	p.F486L	104	79	43.17	188	149	44.21	KIRC	0.038	1	NS
17	41245425	G	T	p.S708Y	65	63	49.22	71	66	48.18	BRCA	0.009	1	NS
17	41245465	C	T	p.D695N	108	129	54.2	57	62	52.1	BRCA	0.004	1	NS
17	41223240	A	G	p.L1564P	41	29	41.43	36	26	41.94	UCEC	0.038	1	NS
17	41244982	A	G	p.Y856H	60	57	48.72	72	64	47.06	STAD	0.033	1	NS
17	41251803	T	C	p.Y179C	145	138	48.76	226	204	47.22	KIRC	0.038	1	NS
17	41243487	T	G	p.N1354T	255	257	50.2	209	192	47.88	BRCA	0.004	1	NS
17	41246215	C	G	p.E445Q	31	55	63.95	63	71	52.99	BRCA	0.009	1	NS
17	41245900	T	G	p.N550H	76	61	44.53	94	69	42.33	KIRC	0.038	1	NS
17	41251803	T	C	p.Y179C	147	165	52.88	211	192	47.64	KIRC	0.038	1	NS
17	41215920	G	A	p.A1708V	134	129	49.05	194	157	44.6	KIRC	0.019	1	NS
17	41245425	G	T	p.S708Y	104	115	52.51	262	227	46.42	BRCA	0.009	1	NS
17	41201142	C	T	p.G1801D	79	78	49.68	41	31	43.06	BRCA	0.004	1	NS
17	41245900	T	G	p.N550H	83	106	56.08	82	72	46.75	BRCA	0.038	1	NS
17	41201211	T	C	p.D1778G	34	37	52.11	41	31	43.06	GBM	0.004	1	NS
17	41244982	A	G	p.Y856H	43	62	58.49	79	67	45.89	BRCA	0.033	1	NS
17	41246014	G	C	p.L512V	81	65	44.52	139	88	38.77	BRCA	0.004	1	NS
17	41243891	C	G	p.E1219D	93	72	43.64	96	58	37.66	KIRC	0.009	1	NS
17	41245132	C	T	p.A806T	49	76	60.8	187	134	41.74	KIRC	0.004	1	NS
17	41223048	A	G	p.M1628T	59	64	52.03	80	52	39.39	LUAD	0.056	1	NS
17	41244664	C	T	p.E962K	75	92	55.09	98	65	39.88	LUAD	0.004	1	NS
17	41244982	A	G	p.Y856H	29	44	60.27	88	58	39.73	STAD	0.033	1	NS
17	41244982	A	G	p.Y856H	66	67	50.38	67	40	37.38	BRCA	0.033	1	NS
17	41243809	C	T	p.V1247I	80	67	45.27	89	50	35.97	LUAD	0.009	1	NS
17	41223048	A	G	p.M1628T	66	63	48.46	176	101	36.46	UCEC	0.056	1	NS
17	41243835	G	A	p.P1238L	198	184	48.04	267	145	35.19	BRCA	0.014	1	NS
17	41228562	T	G	p.K1476T	63	54	46.15	58	28	32.56	PRAD	0.004	1	NS
17	41249283	C	T	p.V191I	45	37	45.12	60	23	27.71	STAD	0.033	1	NS
17	41244094	C	T	p.D1152N	77	76	49.67	95	35	26.92	BRCA	0.004	1	NS
17	41246602	T	C	p.S316G	34	29	46.03	68	18	20.93	BRCA	0.014	1	NS
17	41244493	T	C	p.I1019V	164	151	47.94	144	16	10	BRCA	0.004	1	NS
17	41219694	C	A	p.A1669S	57	30	34.48	54	6	10	OV	0.014	1	NS
17	41243835	G	A	p.P1238L	53	48	47.52	838	88	9.44	LUSC	0.014	1	NS
17	41197711	G	C	p.P1859R	23	28	54.9	59	6	9.23	BRCA	0.009	1	NS
17	41245233	A	G	p.V772A	218	226	50.79	299	30	9.12	LUSC	0.033	1	NS
17	41246670	G	C	p.T293S	93	66	41.51	148	14	8.59	OV	0.004	1	NS
17	41249283	C	T	p.V191I	48	56	53.85	38	3	7.32	BRCA	0.033	1	NS
17	41234513	C	T	p.G1422E	446	435	49.04	746	33	4.23	OV	0.004	1	NS
17	41219694	C	A	p.A1669S	30	28	48.28	64	1	1.54	BRCA	0.014	1	NS

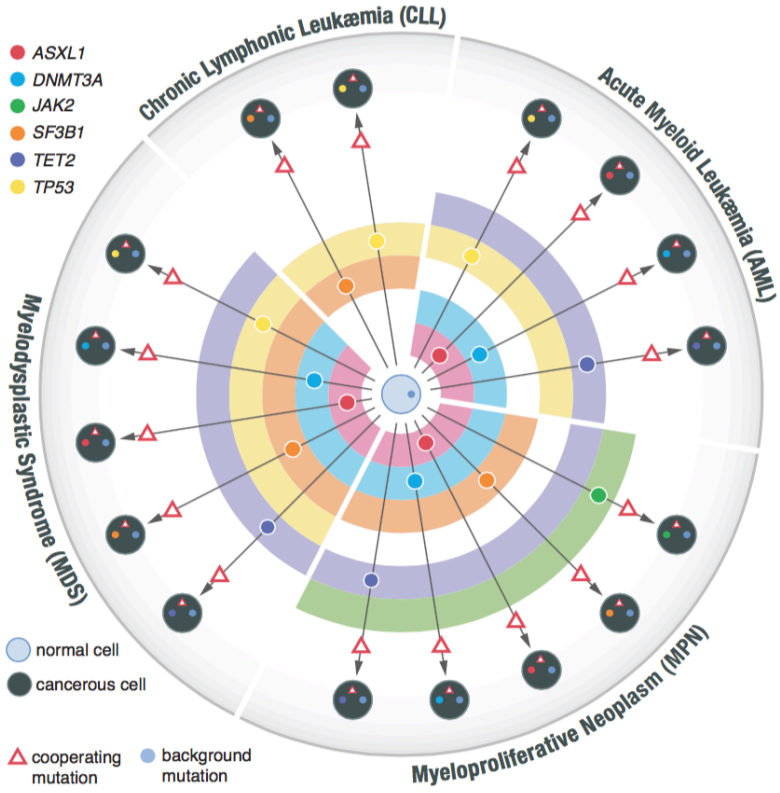
Supplementary Table 2.24: Rare germline missense variants overlapping with recurrent somatic mutations.

Gene	Annotation	Cancer Type	Chr	Position	Reference Allele	Variant Allele	Transcript (Ensembl)
ABCC2	p.E943K	LUAD	10	101590552	G	A	ENST00000370449
ATM	p.R2691C	BRCA	11	108205756	C	T	ENST00000278616
ATP5B	p.S375L	LUSC	12	57033894	G	A	ENST00000552919
BARD1	p.K140N	PRAD	2	215646178	C	A	ENST00000260947
BRCA2	p.E2020K	GBM	13	32914550	G	A	ENST00000380152
CARD11	p.R423Q	STAD	7	2976744	C	T	ENST00000396946
CDH1	p.R224H	PRAD	16	68842735	G	A	ENST00000261769
CDH12	p.D674N	KIRC	5	21752211	C	T	ENST00000382254
CYP2D6	p.H352R	BRCA	22	42523567	T	C	ENST00000360608
CYP2D6	p.H352R	BRCA	22	42523567	T	C	ENST00000360608
CYP2D6	p.H352R	BRCA	22	42523567	T	C	ENST00000360608
DDX11	p.E201K	BRCA	12	31238023	G	A	ENST00000407793
EPHB2	p.G874S	LUAD	1	23236992	G	A	ENST00000400191
EPHB2	p.R392H	BRCA	1	23191577	G	A	ENST00000400191
FANCA	p.R1084C	BRCA	16	89815165	G	A	ENST00000389301
FAT1	p.R806H	UCEC	4	187628565	C	T	ENST00000441802
FGF14	p.T229M	LGG	13	102375254	G	A	ENST00000376131
FLT3	p.R387Q	HNSC	13	28622457	C	T	ENST00000380982
LRP2	p.R3043C	PRAD	2	170044681	G	A	ENST00000263816
LRP2	p.R682C	BRCA	2	170115593	G	A	ENST00000443831
MED12	p.V1588M	HNSC	X	70354597	G	A	ENST00000333646
MORC4	p.R224C	BRCA	X	106228330	G	A	ENST00000355610
NOTCH4	p.T684M	LGG	6	32182003	G	A	ENST00000375023
PARP1	p.A625T	BRCA	1	226564877	C	T	ENST00000366794
PARP1	p.A625T	STAD	1	226564877	C	T	ENST00000366794
PDGFRB	p.S650L	STAD	5	149503887	G	A	ENST00000261799
RICTOR	p.S1101L	BRCA	5	38950648	G	A	ENST00000296782
SETD2	p.E639K	PRAD	3	47164211	C	T	ENST00000409792
TP53	p.R110H	GBM	17	7579358	C	T	ENST00000269305
TP53	p.R158C	GBM	17	7578458	G	A	ENST00000269305
TP53	p.R267Q	LUAD	17	7577138	C	T	ENST00000359597
TP53	p.R175C	GBM	17	7578407	G	A	ENST00000269305
TP53	p.R175C	BRCA	17	7578407	G	A	ENST00000359597
TP53	p.G245V	BRCA	17	7577547	C	A	ENST00000359597

Appendix 2 Supplementary Materials for Chapter 3



Supplementary Figure 3.1: Distribution of blood-specific mutations in *DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *SF3B1*, *GNAS*, and all 31 genes across different age groups. The 91 sites include 77 detected by our processing pipeline and 14 low VAF sites (2 to 10%) identified by read count-based analysis. The total includes all blood-specific mutations in 556 cancer associated genes identified in each age group



Supplementary Figure 3.2: Distinct and common connections among normal blood samples, MPN, MDS, CLL, and AML cases. A combination of precursor, initiating mutations in the normal blood samples may rarely collaborate with subsequent, progression mutations to develop MPN, MDS, CLL, and/or AML.

Supplementary Table 3.1: Sample IDs for the 2,728 TCGA cases included in this study.

The table is too large to display here. So it is hosted by *Nature Medicine* website:
http://www.nature.com/nm/journal/v20/n12/full/nm.3733.html#supplementary-information
http://www.nature.com/nm/journal/v20/n12/extref/nm.3733-S2.xlsx

Supplementary Table 3.2: Samples included in the study and their clinical characteristics.

Cancer Type	Case (N)	Age (Years)	Gender		Vital Status		
			Female (%)	Male (%)	Alive (N)	Dead (N)	NA (N)
BRCA	673	57.9 ± 12.9	99	1	632	41	0
GBM	230	59.5 ± 14.2	37	63	75	155	0
HNSC	243	61.6 ± 12.1	28.3	71.7	131	95	17
KIRC	57	59.5 ± 12.1	38.6	61.4	55	2	0
LGG	210	42.8 ± 13.4	42.4	57.6	163	47	0
LUAD	266	64.4 ± 9.7	51.7	48.3	222	37	7
LUSC	89	67.5 ± 9.4	19.5	80.5	70	17	2
OV	314	59.3 ± 11.6	100	0	151	162	1
PRAD	153	60.1 ± 7.2	0	100	133	1	19
STAD	262	65.6 ± 10.5	39	61	235	16	11
UCEC	231	62.6 ± 10.8	100	0	214	17	0
TOTAL	2728	59.5 ± 13.1	64.4	35.6	2081	590	57

Cancer Type	Ethnicity					
	White (N)	Asian (N)	African American (N)	American Indian/ Alaska Native (N)	Native Hawaiian / Pacific Islander (N)	NA (N)
BRCA	504	48	46	1	0	74
GBM	205	4	17	0	0	4
HNSC	189	6	21	1	0	26
KIRC	50	0	7	0	0	0
LGG	199	0	9	0	0	2
LUAD	191	3	17	0	0	55
LUSC	75	0	3	0	0	11
OV	271	13	13	2	0	15
PRAD	111	2	6	0	0	34
STAD	148	76	1	0	0	37
UCEC	178	13	23	3	7	7
TOTAL	2121	165	163	7	7	265

Supplementary Table 3.3: The distribution of germline variants across 2,728 TCGA samples. TCGA Ovarian counts were collected from the previous report.

Cancer Type	Total Germline Variants	Rare Germline	Missense	Nonsense	Frame-shift Indel	In-frame-shift Indel	Nonstop	Splice site	Silent
BRCA	14070932	426975	255541	4730	5979	5068	272	2863	152522
GBM	4444472	140874	85491	1655	1612	1285	98	978	49755
HNSC	4970133	149952	89998	1712	1715	1410	90	898	54129
KIRC	1166946	45309	27417	651	694	571	33	371	15572
LGG	4167463	113441	68999	1318	1387	1064	59	696	39918
LUAD	5366256	162782	98330	1822	1903	1447	88	954	58238
LUSC	1791732	45243	27868	558	591	405	22	261	15538
OV	NA	137664	131640	2749	1369	NA	NA	1518	388
PRAD	3143264	92097	54992	1046	1118	899	65	539	33438
STAD	5366040	157380	95915	1709	1905	1595	84	961	55211
UCEC	4829789	150768	89441	1713	2002	1764	115	937	54796
TOTAL	49317027	1622485	1025632	19663	20275	15508	926	10976	529505

Supplementary Table 3.4: Somatic mutations in 2,241 TCGA tumor samples included in the study. Somatic mutation data are unavailable for a subset of samples.

The table is too large to display here. So it is hosted by *Nature Medicine* website:
http://www.nature.com/nm/journal/v20/n12/full/nm.3733.html#supplementary-information
http://www.nature.com/nm/journal/v20/n12/extref/nm.3733-S5.xlsx

Supplementary Table 3.5: Somatic mutations in 3,355 TCGA tumor samples from 12 cancer types used for identifying recurrent mutations.

The table is too large to display here. So it is hosted by *Nature Medicine* website:
http://www.nature.com/nm/journal/v20/n12/full/nm.3733.html#supplementary-information
http://www.nature.com/nm/journal/v20/n12/extref/nm.3733-S6.xlsx

Supplementary Table 3.6: Recurrent somatic mutations from 12 TCGA cancer types used for hotspot analysis.

The table is too large to display here. So it is hosted by *Nature Medicine* website:
http://www.nature.com/nm/journal/v20/n12/full/nm.3733.html#supplementary-information
http://www.nature.com/nm/journal/v20/n12/extref/nm.3733-S7.xlsx

Supplementary Table 3.7: 556 cancer-associated genes used in this study.

ABCB1	*ABCC2*	*ABCC4*	*ABCG2*	*ABL1*	*ABL2*	*ACO1*	*ACVR1B*	*ACVR2A*	*ACVR2B*
ADNP	*APOL2*	*ASXL3*	*B4GALT3*	*BCR*	*CASP8*	*CD79B*	*CDKN2A*	*CIC*	*CTNNB1*
AJUBA	*AR*	*ATM*	*BACH1*	*BLM*	*CBFB*	*CDC27*	*CDKN2B*	*CNBD1*	*CUL4A*
AKT1	*AKT2*	*AKT3*	*ALK*	*ALKBH6*	*ALOX12B*	*ALPK2*	*AMER1*	*APC*	*APCDD1*
ARAF	*ARFRP1*	*ARHGAP35*	*ARID1A*	*ARID1B*	*ARID2*	*ARID5B*	*ARL11*	*ASXL1*	*ASXL2*
ATP5B	*ATR*	*ATRX*	*AURKA*	*AURKB*	*AXIN1*	*AXIN2*	*AXL*	*AZGP1*	*B2M*
BAK1	*BAP1*	*BARD1*	*BCL2*	*BCL2L11*	*BCL2L2*	*BCL6*	*BCLAF1*	*BCOR*	*BCORL1*
BRAF	*BRCA1*	*BRCA2*	*BRE*	*BRIP1*	*BRWD3*	*BTG1*	*BTK*	*CAP2*	*CARD11*
CBL	*CBLB*	*CBLC*	*CCND1*	*CCND2*	*CCND3*	*CCNE1*	*CD1D*	*CD70*	*CD79A*
CDC73	*CDH1*	*CDH12*	*CDH18*	*CDK12*	*CDK4*	*CDK6*	*CDK8*	*CDKN1A*	*CDKN1B*
CDKN2C	*CEBPA*	*CENPL*	*CEP76*	*CERS2*	*CHD4*	*CHD8*	*CHEK1*	*CHEK2*	*CHUK*
CNKSR1	*COMT*	*CRBN*	*CREBBP*	*CRIPAK*	*CRKL*	*CRLF2*	*CSF1R*	*CTCF*	*CTNNA1*
CUL4B	*CUX1*	*CYLD*	*CYP17A1*	*CYP1B1*	*CYP2C19*	*CYP2C8*	*CYP2D6*	*CYP3A4*	*CYP3A5*
DAXX	*DNMT1*	*EML4*	*ERBB4*	*EZH2*	*FANCM*	*FGF3*	*FOXA1*	*GNA13*	*H3F3C*
DCAF6	*DNMT3A*	*EMSY*	*ERCC2*	*EZR*	*FAT1*	*FGF4*	*FOXA2*	*GNAQ*	*HAUS3*
DDR1	*DDR2*	*DDX11*	*DDX3X*	*DDX5*	*DIAPH1*	*DIDO1*	*DIS3*	*DLC1*	*DNER*
DOT1L	*DPYD*	*ECSCR*	*EGFR*	*EGR3*	*EIF2S2*	*EIF3A*	*EIF4A2*	*ELF3*	*EME2*
EP300	*EPHA2*	*EPHA3*	*EPHA5*	*EPHB1*	*EPHB2*	*EPHB6*	*EPPK1*	*ERBB2*	*ERBB3*
ERG	*ESR1*	*ESR2*	*ETV1*	*ETV4*	*ETV5*	*ETV6*	*EWSR1*	*EXT1*	*EZH1*
FAM129B	*FAM46C*	*FANCA*	*FANCC*	*FANCD2*	*FANCE*	*FANCF*	*FANCG*	*FANCI*	*FANCL*
FAT3	*FBXW7*	*FCGR1A*	*FCGR2A*	*FCGR3A*	*FGF10*	*FGF12*	*FGF14*	*FGF19*	*FGF23*
FGF6	*FGF7*	*FGFBP1*	*FGFR1*	*FGFR2*	*FGFR3*	*FGFR4*	*FLT1*	*FLT3*	*FLT4*
FOXL2	*FOXQ1*	*FUBP1*	*FZD1*	*GAB2*	*GATA1*	*GATA2*	*GATA3*	*GID4C*	*GNA11*
GNAS	*GNB1*	*GPR124*	*GPS2*	*GRIN2A*	*GRM3*	*GSK3B*	*GSTP1*	*GUCY1A2*	*H3F3A*
HDAC4	*HGF*	*HIF1A*	*HIST1H1C*	*HIST1H1E*	*HIST1H2BD*	*HIST1H3B*	*HIST1H4E*	*HLA-A*	*HLA-B*
HLA-G	*ING1*	*JUN*	*KMT2C*	*MAP2K4*	*MDM4*	*MNDA*	*MYCL1*	*NFE2L2*	*NTN4*
HNF1A	*INHA*	*KDM5A*	*KMT2D*	*MAP3K1*	*MECOM*	*MORC4*	*MYCN*	*NFKBIA*	*NTRK1*
HRAS	*HSP90AB1*	*IDH1*	*IDH2*	*IGF1*	*IGF1R*	*IGF2*	*IKBKE*	*IKZF1*	*IL7R*
INHBA	*INPPL1*	*IPO7*	*IRF4*	*IRS2*	*ITGAV*	*ITPA*	*JAK1*	*JAK2*	*JAK3*
KDM5C	*KDM6A*	*KDR*	*KEAP1*	*KIF5B*	*KIT*	*KLF4*	*KLHL6*	*KMT2A*	*KMT2B*
KRAS	*LIFR*	*LMO1*	*LRP1B*	*LRP2*	*LRRK2*	*MALAT1*	*MAN1B1*	*MAP2K1*	*MAP2K2*
MAP3K13	*MAP3K15*	*MAP4K1*	*MAP4K3*	*MAPK1*	*MAPK8IP1*	*MBD1*	*MC1R*	*MCL1*	*MDM2*
MED12	*MED23*	*MEF2A*	*MEF2B*	*MEN1*	*MET*	*MGA*	*MIR142*	*MITF*	*MLH1*
MPL	*MRE11A*	*MSH2*	*MSH6*	*MTHFR*	*MTOR*	*MUTYH*	*MXRA5*	*MYB*	*MYC*
MYD88	*MYLK*	*MYST3*	*NAV3*	*NBN*	*NBPF1*	*NCOR1*	*NEIL1*	*NF1*	*NF2*
NKX2-1	*NOTCH1*	*NOTCH2*	*NOTCH3*	*NOTCH4*	*NPM1*	*NQO1*	*NRAS*	*NRP2*	*NSD1*
NTRK2	*NTRK3*	*NUP93*	*ODAM*	*OTUD7A*	*PAK3*	*PAK7*	*PALB2*	*PAPD5*	*PARP1*
PARP2	*PDSS2*	*POLD1*	*PRLR*	*RAD50*	*RBM10*	*RPL22*	*SDHC*	*SIRT4*	*SMO*
PARP3	*PHF6*	*POLE*	*PRPF40B*	*RAD51*	*RBMX*	*RPL5*	*SDHD*	*SLC19A1*	*SNX25*
PARP4	*PAX5*	*PBRM1*	*PCBP1*	*PCDH10*	*PDAP1*	*PDCD2L*	*PDGFRA*	*PDGFRB*	*PDK1*
PIK3C2G	*PIK3C3*	*PIK3CA*	*PIK3CG*	*PIK3R1*	*PIK3R2*	*PLCG2*	*PML*	*PMS2*	*PNRC1*
POLQ	*PORCN*	*POU2AF1*	*POU2F2*	*PPM1D*	*PPP2R1A*	*PPP6C*	*PRDM1*	*PRKAR1A*	*PRKDC*
PRSS8	*PTCH1*	*PTEN*	*PTPN11*	*PTPRC*	*PTPRD*	*QKI*	*RAB40A*	*RAC1*	*RAD21*
RAD51B	*RAD51C*	*RAD51D*	*RAD52*	*RAD54L*	*RAF1*	*RALY*	*RARA*	*RASA1*	*RB1*
REL	*RET*	*RHBDF2*	*RHEB*	*RHOA*	*RICTOR*	*RIT1*	*RNF43*	*ROS1*	*RPA1*
RPS14	*RPS15*	*RPS2*	*RPTOR*	*RUNX1*	*RUNX1T1*	*RUNX3*	*RXRA*	*SDHAF2*	*SDHB*
SRC	*SZRD1C*	*TLK2*	*TRAF7*	*UMPS*	*XRCC3*	*SRSF2*	*TAF1*	*TLR4*	*TRRAP*
SERPINB13	*SETBP1*	*SETD2*	*SETDB1*	*SF1*	*SF3B1*	*SGK1*	*SH2B3*	*SIN3A*	*SIRPA*
SLC22A2	*SLCO1B3*	*SMAD2*	*SMAD3*	*SMAD4*	*SMARCA4*	*SMARCB1*	*SMARCD1*	*SMC1A*	*SMC3*
SOCS1	*SOD2*	*SOS1*	*SOX10*	*SOX17*	*SOX2*	*SOX9*	*SPEN*	*SPOP*	*SPRY4*
STAG2	*STAT4*	*STK11*	*STK19*	*STK38*	*STX2*	*SUFU*	*SULT1A1*	*SUZ12*	*SYK*
TBC1D12	*TBL1XR1*	*TBX3*	*TCEB1*	*TCF7L2*	*TET2*	*TFG*	*TGFBR2*	*TIMM17A*	*TIPARP*
TMPRSS2	*TNF*	*TNFAIP3*	*TNFRSF14*	*TOP1*	*TP53*	*TP53BP1*	*TPMT*	*TPX2*	*TRAF3*
TSC1	*TSC2*	*TSHR*	*TSHZ2*	*TSHZ3*	*TYMS*	*TYR*	*U2AF1*	*U2AF65*	*UGT1A1*
USP9X	*VANGL2*	*VEZF1*	*VHL*	*WAC*	*WASF3*	*WISP3*	*WNK1*	*WT1*	*XPO1*
XRCC2	*ZFHX3*	*ZNF217*	*ZNF703*	*ZRANB3*	*ZRSR2*				

Supplementary Table 3.8: 77 blood-specific events detected in 2,728 cases using our standard discovery pipeline.

Gene	Case	Cancer Type	Age (years)	Group	Mutation	Blood Sample				Tumor Sample			
						Ref	Var	VAF	CNV	Ref	Var	VAF	CNV
ASXL1	TCGA-05-4403	LUAD	76	AML	p.R548fs	128	69	35.0	2.0	229	10	4.2	2.0
ASXL1	TCGA-37-4130	LUSC	56	AML	p.Y591fs	142	60	29.7	2.0	247	15	5.7	3.1
ASXL1	TCGA-86-8281	LUAD	75	AML	p.Q575*	108	27	20.0	2.0	179	6	3.2	2.0
ASXL1	TCGA-97-8179	LUAD	72	AML	p.Q733*	36	6	14.3	2.0	52	0	0.0	2.4
ASXL1	TCGA-A5-A0R8	UCEC	81	AML	p.Q733fs	16	6	27.3	2.0	48	4	7.7	2.0
ASXL1	TCGA-BR-8373	STAD	65	AML	p.Y591*	124	27	17.9	2.0	125	2	1.6	2.4
ASXL2	TCGA-DU-6400	LGG	66	AML	e12-2	138	29	17.4	2.0	167	1	0.6	2.2
ATM	TCGA-52-7812	LUSC	68	notAML	p.R337C	51	12	19.1	2.0	46	1	2.1	2.1
AXL	TCGA-97-8174	LUAD	67	AML	p.G517D	161	20	11.1	2.0	75	0	0.0	2.3
BCORL1	TCGA-95-8494	LUAD	null	AML	p.G883E	20	4	16.7	null	101	0	0.0	null
BCORL1	TCGA-EJ-5526	PRAD	56	AML	p.S264*	114	33	22.5	null	128	0	0.0	null
CBL	TCGA-CR-7376	HNSC	83	AML	p.R540*	111	26	19.0	2.0	111	7	5.9	2.0
CDKN2A	TCGA-97-8174	LUAD	67	AML	p.E120*	58	16	21.6	2.0	42	0	0.0	1.8
CREBBP	TCGA-97-8179	LUAD	72	AML	p.A259S	171	24	12.3	2.0	128	0	0.0	2.4
DIDO1	TCGA-97-8174	LUAD	67	notAML	p.H1557R	100	21	17.4	2.0	75	0	0.0	1.9
DNMT3A	TCGA-06-0142	GBM	81	AML	p.R882C	64	12	15.8	2.0	89	3	3.3	2.0
DNMT3A	TCGA-06-2558	GBM	75	AML	p.Y584fs	31	19	38.0	2.0	24	1	4.0	2.1
DNMT3A	TCGA-06-2563	GBM	72	AML	p.E469*	185	48	20.6	2.0	229	2	0.9	2.0
DNMT3A	TCGA-12-3649	GBM	76	AML	e21-2	112	15	11.8	2.0	69	0	0.0	2.1
DNMT3A	TCGA-21-5783	LUSC	76	AML	p.R882H	32	15	31.9	2.0	76	4	5.0	2.7
DNMT3A	TCGA-46-6025	LUSC	71	AML	p.N516fs	98	49	33.3	2.0	132	8	5.7	2.6
DNMT3A	TCGA-55-7815	LUAD	76	AML	e13+1	72	9	11.1	2.0	53	1	1.9	2.0
DNMT3A	TCGA-81-5910	GBM	64	AML	p.R882H	29	16	35.6	2.0	51	1	1.9	2.0
DNMT3A	TCGA-AX-A060	UCEC	77	AML	e22-1	43	22	33.9	null	34	0	0.0	2.0
DNMT3A	TCGA-B2-5636	KIRC	79	AML	e13+1	58	11	15.9	2.0	52	0	0.0	2.0
DNMT3A	TCGA-BH-A0DL	BRCA	64	AML	p.F851fs	252	135	34.9	2.0	465	46	9.0	2.1
DNMT3A	TCGA-BR-7197	STAD	69	AML	p.R882C	101	14	12.2	2.0	62	0	0.0	1.9
DNMT3A	TCGA-BR-A4QL	STAD	75	AML	p.S770*	131	25	16.0	2.0	130	0	0.0	2.0
DNMT3A	TCGA-CR-7370	HNSC	72	AML	p.K577fs	22	7	24.1	2.0	23	0	0.0	1.9
DNMT3A	TCGA-D1-A16G	UCEC	74	AML	p.W314*	106	30	22.1	2.0	165	10	5.7	2.5
DNMT3A	TCGA-D7-A4Z0	STAD	60	AML	p.R882C	67	15	18.3	null	75	5	6.3	2.0
DNMT3A	TCGA-E9-A1RG	BRCA	62	AML	p.R882H	22	6	21.4	2.0	23	0	0.0	2.2
DNMT3A	TCGA-G9-6356	PRAD	60	AML	e12-1	61	34	35.8	2.0	78	3	3.7	2.0
GNAS	TCGA-19-2625	GBM	76	AML	p.R202H	77	13	14.4	2.0	214	1	0.5	2.2
GNAS	TCGA-67-4679	LUAD	69	AML	p.R202H	77	21	21.4	2.0	84	0	0.0	2.0
GNAS	TCGA-D6-A4Z9	HNSC	59	AML	p.R202H	69	9	11.5	2.0	68	1	1.5	2.1
GUCY1A2	TCGA-97-7938	LUAD	76	notAML	p.K643N	15	7	31.8	2.0	27	0	0.0	2.0
HDAC4	TCGA-26-5135	GBM	72	notAML	p.P412R	40	9	18.0	2.0	68	1	1.5	2.0
IDH2	TCGA-D7-8574	STAD	72	AML	p.R140Q	29	18	38.3	2.0	41	0	0.0	2.0
JAK2	TCGA-05-4403	LUAD	76	AML	p.V617F	108	77	41.6	2.0	88	6	6.4	2.0
JAK2	TCGA-06-0240	GBM	57	AML	p.V617F	186	51	21.5	2.0	114	1	0.9	2.0
JAK2	TCGA-26-5135	GBM	72	AML	p.V617F	58	160	73.4	2.0	275	2	0.7	2.0
JAK2	TCGA-99-8025	LUAD	72	AML	p.V617F	76	29	27.6	2.0	139	4	2.8	1.8
JAK2	TCGA-AP-A0LQ	UCEC	59	AML	p.V617F	75	42	35.9	2.0	160	9	5.3	2.0
JAK2	TCGA-B4-5834	KIRC	59	AML	p.V617F	75	30	28.6	2.0	202	3	1.5	2.0
JAK2	TCGA-BK-A139	UCEC	74	AML	p.V617F	124	94	42.9	2.0	156	15	8.8	1.6
JAK2	TCGA-HW-7495	LGG	45	AML	p.V617F	53	10	15.9	2.0	113	0	0.0	2.0
MBD1	TCGA-97-8174	LUAD	67	notAML	p.S63N	102	12	10.5	2.0	66	0	0.0	1.9
MECOM	TCGA-13-0919	OV	52	notAML	p.T982N	56	9	13.9	2.0	82	1	1.2	2.3
MYLK	TCGA-AX-A05T	UCEC	82	AML	p.Q1392*	59	14	19.2	null	70	0	0.0	2.0
NOTCH3	TCGA-BR-	STAD	75	notAML	p.A284T	24	11	31.4	2.0	48	0	0.0	2.0

	A4QL												
PPM1D	TCGA-A7-A0CH	BRCA	79	notAML	p.Q520*	124	68	35.4	null	230	5	2.1	2.0
PPM1D	TCGA-D1-A16D	UCEC	49	notAML	p.S468*	115	31	21.2	2.0	146	5	3.3	2.0
PRKDC	TCGA-26-5135	GBM	72	notAML	p.Q1389R	61	35	36.5	2.0	93	1	1.1	2.0
RICTOR	TCGA-BH-A0DS	BRCA	71	notAML	e9-1	38	9	18.8	2.0	74	0	0.0	2.0
SETBP1	TCGA-26-5135	GBM	72	AML	p.E603A	87	39	31.0	2.0	144	1	0.7	2.0
SF1	TCGA-95-7043	LUAD	63	AML	p.G407V	30	4	11.8	2.0	42	0	0.0	2.5
SF3B1	TCGA-41-2571	GBM	89	AML	p.K700E	87	14	13.9	2.0	175	0	0.0	2.0
SF3B1	TCGA-B0-5698	KIRC	77	AML	p.K700E	45	34	43.0	2.0	90	10	10.0	2.0
SH2B3	TCGA-BH-A0B1	BRCA	66	AML	e3+1	104	30	22.4	2.0	51	3	5.6	null
SNX25	TCGA-13-0919	OV	52	notAML	p.A745T	42	6	12.2	2.0	44	0	0.0	1.9
SOS1	TCGA-95-7043	LUAD	63	notAML	p.G414W	28	4	12.5	2.0	47	0	0.0	2.2
TET2	TCGA-06-2558	GBM	75	AML	p.Q764fs	69	34	33.0	2.0	234	5	2.1	2.1
TET2	TCGA-06-2559	GBM	83	AML	p.F381fs	91	91	50.0	2.0	321	9	2.7	2.0
TET2	TCGA-06-2559	GBM	83	AML	p.Q888*	82	21	20.4	2.0	127	1	0.8	2.0
TET2	TCGA-26-5135	GBM	72	AML	p.T229fs	102	24	19.1	2.0	126	1	0.8	2.0
TET2	TCGA-36-2543	OV	85	AML	p.K889*	135	24	15.1	2.0	132	8	5.7	1.6
TET2	TCGA-69-7764	LUAD	75	AML	p.Q831fs	39	14	26.4	2.0	33	0	0.0	1.5
TET2	TCGA-81-5910	GBM	64	AML	p.H863fs	53	7	11.7	2.0	75	0	0.0	2.1
TET2	TCGA-97-7938	LUAD	76	AML	p.R550*	67	13	16.3	2.0	106	4	3.6	2.0
TET2	TCGA-AK-3433	KIRC	48	AML	p.Q531*	37	5	11.9	2.0	35	0	0.0	2.7
TET2	TCGA-BG-A0W1	UCEC	89	AML	p.Q644*	251	51	16.8	2.0	107	1	0.9	1.9
TP53	TCGA-13-0919	OV	52	AML	p.C275Y	42	7	14.3	2.0	42	0	0.0	1.5
TP53	TCGA-50-5049	LUAD	70	AML	p.R273L	17	9	34.6	2.0	28	2	6.7	1.7
TP53	TCGA-95-8494	LUAD	null	AML	p.Q136*	79	18	18.0	2.0	49	0	0.0	1.5
TP53	TCGA-D7-A6F2	STAD	62	AML	p.Q144*	79	15	16.0	null	37	2	5.1	null
ZRSR2	TCGA-D7-8574	STAD	72	AML	e9+1	29	18	38.3	null	41	0	0.0	null

Supplementary Table 3.9: Low-level blood-specific events detected in *DNMT3A*, *JAK2*, *SF3B1*, *GNAS*, and *IDH2* in TCGA samples.

Gene	Mutation	Case	Fisher P-value	FDR	Normal			Tumor		
					Ref Reads	Var Reads	VAF (%)	Ref Reads	Var Reads	VAF (%)
DNMT3A	R882C	TCGA-A2-A1G1	0.00595	0.02234	44	6	12	30	3	9.09
DNMT3A	R882C	TCGA-BR-8081	0.00722	0.02167	95	10	9.52	46	0	0
DNMT3A	R882C	TCGA-85-7698	0.02635	0.06586	90	7	7.22	83	0	0
DNMT3A	R882C	TCGA-B2-4098	0.05886	0.12616	81	5	5.75	51	1	1.92
DNMT3A	R882C	TCGA-24-1469	0.07955	0.14918	70	4	5.41	102	0	0
DNMT3A	R882H	TCGA-D1-A17H	0.02256	0.10716	57	4	6.56	64	0	0
DNMT3A	R882H	TCGA-FG-7634	0.02542	0.09660	60	4	6.25	63	0	0
DNMT3A	R882H	TCGA-CR-7376	0.06276	0.19875	60	3	4.76	62	2	3.12
IDH2	R140Q	TCGA-09-2053	0.05230	0.18308	102	4	3.77	102	0	0
JAK2	V617F	TCGA-24-1603	0.00376	0.00919	87	8	8.42	126	0	0
JAK2	V617F	TCGA-EJ-5521	0.00606	0.01335	242	16	6.2	197	0	0
JAK2	V617F	TCGA-06-2563	0.01005	0.02011	269	16	5.59	233	0	0
JAK2	V617F	TCGA-AP-A0LD	0.05020	0.09203	96	5	4.95	89	0	0
SF3B1	K700E	TCGA-D1-A0ZS	0.02663	0.08877	127	10	7.3	128	6	4.48

Supplementary Table 3.10: Deep-sequencing based validation of low-level blood-specific events detected in *DNMT3A*, *JAK2*, and *SF3B1* in TCGA samples.

Type	Gene	Mut-ation	Case	Normal			Tumor			Validation Normal			Validation Tumor		
				Ref	Var	VAF (%)	Ref	Var	VAF (%)	Ref	Var	VAF (%)	Ref	Var	VAF (%)
Group1	*DNMT3A*	R882C	TCGA-A2-A1G1	44	6	12	30	3	9.09	525630	65588	11.09	663050	15657	2.31
Group1	*DNMT3A*	R882C	TCGA-24-1469	70	4	5.41	102	0	0	693053	19627	2.75	677489	1123	0.17
Group1	*DNMT3A*	R882H	TCGA-D1-A17H	57	4	6.56	64	0	0	403096	16067	3.83	538616	2068	0.38
Group1	*JAK2*	V617F	TCGA-AP-A0LD	96	5	4.95	89	0	0	622088	23727	3.67	701514	424	0.06
Group1	*SF3B1*	K700E	TCGA-D1-A0ZS	127	10	7.3	128	6	4.48	1413	120	7.83	382	0	0
Group2	*DNMT3A*	R882C	TCGA-A2-A0CW	117	0	0	78	0	0	667273	185	0.03	627418	141	0.02
Group2	*DNMT3A*	R882H	TCGA-A7-A13D	103	0	0	163	0	0	663423	1301	0.2	667017	1101	0.16
Group2	*JAK2*	V617F	TCGA-A2-A04N	122	0	0	75	0	0	616831	755	0.12	680609	1047	0.15
Group2	*SF3B1*	K700E	TCGA-A8-A06Z	133	0	0	187	0	0	32263	2	0.01	102124	6	0.01
Group3	*DNMT3A*	F851fs	TCGA-BH-A0DL	252	135	34.9	465	46	9	7110	4021	36.1	27942	2828	9.19
Group2	*DNMT3A*	W314*	TCGA-D1-A16G	106	30	22.1	165	10	5.71	582397	160832	21.6	517920	43311	7.71
Group2	*DNMT3A*	R882H	TCGA-E9-A1RG	22	6	21.4	23	0	0	375747	43963	10.5	538566	3035	0.56
Group2	*JAK2*	V617F	TCGA-BK-A139	124	94	42.9	156	15	8.77	364000	332335	47.7	591804	82373	12.22
Group2	*SF3B1*	K700E	TCGA-41-2571	87	14	13.9	175	0	0	34723	8830	20.3	62472	10	0.02

*Group1: candidate variants
*Group2: positive variants
*Group3: Negative variants

Supplementary Table 3.11: Truncation and hotspot variants in four prominent genes (*DNMT3A*, *TET2*, *JAK2*, and *ASXL1*) involved in HSPC clonal expansion in 6,503 ESP samples.

RareTruncationVariants

Chr	Start	Ref	Var	Gene	Type	Annotation	European American MAF	African American MAF	Combined MAF	dbSNP_rsID
2	25458593	C	T	*DNMT3A*	Nonsense	p.W860*	0	0.0227	0.0077	rs376830288
2	25459834	C	A	*DNMT3A*	Nonsense	p.E817*	0.0116	0	0.0077	rs373873045
2	25464505	0	G	*DNMT3A*	Frame Shift Ins	p.I669fs	0	0.0234	0.008	NA
2	25466831	T	0	*DNMT3A*	Frame Shift Del	p.P625fs	0.0123	0.0239	0.0163	NA
2	25467468	G	C	*DNMT3A*	Nonsense	p.Y536*	0.0116	0	0.0077	rs370376334
2	25468163	C	A	*DNMT3A*	Nonsense	p.E505*	0.0116	0	0.0077	rs373860660
2	25469488	C	T	*DNMT3A*	Splice Site	e9+1	0.0117	0	0.0077	rs374440649
2	25469922	G	A	*DNMT3A*	Nonsense	p.Q374*	0	0.0227	0.0077	rs369109129
4	106156279	G	0	*TET2*	Frame Shift Del	p.A394fs	0.0121	0	0.008	NA
4	106156313	TTCT	0	*TET2*	Frame Shift Del	p.S407fs	0	0.0234	0.008	NA
4	106156687	C	T	*TET2*	Nonsense	p.Q530*	0.0116	0	0.0077	rs377382567
4	106157505	C	0	*TET2*	Frame Shift Del	p.Q803fs	0.0121	0	0.008	NA
4	106157506	C	T	*TET2*	Nonsense	p.Q803*	0.0116	0	0.0077	rs368508787
4	106157653	G	T	*TET2*	Nonsense	p.E852*	0.0116	0	0.0077	rs374928350
4	106157700	T	G	*TET2*	Nonsense	p.Y867*	0.0116	0	0.0077	rs145844118
4	106157808	C	0	*TET2*	Frame Shift Del	p.Q904fs	0.0363	0	0.024	NA
4	106158114	G	0	*TET2*	Frame Shift Del	p.V1006fs	0.0121	0	0.008	NA
4	106158157	C	T	*TET2*	Nonsense	p.Q1020*	0.0116	0	0.0077	rs375539032
4	106196213	C	T	*TET2*	Nonsense	p.R1516*	0.0314	0	0.0219	rs370735654
4	106196220	C	0	*TET2*	Frame Shift Del	p.S1518fs	0.0159	0	0.0106	NA
4	106197161	C	0	*TET2*	Frame Shift Del	p.L1832fs	0	0.0367	0.0131	NA
9	5022113	T	G	*JAK2*	Nonsense	p.Y42*	0.0116	0	0.0077	rs369748023
20	31021211	C	T	*ASXL1*	Nonsense	p.R404*	0	0.0454	0.0154	rs373145711
20	31021250	C	T	*ASXL1*	Nonsense	p.R417*	0.0116	0	0.0077	rs375215583
20	31021332	C	G	*ASXL1*	Nonsense	p.S444*	0.0116	0	0.0077	rs373126831
20	31022233	A	G	*ASXL1*	Splice Site	e13-2	0.0116	0	0.0077	rs376029425
20	31022288	C	G	*ASXL1*	Nonsense	p.Y591*	0.0116	0	0.0077	rs371369583
20	31022592	C	T	*ASXL1*	Nonsense	p.R693*	0	0.0227	0.0077	rs373221034
20	31022783	C	0	*ASXL1*	Frame Shift Del	p.Q757fs	0.0121	0	0.008	NA

KnownHotSpotVariants

Chr	Start	Ref	Var	Gene	Type	Annotation	European American MAF	African American MAF	Combined MAF	dbSNP_rsID
2	25457242	C	T	*DNMT3A*	Missense	p.R882H	0.0581	0.0908	0.0692	rs147001633
2	25457243	G	A	*DNMT3A*	Missense	p.R882C	0.0465	0	0.0308	rs377577594
9	5073770	G	T	*JAK2*	Missense	p.V617F	0.0233	0.0227	0.0231	rs77375493

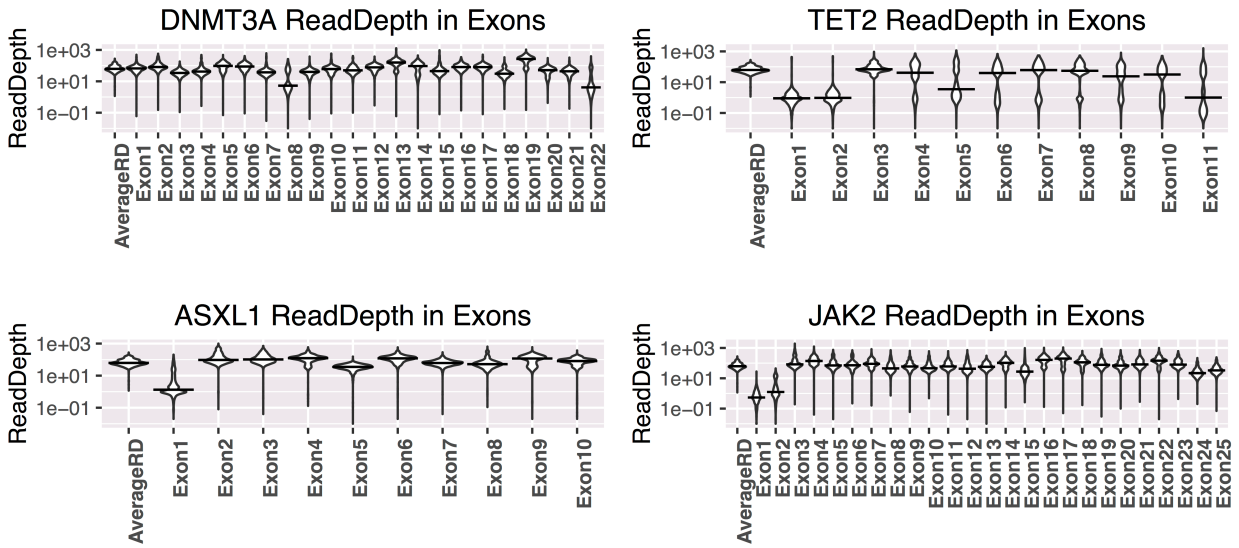
Supplementary Table 3.12: Rare truncation variants and known hotspot variants detected in *DNMT3A*, *TET2*, *ASXL1*, *GNAS*, *JAK2*, *SF3B1*, *IDH1*, and *IDH2* in 557 WHISP samples.

Sample_ID	Chr	Start	Ref	Var	Gene	Type	Annotation
dbGaP-298170-361467	20	31022288	C	G	*ASXL1*	Nonsense	p.Y591*
dbGaP-295088-361485	2	25457242	C	T	*DNMT3A*	Missense	p.R882H
dbGaP-295464-361677	2	25457242	C	T	*DNMT3A*	Missense	p.R882H
dbGaP-298112-361837	2	25457242	C	T	*DNMT3A*	Missense	p.R882H
dbGaP-353698-437130	2	25457242	C	T	*DNMT3A*	Missense	p.R882H
dbGaP-353485-437074	2	25459834	C	A	*DNMT3A*	Nonsense	p.E817*
dbGaP-353587-436912	2	25457243	G	A	*DNMT3A*	Missense	p.R882C
dbGaP-294968-360378	2	25457243	G	A	*DNMT3A*	Missense	p.R882C
dbGaP-295037-360660	2	25457243	G	A	*DNMT3A*	Missense	p.R882C
dbGaP-295149-361236	20	57428858	C	T	*GNAS*	Nonsense	p.Q180*
dbGaP-295493-361708	20	57484421	G	A	*GNAS*	Missense	R202H
dbGaP-294874-361374	9	5073770	G	T	*JAK2*	Missense	p.V617F
dbGaP-298062-361478	9	5073770	G	T	*JAK2*	Missense	p.V617F
dbGaP-295219-361137	9	5073770	G	T	*JAK2*	Missense	p.V617F
dbGap-390083-851772	2	198266834	T	C	*SF3B1*	Missense	p.K700E
dbGaP-297986-360566	2	198266834	T	C	*SF3B1*	Missense	p.K700E
dbGaP-399048-661495	4	106180783	-	G	*TET2*	Frame shift insertion	p.C1271fs

Supplementary Table 3.13: Exome capture sequencing coverage for 11 TCGA cancer types analyzed.

Cancer Type	Minimum Depth 1X (%)	Minimum Depth 5X (%)	Minimum Depth 10X (%)	Minimum Depth 15X (%)	Minimum Depth 20X (%)	Mean Depth (X)
BRCA	88.6 +/- 4.4	75.7 +/- 2.6	71.7 +/- 2.9	68.6 +/- 3.7	65.7 +/- 4.7	96.8 +/- 44.3
GBM	97.0 +/- 3.4	93.6 +/- 6.1	90.9 +/- 7.3	88.6 +/- 8.0	86.4 +/- 8.6	139.4 +/- 41.9
HNSC	98.9 +/- 0.2	96.9 +/- 0.5	94.5 +/- 1.1	92.0 +/- 1.6	89.5 +/- 2.1	89.5 +/- 19.7
KIRC	86.4 +/- 6.5	75.3 +/- 3.3	71.3 +/- 2.8	68.7 +/- 3.3	66.3 +/- 4.0	119.7 +/- 57.7
LGG	98.9 +/- 0.2	96.9 +/- 0.6	94.4 +/- 1.2	92.0 +/- 1.9	89.5 +/- 2.5	99.9 +/- 24.9
LUAD	79.1 +/- 2.2	73.6 +/- 0.8	70.9 +/- 1.2	68.4 +/- 1.7	66.0 +/- 2.2	77.9 +/- 21.1
LUSC	78.7 +/- 1.3	73.8 +/- 0.6	71.2 +/- 1.1	68.8 +/- 1.5	66.6 +/- 2.0	80.5 +/- 26.4
OV	95.3 +/- 3.6	89.0 +/- 6.7	85.3 +/- 7.0	82.5 +/- 7.3	79.9 +/- 7.8	150.2 +/- 70.0
PRAD	79.8 +/- 2.9	74.2 +/- 0.8	71.8 +/- 1.0	69.7 +/- 1.4	67.7 +/- 1.8	95.1 +/- 25.0
STAD	99.1 +/- 0.3	97.2 +/- 0.8	95.0 +/- 1.5	92.6 +/- 2.4	90.2 +/- 3.4	106.3 +/- 27.8
UCEC	86.7 +/- 3.1	75.4 +/- 2.3	72.0 +/- 3.1	69.4 +/- 3.9	67.1 +/- 4.8	128.2 +/- 47.5
TOTAL	90.7 +/- 7.9	83.9 +/- 10.5	80.8 +/- 11.0	78.2 +/- 11.3	75.6 +/- 11.6	107.5 +/- 47.1

Appendix 3 Supplementary Materials for Chapter 4



Supplementary Figure 4.1: Read-depth of the 4 most recurrently mutated genes in the normal blood sample. Violin plot shows the distribution of read depth on each exon across all of the samples included in the study. The horizontal bar indicates the median of read depth.

Supplementary Table 4.1: AML hotspot mutations in normal blood samples

Chr	Start	End	Reference	Variant	Gene	Mutation	Sample
1	115256528	115256528	T	A	*NRAS*	p.Q61H	TCGA-A2-A04R
1	115258744	115258744	C	T	*NRAS*	p.G13D	TCGA-F7-A622
1	115258747	115258747	C	T	*NRAS*	p.G12D	TCGA-36-1581
1	115258747	115258747	C	T	*NRAS*	p.G12D	TCGA-91-A4BD
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-DK-A1AD
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-81-5910
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-21-5783
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-AK-3425
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-E9-A1RG
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-27-2526
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-E8-A44K
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-BJ-A28S
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-DJ-A2QB
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-B1-A654
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-SL-A6JA
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-FS-A4FC
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-D1-A17H
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-FD-A6TK
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-FG-7634
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-24-2280
2	25457242	25457242	C	T	*DNMT3A*	p.R882H	TCGA-DX-A6BB
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-GF-A3OT
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-BR-7197
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-A2-A1G1
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-BR-8081
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-RP-A690
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-85-7698
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-B2-4098
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-B5-A1MZ
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-24-1469
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-D5-6931
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-EY-A1GS
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-B0-5698
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-41-2571
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-AK-3447
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-D1-A0ZS
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-LL-A5YM
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-EE-A29V
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-26-5135
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-EY-A1G8
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-DJ-A2Q2
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-05-4403
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-BK-A139
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-AP-A0LQ
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-BM-6198
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-B4-5834
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-99-8025
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-06-0240
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-A6-2686
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-TQ-A7RQ
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-HW-7495
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-24-1603
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-EJ-5521
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-06-2563

9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-CV-A6JD
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-J9-A8CP
9	5073770	5073770	G	T	*JAK2*	p.V617F	TCGA-AP-A0LD
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-AU-3779
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-D7-8574
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-B5-A11L
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-ER-A194
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-23-1109
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-B5-A3F9
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-EE-A2ML
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-09-2053
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-A6-4105
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-EE-A2MK
15	90631934	90631934	C	T	*IDH2*	p.R140Q	TCGA-IR-A3LA
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-67-4679
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-19-2625
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-EY-A1GU
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-D6-A4Z9
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-HC-A6AN
20	57484421	57484421	G	A	*GNAS*	p.R202H	TCGA-A4-8515

Supplementary Table 4.2: Blood-specific somatic mutation identified in cancer or AML-associated genes

Chr	Start	End	Reference	Alteration	Gene	Mutation	Sample	Gene Type
1	12921594	12921594	C	G	*PRAMEF2*	p.S462*	TCGA-13-1412	AML
1	17256459	17256459	G	A	*CROCC*	p.R157H	TCGA-13-2065	AML
1	17264170	17264170	G	C	*CROCC*	p.E410Q	TCGA-99-8025	AML
1	17275365	17275365	G	A	*CROCC*	p.R927Q	TCGA-KM-8442	AML
1	24384031	24384031	G	T	*MYOM3*	p.D1382E	TCGA-85-A5B5	AML
1	24389994	24389994	C	T	*MYOM3*	p.G1220S	TCGA-24-0968	AML
1	24409143	24409143	C	T	*MYOM3*	p.E679K	TCGA-CI-6620	AML
1	24417427	24417427	C	T	*MYOM3*	p.R432Q	TCGA-C8-A278	AML
1	33134591	33134591	A	G	*RBBP4*	p.I207V	TCGA-BR-8686	AML
1	33138072	33138072	G	A	*RBBP4*	p.E330K	TCGA-37-4130	AML
1	35250428	35250428	G	A	*GJB3*	p.R22H	TCGA-95-7043	AML
1	35250691	35250691	G	A	*GJB3*	p.G110R	TCGA-AX-A2H7	AML
1	36752630	36752630	C	A	*THRAP3*	p.P267T	TCGA-62-8394	AML
1	151263210	151263210	G	A	*ZNF687*	p.R1080H	TCGA-86-6851	AML
1	152281645	152281645	G	T	*FLG*	p.S1906*	TCGA-BR-A4J1	AML
1	152281645	152281645	G	T	*FLG*	p.S1906*	TCGA-UY-A78M	AML
1	154938191	154938191	C	T	*SHC1*	p.R485Q	TCGA-CI-6620	AML
1	154938931	154938931	T	G	*SHC1*	p.Y349S	TCGA-12-0773	AML
1	181547003	181547003	C	T	*CACNA1E*	p.P205L	TCGA-EP-A2KB	AML
1	181700330	181700330	C	T	*CACNA1E*	p.H754Y	TCGA-BS-A0TC	AML
1	181731716	181731716	T	C	*CACNA1E*	p.F1538L	TCGA-AK-3428	AML
1	185878613	185878613	C	T	*HMCN1*	p.P256S	TCGA-CR-7370	AML
1	185891583	185891583	C	T	*HMCN1*	p.R325*	TCGA-CI-6620	AML
1	185891589	185891589	C	T	*HMCN1*	p.P327S	TCGA-12-1095	AML
1	185934959	185934959	G	C	*HMCN1*	p.Q708H	TCGA-05-4403	AML
1	185963964	185963964	C	G	*HMCN1*	p.R1175G	TCGA-29-1776	AML
1	185969187	185969187	A	C	*HMCN1*	p.K1295N	TCGA-AG-3598	AML
1	186030998	186030998	G	A	*HMCN1*	p.R2443Q	TCGA-04-1542	AML
1	186055509	186055509	A	T	*HMCN1*	p.K3006*	TCGA-86-8359	AML
1	186062355	186062355	G	A	*HMCN1*	p.G3326E	TCGA-GV-A3JW	AML
1	186107069	186107069	T	C	*HMCN1*	p.V4630A	TCGA-C8-A1HJ	AML
1	186134259	186134259	G	C	*HMCN1*	p.Q5091H	TCGA-97-8175	AML
1	190068155	190068155	G	T	*FAM5C*	p.Q432K	TCGA-DD-A115	AML
1	190129872	190129872	C	G	*FAM5C*	p.Q370H	TCGA-13-0717	AML
1	190129876	190129876	G	A	*FAM5C*	p.A369V	TCGA-AG-3583	AML
1	190129930	190129930	A	T	*FAM5C*	p.F351Y	TCGA-13-1408	AML
1	190423789	190423789	A	G	*FAM5C*	p.Y78H	TCGA-IA-A40Y	AML
1	201772790	201772790	G	T	*NAV1*	p.S1196I	TCGA-AG-3598	AML
1	201781732	201781732	C	G	*NAV1*	p.L1722V	TCGA-AF-2689	AML
2	21227444	21227444	A	T	*APOB*	p.Y3964*	TCGA-AG-3584	AML
2	21233706	21233706	G	A	*APOB*	p.R2012*	TCGA-D6-A6EM	AML
2	21249752	21249752	C	T	*APOB*	p.A718T	TCGA-36-1580	AML
2	21249752	21249752	C	T	*APOB*	p.A718T	TCGA-AA-3561	AML
2	21255274	21255274	T	G	*APOB*	p.D435A	TCGA-25-1317	AML
2	54120852	54120852	C	T	*PSME4*	p.G1333R	TCGA-13-0717	AML
2	54120852	54120852	C	T	*PSME4*	p.G1333R	TCGA-AA-3538	AML
2	54120852	54120852	C	T	*PSME4*	p.G1333R	TCGA-AA-3549	AML
2	54125056	54125056	A	G	*PSME4*	p.I1186T	TCGA-32-1970	AML
2	54135476	54135476	T	C	*PSME4*	p.K922R	TCGA-EL-A3GQ	AML
2	54150219	54150219	G	T	*PSME4*	p.H649N	TCGA-AA-A01G	AML
2	54164535	54164535	C	T	*PSME4*	p.G230S	TCGA-D8-A143	AML
2	71742787	71742787	C	T	*DYSF*	p.P264L	TCGA-EK-A2PG	AML
2	71797819	71797819	G	T	*DYSF*	p.R1072L	TCGA-F7-A61W	AML

2	71801350	71801350	G	C	*DYSF*	p.G1097A	TCGA-97-7546	AML
2	71801350	71801350	G	C	*DYSF*	p.G1097A	TCGA-EM-A4FO	AML
2	71817373	71817373	T	C	*DYSF*	p.Y1190H	TCGA-AA-3514	AML
2	71838638	71838638	C	G	*DYSF*	p.A1381G	TCGA-AA-A00R	AML
2	71839796	71839797	-	C	*DYSF*	p.I1432fs	TCGA-AA-3955	AML
2	71894538	71894538	C	T	*DYSF*	p.P1776S	TCGA-36-1570	AML
2	71906190	71906190	C	T	*DYSF*	p.S1955F	TCGA-29-1774	AML
2	71906190	71906190	C	T	*DYSF*	p.S1955F	TCGA-CI-6620	AML
2	166912980	166912980	A	C	*SCN1A*	p.I138M	TCGA-13-1408	AML
2	166930064	166930064	G	A	*SCN1A*	p.A23V	TCGA-29-1702	AML
2	210559124	210559124	G	A	*MAP2*	p.D744N	TCGA-13-1497	AML
2	210560193	210560193	A	C	*MAP2*	p.K1100T	TCGA-AA-A00W	AML
2	210570439	210570439	C	T	*MAP2*	p.R1574W	TCGA-CL-5917	AML
2	220342509	220342509	T	G	*SPEG*	NULL	TCGA-A5-A0GB	AML
2	233613710	233613710	T	C	*GIGYF2*	p.L62P	TCGA-AA-A00D	AML
2	233620934	233620934	G	A	*GIGYF2*	p.R90K	TCGA-19-0955	AML
2	233626138	233626138	A	C	*GIGYF2*	p.K175T	TCGA-19-0955	AML
3	48678907	48678907	C	T	*CELSR3*	p.E2964K	TCGA-CI-6620	AML
3	48683549	48683549	G	C	*CELSR3*	p.I2484M	TCGA-AK-3461	AML
3	49694232	49694232	G	A	*BSN*	p.E2415K	TCGA-D5-6540	AML
3	49698124	49698124	G	A	*BSN*	p.R2949Q	TCGA-GC-A3I6	AML
3	54913412	54913412	T	A	*CACNA2D3*	p.Y606*	TCGA-06-6694	AML
3	62484842	62484842	C	T	*CADPS*	p.A901T	TCGA-56-7223	AML
3	62518612	62518612	G	A	*CADPS*	p.A742V	TCGA-25-1625	AML
3	62522124	62522124	C	A	*CADPS*	p.C700F	TCGA-09-1667	AML
3	62522128	62522128	A	T	*CADPS*	p.S699T	TCGA-23-1109	AML
3	62522128	62522128	A	T	*CADPS*	p.S699T	TCGA-36-1570	AML
3	62636542	62636542	C	T	*CADPS*	p.V395M	TCGA-EY-A1GR	AML
3	71026813	71026813	T	C	*FOXP1*	p.Y472C	TCGA-CV-7437	AML
3	85851258	85851258	T	G	*CADM2*	p.I43M	TCGA-AG-3605	AML
3	86010627	86010627	C	A	*CADM2*	p.P260H	TCGA-24-2030	AML
3	86010704	86010704	C	A	*CADM2*	p.L286I	TCGA-AA-3846	AML
3	121725954	121725954	A	G	*ILDR1*	p.F38S	TCGA-AA-3514	AML
3	121725954	121725954	A	G	*ILDR1*	p.F38S	TCGA-AA-A00K	AML
3	123010193	123010193	G	T	*ADCY5*	p.L690M	TCGA-X6-A7WB	AML
3	123014956	123014956	C	T	*ADCY5*	p.R671H	TCGA-D5-6538	AML
3	164716417	164716417	C	T	*SI*	p.R1484H	TCGA-DX-A7EF	AML
3	185644503	185644503	C	T	*TRA2B*	p.R19K	TCGA-CI-6620	AML
4	36115858	36115858	T	A	*ARAP2*	p.K1364*	TCGA-24-1548	AML
4	36148928	36148928	C	G	*ARAP2*	p.E1085Q	TCGA-AG-3608	AML
4	36150079	36150079	A	G	*ARAP2*	p.F983S	TCGA-14-0783	AML
4	48523085	48523085	C	T	*FRYL*	p.A2557T	TCGA-ER-A19H	AML
4	48621344	48621344	C	T	*FRYL*	p.D120N	TCGA-85-8355	AML
4	48636332	48636332	A	C	*FRYL*	p.I32M	TCGA-13-0761	AML
4	89013415	89013415	T	G	*ABCG2*	p.K647Q	TCGA-AA-3534	AML
4	89022391	89022391	T	C	*ABCG2*	p.K453R	TCGA-AG-A016	AML
4	114117617	114117617	A	T	*ANK2*	p.T94S	TCGA-AG-3578	AML
4	114195805	114195805	G	T	*ANK2*	p.K561N	TCGA-AK-3434	AML
4	114203861	114203861	A	G	*ANK2*	p.K638E	TCGA-13-1408	AML
4	151203760	151203760	A	C	*LRBA*	p.L2731V	TCGA-AA-A00W	AML
4	151231390	151231390	T	C	*LRBA*	p.T2625A	TCGA-61-1998	AML
4	151236754	151236754	T	G	*LRBA*	p.D2551A	TCGA-B5-A0K1	AML
4	151520191	151520191	G	A	*LRBA*	p.P2005L	TCGA-AX-A062	AML
4	151749573	151749573	G	C	*LRBA*	p.L1644V	TCGA-LI-A67I	AML
4	151850215	151850215	C	A	*LRBA*	p.L73F	TCGA-AK-3428	AML
4	155156529	155156529	C	G	*DCHS2*	p.S2637T	TCGA-AA-A00J	AML
4	155157548	155157548	A	C	*DCHS2*	p.F2297L	TCGA-B2-4101	AML
4	155241767	155241767	T	C	*DCHS2*	p.N1140S	TCGA-16-1055	AML

4	155241767	155241767	T	C	*DCHS2*	p.N1140S	TCGA-AG-A016	AML
5	13721312	13721312	C	T	*DNAH5*	p.E4026K	TCGA-CI-6620	AML
5	13727662	13727662	C	T	*DNAH5*	p.R3996H	TCGA-DM-A1D6	AML
5	13776574	13776574	C	T	*DNAH5*	p.R3116Q	TCGA-CI-6620	AML
5	13814934	13814934	A	C	*DNAH5*	p.L2337R	TCGA-A8-A094	AML
5	13820505	13820505	C	T	*DNAH5*	p.S2264N	TCGA-B2-4101	AML
5	13820573	13820573	C	A	*DNAH5*	p.W2241C	TCGA-62-8394	AML
5	13870998	13870998	C	T	*DNAH5*	p.E1238K	TCGA-CI-6620	AML
5	13928265	13928265	A	C	*DNAH5*	p.F72C	TCGA-25-1324	AML
5	71493821	71493821	G	T	*MAP1B*	p.E1547*	TCGA-04-1364	AML
5	79031579	79031579	G	A	*CMYA5*	p.E2331K	TCGA-B5-A0JN	AML
5	83476279	83476279	C	T	*EDIL3*	p.R96Q	TCGA-CI-6620	AML
5	137722143	137722144	-	T	*KDM3B*	p.E405fs	TCGA-AX-A06F	AML
5	137727689	137727689	C	T	*KDM3B*	p.P790S	TCGA-CI-6620	AML
5	140263648	140263648	G	A	*PCDHA13*	p.A599T	TCGA-A8-A06Y	AML
5	149786745	149786745	G	T	*CD74*	p.L90M	TCGA-DX-A6YT	AML
5	151775137	151775137	C	T	*NMUR2*	p.V274I	TCGA-AQ-A54N	AML
5	160029624	160029624	C	T	*ATP10B*	p.R1108H	TCGA-13-0761	AML
5	169116314	169116314	A	G	*DOCK2*	p.N274D	TCGA-AA-3514	AML
5	169125401	169125401	A	G	*DOCK2*	p.K335E	TCGA-AA-A00W	AML
5	169507174	169507174	G	T	*DOCK2*	p.R1217M	TCGA-AK-3461	AML
5	172750262	172750262	T	A	*STC2*	p.I156L	TCGA-AG-3598	AML
5	172750273	172750273	T	C	*STC2*	p.N152S	TCGA-24-1105	AML
6	12125879	12125879	A	G	*HIVEP1*	p.K1951E	TCGA-DU-7309	AML
6	12719019	12719019	G	A	*PHACTR1*	p.A15T	TCGA-13-0725	AML
6	13287299	13287299	C	T	*PHACTR1*	p.H142Y	TCGA-CI-6620	AML
6	30884672	30884672	G	C	*VARS2*	p.C260S	TCGA-AG-A016	AML
6	33633644	33633644	T	G	*ITPR3*	p.F481C	TCGA-AA-3514	AML
6	33639851	33639851	C	G	*ITPR3*	p.S925C	TCGA-AK-3428	AML
6	33655931	33655931	C	T	*ITPR3*	p.T2121M	TCGA-GV-A3JX	AML
6	42819837	42819837	A	G	*KIAA0240*	p.K616R	TCGA-13-1408	AML
6	43221065	43221065	G	A	*TTBK1*	p.R47K	TCGA-A6-2681	AML
6	56335969	56335969	C	T	*DST*	p.R7388K	TCGA-B5-A3F9	AML
6	56391358	56391358	T	G	*DST*	p.K5946T	TCGA-13-0801	AML
6	56417676	56417676	T	G	*DST*	p.Q5274P	TCGA-AK-3460	AML
6	56425105	56425105	C	G	*DST*	p.M4778I	TCGA-EB-A42Y	AML
6	56473368	56473368	T	G	*DST*	p.T1987P	TCGA-B2-4101	AML
6	56480676	56480676	G	T	*DST*	p.P2530H	TCGA-AA-3534	AML
6	56481996	56481996	T	C	*DST*	p.Q2090R	TCGA-09-1666	AML
6	56482951	56482951	G	C	*DST*	p.Q1961E	TCGA-26-1440	AML
6	56489323	56489323	A	-	*DST*	p.I1108fs	TCGA-97-7938	AML
6	56492043	56492043	C	T	*DST*	p.E1024K	TCGA-CI-6620	AML
6	56496073	56496073	C	A	*DST*	p.E823*	TCGA-AX-A063	AML
6	56499010	56499010	G	A	*DST*	p.H644Y	TCGA-86-8281	AML
6	56505044	56505044	C	T	*DST*	p.R259Q	TCGA-KV-A6GD	AML
6	56707877	56707877	C	T	*DST*	p.E23K	TCGA-CI-6620	AML
6	72957796	72957796	A	G	*RIMS1*	p.K736R	TCGA-B2-4101	AML
6	72960057	72960057	G	A	*RIMS1*	p.G756R	TCGA-EJ-5531	AML
6	73100438	73100438	A	G	*RIMS1*	p.N1351S	TCGA-24-1105	AML
6	75839883	75839883	C	T	*COL12A1*	p.W2045*	TCGA-41-2571	AML
6	75899528	75899528	C	T	*COL12A1*	p.C133Y	TCGA-68-8251	AML
6	79655838	79655838	C	T	*PHIP*	p.V1504I	TCGA-AA-A00R	AML
6	79688334	79688334	G	A	*PHIP*	p.P955L	TCGA-AA-3514	AML
6	79697925	79697925	C	T	*PHIP*	e21+1	TCGA-06-0188	AML
6	79711697	79711697	A	G	*PHIP*	p.S600P	TCGA-A6-2676	AML
6	79711697	79711697	A	G	*PHIP*	p.S600P	TCGA-AG-3609	AML
6	79711697	79711697	A	G	*PHIP*	p.S600P	TCGA-AK-3428	AML
6	79711781	79711781	A	T	*PHIP*	p.F572I	TCGA-AA-3529	AML

6	79725417	79725417	A	G	*PHIP*	p.I440T	TCGA-B2-4101	AML
6	79735858	79735858	T	G	*PHIP*	p.K208N	TCGA-13-1408	AML
6	79787752	79787752	G	A	*PHIP*	p.R12*	TCGA-CR-7379	AML
6	102307352	102307352	G	A	*GRIK2*	p.R503H	TCGA-06-0142	AML
6	155126538	155126538	T	G	*SCAF8*	p.F366C	TCGA-AA-3534	AML
6	155145471	155145471	C	T	*SCAF8*	p.P743L	TCGA-19-0962	AML
7	34086018	34086018	G	T	*BMPER*	e7+1	TCGA-86-8359	AML
7	34125436	34125436	C	A	*BMPER*	p.L493I	TCGA-DY-A1DD	AML
7	34192817	34192817	C	T	*BMPER*	p.H664Y	TCGA-06-2565	AML
7	43540317	43540317	C	T	*HECW1*	p.R1153W	TCGA-HI-7170	AML
7	77649126	77649126	G	A	*MAGI2*	p.H1292Y	TCGA-FJ-A3ZF	AML
7	77830478	77830478	C	A	*MAGI2*	e11+1	TCGA-AA-A01F	AML
7	83606460	83606460	C	G	*SEMA3A*	p.D569H	TCGA-13-0792	AML
7	83634817	83634817	T	C	*SEMA3A*	p.I400V	TCGA-E2-A574	AML
7	83739800	83739800	C	T	*SEMA3A*	p.G147R	TCGA-AK-3433	AML
7	113517786	113517786	T	C	*PPP1R3A*	p.K1121E	TCGA-12-1098	AML
7	113558612	113558612	C	T	*PPP1R3A*	p.R147Q	TCGA-CG-5717	AML
7	126086275	126086275	G	T	*GRM8*	p.T861K	TCGA-25-1625	AML
7	126249534	126249534	A	T	*GRM8*	p.V459D	TCGA-CH-5750	AML
7	126882877	126882877	C	G	*GRM8*	p.A128P	TCGA-K4-A4AC	AML
7	150643964	150643964	C	T	*KCNH2*	e14+1	TCGA-AK-3447	AML
7	150647322	150647322	C	T	*KCNH2*	p.A778T	TCGA-F5-6810	AML
7	150647343	150647343	G	A	*KCNH2*	p.H771Y	TCGA-AX-A05T	AML
7	151842344	151842344	G	A	*MLL3*	p.R4747*	TCGA-CD-A48A	AML
7	151851121	151851121	T	G	*MLL3*	p.I4141L	TCGA-13-0765	AML
7	151856102	151856102	T	C	*MLL3*	p.K3839R	TCGA-25-1324	AML
7	151873387	151873387	G	C	*MLL3*	p.Q3051E	TCGA-A6-2676	AML
7	151873410	151873410	T	G	*MLL3*	p.E3043A	TCGA-AA-A00K	AML
7	151873423	151873423	G	C	*MLL3*	p.Q3039E	TCGA-AA-A00K	AML
7	151921520	151921520	C	A	*MLL3*	p.W1053L	TCGA-EE-A2MH	AML
7	151927049	151927049	G	A	*MLL3*	p.Q979*	TCGA-AR-A24V	AML
7	151927058	151927058	C	T	*MLL3*	p.A976T	TCGA-06-0157	AML
7	151945038	151945038	T	A	*MLL3*	p.K827N	TCGA-F4-6459	AML
7	151962135	151962135	C	T	*MLL3*	p.C391Y	TCGA-EY-A4KR	AML
7	151962265	151962265	C	T	*MLL3*	p.D348N	TCGA-BR-6801	AML
7	151962265	151962265	C	T	*MLL3*	p.D348N	TCGA-ER-A194	AML
8	3165912	3165912	C	T	*CSMD1*	p.A1250T	TCGA-13-0793	AML
8	75898242	75898242	A	C	*CRISPLD1*	p.E7A	TCGA-23-1024	AML
8	110457104	110457104	G	A	*PKHD1L1*	p.G1669E	TCGA-BR-A4QL	AML
8	110457292	110457292	C	T	*PKHD1L1*	p.P1732S	TCGA-86-8280	AML
8	110477081	110477081	G	A	*PKHD1L1*	p.V2674M	TCGA-B6-A408	AML
8	110477426	110477426	C	T	*PKHD1L1*	p.R2789C	TCGA-CI-6620	AML
8	110523026	110523026	T	G	*PKHD1L1*	p.C3806G	TCGA-AP-A05H	AML
8	113259329	113259329	T	G	*CSMD3*	p.K3381T	TCGA-13-0889	AML
8	113275940	113275940	C	T	*CSMD3*	p.E3264K	TCGA-06-6698	AML
8	113293406	113293406	T	C	*CSMD3*	p.T3169A	TCGA-D5-5538	AML
8	113299440	113299440	G	A	*CSMD3*	p.P3062S	TCGA-DQ-7596	AML
8	113349873	113349873	G	C	*CSMD3*	p.S2247*	TCGA-AG-3598	AML
8	113364731	113364731	T	C	*CSMD3*	p.T2057A	TCGA-AA-A00J	AML
8	113516069	113516069	A	G	*CSMD3*	p.F1678S	TCGA-16-1055	AML
8	113529278	113529278	G	T	*CSMD3*	p.P1581T	TCGA-FD-A6TC	AML
8	113562996	113562996	T	G	*CSMD3*	p.I1490L	TCGA-24-1105	AML
8	113697710	113697710	T	C	*CSMD3*	p.N803D	TCGA-24-1548	AML
8	113841948	113841948	C	T	*CSMD3*	p.G609D	TCGA-CR-5247	AML
8	113960037	113960037	C	A	*CSMD3*	p.R497I	TCGA-BK-A4ZD	AML
8	114326942	114326942	T	A	*CSMD3*	p.N87Y	TCGA-AA-A00W	AML
9	34485158	34485158	G	A	*DNAI1*	p.E34K	TCGA-30-1853	AML
9	73151464	73151464	A	C	*TRPM3*	p.F1537C	TCGA-25-1324	AML

9	73442787	73442787	T	C	*TRPM3*	p.I164V	TCGA-A4-7584	AML
9	86589485	86589485	C	G	*HNRNPK*	p.D93H	TCGA-25-1627	AML
9	117803337	117803337	C	T	*TNC*	p.G1759R	TCGA-CL-5917	AML
9	117825447	117825447	T	G	*TNC*	p.N1261T	TCGA-AG-A016	AML
9	117848674	117848674	G	C	*TNC*	p.R446G	TCGA-36-1577	AML
9	123797174	123797174	T	C	*C5*	e5-2	TCGA-AA-A00W	AML
9	138667190	138667190	A	G	*KCNT1*	p.I719V	TCGA-AA-A00K	AML
9	138667233	138667233	C	T	*KCNT1*	p.A733V	TCGA-MF-A522	AML
9	138683984	138683984	A	G	*KCNT1*	p.T1209A	TCGA-K4-A4AB	AML
9	140772671	140772671	T	C	*CACNA1B*	e1+1	TCGA-PJ-A5Z8	AML
9	140773504	140773504	A	C	*CACNA1B*	e2-1	TCGA-E2-A158	AML
9	140880870	140880870	G	T	*CACNA1B*	p.W593L	TCGA-23-1120	AML
9	140952560	140952560	G	A	*CACNA1B*	p.R1390H	TCGA-FS-A1ZC	AML
9	141010071	141010071	G	A	*CACNA1B*	p.R1907Q	TCGA-24-1845	AML
9	141016381	141016381	C	T	*CACNA1B*	p.T2318I	TCGA-E1-A7YI	AML
10	46965806	46965806	A	C	*SYT15*	p.L297R	TCGA-FS-A4FD	AML
10	73044507	73044507	T	C	*UNC5B*	p.V112A	TCGA-AA-3514	AML
11	1261446	1261446	C	T	*MUC5B*	p.Q1274*	TCGA-QH-A6CX	AML
11	11400736	11400736	C	T	*GALNTL4*	p.R224H	TCGA-62-8394	AML
11	19258879	19258879	C	T	*E2F8*	p.E145K	TCGA-CI-6620	AML
11	30032436	30032436	A	G	*KCNA4*	p.F597S	TCGA-23-1028	AML
11	30033309	30033309	G	A	*KCNA4*	p.A306V	TCGA-A4-A48D	AML
11	30034149	30034149	C	T	*KCNA4*	p.R26Q	TCGA-FG-7637	AML
11	85977208	85977208	G	T	*EED*	p.M270I	TCGA-AG-A016	AML
11	100221518	100221518	G	C	*CNTN5*	p.G1039A	TCGA-50-8460	AML
11	102984369	102984369	T	C	*DYNC2H1*	p.M100T	TCGA-IR-A3LA	AML
11	102987372	102987372	A	G	*DYNC2H1*	p.D232G	TCGA-AK-3460	AML
11	103026079	103026079	A	G	*DYNC2H1*	p.K1198R	TCGA-13-0802	AML
11	103026079	103026079	A	G	*DYNC2H1*	p.K1198R	TCGA-13-1408	AML
11	103026079	103026079	A	G	*DYNC2H1*	p.K1198R	TCGA-24-1604	AML
11	103091352	103091352	C	G	*DYNC2H1*	p.P2983A	TCGA-95-7567	AML
11	103124071	103124071	G	A	*DYNC2H1*	p.R3374H	TCGA-CF-A3MH	AML
11	113281549	113281549	A	C	*DRD2*	p.F413C	TCGA-24-1548	AML
11	120833197	120833197	G	T	*GRIK4*	p.W691C	TCGA-AG-A016	AML
11	120833341	120833341	G	C	*GRIK4*	p.Q739H	TCGA-B2-4101	AML
12	118298112	118298112	C	T	*KSR2*	p.R102H	TCGA-CH-5753	AML
12	125284776	125284776	A	G	*SCARB1*	p.F287S	TCGA-AA-3514	AML
13	29933477	29933477	G	T	*MTUS2*	p.R1005L	TCGA-86-8359	AML
13	30003025	30003025	G	A	*MTUS2*	p.D16N	TCGA-EY-A3L3	AML
13	36428682	36428682	G	A	*DCLK1*	p.S330L	TCGA-36-2547	AML
13	36700129	36700129	G	A	*DCLK1*	p.T49M	TCGA-D1-A17H	AML
13	39261852	39261852	C	T	*FREM2*	p.P124L	TCGA-AK-3444	AML
13	39263661	39263661	G	A	*FREM2*	p.R727H	TCGA-G2-A3VY	AML
13	39264192	39264192	A	G	*FREM2*	p.H904R	TCGA-AA-A00R	AML
13	39450473	39450473	A	G	*FREM2*	p.N2833S	TCGA-A6-6782	AML
13	39452990	39452990	A	-	*FREM2*	p.A2962fs	TCGA-A2-A04U	AML
13	39454448	39454448	C	T	*FREM2*	p.H3012Y	TCGA-AA-3514	AML
13	101714442	101714442	G	A	*NALCN*	p.R1545W	TCGA-BF-A3DL	AML
13	101721047	101721047	C	A	*NALCN*	p.A1444S	TCGA-EE-A2MK	AML
13	101728228	101728228	T	G	*NALCN*	p.K1317T	TCGA-13-0761	AML
13	101735528	101735528	T	C	*NALCN*	p.K1202R	TCGA-B2-4101	AML
13	101944681	101944681	T	G	*NALCN*	p.Q279P	TCGA-AG-3598	AML
13	102047560	102047560	T	A	*NALCN*	p.K89*	TCGA-19-1392	AML
13	102051404	102051404	G	A	*NALCN*	p.S25L	TCGA-DJ-A2Q2	AML
13	114498193	114498193	T	C	*FAM70B*	p.F109L	TCGA-AA-A00K	AML
13	114530118	114530118	C	T	*GAS6*	p.G486D	TCGA-16-1055	AML
14	68040028	68040028	G	A	*PLEKHH1*	p.M588I	TCGA-24-2019	AML
14	79432521	79432521	A	C	*NRXN3*	p.K839T	TCGA-13-1408	AML

14	86089561	86089561	A	C	*FLRT2*	p.K568T	TCGA-AF-3400	AML
14	90650986	90650986	G	T	*KCNK13*	p.W289L	TCGA-AJ-A3TW	AML
15	33925206	33925206	T	C	*RYR3*	p.L975P	TCGA-AP-A051	AML
15	33952554	33952554	C	T	*RYR3*	p.R1518C	TCGA-B6-A0IM	AML
15	34016358	34016358	G	A	*RYR3*	p.R2298H	TCGA-04-1530	AML
15	34042369	34042369	G	A	*RYR3*	p.G2761S	TCGA-A5-A0GA	AML
15	34135707	34135707	T	G	*RYR3*	p.F4410V	TCGA-13-0802	AML
15	85382266	85382266	C	A	*ALPK3*	p.Y322*	TCGA-09-1662	AML
15	85382266	85382266	C	A	*ALPK3*	p.Y322*	TCGA-AG-3598	AML
15	86259127	86259127	A	G	*AKAP13*	p.K1907R	TCGA-AA-A00R	AML
15	86266458	86266458	A	G	*AKAP13*	p.S2222G	TCGA-AR-A2LK	AML
15	86270642	86270642	T	G	*AKAP13*	p.S2349R	TCGA-AA-3538	AML
16	18849461	18849461	C	T	*SMG1*	p.E2430K	TCGA-CI-6620	AML
16	18866161	18866161	T	C	*SMG1*	p.K1434E	TCGA-13-0802	AML
16	18866161	18866161	T	C	*SMG1*	p.K1434E	TCGA-13-1407	AML
16	56947306	56947306	T	A	*SLC12A3*	p.Y1028N	TCGA-AG-A016	AML
16	67575671	67575671	G	T	*FAM65A*	p.E376*	TCGA-CN-5367	AML
16	67579728	67579728	C	T	*FAM65A*	p.R1138W	TCGA-24-1469	AML
16	70935052	70935052	C	T	*HYDIN*	p.W2968*	TCGA-DU-8161	AML
16	71101200	71101200	T	C	*HYDIN*	p.T690A	TCGA-C8-A1HF	AML
16	88653039	88653039	A	T	*ZC3H18*	p.D212V	TCGA-AA-3514	AML
16	88653062	88653062	C	T	*ZC3H18*	p.P220S	TCGA-B2-4101	AML
16	88677735	88677735	G	C	*ZC3H18*	p.E422D	TCGA-F5-6465	AML
16	88690371	88690371	G	A	*ZC3H18*	p.R600Q	TCGA-C8-A27A	AML
17	1562014	1562014	G	A	*PRPF8*	p.Q1728*	TCGA-25-1324	AML
17	1563288	1563288	T	G	*PRPF8*	p.D1598A	TCGA-06-2563	AML
17	1565026	1565026	C	T	*PRPF8*	p.E1361K	TCGA-CI-6620	AML
17	10346811	10346811	G	A	*MYH4*	p.R1901C	TCGA-S9-A6WO	AML
17	10351240	10351240	C	A	*MYH4*	p.K1620N	TCGA-14-3477	AML
17	10351424	10351424	T	C	*MYH4*	p.E1559G	TCGA-EY-A1GJ	AML
17	10354132	10354132	T	C	*MYH4*	p.I1316V	TCGA-BB-7870	AML
17	10354748	10354748	G	A	*MYH4*	p.R1254C	TCGA-29-1703	AML
17	10355280	10355280	A	G	*MYH4*	p.M1239T	TCGA-AF-2689	AML
17	10357189	10357189	A	G	*MYH4*	p.L902S	TCGA-A6-6652	AML
17	10358009	10358009	C	G	*MYH4*	p.E852Q	TCGA-55-8094	AML
17	10358324	10358324	G	A	*MYH4*	p.T790M	TCGA-23-1032	AML
17	10358358	10358358	C	T	*MYH4*	p.E779K	TCGA-MP-A5C7	AML
17	10358529	10358529	A	G	*MYH4*	p.I753T	TCGA-S9-A7QX	AML
17	10366661	10366661	C	-	*MYH4*	p.E267fs	TCGA-29-1703	AML
17	10366663	10366663	A	T	*MYH4*	p.I266N	TCGA-29-1703	AML
17	10370010	10370010	C	T	*MYH4*	p.R18Q	TCGA-CI-6620	AML
17	11573101	11573101	G	A	*DNAH9*	p.V1115I	TCGA-50-5946	AML
17	11583084	11583084	C	A	*DNAH9*	p.L1122M	TCGA-A6-2676	AML
17	11593492	11593492	T	A	*DNAH9*	p.Y1451*	TCGA-16-1055	AML
17	11597310	11597310	G	C	*DNAH9*	p.Q1580H	TCGA-CS-4942	AML
17	11603144	11603144	T	C	*DNAH9*	p.Y1657H	TCGA-13-0901	AML
17	11642317	11642317	C	T	*DNAH9*	p.R1979C	TCGA-ER-A3ET	AML
17	11650904	11650904	G	A	*DNAH9*	p.R2144Q	TCGA-HU-8243	AML
17	11687618	11687618	G	A	*DNAH9*	p.R2608H	TCGA-R8-A6YH	AML
17	11711062	11711062	C	T	*DNAH9*	p.R2812C	TCGA-EJ-7794	AML
17	11711083	11711083	G	T	*DNAH9*	p.G2819*	TCGA-AA-A00Q	AML
17	11784633	11784633	C	T	*DNAH9*	p.A3570V	TCGA-76-6283	AML
17	11786984	11786984	G	T	*DNAH9*	p.A3630S	TCGA-JV-A5VF	AML
17	11790227	11790227	A	G	*DNAH9*	p.Y3686C	TCGA-06-0238	AML
17	11790227	11790227	A	G	*DNAH9*	p.Y3686C	TCGA-DU-7015	AML
17	11835380	11835380	T	C	*DNAH9*	p.F4052S	TCGA-AK-3461	AML
17	11837316	11837316	G	A	*DNAH9*	p.M4139I	TCGA-FD-A5BZ	AML
17	11865389	11865389	T	C	*DNAH9*	p.F4350S	TCGA-AP-A3K1	AML

17	39680397	39680397	T	C	*KRT19*	p.M316V	TCGA-AO-A0JL	AML
17	48646625	48646625	T	G	*CACNA1G*	p.W152G	TCGA-13-0906	AML
17	48699055	48699055	C	A	*CACNA1G*	p.T1987K	TCGA-23-1031	AML
17	51901154	51901154	G	T	*KIF2B*	p.D254Y	TCGA-06-2569	AML
17	72786363	72786363	A	G	*TMEM104*	p.T105A	TCGA-A5-A1OG	AML
17	72832629	72832629	G	A	*TMEM104*	p.G432S	TCGA-AQ-A54N	AML
17	78310077	78310077	C	A	*RNF213*	p.L1476I	TCGA-AK-3428	AML
18	47365663	47365663	C	T	*MYO5B*	p.G138D	TCGA-GU-A42P	AML
18	47389641	47389641	C	A	*MYO5B*	p.Q1300H	TCGA-62-8394	AML
19	8962375	8962375	C	A	*MUC16*	p.E14442*	TCGA-A2-A25F	AML
19	10071461	10071461	T	A	*COL5A3*	p.N1653Y	TCGA-AA-A00R	AML
19	10108812	10108812	T	C	*COL5A3*	p.K375R	TCGA-25-1632	AML
19	13054627	13054628	-	TTGTC	*CALR*	p.K29fs	TCGA-AN-A0XR	AML
19	38935242	38935242	G	A	*RYR1*	p.G186R	TCGA-D7-A6F2	AML
19	38943518	38943518	T	G	*RYR1*	p.L435R	TCGA-AA-A00K	AML
19	38957009	38957009	A	G	*RYR1*	p.Y1050C	TCGA-AG-3609	AML
19	38997019	38997019	T	G	*RYR1*	e55+2	TCGA-A8-A08O	AML
19	39001379	39001379	G	C	*RYR1*	p.S3027T	TCGA-25-1634	AML
19	39008281	39008281	T	C	*RYR1*	p.I3323T	TCGA-AG-A016	AML
19	40367854	40367854	C	T	*FCGBP*	p.C4369Y	TCGA-B5-A11L	AML
19	40411637	40411637	C	T	*FCGBP*	e7+1	TCGA-24-1552	AML
19	42840990	42840990	T	C	*MEGF8*	p.F426L	TCGA-25-1317	AML
19	42854495	42854495	C	T	*MEGF8*	p.P899S	TCGA-25-1317	AML
20	31388652	31388652	G	A	*DNMT3B*	p.W639*	TCGA-16-1055	AML
20	40709527	40709527	A	C	*PTPRT*	p.S1462A	TCGA-A2-A04T	AML
20	40944473	40944473	G	T	*PTPRT*	p.Q677K	TCGA-FD-A3SN	AML
20	41100967	41100967	C	A	*PTPRT*	p.L463F	TCGA-29-1774	AML
20	41408885	41408885	C	T	*PTPRT*	p.E181K	TCGA-CI-6620	AML
21	41452078	41452078	C	T	*DSCAM*	e25+1	TCGA-19-1388	AML
21	41516603	41516603	C	T	*DSCAM*	p.R1025Q	TCGA-CI-6620	AML
21	41725552	41725552	C	T	*DSCAM*	p.W258*	TCGA-AX-A06J	AML
X	34148760	34148760	C	A	*FAM47A*	p.E546*	TCGA-AA-3864	AML
X	63445321	63445321	C	A	*MTMR8*	p.W445C	TCGA-13-1407	AML
X	63555969	63555969	A	G	*MTMR8*	p.F381L	TCGA-25-1321	AML
X	119388243	119388243	A	T	*ZBTB33*	p.K325*	TCGA-DJ-A1QH	AML
X	119388836	119388836	T	G	*ZBTB33*	p.Y522*	TCGA-M7-A721	AML
X	135482269	135482269	G	A	*GPR112*	p.G2857S	TCGA-D1-A0ZZ	AML
X	152811567	152811567	T	C	*ATP2B3*	p.M313T	TCGA-FW-A3I3	AML
X	152814194	152814194	C	T	*ATP2B3*	p.T407M	TCGA-DE-A4MC	AML
X	152825260	152825260	C	G	*ATP2B3*	p.A900G	TCGA-36-1577	AML
X	152830557	152830557	C	T	*ATP2B3*	p.T1113M	TCGA-A6-2684	AML
1	16203071	16203071	G	A	*SPEN*	p.S260N	TCGA-CD-5804	Cancer/AML
1	16256414	16256414	A	G	*SPEN*	p.K1227E	TCGA-13-0717	Cancer/AML
1	162731130	162731130	C	A	*DDR2*	p.P329T	TCGA-AG-A016	Cancer/AML
2	25457176	25457176	G	A	*DNMT3A*	p.P904L	TCGA-MH-A561	Cancer/AML
2	25457176	25457176	G	A	*DNMT3A*	p.P904L	TCGA-VP-A879	Cancer/AML
2	25457192	25457192	G	A	*DNMT3A*	p.R899C	TCGA-06-6694	Cancer/AML
2	25457231	25457231	G	A	*DNMT3A*	p.Q886*	TCGA-G3-A5SL	Cancer/AML
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-06-0142	Cancer/AML
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-D1-A179	Cancer/AML
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-D7-A4Z0	Cancer/AML
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-EO-A3KX	Cancer/AML
2	25457243	25457243	G	A	*DNMT3A*	p.R882C	TCGA-G9-6339	Cancer/AML
2	25457290	25457290	C	A	*DNMT3A*	e22-1	TCGA-AX-A060	Cancer/AML
2	25457290	25457290	C	A	*DNMT3A*	e22-1	TCGA-DA-A1I8	Cancer/AML
2	25457291	25457291	T	C	*DNMT3A*	e22-2	TCGA-EL-A3CR	Cancer/AML
2	25458648	25458648	T	C	*DNMT3A*	p.Q842R	TCGA-E7-A519	Cancer/AML
2	25458696	25458696	T	C	*DNMT3A*	e21-2	TCGA-12-3649	Cancer/AML

2	25462006	25462006	T	C	*DNMT3A*	p.M801V	TCGA-E5-A4TZ	Cancer/AML
2	25462006	25462006	T	G	*DNMT3A*	p.M801L	TCGA-78-7158	Cancer/AML
2	25462068	25462068	A	G	*DNMT3A*	p.I780T	TCGA-CM-6171	Cancer/AML
2	25462068	25462068	A	G	*DNMT3A*	p.I780T	TCGA-NH-A50T	Cancer/AML
2	25462085	25462085	C	T	*DNMT3A*	e19-1	TCGA-A7-A3RF	Cancer/AML
2	25463184	25463184	G	T	*DNMT3A*	p.S770*	TCGA-BR-A4QL	Cancer/AML
2	25463232	25463232	A	C	*DNMT3A*	p.L754R	TCGA-B2-3924	Cancer/AML
2	25463234	25463234	C	T	*DNMT3A*	p.W753*	TCGA-EM-A2OV	Cancer/AML
2	25463235	25463235	C	G	*DNMT3A*	p.W753S	TCGA-CG-4441	Cancer/AML
2	25463242	25463242	A	T	*DNMT3A*	p.F751I	TCGA-K4-A5RI	Cancer/AML
2	25463247	25463247	C	A	*DNMT3A*	p.R749L	TCGA-55-6712	Cancer/AML
2	25463265	25463265	G	C	*DNMT3A*	p.P743R	TCGA-QK-A6VB	Cancer/AML
2	25463266	25463266	G	A	*DNMT3A*	p.P743S	TCGA-DX-A3LT	Cancer/AML
2	25463280	25463280	A	T	*DNMT3A*	p.L738Q	TCGA-HW-A5KL	Cancer/AML
2	25463283	25463283	A	C	*DNMT3A*	p.L737R	TCGA-CV-A45X	Cancer/AML
2	25463287	25463287	G	A	*DNMT3A*	p.R736C	TCGA-C5-A1BE	Cancer/AML
2	25463289	25463289	T	C	*DNMT3A*	p.Y735C	TCGA-EJ-A65F	Cancer/AML
2	25463308	25463308	G	A	*DNMT3A*	p.R729W	TCGA-CN-A6V6	Cancer/AML
2	25463321	25463321	T	C	*DNMT3A*	e18-2	TCGA-85-A5B5	Cancer/AML
2	25463541	25463541	G	C	*DNMT3A*	p.S714C	TCGA-VN-A88Q	Cancer/AML
2	25464439	25464439	G	A	*DNMT3A*	p.Q692*	TCGA-BF-A3DN	Cancer/AML
2	25464460	25464460	C	T	*DNMT3A*	p.G685R	TCGA-B5-A0JX	Cancer/AML
2	25464460	25464460	C	T	*DNMT3A*	p.G685R	TCGA-G9-6367	Cancer/AML
2	25464490	25464490	C	T	*DNMT3A*	p.V675M	TCGA-20-0996	Cancer/AML
2	25464534	25464534	T	C	*DNMT3A*	p.Y660C	TCGA-85-7844	Cancer/AML
2	25464537	25464537	C	T	*DNMT3A*	p.R659H	TCGA-26-5139	Cancer/AML
2	25466796	25466796	A	C	*DNMT3A*	p.V636G	TCGA-E9-A1QZ	Cancer/AML
2	25466797	25466797	C	A	*DNMT3A*	p.V636L	TCGA-EK-A2PG	Cancer/AML
2	25467125	25467125	A	-	*DNMT3A*	p.Y584fs	TCGA-06-2558	Cancer/AML
2	25467145	25467145	T	-	*DNMT3A*	p.K577fs	TCGA-CR-7370	Cancer/AML
2	25467199	25467199	C	T	*DNMT3A*	p.C559Y	TCGA-DM-A28G	Cancer/AML
2	25467408	25467408	C	A	*DNMT3A*	e13+1	TCGA-55-7815	Cancer/AML
2	25467408	25467408	C	T	*DNMT3A*	e13+1	TCGA-S9-A6WP	Cancer/AML
2	25467475	25467475	T	G	*DNMT3A*	p.Q534P	TCGA-X6-A7WD	Cancer/AML
2	25467519	25467519	G	-	*DNMT3A*	p.C520fs	TCGA-CV-7248	Cancer/AML
2	25468129	25468129	T	-	*DNMT3A*	p.N516fs	TCGA-46-6025	Cancer/AML
2	25468202	25468202	C	G	*DNMT3A*	e12-1	TCGA-G9-6356	Cancer/AML
2	25469053	25469053	C	A	*DNMT3A*	p.E469*	TCGA-06-2563	Cancer/AML
2	25469083	25469083	T	A	*DNMT3A*	p.K459*	TCGA-HW-A5KJ	Cancer/AML
2	25469633	25469633	G	A	*DNMT3A*	p.R379C	TCGA-23-2643	Cancer/AML
2	25469646	25469646	C	A	*DNMT3A*	e9-1	TCGA-85-A4JB	Cancer/AML
2	25469924	25469924	A	C	*DNMT3A*	p.L373R	TCGA-53-A4EZ	Cancer/AML
2	25470464	25470464	G	A	*DNMT3A*	p.S337L	TCGA-06-0126	Cancer/AML
2	25470464	25470464	G	C	*DNMT3A*	p.S337*	TCGA-AC-A3W6	Cancer/AML
2	25470480	25470480	C	T	*DNMT3A*	p.G332R	TCGA-62-A46U	Cancer/AML
2	25470480	25470480	C	T	*DNMT3A*	p.G332R	TCGA-A5-A0GB	Cancer/AML
2	25470498	25470498	G	A	*DNMT3A*	p.R326C	TCGA-B6-A0IH	Cancer/AML
2	25470516	25470516	G	A	*DNMT3A*	p.R320*	TCGA-E8-A3X7	Cancer/AML
2	25470516	25470516	G	A	*DNMT3A*	p.R320*	TCGA-EK-A2PI	Cancer/AML
2	25470533	25470533	C	T	*DNMT3A*	p.W314*	TCGA-D1-A16G	Cancer/AML
2	25470588	25470588	C	T	*DNMT3A*	p.V296M	TCGA-19-1789	Cancer/AML
2	25470588	25470588	C	T	*DNMT3A*	p.V296M	TCGA-C8-A278	Cancer/AML
2	25470920	25470920	C	A	*DNMT3A*	p.E281*	TCGA-BJ-A0YZ	Cancer/AML
2	25470993	25470996	GGGG	-	*DNMT3A*	p.P256fs	TCGA-D3-A5GO	Cancer/AML
2	141200153	141200153	G	T	*LRP1B*	p.T3445N	TCGA-BG-A0YV	Cancer/AML
2	141259439	141259439	C	G	*LRP1B*	p.Q2889H	TCGA-AF-2691	Cancer/AML
2	141356297	141356297	T	C	*LRP1B*	p.N2366S	TCGA-19-1388	Cancer/AML
2	141528569	141528569	T	A	*LRP1B*	p.N1836I	TCGA-55-8620	Cancer/AML

2	141641524	141641524	C	G	*LRP1B*	p.W1344S	TCGA-AA-A00W	Cancer/AML
2	141643883	141643883	A	T	*LRP1B*	p.I1263N	TCGA-C8-A278	Cancer/AML
2	141665548	141665548	T	C	*LRP1B*	p.K1140E	TCGA-09-1667	Cancer/AML
2	141680678	141680678	G	T	*LRP1B*	p.Q1059K	TCGA-AA-3514	Cancer/AML
2	141680681	141680681	A	T	*LRP1B*	p.F1058I	TCGA-13-0801	Cancer/AML
2	141680681	141680681	A	T	*LRP1B*	p.F1058I	TCGA-AA-3529	Cancer/AML
2	141773398	141773398	C	A	*LRP1B*	p.R686L	TCGA-DU-6393	Cancer/AML
2	141819834	141819834	A	T	*LRP1B*	p.F341Y	TCGA-13-0901	Cancer/AML
2	198266176	198266176	A	C	*SF3B1*	p.F815C	TCGA-AA-3518	Cancer/AML
2	198266176	198266176	A	C	*SF3B1*	p.F815C	TCGA-AA-3534	Cancer/AML
2	198266834	198266834	T	C	*SF3B1*	p.K700E	TCGA-KU-A66S	Cancer/AML
2	198267359	198267359	C	A	*SF3B1*	p.K666N	TCGA-DM-A28C	Cancer/AML
2	198267359	198267359	C	G	*SF3B1*	p.K666N	TCGA-A6-2681	Cancer/AML
2	198267360	198267360	T	C	*SF3B1*	p.K666R	TCGA-AP-A0LQ	Cancer/AML
2	198267371	198267371	G	C	*SF3B1*	p.H662Q	TCGA-F7-A622	Cancer/AML
2	198267371	198267371	G	T	*SF3B1*	p.H662Q	TCGA-24-2030	Cancer/AML
2	198267480	198267480	T	G	*SF3B1*	p.N626T	TCGA-A6-6653	Cancer/AML
2	198273167	198273167	C	T	*SF3B1*	p.G348E	TCGA-CI-6620	Cancer/AML
2	198273233	198273233	G	A	*SF3B1*	p.T326I	TCGA-AX-A05S	Cancer/AML
3	128200669	128200669	A	T	*GATA2*	p.L379Q	TCGA-FD-A3SM	Cancer/AML
3	128200771	128200771	G	T	*GATA2*	p.A345D	TCGA-X6-A7WB	Cancer/AML
3	128204642	128204642	G	A	*GATA2*	p.P267S	TCGA-AK-3461	Cancer/AML
4	55589839	55589839	T	G	*KIT*	p.Y441D	TCGA-AX-A05T	Cancer/AML
4	55602715	55602715	T	C	*KIT*	p.Y846H	TCGA-AA-3517	Cancer/AML
4	106155310	106155310	A	-	*TET2*	p.N92fs	TCGA-85-8664	Cancer/AML
4	106156242	106156242	C	-	*TET2*	p.F402fs	TCGA-06-2559	Cancer/AML
4	106156654	106156655	-	A	*TET2*	p.F540fs	TCGA-UF-A71B	Cancer/AML
4	106156690	106156690	C	T	*TET2*	p.Q552*	TCGA-AK-3433	Cancer/AML
4	106156747	106156747	C	T	*TET2*	p.R571*	TCGA-97-7938	Cancer/AML
4	106156747	106156747	C	T	*TET2*	p.R571*	TCGA-AM-5820	Cancer/AML
4	106157029	106157029	C	T	*TET2*	p.Q665*	TCGA-BG-A0W1	Cancer/AML
4	106157053	106157053	C	T	*TET2*	p.Q673*	TCGA-DY-A1DC	Cancer/AML
4	106157384	106157385	-	C	*TET2*	p.Q785fs	TCGA-06-2558	Cancer/AML
4	106157588	106157589	-	A	*TET2*	p.Q852fs	TCGA-69-7764	Cancer/AML
4	106157764	106157764	A	T	*TET2*	p.K910*	TCGA-36-2543	Cancer/AML
4	106157824	106157824	C	T	*TET2*	p.Q930*	TCGA-F7-A624	Cancer/AML
4	106157971	106157971	C	T	*TET2*	p.Q979*	TCGA-95-A4VN	Cancer/AML
4	106158431	106158431	T	A	*TET2*	p.L1132*	TCGA-KU-A66S	Cancer/AML
4	106158510	106158510	T	G	*TET2*	e2+2	TCGA-MP-A4TC	Cancer/AML
4	106164793	106164793	T	C	*TET2*	p.C1242R	TCGA-BH-A0WA	Cancer/AML
4	106164917	106164917	G	C	*TET2*	p.R1283P	TCGA-CI-6619	Cancer/AML
4	106164929	106164929	A	G	*TET2*	p.N1287S	TCGA-D8-A1XS	Cancer/AML
4	106193995	106193995	C	G	*TET2*	p.S1507*	TCGA-CA-6715	Cancer/AML
7	55211032	55211032	C	T	*EGFR*	p.A92V	TCGA-NC-A5HD	Cancer/AML
7	55221845	55221845	C	T	*EGFR*	p.R252C	TCGA-16-0861	Cancer/AML
7	55233043	55233043	G	T	*EGFR*	p.G553V	TCGA-DU-7290	Cancer/AML
7	55259466	55259466	A	C	*EGFR*	p.N797H	TCGA-AG-A016	Cancer/AML
7	55273147	55273147	G	T	*EGFR*	p.W1157L	TCGA-AG-A016	Cancer/AML
7	86415889	86415889	C	T	*GRM3*	p.R261*	TCGA-D9-A6EA	Cancer/AML
7	86493647	86493647	C	T	*GRM3*	p.R517*	TCGA-CI-6620	Cancer/AML
7	148504789	148504789	C	A	*EZH2*	p.Q735H	TCGA-A6-2681	Cancer/AML
7	148507431	148507431	T	G	*EZH2*	p.N675H	TCGA-AG-3598	Cancer/AML
7	148512038	148512038	A	C	*EZH2*	p.F547C	TCGA-AG-3578	Cancer/AML
7	148524323	148524323	C	A	*EZH2*	p.D221Y	TCGA-AA-3520	Cancer/AML
7	148526832	148526832	G	T	*EZH2*	p.H158N	TCGA-AA-3514	Cancer/AML
8	12957122	12957122	G	C	*DLC1*	p.D505E	TCGA-AA-3534	Cancer/AML
8	13259044	13259044	C	T	*DLC1*	p.E370K	TCGA-CI-6620	Cancer/AML
8	93017486	93017486	C	A	*RUNX1T1*	p.A163S	TCGA-24-1550	Cancer/AML

8	117864273	117864273	T	C	*RAD21*	p.N7D	TCGA-24-0968	Cancer/AML
8	117869536	117869536	A	G	*RAD21*	p.F220L	TCGA-13-0717	Cancer/AML
8	117878848	117878848	C	A	*RAD21*	p.V41L	TCGA-AU-6004	Cancer/AML
8	128750924	128750924	C	T	*MYC*	p.S139L	TCGA-CI-6620	Cancer/AML
8	144940317	144940317	C	G	*EPPK1*	p.D2369H	TCGA-A1-A0SQ	Cancer/AML
8	144940347	144940347	C	T	*EPPK1*	p.V2359M	TCGA-AR-A0TW	Cancer/AML
8	144940488	144940488	C	T	*EPPK1*	p.D2312N	TCGA-C8-A1HJ	Cancer/AML
8	144940506	144940506	C	T	*EPPK1*	p.A2306T	TCGA-QG-A5Z1	Cancer/AML
8	144940508	144940508	C	T	*EPPK1*	p.R2305H	TCGA-EW-A3E8	Cancer/AML
8	144940509	144940509	G	A	*EPPK1*	p.R2305C	TCGA-E5-A4TZ	Cancer/AML
8	144940517	144940517	G	A	*EPPK1*	p.S2302L	TCGA-J1-A4AH	Cancer/AML
8	144940602	144940602	C	T	*EPPK1*	p.D2274N	TCGA-FD-A5BU	Cancer/AML
8	144942354	144942354	T	C	*EPPK1*	p.I1690V	TCGA-G9-6369	Cancer/AML
8	144945092	144945092	G	A	*EPPK1*	p.T777M	TCGA-K4-A6MB	Cancer/AML
9	5050744	5050744	A	C	*JAK2*	p.E176A	TCGA-24-0968	Cancer/AML
9	5066760	5066760	A	C	*JAK2*	p.N433H	TCGA-13-1408	Cancer/AML
9	98224152	98224152	T	C	*PTCH1*	p.I897V	TCGA-EO-A3KW	Cancer/AML
9	98231259	98231259	T	C	*PTCH1*	p.Y675C	TCGA-AA-A00K	Cancer/AML
9	98232132	98232132	C	T	*PTCH1*	p.E604K	TCGA-CI-6620	Cancer/AML
9	98240434	98240434	T	G	*PTCH1*	p.Q417P	TCGA-AK-3460	Cancer/AML
9	98242767	98242767	T	G	*PTCH1*	p.K284Q	TCGA-AG-A016	Cancer/AML
10	28900784	28900784	C	T	*WAC*	p.T457I	TCGA-23-1021	Cancer/AML
10	112342324	112342324	C	T	*SMC3*	p.S243F	TCGA-CI-6620	Cancer/AML
10	112349676	112349676	A	C	*SMC3*	p.E479A	TCGA-AA-3514	Cancer/AML
11	119148966	119148966	T	C	*CBL*	p.C396R	TCGA-23-1029	Cancer/AML
11	119149334	119149334	G	A	*CBL*	p.E448K	TCGA-B2-4101	Cancer/AML
11	119155953	119155953	C	T	*CBL*	p.R540*	TCGA-CR-7376	Cancer/AML
12	6682257	6682257	T	G	*CHD4*	p.N1875T	TCGA-AK-3460	Cancer/AML
12	6690263	6690263	T	C	*CHD4*	p.E1647G	TCGA-AA-3514	Cancer/AML
12	6701173	6701173	C	T	*CHD4*	p.R1000Q	TCGA-CI-6620	Cancer/AML
12	6702646	6702646	C	T	*CHD4*	p.R817Q	TCGA-J4-A67S	Cancer/AML
12	6704562	6704562	T	G	*CHD4*	p.K687Q	TCGA-AG-A016	Cancer/AML
12	6705266	6705266	A	C	*CHD4*	p.W644G	TCGA-09-1666	Cancer/AML
12	6711550	6711550	T	A	*CHD4*	p.K72*	TCGA-14-0783	Cancer/AML
12	31244669	31244669	A	G	*DDX11*	p.Y369C	TCGA-MU-A5YI	Cancer/AML
12	31247713	31247713	A	G	*DDX11*	p.D480G	TCGA-CV-7252	Cancer/AML
12	112891058	112891058	A	G	*PTPN11*	p.K131R	TCGA-09-0367	Cancer/AML
12	112910785	112910785	G	A	*PTPN11*	p.R265Q	TCGA-13-0803	Cancer/AML
12	112926851	112926851	C	T	*PTPN11*	p.P491S	TCGA-A4-7996	Cancer/AML
12	112926926	112926926	A	G	*PTPN11*	p.M516V	TCGA-G3-A25S	Cancer/AML
13	28592699	28592699	T	A	*FLT3*	p.N816Y	TCGA-AA-3538	Cancer/AML
13	28623571	28623571	G	T	*FLT3*	p.T329N	TCGA-ER-A19E	Cancer/AML
13	28636128	28636128	C	A	*FLT3*	p.E82*	TCGA-24-1560	Cancer/AML
13	28895703	28895703	G	T	*FLT1*	p.A1024E	TCGA-AX-A3FX	Cancer/AML
13	73335542	73335542	C	T	*DIS3*	p.E877K	TCGA-AA-3514	Cancer/AML
13	73351629	73351629	C	A	*DIS3*	p.E195*	TCGA-AX-A3FW	Cancer/AML
15	88476304	88476304	C	A	*NTRK3*	p.G610W	TCGA-DA-A1I8	Cancer/AML
15	88669507	88669507	A	T	*NTRK3*	p.M464K	TCGA-13-0802	Cancer/AML
15	88799240	88799240	C	G	*NTRK3*	p.D49H	TCGA-AG-A016	Cancer/AML
17	7577538	7577538	C	T	*TP53*	p.R248Q	TCGA-AX-A2H7	Cancer/AML
17	7577538	7577538	C	T	*TP53*	p.R248Q	TCGA-DK-A1AD	Cancer/AML
17	7577539	7577539	G	A	*TP53*	p.R248W	TCGA-F5-6702	Cancer/AML
17	7577548	7577548	C	T	*TP53*	p.G245S	TCGA-QK-A6VB	Cancer/AML
17	7578500	7578500	G	A	*TP53*	p.Q144*	TCGA-D7-A6F2	Cancer/AML
17	7578524	7578524	G	A	*TP53*	p.Q136*	TCGA-95-8494	Cancer/AML
17	29496994	29496994	A	C	*NF1*	p.K189Q	TCGA-AA-3558	Cancer/AML
17	29550515	29550515	G	A	*NF1*	p.S592N	TCGA-28-5216	Cancer/AML
17	29560209	29560209	A	G	*NF1*	p.N1229S	TCGA-44-A47F	Cancer/AML

17	29562948	29562948	A	G	*NF1*	p.T1295A	TCGA-14-0871	Cancer/AML
17	29665822	29665822	A	G	*NF1*	p.K2307R	TCGA-C8-A12X	Cancer/AML
17	30315382	30315382	A	C	*SUZ12*	p.Q356P	TCGA-DU-7292	Cancer/AML
17	30325762	30325762	C	T	*SUZ12*	p.R654*	TCGA-68-8251	Cancer/AML
17	74732454	74732454	C	T	*SRSF2*	p.R152H	TCGA-24-2288	Cancer/AML
18	42281637	42281637	C	T	*SETBP1*	p.T109I	TCGA-CQ-A4CE	Cancer/AML
18	42531113	42531113	A	C	*SETBP1*	p.E603A	TCGA-26-5135	Cancer/AML
18	42531913	42531913	G	A	*SETBP1*	p.G870S	TCGA-29-1761	Cancer/AML
19	17943490	17943490	G	A	*JAK3*	p.R840C	TCGA-06-1805	Cancer/AML
20	31021086	31021086	G	T	*ASXL1*	e12-1	TCGA-EB-A5KH	Cancer/AML
20	31021250	31021250	C	T	*ASXL1*	p.R417*	TCGA-CV-7437	Cancer/AML
20	31021295	31021295	C	T	*ASXL1*	p.Q432*	TCGA-77-A5GH	Cancer/AML
20	31021535	31021535	C	T	*ASXL1*	p.Q512*	TCGA-EE-A2GC	Cancer/AML
20	31021535	31021535	C	T	*ASXL1*	p.Q512*	TCGA-NC-A5HD	Cancer/AML
20	31021642	31021643	-	T	*ASXL1*	p.R548fs	TCGA-05-4403	Cancer/AML
20	31022238	31022238	C	T	*ASXL1*	p.Q575*	TCGA-86-8281	Cancer/AML
20	31022263	31022263	G	A	*ASXL1*	p.W583*	TCGA-A6-3808	Cancer/AML
20	31022285	31022286	TT	-	*ASXL1*	p.Y591fs	TCGA-37-4130	Cancer/AML
20	31022288	31022288	C	A	*ASXL1*	p.Y591*	TCGA-BR-8373	Cancer/AML
20	31022288	31022288	C	A	*ASXL1*	p.Y591*	TCGA-EE-A2MC	Cancer/AML
20	31022288	31022288	C	G	*ASXL1*	p.Y591*	TCGA-EC-A1QX	Cancer/AML
20	31022670	31022670	G	-	*ASXL1*	p.E719fs	TCGA-MN-A4N1	Cancer/AML
20	31022712	31022712	C	T	*ASXL1*	p.Q733*	TCGA-97-8179	Cancer/AML
20	31022784	31022784	C	T	*ASXL1*	p.Q757*	TCGA-77-8144	Cancer/AML
20	31022784	31022784	C	T	*ASXL1*	p.Q757*	TCGA-CF-A3MH	Cancer/AML
20	31022817	31022817	C	T	*ASXL1*	p.Q768*	TCGA-EM-A3O3	Cancer/AML
20	31023159	31023159	C	T	*ASXL1*	p.Q882*	TCGA-85-8584	Cancer/AML
20	31023209	31023209	G	A	*ASXL1*	p.W898*	TCGA-NK-A5CX	Cancer/AML
20	31023249	31023249	A	T	*ASXL1*	p.K912*	TCGA-ER-A194	Cancer/AML
20	31024563	31024563	C	T	*ASXL1*	p.Q1350*	TCGA-EK-A2RA	Cancer/AML
21	44514777	44514777	T	G	*U2AF1*	p.Q157P	TCGA-DD-A1EE	Cancer/AML
21	44514777	44514777	T	G	*U2AF1*	p.Q157P	TCGA-EE-A29D	Cancer/AML
22	22314813	22314813	C	T	*TOP3B*	p.A512T	TCGA-AK-3434	Cancer/AML
22	22318590	22318590	G	T	*TOP3B*	p.N347K	TCGA-BR-A4CR	Cancer/AML
X	39933976	39933976	T	G	*BCOR*	p.D208A	TCGA-24-1548	Cancer/AML
X	40996183	40996183	A	G	*USP9X*	p.N188D	TCGA-EU-5907	Cancer/AML
X	41010259	41010259	A	C	*USP9X*	p.K571T	TCGA-13-0906	Cancer/AML
X	41088588	41088588	A	T	*USP9X*	p.K2382*	TCGA-B2-4101	Cancer/AML
X	44879860	44879860	T	A	*KDM6A*	p.I150N	TCGA-24-1545	Cancer/AML
X	44922836	44922836	C	A	*KDM6A*	p.P573H	TCGA-AA-A01G	Cancer/AML
X	44923034	44923034	G	C	*KDM6A*	p.W639S	TCGA-AA-A00K	Cancer/AML
X	44945197	44945197	G	T	*KDM6A*	p.W1181L	TCGA-50-5946	Cancer/AML
X	53407567	53407567	C	T	*SMC1A*	p.E1198K	TCGA-CI-6620	Cancer/AML
X	53436117	53436117	T	G	*SMC1A*	p.N474T	TCGA-AA-3534	Cancer/AML
X	53436124	53436124	C	T	*SMC1A*	p.E472K	TCGA-CI-6620	Cancer/AML
X	53438775	53438775	T	C	*SMC1A*	p.N397S	TCGA-16-0861	Cancer/AML
X	53449458	53449458	A	G	*SMC1A*	p.I31T	TCGA-63-7022	Cancer/AML
X	70346862	70346862	A	G	*MED12*	p.K910R	TCGA-41-3393	Cancer/AML
X	70346871	70346871	A	G	*MED12*	p.D913G	TCGA-23-1024	Cancer/AML
X	70354234	70354234	C	T	*MED12*	p.R1549C	TCGA-23-1109	Cancer/AML
X	123164976	123164976	G	A	*STAG2*	e3+1	TCGA-CC-A5UD	Cancer/AML
X	133547968	133547968	A	T	*PHF6*	p.K235M	TCGA-AA-3520	Cancer/AML
X	133547985	133547985	T	C	*PHF6*	p.Y241H	TCGA-04-1530	Cancer/AML
1	1747227	1747227	C	A	*GNB1*	p.K57N	TCGA-DA-A1I0	Cancer
1	1747227	1747227	C	A	*GNB1*	p.K57N	TCGA-J4-A67O	Cancer
1	1747229	1747229	T	C	*GNB1*	p.K57E	TCGA-19-4068	Cancer
1	1747229	1747229	T	C	*GNB1*	p.K57E	TCGA-A1-A0SE	Cancer
1	1747229	1747229	T	C	*GNB1*	p.K57E	TCGA-A8-A08A	Cancer

1	1747229	1747229	T	C	*GNB1*	p.K57E	TCGA-AK-3444	Cancer
1	1747229	1747229	T	C	*GNB1*	p.K57E	TCGA-EJ-5497	Cancer
1	11259403	11259403	T	C	*MTOR*	p.K1389E	TCGA-AA-3514	Cancer
1	11272439	11272439	T	C	*MTOR*	p.D1164G	TCGA-06-0219	Cancer
1	11288784	11288784	G	C	*MTOR*	p.P991A	TCGA-14-1825	Cancer
1	11308124	11308124	T	A	*MTOR*	p.T290S	TCGA-EB-A41B	Cancer
1	11851367	11851367	T	C	*MTHFR*	p.N591S	TCGA-13-0801	Cancer
1	11861365	11861365	T	A	*MTHFR*	p.M151L	TCGA-09-1668	Cancer
1	16459810	16459810	C	A	*EPHA2*	p.E640*	TCGA-EE-A2ML	Cancer
1	16474936	16474936	G	T	*EPHA2*	p.L254M	TCGA-X6-A7WB	Cancer
1	17359566	17359566	G	C	*SDHB*	p.S92*	TCGA-13-0714	Cancer
1	23222016	23222016	C	T	*EPHB2*	p.S548L	TCGA-CI-6620	Cancer
1	23232579	23232579	A	C	*EPHB2*	p.K622T	TCGA-13-0717	Cancer
1	23232579	23232579	A	G	*EPHB2*	p.K622R	TCGA-12-1095	Cancer
1	23232580	23232580	A	C	*EPHB2*	p.K622N	TCGA-13-1407	Cancer
1	23232580	23232580	A	C	*EPHB2*	p.K622N	TCGA-AA-3561	Cancer
1	23232580	23232580	A	C	*EPHB2*	p.K622N	TCGA-AA-A00Q	Cancer
1	23233400	23233400	A	T	*EPHB2*	p.M696L	TCGA-25-1625	Cancer
1	23239064	23239064	T	A	*EPHB2*	p.F942I	TCGA-A6-2676	Cancer
1	23239065	23239065	T	G	*EPHB2*	p.F942C	TCGA-AF-2691	Cancer
1	23239996	23239996	A	C	*EPHB2*	p.K965T	TCGA-AF-2691	Cancer
1	25254220	25254220	A	C	*RUNX3*	p.V109G	TCGA-A8-A08O	Cancer
1	27094387	27094387	A	G	*ARID1A*	p.E1032G	TCGA-AA-A00A	Cancer
1	27106589	27106589	T	A	*ARID1A*	p.I2067N	TCGA-36-1576	Cancer
1	28240894	28240895	-	A	*RPA2*	p.S20fs	TCGA-B5-A3S1	Cancer
1	46725690	46725690	T	C	*RAD54L*	p.L109P	TCGA-A6-2678	Cancer
1	46725690	46725690	T	C	*RAD54L*	p.L109P	TCGA-AK-3460	Cancer
1	46726639	46726639	C	T	*RAD54L*	p.P240S	TCGA-25-1324	Cancer
1	78422299	78422299	C	T	*FUBP1*	p.A576T	TCGA-AG-A016	Cancer
1	78429830	78429830	C	T	*FUBP1*	p.E341K	TCGA-CI-6620	Cancer
1	93301904	93301904	G	T	*RPL5*	p.G161V	TCGA-FC-7961	Cancer
1	97700536	97700536	G	T	*DPYD*	p.P772T	TCGA-AX-A06J	Cancer
1	120510805	120510805	C	A	*NOTCH2*	p.G387W	TCGA-25-1634	Cancer
1	150923098	150923098	G	C	*SETDB1*	p.C582S	TCGA-13-0793	Cancer
1	150923310	150923310	C	G	*SETDB1*	p.L653V	TCGA-GU-A762	Cancer
1	155210420	155210420	C	T	*GBA*	e2+1	TCGA-EB-A44R	Cancer
1	155874261	155874261	C	G	*RIT1*	p.M107I	TCGA-CN-5367	Cancer
1	156841502	156841502	A	C	*NTRK1*	p.N269H	TCGA-AG-A016	Cancer
1	156845995	156845995	C	T	*NTRK1*	p.A542V	TCGA-AA-3514	Cancer
1	158152713	158152713	G	A	*CD1D*	p.G218D	TCGA-A1-A0SO	Cancer
1	160389236	160389236	C	T	*VANGL2*	p.Q213*	TCGA-24-1105	Cancer
1	161141691	161141691	C	T	*B4GALT3*	p.R366H	TCGA-H4-A2HQ	Cancer
1	168034877	168034877	A	T	*DCAF6*	p.D830V	TCGA-12-1098	Cancer
1	173769598	173769598	A	T	*CENPL*	p.F387L	TCGA-AA-3534	Cancer
1	179089370	179089370	C	T	*ABL2*	p.A334T	TCGA-DK-A6B0	Cancer
1	198711080	198711080	A	T	*PTPRC*	p.H829L	TCGA-16-1055	Cancer
1	198711080	198711080	A	T	*PTPRC*	p.H829L	TCGA-AG-A016	Cancer
1	198719708	198719708	A	C	*PTPRC*	p.K1054Q	TCGA-13-0717	Cancer
1	204515995	204515995	T	C	*MDM4*	p.V298A	TCGA-AA-3514	Cancer
1	206650051	206650051	C	T	*IKBKE*	p.R191*	TCGA-CI-6620	Cancer
1	241663886	241663886	T	C	*FH*	p.K414R	TCGA-19-1392	Cancer
1	241676920	241676920	T	A	*FH*	p.M121L	TCGA-24-1553	Cancer
1	241676968	241676968	C	T	*FH*	p.E105K	TCGA-23-1024	Cancer
1	242035571	242035571	C	T	*EXO1*	p.T502I	TCGA-A6-2686	Cancer
1	243778439	243778439	C	T	*AKT3*	p.E196K	TCGA-CI-6620	Cancer
1	243801011	243801011	C	A	*AKT3*	p.G155C	TCGA-EE-A3AB	Cancer
2	25973025	25973025	G	T	*ASXL2*	p.S467*	TCGA-CN-4729	Cancer
2	25973284	25973284	T	C	*ASXL2*	e12-2	TCGA-DU-6400	Cancer

2	29436851	29436851	G	A	*ALK*	p.R1248*	TCGA-06-5858	Cancer
2	29450539	29450539	C	A	*ALK*	e17-1	TCGA-24-1548	Cancer
2	29551287	29551287	A	T	*ALK*	p.V448D	TCGA-DX-A6YV	Cancer
2	39281948	39281948	A	T	*SOS1*	p.F176Y	TCGA-AA-3514	Cancer
2	39514099	39514099	C	T	*MAP4K3*	p.G410E	TCGA-AA-3520	Cancer
2	42522542	42522542	G	C	*EML4*	p.V459L	TCGA-23-1124	Cancer
2	47698199	47698199	C	A	*MSH2*	p.S586*	TCGA-B2-4101	Cancer
2	47702411	47702411	T	C	*MSH2*	e12+2	TCGA-A6-2676	Cancer
2	48027233	48027233	C	G	*MSH6*	p.A704G	TCGA-DM-A28K	Cancer
2	48030696	48030696	T	C	*MSH6*	p.F1104L	TCGA-13-0802	Cancer
2	58386920	58386920	T	C	*FANCL*	p.M375V	TCGA-AG-3608	Cancer
2	70314915	70314915	C	T	*PCBP1*	p.L14F	TCGA-CI-6620	Cancer
2	96920586	96920586	C	T	*TMEM127*	p.A48T	TCGA-FY-A3NP	Cancer
2	96920631	96920631	C	T	*TMEM127*	p.G33R	TCGA-AA-3553	Cancer
2	111881618	111881618	T	G	*BCL2L11*	p.F99C	TCGA-AK-3460	Cancer
2	128028917	128028917	T	C	*ERCC3*	p.K647R	TCGA-13-0802	Cancer
2	128028917	128028917	T	C	*ERCC3*	p.K647R	TCGA-19-0960	Cancer
2	128028917	128028917	T	C	*ERCC3*	p.K647R	TCGA-AG-3581	Cancer
2	135975075	135975075	C	T	*ZRANB3*	p.E819K	TCGA-CI-6620	Cancer
2	135985387	135985387	C	T	*ZRANB3*	e13+1	TCGA-25-1317	Cancer
2	148657107	148657107	C	G	*ACVR2A*	p.S115C	TCGA-AG-3584	Cancer
2	169780307	169780307	G	T	*ABCB11*	p.A1264D	TCGA-DX-A6YS	Cancer
2	169788945	169788945	A	C	*ABCB11*	p.F1052C	TCGA-13-0717	Cancer
2	170026301	170026301	A	C	*LRP2*	p.F3803C	TCGA-36-1578	Cancer
2	170026302	170026302	A	G	*LRP2*	p.F3803L	TCGA-AA-3534	Cancer
2	170063244	170063244	A	C	*LRP2*	p.F2329C	TCGA-16-1055	Cancer
2	170063244	170063244	A	C	*LRP2*	p.F2329C	TCGA-23-1028	Cancer
2	170068628	170068628	C	T	*LRP2*	p.A2044T	TCGA-23-1027	Cancer
2	170093754	170093754	T	C	*LRP2*	p.D1517G	TCGA-AA-3514	Cancer
2	170093754	170093754	T	C	*LRP2*	p.D1517G	TCGA-AF-2689	Cancer
2	170094807	170094807	C	T	*LRP2*	p.E1434K	TCGA-CI-6620	Cancer
2	170135894	170135894	T	C	*LRP2*	p.D518G	TCGA-04-1364	Cancer
2	187466767	187466767	G	A	*ITGAV*	p.G69R	TCGA-AG-3583	Cancer
2	187466776	187466776	A	G	*ITGAV*	p.K72E	TCGA-25-1633	Cancer
2	187466779	187466779	G	C	*ITGAV*	p.A73P	TCGA-04-1364	Cancer
2	187529366	187529366	G	A	*ITGAV*	p.E691K	TCGA-25-1626	Cancer
2	187534504	187534504	G	A	*ITGAV*	p.R890Q	TCGA-CH-5753	Cancer
2	202149806	202149806	T	A	*CASP8*	p.I416N	TCGA-12-1095	Cancer
2	206610555	206610555	A	T	*NRP2*	p.E576V	TCGA-AA-A00R	Cancer
2	206617580	206617580	C	T	*NRP2*	p.S642L	TCGA-CI-6620	Cancer
2	206641006	206641006	A	T	*NRP2*	p.D826V	TCGA-AG-3598	Cancer
2	212295728	212295728	A	G	*ERBB4*	p.F862S	TCGA-13-0801	Cancer
2	212295786	212295786	C	T	*ERBB4*	p.D843N	TCGA-FD-A5BV	Cancer
2	212522537	212522537	T	C	*ERBB4*	p.S630G	TCGA-AG-3608	Cancer
2	212522549	212522549	T	C	*ERBB4*	p.N626D	TCGA-AG-A016	Cancer
2	212566779	212566779	A	G	*ERBB4*	p.Y468H	TCGA-AG-A016	Cancer
2	212652803	212652803	C	T	*ERBB4*	p.R168Q	TCGA-CN-6992	Cancer
2	233001226	233001226	C	T	*DIS3L2*	p.A269V	TCGA-FS-A1ZM	Cancer
2	233164812	233164812	G	T	*DIS3L2*	p.E574D	TCGA-AR-A24H	Cancer
2	234669075	234669075	C	T	*UGT1A1*	p.Q48*	TCGA-A6-2678	Cancer
2	240111555	240111555	C	T	*HDAC4*	p.E105K	TCGA-CI-6620	Cancer
3	3192694	3192694	G	T	*CRBN*	p.A395E	TCGA-A4-7584	Cancer
3	10106076	10106076	G	T	*FANCD2*	p.D662Y	TCGA-AN-A0FX	Cancer
3	10115047	10115047	G	A	*FANCD2*	e27+1	TCGA-42-2591	Cancer
3	12641731	12641731	C	T	*RAF1*	p.G324S	TCGA-A6-2683	Cancer
3	12660036	12660036	A	G	*RAF1*	p.L62S	TCGA-23-1026	Cancer
3	12660133	12660133	G	A	*RAF1*	p.P30S	TCGA-50-8459	Cancer
3	30732960	30732960	C	A	*TGFBR2*	p.P550T	TCGA-09-1662	Cancer

3	37061893	37061893	T	C	*MLH1*	p.V85A	TCGA-62-A46O	Cancer
3	37092125	37092125	A	G	*MLH1*	p.K510R	TCGA-64-1678	Cancer
3	38182641	38182641	T	C	*MYD88*	p.L273P	TCGA-DY-A1H8	Cancer
3	38182761	38182761	C	T	*MYD88*	p.A313V	TCGA-36-1568	Cancer
3	39104605	39104605	T	A	*WDR48*	p.L38Q	TCGA-12-1098	Cancer
3	39104605	39104605	T	C	*WDR48*	p.L38P	TCGA-24-1556	Cancer
3	39104610	39104610	C	T	*WDR48*	p.P40S	TCGA-24-1556	Cancer
3	39136144	39136144	G	T	*WDR48*	p.L648F	TCGA-B2-4101	Cancer
3	39136154	39136154	A	T	*WDR48*	p.M652L	TCGA-36-1577	Cancer
3	41267225	41267225	A	G	*CTNNB1*	p.K263R	TCGA-12-1095	Cancer
3	41275222	41275222	C	A	*CTNNB1*	p.P456H	TCGA-AA-A00R	Cancer
3	41278078	41278078	G	T	*CTNNB1*	e11-1	TCGA-DX-A6YT	Cancer
3	47125727	47125727	G	A	*SETD2*	p.T1848I	TCGA-AA-A00R	Cancer
3	47155460	47155460	T	G	*SETD2*	p.N1541H	TCGA-DJ-A1QE	Cancer
3	47162630	47162630	C	G	*SETD2*	p.D1166H	TCGA-AJ-A3OK	Cancer
3	49399957	49399957	G	A	*RHOA*	p.Q87*	TCGA-13-0801	Cancer
3	49399957	49399957	G	A	*RHOA*	p.Q87*	TCGA-25-1627	Cancer
3	52438565	52438565	C	T	*BAP1*	p.R385Q	TCGA-CI-6620	Cancer
3	52439212	52439212	T	A	*BAP1*	p.N344Y	TCGA-AF-2689	Cancer
3	52442566	52442566	C	G	*BAP1*	p.R60P	TCGA-AF-2689	Cancer
3	52584823	52584823	G	C	*PBRM1*	p.N1540K	TCGA-29-2431	Cancer
3	52588862	52588862	G	C	*PBRM1*	p.P1496R	TCGA-AA-A01G	Cancer
3	52613181	52613181	C	A	*PBRM1*	p.G1141V	TCGA-AX-A06F	Cancer
3	69987136	69987136	C	G	*MITF*	p.P173R	TCGA-36-1581	Cancer
3	69987136	69987136	C	G	*MITF*	p.P173R	TCGA-AG-3578	Cancer
3	70000970	70000970	T	G	*MITF*	p.I296M	TCGA-AA-A00K	Cancer
3	89259331	89259331	G	A	*EPHA3*	p.D159N	TCGA-A6-2676	Cancer
3	89259331	89259331	G	A	*EPHA3*	p.D159N	TCGA-AA-A00K	Cancer
3	89259331	89259331	G	A	*EPHA3*	p.D159N	TCGA-AF-2689	Cancer
3	100432602	100432602	C	T	*TFG*	p.P25S	TCGA-AG-A016	Cancer
3	105397340	105397340	T	C	*CBLB*	p.H50R	TCGA-06-5858	Cancer
3	119562134	119562134	G	A	*GSK3B*	p.A401V	TCGA-AG-A016	Cancer
3	119595291	119595291	A	T	*GSK3B*	p.F293Y	TCGA-AA-3529	Cancer
3	119631592	119631592	G	A	*GSK3B*	p.P225L	TCGA-AK-3444	Cancer
3	121190975	121190975	G	A	*POLQ*	p.P2194S	TCGA-13-0901	Cancer
3	121215661	121215661	A	C	*POLQ*	p.Y758D	TCGA-24-0968	Cancer
3	121238888	121238888	C	G	*POLQ*	p.R433P	TCGA-AK-3434	Cancer
3	121263625	121263625	C	T	*POLQ*	p.E98K	TCGA-CI-6620	Cancer
3	123348330	123348330	T	A	*MYLK*	p.K1702I	TCGA-24-1548	Cancer
3	123376087	123376087	G	A	*MYLK*	p.Q1392*	TCGA-AX-A05T	Cancer
3	123401120	123401120	C	T	*MYLK*	p.M1201I	TCGA-CI-6620	Cancer
3	123427591	123427591	C	A	*MYLK*	p.W698C	TCGA-09-1662	Cancer
3	124462883	124462883	G	C	*UMPS*	p.M465I	TCGA-13-1408	Cancer
3	142176450	142176450	T	C	*ATR*	p.M2551V	TCGA-CQ-A4CI	Cancer
3	142188181	142188181	T	G	*ATR*	p.K2184Q	TCGA-AK-3431	Cancer
3	142215256	142215256	C	T	*ATR*	p.E1949K	TCGA-EO-A2CH	Cancer
3	142215300	142215300	G	T	*ATR*	p.A1934D	TCGA-X6-A7WA	Cancer
3	142217487	142217487	G	C	*ATR*	p.A1837G	TCGA-24-0968	Cancer
3	142269008	142269008	A	G	*ATR*	p.V981A	TCGA-25-1317	Cancer
3	156395548	156395548	A	G	*TIPARP*	p.D21G	TCGA-29-1784	Cancer
3	156396160	156396160	A	T	*TIPARP*	p.E225V	TCGA-25-1317	Cancer
3	168810807	168810807	T	A	*MECOM*	p.I838F	TCGA-HU-A4GU	Cancer
3	168812911	168812911	G	T	*MECOM*	p.T794N	TCGA-13-0793	Cancer
3	168812911	168812911	G	T	*MECOM*	p.T794N	TCGA-13-0801	Cancer
3	168812911	168812911	G	T	*MECOM*	p.T794N	TCGA-24-1553	Cancer
3	168833978	168833978	G	A	*MECOM*	p.P373L	TCGA-36-1578	Cancer
3	168834045	168834045	G	T	*MECOM*	p.Q351K	TCGA-BH-A0B8	Cancer
3	168834242	168834242	T	C	*MECOM*	p.N285S	TCGA-14-0783	Cancer

3	168834252	168834252	A	C	*MECOM*	p.F282V	TCGA-AA-3534	Cancer
3	168834317	168834317	G	C	*MECOM*	p.T260S	TCGA-A7-A5ZV	Cancer
3	168838970	168838970	G	A	*MECOM*	p.R148W	TCGA-B5-A0JN	Cancer
3	169099094	169099094	T	C	*MECOM*	p.N86D	TCGA-13-1407	Cancer
3	183212063	183212063	T	G	*KLHL6*	p.K385T	TCGA-AA-A00D	Cancer
3	183273381	183273381	C	T	*KLHL6*	p.E21K	TCGA-25-1318	Cancer
3	186502802	186502802	T	A	*EIF4A2*	p.F88Y	TCGA-25-1627	Cancer
3	187443369	187443369	G	T	*BCL6*	p.P586Q	TCGA-25-1634	Cancer
4	1803393	1803393	C	T	*FGFR3*	p.S221L	TCGA-CI-6620	Cancer
4	55133516	55133516	A	T	*PDGFRA*	p.T274S	TCGA-BG-A0M2	Cancer
4	55152092	55152092	G	C	*PDGFRA*	p.D842H	TCGA-AG-3601	Cancer
4	55161371	55161371	G	A	*PDGFRA*	p.E1068K	TCGA-AF-3911	Cancer
4	66286259	66286259	T	G	*EPHA5*	p.K476T	TCGA-AG-3581	Cancer
4	66509098	66509098	C	A	*EPHA5*	p.A77S	TCGA-36-2547	Cancer
4	134072199	134072199	G	C	*PCDH10*	p.G302R	TCGA-13-1497	Cancer
4	134072730	134072730	G	A	*PCDH10*	p.G479S	TCGA-AA-3514	Cancer
4	153250903	153250903	C	T	*FBXW7*	p.C386Y	TCGA-24-1604	Cancer
4	153253796	153253796	T	G	*FBXW7*	p.I313L	TCGA-AA-A01Q	Cancer
4	153332525	153332525	G	C	*FBXW7*	p.T144R	TCGA-95-7043	Cancer
4	153332853	153332853	G	A	*FBXW7*	p.R35C	TCGA-A2-A1FV	Cancer
4	186272686	186272686	G	A	*SNX25*	p.V633M	TCGA-ER-A19H	Cancer
4	187517765	187517765	G	A	*FAT1*	p.P4310L	TCGA-DJ-A13U	Cancer
4	187517870	187517870	C	A	*FAT1*	p.G4275V	TCGA-B2-4101	Cancer
4	187517879	187517879	G	C	*FAT1*	p.S4272C	TCGA-98-8023	Cancer
4	187517919	187517919	T	A	*FAT1*	p.S4259C	TCGA-CV-6943	Cancer
4	187521264	187521264	C	T	*FAT1*	p.G3964E	TCGA-BG-A0VZ	Cancer
4	187540142	187540142	T	A	*FAT1*	p.N2533I	TCGA-L9-A444	Cancer
4	187630170	187630170	G	A	*FAT1*	p.P271L	TCGA-G9-6367	Cancer
4	187630461	187630461	G	A	*FAT1*	p.T174M	TCGA-PQ-A6FN	Cancer
5	1268706	1268706	G	A	*TERT*	p.L777F	TCGA-55-8299	Cancer
5	19571756	19571756	A	T	*CDH18*	p.N395K	TCGA-AG-3587	Cancer
5	21817166	21817166	T	G	*CDH12*	p.D297A	TCGA-AA-A00R	Cancer
5	21817167	21817167	C	G	*CDH12*	p.D297H	TCGA-24-1604	Cancer
5	21975325	21975325	A	T	*CDH12*	p.I134K	TCGA-BH-A0HN	Cancer
5	38954903	38954903	T	G	*RICTOR*	p.K890N	TCGA-13-0793	Cancer
5	38959887	38959887	A	C	*RICTOR*	p.F682C	TCGA-23-1024	Cancer
5	39021206	39021206	T	C	*RICTOR*	p.N44D	TCGA-12-1094	Cancer
5	44388784	44388784	T	C	*FGF10*	p.M1V	TCGA-AG-3605	Cancer
5	56155570	56155570	A	T	*MAP3K1*	p.D221V	TCGA-21-5783	Cancer
5	56168705	56168705	A	T	*MAP3K1*	p.Q520L	TCGA-FS-A4F2	Cancer
5	56178062	56178062	G	T	*MAP3K1*	p.C1012F	TCGA-24-0968	Cancer
5	74848352	74848352	T	C	*POLK*	p.I64T	TCGA-06-1805	Cancer
5	86658429	86658429	G	A	*RASA1*	p.R299H	TCGA-CU-A3QU	Cancer
5	86675579	86675579	G	A	*RASA1*	p.E673K	TCGA-25-1324	Cancer
5	112163683	112163683	G	A	*APC*	p.E536K	TCGA-BB-8601	Cancer
5	131939181	131939181	G	C	*RAD50*	p.Q799H	TCGA-A6-5659	Cancer
5	138253552	138253552	A	G	*CTNNA1*	p.D504G	TCGA-E2-A10B	Cancer
5	138268282	138268282	T	C	*CTNNA1*	p.C772R	TCGA-36-1571	Cancer
5	140960326	140960326	G	A	*DIAPH1*	p.P270L	TCGA-AX-A05T	Cancer
5	149439275	149439275	T	C	*CSF1R*	p.K707R	TCGA-AA-3561	Cancer
5	149439275	149439275	T	C	*CSF1R*	p.K707R	TCGA-AG-3583	Cancer
5	149499120	149499120	G	A	*PDGFRB*	p.P903L	TCGA-06-0939	Cancer
5	149505037	149505037	C	T	*PDGFRB*	p.W593*	TCGA-AF-2689	Cancer
5	149509531	149509531	C	A	*PDGFRB*	p.R456S	TCGA-E1-5304	Cancer
5	176517975	176517975	A	G	*FGFR4*	p.K158R	TCGA-12-1095	Cancer
5	176517975	176517975	A	G	*FGFR4*	p.K158R	TCGA-AA-3514	Cancer
5	176517975	176517975	A	G	*FGFR4*	p.K158R	TCGA-AG-3583	Cancer
5	176520465	176520465	G	A	*FGFR4*	p.R437H	TCGA-AQ-A54N	Cancer

5	176562545	176562545	G	T	*NSD1*	p.K147N	TCGA-24-1105	Cancer
5	176637463	176637463	G	T	*NSD1*	p.R688M	TCGA-36-1570	Cancer
5	176684134	176684134	A	C	*NSD1*	p.N1650H	TCGA-E2-A15G	Cancer
5	176696799	176696799	T	A	*NSD1*	p.Y1834N	TCGA-24-1548	Cancer
5	176709529	176709529	C	T	*NSD1*	p.R1986C	TCGA-23-1122	Cancer
5	176719079	176719079	G	T	*NSD1*	p.G2128V	TCGA-19-0955	Cancer
5	176722279	176722279	G	A	*NSD1*	p.W2637*	TCGA-E2-A1IH	Cancer
5	180041155	180041155	C	T	*FLT4*	p.A1082T	TCGA-AA-3534	Cancer
5	180057056	180057056	C	T	*FLT4*	p.R188Q	TCGA-DK-A2I6	Cancer
6	401729	401729	G	A	*IRF4*	p.E351K	TCGA-A2-A0EY	Cancer
6	17507959	17507959	T	C	*CAP2*	e5+2	TCGA-AA-3514	Cancer
6	18130961	18130961	G	A	*TPMT*	p.R226*	TCGA-CD-8533	Cancer
6	26032033	26032033	G	C	*HIST1H3B*	p.Q86E	TCGA-AA-3514	Cancer
6	26056198	26056198	T	G	*HIST1H1C*	p.K153N	TCGA-KM-8443	Cancer
6	26056365	26056365	C	G	*HIST1H1C*	p.G98R	TCGA-13-0901	Cancer
6	31324465	31324465	C	T	*HLA-B*	p.G115R	TCGA-62-A46O	Cancer
6	31545193	31545193	T	A	*TNF*	p.I194N	TCGA-H7-8502	Cancer
6	33288275	33288275	T	C	*DAXX*	p.K378R	TCGA-13-0802	Cancer
6	35425351	35425351	C	A	*FANCE*	p.Q292K	TCGA-AG-A016	Cancer
6	41904412	41904412	G	A	*CCND3*	p.P149L	TCGA-CG-4449	Cancer
6	44216418	44216418	C	T	*HSP90AB1*	p.Q18*	TCGA-CI-6620	Cancer
6	44220960	44220960	C	T	*HSP90AB1*	p.T637M	TCGA-63-A5MU	Cancer
6	44221316	44221316	G	A	*HSP90AB1*	p.R719H	TCGA-EE-A2MR	Cancer
6	106552908	106552908	G	T	*PRDM1*	p.M291I	TCGA-AA-3514	Cancer
6	106553435	106553435	C	T	*PRDM1*	p.P467L	TCGA-HU-8249	Cancer
6	106554918	106554918	C	T	*PRDM1*	p.R679C	TCGA-16-1055	Cancer
6	111636543	111636543	A	G	*REV3L*	p.F2798S	TCGA-AG-3601	Cancer
6	111688498	111688498	T	A	*REV3L*	p.K2165*	TCGA-25-1634	Cancer
6	111695842	111695842	A	T	*REV3L*	p.V1239D	TCGA-D1-A179	Cancer
6	111696278	111696278	C	T	*REV3L*	p.A1094T	TCGA-GV-A3JX	Cancer
6	111697819	111697819	T	G	*REV3L*	p.K580T	TCGA-AA-A00K	Cancer
6	111710413	111710413	G	T	*REV3L*	p.A253D	TCGA-X6-A7WA	Cancer
6	111714106	111714106	A	C	*REV3L*	p.F212C	TCGA-13-0761	Cancer
6	111714106	111714106	A	C	*REV3L*	p.F212C	TCGA-AG-3600	Cancer
6	117630001	117630001	T	G	*ROS1*	p.Q2175H	TCGA-24-1548	Cancer
6	117641140	117641140	T	C	*ROS1*	p.K1944R	TCGA-AG-3608	Cancer
6	131912498	131912498	G	T	*MED23*	p.A1220D	TCGA-A2-A0EQ	Cancer
6	131915340	131915340	A	G	*MED23*	p.L1050P	TCGA-AA-3514	Cancer
6	134495180	134495180	T	C	*SGK1*	p.Q159R	TCGA-AA-3514	Cancer
6	134495661	134495661	T	A	*SGK1*	p.Y142F	TCGA-FG-7637	Cancer
6	135510953	135510953	C	T	*MYB*	p.H80Y	TCGA-25-1634	Cancer
6	136588182	136588182	G	T	*BCLAF1*	p.S843R	TCGA-AG-3608	Cancer
6	136596792	136596792	T	C	*BCLAF1*	p.H577R	TCGA-06-1801	Cancer
6	136599108	136599108	G	A	*BCLAF1*	p.T304I	TCGA-BR-8286	Cancer
6	136600941	136600941	G	A	*BCLAF1*	p.R22*	TCGA-BF-A1PU	Cancer
6	136600997	136600997	C	T	*BCLAF1*	p.R3H	TCGA-C8-A12T	Cancer
6	138195999	138195999	A	G	*TNFAIP3*	p.M105V	TCGA-AA-A00W	Cancer
6	152332878	152332878	C	T	*ESR1*	p.S395F	TCGA-AA-3517	Cancer
6	157431662	157431662	G	A	*ARID1B*	p.A767T	TCGA-AO-A124	Cancer
6	160103611	160103611	T	A	*SOD2*	p.N149Y	TCGA-AA-3529	Cancer
6	163987758	163987758	G	T	*QKI*	p.V314L	TCGA-GF-A2C7	Cancer
7	6026386	6026390	TTACC	-	*PMS2*	p.S669fs	TCGA-M7-A725	Cancer
7	81331970	81331970	A	C	*HGF*	p.I705S	TCGA-AA-3534	Cancer
7	90895939	90895940	-	AA	*FZD1*	p.P582fs	TCGA-23-1021	Cancer
7	95761141	95761141	G	T	*SLC25A13*	p.P503Q	TCGA-EE-A2MH	Cancer
7	95761187	95761187	T	G	*SLC25A13*	p.K488Q	TCGA-AA-3534	Cancer
7	95799417	95799417	A	C	*SLC25A13*	p.D418E	TCGA-16-1055	Cancer
7	95818962	95818962	T	C	*SLC25A13*	p.K260R	TCGA-12-1095	Cancer

7	98509802	98509802	C	T	*TRRAP*	p.S722F	TCGA-ER-A42L	Cancer
7	98601827	98601827	G	A	*TRRAP*	p.V3428I	TCGA-EE-A2MS	Cancer
7	98609871	98609871	C	T	*TRRAP*	p.L3825F	TCGA-62-8398	Cancer
7	99359784	99359784	T	C	*CYP3A4*	p.K378R	TCGA-AG-3608	Cancer
7	101459313	101459313	G	A	*CUX1*	p.M1I	TCGA-EE-A29S	Cancer
7	106509078	106509078	G	T	*PIK3CG*	p.D358Y	TCGA-DM-A1D0	Cancer
7	116340189	116340189	C	A	*MET*	p.P351T	TCGA-AG-4009	Cancer
7	140500174	140500174	G	A	*BRAF*	p.S323L	TCGA-06-2565	Cancer
8	37697744	37697744	G	A	*GPR124*	p.E873K	TCGA-25-1632	Cancer
8	37698301	37698301	T	C	*GPR124*	p.I894T	TCGA-AA-3518	Cancer
8	37698364	37698364	C	A	*GPR124*	p.P915H	TCGA-AK-3461	Cancer
8	38282218	38282218	C	G	*FGFR1*	e6-1	TCGA-95-8494	Cancer
8	41790240	41790240	G	C	*KAT6A*	p.S1833C	TCGA-AG-A016	Cancer
8	41790454	41790454	T	C	*KAT6A*	p.T1762A	TCGA-76-6280	Cancer
8	41791498	41791498	C	T	*KAT6A*	p.E1414K	TCGA-CI-6620	Cancer
8	41836185	41836185	T	G	*KAT6A*	p.N340H	TCGA-13-0793	Cancer
8	48690378	48690378	T	A	*PRKDC*	p.M3970L	TCGA-F4-6703	Cancer
8	48769761	48769761	T	A	*PRKDC*	p.I2188L	TCGA-B2-4101	Cancer
8	48775016	48775016	T	C	*PRKDC*	p.N1945S	TCGA-SI-A71Q	Cancer
8	48866386	48866386	A	T	*PRKDC*	p.L201Q	TCGA-AK-3428	Cancer
8	55371005	55371005	G	T	*SOX17*	p.G103C	TCGA-AG-3598	Cancer
8	118812120	118812120	C	T	*EXT1*	p.R691H	TCGA-A8-A08Z	Cancer
8	118831937	118831937	G	T	*EXT1*	p.A505D	TCGA-13-0792	Cancer
8	118831940	118831940	G	T	*EXT1*	p.A504E	TCGA-36-1571	Cancer
8	118831946	118831946	A	T	*EXT1*	p.V502E	TCGA-36-1581	Cancer
8	119122623	119122623	G	C	*EXT1*	p.I221M	TCGA-AA-3514	Cancer
9	328039	328039	C	G	*DOCK8*	p.H236Q	TCGA-09-1666	Cancer
9	328039	328039	C	G	*DOCK8*	p.H236Q	TCGA-A6-2678	Cancer
9	328039	328039	C	G	*DOCK8*	p.H236Q	TCGA-AA-A01F	Cancer
9	328039	328039	C	G	*DOCK8*	p.H236Q	TCGA-AG-3609	Cancer
9	377046	377046	G	A	*DOCK8*	p.V691M	TCGA-FS-A1ZD	Cancer
9	382560	382560	C	A	*DOCK8*	p.R817S	TCGA-MP-A4SV	Cancer
9	390549	390549	T	C	*DOCK8*	p.F885L	TCGA-E7-A5KE	Cancer
9	429741	429741	T	G	*DOCK8*	p.F1405V	TCGA-AK-3434	Cancer
9	434893	434893	T	C	*DOCK8*	p.V1566A	TCGA-AA-3514	Cancer
9	441358	441358	C	T	*DOCK8*	p.R1666W	TCGA-BA-6869	Cancer
9	8341773	8341773	T	C	*PTPRD*	p.R1623G	TCGA-DU-7302	Cancer
9	8376030	8376030	C	G	*PTPRD*	p.G1523R	TCGA-25-1626	Cancer
9	8376030	8376030	C	G	*PTPRD*	p.G1523R	TCGA-36-1575	Cancer
9	8376030	8376030	C	G	*PTPRD*	p.G1523R	TCGA-B2-4101	Cancer
9	8484252	8484252	C	T	*PTPRD*	p.G1094R	TCGA-AA-3514	Cancer
9	8485885	8485885	T	C	*PTPRD*	p.T978A	TCGA-62-A46O	Cancer
9	8492883	8492883	G	C	*PTPRD*	p.L816V	TCGA-D1-A16N	Cancer
9	8521421	8521421	C	T	*PTPRD*	p.A273T	TCGA-BR-8289	Cancer
9	21854709	21854709	T	C	*MTAP*	p.F177S	TCGA-AA-3514	Cancer
9	21854709	21854709	T	C	*MTAP*	p.F177S	TCGA-AA-3549	Cancer
9	21854709	21854709	T	C	*MTAP*	p.F177S	TCGA-AA-A00K	Cancer
9	21854709	21854709	T	C	*MTAP*	p.F177S	TCGA-AG-A016	Cancer
9	21854823	21854823	T	C	*MTAP*	p.I215T	TCGA-13-0714	Cancer
9	21971000	21971000	C	A	*CDKN2A*	p.E120*	TCGA-97-8174	Cancer
9	32429452	32429452	T	G	*ACO1*	p.I507S	TCGA-91-8499	Cancer
9	32430482	32430482	A	C	*ACO1*	p.T546P	TCGA-AG-A016	Cancer
9	87356844	87356844	T	C	*NTRK2*	e9+2	TCGA-CV-A6JO	Cancer
9	93641155	93641155	T	-	*SYK*	p.L501fs	TCGA-CN-A63T	Cancer
9	98002969	98002969	G	T	*FANCC*	p.Q103K	TCGA-AG-A016	Cancer
9	101891211	101891211	G	C	*TGFBR1*	p.V58L	TCGA-16-1055	Cancer
9	101894946	101894946	G	A	*TGFBR1*	p.D98N	TCGA-19-1392	Cancer
9	127933374	127933374	A	G	*PPP6C*	p.I91T	TCGA-AA-3514	Cancer

9	130270257	130270257	C	T	*FAM129B*	p.E515K	TCGA-CI-6620	Cancer
9	130271277	130271277	C	T	*FAM129B*	p.R432Q	TCGA-HR-A2OG	Cancer
9	130286093	130286093	G	T	*FAM129B*	p.L152I	TCGA-DQ-7596	Cancer
9	133730245	133730245	C	A	*ABL1*	p.T123N	TCGA-AK-3460	Cancer
9	133738305	133738305	G	T	*ABL1*	p.W254C	TCGA-AG-3578	Cancer
9	133759443	133759443	G	A	*ABL1*	p.R608H	TCGA-DK-A2I1	Cancer
9	133760339	133760339	G	T	*ABL1*	p.D907Y	TCGA-FG-7637	Cancer
9	135781005	135781005	G	C	*TSC1*	p.Q654E	TCGA-CF-A3MI	Cancer
9	137300061	137300061	G	A	*RXRA*	p.V116I	TCGA-A7-A5ZV	Cancer
9	139390560	139390560	T	C	*NOTCH1*	p.Q2544R	TCGA-K6-A3WQ	Cancer
9	139407901	139407901	C	G	*NOTCH1*	p.G766R	TCGA-DK-A3X2	Cancer
10	32307297	32307297	A	C	*KIF5B*	p.L796V	TCGA-AF-2689	Cancer
10	32310198	32310198	C	T	*KIF5B*	p.E684K	TCGA-AA-A00K	Cancer
10	32310204	32310204	C	T	*KIF5B*	p.E682K	TCGA-CI-6620	Cancer
10	43617434	43617434	T	G	*RET*	p.F924C	TCGA-AA-3520	Cancer
10	43619174	43619174	C	T	*RET*	p.P953S	TCGA-AK-3460	Cancer
10	88676981	88676981	G	A	*BMPR1A*	p.E256K	TCGA-DK-A6B0	Cancer
10	90770573	90770573	G	A	*FAS*	e6+1	TCGA-13-1410	Cancer
10	96281821	96281821	T	G	*TBC1D12*	p.F624C	TCGA-13-0802	Cancer
10	96281821	96281821	T	G	*TBC1D12*	p.F624C	TCGA-23-1028	Cancer
10	96281821	96281821	T	G	*TBC1D12*	p.F624C	TCGA-24-1553	Cancer
10	96612671	96612671	A	G	*CYP2C19*	p.*491W	TCGA-36-1577	Cancer
10	101558976	101558976	A	C	*ABCC2*	p.K294Q	TCGA-AA-A01Q	Cancer
10	101559128	101559128	G	C	*ABCC2*	e8+1	TCGA-AG-A016	Cancer
10	101567937	101567937	T	C	*ABCC2*	p.L589P	TCGA-DX-A6YV	Cancer
10	101569918	101569918	A	G	*ABCC2*	p.K615E	TCGA-13-0717	Cancer
10	101954183	101954183	T	C	*CHUK*	p.S609G	TCGA-36-1577	Cancer
10	101981904	101981904	T	C	*CHUK*	p.K117R	TCGA-24-1544	Cancer
10	101981905	101981905	T	C	*CHUK*	p.K117E	TCGA-36-1571	Cancer
10	104359222	104359222	G	A	*SUFU*	p.G315R	TCGA-EX-A1H5	Cancer
10	104595083	104595083	G	A	*CYP17A1*	p.Q122*	TCGA-AX-A06J	Cancer
10	120809315	120809315	T	C	*EIF3A*	p.K886E	TCGA-25-1633	Cancer
10	120816340	120816340	C	G	*EIF3A*	p.R703P	TCGA-AA-A00D	Cancer
10	120816341	120816341	G	C	*EIF3A*	p.R703G	TCGA-AA-A00U	Cancer
10	123256084	123256084	G	T	*FGFR2*	p.Q610K	TCGA-24-1548	Cancer
10	123274681	123274681	G	T	*FGFR2*	p.P414T	TCGA-AG-3598	Cancer
11	534250	534250	G	T	*HRAS*	p.Q25K	TCGA-EE-A2GR	Cancer
11	9450599	9450599	T	G	*IPO7*	p.F483V	TCGA-AK-3428	Cancer
11	9452479	9452479	A	T	*IPO7*	p.K604*	TCGA-AA-3514	Cancer
11	44129500	44129500	C	G	*EXT2*	p.R113G	TCGA-24-1614	Cancer
11	44129650	44129650	T	C	*EXT2*	p.Y163H	TCGA-AA-A00D	Cancer
11	44129650	44129650	T	C	*EXT2*	p.Y163H	TCGA-AF-2689	Cancer
11	64540948	64540948	T	C	*SF1*	p.K189E	TCGA-AA-A00K	Cancer
11	77931484	77931484	G	T	*GAB2*	p.P590T	TCGA-12-1091	Cancer
11	88924547	88924547	G	A	*TYR*	p.D333N	TCGA-AF-2689	Cancer
11	92495335	92495335	C	T	*FAT3*	p.T1328M	TCGA-HU-8245	Cancer
11	92507249	92507249	G	T	*FAT3*	p.G1413V	TCGA-23-1026	Cancer
11	92507249	92507249	G	T	*FAT3*	p.G1413V	TCGA-25-1324	Cancer
11	92577800	92577800	G	A	*FAT3*	p.G91D	TCGA-GM-A2DH	Cancer
11	92623878	92623878	G	T	*FAT3*	p.E760*	TCGA-86-8359	Cancer
11	106558328	106558328	T	C	*GUCY1A2*	p.K747E	TCGA-AA-3518	Cancer
11	106579331	106579331	C	T	*GUCY1A2*	p.R664H	TCGA-AA-A00K	Cancer
11	106680775	106680775	G	C	*GUCY1A2*	p.L546V	TCGA-AA-A00D	Cancer
11	106681104	106681104	C	T	*GUCY1A2*	p.R436Q	TCGA-CI-6620	Cancer
11	106810299	106810299	C	T	*GUCY1A2*	p.E365K	TCGA-97-7552	Cancer
11	108117691	108117691	G	T	*ATM*	p.G301V	TCGA-BG-A0VZ	Cancer
11	108117798	108117798	C	T	*ATM*	p.R337C	TCGA-52-7812	Cancer
11	108141790	108141790	G	C	*ATM*	e18-1	TCGA-A2-A0CS	Cancer

11	108142001	108142001	G	A	*ATM*	p.R982H	TCGA-AK-3456	Cancer
11	108173700	108173700	T	G	*ATM*	p.L1814V	TCGA-26-1440	Cancer
11	108186639	108186639	G	A	*ATM*	NULL	TCGA-CV-7101	Cancer
11	108192065	108192065	G	A	*ATM*	p.E2164K	TCGA-GU-A767	Cancer
11	108201099	108201099	C	T	*ATM*	p.S2489F	TCGA-85-8351	Cancer
11	108213988	108213988	T	G	*ATM*	p.C2770G	TCGA-78-8655	Cancer
11	108216546	108216546	G	A	*ATM*	p.R2832H	TCGA-AO-A0J3	Cancer
11	108218084	108218084	T	C	*ATM*	p.I2888T	TCGA-G2-A2EC	Cancer
11	111228223	111228223	C	A	*POU2AF1*	p.G135W	TCGA-25-1634	Cancer
12	402089	402089	C	T	*KDM5A*	p.A1568T	TCGA-AK-3434	Cancer
12	936232	936232	G	A	*WNK1*	p.M319I	TCGA-D1-A103	Cancer
12	974498	974498	C	T	*WNK1*	p.R788C	TCGA-GM-A3XL	Cancer
12	977270	977270	A	G	*WNK1*	p.N878S	TCGA-EI-6881	Cancer
12	994511	994511	C	T	*WNK1*	p.S2012F	TCGA-CI-6620	Cancer
12	995105	995105	C	T	*WNK1*	p.P2210L	TCGA-06-2569	Cancer
12	1025575	1025575	T	G	*RAD52*	p.Q267P	TCGA-25-1317	Cancer
12	18719891	18719891	A	C	*PIK3C2G*	p.H1304P	TCGA-13-0802	Cancer
12	18719891	18719891	A	C	*PIK3C2G*	p.H1304P	TCGA-25-1324	Cancer
12	40702988	40702988	G	A	*LRRK2*	p.G1424R	TCGA-12-1098	Cancer
12	40714859	40714859	T	G	*LRRK2*	p.I1680R	TCGA-AG-3574	Cancer
12	40734202	40734202	G	A	*LRRK2*	p.G2019S	TCGA-09-0369	Cancer
12	46211492	46211492	A	T	*ARID2*	p.N153I	TCGA-AA-3555	Cancer
12	46298778	46298778	T	C	*ARID2*	p.S1809P	TCGA-09-1662	Cancer
12	50029249	50029249	A	G	*PRPF40B*	p.K401R	TCGA-AG-3594	Cancer
12	52387811	52387811	G	A	*ACVR1B*	p.A479T	TCGA-AG-3608	Cancer
12	56487605	56487605	G	A	*ERBB3*	p.G513D	TCGA-14-0867	Cancer
12	56490607	56490607	C	G	*ERBB3*	p.Q751E	TCGA-AA-3534	Cancer
12	56492290	56492290	A	C	*ERBB3*	p.I116L	TCGA-AA-A00Q	Cancer
12	56492296	56492296	T	A	*ERBB3*	p.W118R	TCGA-13-0717	Cancer
12	56492296	56492296	T	A	*ERBB3*	p.W118R	TCGA-13-0793	Cancer
12	56493451	56493451	C	G	*ERBB3*	p.N194K	TCGA-04-1525	Cancer
12	56495474	56495474	T	C	*ERBB3*	p.Y463H	TCGA-AA-A00K	Cancer
12	57033003	57033003	T	G	*ATP5B*	p.K448T	TCGA-24-1614	Cancer
12	57033003	57033003	T	G	*ATP5B*	p.K448T	TCGA-AA-A00D	Cancer
12	78392238	78392238	G	A	*NAV3*	p.A288T	TCGA-FW-A5DX	Cancer
12	78415643	78415643	G	A	*NAV3*	e9+1	TCGA-FI-A2EY	Cancer
12	78515866	78515866	G	C	*NAV3*	p.G1299A	TCGA-AX-A063	Cancer
12	96131740	96131740	G	C	*NTN4*	p.F219L	TCGA-24-1548	Cancer
12	111884838	111884838	G	A	*SH2B3*	e3+1	TCGA-BH-A0B1	Cancer
12	133202357	133202357	C	G	*POLE*	e9-1	TCGA-23-2078	Cancer
12	133250183	133250183	C	T	*POLE*	p.R457Q	TCGA-HB-A3YV	Cancer
13	20763395	20763408	CCCTTGATGAACTT	-	*GJB2*	p.K105fs	TCGA-D7-A6EV	Cancer
13	25016729	25016729	C	T	*PARP4*	p.R1181K	TCGA-AX-A3GI	Cancer
13	25021323	25021323	A	C	*PARP4*	p.I1039R	TCGA-MN-A4N1	Cancer
13	26911726	26911726	T	A	*CDK8*	p.L51I	TCGA-AA-A00K	Cancer
13	26975712	26975712	T	C	*CDK8*	p.V407A	TCGA-AQ-A54N	Cancer
13	27216449	27216449	C	A	*WASF3*	p.C14*	TCGA-12-1098	Cancer
13	27246071	27246071	A	G	*WASF3*	p.K162R	TCGA-AA-3517	Cancer
13	103506156	103506156	C	T	*ERCC5*	p.T105M	TCGA-13-0751	Cancer
13	103519132	103519132	A	T	*ERCC5*	p.N824Y	TCGA-AA-3514	Cancer
14	21863247	21863247	G	C	*CHD8*	p.F1738L	TCGA-12-0773	Cancer
14	21867802	21867802	C	T	*CHD8*	p.S1627N	TCGA-AP-A05H	Cancer
14	21869607	21869607	C	A	*CHD8*	p.K1376N	TCGA-13-0901	Cancer
14	21869607	21869607	C	A	*CHD8*	p.K1376N	TCGA-AK-3433	Cancer
14	21882512	21882512	T	C	*CHD8*	p.K697R	TCGA-23-1026	Cancer
14	62194229	62194229	G	A	*HIF1A*	p.C234Y	TCGA-86-8054	Cancer
14	64716307	64716307	G	C	*ESR2*	p.H394Q	TCGA-FI-A2F9	Cancer
14	65543274	65543274	C	T	*MAX*	p.D72N	TCGA-AP-A0LL	Cancer

14	95556866	95556866	T	C	*DICER1*	p.K811R	TCGA-KQ-A41R	Cancer
14	95557596	95557596	C	A	*DICER1*	p.G722V	TCGA-AR-A0TP	Cancer
14	95570161	95570161	A	T	*DICER1*	p.L89*	TCGA-24-0968	Cancer
14	95577701	95577701	G	C	*DICER1*	p.P737A	TCGA-ER-A194	Cancer
14	103342756	103342756	A	C	*TRAF3*	p.E155A	TCGA-AG-3578	Cancer
15	31795960	31795960	C	T	*OTUD7A*	p.A319T	TCGA-FI-A3PV	Cancer
15	40488922	40488922	T	C	*BUB1B*	p.F426S	TCGA-AA-3514	Cancer
15	41961438	41961438	G	C	*MGA*	p.D116H	TCGA-25-1318	Cancer
15	41961453	41961453	T	G	*MGA*	p.Y121D	TCGA-24-0968	Cancer
15	41999973	41999973	A	G	*MGA*	p.K746E	TCGA-62-8395	Cancer
15	42002979	42002979	A	G	*MGA*	p.E839G	TCGA-D5-6540	Cancer
15	43701940	43701940	C	T	*TP53BP1*	e25-1	TCGA-19-0960	Cancer
15	43738687	43738687	A	G	*TP53BP1*	p.S980P	TCGA-AA-3561	Cancer
15	43748399	43748399	C	A	*TP53BP1*	p.A803S	TCGA-06-2569	Cancer
15	66727463	66727463	T	G	*MAP2K1*	p.V60G	TCGA-A8-A08O	Cancer
15	67462920	67462920	G	A	*SMAD3*	p.M212I	TCGA-13-1498	Cancer
15	75644493	75644493	G	A	*NEIL1*	p.R159Q	TCGA-CL-5917	Cancer
15	75682073	75682073	C	T	*SIN3A*	p.D981N	TCGA-AX-A06F	Cancer
15	75705208	75705208	G	T	*SIN3A*	p.P218T	TCGA-X6-A7WA	Cancer
15	75722679	75722679	T	C	*SIN3A*	p.Y13C	TCGA-HT-A74K	Cancer
15	80450402	80450402	C	G	*FAH*	p.P28A	TCGA-50-5946	Cancer
15	89811667	89811667	C	T	*FANCI*	p.R265C	TCGA-CI-6620	Cancer
15	89826432	89826432	T	G	*FANCI*	p.L550*	TCGA-19-0962	Cancer
15	91333927	91333927	G	A	*BLM*	p.V958M	TCGA-HE-A5NI	Cancer
15	91346838	91346838	T	A	*BLM*	p.L1149Q	TCGA-AF-3400	Cancer
15	99459339	99459339	C	T	*IGF1R*	p.R659W	TCGA-62-8395	Cancer
15	99491888	99491888	G	A	*IGF1R*	p.V1225I	TCGA-BG-A0VT	Cancer
15	99500663	99500663	A	C	*IGF1R*	p.T1366P	TCGA-24-1419	Cancer
16	2012148	2012149	-	C	*RPS2*	p.T278fs	TCGA-AX-A06J	Cancer
16	2134230	2134230	C	T	*TSC2*	p.S1280L	TCGA-NH-A50T	Cancer
16	3807376	3807376	T	A	*CREBBP*	p.Y1204F	TCGA-24-1463	Cancer
16	3843590	3843590	T	G	*CREBBP*	p.Q338P	TCGA-AG-A016	Cancer
16	3860609	3860609	T	G	*CREBBP*	p.N324H	TCGA-AA-3529	Cancer
16	9857058	9857058	C	G	*GRIN2A*	p.C1448S	TCGA-A2-A04T	Cancer
16	9923491	9923491	A	G	*GRIN2A*	p.F599S	TCGA-19-1392	Cancer
16	9928022	9928022	C	G	*GRIN2A*	p.V573L	TCGA-29-1761	Cancer
16	14015889	14015889	A	G	*ERCC4*	p.E70G	TCGA-AA-3514	Cancer
16	14015897	14015897	A	G	*ERCC4*	p.I73V	TCGA-55-8094	Cancer
16	14029464	14029464	G	A	*ERCC4*	p.G559S	TCGA-13-0802	Cancer
16	50261778	50261778	C	T	*PAPD5*	p.P485L	TCGA-24-1469	Cancer
16	50820790	50820790	A	C	*CYLD*	p.K655N	TCGA-24-1553	Cancer
16	56862935	56862935	G	T	*NUP93*	p.A281S	TCGA-AG-3583	Cancer
16	67645910	67645910	A	C	*CTCF*	p.N280H	TCGA-12-1095	Cancer
16	67645910	67645910	A	C	*CTCF*	p.N280H	TCGA-AG-3584	Cancer
16	72821572	72821572	C	T	*ZFHX3*	p.E3535K	TCGA-CI-6620	Cancer
16	72827585	72827585	G	C	*ZFHX3*	p.A2999G	TCGA-24-1614	Cancer
16	72829588	72829588	T	G	*ZFHX3*	p.K2331N	TCGA-25-1625	Cancer
16	72831356	72831356	G	T	*ZFHX3*	p.A1742E	TCGA-ER-A42L	Cancer
16	72831357	72831357	C	G	*ZFHX3*	p.A1742P	TCGA-ER-A42L	Cancer
16	72923831	72923831	C	G	*ZFHX3*	p.G1083R	TCGA-24-1548	Cancer
16	72923831	72923831	C	G	*ZFHX3*	p.G1083R	TCGA-AA-3514	Cancer
16	72923831	72923831	C	G	*ZFHX3*	p.G1083R	TCGA-AA-A00K	Cancer
16	72923831	72923831	C	G	*ZFHX3*	p.G1083R	TCGA-AA-A00W	Cancer
16	72991497	72991497	G	C	*ZFHX3*	p.L850V	TCGA-A6-2678	Cancer
16	72992414	72992414	G	A	*ZFHX3*	p.S544L	TCGA-24-2254	Cancer
16	89871744	89871744	C	T	*FANCA*	p.C218Y	TCGA-AX-A06F	Cancer
16	89986377	89986378	-	A	*MC1R*	p.G238fs	TCGA-AO-A125	Cancer
16	89986432	89986432	C	T	*MC1R*	p.P256S	TCGA-A5-A2K2	Cancer

17	7217411	7217411	G	T	*GPS2*	p.L129I	TCGA-24-1548	Cancer
17	7976585	7976585	G	A	*ALOX12B*	p.P603S	TCGA-AX-A05U	Cancer
17	7984061	7984061	T	A	*ALOX12B*	p.N189Y	TCGA-AX-A3G6	Cancer
17	7989439	7989439	A	G	*ALOX12B*	p.Y83H	TCGA-AK-3461	Cancer
17	8108584	8108584	G	T	*AURKB*	p.P230T	TCGA-19-0955	Cancer
17	8108590	8108590	C	T	*AURKB*	p.G228R	TCGA-16-1055	Cancer
17	8110150	8110150	A	G	*AURKB*	p.L111S	TCGA-CU-A3QU	Cancer
17	15989715	15989715	C	T	*NCOR1*	p.E1020K	TCGA-AX-A06F	Cancer
17	16004574	16004574	C	T	*NCOR1*	p.E894K	TCGA-CI-6620	Cancer
17	16055265	16055265	G	C	*NCOR1*	p.I279M	TCGA-HT-A74O	Cancer
17	18188820	18188820	G	A	*TOP3A*	p.H538Y	TCGA-FI-A2EW	Cancer
17	18196026	18196026	C	G	*TOP3A*	p.R405P	TCGA-AA-A00R	Cancer
17	33446607	33446607	C	G	*RAD51D*	p.C9S	TCGA-AX-A3FV	Cancer
17	37671989	37671989	T	C	*CDK12*	p.I925T	TCGA-AG-3575	Cancer
17	37687081	37687081	T	C	*CDK12*	p.S1329P	TCGA-25-1623	Cancer
17	37866713	37866713	A	C	*ERBB2*	p.S264R	TCGA-AK-3428	Cancer
17	38508209	38508209	G	C	*RARA*	p.E189Q	TCGA-23-2645	Cancer
17	38508630	38508630	C	G	*RARA*	p.D242E	TCGA-24-1474	Cancer
17	40478145	40478145	T	A	*STAT3*	p.I452F	TCGA-AK-3431	Cancer
17	40478177	40478177	A	G	*STAT3*	p.F441S	TCGA-AA-A00K	Cancer
17	40489511	40489511	G	A	*STAT3*	p.Q247*	TCGA-AX-A05S	Cancer
17	40860072	40860072	G	T	*EZH1*	p.Q452K	TCGA-JV-A5VF	Cancer
17	40870492	40870492	T	A	*EZH1*	p.K234I	TCGA-24-1544	Cancer
17	40872469	40872469	T	C	*EZH1*	e4-2	TCGA-EE-A29E	Cancer
17	40880863	40880863	G	A	*EZH1*	p.Q33*	TCGA-EE-A29E	Cancer
17	41219694	41219694	C	A	*BRCA1*	p.A160S	TCGA-E9-A1NI	Cancer
17	41276054	41276054	T	G	*BRCA1*	p.K20N	TCGA-AG-3605	Cancer
17	41606916	41606916	A	C	*ETV4*	p.W323G	TCGA-25-1634	Cancer
17	41606916	41606916	A	C	*ETV4*	p.W323G	TCGA-A6-2676	Cancer
17	41610714	41610714	G	A	*ETV4*	p.A90V	TCGA-24-1103	Cancer
17	47684672	47684672	G	T	*SPOP*	p.Y259*	TCGA-09-1659	Cancer
17	47684672	47684672	G	T	*SPOP*	p.Y259*	TCGA-24-0968	Cancer
17	47684672	47684672	G	T	*SPOP*	p.Y259*	TCGA-AA-A00A	Cancer
17	47684672	47684672	G	T	*SPOP*	p.Y259*	TCGA-AK-3431	Cancer
17	47684672	47684672	G	T	*SPOP*	p.Y259*	TCGA-AK-3433	Cancer
17	47688785	47688785	C	T	*SPOP*	p.G172D	TCGA-12-1095	Cancer
17	47688789	47688789	A	C	*SPOP*	p.S171A	TCGA-AG-3575	Cancer
17	48453303	48453303	C	A	*EME1*	p.P245Q	TCGA-EY-A3L3	Cancer
17	48453480	48453480	G	T	*EME1*	p.E277*	TCGA-AG-3609	Cancer
17	56056645	56056645	T	G	*VEZF1*	p.K327Q	TCGA-DA-A3F8	Cancer
17	56435456	56435456	G	A	*RNF43*	p.R434W	TCGA-A1-A0SK	Cancer
17	56492733	56492733	A	C	*RNF43*	p.F69C	TCGA-AA-3514	Cancer
17	56492767	56492767	T	C	*RNF43*	p.T58A	TCGA-D1-A0ZP	Cancer
17	57134243	57134243	T	C	*TRIM37*	p.I398V	TCGA-DF-A2KS	Cancer
17	58711254	58711254	C	G	*PPM1D*	p.R248G	TCGA-19-2621	Cancer
17	58734152	58734152	C	G	*PPM1D*	p.Q404E	TCGA-29-1703	Cancer
17	58734323	58734323	T	G	*PPM1D*	p.*431G	TCGA-24-0970	Cancer
17	58740439	58740439	T	-	*PPM1D*	p.L450fs	TCGA-D3-A5GL	Cancer
17	58740498	58740498	C	G	*PPM1D*	p.S468*	TCGA-D1-A16D	Cancer
17	58740529	58740529	C	A	*PPM1D*	p.C478*	TCGA-29-2427	Cancer
17	58740626	58740626	A	C	*PPM1D*	p.K511Q	TCGA-36-1571	Cancer
17	58740653	58740653	C	T	*PPM1D*	p.Q520*	TCGA-A7-A0CH	Cancer
17	58740739	58740739	G	-	*PPM1D*	p.K549fs	TCGA-HD-A6HZ	Cancer
17	58740891	58740891	G	A	*PPM1D*	p.R599K	TCGA-12-0707	Cancer
17	59858270	59858270	T	G	*BRIP1*	p.K575N	TCGA-19-0955	Cancer
17	60558565	60558565	A	G	*TLK2*	p.K27E	TCGA-AG-A016	Cancer
17	63010560	63010560	G	C	*GNA13*	p.H222D	TCGA-09-1662	Cancer
17	63526198	63526198	C	T	*AXIN2*	p.D810N	TCGA-06-2569	Cancer

17	63533601	63533601	T	C	*AXIN2*	p.Y518C	TCGA-AO-A0J2	Cancer
17	63534452	63534452	G	A	*AXIN2*	p.R357C	TCGA-77-A5GH	Cancer
17	66519018	66519018	T	C	*PRKAR1A*	p.I100T	TCGA-24-1558	Cancer
17	78681694	78681694	C	A	*RPTOR*	p.C134*	TCGA-25-1632	Cancer
17	78923324	78923324	C	T	*RPTOR*	p.S1116L	TCGA-CI-6620	Cancer
18	10471569	10471569	C	A	*APCDD1*	p.Y95*	TCGA-BG-A18C	Cancer
18	10487664	10487664	G	A	*APCDD1*	p.A392T	TCGA-FG-7634	Cancer
18	12686307	12686307	T	C	*CEP76*	p.K359R	TCGA-AA-A00K	Cancer
18	31251726	31251726	A	C	*ASXL3*	p.N204T	TCGA-04-1530	Cancer
18	31323321	31323321	A	G	*ASXL3*	p.N1170S	TCGA-EJ-A65F	Cancer
18	39576644	39576644	G	T	*PIK3C3*	p.E312*	TCGA-B2-4101	Cancer
18	48593472	48593472	T	G	*SMAD4*	p.F408C	TCGA-AA-3534	Cancer
18	48604664	48604664	C	T	*SMAD4*	p.R496C	TCGA-CI-6620	Cancer
18	56246661	56246661	A	C	*ALPK2*	p.Y449*	TCGA-AA-A00K	Cancer
19	1440165	1440165	C	G	*RPS15*	p.H86Q	TCGA-EE-A29N	Cancer
19	10599903	10599903	C	T	*KEAP1*	p.G558E	TCGA-86-8359	Cancer
19	11123688	11123688	G	A	*SMARCA4*	p.E780K	TCGA-25-1318	Cancer
19	11123688	11123688	G	A	*SMARCA4*	p.E780K	TCGA-25-1625	Cancer
19	11132584	11132584	A	G	*SMARCA4*	p.K934E	TCGA-AK-3461	Cancer
19	11138599	11138599	C	T	*SMARCA4*	p.R1119C	TCGA-CI-6620	Cancer
19	11169476	11169476	A	G	*SMARCA4*	p.N1548D	TCGA-AR-A0U3	Cancer
19	15273355	15273355	C	T	*NOTCH3*	p.W1945*	TCGA-G9-6339	Cancer
19	15276741	15276741	C	T	*NOTCH3*	p.A1842T	TCGA-AX-A06J	Cancer
19	15276834	15276834	C	T	*NOTCH3*	p.D1811N	TCGA-AF-2689	Cancer
19	15290896	15290896	C	G	*NOTCH3*	p.G1105A	TCGA-AQ-A54N	Cancer
19	15302421	15302421	C	T	*NOTCH3*	p.A284T	TCGA-BR-A4QL	Cancer
19	18273902	18273902	A	G	*PIK3R2*	p.Q412R	TCGA-23-1021	Cancer
19	30312631	30312631	A	C	*CCNE1*	p.E189D	TCGA-13-0889	Cancer
19	30314634	30314634	A	C	*CCNE1*	p.T380P	TCGA-AX-A1CC	Cancer
19	31770488	31770488	C	T	*TSHZ3*	p.E71K	TCGA-CI-6620	Cancer
19	36503951	36503951	T	C	*ALKBH6*	p.K60R	TCGA-13-0717	Cancer
19	36503951	36503951	T	C	*ALKBH6*	p.K60R	TCGA-AG-3608	Cancer
19	40746019	40746019	T	C	*AKT2*	e4-2	TCGA-AA-3821	Cancer
19	40771132	40771132	G	A	*AKT2*	p.R15C	TCGA-13-0792	Cancer
19	41754431	41754431	G	A	*AXL*	p.G249D	TCGA-97-8174	Cancer
19	41758288	41758288	T	C	*AXL*	p.F314L	TCGA-25-1634	Cancer
19	41758862	41758862	A	C	*AXL*	p.D371A	TCGA-AO-A129	Cancer
19	41762429	41762429	C	A	*AXL*	p.Y435*	TCGA-04-1331	Cancer
19	41762429	41762429	C	A	*AXL*	p.Y435*	TCGA-BH-A0E0	Cancer
19	45858928	45858928	T	A	*ERCC2*	p.D513V	TCGA-AP-A0LL	Cancer
19	45871892	45871892	G	T	*ERCC2*	p.P119H	TCGA-EE-A20I	Cancer
19	45873458	45873458	G	T	*ERCC2*	p.P13Q	TCGA-D3-A3C3	Cancer
19	45917294	45917294	T	C	*ERCC1*	e7-2	TCGA-15-0742	Cancer
19	45920082	45920082	C	G	*ERCC1*	p.W200S	TCGA-AA-3514	Cancer
19	52714694	52714694	T	A	*PPP2R1A*	p.F151Y	TCGA-AG-3598	Cancer
20	9561389	9561389	A	T	*PAK7*	p.Y131*	TCGA-L9-A443	Cancer
20	30370135	30370135	G	T	*TPX2*	p.A416S	TCGA-25-1625	Cancer
20	39725965	39725965	T	C	*TOP1*	p.F279S	TCGA-23-1028	Cancer
20	39726949	39726949	G	A	*TOP1*	p.R316Q	TCGA-AX-A05S	Cancer
20	49509761	49509761	T	C	*ADNP*	p.Y497C	TCGA-EE-A2GP	Cancer
20	57484421	57484421	G	A	*GNAS*	p.R844H	TCGA-67-4679	Cancer
20	57484421	57484421	G	A	*GNAS*	p.R844H	TCGA-D6-A4Z9	Cancer
20	57484421	57484421	G	A	*GNAS*	p.R844H	TCGA-EY-A1GU	Cancer
20	57484739	57484739	A	G	*GNAS*	p.D883G	TCGA-CU-A72E	Cancer
20	61537287	61537287	C	T	*DIDO1*	p.V514I	TCGA-AX-A064	Cancer
20	61541236	61541236	C	A	*DIDO1*	p.D326Y	TCGA-AK-3433	Cancer
20	62331866	62331866	C	T	*ARFRP1*	p.G179S	TCGA-62-8398	Cancer
21	30693746	30693746	C	T	*BACH1*	p.R49W	TCGA-AA-3518	Cancer

21	39763607	39763607	T	G	*ERG*	p.K289T	TCGA-14-0783	Cancer
22	21288357	21288357	C	A	*CRKL*	p.P201Q	TCGA-AA-A01G	Cancer
22	22127223	22127223	A	G	*MAPK1*	p.I302T	TCGA-FS-A1ZC	Cancer
22	22142564	22142564	G	T	*MAPK1*	p.P280T	TCGA-36-1568	Cancer
22	22142564	22142564	G	T	*MAPK1*	p.P280T	TCGA-36-1578	Cancer
22	22142591	22142591	T	G	*MAPK1*	p.N271H	TCGA-23-1028	Cancer
22	22142591	22142591	T	G	*MAPK1*	p.N271H	TCGA-36-1578	Cancer
22	29106018	29106019	-	A	*CHEK2*	p.E318fs	TCGA-UF-A7JF	Cancer
22	29695627	29695627	G	T	*EWSR1*	p.G578*	TCGA-BR-A4J2	Cancer
22	29696113	29696113	G	T	*EWSR1*	p.G650C	TCGA-12-0707	Cancer
22	29696120	29696120	A	G	*EWSR1*	p.H652R	TCGA-AA-3514	Cancer
22	30050686	30050686	T	C	*NF2*	p.L163S	TCGA-AA-3514	Cancer
22	41531828	41531828	A	G	*EP300*	p.M514V	TCGA-42-2593	Cancer
22	41568555	41568555	C	T	*EP300*	p.P1502L	TCGA-13-0717	Cancer
22	41573538	41573538	G	T	*EP300*	p.Q1941H	TCGA-AG-A016	Cancer
22	42522940	42522940	C	T	*CYP2D6*	p.E410K	TCGA-BG-A0M6	Cancer
22	42526656	42526657	-	A	*CYP2D6*	p.L47fs	TCGA-C8-A1HO	Cancer
X	15833799	15833799	G	A	*ZRSR2*	e8-1	TCGA-EJ-7330	Cancer
X	15836711	15836711	T	A	*ZRSR2*	p.V258D	TCGA-D9-A1JW	Cancer
X	15836766	15836766	G	C	*ZRSR2*	e9+1	TCGA-D7-8574	Cancer
X	15840902	15840902	T	C	*ZRSR2*	p.L329P	TCGA-CD-5798	Cancer
X	19389468	19389468	C	T	*MAP3K15*	p.D929N	TCGA-A1-A0SO	Cancer
X	19418771	19418771	A	C	*MAP3K15*	p.L451V	TCGA-12-1095	Cancer
X	41205637	41205637	A	G	*DDX3X*	p.K491E	TCGA-CQ-6218	Cancer
X	41205659	41205659	C	A	*DDX3X*	p.T498K	TCGA-AO-A129	Cancer
X	47028716	47028716	G	A	*RBM10*	p.G7D	TCGA-AA-3514	Cancer
X	48549524	48549524	G	T	*WAS*	p.E494*	TCGA-A6-2676	Cancer
X	48650377	48650377	C	T	*GATA1*	p.S116F	TCGA-CI-6620	Cancer
X	53222394	53222394	T	C	*KDM5C*	p.K1480E	TCGA-24-1548	Cancer
X	53222723	53222723	C	T	*KDM5C*	p.E1405K	TCGA-DM-A1HB	Cancer
X	53225129	53225129	C	A	*KDM5C*	p.R1030L	TCGA-24-1548	Cancer
X	66941763	66941763	C	A	*AR*	p.Q803K	TCGA-B2-4101	Cancer
X	70603001	70603001	A	C	*TAF1*	p.K644T	TCGA-13-0793	Cancer
X	70603001	70603001	A	C	*TAF1*	p.K644T	TCGA-36-1578	Cancer
X	70683864	70683864	G	C	*TAF1*	p.A1897P	TCGA-AG-A016	Cancer
X	76912120	76912120	A	G	*ATRX*	p.S1382P	TCGA-AG-A016	Cancer
X	76938973	76938973	G	T	*ATRX*	p.P592H	TCGA-25-1329	Cancer
X	79948443	79948443	C	T	*BRWD3*	p.V1087I	TCGA-AJ-A2QO	Cancer
X	79955530	79955530	C	A	*BRWD3*	p.V957F	TCGA-12-1098	Cancer
X	79973218	79973218	G	T	*BRWD3*	p.N695K	TCGA-AA-3549	Cancer
X	100611209	100611209	T	A	*BTK*	p.K466M	TCGA-36-1568	Cancer
X	100613379	100613379	G	T	*BTK*	p.Q341K	TCGA-AG-3598	Cancer
X	102755278	102755278	A	C	*RAB40A*	p.V136G	TCGA-AR-A0TZ	Cancer
X	106221416	106221416	C	T	*MORC4*	p.R317K	TCGA-CI-6620	Cancer
X	106224209	106224209	A	G	*MORC4*	p.I283T	TCGA-13-0761	Cancer
X	110406153	110406153	C	G	*PAK3*	p.T196R	TCGA-AF-2691	Cancer
X	110439743	110439743	T	G	*PAK3*	p.W464G	TCGA-25-1324	Cancer
X	119666287	119666287	A	C	*CUL4B*	p.M828R	TCGA-A6-2678	Cancer
X	119673152	119673152	T	G	*CUL4B*	p.K589T	TCGA-13-1408	Cancer
X	129147539	129147539	C	G	*BCORL1*	p.S264*	TCGA-EJ-5526	Cancer
X	129148664	129148664	G	A	*BCORL1*	p.R639H	TCGA-DD-A4NJ	Cancer
X	129171468	129171468	C	T	*BCORL1*	p.H1552Y	TCGA-A7-A3J0	Cancer
X	132795817	132795817	T	C	*GPC3*	p.M475V	TCGA-13-0717	Cancer
X	132795817	132795817	T	C	*GPC3*	p.M475V	TCGA-AA-3561	Cancer
X	132887636	132887636	G	T	*GPC3*	p.S302Y	TCGA-26-5135	Cancer
X	153993198	153993198	A	G	*DKC1*	p.K14R	TCGA-JW-A5VK	Cancer
X	153994556	153994556	G	A	*DKC1*	p.R110Q	TCGA-K4-A3WU	Cancer

Supplementary Table 4.3: 26 significantly mutated genes identified in normal blood samples

Gene	Indels	SNVs	Total Mutation	Mutation pMbp	P-value FCPT	P-value LRT	P-value CT	FDR FCPT	FDR LRT	FDR CT
DNMT3A	5	86	91	8.40	0	0	0	0	0	0
JAK2	0	17	17	1.68	0	0	6.76E-21	5.71E-11	0	8.15E-17
ASXL1	3	18	21	1.58	0	0	3.54E-20	4.15E-11	0	2.85E-16
TET2	5	17	22	0.80	0	1.15E-12	3.93E-17	6.01E-09	5.57E-09	2.37E-13
IDH2	0	10	10	2.96	0	0	4.23E-15	4.70E-07	0	2.04E-11
PPM1D	2	8	10	2.41	0	6.21E-10	4.49E-12	8.78E-05	1.87E-06	1.81E-08
SF3B1	0	17	17	1.42	0	1.28E-10	1.25E-11	8.78E-05	4.40E-07	4.31E-08
ZNF318	0	16	16	1.05	0	2.37E-07	4.73E-09	0.0076	0.0003	1.42E-05
MYH4	1	13	14	0.99	0	1.15E-06	5.82E-09	0.0090	0.0010	1.56E-05
PCMTD1	0	7	7	1.17	0	5.71E-09	1.88E-08	0.0400	1.38E-05	4.54E-05
PTN	0	6	6	4.80	0	7.34E-08	5.11E-08	0.1549	0.0001	0.0001
FRG1B	0	8	8	2.50	0	6.06E-08	6.45E-08	0.1080	0.0001	0.0001
GNB1	0	7	7	2.30	0	3.50E-09	8.49E-08	0.1094	9.38E-06	0.0002
EMID2	0	6	6	1.98	0	1.24E-08	3.07E-07	0.2707	2.49E-05	0.0005
TIE1	0	10	10	0.52	0	1.00E-10	3.97E-07	0.3141	4.02E-07	0.0006
EPHB2	0	10	10	0.88	0	8.23E-08	8.06E-07	0.2707	0.0001	0.0011
FRG1	0	8	8	1.85	0	7.49E-06	1.14E-06	0.5440	0.0042	0.0015
ASH1L	0	16	16	0.71	0	3.08E-06	1.23E-06	0.2707	0.0022	0.0016
ZKSCAN4	0	6	6	1.47	0	6.40E-09	2.69E-06	0.9941	1.40E-05	0.0031
TMC1	0	8	8	1.06	0	5.50E-06	2.59E-06	0.6894	0.0035	0.0031
EPPK1	0	10	10	0.78	0	2.94E-07	2.96E-06	0.6894	0.0003	0.0031
SPOP	0	7	7	1.53	0	1.29E-06	2.95E-06	0.9420	0.0011	0.0031
SLC9A4	0	7	7	1.22	0	1.35E-07	4.65E-06	1.0000	0.0002	0.0047
MORC2	0	9	9	0.85	0	1.33E-07	6.99E-06	1.0000	0.0002	0.0066
GBAS	0	7	7	1.75	0	8.27E-06	9.38E-06	1.0000	0.0045	0.0084
TMEM196	2	1	3	3.02	0	2.38E-06	1.10E-05	1.0000	0.0018	0.0091

Supplementary Table 4.4: 58 cancer-associated genes overlapped with blood specific somatic CNVs

Chr	CNV Type	Sample	Genotype:QualityScore		Gene Position	Gene Name
			Blood	Control		
1:120351324-120461178	Deletion	TCGA-D3-A2JA	1:66	0:94,65	1:120457927-120612022	*NOTCH2*
1:120460286-120502127	Deletion	TCGA-AO-A03O	1:99	0:99,78	1:120457927-120612022	*NOTCH2*
1:120461027-120471837	Deletion	TCGA-GN-A268	1:99	0:84,99	1:120457927-120612022	*NOTCH2*
1:120464857-120480635	Deletion	TCGA-EE-A3AH	1:99	0:92,92	1:120457927-120612022	*NOTCH2*
1:120464857-120480635	Deletion	TCGA-ER-A19B	1:99	0:97,80	1:120457927-120612022	*NOTCH2*
1:120483176-120484379	Deletion	TCGA-D3-A1Q5	1:71	0:99,69	1:120457927-120612022	*NOTCH2*
1:120506195-120509114	Deletion	TCGA-99-8028	1:99	0:99,74	1:120457927-120612022	*NOTCH2*
1:120506195-120509114	Deletion	TCGA-DK-A3IK	1:64	0:92,84	1:120457927-120612022	*NOTCH2*
1:120510698-120548213	Deletion	TCGA-D3-A2JC	1:90	0:61,82	1:120457927-120612022	*NOTCH2*
2:47635538-47705660	Deletion	TCGA-AG-3906	1:76	0:72,97	2:47630329-47789452	*MSH2*
2:48025748-48036407	Deletion	TCGA-13-1498	1:75	0:95,90	2:48010307-48034001	*MSH6*
2:48025748-48036407	Deletion	TCGA-A8-A08Z	1:64	0:68,76	2:48010307-48034001	*MSH6*
2:48025748-48046219	Deletion	TCGA-13-1405	1:99	0:86,86	2:48010307-48034001	*MSH6*
2:48025748-48046219	Deletion	TCGA-A8-A06P	1:66	0:87,83	2:48010307-48034001	*MSH6*
2:48025748-48049444	Deletion	TCGA-13-1509	1:99	0:70,97	2:48010307-48034001	*MSH6*
2:202122840-202123107	Deletion	TCGA-41-3392	1:98	0:63,99	2:202098186-202151319	*CASP8*
2:202136237-202153506	Deletion	TCGA-A5-A0GJ	1:98	0:60,77	2:202098186-202151319	*CASP8*
2:202136237-202154579	Deletion	TCGA-BH-A0HX	1:99	0:63,95	2:202098186-202151319	*CASP8*
2:202136237-202173975	Deletion	TCGA-BH-A0HY	1:98	0:63,95	2:202098186-202151319	*CASP8*
2:202137359-202150042	Deletion	TCGA-A8-A07I	1:98	0:70,98	2:202098186-202151319	*CASP8*
2:202137359-202172347	Deletion	TCGA-13-0760	1:89	0:75,88	2:202098186-202151319	*CASP8*
2:202151180-202195247	Deletion	TCGA-B6-A0I8	1:89	0:87,69	2:202098186-202151319	*CASP8*
5:56152425-56189509	Deletion	TCGA-A6-2678	1:97	0:82,76	5:56111399-56189509	*MAP3K1*
7:151876917-151884934	Deletion	TCGA-A2-A0CX	1:81	0:83,93	7:151833915-152132873	*MLL3*
7:151884345-152027829	Deletion	TCGA-23-1809	1:93	0:70,91	7:151833915-152132873	*MLL3*
7:151919656-151945707	Deletion	TCGA-G4-6315	1:86	0:82,91	7:151833915-152132873	*MLL3*
7:151970788-152027829	Deletion	TCGA-19-1385	1:76	0:87,77	7:151833915-152132873	*MLL3*
9:98221880-98718297	Deletion	TCGA-97-8179	1:93	0:71,87	9:98209192-98279104	*PTCH1*
9:135762754-135778176	Deletion	TCGA-CN-5366	1:93	0:89,90	9:135771620-135820010	*TSC1*
12:49436342-49460317	Deletion	TCGA-09-1669	1:85	0:86,83	12:49415561-49449109	*MLL2*
12:49444667-49448811	Deletion	TCGA-25-1314	1:94	0:95,71	12:49415561-49449109	*MLL2*
12:49444667-49449109	Deletion	TCGA-13-0794	1:94	0:66,84	12:49415561-49449109	*MLL2*
12:49446345-49449109	Deletion	TCGA-36-1580	1:71	0:82,94	12:49415561-49449109	*MLL2*
12:49446696-49460486	Deletion	TCGA-BH-A0GZ	1:94	0:92,84	12:49415561-49449109	*MLL2*
12:52369047-52374985	Deletion	TCGA-13-0764	1:94	0:95,71	12:52345526-52387896	*ACVR1B*
12:52385645-52407987	Deletion	TCGA-09-1662	1:94	0:94,79	12:52345526-52387896	*ACVR1B*
13:32928996-32945239	Deletion	TCGA-20-0996	1:89	0:88,85	13:32890596-32972909	*BRCA2*
13:48985523-49037973	Deletion	TCGA-09-1668	1:87	0:94,75	13:48878047-49054785	*RB1*
13:49033822-49281996	Deletion	TCGA-DK-A6B0	1:95	0:85,80	13:48878047-49054785	*RB1*
16:3786035-3860782	Deletion	TCGA-A5-A0RA	1:97	0:84,79	16:3777717-3929919	*CREBBP*
17:15965417-16220149	Deletion	TCGA-UE-A6QU	1:98	0:68,93	17:15935608-16118717	*NCOR1*
17:41243450-41256975	Deletion	TCGA-C5-A1MQ	1:98	0:69,98	17:41197693-41277204	*BRCA1*
18:45368196-45423129	Deletion	TCGA-BH-A18I	1:98	0:90,90	18:45368196-45423129	*SMAD2*
19:11105502-11123790	Deletion	TCGA-13-0764	1:99	0:97,76	19:11094826-11175879	*SMARCA4*
19:11105502-11123790	Deletion	TCGA-13-0794	1:94	0:88,93	19:11094826-11175879	*SMARCA4*
20:30816064-31294564	Deletion	TCGA-DM-A1D6	1:66	0:94,64	20:30946577-31025143	*ASXL1*
1:43806056-43818445	Duplication	TCGA-A8-A08L	2:76	0:99,74	1:43803518-43818445	*MPL*
1:65301077-65309900	Duplication	TCGA-13-0794	2:99	0:80,99	1:65300243-65351949	*JAK1*
1:226173001-226259182	Duplication	TCGA-A6-2678	2:99	0:91,86	1:226252051-226259182	*H3F3A*
2:178081229-178097313	Duplication	TCGA-BH-A0GZ	2:62	0:97,60	2:178092632-178175735	*NFE2L2*
2:198260778-198262842	Duplication	TCGA-AX-A05S	2:78	0:84,98	2:198257025-198299725	*SF3B1*
3:41268697-41277336	Duplication	TCGA-09-0366	2:67	0:92,94	3:41265558-41301589	*CTNNB1*
3:41274830-41280847	Duplication	TCGA-AX-A06F	2:98	0:80,96	3:41265558-41301589	*CTNNB1*

3:41274830-41504746	Duplication	TCGA-13-0794	2:85	0:85,87	3:41265558-41301589	*CTNNB1*
3:178525125-178947232	Duplication	TCGA-13-1497	2:96	0:75,87	3:178916612-178952154	*PIK3CA*
3:178742803-178947232	Duplication	TCGA-AX-A0J1	2:96	0:83,69	3:178916612-178952154	*PIK3CA*
3:178742803-178957854	Duplication	TCGA-BG-A0VT	2:98	0:69,93	3:178916612-178952154	*PIK3CA*
3:178742803-178968780	Duplication	TCGA-AG-3906	2:96	0:79,86	3:178916612-178952154	*PIK3CA*
3:178745432-178957854	Duplication	TCGA-AX-A0IW	2:97	0:79,77	3:178916612-178952154	*PIK3CA*
4:1719877-1804881	Duplication	TCGA-B6-A402	2:97	0:96,65	4:1795660-1809017	*FGFR3*
4:55133454-55133910	Duplication	TCGA-AX-A06F	2:97	.:62,97	4:55106218-55161441	*PDGFRA*
4:55561676-55604725	Duplication	TCGA-30-1891	2:97	0:93,68	4:55524180-55604725	*KIT*
4:55561676-55604725	Duplication	TCGA-32-2616	2:97	0:92,81	4:55524180-55604725	*KIT*
5:149420309-149441233	Duplication	TCGA-B6-A0IB	2:74	0:81,94	5:149433630-149465992	*CSF1R*
7:2946270-2968334	Duplication	TCGA-09-1661	2:95	0:88,89	7:2946270-2998142	*CARD11*
7:54610422-55755619	Duplication	TCGA-D1-A3DG	2:94	0:88,76	7:55086822-55273312	*EGFR*
7:54820102-55273312	Duplication	TCGA-DU-7290	2:94	0:88,74	7:55086822-55273312	*EGFR*
7:55209977-55242515	Duplication	TCGA-A8-A07U	2:71	0:87,78	7:55086822-55273312	*EGFR*
7:55209977-55268108	Duplication	TCGA-B5-A3S1	2:94	0:85,97	7:55086822-55273312	*EGFR*
7:55209977-55496142	Duplication	TCGA-D7-8576	2:94	0:89,88	7:55086822-55273312	*EGFR*
7:55214297-55268108	Duplication	TCGA-AP-A1E0	2:60	0:89,72	7:55086822-55273312	*EGFR*
7:55214297-55863787	Duplication	TCGA-AJ-A3NC	2:94	0:66,74	7:55086822-55273312	*EGFR*
7:116339123-116436180	Duplication	TCGA-09-1668	2:84	0:85,89	7:116335809-116436180	*MET*
7:116380002-116403324	Duplication	TCGA-S9-A6WE	2:94	0:79,91	7:116335809-116436180	*MET*
9:5078304-5090913	Duplication	TCGA-23-1029	2:62	0:73,93	9:4985031-5126976	*JAK2*
9:5078304-5090913	Duplication	TCGA-24-1842	2:94	0:94,67	9:4985031-5126976	*JAK2*
9:5078304-5090913	Duplication	TCGA-30-1855	2:94	0:92,85	9:4985031-5126976	*JAK2*
9:5080227-5090913	Duplication	TCGA-23-1111	2:63	0:84,93	9:4985031-5126976	*JAK2*
9:5123002-5185604	Duplication	TCGA-23-2081	2:94	0:60,93	9:4985031-5126976	*JAK2*
10:123239563-123247629	Duplication	TCGA-13-0794	2:70	0:77,95	10:123237876-123357588	*FGFR2*
13:28608217-28608546	Duplication	TCGA-AX-A064	2:83	0:86,95	13:28578082-28674649	*FLT3*
13:28608217-28608546	Duplication	TCGA-AX-A06B	2:95	0:76,95	13:28578082-28674649	*FLT3*
13:28610070-28624361	Duplication	TCGA-24-1843	2:95	0:92,66	13:28578082-28674649	*FLT3*
14:81606021-81610699	Duplication	TCGA-13-0794	2:68	0:91,95	14:81421331-81610699	*TSHR*
15:90610368-90634878	Duplication	TCGA-HU-A4GY	2:97	0:83,69	15:90627496-90645624	*IDH2*
17:37783644-38080458	Duplication	TCGA-62-8398	2:98	0:65,85	17:37855811-37886681	*ERBB2*
17:37814656-37868302	Duplication	TCGA-55-A493	2:98	0:82,79	17:37855811-37886681	*ERBB2*
17:37880977-38027880	Duplication	TCGA-DK-A2I2	2:98	0:63,91	17:37855811-37886681	*ERBB2*
17:74729373-74739540	Duplication	TCGA-DA-A1IA	2:82	0:86,97	17:74732241-74733244	*SRSF2*
19:3114940-3121319	Duplication	TCGA-CR-7389	2:97	0:98,84	19:3094645-3121434	*GNA11*
19:10251455-10291563	Duplication	TCGA-76-4928	2:99	0:86,80	19:10244025-10311561	*DNMT1*
19:52671309-52705289	Duplication	TCGA-AX-A05U	2:99	0:85,99	19:52693348-52729236	*PPP2R1A*
20:54940141-58491628	Duplication	TCGA-DM-A1D6	2:99	0:82,61	20:57415160-57486249	*GNAS*

Supplementary Table 4.5: Rare *PPM1D* truncation mutations in ExAC database

Chr	Start	ID	Reference	Alternate	Mutation	Annotation	Allele Frequency
17	58734163	.	T	A	p.Cys407Ter	Nonsense	0.00000824
17	58740432	.	C	G	p.Ser446Ter	Nonsense	0.00001648
17	58740438	.	T	-	p.Leu450Ter	Frame-Shift	0.000008238
17	58740467	.	C	T	p.Arg458Ter	Nonsense	0.00001648
17	58740498	.	C	G	p.Ser468Ter	Nonsense	0.00000824
17	58740529	rs146477590	C	A	p.Cys478Ter	Nonsense	0.00003297
17	58740532	.	-	A	p.Ala481SerfsTer8	Frame-Shift	0.000008243
17	58740532	.	A	-	p.Ala481ProfsTer2	Frame-Shift	0.00001649
17	58740601	.	-	A	p.Ser503IlefsTer25	Frame-Shift	0.000008252
17	58740623	.	-	A	p.Asn512LysfsTer16	Frame-Shift	0.000008259
17	58740623	.	A	-	p.Asn512IlefsTer2	Frame-Shift	0.00002478
17	58740653	.	C	T	p.Gln520Ter	Nonsense	0.000008261
17	58740674	.	G	T	p.Glu527Ter	Nonsense	0.000008262
17	58740676	.	AG	-	p.Arg528AsnfsTer7	Frame-Shift	0.000008261
17	58740713	rs138670032	G	T	p.Glu540Ter	Nonsense	0.000008256
17	58740749	.	C	T	p.Arg552Ter	Nonsense	0.000008249
17	58740897	.	TGTT	-	p.Cys603PhefsTer21	Frame-Shift	0.00001681

Curriculum Vitae

Mingchao Xie

Education

Ph.D. in Computational and Systems Biology 08/2012 – present

Washington University in St Louis, MO

M.S. in Biochemistry and Molecular Biology 08/2008 – 08/2010

Pennsylvania State University (University Park), PA

M.S. in Biology 09/2004 – 07/2007

Tsinghua University, Beijing, China

B.S. in Biological Science 09/2000 – 07/2004

Shandong University, Jinan, China

Selected Publication

1. Lu C*, **Xie M***, Wendl MC*, Wang J*, et al. (2015) Patterns and functional implications of rare germline variants across 12 cancer types. ***Nature Communications*** 6 *Contributed equally
2. **Xie M***, Lu C*, Wang J*, McLellan MD, Johnson KJ, Wendl MC, *et al*. (2014) Age-related cancer mutations associated with clonal hematopoietic expansion. ***Nature Medicine*** 20 (12), 1472-1478 *Contributed equally
3. **Xie M***, Hong C*, Zhang B*, Lowdon RF, Xing X, Li D, Zhou X, *et al*. (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. ***Nature Genetics*** 45 (7), 836–841 *Contributed equally
4. **Xie M**, Ai C, Jin X, Liu S, Tao S, Li Z, Wang Z. (2007) Cloning and characterization of chicken SPATA4 gene and analysis of its specific expression. ***Molecular and Cellular Biochemistry*** 306 (1-2), 79-85
5. Jones KB, Barrott JJ, **Xie M**, Haldar M, Jin H, Zhu JF, *et al*. (2016). The impact of chromosomal translocation locus and fusion oncogene coding sequence in synovial sarcomagenesis. ***Oncogene*** doi: 10.1038/onc.2016.38
6. Ye K, Wang J, Jayasinghe R, Lameijer EW, McMichael JF, Ning J, McLellan MD**, Xie M**, Cao S, *et al*. (2016) Systematic discovery of complex insertions and deletions in human cancers. ***Nature Medicine***. 22(1): 97-104
7. **Roadmap Epigenomics Consortium**. (2015) Integrative analysis of 111 reference human epigenomes. ***Nature*** 518 (7539), 317-330
8. Niu B, Ye K, Zhang Q, Lu C, **Xie M**, McLellan MD, Wendl MC, Ding L. (2013) MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. ***Bioinformatics*** 30 (7), 1015-1016
9. Zighelboim I, Mutch DG, Knapp A, Ding L, **Xie M**, Cohn DE, Goodfellow PJ. (2013) High frequency strand slippage mutations in CTCF in MSI-positive endometrial cancers. ***Human Mutation*** 35 (1), 63-65

10. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, **Xie M**, Zhang Q, *et al*. (2013) Mutational landscape and significance across 12 major cancer types. ***Nature*** 502 (7471), 333-339
11. Stevens M, Cheng JB, Li D, **Xie M**, Hong C, Maire CL, Ligon KL, Hirst M, Marra MA, Costello JF, Wang T. (2013) Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. ***Genome Research*** 23 (9), 1541-1553
12. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C, Musselman M, **Xie M**, *et al*. (2012) Comparative epigenomic annotation of regulatory DNA. ***Cell*** 149 (6), 1381-1392
13. Zhou X, Maricque B, **Xie M**, Li D, Sundaram V, Martin EA, Koebbe BC, Nielsen C, Hirst M, Farnham P, *et al*. (2011) The human epigenome browser at Washington University. ***Nature Methods*** 8 (12), 989-990