Journal of
Emerging Investigators

DyGS: a dynamic gene searching algorithm for cancer detection

Jenica Wang[1*], Xue Gong[2]

[1]Henry M. Gunn High School, Palo Alto, CA

*Current institution: Johns Hopkins University, Baltimore, MD

[2]Department of Urology, Stanford, CA

Summary

Cancer is a lethal disease and ranks as the world's second-most prevalent cause of death. So far, biopsy is the most common method conducted to determine the progression of tumors. However, this traditional technique is invasive and cannot be used repeatedly. On the other hand, liquid biopsies, also known as blood-sample tests, have become more and more promising. One critical step in liquid biopsy is to create an effective gene panel that covers the maximum number of cancer cases possible. The hypotheses of this study are that both common genes and genes that are mutated in specific cancer types will be present in the selected gene lists for the 12 cancer types, and some types may have longer lists than others. In this study, we developed a novel dynamic gene-searching algorithm called Dynamic Gene Search (DyGS) to create a gene panel for each of the 12 cancers with the highest annual incidence and death rate. This algorithm generated lists of genes for the 12 cancer types, ranging from 12 to 153 genes, with a median of 47 genes. Notably, many of the genes included in the panel can be targeted by various drugs. Therefore, the gene panels designed by the DyGS algorithm can be used as actionable drug targets for cancer treatment as well as biomarkers in liquid biopsy to help identify the early stages of cancer.

Received: January 16, 2017; **Accepted:** June 5, 2017; **Published:** June 5, 2018

Introduction

Cancer is a life-threatening disease and is the second-leading cause of death in the United States. The estimated number of new cases in 2016 was more than 1.68 million, and the estimated number of deaths was more than 595,000 (1). Even though different types of cancer are similar, in that they involve abnormal cell growth and can invade other parts of the body, a process known as metastasis, they are heterogeneous both at the cellular and molecular levels. Different cancer patients harbor unique mutations, which make cancer detection and treatment more challenging (2, 3). Tissue biopsy is the conventional method used to determine the progression of tumors, but this technique is invasive. On the other hand, liquid biopsy is non-invasive and is a more promising strategy for early cancer detection. Circulating tumor DNA (ctDNA) is released by dead tumor cells and flows into the bloodstream, which can be used for early cancer detection and cancer monitoring (4). Some ctDNA is released from circulating tumor cells (CTCs), which are tumor cells that separate from the tumor and enter the bloodstream (5). Besides CTCs, ctDNA can also be derived from tumor-derived exosomes (6). Recent research conducted at Stanford University has developed a method called CAPP-Seq that evaluates ctDNA to profile lung cancer and has the potential to reduce costs and increase detection sensitivity, which refers to how strong the stimulus must be for the technique to detect it (7). CAPP-Seq, also known as cancer personalized profiling by deep sequencing, is a sequencing method used to quantify cell-free DNA released from tumor cells that entered the bloodstream (8). Currently, more than 40 commercial companies are racing to develop liquid biopsies; however, the gene panels generated have little overlap. Gene panels contain genes associated with the disease under study, cancer in this case, and are utilized to analyze mutations in a given sample. This function benefits cancer diagnosis due to the disease's known association with mutations. Companies such as Guardant Health, Foundation Medicine, and Grail each have their own, specialized gene panels, which can lead to conflicting and even confusing results. To generate consistent and comparable results as well as to reduce the cost of the test, an optimization algorithm based on publicly available data is in great demand.

In addition to liquid biopsy, another technique that is less invasive than traditional biopsy is medical imaging. This method, including CT and MRI scans, can identify "internal" tumor types like pancreatic cancer. Both imaging and liquid biopsies are noninvasive techniques that can aid in diagnosing cancer, identifying its severity, and monitoring its progression. Because medical imaging requires the use of radiation to produce detectable

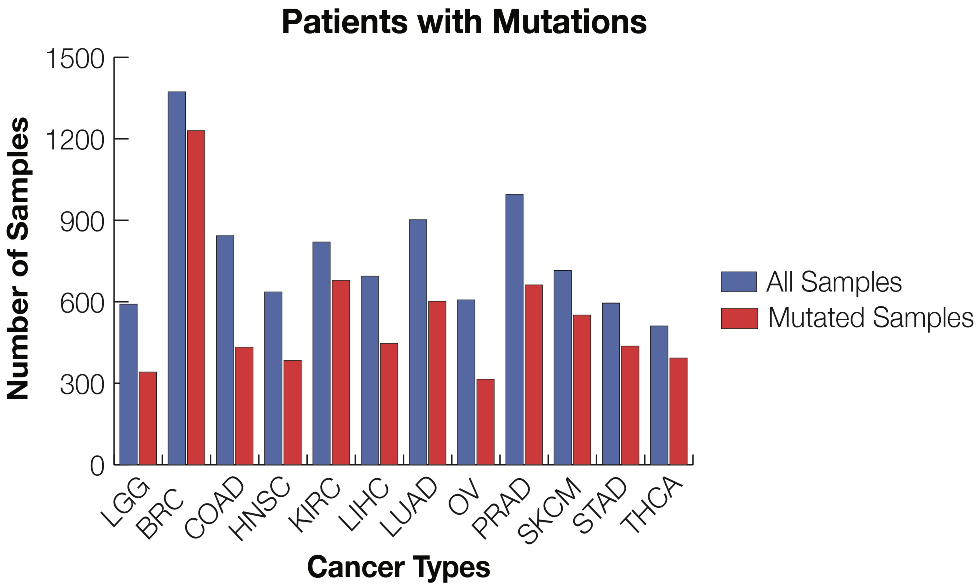


Figure 1. Bar graph displaying the total number of patients (blue bars) and the number of patients with mutations (red bars) in each of the 12 cancer types. LGG for Brain Lower Grade Glioma, BRC for Breast Invasive Carcinoma, COAD for Colorectal Adenocarcinoma, HNSC for Head and Neck Squamous Cell Carcinoma, KIRC for Kidney Clear Cell Carcinoma, LIHC for Liver Hepatocellular Carcinoma, LUAD for Lung Adenocarcinoma, OV for Ovarian Serous Cystadenocarcinoma, PRAD for Prostate Adenocarcinoma, SKCM for Skin Cutaneous Melanoma, STAD for Stomach Adenocarcinoma, and THCA for Papillary Thyroid Carcinoma.

signals, there is a higher chance this technique may not detect the tumor at an early stage due to the tumor's miniscule size. On the other hand, liquid biopsy is highly sensitive and may be more effective at detecting earlier stages of cancer.

Due to the emerging abundance of high-throughput sequencing data (e.g., exon sequencing) for different types of human cancer patients, we can utilize meta-analysis to identify cancer biomarkers using databases such as cBioPortal (9) and The Cancer Genome Atlas (TCGA) (10), each of which has unique features and functions. cBioPortal specializes in gathering different cancer datasets from various projects into a more accessible source for researchers, and TCGA focuses on genomic changes in 33 types of cancer. The high-throughput data of thousands of patients have been generated and deposited in these data repositories. These databases have played a great role in pushing cancer research forward, especially in regards to early cancer detection and prognosis (11). By making the data broadly accessible, more researchers can now use these resources to analyze data and discover new biomarkers that may influence cancer initiation and progression. Identifying what these biomarkers are may lead to the development of therapeutic targets for cancer. On January 12, 2016, President Barack Obama announced the initiation of Cancer Moonshot, a program led by Vice President Joe Biden that is intended to detect cancer at an early stage, prevent cancer, and make more therapies available to patients (12). The large-scale attention cancer research has been given indicates that this big data will revolutionize cancer diagnosis and treatment. To provide answers for the following three hypotheses, we developed an algorithm named the Dynamic Gene Search (DyGS) algorithm:

Hypothesis 1. Some of the well-known, common genes associated with cancer, such as TP53, BRAF, and the RAS gene family, will be present in all or most of the gene lists generated since many of these genes have been reported in many cancer types.

Hypothesis 2. The gene lists of the 12 different cancer types will somewhat overlap due to the hypothesized presence of the common cancer-associated genes. However, some genes may be mutated in specific cancer types, and the prediction is that less than half of the genes in the lists will be like this.

Hypothesis 3. The length of the gene lists for some cancers may be longer than those of others. For example, breast-invasive carcinoma (BRC) and prostate adenocarcinoma (PRAD) may have longer lists than brain lower grade glioma (LGG) because BRC is genetically more heterogeneous, and the genes mutated in PRAD tend to have lower frequencies across patients.

The DyGS algorithm designed innovative gene panels to cover the largest number of patients with the smallest number of gene mutations to facilitate early cancer detection. On average, the algorithm selected 3.5% of mutated genes for inclusion in the gene panels. Therefore, DyGS is a highly effective method to generate gene panels for liquid biopsy. Moreover, the top genes for each of the 12 gene panels included some commonly mutated genes such as TP53, IDH1, APC, BRAF, and MUC16, all of which occur in more than 50% of samples. Other top genes had low frequencies, including NRAS and HRAS, whose frequencies were below 5%. Many of the genes in the gene panels were mutated in only specific cancer types, but some of them were common, such as TP53, BRAF, KRAS, PIK3CA, and EGFR. Also, the gene panels included many druggable genes as annotated in the Drug-Gene Interaction Database (DGIdb) (13). An average median of 40% of each gene panel was druggable. Thus, the 12 gene panels generated by the DyGS algorithm can be used not only as biomarkers for early cancer detection, but also as therapeutic targets for cancer treatment. This study is innovative because data was generated by the state-of-the-art technique exon sequencing, abundant samples from thousands of patients were used, and novel biomarkers with significant drug potential were identified. Therefore, based on these findings, clinicians can use the biomarkers for early cancer detection and personalized treatment, pharmaceutical companies can develop drugs based on the biomarkers identified, and researchers can understand more about cancer biology by conducting further studies.

Results

Dynamic searching algorithm (DyGS) for 12 cancer types

Dynamic Gene Search (DyGS) is a computational algorithm designed to dynamically search for and create an effective mutation gene panel that covers the maximum number of cancer cases possible. Maximum coverage can be achieved by selecting mutations with the highest frequencies in the samples, obtained through downloading cBioPortal datasets, and renewing the matrix, formed through combining the datasets using customized Perl script, until all samples with mutations are covered. In the case where two mutations occur at the same frequency, mutations that are annotated as clinically actionable will be selected. Using this method, the most commonly occurring—and druggable—mutations are covered. DyGS was implemented using both the Perl and R programming languages to analyze the matrices generated, and this algorithm was applied to the 12 cancer types with the highest number of new cases and death rates.

From data collected from various sources, the average number of patient samples is more than 700 **(Figure 1, blue bars)**. As indicated by the red bars, the average of the mutated-sample-to-total-sample ratios is about 70% **(Figure 1, red bars)** with the non-mutated samples resulting from Among these cancer types, colorectal adenocarcinoma (COAD) has the smallest ratio of 51%, while BRC has the largest ratio of 90%.

After applying DyGS to the mutated samples to each cancer type, we generated a list of the most comprehensive genes, whose mutations collectively cover all the patient samples. As shown in Figure 2, the average number of mutated genes for each cancer type is about 1800. For example, BRC has 2343 mutated genes, and stomach adenocarcinoma (STAD) has 2409 mutated genes. Because of the enormous number of mutated genes for each tumor type, it is impractical to include all of them in a gene panel. Therefore, DyGS can effectively reduce the number of genes in the panel. For instance, only about 3.5% of mutated genes are needed to cover all the patients. For COAD, only 12 out of 1781 mutated genes, or about 0.7% of genes, were needed to effectively cover the patients **(Figure 2)**.

The top ten genes with their mutation frequencies across samples are represented in **Table 1**. For each cancer type, there is always a mutated gene that occurs most often. For example, TP53 is the most commonly mutated gene in several cancers, including ovarian (OV) at 87.0% frequency, head and neck (HNSC) at 70.3%, STAD at 48.1%, lung (LUAD) at 45.8%, liver (LIHC) at 32.2%, BRC at 31.9%, and PRAD at 19.6%. For the other cancer types, IDH1 (82.7%), APC (72.5%), BRAF (61.6%), MUC16 (58.3%), and VHL (49.0%) were the most mutated gene for LGG, COAD, papillary thyroid carcinoma (THCA), skin cutaneous melanoma (SKCM), and kidney clear cell carcinoma (KIRC), respectively. It is notable that most of the other mutated genes in the top ten lists occur at very low frequencies of less than 5%. For instance, the second-largest frequency mutation in THCA, NRAS, occurs in 8.7% of the patients, and the

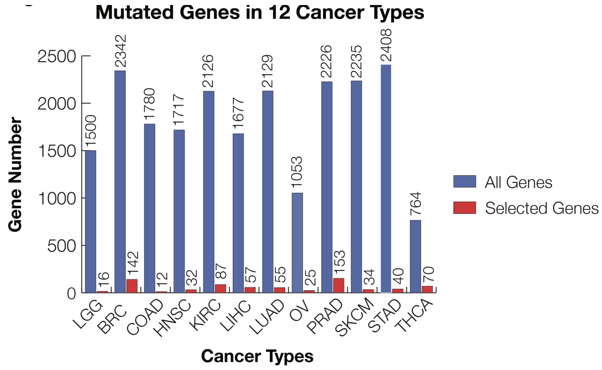
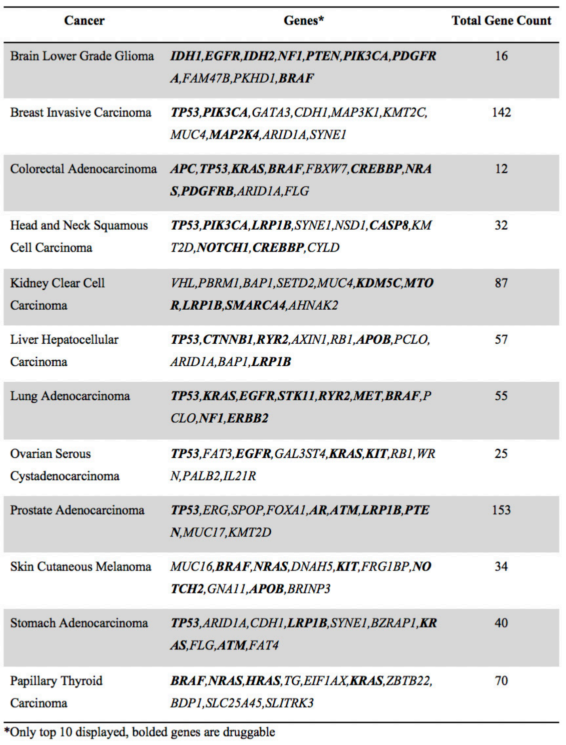


Figure 2. Bar graph displaying the total number of genes (blue bars) and the number of genes selected using the DyGS algorithm (red bars) in each of the 12 cancer types.



Cancer	Genes*	Total Gene Count
Brain Lower Grade Glioma	***IDH1***,***EGFR***,***IDH2***,***NF1***,***PTEN***,***PIK3CA***,***PDGFRA***,*FAM47B*,*PKHD1*,***BRAF***	16
Breast Invasive Carcinoma	***TP53***,***PIK3CA***,*GATA3*,*CDH1*,*MAP3K1*,*KMT2C*,*MUC4*,***MAP2K4***,*ARID1A*,*SYNE1*	142
Colorectal Adenocarcinoma	***APC***,***TP53***,***KRAS***,***BRAF***,*FBXW7*,***CREBBP***,***NRAS***,***PDGFRB***,*ARID1A*,*FLG*	12
Head and Neck Squamous Cell Carcinoma	***TP53***,***PIK3CA***,***LRP1B***,*SYNE1*,*NSD1*,***CASP8***,*KMT2D*,***NOTCH1***,***CREBBP***,*CYLD*	32
Kidney Clear Cell Carcinoma	*VHL*,*PBRM1*,*BAP1*,*SETD2*,*MUC4*,***KDM5C***,***MTOR***,***LRP1B***,***SMARCA4***,*AHNAK2*	87
Liver Hepatocellular Carcinoma	***TP53***,***CTNNB1***,***RYR2***,*AXIN1*,*RB1*,***APOB***,*PCLO*,*ARID1A*,*BAP1*,***LRP1B***	57
Lung Adenocarcinoma	***TP53***,***KRAS***,***EGFR***,***STK11***,***RYR2***,***MET***,***BRAF***,*PCLO*,***NF1***,***ERBB2***	55
Ovarian Serous Cystadenocarcinoma	***TP53***,*FAT3*,***EGFR***,*GAL3ST4*,***KRAS***,***KIT***,*RB1*,*WRN*,*PALB2*,*IL21R*	25
Prostate Adenocarcinoma	***TP53***,*ERG*,*SPOP*,*FOXA1*,***AR***,***ATM***,***LRP1B***,***PTEN***,*MUC17*,*KMT2D*	153
Skin Cutaneous Melanoma	*MUC16*,***BRAF***,***NRAS***,*DNAH5*,***KIT***,*FRG1BP*,***NOTCH2***,*GNA11*,***APOB***,*BRINP3*	34
Stomach Adenocarcinoma	***TP53***,*ARID1A*,*CDH1*,***LRP1B***,*SYNE1*,*BZRAP1*,***KRAS***,*FLG*,***ATM***,*FAT4*	40
Papillary Thyroid Carcinoma	***BRAF***,***NRAS***,***HRAS***,*TG*,*EIF1AX*,***KRAS***,*ZBTB22*,*BDP1*,*SLC25A45*,*SLITRK3*	70

*Only top 10 displayed, bolded genes are druggable

Table 1. Top 10 DyGS-selected genes for each of the 12 cancer types.

third-largest frequency gene HRAS occurs in 3.6% of the patients. Starting from the fourth gene, the mutation frequency falls below 3%.

Mutated genes with cancer-type specificity

To explore cancer-type specificities for mutated genes, unsupervised clustering of the 497 mutated genes across 12 cancer types was performed **(Figure 3)**. Most of the mutated genes were mutated in specific cancer types, represented by the horizontal red bands. For example, about 20% of the mutated genes were exclusive to PRAD and BRC each. To name a few, SPINT3, CXCR3, EWSR, SMO, PPR2, and AMER1 are specific to BRC, while PDK, MAP3K4, OR6C76, MLLT4, EPHB2, and ETV1 are specific to PRAD (14-20). On the other hand, there are some mutated genes that are common to many tumor types, including TP53, BRAF, KRAS, PIK3CA, EGFR, ARID1A, PTEN, and ATM. The unique distribution of mutated genes in each cancer type confirms the approach implemented in this study by investigating each type separately.

Drug-gene networks and clinical implications

A large proportion of the mutated genes are druggable, according to the Drug-Gene Interaction Database (DGIdb). The detailed associations are presented in **Table 1** and **Figure 4**. For example, from this study, LUAD, the most lethal cancer in the United States, has 11 druggable genes: EGFR, BRAF, ERBB2, PIK3CG, AR, ATM, ABL1, MET, NTRK1, KDR, and GRIN2A. Erlotinib, Gefitinib, and Afatinib can target EGFR **(Figure 4)**. The former two are first-line drugs used for treating lung cancer with the EGFR mutation (21). Afatinib has shown clinical benefit in lung cancer patients with brain metastases (22). For the other ten genes, there are 25 drugs that can be used to target these genes. For example, AR can be targeted by five drugs: Nilutamide, Flutamide, Bicalutamide, Fluoxymesterone, and Oxandrolone. Similarly, BRAF can be targeted by four drugs: Sorafenib, Dabrafenib, Vemurafenib, and Trametinib.

These results demonstrate that the gene panels generated from DyGS can be used not only for cancer diagnosis, such as liquid biopsy, but also for cancer treatment since many of these genes included in the panel can be targeted by various drugs. Two potential clinical applications are that non-cancer-related drugs may be used to treat cancer and cancer-type-specific drugs may be used to treat patients of other cancer types.

Discussion

Recent technological advancements in DNA sequencing allow for the further development of personalized medicine by identifying tumorigenic and metastatic gene mutations and pathways and extending therapeutic target opportunities (23). Due to the wave of new technologies available for cancer patient care, tremendous amounts of cancer mutation data have been accumulated on both driver and passenger mutations (24). Driver mutations are those defined to promote cancer development and are therefore sought after as therapeutic targets, while passenger mutations are those that do not directly stimulate cancer initiation and/or progression. Among the cancer treatment options available, liquid biopsy, or blood sample tests, have recently emerged as the most promising strategy due to its more convenient and safer clinical use and lower cost (25). By using effective gene panels, it is unnecessary to sequence the whole human genome to conduct cancer diagnosis; instead, one can evaluate only the biomarkers in the gene panels. Some concerns to liquid biopsy are: 1) the inconsistency due to the abundance of companies developing their own gene panels and 2) the high cost used to accommodate for the enormous number of genes—often several hundred—included in the gene panels.

To address these problems, we developed the DyGS algorithm to optimize gene panels for 12 common cancer types by taking advantage of publicly available high-throughput genomics data. The novel Dynamic

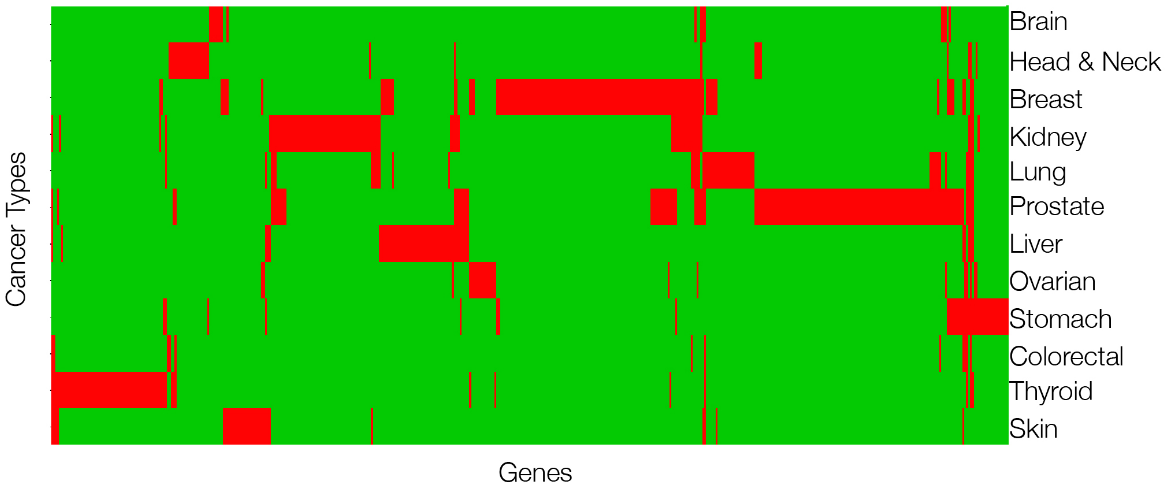


Figure 3. Heatmap representing the presence (red) or absence (green) of DyGS-selected mutated genes in the 12 cancer types. Columns depict genes, and rows depict cancer types.

Gene Search (DyGS) algorithm was designed to cover the maximum number of patients using the minimum number of gene mutations. The 12 gene panels the DyGS algorithm selected used only about 3.5% of the original gene mutation pool, while covering every patient sample. Some of these panels include common, highly mutated genes that have a mutation frequency higher than 50%, while others include genes that have low frequencies of 5% or less. Although these gene mutations do not occur frequently, they were selected because the patients they cover do not have mutations in other, more frequently mutated genes. Many of the genes the DyGS algorithm selected were mutated in specific cancer types; in fact, about 20% of some of the cancer gene panels, such as for BRC, were specific to individual cancer types. An essential future direction of this study is clinical application. According to the Drug-Gene Interaction Database (DGIdb), the gene panels include many genes that are druggable. About 40% of each gene panel is druggable, which indicates that the DyGS-generated gene panels can be used for early cancer detection as well as therapeutic targets in treatment methods.

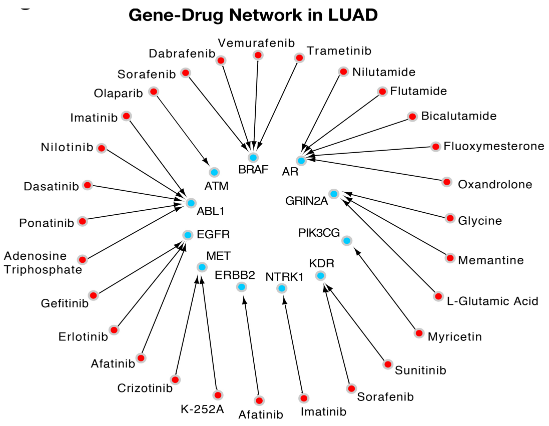


Figure 4. Network displaying the relationships between DyGS-selected genes (blue dots) and drugs targeting them (red dots) retrieved from Drug-Gene Interaction Database (DGIdb) for lung adenocarcinoma (LUAD).

The DyGS algorithm can be applied in many areas of medical research. First, DyGS can be used to analyze additional cancer types other than the 12 investigated in this study. Second, other complicated genetic diseases can use this algorithm, including heart disease, diabetes, and obesity. Third, the DyGS algorithm can be applied to other genetic alterations, such as copy-number variations, epigenetic changes, and gene fusion. However, there are still a few limitations to this algorithm. First, because of the nature of the algorithm and its characteristic of covering all the samples, some cancers' gene panels are larger than others, making them more expensive. For example, PRAD has 153 genes, while COAD only has 12 genes. Second, generating an optimal gene panel for all 12 cancer types is difficult since many of the genes included in the panels occur more frequently in specific cancers, and therefore, the common gene panel would have to be large to cover a significant portion of the patients, which leads to greater costs. Third, the current DyGS algorithm includes missense mutations, which is provided by cBioPortal, but does not take into account other types of mutations, such as insertion/deletion and genomic amplification.

Materials and Methods

Data download for 12 cancer types

cBioPortal for Cancer Genomics is a database originally developed at Memorial Sloan Kettering Cancer Center (MSK) and hosted by the Center for Molecular Oncology at MSK. Currently, it is the most comprehensive data source for cancer genomics data and includes The Cancer Genome Atlas (TCGA). TCGA is a publically available data source that contains comprehensive genomic changes in 33 types of cancer. For this study, mutation datasets of the 12 cancer types with the highest number of new cases and death rates (LGG, BRC, COAD, HNSC, KIRC, LIHC, LUAD, OV, PRAD, SKCM, STAD, THCA) were downloaded from cBioPortal (http://www.cbioportal.org/). These mutations were detected by exome sequencing, which only involves the exons, or the expressed regions, of the DNA. This technique is much less costly than sequencing the entire human genome. For each cancer type, multiple datasets were collected and combined with customized Perl script as a matrix.

Data processing for mutation data

Because the data were downloaded from different sources, it was processed prior to further analysis. For each tumor type, all the samples were combined into a sample file and all the genes were combined into a gene file, both of which were achieved by using the Linux command “cat”. To prepare the data for analysis, (0,1)-matrices were generated from each of the datasets by customized Perl script. The 0 in the matrix represents the absence of the gene mutation in the sample, while 1 indicates the presence of the gene mutation in the sample. In addition, a number matrix of the original files was created. The number matrix denotes the combined number of times the gene mutation appears across samples of all the original files.

Dynamic gene searching algorithm

In order to design a gene panel that covers the greatest number of patients and the smallest number of mutated genes, R and Perl programming languages were employed and used to analyze the matrices generated. We designed an algorithm named Dynamic Gene Search (DyGS) to achieve this goal and applied it to each cancer type.

First, for each missense mutation on the gene, a frequency was calculated by counting the number of samples with the mutated gene divided by the total number of samples. In order to obtain the shortest gene panel possible, the gene list was sorted in descending order of the frequency calculated, and the gene mutation with the maximum patient coverage was selected and added to the gene panel. Second, the matrix was renewed by deleting the selected gene along with the samples it covered. The renewed matrix served as the input for the next round. Third, this process was itinerated until all the samples with mutations were covered. If there was a tie in any step, a tiebreak rule was applied: when a tie occurred, mutations with clinically actionable or functional annotations were selected instead of those that were not. Clinically actionable annotations indicate that the gene is druggable according to the Drug-Gene Interaction Database (DGIdb). Functional annotations represent the genes curated in the Cancer Gene Census database, which is a catalog that includes genes with cancer-causing mutations. So far, it includes about 600 genes. By applying all these conditions, the gene panel output from the program can be considered the optimal gene panel that has the highest patient coverage and smallest number of mutations. The codes used for this study are deposited in GitHub under the repository name “DyGS” at https://github.com/jwang00/DyGS.git.

Drug-gene relationship retrieval

The Drug-Gene Interaction Database (DGIdb) is a web interface that identifies known and potential drug-gene relationships. For each entry, different evidence levels were given by the number of sources and the number of PubMed articles. A higher value in these two areas represents more reliable relationships. In this study, in order to achieve more stringent relationships, a cutoff of three sources was set.

Data analysis and presentation

The bar graphs presented in this study were constructed in Adobe Illustrator Version 18.1.1, the heatmap was originally generated in Cluster 3.0 for Mac OS X (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#ctv) and Java TreeView 1.1.6 (https://sourceforge.net/projects/jtreeview/files/jtreeview/1.1.6/) and edited in Illustrator, and the drug-gene network was created in Cytoscape Version 3.4.0 (http://www.cytoscape.org/download.php) and edited in Illustrator. Cluster and TreeView are software programs that originally analyze and visualize data from DNA microarray experiments but extended to other genomic datasets. Cluster organizes and analyzes the data, while TreeView presents the organized data in a heatmap. Cytoscape is a software platform used for visualizing interaction networks and pathways and integrating them with annotations.

Acknowledgements

I am grateful to Henry M. Gunn High School’s Advanced Authentic Research program (supported by the Palo Alto Unified School District) for the wonderful

research opportunity.

References

1. Society, A.C., Cancer Facts & Figures. 2014.
2. Meacham, C.E. and Morrison, S.J. “Tumour heterogeneity and cancer cell plasticity”. *Nature*, vol. 501, no. 7467, 2013, pp. 328-337.
3. Patel, A.P., et al., “Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma”. *Science*, vol. 433, no. 6190, 2014, pp. 1396-1401.
4. Ignatiadis, M., Lee, M., and Jeffrey S.S. “Circulating Tumor Cells and Circulating Tumor DNA: Challenges and Opportunities on the Path to Clinical Utility”. *Clin Cancer Res*, vol. 21. no. 21, 2015, pp. 4786-4800.
5. Alix-Panabieres, C. and Pantel, K. “Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy”. *Cancer Discov*, vol 6., no. 5., 2016, pp. 479-491.
6. Thakur, B.K., et al. “Double-stranded DNA in exosomes: a novel biomarker in cancer detection”. *Cell Res*, vol. 24, no. 6, 2014, pp. 766-769.
7. Newman, A.M., et al. “An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage”. *Nat Med*, vol. 20, no. 5, 2014, pp. 548-554.
8. Bratman, S.V., et al. “Potential clinical utility of ultrasensitive circulating tumor DNA detection with CAPP-Seq”. *Expert Rev Mol Diagn*, vol. 15, no. 6, 2015, pp. 715-719.
9. Gao, J., et al. “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. *Sci Signal*, vol. 6, no. 269, 2013, p. pl1.
10. Cancer Genome Atlas Research. “The Molecular Taxonomy of Primary Prostate Cancer”. *Cell*, vol. 163, no. 4, 2015, pp. 1011-1025.
11. Tomczak, K., Czerwinska, P., and Wiznerowicz, M. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.” *Contemp Oncol (Pozn)*, vol. 19, no. 1A, 2015, pp. A68-A77.
12. Kaiser, J. and Couzin-Frankel, J. “BIOMEDICAL RESEARCH. Biden seeks clear course for his cancer moonshot”. *Science*, vol. 351, no. 6271, 2016, pp. 325-326.
13. Wagner, A.H., et al. “DGIdb 2.0: mining clinically relevant drug-gene interactions”. *Nucleic Acids Res*, vol. 44, no. D1, 2016, pp. D1036-1044.
14. Alexandrov, L.B., et al. “Signatures of mutational processes in human cancer”. *Nature,* vol. 500, no. 7463, 2013, pp. 415-421.
15. Hodgkinson, A., Chen, Y, and Eyre-Walker, A. “The large-scale distribution of somatic mutations in cancer genomes”. *Hum Mutat*, vol. 33, no. 1, 2012, pp. 136-143.
16. Kan, Z., et al. “Diverse somatic mutation patterns and pathway alterations in human cancers”. *Nature*, vol. 466, no. 7308, 2010, pp. 869-873.
17. Kandoth, C., et al. “Mutational landscape and significance across 12 major cancer types”. *Nature*, vol. 502, no. 7471, 2013, pp. 333-339.
18. Lawrence, M.S., et al. “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. *Nature*, vol. 499, no. 7457, 2013, pp. 214-218.
19. Martincorena, I. and Campbell, P.J. “Somatic mutation in cancer and normal cells”. *Science*, vol. 349, no. 6255, 2015, pp. 1483-1489.
20. Schaefer, M.H. and Serrano, L. “Cell type-specific properties and environment shape tissue specificity of cancer genes”. *Sci Rep*, vol. 6, 2016, p. 20707.
21. Greenhalgh, J., et al. “Erlotinib and gefitinib for treating non-small cell lung cancer that has progressed following prior chemotherapy (review of NICE technology appraisals 162 and 175): a systematic review and economic evaluation”. *Health Technol Assess*, vol. 19, no. 47, 2015, pp. 1-134.
22. Schuler, M., et al. “First-Line Afatinib versus Chemotherapy in Patients with Non-Small Cell Lung Cancer and Common Epidermal Growth Factor Receptor Gene Mutations and Brain Metastases”. *J Thorac Oncol*, vol. 11, no. 3, 2016, pp. 380-390.
23. Rubio-Perez, C., et al. “In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities”. *Cancer Cell*, vol. 27, no. 3, 2015, pp. 382-396.
24. Forbes, S.A., et al. “COSMIC: exploring the world’s knowledge of somatic mutations in human cancer”. *Nucleic Acids Res*, vol. 43, 2015, pp. D805-811.
25. Heitzer, E., Ulz, P., and Geigl, J.B. “Circulating tumor DNA as a liquid biopsy for cancer”. *Clin Chem*, vol. 61, no. 1, 2015, pp. 112-123.