Gene Panel Design for Early Cancer Detection

Jenica Wang[1], Dr. Xue Gong[2]
Henry M. Gunn High School[1], Stanford University[2]

INTRODUCTION

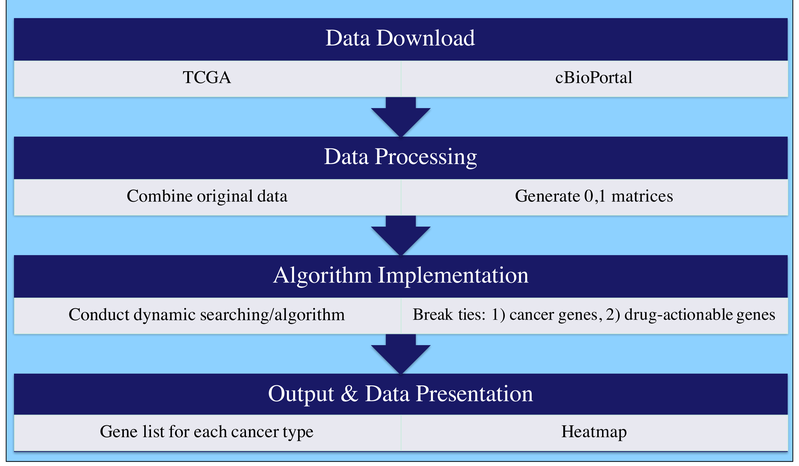
Cancer has become a popular research topic as more and more people die of cancer and the situation involving it becomes graver. A gene panel is set up so that the mutations found in the patients are listed out, compared, and selected to represent the largest number of patients. By finding the mutations correlated to cancer, we can facilitate early cancer detection, potentially saving lives. The development of liquid biopsy and the CAPP-Seq method are more effective than traditional techniques, and through the use of public data from The Cancer Genome Atlas (TCGA) and cBioPortal, we can expand on previous research and create optimal gene panels. DNA can be sequenced in just a few drops of blood through liquid biopsy, and the CAPP-Seq method (Cancer Personalized Profiling by deep Sequencing) can be implemented to more successfully find mutations. Since the traditional technique of (tissue) biopsy involves many risks and factors that may influence the result, most scientists and clinicians have now adopted the newer approaches.

How can optimal gene panels be designed for early cancer detection? Our goal in this study was to generate effective gene panels that cover the smallest number of mutations and the greatest number of patients.

We addressed the cancers with the highest number of new cases and death rates, including lung, breast, prostate, colon, ovarian, kidney, liver, skin, stomach, brain, thyroid, as well as head and neck cancer. For each cancer type, we aimed to design a cancer-specific gene panel.

METHODS

1. Downloaded data of 12 cancers (lung, breast, prostate, colorectal, ovarian, kidney, liver, skin, stomach, brain, thyroid, head & neck) from TCGA/Firehose (https://tcga-data.nci.nih.gov/tcga/, https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata) and cBioPortal (http://www.cbioportal.org/)
2. Processed downloaded data
   - TCGA/Firehose
     - Combined original data into one file for unique genes and one file for samples
   - cBioPortal
     - Combined original data into one file for unique genes and one file for samples
     - Combined all original data
3. Generated matrices
   - TCGA/Firehose
     - 0,1 matrix: 0,1 matrix for all TCGA data
   - cBioPortal
     - Frequency matrix of individual original files put together: gene frequency matrix
     - 0,1 matrix: 0,1 matrix for all cBioPortal data
     - Number matrix of original files: matrix denoting the combined # of times the gene mutation appears across samples of all original files
4. Dynamic searching/algorithm
   - Employed Perl and R programming languages
   - Arranged genes by cover sizes from largest to smallest → select largest
   - Use tiebreaking code set to different conditions to choose specific gene from tie
5. Data Presentation
   - Heatmap & gene-drug network
   - Table of gene lists for each cancer type with identification of cancer genes and druggable genes



RESULTS

1. Genes Related to Cancers (Effective Gene Panel)

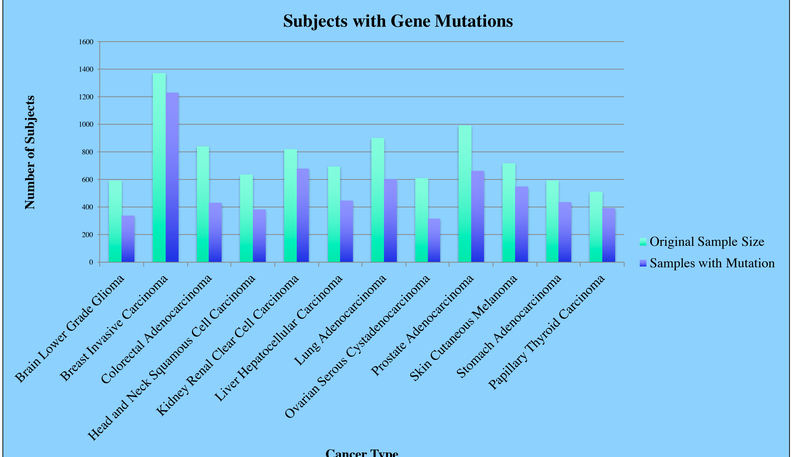
- Bold => documented in COSMIC (Catalogue of Somatic Mutations in Cancer)
- Italicized => documented in DGIdb (The Drug Gene Interaction Database) as druggable
- "Total Gene Count" => number of genes related to cancer types determined through Dynamic Searching Algorithm (DSA)

For complete list of genes, see supplementary materials.

Cancer	Genes (First 20)	Total Gene Count
Brain	***IDH1***,***EGFR***,***IDH2***,***NF1***,***PTEN***,***PIK3CA***,***PDGFRA***,FAM47B,PKHD1,***BRAF***,***NOTCH2***,***ROS1***,**PTPN11**,**MTCP1**,**CDH1**,*G6PD*	16
Breast	***TP53***,***PIK3CA***,**GATA3**,**CDH1**,**MAP3K1**,**KMT2C**,MUC4,***MAP2K4***,**ARID1A**,SYNE1,GOLGA6L2,FAT3,TTN,FLG,MUC12,**SF3B1**,RELN,DMD,***ATM***,***RUNX1***...	142
Colorectal	***APC***,***TP53***,***KRAS***,***BRAF***,**FBXW7**,***CREBBP***,***NRAS***,***PDGFRB***,**ARID1A**,FLG,***CSF3R***,***SMARCA4***	12
Head and Neck	***TP53***,***PIK3CA***,*LRP1B*,SYNE1,**NSD1**,***CASP8***,**KMT2D**,***NOTCH1***,***CREBBP***,**CYLD**,***ALK***,**FBXW7**,**MED12**,USH2A,PTPRT,MUC17,PLEC,PRMT5,***CDKN1B***,***FAS***...	32
Kidney	**VHL**,**PBRM1**,**BAP1**,**SETD2**,MUC4,***KDM5C***,*MTOR*,*LRP1B*,***SMARCA4***,AHNAK2,TCEB1,***TP53***,***PIK3CA***,***PTEN***,*LRRK2*,***APC***,**KMT2C**,PABPC1,OBSCN,ZNF717...	87
Liver	***TP53***,***CTNNB1***,*RYR2*,**AXIN1**,**RB1**,*APOB*,PCLO,**ARID1A**,**BAP1**,*LRP1B*,**TSC2**,**KMT2D**,MKI67,FLG,**SLC45A3**,**KMT2A**,DMD,COL11A1,C17ORF97,***ATM***...	57
Lung	***TP53***,***KRAS***,***EGFR***,***STK11***,*RYR2*,***MET***,***BRAF***,PCLO,***NF1***,***ERBB2***,***ATM***,***GRIN2A***,*ERBB4*,**SETD2**,*PTPRD*,KIAA1109,FRG1BP,***ABL1***,***SMARCA4***,***PTEN***...	55
Ovarian	***TP53***,FAT3,***EGFR***,GAL3ST4,***KRAS***,***KIT***,**RB1**,**WRN**,**PALB2**,**IL21R**,ARSF,COL22A1,MYH4,ABCA9,***GRIN2A***,***NTRK3***,***NOTCH1***,***ITK***,***BRCA1***,**FNBP1**...	25
Prostate	***TP53***,**ERG**,**SPOP**,**FOXA1**,*AR*,***ATM***,*LRP1B*,***PTEN***,MUC17,**KMT2D**,AHNAK2,SYNE1,RP1,***BRCA2***,DCHS2,***PIK3CA***,**KMT2C**,***EP300***,**ETV1**,FAT3...	153
Skin	MUC16,***BRAF***,***NRAS***,DNAH5,***KIT***,FRG1BP,***NOTCH2***,**GNA11**,*APOB*,BRINP3,OR4M1,DSCAM,***NF1***,***RAC1***,**GNAQ**,COL4A5,PCLO,PASD1,FLG,C10ORF120...	34
Stomach	***TP53***,**ARID1A**,**CDH1**,*LRP1B*,SYNE1,BZRAP1,***KRAS***,FLG,***ATM***,FAT4,***CTNNB1***,NAV3,MUC6,***BRD4***,**KMT2C**,*MTNR1B*,RP1,FANCM,HMCN1,**GATA3**...	40
Thyroid	***BRAF***,***NRAS***,***HRAS***,TG,EIF1AX,***KRAS***,ZBTB22,BDP1,SLC25A45,SLITRK3,***ATM***,***TP53***,***DNMT3A***,*MMP24*,VWA2,TENM2,OTUD4,MSI1,POTEE,GLI2...	70

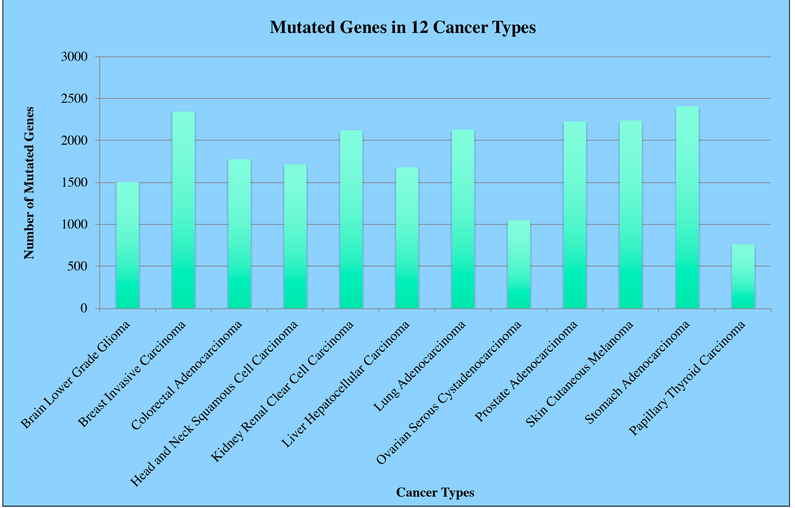
2. Gene Mutations in Subjects

- Samples with gene mutations < original sample size



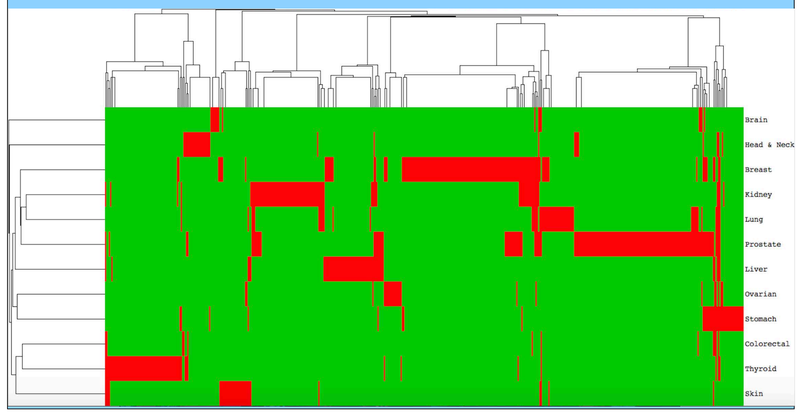
3. Total Number of Mutated Genes for Each Cancer Type

- Obtained number of mutated genes in each cancer type from downloaded data through computer programming



4. Frequency of Gene Mutations in Cancers as Represented in a Heatmap

- **Red** => presence of mutated gene in cancer
- **Green** => absence of mutated gene in cancer
- Some mutated genes are more cancer-specific than others (bands of red)



SUMMARY / CONCLUSIONS

We have designed a gene panel for the 12 cancers we investigated that covers the greatest number of mutated genes and the least number of patients. These genes can then be used as biomarkers in liquid biopsy to help identify the early stages of cancer. Another real-world application includes detecting circulating tumor DNA (ctDNA) in patients' blood to match gene mutations included in the patients to our gene lists, which acts as a potential indicator of cancer.

In clinical applications, digital PCR or targeted sequencing is employed to capture mutations in genes. One consideration is the cost of the assay, which needs to reach a compromise between the number of genes and patients covered and the pricing of the tests. Therefore, further optimization is needed to get the best trade-off between these two factors. This can be achieved by possibly lowering the percentage of coverage (ex. From 100% to 90% or even lower).

We have found a list of mutated genes for each of the 12 cancers addressed in this study. The varying lengths of these lists indicate the presence of heterogeneity in each cancer. For example, in breast cancer, there were 142 genes selected to cover the patients. This is consistent with the biology of the cancer since many subtypes of this cancer are involved, including Luminal A, Luminal B, triple negative/basal-like, and HER2. On the other hand, colorectal cancer's list only consists of 12 genes, suggesting that this cancer is more homogeneous than other cancer types investigated in this study.

In this study, we used dynamic searching and algorithm to find common and rare genes. We can accomplish this by arranging the mutation rates of genes in each cancer. For example, the mutation rates of TP53 in each of the cancer types are: ovarian (274/315=0.8698412) > head and neck (270/384=0.703125) > stomach (210/437=0.4805491) > lung (276/602=0.4584717) > liver (144/447=0.3221476) > breast (392/1230=0.3186991) > prostate (130/662=0.1963746) > colorectal (63/433=0.1454965) > kidney (6/679=0.0088365) > thyroid (2/393=0.005089). Mutations in TP53 are common and cover many patients, but there are some other genes that only cover one patient in each tumor type. This is because mutations of these genes are rare.

To further distinguish among the cancer types, we can categorize genes as druggable or not. For example, FDA has approved the drugs vemurafenib and dabrafenib to treat melanoma patients; these drugs target the gene BRAF at amino acid position number 600 where the normal valine is replaced by glutamic acid. BRAF mutation occurs in 6 out of 12 tumor types investigated (including brain, colorectal, lung, prostate, skin, and thyroid), so we can assume that the drugs can be applied in other cancers as well.

As the cancer dilemma grows, new measures must be taken to prevent and treat the disease. Future research directions include extending what we have done in this project to other cancers, which would allow us to find similarities among the gene mutations and more effectively detect cancer in its early stages. We can also focus on the most common genes; if trends in behavior are found, we can then target the disease.

REFERENCES / ACKNOWLEDGEMENTS

A census of human cancer genes. Nat Rev Cancer. 2004 Mar;4(3):177-83.
DGIdb: mining the druggable genome. Nat Methods. 2013 Dec;10(12):1209-10.
Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646-74.

I would like to thank the Advanced Authentic Research program and Dr. Choe and Ms. Merchant for providing me the opportunity to explore and analyze cancer genes. I would also like to thank my mentor Dr. Xue Gong of Stanford University for the year-long guidance and support.