EXPLORATORY ANALYSIS OF THE NCI-60 DATASET

Abstract:

We conducted an exploratory analysis of the NCI-60 dataset with the purpose of demonstrating how bioinformatical techniques can guide molecular biology research, and uncover previously undetected relationships between genes involved in tumorigenesis. We first conduct dimension reduction via principal component analysis. We then explore the major contributors to a subset of our principal components, followed by hierarchical clustering to explore relationships between different cancer types. Finally, we conduct a mutational and correlation analysis. We have discovered a possible role for A1BG-AS1 in the inhibition of ABCA1 in p53 mutant cells, which provides possible insights into the mechanism by which ABCA1 inhibition increases tumorgenicity (uncovered by previous studies).

Background & Objectives:

The NCI-60 Human Tumor Cell Line database was designed to compile the results of the screening of 60 different human tumor cell lines, to identify and characterize compounds that are able to inhibit growth or induce apoptosis in tumor cells (*National Cancer Institute*, Retrieved March 18, 2018). [i] The project was designed to screen over 3,000 small molecules per year for potential anti-cancer properties, with 60 different tumor cell lines representing common types of cancer such as leukemia, melanoma, and cancers of the brain, lungs, ovary, breast, prostate, colon, and kidneys (*National Cancer Institute*, Retrieved March 18, 2018).

An example of a utility to enhance analytical screening of the NCI-60 dataset is CellMiner, a Bioconductor package which allows queries of up to 150 drugs or genes in the NCI-

60 database [ii] (*Wang, et al.*, 2016; *Corrado & Morine,* 2015). CellMiner also contains annotation data such mutation and gene expression profiles, miRNA expression levels, chromosome specific information, and several other attributes[iii] (Luna, 2017).

Significant strides have been made in data analysis of "omics" data, and the NCI-60 dataset has played a vital role in these advancements; as well as increasing our understanding of cancer. While the National Cancer Institute is in the process of overhauling its tumor cell lines,[iv] (Ledford, 2016), the NCI-60 database remains an important resource for researchers.

Meng, et al., recently studied dimension reduction techniques and integrative analysis using the NCI-60 database[v] (Meng, et al., 2016). Meng and colleagues applied principal component analysis (PCA) to analyze gene expression data of a subset of cell lines from the NCI-60 panel (Meng et al., 2016).

We adopt a similar approach here. We begin with PCA on a subset of cell lines from the NCI-60 panel, and then look to the contribution of individual genes to our principal components. Before conducting a gene-wise analysis, we perform hierarchical clustering on the same subset of data (used for PCA), looking for relationships by cancer type. We then conduct a simple mutational analysis, looking at the number of mutations in common DNA repair genes, as reported by Wood et al.[vi], followed by a correlation study and analysis of results (Wood et al., 2001).

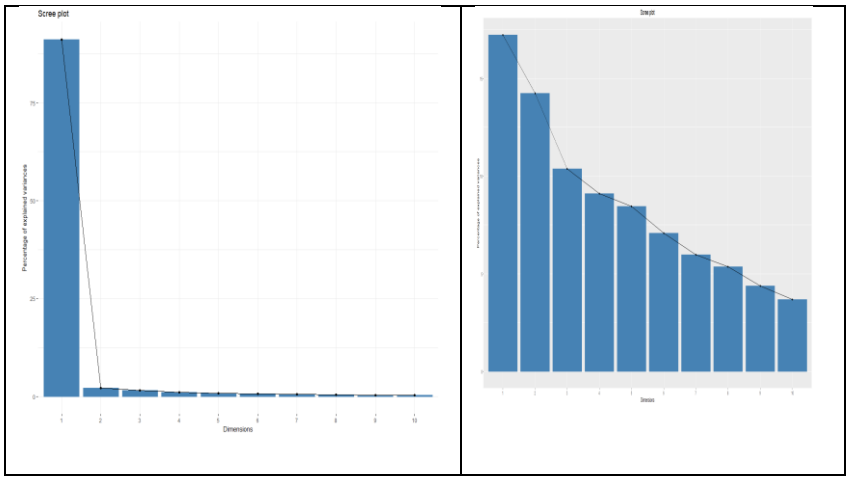
Computational Methods:

All substantive computational work was performed using the R programming language, on the Rstudio IDE.

Results & Discussion:

PRINCIPAL COMPONENT ANALYSIS

PCA is performed on the first 17 columns of data, which represents breast, CNS, and colon cancers. Each column contains expression values taken from microarray analysis of individual patient samples. The data was further truncated by inclusion of only the first 50 rows of data (each row contains gene expression values from individual patient samples).



Scree plots showing the percentage of explained variances for each component. This demonstrates the importance of scaling data prior to performing PCA. According to the first plot (left), the first component accounts for nearly all the variance in the dataset. However, this is characteristic of PCA done without scaling. The second plot (right) shows PCA after scaling (James G. et al., 2013).[vii]

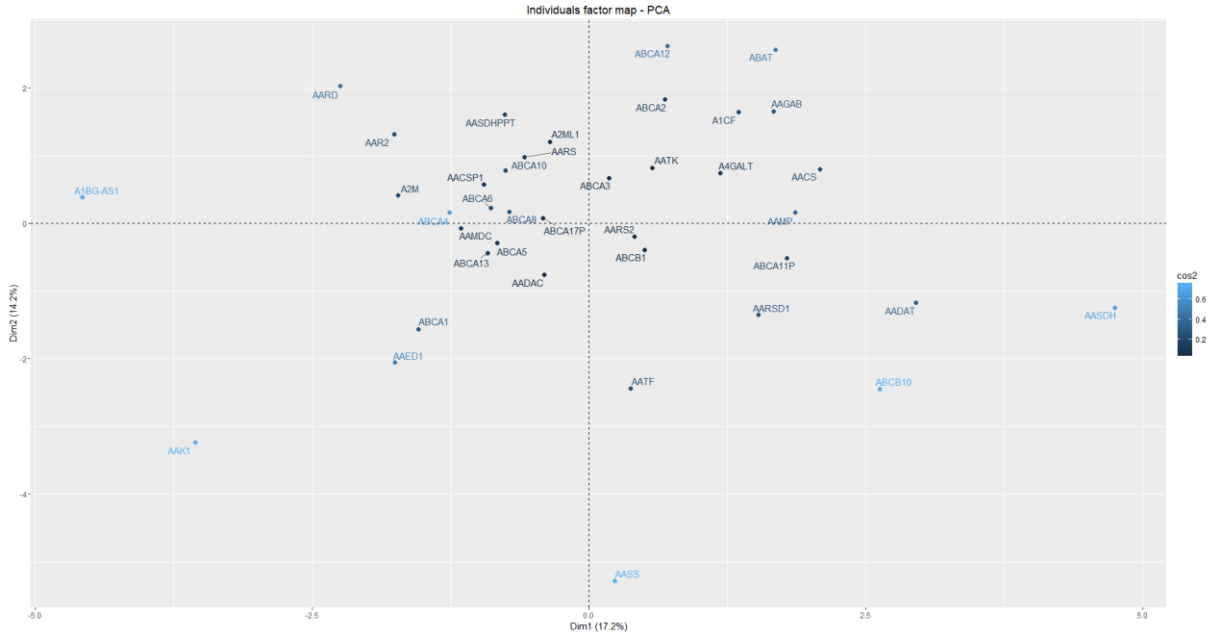
```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6    Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14
Standard deviation     1.7615219 1.6006449 1.3663758 1.2802354 1.2332087 1.12915290 1.0373764 0.98243167 0.88669709 0.81482355 0.79436719 0.64805724 0.60959208 0.59123176
Proportion of Variance 0.1723866 0.1423369 0.1037213 0.0910557 0.0844891 0.07083257 0.0597861 0.05362067 0.04367954 0.03688541 0.03505662 0.02333212 0.02064458 0.01941972
Cumulative Proportion  0.1723866 0.3147235 0.4184448 0.5095005 0.5939896 0.66482216 0.7246083 0.77822893 0.82190847 0.85879388 0.89385051 0.91718263 0.93782721 0.95724693
                          Comp.15    Comp.16     Comp.17     Comp.18
Standard deviation     0.50516167 0.47508585 0.390631058 0.368873546
Proportion of Variance 0.01417713 0.01253925 0.008477368 0.007559316
Cumulative Proportion  0.97142406 0.98396332 0.992440684 1.000000000
```

This summary of PCA results shows the relative importance of each component. PCA allows us to perform dimension reduction; where we retain those components responsible for .85 of the variance. In this case, referring to the "cumulative proportion" row in our summary, we retain components one through ten; allowing us to reduce the number of dimensions by nearly half.

Dimension reduction allows us to reduce the number of variables under consideration (Roweis & Saul, 2000).[viii] Among the subset of 50 genes and 18 different patient samples, taken

across 3 cancer types, representing 900 gene expression values, we were able to reduce our dataset by 8 dimensions and 400 gene expression values.

Next, we investigate the relative contribution of genes in our first two principal components.



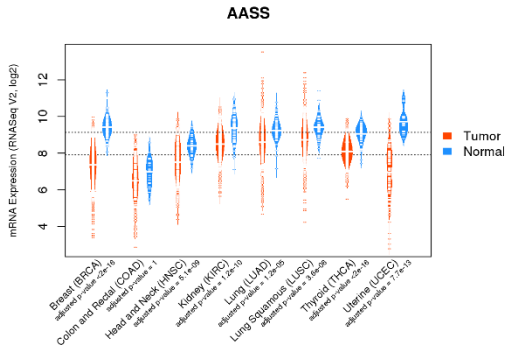
The above plot is a graph of the individual gene expression values identified by principal component analysis for the first two dimensions. Genes with similar expression profiles are grouped together. The plot also provides a color key showing the contribution of each gene.

As shown in the individuals plot, the genes with the highest contributions are as follows:

GENE NAMES	GENE DESCRIPTION
AASDH	**Aminoadipate-semialdehyde dehydrogenase**
ABCB10	**ATP binding cassette subfamily B member 10**
AASS	**Aminoadipate-semialdehyde synthase**
ABCA4	**ATP binding cassette subfamily A member 4**
AAK1	**AP2 associated kinase 1**
A1BG-AS1	**A1BG antisense RNA 1**

Gene descriptions from: *National Center for Biotechnology Information: Gene*. accessed 4 May 2018. Available at: https://www.ncbi.nlm.nih.gov/gene/.

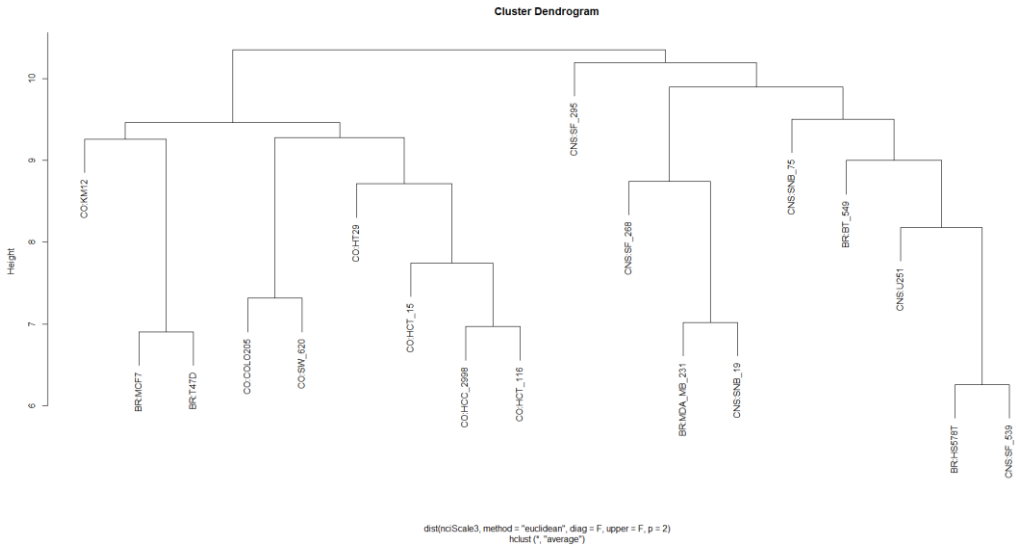
Our investigation of the genes with the highest conribution to our first two components revealed important roles in cancer. For example, ABCB10 appears in the cancer genomes of human breast and colorectal cancers (Liesa et al., 2012).[ix] AAK1 is implicated in colorectal cancer i.e. knockdown of AAK1 inhibits viability in colorectal carcinoma cells through an unknown mechanism (Baldwin et al., 2008).[x] AASDH is commonly found in breast cancer; however, because it's a synonymous variant, it is thought to not play an important role in protein function (Hilbers et al., 2013).[xi] Breast cancer patients with low copy number (deletion) of ABCA4 have lower overall survival probability than patients without them (Havrysh & Kiyamova, 2017).[xii] And finally, AASS is commonly expressed in multiple types of cancer (see image below).



Source: *Cancer Cell Metabolism Gene DB*. accessed 4 May 2018. Available at: https://bioinfo.uth.edu/ccmGDB/gene_search_result.cgi?page=page&type=quick_search&quick_search=10157.

HIERARCHICAL CLUSTER ANALYSIS

Hierarchical clustering was performed on the same subset of data used for PCA; however, the dataset was transposed in order to cluster by cancer type.

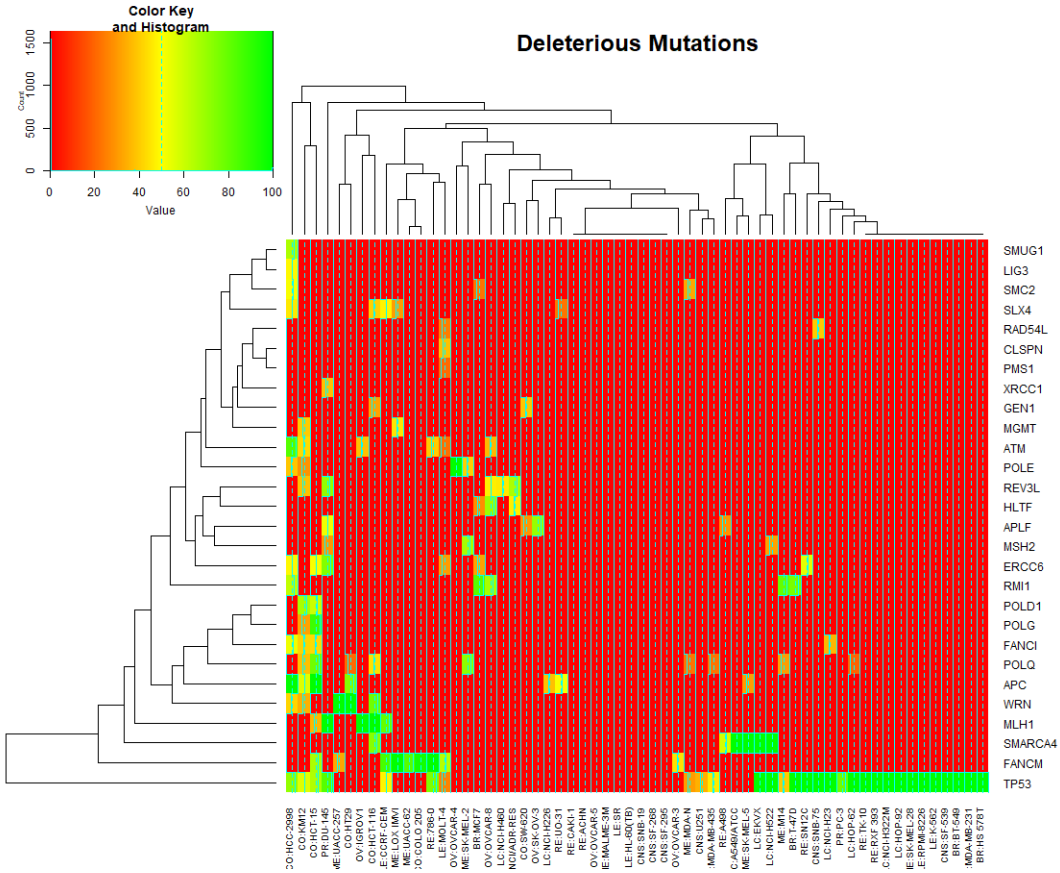


The above shown plot is a dendrogram of hierarchical clustering analysis conducted on the same subset of data used for PCA (transposed). This shows relationship by cancer type. The most interesting relationship we see is between breast and CNS cancers in the right lower portion of the plot. What's particularly striking is this relationship appears twice in the data, and in the former case, the relationship is closer than what we see with respect to cancers of the same type.

Interestingly, we find the closest relationship between two different cancer types, breast and CNS cancers, which was unexpected. Metastases to the central nervous system (commonly the brain) in breast cancer is a common complication (Lin et al., 2013).[xiii] However, it was nonetheless surprising to find a closer link between breast and CNS cancers than between samples from patients with the same cancer type.

MUTATION ANALYSIS

We begin with a simple analysis examining the number of mutations among common DNA repair genes in our dataset.



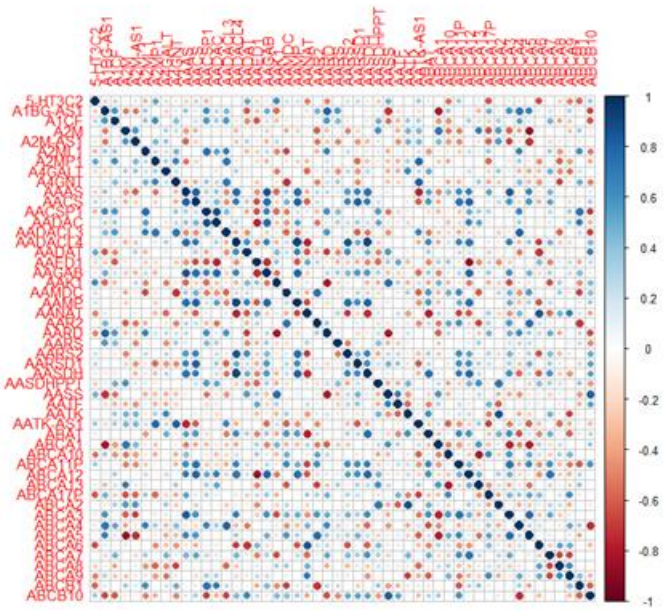
The heatmap shown above provides a tally of the number of mutations for all genes listed on the right most column of the plot. The bottom row lists cancer types. The colors in the heatmap correspond with the color key, showing number ranges for mutation counts for all genes listed, organized by cancer type. For example, we see the highest mutation count with TP53. Approximately half of all cancer types listed exhibit TP53 mutations, with the majority of cancer types showing a high number of TP53 mutations (approximately 80 to 100 mutations).

As shown, TP53 is the most frequently mutated gene among the DNA repair genes analyzed. At the bottom right section of the heatmap, we see high rates of TP53 mutation across several different cancer types; most prominently, breast and CNS cancers. Because we find several points of convergence between breast and CNS cancers, we will now confine our analysis to these two cancer types.

	BR:MCF7	BR:MDA_MB_231	BR:HS578T	BR:BT_549	BR:T47D	CNS:SF_268	CNS:SF_295	CNS:SF_539	CNS:SNB_19	CNS:SNB_75	CNS:U251
BR:MCF7	1.0000000	0.8079494	0.8574417	0.8792642	0.7893461	0.8155008	0.8110618	0.6135896	0.8433769	0.8134806	0.8326105
BR:MDA_MB_231	0.8079494	1.0000000	0.9178530	0.8822749	0.6522693	0.9443339	0.9325476	0.7539262	0.9530054	0.8922111	0.9316257
BR:HS578T	0.8574417	0.9178530	1.0000000	0.9359069	0.6886752	0.9011684	0.9415787	0.7864825	0.9267782	0.9073143	0.9290744
BR:BT_549	0.8792642	0.8822749	0.9359069	1.0000000	0.6743978	0.8953385	0.8909754	0.7875277	0.9082953	0.8906665	0.9123268
BR:T47D	0.7893461	0.6522693	0.6886752	0.6743978	1.0000000	0.6325994	0.6323953	0.4307238	0.6618446	0.6216961	0.6367904
CNS:SF_268	0.8155008	0.9443339	0.9011684	0.8953385	0.6325994	1.0000000	0.8919843	0.7677658	0.9193285	0.8472587	0.8871313

The above table is the "head" of a correlation analysis of breast and CNS cancers. We see high Spearman correlation coefficients across the board, in some case, nearly 0.95, indicating strong correlation in gene expression values across these two cancers.

The head of the correlation analysis conducted on breast and CNS cancers confirms the close relationship between these two cancer types. To explore this relationship further, we truncated our dataset to include only expression values for breast and CNS cancers, and then conducted a correlation analysis using the Spearman method.



Correlation plot using the "corrplot" R package. The color key on the right side of the plot indicates the Spearman correlation coefficient associated with each gene pair shown.

We see numerous gene pairs with correlation coefficients near one (or negative one). A correlation coefficient above 0.90 suggests possible coexpression, while a correlation coefficient near negative one suggests negative regulation (e.g. inhibition). We will confine our analysis to the six genes with the highest contribution to dimensions one and two (from PCA conducted earlier in the study).

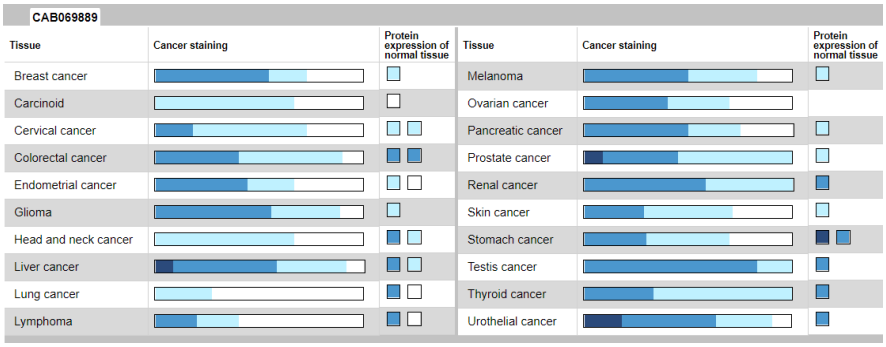
Recall, those six genes are: AASDH, ABCB10, AASS, ABCA4, AAK1, and A1BG-AS1. Below, we provide a table showing gene pair relationships:

GENE NAMES	POSSIBLE COEXPRESSION	POSSIBLE INHIBITION
AASDH	**AADACL4**	
ABCB10		**ABCA4**
AASS		**AARS**
ABCA4		
AAK1		**AARD**
A1BG-AS1		**ABCA1**

The correlation analysis shows several gene pairs with strong positive or negative correlation. The gene pairs which appear in the above table all show a correlation coefficient (approximately) greater than the absolute value of 0.90. In the first case, AASDH/AADACL4, we see a possible coexpression relationship. In four of the five other genes (taken from our principal component analysis) we see a negative correlation above the absolute value of 0.90. This suggest one gene negatively regulates the other e.g. one gene inhibits expression of the other. A survey of the literature shows high expression rates for most of these genes in fat, colon, and brain tissue, along with many other tissue systems (NCBI, retrieved 4 May 2018).[xiv]

A survey of the literature revealed interesting associations between ABCA1, A1BG-AS1, and TP53 (Smith & Land, 2012).[xv] As shown in the table above, we found strong negative

correlation between A1BG-AS1 and ABCA1. As shown in the table below, the ABCA1 gene is expressed in many different cancer types.

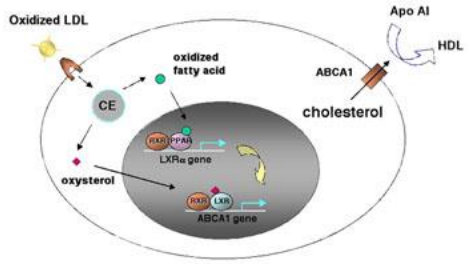


CAB069889

Tissue	Cancer staining	Protein expression of normal tissue	Tissue	Cancer staining	Protein expression of normal tissue
Breast cancer			Melanoma		
Carcinoid			Ovarian cancer		
Cervical cancer			Pancreatic cancer		
Colorectal cancer			Prostate cancer		
Endometrial cancer			Renal cancer		
Glioma			Skin cancer		
Head and neck cancer			Stomach cancer		
Liver cancer			Testis cancer		
Lung cancer			Thyroid cancer		
Lymphoma			Urothelial cancer		

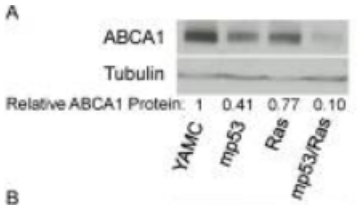
The table shows expression rates for the ABCA1 gene. Source: *The Human Protein Atlas*, accessed 4 May 2018. Available at: http://v16.proteinatlas.org/ENSG00000157426-AASDH/cancer.

ABCA1 is an ATP-binding cassette transporter known as a cholesterol efflux regulatory protein, playing a major role in the regulation of cellular cholesterol and phospholipid homeostasis (Luciani et al., 1994).[xvi] ABCA1 also exhibits anti-cancer activity (Smith & Land, 2012).



Source: Rodriguez, B., *Trafficking in Cholesterol: Investigating the Human ABCA1 Gene*. Genome News Network. 9 July 2001. Available at: http://www.genomenewsnetwork.org/articles/07_01/Trafficking_cholesterol.shtml.

Along with its role in cholesterol regulation, ABCA1 is also involved in tumor inhibition (Smith & Land, 2012). It has been identified as a cooperation response gene (CRG) by virtue of its synergistic regulation by mutant p53 (Smith & Land, 2012). Shown below are the results of western blot, showing murine ABCA1 protein expression in YAMC, p53, and RAS/p53 cells (Smith & Land, 2012).



Source: Smith, B., & Land, H. (2012).

It is thought that ABCA1 can modulate the deleterious effects of mutant p53 by inducing cytochrome C release (Smith & Land, 2012). It has been long established that p53 can induce apoptosis by increasing mitochondrial Cytochrome C release (Schuler et al., 2000).[xvii] As noted above, ABCA1 also induces Cytochrome C release, and this property is responsible for its role in tumor suppression (Smith & Land, 2012). It has also been shown that silencing of ABCA1 in p53 mutant cells increases tumorgenicity (Smith & Land, 2012).

Here we show a possible role for A1BG-AS1 in the inhibition of ABCA1 expression, and we might also infer that this role is enhanced by its occurrence in p53 mutant cells, given the high number of TP53 mutations found in breast and CNS cancers (as revealed by our mutation analysis of DNA repair genes).

Conclusion:

Here we show how an open ended, exploratory analysis of a well curated gene dataset, can lead to valuable insights which can guide human disease research. It is important to note, this analysis was performed on a relatively small subset of the NCI-60 dataset (50 out of 25040 rows, representing gene expression values, and 18 out of 60 columns of data, representing different patient samples organized by cancer type).

To further bolster our findings, it will be necessary to perform a more extensive study. If we can confirm our findings, a confirmatory wet lab study will be needed to elucidate on the role

A1BG-AS1 plays in the inhibition of ABCA1. One possible approach is to conduct a knockout study, silencing expression of A1BG-AS1 in cell culture, and a differential expression analysis to quantify ABCA1 expression both before and after A1BG-AS1 knockdown (e.g. total RNA extraction followed by RNAseq).

References:

[i] In: *National Cancer Institute*, NCI-60 Human Tumor Cell Lines Screen. Last Updated Aug. 26, 2015. Retrieved March 18, 2018. Available at: https://dtp.cancer.gov/discovery_development/nci-60.

[ii] Journal Article: *Sufang Wang et al.*, CellMiner Companion: an interactive web application to explore CellMiner NCI-60 data. Bioinformatics. Vol. 32. Issue 15. pp. 2399-2401. Aug. 2016; *Corrado & Morine*, Analysis of Biological Systems. Imperial College Press. London, UK (2015).

[iii] Journal Article: *Luna, A.A. et al.,* (2017). Package " rcellminerData ." pp. 1–5. Available at: http://www.bioconductor.org.

[iv] Journal Article: *Ledford, H*., US cancer institute to overhaul tumour cell lines. Nature. Vol. 530. p. 391. 25 Feb. 2016. doi: 10.1038/nature.2016.19364.

[v] Journal Article: *Meng et al*., Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics. Vol. 17(4). pp. 628-641. 11 March 2016.

[vi] Journal Article: *Wood et al*. Human DNA Repair Genes. Science. Vol. 291. Issue 5507. pp. 1284 – 1289. doi: 10.1126 (2001).

[vii] Book Chapter: *James, G. et al*., (2013). An Introduction to Statistical Learning. Available at: http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf%5Cnhttp://link.springer.com/10.1007/978-1-4614-7138-7.

[viii] Journal Article: *Roweis, S. T. & Saul, L. K*. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science290. Vol. 290(5500). pp. 2323-2326. doi:10.1126/science.290.550.2323, PMID 11125150.

[ix] Journal Article: *Liesa, M. et al*., (2012). Mitochondrial ABC transporters function: the role of ABCB10 (ABC-me) as a novel player in cellular handling of reactive oxygen species. Biochimica

et Biophysica Acta. Vol. 1823(10). 10.1016/j.bbamcr.2012.07.013. Available at: http://doi.org/10.1016/j.bbamcr.2012.07.013.

[x] Journal Article: *Baldwin, A. et al*., (2008). Kinase requirements in human cells: II. Genetic interaction screens identify kinase requirements following HPV16 E7 expression in cancer cells. Proceedings of the National Academy of Sciences of the United States of America. Vol. 105(43), pp. 16478–16483. Available at: http://doi.org/10.1073/pnas.0806195105.

[xi] Journal Article: *Hilbers, F. S. et al*., (2013). Exome Sequencing of Germline DNA from Non-BRCA1/2 Familial Breast Cancer Cases Selected on the Basis of aCGH Tumor Profiling. PLoS ONE. Vol. 8(1), e55734. http://doi.org/10.1371/journal.pone.0055734.

[xii] Journal Article: *K. Havrysh & R. Kiyamova*, New potential biomarkers for breast cancer prognosis, Annals of Oncology. Vol. 28, Issue suppl_7, 1 October 2017. Available at: mdx508.011, https://doi.org/10.1093/annonc/mdx508.011.

[xiii] Journal Article: Lin, N. et al., *CNS metastases in breast cancer: old challenge new frontiers*. Clinical Cancer Research. Vol. 19. Issue 23. pp. 6404-6418. 1 Dec. 2013. Available at: http://clincancerres.aacrjournals.org/content/19/23/6404.full-text.pdf.

[xiv] In: *National Center for Biotechnology Information: Gene.* Accessed 4 May 2018. Available at: https://www.ncbi.nlm.nih.gov/gene/.

[xv] Journal Article: Smith, B. & Land, H. (2012). *Anti-cancer activity of the cholesterol exporter ABCA1 gene*. Cell Reports. Vol. 2(3). Pp. 580–590. Available at: http://doi.org/10.1016/j.celrep.2012.08.011.

[xvi] Journal Article: Luciani M. F. et al., (May 1994). *Cloning of two novel ABC transporters mapping on human chromosome 9*. Genomics. Vol. 21 (1). pp. 150–9. doi:10.1006/geno.1994.1237. PMID 8088782.

[xvii] Journal Article: Schuler, M. et al., *p53 Induces Apoptosis by Caspase Activation through Mitochondrial Cytochrome C Release*. The Journal of Biochemistry. Vol. 275. No. 10. pp. 7337-7342. 10 March 2000.