Genomic Clustering and Gene Expression in Cancer-Positive Patients

Citadel Datathon
YuHo Hsieh, Xin Rong Chua, Tom Yuz

Non-Technical Executive Summary

Control of gene expression plays a central role in determining cell fate, differentiation, and the maintenance of specific cell lineages. Consequently, advances in gene sequencing seeking to understand gene expression have created an immense amount of biological information. Rapid identification of aberrant gene expression is crucial in preventing defects and disease.

Currently, a number of commonly used methods in genetic data analysis elucidate patterns by evaluating gene expression profiles **regardless of their physical locations** in the genome. These methods include [1], K-means clustering [2], self-organizing maps [3], and support vector machines. However, analysis of **gene clusters** offers a more direct, holistic approach for illuminating order in the entire set of observations. Conceivably, knowledge of gene location supplemented by an understanding of gene expression could help identify new loci associated with a variety of human cancers.

Topic Question: Do genes with a high FPKM in cancer-positive patients cluster around similar chromosomal locations? Can we efficiently make use of expression measurement data to identify useful chemical compounds for the treatment or diagnosis of cancer?

Brief Introduction

A gene cluster is a group of two or more genes found within an organism's DNA that encode for similar proteins, which are often located within a thousand base pairs of each other. If altered patterns of gene expression correlate with abnormalities local to a particular chromosome, then assessment of contiguous genes will rapidly identify potential cancer developments.

For example, several chromosomal regions have been identified for their frequent deletions in various types of cancer. Structural changes at the subchromosomal level such as amplifications or deletions have been demonstrated to affect gene expression and illustrate a common type of genetic defect in this disease. Indeed, chromosomal deletion is an early and frequent somatic genetic alteration during carcinogenesis.

Methodology & Results

The team initiated a focus on breast cancer in order to demonstrate the merit of the clustering approach. The methodology below can be easily extended to all other types of tissue cancers available in the datasets. Data analysis primarily employed the Jupyter notebook and the Pandas and Matplotlib libraries in Python.

We first identify which genes are critical in causing cancer. In the tcga dataset, gene_ids are sorted by their FPKM expression for cancers in each particular diseased tissue (breast, brain, lungs, etc.). The top 100 genes ranked by FPKM are extracted from the sample.

	gene_id	fpkm_expression	sample_number	organ
1393602	ENSG00000136938	187.850785	950	Breast
1928555	ENSG00000115461	135.312476	1112	Breast
454112	ENSG00000211666	112.159984	477	Breast
895269	ENSG00000223601	88.784984	853	Breast
1457161	ENSG00000175895	60.285891	309	Breast
243338	ENSG00000244468	58.631613	1208	Breast
926794	ENSG00000102409	57.958643	722	Breast
1207478	ENSG00000163344	56.092810	201	Breast
1828507	ENSG00000100985	54.880099	618	Breast
1525802	ENSG00000165502	47.813324	1144	Breast

Exhibit A: Ranking of Gene Expression in Breast Cancer patients per FPKM

Overlaying the top 100 genes obtained in Step 1 with the gtex_gene_expression data, we obtain the chromosome_start and chromosome_end locations of each gene.

	gene_id	fpkm_expression	sample_number	organ	chromosome	chromosome_start	chromosome_end	gene_symbol	score	strand_type
1393602	ENSG00000136938	187.850785	950	Breast	chr9	97983360	98015943	ANP32B	586	+
1928555	ENSG00000115461	135.312476	1112	Breast	chr2	216672104	216695525	IGFBP5	654	-
454112	ENSG00000211666	112.159984	477	Breast	chr22	22758903	22759218	IGLV2-14	474	+
895269	ENSG00000223601	88.784984	853	Breast	chr10	22208813	22210021	EBLN1	108	-
1457161	ENSG00000175895	60.285891	309	Breast	chr8	95133803	95156684	PLEKHF2	397	+
243338	ENSG00000244468	58.631613	1208	Breast	chr3	149284781	149333653	RP11-206M11.7	214	+
926794	ENSG00000102409	57.958643	722	Breast	chrX	103215091	103217246	BEX4	579	+
1207478	ENSG00000163344	56.092810	201	Breast	chr1	154924733	154936991	PMVK	528	-
1828507	ENSG00000100985	54.880099	618	Breast	chr20	46008907	46016561	MMP9	431	+
1525802	ENSG00000165502	47.813324	1144	Breast	chr14	49618518	49620685	RPL36AL	649	-

Exhibit B: Identifying Chromosome Start and Chromosome Ends of top 100 Genes

Consequently, we graph each gene's starting position in chromosome and ending position in the chromosome. The method partitions data into subsets by chromosomal location for each gene interrogated by an array. A graphical display is generated by representing each genomic locus with a colored cell that quantitatively reflects its differential expression:

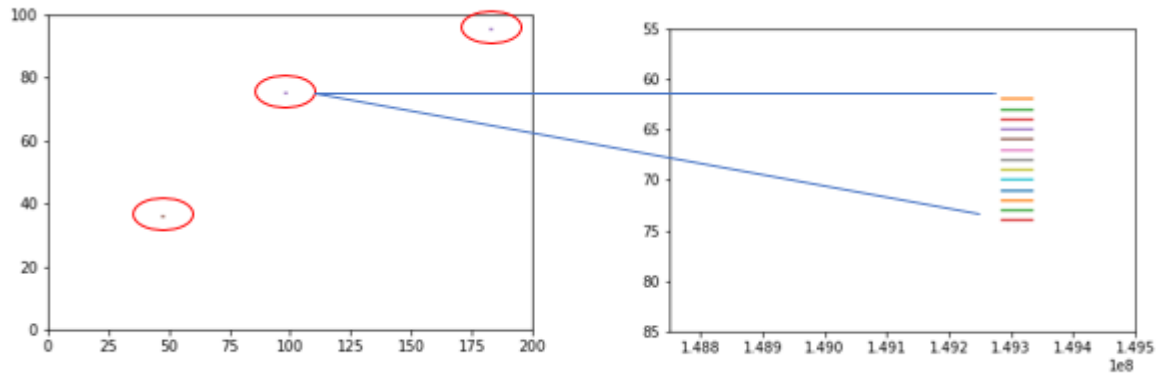

Exhibit C: Clustering of Key Breast Cancer Genes at Particular Chromosomal Areas

For the breast cancer cells shown above, we successfully obtain three key areas of chromosomal overlap. In other words, the genes crucial in causing breast cancer tend to cluster in key chromosomal regions.

In order to ensure that results are robust, it is feasible to assume that similarities in chromosomal location of genes linked to breast cancer should translate into similarity in transcription locations. Hence, turning to the gtex_gen_model dataset, we identify the transcription_start and transcription_end for each of the top 100 genes. Controlling for +/- strands, we visualize the results of location vs gene:

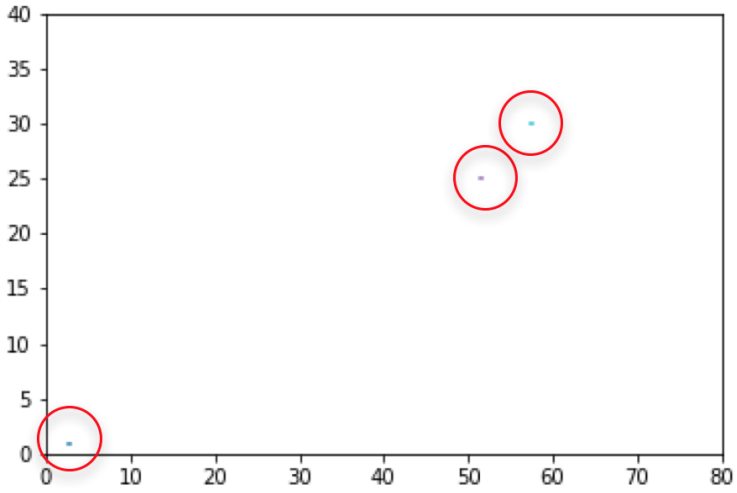

Exhibit D: Clustering of Key Breast Cancer Genes at Particular Transcription Regions

Again we identify the tendency for genes critically linked in the causation of breast cancer to cluster. The above data demonstrates three key clusters corresponding to the coding regions of the genes.

Groups of adjacent and co-regulated genes that are not otherwise functionally related in any obvious way can now be identified by expression profiling. The ability to target a clustered group of genes directly responsible for cancerous mutations would be invaluable in medical treatment. As a result, the team explored the relationship between identified gene clusters and corresponding chemical compounds.

By utilizing the toxicogenomics_diseases dataset, influence_score's of particular chemical compounds on the above-identified gene_id's can be compiled. We extracted 12 genes sharing a similar chromosomal location (obtained from the cluster shown in Exhibit B) and identified the chemical most closely associated with a given gene based on inference score.

gene_id	chemical_nam	chemical_i	gene_form	interactions
ENSG00000165507	Estradiol	D004958	mRNA	Estradiol results in increased expression of C10ORF10 mRNA
ENSG00000115461	Estradiol	D004958	mRNA	Estradiol affects the expression of IGFBP5 mRNA
ENSG00000115461	Estradiol	D004958	mRNA	[Estradiol co-treated with Progesterone] results in decreased expression of IGFBP5 mRNA
ENSG00000115461	Estradiol	D004958	mRNA	[Estradiol co-treated with Tetrachlorodibenzodioxin] results in increased expression of IGFBP5 mRNA
ENSG00000115461	Estradiol	D004958	mRNA	Estradiol results in decreased expression of IGFBP5 mRNA
ENSG00000115461	Estradiol	D004958	mRNA	Estradiol results in increased expression of IGFBP5 mRNA
ENSG00000165502	Estradiol	D004958	mRNA	Estradiol results in decreased expression of RPL36AL mRNA
ENSG00000170099	Estradiol	D004958	protein	ESR1 protein promotes the reaction [Estradiol results in increased expression of SERPINA6 protein]
ENSG00000170099	Estradiol	D004958	protein	Estradiol results in increased expression of SERPINA6 protein
ENSG00000124107	Estradiol	D004958	mRNA	[Estradiol co-treated with Tetrachlorodibenzodioxin] results in decreased expression of SLPI mRNA
ENSG00000124107	Estradiol	D004958	mRNA	Estradiol results in decreased expression of SLPI mRNA
ENSG00000124107	Estradiol	D004958	mRNA	Estradiol results in increased expression of SLPI mRNA

Exhibit E: Interactions between Estradiol and genes associated with breast cancer

As demonstrated above, **5 clustered genes associated with breast cancer correspond to a single chemical compound (estradiol).** In other words, our process first identified the genes most closely linked with breast cancer, organized them based on chromosomal location, and identified a critical chemical compound related to the stimulation of said genes. Monitoring the prevalence of estradiol in the human body offers an early indicator of potential breast cancer developments.

As a matter of fact, estradiol is the #1 breast cancer treatment currently being utilized. Evidently, the genetic clustering process used above was able to ascertain the most useful treatment for breast cancer actually being employed in the world today. The extension of this process to other critical diseases offers an expedited approach to identifying crucial treatments and medical indicators.

Conclusion

Analyzing the clustering of cancer-related genes offers a rapid approach to identifying chemical compounds which serve as early indicators of cancer developments. This study presents an approach that can be adopted to a variety of complex human cancers. Firstly, the genes critically linked to human cancers are identified (in this case breast cancer). Consequently, the identified genes are evaluated based on their chromosomal locations. Our process shows that adjacent pairs of genes show correlated expression associated with a given human cancer. By evaluating gene clusters one can identify key chemical compounds influencing a variety of contiguous genes. These chemical compounds can be used to monitor the potential development of cancer in human patients.

Advances in genome sequencing now allow for precise locations for many expressed genes. Systems that procure map genes based on chromosome location (e.g. DIGIMAP, GENEPLOTTER, and D-Chip) already exist and can be applied to targeting disease incidents. These gene locations serve as ideal “landmarks” for an integrated platform combining a variety of genomic data sources. Alterations in chemical compounds linked to coregulated contiguous gene offers a means of manipulating gene expression linked to cancer.

Future research can demonstrate the robustness of results by extending the methodology to either disease types (brain, lung, kidney, etc.). Further evaluation of similarities within gene clusters can also yield additional chemical compounds previously unlinked to particular cancers.

Thank you to the Citadel Datathon for the opportunity to explore this unique dataset.