**Deciphering Chromosomal Instability: A Comprehensive Analysis of Lamin-Associated Domains and Heterochromatin dynamics in Cancer Genomics**

**Interim Report**

Name: Dylan Haynes-Simmons

Internship location: Janssen Lab, Center for Molecular Medicine, UMC Utrecht

Supervisor: Prof. Aniek Janssen

Daily supervisor: Aditya Dixit

Introduction:

Chromosomal Instability or CIN, a hallmark of many cancer types, is defined by a high frequency of aberrations in the number of whole or partial chromosomes (aneuploidy) and structural chromosomal rearrangements [1]. Given this feature of CIN to compromise the structural integrity of the chromosomes, heterochromatin, the portion of the chromatin that tends to be transcriptionally repressed and accounts for the majority of the chromatin at structurally important sites such as at centromeric, peri-centromeric and telomeric regions, is commonly targeted in cancer [2,3]. Lamina Associated Domains (LADs), nuclear peripheral regions where heterochromatin interacts with the nuclear lamina, are also of significant importance due to their function in gene regulation and nuclear organisation but also get the targeted in tumorigenesis [4]. It is therefore interesting to ask if: *it* *is it possible to identify specific LAD & Heterochromatin-associated aberrations related to certain markers of chromosomal instability in cancer?*

Several studies have been able to identify and classify cancer types based on the analysis of mutational, copy number or other CIN based signatures [5,6,7] but few attempt to relate signs of CIN back to the expression of specific genes within the genome. In this project, by leveraging genomic data curated from public repositories and using other public resources my goal is to do just that with a focus on genes whose dysregulation has the potential to cause structural aberrations at heterochromatin binding regions near the nuclear periphery.

Starting with a core list of candidate proteins that have been selected for based on their association with the nuclear lamina and heterochromatin, a query on a Protein-Protein interaction database such as STRING can be performed in order to curate the interactome of the candidate proteins. Transcription data (RNAseq) for the genes associated with this list of interacting proteins (Gene-set of Interest or Set of Interest) can be used to analyse the transcriptional regulation of these genes across a variety of cancers by accessing this data from The Cancer Genome Atlas (TCGA). After assessing whether the provided interactome represents an interesting set to analyse further, the data can be used to generate expression-based signatures representing different expression patterns in unique types of tumours.

Using the same gene-set of interest, other forms of data for this set but potentially representing underlying effects of chromosome instability can be curated, analysed, and used to generate CIN-based signatures representing different characteristics of CIN in specific cancer types.

The expression-based and CIN-based signatures can then be analysed as one to gain insights into transcriptional mechanisms at the basis of certain aspects of chromosome instability across particular oncogenic landscapes.

**Current State:**

Objective:

To curate a comprehensive genetic interactome for candidate proteins linked to the nuclear lamina and heterochromatin, assess the relevancy of the genes included in the interactome and investigate any collective, tumour-specific expression patterns.

Results:

By querying the STRING PPI database with a list of LAD and heterochromatin associated proteins (Table 1), a gene-set of interest could be obtained, and an analysis of the Set-of-Interest could be carried out to determine the validity of the selection based on collective transcriptional characteristics of the genes within the set.

Given the unique nature of the genetic makeup of the Set-of-Interest, comparing this set to a single randomly generated control set did not seem sufficiently robust, therefore necessitating the generation of multiple control sets for the comparative analyses (see methods).

In the cancerous TCGA PanCan data, the control sets proved to maintain a high degree of transcriptional uniqueness (Friedman’s test: p-value <2.2e-16) despite the large number of controls. Indeed, a contrast analysis comparing the average expressional activity of each control-set to the global transcriptional trend of all other control sets revealed that only 7 of the 547 control-sets did not show a significant difference from the global trend (Pairwise Two-samples Wilcoxon test, adjusted p-value > 0.05). Yet despite their individuality, the Set-of-Interest’s aggregate transcriptional makeup showed a significant difference compared to each of the controls which outweighed any of the differences between any 2 control sets (Fig 1.a).



*Table 1*. Candidate Proteins & Genes and their literature-based characteristics. First two columns are the candidate proteins and genes, respectively. The genes in bold represent the final selection.

Not only was this difference between the Set-of-Interest and each control-set significant (Paired permutation test, p-value = 0 for all 547 comparisons), but the difference between the mean expression of the Set-of-Interest and that of the control sets was also maintained in a firm range between 8 and 10 on a log 2 scale representing an increase in expression in the Set-of-Interest approximately 512 times greater than that within the average control set (Fig 2.a). Furthermore, this difference proved to be persistent and significant in each cancer type when the data was parsed by tumour type (tumour-specific Paired permutation test, p-value = 0), although no type-specific patterns were visible in the aggregate expression data (Fig 3.a). However, when analysing the non-cancerous GTEx data, the Set-of-Interest’s collective transcriptional profile likewise showed a significant difference (paired permutation test, p-value = 0 for all 547 comparisons) compared to each of the controls, greater than the difference between any two control sets (Fig 1.b). Yet on this occasion, the difference between the mean expression of the Set-of-Interest to that of the control sets maintained a slightly lower and wider range between 7 and 10 on a log 2 scale representing an increase in

A screenshot of a graph

Description automatically generatedexpression in the Set-of-Interest around 256 times greater than that displayed within the average control set (Fig 2.b). Here too, the difference proved to be persistent and significant in tissue type when the data was grouped by the latter (tissue-specific Paired permutation test, p-value = 0), although no type-specific patterns were visible in the aggregate expression A screenshot of a graph

Description automatically generateddata (Fig 3.b).

A chart of different types of expressions

Description automatically generated with medium confidence

Methods:

**Data Acquisition**

The initial phase involved curating a list of 23 candidate proteins, known for their association with the nuclear lamina and potential interactions with heterochromatin, based on prior literature and information from UniProt [8]. This list was refined to 19 proteins after evaluating the confidence in their association with the nuclear lamina and chromatin interactions. A custom Python script was developed to leverage this curated list as a query input for the STRING Protein-Protein Interaction (PPI) database [9]. The script specifically requested physical interactions with a minimum confidence score of 0.4, aiming to capture a broad spectrum of potential interactors. The outcome of this query was a comprehensive list of genes (n = 328) encoding the interacting proteins, thus forming the interactome of interest for subsequent analysis.

Parallel to the interactome curation, data was sourced from two extensive repositories: The Cancer Genome Atlas (TCGA) [10] for the cancerous data, and the Genotype-Tissue Expression (GTEx) Portal [11] for non-cancerous data.

Datasets from TCGA:

1. PanCan TPM normalized RNAseq:

* Datatype: PanCan TPM normalized & log2 transformed RNAseq data
* Dimensions: 60,499 gene identifiers across 10,535 samples
* Source: Downloaded via the UCSC Xena browser [12].

Datasets from GTEx:

1. TPM normalised RNAseq:

* Datatype: TPM normalized RNAseq data encompassing
* Dimensions: 56,200 gene identifiers across 15,201 samples

The dual transcription based dataset approach enabled a comparative analysis of gene expression patterns across a wide array of human tissues and cancer types, setting the stage for identifying expression trends uniquely associated with cancer and potentially indicative of CIN.

**Cleaning & Pre-processing**

The data cleaning and pre-processing phase was critical for ensuring the quality and reliability of the dataset used in this study, focusing on the Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) PanCan RNAseq data.

GTEx Data Cleaning:

* Exclusion of Diseased Samples: Samples from patients with long-standing illnesses were removed based on the Death Hardy Score in the metadata, aiming to exclude potential confounders.
* RNA Integrity Number (RIN) Filtering: Samples with an RIN below 6 were excluded, adhering to GTEx guidelines that recommend this threshold for high-quality RNA analysis. The RIN scale ranges from 0 (totally degraded) to 10 (fully intact), and a score of 6 or above is considered suitable for downstream analyses.
* eQTL Analysis Exclusion: Samples marked for exclusion from eQTL analyses were removed to enhance data quality.

TCGA PanCan RNAseq Cleaning:

* Removal of Non-cancerous Samples: Any samples identified as originating from non-cancerous controls or cell lines were excluded, focusing the analysis strictly on cancerous samples.

Combined Filtering:

* Null Value and Gene Identifier Filtering: Samples exhibiting null values across all genes were removed to maintain data integrity. Additionally, gene identifiers lacking a corresponding match in the GTEx or TCGA dataset (based on gene name or ENSEMBL id) were excluded, except for those within the STRING-derived interactome. This process resulted in a curated PanCan TCGA dataset of 9,809 samples and a GTEx dataset of 13,971 samples both across the same 55,487 genes excluding the genes within the Set-of-Interest.

Normalization and Transformation:

* Log2 Transformation: To normalize expression values and facilitate comparative analysis, GTEx expression values were log2 transformed, with zero values adjusted to 0.001 prior to transformation to avoid mathematical errors.

**Implementation of the Set-of-Interest and Control Sets**

Set-of-Interest (SOI) Creation:

* Utilizing STRING’s interactome list to define the SOI. This list was then applied to filter both the TCGA RNAseq and GTEx datasets to isolate expression data exclusively for genes within the SOI. This resulted in two focused datasets: one containing cancerous expression data (9,809 samples × 328 genes) and another with non-cancerous expression data (13,971 samples × 328 genes).

Control Sets for TCGA Data:

* Control sets were constructed by randomly selecting a number of genes equivalent to the SOI gene size (328) from the TCGA PanCan RNAseq dataset, excluding those genes in the SOI. This process was iterated, ensuring each control set contained a unique combination of genes and that every gene in the dataset, not part of the SoI, appeared in at least one control set. The iterative process continued until all non-SOI genes were utilized, with each control set saved as a CSV file, named sequentially as “TCGA\_TPM\_RNA\_Control\_df” plus the iteration number.

Control Sets for GTEx Data:

* The GTEx control sets mirrored the TCGA control sets. For each TCGA control set, gene names and ENSEMBL ids were extracted and used to filter the GTEx RNAseq dataset for matching genes, excluding those in the SOI. This created control datasets for GTEx that paralleled their TCGA counterparts. Each GTEx control set was saved as a CSV file, named “GTEx\_TPM\_RNA\_Control\_df” followed by the corresponding TCGA control set number, ensuring consistency across cancerous and non-cancerous data analyses.

Current Objectives:

**Create Expression-based Signatures by:**

* Using Differential Gene Expression
* Implementing a correlation-based distance matrix
* Clustering into groups
* Standardising the aggregate scores

**Create CIN-based Signatures by**:

* Analysis of variables within each datatype
* Dimensionality reduction given the large number of variables and datatypes
* Implementing a linear or non-linear based clustering method
* Standardising the scores

References:

1. Geigl, J. B., Obenauf, A. C., Schwarzbraun, T., & Speicher, M. R. (2008). Defining 'chromosomal instability'. Trends in genetics : TIG, 24(2), 64–69. <https://doi.org/10.1016/j.tig.2007.11.006>
2. Carone DM, Lawrence JB. Heterochromatin Instability in Cancer: From the Barr Body to Satellites and the Nuclear Periphery. Semin Cancer Biol. 2013 April; 23(2): 99–108.
3. Fischer AH, Zhao C, Li QK, Gustafson KS, Eltoum I-E, Tambouret R, Benstein B, Savaloja LC, Kulesza P. The cytologic criteria of malignancy. J Cell Biochem. 2010; 110:795-811. [PubMed: 20564180]
4. Lochs SJA, Kefalopoulou S, Kind J. Lamina Associated Domains and Gene Regulation in Development and Cancer. Cells. 2019 Mar 21;8(3):271. doi: 10.3390/cells8030271. PMID: 30901978; PMCID: PMC6468596.
5. Steele, C.D., Abbasi, A., Islam, S.M.A. et al. Signatures of copy number alterations in human cancer. Nature 606, 984–991 (2022). <https://doi.org/10.1038/s41586-022-04738-6>
6. Thutkawkorapin, J., Eisfeldt, J., Tham, E. et al. pyCancerSig: subclassifying human cancer with comprehensive single nucleotide, structural and microsatellite mutational signature deconstruction from whole genome sequencing. BMC Bioinformatics 21, 128 (2020). <https://doi.org/10.1186/s12859-020-3451-8>
7. Koh, G., Degasperi, A., Zou, X. et al. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer 21, 619–637 (2021). https://doi.org/10.1038/s41568-021-00377-7
8. The UniProt Consortium , UniProt: the universal protein knowledgebase, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D158–D169, <https://doi.org/10.1093/nar/gkw1099>; <https://www.uniprot.org/>
9. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. V. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 47(D1), D607–D613. https://doi.org/10.1093/nar/gky1131 ; <https://string-db.org>
10. The Cancer Genome Atlas Research Network. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature, 2008.
11. Lonsdale, J., Thomas, J., Salvatore, M. et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585 (2013). <https://doi.org/10.1038/ng.2653>; <https://gtexportal.org>
12. Goldman, M.J., Craft, B., Hastie, M. et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol (2020). <https://doi.org/10.1038/s41587-020-0546-8>; https://xenabrowser.net