The folder contains six folders and each folder represents a different human cell line

CMK: megakaryoblast (cancer blood cell line)

Control HEKa: normal basal epithelial keratinocytes

GM12878: immortalized B lymphocytes

K562: erythroleukemia (cancer blood cell line)

Molm1: megakaryoblast (cancer blood cell line)

THP-1: macrophage-like (cancer blood cell line)

Each folder contains three files: each of the files correspond to interactions called at different thresholds: FDR: 0.1, 0.01 or 0.001. You can choose the threshold you like to start working with (I recommend FDR.0.1).

For each cell type, there are different experimental conditions which I summarise below:

CMK: the cells are treated with two different drugs (carboplatin and gemcitabine). The file contains the interactions (measured in two replicates) in normal state (CN1 and CN2), in treated with drug C (CC1, CC2) and treated with drug G (CG1, CG2)

Control HEKa: two replicates: X20HEKa and X30aHEKa

GM12878: three replicates: GM12878_rep1, GM12878_rep2, GM12878_rep3

K562: the cells are treated with two different drugs (carboplatin and gemcitabine). The file contains the interactions (measured in two replicates) in normal state (KN1 and KN2), in treated with drug C (KC1, KC2) and treated with drug G (KG1, KG2)

Molm1: the cells are treated with two different drugs (carboplatin and gemcitabine). The file contains the interactions (measured in two replicates) in normal state (MN1 and MN2), in treated with drug C (MC1, MC2) and treated with drug G (MG1, MG2)

THP-1: the cells are treated with LPS. The file contains the interactions (measured in two replicates) in normal (untreated) state (THP1.nLPS.rep1, THP1.nLPS.rep3) and in LPS treated state (THP1.wLPS.rep1, THP1.wLPS.rep3)

We perform each experiment at least in two replicates.

Each interaction strength is measured by two parameters: supporting pairs (values in "_SuppPairs" columns) and p_value (values in "_pvalue" columns) in each replicate. The higher the supporting pair, the more likely that there is an interaction between the two regions. The p-value is calculated based on the distance between the loci: since it is more likely to detect interactions between regions that are close to each other in primary sequence, the p-value represent the probability of having the interaction with respect to the distance. The lower the p-value, the more likely that it is an interaction. FDR is calculated using the p-values.

For more information on some of the experiments, please see our publications below:

CMK, K562, Molm1: https://www.life-science-alliance.org/content/7/3/e202302244.abstract

Control HEKa: https://www.sciencedirect.com/science/article/pii/S009167492031410X

THP-1: https://www.sciencedirect.com/science/article/pii/S0008874920303087

Initial Insights from the Dataset
Your dataset contains 37,628 chromatin interactions, each with multiple features related to genes, chromosomal locations, interaction distances, and treatment conditions.

Key Observations:

Chromatin Interactions & Genomic Features
The dataset maps gene-gene interactions using chromosomal locations.
Columns like Feature_Chr, Feature_Start, Interactor_Chr, Interactor_Start, and Interactor_End describe where interactions occur.

Statistical Significance of Interactions
p-values (CG1_p_value, CG2_p_value, CC1_p_value, CC2_p_value, CN1_p_value, CN2_p_value) indicate the significance of interactions across conditions.
Smaller p-values suggest more statistically significant interactions.

Treatment-Specific Interactions
Normal, CarboplatinTreated, and GemcitabineTreated show the presence/absence of interactions under different drug treatments.
This can be used to find interactions specific to drug response.

Chromosomal Distance & Interaction Frequency
distance column represents the genomic distance between interacting elements.
Shorter distances may suggest regulatory interactions, while long distances may indicate long-range chromatin loops.

Next Steps:
Would you like to:

Visualize chromatin interactions by significance? (e.g., a heatmap of p-values)
Analyze treatment-specific interactions? (e.g., differentially significant interactions in drug-treated vs. normal conditions)
Explore interaction distance patterns? (e.g., a scatter plot of distance vs. SuppPairs)
Identify key genes involved in frequent interactions? (e.g., most common interacting genes)

I hope this email finds you well. Based on the discussion that we had in the meeting, I am reaching out regarding some more clarifications related to the genome dataset shared.

1. What exactly are the Supporting pairs and their threshold value? (For CG,CC,CN)
Supporting pairs are the number of read pairs supporting that particular interaction in each sample (experimentally obtained). I believe I used a supporting pair threshold of five for the files I gave you.

2. What are the P-Values and their threshold values?
The p-values indicates how likely to obtain the observed number of supporting pair for the interaction given the linear distance between the interacting regions. If the distance is short, the interaction will have many supporting pairs but that does not mean that it is a functional interaction; rather that they are close in the linear space that is why it has high supporting pairs, such cases will get a high p-value (weak). However, the p-value will be low for the interactions have high supporting pairs given the long linear distance between them. The p-values are calculated based on the background interaction frequency of negative control regions. For more information please read our article here. If the p-value is low, that means that the interaction cannot be explained simply by short distance between the interacting regions, there might be some mechanism that brings those two regions together in 3D space. I believe I gave you three files with three different p-value thresholds (0.05, 0.005,0.0005).

3. Does CG1 and CG2 indicate two iterations?
Yes, each experiment is performed twice.

4. and Should both the values satisfy the Thresholds in order to get output values as 0/1?
Yes, exactly, the supporting pairs and p-values should satisfy the conditions in both replicates to appear in this file.

5. Is there any threshold(Min/Max) value for Distance?
No, but the minimum distance is 1000 bases.

6. What does this "Noflnts" column indicate? (It is the sum of (normal/Carboplatin/Gemcitabine) columns)
How many samples this particular interaction is observed.

7. what does PP/PD indicate in "IndGroup" Column
PP: promoter-promoter interaction
PD: promoter-distal interaction

What Are Cell Lines?

A **cell line** is essentially a population of cells that scientists have learned how to keep alive and growing in the lab indefinitely[1][2]. Imagine if you could take some LEGO blocks and they could make copies of themselves forever - that's what a cell line does, but with actual living cells.

Why are cell lines important?

- They let scientists study diseases without using patients directly
- They're like having a "practice dummy" for testing new treatments

- They're consistent and reliable for experiments

Your dataset contains **six different cell lines**, each representing different types of human cells.

What Is Chromatin and Why Does It Matter?

DNA Folding: The Ultimate Space-Saving Trick

Imagine you have a **6-foot-long piece of string** that you need to fit into a ping-pong ball, but you still need to be able to read specific parts of it quickly16. That's essentially what your cells do with DNA!

Your DNA is about **6 feet long** when stretched out, but it needs to fit inside a cell nucleus that's **10,000 times smaller**17. To do this, DNA gets wrapped around proteins called histones, creating a structure called **chromatin**1819.

Chromatin Interactions: The 3D Organization

Think of chromatin like a **3D filing system**2016. Different parts of your DNA need to "talk" to each other to control which genes get turned on or off. When two distant parts of DNA come close together in 3D space, they form what we call **chromatin interactions** or **chromatin loops**1819.

Why does this matter?

- **Gene regulation**: These interactions control which genes are active1820
- **Cell identity**: They help determine whether a cell becomes skin, blood, or brain17
- **Disease**: When interactions go wrong, it can cause cancer and other diseases

The Problem Statement and Research Context

Your dataset is investigating how **drug treatments change the 3D organization of DNA** in different types of human cells, particularly cancer cells. This research sits at the intersection of **cancer biology, epigenomics, and personalized medicine**

Cell Line	Cell Type	Origin	Cancer Status	Purpose
CMK	Megakaryoblast (blood cells)	Down syndrome patient with acute leukemia	Cancer	Study blood cancer drug response

Cell Line	Cell Type	Origin	Cancer Status	Purpose
Control HEKa	Keratinocytes (skin cells)	Normal human skin	Normal	Baseline comparison
GM12878	B lymphocytes (immune cells)	European female donor	Immortalized	Reference immune cell line
K562	Erythroleukemia (blood cells)	53-year-old woman with chronic leukemia	Cancer	Study blood cancer mechanisms
Molm1	Megakaryoblast (blood cells)	Blood cancer patient	Cancer	Validate CMK findings
THP-1	Macrophage-like (immune cells)	1-year-old boy with acute leukemia	Cancer	Study immune response

The Drug Treatments (The Cleaning Products)

Carboplatin + Gemcitabine Combination

Carboplatin 12:

- **What it is**: A platinum-based chemotherapy drug
- **How it works**: Think of it like a **wrecking ball for DNA**
- **DNA adducts**: It sticks to DNA and creates **DNA damage** (like putting superglue on important documents)
- **Apoptosis**: This damage triggers **programmed cell death** - the cell basically commits suicide when it realizes it's too damaged to function 34

Gemcitabine 567:

- **What it is**: A **fake building block** for DNA
- **Nucleoside analog**: It looks like a real DNA piece but it's actually a **molecular imposter** 7
- **Ribonucleotide reductase**: It blocks an enzyme that makes DNA building blocks 58
- **How it kills**: When cells try to use this fake piece to build new DNA, the construction stops and the cell dies

Why Use Both Together?

- **Synergistic effect**: Like using both a hammer and a screwdriver - they work better together than alone
- **Enhanced cytotoxicity**: They cause more cell death together than separately 910

LPS Treatment

LPS (Lipopolysaccharide) 1112:

- **What it is**: A piece of bacteria wall that tricks immune cells
- **TLR4 pathway**: It activates the cell's "bacterial alarm system" 1112
- **Immune response**: Makes immune cells think there's an infection and go into attack mode

The Hi-C Technology (How We Measure DNA Organization)

Think of Hi-C as **taking a snapshot of which rooms in the house are connected** 131415.

The Process Step-by-Step

Step 1: Crosslinking 1617:

- **Formaldehyde fixation**: Like spraying hairspray on a hairstyle to freeze it in place
- **What it does**: Locks DNA and proteins in their current 3D positions 1617

Step 2: Digestion 181920:

- **Restriction enzymes**: Molecular scissors that cut DNA at specific sequences 1819
- **Like cutting**: Imagine cutting up a tangled ball of yarn at specific colored threads

Step 3: Ligation 1319:

- **DNA ligation**: Gluing DNA pieces back together, but only the ones that were physically close 19
- **Biotinylation**: Adding molecular "tags" to mark the glued joints 1315

Step 4: Sequencing 1315:

- **Paired-end sequencing**: Reading both ends of each glued piece to see which distant parts of DNA were connected

Understanding the Measurements

Supporting Pairs

- **Simple definition**: How many times we saw the same DNA connection
- **Threshold of 5**: Like needing 5 witnesses to believe something really happened
- **Higher numbers**: More confidence that two DNA regions really interact

P-values

- **What it measures**: The chance that a DNA interaction happened by random luck
- **Distance correction**: Far-apart DNA pieces that interact are more interesting than nearby ones that naturally bump into each other
- **Lower p-values**: More likely to be a real, meaningful interaction

False Discovery Rate (FDR) 2122

- **The problem**: When testing thousands of interactions, some will look significant just by chance
- **FDR solution**: Controls how many of your "discoveries" are probably false alarms 2122
- **Thresholds** (0.1, 0.01, 0.001): Like different levels of pickiness - 0.001 is super strict, 0.1 is more relaxed

Distance Categories

- **Short distances (<50kb)**: Like rooms next to each other - not surprising they interact
- **Medium distances (50kb-1Mb)**: Like rooms on different floors - more interesting
- **Long distances (>1Mb)**: Like connecting the basement to the attic - very significant

Interaction Types

PP (Promoter-Promoter) 2324:

- **Promoters**: DNA regions that are like "ON switches" for genes 2325
- **PP interactions**: When two gene switches talk to each other to coordinate gene activity

PD (Promoter-Distal) 2324:

- **Distal elements**: DNA pieces far away that control genes (like **enhancers**) 2324
- **PD interactions**: When a gene switch connects to a distant controller

What Your Dataset Can Tell You

Drug Response Analysis

- **Before vs. After**: Which DNA connections are gained or lost when cells are treated?
- **Cell-type differences**: Do cancer cells and normal cells reorganize their DNA differently?
- **Drug mechanisms**: How do different drugs change DNA organization?

Clinical Implications

- **Personalized medicine**: Understanding which cell types respond best to which drug combinations
- **Drug resistance**: How cancer cells might reorganize their DNA to survive treatment
- **Biomarkers**: DNA interaction patterns that predict treatment success