Bioinformatics Methods and Applications
https://www.sciltp.com/journals/bma

Article

Statistically Dense Intervals in Binary Sequences with Applications to Assessing Local Enrichment in the Human Genome

Ben Galili [1,2,*], Ofri Kutchinsky [1], Shahar Mor [2] and Zohar Yakhini [1,2]

[1] Arazi School of Computer Science, Reichman University, Herzliya 4610101, Israel
[2] Computer Science, Technion-Israel Institute of Technology, Haifa 3200003, Israel
* Correspondence: bgalili@runi.ac.il

How To Cite: Galili, B.; Kutchinsky, O.; Mor, S.; et al. Statistically Dense Intervals in Binary Sequences with Applications to Assessing Local Enrichment in the Human Genome. *Bioinformatics Methods and Applications* **2026**, *1*(1), 3.

Received: 8 September 2025
Revised: 23 March 2026
Accepted: 31 March 2026
Published: 21 April 2026

Abstract: Statistical enrichment tools are highly useful in biological research. Current approaches to statistical enrichment in ranked or ordered lists are either limited to fixed thresholds or, as in GSEA and GOrilla, are limited to the list's suffix (prefix). These methods assess the extreme density of 1s on either side of the binary vector. Statistical significance can be assessed using, e.g., variants of the Wilcoxon Rank-Sum Test and the mHG statistic. In this work, we extend the mHG approach to address enrichment within any index interval of the binary vector. We define and partially characterize related distributions under a uniform null model. Our partial characterization yields useful bounds for extreme events. We provide a software tool to the community that implements the method in Python. Finally, we analyze several example use cases and describe the results. We show, for example, that lung cancer differential expression, comparing ADC to other types, is enriched in a region of Chromosome 3. This example illustrates a typical use case for imHG: assessing enriched intervals for any set of genes of interest. We provide a Python implementation, imHG, for finding and reporting enriched genomic intervals with any given list of genes of interest.

Keywords: statistical enrichment; minimum hypergeometric; genomic location; chromosome

1. Background

Statistical enrichment analysis is a fundamental tool in biological research that identifies over-represented biological properties within a dataset. Standard approaches often require the user to partition a list of sequences or elements into two fixed sets (e.g., "target" vs. "background") or to select an arbitrary cutoff (e.g., the top 100 differentially expressed genes). Determining this partition is often challenging and can bias results if the optimal threshold is unknown.

To address this, "threshold-free" statistical approaches were developed to discover rank-imbalanced elements without committing to a fixed cutoff. The minimum Hypergeometric (mHG) statistic, introduced in [1], assesses enrichment by searching for the optimal partition at the top of a ranked list. The output of mHG consists of the optimal partition and the associated test statistic value. This approach was originally demonstrated for identifying motifs in DNA sequences.

Several additional widely used tools have also adopted similar rank-based strategies to assess enrichment at the extremes of a list. For example, GSEA [2,3] analyzes the enrichment of gene sets—groups sharing biological functions, regulation, or chromosomal locations—within a ranked list of genes. Similarly, GOrilla [4] is a web-based application that identifies enriched Gene Ontology (GO) [5] terms and standardized functional descriptions of gene products at the top of a ranked gene list using the mHG statistic. DRIMUST [6] combines suffix trees with a ranked list approach to find motif enrichment. While powerful, these methods generally restrict their search to the prefix (top) or suffix (bottom) of the list. They effectively assume that the signal of interest is anchored to one of these extremes.

However, biological signals are not always distally located in an ordered list. In many cases, a signal may be concentrated in a specific internal window or "interval". For instance, when genes are ordered by chromosomal position, a cluster of co-regulated genes might appear in the middle of a chromosome rather than at the telomeres. Existing prefix-based tools may miss such local enrichments entirely. In single-cell transcriptomics, as another example, cells are often ordered by 'pseudotime' to model developmental trajectories [7]. Critical differentiation events often occur in intermediate states, effectively 'intervals' along this pseudotime axis, which standard prefix-based enrichment may miss. Similarly, in time-series gene expression [8], regulatory programs may be active only during specific temporal windows. Other approaches, such as SAGO [9], have begun to incorporate spatial arrangements into enrichment analysis, but a general-purpose statistical framework for interval enrichment is an important addition to the analysis toolbox.

In this work, we extend the scope of statistical enrichment to cover enrichment in any continuous interval of indices in an ordered list, generalizing the mHG approach beyond prefixes and suffixes. This is particularly valuable for investigating the co-localization of genomic elements within linear genomic intervals. While eukaryotic genomes fold into complex 3D structures (e.g., Topological Associating Domains, TADs [10]), linear proximity often correlates with regulatory domains and functional clustering [11]. Identifying 1D dense intervals provides a computationally efficient foundation for detecting co-localization before applying complex 3D models. Co-localization in this context has been studied in previous works, including [9, 12–15].

We introduce imHG (interval mHG), a novel algorithm that efficiently identifies significantly enriched intervals within a given ordered list. We provide a methodology to bound the *p*-values of these observed intervals under a uniform null model, addressing the multiple-hypothesis correction required for flexible interval selection. We provide a Python implementation of the tool and demonstrate its utility on simulated data and the GSEA MSigDB human gene set bundle (https://github.com/YakhiniGroup/imhg). Finally, we apply the method to differential expression datasets from lung and breast cancer cohorts to uncover densely statistically significant genomic intervals of varying sizes. Notably, we observed an enrichment of differentially expressed genes on chromosome 3 when analyzing Adenocarcinoma vs. other subtypes.

2. Methods

2.1. Definitions and Notations

A standard threshold-based approach to address statistical enrichment in genomic location would use a fixed target set. For example, the top 100 differentially expressed genes or the genes residing at a given distance from the center of Chromosome 7. In this work, we developed an approach that does not require a fixed target set.

Adopting the notation established by [1] and, for completeness, restating some of the definitions, we define N as the total number of elements in an input binary vector v. B represents the number of elements out of the above N that have a common property of interest, to be termed the foreground elements (the 1s in the vector v). B is sometimes called the weight of v [16, 17]. We use n to denote the size of the target set. For mHG, this is the size of the prefix or suffix under consideration. We use X to denote a random variable counting the number of occurrences of foreground elements in the target set. The null model, as originally described in [1], assumes a uniform distribution over all $\binom{N}{B}$ possible occurrence vectors. Under this model, we have:

$$Prob(X = b) = HG(b; N, B, n) = \frac{\binom{n}{b}\binom{N-n}{B-b}}{\binom{N}{B}} \tag{1}$$

where HG represents the parameterized probability mass function of the hypergeometric distribution. The hypergeometric (right) tail, HGT, is the probability of observing b or more occurrences:

$$Prob(X \geq b) = HGT(b; N, B, n) = \sum_{i=b}^{min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}} \tag{2}$$

In our threshold-free situation, we don't have a strict value or identity for the target set. The studies and methods of [1–4, 6, 18] employ a strategy that seeks a partition for which the statistical enrichment is the most significant and compute the enrichment under that particular partition.

Formally, consider the vector λ that represents ranked or ordered elements and a binary labeling that embodies the property of interest or the foreground elements:

$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \{0, 1\}^N$$

The mHG statistic, a random variable over our null space, consisting of a uniform distribution over a subset of $\{0,1\}^N$, is defined as:

$$mHG(\lambda) = \min_{1 \le n < N} HGT(b_n(\lambda); N, B, n) \tag{3}$$

where $b_n(\lambda) = \sum_{i=1}^{n} \lambda_i$.

While the mHG approach, as well as those of similar methods [2–4,6], address the threshold selection challenge described above, the definition of mHG introduces a new challenge. For mHG, a strong initial anchor is assumed. That is, every partition must start from the first index of the given ordered list. However, we often also want to address cases where the most significant enrichment doesn't necessarily start at the beginning of the observed ordered list. Rather, a different interval can be considered, one that doesn't start at index 0. Such is the case, for example, when the indices represent genomic locations or time points in some process.

To overcome the limitation described above, we developed a novel approach termed imHG (interval mHG). Using the same binary labeling λ defined above, the imHG statistic/score, a random variable over our null space, is defined as:

$$imHG(\lambda) = \min_{1 \le n < k < N} HGT(b_{n,k}(\lambda); N, B, k-n+1) \tag{4}$$

where $b_{n,k}(\lambda) = \sum_{i=n}^{k} \lambda_i$.

In words, the imHG score reflects the tail probability of seeing the observed density of 1's at an interval of length $k-n+1$ at location n of the list, under the null assumption that all configurations of 1's in the vector λ are equiprobable. Figure 1 illustrates the definition.

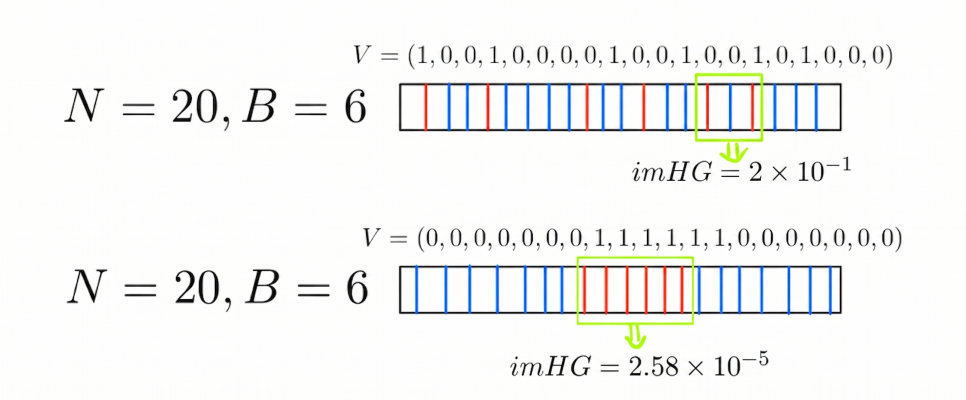

Figure 1. imHG values for two different example vectors and the respective intervals in which the optimum is attained.

To calculate the imHG score for an observed ordered list, we iteratively scan through all the possible (n,k) indices where $1 \le n < k \le N$.

The imHG flexible choice of cutoff introduces a multiple-testing complication and therefore necessitates computing an exact *p*-value or a useful bound. Let Λ be the space of all binary label vectors with B 1's and $N-B$ 0's. Denote $\Lambda = \{0,1\}^{(N-B,B)}$.

Given a vector $\lambda_0 \in \Lambda$, assume that we calculate the observation imHG score, $imHG(\lambda_0) = s$. We formally define the *p*-value of s as:

$$imHG.pval(s) = Prob(imHG(\lambda) \le s) \tag{5}$$

where the probability is under a uniform distribution of all $\binom{N}{B}$ vectors in Λ.

A naïve way to calculate a *p*-value of a given $\lambda_0 \in \Lambda$ is to calculate the imHG scores for all $\lambda \in \Lambda$ exhaustively. This, however, is not a feasible solution due to its exponential nature.

To overcome this problem, we present a useful bound that approximates the true *p*-value.

2.2. Bounding p-Values of Extreme Density Events

Let N be the total number of elements of which B possesses a property of interest. Let Λ be the space of all binary label vectors with B 1's and $N-B$ 0's, that is, $\Lambda = \{0,1\}^{(N-B,B)}$ as denoted above. We consider the null model wherein all $\lambda \in \Lambda$ are equiprobable.

For $\lambda \in \Lambda$, let $r = mHG(\lambda)$ and let $P_1(r)$ be its corresponding *p*-value, that is, $P_1(r) = prob(mHG(\omega) \le r)$.

Let $s = imHG(\lambda)$ and let $P_2(s)$ be its corresponding *p*-value, that is, $P_2(s) = prob(imHG(\omega) \le s)$.

Claim 1. $\forall s \in [0,1]$, $P_2(s) \le N * P_1(s)$

Proof. For a given $\lambda \in \Lambda$, let $s := imHG(\lambda)$.

Define $A, B \subseteq \Lambda$ by:

$$A = \{a : imHG(a) \leq s\}$$

$$B = \{b : mHG(b) \leq s\}$$

Note that $B \subseteq A$.

Let $a \in A - B$. By the definition of imHG, $\exists 0 \leq i < j < N$ s.t. $HGT(b_{i,j}(a); N, B, j - i + 1) \leq s$ where, as above, $b_{i,j}(a) = \sum_{k=i}^{j} a_k$. Moreover, since $a \notin B$, we know that $i > 0$.

Using a shift operation, we will now construct a new vector with the same statistical significance that belongs to the set B. Intuitively, according to HGT properties and by using a shift operation, we will prove that any vector in A has a corresponding vector in B, in which the enrichment of 1's starts at index 0.

Consider $i, j, 0 < i < j < N$, so that $HGT(b_{i,j}(a); N, B, j - i + 1) \leq s$.

Further consider the vector $\sigma = shift(a, i)$ defined by:

$$\begin{aligned}
\sigma(0) &= a(i) \\
\sigma(l) &= a(l + i), 1 \leq l \leq j - i + 1 \\
\sigma(l) &= a(l + i), j - 1 \leq l + i < N \\
\sigma(l) &= a(l - j - 1), j + 1 \leq l < N
\end{aligned}$$

The presented shift operation performs a cyclic rotation to the given vector starting from index i.

We then have:

$$HGT(b_{i,j}(a); N, B, j - i + 1) = HGT(b_{0,j-i}(\sigma); N, B, j - i + 1) \tag{6}$$

By the definition of the shift operation presented above, we can conclude that $\forall v \in A - B, \exists v' \in B$, such that $v' = shift(v, i)$ where i is the start index of the enriched interval.

Consider the inverse shift operation, $shift^{-1}$, which shifts any given vector in a cyclic manner to the right. Further consider a vector $b \in B$ for which we create $N - 1$ shifted replicates as follow: $b_\mu = shift^{-1}(b, \mu), 1 \leq \mu \leq N - 1$. According to (6) we can conclude that every $b_\mu \in A$. Therefore, there are $N - 1$ vectors in set A which originated in the vector b.

We can, therefore, conclude every vector $a \in A$ has a corresponding shifted vector $b \in B$, which has the same statistical significance structure that originates at index 0.

For that reason, $|A| \leq N * |B|$. □

Corollary 1. *Let* $s \in [0, 1]$*, then* $s \leq P_2(s) \leq N * P_1(s)$

To see that this is a useful bound, note that $P_1(s)$ was well characterized in previous work ([1]).

3. Results

3.1. Simulation Analysis

To demonstrate the utility and power of imHG in detecting dense intervals in a clean environment, we performed a series of simulation studies using synthetic binary vectors. We start with vectors filled with zeros, each of length $N = 1000$. For each vector, we uniformly selected an interval location to represent a dense area. Within this interval, we converted each zero to a one independently with a probability of P_{in}. Outside the interval, we flipped each zero to a one with a lower probability of P_{out}, where $P_{out} < P_{in}$. We ran 1000 independent simulations for each combination of interval length and signal strength (density of 1s within the target interval). Figure 2 illustrates the power of the algorithm, defined as the probability of detecting a significant interval ($p < 10^{-4}$) as a function of signal strength. As expected, detection power increases with interval density and length.

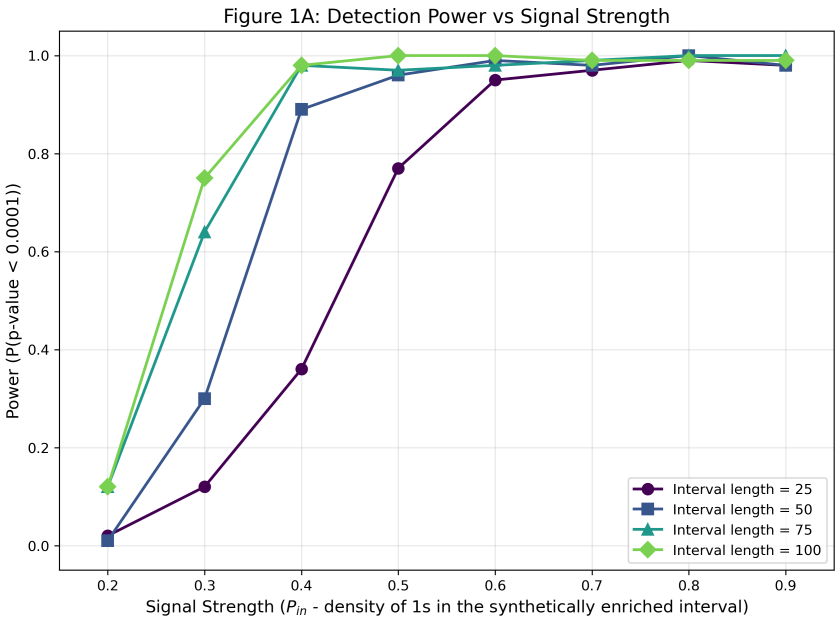

Figure 2. Detection power of the imHG algorithm as a function of signal strength and interval length. The plot demonstrates the algorithm's power to identify statistically dense intervals within binary sequences. Detection power (y-axis) is defined as the probability of identifying a significant interval at a threshold of $p < 0.0001$. The x-axis represents the signal strength, corresponding to the density of 1s (P_{in}) embedded within the target interval (P_{out} is fixed to 0.1). The results indicate that detection power consistently improves as both the signal density increases and the interval length expands (from 25 to 100 elements, within a total length of $N = 1000$).

3.2. imHG Software

We provide a Python implementation package, called imhg, for calculating any vector's $imHG$ score and a bound on its corresponding *p*-value. Our implementation comprises three main components. The first component contains the code that computes the $imhg$ score for a given vector. The second component involves calculating the mHG *p*-value. This value is calculated using a dynamic programming approach first proposed in [1]. Eventually, our final *p*-value bound is calculated by using the mHG *p*-value obtained above and applying Claim 1. The third component presents multiple examples in which we applied our implementation to public datasets to demonstrate different use cases. To apply our statistical tool to biological datasets, we need a method to convert gene symbol names to a chromosome-based binary map on which we will ultimately perform our process. We created a set of 24 binary vectors based on the NCBI GRCh38.p13 assembly to obtain that binary map for human genes. Every component of such a vector represents a gene located on the corresponding chromosome. According to the NCBI database, their indices are determined by their positions on the chromosome, from the outer end of the short arm to the outer end of the long arm (without reference to the centromere). Upon examination of a set of genes of interest, entries representing the genes belonging to the set will contain the digit 1, and the remainder will contain the digit 0. In our package, we offer two options for analyzing input results. The first was applied to the GSEA Human gene set bundle, supporting the use case of analyzing any collection of gene subsets. In this analysis, we receive a dictionary that includes multiple gene sets as input. Each input gene set is associated with a corresponding foreground list of gene symbols. In this type of analysis, no user or dynamic-foreground gene selection is enabled; that is, the imHG tool uses all foreground genes to identify enriched intervals. The second option supports the analysis of, e.g, differential gene expression results. In this approach, the user provides a list of pairs. Each pair consists of the measured gene name and its corresponding differential expression *p*-value (or other quantity of interest). We use this list to form a "new mapping" of chromosomes. The chromosome map is adjusted to include only genes measured in the analyzed study, regardless of their *p*-value. Then, the user provides the attribute set size, representing the number of genes the user defines as positive. The genes are selected in ascending order of *p*-value. Finally, the results are saved into a file, which can later be loaded using the pandas library for further analysis. Further explanation of how we handled and classified positive genes is provided in the following sections.

3.3. Datasets

In the following two sections, we present results from two different datasets. The first is from the Gene Set Enrichment Analysis (GSEA) Human gene set bundle [2,3], available at https://www.gsea-msigdb.org/gsea/index.jsp

(accessed on 21 January 2024).

The second is from the National Center for Biotechnology Information (NCBI) gene expression datasets, which are available at https://www.ncbi.nlm.nih.gov/gene (accessed on 13 March 2024).

3.4. Genome Location Enrichment for GSEA Sets

In this section, we demonstrate how to identify enriched intervals using GSEA's Human gene set bundle, which includes all gene sets in MSigDB. An example of this process is available in our GitHub project. To begin, we generated a list of pairs representing all coded gene names for each of the 24 chromosomes in advance. Next, we downloaded the JSON bundle, which contains a tested gene set name and a corresponding Python dictionary with several metadata attributes for each gene set. We extracted the foreground gene symbols for each gene set, reducing the size of the original JSON bundle. We then created a new chromosomal map, which is a binary vector representing whether the gene is foreground (1) or background (0), according to the analyzed gene set. This resulted in a list of 24 binary vectors, where each vector corresponds to a human chromosome and each entry indicates the location of a gene on that chromosome. For every such binary vector associated with a GSEA set, we applied our $imHG$ calculation method and calculated its *p*-value bound. Due to the large volume of hypothesis tests ($>700,000$ searches), we applied a stringent filtering criterion, discarding enriched intervals with a *p*-value bound less significant than 10^{-10}. For the more significant *p*-values, we saved additional metadata, including the interval's length, the number of foreground genes, the chromosome's total foreground gene count, the chromosome's total gene count, the sum of foreground genes for the set, the set's number of enriched genomic intervals, and the lift. We then save the results to an Excel file, which can be accessed and modified as needed.

In this process, we analyzed 33,591 different sets. For each set, we performed 24 searches, one per chromosome, yielding a total of 772,593 searches. We have identified 681 of them exhibiting genomic location enrichment with an imHG *p*-value less than 10^{-10}. Table 1 presents the most significant results obtained from the analysis.

Table 1. Top Genomic location enrichment for GSEA subsets. The top 10 enrichment results are reported from a total of 681 results with a $-\log$ *p*-value greater than 10. Please see the supplementary for full data. Set partial names are as follows: S1: GOBP-PRODUCTION-OF-MOLECULAR-MEDIATOR-OF-IMMUNE-RESPONSE; S2: GOMF-OLFACTORY-RECEPTOR-ACTIVITY; S3:GOBP-SENSORY-PERCEPTION-OF-SMELL; S4: GOBP-CELL-CELL-ADHESION-VIA-PLASMA-MEMBRANE-ADHESION-MOLECULES; S5: GOCC-DNA-PACKAGING-COMPLEX; S6: GOCC-NUCLEOSOME; S7: GOMF-ODORANT-BINDING; S8: KEGG-OLFACTORY-TRANSDUCTION; S9: GOCC-PROTEIN-DNA-COMPLEX. Column Chr. represents the chromosome in which the top enrichment was observed. The number in brackets indicates the number of additional enriched intervals for this set. Column $\sum B$ represents the total number of foreground genes across all chromosomes for the set. The Lift is calculated by: $\frac{bN}{nB}$.

Set	Chr.	*p*-Value	(n,b,B,N)	ΣB	$\frac{b}{\Sigma B}$	Lift
S1	2	9.035×10^{-71}	(45, 41, 57, 3100)	329	12.46	49.55
S2	1(+6)	8.624×10^{-67}	(59, 42, 63, 4420)	342	12.28	49.94
S3	1(+5)	5.437×10^{-65}	(59, 42, 67, 4420)	456	9.21	46.96
S4	5	1.201×10^{-63}	(41, 41, 67, 2025)	280	14.64	30.22
S5	6(+1)	2.769×10^{-61}	(156, 54, 57, 2515)	164	32.92	15.27
S6	6(+1)	2.769×10^{-61}	(156, 54, 57, 2515)	133	40.60	15.27
S7	11(+1)	9.960×10^{-56}	(101, 50, 70, 2609)	122	40.98	18.45
S8	11(+5)	2.991×10^{-55}	(60, 53, 160, 2609)	388	13.66	14.40
S9	6(+1)	1.734×10^{-54}	(156, 54, 64, 2515)	225	24	13.60
S3	11(+5)	9.455×10^{-54}	(49, 49, 191, 2609)	456	10.74	13.65

The left panel in Figure 3 depicts a pie chart of chromosomes in which the top intervals were found. In the right panel, we see an adjusted view of the same distribution, accounting for chromosome length. Note that whenever we mention the length of an interval or a chromosome, we refer to the number of genes residing therein. We use the following operations to scale. For every chromosome, we compute $\frac{chromosome-length}{\Sigma chromosomes-legnths}$. Then, we divide the number of chromosome-enriched intervals by their relative length, as calculated above. We then normalize this adjusted number to get percentages. In both the normalized and non-normalized pie charts (Figure 3), we observe a relatively high percentage of enriched intervals located on chromosomes 5, 6, and 9. Notably, the strong signal on Chromosome 6 corresponds to the Major Histocompatibility Complex (HLA) region. Since many MSigDB gene

sets are immune-related, identifying this known gene-dense region serves as a biological positive control for the imHG algorithm. Additionally, we notice that the relative frequency of intervals on chromosome 22 doubles after taking into account the lengths. In Figure 4, we can see a histogram of the length of the 681 enriched intervals.

In Figure 5, we depict an ideogram of selected enriched sets.

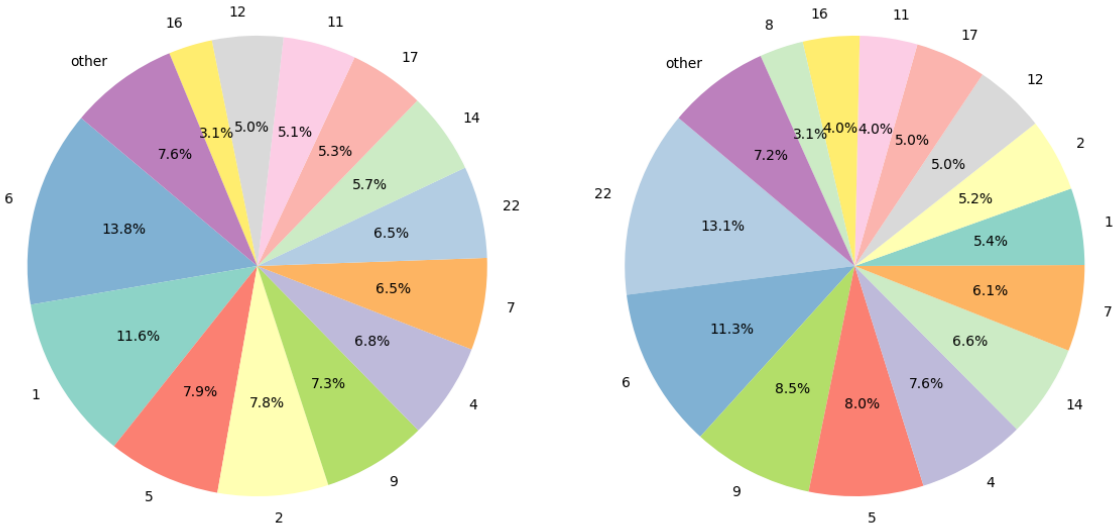

Figure 3. On the (**left**), the chromosomal distribution of GSEA enriched intervals with $-\log(p\text{-value}) > 10$. The (**right**) represents the same data scaled by chromosome length.

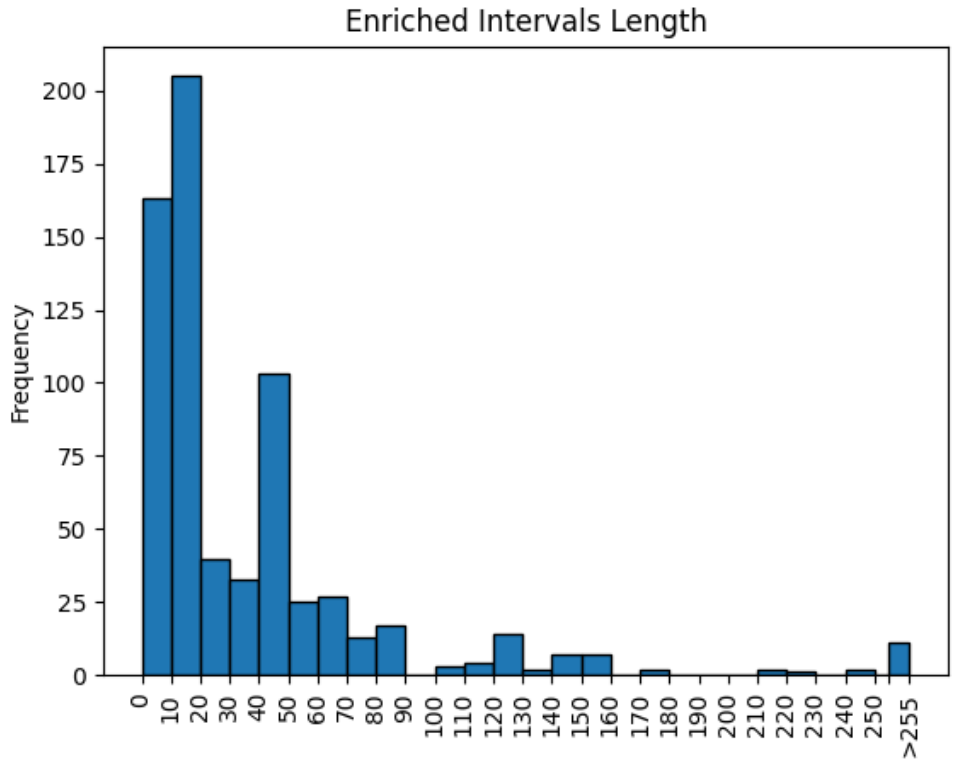

Figure 4. The distribution of GSEA enriched intervals lengths. All enriched intervals in this analysis have a p-value smaller than 10^{-10}, as described at the beginning of this section. Note that all lengths greater than 250 are represented by a single bin.

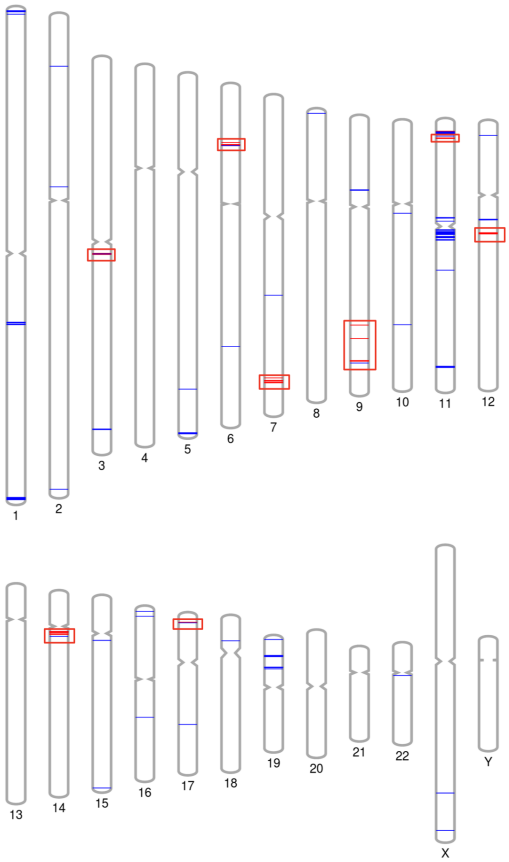

Figure 5. Ideogram for the analysis of the set: REACTOME-OLFACTORY-SIGNALING-PATHWAY, which has 8 significant intervals. The blue lines represent the foreground genes, and the red lines represent the foreground genes in the enriched interval (surrounded by a red box).

3.5. Genome Location Enrichment of Differentially Expressed Genes

3.5.1. Lung Cancer Differential Expression

Lung cancer continues to be the leading cause of cancer-related deaths, accounting for approximately 18% of all cancer deaths, [19]. The two main types of lung cancer are Non-small cell lung cancer (NSCLC) and Small cell lung cancer (SCLC), with NSCLC constituting about 85% of cases. Among NSCLC subtypes, Adenocarcinoma (ADC) is the most prevalent. It is commonly associated with individuals who either smoke or have a history of smoking, although it can also occur in non-smokers, [20].

For our research, we analyzed data from six separate studies that investigated lung cancer mRNA gene expression. Details regarding sample size, platform, and preprocessing for each study are provided in Table 2. These studies identified high-risk factors, efficient gene signatures, [21], and various tumor characteristics. To conduct our analysis, we used these six datasets and categorized them by tumor subtype, focusing on ADC vs. other subtypes. Five of the studies included 54,675 measured genes, and the last included 22,215. Using the GEOparse Python package, we converted each gene ID to its gene symbol. Records that did not map into a single symbol name were dropped. After translating all symbol names and removing corrupted or duplicated records, we created a separate chromosomal map for each series. To identify enriched intervals, we first converted the chromosome map to a binary representation. To do so, we needed to determine what would make the 0's and 1's.

We assign 1 (foreground) to differentially expressed genes. We applied our imHG process using different thresholds to determine differential expression. For a given threshold of T, we chose the T genes with the smallest p-values. In the enrichment process, these genes would serve as the foreground genes. According to the described analysis, we also needed to define a threshold for reporting imHG p-values. Given the characteristics of our analysis, we set the threshold to 10^{-4}. That is, we only report intervals with observed $imHG$ p-value more significant than 10^{-4}. We run this process for different thresholds (T). Table 3 presents the smallest resulting imHG p-values, where T varies across the different sets. We then illustrated in Figure 6 the ideogram, obtained by using Ensemble [22] and PhenoGram Plot [23], of Series GSE37745 [24–28], with $T = 300$ differentially expressed genes (foreground). We want to emphasize two points arising from this ideogram. First, there is a significant enrichment of the interval

on chromosome 3 near the p telomere. Furthermore, a very similar enrichment (in terms of location) was observed when analyzing the data in [29–32]. Our current findings are consistent with prior studies, including [33,34], which established a connection between genes on chromosome 3p and lung cancer pathogenesis. Our findings may suggest that these genetic elements exhibit specificity towards adenocarcinoma (ADC) within lung cancer.

In Table 4, we present the top results for the analysis, with T fixed at 300, 600, or 1000. As we can see, using the same threshold values yields slightly different results (with some intervals left out due to the threshold selection).

Table 2. Summary of Lung Cancer datasets used for differential expression analysis. "Platform" indicates the microarray technology used. Subtype counts indicate the specific samples selected for the "ADC vs Other" comparison.

Study	GEO ID	Platform	Total n	Subtypes (n)	Preprocessing
Zhu et al.	GSE14814	GPL96	133	ADC (71), Other (62)	RMA
Hou et al.	GSE19188	GPL570	91	ADC (45), Other (46)	MAS5.0
Rousseaux	GSE30219	GPL570	307	ADC (85), Other (222)	RMA
Botling	GSE37745	GPL570	196	ADC (106), Other (90)	RMA
Der et al.	GSE50081	GPL570	181	ADC (127), Other (54)	RMA
Lee et al.	GSE8894	GPL570	138	ADC (63), Other (75)	GCRMA

Table 3. Top genomic location enrichment for ADC vs other LUCA differential expression. For every differential expression threshold (T), we report the most significant enrichment as long as it is better than 10^{-4}. The quantities (n, b, B, N) are defined in previous sections. We selected the best T for every dataset analyzed. In each analysis, we examined 16 possible values for T. Therefore, one may apply a Bonferroni correction of 16 to the p-values. Note that the reported p-value is, in fact, the bound from Claim 1. The datasets are: Zhu et al. GSE14814 [35], Hou et al. GSE19188 [29], Rousseax et al. GSE30219 [30], Botling et al. GSE37745 [24], Der et al. GSE50081 [31], Lee et al. GSE8894 [32].

Series	Chr.	p-Value	(n, b, B, N)	T
Zhu et al. [35]	3	1.069×10^{-09}	(78, 33, 77, 663)	800
Hou et al. [29]	3	7.955×10^{-08}	(130, 36, 89, 1030)	1000
Rousseax et al. [30]	3	1.154×10^{-07}	(130, 32, 73, 1030)	600
Botling et al. [24]	3	3.686×10^{-19}	(303, 77, 103, 1030)	600
Der et al. [31]	3	6.628×10^{-15}	(121, 49, 117, 1030)	1000
Lee et al. [32]	3	5.651×10^{-19}	(216, 47, 59, 1030)	380
Botling et al. [24]	14	4.436×10^{-05}	(7, 4, 5, 554)	320

Table 4. Genomic location enrichment for ADC vs other LUCA differential expression. Top 300, 600 and 1000 genes.

Series	Chr.	p-Value	(n, b, B, N)	T
Zhu et al.[35]	3	1.350×10^{-06}	(68, 19, 38, 663)	300
Zhu et al. [35]	3	2.904×10^{-08}	(78, 27, 58, 663)	600
Zhu et al. [35]	3	9.227×10^{-09}	(78, 35, 92, 663)	1000
Hou et al. [29]	3	1.210×10^{-05}	(122, 25, 58, 1030)	600
Hou et al. [29]	3	7.955×10^{-08}	(130, 36, 89, 1030)	1000
Rousseax et al. [30]	3	2.561×10^{-05}	(130, 20, 38, 1030)	300
Rousseax et al. [30]	3	1.154×10^{-07}	(130, 32, 73, 1030)	600
Rousseax et al. [30]	3	1.254×10^{-07}	(175, 48, 111, 1030)	1000
Botling et al. [24]	3	9.026×10^{-16}	(135, 39, 66, 1030)	300
Botling et al. [24]	14	4.436×10^{-05}	(7, 4, 5, 554)	300
Botling et al. [24]	3	3.686×10^{-19}	(303, 77, 103, 1030)	600
Der et al. [31]	3	1.405×10^{-08}	(127, 26, 48, 1030)	300
Der et al. [31]	3	1.523×10^{-12}	(120, 38, 81, 1030)	600
Der et al. [31]	3	6.628×10^{-15}	(121, 49, 117, 1030)	1000
Lee et al. [32]	3	6.216×10^{-19}	(216, 43, 51, 1030)	300
Lee et al. [32]	3	9.754×10^{-10}	(195, 58, 87, 1030)	600
Lee et al. [32]	3	5.594×10^{-10}	(218, 71, 126, 1030)	1000

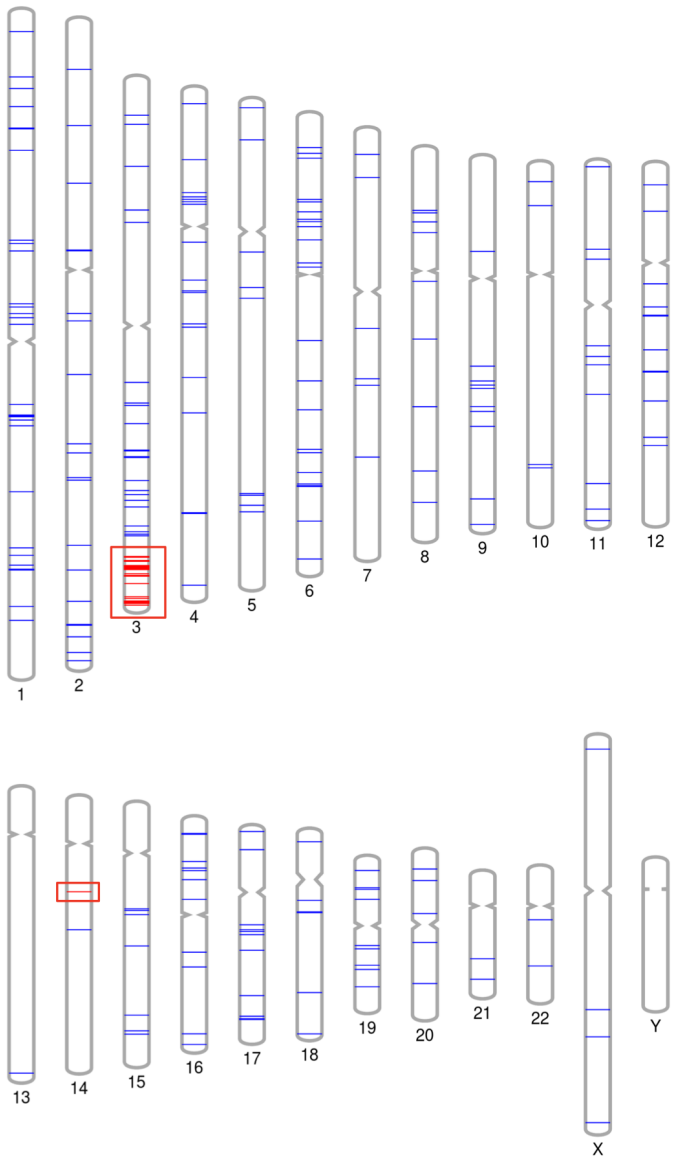

Figure 6. Lung cancer data by Botling et al. [24]. Genomic location enrichment for the 300 most differentially expressed genes, comparing ADC to other types.

3.5.2. Breast Cancer Differential Expression

A similar analysis was conducted on breast cancer data, see Table 5 for the datasets' details. This study utilized publicly available expression data from breast cancer cohorts. The expression data were sourced from various databases, including Gene Expression Omnibus, the European Genome-phenome Archive, ArrayExpress, and TCGA data portals (see data availability for more details). Each study provides information about the characteristics of the breast cancer tumor subtypes. The breast cancer subtypes, Luminal A, Luminal B, Her2-enriched, Basal-like, and Normal-like, were distinguished based on gene expression signatures (e.g., [36–38]). To analyze the data, we split it by tumor subtype, specifically focusing on Luminal A vs. other subtypes. A chromosome map was derived for each set of differentially expressed genes, as was done for the lung cancer analysis. In this analysis, we observed fewer significant signals than in the lung cancer analysis; see Table 6 for the top genomic location enrichment for LumA vs. other. The scarcity of significant intervals suggests that the transcriptomic landscape of Luminal A breast cancer is not heavily driven by localized, 1D chromosomal clustering. Instead, its gene expression profile may rely on broader, trans-acting regulatory networks, demonstrating imHG's ability to highlight differences in spatial genomic architectures across cancer types.

Table 5. Summary of Breast Cancer datasets used for differential expression analysis. "Platform" indicates the microarray or sequencing technology used. Subtype counts indicate the specific samples selected for the "Luminal A vs. Other" comparison.

Study	Accession ID	Platform	Total n	Subtypes (n)	Preprocessing
Aure et al. [39]	GSE58215	Agilent SurePrint G3	283	LumA (121), Other (162)	Log2, Quantile norm.
Haukaas et al. [40]	GSE58215	Agilent SurePrint G3	228	LumA (99), Other (129)	Log2, Quantile norm.
Tinholt et al. [41]	GSE58215	Agilent SurePrint G3	152	LumA (65), Other (87)	Log2, Quantile norm.
Ankill et al. [42]	GSE58215	Agilent SurePrint G3	282	LumA (121), Other (161)	Log2, Quantile norm.
Tekpli et al. [43]	GSE86948/GSE58215	Agilent SurePrint G3	343	LumA (146), Other (197)	Log2, Quantile norm.
Minn et al. [44]	GSE5327	Affymetrix HG-U133A	58	LumA (20), Other (38)	MAS5.0, Log2
Wang et al. [45]	GSE2034	Affymetrix HG-U133A	286	LumA (129), Other (157)	MAS5.0, Scaled

Table 6. Top genomic location enrichment for LumA vs other BRCA differential expression. For every threshold (T), we report the most significant enrichment as long as it is better than 10^{-4}.

Series	Chr.	*p*-Value	(n, b, B, N)	T
Aure et al. [39], Haukaas et al. [40] Tinholt et al. [41], Ankill et al. [42]	13	2.5×10^{-05}	(7, 7, 16, 160)	800
Tekpli et al. [43]	3	9.4×10^{-05}	(133, 14, 15, 536)	180
Minn et al. [44], Wang et al. [45]	9	9×10^{-06}	(9, 8, 25, 339)	800

4. Discussion

In this work, we introduce a novel algorithm for computing p-value bounds for extreme-density events. This bound enables us to assess different events and determine the likelihood of their occurrence under a null model. By applying our algorithm to several datasets, we have identified gene sets with extreme-density events. We hope that our software tool can help researchers discover meaningful events, leading to more novel scientific findings and insights.

It is important to note that, although our bound is not mathematically tight, it can be considered a good practical solution. For example, in our analysis, we applied our algorithm to human-collected gene sets to find enriched intervals in the human genome. Since every chromosome in the human genome comprises no more than 4420 genes, we can conclude that our bounded solution might be, in the worst case, about 10^3 higher than the actual tight solution. One might consider it a significant error; however, it is important to note that for extreme events, such as those presented above, an error of three orders of magnitude for an event with a 10^{-40} likelihood of occurring is almost meaningless. Moreover, we previously presented an upper bound on the actual p-value, but we can also obtain a lower bound as follows:

$$s \leq P_1(s) \leq P_2(s) \leq N * P_1(s) \tag{7}$$

where, for $\lambda \in \Lambda$, $s = mHG(\lambda)$, $P_1(s) = Prob(mHG(\omega) \leq s)$,
and $P_2(s) = Prob(imHG(\omega) \leq s)$.

These bounds are central to this work's contribution, as they enable the efficient and effective use of imHG in practical analysis.

It is also important to note that imHG is designed explicitly to detect localized structural density. Therefore, it evaluates each chromosome (or sequence) independently and does not mathematically combine scores across different chromosomes into a single global metric. If a biological pathway is driven by a highly dispersed network across multiple chromosomes (e.g., trans-acting factors), imHG is not designed to yield a high global rank for that set. Instead, its primary utility lies in identifying specific 1D regional hotspots of activity. Furthermore, the application of imHG to distinct cancer cohorts highlighted its ability to distinguish between different spatial genomic architectures. The strong, localized 1D enrichments observed in lung adenocarcinoma contrast sharply with the sparse localized signals in Luminal A breast cancer. As noted in the results, this scarcity can conceivably reflect a biological reality: Luminal A breast cancer relies heavily on broader, trans-acting regulatory networks (such as the Estrogen Receptor program), where co-regulated genes are dispersed across multiple chromosomes. By successfully identifying these differing landscapes, imHG demonstrates its potential utility in capturing the underlying structural nature of transcriptional dysregulation.

Looking forward, while imHG currently addresses 1D ordered sequences, the mathematical framework can logically be extended to two- or three-dimensional settings. As spatial transcriptomics and highly multiplexed imaging become increasingly prevalent, the ability to identify spatially dense, statistically significant "neighborhoods" of gene expression in 2D tissue slices or 3D histological images presents an exciting future direction for this approach. Finally, while demonstrated here on genomic coordinates, imHG's threshold-free interval detection is applicable to any ordered biological data, paving the way for applications in detecting transient cellular states in single-cell pseudotime trajectories or temporal windows in time-series expression data.

Author Contributions

B.G., S.M. and Z.Y. planned the study, proposed the new theorem, and proved it. S.M. wrote the Python code. B.G. and O.K. helped obtain and process the biological data that was used. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable

Informed Consent Statement

Not applicable

Data Availability Statement

The data for the genome location enrichment of differentially expressed genes was obtained from the National Library of Medicine (NCBI), available at: https://www.ncbi.nlm.nih.gov/gene. Specifically, we used the following series: GSE14814, GSE19188, GSE30219, GSE37745, GSE50081, GSE8894. The data on genome location enrichment for GSEA sets were obtained from https://www.gsea-msigdb.org/gsea/index.jsp. The code used for the analysis is available in our GitHub repository: https://github.com/YakhiniGroup/imhg.

Acknowledgments

We thank the Technion Computer Science Department and the School of Computer Science at IDC Herzliya for their support of the project. We thank the Yakhini Group for useful discussions and comments.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Eden, E.; Lipson, D.; Yogev, S.; et al. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* **2007**, *3*, e39.
2. Subramanian, A.; Tamayo, P.; Mootha, V.K.; et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550.
3. Liberzon, A.; Subramanian, A.; Pinchback, R.; et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740.
4. Eden, E.; Navon, R.; Steinfeld, I.; et al. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinform.* **2009**, *10*, 48.
5. The Gene Ontology Consortium. Available online: https://geneontology.org (accessed on 20 January 2026).
6. Leibovich, L.; Yakhini, Z. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res.* **2012**, *40*, 5832–5847.
7. Street, K.; Risso, D.; Fletcher, R.B.; et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **2018**, *19*, 477.
8. Bar-Joseph, Z.; Gitter, A.; Simon, I. Analyzing time series gene expression data. *Bioinformatics* **2004**, *20*, 2493–2503.

9. Rapoport, R.; Greenberg, A.; Yakhini, Z.; et al. A Cyclic Permutation Approach to Removing Spatial Dependency between Clustered Gene Ontology Terms. *Biology* **2024**, *13*, 175.
10. Dixon, J.R.; Selvaraj, S.; Yue, F.; et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **2012**, *485*, 376–380.
11. Caron, H.; van Schaik, B.; van der Mee, M.; et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **2001**, *291*, 1289–1292.
12. Ben-Elazar, S.; Yakhini, Z.; Yanai, I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome . *Nucleic Acids Res.* **2013**, *41*, 2191–2201.
13. Ben-Elazar, S.; Chor, B.; Yakhini, Z. The Functional 3D Organization of Unicellular Genomes. *Sci. Rep.* **2019**, *9*, 12734.
14. Kariti, H.; Feld, T.; Kaplan, N. Hypothesis-driven probabilistic modelling enables a principled perspective of genomic compartments. *Nucleic Acids Res.* **2023**, *51*, 1103–1119.
15. Golov, A.K.; Gavrilov, A.A.; Kaplan, N.; et al. A genome-wide nucleosome-resolution map of promoter-centered interactions in human cells corroborates the enhancer-promoter looping model. *eLife* **2024**, *13*, RP91596.
16. Roth, R. *Introduction to Coding Theory*; Cambridge University Press: Cambridge, UK, 2006.
17. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
18. Wagner, F. GO-PCA: an unsupervised method to explore gene expression data using prior knowledge. *PloS ONE* **2015**, *10*, e0143196.
19. Sung, H.; Ferlay, J.; Siegel, R.L.; et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249.
20. Niederhuber, J.E.; Armitage, J.O.; Kastan, M.B.; et al. (Eds.) *Abeloff's Clinical Oncology*; Elsevier: Amsterdam, The Netherlands, **2020**.
21. Galili, B.; Tekpli, X.; Kristensen, V.N.; et al. Efficient gene expression signature for a breast cancer immuno-subtype. *PLoS ONE* **2021**, *16*, e0245215.
22. Biomart Datasets, Human Genes (GRCh38.P14). 2023. Available online: https://www.ensembl.org/biomart/martview/d907afbf9d849b84f97a226d8f032d6b (accessed on 30 June 2024).
23. Ritchie Lab Visualization. Available online: http://visualization.ritchielab.org/phenograms/plot (accessed on 30 June 2024).
24. Botling, J.; Edlund, K.; Lohr, M.; et al. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.* **2013**, *19*, 194–204.
25. Jabs, V.; Edlund, K.; König, H.; et al. Integrative analysis of genome-wide gene copy number changes and gene expression in non-small cell lung cancer. *PLoS ONE* **2017**, *12*, e0187246.
26. Lohr, M.; Hellwig, B.; Edlund, K.; et al. Identification of sample annotation errors in gene expression datasets. *Arch. Toxicol.* **2015**, *89*, 2265–2272.
27. Goldmann, T.; Marwitz, S.; Nitschkowski, D.; et al. PD-L1 amplification is associated with an immune cell rich phenotype in squamous cell cancer of the lung. *Cancer Immunol. Immunother.* **2021**, *70*, 2577–2587.
28. Khadse, A.; Haakensen, V.D.; Silwal-Pandit, L.; et al. Prognostic significance of the loss of heterozygosity of KRAS in early-stage lung adenocarcinoma. *Front. Oncol.* **2022**, *12*, 873532.
29. Hou, J.; Aerts, J.; den Hamer, B.; et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE* **2010**, *5*, e10312.
30. Rousseaux, S.; Debernardi, A.; Jacquiau, B.; et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* **2013**, *5*, 186ra66.
31. Der, S.D.; Sykes, J.; Pintilie, M.; et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J. Thorac. Oncol.* **2014**, *9*, 59–64.
32. Lee, E.S.; Son, D.S.; Kim, S.H.; et al. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin. Cancer Res.* **2008**, *14*, 7397–7404.
33. Zabarovsky, E.R.; Lerman, M.I.; Minna, J.D. Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene* **2002**, *21*, 6915–6935.
34. Dehan, E.; Ben-Dor, A.; Liao, W.; et al. Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. *Lung Cancer* **2007**, *56*, 175–184.
35. Zhu, C.Q.; Ding, K.; Strumpf, D.; et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J. Clin. Oncol.* **2010**, *28*, 4417–4424.
36. Enerly, E.; Steinfeld, I.; Kleivi, K.; et al. miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS ONE* **2011**, *6*, e16915.
37. Parker, J.S.; Mullins, M.; Cheang, M.C.; et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **2009**, *27*, 1160–1167.
38. Prat, A.; Pineda, E.; Adamo, B.; et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* **2015**, *24*, S26–S35.
39. Aure, M.R.; Jernström, S.; Krohn, M.; et al. Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome Med.* **2015**, *7*, 21.

40. Haukaas, T.H.; Euceda, L.R.; Giskeødegård, G.F.; et al. Metabolic clusters of breast cancer in relation to gene- and protein expression subtypes. *Cancer Metab.* **2016**, *4*, 12.
41. Tinholt, M.; Vollan, H.K.M.; Sahlberg, K.K.; et al. Tumor expression, plasma levels and genetic polymorphisms of the coagulation inhibitor TFPI are associated with clinicopathological parameters and survival in breast cancer, in contrast to the coagulation initiator TF. *Breast Cancer Res.* **2015**, *17*, 44.
42. Ankill, J.; Aure, M.R.; Bjørklund, S.; et al. Epigenetic alterations at distal enhancers are linked to proliferation in human breast cancer. *NAR Cancer* **2022**, *4*, zcac008.
43. Tekpli, X.; Lien, T.; Røssevold, A.H.; et al. An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* **2019**, *10*, 5499.
44. Minn, A.J.; Gupta, G.P.; Padua, D.; et al. Lung metastasis genes couple breast tumor size and metastatic spread. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6740–6745.
45. Wang, Y.; Klijn, J.G.M.; Zhang, Y.; et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **2005**, *365*, 671–679.