CK.enrich.v2: A Tool to Detect Cancer Genes in the Human Genomic Enriched Zones

Xiaoping Li, Joe Song

April 8, 2017

New Mexico State University

Abstract

Understanding of spatial arrangement of human genome in nucleus provides extra information on gene regulation. Implemented with the R function CKmeans.1d.dp that optimized k means clustering algorithm, we have developed a tool called CK.enrich.v1 that clusters human genome into genometric zones and explores the association between human cancer genes and those genometric zones statistically. The results reported six zones were significant enriched with cancer genes among seven human chromosomes.

Introduction

Static information from DNA sequences is not enough to display the dynamic interactions and regulations of genes in a cell. To understand genome functions, it is important to illustrate how genome is organized in the nucleus spatially [1]. Genome folding and arrangement *in vivo* is rather intricate, but the level of complexity is crucial for DNA replication, gene regulation and genome structure stability [2] [3]. Many diseases such as cancer are caused by chromosomal abnormality in the spatial organization [4]. The conventional method for studying the spatial organization of chromosomes is via visualization by using fluorescent in situ hybridization (FISH) [1]. Recently years, new methods have appeared by using crosslink and intramolecular DNA ligation to capture the spatial loci associations of a whole genome, such as Hi-C, 3C, 4C and 5C [5]. With those methods, it is found that human chromosomes prefer periphery or interior positions in a nucleus which largely due to gene density in quite amount of cell types [6] [7]. Gene-poor regions and gene-rich regions form a polarized arrangement in a chromosome, where gene-rich regions position more centrally in a nucleus whereas gene-poor regions tend to be located at periphery [8]. This radial organization of chromosomes has been reported associated with human diseases such as laminopathies [1] [9]. In addition, Hi-C and 3C-type analysis discovered spatial clustering of active gene domains to interact with each other, and similarly, inactive gene domains tend to associate with inactive gene domains within another cluster in a chromosome [10] [11]. The other feature of gene regulation in human genome is long range. To achieve transcriptional regulation of a target gene distant away from an element, chromosomes has to conform into loops to bring the elements into proximity of the target gene [1]. It is clear that a linear genome is limited to provide a deep understanding of transcriptional regulation mechanism. The configuration of chromosomes can be considered clustered with zones containing genes of a similar type. We suspect that from this arrangement, there will be association between those zones and the occurrence of cancer genes.

K-means clustering is an algorithm of cluster analysis that separates set of n-dimensional points into a set of K clusters. The algorithm does so by finding partitions that minimize the squared errors between the points and the empirical means [12]. The limitation of k-means algorithm lies in repeatability, optimality and operation speed [13]. The appearance of R package CKmeans.1d.dp optimizes k-means algorithm in its repeatability, runtime when the clusters are big in one dimension [13]. Therefore, this function can be quickly implemented to the human genome chromosome datasets.

Results

There are 782 zones formed after clustering. 46 of 1571 cancer genes were shown significantly enriched in six zones over seven chromosomes (p.adjusted < 0.05, alpha = 0.05). Chromosomes and associated zones include chr6(zone 9), chr11(zone 12), chr14 (zone 55), chr18 (zone 12), chr20 (zone 16), chr22 (zone 11) and chrM (zone 6). On chromosome 6, there are 15 genes associated with histone regulation reported on zone No. 9. On chromosome 11, 8 olfactory receptors are clustered in zone 12.

Details see Table 1. Visualization of the enriched genomic zones can be found in Figure 1.

Method

R version and data source:

Three datasets needed for this research. A gff3 format file from human genome (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_26/gencode.v26.annotation.gff3.gz) .

A tsv file with human cancer genes information (http://ncg.kcl.ac.uk/download.php).

A txt file with hg38 cytobands information (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cytoBand.txt.gz).

The R version implemented was R3.3.3 on a 64-bit windows 10 OS.

Statistical analysis:

The distribution of zones in each chromosome was assumed random. The resolution used in CKmeans.1d.dp function set to 1Mb. The probability of each cancer gene falling into a certain zone was calculated with hypergeometric function by using phyper function in R. The p values were further corrected with Bonferroni correction by using p.adjust function in R, method parameter set to “Bonferroni”. Significant level set at 0.05.

R codes:

Step 1: getting the data

```
gff3 <- read.table("gencode.v26.annotation.gff3.gz", sep = "\t", col.names = c("chr", "source", "type", "start", "end", "score", "strand", "phase", "tag"), stringsAsFactors = F)

cancer <- read.table("download_query.tsv", header = TRUE, stringsAsFactors = F, sep = "\t")

cytoband <- read.table("cytoBand.txt.gz", col.names = c("chr", "start", 'end', "name", 'gieStain'), stringsAsFactors = F, sep = "\t")
```

step 2: Using CK.enrich.v1 function on the data

```
enriched <- CK.enrich.v1(gff3, cancer, cytoband)
```

Step 3: Visualization

```
df <- data.frame(chr = unique(enriched$chr), position = c(unique(enriched$zone_start_chr), end = unique(enriched$zone_end_chr)))

library(ggplot2)

ggplot(df, aes(x = chr, y = position, group = chr)) +
    geom_line(size = 5) +
    scale_y_continuous(limits  = c(5000, 100000000)) +
    theme_bw() +
    theme(panel.grid.major = element_line(colour = "grey")) +
    labs(x = "Chromosome", y = "Million bases", title = "Enriched genomic zones") +
    theme(plot.title = element_text(hjust = 0.5)) +
    annotate("text", label = "15/82", x = 6, y = df$position[df$chr == "chr6"][1] + 10000000) +
    annotate("text", label = "12/155", x = 1, y = df$position[df$chr == "chr11"][1] + 10000000) +
    annotate("text", label = "4/27", x = 2, y = df$position[df$chr == "chr14"][1] + 3500000) +
    annotate("text", label = "4/26", x = 3, y = df$position[df$chr == "chr18"][1] + 10000000) +
    annotate("text", label = "4/30", x = 4, y = df$position[df$chr == "chr20"][1] + 10000000) +
    annotate("text", label = "6/54", x = 5, y = df$position[df$chr == "chr22"][1] + 10000000) +
    annotate("text", label = "1/3", x = 7, y = df$position[df$chr == "chrM"][1] + 10000000)
```

Details of CK.enrich.v1 see supplement materials.

Table 1: Cancer gene enriched human genomic zones.

chr	Cytogenetic band	Zone id	Zone #	Start (b)	End (b)	Total gene	Observed gene #	symbols	p-adjusted
Chr6	p36.11	9	57	25e+07	2.7e+07	82	15	HIST1H1C HIST1H1D HIST1H1E HIST1H2AD HIST1H2BD HIST1H2AC HIST1H2BG HIST1H2BF HIST1H2BE HIST1H2BC HIST1H3C HIST1H3B HIST1H4D HIST1H4H HIST1H4B	3.3e-08
Chr11	p34.3-p34.1	12	29	5.2e+07	5.8e+07	155	12	APLNR CLP1 OR5L2 OR4C15 TRIM51 OR4C6 OR5L1 LRRC55 OR10AG1 OR8H2 OR5T1 OR8K1	9.2e-03
Chr14	p21.1-p13.3	55	66	9.5e+07	9.6e+07	27	4	DICER1 TCL6 SERPINA12 SYNE3	0.046
Chr18	p35.3-p35.2	12	22	3.8e+07	4.4e+07	26	4	RIT2 SYT4 SETBP1 EPG5	0.012
Chr20	p34.3	16	24	4e+07	4.3e+07	30	4	PLCG1 TOP1 PTPRT CHD6	0.027
Chr22	p36.11-p35.3	11	22	2.8e+07	3e+07	54	6	EWSR1 NF2 XBP1 CHEK2 KREMEN1 ZNRF3	
ChrM	p36.33	6	14	6609	7957	3	1	COX2	0.03

α=0.05, p.adjust method = Bonferroni, total cancer genes=1571

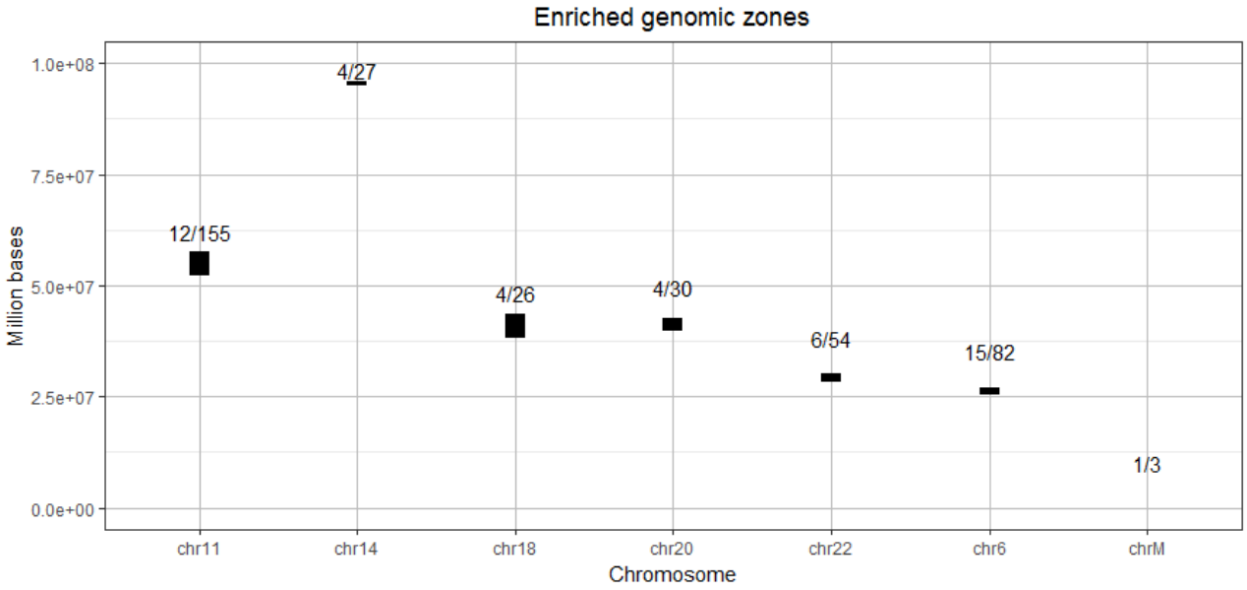

Figure 1: Visualization of enriched genomic zones for the cancer genes.

Discussion

Chromosomes carrying the genome information is not linear, rather they fold into arrangements in 3D to interact [3]. The genome regulation is tightly depend on this spatial organization [1]. From chromosome arrangement, gene clustering was reported where highly expressed genes tend to interact with each other whereas low expressed genes cluster together [11]. Thus, it is reasonable to partition chromosomes into zones and associate those zones with the occurrence of currently reported cancer genes. In this research, we designed a function CK.enrich.v1 that takes advantage of available R package Ckmeans.1d.dp to cluster each chromosomes into zones with k-means algorithm. This function combining with NCG cancer dataset and UCSC cytogenetic data base produces a data frame containing information about statistically significant zones that observed more cancer genes than other.

The results show that among 1571 cancer genes, there are 46 fall into 6 statistically significant zones from 7 human chromosomes. It is noticeable that there is large amount of cancerous histone genes associated with zone No.9 on chromosome 6 and olfactory receptor genes associated with zone No.12 on chromosome 11. In the original cancer data set, it failed to observe genes on chromosome Y.

The results indicate that same type of cancer genes tend to cluster together. Analysis on the enrich zones will provide us new insight to cancer gene regulation *in vivo* and new understanding of relationship between genome organization and disease causes.

Reference

1. Bickmore, W.A., *The Spatial Organization of the Human Genome.* Annual Review of Genomics and Human Genetics, 2013. **14**(1): p. 67-84.
2. Mitelman, F., *Recurrent chromosome aberrations in cancer.* Mutation Research/Reviews in Mutation Research, 2000. **462**(2–3): p. 247-253.
3. Hu, M., et al., *Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data.* Quantitative biology, 2013. **1**(2): p. 156-174.
4. Rowley, J.D., *THE CRITICAL ROLE OF CHROMOSOME TRANSLOCATIONS IN HUMAN LEUKEMIAS.* Annual Review of Genetics, 1998. **32**(1): p. 495-519.
5. de Wit, E. and W. de Laat, *A decade of 3C technologies: insights into nuclear organization.* Genes & Development, 2012. **26**(1): p. 11-24.
6. Bolzer, A., et al., *Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes.* PLOS Biology, 2005. **3**(5): p. e157.
7. Boyle, S., et al., *The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells.* Human Molecular Genetics, 2001. **10**(3): p. 211-220.
8. Solovei, I., et al., *Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution.* Cell. **137**(2): p. 356-368.
9. Meaburn, K.J., et al., *Primary laminopathy fibroblasts display altered genome organization and apoptosis.* Aging Cell, 2007. **6**(2): p. 139-153.
10. Lieberman-Aiden, E., et al., *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.* Science, 2009. **326**(5950): p. 289-293.
11. Sexton, T., et al., *Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.* Cell. **148**(3): p. 458-472.
12. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
13. Wang, H. and M. Song, *Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming.* The R journal, 2011. **3**(2): p. 29-33.