Statistical Determination of Cancer Gene-enriched Human Genomic Zones

Evan Lavelle

Abstract: The object of the project was to create a program in R able to specify which zones on each human chromosome contain a significantly high number of cancer-susceptible genes, a determination made based on the adjusted p-value of every zone. The results, presented here as both a dataframe table and a plot graph, indicate the zones where this value is below the established alpha value cutoff of 0.05. Additionally, information is given as to the cytogenetic bands that correspond to the start and end points of these zones, the chromosomal base indeces of these points, the total number of zones on all chromosomes wherein one or more enriched zones occur, the actual number of cancer genes and total genes within or overlapping every such zone, and the names of the cancer genes meeting this criterion.

Introduction

Biochemical mechanisms exist that allow for synchronous regulation of genes anywhere within a genomic “neighborhood”. In spite of this, it is feasible that instances of cancer-susceptible genes are group in close proximity. To measure this, the chromosomes are segmented into sequences of zones. The null hypothesis to the study will be that cancer genes exhibit a random distribution, and none of the regions are enriched to or past a designated point represented by the conventional alpha value of 0.05. If this threshold is crossed, the hypothesis is confirmed that implies a schema more involved that coincidental placement.

Data Source

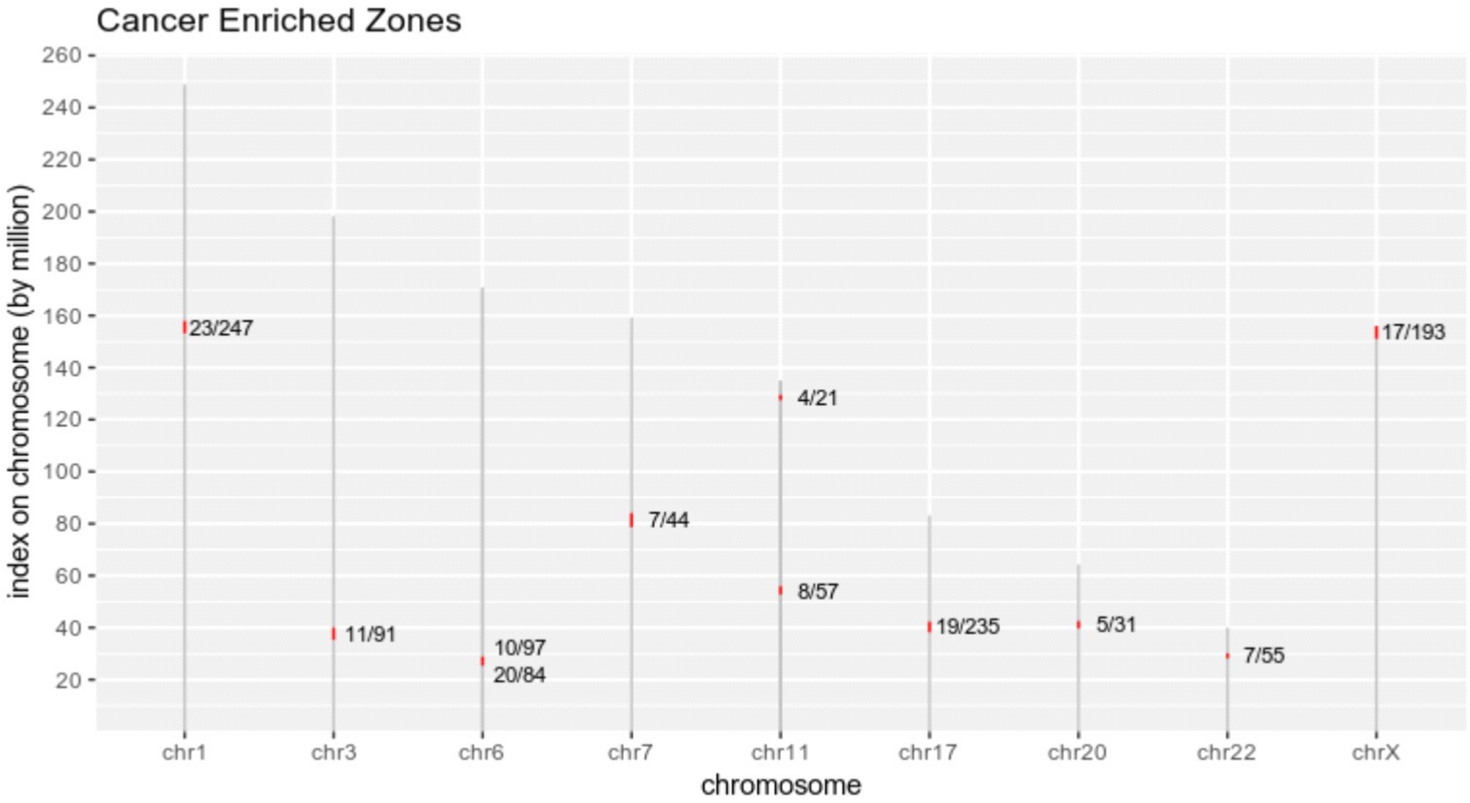
Data on the human genome was obtained from the GENCODE Project’s version 29 release. The list of cancer genes was acquired from NCG6.0, a source maintained by The School of Cancer Studies of King’s College London and The Francis Crick Institute. The annotation of cytogenetic bands used is as published by the University of California Santa Cruz.

Methods

First, the GENCODE file was gleaned to obtain a more manageable data dataframe consisting only of actual genes. Iterating through this data structure, several lists were constructed by reading and matching column content. By this procedure, data on chromosome length, genes per chromosome, gene start and end indeces by chromosome was taken. Using these data, it was possible to find gene cluster centers for each chromosome using the Ckmeans.1d.dp() function. Feeding this portion of the return values into a simple function and appending start and end indeces of each chromosome, vectors of zone boundaries were created. Next, the list of cancer genes was loaded and referenced against the dataframe of all genes. Thusly, chromosomal start and stop indeces of cancer genes were extracted. Comparing this information with the vectors of zone boundaries, the number of cancer genes on each zone was computed. With these data, the phyper() function was ran vector-wise on all zones for each chromosome. Following this, the p-values returned were adjusted by multiplication by a factor representing the average number of zones on each chromosome. A dataframe summarizing pertinent data for zones corresponding to adjusted p-values under 0.05 is made as the program output. This dataframe is then modified to generate a graph plotted with the ggplot2 package. The program has been divided into two files: one, which must be ran first, contains the functions written for the program and tests for some of them, commented out. The second includes the script to produce the output- the files imported must be in the same working directory as the program, or the filepaths specified. The minutiae of the program’s operation are detailed in concise comments throughout both files.

Results

seqid	cband.start	cband.end	zid	nzone	start	end	total	observed	pval	pval.adj
chr1	q21.3	q23.1	22	36	153211512	157883221	247	23	4.02E-05	1.33E-03
chr3	p22.3	p22.1	10	46	35318595	40051828	91	11	2.10E-04	6.94E-03
chr6	p22.2	p22.2	9	57	25401010	27053271	84	20	6.72E-12	2.22E-10
chr6	p22.2	p22.1	10	57	27053271	28804961	97	10	1.37E-03	4.53E-02
chr7	q21.11	q21.11	20	38	78666481	84037746	44	7	2.59E-04	8.57E-03
chr11	p11.11	q12.1	28	71	52685902	55940918	57	8	3.30E-04	1.09E-02
chr11	q24.2	q24.3	68	71	127491958	129491156	21	4	1.06E-03	3.51E-02
chr17	q12	q21.2	10	19	38214458	42326549	235	19	9.34E-04	3.09E-02
chr20	q12	q12	16	24	39663697	42525679	31	5	1.09E-03	3.59E-02
chr22	q12.1	q12.2	10	20	28182734	30067465	55	7	7.00E+00	4.04E-02
chrX	q28	q28	26	26	151067813	156027877	193	17	1.70E+01	1.89E-02



Cancer genes by zone

Chr 1, zone 22:
SDSAF; S100A7; DENND4B; CREB3L4; RPS27; NUP210L; TPM3; ATP8B2; TDRD10; KCNN3; ZBTB7B; MUC1; ASH1L; RIT1; LMNA; SMG5; RHBG; MEF2D; GPATCH4; PRCC; NTRK1; PEAR1; FCRL4; FCRL1

Chr 3, zone 10:
MLH1; VILL; PLCD1; DLEC1; MYD88; ACVR2B; SCN5A; SCN10A; SCN11A; GORASP1; CX3CR1

Chr 6, zone 9:
HIST1H4B; HIST1H3B; HIST1H3C; HIST1H1C; HIST1H2BC; HIST1H2AC; HIST1H1E; HIST1H2BD; HIST1H2BE; HIST1H4D; HIST1H3D; HIST1H2AD; HIST1H2BF; HIST1H4E; HIST1H2BG; HIST1H1D; HIST1H2BH; HIST1H3G; HIST1H4H; BTN3A2

Chr 6, zone 10:
HIST1H2BJ; HIST1H2AG; HIST1H4I; HIST1H2BK; HIST1H3H; HIST1H2AL; HIST1H1B; HIST1H2AM; HIST1H2BO; GPX5

Chr 7, zone 20:
MAGI2; CD36; SEMA3C; HGF; CACNA2D1; PCLO; SEMA3E

Chr 11, zone 28
OR4C46; TRIM48; OR4A16; OR4C15; OR4C6; OR5L1; OR5L2; TRIM51

Chr 11, zone 68
ETS1; FLI1; KCNJ5; ARHGAP32

Chr 17, zone 10
MLLT6; LASP1; MED1; CDK12; ERBB2; IKZF3; WIPF2; RARA; CCR7; SMARCE1; KRT222; TMEM99; KRTAP4-5; KRTAP4-3; KRT13; KRT15; DHX58; STAT5B; STAT3

Chr 20, zone 16
MAFB; TOP1; PLCG1; CHD6; PTPRT

Chr 22, zone 10
CHEK2; XBP1; ZNRF3; KREMEN1; EWSR1; NEFH; NF2

Chr X, zone 26
VMA21; PASD1; MAGEA6; MAGEA1; ATP2B3; CCNQ; DUSP9; ABCD1; ARHGAP4; FLNA; RPL10; FAM50A; PLXNA3; F8; MTCP1; BRCC3; SP

Discussion

After p-value adjustment, a relatively few number of zones (11) were marked as enriched. As visible in the Results, 6 and 11, had two zones each. Interestingly, those on 6 are located back-to-back. Automating plot generation with these considerations in mind was the most challenging aspect of the project. Further work that might done could include running the program for different alpha values and plot the number of enriched zones at each level.

Conclusion

The multiple marked zones are unlikely to have the number cancer genes they do only by chance. Thus, the hypothesis is confirmed. It follows that there is probably some factor similar to these regions that cause them to be prone to the irregular transcription characterizing cancer. In fact, it may be the proximity itself that is responsible for this correlation, and cancer-inducing mutations tend to spread to more proximal genes. However, being that this project is akin to an observational study, the identification as to what this/these attribute/s may be can only be speculated on from within the scope of the project's findings.

References

https://www.gencodegenes.org/human/

http://ncg.kcl.ac.uk/download.php

http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cytoBand.txt.gz