SUPPLEMENTARY DOCUMENT

ZDOG: Zooming In on Dominating Genes with Mutations In Cancer Pathways

Rudi Alberts, Jinyu Chen, Louxin Zhang

National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076

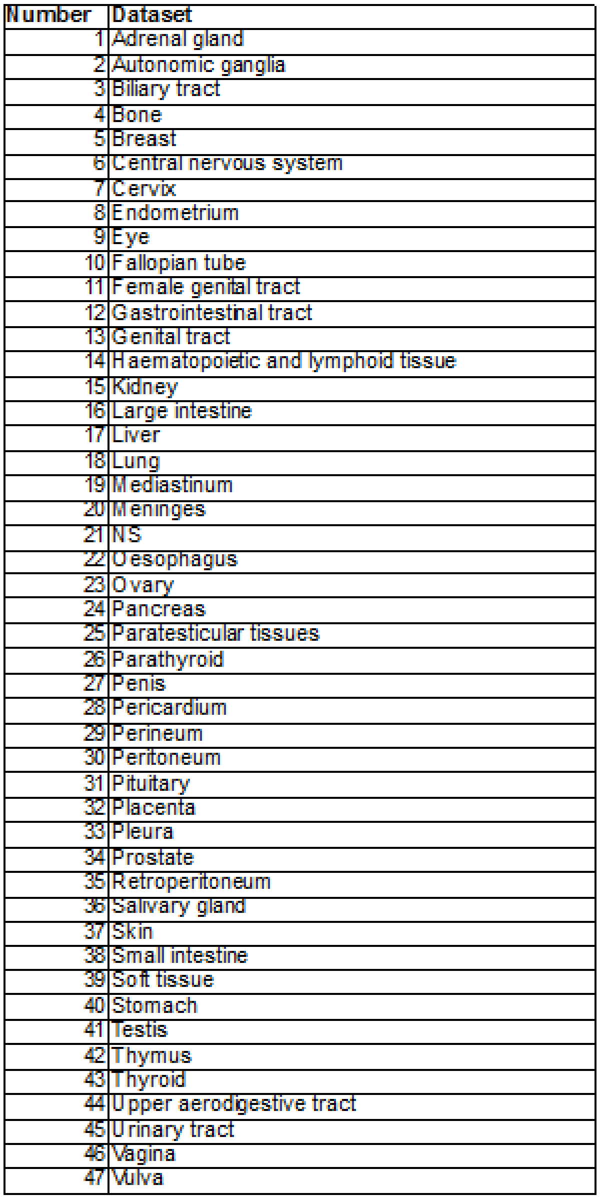
Table of Context

1. Data sources and data preparation

1.1 Catalogue of Somatic Mutations in Cancer (COSMIC)

Catalogue of Somatic Mutations in Cancer data was downloaded from the COSMIC website (COSMIC v87, released 13-NOV-18). Under the Data Downloads page (https://cancer.sanger.ac.uk/cosmic/download) section "COSMIC Mutation Data", we downloaded the tab separated table of 6,581,004 COSMIC coding point mutations from targeted and genome wide screens named CosmicMutantExport.tsv.gz. A home-made bash script was used to extract the relevant columns from this file. We then wrote an R script to summarize mutation data per gene and per dataset. Table 1 shows the division of the data.

Table 1. Datasets derived from the COSMIC database.

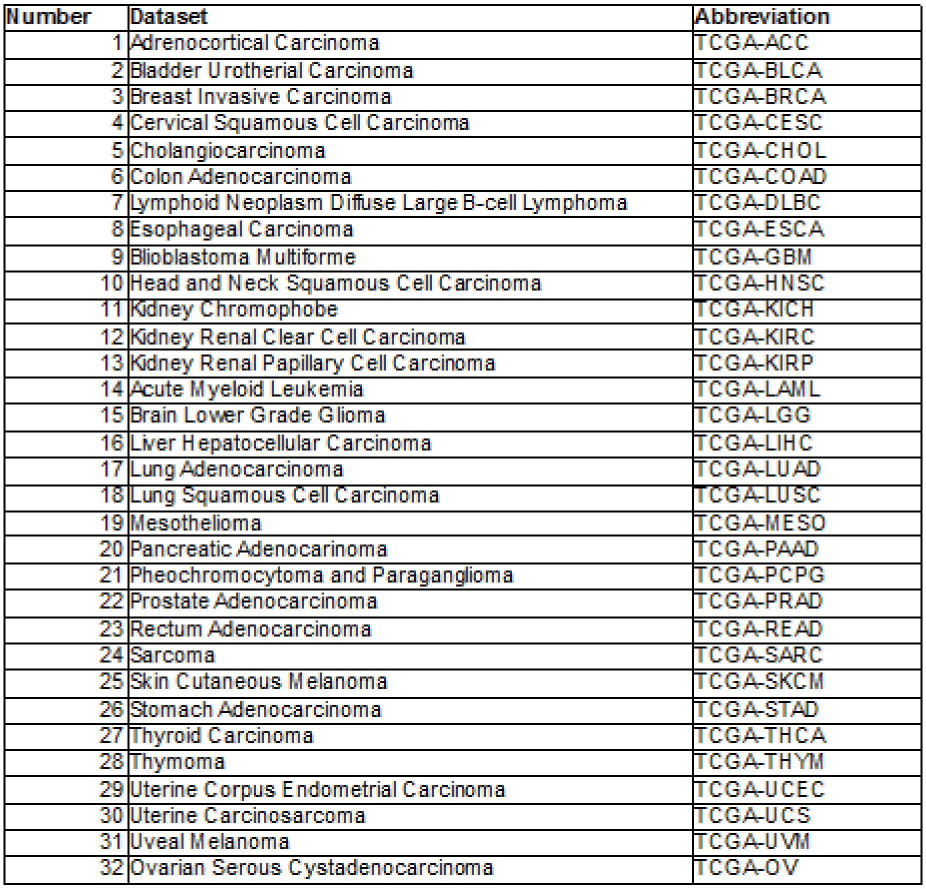


Number	Dataset
1	Adrenal gland
2	Autonomic ganglia
3	Biliary tract
4	Bone
5	Breast
6	Central nervous system
7	Cervix
8	Endometrium
9	Eye
10	Fallopian tube
11	Female genital tract
12	Gastrointestinal tract
13	Genital tract
14	Haematopoietic and lymphoid tissue
15	Kidney
16	Large intestine
17	Liver
18	Lung
19	Mediastinum
20	Meninges
21	NS
22	Oesophagus
23	Ovary
24	Pancreas
25	Paratesticular tissues
26	Parathyroid
27	Penis
28	Pericardium
29	Perineum
30	Peritoneum
31	Pituitary
32	Placenta
33	Pleura
34	Prostate
35	Retroperitoneum
36	Salivary gland
37	Skin
38	Small intestine
39	Soft tissue
40	Stomach
41	Testis
42	Thymus
43	Thyroid
44	Upper aerodigestive tract
45	Urinary tract
46	Vagina
47	Vulva

1.2 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas data was downloaded from the NIH National Cancer Institute GDC Data Portal (https://portal.gdc.cancer.gov/). For each of 32 TCGA projects (Table 2) we downloaded the open access Single Nucleotide Variation data file named like this: TCGA.<dataset abbreviation>.mutect.*.somatic.maf.gz. The TCGA data was processed in the same way as the COSMIC data.

Table 2 Datasets derived from TCGA database



Number	Dataset	Abbreviation
1	Adrenocortical Carcinoma	TCGA-ACC
2	Bladder Urotherial Carcinoma	TCGA-BLCA
3	Breast Invasive Carcinoma	TCGA-BRCA
4	Cervical Squamous Cell Carcinoma	TCGA-CESC
5	Cholangiocarcinoma	TCGA-CHOL
6	Colon Adenocarcinoma	TCGA-COAD
7	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	TCGA-DLBC
8	Esophageal Carcinoma	TCGA-ESCA
9	Blioblastoma Multiforme	TCGA-GBM
10	Head and Neck Squamous Cell Carcinoma	TCGA-HNSC
11	Kidney Chromophobe	TCGA-KICH
12	Kidney Renal Clear Cell Carcinoma	TCGA-KIRC
13	Kidney Renal Papillary Cell Carcinoma	TCGA-KIRP
14	Acute Myeloid Leukemia	TCGA-LAML
15	Brain Lower Grade Glioma	TCGA-LGG
16	Liver Hepatocellular Carcinoma	TCGA-LIHC
17	Lung Adenocarcinoma	TCGA-LUAD
18	Lung Squamous Cell Carcinoma	TCGA-LUSC
19	Mesothelioma	TCGA-MESO
20	Pancreatic Adenocarinoma	TCGA-PAAD
21	Pheochromocytoma and Paraganglioma	TCGA-PCPG
22	Prostate Adenocarcinoma	TCGA-PRAD
23	Rectum Adenocarcinoma	TCGA-READ
24	Sarcoma	TCGA-SARC
25	Skin Cutaneous Melanoma	TCGA-SKCM
26	Stomach Adenocarcinoma	TCGA-STAD
27	Thyroid Carcinoma	TCGA-THCA
28	Thymoma	TCGA-THYM
29	Uterine Corpus Endometrial Carcinoma	TCGA-UCEC
30	Uterine Carcinosarcoma	TCGA-UCS
31	Uveal Melanoma	TCGA-UVM
32	Ovarian Serous Cystadenocarcinoma	TCGA-OV

2. Implementation

2.1 Coloring of allele frequencies for one dataset

Here, we describe how allele frequencies are calculated and colored on genes. First, we check which mutation types are selected for the specific database (COSMIC or TCGA). Next, for each gene in the pathway, we collect all mutations for the selected mutation types. Next, for all selected mutations in the gene we collect the names of the samples that carry the alternative allele. Finally, we count how many *unique* samples are in this collection of names. We do this because the same sample can have the alternative allele for several mutations in the same gene. Now, let d be this amount of unique samples carrying alternative alleles, and let t be the total amount of samples in this dataset. Then, the allele frequency is calculated as $100 \times \frac{d}{t}$.

2.2 Colouring of allele frequencies for multiple datasets

Let d_i be the amount of unique samples carrying an alternative allele for selected mutation types per gene, as calculated in the previous paragraph, in dataset i. Let t_i be the total amount of samples in dataset i. The average allele frequency over multiple datasets N is simply calculated as $100 \times \frac{(\sum_{i=1}^{N} d_i)}{\sum_{i=1}^{N} t_i}$.

2.3 Separate colouring of tumor suppressors and oncogenes

We took the collection of 187 tumor suppressors and oncogenes reported in Supplementary Table 4 of (Sanchez-Vega *et al.*, 2018). We divided them into a list of 63 oncogenes and another list of 124 tumor suppressors. To color genes in ZDOG, we check whether they appear in one of those lists. If the gene is among the tumor suppressors, it gets a blue shade. If the gene is among the oncogenes, it gets a red shade. If the gene in not in the lists, it gets a grey shade.

2.4 Algorithm for computing dominator tree

We implemented the fast algorithm for finding dominators in a flowgraph into ZDOG, originally introduced in 1979 (Lengauer and Tarjan, 1979). After selecting one (and only one) node in the pathway, the user can click “Calculate dominator tree” and the dominator tree will be presented in a new network window in Cytoscape. The previously selected node will be the root of the dominator tree. In this tree, dominating relationships between genes can be directly observed. Also, genes in the tree can be color coded the same way as described above.

Since a gene can appear multiple times in a biological pathway, and since the dominator tree algorithm works on node names and also we are interested in the ‘overall’ dominating relations between genes, before running the dominator tree algorithm, we merge duplicate genes into one gene.

References

Lengauer,T. and Tarjan,R.E. (1979) A fast algorithm for finding dominators in a flowgraph. ACM Trans. Program. Lang. Syst., 1, 121–141.

Sanchez-Vega,F. et al. (2018) Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell, 173, 321–337.