Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identifying pathways affected by cancer mutations☆,☆☆

Prathima Iengar

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

ARTICLE INFO

Keywords:
Organism-level pathways mutated in cancer
Doughnut plots representing mutated pathways and genes in cancer
Sample-wise matrix of gene mutations in cancer
Recurrently mutated genes in cancer
Weighting mutated genes and pathways in cancer
Cancer genes participating in multiple pathways
Gene-wise, chromosome-wise and cancer-wise distribution of cancer mutations

ABSTRACT

Mutations in 15 cancers, sourced from the COSMIC Whole Genomes database, and 297 human pathways, arranged into pathway groups based on the processes they orchestrate, and sourced from the KEGG pathway database, have together been used to identify pathways affected by cancer mutations. Genes studied in ≥ 15, and mutated in ≥ 10 samples of a cancer have been considered recurrently mutated, and pathways with recurrently mutated genes have been considered affected in the cancer. Novel doughnut plots have been presented which enable visualization of the extent to which pathways and genes, in each pathway group, are targeted, in each cancer. The 'organismal systems' pathway group (including organism-level pathways; e.g., nervous system) is the most targeted, more than even the well-recognized signal transduction, cell-cycle and apoptosis, and DNA repair pathway groups. The important, yet poorly-recognized, role played by the group merits attention. Pathways affected in ≥ 7 cancers yielded insights into processes affected.

1. Introduction

Cancer is a complex disease characterized by a large number and variety of mutations in a host of genes. Using powerful second-generation sequencing methods, it is possible to simultaneously identify multiple categories of mutations (e.g., substitutions, chromosomal rearrangements, copy number alterations) in tumour genomes [1]. Many important genes playing a causal role in cancer (cancer genes), the mutations they undergo and the pathways through which they act, have been characterized [2,3]. Mutations in three classes of genes are causal in cancer: oncogenes, tumour suppressors and stability or caretaker genes [3]. While mutations render oncogenes active, they inactivate tumour suppressors; the former drive cancer by accelerating cell-proliferation, the latter by inhibiting cell-death. Stability genes stabilize the genome by minimizing alterations to it; they correct changes to the nucleotide sequence during DNA replication, and control genetic recombination and chromosomal segregation. If stability genes are compromised, mutations in all genes, including oncogenes and tumour suppressors, occur at a higher rate. Insight into mutational processes and genes that drive cancer, afforded by cancer genome sequencing, and cancer progression as an example of an ageing process, have been summarised [4]. Several cancer genes have to be mutated for malignant cancer to develop, each gene contributing to the process by conferring a growth advantage on the cell. Research has shown that while a large number of driver genes are mutated in cancers, the number of pathways targeted is fewer. Somatic mutations characterized in tumour samples from various cancers have shown that RTK/RAS signalling pathways (receptor tyrosine kinases signalling with the RAS/MAPK pathway) are among the most altered across all the cancer types [5]. An 'exclusivity principle' has also been shown to operate in major regulatory pathways (e.g., retinoblastoma tumour suppressor pathway; [6]), wherein it is rare for multiple cancer genes to be mutated in a single pathway, in a single tumour [7,8]; mutating two genes in a pathway would not be of advantage to the tumour if the functional effect of mutating two genes were the same as that of mutating one gene, or, if the functional effects of mutating two genes opposed or interfered with each other. Computational methods have been developed to study the exclusivity principle in cancer exome mutation data [9–13]. A method has also been developed to detect pathways targeted by recurrent copy number alterations in tumours [14].

In the present study, gene mutations in 15 cancers, catalogued in the COSMIC (Catalogue of Somatic Mutations in Cancer) Whole Genomes database, have been extracted, and, for each cancer, human pathways and genes in the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database, affected by the mutations, have been identified. All samples of a cancer have collectively been considered, and recurrently

☆ Financial support: This work and PI have been supported by a grant from the Department of Science and Technology, Government of India, under the Women Scientists Scheme–A (No. SR/WOS-A/LS-458/2012 (G)).

☆☆ Conflict of interest disclosure: The author declares no potential conflict of interest.

E-mail address: prathima@iisc.ac.in.

https://doi.org/10.1016/j.ygeno.2017.12.004
Received 2 August 2017; Received in revised form 23 November 2017; Accepted 10 December 2017

mutated genes in the set of samples (defined as genes studied in ≥ 15, and mutated in ≥ 10 samples) have been identified. Pathways, in which ≥ 1 recurrently mutated genes occur, have been regarded as being affected in the cancer. In KEGG, pathways have been distributed into six groups. Pathways, in each group, affected by mutations, in each cancer, have been identified. Novel doughnut plots have been generated to facilitate visualization of the extent to which each group has been targeted in a cancer. In effect, the study has systematically identified pathways targeted by mutations in order to achieve the hallmarks of cancer [15,16].

2. Methods

A flowchart, presenting a summary of methods, is given at the beginning of Additional file 1: Supplementary Methods.

2.1. Dataset of mutations in cancer genomes

The COSMIC Whole Genomes database, a subset of the COSMIC database (version 71; [17]), was used as the source of mutations in cancer genomes. The set of 15 cancers selected for the study, and the number of samples catalogued for each cancer, in COSMIC Whole Genomes, are listed in Table 1 (smallest sample size, 292). In order to identify affected pathways in cancer, a list of mutated genes is necessary. As, in the database, substitution, insertion and deletion mutations are reported gene-wise, these have been considered in the study; copy number alterations and fusion mutations, for which genes involved are not always reported, have not been considered. In the database, for each cancer sample, genes that have been studied and found to be mutated, and the gene mutation(s), are reported; genes studied and found not to be mutated are also sometimes reported. Synonymous or coding silent mutations in genes have been excluded from the study. All human genes considered in the study (28952), and the chromosome information for each gene, were obtained from COSMIC (Additional file 1: Supplementary Methods (1)).

2.2. Dataset of human pathways

The KEGG database [18,19] was used as the source of 297 human pathways. A file listing the KEGG ID (identity) of each human pathway and, for each pathway, the ID and name of each gene in the pathway, was compiled from the database (Additional file 1: Supplementary Methods (2); Additional file 2: Table S1a). In KEGG, the 297 human pathways have been sub-divided into six groups; in the present study, these groups have been referred to as 'pathway groups', and used. The pathway groups, their sizes, and a brief description of processes orchestrated by each group, are as follows: (i) metabolism (92) – carbohydrate, lipid, nucleotide, amino acid and energy metabolism; (ii) genetic information processing (22) – transcription, translation, replication and repair; (iii) environmental information processing (28) – signal transduction; (iv) cellular processes (15) – cell growth and death, cell membrane functions; (v) organismal systems (69) – immune, endocrine, circulatory, digestive, excretory, nervous and sensory systems; (vi) human diseases (71) – cancers, immune, neurodegenerative, infectious, cardiovascular, endocrine and metabolic diseases (Additional file 3: Table S2).

2.3. Sorting mutations cancer-, sample- and gene-wise

Each line or entry in the COSMIC Whole Genomes dataset is a gene studied in a sample of a cancer, and reports any mutation observed in the gene. To begin with, for each cancer selected for study (Table 1), all gene entries were extracted from the dataset; the entries were then arranged sample-wise. The aim was to examine a sizeable set of samples of a cancer to identify genes that were mutated in multiple samples. A Perl program was written to take as input the entries for each cancer (gene entries sorted sample-wise), and generate a sample-wise matrix of gene mutations (Additional file 2: Table S1b). Each column in the matrix is a cancer sample, and each row is a gene; matrix elements are the number of mutations observed in a gene in a sample. Next, another Perl program was written to summarize the matrix for each cancer; the program counts and reports, for each gene, the number of tumour

Table 1
Fifteen cancers selected for the study, and the number of samples considered for each cancer (columns 1 and 2); the number of recurrently mutated genes, in each cancer, that occur in KEGG pathways (columns 3, 4, 5).

Cancer	Number of samples considered for each cancer	Total number of recurrently mutated genes identified in each cancer	Number of recurrently mutated genes in each cancer with gene names approved by HGNC	Number of recurrently mutated genes in each cancer (with gene names approved by HGNC) which occur in KEGG pathways
Large intestine carcinoma	619	5728	4577	1677
Skin malignant melanoma	649	4538	3358	1337
Lung adenocarcinoma	610	2518	1998	819
Liver carcinoma	880	1439	1005	401
Breast carcinoma	1151	1041	783	331
Lung squamous cell carcinoma	292	860	608	241
Kidney carcinoma	758	330	233	106
Pancreas carcinoma	674	243	151	68
Haematopoietic lymphoid tissue - lymphoid neoplasm	891	116	81	49
CNS glioma (central nervous system – glioma)	896	151	98	48
Haematopoietic lymphoid tissue - haematopoietic neoplasm	525	54	40	23
Ovary carcinoma	532	54	38	21
Prostate carcinoma	446	72	45	13
CNS PNET (central nervous system – primitive neuroectodermal tumour - medulloblastoma)	412	10	6	5
Thyroid carcinoma	413	2	2	2

Column 4 gives the number of recurrently mutated genes in each cancer, when only genes with gene names approved by HGNC (HUGO Gene Nomenclature Committee; [37]) are considered; i.e., genes with names such as ENSG00000093100, BAZ2A_ENST00000379441, etc., which have not been assigned unique names by HGNC, have been left out. The number of recurrently mutated genes in each cancer, with HGNC approved gene names, occurring in KEGG, is given in column 5. The cancers are arranged in descending order of the number of genes in this column.