Life Sciences 211 (2018) 206–214

Contents lists available at ScienceDirect

Life Sciences

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

A pan-cancer study of copy number gain and up-regulation in human oncogenes

YongKiat Wee[a], TianFang Wang[a], Yining Liu[b], Xiaoyan Li[c], Min Zhao[a,*]

[a] School of Engineering, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Queensland 4558, Australia
[b] The School of Public Health, Institute for Chemical Carcinogenesis, Guangzhou Medical University, 195 Dongfengxi Road, Guangzhou 510182, China
[c] Beijing Anzhen Hospital, Capital Medical University, Beijing Institute of Heart, Lung & Blood Vessel Disease, Beijing, China

ARTICLE INFO

Keywords:
Oncogene
Pan-cancer
Copy number variation
Copy number gain
Gene expression

ABSTRACT

Aim: There has been limited research on CNVs in oncogenes and we conducted a systematic pan-cancer analysis of CNVs and their gene expression changes. The aim of the present study was to provide an insight into the relationships between gene expression and oncogenesis.

Main methods: We collected all the oncogenes from ONGene database and overlapped with CNVs TCGA tumour samples from Catalogue of Somatic Mutations in Cancer database. We further conducted an integrative analysis of CNV with gene expression using the data from the matched TCGA tumour samples.

Key findings: From our analysis, we found 637 oncogenes associated with CNVs in 5900 tumour samples. There were 204 oncogenes with frequent copy number of gain (CNG). These 204 oncogenes were enriched in cancer-related pathways including the MAPK cascade and Ras GTPases signalling pathways. By using corresponding tumour samples data to perform integrative analyses of CNVs and gene expression changes, we identified 95 oncogenes with consistent CNG occurrence and up-regulation in the tumour samples, which may represent the recurrent driving force for oncogenesis. Surprisingly, eight oncogenes shown concordant CNG and gene up-regulation in at least 250 tumour samples: *INTS8* (355), *ECT2* (326), *LSM1* (310), *DDHD2* (298), *COPS5* (286), *EIF3E* (281), *TPD52* (258) and *ERBB2* (254).

Significance: As the first report about abundant CNGs on oncogene and concordant change of gene expression, our results may be valuable for the design of CNV-based cancer diagnostic strategy.

1. Introduction

Cancer is a major cause of death worldwide [1] and is a consequence of unlimited cell growth and proliferation. The uncontrolled growth will give rise to a tumour and the cancer cells may spread into healthy tissue (metastasis), and even affect blood and circulatory systems [2]. Cancer is a consequence of gene mutation and arising from the accumulation of somatic genetic alterations [2]. The mutated 'cancer genes' are divided into two groups: 'driver' and 'passenger' [3]. 'Passenger' mutations have neutral effects on the clonal expansion of the cancer cells and do not stimulate growth [3]. Conversely, 'driver' mutations are usually involved in cancer progression and a mutation within those genes such as oncogenes confer a selective growth advantage [4]. These oncogenes often encode proteins for the signalling pathways that maintain normal cell growth. In general, oncogenes arising from the mutations in proto-oncogenes are found in all normal cells and play a role in stimulating excessive cell division. The mutated forms of proto-oncogenes are found in the tumour cells [5]. Gene mutations and translocations can occur during cancer initiation event, whereas amplification normally occurs during progression. Research into oncogenes has been critically important in the treatment of cancer because the outcomes can be applied diagnostically in determining the seriousness and the stages of the disease, and to help in the discovery of potential markers as a guide to future gene therapy [6].

Cancer development involves a sequence of genetic abnormalities including single nucleotide mutations and copy number of variants (CNVs). CNVs are the copies of DNA segments in the human genome, size varies from thousands to millions of DNA bases and can vary in copy-number. Such copy number variations (or CNVs) can result in gene dosage imbalances [7]. There are two categories of CNVs: copy number loss (CNL) denotes the deletion of the gene copies while copy number gain (CNG) denotes the addition of gene copies. It is important to understand CNVs when examining the disease-associated changes and a baseline of human genomic variation needs to be created through

* Corresponding author.
E-mail address: mzhao@usc.edu.au (M. Zhao).

https://doi.org/10.1016/j.lfs.2018.09.032
Received 24 July 2018; Received in revised form 14 September 2018; Accepted 18 September 2018
Available online 19 September 2018

the analysis of the whole-genome CNVs [7,8]. Traditional methods, such as light microscopy for cytogenetic analyses have been used to detect the presence of large fragment deletions and duplications [9]. A large group of copy-number gains or losses has been associated with the development of disease [7,10]. Furthermore, some CNVs were found among the individuals with susceptibility to disease, such as oncogenes and tumour-suppressor genes in cancer [7,10]. The elevated gene expression of oncogene via gene amplification is a common event in human cancer. These amplified genes are required to be overexpressed in order to function as drive alterations. For example, Yamaguchi et al. had conducted an integrative analysis of copy number and gene expression profiling to discover the potential driver genes in 1454 solid tumors. There were 64 known driver oncogenes found in 587 tumors based on their gene expression profiling and CNVs. The authors compared the mRNA expression levels of these 64-known oncogene driver by performing the microarray analysis to assess the fold change between tumors and matched normal tissues in expression levels. The genes with elevated gene expression ≥ 5 fold in tumour tissues were known as overexpressed. The gene expression results of the 12 genes from 64 known oncogenes were then integrated with matched genomic copy number results to explore their relationships between CNVs and gene amplifications. The authors defined the genes with copy number ≥ 6 as overexpressed genes [11]. Ding et al. developed their own hierarchical Bayes statistical model, xseq, to systematically quantify and study the effect of somatic mutations on expression profiles. The authors only focused on the cis-effect impacts of tumour suppressor genes with loss-of-function mutations. The statistical model, xseq is predicted based on the measurable signals from the mutations with functional effects on the transcription in mRNA transcripts. The xseq model applies a precomputed 'influence graph' to integrate initial gene-gene relationship knowledge into its modelling framework [12].

Previous studies have investigated the correlation between CNVs and gene expression across different types of human cancer [13,14]. However, there is limited systematic study of this relationship found in oncogenes. To overcome these constraints, we conducted a pan-cancer CNV analysis on all the human oncogenes in order to explore the overall prospective of the CNV features. From this study, we may also be able to cross-validate some observations from the studies such as the concordance of copy number gain and up-regulation of oncogenes. The results may provide a better understanding of the relationship between CNV and gene expression changes in the progression of cancer.

2. Materials and methods

The 803 curated oncogene data were downloaded from the ONGene (http://ongene.bioinfo-minzhao.org/) [15] database. The format of the data was in plain text and included all the basic information including oncogene IDs and gene symbols. OnGene database is developed based on the curated literature genetic resource of the oncogene-related research. OnGene database contains all the curated genes, literature and functional annotations. This database can be used as a guide to perform a large-scale of genetic screening which related to oncogenes. In addition, it can be also used as a categorised ONG catalogue for experimental validation and integrative analysis of cancer genomics [15]. To perform a series of systematic analyses between CNV and oncogenes, The Cancer Genome Atlas (TCGA) CNVs data [16] were downloaded from the Catalogue of Somatic Mutations in Cancer (COSMIC) database (V78, GRCH 38) [17] and was used to investigate the CNVs in pan-cancer level. TCGA has identified and profiled the molecular alterations of a large number of tumour samples across their DNA, RNA, protein and epigenetic levels [18]. Fig. 1 demonstrates the pipeline for identification of concordant copy number gain and over-expression of oncogenes in human cancer. The CNVs for the oncogenes were extracted based on the official gene symbol. A total of 637 oncogenes overlapped with both the gain and loss gene copies in TCGA cancer samples. The gain-loss ratios for each oncogene were calculated based of the number of gain (CNGs) and loss (CNLs) samples across multiple TCGA cancers. The CNG frequency for each gene was also calculated based on the number of gain divided by the total number of gains and losses across multiple cancers (total number of gain samples/ total number of gain and loss samples). This information was used as to provide cut-off values for filtering purposes: the number samples of CNG ≥ 20; the gain-loss ratio > 2, and; the CNG frequency for CNGs > 0.1. We applied the cut-off value of number samples CNGs > 20 and the ratio of gain/loss > 2 to identify the oncogenes with constant CNGs. To filter out the CNVs in the human population, we defined a threshold-value with duplication frequency > 0.1. Finally, to determine the same oncogenes with concordant CNGs and up-regulation, the threshold value (ratio of Gain_Over/Loss_Under) was set to > 30 samples and 95 genes were generated with consistent CNGs and over-expression. The main reason for this was to identify a reliable gene list with constant CNG and over-expression. We set different cut-off values and we managed to narrow down the gene list to < 100 genes. Therefore, this level of gene list would be performed better for functional analysis.

2.1. CNVs in human population control data

In order to find out the high frequent CNVs in cancer sample, we downloaded the CNVs data in the human population from the DECIPHER v9.11 database [19] at https://decipher.sanger.ac.uk/files/downloads/population_cnv.txt.gz. Using the population CNV file, we aimed to extract the common CNVs and their duplication and deletion frequency; this served as a control data in the analysis. Since the CNV frequencies data in health population were provided in chromosome region, it requires the gene symbols for mapping. Hence, we downloaded the genomic location information for all the Human RefSeq genes from UCSC genome browser database on 20 Oct 2016 at http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz. Using Bedtools (V2.26.0), we performed mapping between the health CNV data and the RefSeq data based on the overlapping of chromosomal locations. The corresponding genomic locations in GRCH 38 were annotated with the control data. As oncogenes are normally known in gain-of-function in cancer development, we only compare the CNG frequency from COSMIC with the duplication frequency in the population control data. The genes with frequent pan-cancer CNGs were defined by filtering common CNVs using the cut-off value of CNG frequency in population control data of > 0.1. As a result, 204 oncogenes were identified from a list of genes with frequent CNGs in comparison with the CNVs in population data. The 204 genes were used for further analysis including functional enrichment and mapping to gene expression data.

2.2. Gene expression analysis in frequent oncogenes with CNGs

All TCGA expression data were downloaded from the COSMIC database (V78) in order to explore the concordant changes of gene expression in oncogenes with the CNVs. The analysis targeted only the gene expression changes in the same TCGA samples with CNGs oncogenes. COSMIC data consists of FPKM, Z-score and RSEM values. Fragments Per Kilobase of transcript per Million mapped reads (FPKM) represents the relative expression of a fragment of transcript. RNA-Seq platform generates the trimmed short reads and the FPKM calculated is based on the reads in the process of gene expression quantification. In addition, RSEM is another well-known measurement value for quantifying the transcripts in RNA-Seq data [20]. The Z-score of the expression data was applied to identify whether an oncogene is over-expression or under-expression. A Z-score for a sample refers to the number of standard deviations away from the mean of expression in the reference and, in the formula below, x represents the expression in tumour sample; μ represents the mean expression in reference samples and δ represents the standard deviation of expression in reference samples: