Somatic Mutaome Profile in Human Cancer Tissues

Nayoung Kim, Yourae Hong, Doyoung Kwon, Sukjoon Yoon*
Department of Biological Sciences, Center for Advanced Bioinformatics and Systems Medicine, Sookmyung Women's University, Seoul 140-742, Korea

Somatic mutation is a major cause of cancer progression and varied responses of tumors against anticancer agents. Thus, we must obtain and characterize genome-wide mutational profiles in individual cancer subtypes. The Cancer Genome Atlas database includes large amounts of sequencing and omics data generated from diverse human cancer tissues. In the present study, we integrated and analyzed the exome sequencing data from ∼3,000 tissue samples and summarized the major mutant genes in each of the diverse cancer subtypes and stages. Mutations were observed in most human genes (∼23,000 genes) with low frequency from an analysis of 11 major cancer subtypes. The majority of tissue samples harbored 20–80 different mutant genes, on average. Lung cancer samples showed a greater number of mutations in diverse genes than other cancer subtypes. Only a few genes were mutated with over 5% frequency in tissue samples. Interestingly, mutation frequency was generally similar between non-metastatic and metastatic samples in most cancer subtypes. Among the 12 major mutations, the TP53, USH2A, TTN, and MUC16 genes were found to be frequent in most cancer types, while BRAF, FRG1B, PBRM1, and VHL showed lineage-specific mutation patterns. The present study provides a useful resource to understand the broad spectrum of mutation frequencies in various cancer types.

Keywords: human tissue samples, metastasis, mutation frequency, TCGA

Introduction

Recent progress in high-throughput sequencing technology has contributed to the generation of genome-wide somatic mutation profiles in diverse cancer samples. The Cancer Genome Atlas (TCGA) is one of largest collaborative efforts to generate multi-level omics data on human cancer tissue samples. Particularly, information on genome-wide somatic mutations has been collectively profiled from exome sequencing data from thousands of patients’ tumor samples. Somatic mutation is a main driving force for cancer development and progression. Thus, many researchers have tried to complete the catalog of somatic mutations in cancer cell lines [1, 2]. Somatic mutation is also known to be involved in key mechanisms for cellular sensitivity or resistance against chemotherapy [3-5].

In our previous study using cancer cell line data, we reported that somatic mutation was a more significant classifier than cancer lineage in predicting the anticancer drug response [6]. Thus, we identified many unknown association patterns between cancer drug response and mutational genotypes in cancer cell lines—e.g., MYC-amp mutation-specific sensitivity of insulin-like growth factor 1 receptor inhibitors. In addition, mutation information provided important clues for us to better interpret the biological relevance of molecular signatures identified from the transcriptome and proteome data of diverse cancer cell lines. The next step should be to find out the clinical application of mutation-specific drug responses or molecular signatures obtained from cell line-based analysis.

Thus, it is important to systematically analyze the mutational genotype (mutaome) of various human tissue samples and identify mutations significantly associated with specific types of tumors. ‘Mutaome’ means the cancer mutational landscape, including mutations in oncogenes and tumor suppressors. In the present study, we organized all sequence-based mutation information into gene-based frequency data. Then, we comparatively determined the major genes of somatic mutations in diverse cancer subtypes and cancer stages (i.e., non-metastatic and metastatic samples). This work will provide practical information for directing in vitro cell line-based mutation-specific phenotypes to clinical
applications in cancer drug discovery and mechanism studies.

Methods

Data acquisition

Somatic mutation data for tumor tissue samples from 2938 patients, harboring 11 cancer types, were obtained from the data portal of TCGA, which were freely available. These data (level 2) for 10 cancer types, except ovarian serous cystadenocarcinoma (OV), provide genome-wide somatic mutations on each sample experimented with the Illumina Genome Analyzer DNA Sequencing platform (Illumina, San Diego, CA, USA). The somatic mutation data for OV were organized, combined with data produced from Illumina and the ABI SOLiD DNA System Sequencing platforms (Applied Biosystems, Foster City, CA, USA). The details, including the full name of each cancer type and updated dates of mutation data, are provided in Table 1.

Together with somatic mutation data, clinical data for each patient were downloaded from the data portal of TCGA. These data were applied to categorize samples into metastasis and non-metastasis. The annotation of ‘pathologic M’ was available to indicate the stage of metastasis in the patient’s tumor samples. The stage of M0’ means that there was no evidence of distant metastasis, and ‘M1’ means that a pathological distant metastasis was found. In this study, the samples from patients annotated as ‘M0’ were classified as non-metastatic, and those with ‘M1’ were classified as metastatic.

Analysis of somatic mutations

The number of detected mutations in each cancer type ranged from tens to hundreds of thousands. We organized the detected point mutations into 23,050 human genes. The number of mutant genes per sample was counted for 3,000 samples. In addition, the mutation frequency and its percentage were calculated for all samples and each cancer type. The major mutant genes that ranked within the top 3 in each cancer type were selected based on the observed frequency in the overall, non-metastatic, and metastatic samples.

The patterns of frequency for the selected major genes were analyzed through hierarchical clustering method. The clustering and its visualization on a heatmap were performed using the software QCanvas [7]. QCanvas can be downloaded freely from the website http://compbio.sookmyung.ac.kr/~qcanvas.

Results and Discussion

Mutation frequency in patients’ tumor samples

From the TCGA database, thousands of patients’ tumor samples were analyzed to detect variants in the whole genome. To quantify the genome-wide mutation profile in diverse tumors, we organized the whole mutant genes into several cancer types and stages, using the annotation information obtained from TCGA. Overall, a total of 871,684 mutations were virtually found in 23,050 human genes from 2,938 patient tumor samples (Table 1). Samples covered 11 diverse cancer types (Table 1). These data were continuously updated and produced by analyzing additional samples. Together with the extended production of multi-level omics data using the same patients’ tissue samples, TCGA provides a useful resource for understanding the role of mutations in cancer progression.

Table 1. Survey of genome-wide somatic mutations in patient tumor samples

| Lineage         | Cancer subtype | Full name of subtype                     | Updated date | No. of samples | No. of detected mutations | No. of mutant genes |
|-----------------|----------------|------------------------------------------|--------------|----------------|----------------------------|---------------------|
| Lung            | LUAD           | Lung adenocarcinoma                      | Apr 30, 2013 | 394            | 179,654                    | 17,307              |
|                 | LUSC           | Lung squamous cell carcinoma             | Apr 30, 2013 | 178            | 65,305                     | 14,873              |
| Ovary           | OV             | Ovarian serous cystadenocarcinoma        | Apr 18, 2013 | 463            | 27,645                     | 12,384              |
| Thyroid         | THCA           | Thyroid carcinoma                       | Apr 25, 2013 | 371            | 21,685                     | 5,798               |
| Head and Neck   | HNSC           | Head and neck squamous cell carcinoma    | May 8, 2013  | 323            | 142,936                    | 15,483              |
| CNS             | GBM            | Glioblastoma multiforme                 | May 10, 2013 | 290            | 21,947                     | 21,553              |
| Uterine         | UCEC           | Uterine corpus endometrioid carcinoma    | May 7, 2013  | 248            | 184,861                    | 19,647              |
| Kidney          | KIRC           | Kidney renal clear cell carcinoma       | May 13, 2013 | 234            | 36,097                     | 10,909              |
| Colon           | COAD           | Colon adenocarcinoma                    | Apr 30, 2013 | 220            | 114,594                    | 17,045              |
| Bladder         | BLCA           | Bladder urothelial carcinoma            | May 7, 2013  | 136            | 51,957                     | 13,985              |
| Rectum          | READ           | Rectum adenocarcinoma                   | May 13, 2013 | 81             | 25,003                     | 10,563              |
| Total           |                |                                          |              | 2,938          | 871,684                    | 23,050              |

Mutations and tissue information are collected and processed from The Cancer Genome Atlas (TCGA) database.
Overall mutation frequency in tissue samples. (A) The distribution of 2,938 patient tissue samples for the number of mutant genes per sample. (B) The distribution of 23,050 mutant genes for % mutation frequency. (C) Box plot of the number of mutant genes per sample for each cancer subtype. (D) Box plot of % sample frequency of individual mutation for each cancer subtype. Number in the box represents median value. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; THCA, thyroid carcinoma; HNSC, head and neck squamous cell carcinoma; GBM, glioblastoma multiforme; UCEC, uterine corpus endometrioid carcinoma; KIRC, kidney renal clear cell carcinoma; COAD, colon adenocarcinoma; BLCA, bladder urothelial carcinoma; READ, rectum adenocarcinoma.

The distribution of mutant genes on each sample showed that a variety of genes are mutated in individual tumor samples. Most samples contained on average 20-80 mutations (Fig. 1A). This means that each single tumor sample has mutations in multiple genes. Various mutations in a cancer sample were already well-characterized and constructed as open source data [8], and the significance of multiple mutations in a single tumor has been constantly suggested [9, 10]. It is appropriate that a tumor may also consist of a heterogeneous collection of cells with different types of mutations. Furthermore, in the aspect of lineage dependency, the amount of mutations in individual samples varied, depending on cancer type (Fig. 1C). Lung adenocarcinoma samples have a wide range of mutation frequencies in individual samples and harbor more mutations than other lineages. The broad range of mutation frequencies in lung cancer was also referred to in Lawrence et al. [11]. In contrast, thyroid carcinoma has relatively few mutations in individual samples. Further studies are required to understand the association with the amounts of mutations and major biological factors in each cancer type. This can be analyzed by comparing with the patient's clinical information, including smoking.

On the other hand, the mutation of each gene was observed at a very low frequency in all samples (Fig. 1B). Most genes contained mutations in less than 1% of samples, and only a few genes contained mutations in over 5% of all samples. Generally, the mutation of a gene showed low frequency in all cancer types, except in uterine corpus endometrioid carcinoma (UCEC) (Fig. 1D). Mutant genes were found in 2.4% of UCEC samples (>2-fold greater than...
other lineages), and sometimes, a gene showed a mutation in >8% of UCEC tumors. In conclusion, there are only a few genes in which mutations are frequently (i.e., >1-2%) found in tumor. Genes with relatively frequent mutations in tumors may have a significant role in cancer progression.

**Comparative analysis of mutations between non-metastatic and metastatic samples**

TCGA provides annotations for the stage of metastasis in each patient sample from the clinical data. According to these data, thousands of samples for 8 cancer types, except for OV, glioblastoma multiforme (GBM), and UCEC, were classified into 89 metastatic and 1341 non-metastatic samples in order to compare the mutation frequency between them. The annotation for metastasis was not provided for the excluded cancer types—OV, GBM, and UCEC. Interestingly, metastatic samples had similar mutation frequencies as non-metastatic samples for most cancer types (Fig. 2). Kidney renal clear cell carcinoma (KIRC) had significant (p < 0.01) difference in frequency between metastatic and non-metastatic samples. This result implied that there is no differential occurrence of mutations between metastatic and non-metastatic samples, except in minor case.

**Table 2.** List of mutant genes showing high mutation frequency in each cancer lineage

| Gene    | Total | LUAD | LUSC | OV  | THCA | HNSC | GBM | UCEC | KIRC | COAD | BLCA | READ |
|---------|-------|------|------|-----|------|------|-----|------|------|------|------|------|
| TP53    | 1,399 | 213  | 146  | 385 | 3    | 231  | 84  | 71   | 13   | 122  | 67   | 64   |
| TTN     | 1,189 | 239  | 141  | 100 | 19   | 170  | 93  | 95   | 73   | 129  | 88   | 42   |
| MUC16   | 758   | 202  | 100  | 29  | 23   | 87   | 64  | 60   | 49   | 74   | 56   | 14   |
| CSMD3   | 546   | 179  | 90   | 30  | 6    | 81   | 11  | 52   | 17   | 48   | 25   | 7    |
| ADAM6   | 539   | 201  | 81   | 41  | 115  | 4    | 52  | 0    | 44   | 0    | 44   | 0    |
| LRP1B   | 461   | 161  | 75   | 15  | 3    | 77   | 8   | 32   | 12   | 40   | 23   | 15   |
| USH2A   | 454   | 159  | 71   | 33  | 1    | 51   | 25  | 41   | 12   | 38   | 15   | 8    |
| PKI3CA  | 403   | 28   | 28   | 6   | 3    | 66   | 32  | 133  | 10   | 60   | 28   | 9    |
| APC     | 332   | 26   | 9    | 12  | 1    | 14   | 2   | 30   | 2    | 160  | 7    | 69   |
| PTEN    | 332   | 7    | 14   | 5   | 2    | 6    | 90  | 161  | 16   | 20   | 5    | 6    |
| BRAF    | 321   | 32   | 9    | 2   | 217  | 5    | 7   | 8    | 1    | 34   | 2    | 4    |
| LOC96610| 309   | 112  | 32   | 0   | 38   | 50   | 1   | 0    | 40   | 0    | 36   | 0    |
| PLEC    | 239   | 50   | 22   | 2   | 5    | 43   | 8   | 22   | 8    | 49   | 15   | 15   |
| FNG1B   | 237   | 61   | 0    | 1   | 51   | 49   | 19  | 3    | 23   | 1    | 29   | 0    |
| PTPTN   | 182   | 48   | 24   | 11  | 1    | 20   | 7   | 19   | 5    | 33   | 7    | 7    |
| PBRM1   | 171   | 8    | 8    | 3   | 2    | 13   | 2   | 12   | 97   | 13   | 10   | 3    |
| VHL     | 130   | 0    | 2    | 0   | 0    | 1    | 1   | 3    | 122  | 1    | 0    | 0    |
| STAG2   | 99    | 15   | 7    | 4   | 1    | 6    | 12  | 25   | 7    | 4    | 17   | 1    |
| LOC653544| 69   | 22   | 1    | 0   | 9    | 12   | 9   | 0    | 14   | 0    | 2    | 2    |
| FAM182B | 46    | 17   | 0    | 0   | 17   | 8    | 0   | 0    | 9    | 0    | 1    | 0    |

A total of 20 mutant genes are listed. These genes were derived from 12, 10, and 15 mutant genes, which are ranked within the top 3 in at least one cancer type based on the % frequency in overall samples, non-metastatic samples, and metastatic samples, respectively. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; THCA, thyroid carcinoma; HNSC, head and neck squamous cell carcinoma; GBM, glioblastoma multiforme; UCEC, uterine corpus endometrioid carcinoma; KIRC, kidney renal clear cell carcinoma; COAD, colon adenocarcinoma; BLCA, bladder urothelial carcinoma; READ, rectum adenocarcinoma.

**Identification of major mutant genes**

In this study, the mutation frequency was analyzed separately in overall, non-metastatic, and metastatic samples. We focused on mutant genes exhibiting high frequency in each sample group. A total of 12, 10, and 15 genes were ranked within the top 3 mutations in at least one cancer type in the overall, non-metastatic, and metastatic samples, respectively (Table 2). Especially, TP53 and TTN showed dominant frequencies for over 1,000 of all samples.

In addition, the pattern of mutation frequency for the selected major mutant genes was analyzed in diverse cancer categories (Fig. 3). Regardless of sample group, MUC16, TTN, and TP53 were found to be frequent in most cancer types. TP53 is a well-known mutant gene, playing an important role in cancer progression [12, 13]. It was reported that TP53 mutation is frequently represented in major cancer lineages [6]. Mutation TTN and MUC16 has not been reported to be critical in cancers. This analysis shows that they may have potential, specific roles in cancer development or progression.

Among 12 major mutant genes selected from the overall samples, BRAF, FNG1B, PBRM1, and VHL had cancer subtype-specific mutation patterns (Fig. 3A). Especially, PBRM1
and VHL showed strong specificity in KIRC tumors. As previously reported [14], the alterations of VHL (a tumor suppressor gene) are clearly dominant in renal cell carcinoma. Together with VHL, PBRM1 was identified as a major gene, frequently mutated in renal carcinoma [15]. An association with the loss of its expression and renal cell carcinoma progression was suggested in previous studies [16].

**Mutant genes dependent on metastasis**

We found that there was no difference in the overall frequency of metastatic and non-metastatic samples (Fig. 2). The nine mutant genes found in non-metastatic samples were all included in the major mutant genes in the overall samples (Fig. 3B). The lineage-dependent frequency in non-metastatic samples was also similar with the overall pattern in Fig. 3A. However, half of the 15 major mutant genes from metastatic samples were different from the mutant genes in non-metastatic or overall samples (Fig. 3C). The mutations of FAM182B, LOC653544, PLEC, STAG2, PTPRT, CSMD3, and LRP1B represented unique frequencies in metastatic samples. Especially, PTPRT is a member of the protein tyrosine phosphatase (PTP) family. The deletion of PTPRD, included in the same PTP family, is frequently seen in metastatic cutaneous squamous cell carcinoma [17]. Further studies are required for other major metastasis-associated mutant genes. In conclusion, the diversity of major mutant genes in metastatic samples is quite different from those in non-metastatic tumors, although the overall mutation frequency is similar between metastatic and non-metastatic tumors. The present study provides a useful resource for understanding the varied frequency of diverse mutations in patients’ tumor samples.

**Acknowledgments**

This research was supported by Sookmyung Women’s University Research Grant 1-1203-0227.

**References**

1. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191-196.

2. Greenman C, Stephens P, Smith R, Dalglish GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153-158.

3. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature
4. Greshock J, Bachman KE, Degenhardt YY, Jing J, Wen YH, Eastman S, et al. Molecular target class is predictive of in vitro response profile. Cancer Res 2010;70:3677-3686.

5. McDermott U, Sharma SV, Dowell L, Greninger P, Montagut C, Lamb J, et al. Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. Proc Natl Acad Sci U S A 2007;104:19936-19941.

6. Kim N, He N, Kim C, Zhang F, Lu Y, Yu Q, et al. Systematic analysis of genotype-specific drug responses in cancer. Int J Cancer 2012;131:2456-2464.

7. Kim N, Park H, He N, Lee HY, Yoon S. QCanvas: an advanced tool for data clustering and visualization of genomics data. Genomics Inform 2012;10:263-265.

8. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 2011;39:D945-D950.

9. Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. Proc Natl Acad Sci U S A 2003;100:776-781.

10. Loeb KR, Loeb LA. Significance of multiple mutations in cancer. Carcinogenesis 2000;21:379-385.

11. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214-218.