Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2018 July ; 2018: 1323–1326. doi:10.1109/EMBC.2018.8512514.

Micro-inversions in Human Cancer Genomes

Li Qu¹, Huaiqiu Zhu², May D. Wang³

¹Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA. She is also with the Dept. of Biomedical Engineering, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, P.R.China 100871.

²Dept. of Biomedical Engineering and Center for Quantitative Biology, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, P.R.China 100871.

³Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA.

Abstract

During the past few years, although scientists and researchers have studied variations in the cancer genomes, our current knowledge about the function of Micro-inversions (MIs) in cancer are still limited. MIs are generally defined as small inversions in DNA segments shorter than 100 bp. To expand our knowledge of their roles in cancer, we analyzed the MIs of 209 samples from four types of cancer, including hepatocellular carcinoma, lung cancer, pancreatic cancer, and bladder cancer. Within all the 209 samples, we identified 2,925 MIs, of which 1,519 (51.93%) are in gene regions. Of the 1,519 MIs in the gene regions, 106 (6.98%) are in the exon regions. We also analyzed 209 healthy samples as the control samples. We further analyzed the distribution of MIs in the four types of cancer among 24 chromosomes. Besides the chromosome preference, different cancers also have different preference for various genes. The MIs preference for different genes among four types of cancer may provide a guidance for the treatment and diagnosis on the four types of cancer. Medical doctors should concentrate more on chromosomes and the genes that MIs prefer to locate on. We also calculated the average count of MIs per individual among each cancer. From this result, we found that the bladder cancer has the most average count of MIs per individual, which means MIs may be more likely to exist in bladder cancer. According to our analysis, MIs play an important role in cancer and should be considered for further analysis.

1. INTRODUCTION

Structural variations (SVs), which generally include insertion, deletion, duplication, inversion, rearrangement, translocation, and copy number variation, are reported to contribute to human diseases including cancer [1]. Due to the development of next-generation sequencing technologies, many new patterns of SVs have been discovered in cancer genomes. Besides, the algorithms using pair-end sequencing data make it possible to identify the patterns of SVs that characterize particular cancer types. These patterns reflect that SVs may only be restricted to a specific cancer or show various prevalence across

cancer types. Although some SVs generated don't necessarily lead to cancer, some may result in critical illness because of altering key genes, enhancers, or promoters [2].

A number of studies have shown that one specific cancer may have SVs in a specific gene. Patients with lung cancer are more likely to have variations on the genes DICER1 and DROSHA [3]; patients with hepatocellular carcinoma are more likely to have variations on the gene SALL4 [4]; Patients with pancreatic cancer are more likely to have SVs on the gene BRAF [5]. Since SVs in different cancers may have various preference of various gene regions, studying the preference of genes among different cancers is especially significant. Besides the difference of SVs among cancers, there are also some regions found highly-centralized in all cancer genomes. Many kinds of tumors were found with the similar replicating region [6]; gene ETV6/RUNX1 were found with deletions in acute lymphoblastic leukemia [7]; lung cancer were found with copy number variations [8]; and cervical cancer were also found with copy number gain [9]. In summary, some SVs took place in only one specific type of cancer, while some SVs are shared by two or more types of cancer.

MIIs are generally defined as small inversions, which are one kind of SVs, in DNA segments shorter than 100 bp. Previous studies have reported discoveries of small inversions in the human and mammalian genomes [10]. However, these inversions are too large, usually of thousands of bp, which don't fit the definition of MIIs studied in this paper. Although previous researches have studied the relationship of SVs and cancer [2], [3], MIIs have not been studied on the human genomes, because of limitations of methods of detecting MIIs. Therefore, the function of MIIs in cancer is not clear. In this paper, we have generated a comprehensive analysis of MIIs among four types of cancer.

2. MATERIALS AND METHODS

The analysis of the samples in this study mainly include sample collection, MI detection and annotation, MI chromosome preference, and top gene analysis (Figure 1). The method details are described as below:

2.1 Sample Collection

To find the features of MIIs among human cancer genomes, we collected BAM files of the unmapped reads of 209 samples from Sequence Read Archive (SRA) database [11]. These samples are from patients of 67 hepatocellular carcinoma, 62 lung cancers, 14 pancreatic cancers, and 66 bladder cancers. The total 209 Illumina samples are all pair-end reads from whole genome sequencing and the length of each read is 151 bp. Besides, all these patients including both males and females in this study are from China, of which the patient age is from 51 to 70 years old. We also downloaded 209 BAM files of unmapped reads of healthy individuals from the recent version of the 1000 Genome Project (1KGP) as the control group. Considering that all the cancer samples are from China, the 209 healthy samples are selected from the populations of CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), and CDX (Chinese Dai in Xishuangbanna, China), which are all Chinese populations.

2.2 MI Detection and Annotation

We used the software MID [12] with default parameters, setting the anchors length at 18 bp and mismatches length at 2 bp. Then, we mapped all the unmapped sequencing reads to the human assembly hg19 [13]. MID is the software that can efficiently detect MIs from unmapped short reads with the bam format through inversion reading and reference genome mapping. The reason that we chose MID as the MI detection tool in this study is, MID is very sensitive to very small size MIs which has length between 10 bp and 100 bp. There are also other inversion detection tools, in which the inversion length is too large that doesn't meet the definition of MIs. After identifying MIs using the MID, we annotated these MIs with the gene function via the GENCODE database [14]. After the annotation process, we classified the MIs into intergenic regions and gene regions. Furthermore, among the MIs in gene regions, we classified the MIs into exon regions and intron regions.

2.3 Average MI Count Per Individual

Since the sample number varies by cancer types, the difference among samples of four types of cancer and healthy samples can be better reflected by the average count of MIs per individual among one cancer. To clarify the question, we define several variables as follows. We define C_i as the count of MIs per sample, N_P as the total number of samples in one specific cancer or the healthy samples, and $\bar{C}_P$ as the average count of MIs per individual among the individuals of one specific cancer. We can calculate the $\bar{C}_P$ using the following equation:

$$\bar{C}_P = \frac{\sum_{i=1}^{N_P} C_i}{N_P} \# \quad (1)$$

2.4 Venn Diagram Analysis

To further investigate the distribution pattern of genes that overlap with MIs among four types of cancer and the healthy samples, we conducted a gene sharing analysis. First, we extracted the MIs that overlapped with genes according to the annotation results from process B. Then, we summarized the genes that overlapped with these MIs among the four types of cancer and the healthy samples. At last, we filtered these genes, removed the overlapped genes, and got the unique genes. We used these unique genes to do a Venn diagram to find the intersection of genes of the four types of cancer and the healthy samples.

3. RESULTS AND DISCUSSIONS

In this study, we analyzed 209 samples from four types of cancer. Using MID, we identified 2,925 MIs, of which 1,519 (51.93%) overlap with gene regions. Of the 1,519 MIs overlapped with gene regions, 106 (6.98%) overlap with the exon regions. Of the 2,925 MIs, 2,060 MIs are unique, which means that 865 MIs overlapped with at least one MI in the 2,060 MIs. We also include 209 samples of healthy people from 1KGP. From the healthy people, we identified only 484 MIs, of which 237 (48.97%) MIs are in gene regions. Of the 237 MIs in gene regions, only 16 (6.75%) are in exon regions. This result implies that MIs are more likely to happen in the cancer samples (2,925 MIs) compared with healthy samples

(484 MIs). However, the frequency of MIs in gene regions of both the cancer samples (51.93%) and healthy samples (48.97%) is high. The details of these results are shown in TABLE I. With these samples, we do the following analysis based on our MI results:

3.1 MI Distribution among 24 Chromosomes

The MI distributions among 24 chromosomes of four cancer samples and normal samples are shown in Figure 2. Generally, the number of MIs correlated with the length of chromosomes. This result is understandable that, longer chromosomes which have more genes and base pairs are more likely to make a mistake. However, the MI number is not strictly related to the chromosome length in some chromosomes. For example, the result in the Figure 2 shows that chromosome 3, 6, 13, 17 in cancer performs differently with other chromosomes. Specifically, chromosome 3 has more MIs than chromosome 2, but the chromosome 2 is longer than chromosome 3. The same phenomenon is also widely happened in the healthy controls (Chromosomes 2–6, 8, 11, 14, 17, and 19). In summary, MIs of cancer and healthy individuals show different prevalence across chromosomes. Medical doctors should concentrate more on the chromosomes that MIs prefer to be located on in cancer samples rather than the healthy samples.

Some studies have shown that one specific cancer may have SVs in a specific gene region on the genome. For example, patients with lung cancer are more likely to have variations on the genes DICER1 and DROSHA [3]. Our results suggest that MIs of four cancers have various preference of locations of chromosomes. Besides, MIs tend to happen more frequently in intron regions than in exon regions.

3.2 Average Count of MIs Per Individual

The average count of MIs per individual among the four types of cancer is shown in Figure 3. In general, the average count of MIs per individual in the normal samples (1.89) is lower than that of any type of cancer (The lowest count of MIs per individual is hepatocellular carcinoma of which is 5.36). Besides, the difference among the average count of MIs per individual in different cancers is very appreciable, varying from 5.36 to 21.98. The bladder cancer has the highest average count of MIs per individual with the C_p of 21.98 and the pancreatic cancer has the lowest average count of MIs per individual with the $\bar{C}_p$ of 3.33. This result shows that MIs are more likely to exist in patients of these four types of cancer compared with healthy people. Moreover, the patients with bladder cancer are the most likely to have more MIs in genomes compared with the other three cancers.

3.3 Venn Diagram of Genes among Four types of cancer

To further investigate the distribution pattern of genes that overlap with MIs among the four types of cancer and healthy samples, we conducted a gene sharing analysis. We used unique genes that overlap with MIs from the four types of cancer including hepatocellular carcinoma, lung cancer, pancreatic cancer, and bladder cancer and healthy samples. The Venn diagram of genes in four types of cancer and healthy samples is shown in Figure 4. In general, the four types of cancer have more specific genes of themselves than the ones shared by at least two cancers. Although no genes are shared by all the four types of cancer, some genes are common in three or two cancers. Accounting that some MIs in cancer

samples also exist in healthy samples, we focused more on the MIs in cancer samples but not in the healthy samples. We infer that MIs in normal samples are neutral and not directly related to cancer induction.

From the Venn diagram, the cancer of bladder cancer, hepatocellular carcinoma, and lung cancer all share two genes, AC079807.4 and EYS genes. However, the AC079807.4 also exist in the normal samples. Thus, we focused on the EYS gene that exist in only cancer samples, which may be associated with the three cancers. The bladder cancer, hepatocellular carcinoma, and pancreatic cancer all have GPC6 gene in common, which doesn't exist in healthy samples. Since GPC6 gene was reported to be related with the ovarian cancer [15], we infer that MIs in GPC6 gene is also related with the incidence of the three cancers, bladder cancer, hepatocellular carcinoma, and pancreatic cancer. The bladder cancer and lung cancer have the most shared genes in common, which is 25 genes. This result shows that the two cancers have many similarities in the MIs on the human genome. Thus, the similar gene therapeutic methods should be designed for bladder cancer and lung cancer.

3.4 Top Genes Overlapped with MIs

We also summarized the top five genes that overlap most MIs among the four types of cancer in TABLE II a), including TNIK, SQSTM1, CTTNBP2, TIAM2, and GNG12. TNIK was reported to be associated with distant recurrence in stage II and III colorectal cancer patients [16]; SQSTM1 was also reported to be related with gastric cancer [17]. We also listed the top three genes in the four types of cancer including hepatocellular carcinoma, lung cancer, pancreatic cancer, and bladder cancer in TABLE II b) – e). In summary, MIs in the four types of cancer have various preference for genes.

4. CONCLUSIONS

In this study, we have generated a comprehensive analysis of MIs. We explored the distribution of MIs among 24 chromosomes in four types of cancer. Comparison of MIs among the four types of cancer and healthy samples were performed in accordance to the different genes that overlap with MIs. Our analysis results show that different cancers have various preference for chromosomes and genes. The results indicate that MIs may be associated with cancer development. The comparative MI analysis of MIs in four different cancers will be the cornerstone of further implementation of precision medicine of individuals and the unveiling of disease mechanisms. One limitation of this study is the small sample size and the number of MIs. Since the number of MIs in each sample is small, only 2,925 MIs from 209 samples are detected. As a next step, we plan to include more samples including more kinds of cancers and detect more MIs to enlarge the cohort of this study. Besides, we will focus more on the MIs that only exist in only one specific cancer. Thus, we will explore the relationship of MIs and a specific cancer.

Acknowledgments

This work was supported in part by grants from the National Science Foundation Award NSF1651360, Microsoft Research, Hewlett Packard, and China Scholarship Council (201606010277). The content is solely the responsibility of the authors and does not necessarily represent the official views of National Science Foundation and China Scholarship Council.

6. REFERENCES

- [1]. Feuk L, Marshall CR, Wintle RF, and Scherer SW, "Structural variants: changing the landscape of chromosomes and design of disease studies,". *Human molecular genetics*, vol. 15, no. suppl_1, pp. 57–66, 2006.
- [2]. Stratton MR, Campbell PJ, and Futreal PA, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009. [PubMed: 19360079]
- [3]. Czubak K, Lewandowska MA, Klonowska K, Roszkowski K, Kowalewski J, Figlerowicz M, Kozłowski P, "High copy number variation of cancer-related microRNA genes and frequent amplification of DICER1 and DROSHA in lung cancer," *Oncotarget*, vol. 6, no. 27, p. 23399, 2015. [PubMed: 26156018]
- [4]. Liu TC, Vachharajani N, Chapman WC, and Brunt EM, "SALL4 immunoreactivity predicts prognosis in Western hepatocellular carcinoma patients but is a rare event: a study of 236 cases," *The American journal of surgical pathology*, vol. 38, no. 7, pp. 966–972, 2014. [PubMed: 24805857]
- [5]. Heestand GM, and Kurzrock R, "Molecular landscape of pancreatic cancer: implications for current clinical trials," *Oncotarget*, vol. 6, no.7, p. 4553, 2015. [PubMed: 25714017]
- [6]. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, Schrock A, Campbell B, Shlien A, Chmielecki J et al., "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden," *Genome medicine*, vol. 9, no. 1, p. 34, 2017. [PubMed: 28420421]
- [7]. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, Alexandrov LB, Loo PV, Cooke SL, Marshall J et al., "RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia," *Nature genetics*, vol. 46, no. 2, pp. 116–125, 2014. [PubMed: 24413735]
- [8]. Stratton MR, "Exploring the genomes of cancer cells: progress and promise," *science*, vol. 331, no. 6024, pp. 1553–1558, 2011. [PubMed: 21436442]
- [9]. Tubio JM, "Somatic structural variation and cancer," *Briefings in functional genomics*, vol. 14, no. 5, pp. 339–351, 2015. [PubMed: 25903743]
- [10]. Chaisson MJ, Raphael BJ, and Pevzner PA, "Microinversions in mammalian evolution," *Proceedings of the National Academy of Sciences*, vol. 103, no. 52, pp. 19824–19829, 2006.
- [11]. Shumway M, Cochrane G, and Sugawara H, "Archiving next generation sequencing data," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D870–D871, 2009. [PubMed: 19965774]
- [12]. He F, Li Y, Tang YH, Ma J, and Zhu H, "Identifying micro-inversions using high-throughput sequencing reads," *BMC genomics*, vol. 17, no. 1, p. 4, 2016. [PubMed: 26818118]
- [13]. 1000 Genomes Project Consortium. "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015. [PubMed: 26432245]
- [14]. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al., "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome research*, vol. 22, no. 9, pp. 1760–1774, 2012. [PubMed: 22955987]
- [15]. Karapetsas A, Giannakakis A, Dangaj D, Lanitis E, Kynigopoulos S, Lambropoulou M, Tanyi JL, Galanis A, Kakolyris S, Trypsianis G et al., "Overexpression of GPC6 and TMEM132D in Early Stage Ovarian Cancer Correlates with CD8," *BioMed Research International*, vol. 2015, 2015.
- [16]. Masuda M, Uno Y, Ohbayashi N, Ohata H, Mimata A, Kukimoto-Niino M, Moriyama H, Kashimoto S, Inoue T, Goto N et al., "TNIK inhibition abrogates colorectal cancer stemness," *Nature communications*, vol. 7, p. 12586, 2016.
- [17]. Cao QH, Liu F, Yang ZL, Fu XH, Yang ZH, Liu Q, Wang L, Wan X, and Fan X, "Prognostic value of autophagy related proteins ULK1, Beclin 1, ATG3, ATG5, ATG7, ATG9, ATG10, ATG12, LC3B and p62/SQSTM1 in gastric cancer," *American journal of translational research*, vol. 8, no.9, p. 3831, 2016. [PubMed: 27725863]

Figure 1.
The flow of MI analysis.

Figure 2.
MI distribution among 24 chromosomes.

Figure 3.
Average count of MIs per individual.

Figure 4.

Venn diagram of genes that overlap with MIs in four cancers. Bladder, hepatocellular, lung, pancreatic, and healthy refers to bladder cancer, hepatocellular carcinoma, lung cancer, pancreatic cancer, and healthy individuals respectively.

Table 1**SUMMARY OF MI RESULTS**

Cancer Type	Sample number	MI indicator		
		<i>MI-num</i> ^a	<i>MI-gen</i> ^b	<i>MI-ave</i> ^c
Hepatocellular carcinoma	67	359	208	5.36
Lung cancer	62	977	520	15.76
Pancreatic cancer	14	138	82	9.86
Bladder cancer	66	1,451	709	21.98
Total	209	2,925	1,519	14.00
Healthy Samples	209	484	237	1.89

^aThe “MI-num” column illustrates the number of Mis for each cancer.

^bThe “MI-gen” column illustrates the number of Mis that overlap with genes for each cancer.

^cThe “MI-Ave” column illustrates the average count of Mis per individual in each cancer.

Table 2**TOP GENES OVERLAPPED WITH MIS**

a)	TOP FIVE GENES AMONG FOUR TYPES OF CANCER		
	Rank	Gene Name	Number of MIs overlap
	1	TNIK	12
	2	SQSTM1	11
	3	CTTNBP2	11
	4	TIAM2	9
	5	GNG12-AS1	9
b)	TOP THREE GENES AMONG BLADDER CANCER		
	Rank	Gene Name	Number of MIs overlap
	1	GNG12-AS1	7
	2	TNIK	7
	3	CTTNBP2	7
c)	TOP THREE GENES AMONG HEPATOCELLULAR CARCINOMA		
	Rank	Gene Name	Number of MIs overlap
	1	TIAM2	9
	2	LPP	7
	3	AC002454.1	6
d)	TOP THREE GENES AMONG LUNG CANCER		
	Rank	Gene Name	Number of MIs overlap
	1	SORCS1	3
	2	AL163953.3	3
	3	RP11-420N3.2	3
e)	TOP THREE GENES AMONG PANCREATIC CANCER		
	Rank	Gene Name	Number of MIs overlap
	1	TNIK	4
	2	SQSTM1	4
	3	C12orf75	2