1 Pan-cancer scale landscape of simple somatic mutations

2
3 Nan Zhou^{1,2}, Blanca Gallego², Jinku Bao^{1,*}, Guy Tsafnat^{2,*}

4
5 ¹ Key Laboratory of the State Ministry of Education for Bio-resources and Ecologic
6 Environment, College of Life Sciences, Sichuan University, Chengdu 610064, China

7 ² Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences,
8 Macquarie University, NSW 2109, Australia

9
10 * To whom correspondence should be addressed

11 Jinku Bao

12 baojinku@scu.edu.cn

13 Tel: +86-28-85415171

14
15 Guy Tsafnat

16 guy.tsafnat@mq.edu.au

17 Tel: +61-2-9850-2430

Abstract

Genome is the carrier of somatic mutations during the development of cancer. The catalogue of simple somatic mutations (SSM) is a subgroup of somatic mutations. It includes single base substitutions, small deletions and insertions of ≤ 200 bp, and multiple base substitutions of ≤ 200 bp. The comprehensive landscape of SSM has not been studied. After analysed 46,692,922 SSM of 10,878 samples, we proposed a pan-cancer scale landscape of SSM for 60 cancer projects in ICGC. In addition, the whole genome sequencing (WGS) and whole exome sequencing (WXS) techniques were compared according to the landscape of SSM. The result indicates numbers of SSM vary dramatically in different cancers. WGS can detect 10 times more single base substitutions and insertions than WXS. In terms of WXS, it called 10 times more deletions than insertions. Multiple base substitutions have not been well studied so they were just observed in a few cancer projects. Cancers generally show high prevalence of C > T substitutions at NpCpG trinucleotide contexts. Skin cancer showed distinct mutational spectra. Breast cancer, bladder cancer, and cervical cancer were found to have similar mutational spectra. Acute myeloid leukemia and lung cancer from South Korea, and colorectal cancer from China show high density of single base substitutions per mega base in chromosome Y. To sum up, our study and findings will be thought provoking in studying SSM in cancer.

Keywords: Somatic mutation, Cancer, Genome, Landscape

Introduction

It is explicit that cancer genomes carry somatic mutations. Somatic mutations are generally categorised into substitutions, insertions and deletions (indels), copy number variations and structural variations [1,2]. Study of mutations of a single gene like p53 in cancer can date back to decades ago [3]. After the advent of high throughput sequencing technology around 2005, genome wide studies have unprecedentedly enriched data of somatic mutations in cancer. For example, projects like TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium) have promoted sequencing of thousands of cancer genomes. In ICGC, somatic mutations are provided as simple somatic mutations (SSM). SSM consists of single base substitutions, deletions of ≤ 200 bp, insertions of ≤ 200 bp, and multiple base substitutions (≥ 2 bp and ≤ 200 bp), which form a subgroup of somatic mutations.

The characteristics of complex somatic mutations in cancers can be clearly shown by mutational landscape [4,5]. For instance, exogenous carcinogenic factors induced somatic mutations in lung cancer and melanoma have been identified by studying the landscape of mutations [6,7]. Nowadays, large scale landscape of somatic mutations of all main cancer types are also available from other studies [8,9]. They indicated that landscape of mutations is an insightful way to study somatic mutations in cancer.

Whole genome sequencing (WGS) and whole exome sequencing (WXS) are two widely used strategies in cancer genomics. The difference between somatic mutations generated by WGS and WXS has not been well studied. Landscape of SSM in ICGC has not been studied yet. In this study, we provided a comprehensive landscape of SSM of all available cancer projects in

ICGC and compared the difference between WGS and WXS in generating SSM at the pan-cancer scale the first time.

Materials and methods

Data collection and preparation

SSM data of cancer genomes used in this study was from release 23 of ICGC [10,11]. It was released on December 7, 2016. It includes 70 cancer projects, 21 cancer primary sites, and 46,693,172 SSM. These data were downloaded according to projects. Because 10 projects lack SSM data, only SSM of 60 cancer projects were downloaded and analysed. Because a mutation can have multiple impacts in the genome, there are duplicated records in the data. Duplications are removed according to the ICGC mutation id. Mutations not found by WGS and WXS were also removed. Finally, the data used for further analysis contains 60 cancer projects, 10,878 samples and 46,692,922 unique SSM in total. Details about the data used in this study were available in Table S1.

Chromosome size

The chromosome size of whole genome was from human reference genome UCSC hg19 [12]. The size of whole exome and length of exons in each chromosome were calculated from human exome of Ensemble release 75 [13]. The exome data was retrieved from Ensembl's MySQL database using "select R.name, E.seq_region_start, E.seq_region_end, E.stable_id from exon as E, seq_region as R where E.seq_region_id=R.seq_region_id". Length of exons in each

chromosome was calculated subsequently. The chromosome size for WGS and WXS was listed in Table S2.

Number of mutations per mega base (Mb) in each chromosome for single base substitutions was calculated. In each chromosome, the total number of mutations was divided by the length in Mb of the corresponding chromosome, either using chromosome size of WGS or WXS.

Mutational spectra

Single base substitutions were firstly classified into six types, C > A, C > G, C > T, T > A, T > C, T > G [14]. Each type was further divided into 16 trinucleotide spectra by integrating a flanking base immediate to the mutated base at 5' and 3'. Finally, single base substitutions are categorised into 96 spectra. Indels and multi base substitutions were classified according to the length of mutated bases. Indels were categorised into four spectra, <= 3 bases, 4-10 bases, 11-30 bases, and >= 31 bases. Multi base substitutions were categorised into three spectra, 1-20 bases, 21-50 bases, and >= 51 bases.

K-S test

In order to see whether a larger chromosome would carry more single base substitutions, Kolmogorov-Smirnov test (K-S test) was used to test if the distribution of mutations across chromosomes was drawn from the same distribution of chromosome length. Because the number of mutations and length of chromosomes are at different scale, they were normalized into range [0, 1] using feature scaling approach $X' = (X - X_{\min}) / (X_{\max} - X_{\min})$, where X is a numeric vector and X' is the normalised X. The K-S test was implemented by R function ks.test,

with parameter alternative set to two.sided [15]. A p value of ≥ 0.05 indicates two distributions are the same, otherwise significantly different.

Results

SSM in cancer

As can be seen in Figure 1, the number of single base substitutions dominates SMM. Deletions and insertions take a relative large portion while multiple base substitutions account for the smallest composition. Numbers of SSM found by different cancer projects vary dramatically. The median of single base substitutions varies from a few to more than 10,000. Compared to WXS, WGS is more likely to detect more insertions and multiple base substitutions. In each cancer project, the number of identified deletions is proportional to the number of insertions. That means if there are more deletions in a cancer, there should be more insertions, and vice versa.

Regarding to WGS (Figure 1a), the majority of cancers have 1,000 to 100,000 single base substitutions. NBL-US contains approximately 10 single base substitutions, whereas at the opposite extreme, about 1/3 of MELA-AU exceed 100,000. Most cancer possess 10 to 1000 deletions and insertions, but THCA-SA and OV-AU only carry about 10 deletions and insertions. Seven WGS projects identified multiple base substitutions. Five of them carry about 10 multiple base substitutions. ORCA-IN possesses more than 10 multiple base substitutions. In MELA-AU, numbers of multiple base substitutions range from 10 to more than 1000.

In Figure 1b, 10 to 1,000 single base substitutions were detected in the vast majority of WXS projects. Samples of LUCS-KR have the largest median (774) of single base substitutions. In CMDI-US and NBL-US, few single base substitutions were identified. WXS detected 10 times more deletions than insertions. The number of deletions in some cancers can exceed 1,000 but the maximum of insertion is just around 100. Multiple base substitutions were detected in six WXS projects. In four of them, there is only one multiple base substitution. In BRCA-UK and COCA-CN, the number increases to 10 and more.

Figure 1. Distribution of SSM in all samples of 60 ICGC cancer projects. Each point denotes a sample. The number of mutations are on the y axis. For better perception, log10 scaling was used on y axis when creating the plot. Substitution explicitly denotes single base substitutions. M-sub represents multiple base substitutions. Black horizontal lines denote the median of mutations of all samples of each project. (a), Projects using WGS strategy. (b), Projects using WXS strategy. The colour code for cancer projects in (a) has no relation with that in (b), which is a routine in the context.

Spectra of SSM

It is clear in Figure 2 that C > T substitutions have the highest opportunity to occur. C > A and T > C are two other substitutions with high prevalence. The vast majority of indels are 1 to 3 bases in length. Multiple base substitutions only present in a few cancers. Multiple base substitutions of ≤ 20 bases are more prevalent than longer ones.

In WGS projects (Figure 2a), few C > A single base substitutions happen at NpCpG trinucleotides (N can be any base) but high prevalence of C > A transversions evenly distribute at other trinucleotides. BRCA-EU and BRCA-FR display similar spectra of substitution, with high probability of C > G and C > T at TpCpA and TpCpT. Unlike other cancers, ESAD-UK mainly has T > G at CpTpT. Two skin cancers, MELA-AU and SKCA-BR have almost the same spectra which are distinct from other cancers. Single base substitutions in the two skin cancers are enriched with C > T transitions that mainly occur at CpCpN (no CpCpG) and TpCpN.

There is an enrichment of C > T substitutions in WXS projects (Figure 2b). The C > T substitutions prefer to happen at NpCpG trinucleotides. Apart from high prevalence of C > T substitutions, BLCA-CN, BLCA-US, BRCA-KR, BRCA-UK, BRCA-US, and CESC-US similarly show high probability of C > G at TpCpA and TpCpT. COCA-CN, READ-US, and UCEC-US additionally have high prevalence of C > A at TpCpT. Less C > T single base substitutions appear in LICA-CN, but it is observed to have dominant T > A at CpTpG trinucleotide.

Figure 2. Spectra of SSM of 60 ICGC cancer projects. Substitution represents single base substitutions. There are 96 trinucleotide spectra for single base substitutions, in order of A[C>A]A/C/G/T, C[C>A]A/C/G/T, ..., T[T>G]A/C/G/T, from bottom to top on y axis of the bottom panels. Indel are deletions and insertions. Indels are categorised as 1 to 3 bases (1to3), 4 to 10 bases (4to10), 11 to 30 bases (11to30), and greater than or equal to 31 bases (ge31). M-sub is multiple base substitutions. Spectra for M-sub are 1 to 20 bases (1to20), 21 to 50 bases (21to50), and greater than or equal to 51 bases (ge51). The heat map shows the percentage a spectrum takes in each category of SSM across cancer projects. (a) Cancer projects of WGS. (b) Cancer projects of WXS.

Substitutions in chromosomes

Figure 3a demonstrates that all WGS cancers show similar distribution of single base substitutions per Mb across the chromosomes, with chromosome 8 per Mb having the most number of mutations and chromosome Y per Mb having the least. It is also clear to see that MELA-AU has an extremely high density of substitutions in every chromosome per Mb. It has at least three times more substitutions than other cancers per Mb.

200

201 The interlacing lines in Figure 3b show no cancer can completely overtake others in terms of
 202 the distribution of single base substitutions per Mb per chromosome. In many WXS cancers,
 203 there is a peak in chromosome X, no matter how many substitutions per Mb in other
 204 chromosomes. COCA-CN, LAML-KR, and LUSC-KR have an obvious increase of
 205 substitutions in chromosome Y. In LAML-KR, the number stands above 3000 and exceeds
 206 other figures.

207

208 *Indels in chromosomes*

209

210 According to Figure 4, the total number of indels in each chromosome varies in cancer projects,
 211 but it is noticeable that all projects display similar distribution of indels across chromosomes.
 212 In WGS projects (Figure 4a), there is a significant peak of number of indels in chromosome X.
 213 In WXS projects (Figure 4b), a peak of number of indels present in chromosome X as well.
 214 What's more, there two other remarkable peaks in chromosome 17 and 19 in WXS cancer
 215 projects.

216

Figure 4. Distribution of total number of indels in 24 chromosomes. (a) Cancer projects of WGS. (b) Cancer projects of WXS.

Discussion

There are only two samples in ALL-US sequenced by WGS. One sample carries four single base substitutions and the other has one deletion. The PAEN-AU in WXS projects only includes three samples, which contribute 14 single base substitutions, 16 deletions, and 6

insertions. The small number of mutations doesn't bear statistical significance. Therefore, the results of WGS ALL-US and WXS PAEN-AU were only presented in the figures but not discussed in the context.

Compared to WXS, WGS detected 10 times more single base substitutions on average (Figure 1). The protein-coding regions constitute about 1% of human genome [16]. It is expected that more mutations should be called from the whole genome. As the ratio of length between genome and exome is not proportional to the ratio between numbers of substitutions, it is inferred that there is higher density of single base substitutions in the exome. This inference may also apply to deletions. WGS detected the same range of numbers of deletions and insertions. WXS detected 10 times more deletions than insertions. This may result from two causes: more insertions happen outside exons or insertions are more difficult to call [17]. In cancer projects which show multiple base substitutions, this mutation type was detected in almost all samples by WGS. WXS only detected multiple base substitutions in a small part of samples. To study this mutation type, the WGS strategy should be used with priority.

BRCA-EU (ER+ve, HER2-ve) and BRCA-FR (subtype defined by HER2) contains different number of WGS samples (560 vs. 64), but they have nearly the same spectra of single base substitutions. Their spectra are also very similar to another breast cancer, BRCA-UK (triple negative/lobular/other, 117 samples) in WXS projects. Their spectra are characterised by C > G and C > T at TpCpA and TpCpT. In two other breast cancers, BRCA-KR (Asian phenotype, 33 samples) and BRCA-US (ductal & lobular, 969 samples) of WXS, their spectra are characterised by C > G at TpCpA and TpCpT, and C > T at NpCpG. This shows that European breast cancers exhibit subtle difference from those outside Europe. This may unveil the difference between histologically different subtypes of breast cancer [18]. On the other hand,

the high prevalence of C > G and C > T at TpCpA and TpCpT may widely exist in all breast cancers.

The most interest finding is that no prevalence of single base substitutions at NpCpG found in WGS BRCA-EU and WXS BRCA-UK. This preference was reported by the original publications of the samples in the two cancers [14,19]. Indeed, ours finding is not contradictory to the original one. In their articles, the fraction of 96 spectra of single base substitutions were normalized according to the prevalence of each trinucleotide in the genome. We didn't normalize it that way. The frequencies of the 96 spectra were calculated as probabilities showing how likely 1 of the 96 substitutions can happen in the cancer. Our method not only shows the truth already known but also discloses the differences ignored previously.

It is also interesting to find similar spectra of single base substitutions between breast cancer (BRCA-EU, BRCA-FR, BRCA-KR, and BRCA-US), bladder cancer (BLCA-CN, BLCA-US), and cervical cancer (CESC-US). Breast cancer and cervical cancer are gender related. More than 90% cervical cancers are caused by human papillomavirus and bladder cancer has also found to be associated with viral infection [20,21]. The finding indicates it is worthwhile to study the common mechanisms underlying these cancers.

Different genomic characterisation exists between esophageal adenocarcinomas and esophageal squamous cell carcinomas [4]. This is evident in Figure 2a. Esophageal adenocarcinoma (ESAD-UK) and esophageal squamous carcinoma (ESCA-CN) show different spectra of single base substitutions. T > G transversions at CpTpT dominate single base substitutions in ESAD-UK whereas in ESCA-CN most substitutions are C > G at TpCpA and TpCpT and C > T at NpCpG.

284

285 WGS projects MELA-AU, SKCA-BR, and WXS SKCM-US are three skin cancers. The
286 subtypes of MELA-AU and SKCM-US are melanoma and cutaneous melanoma. SKCA-BR is
287 from skin and the type is melanoma, but the subtype was not given. Though from different
288 countries, the three skin cancers have the same spectra of single base substitutions (Figure 2).
289 Their distinct spectra from other cancers indicate the cancer originates from an absolutely
290 specific aetiology. This may be attributed to the important role of exogenous carcinogens,
291 mainly environmental factors like ultra violet, in the development of skin cancer [22,23].

292

293 In WGS projects, BRCA-EU and PRAD-UK has multiple base substitutions of more than 50
294 bases. No multiple base substitutions exceed 50 bases in WXS cancer projects. Multiple base
295 substitutions were only detected in a few projects. This indicates this type of somatic mutations
296 has not been well called and studied. Though it is hard to say which method is better at this
297 moment, more studies are worthy of carrying out to study multiple base substitutions.

298

299 After excluding ALL-US, the distribution of single base substitutions per Mb of all WGS
300 projects across chromosomes is drawn from the distribution of chromosome lengths (Table S3).
301 In WXS projects, CDMI-UK, LAML-KR, LUSC-KR, NBL-US, and PAEN-AU don't match
302 the distribution of exons lengths across the chromosomes. This trend is also reflected by
303 interlacing lines in Figure 3b. The different distribution of LAML-KR and LUSC-KR may
304 result from high density of single base substitutions per Mb in chromosome Y. COCA-CN also
305 has a larger number of substitutions per Mb in chromosome Y but doesn't affect the distribution.
306 The large number of substitutions of COCA-CN and LUSC-KR in chromosome Y may reflect
307 the high risk of these cancers in men who have bad behaviours such as smoking and drinking

[24,25]. The reason for many substitutions in chromosome Y of LAML-KR is unclear yet but may implicate certain mutation mechanisms of this cancer in men in South Korea.

Conclusions

In this study, the pan-cancer scale landscape of SSM in 10878 samples of 60 ICGC cancer projects was offered. Many previous work towards mutational landscape only considered single base substitutions. Such limitation was overtaken in this study. Indels and multiple base substitutions were analysed, thus providing a more comprehensive landscape of SSM in human cancer. The difference between WGS and WXS strategies in cancer genome studies was compared according to the landscape of mutation at the pan-cancer scale the first time. Through the mutational landscape of SSM, not only knowledge already known was identified, novel mutational spectra in some cancers are revealed as well. Our study will enhance people's understanding of somatic mutations in cancer genome. It can also help to study the mechanisms of somatic mutations underlying the development of cancer.

Acknowledgements

This work is partly supported by National Natural Science Foundation of China (81373311, 31300674), China Scholarship Council (201506240135), and International Macquarie University Research Excellence Scholarships (2015233).

Additional files

Additional file 1: Table S1, includes details of data used in this study, like cancer projects, cancer types, number of simple somatic mutations, number of samples and sequencing strategy. This is an Excel spreadsheet file (12KB)

Additional file 2: Table S2, includes the size of chromosomes for WGS and WXS. This is an Excel spreadsheet file (12KB).

Additional file 3: Table S3, includes the K-S result that shows whether the distribution of single base substitutions per Mb across chromosomes is drawn from the distribution of chromosome size in the genome. This an Excel spreadsheet file (12KB).

Additional file 4: Includes the statistics of simple somatic mutations that created the figures in this article. This is a compressed file and files inside are text files (zip 554 KB).

References

1. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446:153–8.
2. 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*. Nature Publishing Group; 2010;463:191–6.
3. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. *Science* (80-.). American Association for the Advancement of Science; 1991;253:49–54.
4. Gao Y-B, Chen Z-L, Li J-G, Hu X-D, Shi X-J, Sun Z-M, et al. Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet. Nature Research*; 2014;46:1097–102.
5. Wood LD, Parsons DW, Jones SS, Lin J, Sjöblom T, Leary RJ, et al. The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* (80-.). 2007;318:1108–13.
6. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic

357 landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*. Elsevier;
358 2012;150:1121–34.

359 7. Hodis E, Watson IR, Kryukov G V, Arold ST, Imielinski M, Theurillat J-P, et al. A landscape
360 of driver mutations in melanoma. *Cell*. Elsevier; 2012;150:251–63.

361 8. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science* (80-.).
362 American Association for the Advancement of Science; 2015;349:1483–9.

363 9. Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al.
364 Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*
365 2013;499:214–8.

366 10. https://dcc.icgc.org/releases/release_23. Accessed 25 Jan 2017. DCC ICGC.

367 11. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer
368 Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*. Oxford
369 University Press; 2011;2011:bar026.

370 12. <https://genome.ucsc.edu/goldenpath/help/hg19.chrom.sizes>. Accessed 25 Jan 2017. UCSC.

371 13. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene
372 annotation system. *Database*. Oxford University Press; 2016;2016:baw093.

373 14. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al.
374 Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012;149:979–93.

375 15. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria;
376 2016.

377 16. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture
378 and massively parallel sequencing of 12 human exomes. *Nature*. Nature Publishing Group;
379 2009;461:272–6.

380 17. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al.
381 Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* *Nature*

Research; 2015;47:435–44.

18. Li CI, Malone KE, Daling JR. Differences in breast cancer hormone receptor status and histology by race and ethnicity among women 50 years of age and older. *Cancer Epidemiol. Prev. Biomarkers.* AACR; 2002;11:601–7.

19. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature.* 2016;1–20.

20. Walboomers JMM, Jacobs M V, Manos MM, Bosch FX, Kummer JA, Shah K V, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* 1999;189:12–9.

21. Cancer T, Atlas G, Weinstein JN, Akbani R, Broom BM, Wang W, et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature.* 2014;507:315–22.

22. Garibyan L, Fisher DE. How sunlight causes melanoma. *Curr. Oncol. Rep. Springer;* 2010;12:319–26.

23. Narayanan DL, Saladi RN, Fox JL. Review: Ultraviolet radiation and skin cancer. *Int. J. Dermatol. Wiley Online Library;* 2010;49:978–86.

24. Chen W, Zheng R, Zeng H, Zhang S, He J. Annual report on status of cancer in China, 2011. *Chinese J. cancer Res.* 2015;27:2–12.

25. Oh C-M, Won Y-J, Jung K-W, Kong H-J, Cho H, Lee J-K, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2013. *Cancer Res. Treat. Off. J. Korean Cancer Assoc. Korean Cancer Association;* 2016;48:436.