RESEARCH

Inferring tumor age from multiple neutral evolutionary process

Susana Londono-Munoz[1] and Sergio Pulido-Tamayo[2*†]

Abstract

Background: All cancer arise as a result of a somatic mutation, several models have been developed to understand cancer, it has been seen as an evolutionary process in which somatic mutations are the pieces of the puzzle. Somatic mutation catalogue of the cancer genome has been used to reconstruct cancer phylogeny, in this case somatic mutations without selective pressure will help determine the age of a tumor.

Results: Any type of mutation can have an impact on the cell, the majority of them are neutral, mutations that have less functional impact, that means less selective pressure on the tumor evolution are the ones used as temporal measurement of the tumors. A positive correlation was found between two types of measurements, which shows that an appropriate age can be extracted using both of them.

Conclusion: The use of this method can be coupled with other cancer analysis, for example, along with co-occurrence and mutual exclusivity analysis could bring a better understanding of the tumor, an appropriate tumor classification and with this a more accurate clinical decision, which leads to an efective treatment.

Keywords: Cancer; Tumor age; Tumor evolution; Somatic mutations; Passenger mutations

Background

Every cell in the body is a direct descendant of the fertilized egg from which each of us developed. Through time the DNA sequence of every cell acquires a set of differences from its progenitor, this variations are called somatic mutations [1, 2]. All cancer arise as a result of a somatic mutation, so an approach to unveil the mysteries of cancer is through the catalogue of mutations a cancer genome has [1, 3].

Somatic mutations have been classified depending on certain characteristics, such as location, mutations can occur all over the genome, that can be a simple way to determine if they will affect the protein structure and a basic approach to determine the functional impact the variation will have [4, 5, 6]. In cancer is important to classify mutations as the ones that have been positively selected or "driver" mutations, which confer an advantage to the cell and "passenger" mutations, this type haven't been selected, they just happened to be there [1, 3].

To understand cancer, several models have been developed, it has been seen as an evolutionary process, it is a series events, constant acquisition of mutations and natural selection, these processes mark the genome, and a cancer's life history is encrypted in the somatic mutations that we can find in its genome [7, 8, 2].

Knowing the history of a tumor is important, because as in evolution, it lets us understand the subpopulations of cells present in a tumor and the biological processes, and can be used in a clinical decision [9]. In fact the catalogue of mutations have been used to tell the history of tumors, to reconstruct a phylogeny of the diverse clone subpopulations, which its being recognized to have value making clinical decisions [10, 11].

An important thing in history is time, knowing how long have been the tumor evolving can help in diagnosis, tumor classification, prognosis and treatment. The approach described to determine the age of a tumor based on somatic mutations with low or neutral selective pressure can help to understand mutational process like co-occurrence and mutual exclusivity, that can lead to new functional interactions, which are important not only for understanding cancer, but also in clinical decisions, it can help in the selection of multi targeted anti-tumor therapies, co-mutations suggest combination of drugs might be effective while mutual exclusion indicate combinations likely won't work, this is specially important because treatments focusing in

[*] Correspondence: spulido99@gmail.com
[2] Universidad Eafit, Research group on biodiversity, ecology and evolution, Medellín, Colombia
Full list of author information is available at the end of the article
[†] Thesis assessor

a unique alteration can lead to a single cell that is resistant to therapy proliferates, followed by relapse, increasing the mortality [1, 12, 13].

Studies show that the accumulation of mutations can be accelerated due to genome instability, increased sensitivity to mutagenic agents and breakdown of genomic maintenance [14, 15], according to this we can make the assumption that the mutations present before the tumor started to evolve are negligible in number.

The majority of molecular changes are caused by random fixation of selectively neutral mutants [16], correspondingly in cancer the majority of mutations do not confer advantage, it means they are not detectable under selection [17, 18]. Knowing this we can assume passenger mutations are useful to determine the age of a tumor because most mutations in the cancer genome are of this kind.

As the evolutive history, the age of a tumor can be inferred from the catalogue of mutations, in this paper we will present a method to determine the age of a tumor in terms of the somatic mutations with low or no selection pressure present in the cancer genome.

Methods

Data

All data used in this paper is available on line. The dataset used in this paper is breast invasive carcinoma mutation assessor analysis results [19]. Mutation assessor is a server that predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The functional impact is assessed based on evolutionary conservation of the affected amino acid in protein homologs.

Data preprocess

Data was subjected to an analysis under the scope of the project objectives to select the variables that could be used to determine the age of the tumor. From 356 variables, nine were selected. The variables selected helped us know different characteristics of the mutation including: location, classification and functional impact [6]. The data was analyzed and the outliers were identified by box and whiskers plots, and appropriately managed with the interquartile range (IQR) method, those samples that where below quantile 25 minus 1.5∗IQR or above quantile 75 plus 1.5∗IQR were not taken into account, leaving 800 samples.

Data process

Various measurements based on the count of passenger mutations were performed. Variant classification was the selection criteria to determine the mutations as passenger, the measurements consisted on counting the occurrence of a certain type of mutation by sample. Based on the location of the mutation and the functional impact score (FIS) determined by [6], we took the ones that might have low or neutral functional impact, from the classifications presented in **table 1**. Silent, intergenic region (IGR), intron mutations (int) and missense mutations with less than 1.9 functional impact score (FIS) [6], were selected as the ones that could have least impact on the cell function and have less selective pressure on the tumor evolution, and so the ones to be used as temporal measurement of the tumor.

Variant Classification	Predicted Impact
Nonsense	Non-synonymous medium - high[20]
Missense	Non-synonymous neutral - low - high [6]
Silent	Synonymous neutral - low
Nonstop	Non-synonymous medium - high [21]
In frame indels	Non-synonymous medium - high [21]
Frame shift indels	Non-synonymous medium - high [21]
Start codon indels	Affects translation medium-high
Stop codon indels	Affects translation medium-high
Splice site	Affects translation medium - high
Intergenic regions	Non-coding neutral - low [1]
Intron	Non-coding neutral - low [1]
De Novo start out of frame	Non-synonymous medium-high
De Novo start in frame	Non-synonymous medium-high
5' Flank	Non-coding, gene expression medium-high
3' and 5' Untranscribed regions	Non-coding, gene expression medium-high

Table 1 Variant classifications present in the data with its predicted impact on protein function.

Results and discussion

Although all mutations can have an impact on the cell, the majority of them are neutral [22], so we can assume that those mutations that do not change amino acid in the protein sequence, that are in non coding regions and/or far from known gene control regions, have mostly low or neutral impact. Looking to the types of mutation and its characteristics on **table 1**.

Silent, intergenic region (IGR), intron mutations (int) and missense mutations with less than 1.9 functional impact score (FIS) [6], were selected as the ones that could have least impact on the cell function and have less selective pressure on the tumor evolution, and so the ones to be used as temporal measurement of the tumor.

The measurements to determine the tumor age where used in a breast cancer mutation assessor analysis dataset [19]. For each sample, which has a unique identification total, silent, intergenic region and intron mutations, as well as mutations with low or neutral functional impact (FIS < 1.9) [6] were counted with a python algorithm developed for this purpose.

	Total	IGR	Int	Sil	FI	Sum
Mean	40.35	1.24	1.27	8.7	13.47	11.2
STD	24.05	0.53	0.58	5.6	8.6	5.7
Min	1	0	0	1	2	3
25%	23	1	1	5	7	7
50%	33	1	1	7	11	10
75%	53	1	1	12	18	14
Max	125	3	3	27	44	30

Table 2 General description of the data. Intergenic Regions (IGR), Intron (Int), Silent (Sil), summation of intergenic region, intron and silent (Sum) and low and neutral functional impact (FI) per sample.

General statistics of the measurements, see **table 2**, showed that IGR and intron mutations counts were too low (75% of the samples has 1 mutation), with this counts we don't have enough information to be used as an appropriate temporal measurement by itself, for this reason, the summation of intergenic, intron and silent mutation was taken as a single measurement, leaving us with two different measurements to assess, low FIS (functional impact score below 1.9), sum (summation of IGR, intron and silent mutations).

A simple linear regression was performed between measurements, a positive correlation was found for both regressions. The correlation between silent and low FIS didn't show a lot of difference with the correlation between the other two measurements, because of the similarity, summation of IGR, intron and silent, which includes all the mutations with neutral or low evolutionary impact was used. It's worth to note that the mutations counted for each measurements are only used for that measurement, the correlation between sum and low FIS showed an R value of 0.74, see **figure 1**, from which we can infer that two different types of mutation with the same predicted low or neutral impact on the cell are occurring similarly. A positive correlation was also found between each measurement and the total mutations,see **figure 2** (sum vs total: slope=0.2, R=0.85 and low FIS: slope=0.3, R=0.87) allowing us to use them as an approach to determine the tumor age in terms of neutral and low impact mutations.

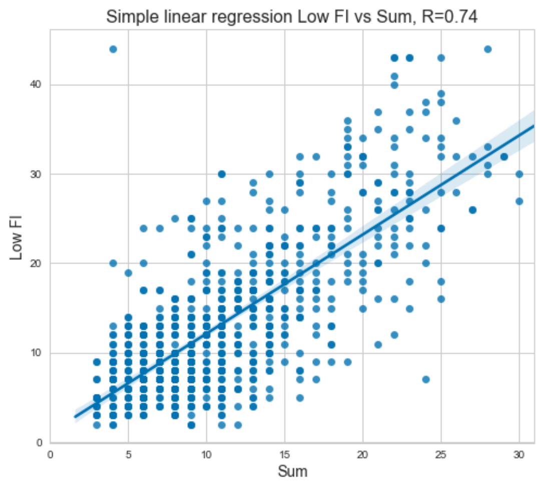

Figure 1 Sum vs Low FIS. Linear regression between measurements, sum and low FI, R=0.74

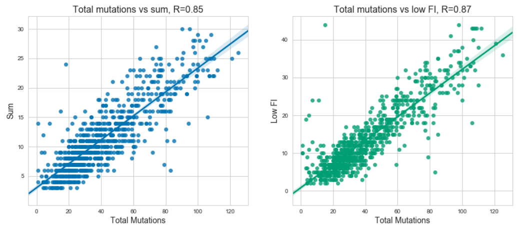

Figure 2 Total mutations linear regressions. Linear regression between each measurement and total mutations for sum R=0.85 and for low FI R=0.87

The tumor with the minimum for each measurement is classified as the youngest, for this dataset it was found that a sample has three for each measurement, which makes it the newest tumor. Difference between the measurements can be found, in some cases it doesn't let us determine the tumor age (difference greater than 12), the oldest tumor has 36 for low FIS and 26 for sum, for the complete list of tumor ages see supplementary data or go to the study repository available on https://github.com/SusanaLondono/TumorAge.

Conclusion

Gigantic global efforts have lead us to understand a little better how cancer develop and how can we fight against it. From this hundreds of gigabytes of data have been put available on-line. Understanding cancer as an evolutionary process is not new [23], but the availability of data is recent and it has to be exploited.

Efforts to uncover the evolutionary history of tumors from the catalogue of mutations, mainly in driver mutations, have been done. This study presents a simple approach to determine the age of a tumor for breast cancer, but it doesn't mean that the method can't be used in other datasets of different types of cancer.

The use of this method can be coupled with other cancer analysis, for example, along with co-occurrence and mutual exclusivity analysis could bring a better understanding of the tumor, an appropriate tumor classification and with this a more accurate clinical decision, which leads to an effective treatment.

Declarations

Availability of data and material

The datasets generated and/or analyzed during the current study are available in the Firehose Broad GDAC repository: https://gdac.broadinstitute.org.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

This article is an undergraduate thesis, SL-M is the student and SP-T is the thesis assessor.

Author details

[1] Universidad Eafit, School of Sciences, Department of biology, Medellín. Colombia. [2]Universidad Eafit, Research group on biodiversity, ecology and evolution, Medellín, Colombia.

References

1. Stratton, M., Campell, P., Futreal, P.: The canger genome. Nature **485**, 719–724 (2009)
2. Yates, L., Campbell, P.: Evolution of the cancer genome. Nature Reviews Genetics **13**, 795–806 (2012)
3. Nik-Zainal, S., Alexandrov, L., Wedge, D., Van Loo, P., Greenman, C., Raine, K., et al: Mutational processes molding the genome of 21 breast cancers. Cell **149**, 979–993 (2012)
4. Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D., Darnell, J.: Mutations: Types and causes. In: Tenney, S. (ed.) Molecular Cell Biology, 4th edn., pp. 53–76. W.H. Freeman and Company, New York (2000)
5. Tate, J., Bamford, S., Jubb, H., Sondka, Z., Beare, D., Bindal, N., et al: Cosmic: the catalogue of somatic mutations in cancer. Nucleic Acids Research **47**, 941–947 (2019)
6. Reva, B., Antiipin, Y., Sander, C.: Predicting the functional impact of mutations: application to cancer genomics. Nucleic Acids Research **39** (2011)
7. Lean, C., Plutynski, A.: The evolution of failure: explaining cancer as an evolutionary process. Biology and philosophy **31**, 39–57 (2015)
8. Nik-Zainal, S., Van Loo, P., Wedge, D., Alexandrov, L., Greenman, C., Wai Lau, K., et al: The life history of 21 breast cancers. Cell **149**, 994–1007 (2012)
9. Fisher, R., Pusztai, L., Swanton, C.: Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer **108**, 479–485 (2013)
10. Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R., Batzoglou, S.: Fast and scalable inference of multi-sample cancer lineages. Genome Biology **16** (2015)
11. Ricketts, C., Popic, V., Toosi, H., Hajirasouliha, I.: Using lichee and bamse for reconstructing cancer phylogenetic trees. Current protocols in bioinformatics, 49 (2018). doi: 10.1002/cpbi.49
12. Ochoa, S., Martinez-Pérez, E., Zea, D., Molina-Vila, M., Marino-Buslje, C.: Co-mutation and exclusion analysis in human tumors, a means for cancer biology studies and treatment design. Human mutation **40**, 413–425 (2019)
13. Tinahai, T., Olson, S., Whitacre, J., Harding, A.: The origins of cancer robustness and evolvability. Integr. Biol. **3**, 17–30 (2011). doi:10.1039/c0ib00046a
14. Hanahan, D., Weinberg, R.: Halmarks of cancer: the next generation. Cell **144**, 646–674 (2011)
15. Negrini, S., Gorgoulis, V., Halazonetis, T.: Genomic inestability an evolving hallmark in cancer. Nature Reviews Molecular Cell Biology **11**, 220–228 (2010)
16. Kimura, M.: The neutral theory of molecular evolution: a review of recent evidence. The japanese journal of genetics **66**, 367–386 (1991)
17. Cannataro, V., Townsed, J.: Neutral theory and the somatic evolution of cancer. Molecular biology and evolution **35**, 1308–1315 (2018)
18. Piraino, S., Furney, S.: Beyond the exome: the role of non-coding somatic mutations in cancer. Annals of oncology **27**, 240–248 (2016)
19. Institute, B., Center, T.G.D.A.: Mutation assessor. Broad Institute of MIT and Harvard (2016). doi:10.7908/C1F18Z2Z
20. Choi, Y., Sims, G., Murphy, S., Miller, J., Chan, A.: Predicting the functional effect of amino acid substitutions and indels. PLoS ONE, 46688 (2012). doi:10.1371/journal.pone.0046688
21. Azia, A., Uversky, V., Horovitz, A., Unger, R.: The effects of mutations on protein function: a comparative study of three databases of mutations in humans. Israel journal of chemistry **53**, 217–226 (2013)
22. Liu, M., Watson, L., Zhang, L.: Classification of mutations by functional impact type: gain of function, loss of function, and switch of function. In: Basu, M., Pan, Y., Wang, J. (eds.) Bioinformatics Research and Applications, pp. 236–242 (2014). Springer, Cham
23. Cairns, J.: Mutation selection and the natural history of cancer. Nature **255**, 197–200 (1975)