Synonymous Somatic Variants in Human Cancer Are Not Infamous: A Plea for Full Disclosure in Databases and Publications

Thierry Soussi, Peter E.M. Taschner, Yardena Samuels

Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications.

Thierry Soussi [1,2,3,4]

Peter E.M. Taschner,[5]

Yardena Samuels, [6]

[1]Sorbonne Université, UPMC Univ Paris 06, F- 75005 Paris, France

[2]INSERM, U1138, Centre de Recherche des Cordeliers, Paris, France

[3]Department of Oncology-Pathology, Karolinska Institutet, Cancer Center Karolinska (CCK) R8:04, Stockholm SE-171 76, Sweden

[5]Generade Centre of Expertise Genomics and University of Applied Sciences Leiden, Zernikedreef 10, Room Z1.026, 2333 CL Leiden, The Netherlands

[6]Molecular Cell Biology Department, Weizmann Institute of Science, Rehovot, 76100, Israel

4: Correspondance: thierry.soussi@ki.se

Keywords: Synonymous variants; database; cancer; pathogenicity prediction

Grant Sponsor: Radiumhemmets Forskningsfonder and the Swedish Cancer Society (Cancerfonden) to TS.

Abstract

Single Nucleotide Variants (SNVs) are the most frequent genetic changes found in human cancer. Most driver alterations are missense and nonsense variants localized in the coding region of cancer genes. Unbiased cancer genome sequencing shows that synonymous SNVs (sSNVs) can be found clustered in the coding regions of several cancer oncogenes or tumor suppressor genes suggesting purifying selection. sSNVs are currently underestimated, as they are usually discarded during analysis. Furthermore, several public databases do not display sSNVs, which can lead to analytical bias and the false assumption that this mutational event is uncommon. Recent progress in our understanding of the deleterious consequences of these sSNVs for RNA stability and protein translation shows that they can act as strong drivers of cancer, as demonstrated for several cancer genes such as *TP53* or *BCL2L12*. It is therefore essential that sSNVs be properly reported and analyzed in order to provide an accurate picture of the genetic landscape of the cancer genome.

Keywords: *TP53*; database; synonymous variants; cancer

Letter

The Next Generation Sequencing (NGS) revolution has opened up a new age in cancer genomics. More than 20,000 cancer genomes/exomes have been sequenced by three large consortia: The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/), a project of the National Cancer Institute and the National Human Genome Research Institute, the UK Cancer Genome Project from the Sanger Institute (http://www.sanger.ac.uk/) and the International Cancer Genome Consortium (ICGC, http://icgc.org/) (Joly et al., 2012, McDermott et al., 2011, Wang et al., 2016). As a result of the decreased costs of new sequencers, data on multiple cancer genomes are also released by individual teams, but, in most cases, only the sequencing of a specific subset of cancer genes panel is performed.

The majority of the genomic data are now available via large-scale repositories and are accessible to the scientific community via web portals such as the ICIG portal (https://dcc.icgc.org/), the cBioPortal for Cancer Genomics (http://www.cbioportal.org/index.do) and COSMIC (http://cancer.sanger.ac.uk/cosmic). Data from COSMIC or the cBioPortal are widely used for multiple studies and it is therefore essential that data from these various repositories are fully described.

Cancer genomes are a complex mixture of somatic genetic and epigenetic modifications, a minority of which are driving alterations (4 to 8) targeting cancer genes and contributing to cancer progression, and several thousand passenger variants that confer no selective growth advantage and that are not subject to positive selection (Chanock and Thomas, 2007).

The protein-centric view of genetic variation prevailing at the end of the twentieth century and the beginning of the twentieth first century has led to a huge bias in data analysis and

interpretation. Only coding regions were screened for variants and exclusively missense or nonsense variants as well as indels were reported, while synonymous variants (sSNV[1]), sometime misnamed neutral mutations, silent mutations or polymorphisms, were neglected. This approach completely ignored the fact that all types of exonic variants can alter splicing (Pagani and Baralle, 2004).

We know now that sSNVs can have multiple consequences for RNA maturation and stability as well as protein translation (Figure 1) (Gartner et al., 2013, Gotea et al., 2015, Holmila et al., 2003, Raponi and Baralle, 2010, Sauna and Kimchi-Sarfaty, 2011, Supek et al., 2014). In addition, tissue-specific and tumor-specific changes in tRNA expression combined with asymmetric tRNA abundance may play a role (Czech et al., 2010). It is therefore essential that these variants are correctly reported and collected. In the absence of a functional assay and in view of the difficulty of assessing their consequences at the RNA level, the discovery of an sSNV as a recurrent event associated with a specific disease will be essential to evaluate. Indeed, it is only once sSNVs are reported in a comprehensive manner, that statistical and bioinformatic tools, such as the ones used for nsSNVs can be applied and developed. The outcome of these analyses will allow investigation of their potential pathogenic role, i.e. disease-causing when other conditions for disease development have been met.

We have noticed significant differences in the descriptions of sSNVs in publications and the various databases. Their absence in publications despite the prevalence of Sanger sequencing is often associated with self-censorship due to the general belief that sSNVs are neutral. The current use of a wide range of stringent filtering processes in NGS can also lead to underestimation of these variants, as they are automatically removed. As

[1] In this article, the term "synonymous variant" (sSNV) refers to a single nucleotide modification localized in the coding region of any gene that does not change the predicted amino acid sequence of the protein, in contrast with nonsynonymous variants (nsSNV) that change the codon significance.

discussed below, the situation is more chaotic in large repositories, resulting in heterogeneous reporting that can be harmful for the scientific community.

Using data from the *TP53* (MIM# 191170) variant database illustrates this problem. The latest issue of the *TP53* gene variant database includes variant data from more than 68,000 patients that have been collected since 1989 (Leroy et al., 2014a). It includes both somatic variants detected in various types of cancer as well as germline variants associated with familial cancer such as Li-Fraumeni syndrome. This database includes 1,982 patients with sSNVs, some of which were detected as both somatic and germline variants. Several of these sSNVs, such as NM_000546.5:c.375G>A (NP_000537.3:p.(Thr125=)) and NM_000546.5:c.672G>A (NP_000537.3:p.(Glu224=)), are last-base exon (LBE) variants that have been shown to impair *TP53* splicing, making them truly pathogenic (Supek et al., 2014, Varley et al., 1998b). Other recurrent sSNVs have been identified, but their pathogenicity remains uncertain.

The cBioPortal developed by the TCGA consortium is one of the most popular portals, comprising the sequences of more than 10,000 tumor genomes including other genomic data such as RNA expression or DNA methylation. However, examination of data concerning *TP53*, as well as other cancer genes, displayed in this database reveals a complete lack of synonymous variants. This is not surprising, as it results from the deliberate choice of the curators of this database not to include sSNVs. Extraction of *TP53* variants from the TCGA raw data available via the synergy web site shows that the classical synonymous variants NM_000546.5:c.375G>A (NP_000537.3: p.(Thr125=)) and NM_000546.5:c.672G>A (NP_000537.3:p.(Glu224=)), as well as other sSNVs, can be identified in all types of cancer, but this information is not reported in cBioPortal. Data from the TCGA consortium are also available via the COSMIC web site and include sSNVs.

Two consortia have sequenced the entire genome of the most common cell lines used in cancer research available from the COSMIC or Cancer Cell Line Encyclopedia (CCLE) portals (Barretina et al., 2012, Ikediobi et al., 2006). Recent studies have shown that major discrepancies can be observed between the two repositories (Hudson et al., 2014, Leroy et al., 2014b). Using the *TP53* status as a criterion for comparison, we observed that 40% of the cell lines have a different *TP53* status (Leroy et al., 2014b). Among the multiple reasons for these variations is the lack of synonymous variants in the CCLE database. Twelve common cell lines containing the deleterious NM_000546.5:c.375G>A (NP_000537.3:p.(Thr=)) variants are correctly labelled in the COSMIC database, but this information is missing from the CCLE. Underestimating this information can be highly misleading, as *TP53* status is an important feature for drug screening.

This situation is not specific for *TP53*, but also concerns other genes with sSNVs. A recent analysis in melanoma revealed an exonic sSNV in the *BLC2L12* gene (NM_138639.1:c. 51C>T, p.Phe7=) that removes the *hsa-miR-671–5p* binding site, leading to increased expression of the protein product that can impair the apoptotic response to TP53 (Gartner et al., 2013). Although this variant is fully described in the COSMIC portal (11 entries from various publications), it is absent from the cBioPortal despite the fact that both portals include data from the same studies.

Beyond the problems associated with heterogeneous variant datasets, neglecting sSNV variants can result in a number of problems.

First, omission of these variants from popular databases could lead to false interpretations with sSNVs mistaken for non-pathogenic variants. As described above, several variants in the *TP53* gene were initially considered to be "neutral SNP" until they were formally

demonstrated to be pathogenic. This can be a critical problem when dealing with germline variants.

Second, omission of these variants could delay the discovery of novel pathogenic mechanisms associated with these sSNVs. Translational research must work in both directions: when clinical practice raises new questions, those need to be resolved by bench work to avoid the frustration of clinicians with variants of unknown significance. Recent studies have detected a second code in the mammalian genome, which is independent of the classic genetic code (Weatheritt and Babu, 2013). Transcription factors can bind within protein-coding regions and this “binding code” may be linked to the various constraints that shape the choice of codons as well as protein evolution. These regions that encode two types of information have been named duons with i) extrinsic information leading to protein synthesis via deciphering of the genetic code of the intermediate mRNA molecule; and ii) intrinsic information acting as a recruitment platform for transcription factors (Stergachis et al., 2013). The mutual or exclusive consequences of sSNVs (as well as nsSNVs) on both codes is currently unknown, but accurate and exhaustive variant databases compared to transcription factor occupancy maps should provide major clues.

It is essential for sSNVs to be reported in publications and databases. The unbiased reporting of genetic alterations in different cancer types provides an extremely powerful data source, that reveals which alterations are being selected for by the tumor. There could be no better starting point for our investigation as to which genetic alterations have a role to play in the disease. Indeed, such databases have allowed the surprising identification and functional evaluation of nsSNV driver variants that occur even at low frequencies (D'Antonio and Ciccarelli, 2013, de Voer et al., 2016, Devarakonda et al., 2013), emphasizing their importance. Such discoveries are expected to increase if more sSNVs are included in publications and databases. In principle, it is feasible to mine raw

NGS data available from public repositories for sSNVs provided that the bioinformatics pipeline does not filter them out. Indeed, in an era where multiple genomic and proteomic platforms are being used in parallel to investigate exactly the same samples, inclusion of sSNVs will allow integrative analysis with additional omic outputs giving an indication as to their possible effects. For example, comprehensive RNAseq and miRNA analyses in parallel to sSNV analyses may immediately indicate as to whether a recurrent sSNV affects binding to its location by a particular miRNA and therefore affects mRNA expression.

It may be difficult to remedy the omission of previous synonymous mutation data, but it should be made increasingly standard practice going forward. Thus, although the filtering process can easily remove very frequent sSNVs that are known to have no disease-causing effect, any novel sSNV detected with a significant frequency in a tumor clone should be considered to be potentially pathogenic until further studies confirm or exclude a pathogenic role. Systematic reporting of sSNVs will be essential to achieve positive progress in our understanding of the full spectrum of functional effects associated with genomic variants.

References

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603-607.

Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z. 2010. A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. J Biol Chem 285: 28741-28748.

Boichard A, Venet L, Naas T, Boutron A, Chevret L, de Baulny HO, De Lonlay P, Legrand A, Nordman P, Brivet M. 2008. Two silent substitutions in the PDHA1 gene cause exon 5 skipping by disruption of a putative exonic splicing enhancer. Mol Genet Metab 93: 323-330.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hébuterne X, Harel-Bellan A, Mograbi B et al. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent

xenophagy in Crohn's disease. Nat Genet 43: 242-245.
Chanock SJ, Thomas G. 2007. The devil is in the DNA. Nat Genet 39: 283-284.
Czech A, Fedyunin I, Zhang G, Ignatova Z. 2010. Silent mutations in sight: co-variations in tRNA abundance as a key to unravel consequences of silent mutations. Mol Biosyst 6: 1767-1772.
D'Antonio M, Ciccarelli FD. 2013. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol 14: R52.
de Voer RM, Hahn MM, Weren RD, Mensenkamp AR, Gilissen C, van Zelst-Stams WA, Spruijt L, Kets CM, Zhang J, Venselaar H, Vreede L, Schubert N et al. 2016. Identification of Novel Candidate Genes for Early-Onset Colorectal Cancer Susceptibility. PLoS Genet 12: e1005880.
Devarakonda S, Morgensztern D, Govindan R. 2013. Clinical applications of The Cancer Genome Atlas project (TCGA) for squamous cell lung carcinoma. Oncology (Williston Park) 27: 899-906.
Friedrich U, Datta S, Schubert T, Plössl K, Schneider M, Grassmann F, Fuchshofer R, Tiefenbach KJ, Längst G, Weber BH. 2015. Synonymous variants in HTRA1 implicated in AMD susceptibility impair its capacity to regulate TGF-β signaling. Hum Mol Genet 24: 6361-6373.
Gartner JJ, Parker SC, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, Gotea V, Teer JK et al. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci U S A 110: 13481-13486.
Gotea V, Gartner JJ, Qutob N, Elnitski L, Samuels Y. 2015. The functional relevance of somatic synonymous mutations in melanoma and other cancers. Pigment Cell Melanoma Res 28: 673-684.
Griseri P, Bourcier C, Hieblot C, Essafi-Benkhadir K, Chamorey E, Touriol C, Pagès G. 2011. A synonymous polymorphism of the Tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to Herceptin treatment in breast cancer patients. Hum Mol Genet 20: 4556-4568.
Holmila R, Fouquet C, Cadranel J, Zalcman G, Soussi T. 2003. Splice mutations in the p53 gene: case report and review of the literature. Hum Mutat 21: 101-102.
Hudson AM, Yates T, Li Y, Trotter EW, Fawdar S, Chapman P, Lorigan P, Biankin A, Miller CJ, Brognard J. 2014. Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. Cancer Res 74: 6390-6396.
Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, Santarius T, Avis T, Barthorpe S, Brackenbury L, Buck G, Butler A et al. 2006. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther 5: 2606-2612.
Joly Y, Dove ES, Knoppers BM, Bobrow M, Chalmers D. 2012. Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). PLoS Comput Biol 8: e1002549.
Jung H, Lee D, Lee J, Park D, Kim YJ, Park WY, Hong D, Park PJ, Lee E. 2015. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat Genet 47: 1242-1248.
Lazrak A, Fu L, Bali V, Bartoszewski R, Rab A, Havasi V, Keiles S, Kappes J, Kumar R, Lefkowitz E, Sorscher EJ, Matalon S et al. 2013. The silent codon change I507-ATC->ATT contributes to the severity of the ΔF508 CFTR channel dysfunction. FASEB J 27: 4630-4645.
Leroy B, Anderson M, Soussi T. 2014a. TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum Mutat 35: 672-688.
Leroy B, Girard L, Hollestelle A, Minna JD, Gazdar AF, Soussi T. 2014b. Analysis of TP53 mutation status in human cancer cell lines: a reassessment. Hum Mutat 35: 756-765.
McDermott U, Downing JR, Stratton MR. 2011. Genomics and the continuum of cancer care. N Engl J Med 364: 340-350.
Montera M, Piaggio F, Marchese C, Gismondi V, Stella A, Resta N, Varesco L, Guanti G, Mareni C. 2001. A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. J Med Genet 38: 863-867.
Pagani F, Baralle FE. 2004. Genomic variants in exons and introns: identifying the splicing

spoilers. Nat Rev Genet 5: 389-396.
Raponi M, Baralle D. 2010. Alternative splicing: good and bad effects of translationally silent substitutions. FEBS J 277: 836-840.
Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12: 683-691.
Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, Stamatoyannopoulos JA. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342: 1367-1372.
Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. Cell 156: 1324-1335.
Varley JM, Chapman P, McGown G, Thorncroft M, White GR, Greaves MJ, Scott D, Spreadborough A, Tricker KJ, Birch JM, Evans DG, Reddel R et al. 1998a. Genetic and functional studies of a germline TP53 splicing mutation in a Li-Fraumeni-like family. Oncogene 16: 3291-3298.
Varley JM, McGown G, Thorncroft M, White GR, Tricker KJ, Kelsey AM, Birch JM, Evans DG. 1998b. A novel TP53 splicing mutation in a Li-Fraumeni syndrome family: a patient with Wilms' tumour is not a mutation carrier. Br J Cancer 78: 1081-1083.
Wang Z, Jensen MA, Zenklusen JC. 2016. A Practical Guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol 1418: 111-141.
Weatheritt RJ, Babu MM. 2013. Evolution. The hidden codes that shape protein evolution. Science 342: 1325-1326.

Figure 1: Multiple consequences of sSNV on RNA and protein. See Table 1 for more information. ESE: Exon splicing Enhancer; ESS: Exon splicing Silencer; LBE: Last-base exon