6. Identification of oncogenic driver mutations

David Tamborero, Abel Gonzalez-Perez, Nuria Lopez-Bigas

The boost of sequencing technology applied to tumor genomes profiling has revealed cancer as a disease as heterogeneous at a molecular level as it is clinically. The identification of the events responsible for the phenotypes that the tumor cell acquires at the different stages of the disease is required for any downstream analysis aimed to better understand the process and develop novel therapeutical interventions. The present review covers the description of the computational methods able to analyse the vast amount of data generated by the current re-sequencing projects and identify mutations, genes and pathways driving the tumorigenesis.

Introduction

Second-generation sequencing technologies opened the door to the comprehensive characterization of the somatic alterations in the genome of tumor cells, thus transforming the study of cancer [1]. Cancer samples, even from the same tumor type, have been revealed as highly heterogeneous in their somatic abnormalities [2]. The rate and pattern of somatic mutations are shaped by factors like the number of replication rounds accumulated by cells, the presence of defects in the DNA maintenance mechanisms and the exposure to environmental insults [3]. Only a small subset of the somatic mutations found in cancer cells are responsible of the tumorigenesis, whereas the remaining are sporadic events secondary to the genomic instability caused by the cancer [4]. The need to distinguish the former –driver mutations– from the latter –passengers– is one of the most important tasks in cancer biology research as well as in the development of novel therapeutic interventions.

Despite big advances in recent years, the interpretation of the vast amount of data provided by tumor genomes re-sequencing is still challenging in many ways. Here, we review the state-of-the-art of methods aimed to identify drivers of cancer and discuss some of their main results. Bear in mind that in this article we cover only somatic singe nucleotide variants and short frameshifts. In addition, we focus only on strategies that evaluate the effect of mutations in the exome; the role of alterations in non-coding regions is still a more immature field that will certainly explode in coming years thanks to the decreasing costs of whole genome

[Keywords]
cancer drivers, somatic mutations, computational methods

David Tamborero[1] /Abel Gonzalez-Perez[1] /Nuria Lopez-Bigas[1] [2]
Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra[1] /Institució Catalana de Recerca i Estudis Avançats (ICREA)[2]

sequencing.

1 Identifying the impact of cancer mutations

The first step to assess the role of a somatic mutation detected by tumor genome re-sequencing is to annotate it in the context of genomic elements. For coding sequences, this means to identify the protein-coding gene overlapping the mutation and assess its consequence. Mutations that truncate the protein product (stop gained or frameshift mutations) probably result in inactivation of the protein; in comparison, synonymous mutations are much milder to protein function. Between these two extremes, non-synonymous mutations are the subject of computational methods to assess their impact on protein function (**Table 1**) [5]. Tools initially designed to be used on germ-line mutations, mutations, such as SIFT [6], Polyphen-2 [7] and Mutation Assessor [8], use different metrics of amino acid conservation to infer the extent of protein impairment. Condel integrates the output of these tools into a consensus deleteriousness score, based on a weighted average of the normalized original scores, which outperform them [9]. Recently, some bioinformatics tools have been specifically developed to rank tumor somatic mutations. FATHMM distinguishes driver mutations through sequence conservation data within hidden Markov models weighted with cancer data [10]. CHASM uses a Random Forest algorithm trained with a set of driver and synthetically-created passenger mutations set, which using several features on top of conservation, such as specific amino acid substitutions and predicted protein structure change properties [11]. Finally, transFIC takes into account the differences in baseline tolerance to functional variants between groups of genes to refine the score of some of the aforementioned methods [12]. The analysis of mutations with other consequence types, such as splice-site mutations requires other specific tools [5].

2 Identifying signals of positive selection

A driver mutation confers selective growth advantage to a cell and is thus selected during the clonal evolution of the tumor, whereas a passenger mutation is propagated as a bystander. Genes involved in tumorigenesis will therefore exhibit signals of positive selection across a cohort of tumor samples. Several signals of positive selection have been exploited to identify putative driver genes (**Figure 1, Table 1**). The most intuitive one consists in evaluating whether a gene is mutated more frequently than one would expect given the background mutation rate. This has been implemented in methods such as MuSiC, developed by the Washington University [13], and MutSig, developed by the Broad Institute [14]. The estimation of the background mutation rate takes into account features such as gene length, the type of mutation and the nucleotide context, which shapes the baseline probability of the somatic mutation to occur within each particular genomic position. Other factors that influence the mutation rate -as the replication time of the DNA region and the gene expression level- have been incorporated as covariates in the statistical framework of the latest implementation of MutSig -termed MutSigCV- to increase its accuracy [14]. Although these methods have been demonstrated to be successful in detecting those genes that are more frequently mutated in cancer, this approach rarely detects lowly recurrent drivers, which are crucial to understand the whole picture of tumorigenesis, since further precision in the estimation of the background mutation rate and/or larger sample sizes are required for this.

A second approach that does not depend on the mutation burden relies on the evaluation of the functional impact of the mutations of each gene across the samples cohort. Driver mutations in coding genes must impact the function of the encoded protein, as opposed to passengers that are randomly distributed. Therefore, the identification of genes bearing mutations biased towards higher functional impact is an indication of positive selection and can be used to distinguish driver genes. Because this approach does not rely on the estimation of the background mutation rate, it is suited to detect drivers regardless of their mutation frequency. This is implemented by OncodriveFM [15], which uses several metrics to estimate the functional impact of each mutation per gene and computes their deviation with respect to the background.

The clustering of the mutations in particular positions of the protein primary structure is a third signal of positive selection. Those mutations that accumulate in certain positions of the protein should correspond to events targeted by cancer. This idea is implemented by OncodriveCLUST [16], which takes into account that the probability of mutations is not homogeneous across the protein sequence. This method uses silent muta-

Table 1

	Method	Description
Estimation of the functional impact of a single non-synonymous variants	SIFT [6]	Builds a MSA of similar proteins according to a database defined by the user and calculates normalized probabilities for all possible substitutions at all positions of the alignment. Based on these probabilities, SIFT classifies observed substitutions as likely neutral or deleterious.
	Polyphen-2 [7]	Naïve Bayes classifier trained from two data sets that contain both deleterious and neutral amino acid changes. Eight sequence-based and three structure-based predictive features, most of them involving comparison of a given property of the wild-type amino acid and its mutated counterpart are the properties used to build the classifier.
	Mutation Assessor [8]	A prediction of the functional impact of nsSNVs is based on the assessment of evolutionary conservation of amino acid residues. It exploits the evolutionary conservation in protein subfamilies, which are determined by clustering MSAs of homologous sequences on the background of conservation of overall function.
	Condel [9]	Condel (Consensus deleteriousness score) is an approach to combine the functional impact scores of nsSNVs. It uses values extracted from the complementary cumulative distributions of the scores produced by individual tools on a dataset of deleterious and neutral nsSNVs as weights to combine them.
	FATHMM [10]	Predicts the functional effects of cancer somatic mutations combining sequence conservation with hidden Markov models representing the alignment of homologous sequences and conserved protein domains with cancer "pathogenicity weights" representing the tolerance of the corresponding model to cancer mutations
	CHASM [11]	A random forest classifier is trained on a curated set of driver mutations derived from COSMIC and randomly simulated passenger mutations. It uses eighty-six diverse features (available at SNVBox database), including physio-chemical properties of amino acid residues, scores derived from MSAs of protein or DNA, region-based amino acid sequence composition, predicted properties of local protein structure and annotations from the UniProtKB feature tables.
	transFIC [12]	transFIC (for transformed functional impact scores for cancer) takes the Functional Impact Score produced by any method aimed at evaluating the impact of a mutation on the functionality of a protein and transforms it, taking into account the baseline tolerance of similar proteins to functional impacting variants. The transformation can be interpreted as an adjustment for the impact of the somatic variant on cell operation.
Identification of driver genes across tumor cohorts	MuSiC [13]	Identifies genes more mutated than expected by chance taking into account several features that shape the baseline probability of each observed mutation –e.g. sequencing coverage, gene length and nucleotide change type.
	MutSigCV [14]	Identifies genes recurrently mutated; it incorporates additional data related with the mutation burden expected in each gene to construct the background model. These covariates includes, among others, data of gene expression collected from the RNAseq observed in cancer cell lines and DNA replication time measured in HeLa cells.
	OncodriveFM [15]	Identifies genes biased towards the accumulation of mutations with a larger functional impact measured by the combination of several metrics.
	OncodriveCLUST [16]	Identifies genes whose mutations are clustered more than expected by a baseline model constructed by using silent mutations
	Active Driver [17]	Identifies genes whose mutations tend to accumulate in or around phosphosites
Identification of driver modules across tumor cohorts	MEMo [18]	Finds gene modules connected according a priori pathway data exhibiting sample mutually exclusive alterations
	HotNet [19]	Finds gene modules in which a certain metric –e.g. the mutation frequency or the functional impact score– is accumulated according to a heat diffusion model propagated via a priori pathway data connected genes

MSA：multiple sequence alignment, nsSNVs：non-synonymous single nucleotide variants

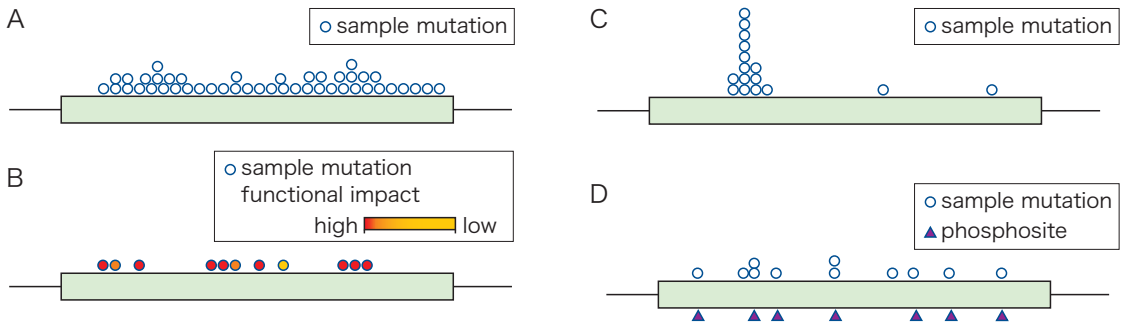

Figure 1 Signals of positive selection

Representation of four genes showing different signals of positive selection across a cohort of cancer samples. Each dot depicts the gene position of a somatic mutation in a different sample of that cohort. A) Accumulation of a high number of mutations that is larger than the expected by chance, thus the gene would be identified as frequently mutated for that tumor cohort. Methods need to take into account multiple features known to influence the baseline mutation probability. B) Mutations observed in the gene are biased towards those that cause a high functional impact, indicating that this type of mutations have been selected and thus that the gene is a driver. The ability of the metrics aimed to score the functional impact shapes the performance of this approach -e.g. mutations that change residue 1047 of PIK3CA are known to be oncogenic but are underestimated by functional impact measurement based on conservation, since this particular residue is highly variable across species [16]. C) Mutations observed in the gene tends to accumulate in very specific regions, as highlighted by a method based on clustering criteria. This approach better captures gain-of-function mutations, such as the BRAF oncogenic mutations in residue 600 [16], since truncating mutations leading to loss-of-function are more distributed across the gene. D) In this case, mutations tend to occur in phosphosites of the protein. This is highlighted by a method aimed to identify cancer events targeting protein phosphorylation.

tions as an estimation of the baseline clustering of mutations in a tumor. Genes with regions harboring an accumulation of non-synonymous mutations above this baseline are detected by the method.

Additional approaches have been also developed to identify other signals of positive selection. For instance, ActiveDriver detects genes whose somatic mutations preferentially occur in or around phosphosites, and thus it is aimed at driver events disrupting phosphorylation networks [17]. Other methods examine mutations in gene modules: MEMo uses pathway data to detect cliques of connected genes whose mutations follow a pattern of mutual exclusivity in samples. The rationale behind it is that a mutation on a second gene of an already mutated pathway either does not confer further selective advantage to the tumor cell or can cause synthetic lethality [18]. On the other hand, the HotNet algorithm groups connected genes enriched for a particular metric –e.g. the mutation frequency– following a heat diffusion model through a gene interaction map [19].

Needless to say, the performance of any method is shaped by the inability of constructing a statistical framework taking into account the complexity of all the potential factors and also by the limitations of the criteria itself. For instance, methods based on mutation frequency tend to overlook lowly recurrent drivers; the assessment of functional impact is clearer for loss-of-function events; and the identification of clustered mutations better identifies oncogenes. Therefore, the combination of several approaches should be the best option to balance their pros and contras and to obtain the most comprehensive and confident list of driver genes. We recently analyzed 3,205 samples from 12 different cancer types following this idea [20]. The resulting list of 291 putative mutational drivers showed that the retrieval of *bona fide* cancer genes is improved by using a combination of methods that examine complementary signals of positive selection.

Finally, it is important to stress that not all mutations occurring in a gene identified as cancer driver are necessarily involved in the tumorigenesis. In other words, driver genes have the potential to cause tumor phenotypes and tend to accumulate driver mutations, but they can also bear passenger events. This should be kept in mind when evaluating mutations of an indi-

vidual tumor.

3 Towards a full catalog of cancer drivers

Mutations in driver genes do not fully explain tumorigenesis in all the cases: some tumor samples –specially those with a lower mutation burden– show few or no mutations in driver genes, although common knowledge expects that the disease be caused by several alterations acquired during its progression. This can be due to several reasons: first, the failure in detecting all mutational driver coding genes. Second, some driving events may occur in non-coding regions. Third, the tumorigenesis can be driven by mechanisms other than mutations (e.g. translocations, copy number alterations, hypermethylation). However, we expect that the analysis of mutations from large tumor cohorts should be able to retrieve a comprehensive catalog of cancer drivers, as it is likely that most genes driving tumorigenesis through other mechanism can also be drivers upon point mutations. For instance, drivers targeted by epigenomic silencing or gene deletion are likely to be targeted also by truncating mutations; those with driver copy number amplification or gene fusion, can also act as drivers through activating mutations [21].

Several recent studies have published catalogs of driver genes acting across large data sets provided, primarily, by The Cancer Genome Atlas and the International Cancer Genome Consortium initiatives [2) 20) 22) 23)]. All of them showed that few genes are mutated at relatively high frequency, whereas the landscape of cancer is dominated by a long tail of lowly recurrent drivers whose detection is less reliable [2]. It is likely that the discovery of genes that are more frequently mutated is close to be completed [23]–except for those related with rare cancer types not yet systematically studied – whereas the catalog of lowly recurrent drivers will be extended as the number of included cases increases. It is still unknown why some genes are much more commonly targeted by cancer, but this is a matter of the highest interest to better understand the tumorigenic process and the addiction of the tumor cells to certain mechanisms.

4 IntOGen-mutations

One of the major hurdles in cancer genomics research was the lack of bioinformatics pipelines able to easily analyze the large catalogs of mutations obtained by cancer re-sequencing projects. To address this question, we created IntOGen-mutations (http://www.intogen.org/mutations), a web-based platform aimed to identify cancer drivers from datasets of tumor sample cohorts, as well as to browse the results of systematic analyses of currently available cancer projects provided by large international consortia and individual laboratories. The initial release of IntOGen-mutations analyzed data from more than 4,600 tumor samples from 31 large cancer projects collected from different sources [24]. The analysis includes the assessment of signals of positive selection across each tumor cohort and the results obtained at a level of gene, pathway and tissue. In addition, links to external databases of interest are included. IntOGen-mutations will be regularly updated with new cancer genome resequencing data, and the next release will cover the analysis of more than 7,000 samples. In addition to browsing the results of already analyzed datasets the user can analyze either their own cohort of tumor samples or mutations in a single individual. The analysis pipeline can be run online in our servers or locally on the user's computer. Future version of IntOGen-mutations will include information of targeted drugs to further aid therapeutic decision-making.

Final remarks

Cancer is a heterogeneous disease characterized by numerous somatic mutations and its understanding requires separating driver events from passengers. Next-generation sequencing technologies allow analyzing large cohorts of sequenced tumors to detect the signals of positive selection that occur in driver genes. Each method aimed for that purpose presents specific caveats that should be taken into account when interpreting their results, and the best option is to combine the results of complementary approaches to balance the pros and cons of each of them [20]. The use of this strategy has recently allowed to confirm the role of known cancer genes, to extend their implication in other tumors and to discover novel candidates and biological processes involved in tumor evolution. In this regard, the identification of drivers more commonly targeted by the major human cancers are probably almost completed, whereas the discovery of the long tail of lowly recurrent drivers that occur in cancer will be further extended with the analysis of larger datasets. The retrieval of a comprehensive catalog of driver genes is the first step for any downstream anal-

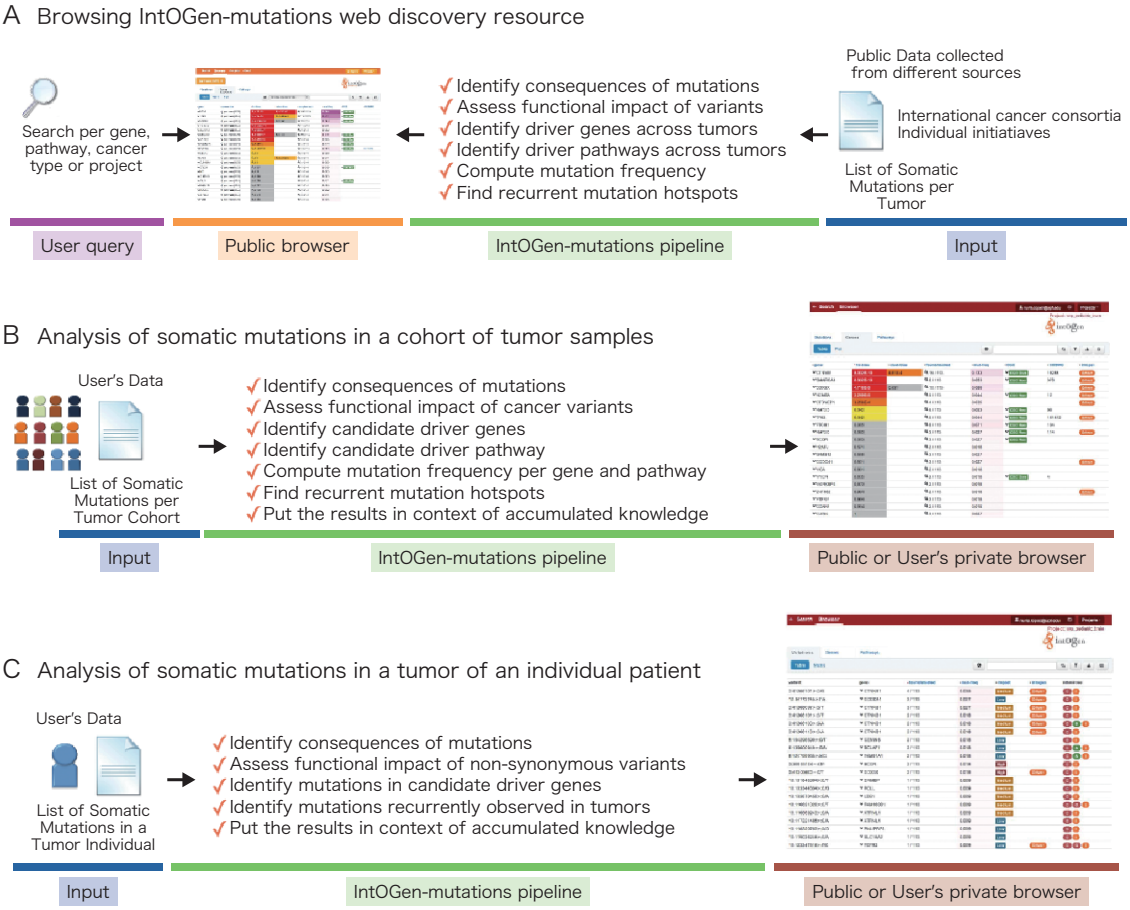

Figure 2 Representation of IntOGen-mutations use cases
The resource allows to browse the results of the systematic analysis of currently available cancer re-sequencing projects (A), to analyse novel somatic mutation data of a tumor cohort (B) and to analyse the somatic mutations of a single tumor (C). 〔Reproduced from (24)〕

ysis aimed to better understand tumorigenesis and to develop novel therapeutic strategies tailored for each individual patient. This will require to understand the action of driver events in both time and space and to identify the specific vulnerabilities of tumor cells in order to design selective therapeutic interventions that will take into account their clonal contents and its interaction with normal cells. These results, which ultimately will require experimental validation, represent the cornerstone of a novel generation of strategies that will lead to a more rational and efficient management of cancer, together with the improvement of personalized medicine via targeted drugs, the use of immunotherapy and the development of better tools for early detection.

References

1) Stratton MR : Science, 331 : 1553–1558, 2011
2) Vogelstein B, et al : Science, 339 : 1546–1558, 2013
3) Alexandrov LB, et al : Nature, 500 : 415–421, 2013
4) Stratton MR, et al : Nature, 458 : 719–724, 2009
5) Gonzalez-Perez A, et al : Nat Methods, 10 : 723–729, 2013
6) Ng PC & Henikoff S : Nucleic Acids Res, 31 : 3812–3814, 2003
7) Adzhubei IA, et al : Nat Methods, 7 : 248–249, 2010
8) Reva B, et al : Nucleic Acids Res, 39 : e118, 2011
9) González-Pérez A & López-Bigas N : Am J Hum Genet,

88 : 440–449, 2011
10) Shihab HA, et al : Bioinformatics, 29 : 1504–1510, 2013
11) Carter H, et al : Cancer Res, 69 : 6660–6667, 2009
12) Gonzalez-Perez A, et al : Genome Med, 4 : 89, 2012
13) Dees ND, et al : Genome Res, 22 : 1589–1598, 2012
14) Lawrence MS, et al : Nature, 499 : 214–218, 2013
15) Gonzalez-Perez A & Lopez-Bigas N : Nucleic Acids Res, 40 : e169, 2012
16) Tamborero D, et al : Bioinformatics, 29 : 2238–2244, 2013
17) Reimand J & Bader GD : Mol Syst Biol, 9 : 637, 2013
18) Ciriello G, et al : Genome Res, 22 : 398–406, 2012
19) Vandin F, et al : Genome Res, 22 : 375–385, 2012
20) Tamborero D, et al : Sci Rep, 3 : 2650, 2013
21) Tamborero D, et al : PLoS One, 8 : e55489, 2013
22) Kandoth C, et al : Nature, 502 : 333–339, 2013
23) Lawrence MS, et al : Nature, 505 : 495–501, 2014
24) Gonzalez-Perez A, et al : Nat Methods, 10 : 1081–1082, 2013

Nuria Lopez-Bigas : Dr. Lopez-Bigas is the head of the Biomedical Genomics group at the Universitat Pompeu Fabra in Barcelona. She completed her PhD degree in 2002 on the molecular basis of deafness at the Oncologic Research Institute in Barcelona. She then moved to the European Bioinformatics Institute in Hinxton (Cambridge, UK) to participate in a project on the computational study of disease and cancer genes. Since 2006 she has been a group leader at the University Pompeu Fabra in Barcelona focusing on cancer genomics and bioinformatics, in 2011 she was appointed as ICREA research professor.

本総説の日本語訳は，下記書籍にてご覧いただけます．

実験医学増刊　Vol.32 No.12

個別化医療を拓く がんゲノム研究

一解き明かされるがんの本質と分子診断・治療応用への展開

柴田龍弘／編

B5判　231ページ　ISBN978-4-7581-0340-4

http://www.yodosha.co.jp/jikkenigaku/book/9784758103404/

[発行元]

株式会社　羊　土　社
〒101-0052
東京都千代田区神田小川町2-5-1
TEL　03(5282)1211(代表)
FAX　03(5282)1212
E-mail　eigyo@yodosha.co.jp
URL　http://www.yodosha.co.jp/