COSMIC, curating the cancer variome.

Simon A. Forbes[1], Gurpreet Tang[1], Jon Teague[1], Andrew Futreal[*1], Mike Stratton[1]

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

* Corresponding author email : paf @sanger.ac.uk

Abstract

Background. COSMIC (http://www.sanger.ac.uk/cosmic) is a system designed to curate the world's literature on somatic mutations in known cancer genes. Initially conceived to capture the mutation spread in point-mutated genes, COSMIC has now grown to encompass gene fusion products of genome rearrangement events which generate completely novel transcripts, together with all the somatic mutation data from candidate gene screens at the Cancer Genome Project, UK (CGP), covering almost 5000 genes of potential interest in cancer genetics.
Results. The latest release of COSMIC (version 37; July 2008) now holds full and up-to-date curation of over 5,900 scientific papers, examining over 268,000 tumours, in which over 59,000 mutations are detailed through 60 point-mutated genes. Fusion gene products have been curated for 16 pairs of genes, described through over 4200 tumours. 2246 papers were rejected during manual curation, usually due to significant inconsistencies in the publication. A relational database holds the captured information, which is warehoused for each release. The information is presented on the internet with a series of graphical and tabulated views aiding navigation and interpretation.
Conclusions. The current version of COSMIC is close to fulfilling its original intentions, with curation of most point-mutated genes in cancer complete. However, new challenges are emerging with the need to calculate the effect of high numbers of observed sequence changes to identify those driving tumour formation, and the need to meaningfully handle the increasing quantities of data from high-throughput screens and next-generation sequencing technologies.

Background

The scientific literature contains tens of thousands of publications describing the involvement of somatic gene mutations in a wide range of human cancer phenotypes and this is enhanced by a number of online resources. However, these online resources tend to be focused on individual loci (e.g. IARC p53 database, [1]), not providing genome-wide information on mutation combinations within tumours, or very broadly focused, storing minimal information, usually only on high-frequency mutant alleles, thus losing much context detail (e.g. OMIM, [2]).

The wide distribution of the data in the literature, together with the broad spread of available details and formats makes it very difficult to investigate for aggregate statistical interpretation. COSMIC is designed to overcome this limitation by curating all this data, in as much detail as possible, into one repository which can be examined easily on the internet. Literature curation data are presented in tandem with somatic mutation screening derived from ongoing CGP studies. Large datasets are therefore made available for deep datamining, whilst maintaining sample sizes which can still achieve good statistical significance.

The CGP maintains a listing of all genes which are proved to have an involvement in causing cancer when mutated, called the Cancer Gene Census (CGC, http://www.sanger.ac.uk/genetics/CGP/Census/ [3]).

Describing over 360 genes, this resource has led the curation efforts of the COSMIC project. Genes with small somatic intragenic mutations have been prioritised, followed by genes involved in novel oncogenic fusions. All the genes released in COSMIC are periodically revisited to ensure their mutation data are maintained up-to-date,

All the data in COSMIC is manually curated, so as to maximise the precision of the data as it is interpreted from published sources. Manual curation also provides significant feedback on the quality of data being curated as well as the methods used in its interpretation; to ensure the quality of COSMIC's data, 2246 papers have been rejected during the curation process due to the absence of mandatory datapoints or significant inconsistencies in the publication. In addition to this manual curation, COSMIC's subproject, the "CGP Cancer Cell Line Project" evaluates each observed variant for its likely impact in tumour formation, specifically releasing only those variants considered to have a deleterious effect.

Results

Curation of the literature for 62 genes with oncogenic point mutations is now nearing completion (with the notable exception of TP53 which is independently curated; IARC TP53 database, [1]). Additionally, 15 genes have been assessed for their involvement in fusion events. Whilst the literature curation domain in COSMIC aims to completely capture the mutation data on a small set of known cancer genes resulting in large sample counts, the CGP domain aims to examine a small number of tumours through a large set of almost 4,800 candidate genes searching for novel oncomutations. COSMIC has now captured results on a total of 268,938 tumours which have been examined through 4,773 genes in various combinations representing 1,019,304 individual experiments. 5,902 publications, together with unpublished CGP contributions have described 59,187 small intragenic mutations and 2,266 instances of novel fusion mutations.

COSMIC's website aims to make searching, navigation and interpretation as easy as possible by providing much graphical summary with the contents of each image reflected in a table with links to further details (figure 1). Detailed descriptions of the system's concepts, contents and usage have recently been published [4].

Discussion

The extraction of complex mutation annotation together with clinical and phenotypic details about tumour samples has proved beyond any retrieval method other than manual. Building the COSMIC dataset has taken a team of manual curators eight years and is still ongoing. Attempts to automatically text-mine scientific publications could retrieve simple data components, but rarely captured the rich contextual information that makes COSMIC so successful in cancer genetics.

With the advent of systematic candidate gene screening and next generation sequencing technologies, the size and scope of screens is increasing. In one of the largest analysis to date, 22 tumours were examined through over 18,000 genes in an attempt to find novel oncogenic variants [5]. Due to the size of these studies, finding and interpreting the supplementary datasheets where the results were stored is a complex and time-consuming procedure, not suited to text-mining programs or humans; a semi-manual/semi-automated approach is now being explored; perhaps sample context can be retrieved manually and simple lists of genes and mutations automatically interpreted. With the recent advent of whole-genome sequencing technologies these complex contextual problems will become even more problematic, increasingly likely to require such an approach. The CGP has reported its first successful examinations of whole genomes using next-generation sequencers [6]; tumours can now receive annotations across their entire genome, not just of mutations in putative cancer genes, but of genome mutations and rearrangements regardless of their position relative to known coding or regulatory sequences. With total mutation numbers in the thousands, the storage and navigation of these data in COSMIC will need to be radically altered, offering novel navigation and graphical overviews. Integrating this into the current system is an upgrade challenge currently underway, keeping COSMIC up-to-date as mutation detection technology improves apace.

COSMIC's presentation of complex mutation data in a phenotypic context has proved very successful, with the website consistently registering around 400,000 page impressions per week in 2008. However, imparting meaning to each variant has become an immensely complex proposition, especially with the larger systematic screens examining anonymous genes of unknown function. Most of COSMIC makes very little distinction between mutations known to cause cancer and passenger mutations, unless this is discussed in the publication being curated, which is rare. Whilst it is easy to define, for instance, a frameshift mutation in a tumour suppressor gene as an oncogenic variant, it can be very difficult to determine the oncogenic consequence of a novel missense mutation. CGP's "Cancer Cell Line Project", displayed on COSMIC's green pages is the only part of the system to make this causal distinction. With highly unstable genomes, cell lines provide mutation hunters with many variants, many of which will have no significant oncogenic consequence. COSMIC's green pages therefore, only record and display those with obvious or previously characterised oncogenic characteristics; any variants not obviously oncogenic are not shown. How this consequence is measured, is again a manual process largely based upon previous published evidence that the

variant has been observed somatic before or that it has clear functional consequences [7].

Many of the well known mutations in cancer have been fully characterised; for instance, mutations at codons 12,13 and 61 in KRAS are known to interfere with the binding of RAS and its inactivator, thus extending the molecule's signalling, upregulating cellular growth via the MAPK/ERK signalling cascade [8]. However, missense mutations are recorded at over 20 other sites in this gene and the meaning of these is much less clear, even when the positions of these residues in the 3D structure of the molecule is easily identifiable. Again, software is available (e.g. SIFT [9], CanPredict [10]) to help with this interpretation and a statistical evaluation of the output of many such prediction programs may well be the way forward in helping sift for the significant oncomutations requiring further examination. Calculating the mutation consequence for the larger number of variants from systematic screens is an immediate challenge in cancer genetics. The majority of sequence changes identified in such screens are likely to be passengers rather than drivers, and the ability to differentiate the two with high-throughput methodologies is becoming a necessity. COSMIC's high-quality systematic curation and storage of cancer mutations provides the ideal framework and test datasets on which to found an effort to develop these new standards and algorithms.

References

1. Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian,S.V., Hainaut, P., Olivier, M. 2007. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat. 28:622-29.

2. Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., McKusik, V.A. 2002. Online Mendelian Inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl. Acids Res. 30:52-55.

3. Futreal, P.A, Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R. 2004. A census of human cancer genes. Nature Reviews Cancer 4:177-183.

4. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008, Chapter 10:Unit 10.11.

5. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. The genomic landscapes of human breast and colorectal cancers. Science. 2007,318:1108-13

6. Campbell P.J., Stephens P.J., Pleasance E.D., O'Meara S., Li H., Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA. 2008. Nat Genet 40:685-686.

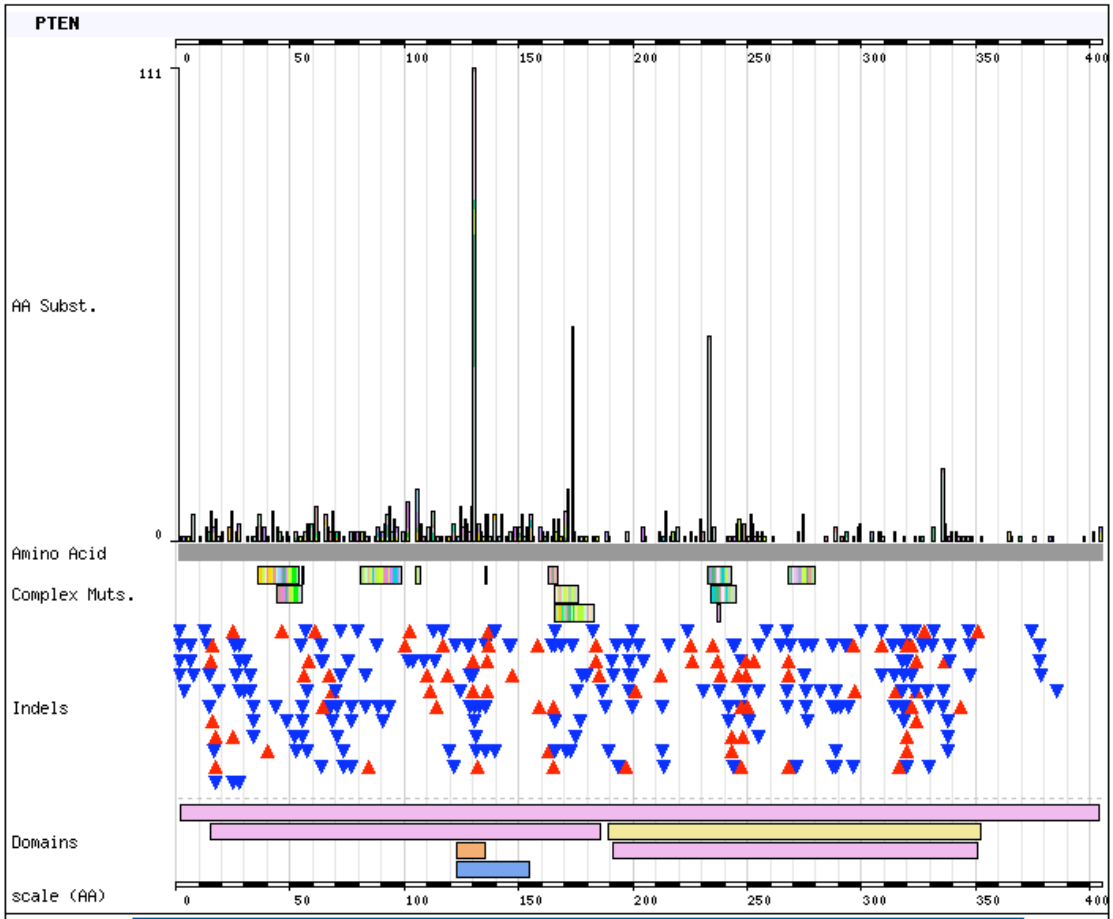
7. Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, Santarius T, Avis T, Barthorpe S, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Hunter C, Jenkinson A, Jones D, Kosmidou V, Lugg R, Menzies A, Mironenko T, Parker A, Perry J, Raine K, Richardson D, Shepherd R, Small A, Smith R, Solomon H, Stephens P, Teague J, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Reinhold W, Weinstein JN, Stratton MR, Futreal PA, Wooster R. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006, 5:2606-12.

8. Scheffzek K, Ahmadian MR, Kabsch W, Wiesmuller L, Lautwein A, Schmitz F, Wittinghofer A. The Ras-RasGAP complex:structural basis for GTPase activation and its loss in oncogenic RAS mutants. 1997. Science 277: 333-338.

9. Ng PC, Henikoff S. 2003. Nucleic Acids Res 31:3812-4

10. Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. 2007. Nucleic Acids Res. 35:W595-8.

Figures



Details for PTEN

Primary Tissue	Mutated Samples	% Mutated	All Samples	Mutation Data
NS	0	0%	22	More Details
adrenal gland	0	0%	60	More Details
autonomic ganglia	2	3%	64	More Details
biliary tract	1	17%	6	More Details
bone	5	7%	74	More Details
breast	34	6%	561	More Details
central nervous system	556	20%	2758	More Details

Figure 1. Upon selection of a gene and a cancer phenotype the histogram page, probably the core of the system, displays the mutation spectrum in a variety of accessible ways.