bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations

David Tamborero[1,2], Carlota Rubio-Perez[1], Jordi Deu-Pons[1,2], Michael P Schroeder[1,3], Ana Vivancos[4], Ana Rovira[5], Ignasi Tusquets[5,6], Joan Albanell[5], Jordi Rodon[4], Josep Tabernero[4], Carmen de Torres[7], Rodrigo Dienstmann[4], Abel Gonzalez-Perez[1,2] and Nuria Lopez-Bigas[1,2,8]

[1] Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical, Research Institute and Pompeu Fabra University, Barcelona, Spain

[2] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

[3] Charité – Universitätsmedizin Berlin, Germany

[4] Vall d´Hebron Institute of Oncology, Vall d'Hebron University Hospital, Autonomous University of Barcelona, Barcelona, Spain

[5] Medical Oncology Service, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

[6] Autonomous University of Barcelona, Barcelona, Spain

[7] Developmental Tumor Biology Laboratory, Institut de Recerca Pediàtrica- Hospital Sant Joan de Déu Barcelona, Spain

[8] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract

While tumor genome sequencing has become widely available in clinical and research settings, the interpretation of tumor somatic variants remains an important bottleneck. Most of the alterations observed in tumors, including those in well-known cancer genes, are of uncertain significance. Moreover, the information on tumor genomic alterations shaping the response to existing therapies is fragmented across the literature and several specialized resources. Here we present the Cancer Genome Interpreter (http://www.cancergenomeinterpreter.org), an open access tool that we have implemented to annotate genomic alterations and interpret their possible role in tumorigenesis and in the response to anti-cancer therapies.

New computational tools to support the interpretation of tumor genomes are needed

Cancer is predominantly a genetic disease, caused by the accumulation of so-called "driver" genomic alterations that confer cells tumorigenic capabilities[1]. Thousands of tumor genomes are sequenced every year in research projects and clinical settings around the world. In some cases the whole-genome is sequenced while other focus on the exome or a panel of selected genes. In all cases, the sequencing is followed by the necessity to annotate which of the somatic mutations identified have a possible role in tumorigenesis and treatment response. We call this process 'the interpretation of cancer genomes' and it is currently a tedious

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

procedure. One of its major bottlenecks is identifying the driver alterations. A widely employed approach to solve this hurdle consists in focusing on the mutations affecting known cancer genes, i.e., tumor suppressors and oncogenes. These were initially identified through experimentation, giving rise over the past 40 years to a census of human cancer genes[2]. More recently, large re-sequencing projects have provided the opportunity to systematically identify the genes involved in tumorigenesis by detecting signals of positive selection in their alterations pattern across about two dozen malignancies[3–6]. Nevertheless, many somatic variants in tumors, even those in cancer genes, still have uncertain significance and thus it is not clear whether or not they are drivers. Another hurdle in the interpretation of cancer genomes concerns one of its crucial aims: the identification of tumor alterations that may affect treatment options. Unstructured information on the effectiveness of therapies targeting specific cancer drivers is continuously generated by clinical trials and pre-clinical experiments. In summary, novel computational tools are required to address the two aforementioned critical challenges. This includes, on the one hand, methods to estimate the oncogenic effect of the variants observed in a tumor (i.e., identifying validated driver variants and providing some estimation for variants of unknown significance), and on the other, resources that systematically gather the information on biomarkers of drug response and organize them according to distinct use requirements.

The Cancer Genome Interpreter

Here, we describe the Cancer Genome Interpreter (CGI), a platform that systematizes the interpretation of cancer genomes and makes it automatic. The specific aim of the CGI is to determine which alterations observed in a tumor are more likely to be drivers and identify those that may constitute biomarkers of response to therapies (Fig. 1; details in Supp. Note I). CGI relies on existing knowledge collected from several resources and on computational methods that annotate the alterations in a tumor according to distinct levels of evidence. The tool is a freely available web-resource under an open license, which is intended to facilitate its use by cancer researchers and medical oncologists (http://cancergenomeinterpreter.org). In the following sections we present a blueprint for the interpretation of cancer genomes and describe its implementation in the CGI.

A comprehensive catalog of cancer genes across tumor types

One of the main aims of the interpretation of cancer genomes is to identify the alterations responsible for oncogenic traits. We propose that this process begins with a focus on alterations that affect the genes capable of driving the growth of a particular tumor type. Therefore, we compiled a catalog of genes involved in the onset and progression of different types of cancer, obtained via different methods and from different sources (Supp. Note II). First, we collected genes that have been experimentally or clinically verified to drive tumorigenesis from manually annotated resources[2,7–10] and the literature. Second, we exploited the bioinformatics results from the analysis of large tumor cohorts re-sequenced by international efforts such as The Cancer Genome Atlas and the International Cancer Genome Consortium[11,12]. On detail, we identified genes whose somatic alterations exhibit signals of positive selection across 6,729 tumors representing 28 types of cancer[4]. In addition, we retrieved the mode of action of each of these cancer genes (i.e., whether they function as an oncogene or a tumor suppressor), curated following state-of-the art knowledge when available and otherwise estimated *in silico*[13]. The resulting Catalog of Cancer Genes currently comprises 837 genes with some evidence of being drivers in 193 different cancer types (Fig. 2a). We annotated each of these

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes, identifying (i) the malignancies it drives, organized according to available evidence; (ii) the types of alterations involved (mutations, copy number alterations and/or gene translocations); (iii) the original source(s) reporting it; (iv) the context (germline or somatic) in which these alterations are tumorigenic; and (v) its mode of action as appropriate. The Catalog is available for download through the CGI website (https://www.cancergenomeinterpreter.org/genes).

Most mutations affecting cancer genes are of uncertain significance

A key aspect of assessing the mutations observed in cancer genes is the tumorigenic potential of each individual variant, as not all of them are necessarily capable of driving tumorigenesis. Therefore, the CGI next focuses on protein affecting mutations (PAMs) that occur in genes of the Catalog of Cancer Genes. Validated tumorigenic mutations may confidently be labeled as drivers when detected in a tumor. We compiled an inventory that currently contains 3,939 such validated driver or cancer predisposing variants from dedicated resources[7–10,14] and the literature (Fig. 2B and Supp Note III). This Catalog of Validated Oncogenic Mutations is available for download through the CGI website (https://www.cancergenomeinterpreter.org/mutations). In the pan-cancer cohort of 6,792 sequenced tumors[4] only 4,142 (630 unique variants) of the 44,648 PAMs found in cancer genes appear in this Catalog. In other words, 90.7% of all PAMs that affect cancer genes in this cohort are currently of uncertain significance for tumorigenesis, a proportion that varies widely per gene and tumor type (Fig. 2c and Supp Note VII). This highlights the need for a means to estimate the tumorigenic potential of these variants. We reasoned that several features of each specific mutation as well as of the genes affected by them could help address this question. Moreover, we propose that some of these features of interest can be extracted from the analyses of large sequenced cohorts of healthy and tumor tissue[4,15]. Examples of relevant attributes include the following: i) the tumorigenic mode of action of the gene in that cancer (oncogene or tumor suppressor); ii) the consequence type of the mutation (e.g. synonymous, missense or truncating); iii) its position within the transcript; iv) whether it falls in a mutational hotspot or cluster; v) its predicted functional impact; vi) its frequency within the human population; and vii) whether it occurs in a domain of the protein that is depleted of germline variants. The CGI assesses the tumorigenic potential of the variants of unknown significance via OncodriveMUT, a rule-based approach that combines the values of these features (Fig. 1C; Supp. Note IVa). To assess the performance of OncodriveMUT in the task of classifying driver and passenger mutations, we used the Catalog of Validated Oncogenic Mutations (n=3,939) and a collected set of neutral PAMs affecting cancer genes (n=1,247). We found that OncodriveMUT separates the variants of these two data sets with 91% of accuracy (Matthews correlation coefficient, 0.78) (Supp Note IVb). Furthermore, the predictions of OncodriveMUT exhibited a high concordance with the results of experiments assessing the tumorigenic effect of other mutations that are uncommonly seen in cancer[16–19] (Supp Note IVb). In summary, the CGI annotates the mutations affecting cancer genes with features relevant to their potential role in cancer to facilitate the user's review, identifying validated drivers and classifying the most likely drivers among the variants of unknown significance.

A database of genomic determinants of anti-cancer drug response

The second major aim of the effort to interpret cancer genomes is to identify which of the tumor alterations may shape the response to anti-cancer therapies. Findings about the influence of genomic alterations on drug

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

response are continuously generated and reported through publications, clinical trials and conference communications. The challenge resides in gathering relevant results into an easy-to-use resource, and organizing them according to the needs of different users. The CGI employs two resources to explore the associations between gene alterations and drug responses. The first is the Cancer Biomarkers database, an extension of a previous collection of genomic biomarkers of anti-cancer drug response[8], which currently contains information on 1,574 genomic biomarkers of response (sensitivity, resistance or toxicity) to 221 drugs across 79 types of cancer. Negative results of clinical trials, e.g. the unsuccessful use of BRAF V600 inhibitors as a single therapeutic agent in colorectal cancers bearing that mutation, are also included in the database. Importantly, these biomarkers are organized according to the level of clinical evidence supporting each one, ranging from results of pre-clinical data, case reports, and clinical trials in early (I/II) and late phases (III/IV) to standard-of-care guidelines. The database is under continuous update by a board of medical oncologists and cancer genomics experts (Fig. 3A and Supp. Note V). The second resource is the Cancer Bioactivities database, which currently contains information of 20,243 chemical compounds-protein product interactions that may support novel research applications. We built this database by compiling a catalog of available results from bioactivity assays of small molecules interacting with cancer genes. This information was obtained by querying several external databases (Supp. Note VI). The CGI matches biomarkers or target genes in these databases to alterations observed in tumors. Of note, it reports co-occurring alterations that affect the response to a given treatment. This includes the co-existence of biomarkers of resistance and sensitivity to the same drug, and biomarkers of drug sensitivity that depend upon simultaneous genomic events.

In summary, these two databases constitute comprehensive repositories of genome-guided therapeutic actionability in cancer according to current supporting evidences. Both resources are available for download through the CGI website (https://www.cancergenomeinterpreter.org/biomarkers, https://www.cancergenomeinterpreter.org/bioactivities). The integration of these two databases with those developed in parallel by other institutions with similar purposes is currently being undertaken within the framework of the Global Alliance for Genomics & Health[20], described below.

Current applications and future prospects of the CGI

The CGI (and the databases gathered for its implementation) are under open license, and the resource can be accessed via the web resource and an Application Programming Interface (API; see Supp. Note Ic and Id). The use of the CGI to automatically interpret cancer genomes has broad potential applications, ranging from basic cancer genomics to the translational setting. One feature of the CGI that makes it particularly suitable to different types of applications is its flexibility. The user can input tumor alterations by uploading files following different standards and/or by typing them in a free-text box. The system is prepared to automatically recognize and re-map as necessary different formats, such as genomic, transcript or protein-based coordinates for mutations (Supp. Note Ib). The use of the CGI can help addressing questions raised in different oncology research settings. A newly sequenced group of tumors may be readily interpreted, as exemplified with the pan-cancer cohort presented in this article. The application of the CGI to the mutations profiled across the whole exomes of these tumors delivered a catalog of putative driver alterations across its

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28 cancer types (made available through http://www.intogen.org) (Suppl Note VII). The potential of a comprehensive analysis of individual alterations is illustrated by the identification of uncommon events that may be exploited by drug repurposing opportunities (Figure 3B and Supp Note VII). Overall, the CGI identified 5.2% and 3.5% of the samples in the cohort with genomic alterations that are biomarkers of drug sensitivity used in the clinical practice (FDA-approved or international guidelines) or reported in late (phases III-IV) clinical trials, respectively. When considering biomarkers supported by lower levels of clinical relevance, a total of 62% of the tumors exhibited at least one potentially actionable alteration, a number that largely varied across cancer types (Figure 3C and Supp Note VII). However, this cohort mostly includes samples sequenced at diagnosis and thus they may not reflect the type of tumors that are evaluated by molecular oncology boards at present. We also applied the CGI to the sequencing data of 17,642 tumors recently released by the GENIE project, which gathers more advanced cancers profiled by targeted panels[21]. The CGI identified 8% and 6% in that cohort exhibiting biomarkers of drug sensitivity used in clinical practice or reported in late clinical trials, and overall 72% of these tumors exhibited at least one actionable alteration supported by any level of evidence (Figure 3D and Supp Note VII). In addition, the GENIE cohort exhibited more genomic biomarkers of drug resistance, as expected from tumors with a higher proportion of recurrent/relapse patients (Supp Note VII). These analyses provide a comprehensive state-of-the-art snapshot of the putative genomic drivers of cancer and the landscape of genomic guided therapies as it stands today.

On the other hand, the application of the CGI to analyze the results of drug response observed in tumors with different genomic architecture could contribute to the discovery of novel genomic biomarkers of drug sensitivity or resistance. On detail, the distinction between driver and passenger events allows the development of better predictive models[22]. In the clinical setting, application of the CGI to analyze the list of alterations detected in a patient's tumor could support decision-making in multiple scenarios, especially in cases of variants of unknown significance that may have implications for response to therapy. Early clinical adopters of the CGI used the resource to support the final decision of the most appropriate clinical trial to enroll cancer patients or explore potential drug re-purposing opportunities for pediatric tumors (see Supp. Note VIII).

Crucial to the performance of the CGI are the maintenance and further development of its two distinct types of resources: the repositories of accumulated knowledge and the bioinformatics methods. As new tumor cohorts are re-sequenced and analyzed, our medium-term plans include further development of the catalogs of cancer genes and oncogenic mutations, including both new malignancies and new genomic elements. In particular, the possibility to identify non-coding cancer drivers[23] from currently generated whole-genome mutation data will open up the opportunity to explore the actionability of non-coding genomic alterations (https://dcc.icgc.org/pcawg). With respect to the aggregation, curation and interpretation of databases of cancer biomarkers and bioactivities, our team follows the standard operating procedures developed under the umbrella of the H2020 MedBioinformatics (http://www.medbioinformatics.eu/) project, thus ensuring the mid-term maintenance of these resources. The feedback from the community is also facilitated through the CGI web interface. Access to this type of cancer data is crucial for the advance of precision medicine, but is highly complex and difficulty for a single institution to comprehensively manage and update. Multiple efforts with similar purposes are currently underway, including My Cancer Genome,

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://www.mycancergenome.org; PMKB, https://pmkb.weill.cornell.edu/; PCT, https://pct.mdanderson.org; OncoKB, http://oncokb.org; CIViC, https://civic.genome.wustl.edu; and JAX-CKB https://ckb.jax.org. Within the Global Alliance for Genomics & Health framework[20], the Variant Interpretation for Cancer Consortium (http://ga4gh.org/#/vicc) was recently launched with the aim to unify the curation efforts of several institutions, including our own. We envision that individual databases will continue to be maintained to fulfill specific needs[24], but our long-term impact will largely rely, first, on the establishment of international standards for the collection of data relevant to associations between cancer variant-clinical outcome and, second, on our collective success in encouraging the community to share such knowledge.

In summary, the CGI is a versatile platform that automates the steps we propose for the interpretation of cancer genomes, annotating the potential of the alterations detected in human tumors as cancer drivers and their possible effect on treatment response, according to current levels of evidence. The characteristics of the CGI, and the commitment to maintain it as part of a community effort to keep the resource up-to-date with evolving knowledge, allow its establishment as a widely disseminated, easy-to-use tool for both pre-clinical and translational cancer research settings.

Acknowledgements

This project has received funding from Fundació La Marató de TV3, the European Union's Horizon 2020 research and innovation programme 2014-2020 under Grant Agreement No 634143, and by the European Research Council (Consolidator Grant 682398). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO) (Government of Spain) and is supported by CERCA (Generalitat de Catalunya). DT is supported by project SAF2015-74072-JIN funded by the Agencia Estatal de Investigacion (AEI) and Fondo Europeo de Desarrollo Regional (FEDER). CR-P is funded by FPI MINECO grant (BES-2013-063354). AG-P is supported by a Ramon y Cajal fellowship (RYC-2013-14554). We appreciate the support provided by Wanding Zhou for the use of the TransVar method and the work of Elaine Lilly in the edition of the text.

Figure 1

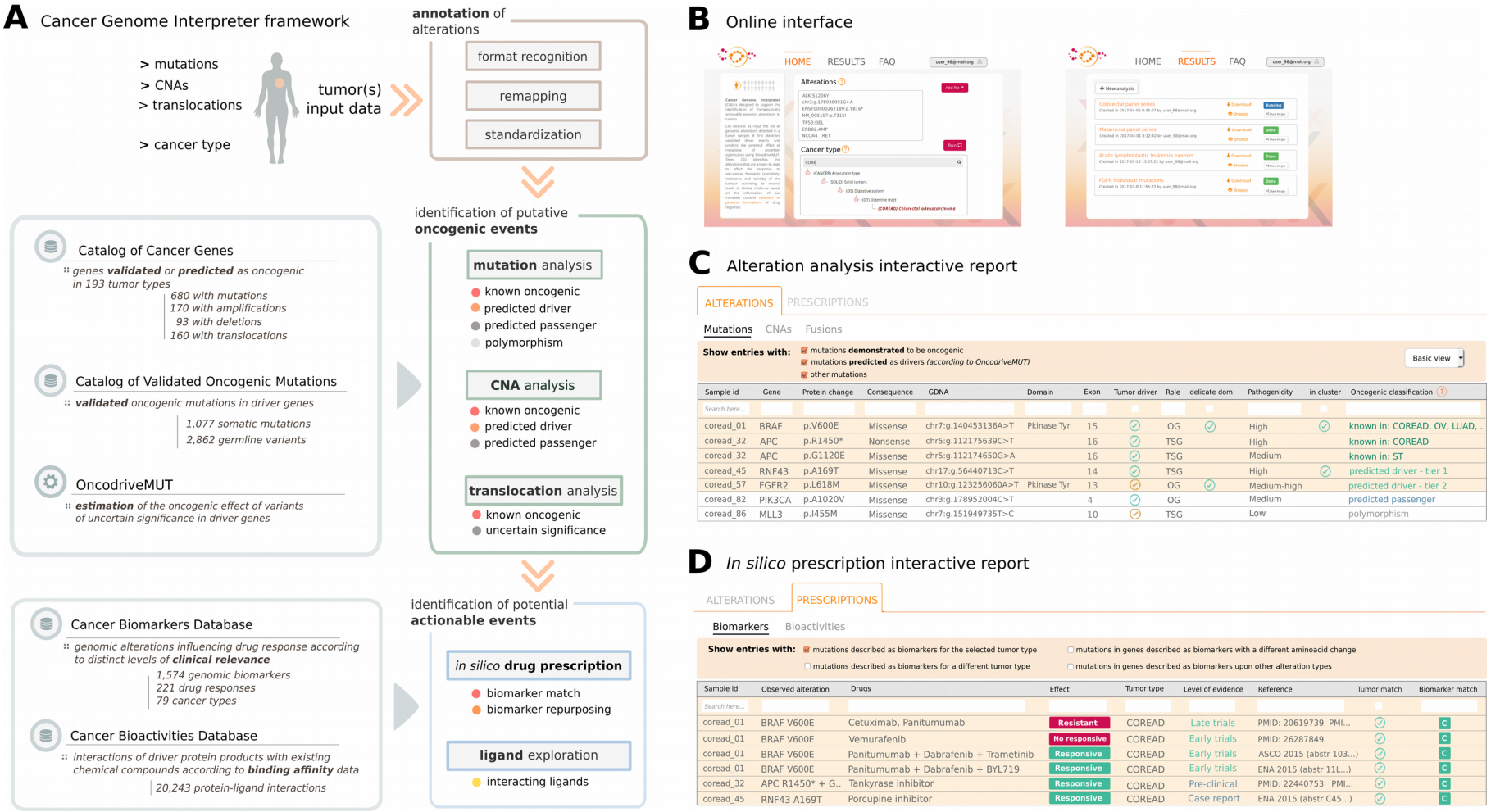
A Cancer Genome Interpreter framework
> mutations
> CNAs
> translocations
> cancer type
tumor(s) input data
annotation of alterations
format recognition
remapping
standardization
identification of putative oncogenic events
Catalog of Cancer Genes
genes validated or predicted as oncogenic in 193 tumor types
680 with mutations
170 with amplifications
93 with deletions
160 with translocations
Catalog of Validated Oncogenic Mutations
validated oncogenic mutations in driver genes
1,077 somatic mutations
2,862 germline variants
OncodriveMUT
estimation of the oncogenic effect of variants of uncertain significance in driver genes
mutation analysis
known oncogenic
predicted driver
predicted passenger
polymorphism
CNA analysis
known oncogenic
predicted driver
predicted passenger
translocation analysis
known oncogenic
uncertain significance
identification of potential actionable events
Cancer Biomarkers Database
genomic alterations influencing drug response according to distinct levels of clinical relevance
1,574 genomic biomarkers
221 drug responses
79 cancer types
Cancer Bioactivities Database
interactions of driver protein products with existing chemical compounds according to binding affinity data
20,243 protein-ligand interactions
in silico drug prescription
biomarker match
biomarker repurposing
ligand exploration
interacting ligands
B Online interface
HOME
RESULTS
FAQ
Alterations
Cancer type
C Alteration analysis interactive report
ALTERATIONS
PRESCRIPTIONS
Mutations
CNAs
Fusions
Show entries with:
mutations demonstrated to be oncogenic
mutations predicted as drivers (according to OncodriveMUT)
other mutations
Basic view
Sample id | Gene | Protein change | Consequence | GDNA | Domain | Exon | Tumor driver | Role | delicate dom | Pathogenicity | in cluster | Oncogenic classification
coread_01 | BRAF | p.V600E | Missense | chr7:g.140453136A>T | Pkinase Tyr | 15 | OG | High | known in: COREAD, OV, LUAD, ..
coread_32 | APC | p.R1450* | Nonsense | chr5:g.112175639C>T | 16 | TSG | High | known in: COREAD
coread_32 | APC | p.G1120E | Missense | chr5:g.112174650G>A | 16 | TSG | Medium | known in: ST
coread_45 | RNF43 | p.A169T | Missense | chr17:g.56440713C>T | 14 | TSG | High | predicted driver - tier 1
coread_57 | FGFR2 | p.L618M | Missense | chr10:g.123256060A>T | Pkinase Tyr | 13 | OG | Medium-high | predicted driver - tier 2
coread_82 | PIK3CA | p.A1020V | Missense | chr3:g.178952004C>T | 4 | OG | Medium | predicted passenger
coread_86 | MLL3 | p.I455M | Missense | chr7:g.151949735T>C | 10 | TSG | Low | polymorphism
D In silico prescription interactive report
ALTERATIONS
PRESCRIPTIONS
Biomarkers
Bioactivities
Show entries with:
mutations described as biomarkers for the selected tumor type
mutations described as biomarkers for a different tumor type
mutations in genes described as biomarkers with a different aminoacid change
mutations in genes described as biomarkers upon other alteration types
Sample id | Observed alteration | Drugs | Effect | Tumor type | Level of evidence | Reference | Tumor match | Biomarker match
coread_01 | BRAF V600E | Cetuximab, Panitumumab | Resistant | COREAD | Late trials | PMID: 20619739 PMI... | C
coread_01 | BRAF V600E | Vemurafenib | No responsive | COREAD | Early trials | PMID: 26287849. | C
coread_01 | BRAF V600E | Panitumumab + Dabrafenib + Trametinib | Responsive | COREAD | Early trials | ASCO 2015 (abstr 103...) | C
coread_01 | BRAF V600E | Panitumumab + Dabrafenib + BYL719 | Responsive | COREAD | Early trials | ENA 2015 (abstr 11L...) | C
coread_32 | APC R1450* + G.. | Tankyrase inhibitor | Responsive | COREAD | Pre-clinical | PMID: 22440753 PMI... | C
coread_45 | RNF43 A169T | Porcupine inhibitor | Responsive | COREAD | Case report | ENA 2015 (abstr C45...) | C

Figure 2

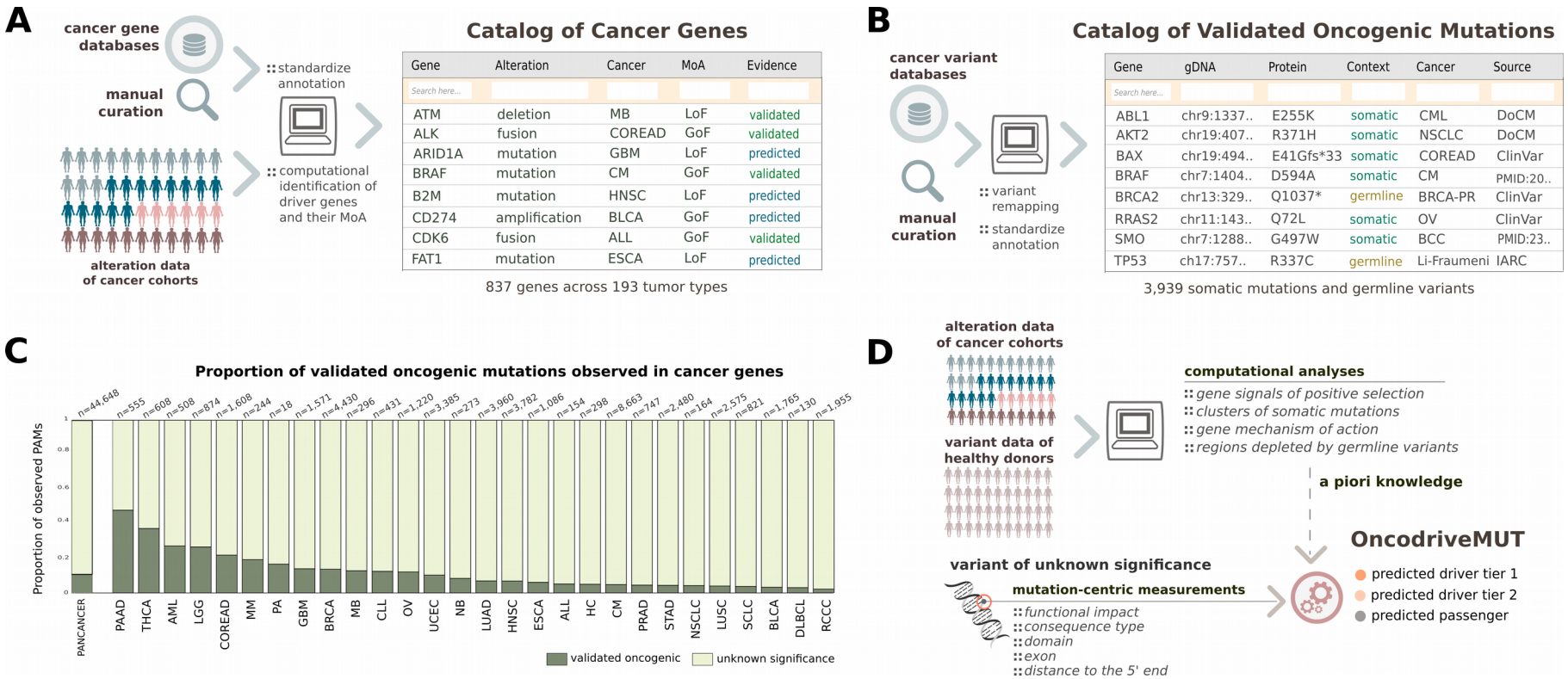
A
cancer gene databases
manual curation
alteration data of cancer cohorts
:: standardize annotation
:: computational identification of driver genes and their MoA
Catalog of Cancer Genes
Gene | Alteration | Cancer | MoA | Evidence
ATM | deletion | MB | LoF | validated
ALK | fusion | COREAD | GoF | validated
ARID1A | mutation | GBM | LoF | predicted
BRAF | mutation | CM | GoF | validated
B2M | mutation | HNSC | LoF | predicted
CD274 | amplification | BLCA | GoF | predicted
CDK6 | fusion | ALL | GoF | validated
FAT1 | mutation | ESCA | LoF | predicted
837 genes across 193 tumor types
B
cancer variant databases
manual curation
:: variant remapping
:: standardize annotation
Catalog of Validated Oncogenic Mutations
Gene | gDNA | Protein | Context | Cancer | Source
ABL1 | chr9:1337.. | E255K | somatic | CML | DoCM
AKT2 | chr19:407.. | R371H | somatic | NSCLC | DoCM
BAX | chr19:494.. | E41Gfs*33 | somatic | COREAD | ClinVar
BRAF | chr7:1404.. | D594A | somatic | CM | PMID:20..
BRCA2 | chr13:329.. | Q1037* | germline | BRCA-PR | ClinVar
RRAS2 | chr11:143.. | Q72L | somatic | OV | ClinVar
SMO | chr7:1288.. | G497W | somatic | BCC | PMID:23..
TP53 | ch17:757.. | R337C | germline | Li-Fraumeni | IARC
3,939 somatic mutations and germline variants
C
Proportion of validated oncogenic mutations observed in cancer genes
Proportion of observed PAMs
PANCANCER PAAD THCA AML LGG COREAD MM PA GBM BRCA MB CLL OV UCEC NB LUAD HNSC ESCA ALL HC CM PRAD STAD NSCLC LUSC SCLC BLCA DLBCL RCCC
validated oncogenic
unknown significance
D
alteration data of cancer cohorts
variant data of healthy donors
computational analyses
:: gene signals of positive selection
:: clusters of somatic mutations
:: gene mechanism of action
:: regions depleted by germline variants
a piori knowledge
OncodriveMUT
predicted driver tier 1
predicted driver tier 2
predicted passenger
variant of unknown significance
mutation-centric measurements
:: functional impact
:: consequence type
:: domain
:: exon
:: distance to the 5' end

Figure 3

A
curators
community
biomarker required fields
Biomarker
:: mutation 1
:: CNA
:: fusion
:: others 2
Drug
:: drug name 3
:: drug family
Effect
:: responsive
:: no responsive
:: resistance
:: toxicity
Cancer
:: tumor type
:: tumor subtype 3
Evidence
:: clinical practice
:: clinical trials III-IV
:: clinical trials I-II
:: case reports
:: pre-clinical data
Reference
:: guidelines 4
:: publication
:: conference abstract
1 may also include the wildtype allele
2 includes multi-marker events
3 when available
4 include international clinical recommendations
annotation
syntax standardization
remapping
links to references
duplicated entries check
1,574 genomic biomarkers
221 anti-cancer drugs
79 cancer types
Cancer Biomarkers database
mutations
CNAs
fusions
others
pre-clinical
case reports
early trials
late trials
guidelines
B
Gene mutations offering drug tumor type - repurposing opportunities
number of tumors
cancer type
ALL
AML
BLCA
BRCA
CLL
CM
COREAD
DLBC
ESCA
GBM
HC
HNSC
LGG
LUAD
LUSC
MB
MM
NB
NSCLC
OV
PA
PAAD
PRAD
RCCC
SCLC
STAD
THCA
UCEC
chemotherapeutic agents
supporting evidence
clinical guidelines/ recommendations
late trials
early trials
case reports
preclinical data
C
Proportion of tumors with genomic biomarkers of drug sensitivity
(6,792 patients with exome sequencing)
Number of analysed samples
total
mutations
CNAs
fusions
data type
D
Proportion of tumors with genomic biomarkers of drug sensitivity
(17,642 patients with gene panel sequencing)
total
mutations
CNAs
fusions

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. Cancer Genome Interpreter

(a) Outline of the CGI workflow. With a list of genomic alterations in a tumor of a given cancer type as input, the CGI automatically recognizes the format, remaps the variants as needed and standardizes the annotation for downstream compatibility. Next, it identifies known driver alterations and annotates and classifies the remaining variants of unknown significance. Finally, alterations that are biomarkers of drug effect are identified. **(b)** The CGI may be run via web at http://cancergenomeinterpreter.com (left panel), or through an API. The web results can be stored in a private repository (right panel) for their management. The results of the CGI are provided via interactive reports: **(c)** Mutation analysis report (example). It contains the annotations of all mutations, which empowers the user's review, and the labels for those known or predicted to be drivers by OncodriveMUT. **(d)** Biomarkers-match report (example). It contains the putative biomarkers of drug response found in the tumor organized according to distinct levels of clinical relevance. These web reports are interactive and configurable by the user.

Figure 2. Annotating mutations in cancer genes

(a) Catalog of Cancer Genes. Genes that drive tumorigenesis via mutations, copy number alterations and/or translocations are annotated with their mode of action (MoA). **(b)** Catalog of Validated Oncogenic Mutations. Clinically or experimentally validated driver mutations were gathered from manually annotated resources and the cancer literature. **(c)** Proportion of validated mutations observed across the cancer genes of 6,792 tumors. Cancer types acronyms: acute lymphocytic leukemia (ALL); acute myeloid leukemia (AML); bladder carcinoma (BLCA); breast carcinoma (BRCA); chronic lymphocytic leukemia (CLL); cutaneous melanoma (CM); colorectal adenocarcinoma (COREAD); diffuse large B cell lymphoma (DLBC); esophageal carcinoma (ESCA); glioblastoma multiforme (GBM); hepatocarcinoma (HC); head and neck squamous cell carcinoma (HNSC); lower grade glioma (LGG); lung adenocarcinoma (LUAD); lung squamous cell carcinoma (LUSC); medulloblastoma (MB); multiple myeloma (MM); neuroblastoma (NB); non small cell lung carcinoma (NSCLC); serous ovarian adenocarcinoma (OV); pilocytic astrocytoma (PA); pancreas adenocarcinoma (PAAD); prostate adenocarcinoma (PRAD); renal clear cell carcinoma (RCC); small cell lung carcinoma (SCLC); stomach adenocarcinoma (STAD); thyroid carcinoma (THCA) and uterine corpus endometrioid carcinoma (UCEC). **(d)** OncodriveMUT schema to estimate the oncogenic potential of the variants of unknown significance. A set of heuristic rules combines the annotations obtained for a given mutation with the knowledge about the genes (or regions thereof) in which it is observed, as retrieved from the computational analyses of sequenced cohorts.

Figure 3: Cancer Biomarkers Database

(a) A board of clinical and research experts gather the genomic biomarkers of drug response to be included in the Cancer Biomarkers database through periodic updates. Upper part of the panel displays the simplified schema of the data model. The clinical/research community is encouraged to provide feedback to edit an existing entry or add a novel one by using the comment feature available in the web service. Any suggestion is subsequently evaluated by the scientific team and incorporated as appropriate. A semi-automatic pipeline annotates any novel entry to ensure the consistency of the attributes, including the variant re-mapping from protein to genomic coordinates when necessary. Lower part of the panel displays some of the 1,574

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

biomarkers that have been collected in the current version of the database, and the left pie charts summarize the content.

(b) CGI analyses detect putative driver mutations in individual tumors that are rarely observed in the corresponding cancer type. When these variants are known targets of anti-cancer therapies, they may constitute tumor type repurposing opportunities. The graph summarizes some of these potential opportunities detected by the CGI on 6,792 tumors with exome-sequencing data, which are currently unexplored. The barplots display the overall number of tumor samples (separated by cancer type) in which they were observed. The acronym of the cancer type in which the genomic event is demonstrated to confer sensitivity to the drug is shown in parenthesis following the name of the drug, and the clinical evidence of that association is represented through color circles (note that the clinical guidelines/recommendations label refers to FDA-approved or international guidelines). Targeted drugs and chemotherapies are shown separately. Cancer acronyms that are not included in the Figure 2 legend: RA: renal angiomyolipoma; BCC: basal cell carcinoma; GCA: giant cell astrocytoma; G: glioma; MCL: mantle cell lymphoma; MRT: malignant rhabdoid tumor; and R: renal; CH: chollangiocarcinoma.

(c) Therapeutic landscape of 6,792 tumors with exome-sequencing data. Fraction of tumors with genomic alterations that are biomarkers of drug response in each cancer type. Colors in the bars denote the clinical evidence supporting the effect of biomarkers in that disease (see evidence colors in panel B). Note that the event with evidence closest to the clinic is given priority when several biomarkers of drug response co-occur in the same tumor sample. The lower part of the graph indicates the total number of samples per cancer type, detailing the number of samples in which mutation, copy number alteration (CNA) and/or fusion data was analysed. Cancer acronyms as in Figure 2 legend.

(d) Same as panel C for a cohort of 17,462 tumors sequenced by targeted panels and gathered by the GENIE project. Tumors were grouped according to the most specific disease subtype available in the patient information. Cancer acronyms that are not included in the Figure 2 legend are detailed in the Suppl. Material.

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Weinstein, I. B. Cancer. Addiction to Oncogenes--the Achilles Heal of Cancer. *Science (80-.).* **297,** 63–64 (2002).

2. Futreal, P. A. *et al*. A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–83 (2004).

3. Tamborero, D. *et al*. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).

4. Rubio-Perez, C. *et al*. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27,** 382–396 (2015).

5. Lawrence, M. S. *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

6. Kandoth, C. *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9 (2013).

7. Landrum, M. J. *et al*. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862-8 (2015).

8. Dienstmann, R., Jang, I. S., Bot, B., Friend, S. & Guinney, J. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov.* **5,** 118–123 (2015).

9. Petitjean, A. *et al*. 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,** 622–629 (2007).

10. Ainscough, B. J. *et al*. DoCM: a database of curated mutations in cancer. *Nat Meth* **13,** 806–807 (2016).

11. Hudson, T. J. *et al*. International network of cancer genome projects. *Nature* **464,** 993–8 (2010).

12. Parsons, D. W. An integrated genomic analysis of human glioblastoma multiforme. *Science (80-.).* **321,** 1807–1812 (2008).

13. Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. in *Bioinformatics* **30,** (2014).

14. Martelotto, L. G. *et al*. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol.* **15,** 484 (2014).

15. Lek, M., Tewksbury, J. & Services, H. Analysis of protein-coding genetic variation in 60,706 humans. *Nat. Publ. Gr.* **536,** 1–26 (2014).

16. Kato, S. *et al*. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A* **100,** 8424–8429 (2003).

17. Dogruluk, T. *et al*. Identification of variant-specific functions of PIK3CA by rapid phenotyping of rare mutations. *Cancer Res.* **75,** 5341–5354 (2015).

18. Kim, E. *et al*. Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. *Cancer Discov.* **2641,** 617–632 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/140475; this version posted June 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Berger, A. H. *et al*. High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* (2015). doi:10.1016/j.ccell.2016.06.022

20. Global Alliance for Genomics and Health, T. G. A. for G. and *et al*. GENOMICS. A federated ecosystem for sharing genomic, clinical data. *Science* **352,** 1278–80 (2016).

21. AACR Project GENIE: Powering Precision Medicine Through An International Consortium. *Cancer Discov*. (2017).

22. Iorio, F. *et al*. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166,** 740–754 (2016).

23. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. **17,** 128 (2016).

24. Griffith, M. *et al*. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* **49,** 170–174 (2017).