D992–D999 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 29 October 2015*
doi: 10.1093/nar/gkv1123

NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings

Omer An, Giovanni M. Dall'Olio, Thanos P. Mourikis and Francesca D. Ciccarelli*

Division of Cancer Studies, King's College London, London SE11UL, UK

Received September 14, 2015; Revised October 12, 2015; Accepted October 14, 2015

ABSTRACT

The Network of Cancer Genes (NCG, http://ncg.kcl.ac.uk/) is a manually curated repository of cancer genes derived from the scientific literature. Due to the increasing amount of cancer genomic data, we have introduced a more robust procedure to extract cancer genes from published cancer mutational screenings and two curators independently reviewed each publication. NCG release 5.0 (August 2015) collects 1571 cancer genes from 175 published studies that describe 188 mutational screenings of 13 315 cancer samples from 49 cancer types and 24 primary sites. In addition to collecting cancer genes, NCG also provides information on the experimental validation that supports the role of these genes in cancer and annotates their properties (duplicability, evolutionary origin, expression profile, function and interactions with proteins and miRNAs).

INTRODUCTION

Cancer genome projects, including The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/) and the International Cancer Genome Project (ICGC, https://dcc.icgc.org/) have so far mapped DNA alterations in more than 13 000 cancer samples. These massive sequencing efforts show that somatic modifications vary greatly between and within cancer types (1–3). Only some of the acquired alterations, however, confer a selective advantage that promotes cancer development (*driver alterations*). The large majority of alterations have no or little role in cancer and are fixed in the cancer genome as a by-product of the selection acting on drivers (*passenger alterations*). One of the challenges of cancer genomics is to effectively distinguish between driver and passenger alterations in order to identify the molecular determinants of cancer. Most known driver alterations modify protein-coding genes (*cancer genes*). The ability to identify cancer genes among the wealth of mutated genes is crucial to better understand cancer biology and to empower the development of innovative anti-cancer therapy.

Network of Cancer Genes (NCG) is a database launched in 2010 with the aim to collect cancer genes from the literature. Curators constantly review cancer mutational screenings and annotate altered genes that either have well-established cancer functions (*known cancer genes*) or are putative cancer drivers (*candidate cancer genes*). Originally (4), NCG collected data from only five mutational screenings and annotated most known cancer genes from the Cancer Gene Census (CGC) (5). The last five years have seen the rapid accumulation of cancer genomic data from thousands of samples, with almost all human genes mutated in at least one sample (6,7). Due to this overwhelming amount of data and to avoid the inclusion of mutated genes with no role in cancer, in this release we have substantially reviewed the procedure to identify cancer genes. NCG now collects 1571 cancer genes, 518 of which are known cancer genes. The remaining 1053 genes are candidate cancer genes whose driver role has been predicted in the original publication using a variety of methods (Supplementary Table S1). Given the importance of a robust experimental support for the cancer activity of candidate cancer genes, NCG now collects additional literature describing available orthogonal validations. NCG also annotates various properties of cancer genes such as the presence of extra copies in the genome (gene duplicability), the evolutionary origin, the connectivity of the encoded proteins in the protein–protein and miRNA interaction networks, and the comprehensive gene expression profile across 38 human tissues and 1543 cancer cell lines.

The manual curation of the literature to extract cancer driver genes and the annotation of a large number of additional properties make NCG a comprehensive and updated resource to navigate the overwhelming amount of cancer data with a particular focus on the genetic determinants of cancer.

MANUAL ANNOTATION OF CANCER GENES

In this release of NCG, the procedure for the inclusion of cancer genes in NCG has been reviewed and standardized

*To whom correspondence should be addressed. Tel: +44 20 7848 6616; Fax: +44 20 7848 6220; Email: francesca.ciccarelli@kcl.ac.uk

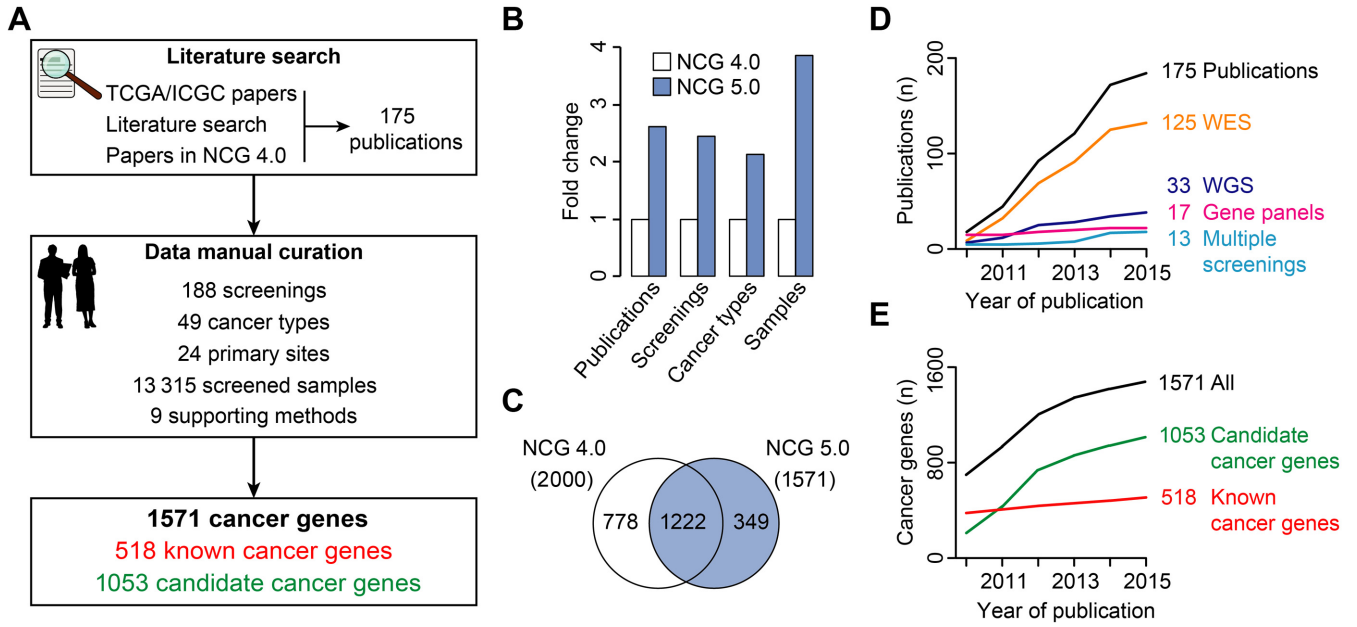

Figure 1. Curation procedure and comparison between NCG 5.0 and NCG 4.0: (**A**) Flowchart of the curation procedure used in NCG. After the identification of relevant publications describing cancer mutational screenings, two independent curators extract cancer genes and related information on types of screening and cancer, primary sites, screened samples and supporting methods. (**B**) Number of publications, screenings, cancer types and screened samples in NCG 5.0 as compared to NCG 4.0. (**C**) Venn diagram of cancer genes in NCG 4.0 and NCG 5.0. The reasons for the removal of 778 genes from the database are detailed in Supplementary Table S2. (**D–E**) Growth of NCG data in time. Shown are the number of publications, screenings and cancer genes starting from 2010, year of the first release of NCG. All screenings that were published prior of 2010, were collapsed.

(Figure 1A). The first difference with previous versions is to restrict the inclusion only to studies that describe mutational screenings of cancer samples and that distinguish between cancer genes and genes with passenger mutations. This led to the identification of 119 new publications. To be consistent with these inclusion criteria, all 68 studies present in the previous release were re-analysed. Twelve of them were excluded because they screened cancer cell lines rather than cancer samples or used no methods to identify cancer genes among all mutated genes. As a result of this extensive literature search, NCG 5.0 currently collects 175 studies (Supplementary Table S1). Two curators reviewed independently each publication to extract cancer genes and complementary information, such as the screening and the cancer types, the primary sites, the number of sequenced samples and the methods that were applied to identify cancer genes (Figure 1A). This manual curation resulted in 1260 cancer genes, 207 of which were annotated as known cancer genes in CGC. The remaining 1053 genes were candidate cancer genes identified in the original study using one or more methods (Supplementary Table S1). Additional known cancer genes were also added from CGC (February 2014), leading to a total of 1571 cancer genes. If information was available, cancer genes were further annotated as dominant (mostly oncogenes) or recessive (mostly tumour-suppressors) genes.

As compared to NCG 4.0 (8), NCG 5.0 now collects information from more than the double number of publications, screenings and cancer types and from four times more cancer samples (Figure 1B). Despite this substantial increase of data, the number of cancer genes decreased from 2000 to 1571 (Figure 1C), because of the more restrictive criteria. In particular, 612 genes were removed because the original publication was excluded and 166 genes because they had no support as cancer drivers (Supplementary Table S2). Overall, the studies in NCG 5.0 describe 188 mutational screenings, including 125 whole exome sequencings, 33 whole genome sequencings, 17 screenings of selected gene panels and 13 screenings based on multiple approaches (Figure 1D). Interestingly, the number of cancer genes with a well-documented role in cancer increases at a much slower pace as compared to candidate cancer genes (Figure 1E). This highlights the currently unmet need of efficient experimental assays that support the predicted role of candidate genes in cancer.

Almost all mutational screenings collected in NCG 5.0 applied only one method to identify cancer genes (Supplementary Table S1). The most common was the recurrence of mutation of a given gene across samples, which was taken as a sign of functional selection (Figure 2A and Supplementary Table S1). Other commonly used methods included MutSig (6) and MuSiC (9) (Figure 2A and Supplementary Table S1). Interestingly, the majority of known cancer genes (67%) had the support of at least two methods (Figure 2B), while most candidate cancer genes (78%) have been predicted by only one method (Figure 2C). In agreement with this, known cancer genes were overall identified as drivers across a higher number of mutational screenings and primary cancer sites as compared to candidate cancer genes (Figure 2D). The tendency of candidate cancer genes to be cancer specific was also reflected by the lower overlap between methods that support them as compared to those that support known cancer genes (Figure 2E). Cases where the overlap was higher (i.e. between MutSig and Invex, Figure

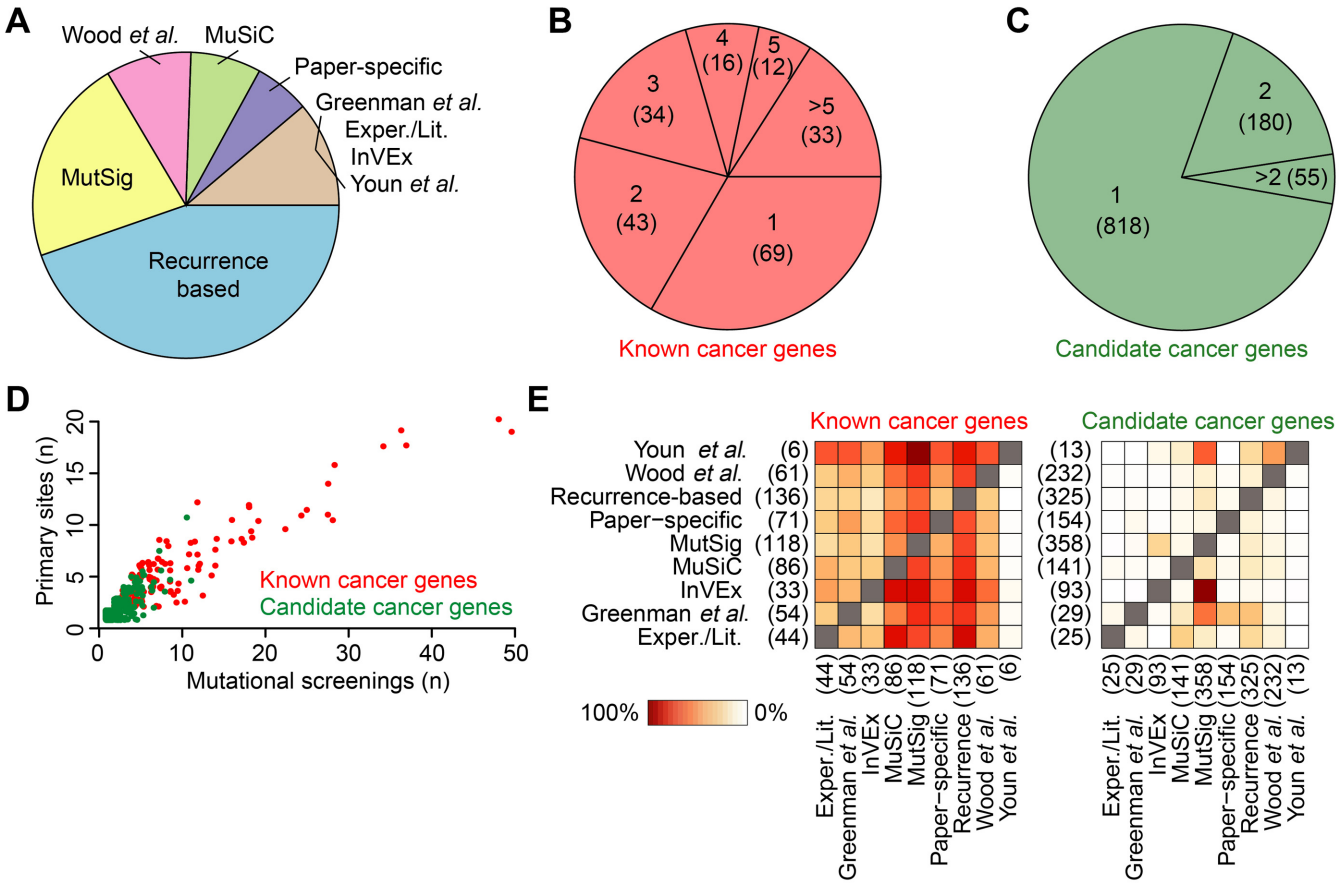

Figure 2. Overview of data in NCG 5.0: (**A**) Cancer mutational screenings divided according to the method that was applied to identify cancer genes in the original publication. Methods and corresponding screenings are described in Supplementary Table S1. (**B–C**) Fractions of known and candidate cancer genes supported by one or more methods. Gene counts are reported in brackets. (**D**) Number of mutational screenings and primary sites where each cancer gene has been reported as a driver. *TP53* is an outlier and has been excluded from the analysis because it has been identified in 113 screenings across 22 primary sites. (**E**) Heatmaps of the overlap between methods identifying known and candidate cancer genes. Each box represents the percentage of cancer genes identified with one method that are also supported by another. For each method, the total number of associated cancer genes is reported in brackets.

2E) corresponded to screenings where both methods were used (Supplementary Table S1).

EXPERIMENTAL VALIDATION OF CANDIDATE CANCER GENES

Candidate cancer genes that are identified using computational methods often lack additional experimental validation of their cancer driver role. The main reason is that functional follow-ups are often cumbersome and require *ad hoc* design for individual genes. The experimental proof of predicted driver role is however crucial for the translatability of potentially relevant discoveries into increased knowledge and novel treatments.

In this release of NCG, we have extensively reviewed the literature to search for experimental validations of candidate cancer genes. NCG now annotates available orthogonal experiments that have been performed in the original study or in follow-up studies for 120 out of 1053 candidate cancer genes (11% of the total, Table 1 and Supplementary Table S3). Most commonly used approaches measure the effect of gene silencing or gene overexpression in cell lines (Figure 3A and Supplementary Table S3) and the majority of candidate genes (83 out of 120) have been validated through multiple assays (Figure 3B).

An interesting case is *CSMD3,* the gene associated with benign adult familial myoclonic epilepsy (10) that encodes a long multi-repeat protein (Figure 3C). *CSMD3* has been found recurrently mutated across several cancer types and, therefore, has been predicted as a cancer driver by several methods (Figure 3D). Because of its length, sequence composition and location in proximity of fragile sites of the genome, *CSMD3* was regarded as a possible false positive in NCG 4.0. The fact that *CSMD3* is constitutionally not expressed in many tissues where it is mutated (Figure 3E) also supports the passenger role of the acquired mutations. Despite this, however, the stable knockout of *CSMD3* in immortalized epithelial cells has been reported to increase cell proliferation (11), thus suggesting a tumour-suppressor role for this gene. This example highlights the difficulty to correctly predict the driver role of mutated genes and the need of multiple independent pieces of evidence to assess the role of mutations in cancer.

Table 1. Experimental validation of candidate cancer genes

Experimental validation	Candidate cancer genes (n)	Publications (n)
Gene overexpression	60	74
Transient RNA interference	58	52
Mutagenesis	31	41
Immunostaining	25	26
Stable gene knockout	23	22
Survival analysis	20	21
Protein activity assay	19	20
Drug response assay	15	17
In silico protein modelling	12	14
Xenograft	10	11
Rhotekin pull-down	2	5
Total	275 (120 unique genes)	303 (166 unique publications)

For each type of experimental validation, the numbers of validated candidate genes and corresponding publications are shown. The complete gene list with references to the original papers is given in Supplementary Table S3.

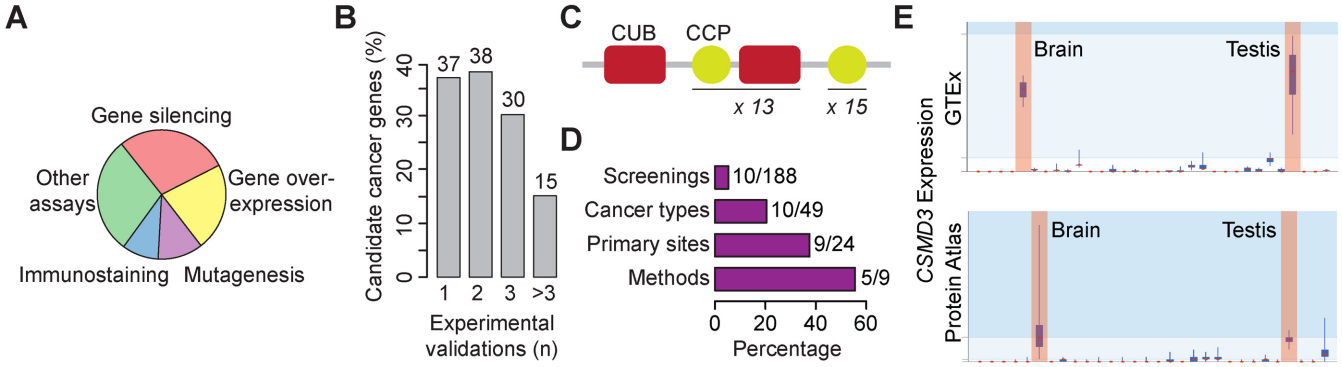

Figure 3. Validation of candidate cancer genes and alteration spectrum of *CSMD3*: (**A**) Fractions of validated candidate cancer genes according to the used experimental assay. Gene silencing refers to stable knockout or transient knockdown via RNA interference. Other assays include *in silico* protein modelling, survival analysis, drug response, protein activity, rhotekin pull-down and xenograft cancer models. (**B**) Percentage of candidate cancer genes that have been validated using one or more experimental approaches. The corresponding number of genes is shown above each bar. The full list of experiments and genes is reported in Supplementary Table S3. (**C**) Protein domain architecture of CSMD3 according to the SMART database (32). (**D**) Percentage of mutational screenings, cancer types, primary sites and methods that support the cancer driver role of *CSMD3*. Corresponding numbers are provided. (**E**) Expression profile of *CSMD3* in normal human tissues. Tissues where the gene is expressed in GTEx and Protein Atlas are highlighted in red.

ANNOTATION OF CANCER GENE PROPERTIES

To annotate the properties of cancer genes, original data on human genes, orthology, protein–protein and miRNA interactions and gene expression have been updated (Table 2).

Applying the previously described method (12), protein sequences from RefSeq v.63 (13) were aligned to the human genome assembly Hg19 to identify unique gene loci. These included 1525 of the 1571 cancer genes (13 cancer genes did not have RefSeq entries and 33 had no match in Hg19 or were gene isoforms). Cancer genes confirm their lower duplicability as compared to non-cancer genes and the signal derives from recessive cancer genes (P-value = 0.02, chi-square test, Table 2).

Orthology information from EggNOG v.4 (14) was used to trace the evolutionary origin of 1501 cancer genes, as described earlier (15). In line with previous reports (15–17), a higher fraction of cancer genes have orthologs in premetazoan species as compared to other human genes (P-value = 0.03, chi-square test, Table 2).

Four sources of primary interaction data (BioGRID v.3.4.125 (18); MIntAct v.190 (19); DIP (April 2015) (20); HPRD v.9 (21)) were integrated to rebuild the human protein–protein interaction network. This network included 1332 cancer proteins, which encode a higher fraction of hubs (defined as 25% most connected nodes of the network) as compared to other human proteins (P-value = 2.7×10^{-56}, chi-square test, Table 2). We verified that cancer genes encode a higher fraction of protein hubs also in the network derived from high-throughput screenings (P-value = 7.7×10^{-13}, chi-square test, Table 2). This excludes biases due to the higher number of single-gene experiments involving cancer proteins.

To complete the annotation of protein–protein interactions, NCG now collects also information on 752 cancer proteins involved in complexes as gathered from three resources (CORUM (February 2012) (22), HPRD v.9 (21), Reactome v.53 (23)). Supporting the signal from the overall protein–protein interaction network, a higher percentage of cancer proteins engage in complexes as compared to non-cancer proteins (P-value = 3.0×10^{-67}, chi-square test, Table 2).

Interactions between 324 miRNAs and 1101 cancer genes were derived from miRTarBase v.4.5 (24) and miRecords (April 2013) (25). Similarly to the protein–protein interaction network, also in the miRNA network a significantly larger fraction of cancer genes are target of miRNAs as

Table 2. Data and properties of cancer genes in NCG 5.0

Data sets in NCG 5.0		All cancer genes (1571)	Known cancer genes (518)		Candidate cancer genes (1053)	Other human genes
			Dominant (395)	Recessive (112)		
Human genes	All genes	1525	382	112	1020	17 489
	Duplicated genes (%)	280 (18%)	76 (20%)	12 (11%)	187 (18%)	3520 (20%)
Orthology	All genes	1501	379	110	1001	16 618
	Pre-metazoan genes (%)	992 (66%)	233 (61%)	80 (72%)	672 (67%)	10 516 (63%)
Protein–protein interactions	All nodes	1332	371	110	840	13 262
	Hubs (%)	558 (42%)	213 (57%)	78 (71%)	257 (31%)	2970 (22%)
	All nodes in HT network	1177	339	108	720	11 481
	Hubs in HT network (%)	386 (33%)	148 (44%)	52 (48%)	177 (25%)	2681 (23%)
Protein complexes	Proteins (%)	752 (49%)	238 (62%)	87 (78%)	418 (41%)	4917 (28%)
miRNA interactions	miRNA target genes (%)	1101 (72%)	332 (87%)	99 (88%)	662 (65%)	10 643 (61%)
	miRNAs	324	247	163	250	438
Expression in normal tissues	All genes in GTEx	1513	379	111	1012	16 818
	Ubiquitous genes (%)	965 (64%)	301 (79%)	98 (88%)	555 (55%)	11 077 (66%)
	Tissue-specific genes (%)	62 (4%)	5 (1%)	0 (0%)	57 (6%)	726 (4%)
	All genes in Protein Atlas	1517	378	112	1016	16 889
	Ubiquitous genes (%)	831 (55%)	278 (74%)	95 (85%)	447 (44%)	9492 (56%)
	Tissue-specific genes (%)	90 (6%)	11 (3%)	1 (1%)	78 (8%)	1042 (6%)
Expression in cancer cell lines	Cancer cell line encyclopedia	1426	367	106	942	15 158
	COSMIC Cancer Lines	1398	358	105	924	14 788
	Genentech data set	1524	381	112	1020	17 164

Of the 518 known cancer genes derived from CGC, 391 are annotated as dominant (mostly oncogenes), 108 as recessive (mostly tumour-suppressors), four as both as dominant and recessive and 15 have no specified mode of action. Duplicated genes have one or more duplicated loci in the genome covering ≥60% of their length (12). Pre-metazoan genes originated in the Last Universal Common Ancestor, Eukaryotes or Opisthokonts. Ubiquitously expressed genes are expressed in ≥95% tissues (29 tissues in GTEx and 30 tissues in Protein Atlas). HT = high throughput (publications reporting ≥100 interactions).

compared to other human genes (P-value = 3.0×10^{-18}, chi-square test, Table 2).

This release of NCG provides information on the expression of cancer genes in normal tissues and in cancer cell lines. For normal tissues, NCG relies on GTEx v.1.1.8 (26) and Protein Atlas (April 2015) (27), which both derive gene expression from RNASeq data in a total of 38 tissues. Expression values (FPKM for GTEx and RPKM for Protein Atlas) were used to derive expression categories (low, medium and high expression) for each gene and to calculate the distribution of gene expression across samples in each tissue. In both data sets, larger fractions of known cancer genes, but not of candidate cancer genes, are ubiquitously expressed (expression in >95% of all tissues) as compared to other genes (P-value = 1.3×10^{-13} and $P = 1.3 \times 10^{-19}$ for GTEx and Protein Atlas, respectively, chi-square test, Table 2). Conversely, significantly lower fractions of known cancer genes, but not of candidate cancer genes, are tissue specific (P-value = 4.2×10^{-4} and P-value = 6.9×10^{-4}, for GTEx and Protein Atlas, respectively, chi-square test, Table 2).

Three data sets (Cancer Cell Lines Encyclopedia (28), COSMIC Cancer Lines Project (29) and the recently released Genentech data set (30)) were used to derive gene expression in a total of 1543 cancer cell lines (Table 2). For each cancer gene, NCG provides the original expression value in each cell line as well as the normalized expression score, calculated as previously reported (31).

DATA ACCESS

NCG web interface has been reorganized, with particular focus on the summary of gene information and on the visualization of gene expression profiles. The gene summary now includes additional cross-references to external resources on protein domain architecture (32), drug and compound interactions (33,34) and protein druggability (35). For each cancer gene, the type of mutational screenings, the supporting methods and any experimental validation are detailed. Gene expression profiles are now shown as interactive graphs reporting the distribution of expression levels in each normal tissue and as summary tables in cancer cell lines.

NCG website provides overview statistics of the data contained in the database, including the list of 49 cancer types and corresponding 24 primary sites, the distribution of known and candidate cancer genes per primary sites, and information on 48 possible false positives. These include 14 genes derived from the literature (6), 4 additional genes that likely accumulate a high number of alterations due to their length and 30 olfactory receptor genes. All data contained in the database can be exported in batch using the advanced search option.

NCG USAGE

NCG offers a multi-level annotation of cancer genes that can be queried to gain insights on mutation status, properties, function and expression profiles of cancer genes (Figure 4A). This information facilitates the characterization of cancer genes and associated features. For example, gene duplicability has been exploited to extract duplicated tumour suppressor genes and to verify the occurrence of negative epistasis between them and their paralogs (36). Another useful feature of NCG is the comprehensive overview of gene expression profiles across a vast range of normal tissues and cancer cell lines. This can guide the selection of the most adequate cell systems for planning *in vitro* experiments (Figure 4B).

NCG is exploited widely as a repository of cancer genes (17,37–50). Examples include the use of NCG to test for the proximity of cancer genes to retrovirus insertion sites (48) and to evaluate the features of cancer classification methods (41). NCG also facilitates the interpretation of cancer mutational screenings by annotating the properties of mutated genes (Figure 4C) overall and in selected cancer types

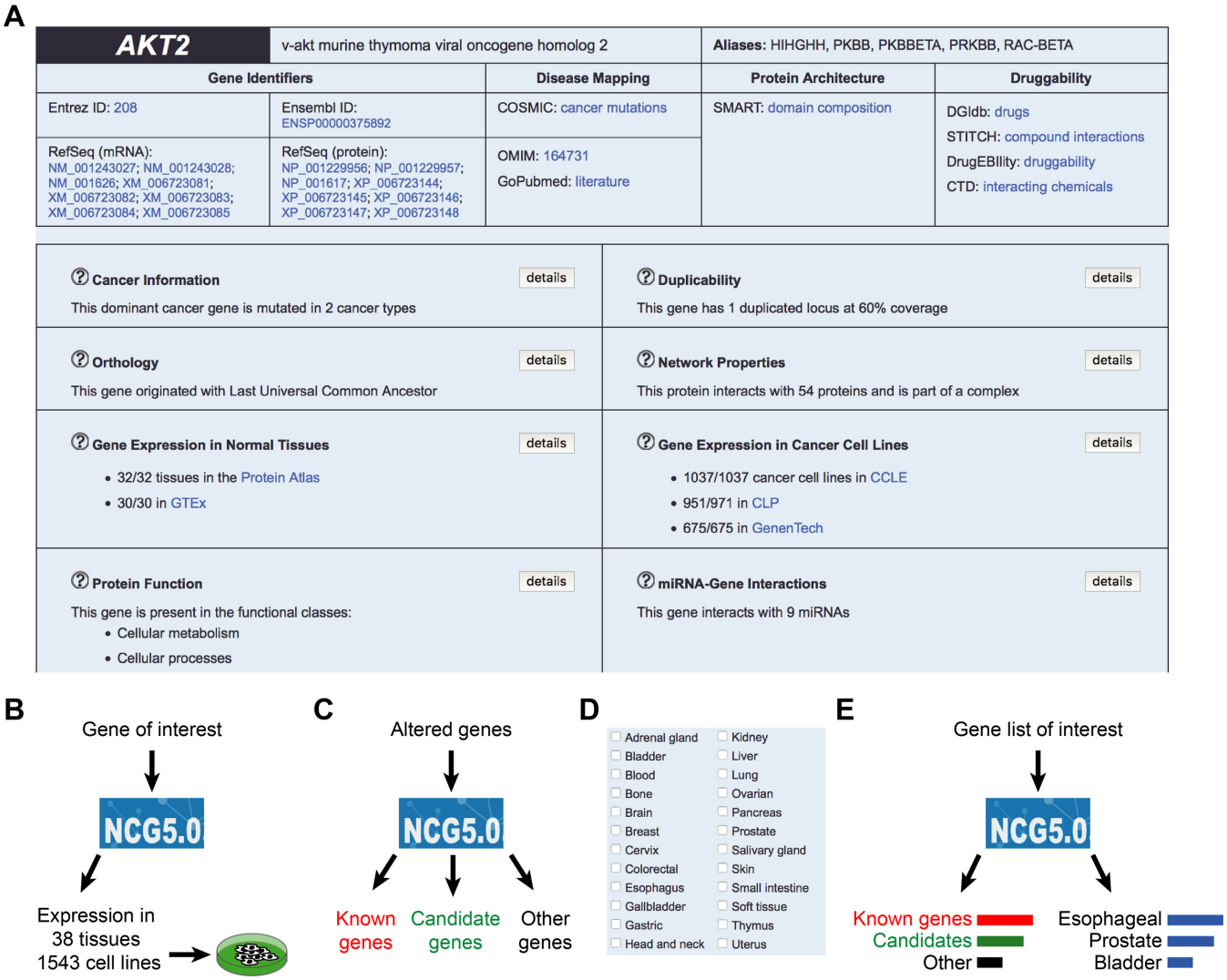

Figure 4. Examples of NCG usage: (**A**) Example of information available in NCG for a given cancer gene, in this case the oncogene *AKT2*. NCG summarizes the gene mutation profile across cancer types, information on duplicability, orthology, protein–protein and miRNA interactions and gene expression (**B**) NCG can facilitate the selection of the best cell systems for experimental assays by providing the expression profile of the gene of interest in several tissues and cell lines. (**C**) NCG can be used to annotate altered genes from mutational screenings. (**D**) The advanced search interface of NCG allows the identification of drivers in a variety of cancer types. (**E**) NCG can be integrated in gene enrichment analysis pipelines as a source of cancer genes.

(Figure 4D). For example, NCG has been used to verify whether genes undergoing copy number variations in familial breast cancer were already known cancer genes (49). Finally, NCG can be easily integrated into more complex analytical pipelines (Figure 4E). In the method developed by Zeller *et al.*, NCG provides a source of true cancer genes to prioritize drivers (50). In the DOSE bioconductor package, NCG is implemented as a source of cancer genes to perform enrichment analysis (51).

FUTURE WORK

It is expected that mutational screenings of cancer samples will continue to produce large amounts of data in the next years. The launch of personal genome initiatives ((52) and www.genomicsengland.co.uk) and the delivery of pan-cancer projects will substantially enlarge the spectrum of cancer types and samples with available mutational profiles. This will allow the discovery of novel cancer genes, particularly of those that recur in few samples and are currently difficult to identify. In parallel, the development of novel approaches for high-throughput functional screenings (e.g. based on the CRISPR-Cas technology (53–56)) promises to improve the efficiency of experimental validation assays.

In this exciting scenario, NCG will continue in its commitment to manually curate the literature to extract cancer genes and annotate available orthogonal supports. NCG will also expand to include other types of cancer driver alterations, such as copy number variations, gene rearrangements and non-coding modifications (57,58). In addition to enlarge the repertoire of cancer drivers, NCG will integrate new properties, e.g. the epigenetic regulation of cancer genes and their germline mutations.

As data become available, NCG will include the clinical relevance of cancer genes, such as their actionability

as pharmacological targets (59) and their applicability as biomarkers of cancer progression. All these efforts will contribute towards a more complete characterization of the molecular determinants of cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank Alex Mastrogiannopoulos for his help in the manual curation of the experimental validation of candidate cancer genes and all members of the Ciccarelli lab for providing suggestions to improve NCG.

FUNDING

European Union's Seventh Framework Programme [(FP7/2007-2013) under grant agreement No. 259743] (MODHEP consortium). The authors acknowledge support from the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.
Conflict of interest statement. None declared.

REFERENCES

1. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
2. Garraway,L.A. and Lander,E.S. (2013) Lessons from the cancer genome. *Cell*, **153**, 17–37.
3. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
4. Syed,A.S., D'Antonio,M. and Ciccarelli,F.D. (2010) Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Res.*, **38**, D670–D675.
5. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
6. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
7. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.
8. An,O., Pendino,V., D'Antonio,M., Ratti,E., Gentilini,M. and Ciccarelli,F.D. (2014) NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, **2014**, bau015.
9. Dees,N.D., Zhang,Q., Kandoth,C., Wendl,M.C., Schierding,W., Koboldt,D.C., Mooney,T.B., Callaway,M.B., Dooling,D., Mardis,E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
10. Shimizu,A., Asakawa,S., Sasaki,T., Yamazaki,S., Yamagata,H., Kudoh,J., Minoshima,S., Kondo,I. and Shimizu,N. (2003) A novel giant gene CSMD3 encoding a protein with CUB and sushi multiple domains: a candidate gene for benign adult familial myoclonic epilepsy on human chromosome 8q23.3-q24.1. *Biochem. Biophys. Res. Commun.*, **309**, 143–154.
11. Liu,P., Morrison,C., Wang,L., Xiong,D., Vedell,P., Cui,P., Hua,X., Ding,F., Lu,Y., James,M. *et al.* (2012) Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, **33**, 1270–1276.
12. Rambaldi,D., Giorgi,F.M., Capuani,F., Ciliberto,A. and Ciccarelli,F.D. (2008) Low duplicability and network fragility of cancer genes. *Trends Genet.*, **24**, 427–430.
13. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
14. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldon,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
15. D'Antonio,M. and Ciccarelli,F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.*, **7**, e1002029.
16. Domazet-Loso,T. and Tautz,D. (2008) An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.*, **25**, 2699–2707.
17. Domazet-Loso,T. and Tautz,D. (2010) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.*, **8**, 66.
18. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
19. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
20. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
21. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
22. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.W. (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, **38**, D497–D501.
23. Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, **4**, 1180–1211.
24. Hsu,S.D., Tseng,Y.T., Shrestha,S., Lin,Y.L., Khaleel,A., Chou,C.H., Chu,C.F., Huang,H.Y., Lin,C.M., Ho,S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.
25. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
26. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
27. Uhlén,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
28. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
29. Garnett,M.J., Edelman,E.J., Heidorn,S.J., Greenman,C.D., Dastur,A., Lau,K.W., Greninger,P., Thompson,I.R., Luo,X., Soares,J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
30. Klijn,C., Durinck,S., Stawiski,E.W., Haverty,P.M., Jiang,Z., Liu,H., Degenhardt,J., Mayba,O., Gnad,F., Liu,J. *et al.* (2015) A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.*, **33**, 306–312.

31. D'Antonio,M. and Ciccarelli,F.D. (2013) Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.*, **14**, R52.
32. Letunic,I., Doerks,T. and Bork,P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.
33. Griffith,M., Griffith,O.L., Coffman,A.C., Weible,J.V., McMichael,J.F., Spies,N.C., Koval,J., Das,I., Callaway,M.B., Eldred,J.M. *et al.* (2013) DGIdb: mining the druggable genome. *Nat. Methods*, **10**, 1209–1210.
34. Kuhn,M., Szklarczyk,D., Pletscher-Frankild,S., Blicher,T.H., von Mering,C., Jensen,L.J. and Bork,P. (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**, D401–D407.
35. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
36. D'Antonio,M., Guerra,R.F., Cereda,M., Marchesi,S., Montani,F., Nicassio,F., Di Fiore,P.P. and Ciccarelli,F.D. (2013) Recessive cancer genes engage in negative genetic interactions with their functional paralogs. *Cell Rep.*, **5**, 1519–1526.
37. Cheng,F., Jia,P., Wang,Q., Lin,C.C., Li,W.H. and Zhao,Z. (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol. Biol. Evol.*, **31**, 2156–2169.
38. Haemmerle,R., Phaltane,R., Rothe,M., Schroder,S., Schambach,A., Moritz,T. and Modlich,U. (2014) Clonal dominance with retroviral vector insertions near the ANGPT1 and ANGPT2 genes in a human xenotransplant mouse model. *Mol. Ther. Nucleic Acids*, **3**, e200.
39. Liu,W. and Xie,H. (2013) Predicting potential cancer genes by integrating network properties, sequence features and functional annotations. *Sci. China Life Sci.*, **56**, 751–757.
40. Liu,Y., Tian,F., Hu,Z. and DeLisi,C. (2015) Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers. *Sci. Rep.*, **5**, 10204.
41. List,M., Hauschild,A.C., Tan,Q., Kruse,T.A., Mollenhauer,J., Baumbach,J. and Batra,R. (2014) Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinform.*, **11**, 236.
42. Nayak,L., Tunga,H. and De,R.K. (2013) Disease co-morbidity and the human Wnt signaling pathway: a network-wise study. *OMICS*, **17**, 318–337.
43. Phaltane,R., Haemmerle,R., Rothe,M., Modlich,U. and Moritz,T. (2014) Efficiency and safety of O(6)-methylguanine DNA methyltransferase (MGMT(P140K))-mediated in vivo selection in a humanized mouse model. *Hum. Gene Ther.*, **25**, 144–155.
44. Saadatian,Z., Masotti,A., Nariman Saleh Fam,Z., Alipoor,B., Bastami,M. and Ghaedi,H. (2014) Single-nucleotide polymorphisms within micrornas sequences and their 3' UTR target sites may regulate gene expression in gastrointestinal tract cancers. *Iran Red Crescent Med. J.*, **16**, e16659.
45. Srivastava,A., Kumar,S. and Ramaswamy,R. (2014) Two-layer modular analysis of gene and protein networks in breast cancer. *BMC Syst. Biol.*, **8**, 81.
46. Shangguan,H., Tan,S.Y. and Zhang,J.R. (2015) Bioinformatics analysis of gene expression profiles in hepatocellular carcinoma. *Eur. Rev. Med. Pharmacol. Sci.*, **19**, 2054–2061.
47. Yu,H., Mitra,R., Yang,J., Li,Y. and Zhao,Z. (2014) Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. *Sci. China Life Sci.*, **57**, 1090–1102.
48. Olszko,M.E., Adair,J.E., Linde,I., Rae,D.T., Trobridge,P., Hocum,J.D., Rawlings,D.J., Kiem,H.P. and Trobridge,G.D. (2015) Foamy viral vector integration sites in SCID-repopulating cells after MGMTP140K-mediated in vivo selection. *Gene Ther.*, **22**, 591–595.
49. Masson,A.L., Talseth-Palmer,B.A., Evans,T.J., Grice,D.M., Hannan,G.N. and Scott,R.J. (2014) Expanding the genetic basis of copy number variation in familial breast cancer. *Hered. Cancer Clin. Pract.*, **12**, 15.
50. Zeller,M., Magnan,C.N., Patel,V.R., Rigor,P., Sender,L. and Baldi,P. (2014) A genomic analysis pipeline and its application to pediatric cancers. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **11**, 826–839.
51. Yu,G., Wang,L.G., Yan,G.R. and He,Q.Y. (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.
52. Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.
53. Shalem,O., Sanjana,N.E., Hartenian,E., Shi,X., Scott,D.A., Mikkelsen,T.S., Heckl,D., Ebert,B.L., Root,D.E., Doench,J.G. *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.
54. Gilbert,L.A., Horlbeck,M.A., Adamson,B., Villalta,J.E., Chen,Y., Whitehead,E.H., Guimaraes,C., Panning,B., Ploegh,H.L., Bassik,M.C. *et al.* (2014) Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, **159**, 647–661.
55. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
56. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
57. Borah,S., Xi,L., Zaug,A.J., Powell,N.M., Dancik,G.M., Cohen,S.B., Costello,J.C., Theodorescu,D. and Cech,T.R. (2015) Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer. *Science*, **347**, 1006–1010.
58. Vinagre,J., Almeida,A., Populo,H., Batista,R., Lyra,J., Pinto,V., Coelho,R., Celestino,R., Prazeres,H., Lima,L. *et al.* (2013) Frequency of TERT promoter mutations in human cancers. *Nat. Commun.*, **4**, 2185.
59. McGranahan,N. and Swanton,C. (2015) Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, **27**, 15–26.