Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

Database update

NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes

Omer An[1], Vera Pendino[1], Matteo D'Antonio[1], Emanuele Ratti[1], Marco Gentilini[1] and Francesca D. Ciccarelli[1,2,*]

[1]Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy and [2]Division of Cancer Studies, King's College London, London SE1 1UL, UK

***Corresponding author:** Tel: +44 (0)20 7848 6616; Fax: +44 (0)20 7848 6220; Email: francesca.ciccarelli@kcl.ac.uk

Submitted 29 November 2013; Revised 10 January 2014; Accepted 2 February 2014

Citation details: An,O., Pendino,V., D'Antonio,M., *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database* (2014) Vol. 2014: article ID bau015; doi:10.1093/database/bau015.

NCG 4.0 is the latest update of the Network of Cancer Genes, a web-based repository of systems-level properties of cancer genes. In its current version, the database collects information on 537 known (i.e. experimentally supported) and 1463 candidate (i.e. inferred using statistical methods) cancer genes. Candidate cancer genes derive from the manual revision of 67 original publications describing the mutational screening of 3460 human exomes and genomes in 23 different cancer types. For all 2000 cancer genes, duplicability, evolutionary origin, expression, functional annotation, interaction network with other human proteins and with microRNAs are reported. In addition to providing a substantial update of cancer-related information, NCG 4.0 also introduces two new features. The first is the annotation of possible false-positive cancer drivers, defined as candidate cancer genes inferred from large-scale screenings whose association with cancer is likely to be spurious. The second is the description of the systems-level properties of 64 human microRNAs that are causally involved in cancer progression (oncomiRs). Owing to the manual revision of all information, NCG 4.0 constitutes a complete and reliable resource on human coding and non-coding genes whose deregulation drives cancer onset and/or progression. NCG 4.0 can also be downloaded as a free application for Android smart phones.

Database URL: http://bio.ieo.eu/ncg/

Introduction

Sequencing of exomes and genomes from thousands of cancer samples led to the identification of an increasing number of mutated genes that may contribute to driving human cancer (1–3). Owing to the massive amount of information derived from these studies, it is often difficult to get an overview of the genes that play a driver role in cancer on mutation (cancer genes). Since 2010, the Network of Cancer Genes (NCG) has been collecting information on a manually curated list of known and candidate cancer genes (4, 5). Known cancer genes have robust experimental support on their role in cancer onset and progression. Candidate cancer genes instead derive from large-scale mutational screenings of cancer samples and have been identified using statistical methods with poor or no experimental follow-up. Candidate cancer genes are thus prone to include false positives as a consequence of the difficult discrimination between passenger and driver mutations (6, 7). To account for this, NCG 4.0 reports a list of candidate cancer genes whose association with cancer is likely to be spurious owing to function, length and literature evidence.

For each known and candidate cancer gene, NCG 4.0 annotates a series of systems-level properties, defined as features that distinguish a group of genes (in this case, cancer-related genes) from the rest, and that cannot be

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

ascribed to the function of the single gene alone (8). Systems-level properties currently reported in NCG are of evolutionary origin and duplicability, primary and secondary interaction network of the encoded proteins and miRNA regulatory networks. In addition, NCG 4.0 provides information on gene expression in 109 human tissues and on their functional characterization based on Gene Ontology (9). Owing to the increasing evidence of the primary role of microRNA (miRNA) deregulation in the onset of human cancer (10, 11), NCG 4.0 also annotates the systems-level properties of 64 cancer-related miRNAs (oncomiRs) manually derived from the literature.

Compared with other databases collecting all cancer mutations, such as COSMIC (12), ICGC (13) and CGAP (14), NCG 4.0 provides the community with a manually reviewed and constantly updated repository only of cancer drivers. In addition, it also annotates the properties of these genes, thus resulting useful to address different types of questions regarding cancer determinants (Figure 1) and to mine the increasing amount of information on cancer mutations.

Database Description and Updates

Manual collection of cancer genes

NCG 4.0 annotates the properties of 2000 cancer genes, defined as genes that contribute in promoting the onset and/or the development of human cancer. This list is derived from the union of two datasets. The first combined a literature-based repository of 484 genes from the Cancer Gene Census (377 dominant, 111 recessive and 4 genes that can act as both dominant and recessive, as frozen in January 2013) (15) with 77 genes whose amplification is causally implicated in cancer (16). This led to 537 experimentally supported cancer genes, which we defined as 'known cancer genes'. The second dataset consisted of 1463 genes that are likely to be involved in cancer development on mutation, which we defined as 'candidate

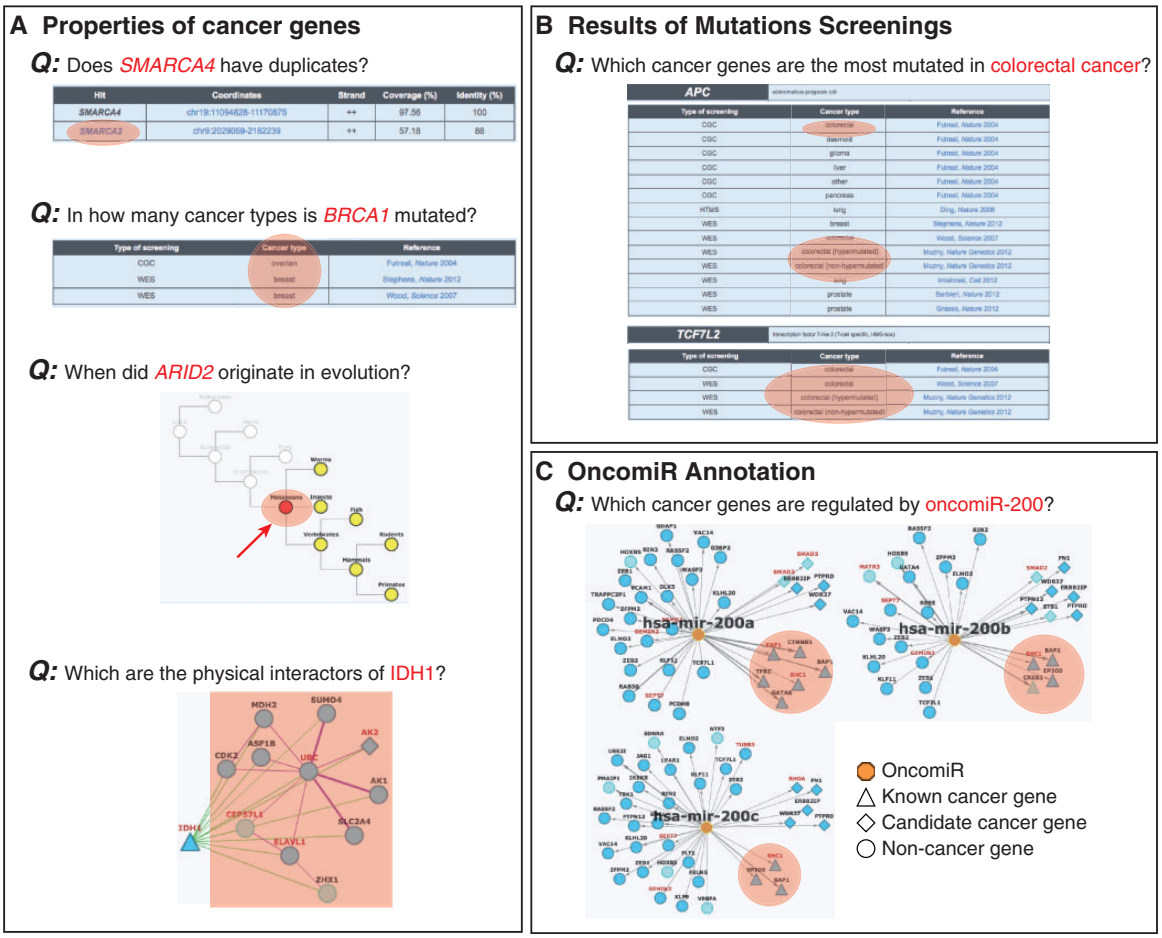

Figure 1. Examples of queries that can be done in NCG. Information stored in NCG can be used to address different queries regarding the properties of (**A**) individual cancer genes, (**B**) cancer types and (**C**) oncomiRs. Relevant information to address the specific queries is highlighted in orange.

cancer genes'. These genes derived from the manual revision of 67 publications corresponding to 77 re-sequencing screenings of the whole exomes (49 screenings), the whole genomes (19 screenings) and selected gene sets (9 screenings), conducted on 3640 samples from 23 cancer types (Supplementary Table S1) (17–83). These papers represented a comprehensive set of high-throughput cancer re-sequencing screenings.

Compared with the previous version, NCG 4.0 appreciably increased the number of cancer genes, particularly candidates, and of sequenced samples (Figure 2A). Such accretion of knowledge reflects the current massive worldwide efforts to characterize cancer mutational landscapes in detail. Although we are expected to reach a plateau in the discovery of new driver genes because genes frequently (and significantly) mutated in some cancer types are also mutated at low frequency in other cancer types (1), our data show that we are still in the growing phase. In particular, for most cancer types the number of new candidate cancer genes increases with the number of sequenced samples (Figure 2B). As already noticed (1, 6), most cancer genes, and in particular candidates, are specific for a given cancer type, and only few known cancer genes recurrently mutate in several cancers (Figure 2C). This observation once again confirms the heterogeneity of cancer mutation landscape (3).

Human gene set and orthology information

To identify the list of unique human genes, we aligned 33 427 protein sequences from RefSeq v.51 (84) to the reference human genome Hg19, using a method previously developed by our group (5, 8). This led to the identification of 19 045 unique gene loci, including 1961 of the 2000 cancer genes. Of the remaining 39 cancer genes, 29 did not have RefSeq protein entries and 10 were discarded because their protein sequences aligned to the genome for <60% of their length. For each cancer gene we retrieved

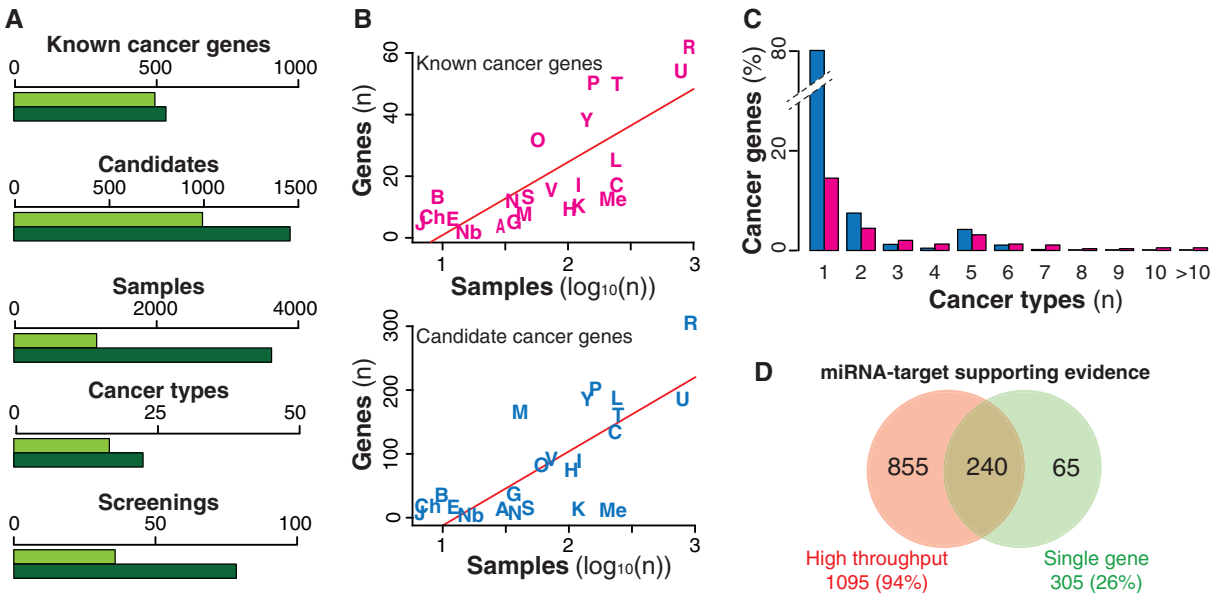

Figure 2. Overview of the data collected in NCG 4.0. (**A**) Comparison of data stored in NCG 3.0 and NCG 4.0. (**B**) Linear regression curves between the number of known and candidate cancer genes and the number of sequenced samples in each cancer type. Some cancer types deviate from linearity and this can be due to different reasons. For example, melanoma has a high number of candidate cancer genes (169) despite the low number of sequenced samples (41). In this case, the most likely explanation is that most of these candidate genes derive from two screenings (61, 75) that did not apply any methods to identify cancer drivers (Table 1, Supplementary Table S1). In the case of medulloblastoma, candidate and known cancer genes are only 25 despite 211 samples having been screened. This likely depends on the low mutation frequency of medulloblastoma [<1 mutation/Mb (40, 57, 64, 67)]. (**C**) Recurrence of known and candidate cancer genes in different cancer types. The only cancer genes that have been found mutated in more than 10 different cancer types are *TP53* (20 cancer types), *PIK3CA* (13 cancer types) and *PTEN* (12 cancer types). (**D**) Comparison of cancer miRNA targets that have been identified using single gene (i.e. reporter assay, western blot) and high throughput approaches (i.e. microarray, proteomic experiments and next-generation sequencing).

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

duplicability, evolutionary origin, functional annotation, gene expression profile, protein–protein interaction and gene-microRNA interaction.

We assessed gene duplicability by the presence of one or more additional hits on the genome covering at least 60% of cancer protein length (8). Of the 1961 cancer genes, 325 (17%) had at least one extra copy on the genome. This was a significantly lower fraction compared with the rest of human genes (21%, P-value = 7.8×10^{-06}, chi-square test), thus confirming the tendency of cancer genes to preserve a singleton status in the genome (8).

We assessed orthology relationships for 1978 of the 2000 cancer genes annotated in EggNOG v.3.0 (85) and used this information to infer the evolutionary origin of each cancer gene, defined as the most ancient node of the tree of life where the ortholog for that gene could be found (86). As already reported (86, 87), we confirmed that the fraction of old cancer genes that originated in prokaryotes and unicellular eukaryotes (1500, 76% of the total) was higher than in the rest of human genes (68%, P-value = 6.1×10^{-13}, chi-square test). Moreover, we also confirmed that recessive cancer genes are older than dominant cancer genes (4). The vast majority of recessive cancer genes (87/111, 78%) originated already with the last universal common ancestor or with unicellular eukaryotes, compared with only 67% of dominant cancer genes (P-value = 0.03, chi-square test).

Protein–protein and miRNA-target interaction networks

We rebuilt the human protein–protein interaction network integrating direct experimental evidence from five sources: HPRD (frozen on 13 April 2010) (88), BioGRID v.3.2.96 (89), IntAct v.159 (frozen on 14 December 2012) (90), MINT (frozen on 26 October 2012) (91) and DIP (frozen on 10 October 2010) (92). This resulted in a global network of 16 241 proteins (nodes) and 164 008 binary interactions (edges), supported by 33 497 independent publications. Interaction data were available for 1706 cancer proteins, and hubs (defined as proteins with at least 15 interactions) constituted 45% of all cancer genes, compared with 30% of the rest of human genes (P-value = 3.60 10^{-38}, chi-square test).

The interaction network between miRNAs and cancer genes relied on experimental data extracted from three different sources: TarBase v.5.0 (93), miRecords v.4.0 (94) and miRTarBase v.4.4 (95). The integration of these data led to 1160 cancer targets of miRNAs (58% of the total). This was a significantly higher proportion compared with the rest of human genes (48%, P-value = 1.02×10^{-17}, chi-square test) and confirmed the tendency of cancer genes to be regulated by miRNAs (4). This enrichment may reflect the fact that cancer genes are overall better characterized and thus more information is available on them. However, >70% of miRNA targets have been identified through high-throughput screenings (such as microarray, mass spectrometry and sequencing, Figure 2D), thus partially reducing the bias. Finally, we also updated the list of cancer genes that host miRNAs within their genomic loci (87 genes, 4.4% of the total).

Novel Features of NCG 4.0

Identification of possible false cancer genes

With the increasing evidence of an overwhelming number of mutations acquired during cancer progression (most of which with no role in the disease), a number of statistical methods have been developed to identify cancer drivers within the whole set of mutated genes. These methods take into account several features including the tendency of the same gene to be mutated across many samples, the cancer-specific background mutation rate, the gene length and expression and the mutation effect on the encoded protein (Table 1, Supplementary Table S1). Despite all efforts to refine the identification of driver mutations, current approaches are still prone to false positives, i.e. mutated genes that are erroneously identified as cancer drivers (6, 7). For example, genes encoding olfactory receptors are often included in the list of candidates, because they tend to mutate although the biological function and expression pattern of these genes strongly dismiss a possible functional role in the disease. Similarly, overly long genes are also probable false positives because their recurrent mutations in several samples are most likely due to their length more than to their function (6, 7). Because the main goal of NCG is to annotate the properties of cancer genes, we decided to collect all putative cancer genes from primary data without removing possible false positives. However, we added a warning concerning the possible spurious cancer associations for 60 genes (39 olfactory receptors, 14 genes with long exons and/or introns and 7 additional false positives derived from literature (7) (Figure 3A, Table 1). Although gene length by itself does not imply spurious associations, we derived the length distributions of all candidate cancer genes and considered genes with long introns (Figure 3B) or long exons (Figure 3C) as possible false positives.

Gene expression profiles

To complete the functional annotation of cancer genes, we derived expression levels for 1528 of them from two high-throughput gene expression experiments on 109 human tissues (99, 100). We normalized and processed the raw CEL files obtained from the corresponding Gene Expression Omnibus series (GSE2361 and GSE1133) using the MAS5 algorithm of the R *affy* package (101, 102). Because more than one probe can be associated with one gene, the expression level of each cancer gene in a given

Table 1. Methods used to identify candidate cancer genes and possible false positives

Method	MuSiC (96)	Mutsig (7)	Wood *et al.* (80, 97)	Greenman *et al.* (98)	Paper-specific	Recurrence-based	None
Candidate cancer genes	Genes that mutate with significantly higher rate than the background, considering multiple mutational mechanisms. It allows for pathway and proximity analysis, clinical correlation test and PFAM/OMIM query	Genes that mutate more often than expected, given the background mutation rate. It clusters mutations in hotspots and considers the functional impact and the conservation of the genomic site. The latest version takes into account patient and genomic mutation patterns	Genes that (a) mutate in both discovery and validation screens; (b) whose mutations exceed a certain threshold and; (c) mutate at a frequency higher than the passenger mutation rate	Genes that mutate at higher frequency than expected. Expectation is estimated using silent mutations	*Ad hoc* methodology developed for the specific set of samples and cancer type analyzed in the paper	Recurrence of mutations in a gene within samples is taken as evidence of its causal involvement in disease onset. Particularly used when few samples and/or cancer types with low mutation instability are analyzed	Often associated to whole genome screening, when only one or very few samples are sequenced. In such cases, all mutated genes are retained as possible candidates
Number of screenings	5	17	13	3	10	17	12
Possible false positives	*6 (LRP1B, OR6A2, OR11L1, OR5B17, OR10G7, RYR2)*	*17 (CNTNAP2, CSMD3, LRP1B, ORC4C15, OR8H2, OR8K1, OR6K3, OR5L2, OR2T33, PCLO, LRP2, MUC4, NEB, RYR2, SYNE1, SYNE2, TTN)*	*9 (CCDC168, CNTNAP2, CSMD3, EYS, LRP2, MUC16, OR2L13, OR51E1, TTN)*	*None*	*15 (CSMD3, CNTNAP2, OR4L1, OR10G9, OR5L1, OR4K14, OR4C13, OR4C6, OR51L2, OR1M1, OR2A42, OR10AG1, OR2A2, OR4K1, OR52E8)*	*13 (CNTN5, CSMD3, DMD, LRP1B, FSIP2, OR2M4, OR10R2, OR1L8, OR4C46, PCLO, RYR2, SYNE1, TTN)*	*17 (CTNNA3, CNTN5, CSMD1, OR2T11, OR2T34, OR5M5P, OR4S2, OBSCN, OR1J2, OR4D11, OR5H2, OR4Q3, OR4N5, OR52A5, RYR3, SYNE2, TTN)*

For each method used to identify candidate cancer genes (i.e. new possible cancer drivers) in the 77 screenings, reported are a brief description of the procedure, the number of screenings that relied on it and the associated possible false positives.

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

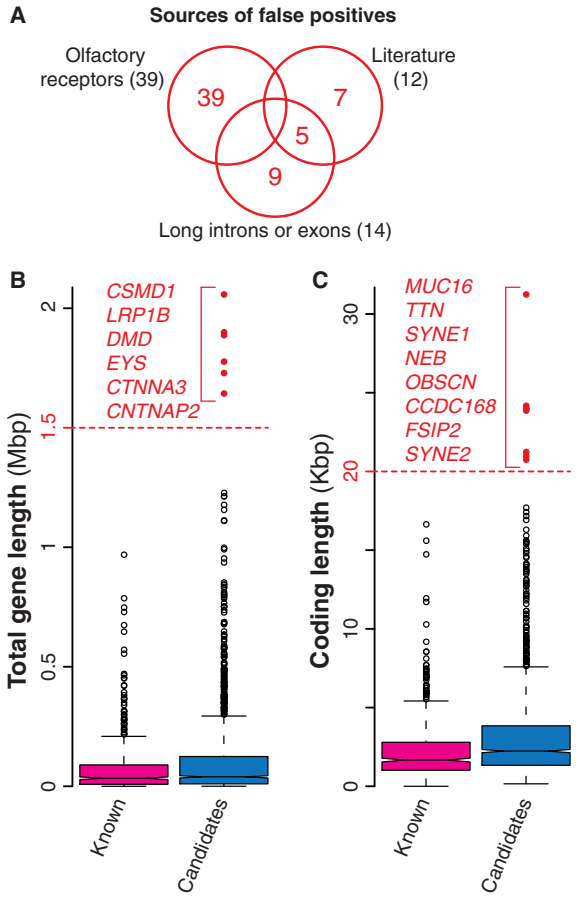

Figure 3. Possible false positives among candidate cancer drivers. (**A**) Venn diagram of the three groups of possible false positives. In total, we identified 60 genes, 65% of which were olfactory receptors, 23% were long genes and the remaining 20% were derived from literature (7). (**B**) Distribution of the total length for known and candidate cancer genes. Total gene length was measured as total number of nucleotides spanning the entire gene locus, including exons and introns. Red dots indicate possible false positives (gene longer than 1.5 Mb). (**C**) Length distribution of the coding regions for known and candidate cancer genes computed as the number of nucleotides covering the coding exons. Genes longer than 20 Kb (red dots) were considered as possible false positives.

tissue was defined as the mean expression levels of all probes with detection $P < 0.05$. If all probes for a gene had detection $P > 0.05$, the gene was considered as not expressed.

To make a comparative assessment of the expression levels of a cancer gene i in a given tissue t with those of all other genes in the same tissue, we first calculated the expression levels of all human genes in that tissue. We then derived the normalized expression level n of the cancer gene i in the tissue t, measured as:

$$n_{i,t} = \frac{(e_{i,t} - E_t)}{(e_{i,t} + E_t)}$$

where $e_{i,t}$ was the expression level of the cancer gene i in tissue t and E_t was the median expression level of all genes in tissue t. Normalized expression levels allowed a direct comparison of the expression of all genes in each given tissue.

Manual collection of miRNAs involved in human cancer (oncomiRs)

We manually gathered the list of oncomiRs from the literature and included only miRNA families (i.e. miRNAs with high sequence similarity) and miRNA clusters (i.e. miRNAs that are neighbors in the genome and co-transcribed) whose role in cancer was well described and experimentally supported (103–108). This led to 64 oncomiRs involved in 27 cancer types. Similarly to protein-coding genes we retrieved details on duplicability, evolutionary origin and interaction network for all these oncomiRs.

To infer oncomiR duplicability, we downloaded 1424 human miRNAs from miRBase v.17 (109) and considered all mature miRNAs with the same seed (i.e. the 6–8 nt-long region at the 5′-end of the sequence) as duplicated miRNAs. The rationale for this choice was that, because seeds determine the specificity in target recognition, their sequences are the most conserved among homologous miRNAs (110). Among 64 oncomiRs, 51 (79%) were duplicated compared with 33% other duplicated human miRNAs ($P = 4.5 \times 10^{-16}$, chi-square test). Therefore, unlike protein-coding cancer genes that maintain a singleton status in the genome, oncomiRs tend to have additional copies that share the site of recognition of the RNA targets.

To pinpoint when oncomiRs appeared in evolution, we developed a procedure similar to that used for protein-coding genes and traced the most ancient miRNA ortholog. We first retrieved the orthologs of 835 human miRNAs for which miRNA families were available in miRBase (including all 64 oncomiRs). We then assigned the origin of each miRNA as the most ancient ortholog within the corresponding family. Sixty oncomiRs (94% of the total) had orthologs in vertebrates, compared with only 19% of the rest of human miRNAs, thus suggesting that oncomiRs originated earlier than the rest of human miRNAs. It is worth noticing that the marked differences in duplicability and origin between oncomiRs and other human miRNAs are at least partly inflated by the high interest in oncomiRs that boosted the search of their paralogs and orthologs in other species.

Web Interface, Implementation and Data Availability

NCG 4.0 runs on an Apache web server and data are stored in a MySQL database. The web interface was developed in

PHP and network visualization was implemented in Cytoscape Web (http://cytoscapeweb.cytoscape.org/) (111).

We modified NCG 4.0 web interface to enhance functionalities and facilitate the retrieval of the properties of cancer genes and oncomiRs. In addition to searching for single genes or list of genes of interest, the user can now visualize and browse all 2000 cancer genes, as well as retrieve cancer genes based on specific filters. NCG 4.0 also provides a detailed report on the cancer types and the corresponding publications where it was found mutated. Similar types of searches can be done on the 64 oncomiRs.

All data stored in NCG 4.0 are summarized in the statistics section that provides an overview on the properties of cancer genes and oncomiRs. For example, it is possible to compare mutation frequency, number of cancer genes and oncomiRs as well as their recurrence across the different cancer types and screenings. The bulk content of the database as well as the list of cancer genes, false positives and oncomiRs can be downloaded as text files. We developed a mobile phone application for NCG 4.0 that is freely available from the Web site.

Supplementary Data

Supplementary data are available at *Database* online.

Acknowledgements

The authors thank all members of the Ciccarelli laboratory for testing the database and providing useful suggestions to improve it, and Alessandro Ogier for his help in implementing the web interface and the mobile application.

Funding

Associazione Italiana Ricerca sul Cancro [AIRC-IG 12742] and Italian Ministry of Health [Grant Giovani Ricercatori 2010] to F.D.C. Funding for open access charge: Associazione Italiana Ricerca sul Cancro [AIRC-IG 12742].

Conflict of interest. None declared.

References

1. Garraway,L.A. and Lander,E.S. (2013) Lessons from the cancer genome. *Cell.*, **153**, 17–37.
2. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature.*, **458**, 719–724.
3. Vogelstein,B., Papadopoulos,N., Velculescu,V.E. *et al*. (2013) Cancer genome landscapes. *Science.*, **339**, 1546–1558.
4. D'Antonio,M., Pendino,V., Sinha,S. *et al*. (2012) Network of cancer genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res.*, **40**, D978–D983.
5. 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.
6. D'Antonio,M. and Ciccarelli,F.D. (2013) Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.*, **14**, R52.
7. Lawrence,M.S., Stojanov,P., Polak,P. *et al*. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.*, **499**, 214–218.
8. Rambaldi,D., Giorgi,F.M., Capuani,F. *et al*. (2008) Low duplicability and network fragility of cancer genes. *Trends Genet.*, **24**, 427–430.
9. The Gene Ontology Consortium, (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
10. Calin,G.A. and Croce,C.M. (2006) MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res.*, **66**, 7390–7394.
11. Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, **10**, 704–714.
12. Forbes,S.A., Bindal,N., Bamford,S. *et al*. (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
13. Zhang,J., Baran,J., Cros,A. *et al*. (2011) International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database.*, **2011**, bar026.
14. Riggins,G.J. and Strausberg,R.L. (2001) Genome and genetic resources from the cancer genome anatomy project. *Hum. Mol. Genet.*, **10**, 663–667.
15. Futreal,P.A., Coin,L., Marshall,M. *et al*. (2004) A census of human cancer genes. *Nat. Rev. Cancer.*, **4**, 177–183.
16. Santarius,T., Shipley,J., Brewer,D. *et al*. (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer.*, **10**, 59–64.
17. Agrawal,N., Frederick,M.J., Pickering,C.R. *et al*. (2011) Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science.*, **333**, 1154–1157.
18. Banerji,S., Cibulskis,K., Rangel-Escareno,C. *et al*. (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature.*, **486**, 405–409.
19. Barbieri,C.E., Baca,S.C., Lawrence,M.S. *et al*. (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.*, **44**, 685–689.
20. Barretina,J., Taylor,B.S., Banerji,S. *et al*. (2010) Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat. Genet.*, **42**, 715–721.
21. Berger,M.F., Hodis,E., Heffernan,T.P. *et al*. (2012) Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature.*, **485**, 502–506.
22. Berger,M.F., Lawrence,M.S., Demichelis,F. *et al*. (2011) The genomic complexity of primary human prostate cancer. *Nature.*, **470**, 214–220.
23. Bettegowda,C., Agrawal,N., Jiao,Y. *et al*. (2011) Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science.*, **333**, 1453–1455.
24. Biankin,A.V., Waddell,N., Kassahn,K.S. *et al*. (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature.*, **491**, 399–405.
25. Chapman,M.A., Lawrence,M.S., Keats,J.J. *et al*. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature.*, **471**, 467–472.

26. Clark,M.J., Homer,N., O'Connor,B.D. *et al*. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
27. Dalgliesh,G.L., Furge,K., Greenman,C. *et al*. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature.*, **463**, 360–363.
28. Ding,L., Ellis,M.J., Li,S. *et al*. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature.*, **464**, 999–1005.
29. Ding,L., Getz,G., Wheeler,D.A. *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature.*, **455**, 1069–1075.
30. Fujimoto,A., Totoki,Y., Abe,T. *et al*. (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.*, **44**, 760–764.
31. Grasso,C.S., Wu,Y.M., Robinson,D.R. *et al*. (2012) The mutational landscape of lethal castration-resistant prostate cancer. *Nature.*, **487**, 239–243.
32. Greif,P.A., Eck,S.H., Konstandin,N.P. *et al*. (2011) Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia.*, **25**, 821–827.
33. Gui,Y., Guo,G., Huang,Y. *et al*. (2011) Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat. Genet.*, **43**, 875–878.
34. Guichard,C., Amaddeo,G., Imbeaud,S. *et al*. (2012) Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.*, **44**, 694–698.
35. Guo,G., Gui,Y., Gao,S. *et al*. (2012) Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat. Genet.*, **44**, 17–19.
36. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature.*, **489**, 519–525.
37. Huang,J., Deng,Q., Wang,Q. *et al*. (2012) Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat. Genet.*, **44**, 1117–1121.
38. Imielinski,M., Berger,A.H., Hammerman,P.S. *et al*. (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell.*, **150**, 1107–1120.
39. Jiao,Y., Shi,C., Edil,B.H. *et al*. (2011) DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science.*, **331**, 1199–1203.
40. Jones,D.T., Jager,N., Kool,M. *et al*. (2012) Dissecting the genomic complexity underlying medulloblastoma. *Nature.*, **488**, 100–105.
41. Jones,S., Zhang,X., Parsons,D.W. *et al*. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.*, **321**, 1801–1806.
42. Kan,Z., Jaiswal,B.S., Stinson,J. *et al*. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature.*, **466**, 869–873.
43. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature.*, **490**, 61–70.
44. Le Gallo,M., O'Hara,A.J., Rudd,M.L. *et al*. (2012) Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat. Genet.*, **44**, 1310–1315.
45. Lee,W., Jiang,Z., Liu,J. *et al*. (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature.*, **465**, 473–477.
46. Ley,T.J., Mardis,E.R., Ding,L. *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature.*, **456**, 66–72.
47. Li,M., Zhao,H., Zhang,X. *et al*. (2011) Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.*, **43**, 828–829.
48. Lilljebjorn,H., Rissler,M., Lassen,C. *et al*. (2012) Whole-exome sequencing of pediatric acute lymphoblastic leukemia. *Leukemia.*, **26**, 1602–1607.
49. Lohr,J.G., Stojanov,P., Lawrence,M.S. *et al*. (2011) Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA.*, **109**, 3879–3884.
50. Love,C., Sun,Z., Jima,D. *et al*. (2012) The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.*, **44**, 1321–1325.
51. Mardis,E.R., Ding,L., Dooling,D.J. *et al*. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.*, **361**, 1058–1066.
52. Morin,R.D., Mendez-Lago,M., Mungall,A.J. *et al*. (2011) Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature.*, **476**, 298–303.
53. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature.*, **487**, 330–337.
54. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.*, **455**, 1061–1068.
55. Ong,C.K., Subimerb,C., Pairojkul,C. *et al*. (2012) Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nat. Genet.*, **44**, 690–693.
56. Parsons,D.W., Jones,S., Zhang,X. *et al*. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science.*, **321**, 1807–1812.
57. Parsons,D.W., Li,M., Zhang,X. *et al*. (2010) The genetic landscape of the childhood cancer medulloblastoma. *Science.*, **331**, 435–439.
58. Pasqualucci,L., Trifonov,V., Fabbri,G. *et al*. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.*, **43**, 830–837.
59. Peifer,M., Fernandez-Cuesta,L., Sos,M.L. *et al*. (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.*, **44**, 1104–1110.
60. Piazza,R., Valletta,S., Winkelmann,N. *et al*. (2013) Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat. Genet.*, **45**, 18–24.
61. Pleasance,E.D., Cheetham,R.K., Stephens,P.J. *et al*. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature.*, **463**, 191–196.
62. Pleasance,E.D., Stephens,P.J., O'Meara,S. *et al*. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature.*, **463**, 184–190.
63. Puente,X.S., Pinyol,M., Quesada,V. *et al*. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature.*, **475**, 101–105.
64. Pugh,T.J., Weeraratne,S.D., Archer,T.C. *et al*. (2012) Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature.*, **488**, 106–110.
65. Quesada,V., Conde,L., Villamor,N. *et al*. (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 47–52.

66. Richter,J., Schlesner,M., Hoffmann,S. *et al*. (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.*, **44**, 1316–1320.
67. Robinson,G., Parker,M., Kranenburg,T.A. *et al*. (2012) Novel mutations target distinct subgroups of medulloblastoma. *Nature.*, **488**, 43–48.
68. Rudin,C.M., Durinck,S., Stawiski,E.W. *et al*. (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.*, **44**, 1111–1116.
69. Sausen,M., Leary,R.J., Jones,S. *et al*. (2012) Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat. Genet.*, **45**, 12–17.
70. Schwartzentruber,J., Korshunov,A., Liu,X.Y. *et al*. (2012) Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature.*, **482**, 226–231.
71. Shah,S.P., Morin,R.D., Khattra,J. *et al*. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature.*, **461**, 809–813.
72. Stephens,P.J., Tarpey,P.S., Davies,H. *et al*. (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature.*, **486**, 400–404.
73. Stransky,N., Egloff,A.M., Tward,A.D. *et al*. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science.*, **333**, 1157–1160.
74. Totoki,Y., Tatsuno,K., Yamamoto,S. *et al*. (2011) High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.*, **43**, 464–469.
75. Turajlic,S., Furney,S.J., Lambros,M.B. *et al*. (2012) Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Res.*, **22**, 196–207.
76. Varela,I., Tarpey,P., Raine,K. *et al*. (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature.*, **469**, 539–542.
77. Wang,K., Kan,J., Yuen,S.T. *et al*. (2011) Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat. Genet.*, **43**, 1219–1223.
78. Wang,L., Tsutsumi,S., Kawaguchi,T. *et al*. (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. *Genome Res.*, **22**, 208–219.
79. Wei,X., Walia,V., Lin,J.C. *et al*. (2011) Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.*, **43**, 442–446.
80. Wood,L.D., Parsons,D.W., Jones,S. *et al*. (2007) The genomic landscapes of human breast and colorectal cancers. *Science.*, **318**, 1108–1113.
81. Yan,X.J., Xu,J., Gu,Z.H. *et al*. (2011) Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat. Genet.*, **43**, 309–315.
82. Yoshida,K., Sanada,M., Shiraishi,Y. *et al*. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature.*, **478**, 64–69.
83. Zang,Z.J., Cutcutache,I., Poon,S.L. *et al*. (2012) Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.*, **44**, 570–574.
84. Pruitt,K.D., Tatusova,T., Brown,G.R. *et al*. (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
85. Powell,S., Szklarczyk,D., Trachana,K. *et al*. (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
86. D'Antonio,M. and Ciccarelli,F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.*, **7**, e1002029.
87. 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.
88. Keshava Prasad,T.S., Goel,R., Kandasamy,K. *et al*. (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
89. Chatr-Aryamontri,A., Breitkreutz,B.J., Heinicke,S. *et al*. (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
90. Kerrien,S., Aranda,B., Breuza,L. *et al*. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
91. Ceol,A., Chatr Aryamontri,A., Licata,L. *et al*. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
92. Salwinski,L., Miller,C.S., Smith,A.J. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
93. Papadopoulos,G.L., Reczko,M., Simossis,V.A. *et al*. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
94. Xiao,F., Zuo,Z., Cai,G. *et al*. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
95. Hsu,S.D., Lin,F.M., Wu,W.Y. *et al*. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
96. Dees,N.D., Zhang,Q., Kandoth,C. *et al*. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
97. Sjoblom,T., Jones,S., Wood,L.D. *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science.*, **314**, 268–274.
98. Greenman,C., Wooster,R., Futreal,P.A. *et al*. (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics.*, **173**, 2187–2198.
99. Ge,X., Yamamoto,S., Tsutsumi,S. *et al*. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics.*, **86**, 127–141.
100. Su,A.I., Wiltshire,T., Batalov,S. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA.*, **101**, 6062–6067.
101. Gautier,L., Cope,L., Bolstad,B.M. *et al*. (2004) affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics.*, **20**, 307–315.
102. Hubbell,E., Liu,W.M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics.*, **18**, 1585–1592.
103. Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer.*, **6**, 259–269.
104. Kent,O.A. and Mendell,J.T. (2006) A small piece in the cancer puzzle: microRNAs as tumor suppressors and oncogenes. *Oncogene.*, **25**, 6188–6196.
105. Lujambio,A. and Lowe,S.W. (2012) The microcosmos of cancer. *Nature.*, **482**, 347–355.

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

106. Manikandan,J., Aarthi,J.J., Kumar,S.D. *et al*. (2008) Oncomirs: the potential role of non-coding microRNAs in understanding cancer. *Bioinformation.*, **2**, 330–334.

107. Spizzo,R., Nicoloso,M.S., Croce,C.M. *et al*. (2009) Snapshot: microRNAs in cancer. *Cell.*, **137**, 586–586.

108. Yang,D., Sun,Y., Hu,L. *et al*. (2013) Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell.*, **23**, 186–199.

109. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

110. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.*, **116**, 281–297.

111. Lopes,C.T., Franz,M., Kazi,F. *et al*. (2010) Cytoscape web: an interactive web-based network browser. *Bioinformatics.*, **26**, 2347–2348.