Published in final edited form as:

Curr Opin Genet Dev. 2014 February ; 24: 92–98. doi:10.1016/j.gde.2013.12.003.

Collection, Integration and Analysis of Cancer Genomic Profiles: From Data to Insight

Jianjiong Gao, Giovanni Ciriello, Chris Sander, and Nikolaus Schultz

Computational Biology Center, Memorial Sloan-Kettering Cancer Center, Box 460, New York, NY 10065

Abstract

The recent deluge of cancer genomics data provides a tremendous opportunity for the discovery of detailed mechanisms of tumorigenesis and the development of therapeutics. However, identifying the functionally relevant genomic alterations ('drivers') among the many non-oncogenic events ('passengers') presents a major challenge. Several new methods have been developed over the past few years that identify recurrently altered genes. Mapping the recurrent genomic alterations, such as somatic mutations and focal DNA copy-number alterations, onto individual tumor samples as tumor-specific event calls facilitates the identification of altered processes and pathways. The resulting reduction in complexity makes cancer genomics data more easily interpretable by cancer researchers and is now driving the development of powerful yet intuitive web-based analysis tools.

Introduction

Large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and several efforts led by individual institutions, have recently generated an unprecedented amount of genomic data on tumor samples. [1-22] While early cancer genomics projects initially focused on array-based mRNA expression and then DNA copy-number data, most projects now employ some form of high-throughput sequencing, e.g., all RNA, whole exome, and/or whole genome sequencing. To date, these projects have explored somatic mutations in the coding regions of all genes in more than 15,000 tumors from more than 30 tumor types, and many have also generated detailed maps of DNA copy-number alterations, DNA methylation changes, and mRNA expression changes.

These efforts have led to the discovery of novel cancer genes, such as isocitrate dehydrogenase (IDH1) [23] and polymerase epsilon (POLE) [24], and they have elucidated the involvement of a number of biological processes and signaling pathways in tumor initiation and progression [25-28]. Many of these pathways tend to be altered in the majority of tumors, but the exact genomic mechanisms of dysregulation differ. As a result, we now

© 2013 Elsevier Ltd. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

have a better understanding of the heterogeneity of alterations within tumor types as well as an appreciation of similarities across tumor types [28].

However, many challenges remain. As the field is moving away from array-based technologies and Sanger sequencing, new software and algorithms for sequence analysis need to be developed (reviewed in [29] and [30]). While sequence alignment and mutation calling methods are evolving rapidly, their performance is further complicated by varying degrees of tumor heterogeneity, tumor purity and uneven sequence coverage.

This review covers the current state of downstream data collection, integration and analysis. One of the main challenges is to make these complex data easily accessible and interpretable. Experience from the last few years has shown that this is best achieved by first *distilling* the genomic data to a set of likely functional alteration events (mutations, copy-number changes, methylation events, significant over- and under-expression) (Fig. 1). These candidate functional events can be mapped onto individual tumor samples. The resulting simplified maps of genomic alterations (event maps) can be more easily used to identify commonly targeted pathways and to identify potential treatment options down to the level of a single patient (Fig. 1).

Repositories for cancer genomics data

There are several online databases that host cancer genomics data. For reasons of practicability and access control, normalized gene-level data and raw sequencing data are usually stored in separate repositories. All these data are freely available to the public, but access to raw sequence data requires authorization by the individual projects' data access committees. A full listing of all available resources that serve TCGA, the ICGC, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative, the Cancer Cell Line Encyclopedia (CCLE) and other projects is shown in Table 1.

Detecting recurrent genomic alterations to find cancer drivers

Part of the art in interpreting complex genomic data from tumor samples is to separate the signal from the noise, i.e. identify specific genomic alterations that contribute to the development and growth of a tumor (so-called *drivers*) within a background of a large number of alterations that do not confer a selective advantage for the tumor (*passengers*). Several methods have been developed for the identification of somatic mutations or DNA copy-number alterations that, across a set of tumors, occur at a higher rate than expected by chance (recurrent events).

The methods that identify recurrently mutated genes typically take into account factors such as the number and types of mutations in a gene, the length of the gene, the background mutation rate of a tumor and gene, DNA sequence conservation and recurrence at specific positions (hotspots). The most commonly used methods are MutSig [31], MUSIC [32], and InVex [33]. More recently, the functional impact of mutations, as predicted by tools such as SIFT [34], PolyPhen-2 [35], and MutationAssessor [36], has also been considered (OncodriveFM [37]) as well as the clustering of mutations along the protein sequence of a

gene (MUSIC [32] and OncodriveCLUST [38]). However, since these methods rely on recurrence, they cannot identify rare driver mutations. Some of these mutations may be common in certain cancer types, but others may be so rare that they cannot be detected by even the most sophisticated recurrence methods.

Recurrence-based methods have also been developed to identify genes that are altered by copy-number changes, e.g. GISTIC2.0 [39] and RAE [40]. These methods include amplitude and focality. Many of the recurrently altered regions (referred to as Regions of Interest, ROIs) contain no known oncogenes or tumor suppressors [41], and most contain multiple genes. Correlation with mRNA expression can be used to exclude from downstream analyses the genes that are not expressed or not sensitive to changes in DNA copy number. The impact of copy number changes on expression has been considered for driver genes in Oncodrive-CIS [42].

Similar methods can be applied to DNA methylation data to identify recurrently silenced genes, especially when coupled to mRNA expression data. Outlier expression analysis has been successfully applied to identify ETS family gene fusions in prostate cancer [43], but these methods are not yet commonly used to identify genes with unusual (e.g., bimodal) expression patterns in cancer genomics data sets. Expression data from RNA sequencing now makes it possible to detect fusion genes. Several software tools for fusion detection exist, like DeFuse [44], FusionSeq [45], or BreakFusion [46], but no consensus method has emerged yet. The role of aberrant splicing events can also be explored, e.g. by using JuncBASE [47].

Pathway analysis: Understanding oncogenic biological processes

Distilling the large number of genomic alterations in tumor samples down to recurrent or known oncogenic events greatly reduces the complexity of the data and thereby makes it easier to identify commonly altered signaling pathways and biological processes. While dozens of tools for the analysis of altered pathways have been developed in the last decade (reviewed in [48]), this review only covers those that use recurrent genomic alteration events.

The first generation of pathway analysis methods that use recurrent genomic events, such as HotNet [49], Netbox [50], Ingenuity Pathway analysis, and ontology enrichment approaches, do not take sample-specific information into account. However, the recurrent alteration events identified by multiple analysis platforms can be integrated and mapped onto individual tumor samples. Discrete genomic alteration events per tumor (e.g., mutated, amplified, deleted, over- or under-expressed) allow the exploration of co-occurrence and mutual exclusivity between events across a set of samples in a straightforward way, which has been shown to be a powerful way to identify commonly altered pathways, as first shown using the MEMo [51] and Dendrix [52,53] methods.

Figure 2 shows a step-by-step example of the identification of an altered signaling pathway (IGF2/PI3K signaling) via genomic alterations that were recurrent in the TCGA colorectal cancer data set. In this data set, IGF2 overexpression was identified in 20% of samples by a global search for genes with bimodal expression patterns. The MEMo algorithm then

identified the mutually exclusive pattern of IGF2 overexpression or alterations in the PI3K-signaling axis, suggesting that IGF2 activity is an alternative way of activating the PI3K pathway.

The next generation of tools and software packages will look across tissue boundaries and explore relationships of alterations in multiple different cancer types. Recurrent alteration events from multiple genomic platforms have recently been used to identify genomic subtypes in a dataset of twelve different cancer types from TCGA [28].

Interactive analysis and visualization tools: Interpreting complex data

Tools for the interactive analysis and visualization of multi-dimensional genomic data have come a long way in reducing the gap between complex genomic data and researchers without training in bioinformatics. Cancer researchers can now quickly visualize and explore complex data and test hypotheses. A full listing of these tools is presented in Table 2. Below, we highlight the most widely used non-commercial tools and explain how they were used in the example provided in Figure 2:

- The Broad Firehose is a pipeline that aggregates all TCGA data and performs many of the most commonly used analysis tools and algorithms automatically (e.g., GISTIC2.0 and MutSig). The results are made available via interactive reports on a user-friendly website (<http://gdac.broadinstitute.org>). In the example in Figure 2, all data was initially processed by the Firehose, and recurrently altered genes were identified via MutSig and GISTIC.
- COSMIC, the catalogue of somatic mutations in cancer [54], stores carefully curated somatic mutations and gene fusions from the literature and has become the main source of information about mutations in cancer. In the example in Figure 2, data from COSMIC was used to validate the recurrence of individual mutations in PIK3CA.
- The Integrative Genomics Viewer (IGV, **21221095**) and the UCSC Cancer Genomics Browser [55] can be used to visualize and explore copy-number, expression, methylation and mutation data, even in the context of clinical annotation. Both tools are genome centered, i.e. data is primarily visualized in the context of genomic position, although more recent additions allow the simultaneous visualization of multiple genes in separate genomic regions. In the example in Figure 2, IGV was used to visualize and validate the focality of the recurrent PTEN deletions.
- The cBioPortal for Cancer Genomics [56,57] integrates data sets such as mutation, copy number, gene expression, and clinical information from TCGA (via the Broad Firehose), ICGC, and other published data sets. The key simplifying concept of the cBioPortal is to use alteration events, like DNA mutations, gene amplifications, deletions, and gene over- or under-expression, and to map them onto individual samples. This concept enables the integration of multiple data types, queries for specific alteration events in samples, and downstream analyses such as survival analysis and pathway analysis. In the example in Figure 2, the cBioPortal was used

to visualize the bimodal expression pattern of IGF2, and to draw the OncoPrint that highlights the mutual exclusivity between the alterations.

Conclusions

The recent advances in nucleic acid sequencing have allowed the systematic analysis of the genomic alterations that are responsible for tumor initiation and progression. A key step was the realization that most tumors have alterations in a finite set of biological pathways and processes. The mechanisms by which these are altered tend to differ from tumor to tumor, but hundreds of events are recurrent, i.e. they can, if enough tumors are analyzed, be discovered in multiple tumor samples. With a primary focus on such recurrent events, we are developing a better understanding of the similarities and differences between tumor types and individual tumors, with explicit consequences for the development of personalized treatment.

However, there is an urgent need for further improvement of the analysis tools. To keep up with the increased throughput of DNA sequencers, novel analysis methods and software tools are being developed to address new opportunities. For example, an increase in sequencing coverage makes it possible to identify sub-clonal events, and small insertions and deletions may become detectable more reliably. A key challenge is to increase the ability to detect infrequent functional alterations (often referred to as the long tail). Simply sequencing more tumors will help to increase power to distinguish between functional and non-functional events, but this can be enhanced by improved methods that take into account, for example, protein sequence conservation, recurrence in adjacent positions in three-dimensional protein structure, as well as prior information about the known effect of a specific mutation.

Moreover, we are just beginning to develop the bioinformatics of systematically identify functionally relevant epigenetic alteration events or other genetic events ('cis' or 'trans') that change the expression of oncogenic genes. For example, promoter mutations were recently found to cause overexpression of the TERT gene, but no causative mechanism is currently known for genes such as IGF2 or AKT3, which are significantly overexpressed in subsets of various cancer types without DNA copy number alteration.

Improvements of such computational tools will ultimately benefit cancer patients, such that their treatment will be tailored to the specific combination of genomic alterations found in their tumor. The next several years will likely see the development of new algorithms and software tools that serve up genomic alteration data to oncologists, who will be able to use them alongside histological, pathological, and other clinical annotation to inform treatment decisions. An urgent challenge is the development of novel therapeutic targets for alterations that are currently not targetable, and to find ways to beat or prevent drug resistance that almost inevitably develops even after successful targeted therapy.

Acknowledgments

We thank Debra Bemis for critical reading and editing of the manuscript. This work was supported by the US National Cancer Institute as part of the TCGA Genome Data Analysis Center grant (NCI-U24CA143840), a Stand

Up To Cancer Dream Team Translational Research grant (a Program of the Entertainment Industry Foundation, grant number SU2C-AACR-DT0209), the National Resource for Network Biology (NIH National Center for Research Resources grant (NIH-GM103504) and the Starr Cancer Consortium (15-A500).

References

1. Ho AS, Kannan K, Roy DM, Morris LG, Ganly I, Katabi N, Ramaswami D, Walsh LA, Eng S, Huse JT, et al. The mutational landscape of adenoid cystic carcinoma. *Nat Genet.* 2013; 45:791–798. [PubMed: 23685749]
2. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, Payne-Turner D, Churchman M, Andersson A, Chen SC, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet.* 2013; 45:242–252. [PubMed: 23334668]
3. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature.* 2012; 486:395–399. [PubMed: 22495314]
4. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature.* 2012; 486:405–409. [PubMed: 22722202]
5. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature.* 2012; 486:400–404. [PubMed: 22722201]
6. Network TCGA. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012; 490:61–70. [PubMed: 23000897]
- 7*. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* 2012; 483:603–607. Detailed genomic and drug sensitivity profiling of 1,000 cancer cell lines. [PubMed: 22460905]
8. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* 2012; 483:570–575. [PubMed: 22460902]
9. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Bea S, Pinyol M, Martínez-Trillos A, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet.* 2012; 44:47–52. [PubMed: 22158541]
10. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, Chaudhuri S, Guan Y, Janakiraman V, Jaiswal BS, et al. Recurrent R-spondin fusions in colon cancer. *Nature.* 2012; 488:660–664. [PubMed: 22895193]
11. Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 2012; 487:330–337. [PubMed: 22810696]
12. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, Bandla S, Imamura Y, Schumacher SE, Shefler E, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet.* 2013; 45:478–486. [PubMed: 23525077]
13. Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, Jia W, Li Z, He M, Sun L, et al. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat Genet.* 2012; 44:17–19. [PubMed: 22138691]
14. Network TCGA. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* 2013; 368:2059–2074. [PubMed: 23634996]
15. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet.* 2012; 44:760–764. [PubMed: 22634756]
16. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell.* 2012; 150:1107–1120. [PubMed: 22980975]

17. Network TCGA. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012; 489:519–525. [PubMed: 22960745]
18. Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, Carneiro MO, Carter SL, Cibulskis K, Erlich RL, et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*. 2012; 488:106–110. [PubMed: 22820256]
19. Jones DT, Jager N, Kool M, Zichner T, Hutter B, Sultan M, Cho YJ, Pugh TJ, Hovestadt V, Stutz AM, et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*. 2012; 488:100–105. [PubMed: 22832583]
20. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, Phoenix TN, Hedlund E, Wei L, Zhu X, et al. Novel mutations target distinct subgroups of medulloblastoma. *Nature*. 2012; 488:43–48. [PubMed: 22722829]
21. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, Hamdi M, van Nes J, Westerman BA, van Arkel J, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*. 2012; 483:589–593. [PubMed: 22367537]
22. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, et al. The genetic landscape of high-risk neuroblastoma. *Nat Genet*. 2013; 45:279–284. [PubMed: 23334666]
23. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*. 2008; 321:1807–1812. [PubMed: 18772396]
- 24*. Network TCGA. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013; 497:67–73. Integrated genomic data was used to identify four distinct subtypes of endometrial carcinoma, with possible clinical implications. [PubMed: 23636398]
- 25*. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell*. 2013; 153:17–37. An extensive review of the recent findings of large-scale cancer genomic studies. [PubMed: 23540688]
26. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011; 144:646–674. [PubMed: 21376230]
27. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YS, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013; 45:1113–1120. [PubMed: 24071849]
- 28*. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*. 2013; 45:1127–1133. Tumor-specific alteration events were used to identify distinct genomic subtypes that span twelve different cancer types from the TCGA. [PubMed: 24071851]
29. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*. 2010; 11:685–696. [PubMed: 20847746]
30. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet*. 2013; 14:333–346. [PubMed: 23594911]
- 31*. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499:214–218. A systematic analysis of recurrently mutated genes and nucleotide change spectra across cancer types. [PubMed: 23770567]
32. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res*. 2012; 22:1589–1598. [PubMed: 22759861]
33. Hodis E, Watson IR, Kryukov GV, Arolt ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, et al. A landscape of driver mutations in melanoma. *Cell*. 2012; 150:251–263. [PubMed: 22817889]
34. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009; 4:1073–1081. [PubMed: 19561590]
35. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–249. [PubMed: 20354512]

36. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011; 39:e118. [PubMed: 21727090]
37. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* 2012; 40:e169. [PubMed: 22904074]
38. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics.* 2013; 29:2238–2244. [PubMed: 23884480]
39. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 2011; 12:R41. [PubMed: 21527027]
40. Taylor BS, Barretina J, Socci ND, Decarolis P, Ladanyi M, Meyerson M, Singer S, Sander C. Functional copy-number alterations in cancer. *PLoS One.* 2008; 3:e3179. [PubMed: 18784837]
- 41*. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet.* 2013; 45:1134–1140. An up-to-date description of recurrent copy-number alterations across cancer types. Many of these alterations contain no known driver genes. [PubMed: 24071852]
42. Tamborero D, Lopez-Bigas N, Gonzalez-Perez A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS One.* 2013; 8:e55489. [PubMed: 23408991]
43. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science.* 2005; 310:644–648. [PubMed: 16254181]
44. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol.* 2011; 7:e1001138. [PubMed: 21625565]
45. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.* 2010; 11:R104. [PubMed: 20964841]
46. Chen K, Wallis JW, Kandath C, Kalicki-Veizer JM, Mungall KL, Mungall AJ, Jones SJ, Marra MA, Ley TJ, Mardis ER, et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics.* 2012; 28:1923–1924. [PubMed: 22563071]
47. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR. Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res.* 2011; 21:193–202. [PubMed: 20921232]
- 48*. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012; 8:e1002375. A very detailed review of the different flavors of pathway analysis that have emerged over the past decade. [PubMed: 22383865]
49. Vandin F, Clay P, Upfal E, Raphael BJ. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput.* 2012:55–66. [PubMed: 22174262]
50. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One.* 2010; 5:e8918. [PubMed: 20169195]
51. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 2012; 22:398–406. [PubMed: 21908773]
52. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res.* 2012; 22:375–385. [PubMed: 21653252]
53. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol.* 2013; 9:e1003054. [PubMed: 23717195]
54. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 2011; 39:D945–950. [PubMed: 20952405]

55. Zhu J, Sanborn JZ, Benz S, Szeto C, Hsu F, Kuhn RM, Karolchik D, Archie J, Lenburg ME, Esserman LJ, et al. The UCSC Cancer Genomics Browser. *Nat Methods*. 2009; 6:239–240. [PubMed: 19333237]
56. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013; 6:pl1. [PubMed: 23550210]
57. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012; 2:401–404. [PubMed: 22588877]
58. Omberg L, Ellrott K, Yuan Y, Kandath C, Wong C, Kellen MR, Friend SH, Stuart J, Liang H, Margolin AA. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet*. 2013; 45:1121–1126. [PubMed: 24071850]
59. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods*. 2013
60. Perez-Llamas C, Lopez-Bigas N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One*. 2011; 6:e19541. [PubMed: 21602921]

Figure 1. Cancer genomics data processing and analysis: From raw data to biological insight

A key step towards extracting biological insights from complex cancer genomic data is the identification of the genomic alterations that contribute to tumorigenesis. The most likely candidates are the (statistically re-ranked) recurrent events, which, when mapped onto individual tumor samples, can be used to identify commonly altered pathways and potentially inform treatment decisions.

Figure 2. From recurrent events to altered pathways: PI3K signaling in colorectal cancer

This example demonstrates how activation of PI3K signalling by several different mechanisms, including overexpression of IGF2, was identified as a common pathway alteration in colorectal cancer. In the first step, recurrent alterations were identified separately in the mutation data (*MutSig* identified PIK3CA, among others, as frequently mutated), copy number data (*GISTIC* identified PTEN as focally deleted), and mRNA expression data (IGF2 is overexpressed in 20% of tumors). In the next step (middle panel), MEMo was used to identify recurrently altered gene sets, and a module consisting of IGF2, PIK3CA, PIK3R1, and PTEN was identified as the most significant result (only samples with an alteration in the four genes are shown). IGF2 is connected to PI3K signaling via IGF1R (bottom panel).

Table 1
Public repositories for cancer genomics data

cgHub https://cghub.ucsc.edu	Raw sequencing data for TCGA, the Cancer Cell Line Encyclopedia (CCLE) [7], and other projects.
dbGaP http://ncbi.nlm.nih.gov/gap	Raw sequencing data for TARGET and several other smaller sequencing studies.
EGA https://www.ebi.ac.uk/ega	Raw sequencing data for ICGC and other projects.
ArrayExpress http://www.ebi.ac.uk/arrayexpress	mRNA expression and DNA copy-number data. Used by several smaller cancer genomics studies.
GEO http://ncbi.nlm.nih.gov/geo	mRNA expression and DNA copy-number data. Used by several smaller cancer genomics studies.
CCLE http://broadinstitute.org/ccle/	Mutation, copy number and mRNA expression data for ~1000 cancer cell lines.
ICGC Data Portal http://dcc.icgc.org	Molecular profiling data and clinical information from participating projects. In addition to data from TCGA, it currently contains data from more than 1800 samples from 23 projects. The ICGC has set a goal of collecting and profiling more than 25,000 tumor samples from 50 projects.
TARGET Data Portal http://target.nci.nih.gov	Molecular profiling data and clinical information from TARGET projects, focusing on childhood cancers.
TCGA Data Portal https://tcga-data.nci.nih.gov	Molecular profiling data and clinical information generated by TCGA. It contains data for more than 8,000 samples from 30 tumor types (as of September 23, 2013). It is expected to grow by several thousand more samples by the end of 2014.
Synapse https://www.synapse.org	Curated TCGA data sets, including from the pan-cancer analysis project [58].
Broad GDAC Firehose http://gdac.broadinstitute.org	Aggregated and processed TCGA data sets, including automated standard analyses (recurrence, clustering, correlations, etc.). GDAC = Genome Data Analysis Center

Table 2
Web-based analysis resources for cancer genomics data

Broad Firehose http://gdac.broadinstitute.org	Aggregates all TCGA data and performs many of the most commonly used analysis tools and algorithms automatically (e.g., GISTIC2.0 and MutSig). The results are made available via interactive reports on a user-friendly website.
cBioPortal for Cancer Genomics http://cbiportal.org	Integrates data sets such as mutation, copy number, gene expression and clinical information from TCGA (via the Broad Firehose), ICGC, and other published data sets, maps alteration events back to tumor samples and make them queryable.
COSMIC http://cancer.sanger.ac.uk	Stores carefully curated somatic mutations and gene fusions from the literature and has become the main source of information about mutations in cancer.
ICGC Data Portal http://dcc.icgc.org	The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.
Integrative Genomics Viewer (IGV) http://broadinstitute.org/igv	Visualizes copy-number, expression, methylation and mutation data, even in the context of clinical annotation.
IntOGen http://intogen.org	IntOGen-mutations [59] can be used to query cancer drivers predicted by OncodriveFM [37] and OncodriveCLUST [38]. IntOGen-Arrays can be used to query gene expression and copy-number changes across tumor types. Datasets from IntOGen can be loaded directly into Gitoos [60] for visualization.
Oncomine http://oncomine.org	OncoMine Research Edition was primarily designed for the analysis of mRNA expression data in tumor samples. Additional features and the capability to analyze DNA copy-number changes and mutations are available in a Premium edition and in the Oncomine Gene Browser, which carry an annual license fee.
Regulome Explorer http://explorer.cancerregulome.org	Supports association analysis between genomic features from multiple data types, including continuous data like expression, discrete data like mutations, and also clinical attributes.
UCSC Cancer Genomics Browser https://genome-cancer.ucsc.edu	Visualizes copy-number, expression, methylation and mutation data, even in the context of clinical annotation for TCGA datasets.