Published in final edited form as:

Curr Protoc Mol Biol. ; 121: 19.14.1–19.14.13. doi:10.1002/cpmb.49.

Making use of cancer genomic databases

Chad J. Creighton^{1,2,3,4}

¹Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA.

²Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

³Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA.

⁴Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

Abstract

The vast amounts of genomic data now deposited in public repositories represent rich resources for cancer researchers. Large-scale genomics initiatives such as The Cancer Genome Atlas have made available data from multiple molecular profiling platforms (e.g. somatic mutation, RNA and protein expression, and DNA methylation) for the same set of over 10,000 human tumors. There has been much collective effort toward providing user-friendly software tools for biologists lacking computational skills to ask questions of large-scale genomic datasets. At the same time, there remains a clear need for skilled bioinformatics analysts to answer the types of questions that cannot easily be addressed using the public user-friendly software tools. This overview introduces the reader to the many resources available for working with cancer genomic databases.

Keywords

cancer genomics; cancer bioinformatics; databases; analysis software; The Cancer Genome Atlas (TCGA)

Introduction

With the rise of advanced technologies involving DNA sequencing and microarrays, the generation of high throughput molecular datasets (including data on mutations or gene and protein expression) has greatly outpaced our ability to analyze and digest the results within a single study. The vast amounts of genomic data being deposited in public database repositories, such as the Gene Expression Omnibus (GEO) and the NCI Genomic Data Commons (GDC), represent invaluable resources for investigators to re-analyze datasets to explore different questions from those posed in the original studies. In cancer research, large scale genomic datasets can be generated from human tumor specimens or from experimental model systems. While data from experimental models can help establish cause-and-effect relationships between genes and pathways, studies of human tumor specimens are also

needed in order to help ground experimental observations as being relevant in the setting of human patients. For researchers who do not have strong computational skills, a large number of freely available web interfaces can address a discrete set of questions based on high throughput molecular and genomic datasets. At the same time, there remains a clear need for skilled bioinformatics analysts, versatile in handling and querying genomic data in their various forms, to answer questions not easily addressed using the public user-friendly software tools.

This unit outlines resources for making use of publicly available data on cancer genomics. These resources involve a range of large scale datasets and analysis tools, and can span differing levels of computational expertise among users. The field of cancer bioinformatics is broadly described here, a field involving applied analysis work in addition to the development of new software tools and algorithms. The various levels of genomic data—from raw data to insights—are also described, along with how, in general, each level may be processed and analyzed. Some recent initiatives involving the generation of large-scale cancer genomics datasets, including The Cancer Genome Atlas (TCGA), are noted. Some cancer genomic databases that are commonly used by both molecular biologists and bioinformatics experts are also described. Finally, some pointers in getting less experienced researchers started on acquiring bioinformatics expertise are provided.

Cancer Bioinformatics

Bioinformatics is an interdisciplinary field that utilizes computer science and statistics in the analysis of large scale biomolecular datasets (Figure 1A). In the context of bioinformatics, computer science and statistics represent tools, while a good understanding of molecular biology provides the motivation for utilizing these tools. A scientist with strong skills in computer science or statistics, but having a weak level of knowledge of molecular biology, will be limited in carrying out bioinformatics research, at least without collaborating with a molecular biologist. Expertise in molecular biology, including an appreciation for cell biology and genetics, is critical to formulate the right questions to ask of available genomic datasets. For the subspecialty of cancer bioinformatics, a good understanding of cancer as a disease process is also needed.

Within the field of bioinformatics, there are different areas of specialty, and bioinformatics scientists are likely to be particularly strong in one area. Areas of specialty in bioinformatics include algorithm development, development of software and database tools, and analysis (Figure 1B). Algorithm development involves finding new analytical ways of answering questions or of drawing out new types of information from molecular datasets. Development of software and database tools could serve either “high powered” users (e.g. those with strong analytical and coding skills) or a wider audience of users that do not know programming but are able to use “point-and-click” interfaces, which represents a large portion of the molecular biology research community. Analysis applies existing tools and algorithms to draw insights from large molecular datasets. While much of the emphasis in bioinformatics research focuses on developing new algorithms or of developing new software tools for biologists, more attention might be given to the need for more scientists with expertise in meeting practical analysis needs for a given project, or “applied

bioinformatics.” As the techniques for generating large scale datasets have matured, there is a greater need for high level analyses to make sense of the results. While biologist-friendly software tools may provide some views of the data, there are limitations on the types of questions such tools can answer.

Levels of Genomic Data

We can think of genomic data as having several levels, ranging from raw data as initially generated by the instrument, to processed and normalized data ready for higher-level analyses, to results and findings that provide real insight into the biological system under study. Here, we follow a convention of delineating levels of genomic data according to designated Levels 1 through 5 (Figure 2A). This convention would be along the lines of what other large scale projects such as TCGA have followed, though other projects may elect to use somewhat different nomenclature and categorization from what is put forth here. Level 1 represents raw data files from the instrument; examples include fastq files from sequencing and image files from microarrays. Level 2 data involves essential information extracted from the raw data; examples include BAMs for sequencing data and probe-level intensity values for microarray data. Level 3 data are ready for platform-level analyses, such as a set of mutation variant calls for DNA sequencing or a normalized data table for gene expression profiling (by either RNA sequencing or microarrays). Level 4 data involve results from platform-centric analyses; for example, analysis of mutation variants for a set of cancer samples may involve identifying significantly mutated genes (Lawrence et al., 2014), and analysis of gene expression profiling data often involves identification of differentially expressed genes (Storey and Tibshirani, 2003). Level 5 involves taking things further from platform-specific results, which can involve integration of these results with results from other data platforms or with results from external datasets, as well using outside subject matter expertise (e.g. knowledge of molecular biology or of cancer) in the meaningful interpretation of the data.

For sufficiently mature genomic data platforms, automated software pipelines can be developed for many low-level analyses. However, there are inherent limitations on what analysis tasks can be fully automated, particularly regarding “Level 5” types of analyses (Figure 2B). Because they are relatively well defined regardless of project specifics, many tasks involving data Levels 1–4 lend themselves to automation, while Level 5 (involving data integration and domain expertise) typically requires more creativity and thought by analysts. Level 4 also involves a number of decision points (e.g. choice of analytical approach or use of statistical cutoffs), and so may also involve direct involvement by analysts.

There is also ample room for new algorithms for mining new types of information from raw data. Basic and routine service, stable pipelines, and application development tend to be involved in fully automated tasks, while advanced and more complex needs, rapid upgrades, new method development, and unique analytical approaches are more time consuming and project specific. As some analyses performed manually by analysts eventually become more automated, new questions and new approaches to the data can continue to be explored. Data integration, involving different data platforms and external datasets such as those from

TCGA, provides a wide universe of possibilities. Bioinformatics research is not a fully automated process, whereby all possible insights into the data would be obtainable from a few clicks.

There has been much collective effort to provide user-friendly software for biologists lacking computational skills to interrogate large scale genomic datasets. While such software tools have done great service in allowing a large segment of the research community some level of access to the data, there are inherent limitations of such tools in terms of the types of questions that can be addressed. There is an inherent trade-off with software analysis tools involving general usability versus flexibility (Figure 2C). Software tools considered highly usable by a wide audience (e.g. software with point-and-click interfaces) have limited flexibility, in the sense that they can answer only a select set of predetermined questions. Such biologist-friendly tools that have found a large audience include OncoPrint and cBioPortal (Table 1).

On the other hand, software tools offering the flexibility to answer a wide range of questions require a high degree of expertise, and therefore have a smaller audience of users. Software that is highly flexible and versatile includes the R package, which requires coding skills and a good understanding of statistics. Bioinformatics analysts should be highly flexible in their ability to answer questions using genomic datasets, beyond what less specialized scientists would be capable of. Invariably, when using biologist-friendly software tools with fixed and inflexible user interfaces, one will “hit a wall,” whereby a question is asked that is outside of what the software can perform; these situations represent opportunities for collaboration between molecular biologists and bioinformatics analysts.

Large-Scale Cancer Genomics Initiatives

Over the last decade, a number of large scale cancer genomics initiatives have generated rich genomic and molecular profiling datasets for use by the broader scientific community. A few of these initiatives profiling large numbers of human tumors—TCGA, ICGC, and PCAWG—are described below. Other large-scale cancer datasets include the Cancer Cell Line Encyclopedia (CCLE)(Barretina et al., 2012), the Genomics of Drug Sensitivity in Cancer (<http://www.cancerrxgene.org/>)(Garnett et al., 2012), The BROAD LINCS data set (<http://www.lincscloud.org/>), the Fantom Consortium data sets (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014), the ENCyclopedia Of DNA Elements (ENCODE) project (ENCODE_Project Consortium, 2012), and the Epigenome Roadmap (Roadmap_Epigenomics_Consortium et al., 2015).

The Cancer Genome Atlas (TCGA)

Beginning around 2006 and winding down as of 2017, TCGA was a large-scale, multi-institutional collaborative effort supported by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to systematically characterize the genomic changes that occur in cancer. For a common set of over 10,000 human tumors representing 32 different cancer types (Figure 3A), TCGA generated molecular profile data at the levels of gene expression, miRNA expression, protein expression (by Reverse Phase Protein Array or RPPA), DNA copy (by SNP array), DNA methylation, and somatic

mutation. For most of the cancer types, a TCGA-sponsored Analysis Working Group (AWG), typically consisting of a team of 20–30 scientists, extensively analyzed the molecular data for the given cancer type over a period of 1–3 years. The end result of the AWG's efforts was a publication, or “marker paper,” highlighting insights made into the molecular basis of the disease. In addition to studies focusing on a particular cancer type, “pan-cancer” molecular studies of TCGA data explore commonalities across cancer types as well as differences between cancer types (Cancer_Genome_Atlas_Research_Network et al., 2013). For example, pan-cancer genomic analysis of an initial set of 12 different cancer types in TCGA—combining expression data with DNA methylation and with DNA copy data (Hoadley et al., 2014)—found these cancers to segregate largely on the basis of either cancer type (as defined by tissue of origin) or of squamous histology.

The raw data generated by TCGA has been provided to the research community without publication restrictions, and represents an important resource for future studies. Independent studies subsequent to the TCGA-led efforts have used multiplatform-based molecular datasets from TCGA involving multiple cancer types (Figure 3B). For example, recent pan-cancer studies have respectively examined epithelial-mesenchymal transition (EMT) marker expression (Gibbons and Creighton, 2017) and examined PI3K/AKT/mTOR pathway alterations (Zhang et al., 2017) across all 32 cancer types in TCGA. Many web-based, biologist-friendly resources are available for downloading, visualizing, and exploring TCGA datasets, some of which are listed in Table 1 and described further below. Some notable advantages that TCGA datasets have over other publicly available cancer genomic datasets include: 1) very large sample size, 2) multiple data platforms for the same samples (e.g. protein expression and DNA methylation in addition to somatic mutations), and 3) unified data generation, low-level analysis, and calling across all samples and cancer types profiled. TCGA data generation formally ended around the year 2015, with the data now residing in the NCI Genomic Data Commons data portal (<https://gdc.cancer.gov/>).

International Cancer Genomics Consortium (ICGC)

In some ways, ICGC might be thought of as the international equivalent of TCGA, though large scale cancer genomics initiatives sponsored in the United States have some level of coordination with ICGC. ICGC was launched in 2008 to coordinate large-scale cancer genome studies in tumors from some 50 different cancer types, with research participants from all over the world (International_Cancer_Genome_Consortium et al., 2010). Working groups were created to develop strategies and policies that would form the basis for participation in the ICGC, in a number of different areas including a bioethical framework for generating and sharing genomic data, consistency of sample processing, guidelines about the use of common definitions and data standards, study design and statistical issues, and data release and intellectual property policies. The ICGC Data Portal (<https://dcc.icgc.org/>) provides tools for visualizing, querying, and downloading the data released quarterly by the consortium's member projects. As of June 2017, the Data Portal housed molecular data from 17,570 donors, representing 76 different cancer projects (including TCGA-sponsored projects) involving 21 primary cancer sites.

Pan-Cancer Analysis of Whole Genomes (PCAWG)

While most cancer samples sequenced as part of TCGA and ICGC efforts involved Whole Exome Sequencing (WXS), which primarily sequences the ~1% of the genome that encodes proteins, a subset of samples underwent Whole Genome Sequencing (WGS). The large number of samples subjected to WGS by both TCGA and ICGC offers the opportunity to study the genomic landscape of cancers beyond their protein-coding exomes. Beyond providing insights into how mutations affect regulatory regions, WGS can identify genomic rearrangements, which can have major effects on both gene copy number and gene expression (Davis et al., 2014; Yang et al., 2016). As compared to WXS or SNP arrays, WGS offers much better resolution in the inference of noncoding mutations and of structural variants (SVs) resulting from genomic rearrangements (Weischenfeldt et al., 2017; Yang et al., 2016).

The Pan-cancer Analysis of Whole Genomes (PCAWG) initiative is an international consortium involving both TCGA and ICGC, aiming to comprehensively analyze more than 2,800 cancer whole genomes shared between the two initiatives. PCAWG WGS profiles were collected from multiple TCGA and ICGC studies and represent a wide range of cancer types (Figure 3B). PCAWG data involve a comprehensive and unified identification of noncoding somatic substitutions, indels, and SVs, based on “consensus” calling across multiple independent algorithmic pipelines, together with initial basic filtering, quality checks, and merging (Campbell et al., 2017; Yung et al., 2017). Numerous individual studies of PCAWG data are ongoing or close to publication. A PCAWG “Landing Page” (<http://docs.icgc.org/pcawg/>) highlights biologist-friendly and publicly-available resources for downloading, visualizing, and exploring PCAWG data: including The ICGC Data Portal, UCSC Xena, the EBI Expression Atlas, and PCAWG-Scout. These resources have the PCAWG data pre-loaded, providing easy online access for dynamically querying these complex genomics data and exploring the molecular landscapes of the various cancer types. For example, Xena (<http://xena.ucsc.edu/>) allows the user to visualize multiple types of data by sample-level, including coding and non-coding mutations, structural variants, and gene co-expression patterns; for the specific group of samples (color-coded by cancer type), the corresponding patterns for the sequencing-based data types are represented as colorgrams.

Cancer Genomic Database Resources

Some of the more commonly used cancer genomic databases are described here, though this list is not comprehensive. See Table 1 for access information to these resources, all of which provide a web interface for accessing data or results. As described below and in Table 1, many of these resources allow the user to visualize results for either a select set of top-altered genes or for an inputted list of genes, and other resources allow for bulk download of Level 3 or Level 4 data from TCGA and from other large-scale public datasets. Other relevant cancer genomic databases include those from ICGC and PCAWG, as described above.

Gene Expression Omnibus (GEO)

One of the most popular repositories for gene expression and DNA copy data from published studies is GEO (<http://www.ncbi.nlm.nih.gov/geo/>), which makes tens of thousands of molecular profile datasets publicly available (as of 2017, over 85,000 datasets). In addition to GEO (maintained by the United States' National Center for Biotechnology Information or NCBI), the ArrayExpress database (<https://www.ebi.ac.uk/arrayexpress/>), maintained by The European Bioinformatics Institute or EMBL-EBI performs a similar function and is of comparable scale as GEO, making profiling data from an additional tens of thousands of studies available. When publishing studies using molecular profiling data, authors are encouraged by journals, by funding mechanisms, or by the scientific community, to deposit their data in a public repository. GEO (as well as ArrayExpress) includes gene expression and copy profiles of both human tumors and experimental cancer models. These datasets include, for instance, ones in which a specific gene may be altered in cells (e.g. by knockout or knockdown or overexpression), in order to observe corresponding effects on transcriptional targets. Such datasets are valuable in deriving transcriptional signatures to be associated with a given pathway that is driven by an oncogene or tumor suppressor. Most any gene that one might consider has been manipulated in some model system in a published molecular profiling dataset.

GEO primarily facilitates upload and download of raw and processed expression profiling datasets. Download of datasets can be carried out using the web interface or the Bioconductor package in the R statistical program. From a given GEO accession entry on the GEO web site, one can specify how much data to be downloaded and in what format (using the "Scope," "Format," and "Amount" fields, described in detail at the web interface), and then hit the "GO" button for download. For using R Bioconductor, sample code for GEO download is provided with Bioconductor documentation. In addition, "GEO2R" (<https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html>) is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. GEO2R provides a simple interface that allows users to perform R statistical analysis without command line expertise. To use GEO2R, the user enters a GEO series accession number (referring to a specific dataset to be analyzed), defines sample groups to be compared, assigns samples to each group, then runs the comparison test for each gene (using default or user-specified parameters). Results are presented in the browser as a table of the top 250 genes ranked by P-value, while the complete results table may be downloaded.

NCI Genomic Data Commons (GDC)

The GDC (<https://gdc.cancer.gov/>) is the official unified data repository for several NCI-sponsored cancer genome programs, including TCGA and Therapeutically Applicable Research to Generate Effective Treatments (TARGET). Future or ongoing NCI-sponsored large scale genomics projects will likely deposit data in GDC. The GDC Data Portal allows querying and downloading of the complete data. The GDC also provides a Data Transfer Tool and an API for programmatic access. Both raw data and processed data may be downloaded from GDC. Permissions are needed for downloading raw sequencing data and

other "Controlled-access" protected data, in order to protect the patient's DNA sequence information; GDC provides links with instructions for obtaining controlled access. In addition, the GDC provides user-friendly Data Analysis, Visualization, and Exploration (DAVE) Tools, which allow users to visualize results involving frequently mutated genes for a given data project.

For somatic mutation data of TCGA tumors in particular, the best variant calling dataset for carrying out pan-cancer analyses, is currently the MC3 (Multi-Center Mutation Calling in Multiple Cancers) set of mutation calls, which is described in detail at <https://www.synapse.org/#!Synapse:syn7214402>. The MC3 dataset was produced using six different variant calling algorithms from four centers on over 10,000 tumor/normal pairs in TCGA. The MC3 dataset consists of a single file (in Mutation Annotation Format, or MAF) with each row representing a somatic mutation event within a given TCGA tumor sample. Proteomic data generated by RPPA across 7663 patient tumors are also available at The Cancer Proteome Atlas database (<http://tcpaportal.org/tcpa/>), which features Level 4 data, with batch corrections allowing for comparisons across cancer types.

Broad GDAC Firehose

Over the years, The Broad Institute GDAC "Firehose" web site (<http://gdac.broadinstitute.org/>) has provided a convenient way to download Level 3 and 4 data from TCGA. Firehose executes thousands of analysis pipelines on the entire TCGA dataset and makes analysis results and high-level standardized data tables available for download, including version-stamped, standardized datasets, packages of standard scientific analysis results, and biologist-friendly reports. Examples of Level 4 analysis include GISTIC (Mermel et al., 2011) analysis to identify significant regions of copy alteration, and MutSig (Lawrence et al., 2014) analysis to identify significantly mutated genes. TCGA data generation and processing formally ended in 2016, with the data since migrated to the NCI GDC; no further updates to TCGA datasets and their analysis pipelines from Firehose will be forthcoming. At the Firehose web site, "Firebrowse" user-friendly tools allow viewing the expression profile for a gene across cancer types, or for viewing top mutated or copy-altered genes for a given cancer type.

cBioPortal

cBioPortal (<http://www.cbioportal.org/>) provides visualization, analysis results, and downloads of large-scale cancer genomics data sets (Gao et al., 2013). cBioPortal focuses primarily on data from gene sequencing cancer studies, including TCGA. As of 2017, cBioPortal contains data from 162 cancer studies, including data from the recently published "MSK-IMPACT" Clinical Sequencing Cohort (Zehir et al., 2017) from Memorial Sloan Kettering Cancer Center (MSKCC), with compiled tumor and matched normal sequence data on 410 genes, from a unique cohort of more than 10,000 patients with advanced cancer and available pathological and clinical annotations.

For a particular gene or a set of input genes, one can observe mutation and copy alteration patterns across samples within a given study as "oncprints," which visualize multiple genomic alteration events by heatmap representation. One can also visualize mutations

occurring within a gene as a “lollipop” plot using the “MutationMapper” tool, and generate oncoprints with inputted data using the “OncoPrinter” tool; for each of these tools, the user inputs a set of tab-delimited genomic alteration events within the specified format (examples being provided). Data represented in cBioPortal are available for download, including data from each of the individual platforms represented in TCGA. Source code for cBioPortal is publicly available, and users with sufficient technical skills can download and install a local version of cBioPortal on their own servers.

Oncomine

Oncomine (<https://www.oncomine.org>) is a bioinformatics initiative aimed at collecting, standardizing, analyzing, and delivering cancer transcriptome data to the biomedical research community (Rhodes et al., 2007). The first version of Oncomine was released in 2003, with 40 microarray data sets and nearly 100 differential expression analyses, allowing users to query differential expression results for a gene of interest across collected datasets. Over time, the number of datasets in Oncomine has been considerably expanded, with 715 datasets and over 80,000 samples as of 2017. These datasets mainly consist of gene expression profiling, though some datasets are of copy number. The academic version of Oncomine allows one to input a specific gene and observe its differential expression patterns (e.g. cancer vs normal) across datasets and studies. Many more features are available in the commercial version, including the ability to download publication-quality figures of analysis results and the ability to carry out other types of comparisons (e.g. high versus low stage or grade). “Concept maps” allow one to explore relationships involving differentially expressed genes, but these are only available in the commercial version. Datasets available in Oncomine are not available for download via the Oncomine web site. Though Oncomine does get substantial use by the research community, limitations on available analysis features in the academic version, as well as caveats with interpreting cancer versus normal comparisons (where “normal” tissue represents a collection of different cell types), are considerable.

UALCAN

UALCAN (<http://ualcan.path.uab.edu/>) is an easy to use, interactive web-portal to facilitate the study of gene expression variations and survival associations across tumors, using TCGA gene expression data (Chandrashekar et al., 2017). For a given cancer type represented in TCGA, UALCAN users can do the following: 1) analyze relative expression of a query gene or genes across tumor and normal samples, as well as in various tumor sub-groups based on individual cancer stages, tumor grade, race, body weight, or other clinicopathologic features; 2) estimate the effect of gene expression level and clinicopathologic features on patient survival; and 3) identify the top over- and under-expressed genes for the given cancer type. For example, a user can input a list of genes, and for each gene a boxplot is provided, e.g. comparing cancer versus normal or comparing differences according to cancer stage (the user can select the box plot of interest); Kaplan Meier plots evaluating survival correlations for each gene are also provided. The analysis results (box plots, Kaplan Meier plots, and heat maps) can be printed directly or downloaded in several formats including PNG (Portable Network Graphics), JPEG (Joint Photographic Experts Group), PDF (Portable Document Format), and SVG (Scalable Vector Graphics). Concerning survival analyses

using TCGA data, it should be noted that for some cancer types represented in TCGA, the patient follow-up times are relatively short, which results in less statistical power in being able to identify robust survival correlates.

KM plotter

KM plotter (<http://kmplot.com>) assesses the effect of input genes on survival in specific cancer types, including breast, ovarian, lung and gastric cancer patients (Szász et al., 2016). The gene expression profiling datasets in KM plotter have been compiled from a number of smaller datasets from individual studies, with the combined datasets offering much more statistical power. Gene expression data and relapse free and overall survival information are downloaded from GEO, European Genome-phenome Archive (EGA), and TCGA. KM plotter datasets represent over 50,000 gene transcripts and over 10,000 cancer samples, including over 5,000 breast, over 1,800 ovarian, over 2,400 lung, and over 1,000 gastric cancer patients. For individual genes or for a multigene classifier, KM plotter draws a Kaplan-Meier (KM) plot of samples stratified by expression and assesses the significance of survival differences. To analyze the prognostic value of a particular gene, the patient samples are split into two groups according to various quantile expressions of the proposed biomarker. The two patient cohorts are compared by a KM survival plot, and the hazard ratio with 95% confidence intervals and the log-rank P-value are calculated. Each expression dataset is updated biannually.

Acquiring Bioinformatics Analysis Expertise

While most of the cancer genomic databases described above feature user-friendly interfaces accessible to most researchers, questions may be asked of the underlying data outside the range of what user-friendly tools can readily answer. For example, recent studies from my own group exploring TCGA data (Chen et al., 2017; Chen et al., 2016a; Chen et al., 2016b; Gibbons and Creighton, 2017; Zhang et al., 2017)—studies that involved integration of results between various genomic data platforms and spanning multiple cancer types—were not achievable using just the available user-friendly web-based interfaces for querying TCGA. Instead, a high level of expertise from my own group in using more flexible tools was required, involving some programming skills, manual integration of disparate data tables, and effective visual communication using Adobe Illustrator. As noted above, Level 3 data from TCGA and from other public resources are available for download and to interrogate with much greater flexibility. For digging deeper into genomic datasets, a high level of bioinformatics analysis expertise is needed. There is ample opportunity for collaborations between molecular biologists—who can provide new datasets of experimental models as well as domain expertise—and bioinformatics analysts—who can provide the analytical and computational expertise. Researchers who may have started out as molecular biologists can also acquire bioinformatics analysis skills. Younger researchers, particularly those still in training, may have both the time and aptitude needed to learn skills for coding and integrating data tables.

Different bioinformatics analysts have different skill sets, and each may have his or her preferences in how to handle a particular dataset. While the entire range of analysis tools for

expert analysis cannot be fully explored here, we note some flexible analysis tools that readers in the training phase may consider as a starting point. Bioconductor (Gentleman et al., 2005) is an open source software project based on the R programming language (<https://www.r-project.org/>) that provides tools for the analysis of high-throughput genomic data. Bioconductor and R are very popular among bioinformatics researchers and versatile in carrying out many types of analyses. R is extremely useful in carrying out any number of statistical analyses and in generating graphics. R uses a scripting language, with its functions and libraries being well-described in its documentation as well as on the web, and one can readily find example code for carrying out specific tasks. While it may not get as much attention from the bioinformatics community at large, Microsoft Excel can serve as a high-powered analysis tool that is very useful in integrating results between data tables. Excel formulas such as “MATCH” and “INDEX” allow one to look up elements from one data table within a second data table, and learning the shortcut keys in Excel will save one a great deal of time versus clicking or scrolling with the mouse. Effective visualization of results from complex molecular datasets is a critical component of bioinformatics analysis (Creighton and Huang, 2015). Visualization of differential expression patterns using heat maps can be performed, using either JavaTreeview (Saldanha, 2004), matrix2png (Pavlidis and Noble, 2003), or R. Adobe Illustrator is a highly useful software tool for visualizing and presenting various results together as publication-quality figures.

Concluding Remarks

While the generation of large-scale genomic datasets is being met with fewer technical hurdles, the higher-level analyses of such datasets remain challenging, though doable, requiring scientists with unique skillsets. In the case of the Human Genome Project (HGP), data generation itself was the primary goal, but current and future large scale initiatives aimed at a better understanding of cancer and other diseases will need a strong analysis component to generate new insights from genomics data. User-friendly tools and database interfaces will remain important in giving scientists lacking computational skills at least some level of access to large scale cancer genomic datasets. However, academic institutions and funding mechanisms could do more to encourage expertise in bioinformatics analysis, in addition to software and algorithm development. Of the next generation of scientists, more and more are likely to acquire bioinformatics analysis expertise, allowing them to directly interface genomic data and probe deeper into questions concerning the molecular biology of cancer.

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) grant P30CA125123 (C. Creighton) and Cancer Prevention and Research Institute of Texas (CPRIT) grant RP120713 C2 (C. Creighton).

Literature Cited

- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin A, Kim S, Wilson C, Lehár J, Kryukov G, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483:603–607. [PubMed: 22460905]
- Campbell, P., Getz, G., Stuart, J., Korb, J., Stein, L. Pan-cancer analysis of whole genomes. *BioRxiv*. 2017. <https://doi.org/10.1101/162784>

- Weinstein J, Collisson E, Mills G, Shaw K, Ozenberger B, Ellrott K, Shmulevich I, Sander C, Stuart J. Cancer_Genome_Atlas_Research_Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*. 2013; 45:1113–1120. [PubMed: 24071849]
- Chandrashekar D, Bashel B, Balasubramanya S, Creighton C, Ponce-Rodriguez I, Chakravarthi B, Varambally S. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia*. 2017; 19:649–658. [PubMed: 28732212]
- Chen F, Zhang Y, Bossé D, Lalani A, Hakimi A, Hsieh J, Choueiri T, Gibbons D, Ittmann M, Creighton C. Pan-urolologic cancer genomic subtypes that transcend tissue of origin. *Nat Commun*. 2017; 8:199. [PubMed: 28775315]
- Chen F, Zhang Y, Parra E, Rodriguez J, Behrens C, Akbani R, Lu Y, Kurie J, Gibbons D, Mills G, et al. Multiplatform-based Molecular Subtypes of Non-Small Cell Lung Cancer. *Oncogene*. 2016a E-pub Oct 24.
- Chen F, Zhang Y, enbabao lu Y, Ciriello G, Yang L, Reznik E, Shuch B, Micevic G, De Velasco G, Shinbrot E, et al. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Rep*. 2016b; 14:2476–2489. [PubMed: 26947078]
- Creighton C, Huang S. Reverse phase protein arrays in signaling pathways: a data integration perspective. *Drug Des Devel Ther*. 2015; 9:3519–3527.
- Davis C, Ricketts C, Wang M, Yang L, Cherniack A, Shen H, Buhay C, Kang H, Kim S, Fahey C, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*. 2014; 26:319–330. [PubMed: 25155756]
- ENCODE_Project_Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
- Forrest A, Kawaji H, Rehli M, Baillie J, de Hoon M, Lassmann T, Itoh M, Summers K, Suzuki H, et al. FANTOM_Consortium_and_the_RIKEN_PMI_and_CLST_(DGT). A promoter-level mammalian expression atlas. *Nature*. 2014; 507:462–470. [PubMed: 24670764]
- Gao J, Aksoy B, Dogrusoz U, Dresdner G, Gross B, Sumer S, 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]
- Garnett M, Edelman E, Heidorn S, Greenman C, Dastur A, Lau K, Greninger P, Thompson I, Luo X, Soares J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012; 483:570–575. [PubMed: 22460902]
- Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S. *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer Science+Business Media, Inc.; 2005.
- Gibbons D, Creighton C. Pan-cancer survey of epithelial-mesenchymal transition markers across The Cancer Genome Atlas. *Dev Dyn*. 2017 E-pub 2017 Jan10.
- Hoadley K, Yau C, Wolf D, Cherniack A, Tamborero D, Ng S, Leiserson M, Niu B, McLellan M, Uzunangelov V, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*. 2014; 158:929–944. [PubMed: 25109877]
- Hudson T, Anderson W, Artez A, Barker A, Bell C, Bernabé R, Bhan M, Calvo F, Eerola I, et al. International_Cancer_Genome_Consortium. International network of cancer genome projects. *Nature*. 2010; 464:993–998. [PubMed: 20393554]
- Lawrence M, Stojanov P, Mermel C, Robinson J, Garraway L, Golub T, Meyerson M, Gabriel S, Lander E, Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014; 505:495–501. [PubMed: 24390350]
- 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 biology*. 2011; 12:R41. [PubMed: 21527027]
- Pavlidis P, Noble W. Matrix2png: A Utility for Visualizing Matrix Data. *Bioinformatics*. 2003; 19:295–296. [PubMed: 12538257]
- Rhodes D, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs B, Barrette T, Anstet M, Kincaid-Beal C, Kulkarni P, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*. 2007; 9:166–180. [PubMed: 17356713]

- Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Roadmap_Epigenomics_Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518:317–330. [PubMed: 25693563]
- Saldanha AJ. Java Treeview--extensible visualization of microarray data. *Bioinformatics*. 2004; 20:3246–3248. [PubMed: 15180930]
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 2003; 100:9440–9445. [PubMed: 12883005]
- Szász A, Lániczky A, Nagy Á, Förster S, Hark K, Green J, Boussioutas A, Busuttil R, Szabó A, Gy rffy B. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget*. 2016; 7:49322–49333. [PubMed: 27384994]
- Weischenfeldt J, Dubash T, Drainas A, Mardin B, Chen Y, Stütz A, Waszak S, Bosco G, Halvorsen A, Raeder B, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature genetics*. 2017; 49:65–74. [PubMed: 27869826]
- Yang L, Lee M, Lu H, Oh D, Kim Y, Park D, Park G, Ren X, Bristow C, Haseley P, et al. Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. *Am J Hum Genet*. 2016; 98:843–856. [PubMed: 27153396]
- Yung, C., O'Connor, B., Yakneen, S., Zhang, J., Ellrott, K., Kleinheinz, K., Miyoshi, N., Raine, K., Royo, R., Saksena, G., et al. Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *BioRxiv*. 2017. <https://doi.org/10.1101/161638>
- Zehir A, Benayed R, Shah R, Syed A, Middha S, Kim H, Srinivasan P, Gao J, Chakravarty D, Devlin S, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*. 2017; 23:703–713. [PubMed: 28481359]
- Zhang Y, Kwok-Shing Ng P, Kucherlapati M, Chen F, Liu Y, Tsang Y, de Velasco G, Jeong K, Akbani R, Hadjipanayis A, et al. A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell*. 2017 E-pub May 8.

Figure 1. Overview of bioinformatics as a discipline

(A) Bioinformatics involves computer science, statistics, and molecular biology. Computer science and statistics represent tools for the analysis of large molecular datasets, while a good understanding of molecular biology provides the motivation for drawing insights from the data. (B) Areas of specialty in bioinformatics include algorithm development, development of software and database tools, and analysis. Analysis applies existing tools and algorithms to draw insights from large molecular datasets.

Figure 2. Overview of the different levels of molecular profiling data

(A) Levels of molecular profiling data can range from Level 1 (raw data) to Level 5 (results yielding insights from the data). Examples (“Ex”) for each data level are listed. BAM, binary format for storing sequence data; expr. table, expression data table; SMGs, significantly mutated genes; Diff Ex., differentially expressed genes. (B) There are practical limitations with automating analysis tasks involving molecular profiling data. Many tasks involving data Levels 1–4 may lend themselves to automation, while Level 5 (involving data integration and domain expertise) typically requires more creativity and thought involving human analysts. Basic and routine service, stable pipelines, and application development tend to be

involved in fully automated tasks (where automation can save one considerable time), while advanced and highly complex needs, rapid upgrades, new method development, and unique analytical approaches are more time consuming and would be project specific. (C) There is an inherent trade-off with software analysis tools. As general usability increases, flexibility decreases and vice versa. A tool that is usable by a large audience can answer a limited set of questions, while a tool providing maximum flexibility to answer a wide range of questions would require specialized expertise to use.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Overview of data and samples available as part of The Cancer Genome Atlas (TCGA)

(A) For profiled TCGA cases, distributions by cancer type, with number of cases for each cancer type indicated. Tumors spanned 32 different TCGA projects, each project representing a specific cancer type, listed as follows: LAML, Acute Myeloid Leukemia; ACC, Adrenocortical carcinoma; BLCA, Bladder Urothelial Carcinoma; LGG, Brain Lower Grade Glioma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CRC, Colorectal adenocarcinoma (combining COAD and READ projects); ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular Germ Cell Tumors; THYM, Thymoma; THCA, Thyroid carcinoma; UCS, Uterine Carcinosarcoma; UCEC, Uterine

Corpus Endometrial Carcinoma. Numbers based off of a recent pan-cancer study of the PI3K/AKT/mTOR pathway (Zhang et al., 2017). **(B)** For 11232 cancer cases in TCGA, availability of data by molecular platform (WXS, whole exome sequencing; MC3, Multi-Center Mutation Calling in Multiple Cancers; RPPA, Reverse Phase Protein Array) and by inclusion in recent studies involving data from multiple TCGA projects (Campbell et al., 2017; Chen et al., 2017; Chen et al., 2016a; Chen et al., 2016b; Gibbons and Creighton, 2017; Zhang et al., 2017). PCAWG, Pan-cancer analysis of whole genomes project; WGS, whole genome sequencing.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Selected Cancer Genomic Database Resources

name	URL	description	bulk data retrieval	analysis features of note
Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/	A public functional genomics data repository, including array- and sequence-based data, representing primary datasets from tens of thousands of published studies.	yes	An interactive web tool, GEO2R, allows users to compare two or more groups of samples in order to identify genes that are differentially expressed.
NCI Genomic Data Commons (GDC)	https://gdc.cancer.gov/	Unified data repository for several NCI-sponsored cancer genome programs, including TCGA and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).	yes	Data Analysis, Visualization, and Exploration (DAVE) Tools allow users to visualize results involving frequently mutated genes for a given project.
Broad GDAC Firehose	http://gdac.broadinstitute.org/	Executes thousands of analysis pipelines on the entire TCGA dataset and makes analysis results and high-level standardized data tables available for download.	yes	Firebrowse tools allow for viewing the expression profile for a gene across cancer types, or for viewing top mutated or copy-altered genes for a given cancer type.
cBioPortal	http://www.cbioportal.org/	Provides visualization, analysis and download of large-scale cancer genomics data sets. Gene sequencing cancer studies, including TCGA, are represented.	yes	For a set of input genes, observe mutation and copy alteration patterns across samples within a given study.
Oncomine	https://www.oncomine.org	Database of molecular profiling data (primarily gene expression) from published cancer studies, with web-based data mining tools for exploring specific genes.	no	Academic version allows one to input a specific gene and observe its differential expression patterns (e.g. cancer vs normal) across datasets and studies. More features are available in the commercial version.
UALCAN	http://ualcan.path.uab.edu/	Provides easy access to publicly available cancer transcriptome data from TCGA, with publication quality graphs and plots depicting gene expression differences and patient survival associations.	no	For individual genes and specified cancer type, generates kaplan-meier (KM) plot of samples stratified by expression and assesses the expression differences between comparison groups (e.g. cancer vs normal).
KM plotter	http://kmplot.com	Assesses the effect of input genes on survival in specific cancer types including breast, ovarian, lung and gastric cancer patients.	no	For individual genes or for multigene classifier, draws a kaplan-meier (KM) plot of samples stratified by expression and assesses the significance of survival differences.

NCI, National Cancer Institute; GDAC, Genome Data Analysis Center; TCGA, The Cancer Genome Atlas