UCSC Genome Browser: Possible Contributions in Sars-Cov-2 Diagnosis

Sanskrita Sukla

Amity Institute of Biotechnology, Amity University, Noida

---***---

Abstract - In recent times, the approach for genomic analysis has been revolutionised. The venture of enquiry and study no longer requires long hours of preparing, testing and waiting for results nor does it require a long wait time to obtain pre-analysed support data for further experimentation. Genome Browsers with their underlying database housing colossal amounts of documented, pre-computed data have provided a sturdy stepping ground for the exponential rate at which new research has been progressing. The UCSC Genome Browser with its central and offsite databases, an array of gateway interfaces and tools, secure data hubs for personal and public circulation— forms one of the most far-reaching assets for sequencing and further investigation. The paper discusses the procedural work put in to develop such massive data repositories, their distribution and establishment of a browser that queries them for graphical results. The paper pursues the examination of the specifics of the UCSC Genome Browser architecture and its tools. Along the way, a few segments digress to provide insights about addition of new assemblies and pertinent projects like, ENCODE. Finally, wherever relevant, the paper reviews a few instrumental studies to which UCSC Genome Browser has played a pivotal role.

***Key Words*:** BLAT, database, genomic data, SARS-CoV-2, sequencing, UCSC genome browser

1. INTRODUCTION TO DATA REPOSITORIES AND GENOMIC DATA ANALYSIS

In 1990, the advent of The Human Genome Project (HGP) popularised the idea of mapping of the information contained in and associated to the nucleotides. The project went on to create the first extensively sequenced catalogue of the human genome. It mostly documented the chromosomal regions excluding the primary and secondary constriction areas, which only made up about 8% of the genome. [1]

This massive project created the first real need for a properly architectured genome database and associated browser tools. The sequenced data was reserved in GenBank (NCBI) and shared with contemporary databases like DNA DataBank of Japan (National Institute of Genetics) and European Molecular Biology laboratory. [2]

The process of putting together a Genome Database or Browser is a long, ongoing and intricate process, especially with the exponential rate with which the data is being evaluated. For this to be accessible to various research opportunities, it needs to be organised and properly 'tagged', with all the information available minimal clicks away. However, it is not possible to store such varied data, say— loci positions, transcript data i.e., mRNA data of a particular loci, protein structure or biochemical function or role, in a single, independent, horizontally-scaling database. [2]

Genome data storage and querying needs to involve several interconnected and compartmentalised databases. The employed database depends on a genome repository that caters to the need of a wide array of Genome Browsers. For example, NCBI DataViewer, UCSC Genome Browser and Ensembl Browser share their resources. [2]

1.1. Establishing a Sequence data Repository

The creation of any Genome Data Repository more or else follows a basic rough outline— sequencing, assembling and gene positioning in the genome (usually against a reference genome), aligning transcript data (mRNA transcribed by the genes), ascertaining the gene model (making clear regional demarcations of the parts of the genome producing coding RNA or ncRNA that regulate transcriptional or post-transcriptional activities), constructing a data model that is able to house complex interrelated data and most essentially maintenance and timely revision and updation of the data. [2]

Sequencing is the process of computing the order of nucleobases in an organism's genome. A sequence would ideally have a complete ordered list of all the nucleobases in the genome. [3]

DNA sequencing involves fragmenting the target DNA and its subsequent attachment to a short, chemically synthesised oligonucleotide called a linker or an adapter on both sides. These unordered *fragments* together form a *sequencing library*. Once the nucleobases are inferred and sequenced into the ordered lists they are known as a set of *reads*. [3]

An important part of this process is *annotation* of the sequence. Annotating a sequence refers to tagging a gene locus with various genomic information. There are three levels of annotation— nucleotide level, protein level and process level. The protein level and process level form the *functional* portion of the annotation data. Nucleotide level of annotation or the *structural* portion, recognises and names the sequences that act as coding portions of the genome, thus tagging the gene with the Transcript data. Protein level of annotation helps with recognising the significance of genes with reference to their physical products. The process level of annotation identifies the biochemical pathways and various other functional roles a particular segment is involved in. [4][5]

Now, usually even an entire set of non-overlapping sequence read lengths for a particular organism still falls short of the actual genome length. Hence, while working on genome analysis for an organism, the sequencing of reads is followed by arranging the reads so as to form a scaffolding. Over the course of analysis, error correction and gap filling are done to accommodate newly recognised or sequenced fragments. This is the process of the *Genome Assembly*. [6]

The sequenced nucleobases are put in proper order and hence it acts like a reference for further assembly or querying. For example, in case a new analysis yields copies of the same sequence or same sequences but with varying nucleotide repeat numbers, it is easy to place them in the Genome Assembly. [2]

1.2. Sequence Repository: Unification and Challenges

The sheer volume of this unmitigated data is dumped in *Sequence repositories* all over the world. These Repositories accept patented data, data from personal research or Sequencing projects and store it in databanks like NCBI, DNA Databank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) and so on. Since the late 2000s, these databases have tried to compose a master databank of sorts, the International Nucleotide Sequence Database Collaboration (INSDC). [7]

Since the time of its creation the sheer volume of data has grown many folds with an assortment of complex and non-uniform layouts hence pushing for more organisational efforts. Although INSDC enforces a few protocols to establish that the content is formatted similarly for all the data repositories and is flexible with respect to syntax changes. [8]

These ‘libraries’ still pose a serious challenge to proper inquiry and investigation due to— the data more often than not being inaccurate, inadequate, repetitive or overlapped rendering the shared efforts invalid. The repository data is also not curated or annotated properly, more often than not. [7]

The attempts at revamping and restructuring the empirical data has led to the creation of specialised databases that support a querying and systematic display of a certain category of data, say— genome databases, bio-molecular or transcriptional information specific databases. [7]

2. GENOME BROWSERS AND THEIR CLASSIFICATION

The aforementioned databases are rooted in browsers that make querying of much data easier. By a methodical assimilation of genome assemblies and annotations from various sources the Genome Browser-databases form an exhaustive source of information regarding various species and their genetic information. [7][9]

Genome Browsers allow querying of genomic data and viewing it at varying levels of abstraction or complexities. These browsers display annotated data graphically, supporting annotations from multiple platforms and selective viewing of the available directory of annotations. The GUI allows traversal through the entire lengths of genome for various species and also has tools for analysing them at various levels for annotation. [2][9]

Web-based Genome Browsers are preferred over the other available browsers because of their high accessibility, high-speed performance and up-to-date data quality— usually supplied by dedicated research units. [9]

Two kinds of web-based genomic browsers are currently in use: Multi-species variants and Species-specific variants.

The *species-specific variants*, as the name suggests, have genomic data and respective annotations for a single organism. Examples of species-specific browsers include: RGD for a rat species *Rattus norvegicus*, Xenbase for a frog species *Xenopus tropicalis*, Zfin for an Actinopterygii-an *Danio rerio* or Zebra Fish, Flybase for *Drosophila*, etc. [9]

The *Multi-species variants* are used more extensively for their ability to provide a comparative analysis among various subjects. Multi-species variants can provide genome sequences of evolutionary predecessors or relatives as the reference genome for the sequencing of newer initiatives. The most important element of these browsers, however, is that multi-species variants can provide a truly unique and holistic perspective of a certain gene loci or a transcript or some protein, relating it to evolutionary, taxonomic, species data, expression profiling etcetera. These browsers draw from the repository a huge bundle of transcript data, information about regulatory sequences and also various theories and opinions about their actual roles in pathways or an expression profile. NCBI MapViewer, UCSC Genome Browser, VISTA, Phytozome (plant species browser), Gramene (plant species browser), Genome Projector (for bacterial species), JBrowse, ABrowse, Anno-j etcetera are some of the more commonly used Browsers. [9]

3. THE UCSC GENOME BROWSER

The UCSC Genome Browser includes sequenced data for multiple species of superphylum Deuterostomia (including Phylum Echinodermata, Hemichordata and Chordata) with a large number of those belonging to subphylum Vertebrata, class Insecta, phylum Nematoda and certain species of phylum Mollusca and Fungi. [9] As of the September 2019 update, the underlying database has data worth a whopping 195 assemblies for 105 odd species. [10]

The UCSC Genome Browser was put together during and after the completion of the first and one of the most ambitious sequencing projects till date— The Human Genome Project. Despite its vast archive, only about 20% of its resources are dedicated to the other vertebrate or invertebrate models. [11]

3.1. Specifics of the underlying Database

The underlying database engages the *relational model*, more specifically MySQL RDS, where the data is stored in tables with intersecting rows and columns maintaining records and attributes respectively. These records need to have a well-established and acceptable genomic relation, when tabulated together. The aim is to perfectly correlate the annotation with the sequenced genome. [2]

The database is *distributed* in the sense that only the data pertaining to a single or few organisms is in a site. This makes it easier to construct sites for newer organisms. These ‘locations’ are still related to a single central server and hence backup and security issues are minimal. [2][9]

The browser server confined Sequence data is housed in text files (like FASTA) worth a total of 8 Tb of data and other file formats house subsume about 3 Tb worth annotation and track data. [12]

3.1.1. Data Tabulation for task-specific efficacy

The data tabulation is done based on commonly accepted Genomic Coordinate systems, referred to as Positional tabulation or can be Non-Positional. [12]

Positional Tabulation hooks the genomic features or the functional gene segments to certain numbered regions of the nucleotide in sequenced genomes. The UCSC Genome Database and Browser use two major Genomic Coordinate Systems— the InterBase Coordinate System and the Base Coordinate System. The *InterBase Coordinate System* or *zero-based, half-open* (also known as *space based*) hooks the genomic feature between the nucleotides. The coordinates start from 0 and the numbers are assigned to the start and end of each nucleobase and not to the nucleobase itself. The space between two numbers depicts the nucleotides. This system is used in most of the tools, flat files and the data storage. The *Base Coordinate System* or the *one-based, fully closed* is the one used in the browser displays. The numbers are assigned to each consecutive nucleobase and each count of a number stands for a nucleotide. The InterBase System is usually preferred because of the ability to easily compute sequence length, tag inter-nucleotide features and other features. [12][13]

For relatively smaller tables, like tables containing data about Cysteine-Guanine nucleotide repeats, Sequence Tagged Sites, intron-exon boundaries and tandem repeat areas, the indexing is done at start of coordinate of specific chromosome and the end of the coordinate and the related data is sorted with respect to the chromosome number and start position of nucleotide. [12]

For larger tables that contain expression data and transcript data about alignment or sources of translated proteins, mRNA or EST data, using the same optimization methods would be futile. Hence, the data sets are made for each chromosome. The tables are divided by the chromosome number and indexing, and sorting is done for the data anchored to the chromosome. This decreases the size of the lookup table used by the browser to retrieve data when queried. [12]

For larger positional tables a certain organisational binning scheme is employed by the browser to make these tables faster to query and give results. It was suggested by L. Stein and R. Durbin. They used the size of the end of chromosome coordinate to determine the size and number of bins used. Usually bin size for a chromosome begins from 512 Mb and is further divided into bins of size 64 Mb, 8 Mb, 1 Mb and 128 Kb in succession. If the coordinate for the end of the chromosome is greater than a value of 2^{29} then the binning scheme starts with upto 2 Gb. [14]

The bins contain contiguous genomic features corresponding to the coordinates and while querying they pick up ranges of bins to find the results. [14]

Non-Positional tables assign the annotation to some names that are commonly used for a particular gene or sequence, the researcher(s) or a particular keyword. Some of the non-positional tables could be self-referential, by tabulating data regarding the database and its sources. [12]

3.1.2. Addition of New Assemblies to the Database

Newly sequenced assemblies are uploaded to the NCBI database. The UCSC sources a text based FASTA formatted file, that contains the name of the species, the identification codes and a few more details along with lines of nucleobase sequences. Along with that a file containing sequence descriptors of the consensus region of DNA is also retrieved. This is called an AGP or A Golden Path file. This is compressed in a manner that the nucleobases each occupy two bits of space rather than the one-byte memory taken up by nucleobases in FASTA format. This compressed version is then used by the browser to deliver results. [11]

The improved nomenclature of UCSC database specific assemblies follows the gggSss*n* format. The initial three letters of genus name(ggg) followed by the initial three letters of the species name (Ss) and a numbered code(*n*). The three-letter format of nomenclature can be seen in the assemblies of yeast (Saccharomyces cerevisiae, sacCer##) and Zebrafish (Danio rerio, danRer##). The human genome assemblies are labelled hg*nn* as it is the first of the assemblies to be put together and it stands for **h**uman **g**enome followed by a numbered code(*nn*). The older assembly models for fruit flies (Drosophila melanogaster, dm##), rats (Rattus norvegicus, rn##) and mice follow the two letter models. [11]

These names are also used for referring to the database names where these sequences along with their annotations are stored for active access.

Once the new assemblies are named, they are associated to the central databases by tables referencing their storage data and id, the basic initialising requires a minimum of five tables. These tables include information about their search framework from the Browser, display preferences available for the user, separate tables for contig (known sequence of nucleobases) and gap (portion of genome where order is undecided) sequences and tabulation of the length and identification code of the genomes. The display preferences and search framework tables have self-referential data to the Browser and database connections they have. [11]

3.1.3. Compiling Track Data for the Browser

Once the assembly is uploaded it can be further linked to various kinds of tracks. *Tracks* refer to annotation data that either share the same analysis source or an innate category. These tracks are displayed graphically along the y-axis with respective coordinates along the x-axis, when the browser renders them. [12]

The tracks displayed in UCSC Browser can be pre-computed in the central database and hosted there. The other options are to analyse and compile it elsewhere and display it on the browser or to compute it remotely as well as display it elsewhere but have links embedded in the result obtained. [11] The annotation data processed and computed by UCSC in the database includes mRNA and fragmented cDNA data, information about inherited familial genes, protein to nucleotide alignment data using tBLASTn (especially human proteins against other vertebrate genomes), identified conserved genes in phylogenetic trees and their similarity index. [15] The data pre-processed in the UCSC database, prior to display, forms a limited portion of the tracks. [11]

Third party tracks contribute majorly to the source of the annotation data in the central database that is then displayed in Browser in the same format as the data processed in the browser. A few examples of such tracks include expression data, nucleotide repeats and variation of the repeat of gene copies in different individuals, dysmorphology data regarding disorders and more. A few annotation sets, although sourced from other databases, still undergo variations and further processing to be displayed as tracks. For example, dbSNP data derived is further split into tracks that show variations in single nucleobase (SNPs), variations flagged as probable disease-causing polymorphisms and repeated occurance of a certain SNP. [11]

Sometimes data sets are pre-sorted remotely and are available for actual depiction off the Browser. The central database does not contain the relevant annotation data rather it just presents some information that links the address of the host, to which the users can navigate to— to view the data. [11]

The UCSC Browser has direct gateway links to a few important source projects like ENCODE, DECIPHER etc. UCSC provides a data warehouse for the data from these projects after proper quality and accuracy testing. The data is established in the public domain after making it compatible to the GUI of the Browser. [11]

The ENCODE Project: A major source of Track Data

High resourcing projects for the Browser like ENCODE call for development of specialised tools for data from the particular project and easy wielding of basic Browser tools as well. [16]

The ENCODE Project was initiated in the year 2003 as a successor to the Human Genome Project. Over the years it has been one of the most crucial contributors to the repository of the functional portion of the genome. The initial decision about experimentation and analysis modus was made in the pilot phase of the project by using about 1% of the sequenced genome. About 30 million base pairs worth sequences were picked by hand or randomly and biochemical and specific biological roles were determined for each. The phase helped zero in on the technology to be used for further analyzing the rest of the genome descriptively. [16] The second, third and fourth phases have focused on extending the same amount of thoroughness to the rest of the Human Genome. [17]

The annotation data generated by ENCODE produces a huge number of tracks. Hence, the data is broken down into six ENCODE-specific track groups that provide data about: genomic regions, structures of various genomic units, transcript relation to gene (chromatin immunoprecipitation data or ChIP tracks), abundance of certain transcripts throughout the genome (transcript level data), several species homology/analogy data tracks and tracks showcasing variations. [11] In the recent updates the track display can be modeled more to the users' needs by entering specific keywords. [10]

The tools rooted in UCSC also parse all the submitted tracks and display information gaps present in the study, in-genome comparisons for functional similarities with the use of sequence alignment algorithms and analytical tools for species to species comparative study of similar genomic elements. [17]

These data tracks can also be viewed in a tabular format using the Table Browser of UCSC. The graphical format provides a good understanding of the data itself while tabular format describes its assembly and organisation on a per chromosome basis. For whole-genome annotation analysis, there is an option to download the data via a file transfer network. [17]

The file formats used by ENCODE project for various purposes include FASTQ (for text based files containing nucleobase and accuracy information), interconvertible Binary and Sequence Alignment/Mapping (contains sequenced and aligned data), bigBed file (annotation data obtained upon analysis), bigWig files (visualisation details for contiguous mapped reads) and more. [17][16]

3.2. Engineering the UCSC Genome Browser

The UCSC Genome Browser adheres to a *client-server model* i.e., the client or the user and the server have a continuous conversation over a network. The client side requests data and the server host interprets the input data and provides the resources; however, the user doesn’t share its own data.

The UCSC Genome Browser is a *traditional browser* by virtue of the logic behind the functioning of the browser with its serving database. It was built on classic web technologies. [9]

To actually aid the database-to-browser bridge, an automated SQL program makes use of input data to create a C language native, composite data type— *structure* which defines manually grouped data under a particular name in the memory. The data is tabulated using a create table statement in SQL and a number of C functions are used for the final data results. [12]

It has a *full duplex*, *synchronised* form of data transmission. The data is sent in the form of blocks and frames from the server-side in a contiguous manner. A local copy of the data is present on the server disk of the browsers, to boost speed and performance and along with multiple other optimization strategies to help with standardizing the page load time for results of differing volumes. [9]

The UCSC Genome browser handles data delivery to the user in an impartial manner, using *flat web pages* or *static web pages*, disregarding the context and subject. The data is released exactly as it is stored, and no changes are made to it. [9] If the web-applications were present on-site in the web pages they could perform real-time customized production of data with client specific parameters rather than carrying the request back to the database to form a display result. Although it is somewhat outdated, static web pages are easier to host, index in search engines, quicker to develop and faster to transfer in slow connections.

The images are also rendered on the *server side* before transfer. This allows the page to be accessible with even the most basic of tech and also reduces the page load time. [9]

In some of its aspects, the Browser lags behind in what they could offer to the computational biology and genetics but there are continual attempts to replicate features offered by next-gen sequencing and querying web browsers. However, it is also imperative to appreciate that these 'traditional' properties actually help ensure that large volumes of data are transferred safely and, in their entirety, and the user achieves a comprehensive experience.

The Browser aims to provide high quality and thorough data wrt to annotations markers, models, transcriptional and post-transcriptional regulatory data, epigenomic data and data regarding constrictions on the chromosomes. However, many times the depicted data sets contravene themselves when placed together and the browser doesn't make it an assignment to assess the best results. It is up to the user to make the decision. [11]

The browser shows data in various grades of resolution from one nucleotide to entire evolutionary lines of genome. It acts as a web page with several hyperlinks to papers or journals discussing various models from other databases and various transcript evidences. Rather than requiring the users to learn the nuances of all the Browsers it allows transfer to the sorted information directly. [18]

As further gap-filling and scaffolding is done with respect to reference genomes, new assemblies are updated for species. The older versions can still be accessed as easily for projects that have already been implementing and have invested in the particular assembly. [11]

4. THE GRAPHICAL BROWSER AND OTHER TOOLS

The Browser itself is a 'coordinate dependent map navigation system' which works not very different from navigator apps. There are multiple other useful applications built around it, some of it based on the suggestions of the developers and the users. There are more than thirty different applications, built around the Display Engine that are based on Common Gateway Interface programs. [11][18]

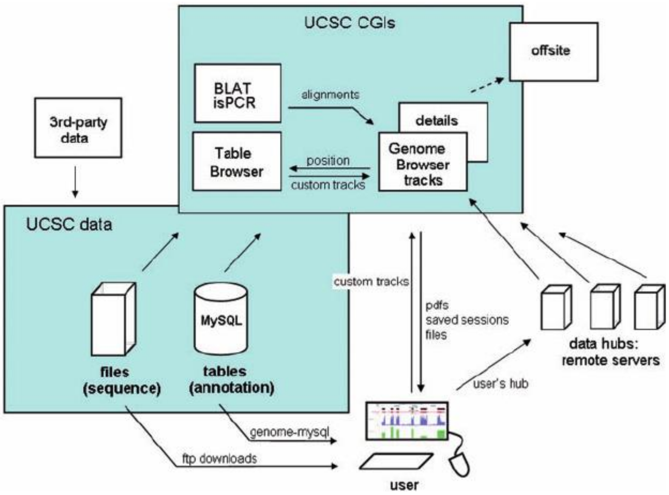


Figure 3.2.1. A general scheme of the UCSC Browser, local and remote database and accessory tools

Source: *The UCSC genome browser and associated tools*; Kent W.J., Kuhn R.M., Haussler D.

4.1. The Display Engine

The Display Engine forms approximately 55% for the browser traffic. The Interface programs front-end the database i.e, help with conversion of hierarchically stored and flat file data to a graphical format. [11]

The *hgGateway interface* etches a graphical interface on which the user makes various selections like species, respective genome assembly, tools display and more. [11]

Once the user sends requests using a certain position or a keyword, the *hgTracks interface* quickly queries the database for answering tracks and renders them graphically onto the screen. The rendering is done such that each slice of track on the y-axis becomes a hyperlink to further detailed and precomputed alignment data for the user. The binning scheme described earlier in the paper, greatly improves the speed and efficiency of the entire process. In the most recent of advances the search engine can easily take coordinate formats or keywords used by other genome browsers. [11]

The *hgGenes interface* draws the graphical interface that contains further information about the data tracks and can redirect the users to other databases or repositories and affiliated journals and studies. [11]

Some of the outsourcing is done from PubMed, UniProtKB, database of Online Mendelian Inheritance in Man, NCBI directories etc. [12]

The Browser also mediates seamless transition between details at various degrees of magnification. The longest chromosome has approximately 250 million base pairs and the browser aims to provide display resizing options from a sequence that is few bases long to the entire chromosome. The Browser has options that facilitate zooming in and out smoothly— with only a few clicks and provides concurrent scaling information along with related feature data. The amount and type of data displayed alongside the —sequence depends on the resolution of the genome. While displaying an entire chromosome, the data could be limited to only important landmarks and certain biologically important transcript data. However, querying data regarding a certain gene element or a sequence could allow a more exhaustive data output. [18][11]

4.2. The Table Browser

The Table Browser accounts for less than 15% of the server traffic. It serves as a very adaptable interface between the relational database and the display engine. It has multiple menu options to choose from and allows quick and easy changing of the parameters by the users. The graphical display lies atop the Table Browser. This means that the annotation data visualised graphically can be retraced back to tabulated track data. The rendering is facilitated by the table browser under any set of parameters along with the option to compare various tracks together. This helps with gaining results in the form of various data sets like— common data for all chosen parameters, data that fulfill at least one chosen parameter or data analysed comparatively. These result sets can be obtained in the user-called format. The Table Browser proves useful in providing other information regarding related data and metadata of the tables being queried. [19][11]

4.2.1. Use in Diagnosis of SARS-CoV-2

The Table Browser is an easy querying tool to use to obtain large amounts of tabulated textual data. [11] Figure 4.2.1. depicts use of Table Browser to obtain possible CRISPR target sites in SARS-CoV-2. The CRISPR sites hold possible diagnostic and prognostic features as they can be treated with an associated enzyme, Cas13 to trigger the dismantling of the specific RNA segment, achieving a state of 'transient knockdown'. In certain cases, the enzyme goes 'rogue' and starts cutting through adjacent RNA that don't have the CRISPR repeat region— producing a mass signaling effect. [20]

The table browser can be used for locating such sites in any genome. The Browser allows the user to choose from 6 different monophyletic groups and thereafter to pick the species and related genome assembly. In figure 4.2.1., the only assembly available for SARS-Cov-2 is selected. Although other more thoroughly studied organisms have multiple older assemblies in the database too.

Now, the user can either narrow down the search options by specifying the group that contains the desired tracks and annotation data or just apply the 'All Tracks' parameter. Various genome feature data groups can be found in the track pull-down list. The user can hereafter choose to either display the selected track data for either the entire genome or a single coordinate. In figure 4.2.1., the Browser has values that will display CRISPR region annotation data for the entire SARS-CoV-2 genome.

The Browser also has other filtering and comparison features which can refine the result based on coordinate or memory ranges and the number of other databases vouching for the same data.

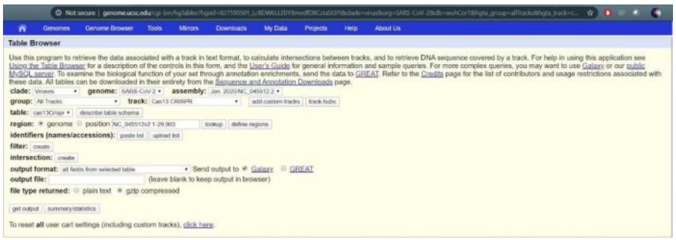

Not secure | genome.ucsc.edu/cgi-bin/hgTables?hgsid=837590501_LrBDWiLLEIYXmedfDXCzta5I3P&clade=virus&org=SARS-CoV-2&db=wuhCor1&hgta_group=allTracks&hgta_track=c...
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Projects
Help
About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, and the User's Guide for general information and sample queries. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

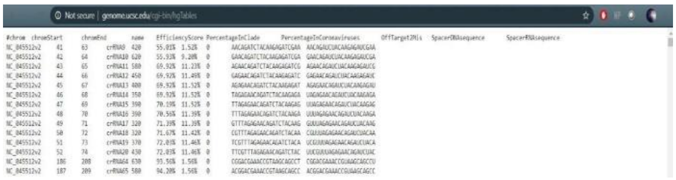
clade: Viruses **genome:** SARS-CoV-2 **assembly:** Jan. 2020/NC_045512.2
group: All Tracks **track:** Cas13 CRISPR [add custom tracks] [track hubs]
table: cas13Crispr [describe table schema]
region: ● genome ○ position NC_045512v2:1-29,903 [lookup] [define regions]
identifiers (names/accessions): [paste list] [upload list]
filter: [create]
intersection: [create]
output format: all fields from selected table Send output to ☑ Galaxy □ GREAT
output file: (leave blank to keep output in browser)
file type returned: ○ plain text ● gzip compressed

[get output] [summary/statistics]

To reset **all** user cart settings (including custom tracks), click here.

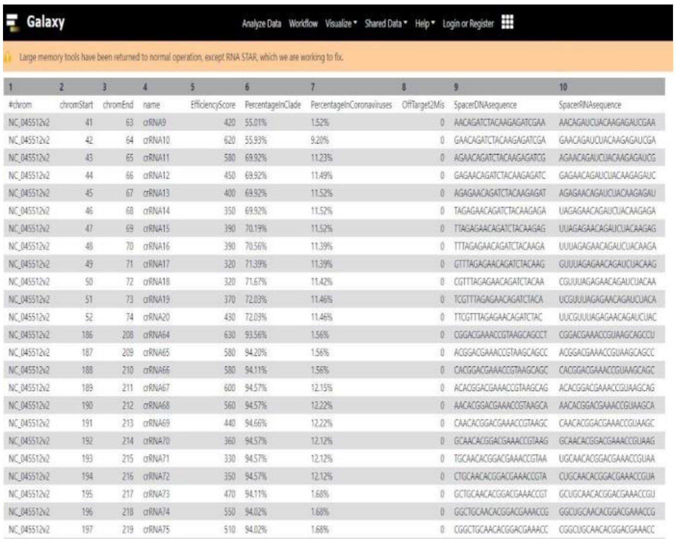
Figure 4.2.1. The Table Browser at UCSC Genome Browser is set to a certain combination of parameters to display data related to the CRISPR sites that can be targeted by Cas13 and associated enzymes to provide a diagnosis for SARS-CoV-2 infected patients

The query in figure 4.2.1. does not have any special 'filtering' or 'intersection' requests. The user may choose to view only certain attributes of the table in the output if they wish. Once desired selections are made the user can view the output directly via the table browser which gives the data objects in a single tab-spaced table (figure 4.2.2a). The tables could have an unstable look to them, hence using the GalaxyProject platform the output can be viewed in an organised tabular format (figure 4.2.2b).



```
#chrom  chromStart      chromEnd        name    EfficiencyScore PercentageInClade       PercentageInCoronaviruses       OffTarget2Mis   SpacerDNAsequence       SpacerRNAsequence
NC_045512v2     41      63      crRNA9  420     55.01%  1.52%   0       AACAGATCTACAAGAGATCGAA  AACAGAUCUACAAGAGAUCGAA
NC_045512v2     42      64      crRNA10 620     55.93%  9.20%   0       GAACAGATCTACAAGAGATCGA  GAACAGAUCUACAAGAGAUCGA
NC_045512v2     43      65      crRNA11 580     69.92%  11.23%  0       AGAACAGATCTACAAGAGATCG  AGAACAGAUCUACAAGAGAUCG
NC_045512v2     44      66      crRNA12 450     69.92%  11.49%  0       GAGAACAGATCTACAAGAGATC  GAGAACAGAUCUACAAGAGAUC
NC_045512v2     45      67      crRNA13 400     69.92%  11.52%  0       AGAGAACAGATCTACAAGAGAT  AGAGAACAGAUCUACAAGAGAU
NC_045512v2     46      68      crRNA14 350     69.92%  11.52%  0       TAGAGAACAGATCTACAAGAGA  UAGAGAACAGAUCUACAAGAGA
NC_045512v2     47      69      crRNA15 390     70.19%  11.52%  0       TTAGAGAACAGATCTACAAGAG  UUAGAGAACAGAUCUACAAGAG
NC_045512v2     48      70      crRNA16 390     70.56%  11.39%  0       TTTAGAGAACAGATCTACAAGA  UUUAGAGAACAGAUCUACAAGA
NC_045512v2     49      71      crRNA17 320     71.39%  11.39%  0       GTTTAGAGAACAGATCTACAAG  GUUUAGAGAACAGAUCUACAAG
NC_045512v2     50      72      crRNA18 320     71.67%  11.42%  0       CGTTTAGAGAACAGATCTACAA  CGUUUAGAGAACAGAUCUACAA
NC_045512v2     51      73      crRNA19 370     72.03%  11.46%  0       TCGTTTAGAGAACAGATCTACA  UCGUUUAGAGAACAGAUCUACA
NC_045512v2     52      74      crRNA20 430     72.03%  11.46%  0       TTCGTTTAGAGAACAGATCTAC  UUCGUUUAGAGAACAGAUCUAC
NC_045512v2     186     208     crRNA64 630     93.56%  1.56%   0       CGGACGAAACCGTAAGCAGCCT  CGGACGAAACCGUAAGCAGCCU
NC_045512v2     187     209     crRNA65 580     94.20%  1.56%   0       ACGGACGAAACCGTAAGCAGCC  ACGGACGAAACCGUAAGCAGCC
```

Figure 4.2.2a. The data as displayed directly by the output from the Table Browser



Large memory tools have been returned to normal operation, except RNA STAR, which we are working to fix.

1	2	3	4	5	6	7	8	9	10
#chrom	chromStart	chromEnd	name	EfficiencyScore	PercentageInClade	PercentageInCoronaviruses	OffTarget2Mis	SpacerDNAsequence	SpacerRNAsequence
NC_045512v2	41	63	crRNA9	420	55.01%	1.52%	0	AACAGATCTACAAGAGATCGAA	AACAGAUCUACAAGAGAUCGAA
NC_045512v2	42	64	crRNA10	620	55.93%	9.20%	0	GAACAGATCTACAAGAGATCGA	GAACAGAUCUACAAGAGAUCGA
NC_045512v2	43	65	crRNA11	580	69.92%	11.23%	0	AGAACAGATCTACAAGAGATCG	AGAACAGAUCUACAAGAGAUCG
NC_045512v2	44	66	crRNA12	450	69.92%	11.49%	0	GAGAACAGATCTACAAGAGATC	GAGAACAGAUCUACAAGAGAUC
NC_045512v2	45	67	crRNA13	400	69.92%	11.52%	0	AGAGAACAGATCTACAAGAGAT	AGAGAACAGAUCUACAAGAGAU
NC_045512v2	46	68	crRNA14	350	69.92%	11.52%	0	TAGAGAACAGATCTACAAGAGA	UAGAGAACAGAUCUACAAGAGA
NC_045512v2	47	69	crRNA15	390	70.19%	11.52%	0	TTAGAGAACAGATCTACAAGAG	UUAGAGAACAGAUCUACAAGAG
NC_045512v2	48	70	crRNA16	390	70.56%	11.39%	0	TTTAGAGAACAGATCTACAAGA	UUUAGAGAACAGAUCUACAAGA
NC_045512v2	49	71	crRNA17	320	71.39%	11.39%	0	GTTTAGAGAACAGATCTACAAG	GUUUAGAGAACAGAUCUACAAG
NC_045512v2	50	72	crRNA18	320	71.67%	11.42%	0	CGTTTAGAGAACAGATCTACAA	CGUUUAGAGAACAGAUCUACAA
NC_045512v2	51	73	crRNA19	370	72.03%	11.46%	0	TCGTTTAGAGAACAGATCTACA	UCGUUUAGAGAACAGAUCUACA
NC_045512v2	52	74	crRNA20	430	72.03%	11.46%	0	TTCGTTTAGAGAACAGATCTAC	UUCGUUUAGAGAACAGAUCUAC
NC_045512v2	186	208	crRNA64	630	93.56%	1.56%	0	CGGACGAAACCGTAAGCAGCCT	CGGACGAAACCGUAAGCAGCCU
NC_045512v2	187	209	crRNA65	580	94.20%	1.56%	0	ACGGACGAAACCGTAAGCAGCC	ACGGACGAAACCGUAAGCAGCC
NC_045512v2	188	210	crRNA66	580	94.11%	1.56%	0	CACGGACGAAACCGTAAGCAGC	CACGGACGAAACCGUAAGCAGC
NC_045512v2	189	211	crRNA67	600	94.57%	12.15%	0	ACACGGACGAAACCGTAAGCAG	ACACGGACGAAACCGUAAGCAG
NC_045512v2	190	212	crRNA68	560	94.57%	12.22%	0	AACACGGACGAAACCGTAAGCA	AACACGGACGAAACCGUAAGCA
NC_045512v2	191	213	crRNA69	440	94.66%	12.22%	0	CAACACGGACGAAACCGTAAGC	CAACACGGACGAAACCGUAAGC
NC_045512v2	192	214	crRNA70	360	94.57%	12.12%	0	GCAACACGGACGAAACCGTAAG	GCAACACGGACGAAACCGUAAG
NC_045512v2	193	215	crRNA71	330	94.57%	12.12%	0	TGCAACACGGACGAAACCGTAA	UGCAACACGGACGAAACCGUAA
NC_045512v2	194	216	crRNA72	350	94.57%	12.12%	0	CTGCAACACGGACGAAACCGTA	CUGCAACACGGACGAAACCGUA
NC_045512v2	195	217	crRNA73	470	94.11%	1.68%	0	GCTGCAACACGGACGAAACCGT	GCUGCAACACGGACGAAACCGU
NC_045512v2	196	218	crRNA74	550	94.02%	1.68%	0	GGCTGCAACACGGACGAAACCG	GGCUGCAACACGGACGAAACCG
NC_045512v2	197	219	crRNA75	510	94.02%	1.68%	0	CGGCTGCAACACGGACGAAACC	CGGCUGCAACACGGACGAAACC

Figure 4.2.2b. The data as viewed on the GalaxyProject data integration platform

Research groups at McGovern Institute and MIT have speculated the use of a paper-based test— Specific High Sensitivity Reporter unLOCKing (SHERLOCK) that follows the Cas13-CRISPR slicing principle for quick identification of infected patients. The test requires use of paper strips dipped in the processed sample (containing RNA) obtained from patients, can be treated with species specific CRISPR-associated Cas13, Cas12a and other enzymes to obtain a clear signaling readout. [21]

Diagnostic research like this requires data bulk sequence-annotation data that has been assorted into various tracks and tables. Thus, the Table Browser serves as a powerful tool for this research and is rooted in up-to-date and periodically refurbished data warehouses.

4.3. BLAT: BLAST-like alignment tool

BLAT— *BLAST-like Alignment Tool* was developed by W. James Kent to reduce the time required by BLAST to align sequences or protein. [22]

Both the algorithms serve to align transcript data, DNA sequences and protein Amino Acid sequences against whole genome/exome sequences. BLAT algorithm has improved efficiency and can deliver outputs more intuitively. According to the available information at UCSC FAQ, to trigger the alignment of DNA (at least 40 bases long) the algorithm needs a similarity index of about 95% and to trigger polypeptide/protein alignment (of more than 20 Amino Acids) against the genome sequence, a similarity index of 80%. [22]

BLAT as developed by Kent, indexes the database and then traverses through the suspected sequence rather than indexing the sequence itself (as in case of BLAST). This improves the speed greatly, however indexing an entire database is a time-consuming process. Hence, the hosting server keeps a copy of the indexed database at ready. A direct visit to the server solves that issue. The interactive server can be accessed at the UCSC Genome Browser interface.

The BLAT algorithm also allows for display of high scoring sequences sewed together using multiple perfect or near perfect alignment results. When comparing DNA sequences of two close species, the regions that match homologously are also displayed as a complete alignment rather than a list of matching sequences. BLAT algorithm is effectively more intuitive as it is able to virtually recreate precursor mRNA with accurately positioned splice sites, thus giving an unbroken mRNA transcript of the genome for further alignment purposes. [22][23]

BLAT can perform inter-species genome-genome alignments, protein-genome alignments and also at times protein-protein alignments. The efficiency of the tool starts tanking at a little under 2000,000 nucleobases, hences short reads queries are preferred. [22]

Notably it was composed for mRNA transcript and random sub-sequence transcript alignments and has been used for exon-intron site identification, identification of exomes and more. [23]

4.4. Custom Tracks

This is one of the Browser tools equipped to upload personal annotation data to the Browser and to assimilate tracks for comparison and analysis. It can be used for provisional display or become a permanent track group. The Browser allows the individual users to upload data and tracks for integration in the underlying database. The data file can be uploaded after specifying the species and assembly the data has been investigated with reference to. The user can either directly submit the data in the text box, or upload a file of any one of the compatible formats or the link can be posted in the text box. BigBed, BigWig, WIG, BAM, BarChart, BED, bedGraph, GFF, GTF are some of the compatible file formats. [11]

Once incorporated, these tracks are listed under User Tracks groups that have similar graphical interactive potential as the native tracks. [11]

4.5. Track Data Hubs

Track Data Hub feature was introduced in the 2011 update as part of the data decentralising initiative for making the database more expandable and adaptable. [24] The Track Hubs serve a similar purpose as the Custom Tracks tool. The Track Hubs also allows the users to upload their own data and graphically visualise it alongside the native Browser data.The Custom Tracks have limitations due to their distributional and configurational disparity leading to hitches in the graphical rendering for the users. [11]

The data hub cataloguing the data set files, is present outside the central data repository, in a server remote to it. The hub is accessible to the Browser via a web network and URL and doesn’t require a DAS server. This remote repository requires a few metadata files with the sequence-annotation data and corresponding descriptor files. These files define the attributes of the Track hubs, the attributes of the datasets and their rendering parameters as well as species and assembly data. [25]

The data— stored in compressed files in binary format— within the hubs follow the syntax and arrangement of the native data and can hence use all the tools and interfaces in exactly the same manner. It is also indexed similarly to keep the speed and efficiency optimal. The server doesn’t store a reserve copy of the tracks in the data hub at hand. All the queries are sent to the remote location and the pertaining information is sent back to the browser for the client. A cache copy could be stored for rapid output if the query is sent again. These features eliminated the issue of a glitchy graphical experience for the user. [25]

UCSC requires these data hubs to be identified and registered by the hub designers, who can choose to make either publicly (can be found in the ‘Public Hubs’ option) or semi-publicly (can be found under the ‘My Hubs’ option and can be shared in a particular organisation or research group) available. Once certified, the data is accessible to the table browser in the same way the central database is. [25][11]

Like Track Hubs, UCSC also supports an **Assembly Hub** feature which hosts whole genome sequence data off-site the central database and is accessible to the users with the same adaptability. [25]

4.6. Track Collection Builder

It is one of the most recent additions in the Browser tools. The Collection Builder, true to its name, allows users to group Browser tracks, Custom tracks, tracks from data hubs and graph tracks under a single named collection. These tracks can then be subjected to uniform display parameters and comparative scales and sorted according to need. [11][25]

4.7. *in silico* PCR

The *is*PCR tool allows the user to view the size (in base pairs) and sequence results of a theoretical Polymerase Chain Reaction. [11] The user can enter the *forward* and *reverse* primer sequences for an organism and corresponding assembly of their choice. The user must be careful about choosing the length of the primer itself. Care should also be taken to note the directionality of the reverse primer. The tool interface also provides an option to flip the reverse primer if it is ordered from 5’ end to 3’ end. The user may restrict the size of sequence the amplicon is to be tallied against, but that is optional.

Once data is submitted in the form of a query, the browser returns possible resultant PCR amplicons based on the target (RNA/DNA) chosen. The results contain the length of the predicted sequence in base pair units and the name of the assembly region based on the nucleobase order. The name of the assembly or gene is present as a hyperlink that redirects to the browser itself for graphical rendition.

The *is*PCR tool helps in ascertaining that use of the specific primer will not amplify in a manner that resembles any other sequence in the assembly. [11]

5. USE OF THE UCSC GENOME BROWSER IN RESEARCH TODAY

There have been many projects to which UCSC Genome Browser has extended its database and tools, for the needs of the users. TCGA (The Cancer Genome Atlas) Project is an integrated effort to evaluate experimental genomic data derived from various cancer patients. It aims to find homology in the origin of varied cancerous forms and thus understand cancer biology better.

The UCSC Xena Browser supports graphical rendering of personal and public genomic data available on data hubs. This has been a preferred tool among many cancer biologists due to the easy and intuitive availability of private host hubs. The Xena initiative is taken to understand cellular, biochemical and genomic data interrelatedness more transparently and logically. [26]

It is unique in the sense that it allows comparison of data tracks and data sets that are rendered differently by virtue of their storage format. The data can be uploaded and compared against the database tracks and can also be shared among a select group of people if required. [27]

Trends in Systems Biology for cancer research have pointed towards the use of specific gene data, whole genome data and biomolecular data, for a holistic view. UCSC Xena Browser supports the visualisation of specific gene data and genomic regions related to cancer associated annotations data. The Browser helps in viewing data like variations in nucleotide repeat sequences, point mutations (insertion, duplication and deletion data), allelic variations, gene markers, genetic and epigenetic variant permutations, gene expressions, data regarding long stretches of non-coding RNA, occurring in genomes of affected individuals. [28][27] The Browser also differentiates between cell/tissue specific response of all the variations and mutations that cause tumorous growth. Biologically and clinically relevant data for all subtypes of data enables the users to study cancer biology in context of prognosis. [28]

The PCAWG (Pan-Cancer Analysis of Whole Genomes) initiative is a very important initiative in the right direction. The host, International Cancer Genome Consortium has set out to make history by studying nearly 3000 cancer-genomes, sprawling over more or less 40 tumour types and their healthy counterparts. [26] UCSC Xena lends itself to the analysis by providing a primary visualisation interface for a variety of data including: CNVs, expression models, signature aberrations and mutations, SNP data, corresponding phenotypic data, differential cell/tissue reactions etc.

Kaplan–Meier plots, more specific graphical depictions like histograms, ideograms etcetera, on comparative fronts to provide an in-depth understanding of the nuances of origin and spread of tumours. [29]

The Xena browser, as of July 2018, has data hubs worth more than 1500 tracks of about 35 different cancer targets. [28]

6. CONCLUSION AND FUTURE PLANS OF ADVANCEMENTS

Above all, the browser aims to amass sequence assembly data from a range of species to make genomic enquiry more complete in its truly phylogenetic form.

Sequenced genomes provide raw data for exhaustive exploration of annotation data. This serves as a basis for gap-filling, resolution enhancement, comparative tracks and investigation of more annotation data.

When the Browser was first instantiated, the creators aimed to make analysis of data, especially in human and model organism genomes more flexible and more universal. [11] The software is frequently amended to meet the demands of increasingly efficient high throughput sequencing techniques. The developers continually work on emulating the same resolution and clarity to long range chromosome data as possessed by point data related to genes. A number of online resources and tutorials are available for users of all levels of proficiency.

The study of genomics at this scale, precision and ingenuity confers a better understanding of human diseases and ailments and motivates us to come up with better, more permanent solutions for them.

In future iterations of upgrades, clinical and medical data—with tracks associating gene or nucleotide variations to phenotypic expression or diagnostic data, will be available. [19]

REFERENCES

1. Genoscope (archived October 2013); *The Human Genome Project FAQ* (retrieved February 2015)
2. Schattner P. (2008); "*Genome, Browsers and Databases: Data-Mining tools for Integrated Genomic Databases*"; Chapter 1: The Molecular Biology Data Explosion
3. Breda G. (2016); "*Sequencing library: what is it?*"
4. Azevedo Vasco, Barh Debmalya, da Silva A.L.da Costa, das Graças D.A., de Sá Pablo H.C.G., Guimarães L.C., Ramos R.T.J., Veras A.de Oliveira (2018); "*Omics Technologies and Bio-Engineering: Towards Improving Quality of Life*"; Chapter 11: Next-Generation Sequencing and Data Analysis: Strategies, Tools, Pipelines and Protocols
5. Huh S., Kong J., Kim B., Kim K., Won J., Yoon J. (June 2019); "*GAAP: A Genome Assembly + Annotation Pipeline*"
6. Foxman B. (2012); "*Molecular Tools and Infectious Disease Epidemiology*"; Chapter 5: A Primer of Microbiology
7. Karolchik D, Lathe III W.C, Mangan M.E., Williams J.M. (2008); "*Genomic Data Resources: Challenges and Promises*"
8. Brunak S., Danchin A., Hattori M., Matise T., Nakamura H., Preuss D., Shinozaki K. (2002); "*Nucleotide Sequence Database Policies*"; Policies of the International Nucleotide Sequence Database Collaboration
9. Gao G., Kong L., Luo J., Wang J. (revised 2013); "*A brief introduction to web-based genome browsers*"
10. Haeussler M., Zweig A.S., Tyner C., Speir M.L., Rosenbloom K.R., Raney B.J., Lee C.M., Lee B.T., Hinrichs A.S., Gonzalez J.N., Gibson D., Diekhans M., Clawson H., Casper J., Barber G.P., Haussler D., Kuhn

R.M., Kent W.J. (2019); "*The UCSC Genome Browser database: 2019 update*"

11. Kent W.J., Kuhn R.M., Haussler D. (2013); "*The UCSC genome browser and associated tools*"
12. Baertsch R., Diekhans M., Furey T.S., Haussler D., Hinrichs A., Karolchik D., W. J. Kent, Lu Y.T., Roskin K.M., Schwartz M., Sugnet C.W., Thomas D.W., Weber R.J. (2003); "*The UCSC Genome Browser Database*"
13. Samarjiva S. (July 2014); "*Genome Browsers*"
14. Furey T.S., Haussler D., Kent W.J., Pringle T.H., Roskin K.M., Sugnet C.W.,Zahler A.M. (2002); "*The Human Genome Browser at UCSC*"
15. Altschul S., Lipman D., Madden T., Miller M., Schäffer A., Zhang J., Zhang Z. (1997); "*Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*"
16. Barber G.P., Clawson H., Harte R.A., Huassler D., Hillman-Jackson J., Hinrichs A.H., Karolchik D., Kent J., Kuhn R., Raney B.J., Rhead B.L., Rosenbloom K.R., Smith R., Thakkapallayil A., Thomas D.J., Trumbower H., Zweig A.S. (2008); "*The ENCODE Project at UC Santa Cruz*"
17. Barber G.P., Dreszer T.R., Fujita P.A., Haussler D., Hinrichs A.S., Karolchik D., Kent W.J., Kuhn R.M., Learned K., Meyer L.R., Pheasant M., Pohl A., Raney B.J., Rhead B., Rosenbloom K.R., Smith K.E., Wang T., Zweig A.S. (2012); "*ENCODE whole-genome data in the UCSC Genome Browser*"
18. Haussler D., Karolchik D., Kuhn R. M., Zweig A.S. (2008); "*UCSC genome browser tutorial*"
19. Furey T.S., Haussler D., Hinrichs A.S., Karolchik D., Kent W.J., Roskin K.M., Sugnet C.W. (2004); "*The UCSC Table Browser data retrieval tool*"
20. Trafton A. (February 2018); "*Researchers advance CRISPR-based tool for diagnosing disease*"
21. McGovern Institute (February 2020); "*Press Enabling coronavirus detection using CRISPR-Cas13: An open-access SHERLOCK research protocol*"
22. Kent W.J. (2002); "*BLAT—The BLAST-Like Alignment Tool*"
23. Bhagwat M., Robison R.R., Young L. (2012); "*Using BLAT to Find Sequence Similarity in Closely Related Genomes*"
24. Barber G.P., Clawson H., Cline M.S., Diekhans M., Dreszer T.R., Fujita P.A., Giardine B.M., Goldman M., Guruvadoo L., Harte R.A., Haussler D., Hinrichs A.S., Hsu F., Karolchik D.,Kent W.J., Kirkup V., Kuhn R.M., Learned K., Li C.H., Malladi V.S., Meyer L.R., Pohl A., Raney B.J., Rhead B., Roe G., Rosenbloom K.R., Sloan C.A., Wong M., Zweig A.S. (2011); "*The UCSC Genome Browser database: extensions and updates 2011*"
25. Barber G.P., Clawson H., Dreszer T.R., Fujita P.A., Karolchik D., Kent W.J., Nguyen N., Paten B., Raney B.J., Wang T., Zweig A.S. (2014); "*Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser*"
26. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (formed by collective efforts of 800 different university and research institutes) (February 2020); "*Pan-cancer analysis of whole genomes*"
27. Craft B., Goldman M., Haussler D., Zhu J. (2018); "*Cancer genomics visualization and interpretation using UCSC Xena*" [abstract]
28. Brooks A.N., Craft B., Goldman M., Hastie M., Haussler D., Kamath A., McDade F., Repečka K., Rogers D., Zhu J. (March 2019); "*The UCSC Xena platform for public and private cancer genomics data visualization and interpretation*"
29. Cline M.S., Craft B., Goldman M., Haussler D., Ma S., Swatloski T., Zhu J. (2013); "*Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser*"

Hinrich A.S., Karolchik D., Kent J.W. (2009); "*The UCSC Genome Browser*"