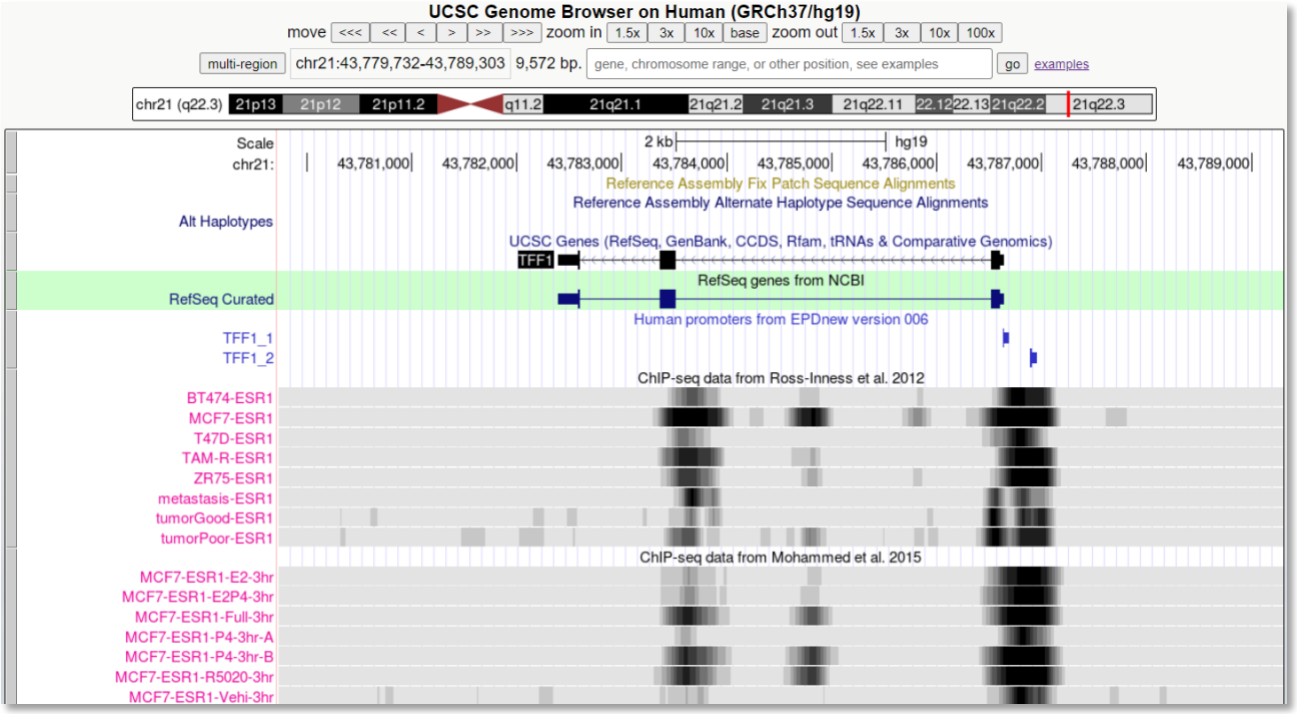
BREAST CANCER EPIGENOMICS TRACK HUB BEGINNER'S TUTORIAL (v1.0)

The Breast Cancer (BC) epigenomics track hub (or BC hub, https://bchub.epfl.ch/) is embedded in the UCSC genome browser (https://genome.ucsc.edu/) and integrates multiple breast cancer-related datasets for visualization of transcription factor (TF) cistromes and epigenomics datasets by means of a standardized pipeline that allows direct comparison among datasets. All datasets present in the BC hub are gathered from published studies and include epigenomics data from breast cancer cell lines, surgically-removed breast tumors, patient-derived xenografts, and normal breast tissue specimens. Additional tracks add further information about the genomic context such as annotated promoter regions, chromatin accessibility, TFs footprints, and sequence conservation.

Navigating the genome browser

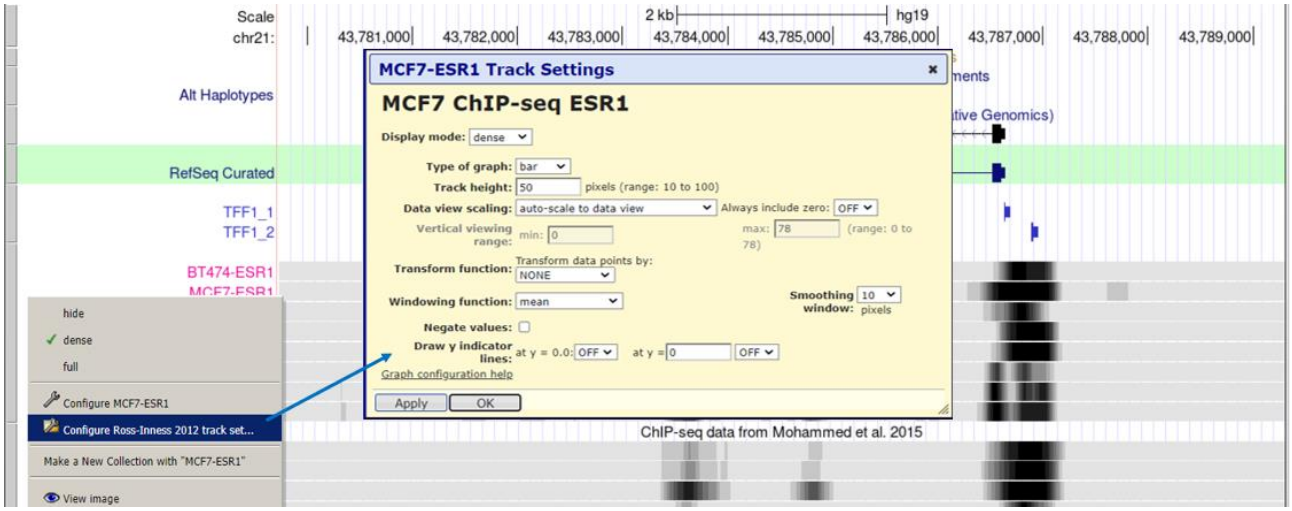
The BC hub is integrated in the UCSC genome browser and can be accessed from here -> https://bchub.epfl.ch/.

Located on the top right corner of the main browser page, the search bar can be used to type in specific genes or indicate the genomic coordinates of a desired region. For example, typing the following text "chr1:1,000,000-1,000,500" will make the browser display to a 500 bp wide region starting at chromosome 1 position 1,000,000. Regions of interest can be moved by click-and-drag in the browser window, or zoomed in and out using the zoom buttons. For further general instructions on how to use the UCSC genome browser, please refer to the UCSC genome browser user's guide found at https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html.



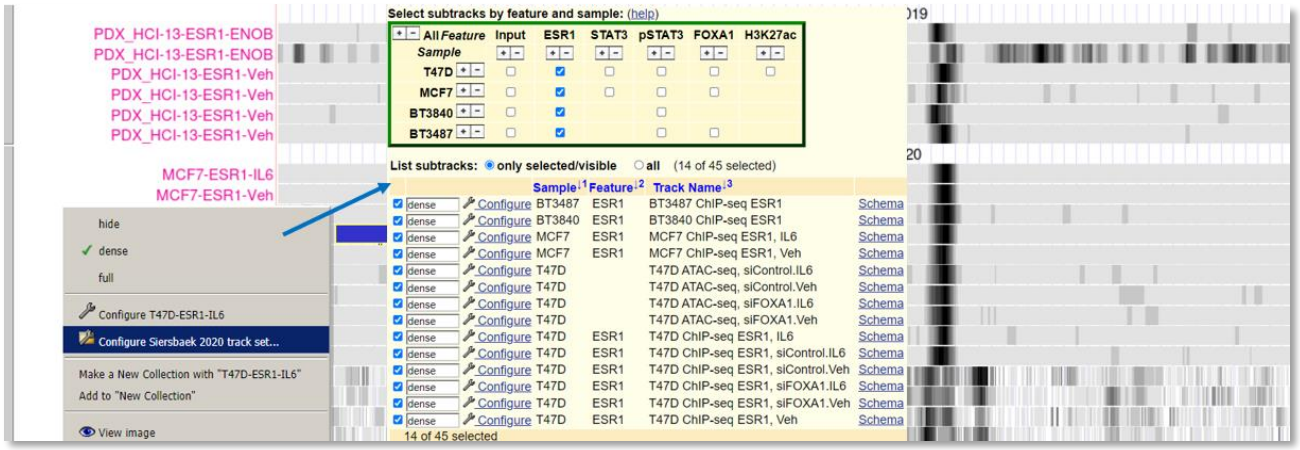
Customizing the track order and view

Tracks can be moved up or down to a desired position by dragging the left side of each track. Display settings for each data track can be individually changed by right-clicking on the track itself. By choosing to configure a single track, a pop-up window will appear (as shown in the snapshot here below) to allow users to select the desired display parameters. Individual tracks can be viewed in full or dense mode, or can be hidden from the genome browser (they can be retrieved at any point from the full list of available tracks). Further display settings such as the track height, scale, and smoothing window can be changed in the track configuration panel. More information about track settings can be found at https://genome-euro.ucsc.edu/goldenPath/help/hgWiggleTrackHelp.html



Configuring the track list

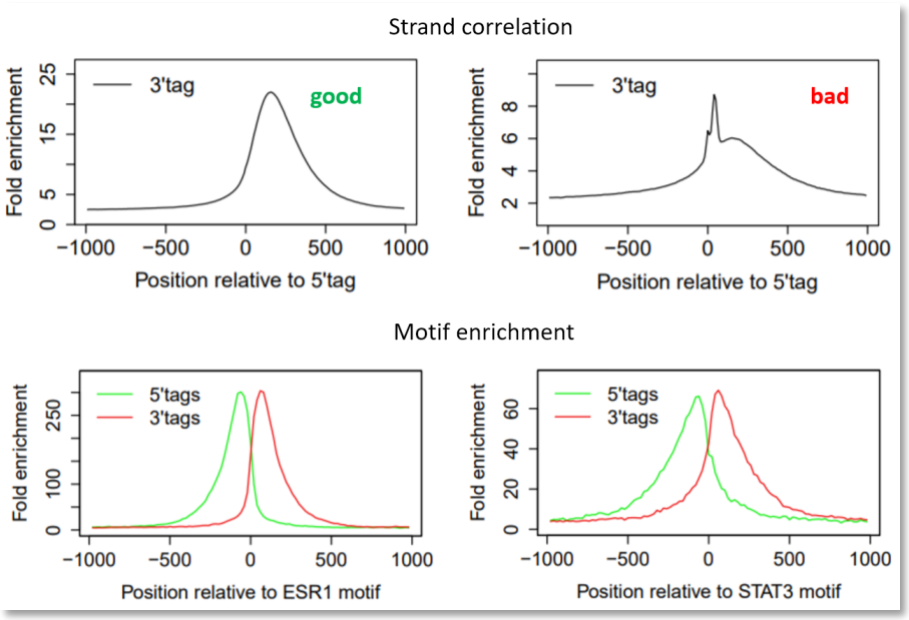
To configure an entire track set, users can select the "configure track set" option in the pop-up window. This will open a new page showing the full track setting options as well as a description of the data track. Sub-tracks from a given dataset can be selected in the track set configuration settings. Tracks can be selected/unselected individually in the bottom panel, or selected based on properties such as sample type and/or feature in the upper panel.



Retrieving sample information and quality control

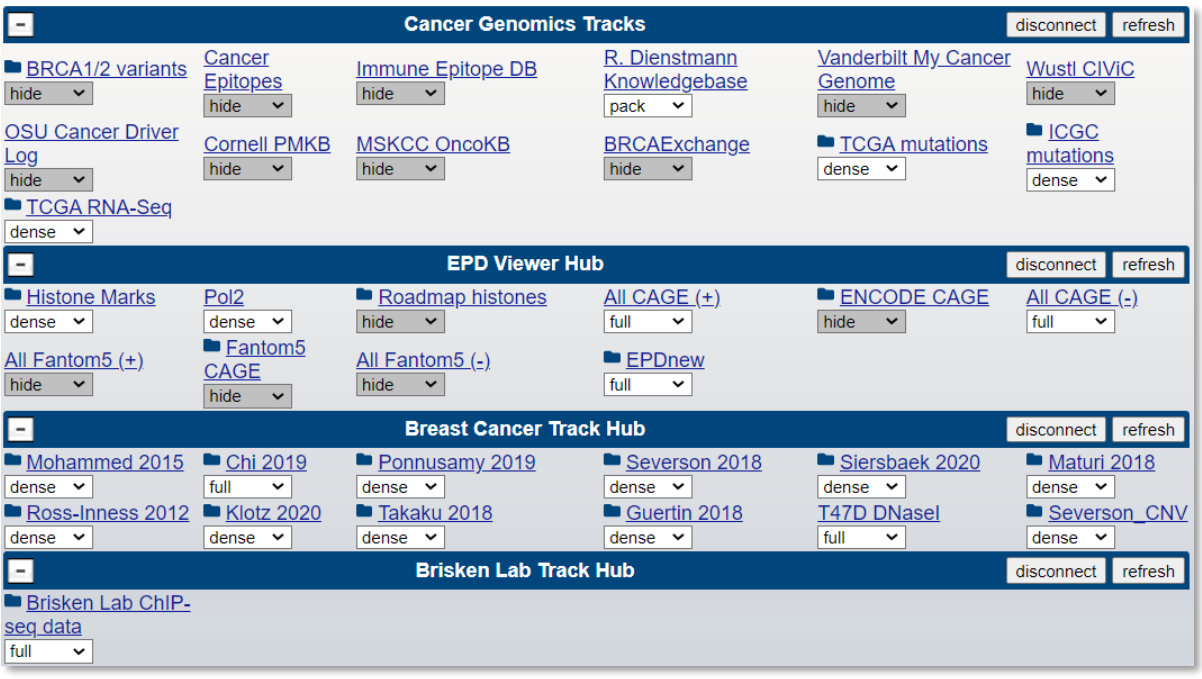
The description of the experiments, the pipeline used for data processing, and references to the original study can be found by scrolling down the track list configuration panel. Furthermore, quality control reports (according to ENCODE guidelines. PMID: 22955991 https://pubmed.ncbi.nlm.nih.gov/22955991/) can be viewed and downloaded for all the samples processed via the BC hub standard pipeline. The strand

correlation plots represent the distribution of the plus and minus strand reads respectively up-stream and down-stream of the protein-bound regions. The cross-correlation plot typically produces two peaks: a peak of enrichment corresponding to the predominant fragment length (highest correlation value) and a peak corresponding to the read length ("phantom" peak). Ideal distributions are narrow and sharp, while the presence of shoulders or less-defined shape of the distribution might reflect poor sample quality or technical issues (see examples below). The second quality control report, often named motif enrichment analysis, displays the distribution of positions of the plus and minus strand fragments relative to the sequence motif of the immunoprecipitated transcription factor (for ChIP-seq data). Failed experiment and inputs tend to have flat or noisy distributions.

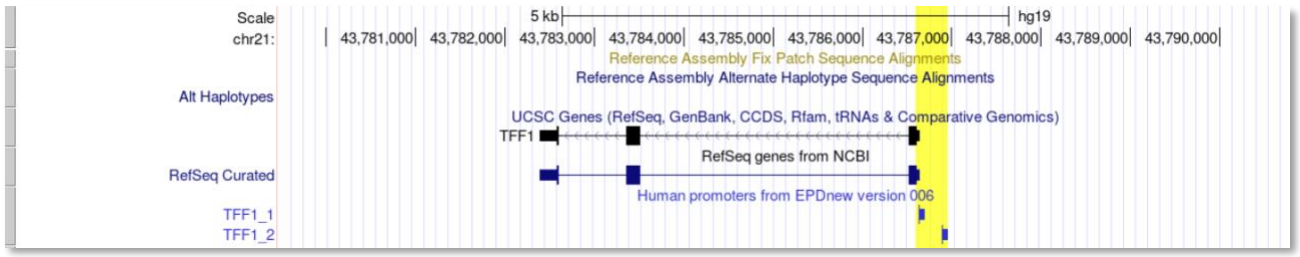


UCSC integrated tracks

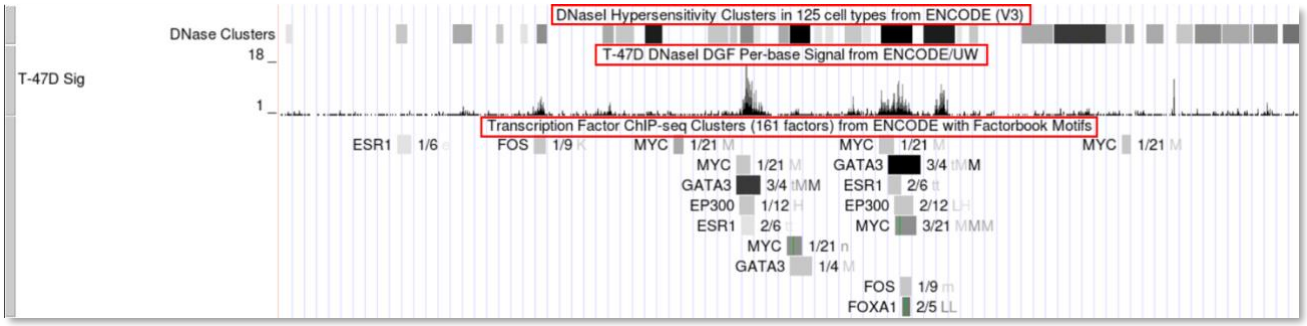
Additional tracks from various sources are included in the BC hub and can be toggled via the UCSC track manager menu (shown here below), which is located in the main browser page just below the displayed track sets.



- Eukaryotic Promoter Database (EPD) track (https://epd.epfl.ch//index.php https://academic.oup.com/nar/article/41/D1/D157/1070274) includes a collection 29598 human promoters that have been experimentally validated.



- ENCODE tracks (https://genome.ucsc.edu/ENCODE/): The Encyclopedia of DNA Elements (ENCODE), loaded by default in the UCSC genome browser, is an integrated database of functional regulatory elements in the human genome. BC hub sessions includes tracks from the ENCODE project. The "Transcription factors ChIP-seq clusters" tracks display transcription factor (TF) bindings derived from ChIP-seq experiments. Binding sites are represented by rectangular boxes, color-coded in grey scale according to the binding strength. A green vertical line appears in the TF binding box when the predicted sequence motif of the TF is recognized in the bound region, implying that the DNA binding is likely to be direct, rather than tethered by other proteins. Right-clicking on the TF binding box redirects the browser to the page containing metadata such as sample information, experimental condition, and sequence motif, if present. TFs displayed in the track can be filtered in the track configuration window. Other ENCODE tracks, such as "DNase Clusters" and "T-47D Sig" tracks, provide information on chromatin accessibility based on hypersensitivity to DNase I cuts. Peaks represent more accessible genomic areas.



- Conservation track (PhastCons): the sequence conservation track displays the degree of sequence conservation among 100 vertebrates species. Highly conserved sequences in non-coding regions are more likely to carry essential functions in genes regulation, thus are areas of interest in epigenetics studies.

What can you do with the Breast Cancer hub?

Here are few of the many biological questions that can be addressed by exploring the BC hub:

- *Is TF binding consistent across different breast cancer cell lines and/or tumor samples?*
- *Is TF binding dependent on hormone treatment?*
- *Are TF binding events shared between normal breast tissue and tumor, or between male and female breast cancer patients?*
- *Does the binding region correspond to an experimentally-validated promoter?*
- *Are there other TFs associated with the genomic region(s) of interest? Are there sequence motifs within the region(s) of interest?*
- *Is chromatin accessible or not in tissues or cell lines of interest?*
- *Are the binding regions of interest conserved across vertebrates (hinting at possible biological functions)?*