Published online 8 November 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D951–D959
doi:10.1093/nar/gkq1113

# The UCSC cancer genomics browser: update 2011

J. Zachary Sanborn[1], Stephen C. Benz[1], Brian Craft[1], Christopher Szeto[1], Kord M. Kober[1], Laurence Meyer[1], Charles J. Vaske[1], Mary Goldman[1], Kayla E. Smith[1], Robert M. Kuhn[1], Donna Karolchik[1], W. James Kent[1], Joshua M. Stuart[1], David Haussler[1,2,*] and Jingchun Zhu[1,*]

[1]Department of Biomolecular Engineering, Center for Biomolecular Science and Engineering and [2]Howard Hughes Medical Institute, University of California at Santa Cruz, Santa Cruz, CA 95064, USA

Received October 1, 2010; Accepted October 18, 2010

## ABSTRACT

**The UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu) comprises a suite of web-based tools to integrate, visualize and analyze cancer genomics and clinical data. The browser displays whole-genome views of genome-wide experimental measurements for multiple samples alongside their associated clinical information. Multiple data sets can be viewed simultaneously as coordinated 'heatmap tracks' to compare across studies or different data modalities. Users can order, filter, aggregate, classify and display data interactively based on any given feature set including clinical features, annotated biological pathways and user-contributed collections of genes. Integrated standard statistical tools provide dynamic quantitative analysis within all available data sets. The browser hosts a growing body of publicly available cancer genomics data from a variety of cancer types, including data generated from the Cancer Genome Atlas project. Multiple consortiums use the browser on confidential prepublication data enabled by private installations. Many new features have been added, including the hgMicroscope tumor image viewer, hgSignature for real-time genomic signature evaluation on any browser track, and 'PARADIGM' pathway tracks to display integrative pathway activities. The browser is integrated with the UCSC Genome Browser; thus inheriting and integrating the Genome Browser's rich set of human biology and genetics data that enhances the interpretability of the cancer genomics data.**

## INTRODUCTION

Cancer is a disease with both genetic and epigenetic causes. Cancer-associated alterations exploit many different molecular mechanisms that disrupt cellular pathways and result in uncontrolled cell proliferation (1–10). In recent years, development of high-throughput genomic technologies has propelled the cancer genomics field into a rapidly evolving discipline. Large genomic projects, such as The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) (11), generate comprehensive genome-wide data sets including DNA sequence variants, copy-number alterations (CNA), epigenetic and transcriptomic changes. Investigators attempt to combine genomics with clinical information on cancer samples to catalog genome alterations associated with human cancers, describe the genomic fingerprints associated with specific cancer types (12–18), predict response to therapies (19,20), and develop adaptive strategies to improve patient care (21,22).

Cancer genomics resources are growing at an unprecedented pace (23). However, a comprehensive analysis of the cancer genome remains a daunting challenge. This is in part due to the limitations in current technologies to visualize, integrate, compare and analyze cancer genomics data. Such limitations prevent investigators from truly appreciating the breadth and depth of these genomics and epigenomics resources. Just like cancer itself, cancer genomics data are complex and heterogeneous. Careful statistical and algorithmic considerations are required to integrate the information provided by the large volume and variety of data alongside clinical annotations. Ultimately, these data and the specific conclusions they support must be presented in a coherent system for display and analysis accessible to the scientific and medical communities.

*To whom correspondence should be addressed. Tel: +831 459 5232; Fax: +831 459 1809; Email: jzhu@soe.ucsc.edu
Correspondence may also be addressed to David Haussler. Tel: +831 459 2105; Fax: +831 459 1809; Email: haussler@soe.ucsc.edu

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

We have developed an open-access web-based tool called the UCSC Cancer Genomics Browser to facilitate the integrative, interactive and versatile display and the comprehensive analysis of cancer genomics and clinical data (24). This browser displays a whole-genome-oriented view of genome-wide experimental measurements for individual and sets of samples/patients as heatmaps. Clinical features are displayed alongside the genomics data in a separate but coordinated heatmap. Investigators interact with the browser to order, filter, aggregate and display data according to clinical features, annotated biological pathways or user-contributed collections of genes. Statistical analyses can be applied within data sets and displayed graphically on the browser. A rapidly expanding body of publicly available cancer genomics studies are organized into tracks of data on our public website. Investigators may use the browser over the web or install it locally on their servers, where security access controls allow user authentication and restricted access to underlying features of data sets when it is appropriate to protect patient privacy or when investigators wish to control access to their data.

An increasing number of public data sets are curated and hosted on the website. A significant proportion of new genomics data comes from the TCGA project, which is producing a comprehensive knowledge base of cancer-specific genomic aberrations for the oncology community. These data will facilitate novel translational approaches and ultimately accelerate the development of new cancer diagnostic, prevention and treatment strategies. The full-scale project aims to generate data on approximately 25 different types of cancers. Currently, we host data from the open access tier consisting of gene expression, DNA copy number, DNA methylation, miRNA, somatic mutation tracks as well as the associated clinical data for 467 TCGA glioblastoma multiforme (GBM) samples (282 tumor, 149 blood-normal, 36 solid-normal) and 1081 serous ovarian cystadenocarcinoma (OV) samples (543 tumor, 410 blood-normal, 128 solid-normal). PARADIGM pathway analysis was developed at UCSC (25) to integrate multi-dimensional data such as those from TCGA. The cancer browser hosts PARADIGM pathway tracks for TCGA GBM and OV samples on the public portal.

Since the cancer browser's initial launch, we have implemented several new features and upgrades to enable a more in-depth investigation of additional data linked to the cancer samples. These features include an image viewer called hgMicroscope and a genomic signature calculator and visualization feature called hgSignature. Several visualization improvements are also implemented to provide more versatile browsing capabilities such as a drag-to-zoom chromosomal view, probe-level mean normalization, a new summary view and custom tracks.

A video tutorial and user's guide provide an introduction on using the UCSC Cancer Genomics Browser, both accessible on the browser web site.

## NEW DATA

### TCGA open-access data tracks

The UCSC Cancer Genomics Browser provides a public portal to visualize, analyze and access TCGA data. We obtain Level 3 genomic data as well as associated clinical information from the TCGA Data Repository (http://tcga-data.nci.nih.gov/tcga/). These data are processed and stored in the UCSC Cancer Genomics Database from which cancer browser tracks are built. After quality assurance evaluation, the tracks are published on the cancer browser portal. We are working towards an automated processing pipeline that will support monthly data updates from TCGA Data Repository.

Each cancer browser track has two components: a single type of genomic data, such as array-based gene expression measurements, on a set of samples and the associated sample clinical information. For example, the 'TCGA OV Hudson Alpha 1M Duo Copy Number' track is the copy number variation data of 512 TCGA ovarian tumor samples assayed using the Illumina 1M-Duo DNA Analysis BeadChip at Hudson Alpha and accompanied by the corresponding clinical information for these samples (Figure 1). The public browser as of September 2010 hosts TCGA data consisting of 14 ovarian tracks (2 gene expression, 4 gene expression tumor versus normal, 5 copy number variation, 1 DNA methylation, 1 miRNA, 1 PARADIGM activities and 42 clinical variables) and 7 GBM tracks (1 gene expression tumor, 1 gene expression tumor versus normal, 2 copy number variation, 1 DNA methylation, 1 miRNA, 1 PARADIGM activities and 26 clinical variables). The information is summarized in Table 1. Detailed information about each track, such as source data, sample number and available clinical data, can be viewed on the track detail page accessed by following the track's hyperlink. All available tracks are listed at the bottom of the browser page in a fashion similar to the UCSC Genome Browser, where individual tracks can be toggled on and off manually.

In addition to displaying TCGA Level 3 genomic data, we also produce tumor versus normal gene expression data tracks. For each tumor type, the TCGA project has transcriptomic data on a handful of adjacent normal samples that match the tissue of origin of the tumor type. For ovarian tumor and GBM, these normal samples are designated as 'solid tissue normal' (code 11 in tcga-data.nci.nih.gov/datareports/codeTablesReport.htm#codeTables). Gene expression data on both types of samples are generated by the same technology (e.g. array platform) and by the same genome characterization center (GCC). By removing the tissue-specific gene expression pattern, we can focus on the transcriptional differences in the tumor. To accomplish this, we first generate a normal expression profile for each gene expression profiling method (same platform and GCC) by taking the median of each gene's expression level measured in the solid tissue normal data set. Then, we use the appropriate normal profile to normalize the tumor expression data for each gene. Five tumor versus normal gene expression tracks for ovarian tumor and one for GBM are

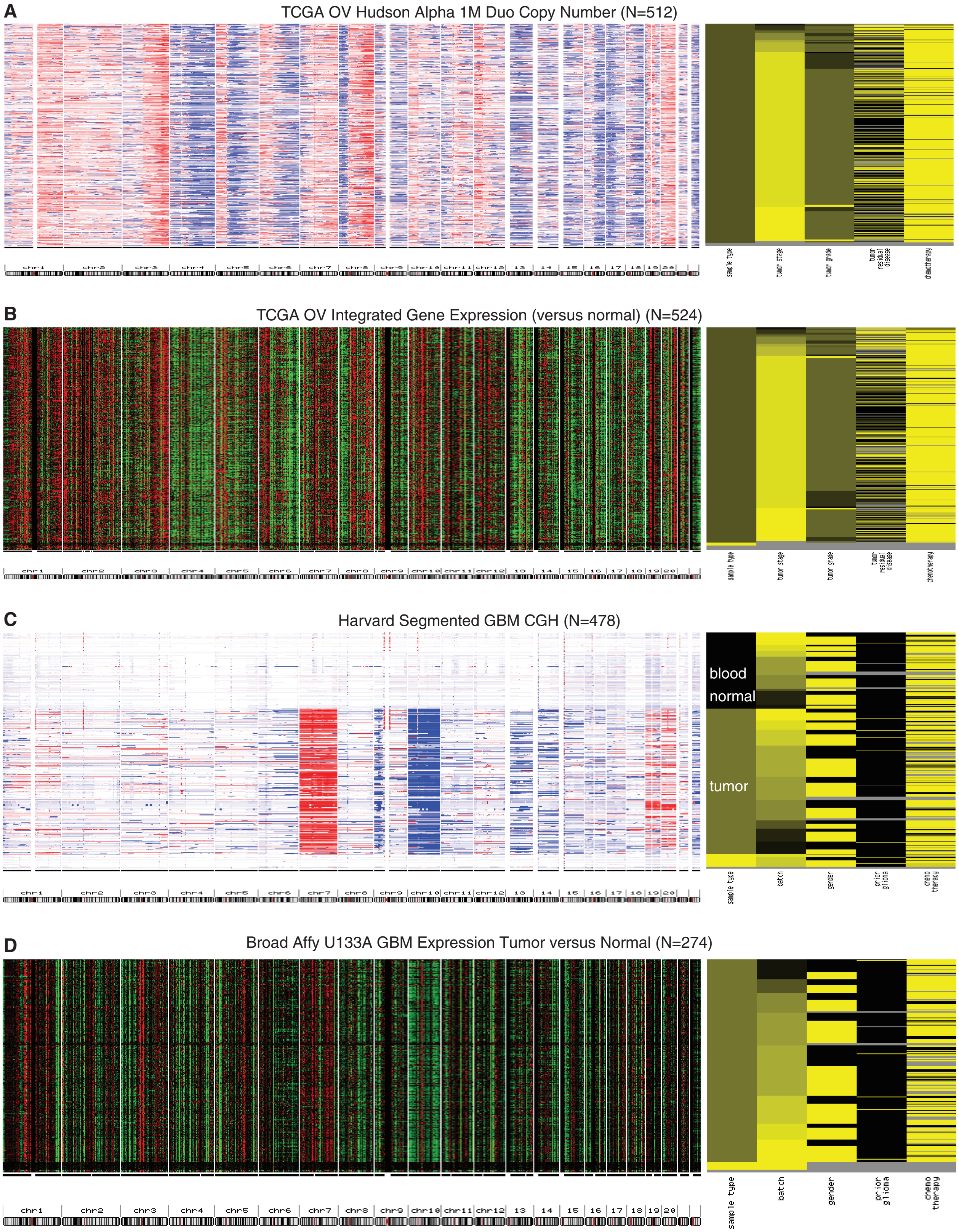


**Figure 1.** TCGA GBM and ovarian tumor gene expression and DNA copy number tracks. Copy number tracks by default use red and blue to represent amplification and deletion, respectively. Gene expression tracks by default use red and green to represent over- and under-expression, respectively. (**A**) TCGA ovarian copy number. (**B**) Ovarian tumor gene expression normalized by solid normal controls. (**C**) GBM copy number. (**D**) GBM gene expression normalized by solid normal controls. Accompanying clinical information is shown on the right of each genomics heatmap. Clinical values are coded in color and displayed as a yellow-black heatmap. The user can sort samples according to either the genomic or clinical heatmap by clicking on the feature of interest. For example, in (C) GBM copy number track, GBM samples are sorted by sample type in the order: blood normal, tumor and solid normal. The GBM genomic heatmap is organized according to the same clinical order, showing copy number abnormality in tumors, while such abnormality is mostly absent in blood normals and solid normals.

**Table 1.** UCSC Cancer Genomics Browser data track summary

| Cancer type | Gene expression | Tumor versus normal gene expression | Copy number | Somatic mutation | DNA methylation | miRNA exp | PARADIGM pathway |
|---|---|---|---|---|---|---|---|
| TCGA GBM | 1 (274) | 1 (274) | 2 (911) | | 1 (265) | 1 (228) | 1 (230) |
| TCGA ovarian | 2 (1050) | 4 (2104) | 5 (3685) | 1 track to be released | 1 (563) | 1 (295) | 1 (489) |
| Breast | | | 5 (514) | | | | |
| Brain | 9 (1584) | | 2 (217) | | | | |
| Colon | 1 (105) | | | | | | |
| Leukemia/Lymphoma | 2 (432) | | 3 (363) | | | | |
| Lung | 2 (205) | | 1 (383) | | | | |
| Melanoma | 1 (95) | | 1 (101) | | | | |
| Ovarian | 1 (285) | | 1 (118) | | | | |
| Pancreas | 1 (107) | | 2 (52) | | | | |
| Multi-tissue | | | 1 (302) | | | | |
| COSMIC | | | | 2 (76 tissues) | | | |
| NCI60 | 1 (60) | | 1 (60) | | | | |
| Mouse | | | 2 (142) | | | | |

Number of tracks by cancer type and data type; number of samples is in parenthesis.

available on the cancer browser. Figure 1 illustrates the whole-genome view for the DNA copy number and gene expression (tumor versus normal) data on TCGA ovarian and GBM tumor samples. The genome-wide copy number profile is strikingly different between the two types of cancers: the ovarian tumor shows large-scale copy number abnormalities distributed over the entire genome, while in GBM such abnormalities are largely localized to certain arms, especially chromosome 7, 10, 13q, 14q and 9p (11). DNA copy number and gene expression data are visibly correlated at the level of whole chromosomal arms in ovarian tumor samples. This can be seen as an amplification/deletion pattern (red/blue) in the copy number track correlated with the over/under expression pattern (red/green) in the gene expression track.

The browser will be updated with data from several other cancer types as they are generated by TCGA. The next several cancer types expected are acute myeloid leukemia, breast carcinoma, colon adenocarcinoma, lung adenocarcinoma and lung squamous cell carcinoma.

## PARADIGM pathway analysis tracks

A major challenge in interpreting high-throughput multianalyte genomic data sets such as those produced by TCGA is data integration and/or data interpretation in the context of biological pathways. In this context, PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models) was developed to infer the activities of genetic pathways by integrating any number of functional genomic data sets in a patient sample in the context of genetic pathways (25). Using the PARADIGM framework one can virtually encode any genetic pathway and its underlying interactions. Multianalyte genomics data within each gene are then integrated by statistical modeling of the central dogma, i.e. DNA copy number controls RNA expression, which in turn controls protein level and activity. Genetic interactions between genes and complexes are also encoded to capture the information flow between molecular entities. PARADIGM implements a standard factor graph inference method to compute sample-specific activities for each pathway entity.

We have developed an automated pipeline to run PARADIGM analyses using data stored in the cancer browser database. The activities from each PARADIGM analysis are stored in a database and represented as an individual track in the cancer browser. On the public portal, we currently host two PARADIGM activity tracks for the TCGA project, one GBM and one OV, which were generated using copy number and gene expression (tumor versus normal) data sets as input data. Each track has around 8000 activities computed from 135 pathways derived from the NCI Pathway Interaction Database (26). Currently, PARADIGM tracks can be viewed using specialized PARADIGM genesets that list all entities in a pathway. Users can find all PARADIGM genesets by searching for the keyword 'paradigm' in the existing genesets search interface. Red or blue color is used to represent active or repressed activity levels, respectively. For example, 'paradigm44_HIF-1-alpha transcription factor network' is shown to be highly active in a large number of the TCGA GBM tumor samples (Supplementary Figure S1A).

Efforts are currently underway to further automate this pipeline enabling all appropriate data sets to be processed by PARADIGM on a regular basis as the software and pathway database are updated.

## Additional tracks

The public cancer browser hosts a large collection of data tracks in addition to the TCGA data sets. Breast cancer studies are currently the most prevalent in our collection (14 tracks), including three multianalyte studies each with a pair of copy number and gene expression tracks on the sample cell lines or tumor samples (27–30). Other tracks of note include the 'Hess Chemo Gene Exp' study that focuses on chemotherapy response prediction (20), the 'Miller TP53 Gene Exp' study on TP53 mutation associated gene expression profiles (31), and five studies on breast cancer prognostic predictions (31–35). The public portal also hosts a number of lung, blood, skin, brain, ovarian, pancreatic cancer tracks, two COSMIC tracks (mutation frequency and count per tissue), a pair of NCI60 tracks (gene expression and copy number) and two mouse study tracks where the mouse genomics data were translated into human coordinates for display and analysis. The detailed summary of tracks and samples by cancer type and data type is described in Table 1.

## NEW FEATURES

### hgMicroscope

hgMicroscope a large tumor image viewer based on VisiGene from the UCSC Genome Browser (36). hgMicroscope allows users to zoom and pan across detailed tumor histology images interactively in real time by preprocessing the large tumor section images into layered tiles and using a javascript website to provide tiles on a service-on-demand basis. This provides a Google Maps-like browsing experience of pathology samples to enable real-time inspection of high-resolution images over the web. The novelty of hgMicroscope compared to other image viewers is its capability to search and view tumor images according to clinical variables. All images that fit the search criteria are listed, and users can quickly click through to view an individual image. Our image database stores these high-resolution pathology images as a set of image tiles across seven levels of abstraction, from low-resolution single image tiles all the way to full-resolution image tiles. We currently process all TCGA GBM and OV tumor images (Aperio-compatible files) downloadable from the TCGA DCC and store image tiles locally in our database. Users can access hgMicroscope and the TCGA images by clicking the 'Tumor Images' button at browser top banner. In addition, these TCGA tumor images can be accessed from any TCGA genomics track. After a user defines a subset of samples in the cancer browser, a single click takes the user to hgMicroscope displaying histology images from that specific subset of samples.

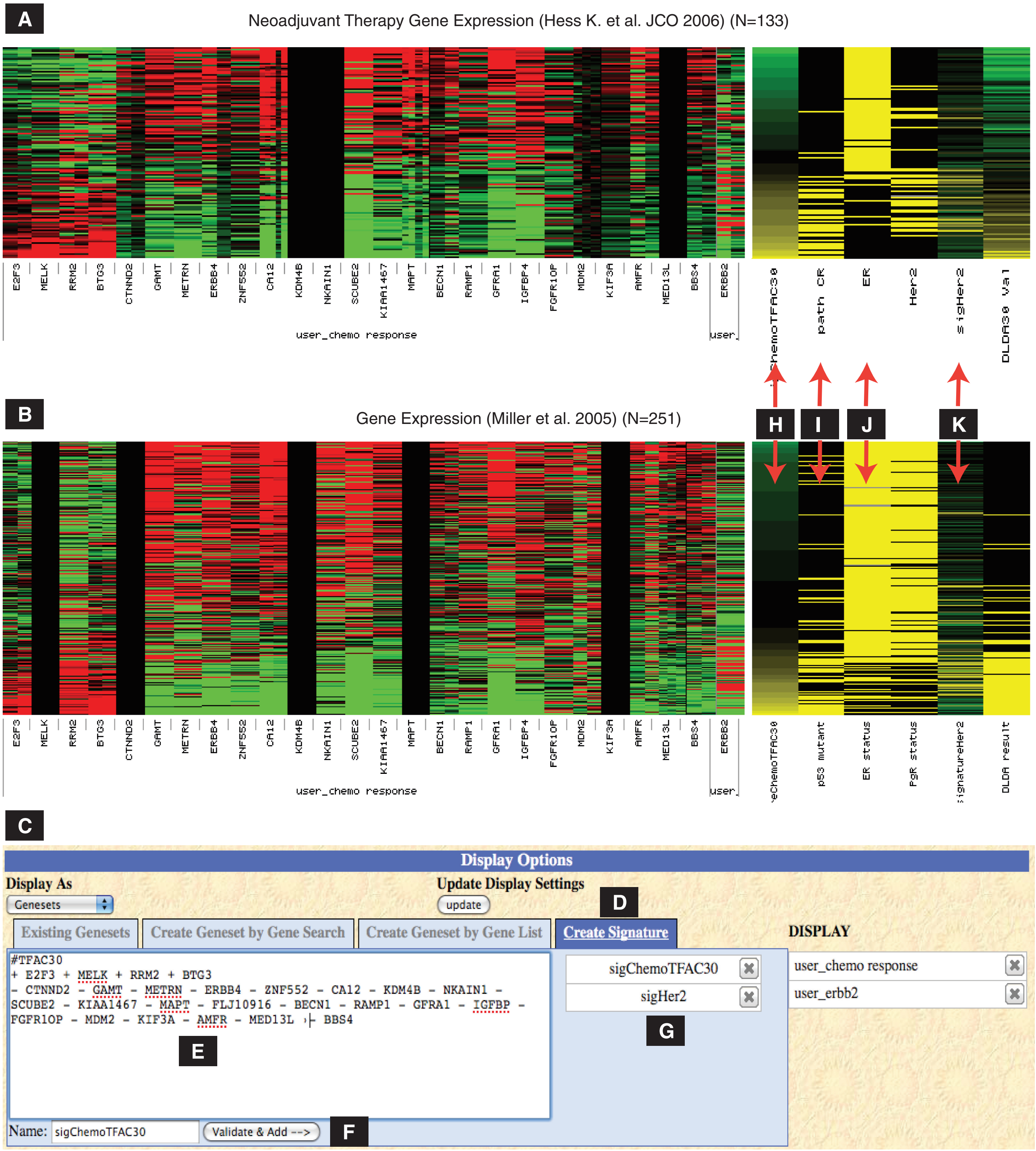
### hgSignature

Genomic technologies such as DNA microarrays measure levels of tens of thousands of genes in a single experiment. Results from such experiments have been used to uncover molecular signatures that could influence clinical care such as those for disease subtype, response to chemotherapy or disease outcome. Breast cancer researchers have developed multiple genomic signatures, mostly identified by microarray gene expression profiling (20,31–35,37,38). Many of these signatures are further developed into commercial products, such as Oncotype DX (Genomic Health) and MammaPrint (Agendia) (38).

hgSignature is a new cancer browser feature driven by the need to evaluate genomic signatures in real-time on any sample or study of interest. Genomic signatures are entered by users as mathematical formulas, such as '0.83*TP53—2*ESR1′. Here the gene names represent numerical values associated with these genes in a user-specified data set that is currently being displayed on the browser. Mathematical operators such as addition, subtraction, multiplication, division, power and inverse are implemented enabling a wide range of algebraic expressions. hgSignature computes the signature value on the fly for all samples in tracks that are being displayed in the cancer browser. The signature becomes a new clinical feature that can be sorted, grouped and analyzed by statistical tests or compared to other clinical features and genomic data. Figure 2 shows a browser screen shot of a chemotherapy response signature (supplementary data), derived from the Hess *et al.* study (20), and applied to two gene expression tracks. As a positive control, when we applied this signature to data from the original publication (Hess track, 2A), the signature score is highly correlated with path CR (pathologic complete response) in 2H, I. When applied to a second track, the Miller track in 2B (31), the derived signature feature also visually correlates with the TP53 mutation status (2H, I). The signature scores also correlate well with estrogen receptor status in both studies (2J). A quick browser exploration strongly indicates that these clinical variables (TP53 somatic mutations, chemo response, ER status and the genomic signature) are highly correlated. One can further subgroup the patients into high and low signature score subgroups, and use the statistical tools on the browser to ask what genes are differentially expressed between the two subgroups (Supplementary Figure S2).

hgSignature achieves four goals. First, it provides users with the capability to upload and evaluate a genomic signature in real-time without the burden of tedious bioinformatics data preprocessing. Second, it allows exploration of relationships between signatures and other clinical variables, e.g. the correlation of a prognostic signature with TP53 mutation status. Third, it facilitates comparisons of different signatures on the same samples or the same signature across multiple studies. Finally, it assists in genome-wide identification of differential genes based on a signature.

Genomic signature predictions have been shown to be sensitive to assay platform, experimental protocol, patient cohort and data processing methods (39). Signatures that are used commercially or are under clinical investigation typically use strict protocols to control for systematic biases as much as possible, from sample preparation to data preprocessing, before signature values are computed. For this reason, caution must be used when interpreting the results from user-defined signatures. They are not intended to be a replacement for well-developed, carefully controlled and clinically validated diagnostic signatures. Instead they provide a convenient, fast and interactive means to formulate and explore hypotheses prior to a more rigorous validation, taking advantage of the genomics data that are already curated on the browser portal. As hgSignature is expanded to interact with various statistical and machine learning methods that can refine existing signatures or build their own signatures according to user-defined



```
#TFAC30
+ E2F3 + MELK + RRM2 + BTG3
- CTNND2 - GAMT - METRN - ERBB4 - ZNF552 - CA12 - KDM4B - NKAIN1 -
SCUBE2 - KIAA1467 - MAPT - FLJ10916 - BECN1 - RAMP1 - GFRA1 - IGFBP -
FGFR1OP - MDM2 - KIF3A - AMFR - MED13L - BBS4
```

Name: sigChemoTFAC30

sigChemoTFAC30

sigHer2

user_chemo response

user_erbb2

**Figure 2.** hgSignature screen shot showing application of a user-defined signature to two gene expression tracks in the browser. (**A**) Hess track. (**B**) Miller track. (**C**) hgSignature user interface, under Genesets view. (**D**) Clicking the 'Create Signature' tab to enter hgSignature. (**E**) Text input signature, showing an example. (**F**) Naming the signature and clicking 'Validate and add' to add the signature. (**G**) List of available signatures, clicking to recall the content of a specific signature. (**H**) Chemo response signature score is automatically computed and added to the track clinical heatmap, and used as a clinical feature. (**I**) Signature score in H correlates with pathologic complete response (top) and TP53 status (bottom). (**J**) Signature score correlates with ER status (top and bottom). (**K**) Use hgSignature (signature = ERBB2) to pull out ERBB2 gene expression as a new clinical variable, which can be used to substitute Her2 Status when clinical status is not available. One can also subgroup samples based on ERBB2 gene expression or any genomic data.

goals, we expect this facility to become a powerful tool for cancer genome data exploration.

### Proportional summary view

The browser provides three display modes for each data track: heatmap view, box plot summary view and proportional summary view. In proportional summary view, genomic data under each probe/gene or clinical feature are sorted in ascending order (Supplementary Figure S3A). Proportion view is a simple yet powerful visualization modality. For example, the focal amplification on chromosomal 7 is readily visible in TCGA GBM copy number data under the whole genomic view, and zooming to chr7p11 confirms that it is in fact the focal amplification of the region containing EGFR (Supplementaary Figure S3B).

### Chromosomal view navigation

Exploring data in chromosomal view has been improved by replacing the previous click-to-zoom method with drag-to-zoom that allows users to quickly zoom to a region of interest (Supplementary Figure S3A). The clicking operation is now reserved for sorting genomic data. We also provide a search by gene name to jump to the gene's locus within chromosome view (Supplementary Figure S3C).

### Heatmap view customization

We have implemented two new heatmap settings to assist users in customizing track visualization. The 'gain' setting allows the user to change the heatmap color saturation in real time. Each track begins at an optimized gain setting determined during data wrangling and curation, but users may adjust the gain to produce brighter or dimmer heatmaps. As an added benefit, the gain parameter can be used to gauge the overall genomic signal for each study: arrays with less overall signal tend to produce dimmer heatmap images relative to others with equal gain settings. The second display setting is a probe-level data normalization. A subset of microarray gene expression data available in the public domain is provided without probe-level normalization, in particular studies that utilize the Affymetrix array platform. Without probe normalization, gene expression tracks often show vertical red and green stripes, representing both biological and array probe biases. Without appropriate control samples such as the solid normal samples used in the TCGA project, one cannot differentiate biologically relevant patterns from patterns introduced by probe design or hybridization efficiency. A commonly used practice to remove these biases is probe-level normalization. To this end, we have implemented the ability to perform dynamic mean-normalization across each probe. Probe-level normalization provides a simple way for users to more accurately compare data between data sets on the browser. These heatmap settings are easily accessed by clicking on a gear symbol next to the track title (Supplementary Figure S3D).

### Statistical tests

Eight statistical tests are now available for comparing genomic data between two subgroups of samples, including student's *t*-test, Wilcoxon test and Fisher's Exact Test. Detailed information on each statistical test can be found at https://genome-cancer.soe.ucsc.edu/cancerGenomics/stats-info/. In association with these tests, multiple hypotheses *P*-value adjustments have been implemented including Bonferonni and Benjamini-Hochberg false discovery rate (FDR) correction.

### Custom tracks

The cancer browser now provides custom track support via the integrated UCSC Human Genome Browser custom track mechanism. Users can upload their own genomic data in UCSC microarray data format (BED15) by clicking on the 'Custom Tracks' tab on the top banner. After reloading the cancer browser, the custom data track is automatically displayed as one of the tracks in the cancer browser. An example custom track BED15 file is in the supplementary data. Due to patient privacy concerns, custom clinical data cannot be uploaded to the public portal at this time.

### Architecture of combined common and local databases

In addition to the public portal, the browser can be installed for confidential or private data for controlled, collaborative access. As examples, the browser has been deployed for the Stand Up to Cancer (SU2C) breast cancer dream team (https://genome-cancer-su2c.soe.ucsc.edu/) and the I-SPY consortium (http://tr.nci.nih.gov/iSpy). Potentially patient-identifiable information such as germline mutations, confidential clinical information and the nature of pre-publication data makes such private installations inevitable. The proliferation of private installations makes administration of databases supporting each private browser increasingly impractical. We have implemented a new distributed database architecture that stores public browser data in a centralized database accessible by all private cancer browser installations at UCSC while still allowing private data to be stored in separate, secure databases. On each private cancer browser, all public portal tracks are displayed seamlessly next to private data, allowing users to compare their data to the genomics data sets available on the public browser.

## FUTURE DIRECTIONS

We will continue to incorporate new and updated data from TCGA project and other studies. We will develop the cancer browser to meet the need to efficiently store, view and analyze next-generation sequencing data including those generated by the TCGA project. We will also develop new viewing capabilities that integrate data across tracks for multianalyte data, for example view copy number, gene expression, somatic mutation, DNA methylation and clinical data from the same set of samples side by side. We will implement additional statistical tools such as generation of Kaplan–Meier survival plots for selected subsets of patients.

## CONTACTING US

We have one public moderated mailing list for user support of the UCSC cancer browser: genome-cancer@soe.ucsc.edu

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Eric Weiler, Victoria Lin and Jorge Garcia for computer support, the UCSC Genome Browser staff for advice on integration and support of the UCSC Genome Browser, database and source code library, Dr Eric Collisson and Dr Laura Esserman for their advice on the cancer browser development.

## FUNDING

National Cancer Institute (TCGA 5R21CA135937-02, 5U24CA143858-02, 3U24CA143858-02S1); Stand Up To Cancer; California Institute of Quantitative Biosciences; Howard Hughes Medical Institute (to D.H.). J.Z.S. is an ARCS (Achievement Rewards for College Students) scholar. Funding for open access charge: The Howard Hughes Medical Institute.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Hahn,W.C. and Weinberg,R.A. (2002) Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer*, **2**, 331–341.
2. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
3. Hunter,T. (2000) Signaling–2000 and beyond. *Cell*, **100**, 113–127.
4. Levine,A.J. (1997) p53, the cellular gatekeeper for growth and division. *Cell*, **88**, 323–331.
5. Levine,A.J. and Broach,J.R. (1995) Oncogenes and cell proliferation. *Curr. Opin. Genet. Dev.*, **5**, 1–4.
6. Sherr,C.J. (1996) Cancer cell cycles. *Science*, **274**, 1672–1677.
7. Sherr,C.J. (2004) Principles of tumor suppression. *Cell*, **116**, 235–246.
8. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.
9. Weinberg,R.A. (1994) Oncogenes and tumor suppressor genes. *CA Cancer J. Clin.*, **44**, 160–170.
10. Brena,R.M. and Costello,J.F. (2007) Genome-epigenome interactions in cancer. *Hum. Mol. Genet.*, **16**, R96–R105.
11. McLendon,R., Friedman,A., Bigner,D., Van Meir,E.G., Brat,D.J., Mastrogianakis,M., Olson,J.J., Mikkelsen,T., Lehman,N., Aldape,K. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
12. Boehm,J.S., Zhao,J.J., Yao,J., Kim,S.Y., Firestein,R., Dunn,I.F., Sjostrom,S.K., Garraway,L.A., Weremowicz,S., Richardson,A.L. *et al.* (2007) Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell*, **129**, 1065–1079.
13. Brown,L.A., Hoog,J., Chin,S.F., Tao,Y., Zayed,A.A., Chin,K., Teschendorff,A.E., Quackenbush,J.F., Marioni,J.C., Leung,S. *et al.* (2008) ESR1 gene amplification in breast cancer: a common phenomenon? *Nat. Genet.*, **40**, 806–807; author reply 810–802.
14. Garraway,L.A., Widlund,H.R., Rubin,M.A., Getz,G., Berger,A.J., Ramaswamy,S., Beroukhim,R., Milner,D.A., Granter,S.R., Du,J. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.
15. Harada,T., Chelala,C., Bhakta,V., Chaplin,T., Caulee,K., Baril,P., Young,B.D. and Lemoine,N.R. (2008) Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene*, **27**, 1951–1960.
16. Haverty,P.M., Fridlyand,J., Li,L., Getz,G., Beroukhim,R., Lohr,S., Wu,T.D., Cavet,G., Zhang,Z. and Chant,J. (2008) High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer*, **47**, 530–542.
17. Parsons,D.W., Jones,S., Zhang,X., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Siu,I.M., Gallia,G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
18. Weir,B.A., Woo,M.S., Getz,G., Perner,S., Ding,L., Beroukhim,R., Lin,W.M., Province,M.A., Kraja,A., Johnson,L.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.
19. Climent,J., Dimitrow,P., Fridlyand,J., Palacios,J., Siebert,R., Albertson,D.G., Gray,J.W., Pinkel,D., Lluch,A. and Martinez-Climent,J.A. (2007) Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. *Cancer Res.*, **67**, 818–826.
20. Hess,K.R., Anderson,K., Symmans,W.F., Valero,V., Ibrahim,N., Mejia,J.A., Booser,D., Theriault,R.L., Buzdar,A.U., Dempsey,P.J. *et al.* (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, **24**, 4236–4244.
21. Chin,L. and Gray,J.W. (2008) Translating insights from the cancer genome into clinical practice. *Nature*, **452**, 553–563.
22. Esserman,L. (2004) Neoadjuvant chemotherapy for primary breast cancer: lessons learned and opportunities to optimize therapy. *Ann. Surg. Oncol.*, **11**, 3S–8S.
23. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Varambally,R., Yu,J., Briggs,B.B., Barrette,T.R., Anstet,M.J., Kincead-Beal,C., Kulkarni,P. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
24. Zhu,J., Sanborn,J.Z., Benz,S., Szeto,C., Hsu,F., Kuhn,R.M., Karolchik,D., Archie,J., Lenburg,M.E., Esserman,L.J. *et al.* (2009) The UCSC Cancer Genomics Browser. *Nat. Methods*, **6**, 239–240.
25. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.
26. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
27. Neve,R.M., Chin,K., Fridlyand,J., Yeh,J., Baehner,F.L., Fevr,T., Clark,L., Bayani,N., Coppe,J.P., Tong,F. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**, 515–527.
28. Chin,K., DeVries,S., Fridlyand,J., Spellman,P.T., Roydasgupta,R., Kuo,W.L., Lapuk,A., Neve,R.M., Qian,Z., Ryder,T. *et al.* (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.
29. Chin,S.F., Teschendorff,A.E., Marioni,J.C., Wang,Y., Barbosa-Morais,N.L., Thorne,N.P., Costa,J.L., Pinder,S.E., van de Wiel,M.A., Green,A.R. *et al.* (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, **8**, R215.
30. Naderi,A., Teschendorff,A.E., Barbosa-Morais,N.L., Pinder,S.E., Green,A.R., Powe,D.G., Robertson,J.F., Aparicio,S., Ellis,I.O., Brenton,J.D. *et al.* (2007) A gene-expression signature to predict

survival in breast cancer across independent data sets. *Oncogene*, **26**, 1507–1516.

31. Miller,L.D., Smeds,J., George,J., Vega,V.B., Vergara,L., Ploner,A., Pawitan,Y., Hall,P., Klaar,S., Liu,E.T. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.
32. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
33. van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.
34. Wang,Y., Klijn,J.G., Zhang,Y., Sieuwerts,A.M., Look,M.P., Yang,F., Talantov,D., Timmermans,M., Meijer-van Gelder,M.E., Yu,J. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.
35. Desmedt,C., Piette,F., Loi,S., Wang,Y., Lallemand,F., Haibe-Kains,B., Viale,G., Delorenzi,M., Zhang,Y., d'Assignies,M.S. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.
36. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E., Siepel,A. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
37. Paik,S., Shak,S., Tang,G., Kim,C., Baker,J., Cronin,M., Baehner,F.L., Walker,M.G., Watson,D., Park,T. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.
38. Sotiriou,C. and Pusztai,L. (2009) Gene-expression signatures in breast cancer. *N. Engl. J. Med.*, **360**, 790–800.
39. Benito,M., Parker,J., Du,Q., Wu,J., Xiang,D., Perou,C.M. and Marron,J.S. (2004) Adjustment of systematic microarray data biases. *Bioinformatics*, **20**, 105–114.