nature | methods

The UCSC Cancer Genomics Browser

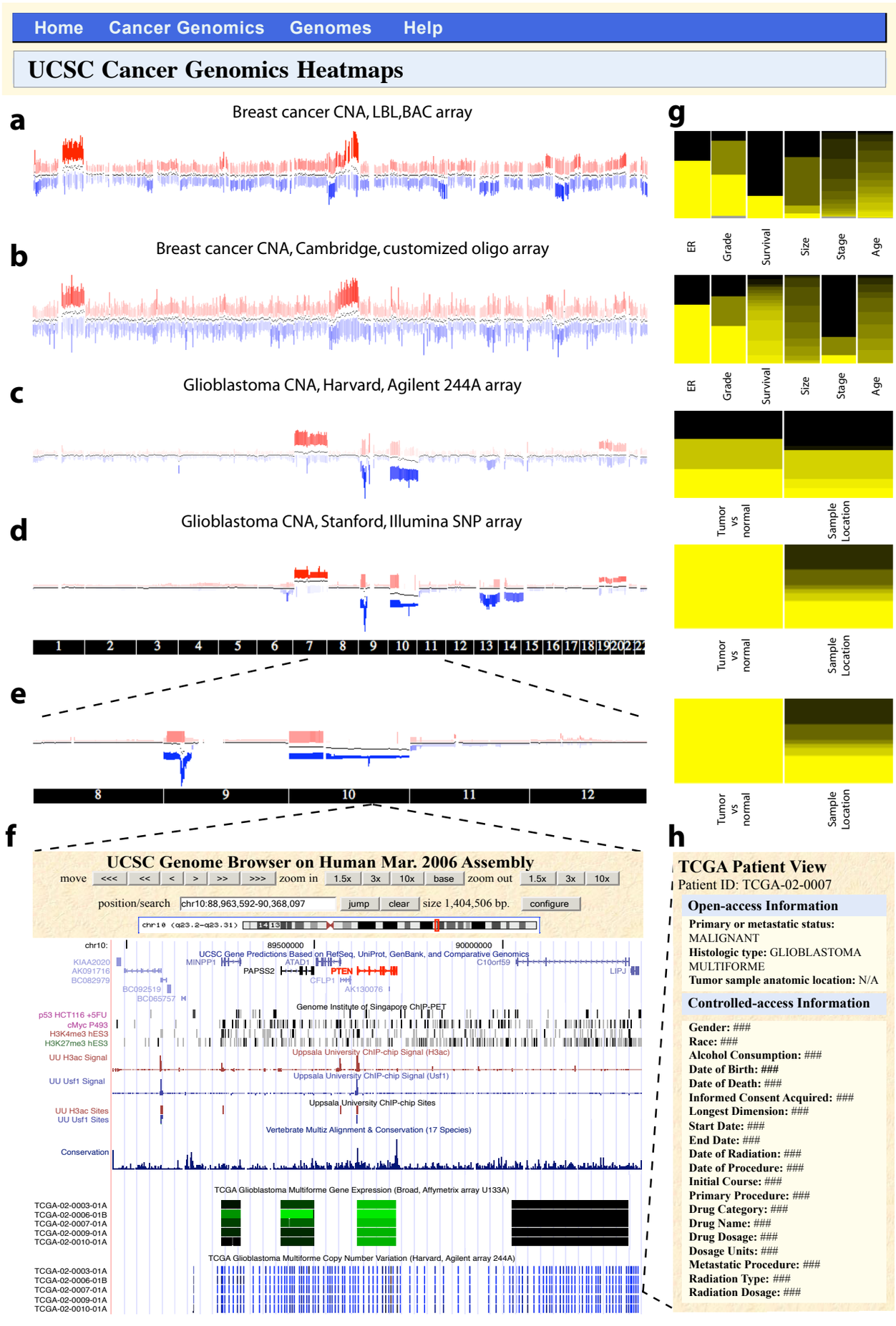
Jingchun Zhu, J Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M Kuhn, Donna Karolchik, John Archie, Marc E Lenburg, Laura J Esserman, W James Kent, David Haussler &Ting Wang

Supplementary figures and text:

Supplementary Figure 1	Genome Heatmap gateway and summary view
Supplementary Figure 2	Gateway for selection of studies
Supplementary Figure 3	The clinical feature configuration panel
Supplementary Figure 4	Controlling genesets in “genesets” or “pathway” view
Supplementary Figure 5	Comparison of outcome predictors with Pathway Sorter and Feature Sorter
Supplementary Figure 6	Outcome predictors are highly correlated with estrogen receptor (ER) status
Supplementary Figure 7	Comparison of three outcome predictors across studies
Supplementary Figure 8	Deletions, duplication and mutations disrupt interacting pathways of genes
Supplementary Table 2	Publicly available cancer genomics studies in the UCSC Cancer Genomics Browser
Supplementary Notes	
Supplementary Methods	
Supplementary Tutorial	

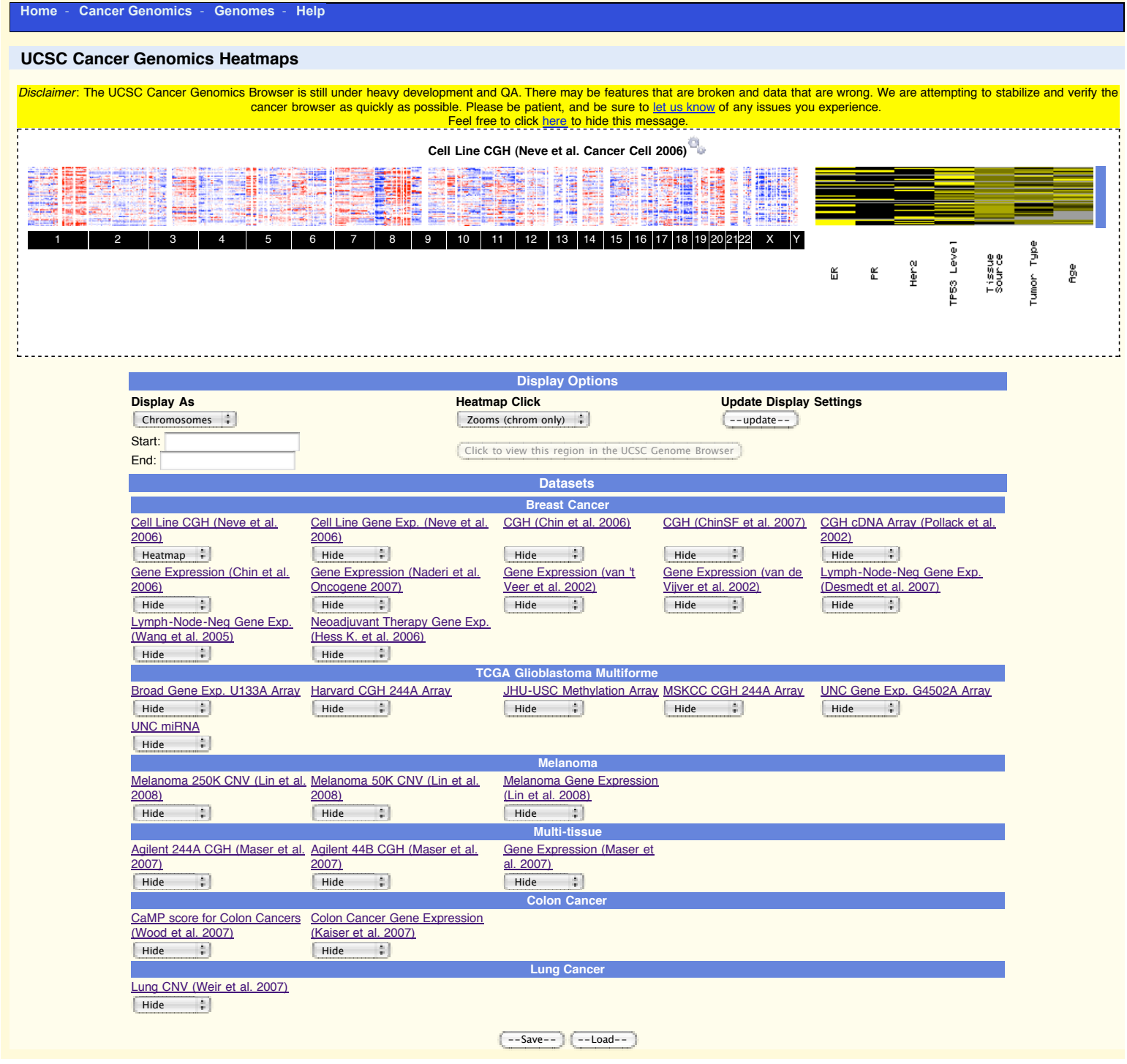
Note: Supplementary Table 1 is available on the Nature Methods website.

Supplementary Fig. 1: Genome Heatmap summary view



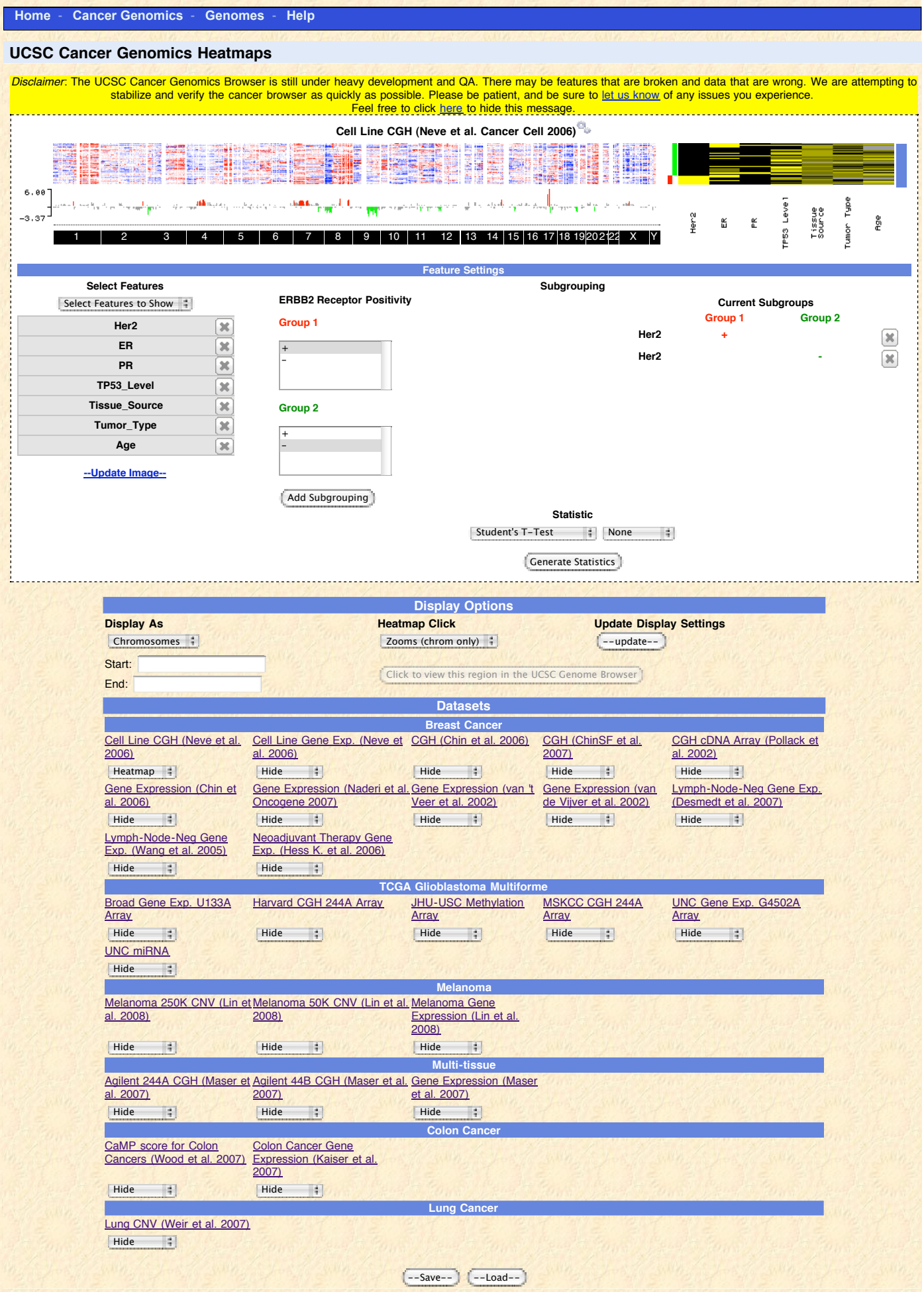
Genome heatmaps are summarized using boxplots (i.e., box-and-whisker plot) showing the data distribution on each probe. The black dots are the medians of data distribution, and the red and blue lines are the "whiskers" as described in a standard boxplot[32]. The red/blue color is scaled according to midpoint values of the whisker lines, therefore having the effect of using brighter colors to emphasize the genomic locations with more extreme array measurements. Panels (a)-(d) are screenshots of four whole-genome oriented summary views of copy number variation in two different tumor types. Panels (a) and (b) are copy number variation data of adenocarcinoma of the breast from a cohort of California patients assessed using BAC array CGH [5] (a), and a cohort of European patients using a customized Agilent oligo array [6] (b). Data in panels (c) and (d) are glioblastoma patient cohorts made available through the TCGA consortium and assayed on different experimental platforms, Agilent 244A CGH array (c) and Illumina 550K SNP array (d). These panels reveal strong similarities in global copy number variation signatures within a cancer type and differences between cancer types. Panel (e) illustrates a zoomed-in summary view of the data from panel (d) on chromosome 8 to 12. The summary view is available under the chromosomal display mode. An alternative view of the same array data, heatmap view, is illustrated in Fig. 1. Fig. 1 panels (c) and (d) display the subset of TCGA GBM samples whose clinical parameters are available, whereas panels (c) and (d) in Supplementary Fig. 1 are calculated using all GBM samples, many of which lack clinical parameters (illustrated in gray in the clinical heatmaps). Panel (f) illustrates a further zoomed-in view that takes you to the UCSC Genome Browser[8]. Here the CNA data appear as a composite data track, with each subtrack corresponding to one patient or sample (only 5 samples are shown due to space limitations). Panel (g) illustrates the clinical heatmap, and (h) shows clinical values associated with one sample.

Supplementary Fig. 2: UCSC Cancer Genomics Browser track control



Datasets from various studies are grouped according to cancer type. Each dataset can be displayed as a genome heatmap, or as a summary.

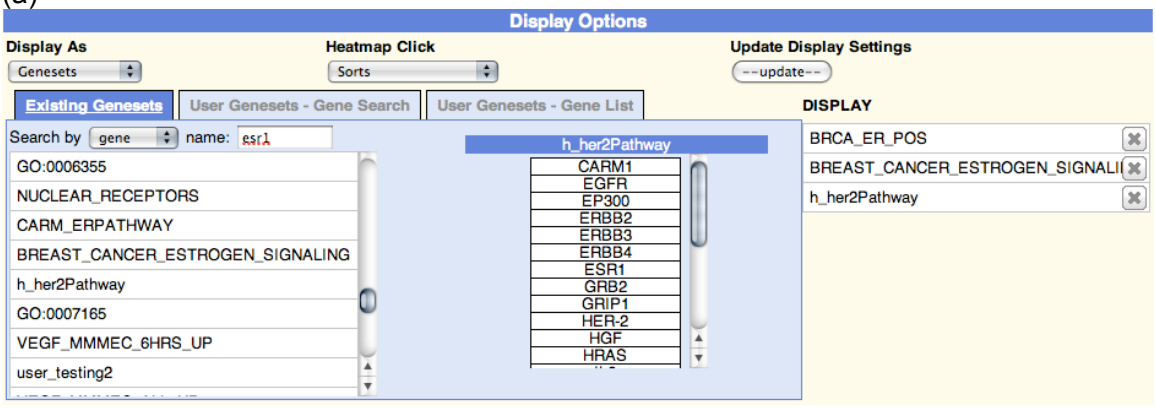
Supplementary Fig. 3: The clinical feature configuration panel



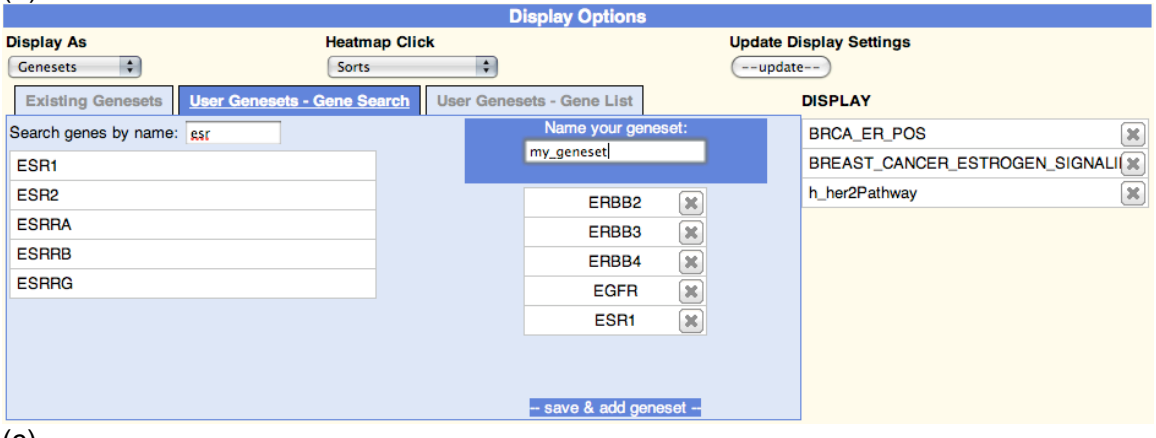
To control the clinical heatmap, the user must click on the blue bar to the right of the clinical heatmap to bring up the “Feature Settings” panel shown above. From here, the user may select and rearrange the features drawn in the clinical heatmap, define subgroups according to those features, as well as perform various statistics on the defined subgroups.

Supplementary Fig. 4: Controlling genesets in "Geneset" view

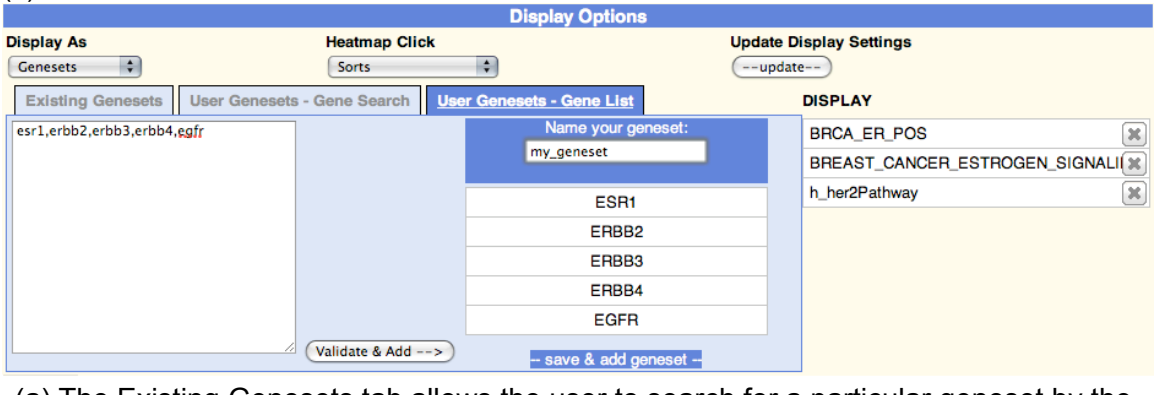
(a)



(b)

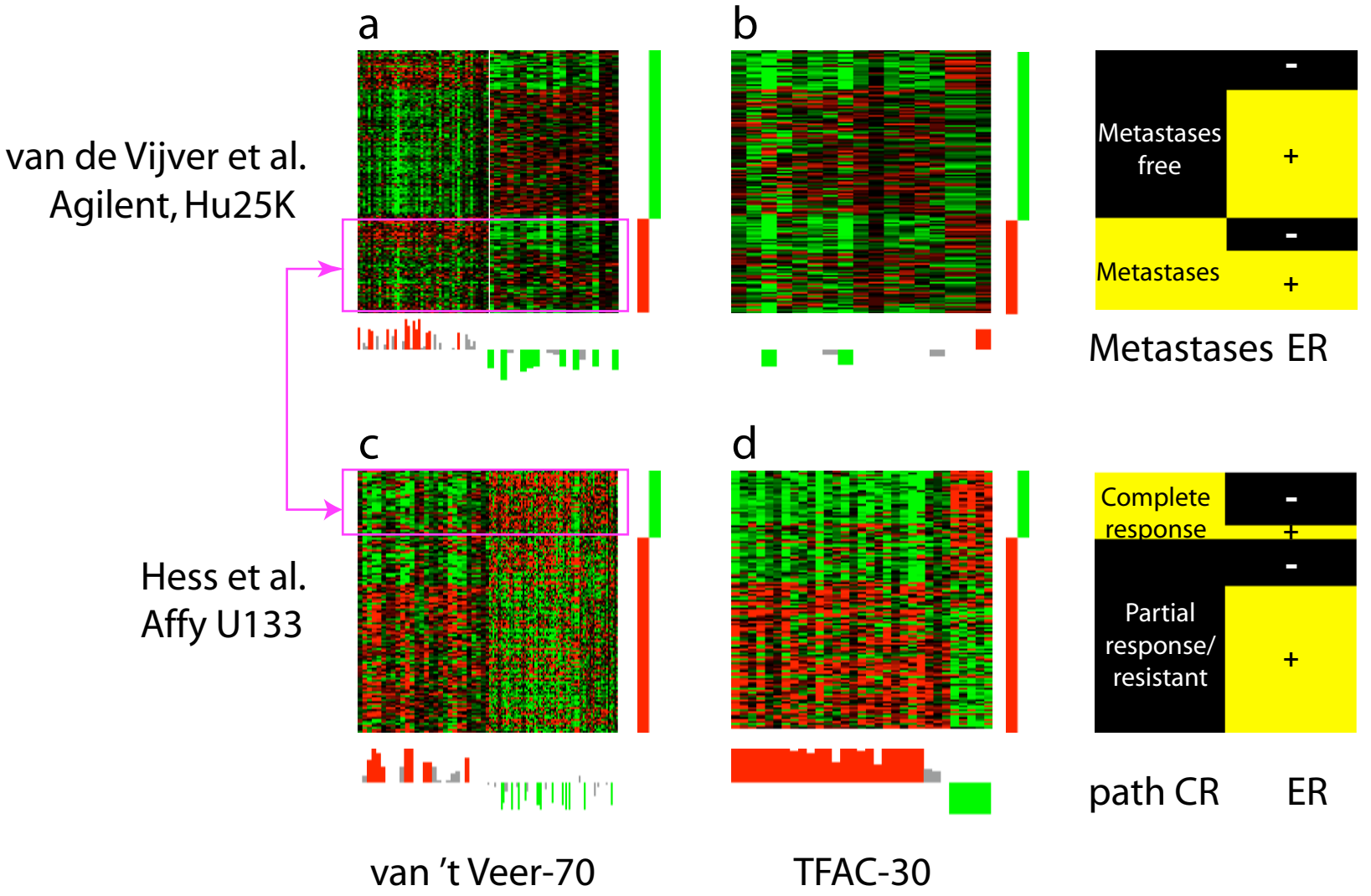


(c)



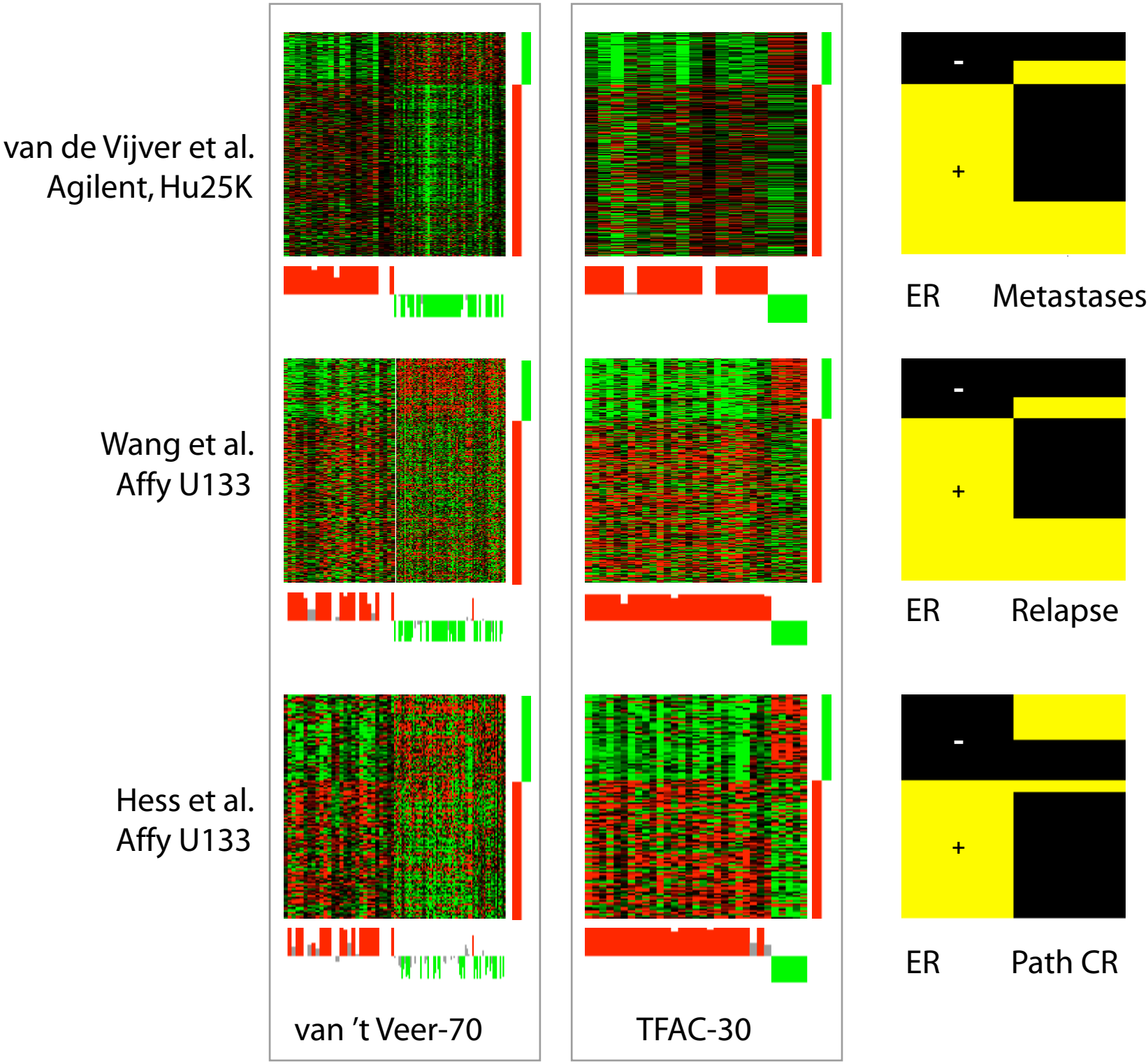
(a) The Existing Genesets tab allows the user to search for a particular geneset by the geneset's name or by a gene belonging to the geneset. When the desired geneset is found, the user may click on it to view a listing of the genes in that set as well as add it to the list of genesets to be displayed. The search will also look at any genesets the user may have previously saved in our database. (b) The User Genesets – Gene Search tab allows users to define their own genesets by searching for genes one at a time. The user enters a partial or full gene name in the search area, which returns a list of matching HUGO gene symbols. The user selects the desired gene to add it to the geneset being defined. (c) The User Genesets – Gene List tab allows also users to define their own genesets by inputting a comma-separated list of genes. This list of genes is validated against the HUGO symbols in our database and can then be stored in our database. When the geneset is fully defined after (b) or (c), the user can store the geneset in our database and add it to the list of genesets being displayed.

Supplementary Fig. 5: Comparison of outcome predictors with Pathway Sorter and Feature Sorter



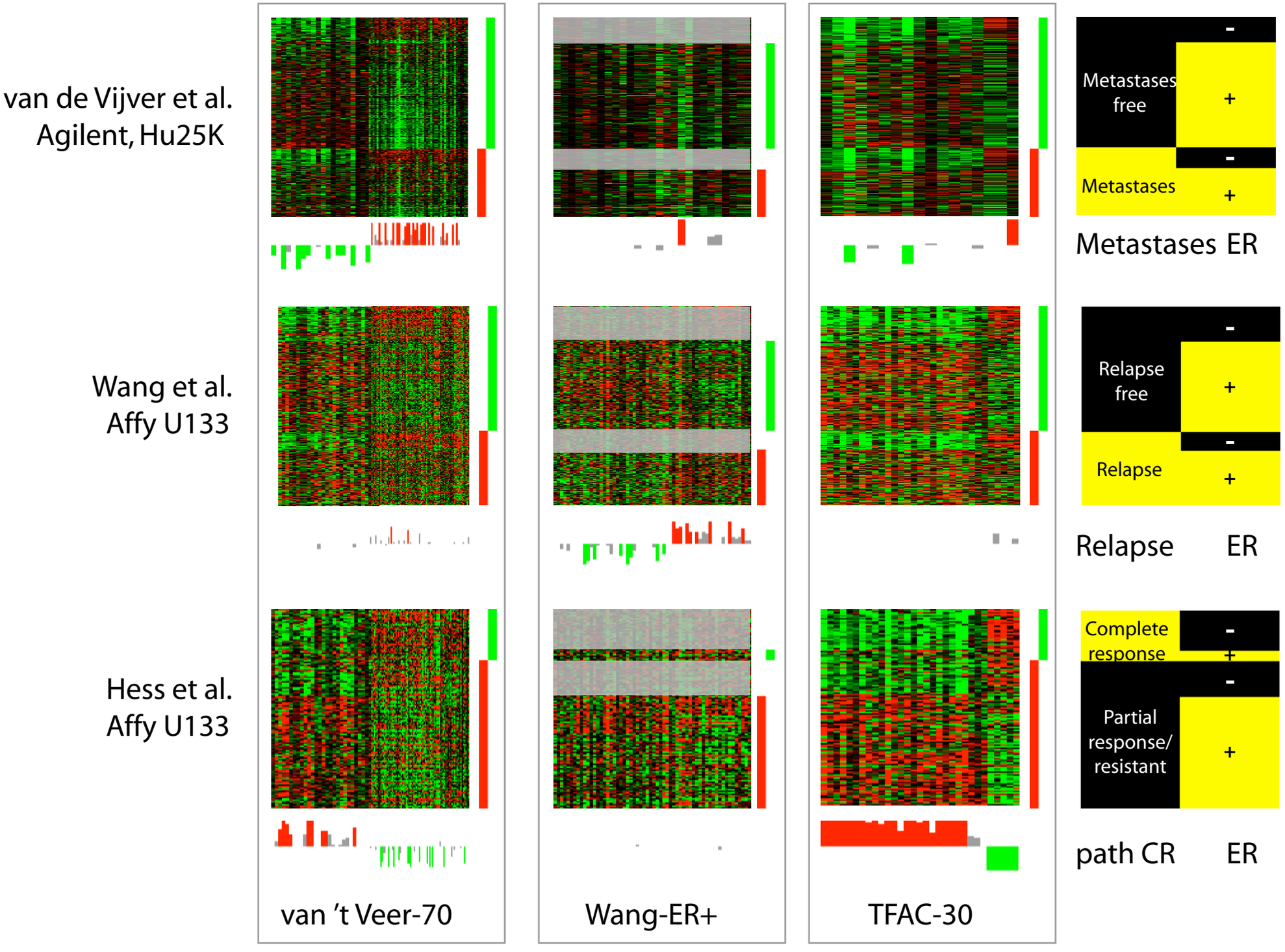
The 2x2 panel is a breast cancer gene expression profile from two clinical studies[27-29] compared using two outcome predictors [26,27,29]. Each row is a clinical study, and each column is an outcome predictor. Clinical outcome and tumor estrogen receptor status (ER) from the corresponding studies are color-coded in yellow (positive) and black (negative) and displayed alongside expression data. This cross-study and cross-gene set comparison is generated using the UCSC Cancer Genomics Browser Pathway Sorter and Feature Sorter. The Feature Sorter sorts tumor samples first by outcome, then by ER. The Pathway Sorter organizes gene expression data into two gene sets, in this case the two outcome predictors (van 't Veer-70 and TFAC-30). The Feature Sorter divides tumor samples into two subgroups based on clinical outcomes, demarcated by the red and green vertical bars on the right of each panel. Values measuring differential gene expression between the red and green subgroups are displayed as statistical tracks below the panel. The height of the statistics track bars is Bonferroni corrected Wilcoxon *p*-values (-log10(*p*)). The track is colored in red or green when $p < 0.05$, otherwise in gray. Green or downward track bars indicate relative higher expression in the green subgroup of tumors. Red or upward track bars indicate relative higher expression in the red subgroup of tumors.

Supplementary Fig. 6: Outcome predictors are highly correlated with estrogen receptor (ER) status



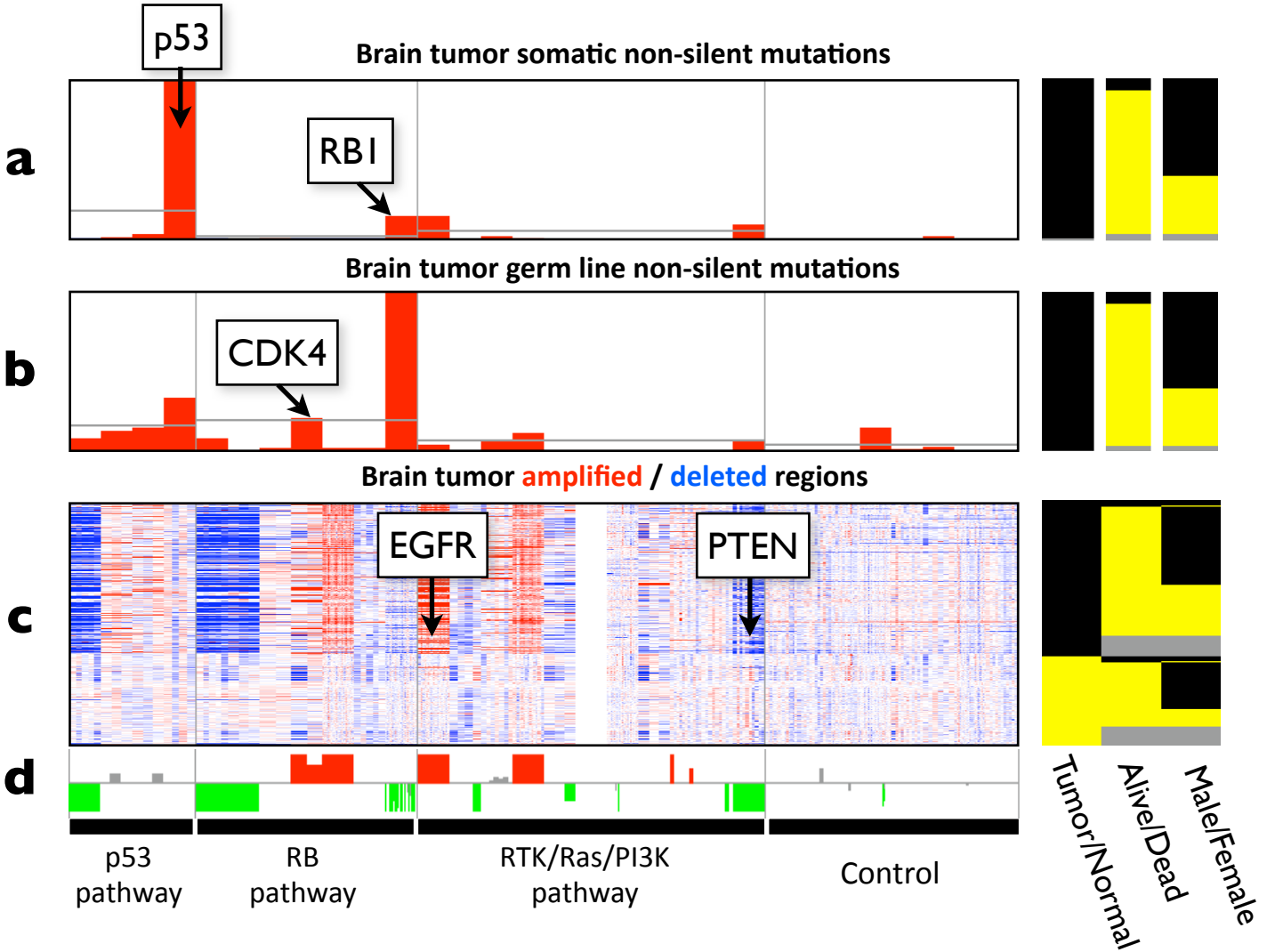
The 3-by-2 panel compares two outcome predictors (van 't Veer-70, TFAC-30)[26,27] with ER status using breast cancer gene expression profile from three clinical studies[26,27,29]. Each row is a clinical study, and each column is an outcome predictor. Tumor ER status and clinical outcome from the corresponding studies are color-coded in yellow (positive) and black (negative). Tumor samples are divided into two subgroups demarcated by the red (ER+) and green (ER-) vertical bars on the right of the panel. Expression profiles of genes in the two outcome predictors appear highly correlated with ER status. To quantify the visual pattern, Wilcoxon rank-sum test is used to measure the significance of differential gene expression between the ER+ and ER- subgroups. The p-values are displayed as statistical tracks below the panel. The height of the statistics track is Bonferroni corrected p-values (-log10(p)). The track is colored in red or green when corrected $p < 0.05$, otherwise in gray. Green or downward track bars indicate relatively higher expression in the green subgroup of tumors versus the red subgroup. Red or upward track bars show relatively higher expression in the red subgroup versus the green subgroup.

Supplementary Fig. 7: Three-way comparison of expression-based outcome predictors



This figure is similar to Supplementary Fig. 5 (refer to its legend), except that another study, Wang et al.[29], is chosen to construct a three-way comparison. Wang et al. identified a 76-gene signature consisting of 60 genes for patients with estrogen-receptor (ER) positive tumors and 16 genes for ER-negative tumors to predict distant metastasis of lymph-node-negative primary breast cancer, treated or untreated (here we use the 60-gene set, Wang-ER+). van't Veer-70 predictor overlaps with Wang-ER+ by 4 genes. There is no gene in common across all three predictor sets. The Wang samples are classified based on their relapse status. For clearer illustration, data from ER negative samples is masked in the column, because they are irrelevant to the Wang-ER+ classifier.

Supplementary Fig. 8: Integrative visualization of TCGA Glioblastoma Multiforme (GBM) genomic data



(a,b) Histogram of non-silent mutations (e.g. missense, insertion, deletions, etc.) in GBM somatic and germ line tissues, respectively. (c) Copy number alterations in GBM tumor and normal samples. Red marks represent amplified genes in the sample tissue, while blue marks represent deleted genes. Note that the clinical data to the right of (c) is sorted by tissue type: tumor (black) vs. normal (yellow). The vast majority of copy number alterations seen in the heatmap occur in the tumor samples, as expected. (d) A Bonferroni-corrected t-test comparing the distribution of copy number alterations in tumor versus normal samples. A green bar represents a significantly deleted gene ($p < 0.05$, after Bonferroni correction) and a red bar represents a significantly amplified gene in the tumor samples.

Table 2: Publicly available cancer genomics studies in the UCSC Cancer Genomics Browser

Study	Reference	Tissue Type	Type of Data	Platform	Number of Samples	Number of Features
Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.	(1)	Breast	Microarray Expression	Affymetrix HG-U133A	133	26
A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.	(2) PMID: 17157791	Breast	Microarray Expression and Copy Number Variation	Affymetrix HG-U133A / OncoBAC	51 / 53	9
Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.	(3) PMID: 15721472	Breast	Microarray Expression	Affymetrix HG-U133A	286	4
Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.	(4) PMID: 17157792	Breast	Microarray Expression and Copy Number Variation	Affymetrix Genechip HTA / OncoBAC	118 / 145	34
Gene expression profiling predicts clinical outcome of breast cancer.	(5) PMID: 11823860	Breast	Microarray Expression	?	117	10
High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.	(6) PMID: 17925008	Breast	Copy Number Variation	Custom 30K chip	220	20
Microarray analysis reveals	(7) PMID:	Breast	Copy Number	?	46	0

a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors	12297621		Variation			
A gene-expression signature to predict survival in breast cancer across independent data sets	(8) PMID: 16936776	Breast	Microarray Expression	Agilent Human 1A	135	20
Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series	(9) PMID: 17545524	Breast	Microarray Expression	Affymetrix HG-U133A	198	24
A gene-expression signature as a predictor of survival in breast cancer.	(10) PMID: 12490681	Breast	Microarray Expression	?	295	14
Modeling genomic diversity and tumor dependency in malignant melanoma.	(11) PMID: 18245465	Melanoma	Microarray Expression and Copy Number Variation(x2)	Affymetrix HG-U133A / Affymetrix 250K StyI / Affymetrix 50K XbaI	95 / 70 / 31	5
Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers.	(12) PMID: 17515920	Mixed (T-ALL, Pancreas, Colorectal, GBM & Melanoma)	Microarray Expression and Copy Number Variation(x2)	Agilent Human 1A / Agilent 44B / Agilent 244A	204 / 86 / 8	1
Characterizing the cancer genome in lung adenocarcinoma.	(13) PMID: 17982442	Lung	Copy Number Variation	Affymetrix 500K StyI	383	16
Comprehensive genomic characterization defines	(14) PMID: 18772890	Glioblastoma Multiforme	Gene Expression (x2),	Affymetrix HG-U133A / Agilent	188 / 244 / 432 / 341 /	3

human glioblastoma genes and core pathways.			Copy Number Variation (x2), Methylation, miRNA	G4502A / Agilent 244A / Agilent 244A / Illumina OMA003 / Agilent G4470B	237 / 228	
Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer.	(15) PMID: 17615082	Colon	Expression	Affymetrix HG-U133Plus 2	105	17

(1) Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol. 2006 Sep 10; 24(26):4236-44. Epub 2006 Aug 8.

(2) Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006 Dec;10(6):515-27. PMID: 17157791

(3) Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005 Feb 19-25;365(9460):671-9. PMID: 15721472

(4) Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006 Dec;10(6):529-41. PMID: 17157792

(5) van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-6. PMID: 11823860

(6) Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavaré S, Brenton JD, Ylstra B, Caldas C. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol. 2007;8(10):R215. PMID: 17925008

(7) Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12963-8. Epub 2002 Sep 24. PMID: 12297621

(8) Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, Robertson JF, Aparicio S, Ellis IO, Brenton JD, Caldas C. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene. 2007 Mar 1;26(10):1507-16. Epub 2006 Aug 28. PMID: 16936776

(9) Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris AL, Klijn JG, Foekens JA, Cardoso F, Piccart MJ, Buyse M, Sotiriou C; TRANSBIG Consortium. 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. 2007 Jun 1;13(11):3207-14. PMID: 17545524

(10) van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 Dec 19;347(25):1999-2009. PMID: 12490681

(11) Lin WM, Baker AC, Beroukhim R, Winckler W, Feng W, Marmion JM, Laine E, Greulich H, Tseng H, Gates C, Hodi FS, Dranoff G, Sellers WR, Thomas RK, Meyerson M, Golub TR, Dummer R, Herlyn M, Getz G, Garraway LA. Cancer Res. 2008 Feb 1;68(3):664-73. PMID: 18245465

(12) Maser RS, Choudhury B, Campbell PJ, Feng B, Wong KK, Protopopov A, O'Neil J, Gutierrez A, Ivanova E, Perna I, Lin E, Mani V, Jiang S, McNamara K, Zaghlul S, Edkins S, Stevens C, Brennan C, Martin ES, Wiedemeyer R, Kabbarah O, Nogueira C, Histen G, Aster J, Mansour M, Duke V, Foroni L, Fielding AK, Goldstone AH, Rowe JM, Wang YA, Look AT, Stratton MR, Chin L, Futreal PA, DePinho RA. Nature. 2007 Jun 21;447(7147):966-71. Epub 2007 May 21. PMID: 17515920

(13) Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, Mardis ER, McPherson JD, Minna JD, Morgan MB, Nadel M, Orringer MB, Osborne JR, Ozenberger B, Ramos AH, Robinson J, Roth JA, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz MR, Tsao MS, Twomey D, Verhaak RG, Weinstock GM, Wheeler DA, Winckler W, Yoshizawa A, Yu S, Zakowski MF, Zhang Q, Beer DG, Wistuba II, Watson MA, Garraway LA, Ladanyi M, Travis WD, Pao W, Rubin MA, Gabriel SB, Gibbs RA, Varmus HE, Wilson RK, Lander ES, Meyerson M. Nature. 2007 Dec 6;450(7171):893-8. Epub 2007 Nov 4. PMID: 17982442

(14) The Cancer Genome Atlas Research Network. Nature. Epub 2008 Sept 4. PMID: 18772890

(15) Kaiser S, Park YK, Franklin JL, Halberg RB, Yu M, Jessen WJ, Freudenberg J, Chen X, Haigis K, Jegga AG, Kong S, Sakthivel B, Xu H, Reichling T, Azhar M, Boivin GP, Roberts RB, Bissahoyo AC, Gonzales F, Bloom GC, Eschrich S, Carter SL, Aronow JE, Kleimeyer J, Kleimeyer M, Ramaswamy V, Settle SH, Boone B, Levy S, Graff JM, Doetschman T, Groden J, Dove WF, Threadgill DW, Yeatman TJ, Coffey RJ Jr, Aronow BJ. Genome Biol. 2007;8(7):R131. PMID: 17615082

Supplementary Notes

Availability

To access the public site, please visit http://genome-cancer.ucsc.edu/. For general questions regarding the Browser, please contact user support via email at genome-cancer@soe.ucsc.edu. You may contact the authors to obtain a copy of the software to install locally.

Current popular tools for cancer genomics analysis

Several existing tools support large genomic data analysis. One that perhaps most closely fulfills some of the goals of cancer genomics is the Genboree Discovery System (www.genboree.org). This internet-based system can be configured to support genome-centric discovery processes such as array-CGH studies and genome re-sequencing. Another powerful tool is GenePattern[1]. Originally developed to allow biomedical researchers to perform custom gene expression analysis, GenePattern has expanded its capabilities to support other genomics studies including proteomics and SNP analysis by fostering close to 90 computational and visualization modules. Additional resources under development include the Integrative Genomics Viewer developed at Broad Institute (http://www.broad.mit.edu/igv/) and the Cancer Molecular Analysis Portal developed at NCI (https://cma.nci.nih.gov/cma/).

Components of UCSC Cancer Genomics Browser

(1) Genome Heatmap: visualizing genomic data with a genome heatmap

The main panel of the Cancer Genomics Browser, Genome Heatmap, contains a whole-genome-oriented view of genome-wide experimental measurements for individual and sets of samples/patients. Heatmaps, graphical representations of data where the values taken by a variable in a two-dimensional map are represented as colors, have long been used to display high-throughput, genome-wide experimental data such as those generated by microarrays[2]. The main advantage of a heatmap is that, in addition to the two dimensions of the plane into which information can be coded, color is used to add a third dimension of information. The basic heatmap concept is implemented in Genome Heatmap to simultaneously display thousands of whole-genome data points of many patients or samples, such as transcriptome data measured by a genome tiling array and copy number alteration (CNA) data measured by a CGHarray (**Fig. 1b-e** in main text). In each Genome Heatmap panel, the x-axis represents genomic coordinates, i.e. a concatenation of all human chromosomes. The y-axis is an ordered stack of genome-wide measurements, each row representing data of one tumor or one sample. Various color gradients are used to encode the actual experimental measurements. For example, a red-black-green gradient is used to display expression data, with red and green defining up- and down-regulated genes respectively, and an alternative red-white-blue color gradient is used for CNA data, where red indicates amplification and blue indicates deletion (**Fig. 1** in main text).

Supplementary Fig. 1 displays a summary view for the CNA data from two different tumor types, glioblastoma of the brain and carcinoma of the breast, which are visualized as genome heatmaps in **Fig. 1**. Both views reveal immediately that these two tumor types exhibit coherent but highly distinct global patterns of genomic copy number variation (Fig. 1a-d, and Supplementary Fig. 1a-d)[3-7]. A configuration panel provides a gateway for investigators to select datasets and display options (**Supplementary Fig.2**), enabling interactive discovery of these patterns. To explore detailed data patterns in a particular genomic region, the Genome Heatmap allows users to zoom and pan (**Supplementary Fig. 1e**), ultimately bringing them to the general UCSC Genome Browser[8,9], where it

is easy to examine data from individual samples in small regions (**Supplementary Fig. 1f**). For example, *PTEN* is a tumor suppressor that is commonly mutated in a variety of cancers[10-12]. Zooming in to the region containing the *PTEN* gene from a heatmap containing tumor-derived glioblastoma CNA data (**Supplementary Fig. 1e**) shows a clear tendency in tumors toward genomic deletion in this region, and allows one to identify particular samples exhibiting this deletion (**Supplementary Fig. 1f**). Other features of this gene and its genomic region are easily accessible through the browser at this point, including comparative genomics annotations and tracks of genome-wide chromatin immunoprecipitation (ChIP) data in different cell lines for various histone modifications and transcription factor binding sites, such as the binding sites for c-MYC and p53. These are a sampling of the myriad such tracks with basic biological information of relevance to cancer that will become available through the ENCODE project[13] and related epigenomic initiatives.

(2) Feature Sorter: visually integrating clinical data with a feature sorter

Clinical risk factors are commonly used to assess the likelihood of cancer progression. The Feature Sorter panel on the right-hand side of the Cancer Genomics Browser allows researchers to visually examine the relationship between clinical and genomic measurements by placing a heatmap representing clinical data for each sample in a secondary panel beside the heatmaps representing genomic data from these same samples. The secondary heatmap illustrates any or all clinical features available to the user, based on their authorized level of data access (**Fig. 1g** and **Supplementary Fig. 1g**). Selection and rearrangement of the vertical (patient) order of the samples in the clinical and genomic heatmaps can be accomplished by simultaneously sorting based on a clinical feature or combination of features. All clinical features are encoded numerically (e.g. tumor stage or tumor response: see **Supplementary Method**) according to specific data agreement and made available to the Feature Sorter to facilitate this sorting. After numeric encoding, feature values are displayed using a color range of green (negative), black (zero) and yellow (positive) in the clinical heatmap. For example, the breast cancer data in **Fig. 1** are sorted by estrogen receptor status[5,6], with ER- coded as 0 (black) and ER+ as 1 (yellow); the brain cancer samples, which contain normal controls, are sorted on normal (0, black) vs. tumor (1, yellow). The features displayed at any given time by the Feature Sorter are defined by a Feature Settings control panel (**Supplementary Fig.3**). We used the I-SPY TRIAL dataset as a use case to develop this tool, taking advantage of its rich diversity of clinical, pathological, biochemical and molecular data and the fact that all data were aggregated in a single source via caIntegrator (https://cabig.nci.nih.gov/tools/caIntegrator).

With selection and vertical sorting, the Feature Sorter provides a means to visually explore the correlation of clinical and genomic data across patients, using the pattern-recognition power of the human eye to enable the exploration of vast amounts of data without the need to perform complex computational analyses. For example, we see immediately that there is a striking difference between the genomic content of these two sample types when the data in **Fig. 1d** are sorted on normal vs. tumor, with the normal samples exhibiting almost no large-scale copy number variation and the tumors rife with CNAs. When a combination of features is sorted, the first feature acts as the primary sort field; any ties between patients with identical values for the first feature are broken by the subsequent sorting features. The sorting type (ascending or descending) is adjusted by clicking on the feature's column in the heatmap, while the sorting order can be adjusted on the configuration panel (**Supplementary Fig. 3**). Data from the entire dataset or any subset defined by a combination of features can be alternatively viewed in a summary or aggregate mode (**Supplementary Fig. 1**). Apart from caveats noted in Homer et al.[14] such data aggregation affords similar interpretation of the data and often makes trends more transparent, while maintaining the anonymity of individual patients.

Finally, a "patient view" page, such as the one shown in **Supplementary Fig. 1h**, shows all the clinical parameters and metadata associated with this patient in text format. When part of the clinical data requires access control, only authorized investigators are able to obtain the actual values of these clinical parameters. For these investigators, the cancer genomic and clinical data are

available with all the other genomic datasets that cohabitate the Genome Browser, from which a large amount of biological and biomedical information can be obtained, correlated, or intersected using the existing Genome Browser tools[8].

(3) Pathway Sorter: zooming into pathways with a pathway sorter

While the Genome Heatmap provides a means to visualize cancer genomics data at the whole-genome or chromosomal level and the UCSC Genome Browser displays an integration of such data at individual genomic loci, it is becoming increasingly clear that genetic pathways rather than individual genes govern the course of tumorigenesis[15-18]. Mutations or expression changes in any of several genes in the same pathway can cause equivalent disturbance of the cellular dynamics. Pathways therefore provide a more robust and biologically meaningful way to summarize genomic data by grouping genes that may act in a concerted manner. Furthermore, pathways can include genes that do not themselves exhibit significant expression change. These genes may be downstream targets of an unperturbed, alternative path or, as found in a study on cancer subnetworks[19], these particular genes may function as bridges connecting different subnetworks within the overarching pathway.

The Pathway Sorter provides the ability to visualize cancer genomic data within the context of pathways by organizing the placement of data into sets of genes according to individual pathways as opposed to chromosomal gene location. Such an organization enables the visualization of genomic data of a genetic regulatory pathway, a signaling pathway, a set of genes in a given Gene Ontology (GO) category, or any user-defined set of genes. Used in conjunction with the Feature Sorter, the Pathway Sorter can help researchers discover perturbations of certain pathways that segregate a subset of patients with a particular clinical status.

The Pathway Sorter features a number of predefined genesets that group genes based on GO terms, those participating in genetic pathways stored in the KEGG[20] and Reactome[13,21] databases, those sharing expert-curated "gold standard" functional annotations[22], as well as other reliable means of categorizing related genes (**Supplementary Table 1**, and **Supplementary Fig. 4** for configuration panel). Researchers can also define their own desired genesets.

(4) Statistical analysis features

The Genome Heatmap, Feature Sorter and Pathway Sorter provide an interface to visually explore cancer genomics datasets. However, statistical analysis of these data is needed to provide a quantitative assessment of observed patterns and to furthermore reveal significant associations that do not lend themselves to visual identification. The UCSC Cancer Genomics Browser provides basic statistical support and additional functionality is under development. For example, a user can define subgroups within a dataset, compare their differences statistically using a Student's t-test or Wilcoxon test, or generate an aggregate for each subgroup and compare the difference in a similar fashion (see Supplementary Text for a detailed procedure). In addition, the browser allows the visualization of results generated from standard and advanced statistical analysis tools (e.g. the statistic track in **Supplementary Fig. 5**), such as publicly available software, statistical methods implemented by our group, or tools developed by the user to analyze cancer genomic datasets.

Example: using the browser utilities to compare gene expression classifiers

In recent years, genome-wide measurements of gene expression have been used to identify patterns of gene activity that can be used as diagnostic and prognostic markers[23]. These markers are selected to discriminate between different classes of disease with an aim to provide a better means for individual risk assessment and outcome prediction in patients. Biological properties of the components of these markers can also shed light on pathogenic processes and suggest potential

intervention strategies. A number of groups have developed multi-gene prognostic markers for several cancers. While there is often little overlap between the genes that have outcome-predicting potential between studies, several studies[24,25] have shown that concordant predictions are made by these various biomarkers. The ability to compare biomarkers across clinical scenarios and across platforms using a tool such as the UCSC Cancer Genomics Browser is likely to provide further insights into the similarities and differences between various biomarkers.

Consider two recent outcome predictors based on gene expression for breast cancer patients. van't Veer et al.[26] found a 70-gene expression signature that is predictive of distant metastasis (poor prognosis) in the absence of treatment for lymph-node-negative patients (van't Veer-70); Hess et al.[27] developed a 30-probe predictor of pathologic complete response (path CR) to preoperative weekly paclitaxel and fluorouracil-doxorubicin-cyclophosphamide (T/FAC) chemotherapy (TFAC-30). The prognostic power of the van't Veer-70 marker was validated in a large consecutive-patient cohort in a subsequent study by van de Vijver et al.[28]. These two studies aimed at different clinical outcomes (metastasis without treatment versus response to chemotherapy), used different expression platforms (Agilent Hu25K versus Affymetrix U133), and shared little commonalities. The two predictor sets overlap by one gene.

To use the Cancer Genomics Browser to explore these datasets, we group genes corresponding to van't Veer-70 and TFAC-30 into custom-defined genesets and visualize expression data from both studies within these genesets (**Supplementary Fig. 5**). All samples are secondarily sorted by their ER status. Expression data are displayed in rows, while the two columns represent two predictor genesets, as shown in **Supplementary Fig. 5** in a roughly two-by-two, four panel layout. Since we know the clinical outcome, we classify the van de Vijver samples (lymph-node negative patients only) based on their metastatic status, and the Hess samples based on their path CR status. The expression patterns of genes within each predictor geneset can be directly compared between patients of different clinical outcome. Samples were assigned to groups as indicated by the red or green bars. A Wilcoxon test is performed for each gene between differentially classified samples, and a p-value (corrected for multiple hypotheses) is plotted in a statistic track below each expression panel. For example, in panel (a), samples are classified into two groups. The upper group labeled by the green bar contains patients that did not have distant metastasis, while the lower group labeled by the red bar contains patients that had metastasis. Expression values of genes that correspond to van't Veer-70 are displayed, and the statistic track below the expression panel suggests that more than half of the genes are significantly differentially expressed between these two patient groups.

Several interesting observations can be made based on this simple display. For example, van de Vijver patients show significant expression differences in several genes that predict response to chemotherapy (panel b); Hess patients show consistent differential expression patterns in genes that predict distant metastasis (panel c). This may not be surprising since the clinical outcomes measured in each study are different but highly correlated, and genes in each marker set may be related as well – each marker set samples only parts of common pathways that are central to breast cancer. In particular, the expression pattern of Hess patients who achieved complete pathologic response to chemotherapy correlates positively with van de Vijver patients who developed distant metastases, as highlighted in **Supplementary Fig. 5** by the two connected purple boxes, indicating an intrinsic connection between these two clinical outcomes. Such a correlation between two different patient cohorts and two different clinical measurements suggests a greater biological robustness to the conclusion than mere replication of the same result on independent patient cohorts. This correlation suggests that cases with poor prognosis may have better response to chemotherapy, a conclusion also supported by similar findings from the I-SPY consortium (unpublished data). If corroborated, such insights may eventually be used to tailor individual patient care.

Statistical analysis of individual predictors can also reveal the extent to which they depend on certain key clinical parameters. For example, if we sort patients based on their ER status (**Supplementary Fig. 6**), it becomes obvious that TFAC-30 and van't Veer-70 are highly correlated with ER gene expression. One may wonder whether these predictors are actually more valuable than

ER status alone. Using the mutual information statistic, we can easily quantify the amount of additional information about outcome these predictors contribute beyond that given by ER status alone. For TFAC-30, ER status alone provides 0.160 bits of information per patient toward predicting path CR (pathological complete response), but given this information from ER status, the TFAC-30 predictor provides significantly more information toward predicting path CR (an additional 0.106 bits per patient), indicating that it is definitely a more valuable biomarker than ER status alone. Similar trends can be obtained for van't Veer-70, as illustrated in the following table (**Supplementary Methods**):

Table: Mutual information (*I*) of ER, outcome, and classifier

(a)

	Mutual Information	van de Vijver et al.[28]
A	*I* (classifier; ER)	0.110
B	*I* (metastases; classifier)	0.081
C	*I* (metastases; ER)	0.011
D	*I* (metastases; classifier | ER)	0.077
E	*I* (metastases; ER | classifier)	0.007

(b)

	Mutual Information	Hess et al.[27]
A	*I* (classifier; ER)	0.370
B	*I* (path CR; classifier)	0.256
C	*I* (path CR; ER)	0.160
D	*I* (path CR; classifier | ER)	0.106
E	*I* (path CR; ER | classifier)	0.009

In a slightly larger, three-way comparison we included data from Wang et al.[29] that shared a similar clinical outcome with van de Vijver et al. and used the same platform as Hess et al. (**Supplementary Fig. 7**). While these marker sets differentiate patients in their respective studies, they do not seem to work well on other patient cohorts. This is largely due to many differences between studies. However, it also highlights the difficulty in identifying key genes and pathways that have important, generalizable prognostic and diagnostic value. Perhaps a solution may be to look at it from a pathway-driven perspective, so that lack of overlap of specific genes may not matter if one uses pathways instead of specific genes [17-19,30].

Example: using the browser utilities to identify disturbed pathways in glioblastoma

In this example we demonstrate the power of using the geneset view to visually assess the state of biological pathways across different data types (**Supplementary Fig. 8**). GBM mutation and copy number alteration data are shown for the three pathways identified by the TCGA Research Network and a control pathway containing a collection of sequenced genes on chromosome 18. The three pathways are: (1) p53 pathway (CDKN2A, MDM2, MDM4 andTP53); (2) RB pathway (CDKN2A, CDKN2B, CDKN2C, CDK4, CDK6, CCND2 and RB1); (3) RTK/Ras/PI3K pathway (EGFR, ERBB2, PDGFRA, MET, SPRY2, RAS, NF1, AKT1, PIK3CA, PIK3R1 and PTEN). The control pathway contains the following genes: KIAA1632, PTPN2, ROCK1, PHLPP, SMAD2, SMAD4, BCL2 and C18orf25.

What should be readily apparent is that the three pathways experienced a much greater number of non-silent mutations (in both germ line and somatic tissues) and copy number alterations compared to the control pathway. By presenting the genomic data in such a way, it is much easier to gain an appreciation of the level of disturbance that these pathways undergo during/after tumorigenesis.

Data security and user access control

The UCSC Cancer Genomics Browser is currently installed on several web server systems that serve different collaboration projects. All these servers (except the public server genome-cancer.ucsc.edu) are implemented with the industry standard HTTPS protocol with user login and password control. Each project has its own dedicated web server with specific authorized web users (granted by responsible persons of individual projects), hence the private confidential control-access data can be accessed by their corresponding authorized users only.

We are also exploring the possibility of adopting a universal authentication system, InCommon, for some server(s) to support large user communities, e.g.TCGA.

As the number of requests for collaboration grows, we find ourselves running out of available physical machines for web servers; hence we adopted the Virtual Machine (VM) technology that enables us to create multiple VMs (one VM for one web server) on a single computer hardware system. This also allows us to update our code base to enable individual VMs to share a large number of common public datasets (e.g. the human genome with its massive amount of public annotation data) without duplicating them on individual systems.

On the public site, we support limited user authentication for collaborations and saved custom settings. The user is authenticated with a secured token-exchange mechanism that provides the greatest level of non-encrypted authentication available. Passwords are encrypted with industry-standard one-way encryption and are never sent in plain text. User accounts can be requested from the authors. When a collaboration is established, pre-publication non-clinical data can be uploaded to the browser and visualization of that data can be limited to one or more user accounts.

Supplementary Methods

Mapping genomics data to the human genome assembly

In order to display microarray data in either chromosome or geneset views, the UCSC Cancer Genomics Browser requires that each microarray probe be mapped to the latest human genome assembly (build 36, March 2006, hg18). Most recent studies and modern microarray platforms already provide this information. However, for those datasets that do not provide probe mappings, we use blat[31] to align the oligonucleotide probes to the human genome. Valid mappings require at least a 95% sequence identity across the full length of the probe. If a probe maps to multiple genomic regions under these criteria, the highest-scoring alignment is chosen for the probe mapping.

Encoding clinical features

The clinical feature heatmap is a graphical representation of a diverse set of clinical data, but in order to display any clinical data we must be able to transform it into numeric form. For some clinical measurements, such as Tumor Size, no transformation is required. Coded measurements are enumerated by assigning each code a numeric value. For example, a measurement such as ER Status may have three codes [“negative”, “positive”, “indeterminate”] that we numerically represent as “negative” = 0, “positive” = 1, and “indeterminate” = 2. To draw the clinical heatmap, each transformed clinical datapoint is assigned a color from a green-black-yellow colorspace, where negative numbers

are given shades of green, zero is black, and positive numbers are given shades of yellow. The colorspace is scaled according to the minimum and maximum numeric value for each clinical feature separately. Clinical features that cannot be reasonably encoded via any means are not displayed in the clinical heatmap.

Assessing relationship between ER status and classifiers in outcome prediction

Mutual information of ER, classifier and outcome is used to evaluate the power of estrogen receptor status and/or classifier to predict clinical outcome.

$I(X;Y)$ is the amount of mutual information between variable X and Y. $I(X; Y \mid Z)$ is the amount of additional information Y contributes to X given Z.

In the Hess et al. dataset, the clinical outcome is path CR (pathologic complete response). Mutual information is computed using the original classifier to path CR the dataset has produced. In the van de Vijver et al. dataset, the outcome is metastases, and mutual information is computed using a classifier of distant metastases developed by van 't Veer et al.[26]

In both studies, the classifier is more correlated with estrogen receptor status (ER) than with outcome (Row A, B). However, classifier is still a better predictor to outcome compared to ER (Row B, C). Classifier contributes a significant amount of additional information at predicting outcome than ER alone (Row D).

Mutual information is measured in bits. Logarithms to the base 2 are used in the calculation. Pseudo-counts amount to 5% of the total number of samples in each comparison.

Implementing the UCSC Cancer Genomics Browser

Database and display

To produce the genomic or geneset heatmaps displayed in the UCSC Cancer Genomics Browser, microarray data stored in large bed15-formatted MySql tables must be condensed into a color image 700 pixels wide. Clearly, when displaying whole-genome data or a dense geneset in such a constrained space, many microarray probes will fall under the same pixel. Regardless of the view mode, all probe data located in the same pixel will be averaged together and that average value is used for that pixel's color.

For "chromosome-view", the UCSC Cancer Genomics Browser uses down-sampled versions of all microarray data tables in order to reduce the time necessary to draw the image, thereby improving the tool's responsiveness. Down-sampled tables are created by breaking up the whole genome into approximately 3,000 bins and then averaging together probe data that fall within the same bin. The down-sampled tables are used in lieu of the original "high-resolution" table when the image is currently displaying more than one chromosome. When the image is zoomed in to a single chromosome, the high-resolution table is used to draw the image.

When the display is set to "geneset-view", the image is split into distinct sections for each geneset. The width of each section is proportional to the total number of genes in its corresponding geneset, with a minimal width of 20 pixels so that small genesets are still readily visible. Each section is then divided equally among all genes in the geneset, where probes assigned to the same pixel are averaged together as before. Alternatively, some microarray datasets may have multiple probes that map to the same gene, and in these cases the region allocated for the gene is divided equally among the multiple probes.

Use of Web2.0 technology

The UCSC Cancer Genomics Browser heavily utilizes dynamic HTML and JavaScript, and as a result requires a modern browser to operate correctly. Firefox 3.0+ and Safari 3.0+ are fully supported on

both Windows and OS X platforms. Some features of the Cancer Browser may not function correctly when using Internet Explorer at this time.

Saving a session

With the Web 2.0 framework, a new method is required to store a user's session. Because the state of the user's session is maintained entirely on the client computer, the client must communicate its state to the server to be stored in the server's MySql database. In the current implementation, when the user clicks the "Save" button, the client sends the server a string formatted in JavaScript Object Notation (JSON) that fully describes the client's state. The user can then recall the saved session by pressing the "Load" button, at which point all of the user's settings from the saved session are loaded into the client's current workspace.

There are a few limitations with the current saving/loading mechanism. Only a single session can be stored per user in the database. Also, since the session is stored in a database using the UCSC Genome Browser user ID, the same computer and web browser on which the session was initially saved must be used to load the session at a later time. Both of these limitations will be addressed in the near future, adding the ability to save and load multiple user sessions and easily share any saved session with collaborators.

Current statistics available on the browser:

The UCSC Cancer Genomics Browser has several statistical methods implemented currently: difference of means, Student's t-test, Wilcoxon rank-sum test, Fisher's exact test, Fisher linear discriminant, Levene's and Brown-Forsythe tests of homogeneity of variance, and the Jarque-Bera test of normality.

Difference of means computes the difference between the means of the two subgroups for each probe being compared. The Student's t-test is a two-tailed test mapped to the Student's t distribution. It displays the p-value of each probe being derived from different underlying distributions in the two subgroups. The Wilcoxon rank-sum test is a non-parametric alternative to the Student's t-test that also displays p-values of the difference between the subgroups. The Fisher's exact test analyzes the association of subgroups. It calculates the p-value of getting the observed level of unbalance between subgroups based on a hypergeometric distribution. The Fisher's linear discriminant function finds the residual around the mean after subgrouping, then empirically calculates the p-value of obtaining such a residual by comparing to 100 permutations irrespective of subgrouping. Levene's test finds the p-value (based on an a Student's t distribution) of the groups having their observed variances if they are from the same distribution. The Brown-Forsythe test is similar to Levene's test but is calculated around the median rather than mean, so is more suited to measure variance differences in skewed distributions. The Jarque-Bera test is a goodness-of-fit test that calculates the p-value the data in each subgroup is normally distributed given their skewness and kurtosis. The p-value of the subgroup that is most non-normal is displayed. All p-values are displayed in log scale and coloring is based on the specific significance cutoff of 0.05. Bonferroni correction is available for all statistical tests and will scale the p-values based on the number of tests being performed.

Current publically available studies

The public site contains a rapidly growing body of publicly available cancer genomic data, including 21 published studies, datasets from the TCGA consortium, and others. These correspond to roughly 4,343 genome-wide experiments from 3,033 samples (**Supplementary Table 2**).

Acknowledgements

We would like to thank the many collaborators who have contributed data to our project, especially the I-SPY TRIAL Consortium and the TCGA Consortium for allowing us to use prepublication data to tune our tools. We thank members of the UCSC Genome Browser bioinformatics team for comments and support. We acknowledge the dedicated system administrators who have provided an excellent computing environment: Jorge Garcia, Erich Weiler, Chester Manuel and Victoria Lin. We would also like to thank Meredith Buxton and Sarah Davis for coordinating the I-SPY TRIAL. We thank Laura van't Veer and Joseph Costello for insightful comments. D.H. is a Howard Hughes Medical Institute investigator. T.W. is a Helen Hay Whitney Fellow. We acknowledge support from the NIH Training Grant T32 GM070386, the I-SPY Consortium, California QB3 and its INSTINCT program.

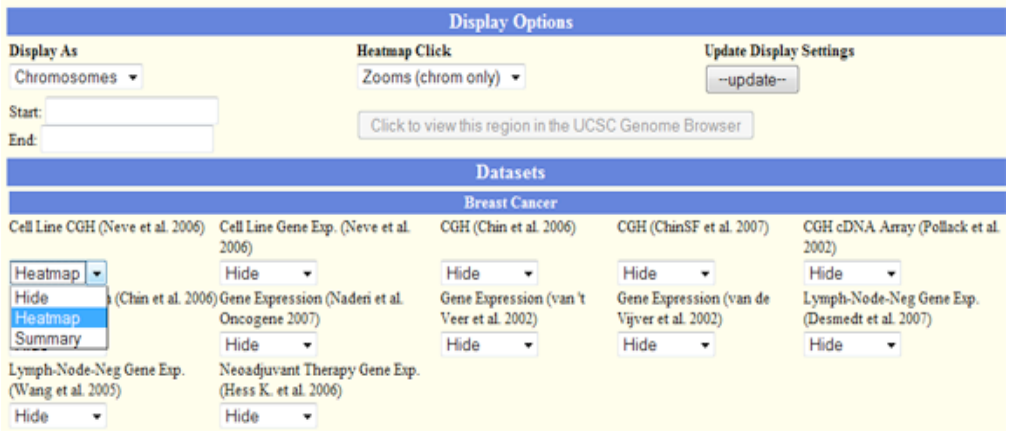
References

1. M. Reich, T. Liefeld, J. Gould et al., *Nat Genet* **38** (5), 500 (2006).
2. M. B. Eisen, P. T. Spellman, P. O. Brown et al., *Proc Natl Acad Sci U S A* **95** (25), 14863 (1998).
3. L. A. Brown, J. Hoog, S. F. Chin et al., *Nat Genet* **40** (7), 806 (2008).
4. B. Cadieux, T. T. Ching, S. R. VandenBerg et al., *Cancer Res* **66** (17), 8469 (2006).
5. K. Chin, S. DeVries, J. Fridlyand et al., *Cancer Cell* **10** (6), 529 (2006).
6. S. F. Chin, A. E. Teschendorff, J. C. Marioni et al., *Genome Biol* **8** (10), R215 (2007).
7. Y. Kotliarov, M. E. Steed, N. Christopher et al., *Cancer Res* **66** (19), 9428 (2006).
8. D. Karolchik, R. M. Kuhn, R. Baertsch et al., *Nucleic Acids Res* **36** (Database issue), D773 (2008).
9. W. J. Kent, C. W. Sugnet, T. S. Furey et al., *Genome Res* **12** (6), 996 (2002).
10. H. Ohgaki and P. Kleihues, *Am J Pathol* **170** (5), 1445 (2007).
11. X. Fan, Y. Aalto, S. G. Sanko et al., *Int J Oncol* **21** (5), 1141 (2002).
12. J. Li, C. Yen, D. Liaw et al., *Science* **275** (5308), 1943 (1997).
13. I. Vastrik, P. D'Eustachio, E. Schmidt et al., *Genome Biol* **8** (3), R39 (2007).
14. N. Homer, S. Szelinger, M. Redman et al., *PLoS Genet* **4** (8), e1000167 (2008).
15. B. Vogelstein and K. W. Kinzler, *Nat Med* **10** (8), 789 (2004).
16. L. D. Wood, D. W. Parsons, S. Jones et al., *Science* **318** (5853), 1108 (2007).
17. R. McLendon, A. Friedman, D. Bigner et al., (2008).
18. S. Jones, X. Zhang, D. W. Parsons et al., (2008).
19. H. Y. Chuang, E. Lee, Y. T. Liu et al., *Mol Syst Biol* **3**, 140 Epub 2007 Oct 16 (2007).
20. M. Kanehisa, S. Goto, M. Hattori et al., *Nucleic Acids Res* **34** (Database issue), D354 (2006).
21. G. Joshi-Tope, M. Gillespie, I. Vastrik et al., *Nucleic Acids Res* **33** (Database issue), D428 (2005).
22. C. L. Myers, D. R. Barrett, M. A. Hibbs et al., *BMC Genomics* **7**, 187 (2006).
23. L. J. Esserman, Y. Shieh, J. W. Park et al., *Expert Rev Mol Diagn* **7** (5), 533 (2007).
24. C. Fan, D. S. Oh, L. Wessels et al., *N Engl J Med* **355** (6), 560 (2006).
25. K. Shedden, J. M. Taylor, S. A. Enkemann et al., *Nat Med* **14** (8), 822 (2008).
26. L. J. van 't Veer, H. Dai, M. J. van de Vijver et al., *Nature* **415** (6871), 530 (2002).
27. K. R. Hess, K. Anderson, W. F. Symmans et al., *J Clin Oncol* **24** (26), 4236 (2006).
28. M. J. van de Vijver, Y. D. He, L. J. van't Veer et al., *N Engl J Med* **347** (25), 1999 (2002).
29. Y. Wang, J. G. Klijn, Y. Zhang et al., *Lancet* **365** (9460), 671 (2005).
30. D. W. Parsons, S. Jones, X. Zhang et al., (2008).
31. W. J. Kent, *Genome Res* **12** (4), 656 (2002).
32. John W. Tukey, *Addison-Wesley, Reading, MA* (1977).

Supplementary Tutorial: Jump start on the UCSC Cancer Genomics Browser

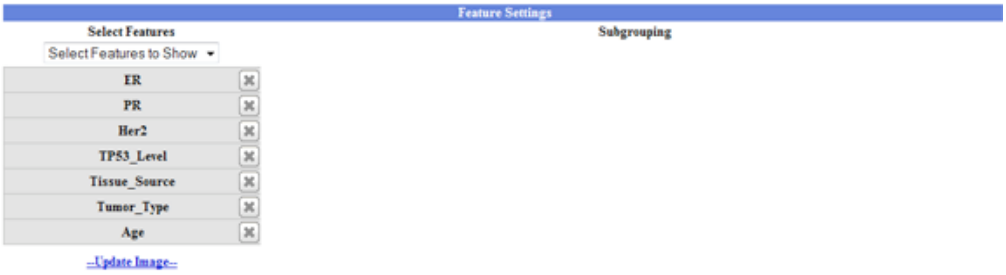
Step-by-step instructions to start a Cancer Browser session

1. Begin by choosing a dataset you would like to view, and select “Heatmap” from the dropdown menu. Here we’ve selected the Neve et al. cell line CGH dataset:



The heatmap will load and appear above the list of datasets. The heatmap on the left represents the copy number values of each probe across the genome (labeled below) with each pixel row representing a sample. The feature information of each sample is represented by a heatmap on the right, with each feature labeled below it. Clicking on a feature will sort the samples (i.e. rows) accordingly. Secondary sorting using shift-clicking breaks ties in the primary sort.

2. To access feature configuration options, click on the blue bar to the right of the feature information. This will show a new panel below the dataset:



The currently displayed features are listed and can be removed by clicking the “X” symbol to the right of the text. They can be reorganized by dragging and dropping the features in order. Additional available features can be added through the dropdown menu above the list. Clicking “Update Image” will refresh the feature heatmap above with the current features in the list in order. For example, if we remove all but “Her2”, the feature heatmap will show only that feature:

3. To define subgroups, click the name of a feature in the list, which will display subgrouping controls for that feature (the red and green selection boxes in the image above). By selecting the subset of features in each box and clicking the "Add Subgrouping" button, we can define two subgroups based on Her2 status:

You can see the current subgrouping defined in the list on the right, and a new red/green bar appears next to the feature information allowing you to identify which samples belong to which subgroup. The option to run statistics is displayed at the bottom.

4. By selecting Student's t-test from the statistics dropdown and Bonferroni from the next dropdown, we are able to run genome-level statistics to determine if there are any

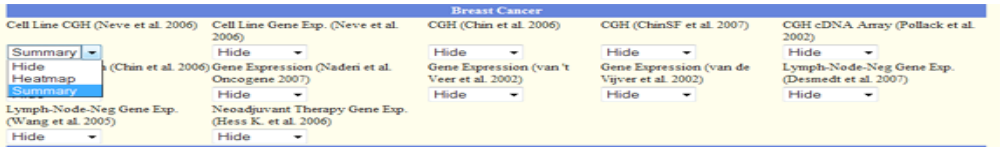
regions of particular interest (with multiple hypothesis correction).

5. We can see a small significant peak in chromosome 17. By clicking on the heatmap at that position, we are able to zoom in to investigate the region further.

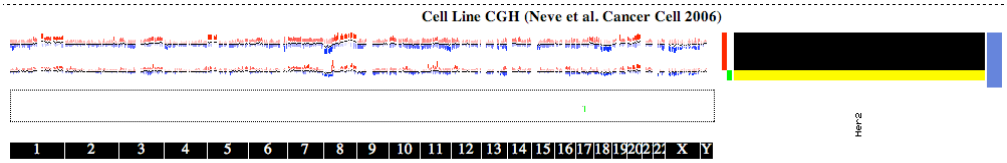
Below the heatmaps, we can see the current coordinates of the zoomed image, as well as a button to jump directly to the UCSC Genome Browser:

If we click into the UCSC Genome Browser, we can see that the Her2 amplicon is represented in that region, which explains the Her2 significance between the subgroups of samples we selected.

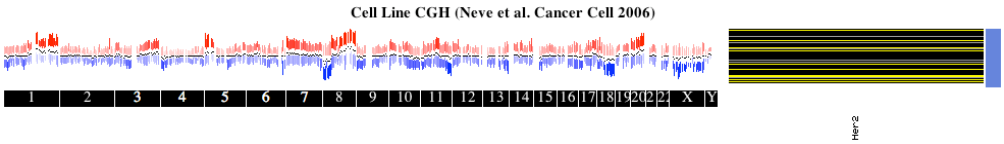
6. Shift-clicking allows us to zoom back out to the full genome level, and we can switch between heatmap and summary view for the dataset:



The summary view allows us to view the distribution of amplifications and deletions across the genome for a dataset. Since we have subgroups defined, we will see a summary image which is split top and bottom for the two subgroups, as well as the statistic track below:



We can remove the subgrouping and see a single summary view for the dataset:



These views allow us to quickly display large-scale genomic amplifications or deletions, as well as examine possible amplifications or deletions for specific subgroups of samples.