

Subtype and pathway specific responses to anti-cancer compounds in breast cancer

Heiser, et al.

Supplementary Information

Association of growth rate and response to therapeutic agents

In general, we found that luminal subtype cell lines grew more slowly than basal or claudin-low cells (Kruskal-Wallis test $p = 0.006$, **Fig. S3A and Table S1**) and the range of doubling times was broad (18 to 300 hours). This raised the possibility that the most sensitive cell lines were those that grew most rapidly. We tested this hypothesis by assessing the effects of subtype and doubling time simultaneously with an ANCOVA (see details below) and found that 20 of 23 subtype-specific compounds had better associations with subtype than with doubling time (mean log ratio of p -values = 0.87, standard deviation 1.09). Moreover, 11 of 23 subtype-specific compounds were most effective in the most slowly growing luminal cell lines (**Table 1**). One agent, 5-fluorouracil, was not significant in the subtype test alone but showed strong significance in the ANCOVA model for both class and doubling time. (**Fig. S3B**). We conclude that in most cases, the 3-day proliferation assay detects molecular signature-specific responses that are not strongly influenced by growth rate.

To assess the effects of cell line subtype and growth rate on drug sensitivity, we performed a set of 2-way analysis of covariance (ANCOVA) tests, one for each of the three cell line classification schemes: *i*) luminal vs. basal vs. claudin-low; *ii*) luminal vs. basal + claudin-low; and *iii*) ERBB2^{AMP} vs. non-ERBB2^{AMP}. This yielded 6 sets of p -values (2 main effects x 3 classification schemes); we separately adjusted the two sets of main effect p -values for multiple comparisons. We used the R functions `lm` and `Anova` (available as part of the `car` package).

Integrated Pathway Analysis

Integration of copy number, gene expression and pathway interaction data was performed using the PARADIGM software(1). Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single cell line or patient sample. TCGA breast cancer data was obtained from the TCGA DCC on November 7, 2010. TCGA and cell line gene expression data were median probe centered within each data set separately. All of the values in each of these datasets (either the cell lines or TCGA tumor samples) were rank transformed and converted to $-\log_{10}$ rank ratios before supplying to PARADIGM. Pathways were obtained in BioPax Level 2 format on October 13, 2010 from <http://pid.nci.nih.gov/> and included NCI-PID, Reactome, and BioCarta databases. Interactions were combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. “cell cycle”) were retained as pathway concepts. Before merging gene concepts, all gene identifiers were translated into HUGO nomenclature. All interactions were included and no attempt was made to resolve conflicting influences. A breadth-first undirected traversal starting from P53 (the most

connected component) was performed to build one single component. The resulting merged pathway structure contained a total of 8768 concepts representing 3491 proteins, 4757 complexes, and 520 processes. Expectation-Maximization parameters for PARADIGM were trained on the cell line data and then applied to the TCGA samples. Data from the cell lines and tumor samples were then combined into a single data matrix. Any entry without at least 1 value above 0.5 IPL in either the data from cell lines or tumor samples was removed from further analysis.

TCGA and cell line clustering

Using PARADIGM IPLs, cell lines were clustered together with TCGA tumor samples to determine if cell lines were similar to tumor samples of the same subtype. Well-studied areas of the SuperPathway contain genes with many interactions (hubs) and large signaling chains of many intermediate complexes and abstract processes for which no direct data is available. To avoid bias toward due to the presence of hubs, pathway concepts with highly correlated vectors (Pearson correlation coefficient > 0.9) across both the cell line and tumor samples were unified into a single vector prior to clustering. This unification resulted in 2351 non-redundant vectors from the original 8768 pathway concepts.

Samples were clustered using the resulting set of non-redundant concepts. The matrix of inferred pathway activities for the 46 breast cancer cell lines and 183 TCGA tumor samples was clustered using complete linkage hierarchical agglomerative clustering implemented in the Eisen Cluster software package version 3.0(2). Uncentered Pearson correlation was used as the metric for the pathway concepts and Euclidean distance was used for sample metric (**Fig. S4**).

To quantify the degree to which cell lines clustered with tumor samples of the same subtype, we compared two distributions of t-statistics derived from Pearson correlations (**Fig. S5**). Let C_s be the set of cell lines of subtype s . Similarly, let T_s be the set of TCGA tumor samples of subtype s . For example, C_{basal} and T_{basal} are the set of all basal cell lines and basal tumor samples respectively. The first distribution was made up of t-statistics derived from the Pearson correlations between every possible pair containing a cell line and tumor sample of the same subtype; i.e. for all subtypes s , every pairwise correlation t-statistic was computed between a pair (c, t) such that $c \in C_s$ and $t \in T_s$. The second distribution was made of correlation t-statistics between cell lines of different subtypes; i.e. computed over pairs (c, c') such that $c \in C_s$ and $c' \in C_{s'}$, and $s \neq s'$. We performed a Kolmogorov-Smirnov test to compare the distributions. We repeated this analysis using samples from the same source (cell line or tumor) to verify that cells of the same subtype have overall pathway activities that are more similar than cells of different subtypes. As above, the first distribution was made up of t-statistics between pairs of samples of the same subtype and the same origin (cell line or tumor). The second distribution was made of correlation t-statistics between samples of different subtypes again from the same origin.

We assessed the significance of the subpathways by comparing to subnetworks generated from a background model. Specifically, we measured how likely it would be to find the

identified subnetworks with the observed sizes by chance. To this end, we constructed a background set of subnetworks computed from 1000 simulations in which samples were randomly partitioned into two equal bins, which simulated groupings of cancer cell lines reflecting no biological relevance. Statistical Analysis of Microarrays (SAM) then was used to compute differential pathway activity scores for each concept and each random partitioning using the same thresholds as with the original subtype definitions. Histograms of the subpathway sizes obtained from the random partitions were compared against the subpathway sizes derived from the original subtypes to gauge significance. Our analysis shows that subnetworks derived from the original basal, luminal, claudin-low, and ERBB2^{AMP} definitions are significantly larger than those subnetworks obtained from random partitionings. This is true for both the size of the entire subpathway (total number of nodes in the graph) as well as for the largest connected component. Histograms and empirical Z-scores were collected from this random partitioning analysis (see **Fig. S6**).

Supplementary Figure and Table Legends

Figure S1. Genomic and transcriptional profiles of the breast cancer cell lines. A. Hierarchical consensus clustering matrix for 55 breast cancer cell lines showing 3 clusters (claudin-low, luminal, basal) based on gene expression signatures. For each cell line combination, color intensity is proportional to consensus. **B.** DNA copy number aberrations for 43 breast cancer cell lines are plotted with $\log_{10}(\text{FDR})$ of GISTIC analysis on the y-axis and chromosome position on the x-axis. Copy number gains are shown in red with positive $\log_{10}(\text{FDR})$ and losses are shown in green with negative $\log_{10}(\text{FDR})$.

Figure S2. GI50 calculations are highly reproducible. A. Each bar represents a count of the frequency of replicated drug/cell line combinations. Most cell lines were tested only one time against a particular compound, but some drug/cell line combinations were tested multiple times. **B.** Each boxplot represents the distribution of median absolute deviations for drug/cell line pairs with 3 or 4 replicates. **C.** Example drug response curves for HCC1395 treated with cisplatin. Data from three experiments are shown, each plotted in a unique color. Each dot represents the growth inhibition following three days of treatment with one of 10 concentrations of cisplatin. For each dose of each experiment, measurements are performed in triplicate. The x-axis represents increasing cisplatin concentration; the y-axis indicates growth inhibition following treatment. A single curve is fit to the set of 30 data points (3 untreated and 27 treated). The vertical line represents GI50, which is extrapolated from the fitted curve. Across multiple experimental replicates, the dose-response curve is highly reproducible. **D, E, F.** Example drug response curves for three other cell lines, each treated with a different compound. Convention as in C.

Figure S3. Doubling time varies across cell line subtype. A. Growth rate, computed as the median doubling time in hours, of the breast cancer cell lines subtypes are shown as box-plots. The basal and claudin-low subtypes have shorter median doubling time as

compared to luminal and ERBB2^{AMP} subtypes, Kruskal-Wallis p value ($p = 0.006$). **B.** The ANCOVA model shows strong effects of both subtype and growth rate on response to 5-FU. Luminal (black) and basal/claudin-low (red) breast cancer lines each show significant associations to growth rate but have distinct slopes.

Figure S4. Heatmap of non-redundant PARADIGM activities for both cell line and TCGA samples. Cluster dendrogram represents Euclidian distance between samples and was created using Eisen Cluster and drawn using Java Treeview. Each row represents a network feature, each column represents a sample (tumor or cell line). Colored bars below dendrogram indicate sample subtype (upper) and sample cohort (lower). Overall, cell lines and tumors are intermixed, and subtypes tend to cluster together, indicating that the tumors and cell lines share many of the same network features.

Figure S5. Inferred pathway activities are more strongly correlated within subtypes than within cohorts. **A.** Histogram of t-statistics derived from Pearson correlations computed between cell lines and TCGA samples of the same subtype (red) compared to t-statistics of Pearson correlations between cell lines of different subtypes (black). X-axis corresponds to the Pearson correlation t-statistic; y-axis shows the density of (cell-line, cell-line) or (cell-line, TCGA sample) pairs. K-S test ($p < 1 \times 10^{-22}$) indicates cell lines and TCGA samples of the same subtype are more alike than cell lines of other subtypes. **B.** Pairwise Pearson correlations of cell line samples within the same subgroup (red) versus cell line samples in different subgroups (black). **C.** Pairwise correlations of TCGA breast cancer samples within the same subgroup (red) versus cell line samples in different subgroups (black).

Figure S6. The cell line networks are highly significant. The significance of the subpathways identified by our method was assessed by comparing the size of our subpathways to the size of the subpathways generated from a background model in which cells were randomly partitioned into groups, rather than in the original subtype definitions. The subpathway sizes were measured in two ways, the total number of nodes in the subpathway (A,C,E,G) and the number of nodes in the largest connected component of the subpathway (B,D,F,H). The luminal (A,B), ERBB2^{AMP} (C,D), claudin-low (E,F), and basal (G,H) subpathway sizes are shown as red dotted lines compared against the distribution of null subpathway sizes. In all cases the subpathway sizes for the true subtype partitioning are significantly larger than the subpathway sizes for the background model.

Dataset S1. Transcriptional, genomic and phenotypic characteristics of cell lines in the panel.

Dataset S2. Drug response data for each cell line tested against 77 therapeutic compounds. Data are $-\log_{10}$ transformed. These data were used to determine subtype specific responses.

Dataset S3. Pearson correlations between drug responses for all compound pairs.

Dataset S4. Subtype associations for all therapeutic compounds. Both raw p -values and FDR-corrected q -values are shown.

Dataset S5. Censored drug response data. GI50 values that are same as maximum experimental concentration used for different drugs were removed. Data are $-\log_{10}$ transformed. These data were used to identify responses associated with copy number aberrations.

Dataset S6. Non-redundant PARADIGM activities for cell lines and TCGA samples share similar subtype enrichment. Non-redundant pathway entities are along the rows, with subtype-specific SAM scores along the columns, for cell lines alone, TCGA samples alone, and the combined cell line and tumor samples (“BOTH”) as shown in Figure S4. Each SAM score has been ranked within the source-specific subtype and average ranks were computed between the cell lines and TCGA samples for the same subtype (luminal, basal, claudin-low and ERBB2^{AMP}).

Dataset S7. Subtypes for TCGA breast tumor samples, as determined by the TCGA AWG using hierarchical clustering of an intrinsic gene list. Data were obtained from the TCGA on November 6, 2010.

Waterfall plots

Waterfall plots of breast cancer subtypes and anti-cancer compounds follow after the supplementary figures. Association of clinical subtypes of breast cancer cell lines with 74 anti-cancer compounds. Each bar represents response sensitivity for one cell line, cell lines are ordered by sensitivity ($-\log_{10}(\text{GI}_{50})$) and colored to indicate subtype.

PARADIGM Network files

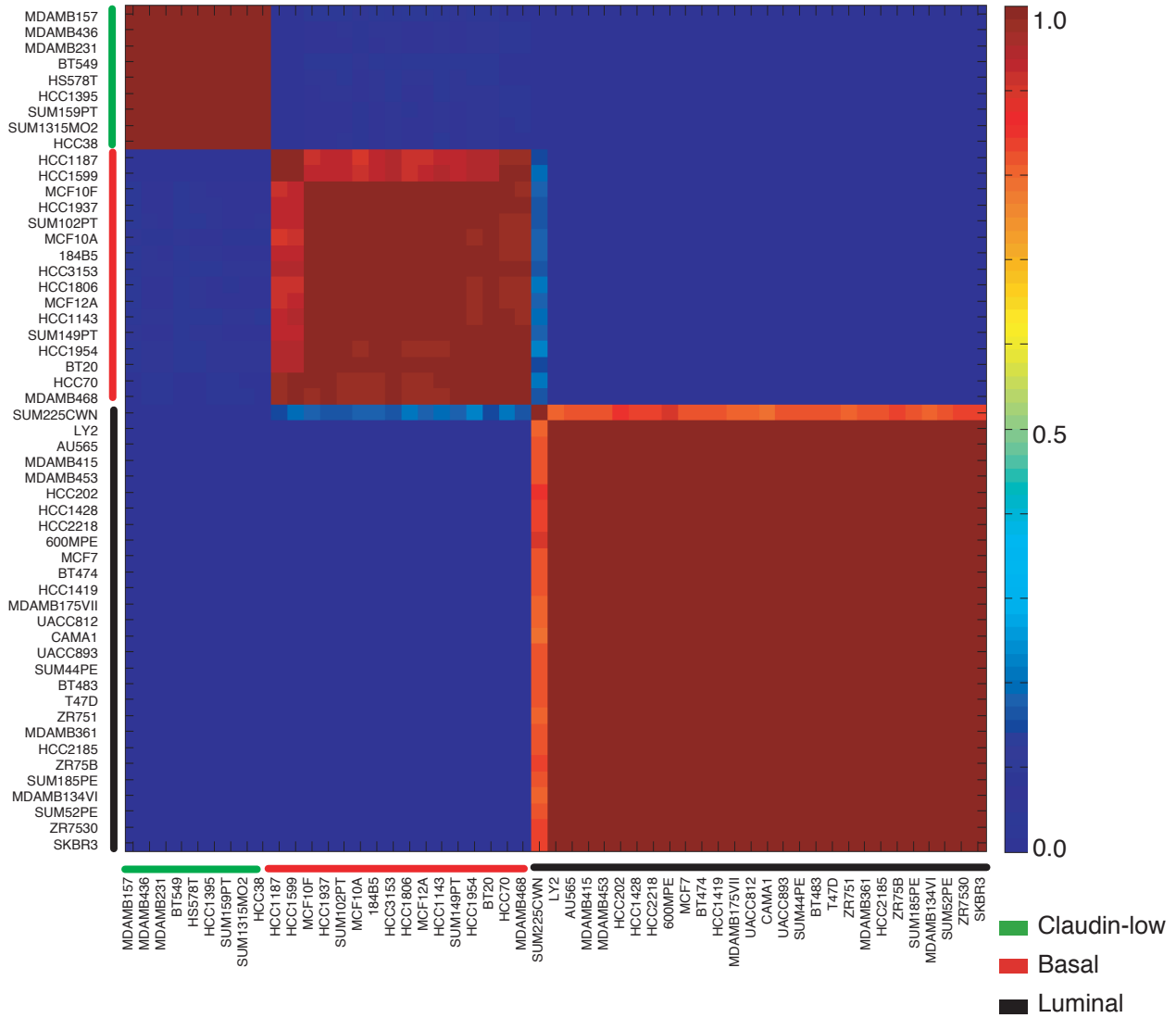
Cytoscape files for the PARADIGM breast cancer cell line networks are available at: <http://users.soe.ucsc.edu/~jstuart/heiser2011/>

References

1. Vaske CJ, *et al.* (Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26(12):i237-245.
2. de Hoon MJ, Imoto S, Nolan J, & Miyano S (2004) Open source clustering software. *Bioinformatics* 20(9):1453-1454.

Figure S1

A



B

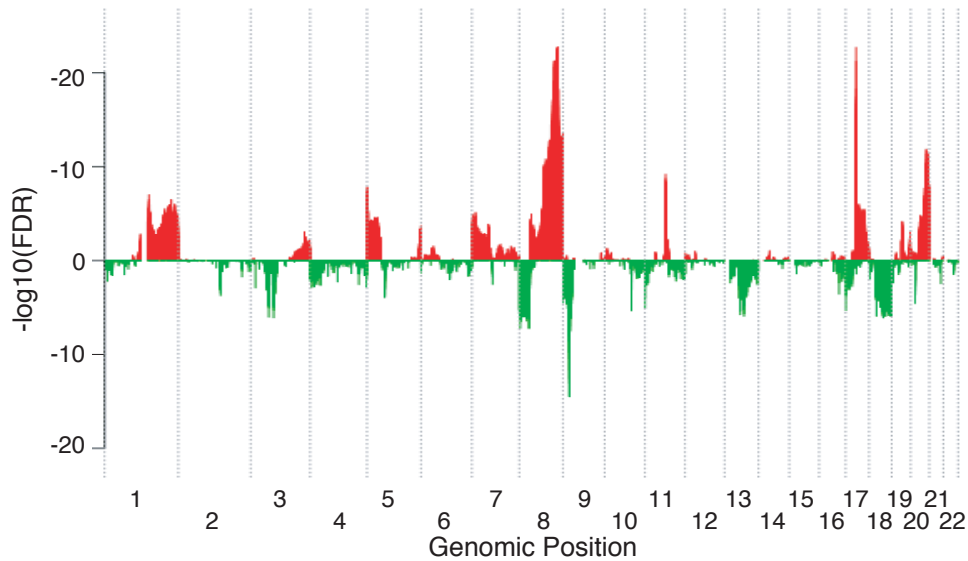


Figure S2

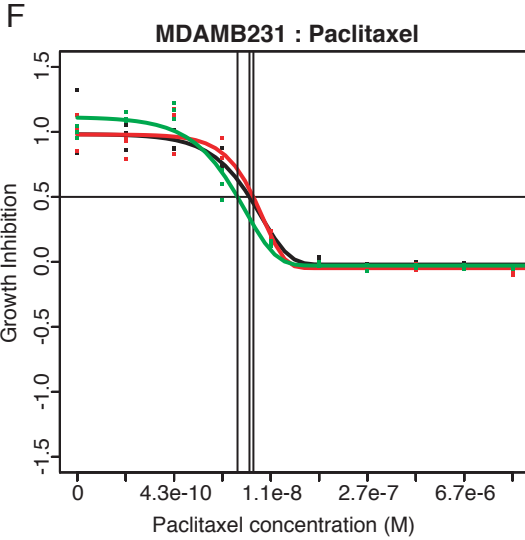
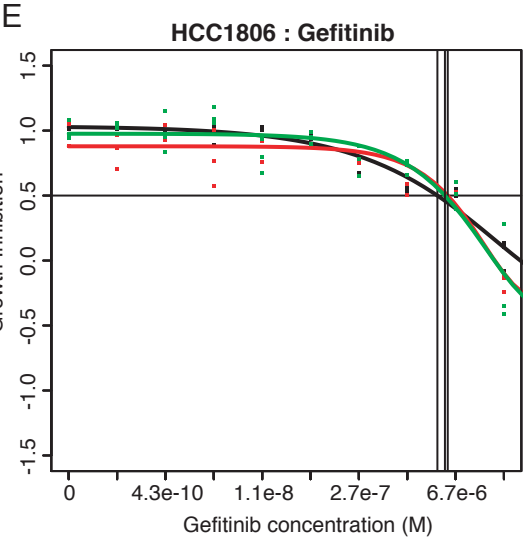
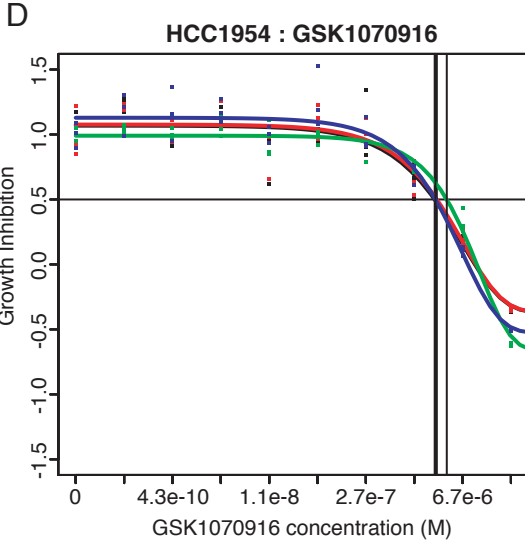
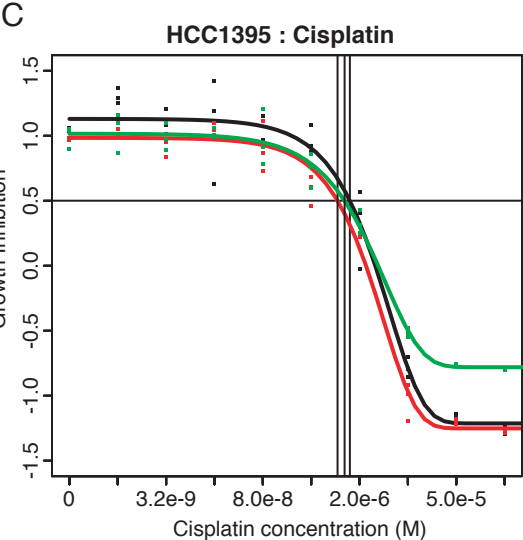
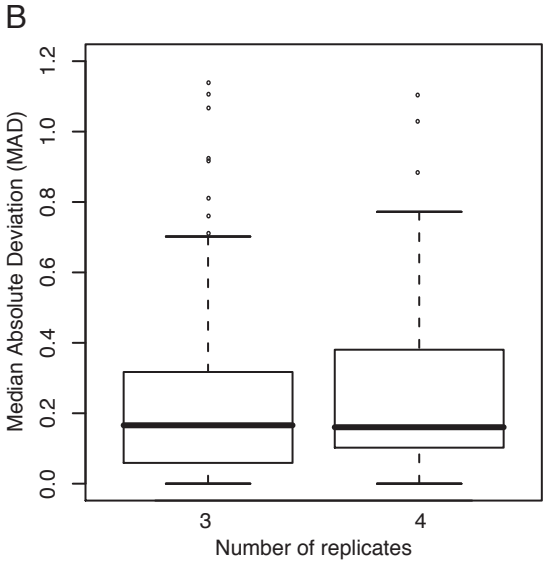
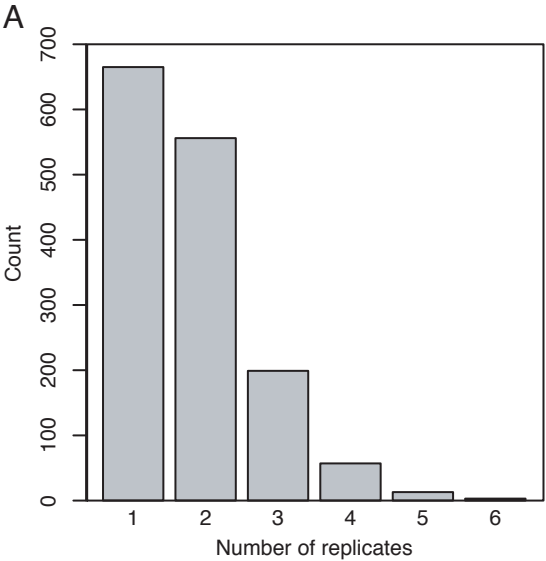


Figure S3

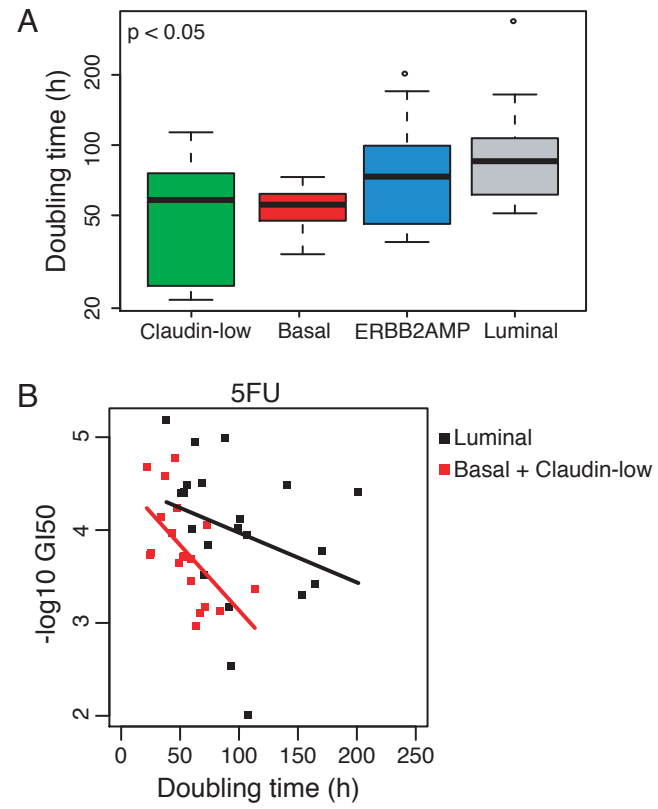


Figure S4

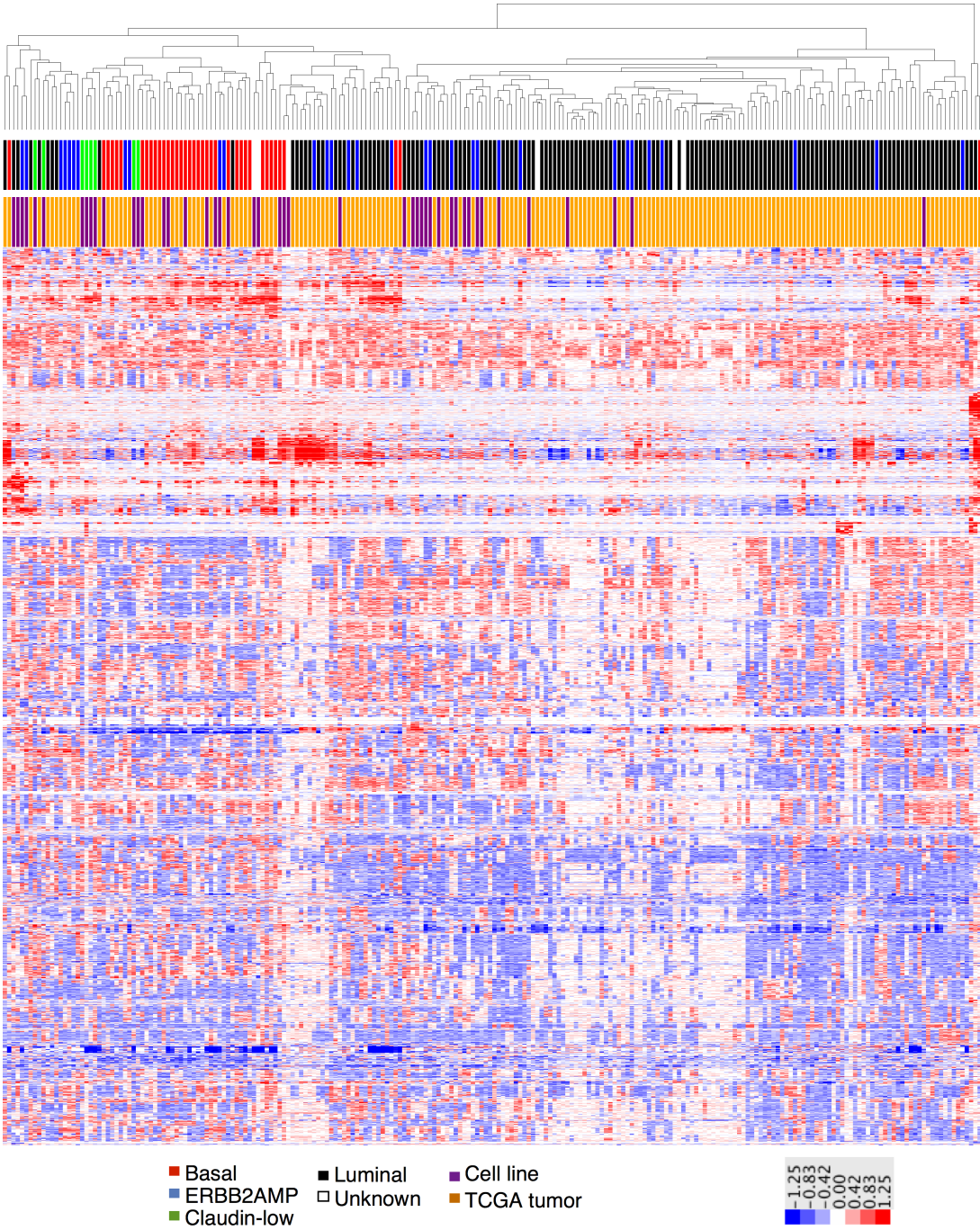


Figure S5

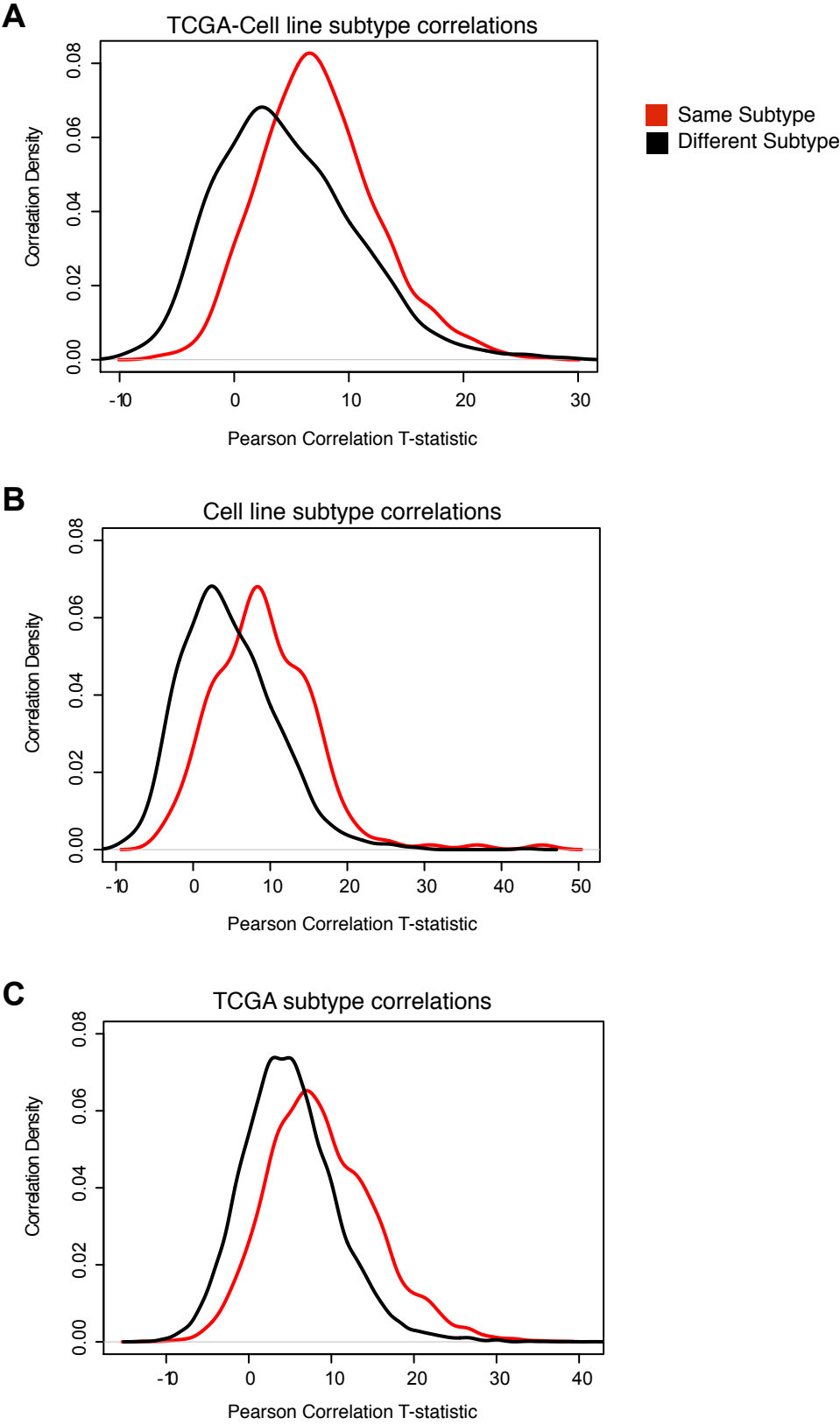
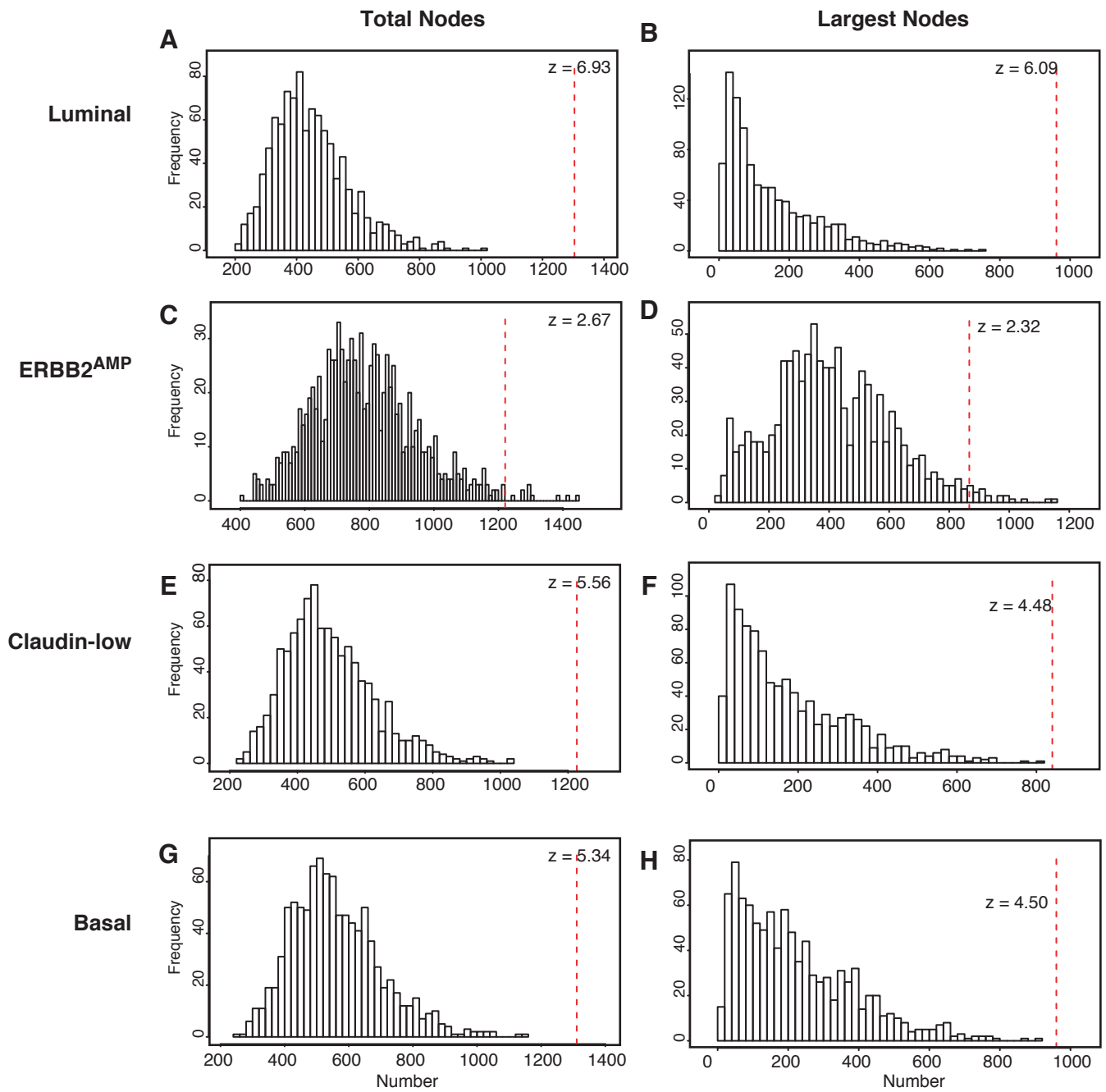
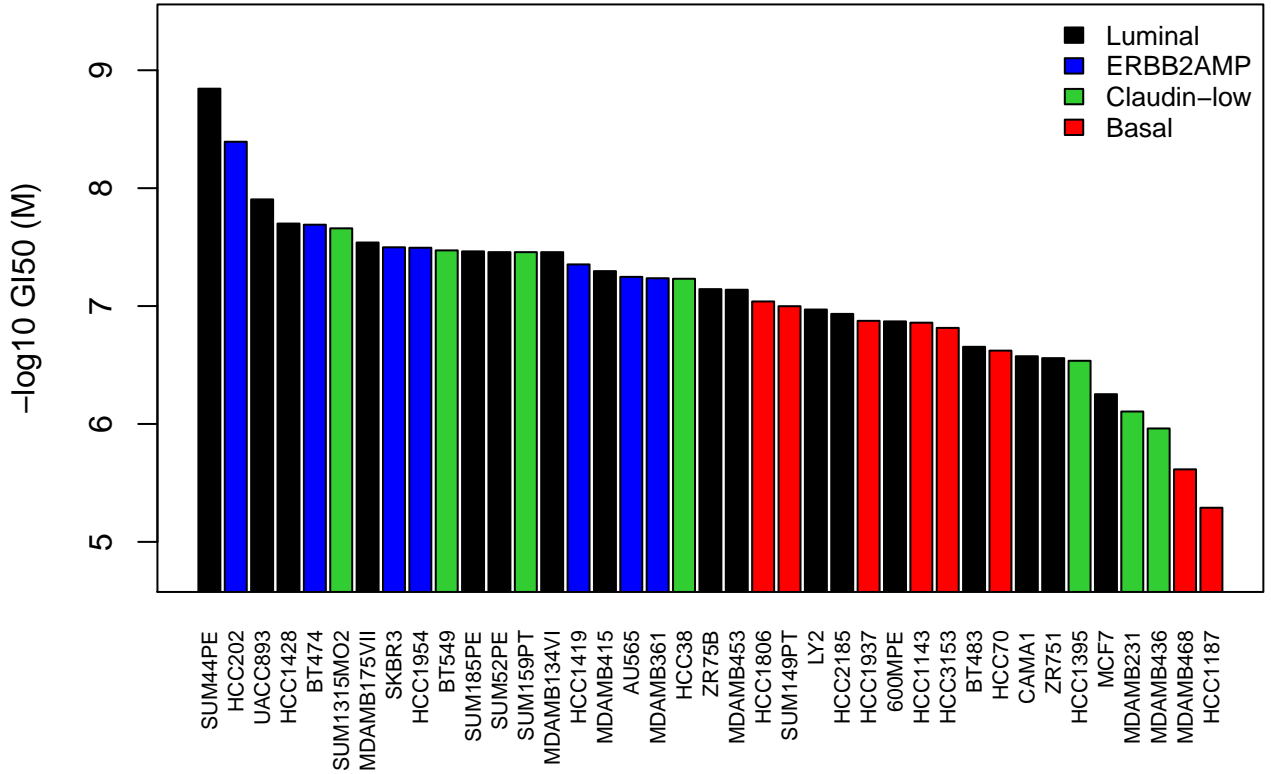


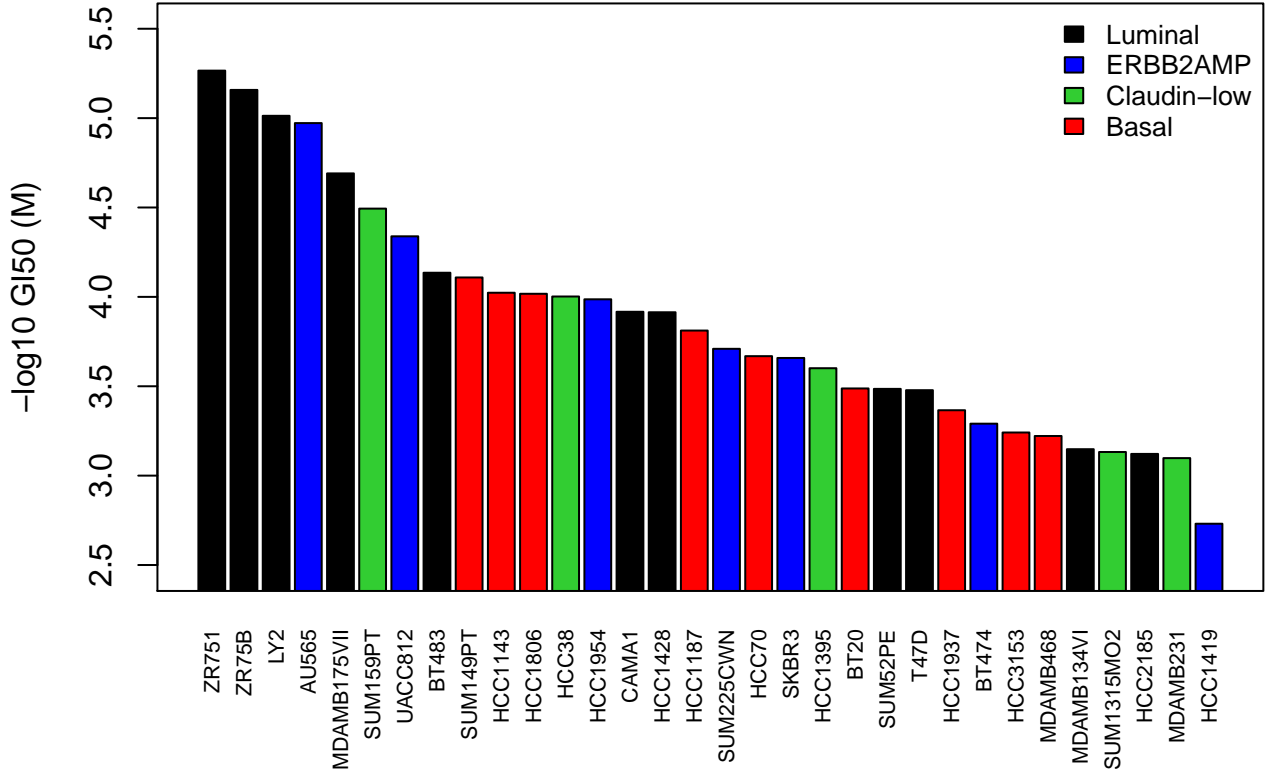
Figure S6



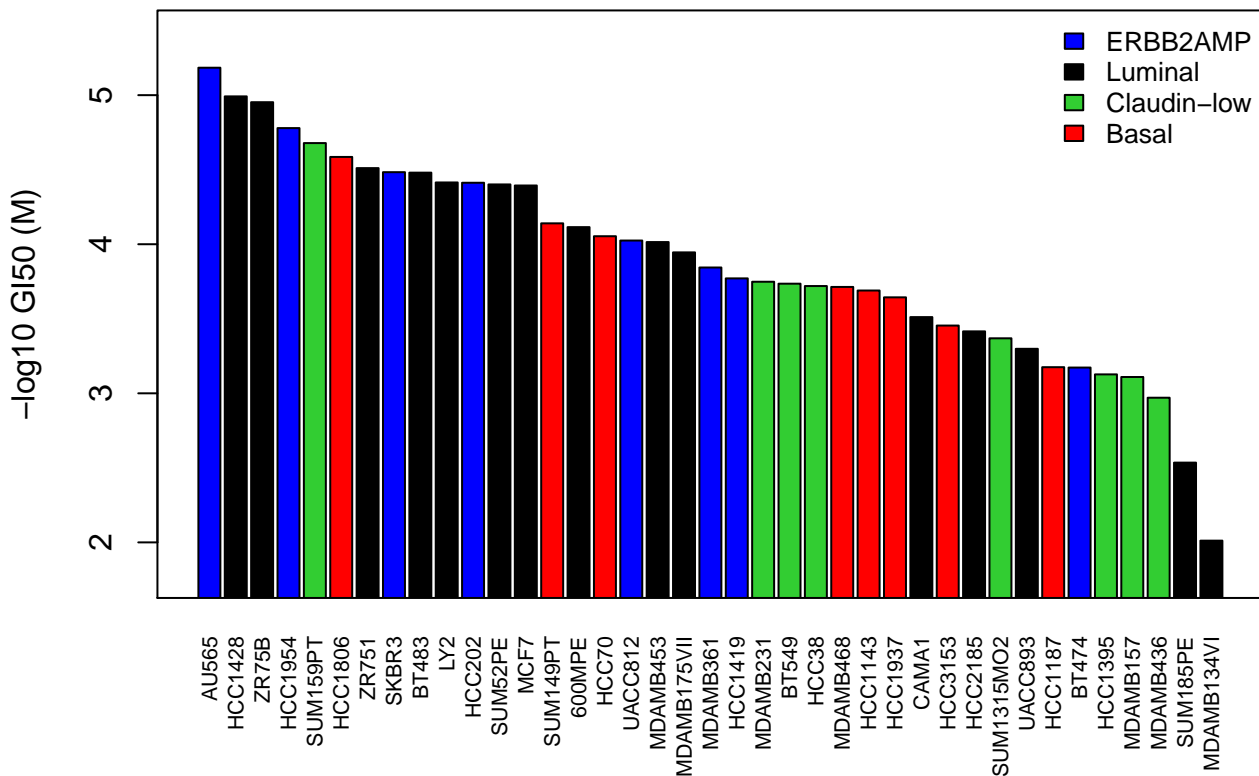
17-AAG (Hsp90)



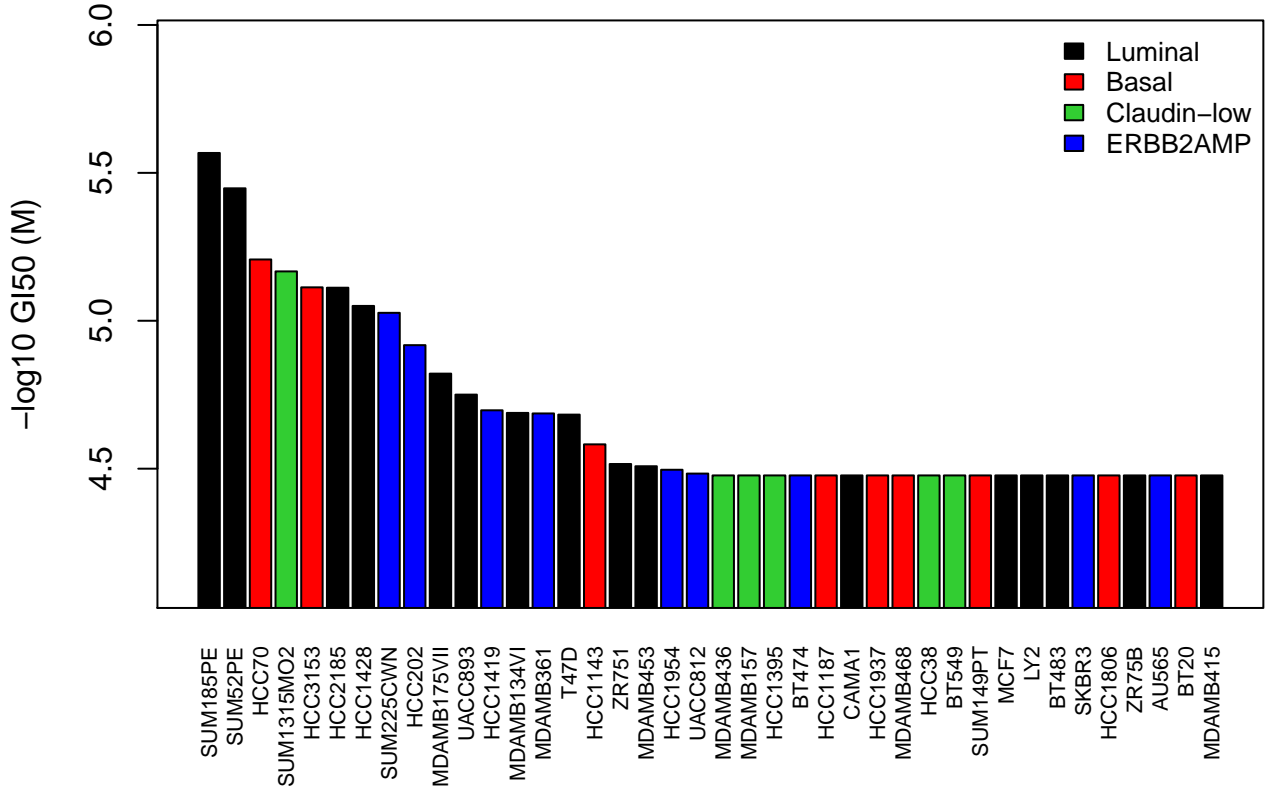
5-FdUR (DNA)



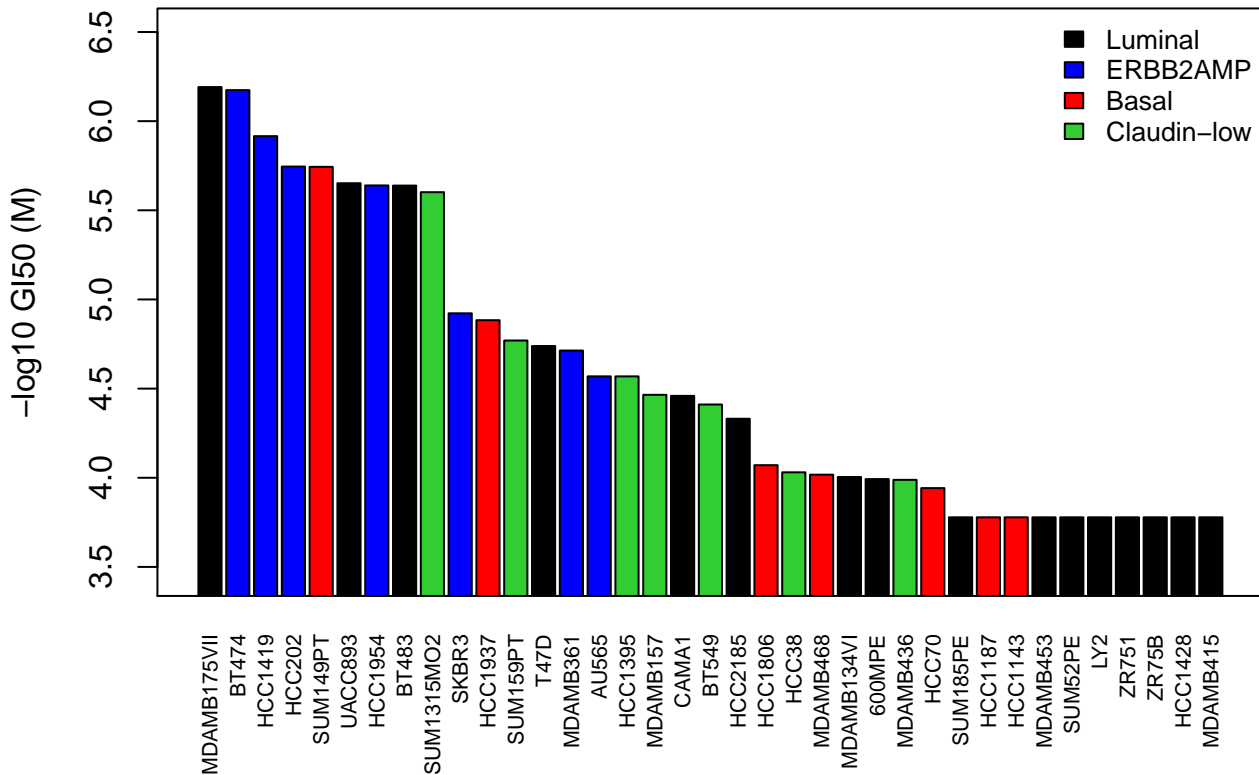
5-FU (pyrimidine analog, thymidylate synthase)



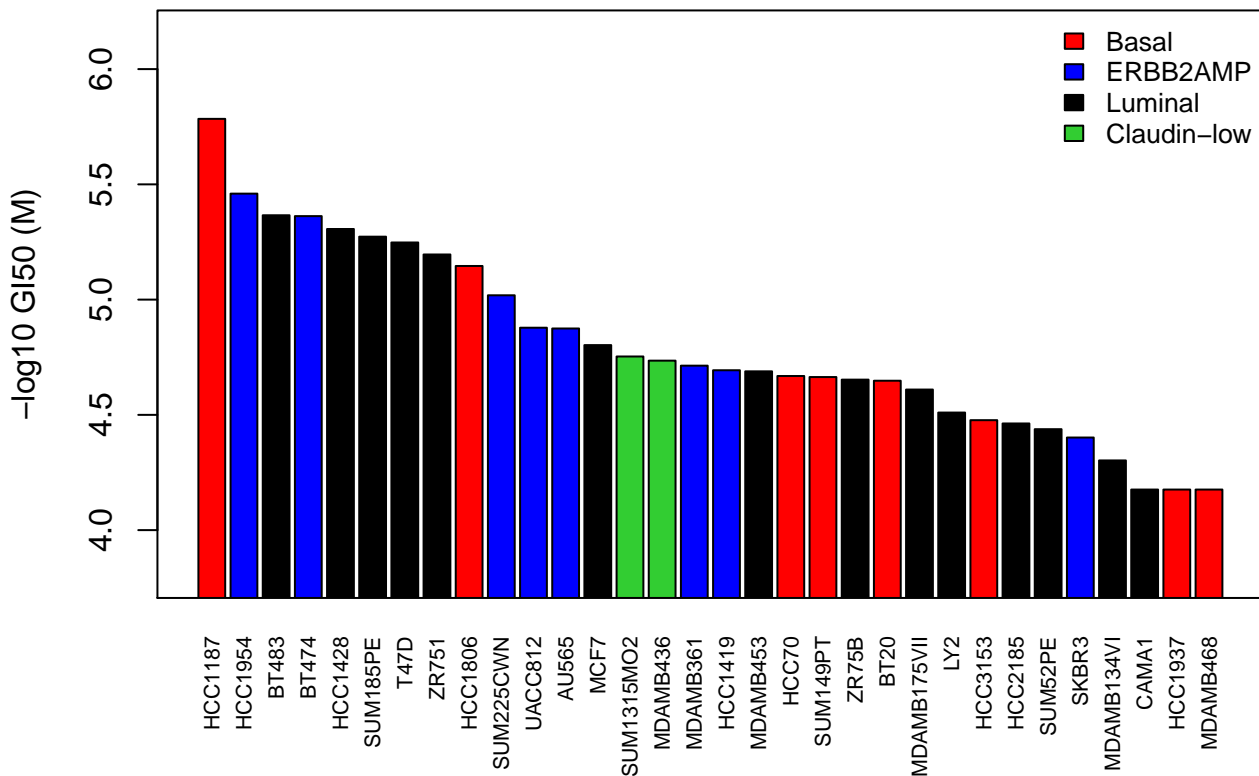
AG1024 (IGF1R)



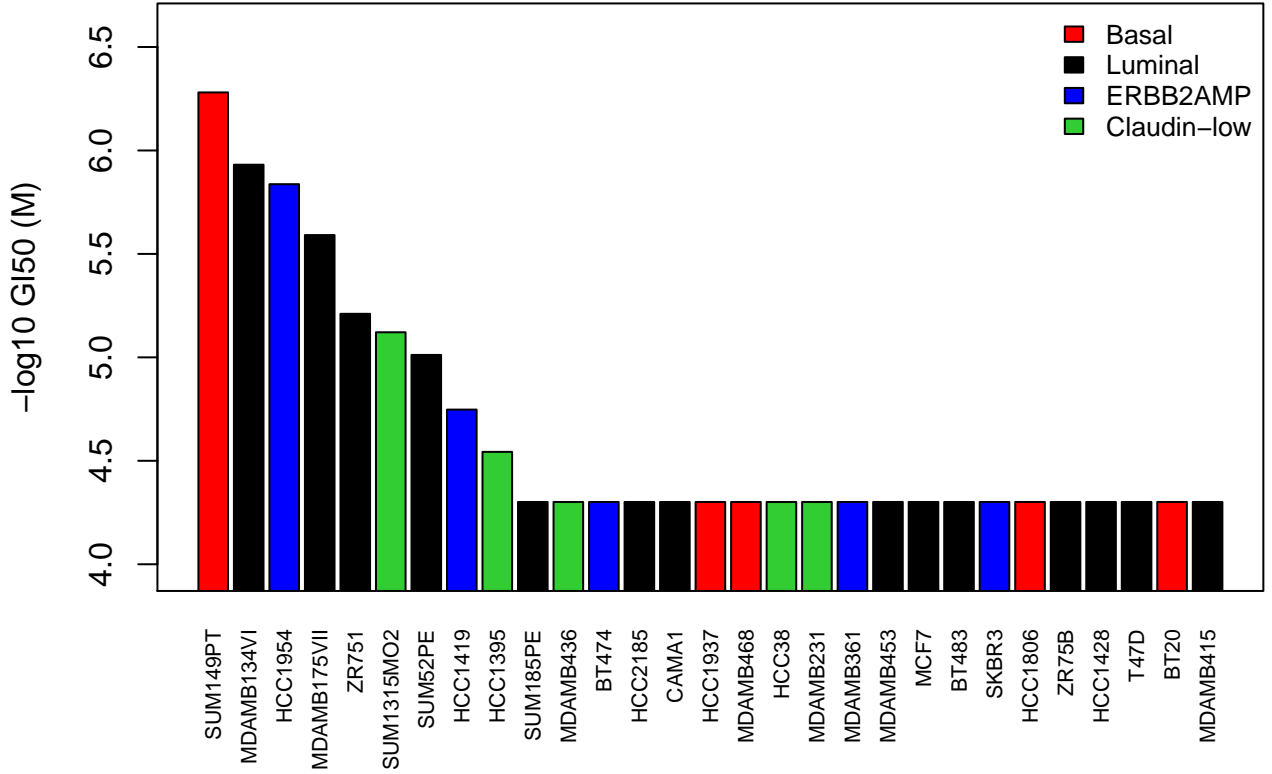
AG1478 (EGFR)



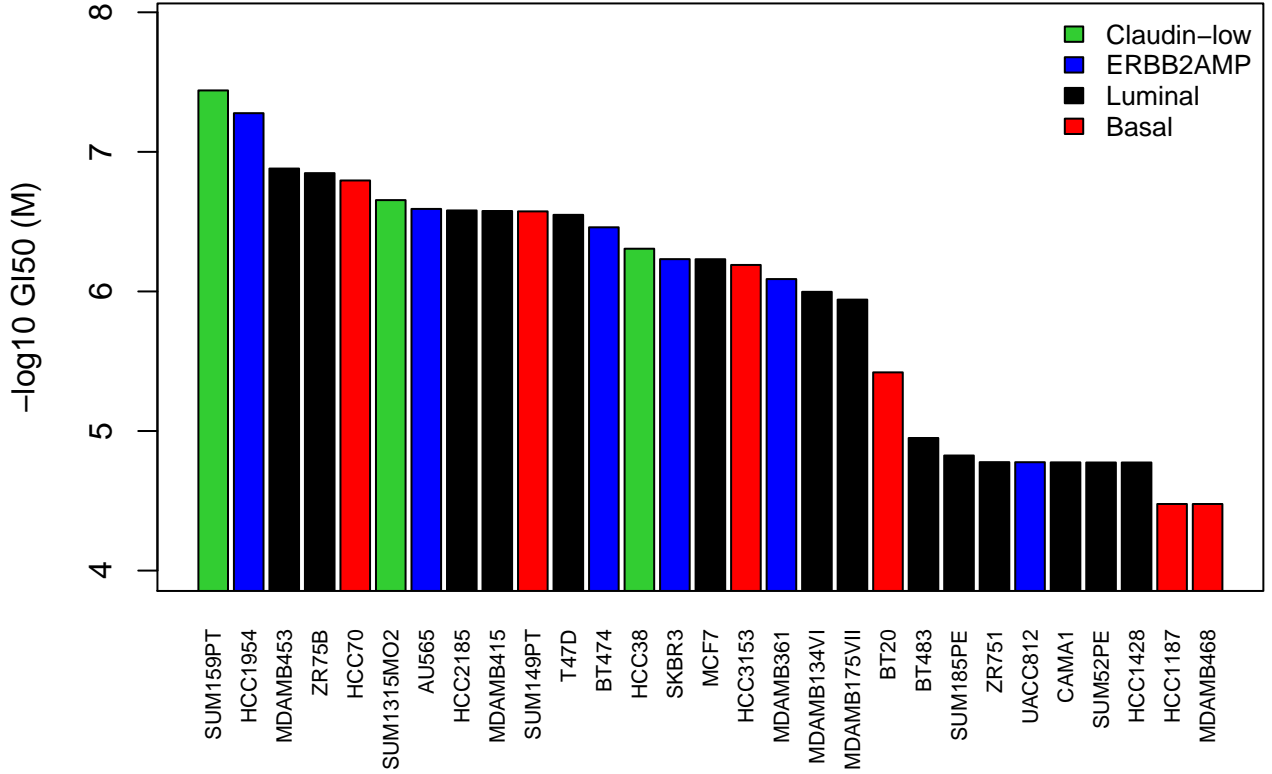
AS-252424 (PI3K gamma)



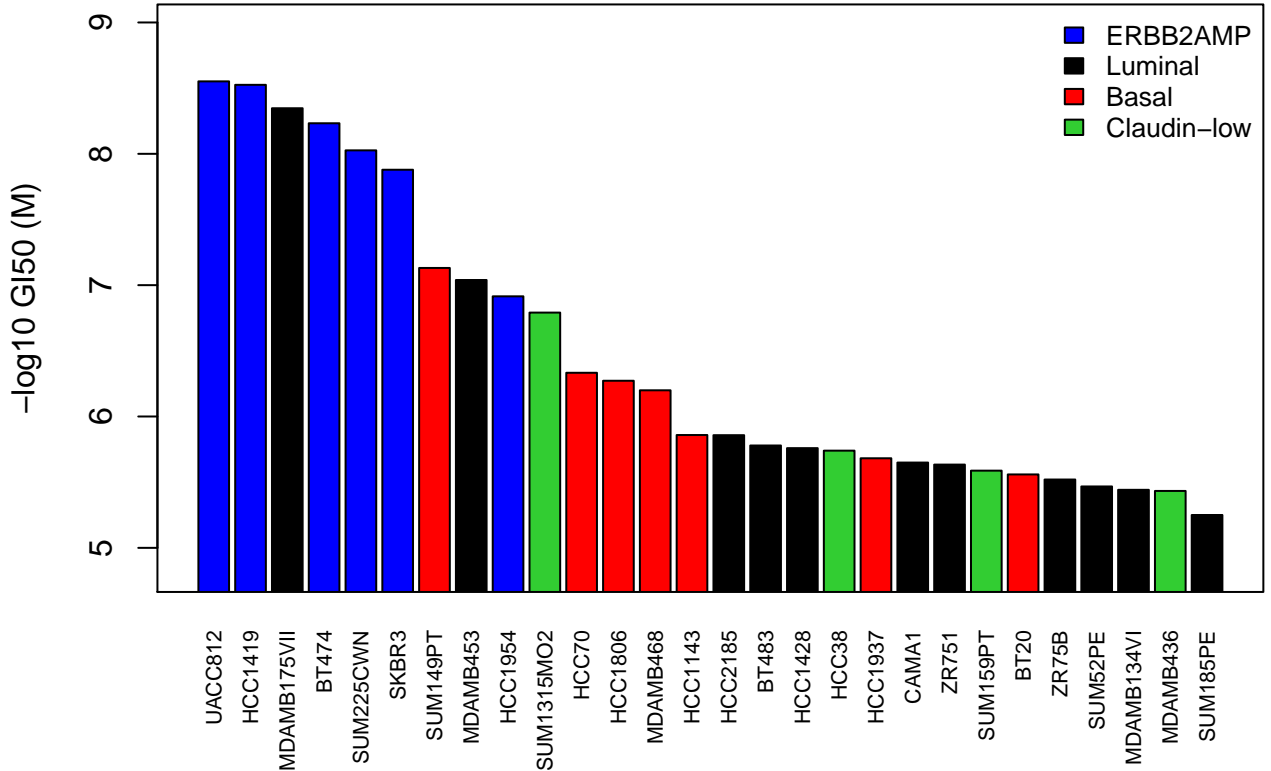
AZD6244 (MEK)



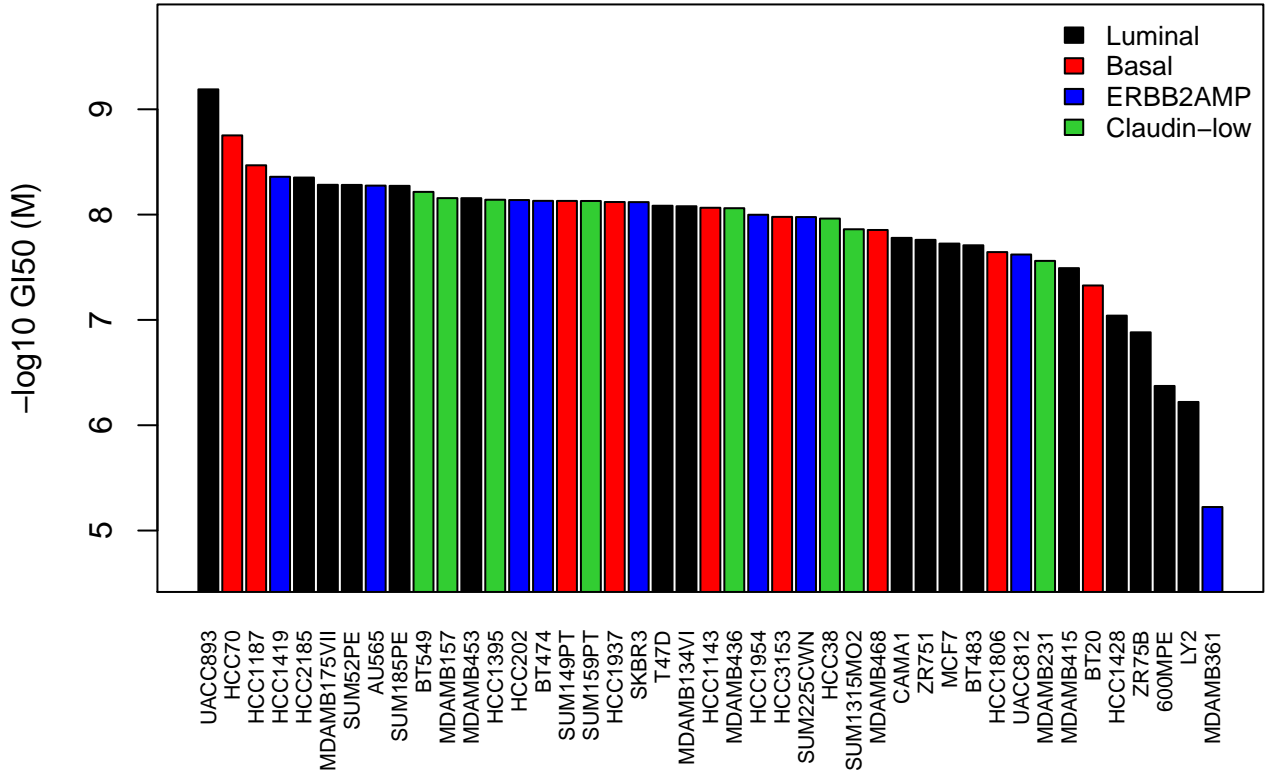
BEZ235 (PI3K)



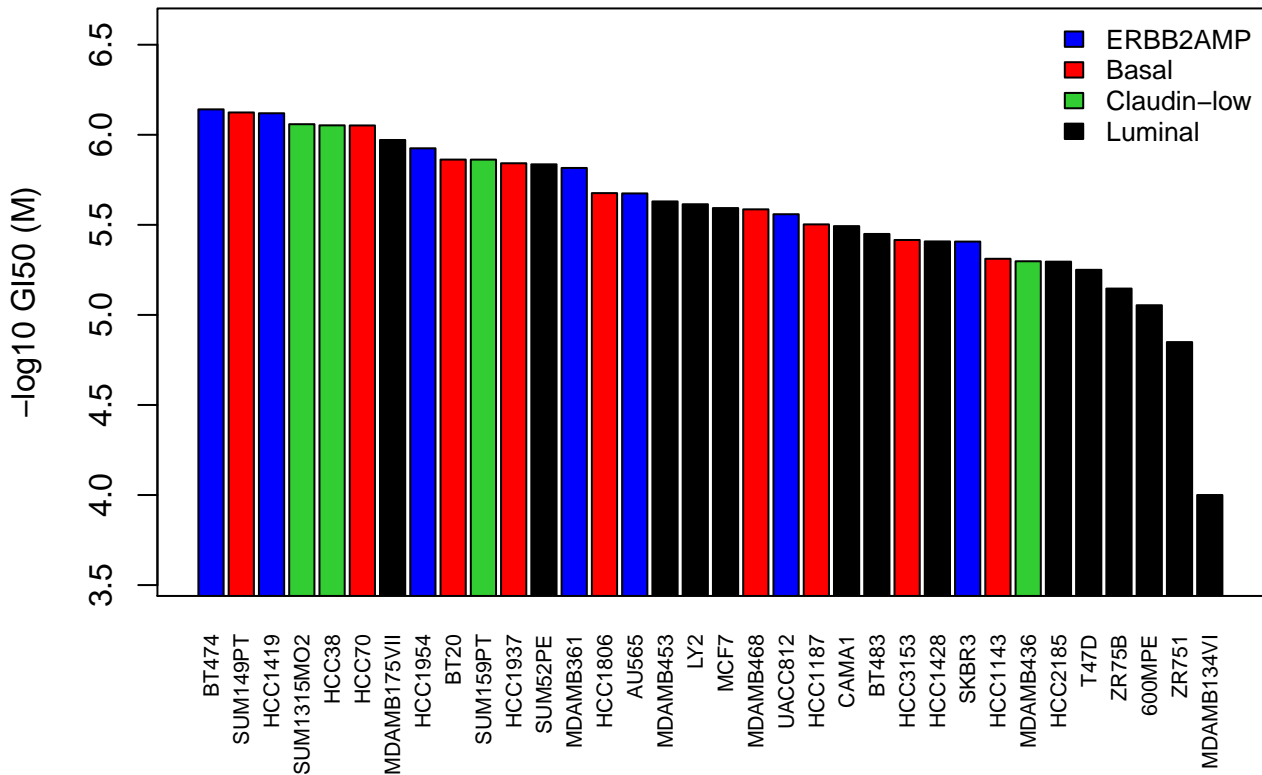
BIBW 2992 (EGFR and HER2 inhibitor)



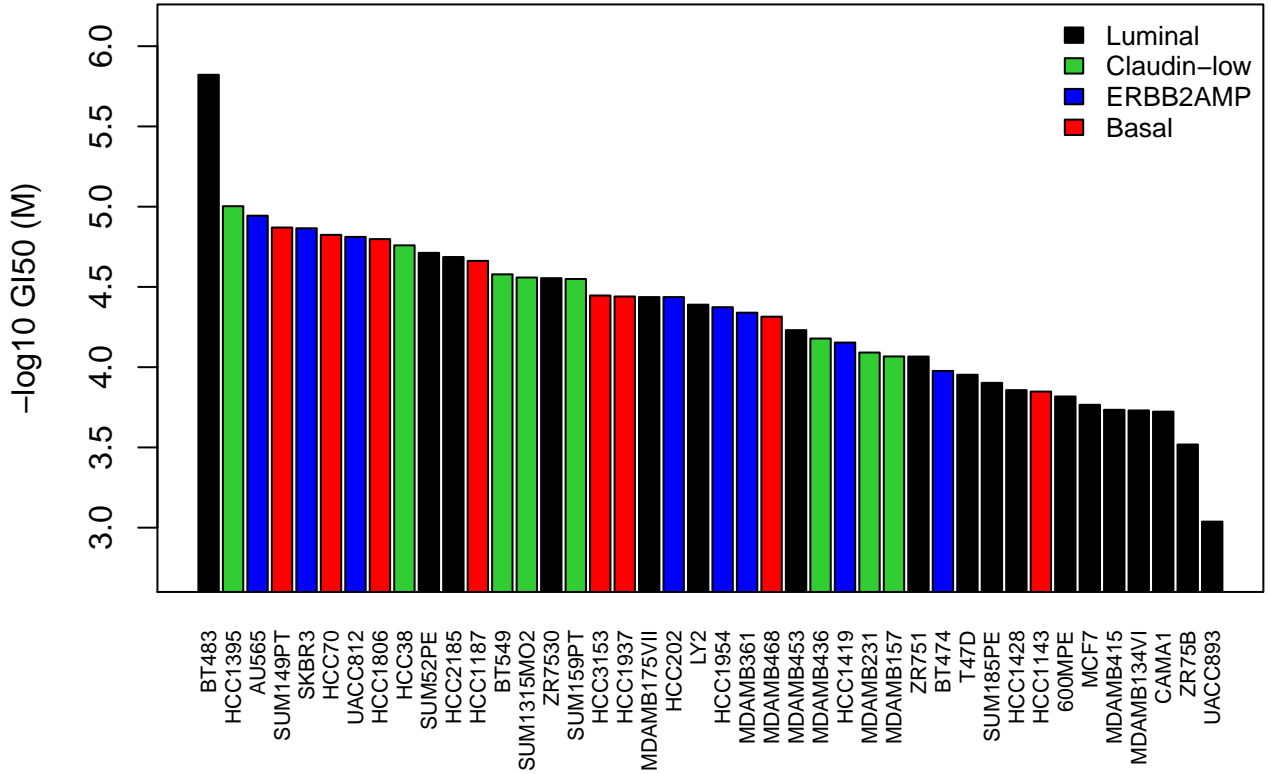
Bortezomib (Proteasome, NFkB)



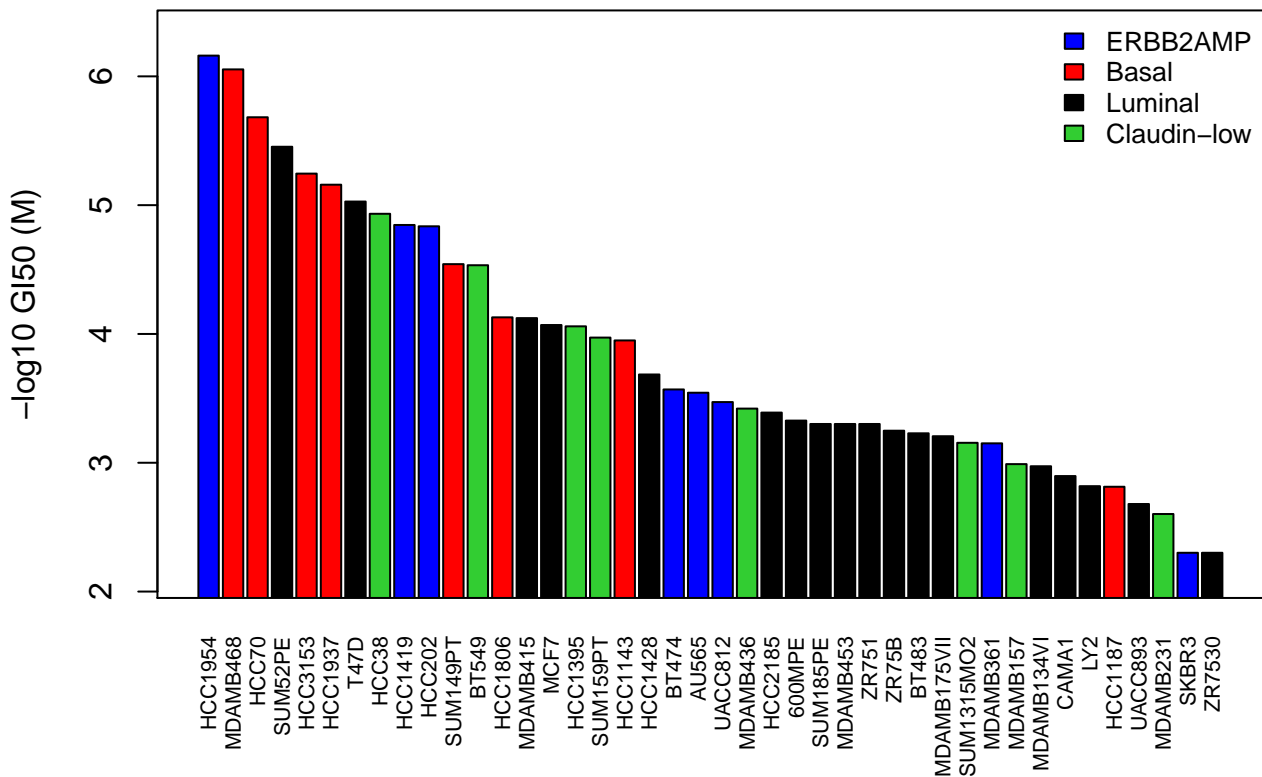
Bosutinib (Src)



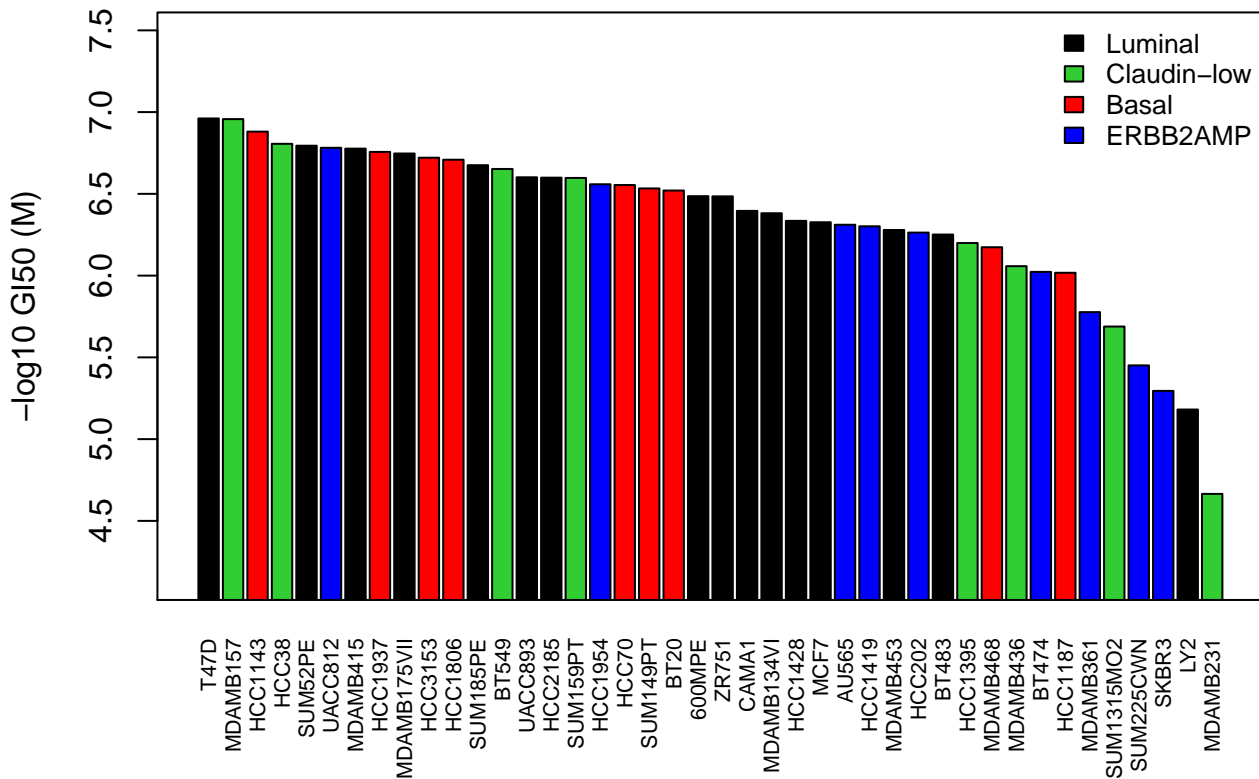
Carboplatin (DNA cross-linker)



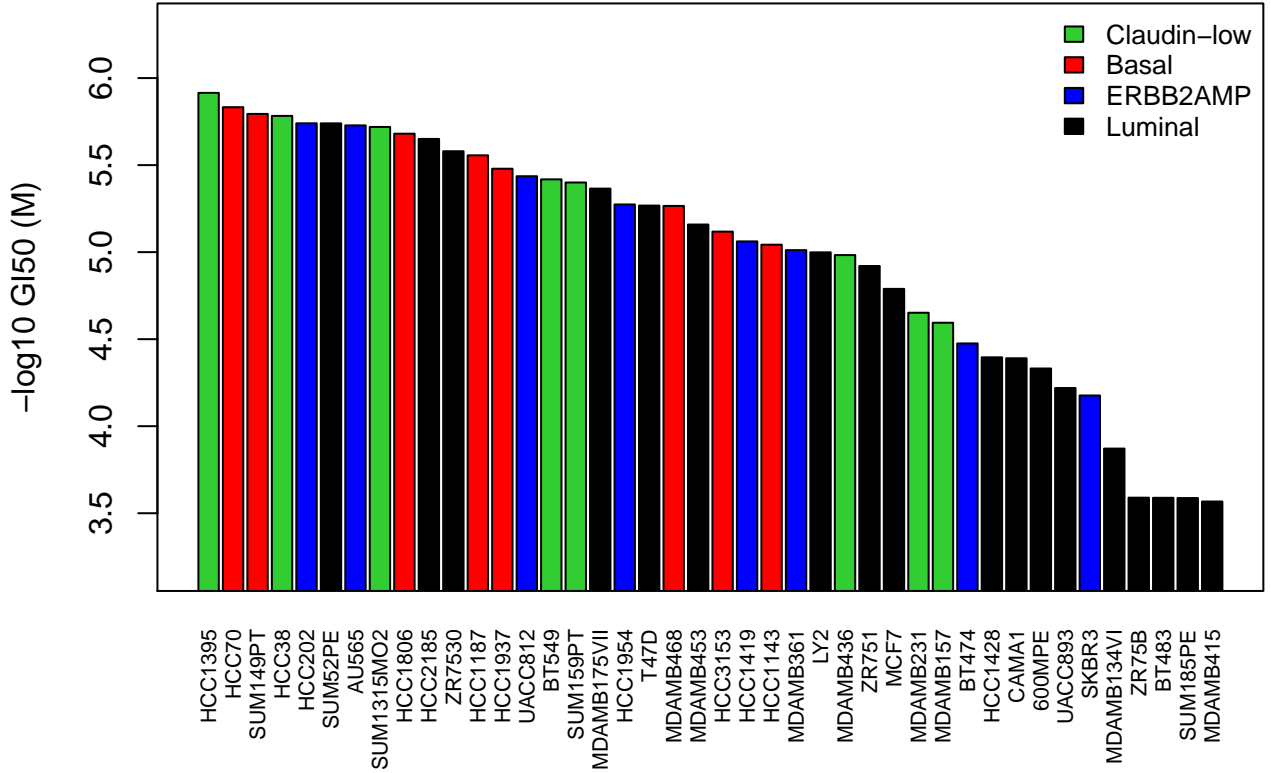
CGC-11047 (polyamine analogue)



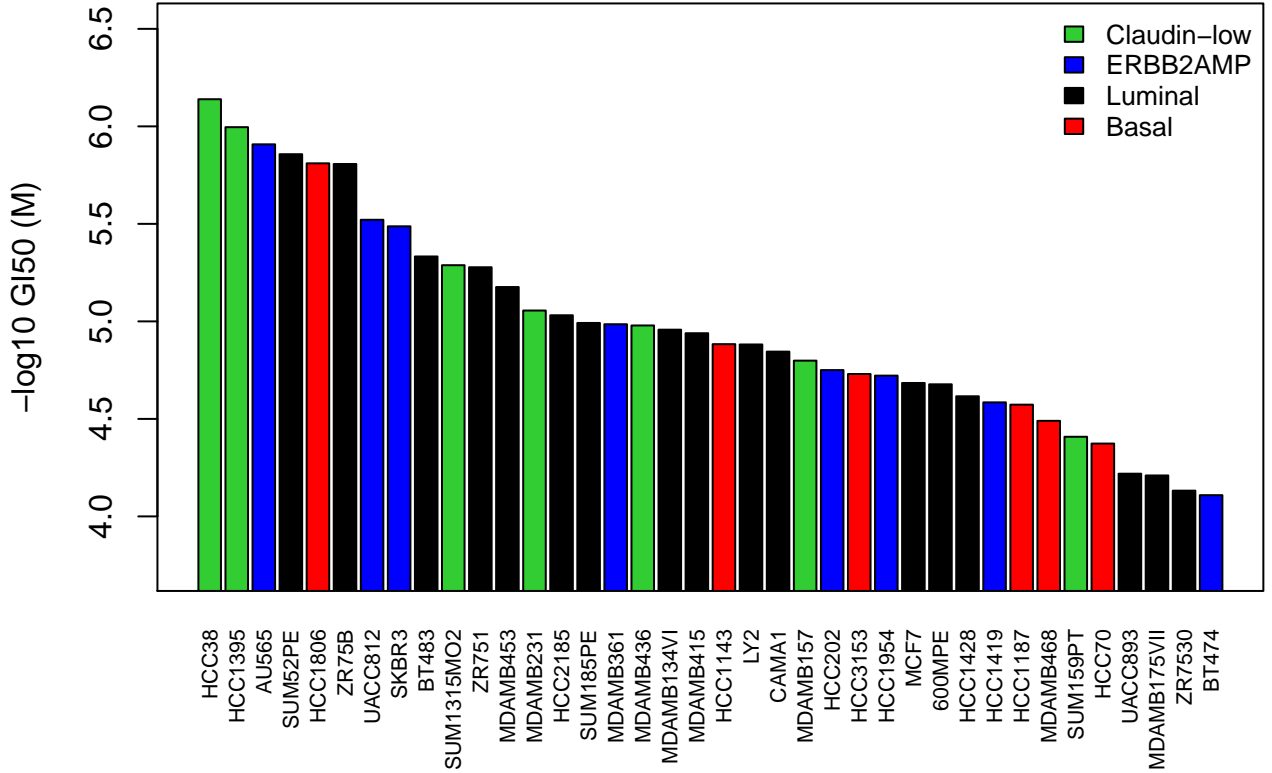
CGC-11144 (polyamine analogue)



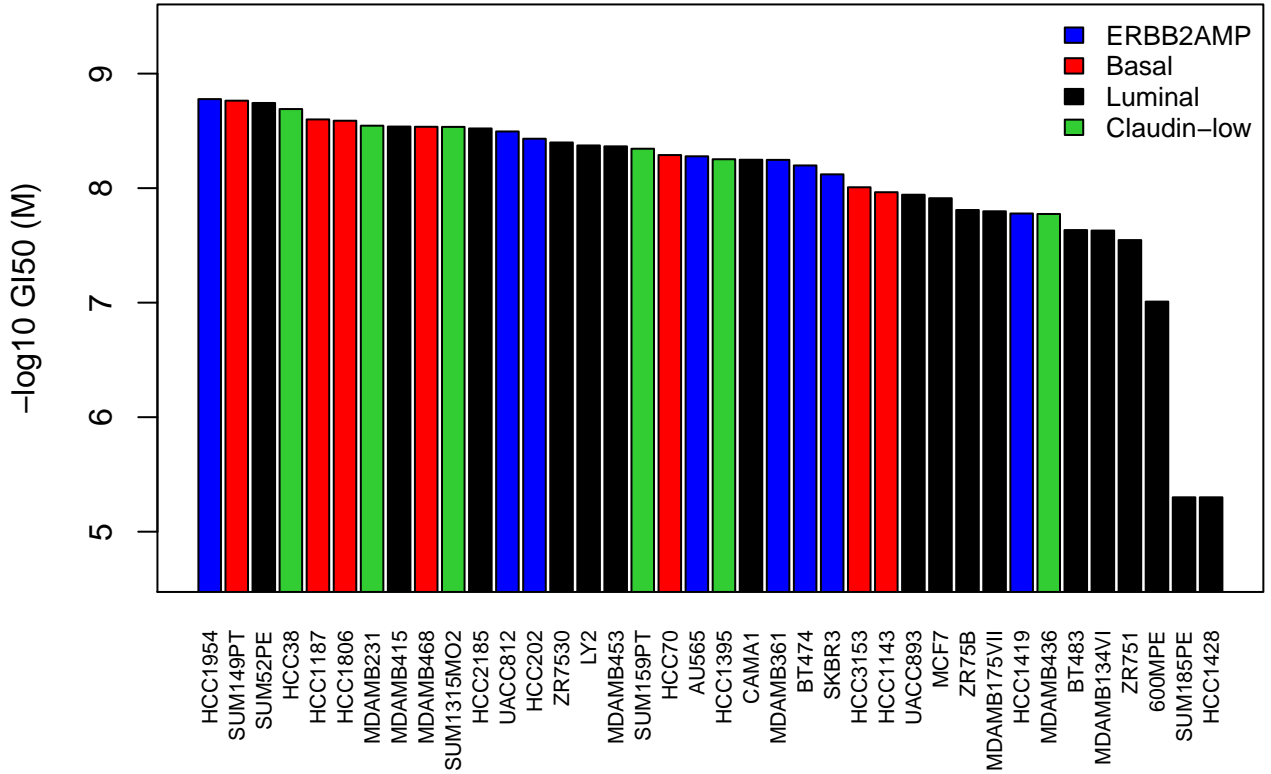
Cisplatin (DNA cross-linker)



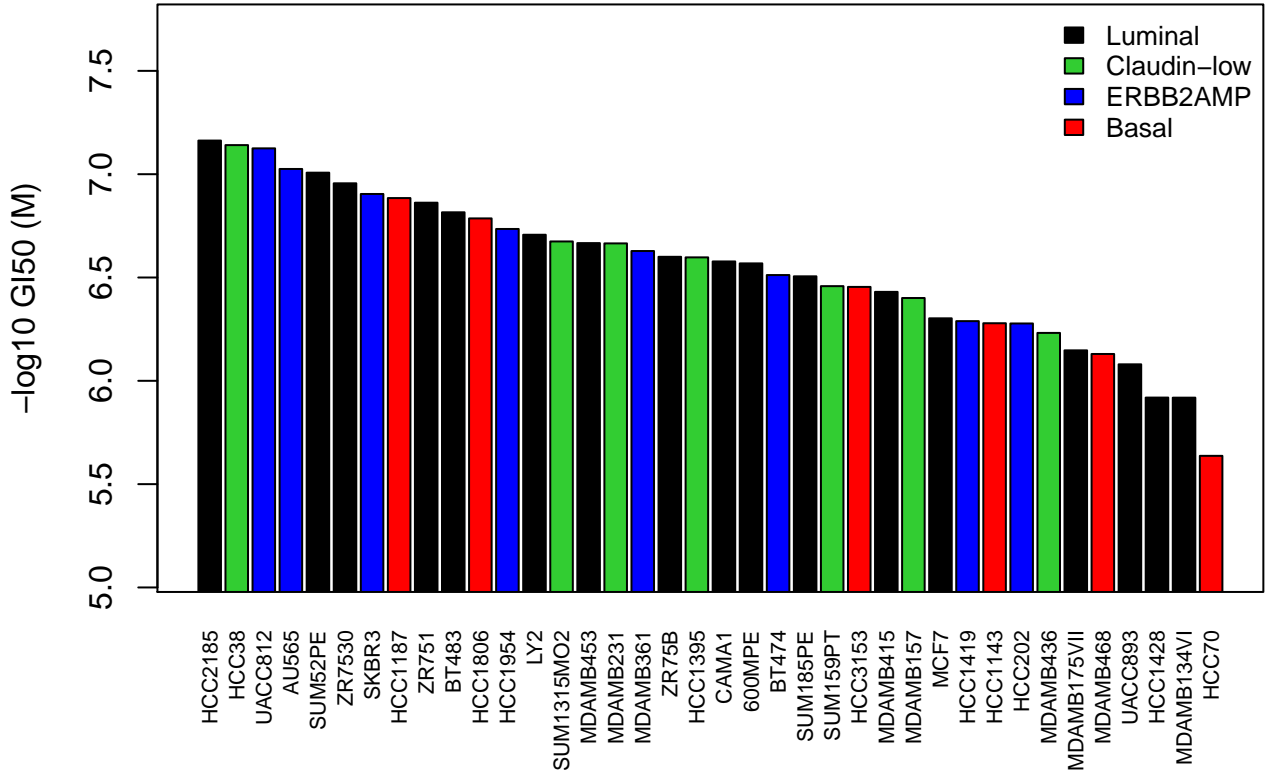
CPT-11 (Topoisomerase I)



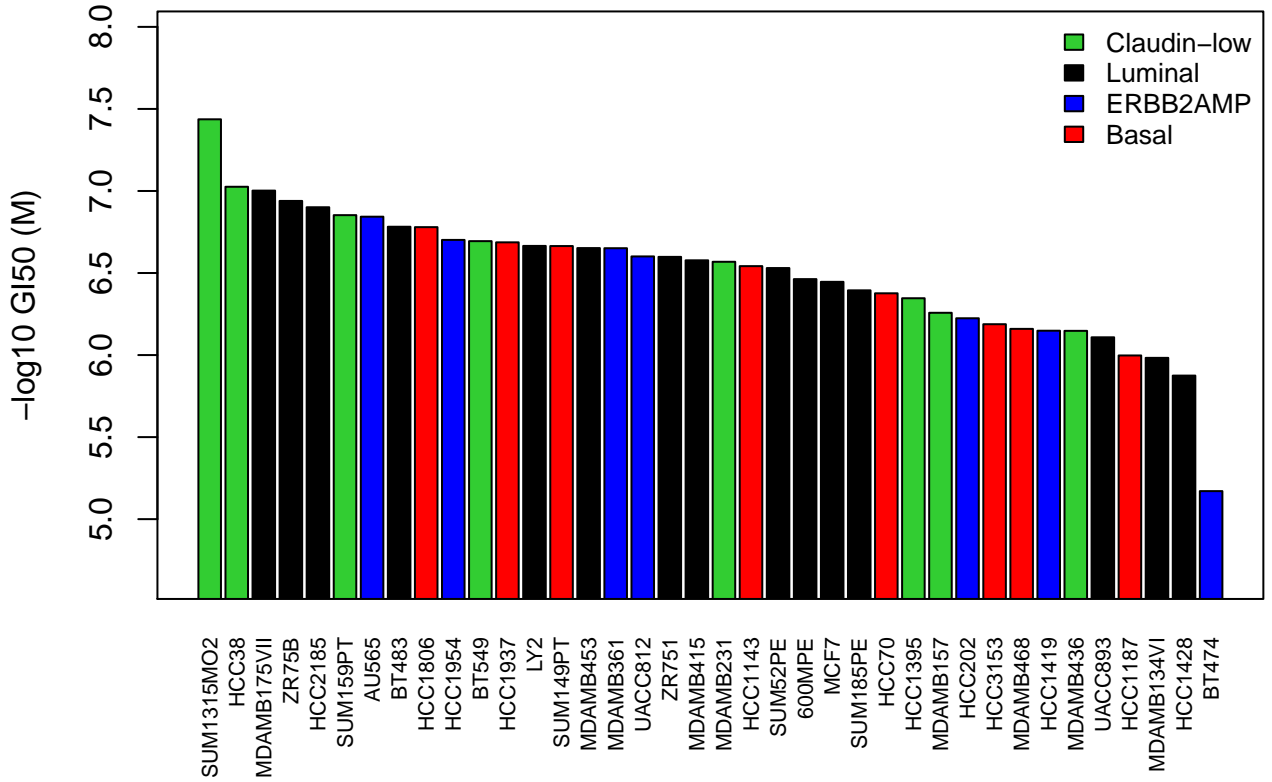
Docetaxel (Microtubule)



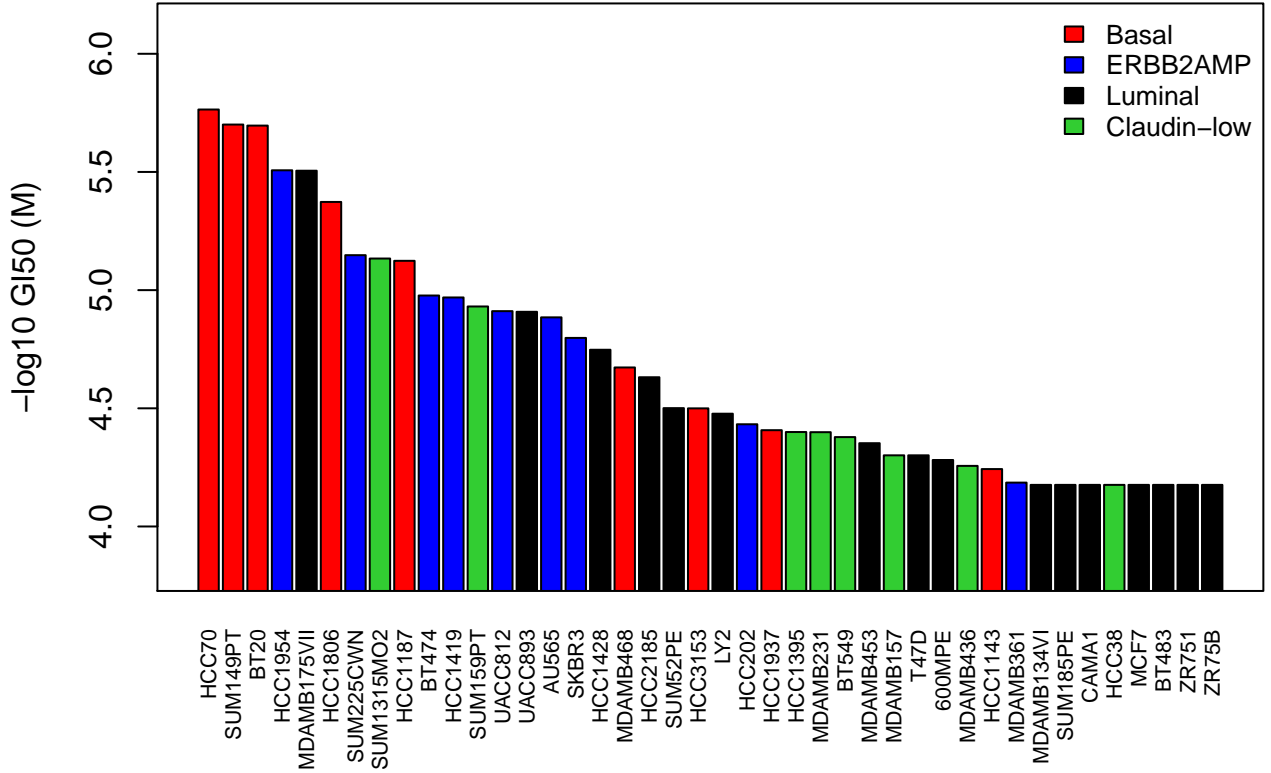
Doxorubicin (Topoisomerase II)



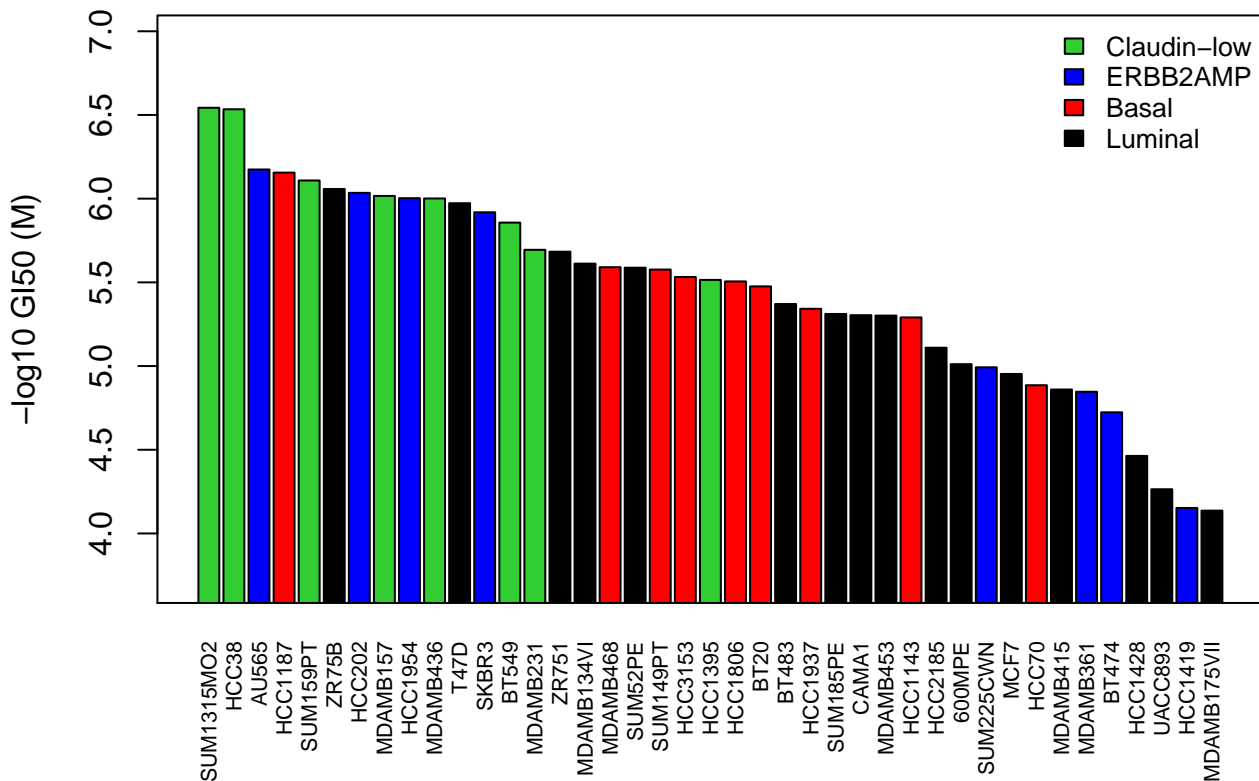
Epirubicin (Topoisomerase II)



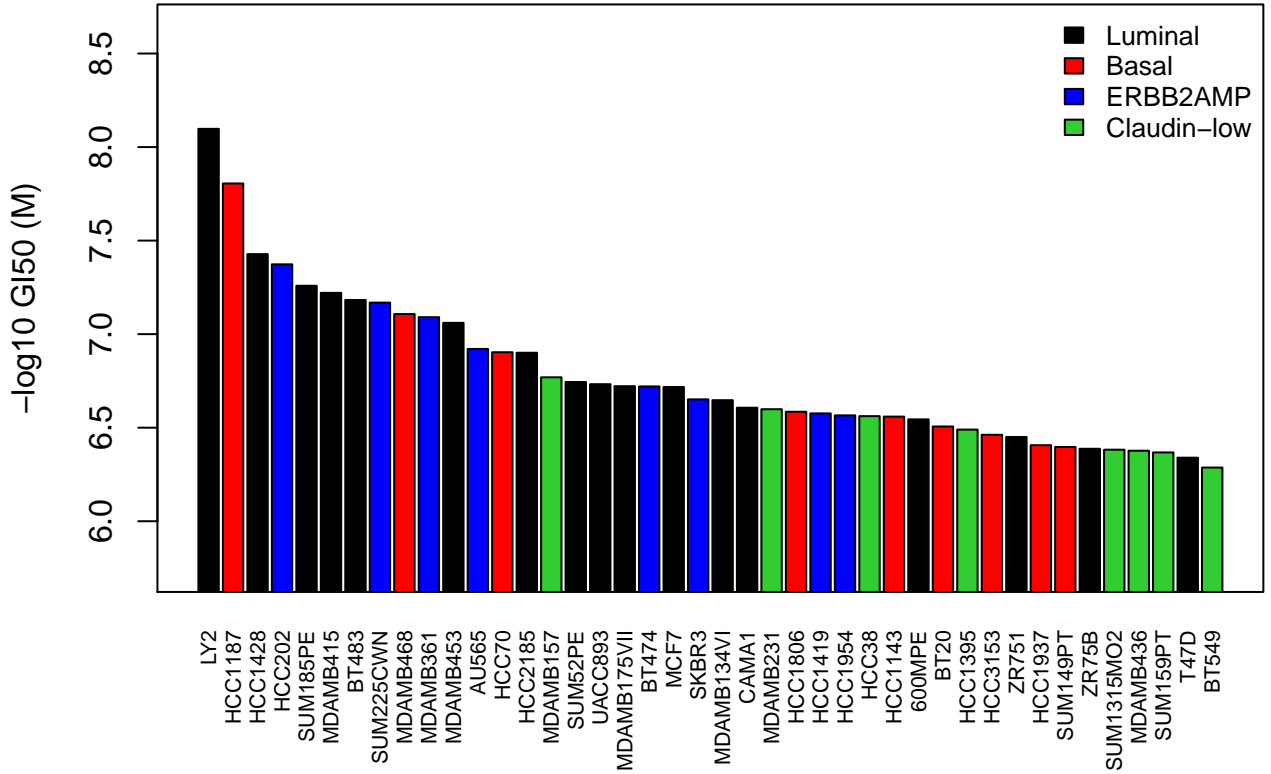
Erlotinib (EGFR)



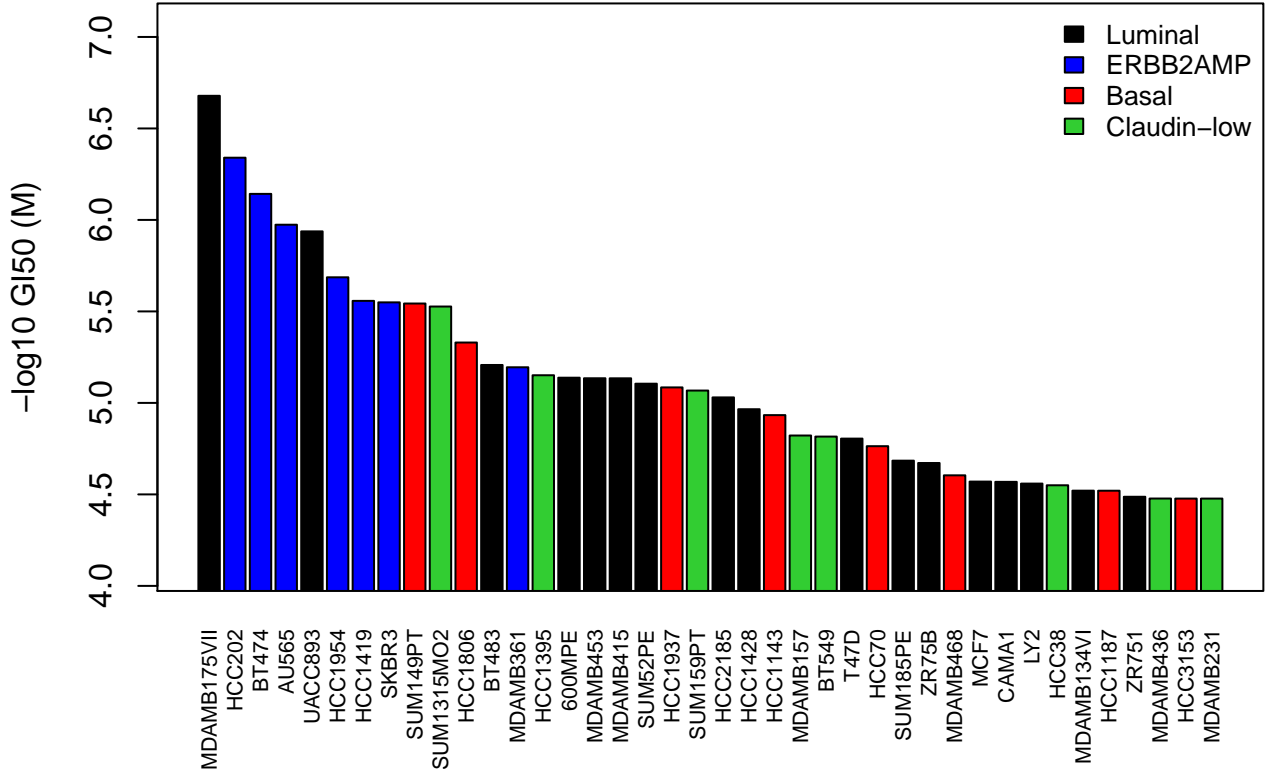
Etoposide (Topoisomerase II)



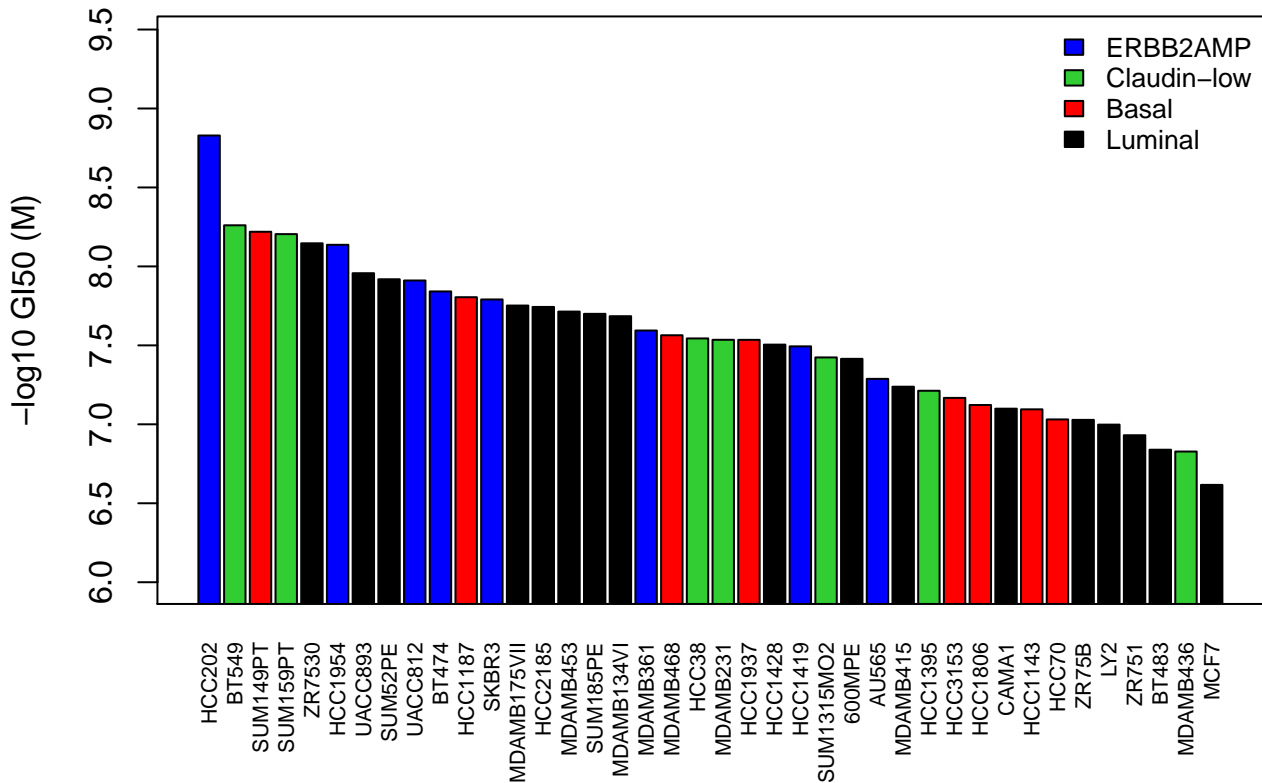
Fascaplysin (CDK)



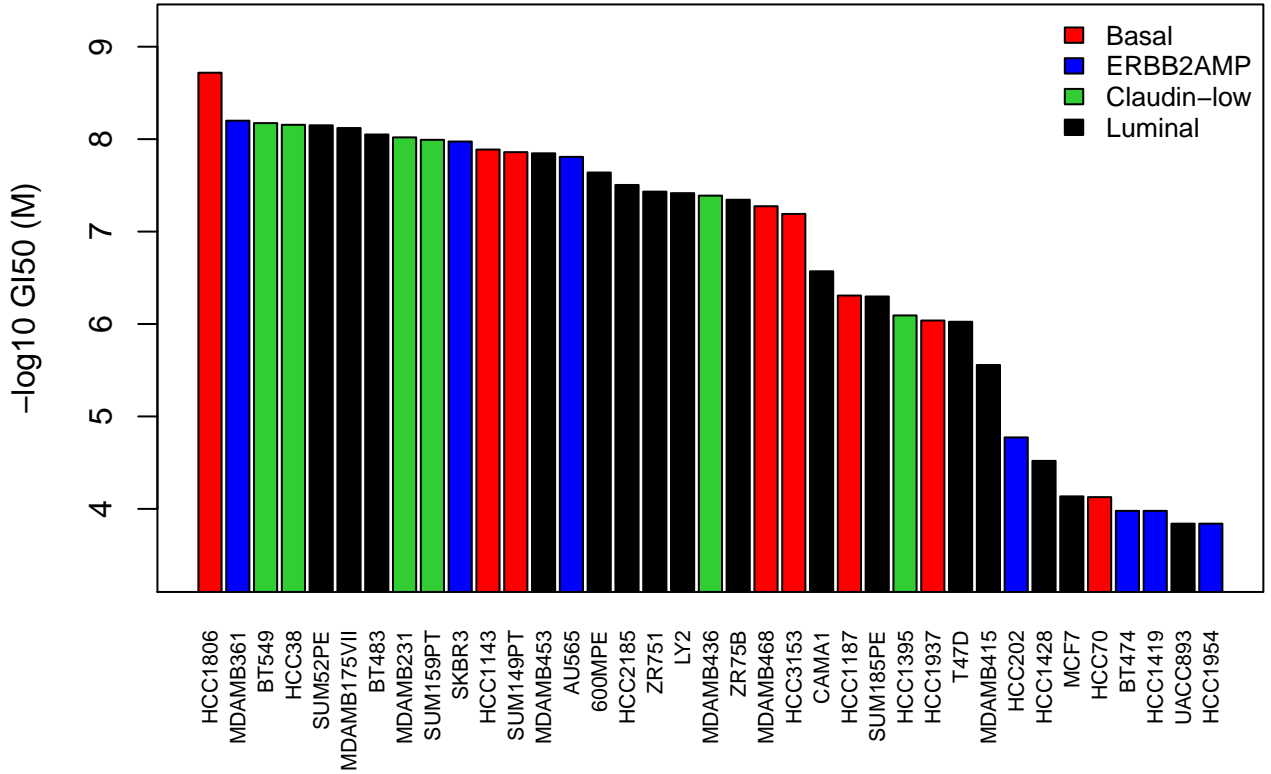
Gefitinib (EGFR)



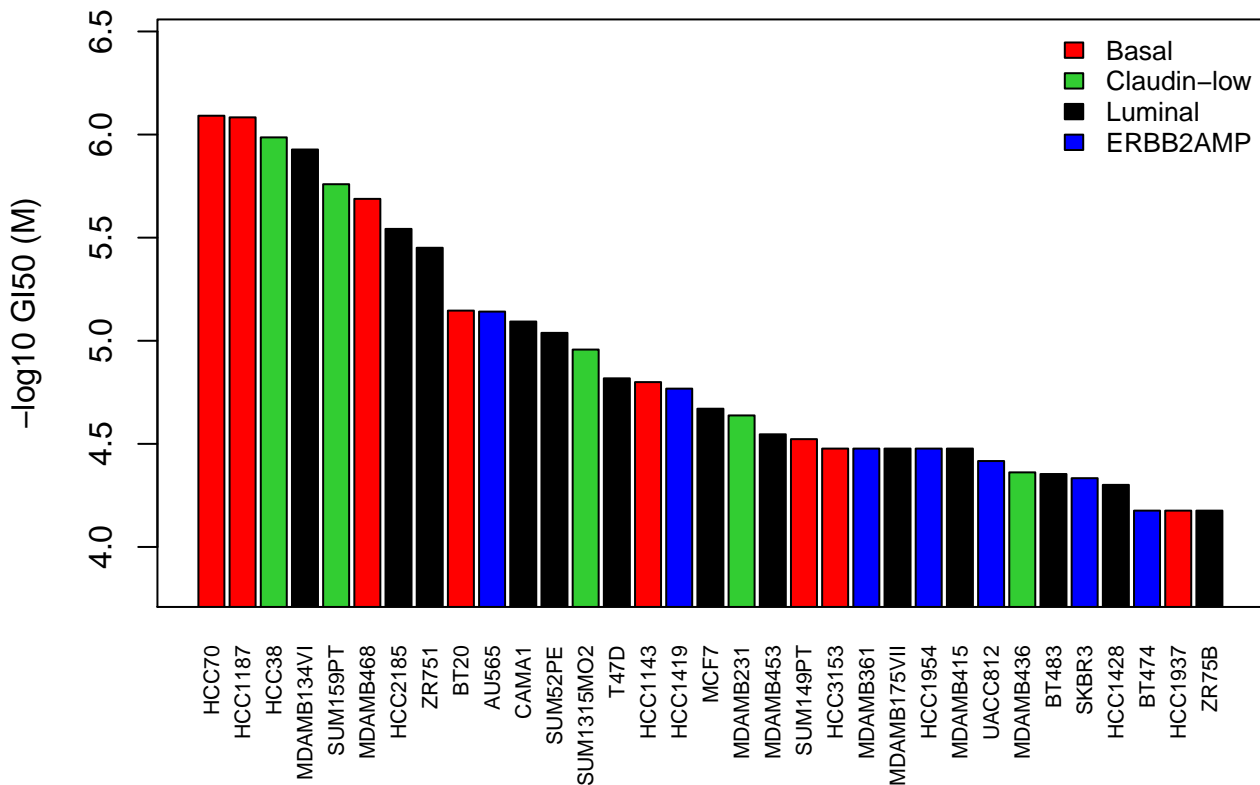
Geldanamycin (Hsp90)



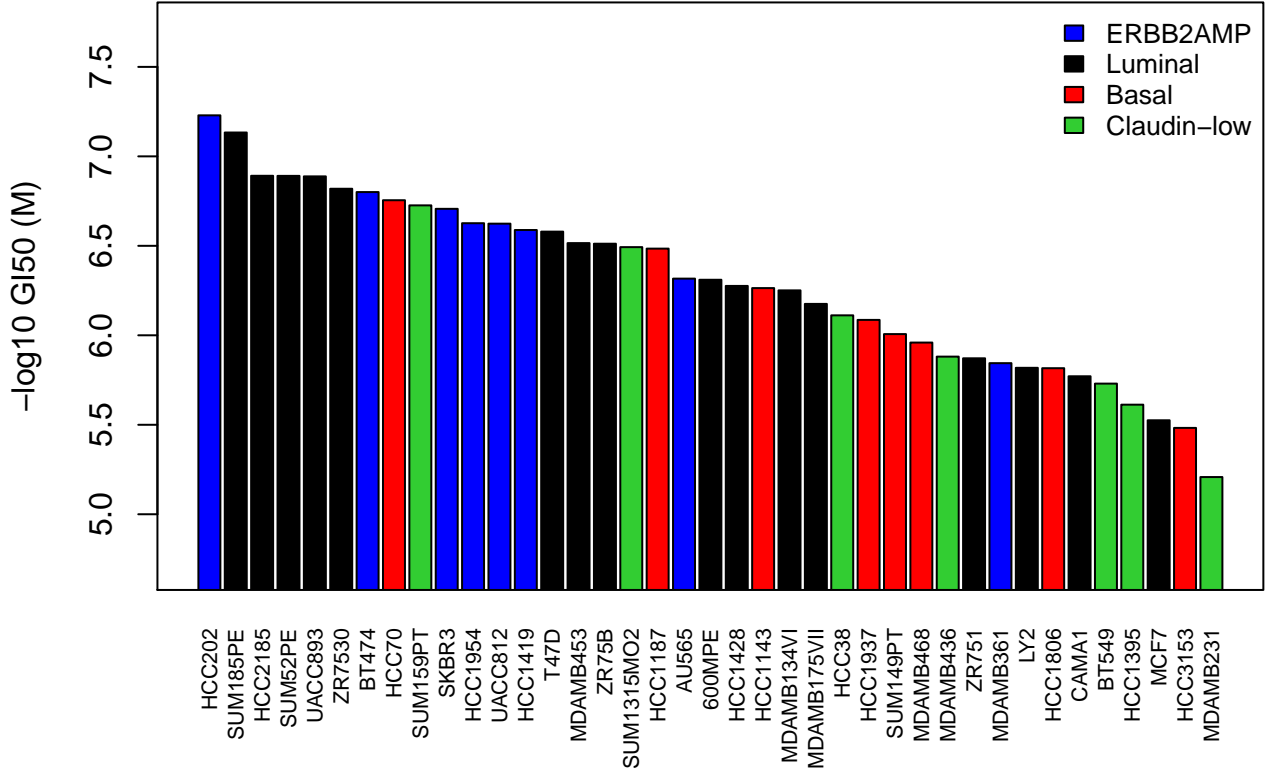
Gemcitabine (pyrimidine animetabolite)



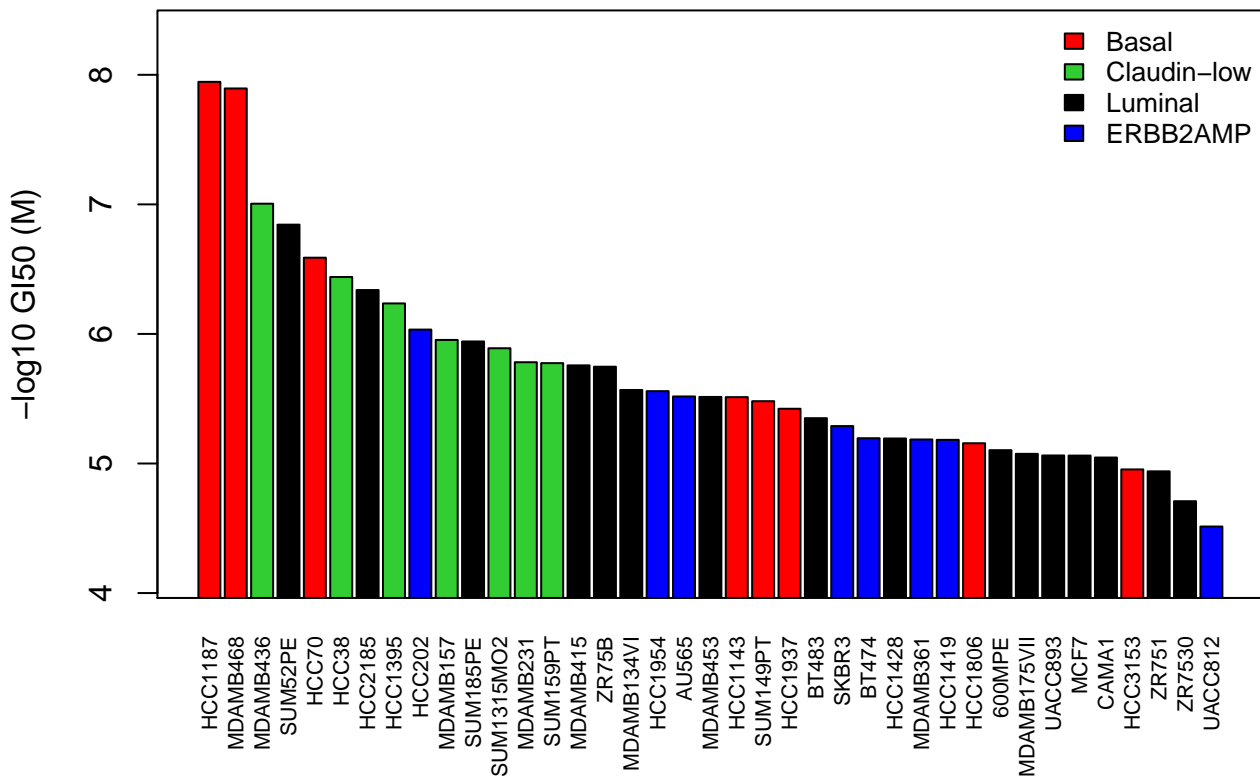
Glycyl-H-1152 (Rho kinase)



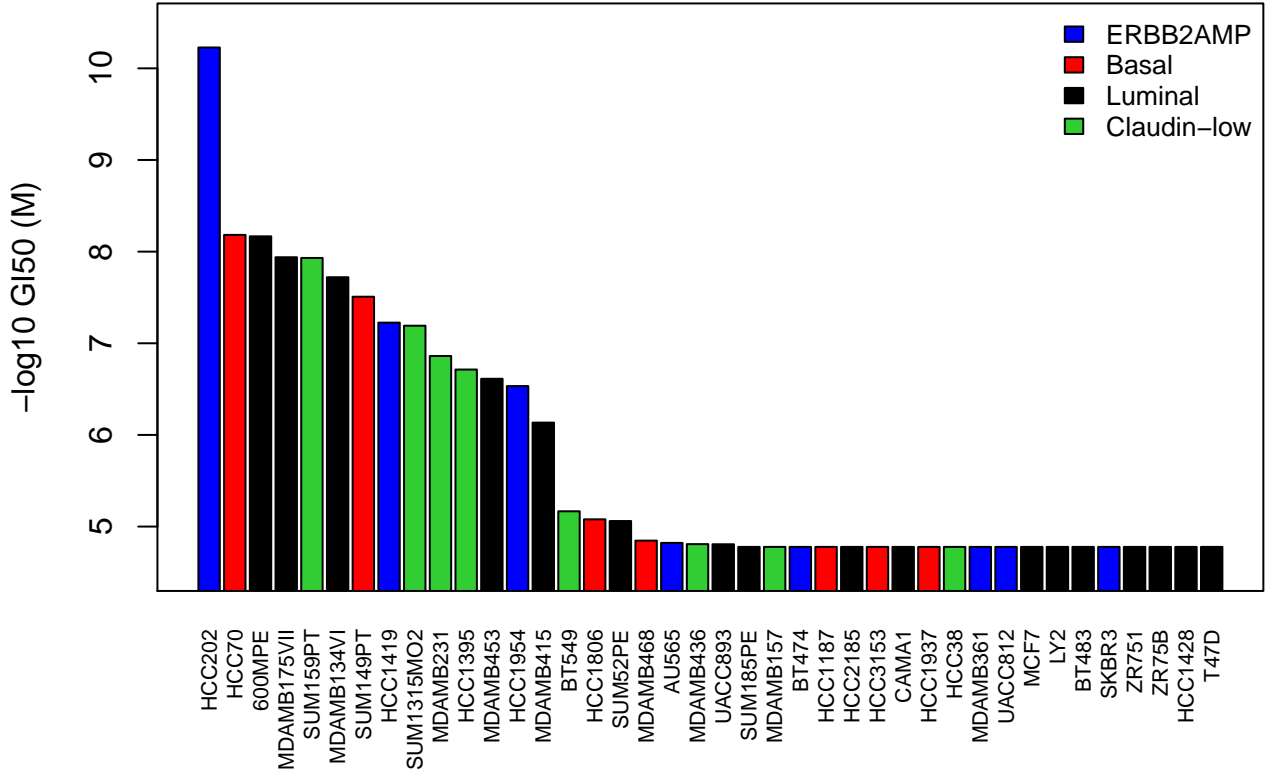
GSK1059615 ()



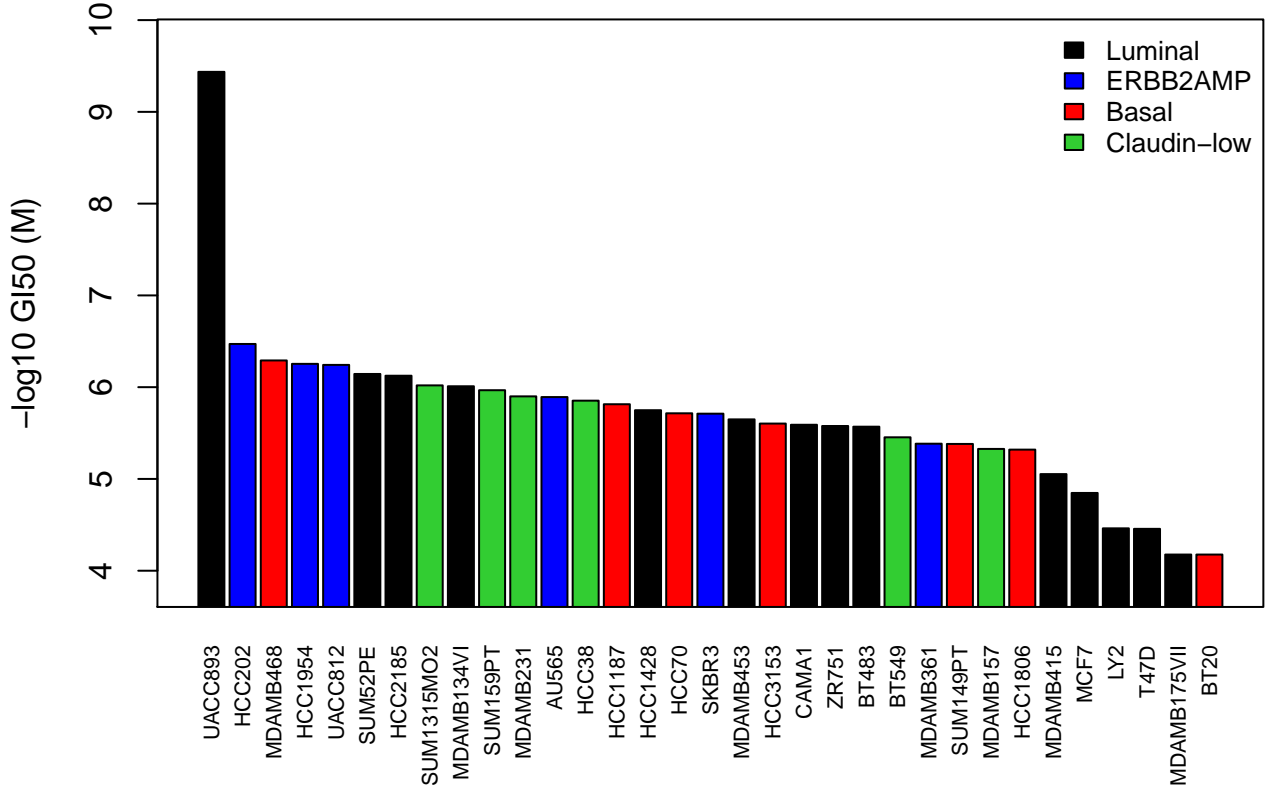
GSK1070916 (aurora kinase B & C)



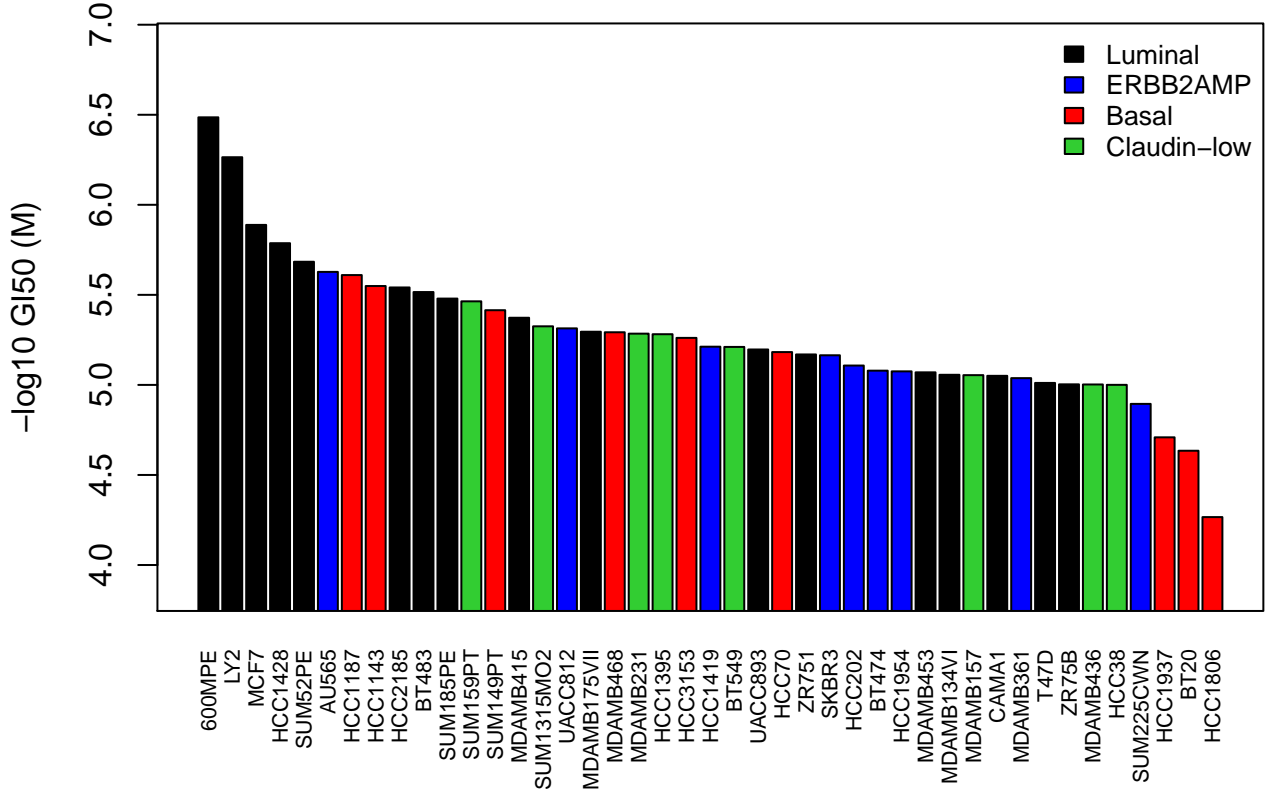
GSK1120212 ()



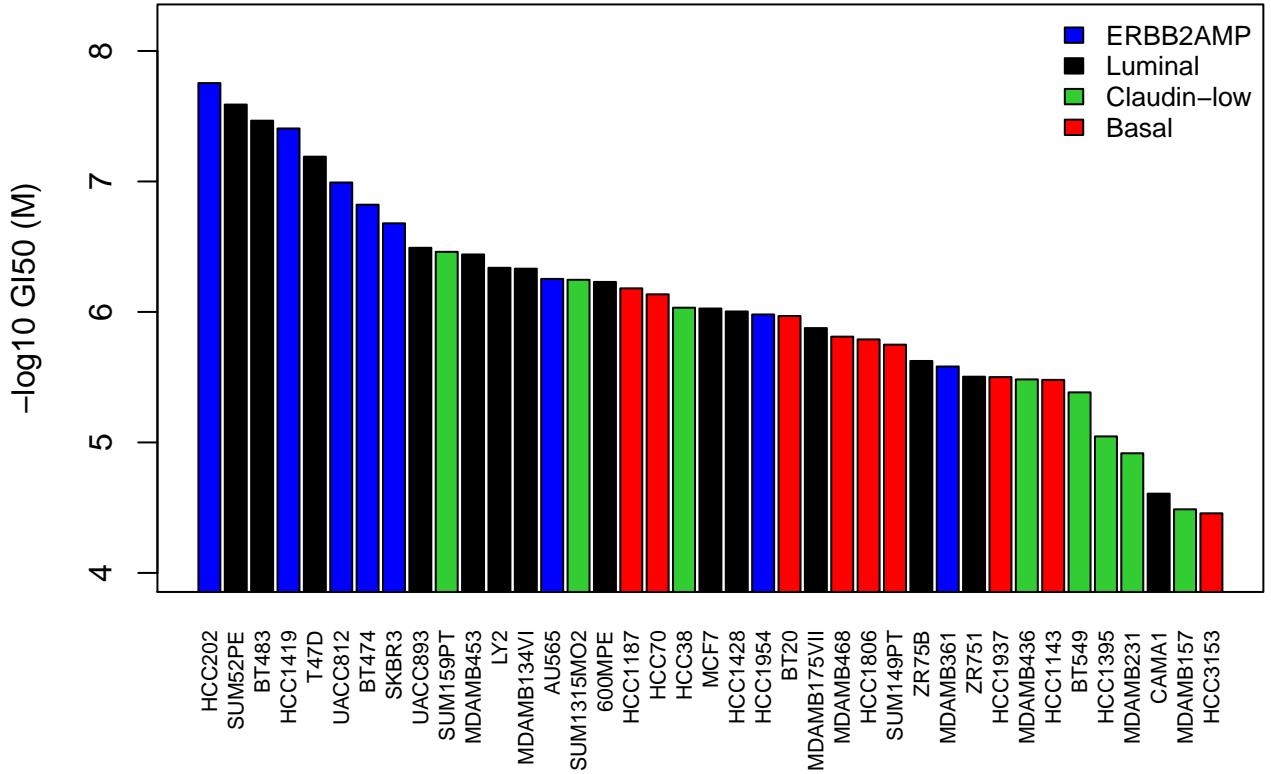
GSK1487371 ()



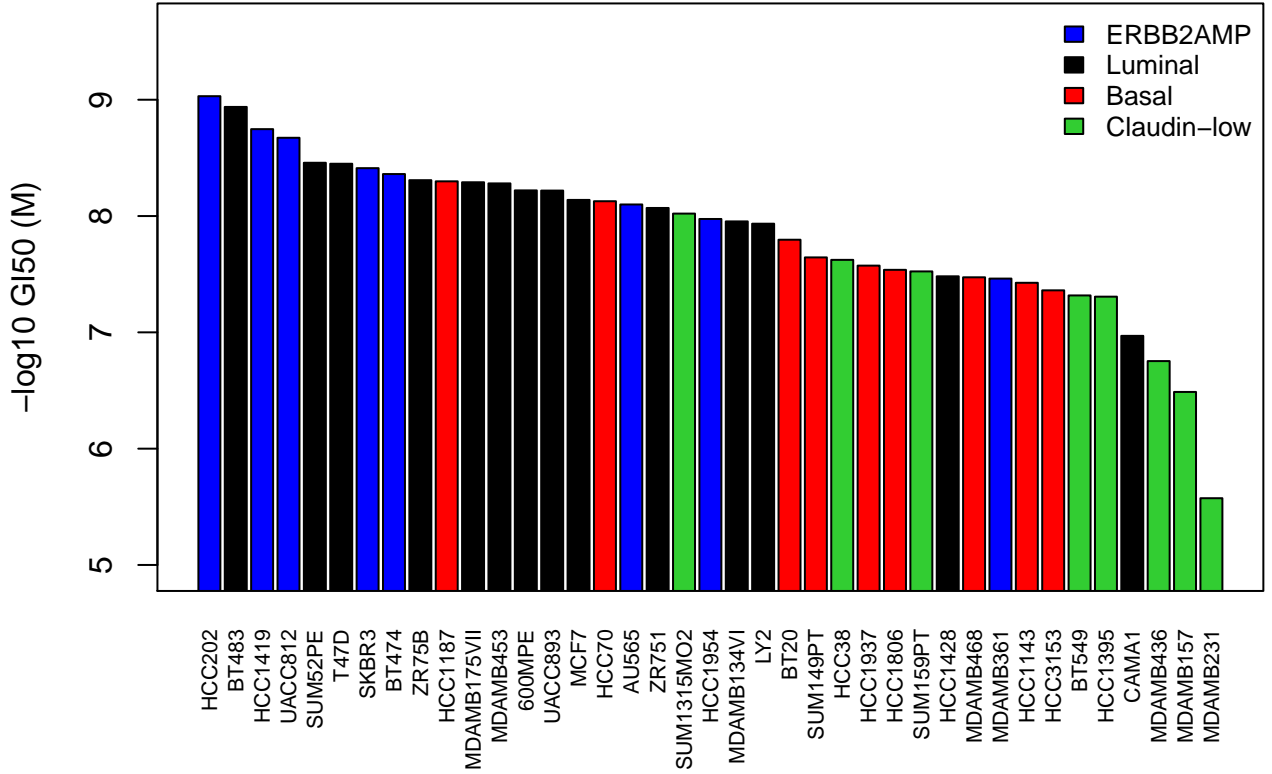
GSK1838705 ()



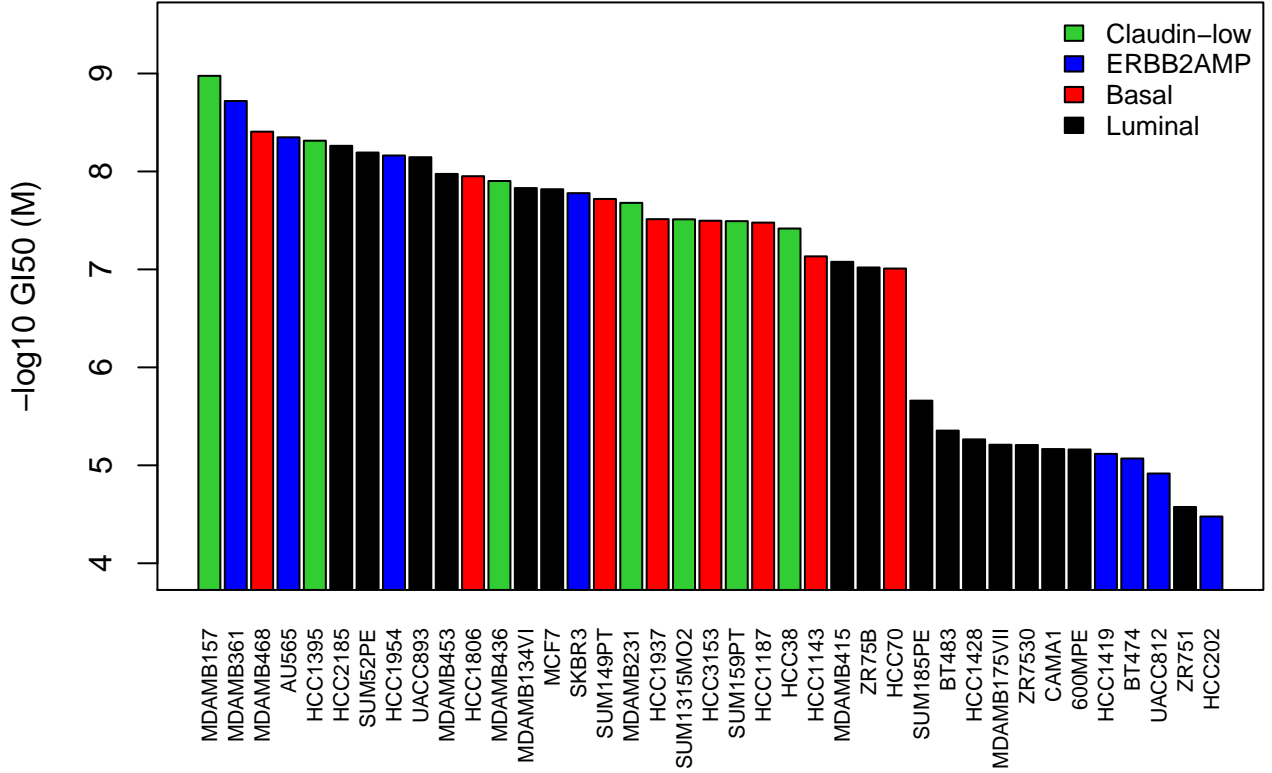
GSK2119563 ()



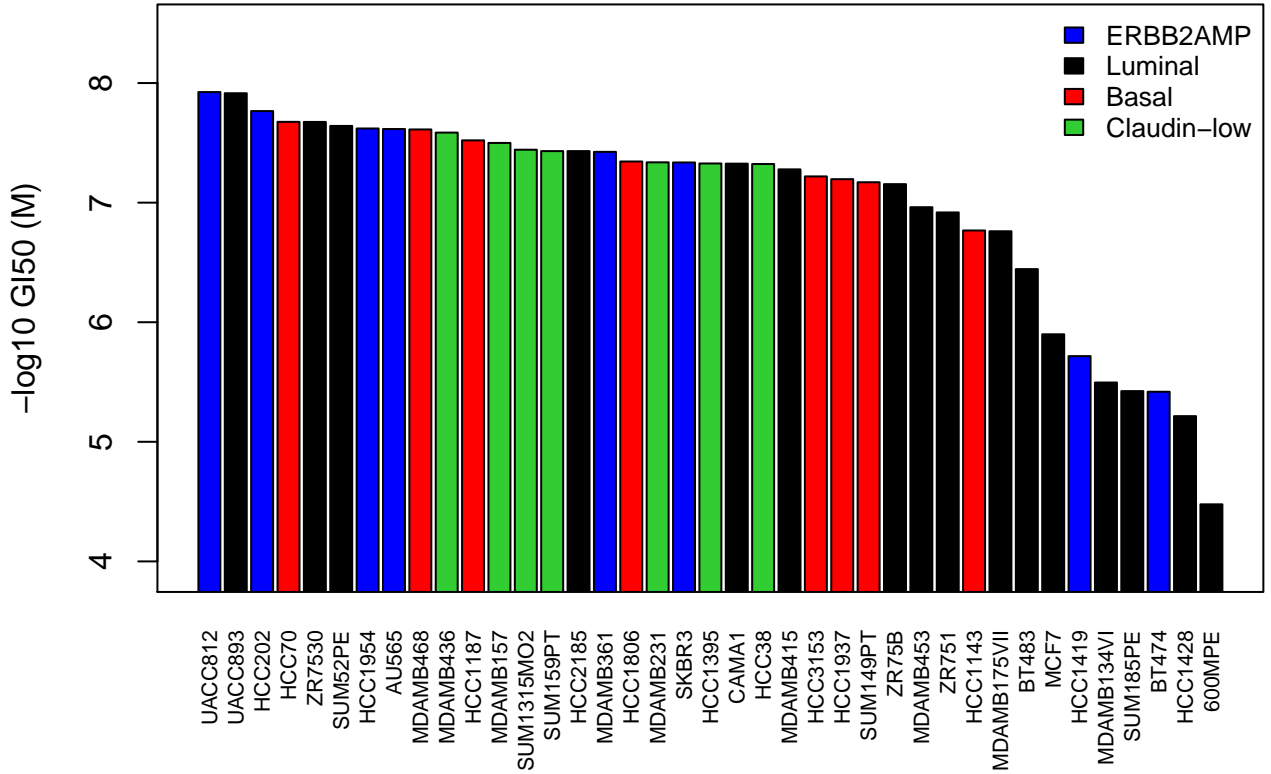
GSK2126458 ()



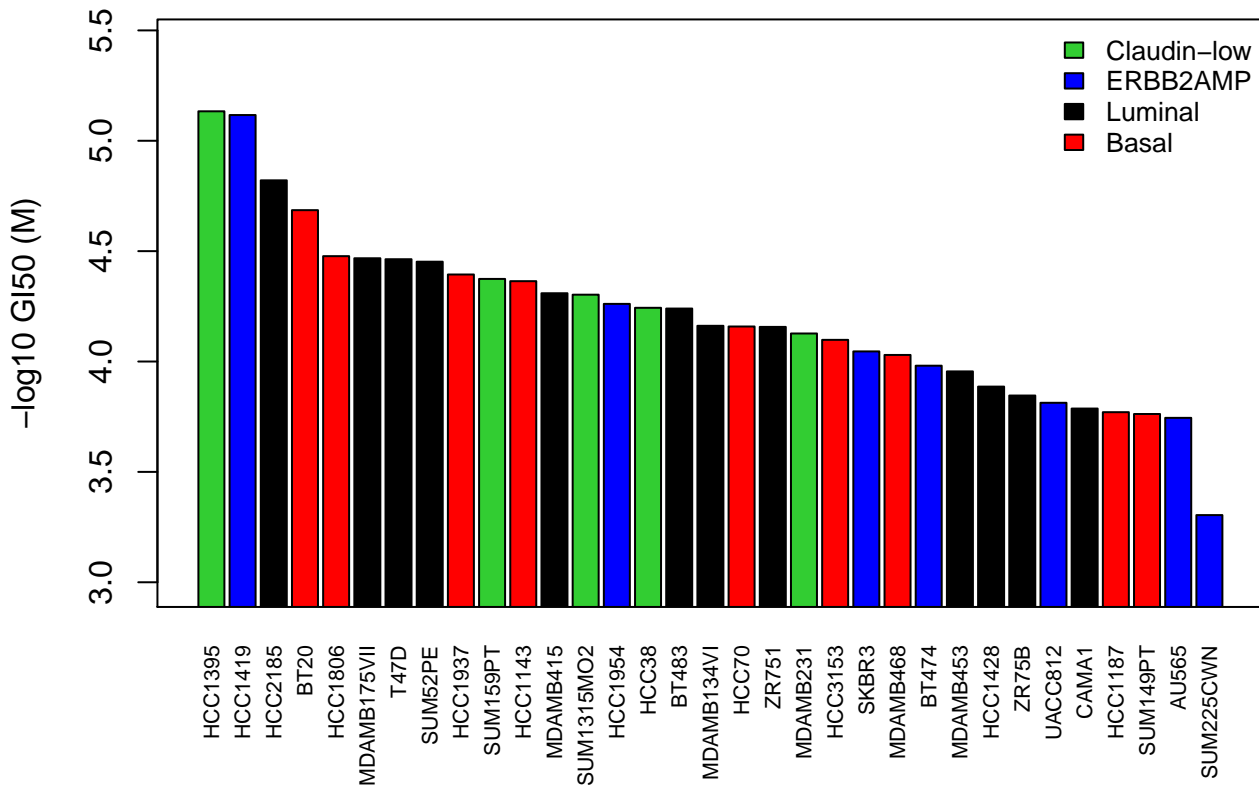
GSK461364 ()



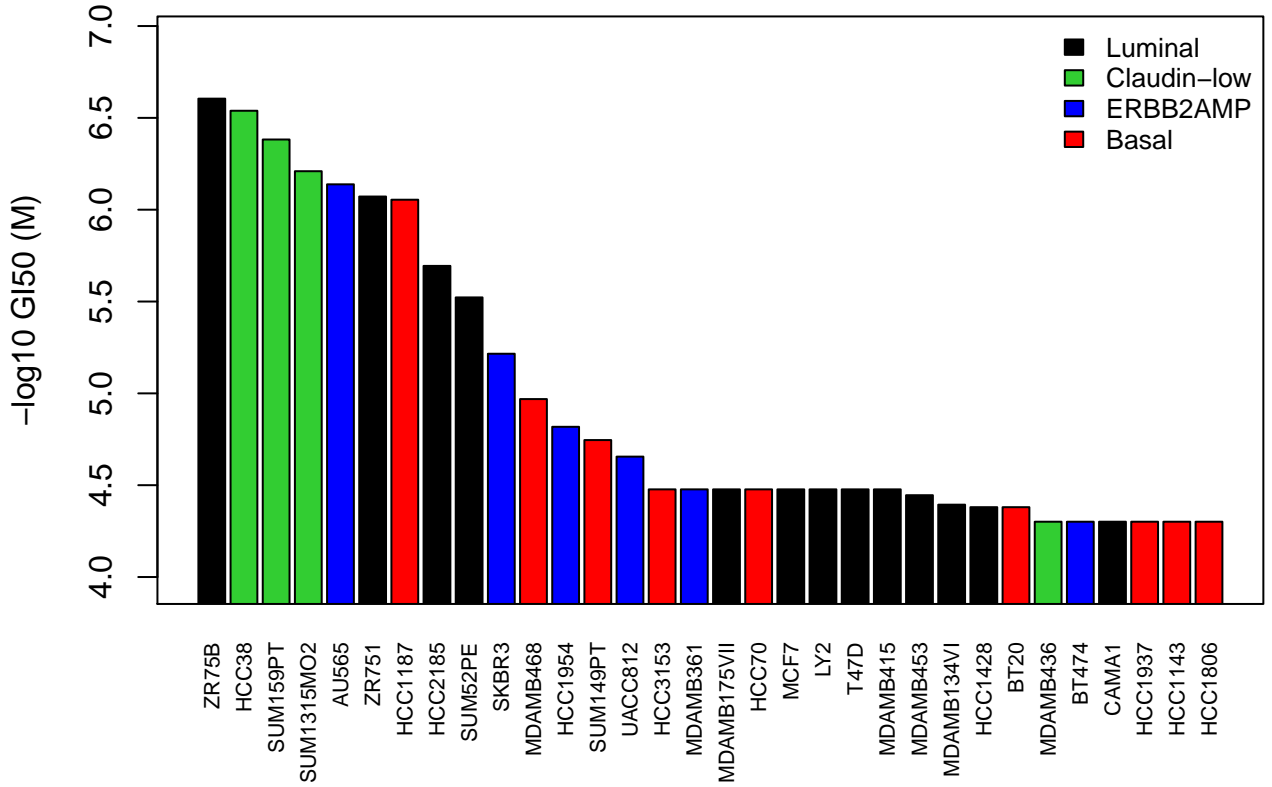
GSK923295 (CENPE)



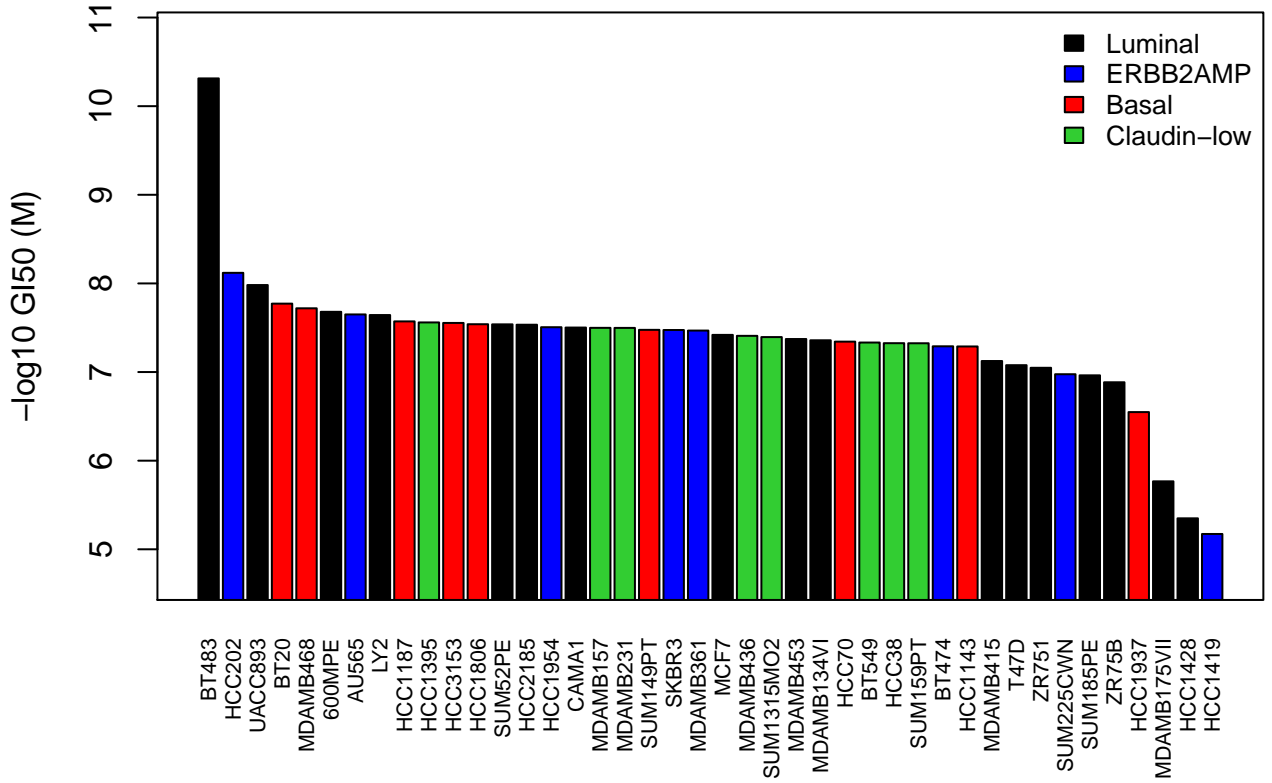
Ibandronate sodium salt (farnesyl diphosphate synthase)



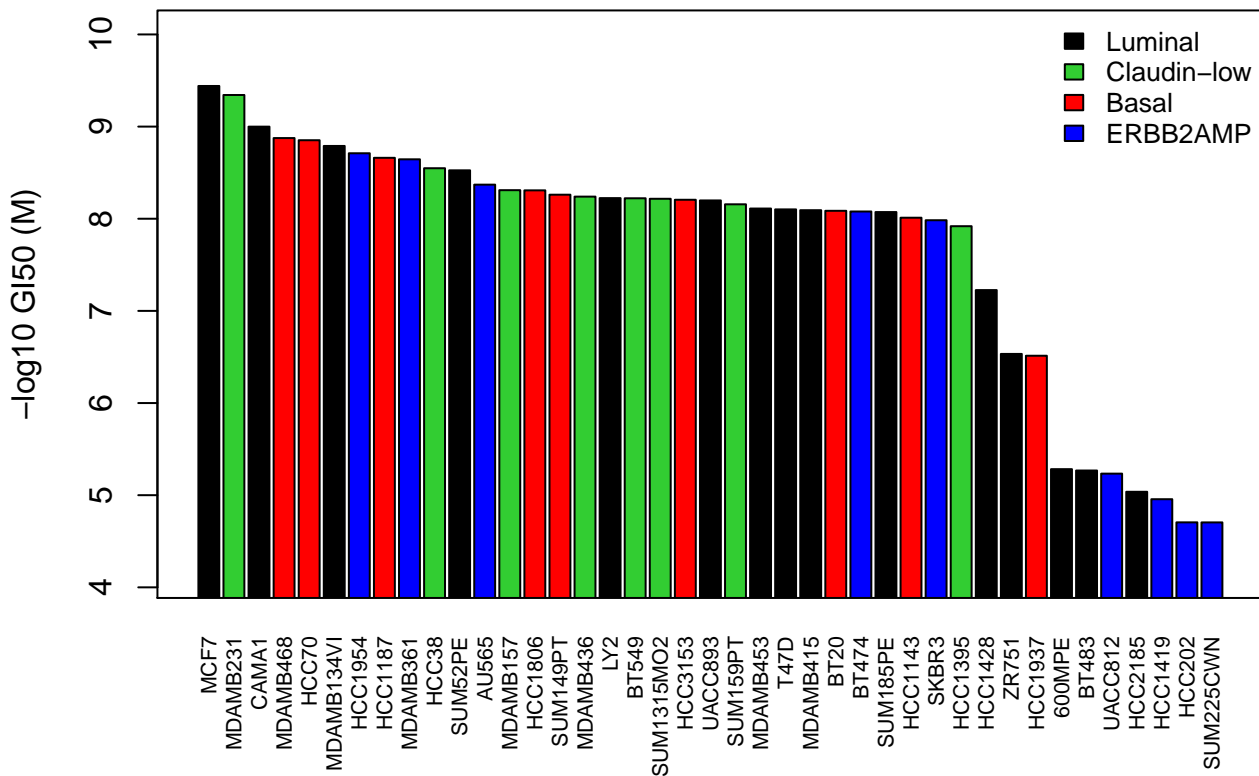
ICRF-193 (PLK1, topo II)



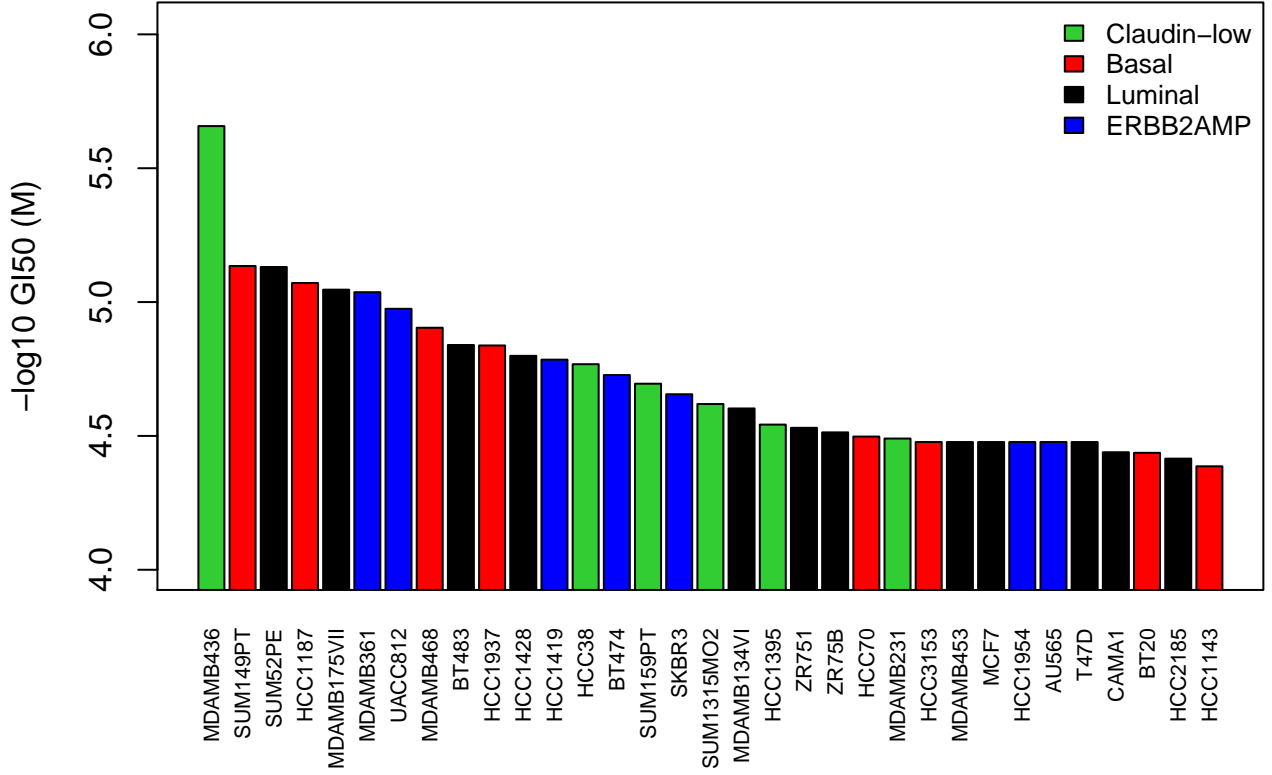
Ispinesib (Kinesin)



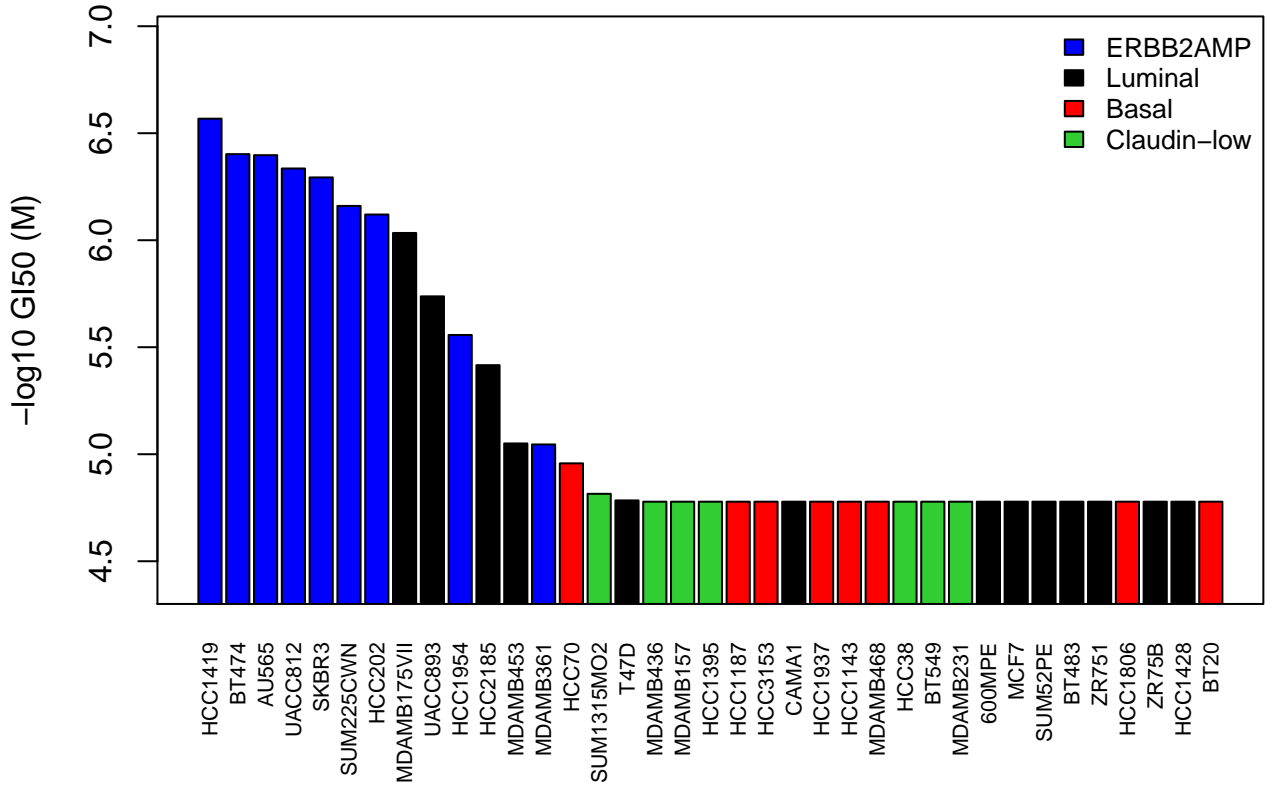
Ixabepilone (Microtubule)



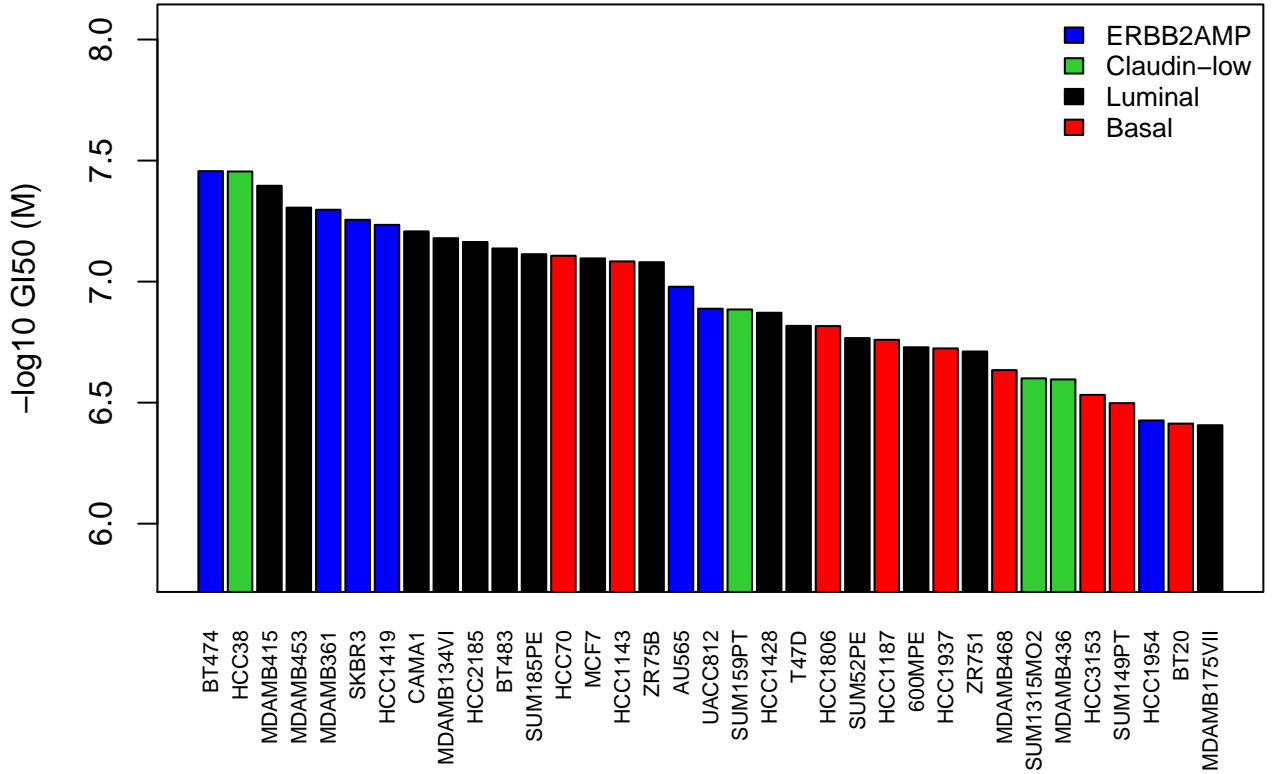
L-779450 (B-raf)



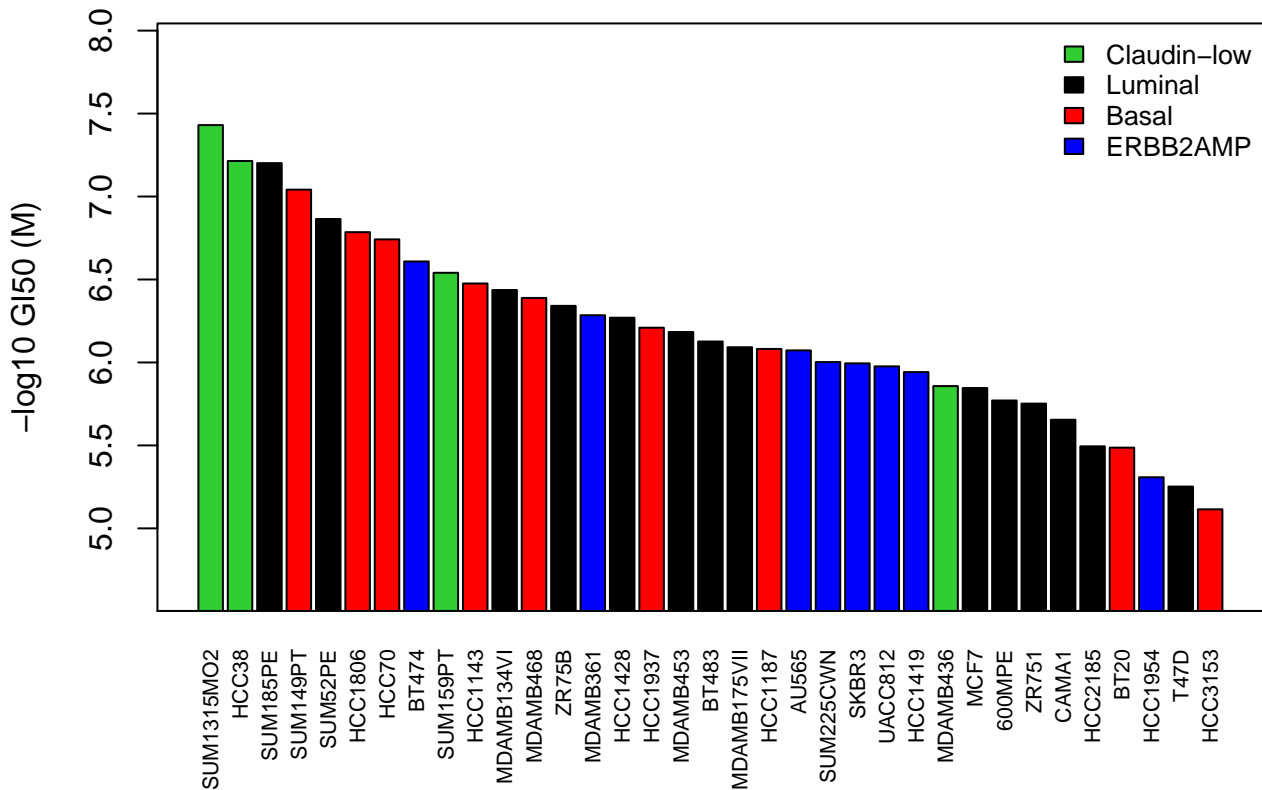
Lapatinib (ERBB2, EGFR)



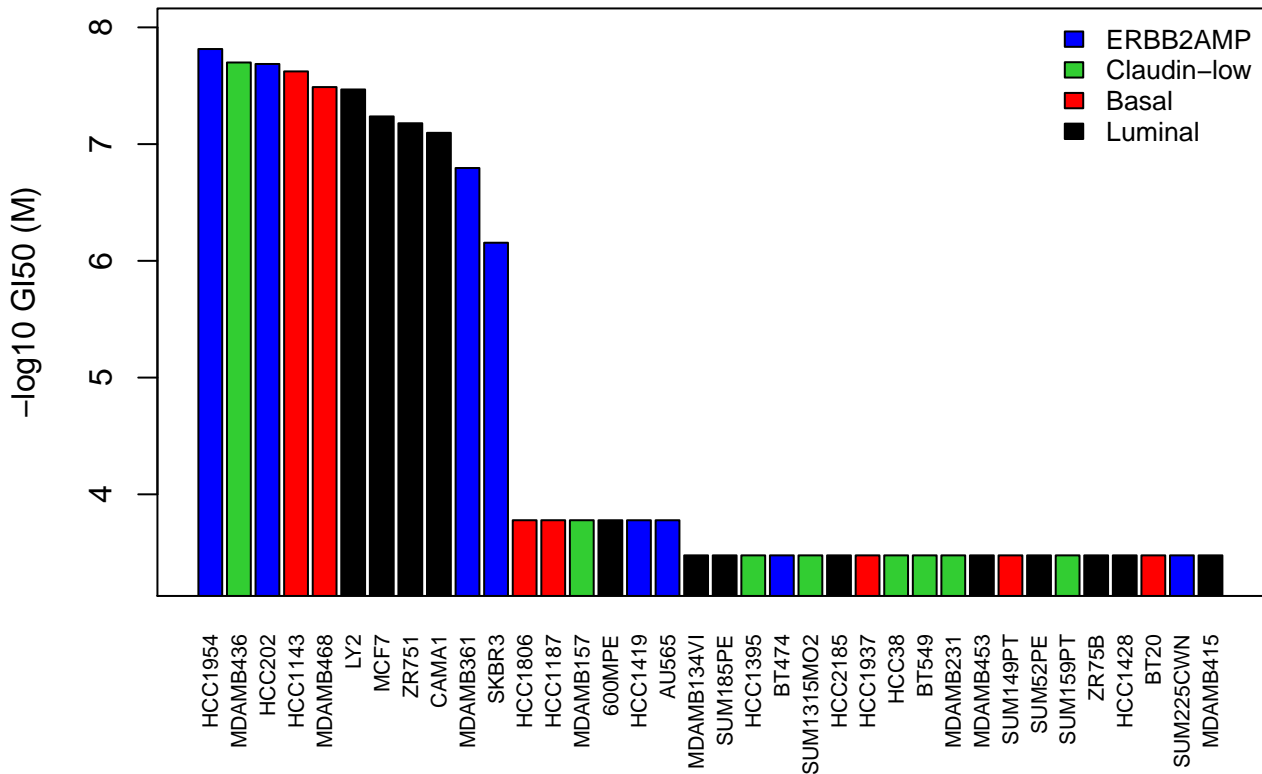
LBH589 (HDAC, pan inibitor)



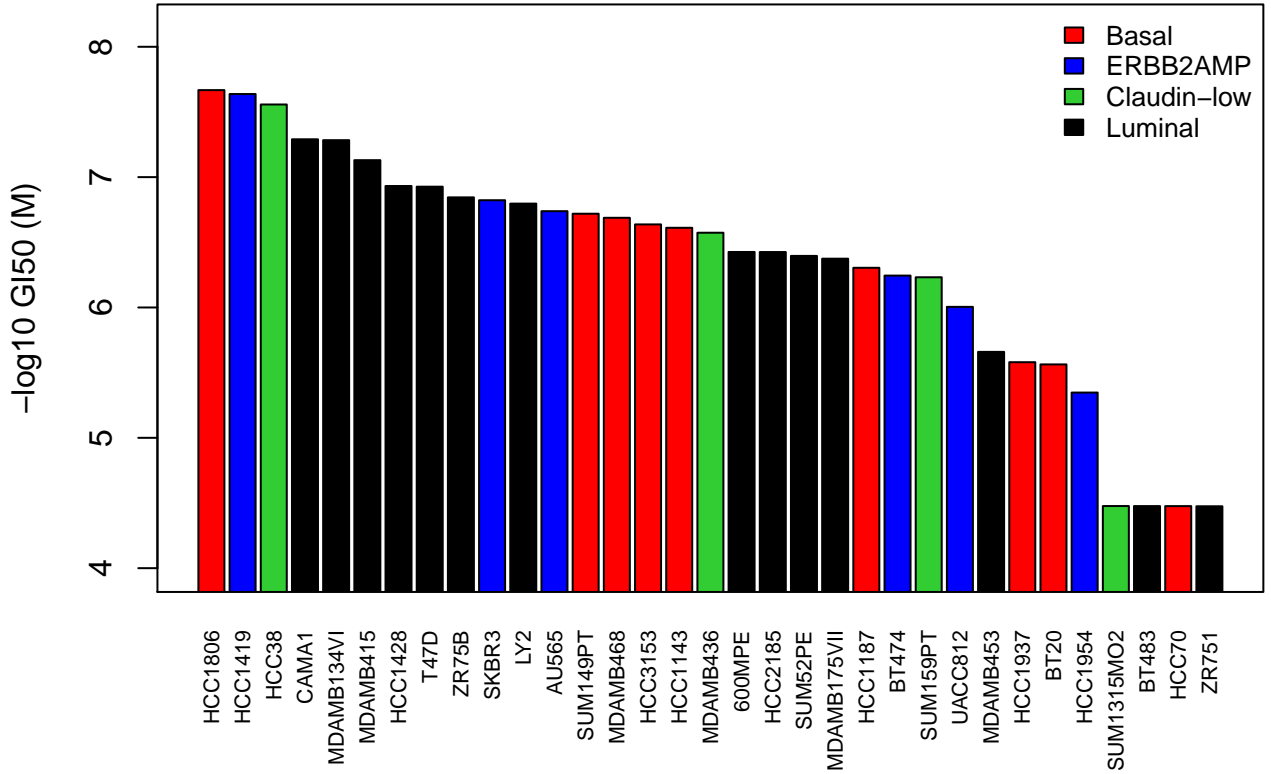
Lestaurtinib (FLT-3, TrkA)



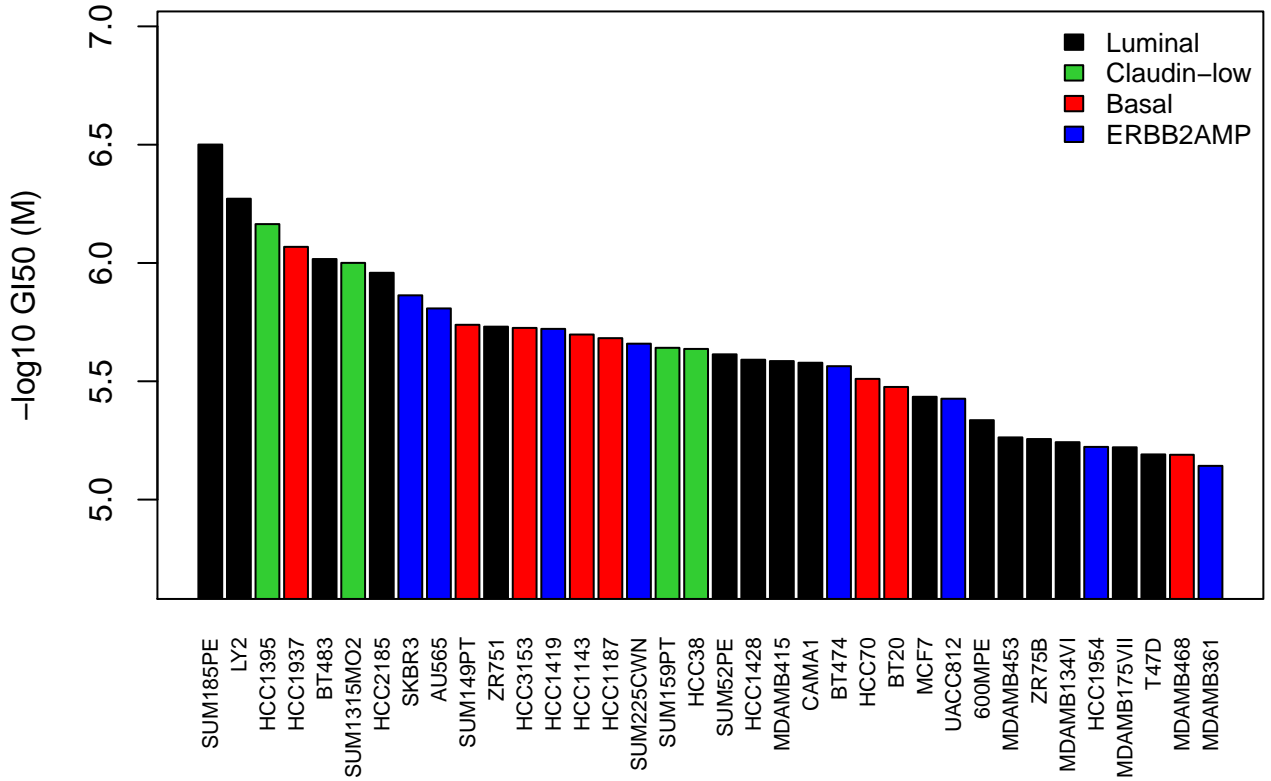
Methotrexate (DHFR)



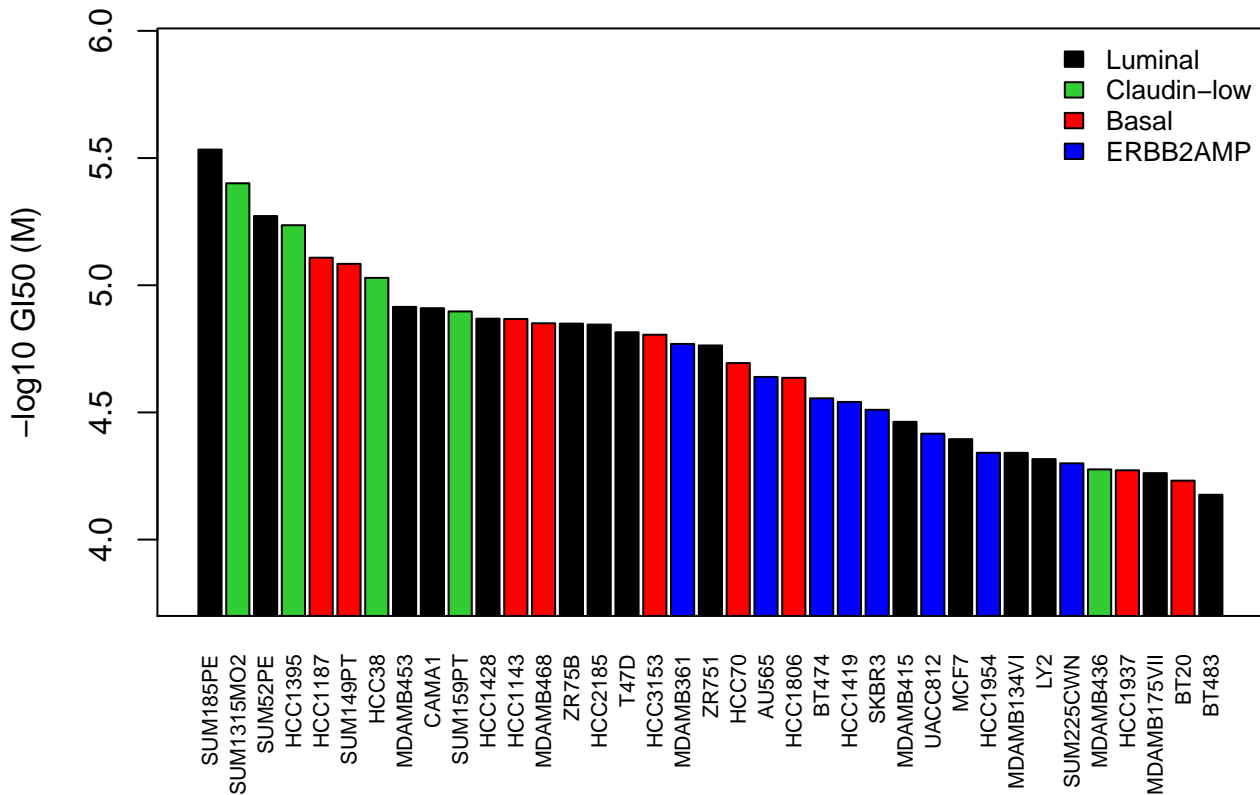
MLN4924 (NAE)



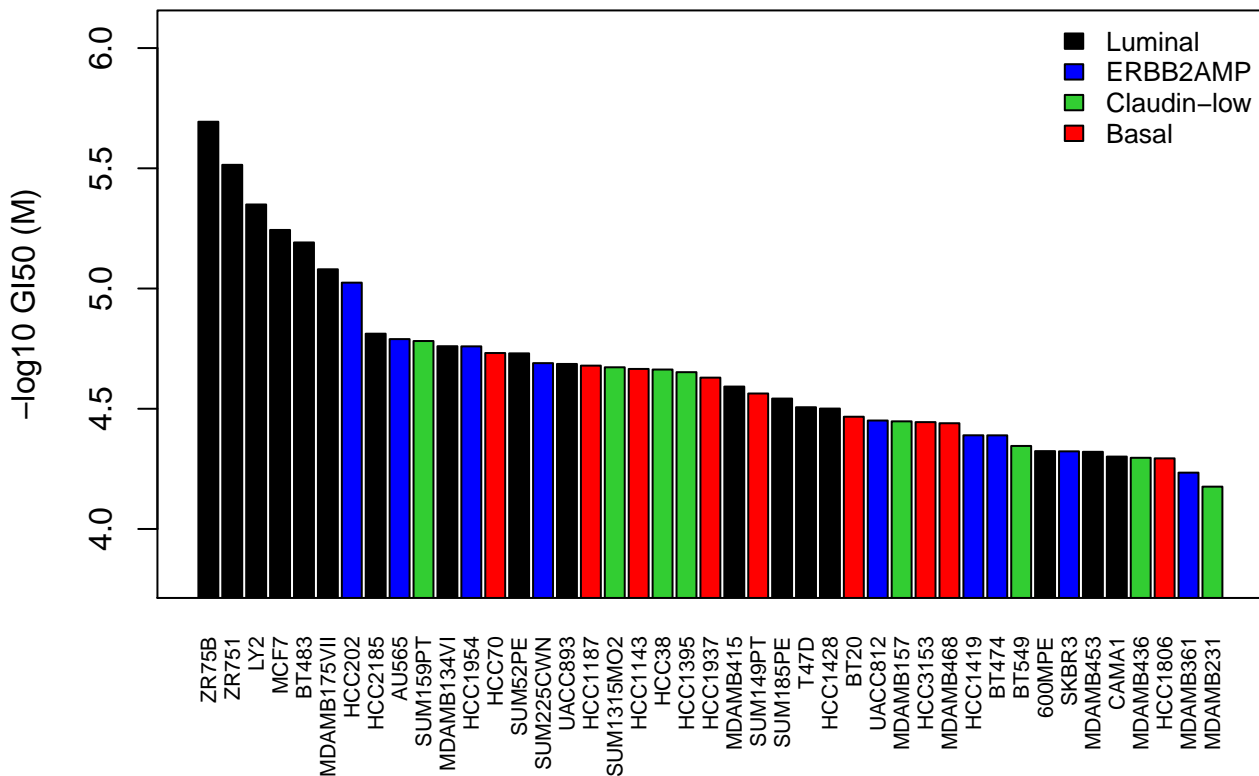
NSC 663284 (cdc25s)



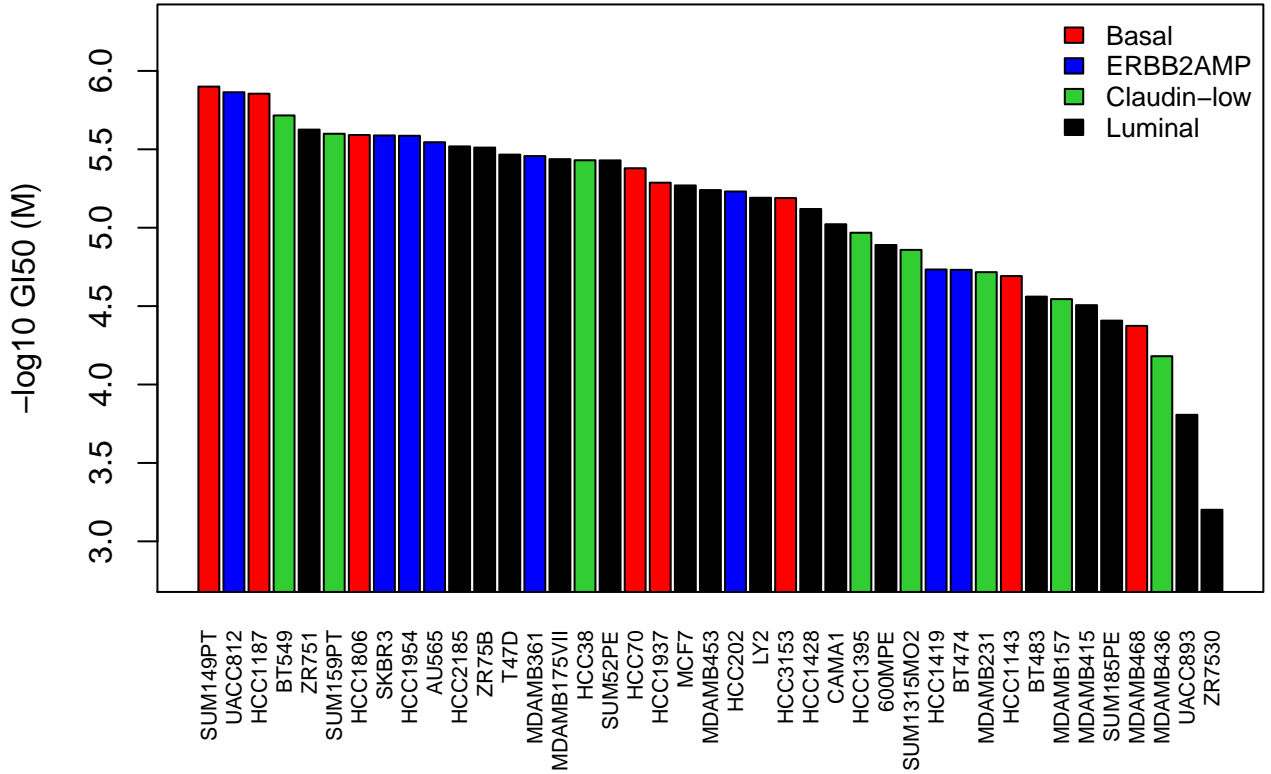
NU6102 (CDK1/CCNB)



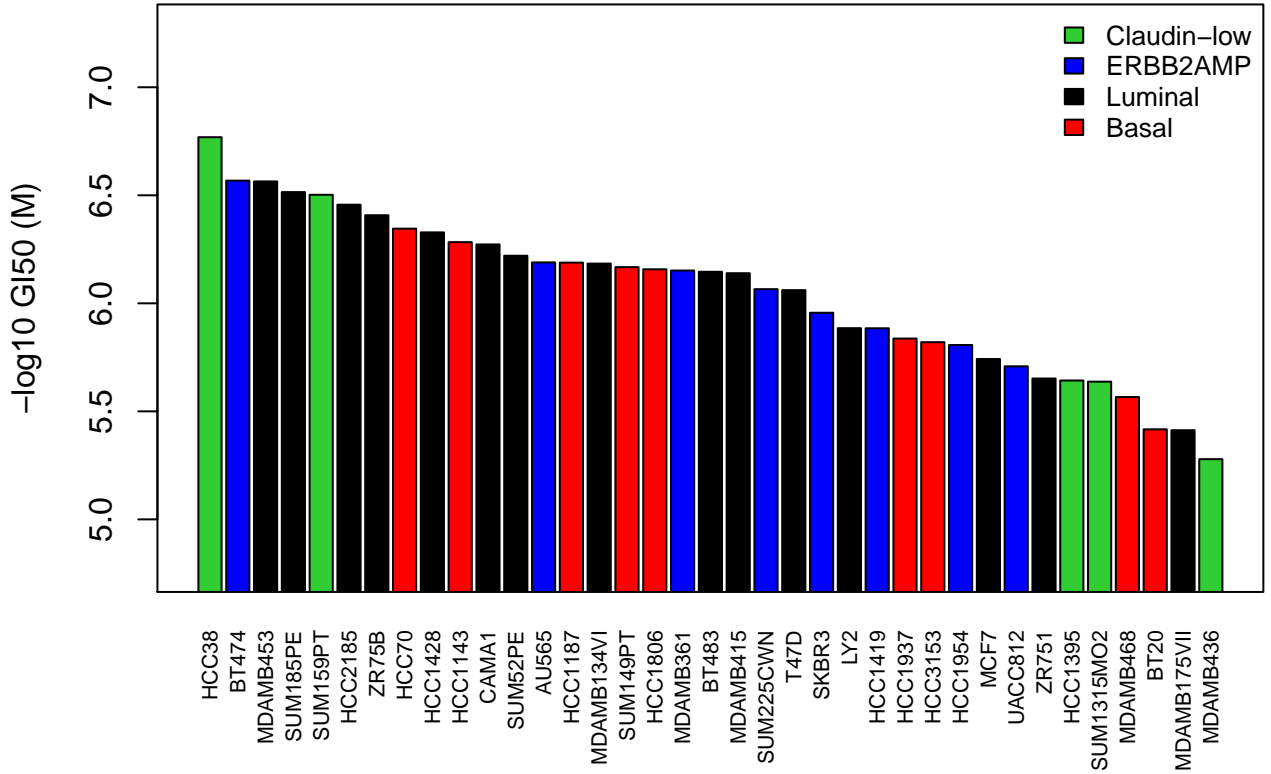
Nutlin 3a (MDM2)



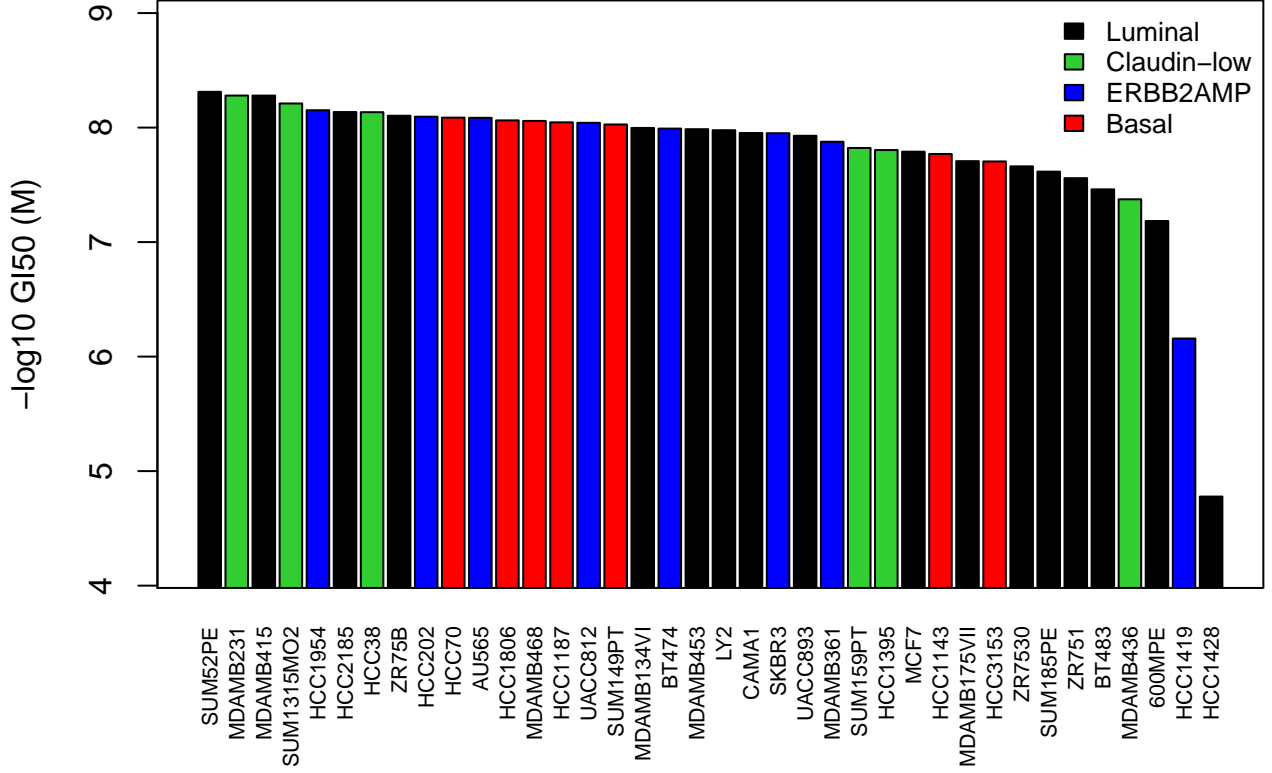
Oxaliplatin (DNA cross-linker)



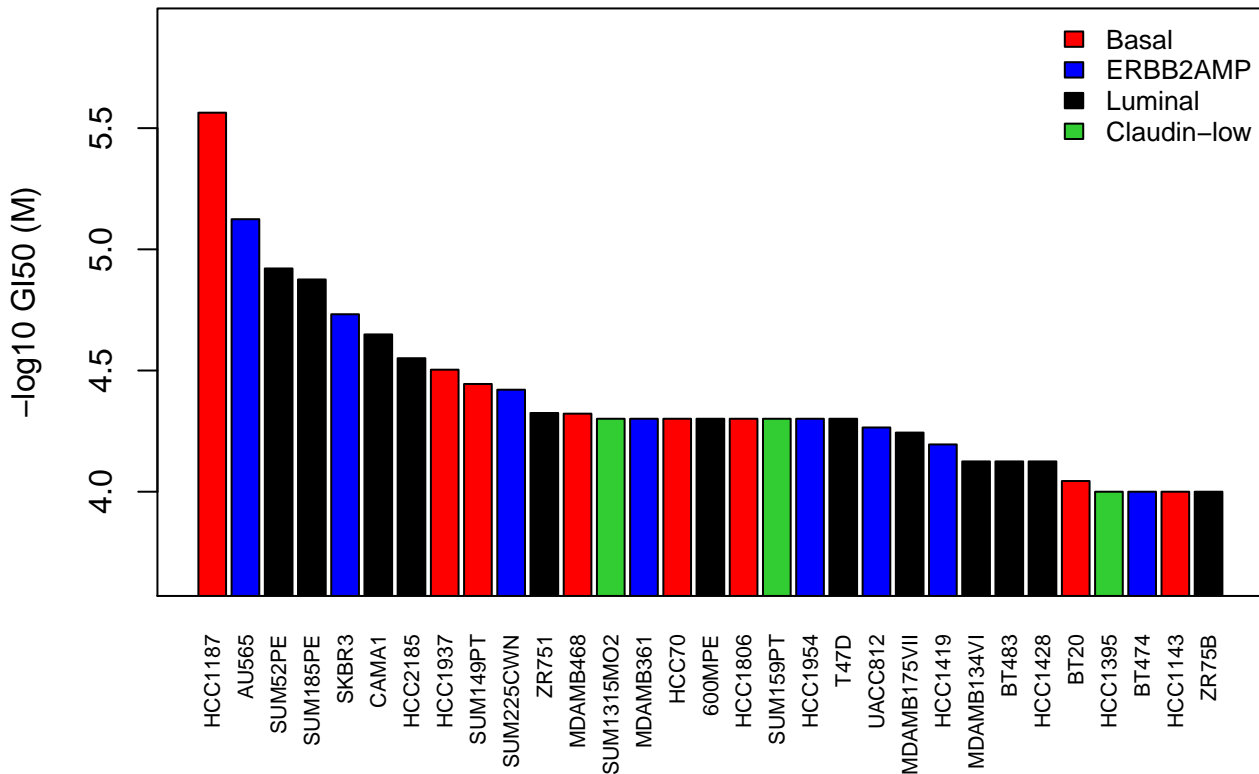
Oxamflatin (HDAC)



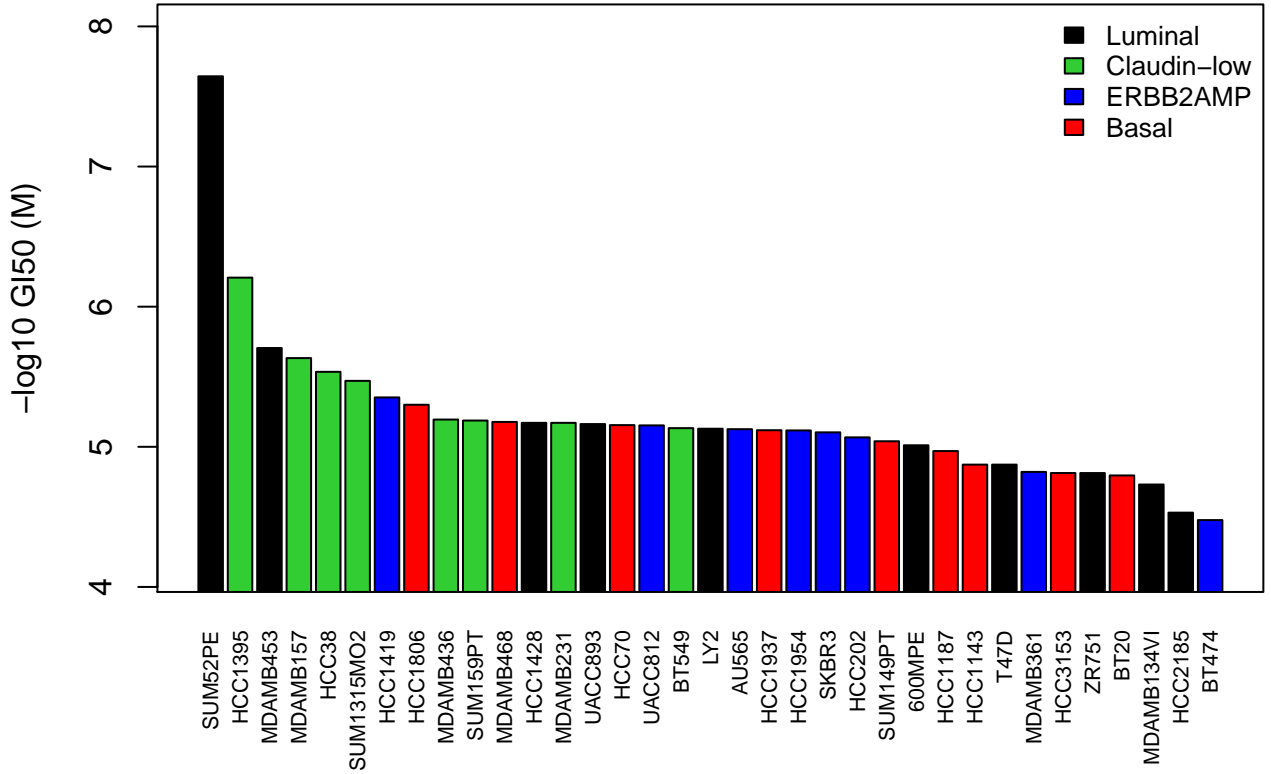
Paclitaxel (Microtubule)



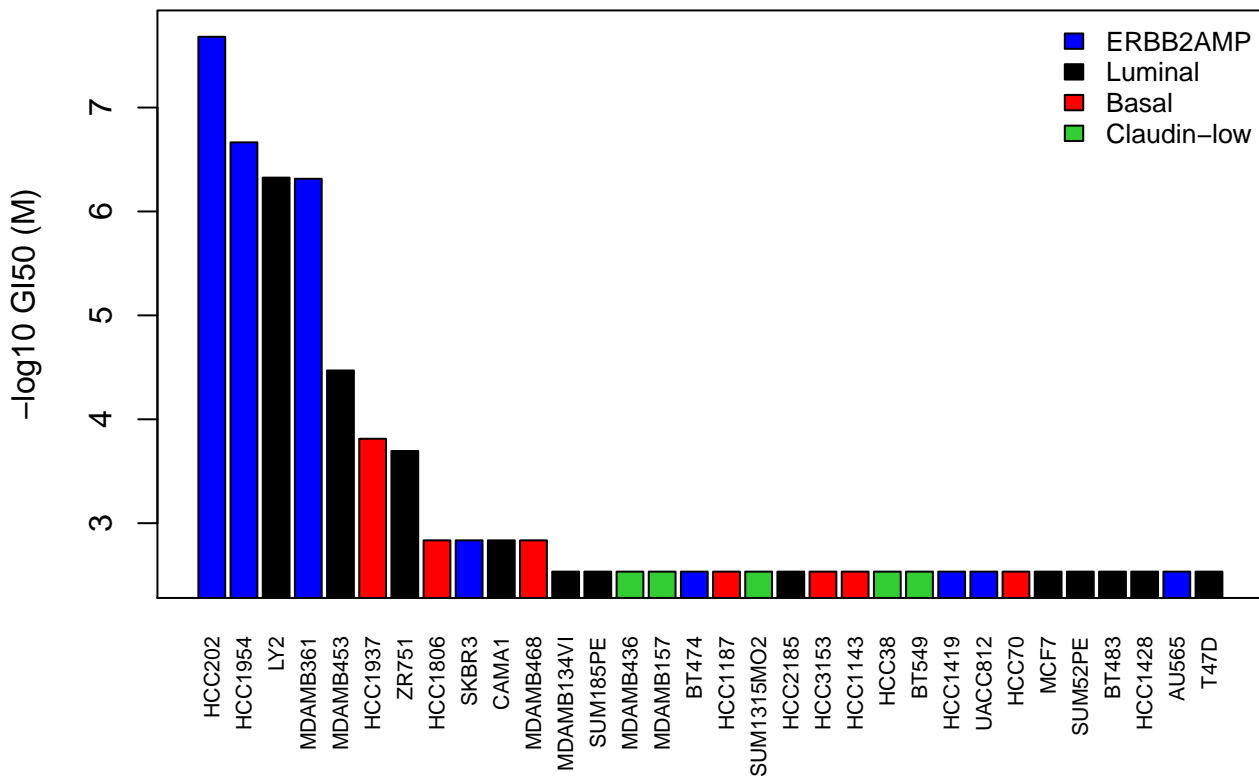
PD 98059 (MEK)



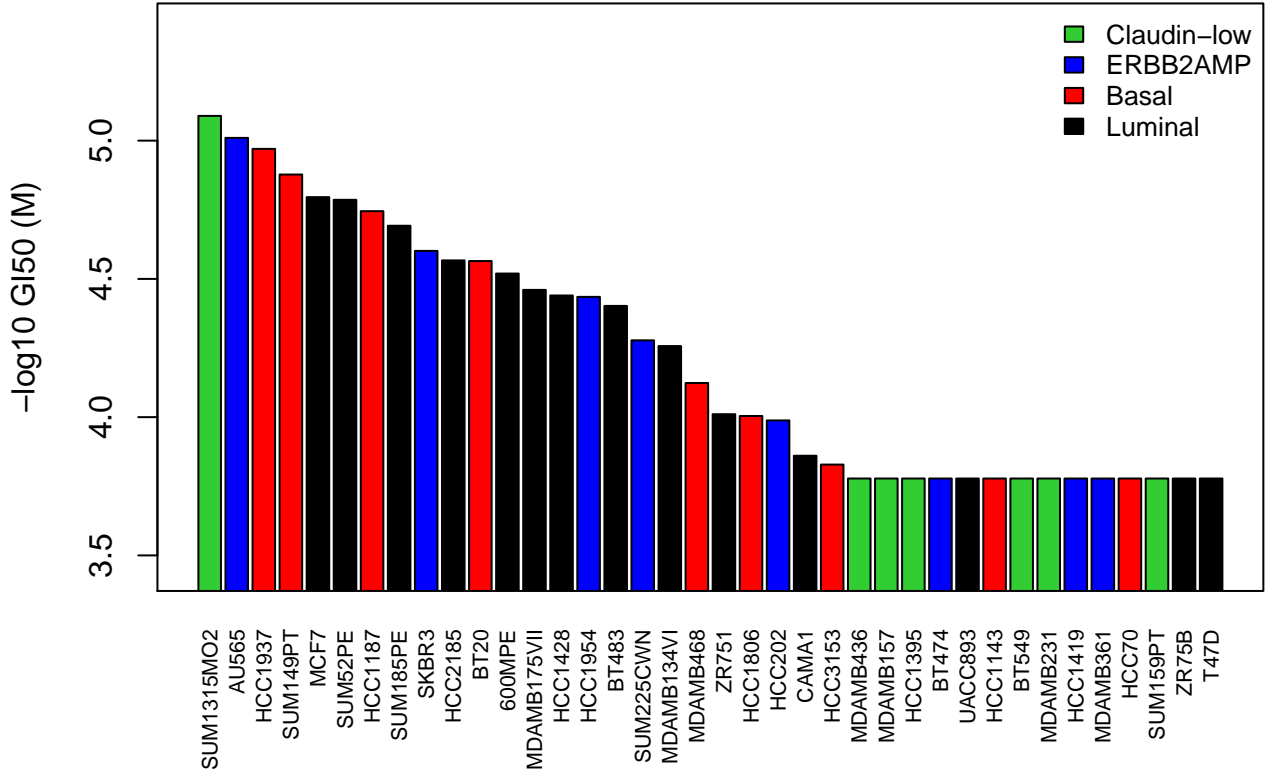
PD173074 (FGFR3)



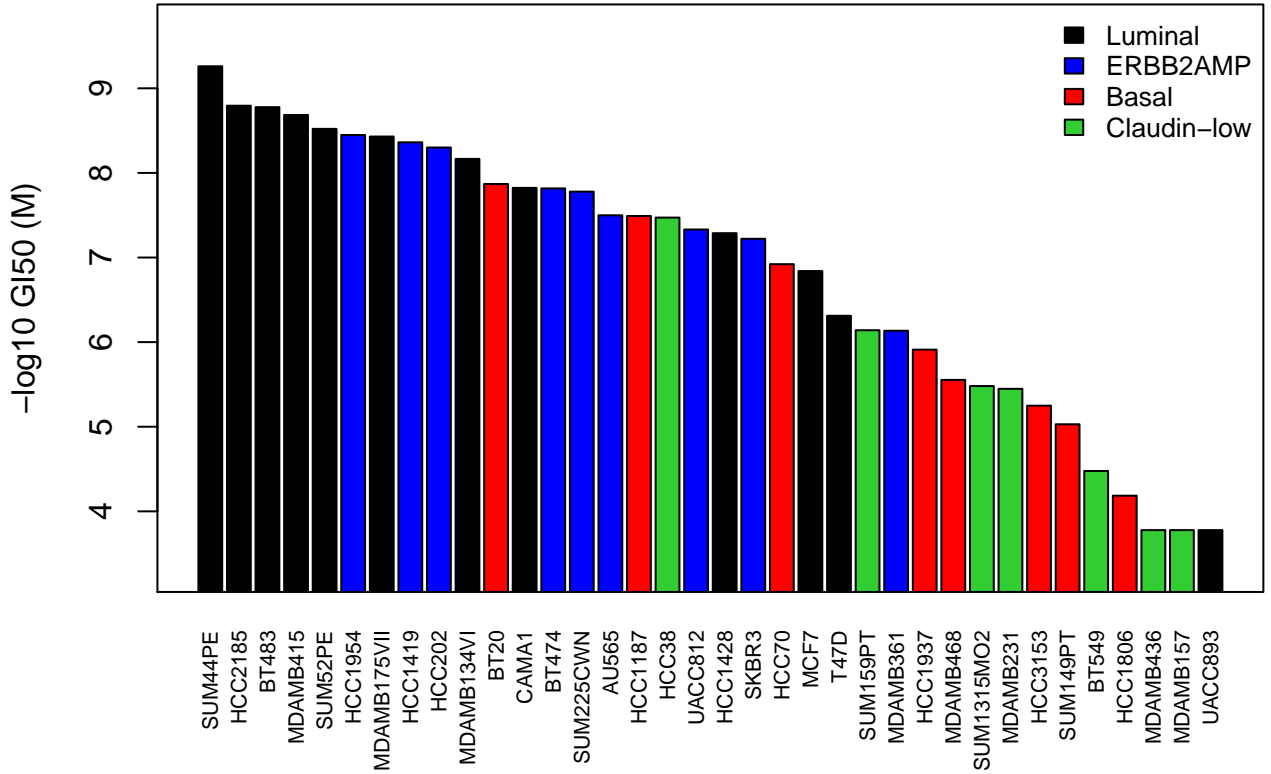
Pemetrexed (DNA synthesis/repair)



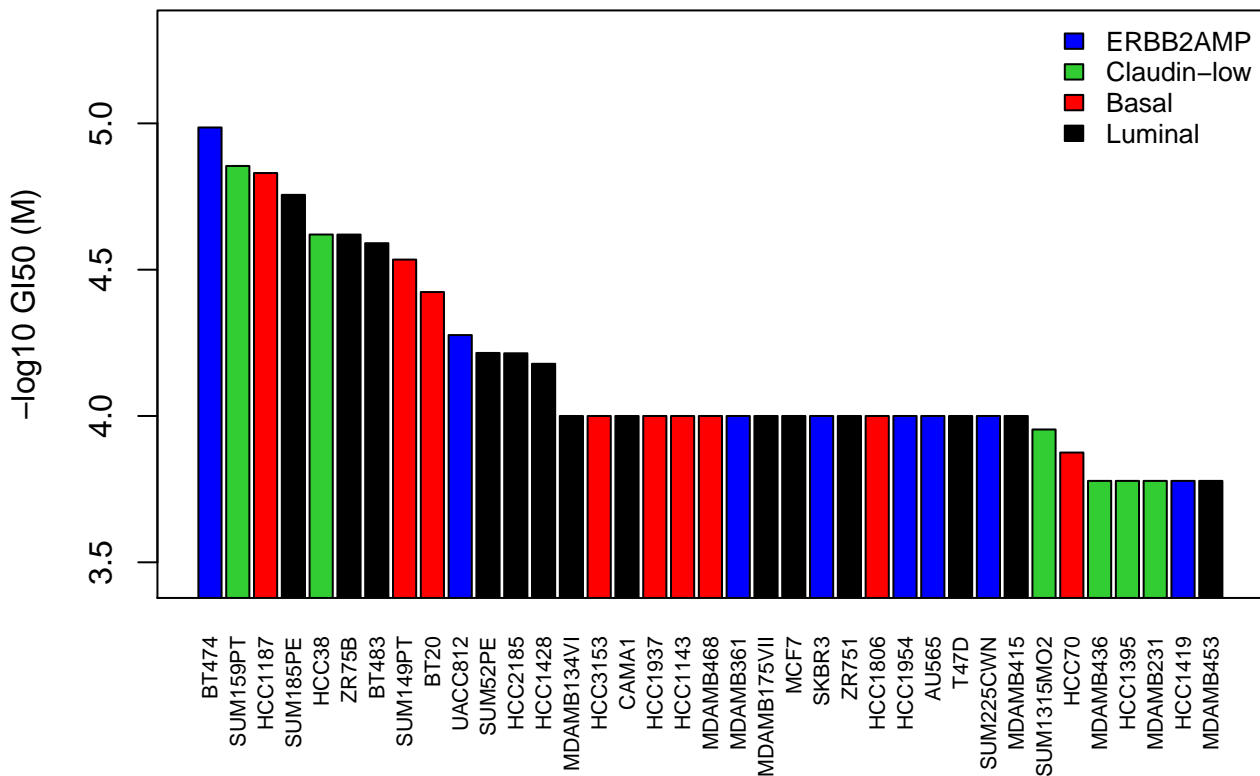
Purvalanol A (CDK1)



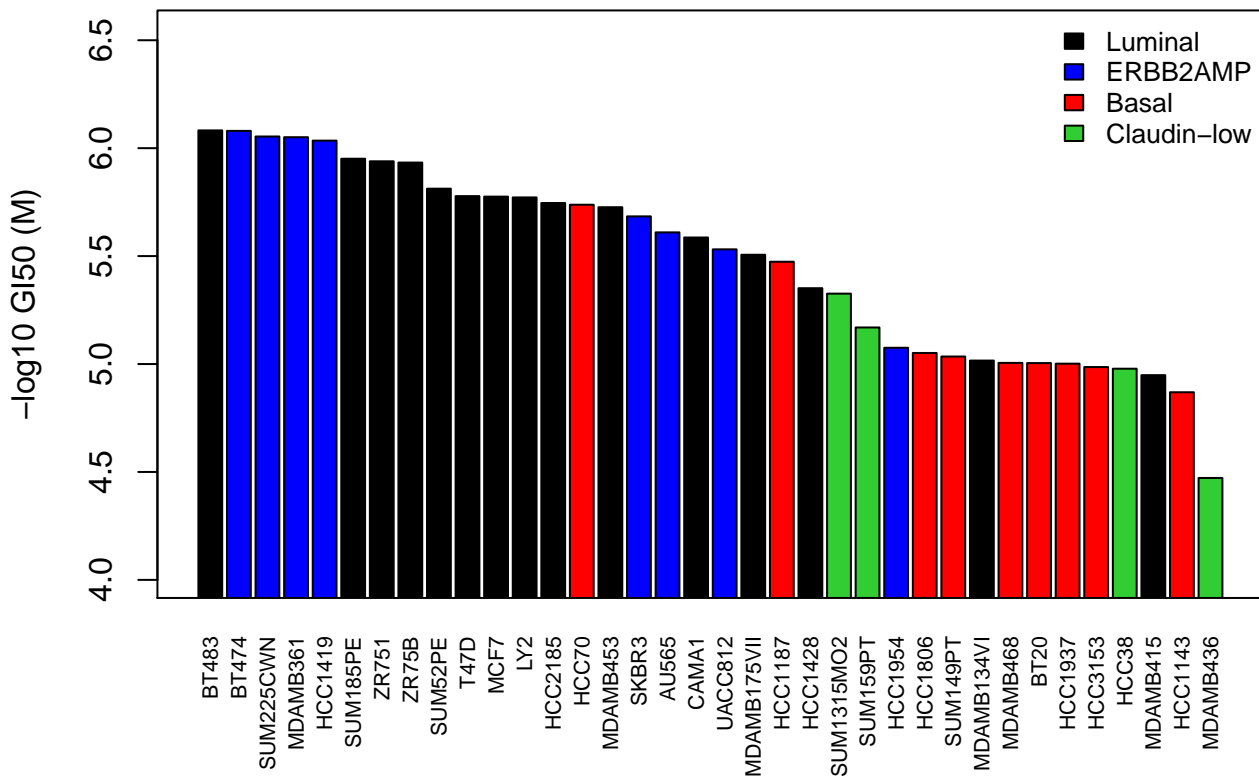
Rapamycin (mTOR)



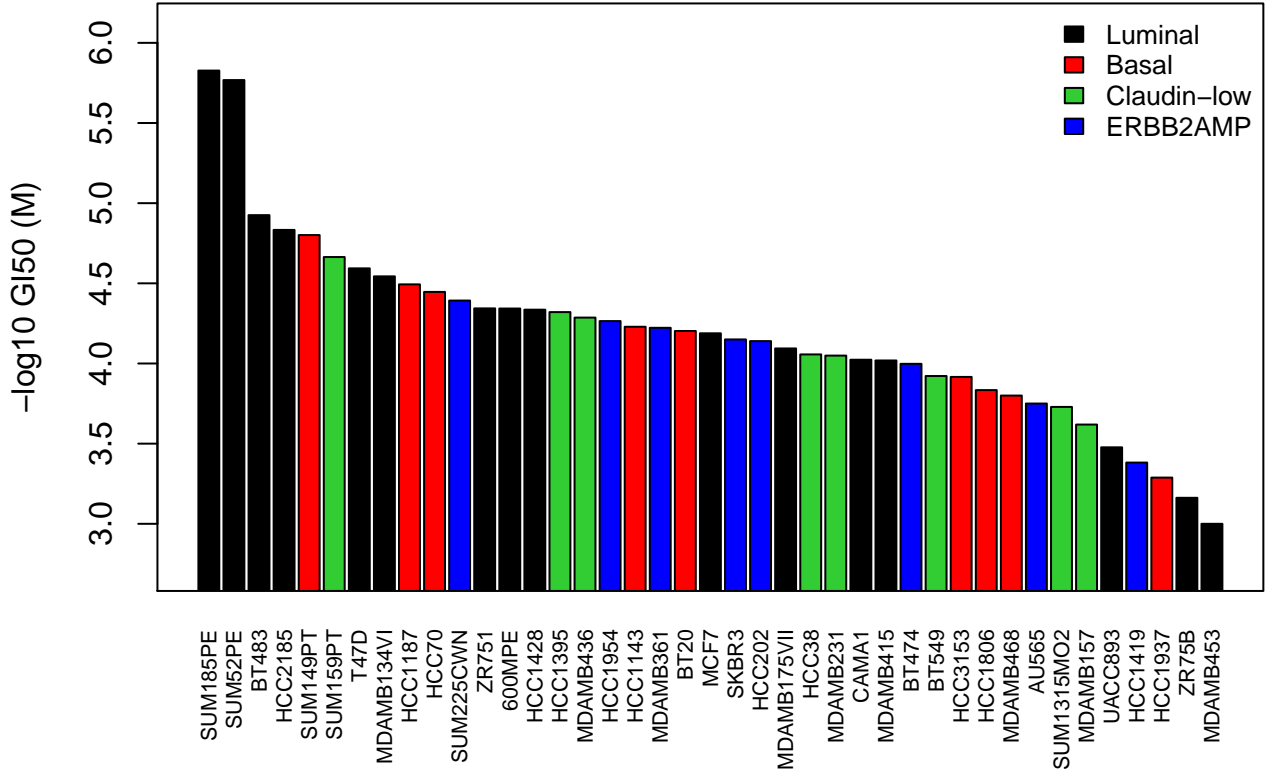
SB-3CT (MMP2, MMP9)



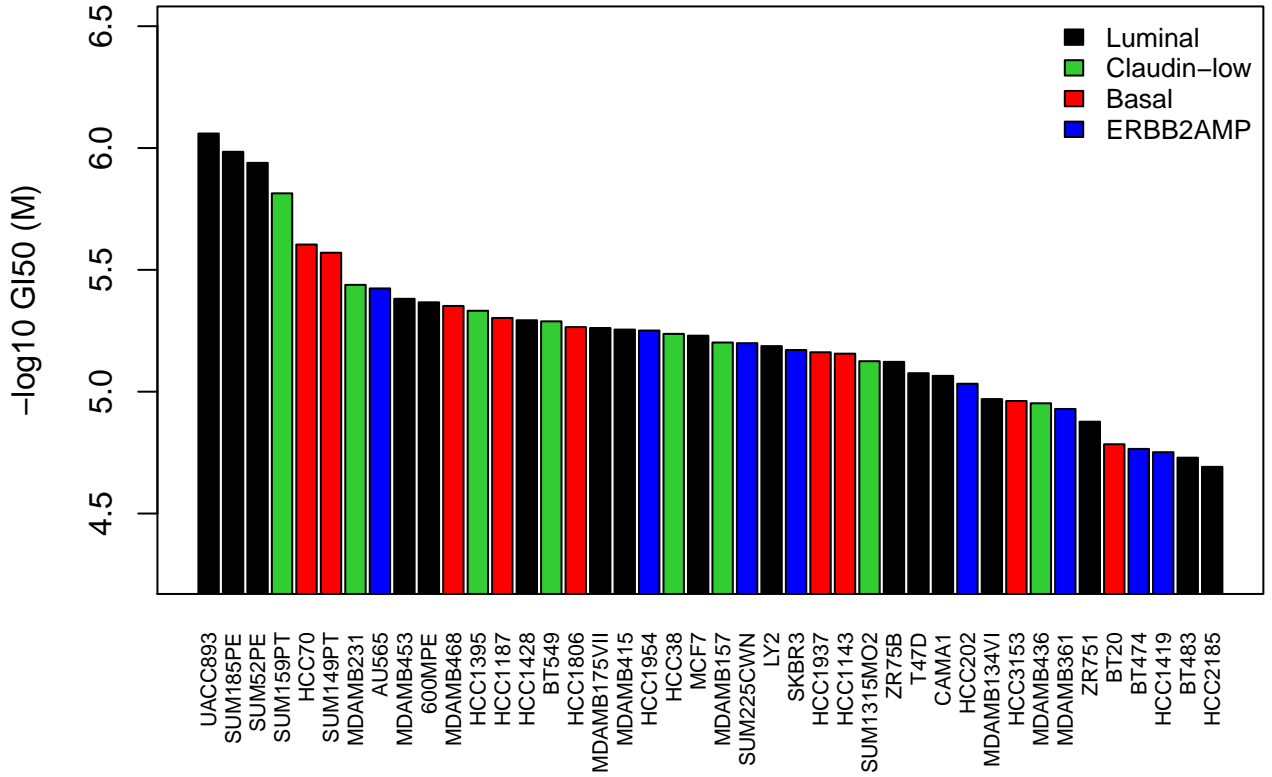
Sigma AKT1-2 inhibitor (Akt 1/2)



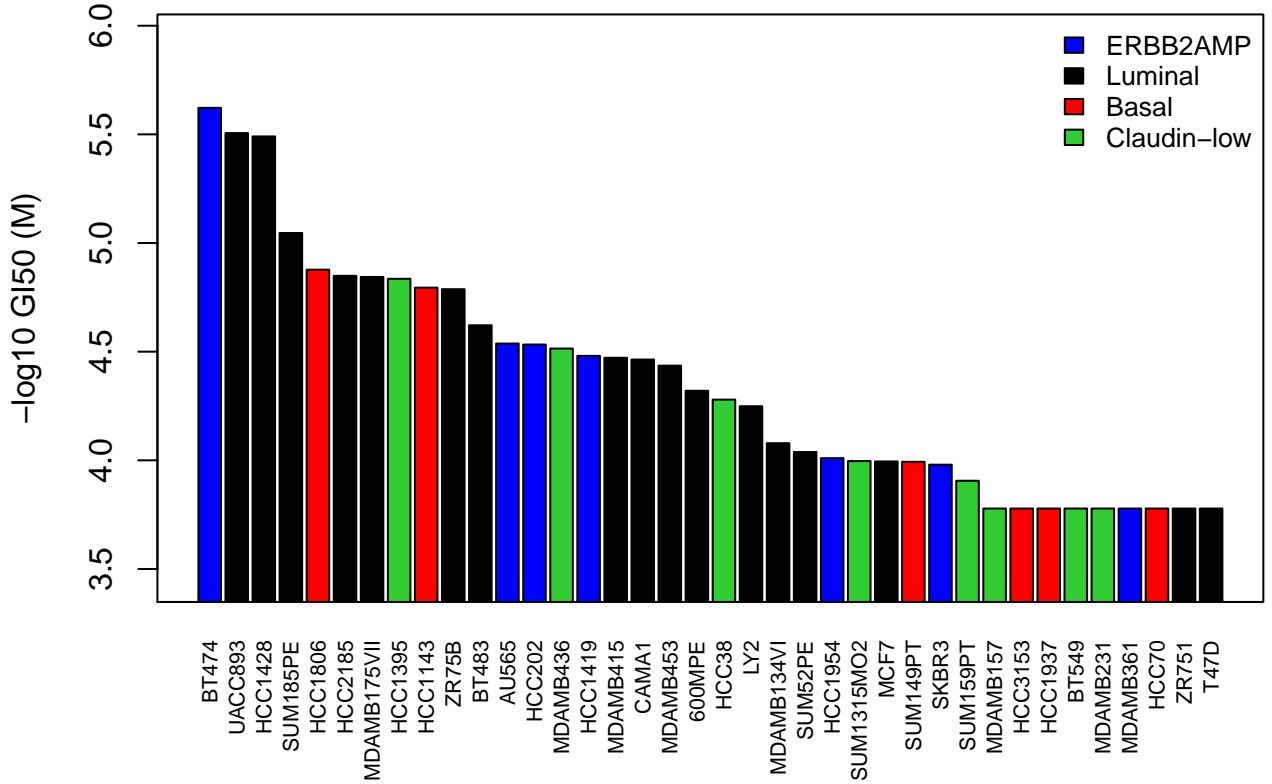
Sorafenib (VEGFR)



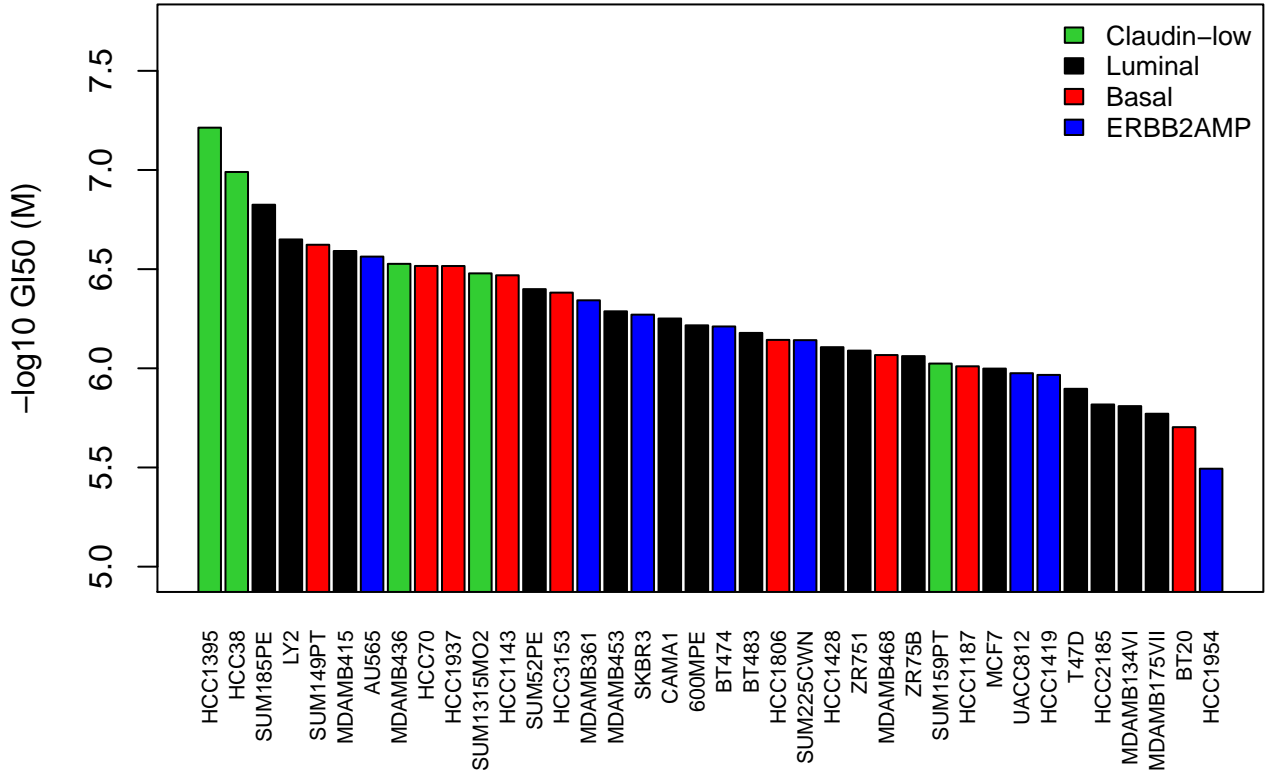
Sunitinib Malate (VEGFR)



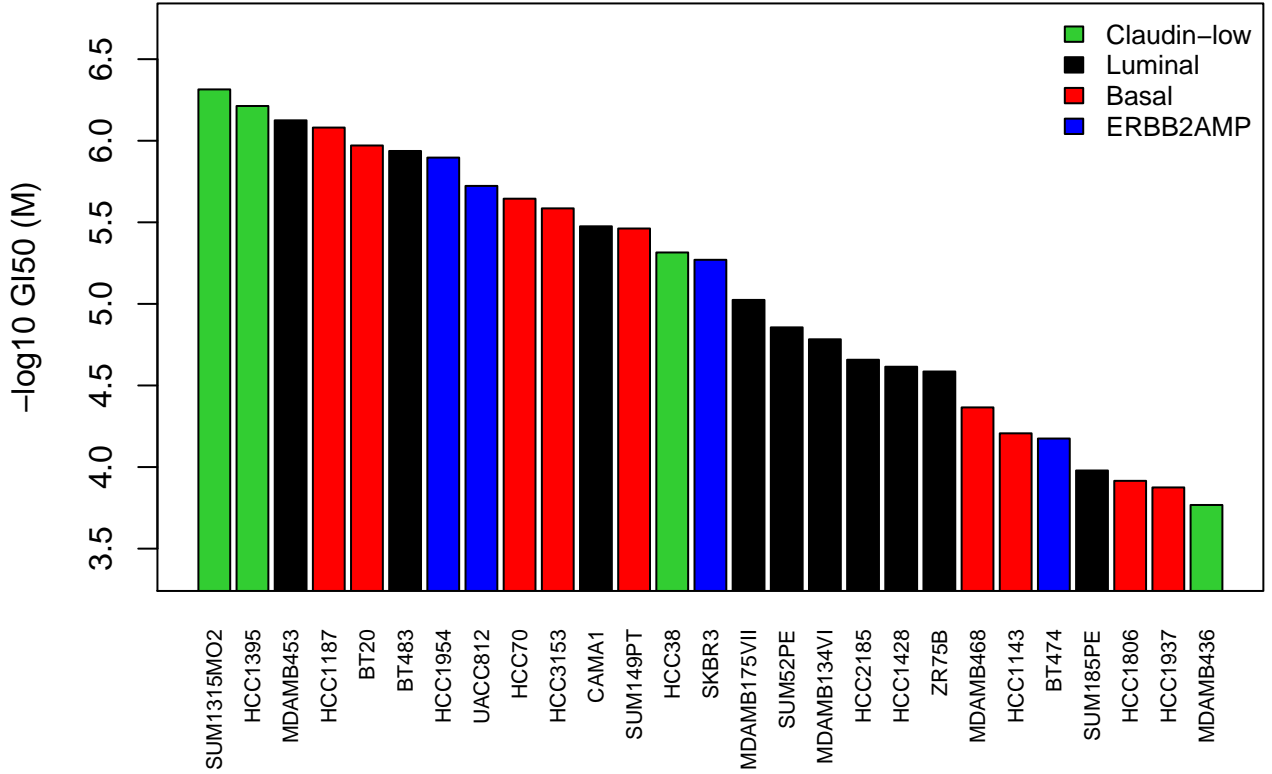
Tamoxifen (ESR1)



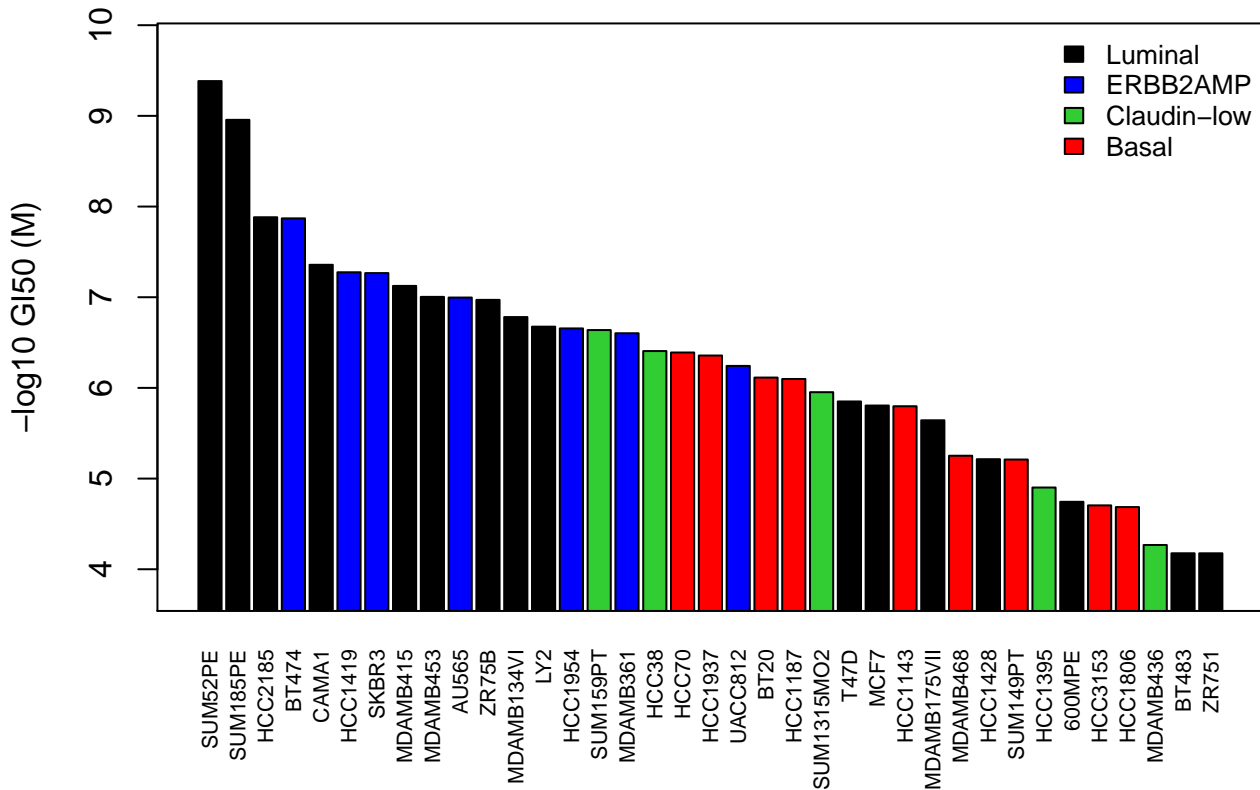
TCS 2312 dihydrochloride (chk1)



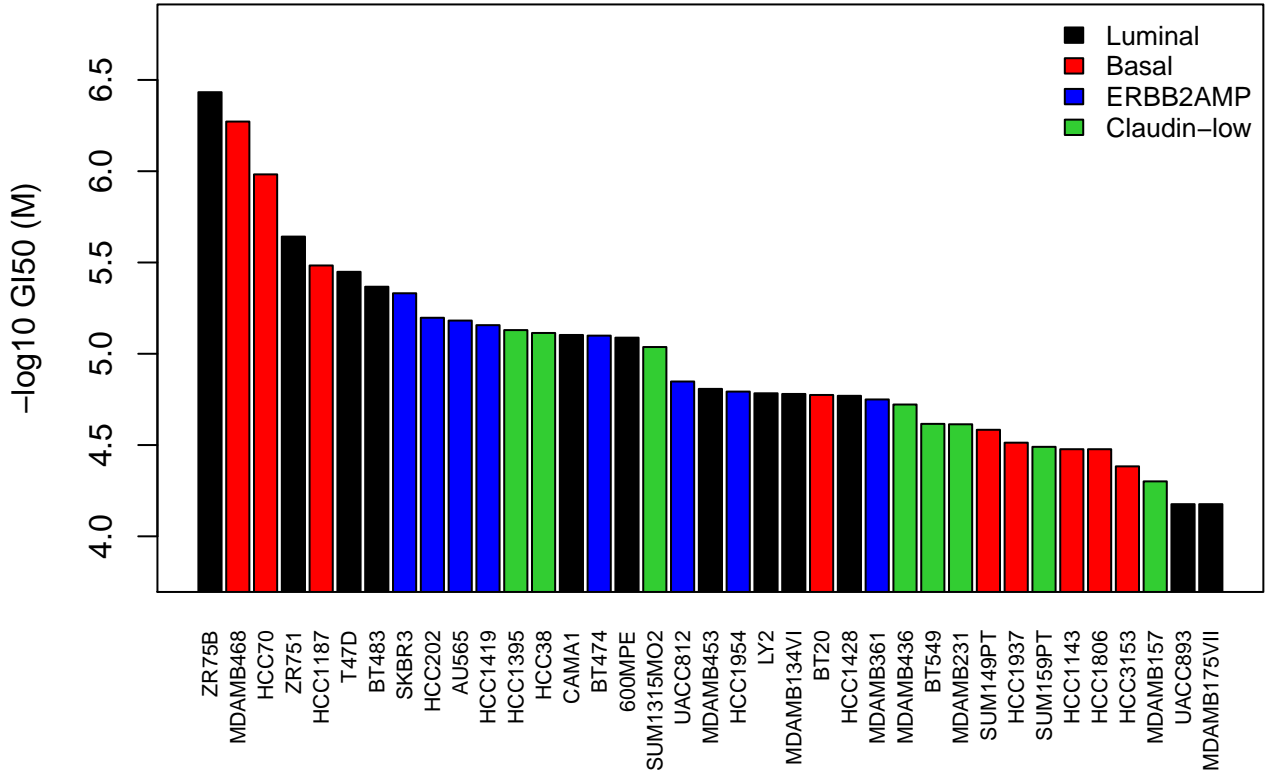
TCS JNK 5a (JNK)



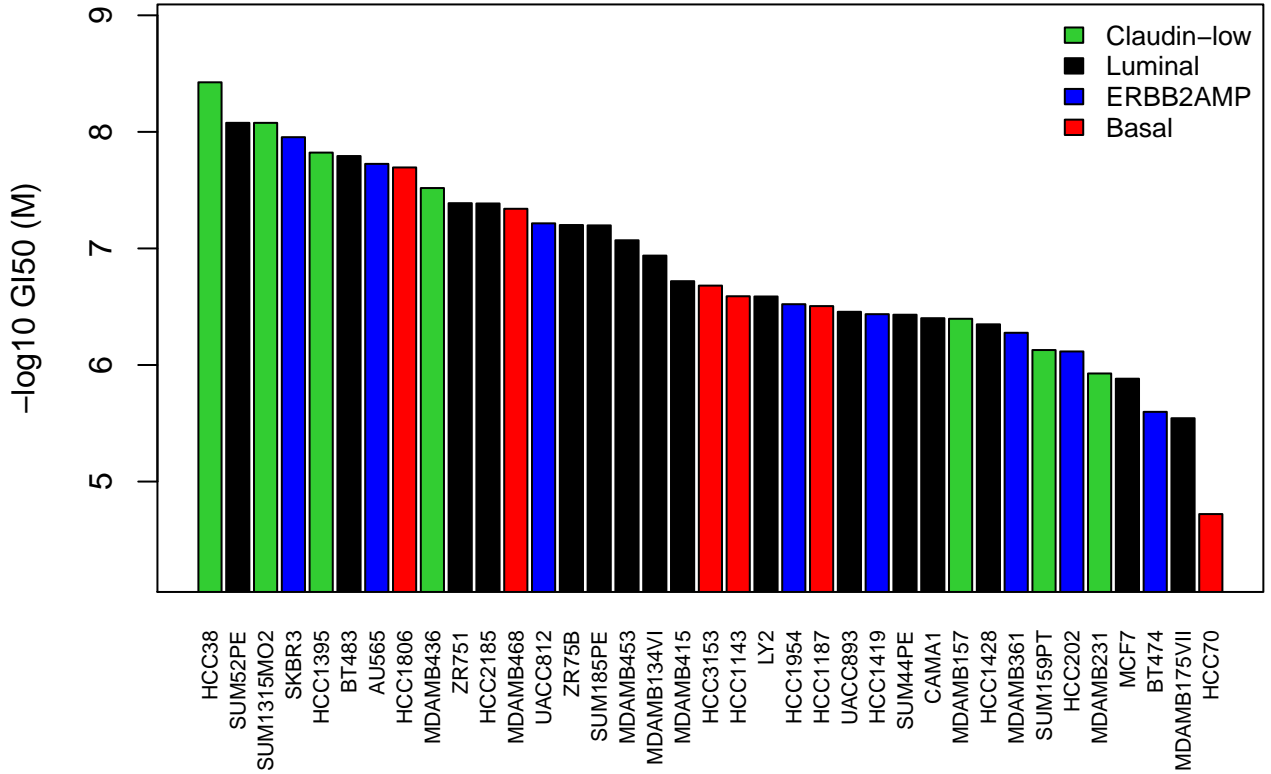
Temsirolimus (mTOR)



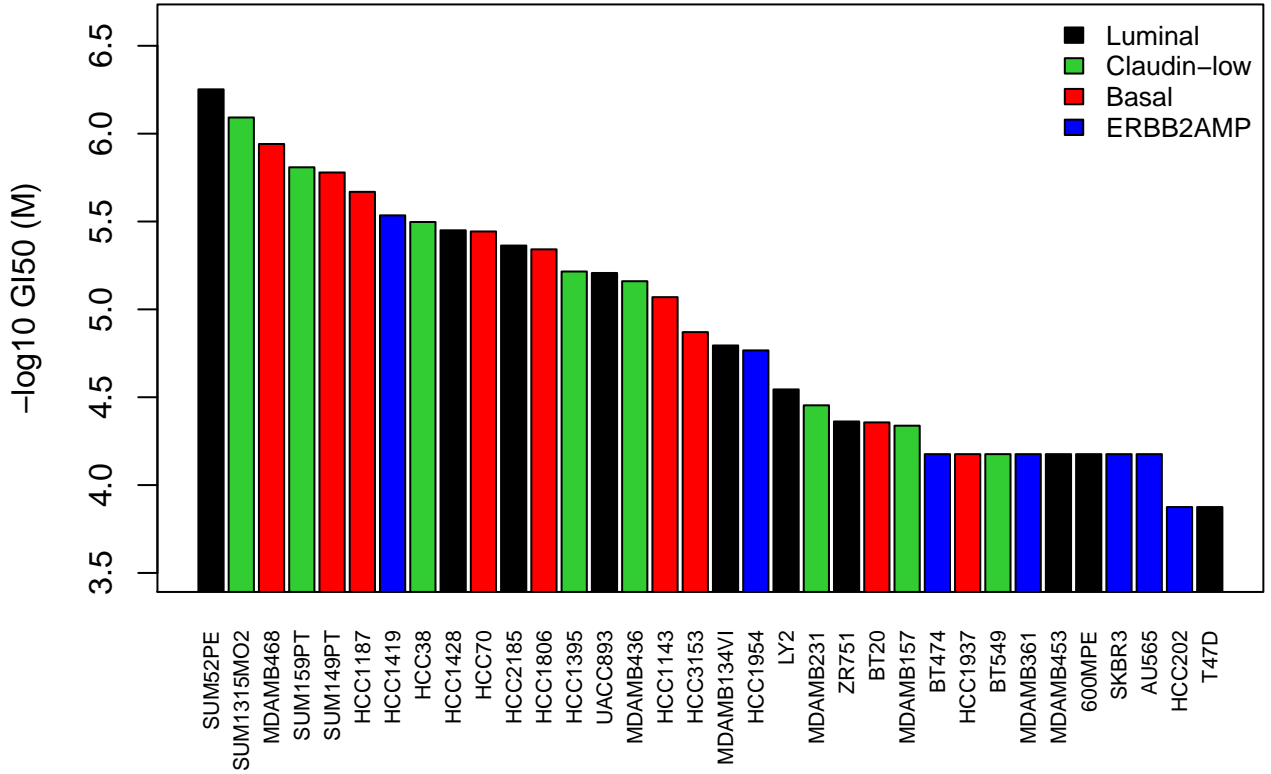
TGX-221 (PI3K, beta selective)



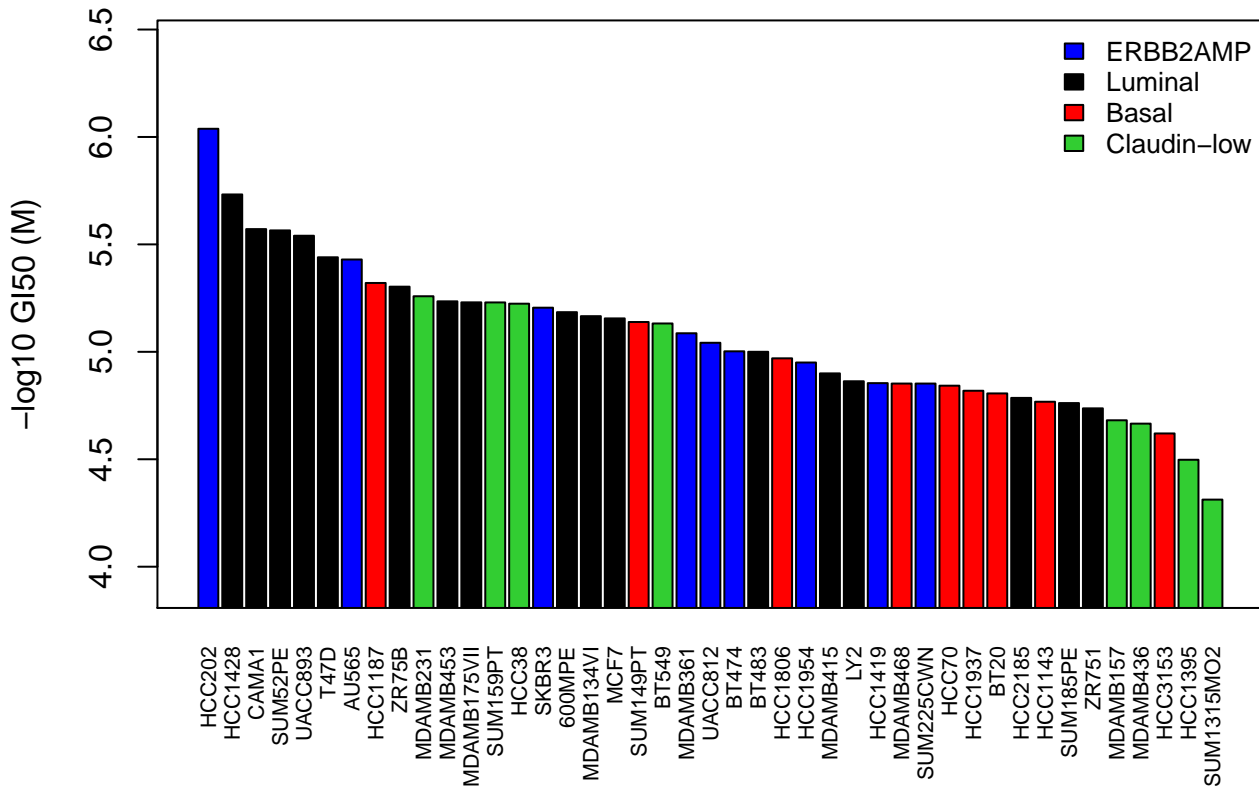
Topotecan (Topoisomerase I)



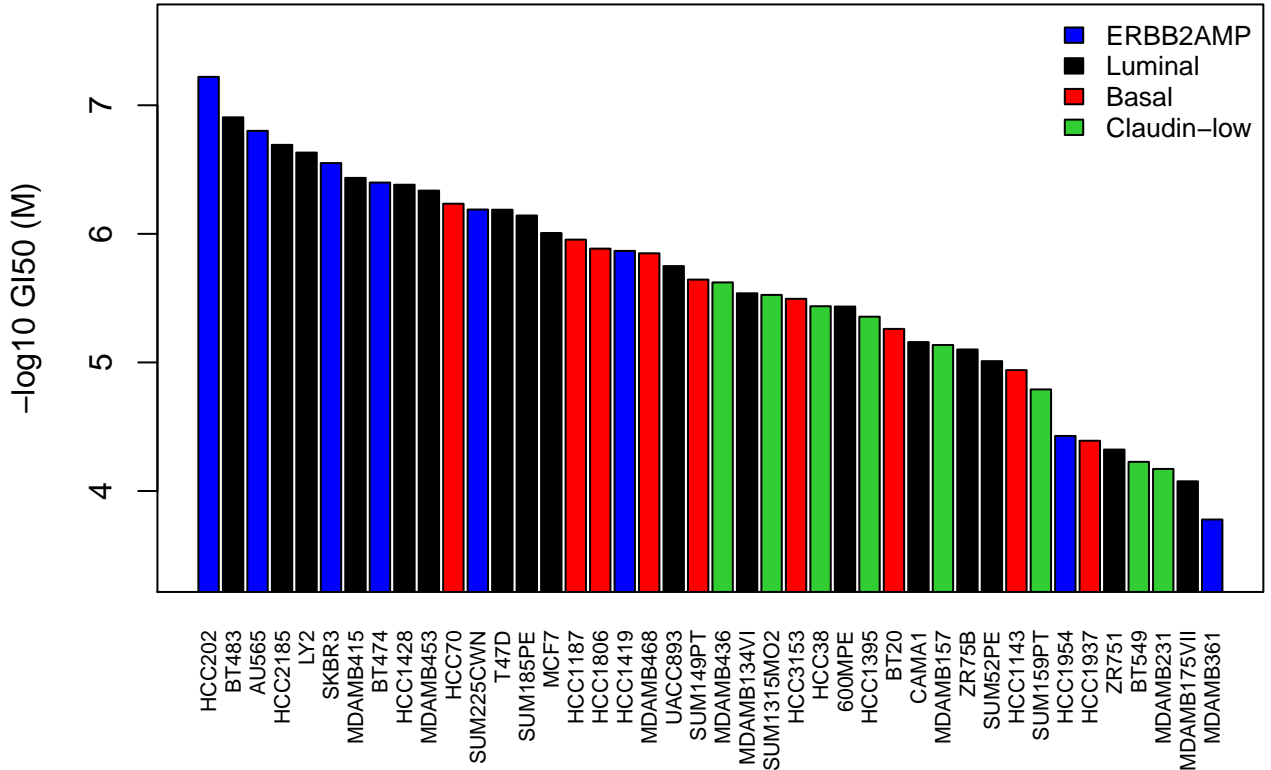
TPCA-1 (IKK2 (I κ B kinase 2))



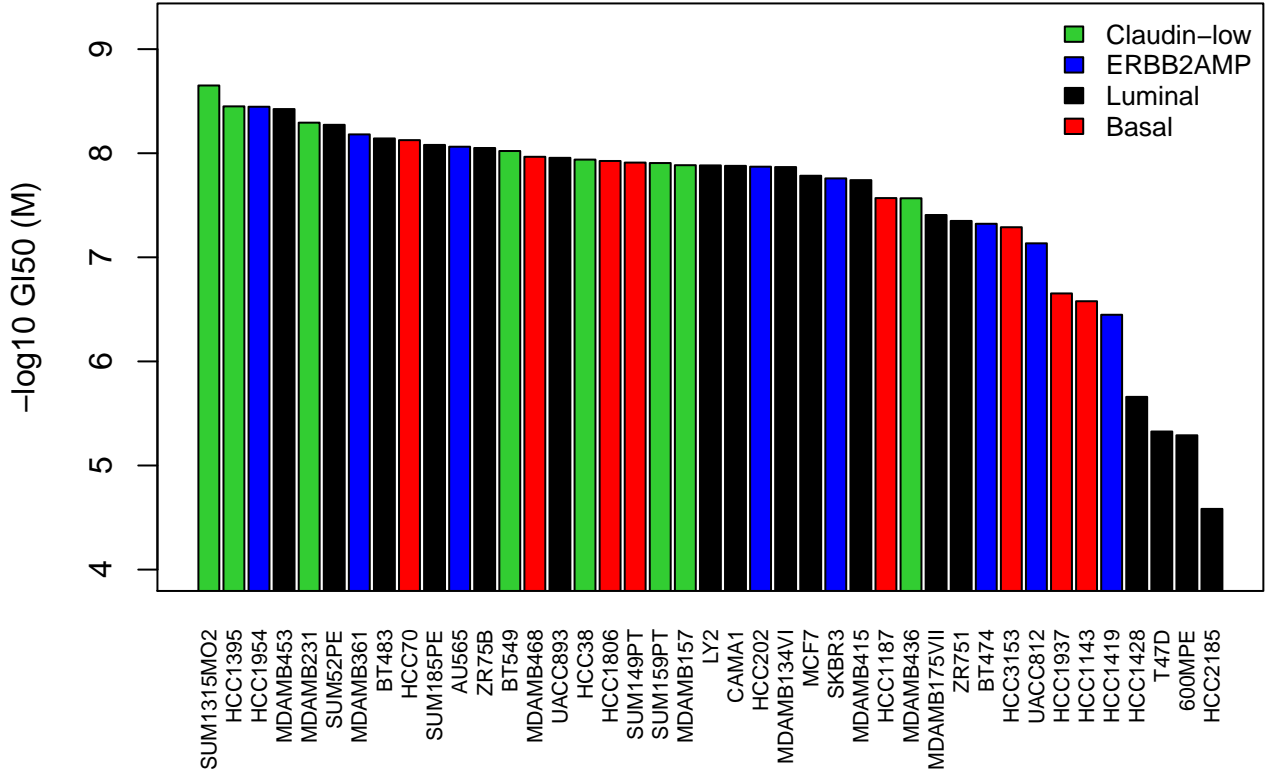
Trichostatin A (Histone deacetylase)



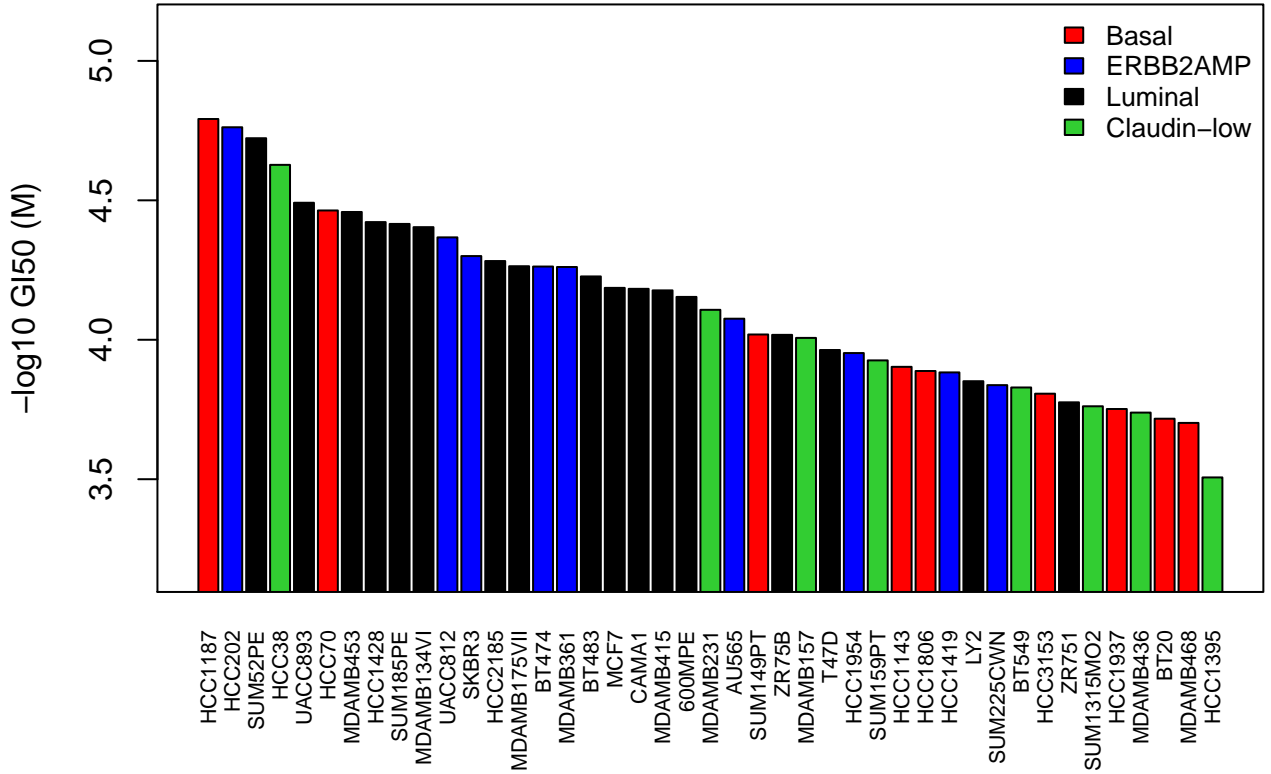
Triciribine (AKT, ZNF217 amplification)



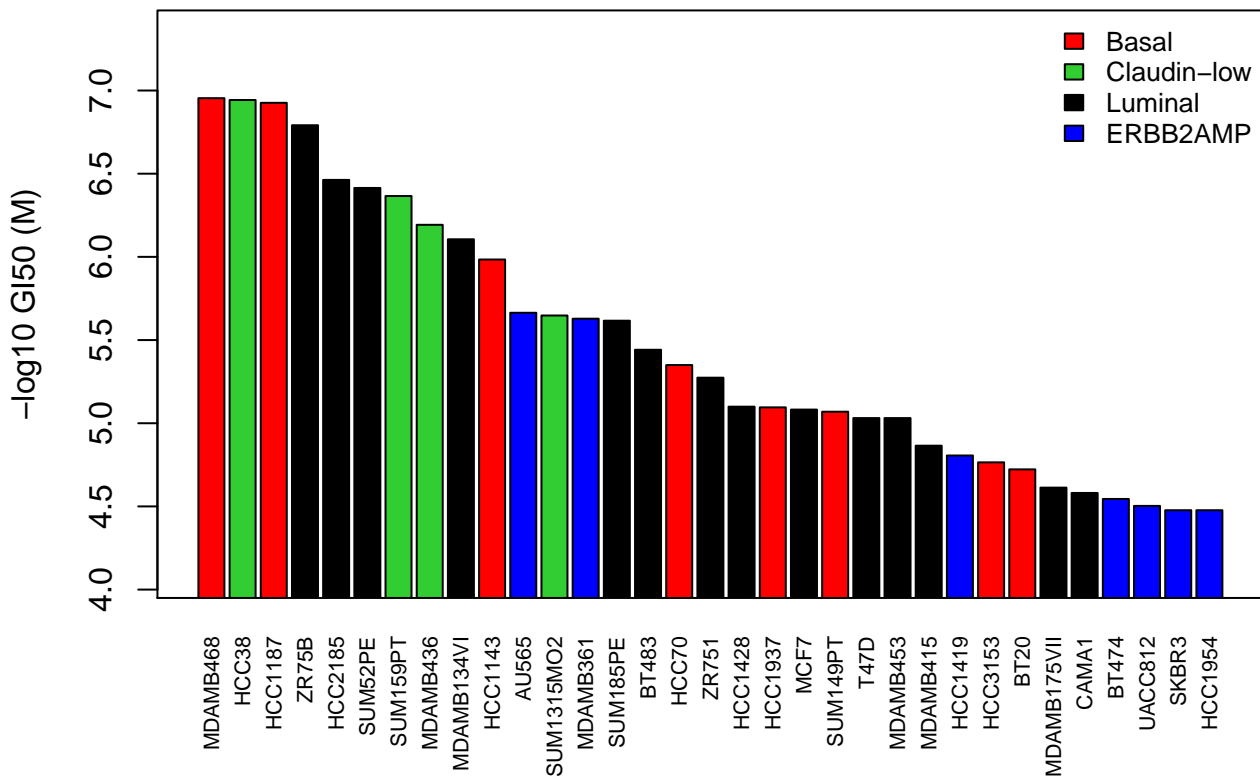
Vinorelbine (Microtubule)



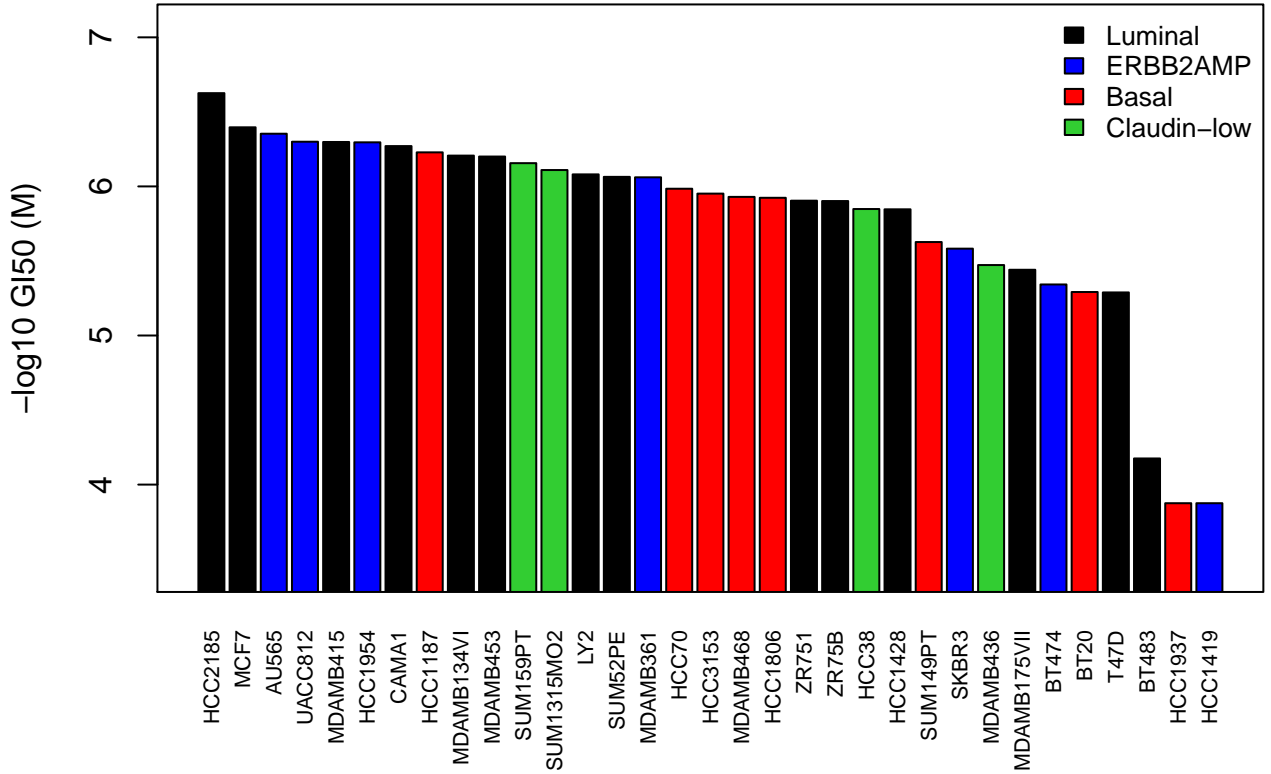
Vorinostat (Histone deacetylase)



VX-680 (aurora kinase)



XRP44X (Ras-Net (Elk-3))



ZM 447439 (AURKA)

