# KnowledgePearls: Provenance-Based Visualization Retrieval Supplementary Material

Holger Stitz, Samuel Gratzl, Harald Piringer, Thomas Zichner, and Marc Streit

\_\_\_\_ **♦** \_\_\_\_\_

## **1** VEGA INTEGRATION

Listing 1. Vega signal declaration with additional track and search properties for the retrieval.

## 2 CASE STUDY: VISUAL ANALYSIS

The following figures show the user interaction with Ordino in chronological order. We describe the user's insight and rationale in each figure caption to make the interaction more transparent.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

🛟 Ordino			
×			
Predefined Datasets			
Cell Line Panels Gene Sets Tissue Panels			
Search Gene		Go Save	
Predefined Sets	🛔 My Sets	Public Sets	
<ul> <li>&gt; All</li> <li>&gt; Cancer Gene Census</li> <li>&gt; DRIVE Genes</li> <li>&gt; Driver Kinases Doi:10.1038/Nrc.2015.18</li> <li>&gt; Normal Chromosome Protein Coding Human Generation</li> </ul>	ies		
<ul> <li>Uploaded Datasets</li> <li>Temporary Sessions</li> </ul>			
Persistent Sessions			

Fig. S 1. The user is interested in the breast cancer cell line *BT-20* and wants to know which cancer gene(s) could be important for growth in that cell line? In Ordino, users can choose between predefined datasets or upload custom datasets. For the case study the user starts with the *Cancer Gene Census* dataset.

## ి Ordino

### Genes Genes

### Showing 563 of 563 Genes

Rank	S	Symbol	Ensembl	Name	Chromosome	Biotype	TPM of BT-20	<ul> <li>Relative Copy Number of BT-20</li> </ul>	* +	8 1
1		EGFR	ENSG00000146648	epidermal growth factor receptor [S	7	protein_coding				
2		MAF	ENSG00000178573	v-maf musculoaponeurotic fibrosarc	16	protein_coding				
3		FBXW7	ENSG00000109670	F-box and WD repeat domain contai	4	protein_coding				
4		CDH1	ENSG0000039068	cadherin 1, type 1, E-cadherin (epith	16	protein_coding				
5		NRAS	ENSG00000213281	neuroblastoma RAS viral (v-ras) on	1	protein_coding				
6		FAM46C	ENSG00000183508	family with sequence similarity 46, ı	1	protein_coding				
7		TRIM33	ENSG00000197323	tripartite motif containing 33 [Sourc	1	protein_codinc				
8		ATP1A1	ENSG00000163399	ATPase, Na+/K+ transporting, alpha	1	protein_codinc				
		NOTCH2	ENSG00000134250	notch 2 [Source:HGNC Symbol;Acc:	1	protein_codinc				
10		ZFHX3	ENSG00000140836	zinc finger homeobox 3 [Source:HG	16	protein_codinc				
11		ZMYM2	ENSG00000121741	zinc finger, MYM-type 2 [Source:HG	13	protein_codinc				
12		IL7R	ENSG00000168685	interleukin 7 receptor [Source:HGN(	5	protein_codinc				
13		ROS1	ENSG00000047936	c-ros oncogene 1, receptor tyrosine	6	protein_coding				
14		FOXO3	ENSG00000118689	forkhead box 03 [Source:HGNC Syn	6	protein_coding				
15		PRDM1	ENSG00000057657	PR domain containing 1, with ZNF c	6	protein_codinc				
16		GOPC	ENSG00000047932	golgi-associated PDZ and coiled-coi	6	protein_coding				
17		WHSC1	ENSG00000109685	Wolf-Hirschhorn syndrome candidat	4	protein_codinc				
18		MLLT10	ENSG00000078403	myeloid/lymphoid or mixed-lineage	10	protein_coding				
19		CBFB	ENSG0000067955	core-binding factor, beta subunit [Se	16	protein_coding				
20		NFE2L2	ENSG00000116044	nuclear factor (erythroid-derived 2)-	2	protein_codinc				
		HOXD13	ENSG00000128714	homeobox D13 [Source:HGNC Syml	2	protein_coding				
22		HOXD11	ENSG00000128713	homeobox D11 [Source:HGNC Syml	2	protein_coding				
23		DROSHA	ENSG00000113360	drosha, ribonuclease type III [Sourc	5	protein_coding				
24		SDHA	ENSG0000073578	succinate dehydrogenase complex,	5	protein_coding				
25		TERT	ENSG00000164362	telomerase reverse transcriptase [S	5	protein_coding				
26		CTCF	ENSG00000102974	CCCTC-binding factor (zinc finger p	16	protein_coding				
27		BCORL1	ENSG0000085185	BCL6 corepressor-like 1 [Source:HG	Х	protein_coding				
		STAG2	ENSG00000101972	stromal antigen 2 [Source:HGNC Sy	Х	protein_coding				
29		ELF4	ENSG00000102034	E74-like factor 4 (ets domain transc	Х	protein_coding				
30		PHF6	ENSG00000156531	PHD finger protein 6 [Source:HGNC	Х	protein_coding				
31		GPC3	ENSG00000147257	glypican 3 [Source:HGNC Symbol;A	Х	protein_coding				
32		MITF	ENSG00000187098	microphthalmia-associated transcri	3	protein_coding				
33		FOXP1	ENSG00000114861	forkhead box P1 [Source:HGNC Syn	3	protein_coding				
34		SBDS	ENSG00000126524	Shwachman-Bodian-Diamond syndr	7	protein_codinc				
35		ABI1	ENSG00000136754	abl-interactor 1 [Source:HGNC Syml	10	protein_codinc				
36		KIF5B	ENSG00000170759	kinesin family member 5B [Source:I		protein_codinc				
37		ERCC5	ENSG00000134899	excision repair cross-complementin		protein_codinc				
		RBM15	ENSG00000162775	RNA binding motif protein 15 [Sour		protein_coding				
39		CDH11	ENSG00000140937	cadherin 11, type 2, OB-cadherin (០៖		protein_coding				
40		CBFA2T3	ENSG00000129993	core-binding factor, runt domain, alr		protein_codinc				
41		SF3B1	ENSG00000115524	splicing factor 3b, subunit 1, 155kD		protein_codinc				
42		CASP8	ENSG0000064012	caspase 8, apoptosis-related cysteii		protein_codinc				
43		PDE4DIP	ENSG00000178104	phosphodiesterase 4D interacting p	1	protein_coding				

Fig. S 2. Adding the two score columns *TPM* (gene expression level) and *Relative Copy Number* for the cell line *BT-20* and sorting the ranking by the relative copy number column reveals the gene *EGFR* as an amplified and also highly expressed candidate. Likely, it is important for cell growth in *BT-20*.

ి Ordino	
🕻 Ordino	

Genes Copy Number

		×	General	Data Source	Cell Line 🔻	Data Subtype Rel	lative Copy Num	ber 🔻	Filter: 58 / 1009		
			Database Info								
Rank	S Symbol		Sample overview	Rank S	Name	Tumor Type	Organ	Gender	▼ EGFR	* +	E 1
1 🗹			Copy Number	1 🗆	MDA-MB-468	breast carcino	breast	female		1	
2				2 🗆	BT-20	breast carcino	breast	female		J	
3 🗆			Expression	3 🗆	CAL-85-1	breast carcino	breast	female			
4 🗆				4 🗆	AU565	breast carcino	breast	female			
5 🗆			Mutation	5 🗆	SK-BR-3	breast carcino	breast	female			
6 🗆				6 🗆	HCC1143	breast carcino	breast	female			
7			Combined View	7 🗆	HDQ-P1	breast carcino	breast	female			
8 🗆				8 🗆	BT-483	breast carcino	breast	female			
9 🗆			OncoPrint	9 🗆	HCC70	breast carcino	breast	female			
10				10 🗆	HCC1187	breast carcino	breast	female			
11 🗆			Visualization	11 🗆	CAL-148	breast carcino	breast	female			
12			Co-Expression	12 🗆	MDA-MB-157	breast carcino	breast	female			
13 🗆			OU EXPICISION	13 🗆	MDA-MB-175-VII	breast carcino	breast	female			
14 🗆			Expression vs. Copy Number	14 🗆	BT-549	breast carcino	breast	female			
15 🗆			Expression vs. copy Number	15 🗆	HCC202	breast carcino	breast	female			
16 🗆			External resources	16 🗆	BT-474	breast carcino	breast	female			
17 🗆	) WHSC1		External resources	17 🗆	T-47D	breast carcino	breast	female			
18 🗆	) MLLT10		canSAR	18 🗆	UACC-812	breast carcino	breast	female			
19 🗆				19 🗆	HCC1419	breast carcino	breast	female			
20 🗆	NFE2L2		Ensembl	20 🗆	UACC-893	breast carcino	breast	female			
21 🗆	) HOXD13			21 🗆	HCC1937	breast carcino	breast	female			
22 🗆	) HOXD11		Human Protein Atlas	22 🗆	HCC38	breast carcino	breast	female			
23 🗆	DROSHA			23 🗆	MDA-MB-436	breast carcino	breast	female			
24 🗆	) SDHA		Open Targets	24 🗆	JIMT-1	breast carcino	breast	female			
25 🗆	) TERT			25 🗆	HCC1428	breast carcino	breast				
26 🗆	CTCF		PubMed	26 🗆	Hs 578T	breast carcino	breast	female			
27 🗆	BCORL1			27 🗆	HCC1395	breast carcino	breast	female			
28 🗆	STAG2		UniProt	28 🗆	DU4475	breast carcino	breast	female			
29 🗆	ELF4			29 🗆	HCC1954	breast carcino	breast	female			
30 🗆	PHF6			30 🗆	Hs 739.T	breast carcino	breast	female			
31 🗆	GPC3			31 🗆	MDA-MB-361	breast carcino	breast	female			
32 🗆	MITF			32 🗆	MDA-MB-231	breast carcino	breast	female			
33 🗆	FOXP1			33 🗆	HCC1500	breast carcino	breast	female			
34 🗆	SBDS			34 🗆	HCC2218	breast carcino	breast	female			
35 🗆	) ABI1			35 🗆	CAMA-1	breast carcino	breast	female			
36 🗆	) KIF5B			36 🗆	Hs 274.T	breast carcino	breast	female			
37 🗆	ERCC5			37 🗆	ZR-75-30	breast carcino	breast	female			
38 🗆	RBM15			38 🗆	CAL-120	breast carcino	breast	female			
39 🗆	CDH11			39 🗆	CAL-51	breast carcino	breast	female			
40 🗆	CBFA2T3			40 🗆	EFM-19	breast carcino	breast	female			
41 🗆	SF3B1			41 🗆	Hs 742.T	breast carcino	breast	female			
42	CASP8			42 🗆	Hs 281.T	breast carcino	breast	female			
43	PDE4DIP			43 🗆	MDA-MB-453	breast carcino	breast	female	-		

Fig. S 3. Next, the user wants to know, if there are other breast cancer cell lines with *EGFR* amplification? Selecting the gene *EGFR* from the ranking brings up a list of further detail views. The user selects the *Copy Number* view, filters the ranking by tumor type *breast cancer*, and sorts it by the *EGFR* copy number values. The insight is that two breast cancer cell lines have a clear *EGFR* copy number gain (CN > 4). There is one cell line, *MDA-MB-468*, that has a higher *EGFR* copy number than *BT-20*.

🛟 Ordino

#### Genes 6

Showing 563 of 563 Genes; 1 selected

Rank S	Symbol	Ensembl	Name	Chromosome	Biotype	TPM of BT-20	Relative Copy Number of BT-20	<ul> <li>Relative Copy</li> <li>Number &gt; 4</li> </ul>	* +	e 1
	ERBB2	ENSG00000141736	v-erb-b2 erythroblastic leukemia vira	17	protein_codinc					
2 🗆	CDK12	ENSG00000167258	cyclin-dependent kinase 12 [Source	17	protein_coding					
3 🗆	EXT1	ENSG00000182197	exostosin 1 [Source:HGNC Symbol;,	8	protein_coding					
4 🗆	MYC	ENSG00000136997	v-myc myelocytomatosis viral onco	8	protein_coding					
5 🗆	RAD21	ENSG00000164754	RAD21 homolog (S. pombe) [Sourc	8	protein_coding					
	PPM1D	ENSG00000170836	protein phosphatase, Mg2+/Mn2+ d	17	protein_coding					
7 🗆	LASP1	ENSG0000002834	LIM and SH3 protein 1 [Source:HGN	17	protein_coding					
8 🗆	NBN	ENSG00000104320	nibrin [Source:HGNC Symbol;Acc:76	8	protein_coding					
9 🗆	CCND1	ENSG00000110092	cyclin D1 [Source:HGNC Symbol;Ac	11	protein_coding					
10 🗆	GNAS	ENSG0000087460	GNAS complex locus [Source:HGNC	20	protein_coding					
11 🗆	FGFR1	ENSG00000077782	fibroblast growth factor receptor 1	8	protein_coding					
12 🗆		ENSG00000168646	axin 2 [Source:HGNC Symbol;Acc:9	17	protein_coding					
13 🗆	CLTC	ENSG00000141367	clathrin, heavy chain (Hc) [Source:	17	protein_coding					
	COL1A1	ENSG00000108821	collagen, type I, alpha 1 [Source:HG	17	protein_coding					
15 🗆		ENSG00000108292	myeloid/lymphoid or mixed-lineage	17	protein_coding	1				
		ENSG00000105173	cyclin E1 [Source:HGNC Symbol;Ac	19	protein_coding					
	WHSC1L1	ENSG00000147548	Wolf-Hirschhorn syndrome candidat	8	protein_coding					
18 🗆	BRIP1	ENSG00000136492	BRCA1 interacting protein C-termina	17	protein_coding					
19 🗆		ENSG00000121067	speckle-type POZ protein [Source:H	17	protein_coding					
20 🗹		ENSG00000146648	epidermal growth factor receptor [S	7	protein_coding	<mark>32</mark> 3	5.03	0.0351	]	
21 🗆		ENSG00000164683	hairy/enhancer-of-split related with	8	protein_coding					
	NCOA2	ENSG00000140396	nuclear receptor coactivator 2 [Sou		protein_coding					
	SALL4	ENSG00000101115	sal-like 4 (Drosophila) [Source:HGN	20	protein_coding					
	KAT6A	ENSG0000083168	K(lysine) acetyltransferase 6A [Sou	8	protein_coding					
	TRIM33	ENSG00000197323	tripartite motif containing 33 [Sourc	1	protein_coding					
	EIF3E	ENSG00000104408	eukaryotic translation initiation fact	8	protein_codinç					
	RUNX1T1	ENSG00000079102	runt-related transcription factor 1; ti	8	protein_coding					
28 🗆	MDM2	ENSG00000135679	Mdm2, p53 E3 ubiquitin protein liga	12	protein_coding					
29 🗆	DDX3X	ENSG00000215301	DEAD (Asp-Glu-Ala-Asp) box polype	Х	protein_coding					
	RBM10	ENSG00000182872	RNA binding motif protein 10 [Source	Х	protein_coding					
31 🗆	RARA	ENSG00000131759	retinoic acid receptor, alpha [Source	17	protein_coding					
	CDK4	ENSG00000135446	cyclin-dependent kinase 4 [Source:	12	protein_coding					
	LM02	ENSG00000135363	LIM domain only 2 (rhombotin-like	11	protein_coding					
34 🗆		ENSG00000141027	nuclear receptor corepressor 1 [Sou	17	protein_coding					
35 🗆	SS18L1	ENSG00000184402	synovial sarcoma translocation gen	20	protein_coding					
36 🗆	AKT1	ENSG00000142208	v-akt murine thymoma viral oncoge		protein_codinc					
	UBR5	ENSG00000104517	ubiquitin protein ligase E3 compone	8	protein_coding					
	MYOD1	ENSG00000129152	myogenic differentiation 1 [Source:	11	protein_coding					
	MYD88	ENSG00000172936	myeloid differentiation primary resp	3	protein_coding					
40 🗆		ENSG00000197157	staphylococcal nuclease and tudor	7	protein_coding					
	BCL2	ENSG00000171791	B-cell CLL/lymphoma 2 [Source:HG	18	protein_codinc					
41 🗆										
	МҮВ	ENSG00000118513	v-myb myeloblastosis viral oncogen	6	protein_codinc					

Fig. S 4. The user has a new thought and wants to know which cancer genes are more frequently amplified in breast cancer than *EGFR*? The previous detail view is closed and another score column with the parameters tumor type *breast cancer*; data type *Relative Copy Number*; aggregation *Frequency*; cut-off > 4 is added to the ranking. Sorting this column reveals that *EGFR* is only ranked on place 20 in the list. Most frequently amplified in the assessed breast cancer cell lines is *ERBB2*, which is amplified in about 25% of cell lines.

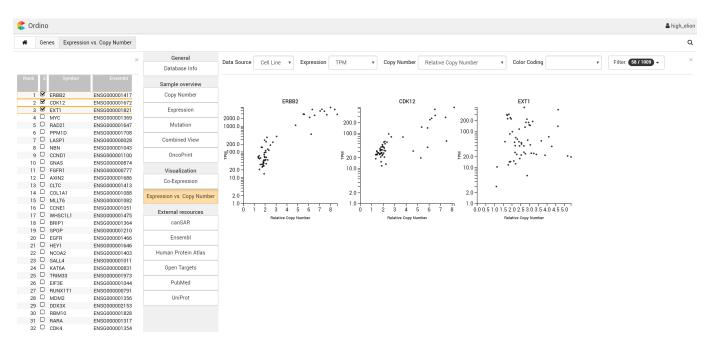


Fig. S 5. Next the user wants to know, if the amplifications of the top genes also lead to a high gene expression? Selecting the first three genes (*ERBB2, CDK12,* and *EXT1*) and opening the detail view *Expression vs. Copy Number* (still filtered for *breast cancer* cell lines) shows a scatterplot for every selected gene. It is obvious that the gene expression for all *ERBB2* amplified cell lines is very high. This correlation is also present in a weaker form for *CDK12* and seems to be absent for *EXT1*.

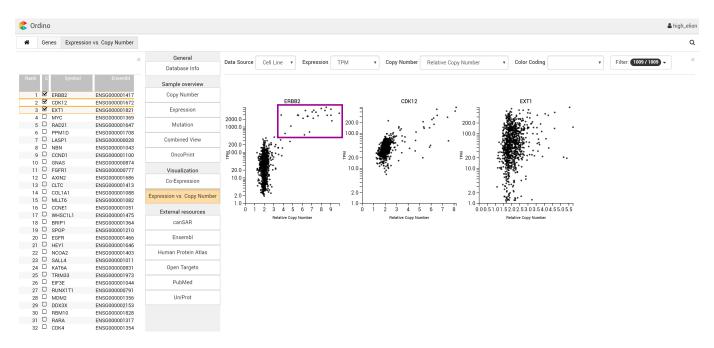


Fig. S 6. The user removes the filter to explore if the expression to copy number correlation is also present in other tumor types. The result is that all cell lines with a very high *ERBB2* copy number have high *ERBB2* expression.

🛟 Or	dino					
ñ	Genes	Expression vs. Copy Number	Database Info			
			Ge	eneral		
				ase Info		
		4	Gene	overview	Field Name	NCI-H2170
		ERBB2	Сору	Number	celllinename	NCI-H2170
			Exp	ression	species	human
2000.0 1000.0			Mu	tation	organ	lung
			Combined View		metastatic_site	
200.0 ឆ្នា 00.0					histology_type	squamous cell carcinoma
	E 🕈			I resources	morphology	
20.0 10.0					tumortype	NSCLC
					growth_type	
2.0 1.0	- *				gender	male
	0 1 2	3 4 5 6 7 8 9			age_at_surgery	
		Relative Copy Number				

Fig. S 7. Selecting the most right point in the *ERBB2* scatterplot and opening the *Database Info* detail view, reveals that the cell line with the highest *ERBB2* amplification is *NCI-H2170* and belongs to *non-small-cell lung cancer* (NSCLC).

🛟 Ordino		
Genes		

Showing 563 of 563 Genes

Rank S	Symbol	Ensembl	Name	Chromosome	Biotype	TPM of BT-20	Relative Copy Number of BT-20	Relative Copy Number > 4	<ul> <li>Relative Copy</li> <li>Number &gt; 4</li> </ul>	초 🖻 + 1
	MYC	ENSG00000136997	v-myc myelocytomatosis viral onco	8	protein_coding					
2 🗆	EGFR	ENSG00000146648	epidermal growth factor receptor [S	7	protein_coding					
3 🗆	NKX2-1	ENSG00000136352	NK2 homeobox 1 [Source:HGNC Sy	14	protein_coding					
4 🗆	CDK4	ENSG00000135446	cyclin-dependent kinase 4 [Source:H	12	protein_coding					
5 🗆	MECOM	ENSG0000085276	MDS1 and EVI1 complex locus [Sou	3	protein_coding					
6 🗆	KAT6A	ENSG0000083168	K(lysine) acetyltransferase 6A [Sou	8	protein_coding					
7 🗆	KIT	ENSG00000157404	v-kit Hardy-Zuckerman 4 feline sarc	4	protein_coding					
8 🗆	LZTR1	ENSG0000099949	leucine-zipper-like transcription regu	22	protein_coding					
9 🗆	ARNT	ENSG00000143437	aryl hydrocarbon receptor nuclear ti	1	protein_coding					
10 🗆	ERBB2	ENSG00000141736	v-erb-b2 erythroblastic leukemia vira	17	protein_coding					
11 🗆	MET	ENSG00000105976	met proto-oncogene (hepatocyte gr	7	protein_coding					
12 🗆	MLLT11	ENSG00000213190	myeloid/lymphoid or mixed-lineage	1	protein_coding					
13 🗆	FOXA1	ENSG00000129514	forkhead box A1 [Source:HGNC Syn	14	protein_coding					
14 🗆	KRAS	ENSG00000133703	v-Ki-ras2 Kirsten rat sarcoma viral o	12	protein_coding					
15 🗆	BCL9	ENSG00000116128	B-cell CLL/lymphoma 9 [Source:HG	1	protein_coding					
16 🗆	CDK12	ENSG00000167258	cyclin-dependent kinase 12 [Source	17	protein_coding					
17 🗆	MAPK1	ENSG00000100030	mitogen-activated protein kinase 1	22	protein_coding					
18 🗆	TPM3	ENSG00000143549	tropomyosin 3 [Source:HGNC Symb	1	protein_coding					
19 🗆	NRAS	ENSG00000213281	neuroblastoma RAS viral (v-ras) on	1	protein_coding					
20 🗆	NRG1	ENSG00000157168	neuregulin 1 [Source:HGNC Symbol	8	protein_codinc					
21 🗆	PDGFRA	ENSG00000134853	platelet-derived growth factor recep	4	protein_codinc					
22 🗆	TERT	ENSG00000164362	telomerase reverse transcriptase [S	5	protein_codinc					
23 🗆	TRIM33	ENSG00000197323	tripartite motif containing 33 [Sourc	1	protein_codinc					
24 🗆	CRLF2	ENSG00000205755	cytokine receptor-like factor 2 [Sour	х	protein_codinc					
25 🗆	FAM46C	ENSG00000183508	family with sequence similarity 46, I	1	protein_coding					
26 🗆	FANCD2	ENSG00000144554	Fanconi anemia, complementation (	3	protein_coding					
27 🗆	FGFR1	ENSG00000077782	fibroblast growth factor receptor 1	8	protein_coding					
28 🗆	FIP1L1	ENSG00000145216	FIP1 like 1 (S. cerevisiae) [Source:F	4	protein_coding					
29 🗆	BCL7A	ENSG00000110987	B-cell CLL/lymphoma 7A [Source:HI	12	protein_coding					
30 🗆	FUBP1	ENSG00000162613	far upstream element (FUSE) bindir	1	protein_coding					
31 🗆	HMGA2	ENSG00000149948	high mobility group AT-hook 2 [Sour	12	protein_coding					
32 🗆	HNF1A	ENSG00000135100	HNF1 homeobox A [Source:HGNC S	12	protein_coding					
33 🗆	HOXC11	ENSG00000123388	homeobox C11 [Source:HGNC Symł	12	protein_coding					
34 🗆	HOXC13	ENSG00000123364	homeobox C13 [Source:HGNC Symł	12	protein_coding					
35 🗆	IKBKB	ENSG00000104365	inhibitor of kappa light polypeptide		protein_coding					
36 🗆	AR	ENSG00000169083	androgen receptor [Source:HGNC S	Х	protein_coding					
37 🗆	KDM5A	ENSG0000073614	lysine (K)-specific demethylase 5A	12	protein_coding					
38 🗆	BCR	ENSG00000186716	breakpoint cluster region [Source:H	22	protein_coding					
39 🗆	CCND1	ENSG00000110092	cyclin D1 [Source:HGNC Symbol;Ac	11	protein_coding					
40 🗆	ABI1	ENSG00000136754	abl-interactor 1 [Source:HGNC Syml	10	protein_codinç					

Fig. S 8. With the next step the user wants to investigate, which genes are typically amplified in *non-small-cell lung cancer* (NSCLC)? He switches back to the initial gene ranking, adds another cell line score column with the parameters tumor type *NSCLC*; data type *Relative Copy Number*; aggregation *Frequency*; cut-off > 4, and sorts the ranking by this column. Interestingly, the gene *EGFR* is ranked second this time after the gene *MYC*.

🛟 Ordino									
Genes	Copy Number								
		× General							
		Database Info	Data Source	Cell Line 🔻	Data Subtype	Relative Copy Number	•	Filter: 127 / 1009	Showing 127 of 127 Cell Lines
Rank S	Symbol Ensemb	Sample overview	Rank S	Name	Tumor Ty	be Organ	Gender	- EGFR	± ≞ + ±
1 🗆 мус	ENSG0000013	Copy Number	10	HCC-827	NSCLC	lung	female		+ ±
2 🗹 EGFI				NCI-H3255	NSCLC	lung	female		
3 🗆 NKX			3 🗆	NCI-H1573	NSCLC	lung	female		
4 CDK			4 🗆	NCI-H1838	NSCLC	lung	female		
5 🗆 MEC			5 🗆	NCI-H1568	NSCLC	lung	female		
6 🗆 KATO			6 🗆	HCC-2279	NSCLC	lung	female		
7 🗆 кіт	ENSG0000015		7 🗆	LUDLU-1	NSCLC	lung	male		
8 D LZTF			8 🗆	HCC4006	NSCLC	lung	male		
9 🗆 ARN			9 🗆	LU65	NSCLC	lung	male		
10 C ERBI			10 🗆	NCI-H596	NSCLC	lung	male		
11 O MET			11 🗆	NCI-H1623	NSCLC	lung	male		
12 O MLL		31	12 🗆	PC-14	NSCLC	lung	marc		
13 C FOX		Co-Expression	13 🗆	NCI-H2073	NSCLC	lung	female		
14 D KBA		27	14 🗆	LOU-NH91	NSCLC	lung	female		
15 D BCL		Expression vs. Copy Number	15 🗆	KNS-62	NSCLC	lung	male		
16 CDK			16 🗆	RERF-LC-AI	NSCLC	lung	male		
17 D MAP		External recources	17 0	ChaGo-K-1	NSCLC	lung	male		
18 O TPM			18 🗆	NCI-H1703	NSCLC	lung	male		
19 🗆 NRA			19 🗆	NCI-H2228	NSCLC	lung	female		
20 D NRG			20	LC-1F	NSCLC	lung	male		
21 D PDG			20 0	Calu-3	NSCLC	lung	male		
21 C FDG 22 TER			21 0	HCC2935	NSCLC	lung	male		
23 C TER			22 0	NCI-H2444	NSCLC	lung	male		
23 C TRIN 24 C CRLI			23 0	NCI-H2444 NCI-H1975	NSCLC	lung	_		
24 C CRL			24 0	NCI-H1975 NCI-H2291	NSCLC	lung	female		
26 G FAN			26	NCI-H2291	NSCLC	lung	male male		
27 G FGFI			20 0	NCI-H1303	NSCLC		male		
28 G FIP1			27 0	NCI-H2087	NSCLC	lung	male		
29 D BCL			29 🗆	NCI-H1792 NCI-H2342	NSCLC		_		
30 D FUB			30	LXF-289	NSCLC	lung	male male		
30 C FOB			30 0	NCI-H2085	NSCLC	lung	_		
31 U HMG 32 U HNF			31 0	NCI-H2085 NCI-H522	NSCLC	lung	male male		
32 U HNF 33 U HOX			32		NSCLC	lung	female		
33 U HOX 34 U HOX			33 0	NCI-H2023	NSCLC	lung lung			
34 C HOX 35 C IKBK			34 0	ABC-1	NSCLC		male male		
35 G IKBK	ENSG0000010 ENSG0000016		35	NCI-H1650	NSCLC	lung			
36 U AR 37 U KDM			36	NCI-H1650 NCI-H1651		lung	male		
			-		NSCLC	lung	male		
_				NCI-H292	NSCLC	lung	female		
			39 🗆	NCI-H2110	NSCLC	lung	Consult.		
40 🗆 ABI1	ENSG0000013		40 🗆	NCI-H1734	NSCLC	lung	female		

Fig. S 9. The user is still interested in *EGFR* and wants to know which *NSCLC* cell lines have an *EGFR* amplification? He selects the gene *EGFR* from the ranking and opens the *Copy Number* detail view. The copy number value for *EGFR* is added as additional column to the following cell line ranking. Additionally the ranking is filtered by *NSCLC* cell lines only and sorted by the relative copy number. The top hit is the cell line *HCC-827*.

## 3 CASE STUDY: SEARCH AND CONTINUATION OF ANALYSIS

The initial question is: which analysis states that are similar to the last one were done previously?

Genes Copy Nur	nber									d Current Session History $\mathbf{T}$ ×	Q Search in Current Session
×	General	Data Source	Cell Line 🔻	Data Subtyne	Relative Copy Num	har v	Filter: 127 / 1009	-	×	0	Search for attribute, selection,
	Database Info	butu oouroc	Gen Enice	butu oubtype	neiauve copy nun	bei ·				Selected ENSG00000146648 (1 Ensembls)	
S Symbol 🔺	Sample overview	Rank S	Name	Tumor Typ	e Organ	Gender	<ul> <li>EGFR</li> </ul>	± 8	*	e 📶 + Add Copy Number	× ENSG00000146648 ×
O MYC								+ ±		• 📶 Set Parameter "filter"	
2 EGFR	Copy Number		HCC-827 NCI-H3255	NSCLC	lung	female				Change Sort Criteria	
3 D NKX2-1	Expression		NCI-H3255 NCI-H1573	NSCLC NSCLC	lung	female female				Add Aggregated Cell Line Score	11 Provenance States found
4 C CDK4	Expression		NCI-H1838	NSCLC	lung	female	_			Change Sort Criteria	
5 MECOM	Mutation		NCI-H1568	NSCLC	lung	female				1	Selected ENSG00000146648 (1 Ensembls) •
6 C KAT6A		6 🗆	HCC-2279	NSCLC	lung	female				1	Views: Genes
7 🗆 KIT	Combined View		LUDLU-1	NSCLC	lung	male				1	Selected Genes: ENSG00000146648
8 C LZTR1			HCC4006	NSCLC	lung	male				1	
9 🗆 ARNT	OncoPrint			NSCLC	lung	male				1	Selected ENSG00000146648 (1 Ensembls)
10 🗆 ERBB2			NCI-H596	NSCLC	lung	male				Selected ENSG00000146648 (1 Ensembls)	Views: Genes
11 🗆 MET	Visualization		NCI-H1623	NSCLC	lung	male				all+ Add Copy Number	Selected Genes: ENS000000146648
12 O MLLT11	Co-Expression	12 🗆		NSCLC	lung					• 📶 Set Parameter "filter"	Contente office.
13 D FOXA1			NCI-H2073	NSCLC	lung	female				• Change Sort Criteria	
14 C KRAS	Expression vs. Copy Number		LOU-NH91	NSCLC	lung	female					
15 D BCL9			KNS-62	NSCLC	lung	male					
16 CDK12 17 C MAPK1	External resources		RERF-LC-AI ChaGo-K-1	NSCLC	lung	male					
18 D TPM3	canSAR		NCI-H1703	NSCLC	lung	male	-				
19 D NRAS	CaliSAN		NCI-H1703 NCI-H2228	NSCLC NSCLC	lung	female					
20 D NRG1	Ensembl	20		NSCLC	lung	male	-				
21 D PDGFRA	Enaction	20 0		NSCLC	lung	male					
22 D TERT	Human Protein Atlas		HCC2935	NSCLC	lung	male	-				
23 C TRIM33				NSCLC	lung	male					
24 CRLF2	Open Targets		NCI-H1975	NSCLC	lung	female					
25 🗆 FAM46C		25 🗆	NCI-H2291	NSCLC	lung	male					
26 G FANCD2	PubMed		NCI-H1563	NSCLC	lung	male					
27 GFGFR1			NCI-H2087	NSCLC	lung	male					
28 🗆 FIP1L1	UniProt		NCI-H1792	NSCLC	lung	male					
29 D BCL7A			NCI-H2342	NSCLC	lung	male					
0 D FUBP1				NSCLC	lung	male					
1 HMGA2				NSCLC	lung	male					
32 HNF1A 33 HOXC11			NCI-H522	NSCLC	lung	male					
33 U HOXCII 34 U HOXCI3				NSCLC	lung	female					
4 □ HOXCI3 5 □ IKBKB		34 0		NSCLC	lung	male					
6 D AR			ABC-1 NCI-H1650	NSCLC NSCLC	lung	male					
87 🗆 KDM5A				NSCLC	lung	male					
38 D BCB				NSCLC	lung	female	-				
19 C CCND1			NCI-H292 NCI-H2110	NSCLC	lung	Ternate					
			NCI-H1734	NSCLC	lung	female					
1 D LRIG3				NSCLC	lung	- I OTTAILO					
12 D LYL1			RERF-LC-KJ	NSCLC	lung	male					
43 CCNE1			NCI-H1755	NSCLC	lung	female					
44 🗆 AMER1			NCI-H2106	NSCLC	lung	male					
5 CD274		45 🗆	NCI-H650	NSCLC	lung	male				🚍 Data	
6 ARID2		46 🗆	NCI-H810	NSCLC	lung	male				E Data	
7 🗆 LM02			LC-1/sq-SF	NSCLC	lung	male					
18 🗆 MLLT6		48 🗆	NCI-H1693	NSCLC	lung	female				Selections	
49 🗆 MSN 👻			SK-MES-1	NSCLC	lung	male				Layout	
		50 0	HCC-15	NSCLC	luna	male			-	Analysis	

Fig. S 10. Filtering the search suggestions to properties of the active state and selecting the only gene *EGFR* with the identifier *ENSG00000146648* as search term results in two sequences where this gene was selected. When hovering over the search results the corresponding states in the provenance graph are highlighted.

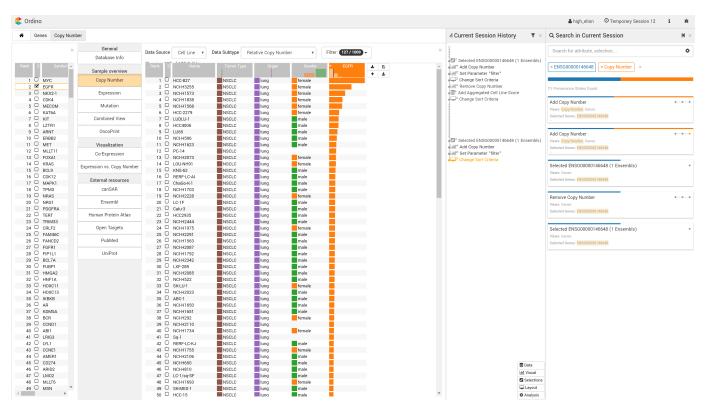


Fig. S 11. Refining the search with the *Copy Number* detail view results in five shorter sequences, where two of them are matching both search terms. The latter sequence contains the currently active state.

Genes Copy Num	nber					d Current Session History $\mathbf{T}$ ×	Q Search in Current Session
×	General	Data Source Cell Line v	Data Subtype Relative Copy Nu	mber v Filter: 58/10		× *	Search for attribute, selection,
	Database Info	Data Source Cell Line V	Helative Copy Nu	mber v Filter. 05710		Selected ENSG00000146648 (1 Ensembls)	
S Symbol 🔺		Rank S Name	Tumor Type Organ	Gender - EGFR	4 8	Selected ENSCOULOUT40048 (TEnsembls)	× ENSG00000146648 × Copy Number × breast carcinoma
	Sample overview				+ 1	• Info Set Parameter "filter"	
EGFR EGFR	Copy Number	1 D MDA-MB-468	breast carcino breast	female		📭 🖓 Change Sort Criteria	
MAF		2 D BT-20	breast carcino breast	female		e III * Remove Copy Number	12 Provenance States found
FBXW7	Expression	3 🗆 CAL-85-1	breast carcino breast	female		Add Aggregated Cell Line Score	
C CDH1		4 🗆 AU565	breast carcino breast	female		Change Sort Criteria	Set Parameter "filter"
FAM46C	Mutation	5 C SK-BR-3 6 C HCC1143	breast carcino breast	female			Views: Copy Number, Genes
TRIM33	Combined View	7 D HDQ-P1	breast carcino breast	female female		e dill+ Add Expression vs. Copy Number	Selected Genes: ENSG00000146648
C ATPIA1	Combined view	8 D BT-483	breast carcino breast breast carcino breast	female			
O NOTCH2	OncoPrint	9 C HCC70	breast carcino breast	female			Sequence of matching states
C ZFHX3	oncor mit	10 HCC1187	breast carcino breast	female			Set Parameter "filter"
ZMYM2	Visualization	11 CAL-148	breast carcino breast	female		Selected ENSG00000146648 (1 Ensembls)	Set Parameter Tilter
□ IL7R		12 MDA-MB-157	breast carcino breast	female		• HI+ Add Copy Number	Change Sort Criteria
BOS1	Co-Expression	13 MDA-MB-175-VII	breast carcino breast	female		e dillo Set Parameter "filter"	
FOXO3		14 D BT-549	breast carcino breast	female		Change Sort Criteria	
PRDM1	Expression vs. Copy Number	15 C HCC202	breast carcino breast	female			Add Copy Number
GOPC		16 D BT-474	breast carcino breast	female			Views: Copy Number, Genes
WHSC1	External resources	17 🗆 T-47D	breast carcino breast	female			Selected Genes: ENSG00000146648
MLLT10	canSAR	18 UACC-812	breast carcino breast	female			Selected Genes: ENSGUUUUT40048
CBFB		19 HCC1419	breast carcino breast	female			
NFE2L2	Ensembl	20 🗆 UACC-893	breast carcino 📕 breast	female			Add Copy Number +
HOXD13		21 HCC1937	breast carcino breast	female			Views: Copy Number, Genes
HOXD11	Human Protein Atlas	22 D HCC38	breast carcino 📕 breast	female			Selected Genes: ENSG00000146648
DROSHA		23 D MDA-MB-436	breast carcino breast	female			
SDHA	Open Targets	24 🗆 JIMT-1	breast carcino breast	female			
C TERT	PubMed	25 C HCC1428 26 Hs 578T	breast carcino breast				Selected ENSG00000146648 (1 Ensembls)
BCORL1	PubMed	26 U Hs 5781 27 U Hcc1395	breast carcino breast breast carcino breast	female female			Views: Genes
STAG2	UniProt	28 D DU4475	breast carcino breast	female			Selected Genes: ENSG00000146648
C ELF4	OTHEROC	28 C D04475 29 C HCC1954	breast carcino breast	female			
D PHF6		30 Hs 739.T	breast carcino breast	female			Bemove Copy Number
GPC3		31 D MDA-MB-361	breast carcino breast	female			Remove Copy Number   Views: Genes
MITE		32 MDA-MB-231	breast carcino breast	female			
G FOXP1		33 HCC1500	breast carcino breast	female			Selected Genes: ENSG00000146648
SBDS		34 D HCC2218	breast carcino breast	female			
C ABI1		35 🗆 CAMA-1	breast carcino breast	female			Selected ENSG00000146648 (1 Ensembls)
C KIF5B		36 🗆 Hs 274.T	breast carcino 📕 breast	female			Views: Genes
ERCC5		37 🗆 ZR-75-30	breast carcino 📕 breast	female			Selected Genes: ENSG00000146648
RBM15		38 CAL-120	breast carcino 📕 breast	female			
CDH11		39 🗆 CAL-51	breast carcino 📕 breast	female			
CBFA2T3		40 EFM-19	breast carcino 📕 breast	female			Add Expression vs. Copy Number
SF3B1		41 Hs 742.T	breast carcino breast	female			Views: Genes, Expression vs. Copy Number
CASP8		42 Hs 281.T	breast carcino breast	female			Selected Genes: ENSG00000141736, ENSG00000167258, ENSG00000182197
D PDE4DIP		43 MDA-MB-453	breast carcino breast	female			
SRC SDC4		44 C HCC1569 45 HCC1806	breast carcino breast	female			
D SDC4		45 U HCC1806 46 U MDA-MB-415	breast carcino breast	female		🛢 Data	
D TOP1		46 U MDA-MB-415 47 U MDA-MB-134-VI	breast carcino breast	female		Lat Visual	
D PLCG1		47 U MDA-MB-134-VI 48 U Hs 606.T	breast carcino breast breast carcino breast	female		Selections	
D SALL4		48 □ Hs 606.T 49 □ ZR-75-1	breast carcino breast	female female		Layout	
SALL4		49 U 2H-75-1 50 U Hs 343.T	breast carcino breast	female		<ul> <li>Analysis</li> </ul>	

Fig. S 12. Refining the search results with *breast cancer* tumor type as third search term shows that only one sequence matches all search terms. Expanding the sequence and jumping to the last state shows that the *EGFR* copy number was assessed for *breast cancer* cell lines with *MDA-MB-468* as the top hit.

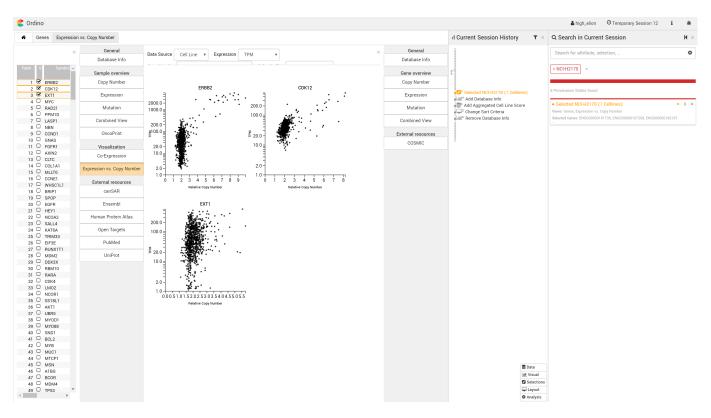


Fig. S 13. The user searches for the cell line *NCI-H2170* to recall if that cell line was used in the analysis. Indeed, the cell line was found in one state sequence and was selected in the *Expression vs. Copy Number* detail view.

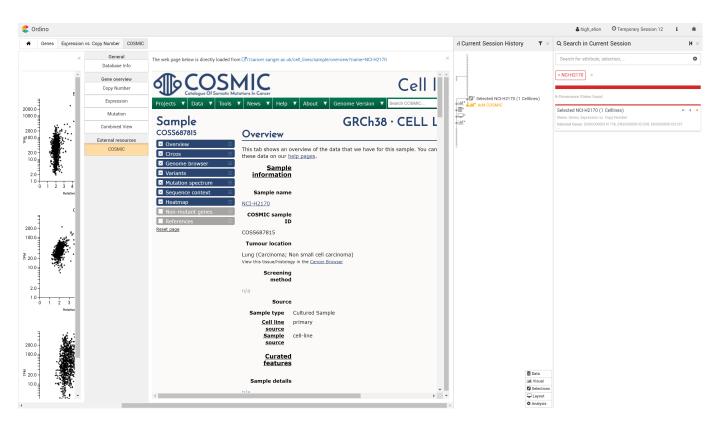


Fig. S 14. Following up with the analysis of cell line NCI-H2170 the user opens the COSMIC detail view and browses the information that is available about this cell line in the COSMIC data base.