

# 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.

```
"signals": [{  
  "name": "xField",  
  "value": "gdp",  
  "bind": {  
    "input": "select",  
    "options": ["gdp", "population", "life_expect",  
               "fertility", "child_mortality"]  
  },  
  "track": {  
    "title": "X = {{value}}",  
    "category": "data",  
    "operation": "update"  
  },  
  "search": {  
    "type": "category",  
    "title": "{{value}}",  
    "group": "Attributes"  
  }  
}]
```

## 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.



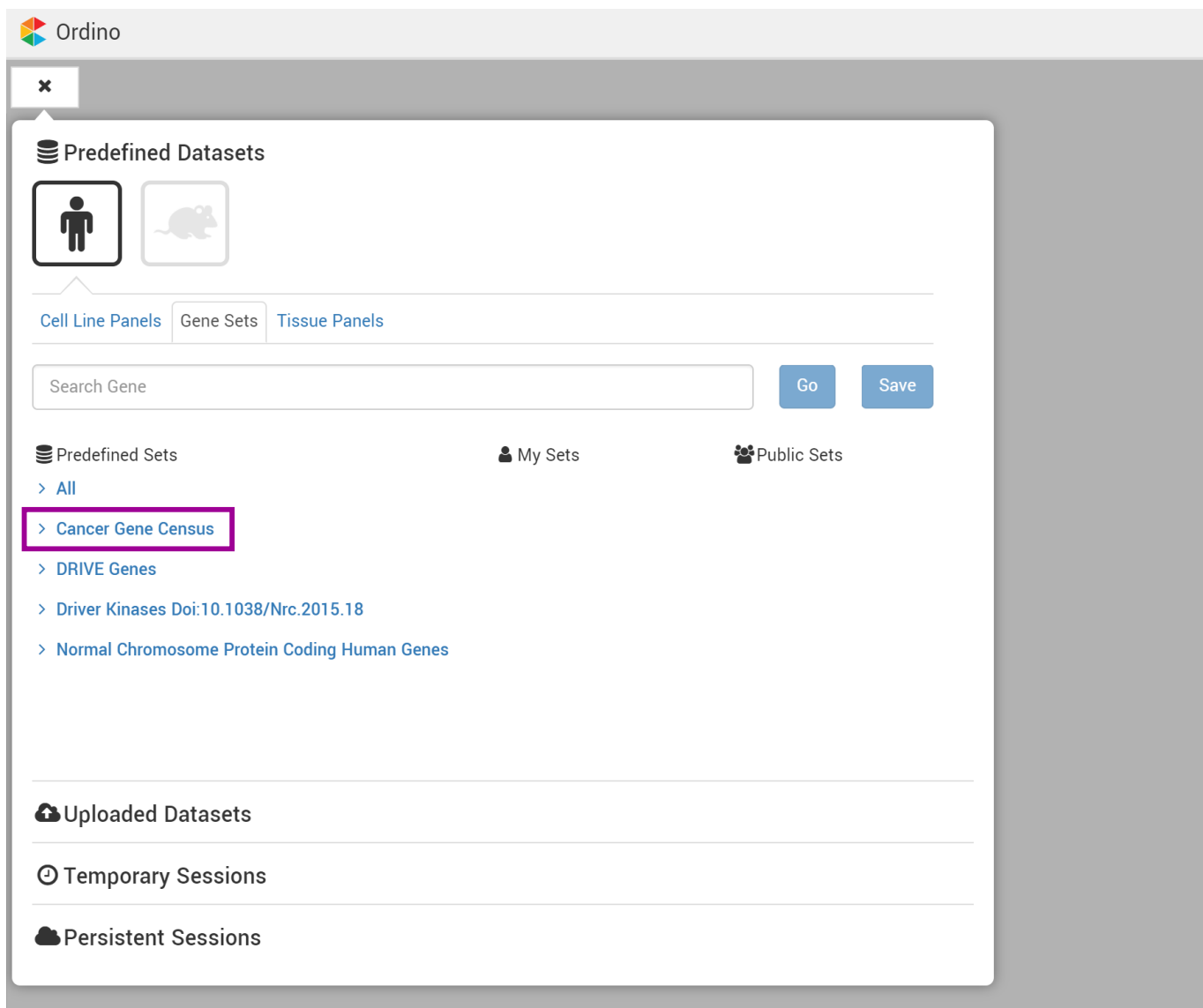


Fig. S 1. The user is interested in the breast cancer cell line *BT-20* and wants to know which cancer gene(s) could be important for growth in that cell line? In Ordino, users can choose between predefined datasets or upload custom datasets. For the case study the user starts with the *Cancer Gene Census*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

Ordino

Genes

Showing 563 of 563 Genes

Rank	S	Symbol	Ensembl	Name	Chromosome	Biotype	TPM of BT-20	Relative Copy Number of BT-20		
1	<input type="checkbox"/>	EGFR	ENSG00000146648	epidermal growth factor receptor [S	7	protein_coding				
2	<input type="checkbox"/>	MAF	ENSG00000178573	v-maf musculoaponeurotic fibrosar	16	protein_coding				
3	<input type="checkbox"/>	FBXW7	ENSG00000109670	F-box and WD repeat domain contai	4	protein_coding				
4	<input type="checkbox"/>	CDH1	ENSG00000039068	cadherin 1, type 1, E-cadherin (epith	16	protein_coding				
5	<input type="checkbox"/>	NRAS	ENSG00000213281	neuroblastoma RAS viral (v-ras) on	1	protein_coding				
6	<input type="checkbox"/>	FAM46C	ENSG00000183508	family with sequence similarity 46, i	1	protein_coding				
7	<input type="checkbox"/>	TRIM33	ENSG00000197323	tripartite motif containing 33 [Sourc	1	protein_coding				
8	<input type="checkbox"/>	ATP1A1	ENSG00000163399	ATPase, Na+/K+ transporting, alpha	1	protein_coding				
9	<input type="checkbox"/>	NOTCH2	ENSG00000134250	notch 2 [Source:HGNC Symbol;Acc:	1	protein_coding				
10	<input type="checkbox"/>	ZFXH3	ENSG00000140836	zinc finger homeobox 3 [Source:HG	16	protein_coding				
11	<input type="checkbox"/>	ZMYM2	ENSG00000121741	zinc finger, MYM-type 2 [Source:HG	13	protein_coding				
12	<input type="checkbox"/>	IL7R	ENSG00000168685	interleukin 7 receptor [Source:HGNC	5	protein_coding				
13	<input type="checkbox"/>	ROS1	ENSG00000047936	c-ros oncogene 1, receptor tyrosine	6	protein_coding				
14	<input type="checkbox"/>	FOXO3	ENSG00000118689	forkhead box O3 [Source:HGNC Syn	6	protein_coding				
15	<input type="checkbox"/>	PRDM1	ENSG00000057657	PR domain containing 1, with ZNF c	6	protein_coding				
16	<input type="checkbox"/>	GOPC	ENSG00000047932	golgi-associated PDZ and coiled-coi	6	protein_coding				
17	<input type="checkbox"/>	WHSC1	ENSG00000109685	Wolf-Hirschhorn syndrome candida	4	protein_coding				
18	<input type="checkbox"/>	MLLT10	ENSG000000078403	myeloid/lymphoid or mixed-lineage	10	protein_coding				
19	<input type="checkbox"/>	CBFB	ENSG000000067955	core-binding factor, beta subunit [S	16	protein_coding				
20	<input type="checkbox"/>	NFE2L2	ENSG00000116044	nuclear factor (erythroid-derived 2)-	2	protein_coding				
21	<input type="checkbox"/>	HOXD13	ENSG00000128714	homeobox D13 [Source:HGNC Sym]	2	protein_coding				
22	<input type="checkbox"/>	HOXD11	ENSG00000128713	homeobox D11 [Source:HGNC Sym]	2	protein_coding				
23	<input type="checkbox"/>	DROSHA	ENSG00000113360	drosha, ribonuclease type III [Sourc	5	protein_coding				
24	<input type="checkbox"/>	SDHA	ENSG000000073578	succinate dehydrogenase complex,	5	protein_coding				
25	<input type="checkbox"/>	TERT	ENSG00000164362	telomerase reverse transcriptase [S	5	protein_coding				
26	<input type="checkbox"/>	CTCF	ENSG00000102974	CCCTC-binding factor (zinc finger p	16	protein_coding				
27	<input type="checkbox"/>	BCORL1	ENSG000000085185	BCL6 corepressor-like 1 [Source:HG	X	protein_coding				
28	<input type="checkbox"/>	STAG2	ENSG00000101972	stromal antigen 2 [Source:HGNC Sy	X	protein_coding				
29	<input type="checkbox"/>	ELF4	ENSG00000102034	E74-like factor 4 (ets domain transc	X	protein_coding				
30	<input type="checkbox"/>	PHF6	ENSG00000156531	PHD finger protein 6 [Source:HGNC	X	protein_coding				
31	<input type="checkbox"/>	GPC3	ENSG00000147257	glypican 3 [Source:HGNC Symbol;A	X	protein_coding				
32	<input type="checkbox"/>	MITF	ENSG00000187098	microphthalmia-associated transcri	3	protein_coding				
33	<input type="checkbox"/>	FOXP1	ENSG00000114861	forkhead box P1 [Source:HGNC Syn	3	protein_coding				
34	<input type="checkbox"/>	SBDS	ENSG00000126524	Shwachman-Bodian-Diamond syndr	7	protein_coding				
35	<input type="checkbox"/>	ABI1	ENSG00000136754	abl-interactor 1 [Source:HGNC Sym]	10	protein_coding				
36	<input type="checkbox"/>	KIF5B	ENSG00000170759	kinesin family member 5B [Source:!	10	protein_coding				
37	<input type="checkbox"/>	ERCC5	ENSG00000134899	excision repair cross-complementin	13	protein_coding				
38	<input type="checkbox"/>	RBM15	ENSG00000162775	RNA binding motif protein 15 [Sour	1	protein_coding				
39	<input type="checkbox"/>	CDH11	ENSG00000140937	cadherin 11, type 2, OB-cadherin (o	16	protein_coding				
40	<input type="checkbox"/>	CBFA2T3	ENSG00000129993	core-binding factor, runt domain, al	16	protein_coding				
41	<input type="checkbox"/>	SF3B1	ENSG00000115524	splicing factor 3b, subunit 1, 155kD	2	protein_coding				
42	<input type="checkbox"/>	CASP8	ENSG00000064012	caspase 8, apoptosis-related cystei	2	protein_coding				
43	<input type="checkbox"/>	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*.

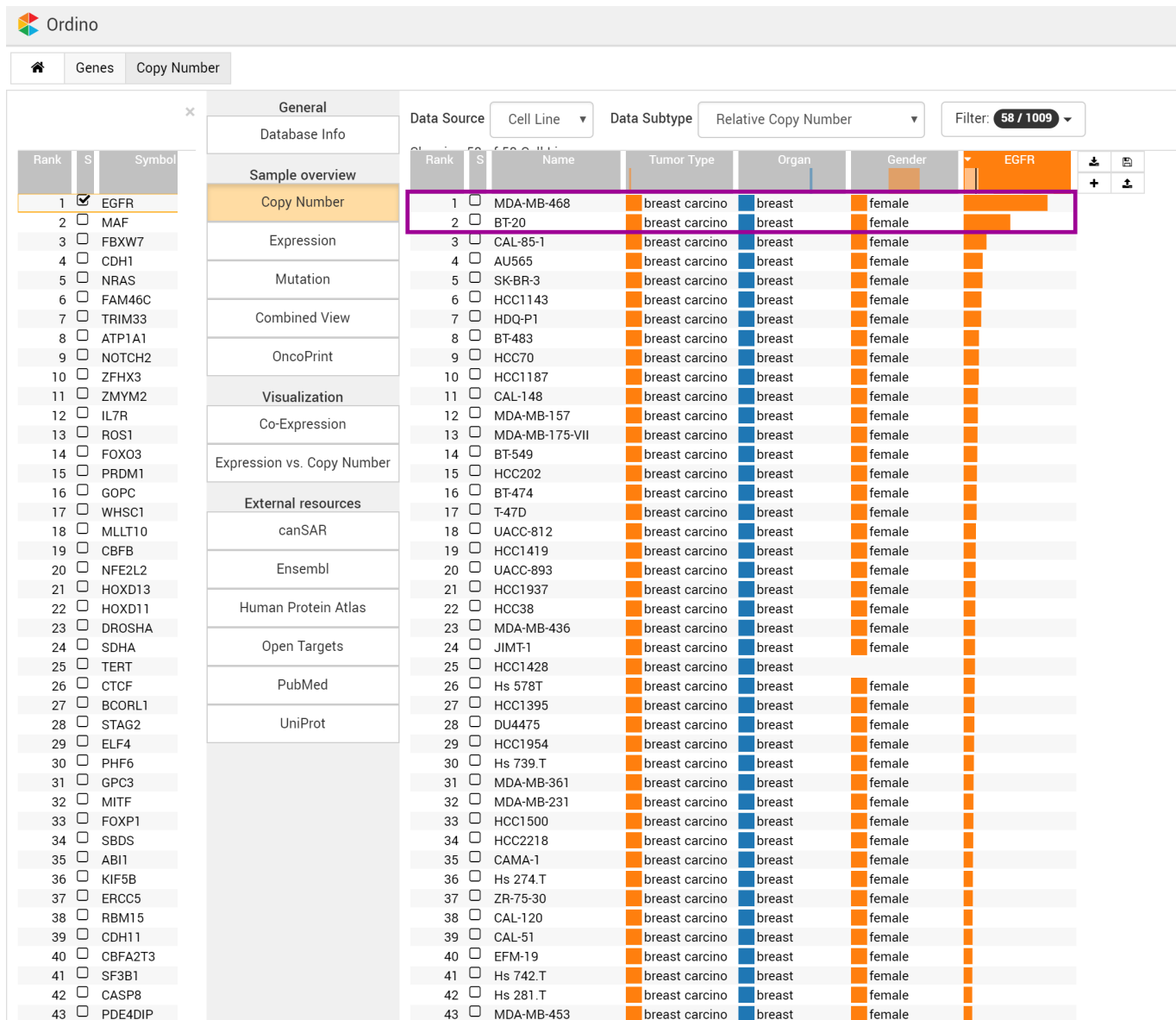


Fig. S 3. Next, the user wants to know, if there are other breast cancer cell lines with *EGFR* amplification? Selecting the gene *EGFR* from the ranking brings up a list of further detail views. The user selects the *Copy Number* view, filters the ranking by tumor type *breast cancer*, and sorts it 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										
Showing 563 of 563 Genes; 1 selected										
Rank	S	Symbol	Ensembl	Name	Chromosome	Biotype	TPM of BT-20	Relative Copy Number of BT-20	Relative Copy Number > 4	
1	<input type="checkbox"/>	ERBB2	ENSG00000141736	v-erb-b2 erythroblastic leukemia viri	17	protein_coding				
2	<input type="checkbox"/>	CDK12	ENSG00000167258	cyclin-dependent kinase 12 [Source	17	protein_coding				
3	<input type="checkbox"/>	EXT1	ENSG00000182197	exostosin 1 [Source:HGNC Symbol;	8	protein_coding				
4	<input type="checkbox"/>	MYC	ENSG00000136997	v-myc myelocytomatosis viral onco	8	protein_coding				
5	<input type="checkbox"/>	RAD21	ENSG00000164754	RAD21 homolog (S. pombe) [Sourc	8	protein_coding				
6	<input type="checkbox"/>	PPM1D	ENSG00000170836	protein phosphatase, Mg2+/Mn2+ d	17	protein_coding				
7	<input type="checkbox"/>	LASP1	ENSG00000002834	LIM and SH3 protein 1 [Source:HG	17	protein_coding				
8	<input type="checkbox"/>	NBN	ENSG00000104320	nibrin [Source:HGNC Symbol;Acc:7	8	protein_coding				
9	<input type="checkbox"/>	CCND1	ENSG00000110092	cyclin D1 [Source:HGNC Symbol;Ac	11	protein_coding				
10	<input type="checkbox"/>	GNAS	ENSG000000087460	GNAS complex locus [Source:HGNC	20	protein_coding				
11	<input type="checkbox"/>	FGFR1	ENSG00000077782	fibroblast growth factor receptor 1	8	protein_coding				
12	<input type="checkbox"/>	AXIN2	ENSG00000168646	axin 2 [Source:HGNC Symbol;Acc:9	17	protein_coding				
13	<input type="checkbox"/>	CLTC	ENSG00000141367	clathrin, heavy chain (Hc) [Source:	17	protein_coding				
14	<input type="checkbox"/>	COL1A1	ENSG00000108821	collagen, type I, alpha 1 [Source:HG	17	protein_coding				
15	<input type="checkbox"/>	MLLT6	ENSG00000108292	myeloid/lymphoid or mixed-lineage	17	protein_coding				
16	<input type="checkbox"/>	CCNE1	ENSG00000105173	cyclin E1 [Source:HGNC Symbol;Ac	19	protein_coding				
17	<input type="checkbox"/>	WHSC1L1	ENSG00000147548	Wolf-Hirschhorn syndrome candidat	8	protein_coding				
18	<input type="checkbox"/>	BRIP1	ENSG00000136492	BRCA1 interacting protein C-termin	17	protein_coding				
19	<input type="checkbox"/>	SPOP	ENSG00000121067	speckle-type POZ protein [Source:H	17	protein_coding				
20	<input checked="" type="checkbox"/>	EGFR	ENSG00000146648	epidermal growth factor receptor [S	7	protein_coding	323	5.03	0.0351	
21	<input type="checkbox"/>	HEY1	ENSG00000164683	hairy/enhancer-of-split related with	8	protein_coding				
22	<input type="checkbox"/>	NCOA2	ENSG00000140396	nuclear receptor coactivator 2 [Sou	8	protein_coding				
23	<input type="checkbox"/>	SALL4	ENSG00000101115	sal-like 4 (Drosophila) [Source:HG	20	protein_coding				
24	<input type="checkbox"/>	KAT6A	ENSG000000083168	K(lysine) acetyltransferase 6A [Sou	8	protein_coding				
25	<input type="checkbox"/>	TRIM33	ENSG00000197323	tripartite motif containing 33 [Sour	1	protein_coding				
26	<input type="checkbox"/>	EIF3E	ENSG00000104408	eukaryotic translation initiation fact	8	protein_coding				
27	<input type="checkbox"/>	RUNX1T1	ENSG000000079102	runt-related transcription factor 1; t	8	protein_coding				
28	<input type="checkbox"/>	MDM2	ENSG00000135679	Mdm2, p53 E3 ubiquitin protein liga	12	protein_coding				
29	<input type="checkbox"/>	DDX3X	ENSG00000215301	DEAD (Asp-Glu-Ala-Asp) box polype	X	protein_coding				
30	<input type="checkbox"/>	RBM10	ENSG00000182872	RNA binding motif protein 10 [Sour	X	protein_coding				
31	<input type="checkbox"/>	RARA	ENSG00000131759	retinoic acid receptor, alpha [Sourc	17	protein_coding				
32	<input type="checkbox"/>	CDK4	ENSG00000135446	cyclin-dependent kinase 4 [Source:	12	protein_coding				
33	<input type="checkbox"/>	LMO2	ENSG00000135363	LIM domain only 2 (rhombotin-like	11	protein_coding				
34	<input type="checkbox"/>	NCOR1	ENSG00000141027	nuclear receptor corepressor 1 [Sou	17	protein_coding				
35	<input type="checkbox"/>	SS18L1	ENSG00000184402	synovial sarcoma translocation gen	20	protein_coding				
36	<input type="checkbox"/>	AKT1	ENSG00000142208	v-akt murine thymoma viral oncogen	14	protein_coding				
37	<input type="checkbox"/>	UBR5	ENSG00000104517	ubiquitin protein ligase E3 compone	8	protein_coding				
38	<input type="checkbox"/>	MYOD1	ENSG00000129152	myogenic differentiation 1 [Source:	11	protein_coding				
39	<input type="checkbox"/>	MYD88	ENSG00000172936	myeloid differentiation primary resp	3	protein_coding				
40	<input type="checkbox"/>	SND1	ENSG00000197157	staphylococcal nuclease and tudor	7	protein_coding				
41	<input type="checkbox"/>	BCL2	ENSG00000171791	B-cell CLL/lymphoma 2 [Source:HG	18	protein_coding				
42	<input type="checkbox"/>	MYB	ENSG00000118513	v-myb myeloblastosis viral oncogen	6	protein_coding				
43	<input type="checkbox"/>	MUC1	ENSG00000185499	mucin 1, cell surface associated [S	1	protein_coding				

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.

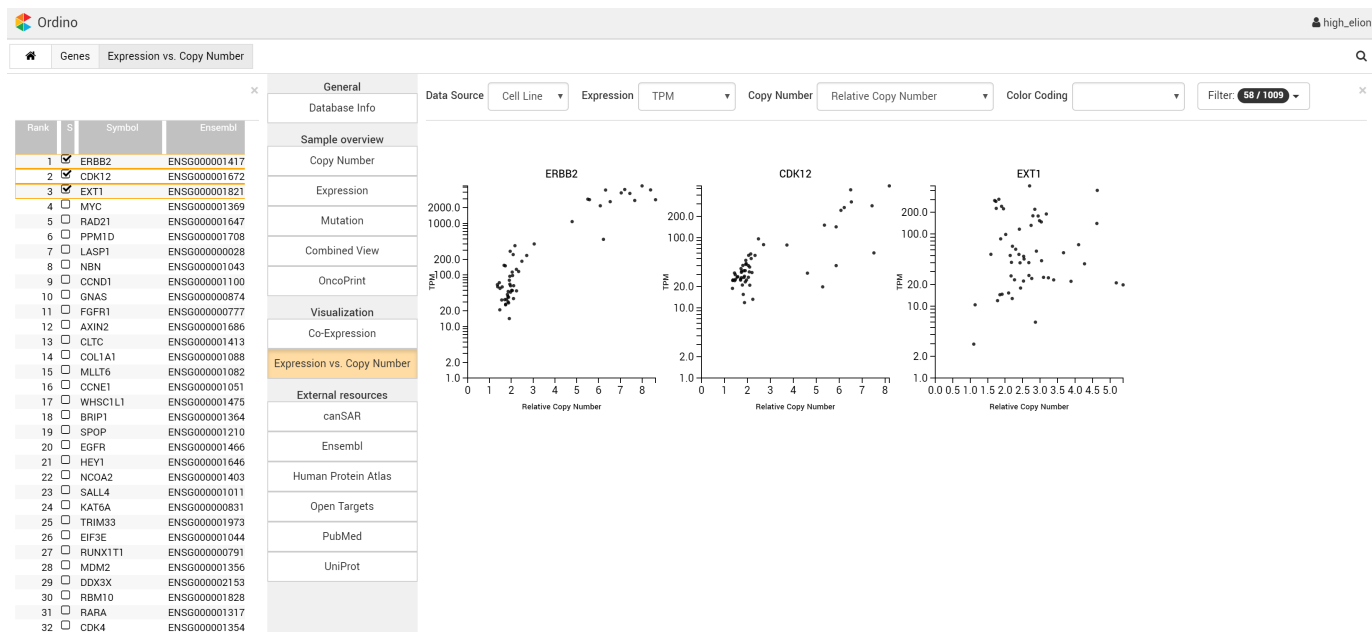


Fig. S 5. Next the user wants to know, if the amplifications of the top genes also lead to a high gene expression? Selecting the first three genes (*ERBB2*, *CDK12*, and *EXT1*) and opening the detail view *Expression vs. Copy Number* (still filtered for *breast cancer* cell lines) shows a scatterplot 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*.

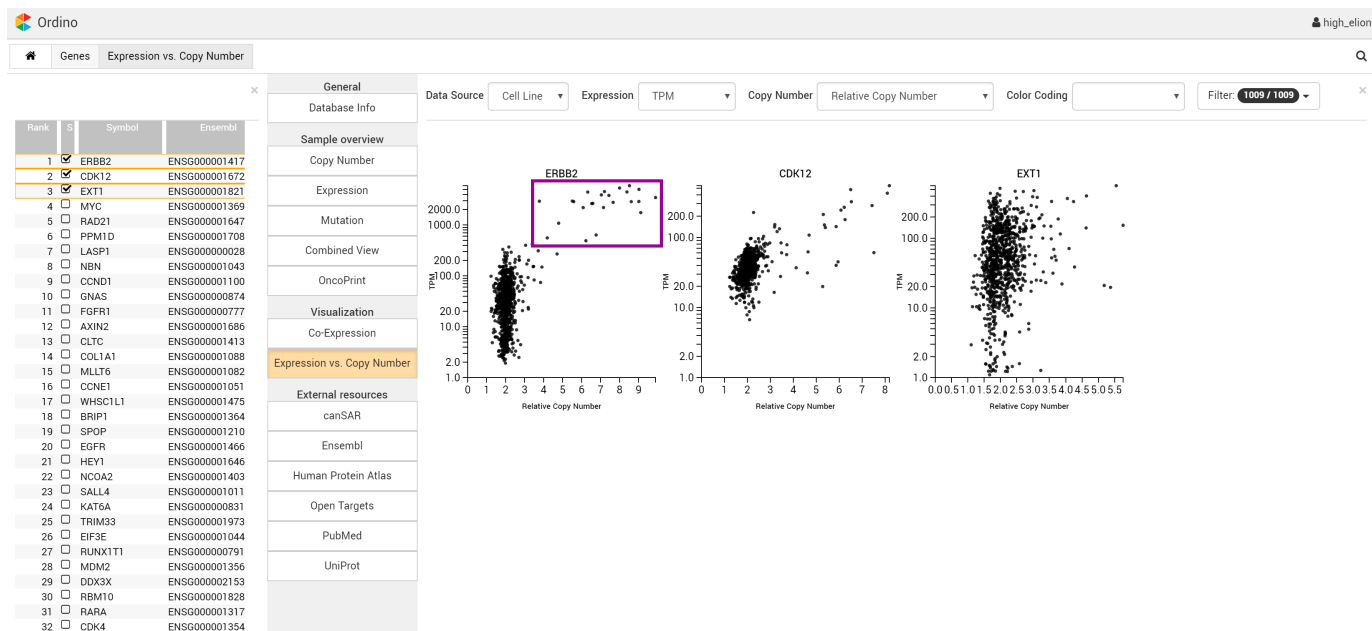


Fig. S 6. The user removes the filter to explore if the expression to copy number correlation is also present in other tumor types. The result is that all cell lines with a very high *ERBB2* copy number have high *ERBB2* expression.

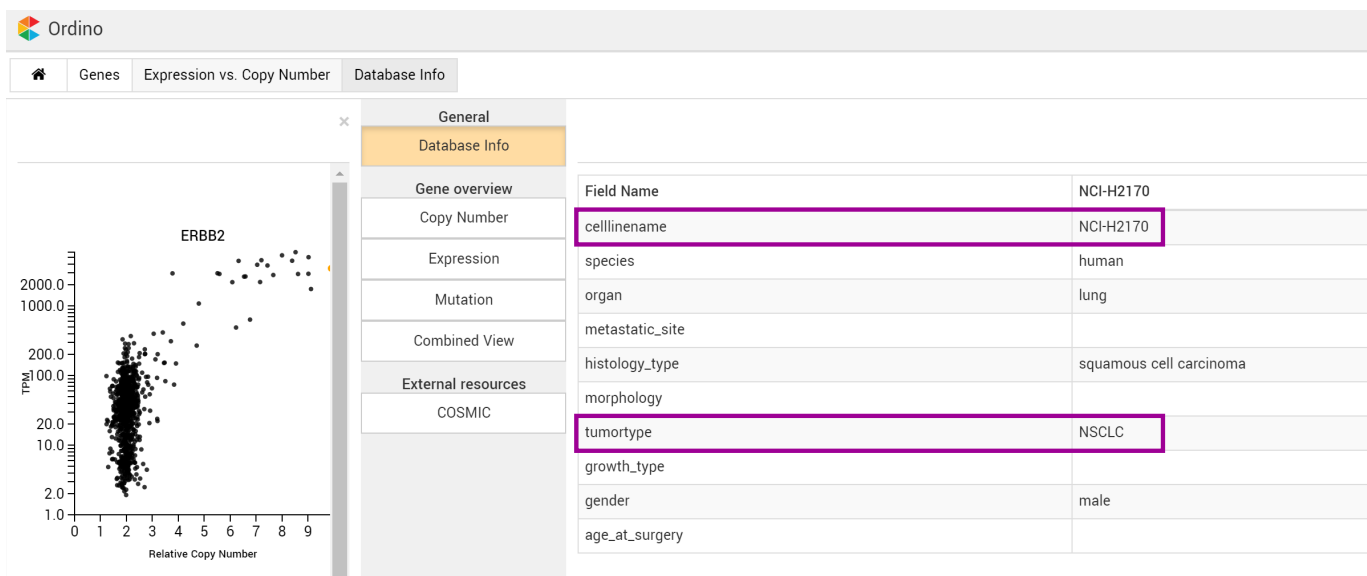


Fig. S 7. Selecting the most right point in the *ERBB2* scatterplot and opening the *Database Info* detail view, reveals that the cell line with the highest *ERBB2* amplification is *NCI-H2170* and belongs to *non-small-cell lung cancer* (NSCLC).



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*.

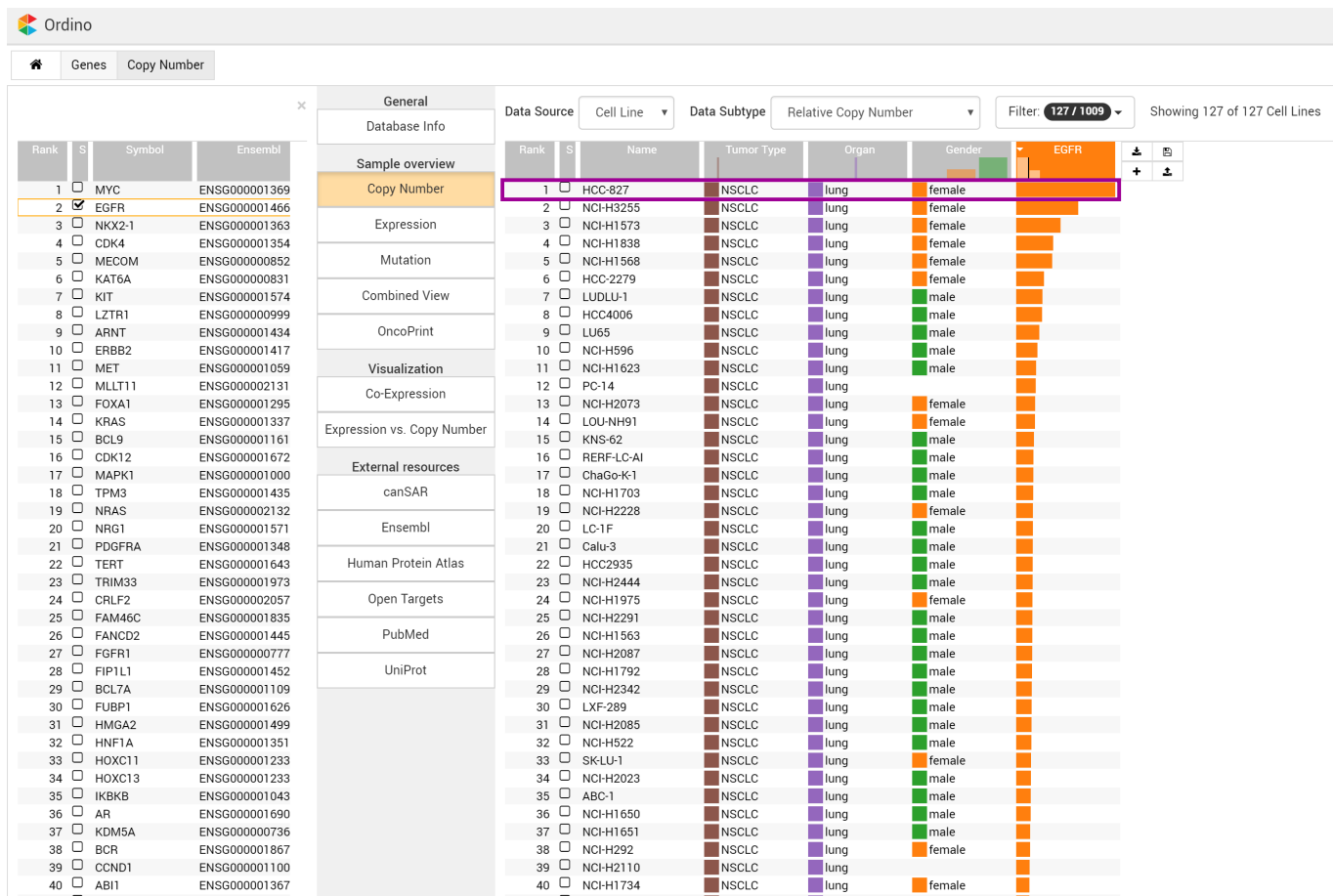


Fig. S 9. The user is still interested in *EGFR* and wants to know which *NSCLC* cell lines have an *EGFR* amplification? He selects the gene *EGFR* from the ranking and opens the *Copy Number* detail view. The copy number value for *EGFR* is added as additional column to the following cell line ranking.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

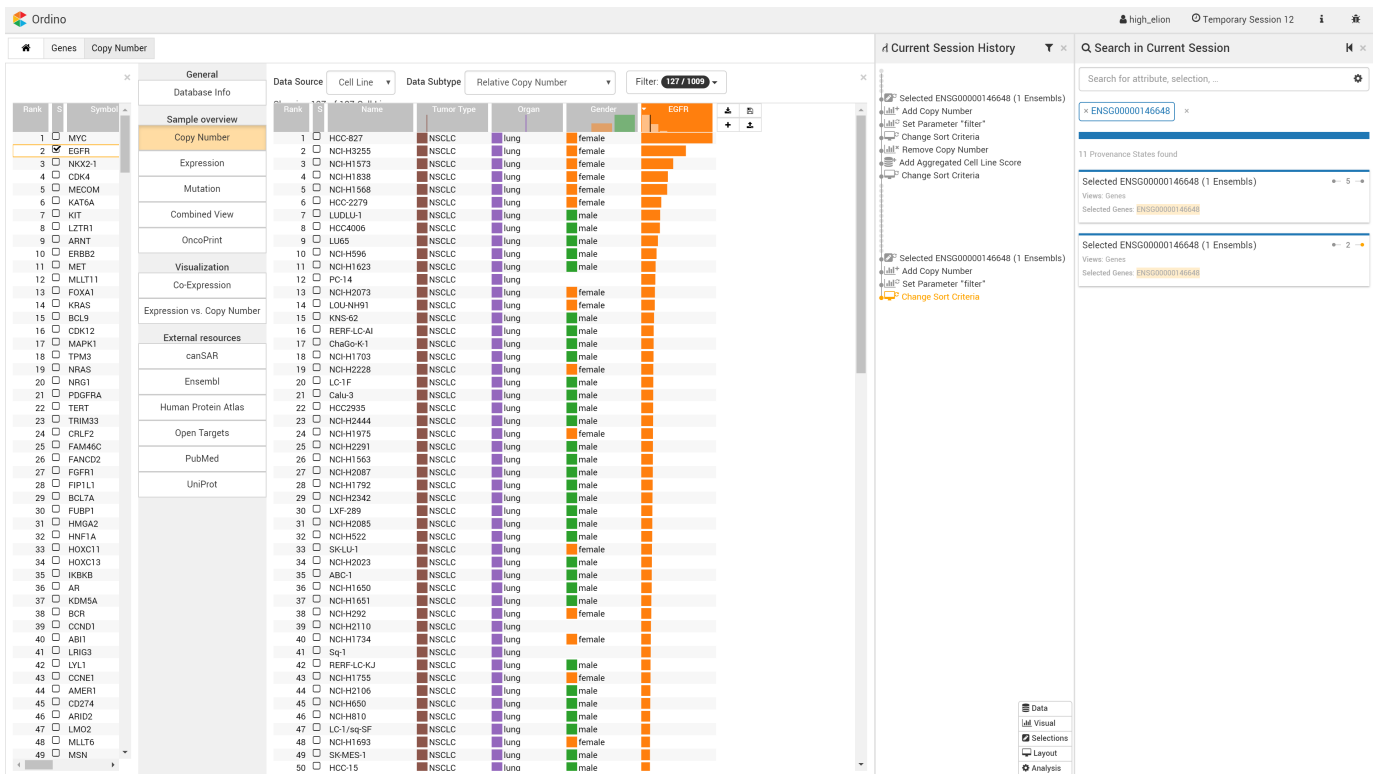


Fig. S 10. Filtering the search suggestions to properties of the active state and selecting the only gene *EGFR* with the identifier *ENSG00000146648* as search term results in two sequences where this gene was selected. When hovering over the search results the corresponding states in the provenance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

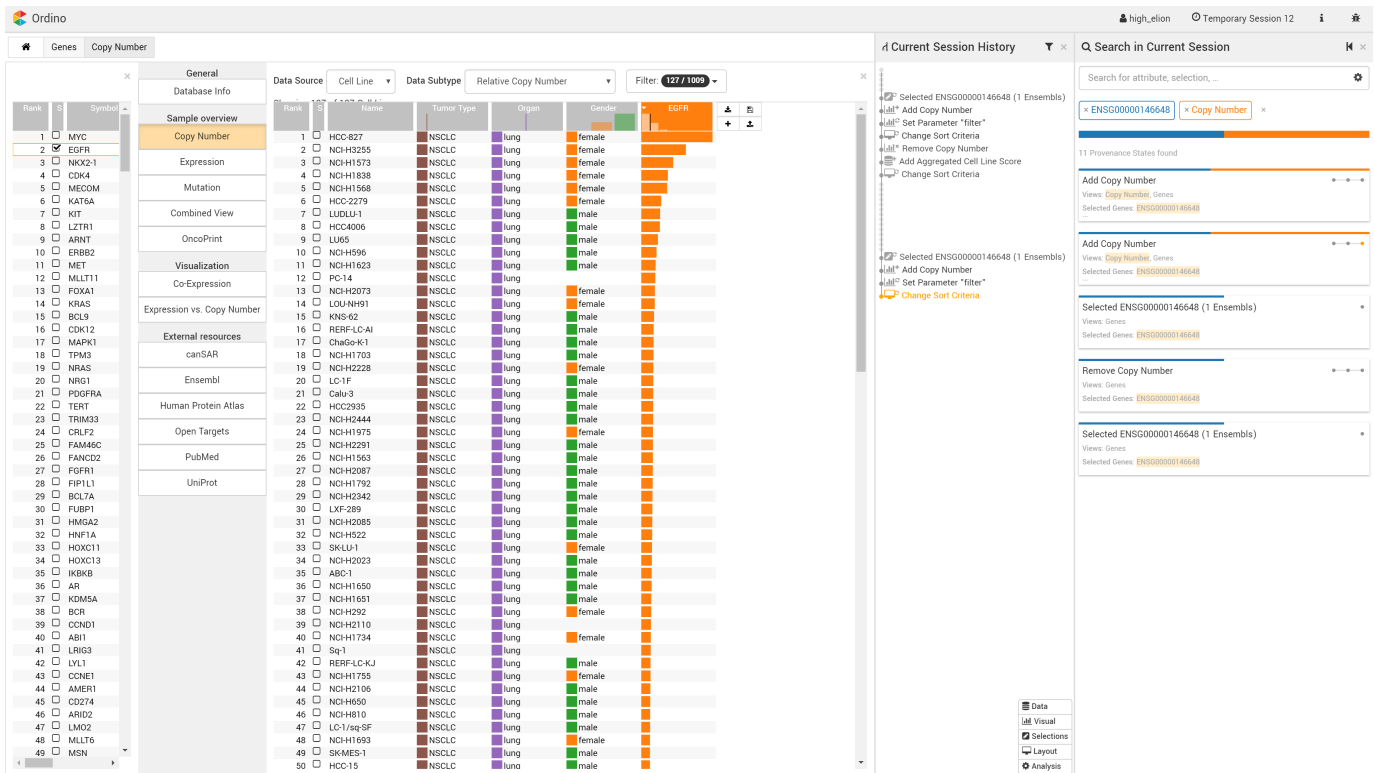


Fig. S 11. Refining the search with the *Copy Number* detail view results in five shorter sequences, where two of them are matching both search terms. The latter sequence contains the currently active state.

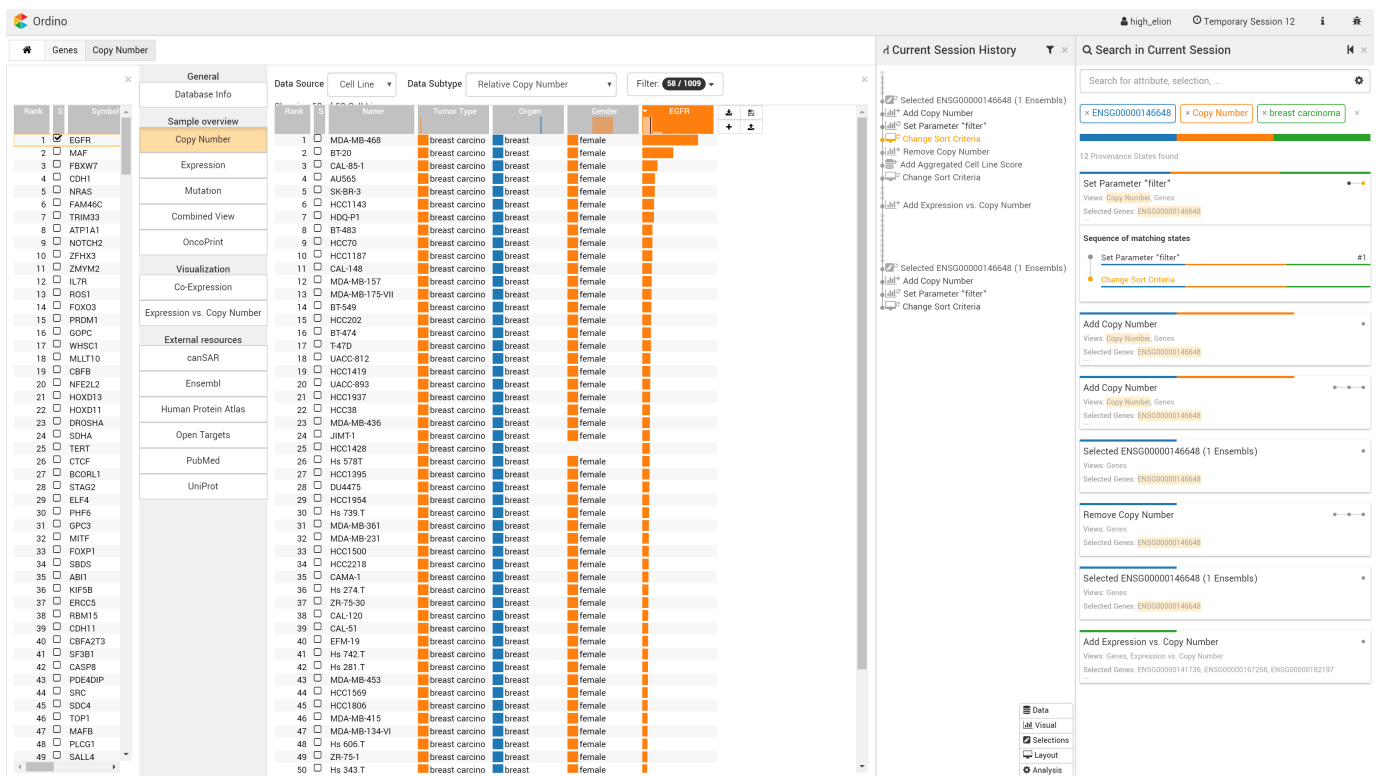


Fig. S 12. Refining the search results with *breast cancer* tumor type as third search term shows that only one sequence matches all search terms. Expanding the sequence and jumping to the last state shows that the *EGFR* copy number was assessed for *breast cancer* cell lines with *MDA-MB-468* as the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

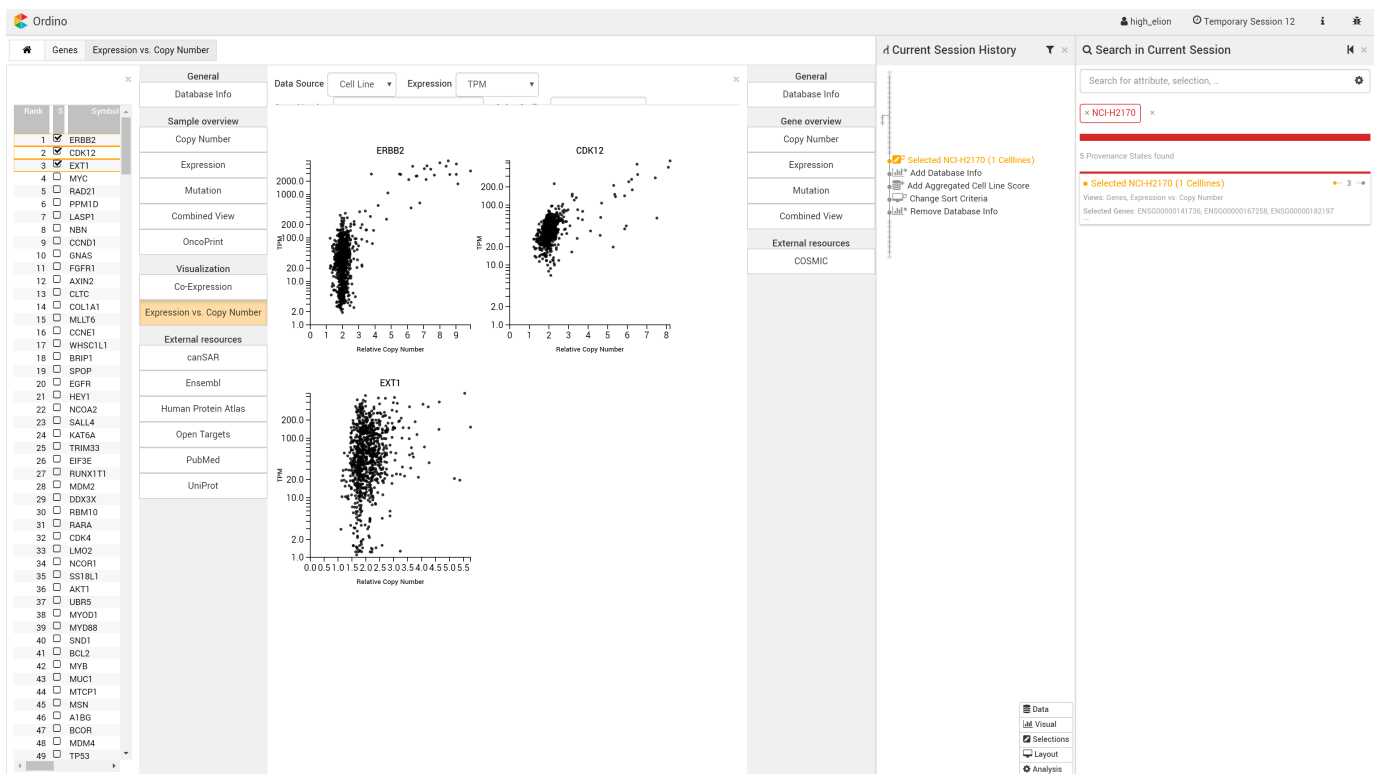


Fig. S 13. The user searches for the cell line *NCI-H2170* to recall if that cell line was used in the analysis. Indeed, the cell line was found in one state sequence and was selected in the *Expression vs. Copy Number* detail view.

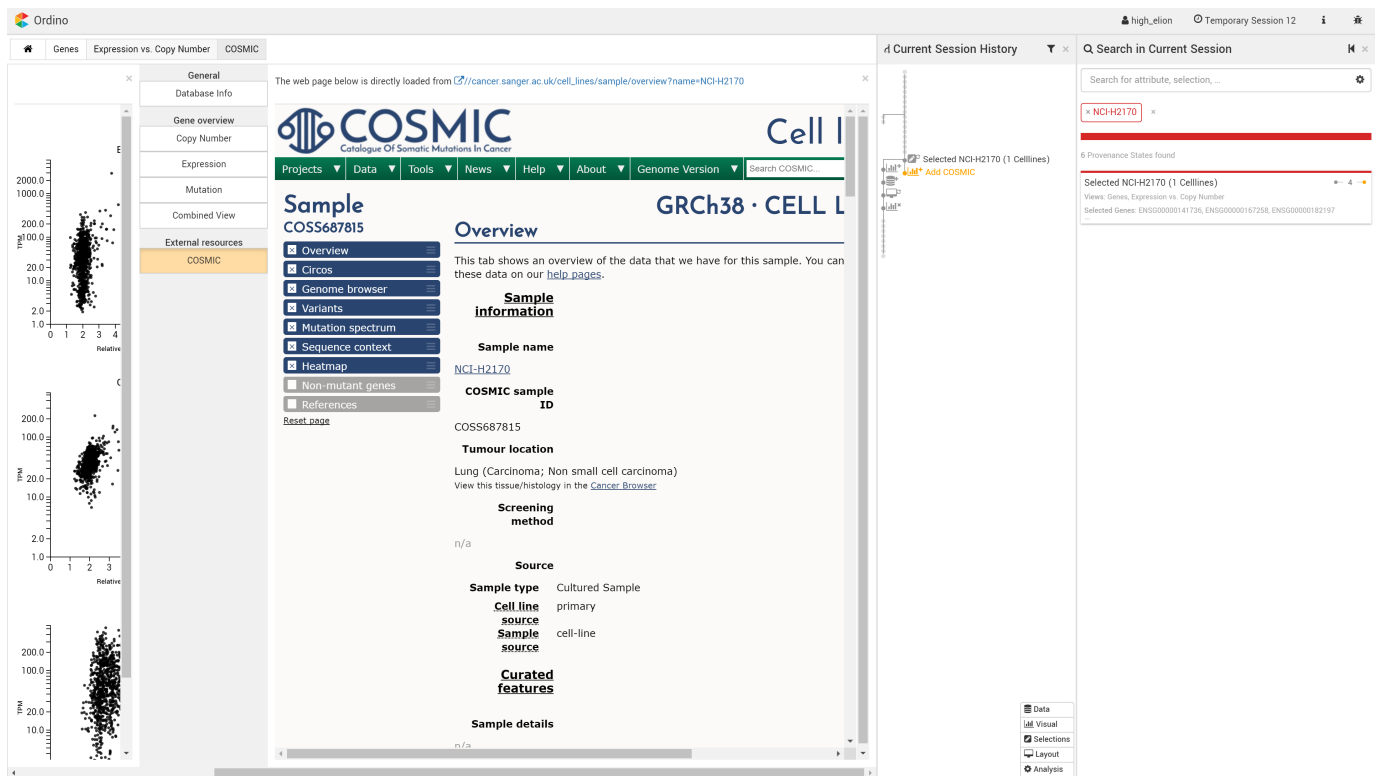


Fig. S 14. Following up with the analysis of cell line *NCI-H2170* the user opens the *COSMIC* detail view and browses the information that is available about this cell line in the *COSMIC* data base.