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.

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

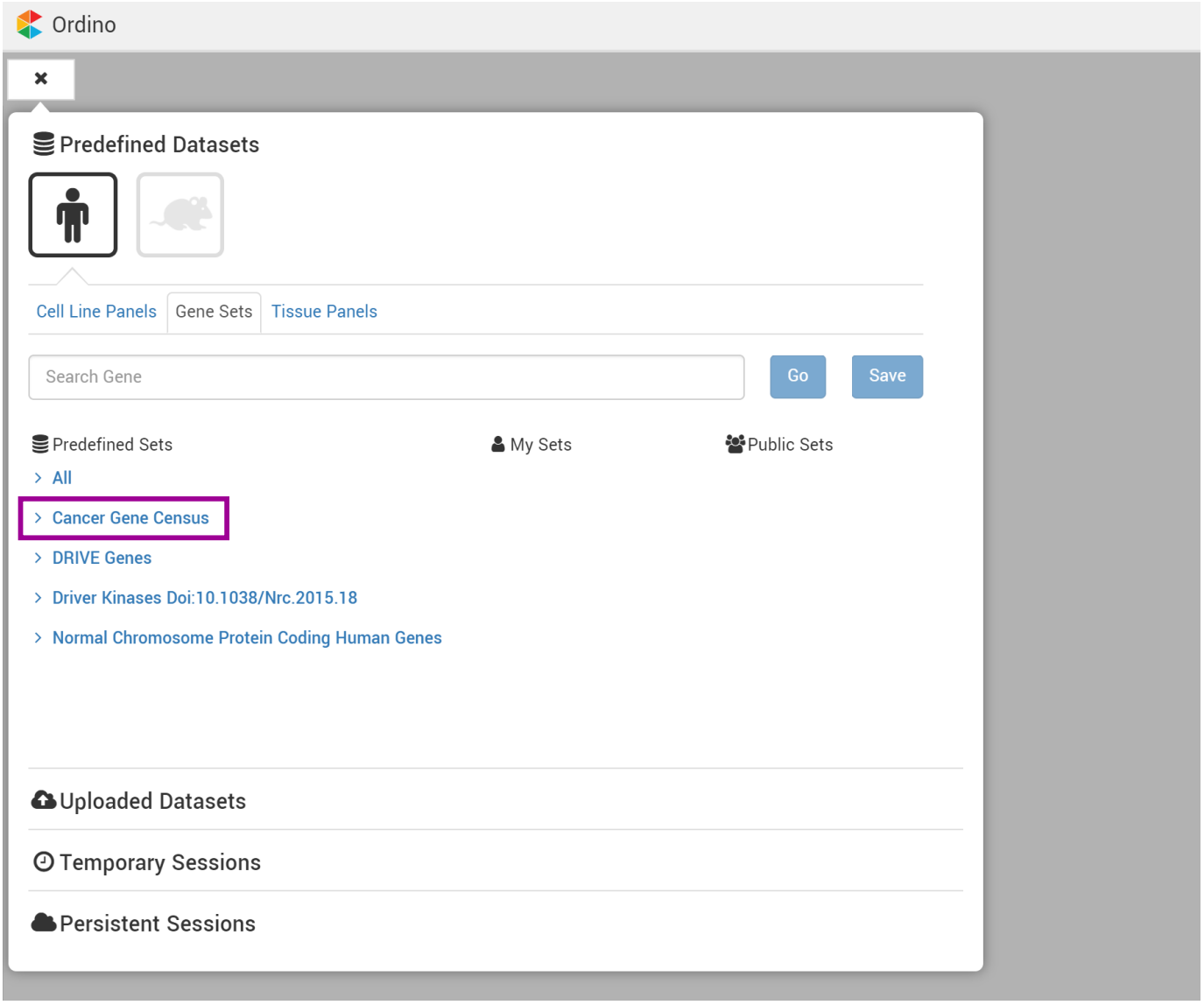

Fig. S 1. The user is interested in the breast cancer cell line *BT-20* and wants to know which cancer gene(s) could be important for growth in that cell line? In Ordino, users can choose between predefined datasets or upload custom datasets. For the case study the user starts with the *Cancer Gene Census*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

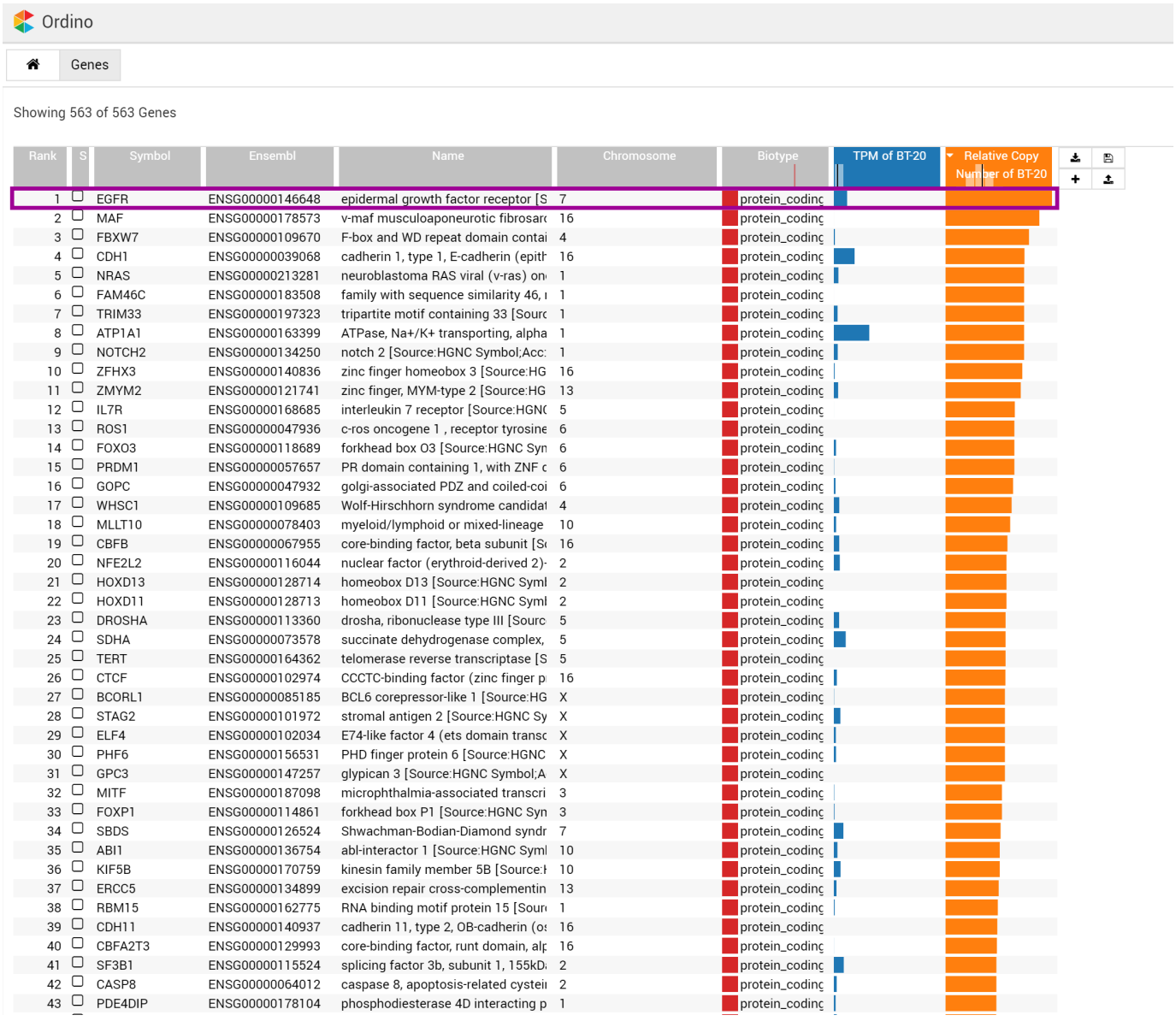

Fig. S 2. Adding the two score columns *TPM* (gene expression level) and *Relative Copy Number* for the cell line *BT-20* and sorting the ranking by the relative copy number column reveals the gene *EGFR* as an amplified and also highly expressed candidate. Likely, it is important for cell growth in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

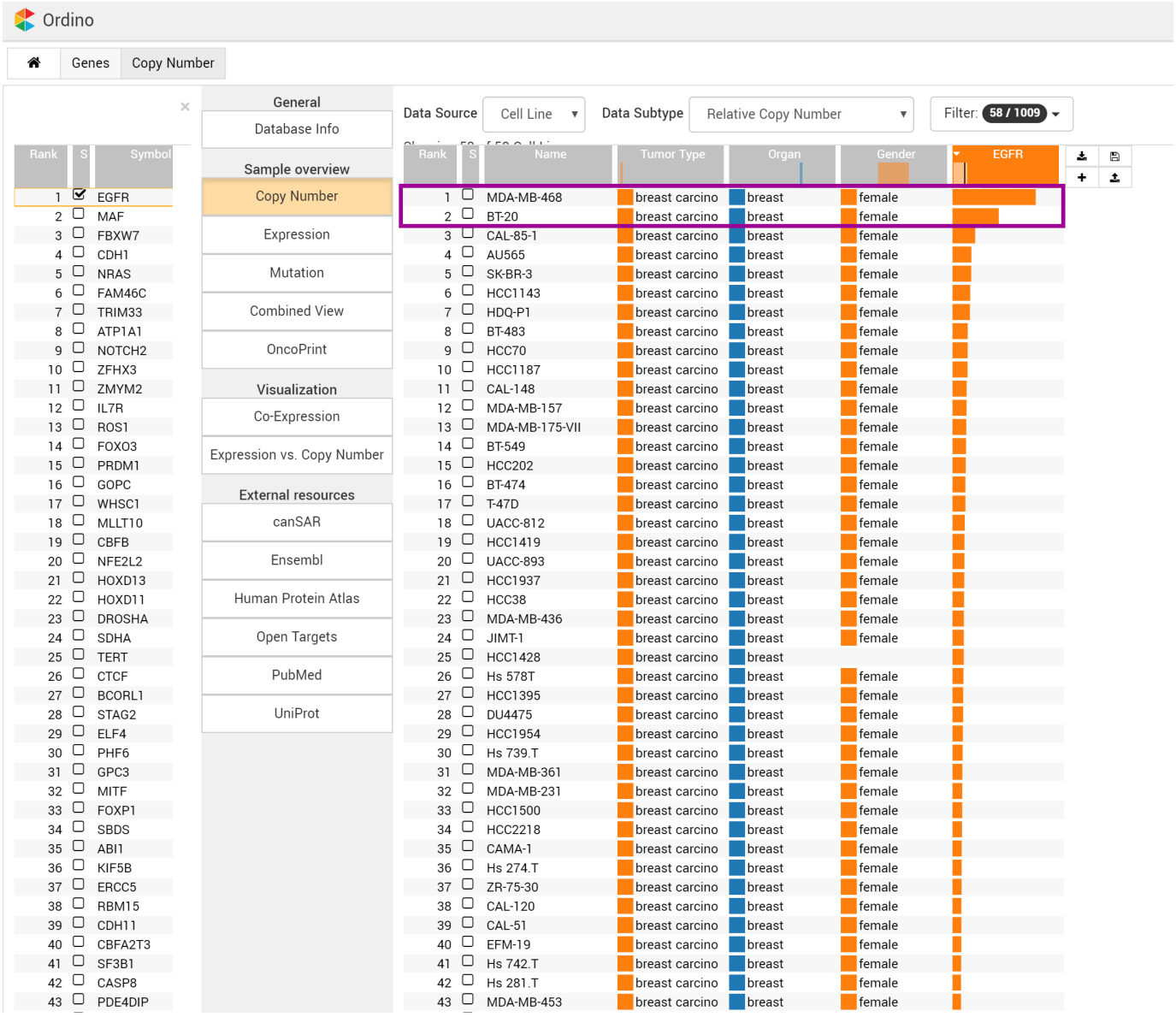

Fig. S 3. Next, the user wants to know, if there are other breast cancer cell lines with *EGFR* amplification? Selecting the gene *EGFR* from the ranking brings up a list of further detail views. The user selects the *Copy Number* view, filters the ranking by tumor type *breast cancer*, and sorts it 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*.

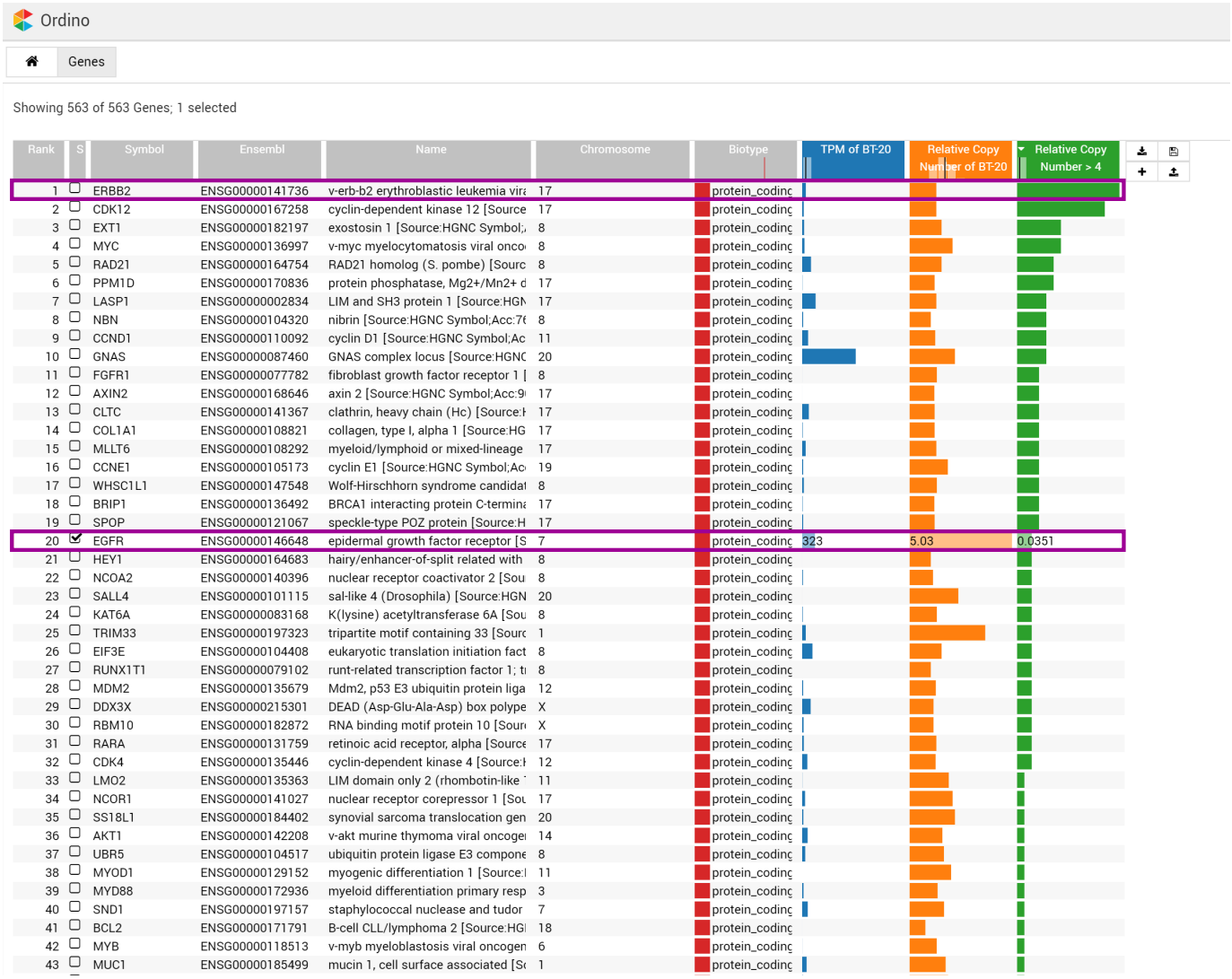

Fig. S 4. The user has a new thought and wants to know which cancer genes are more frequently amplified in breast cancer than *EGFR*? The previous detail view is closed and another score column with the parameters tumor type *breast cancer*; data type *Relative Copy Number*; aggregation *Frequency*; 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.

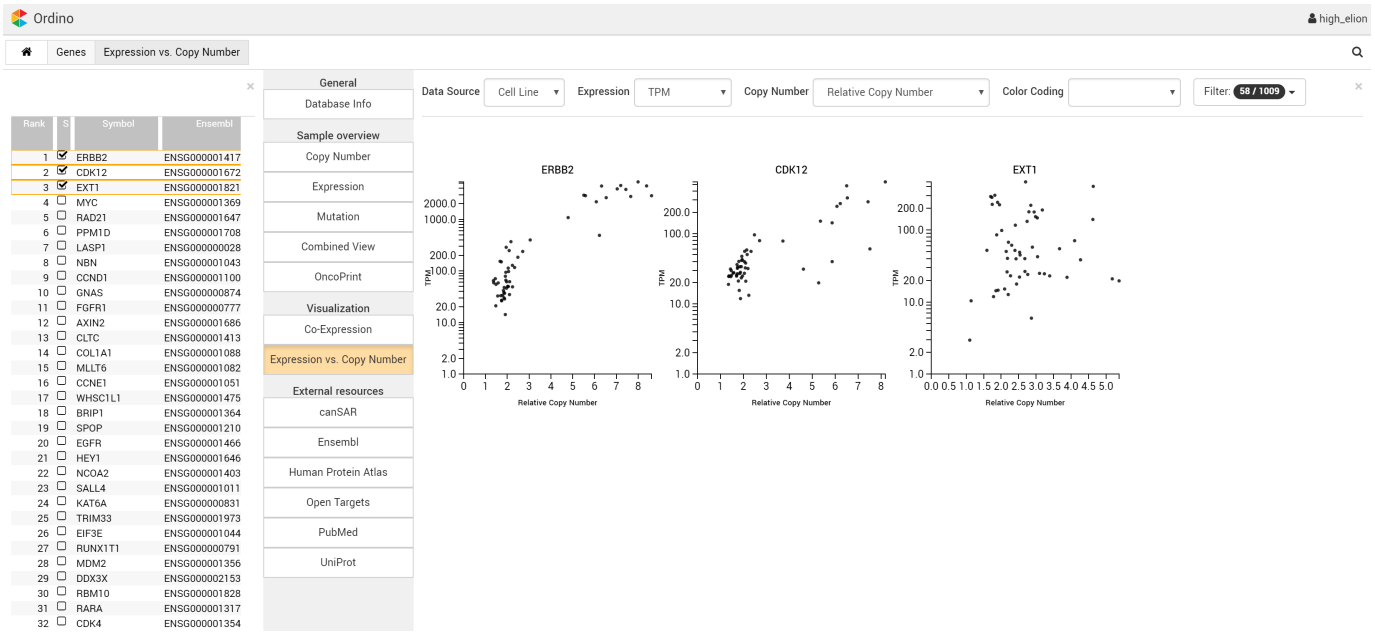

Fig. S 5. Next the user wants to know, if the amplifications of the top genes also lead to a high gene expression? Selecting the first three genes (*ERBB2*, *CDK12*, and *EXT1*) and opening the detail view *Expression vs. Copy Number* (still filtered for *breast cancer* cell lines) shows a scatterplot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

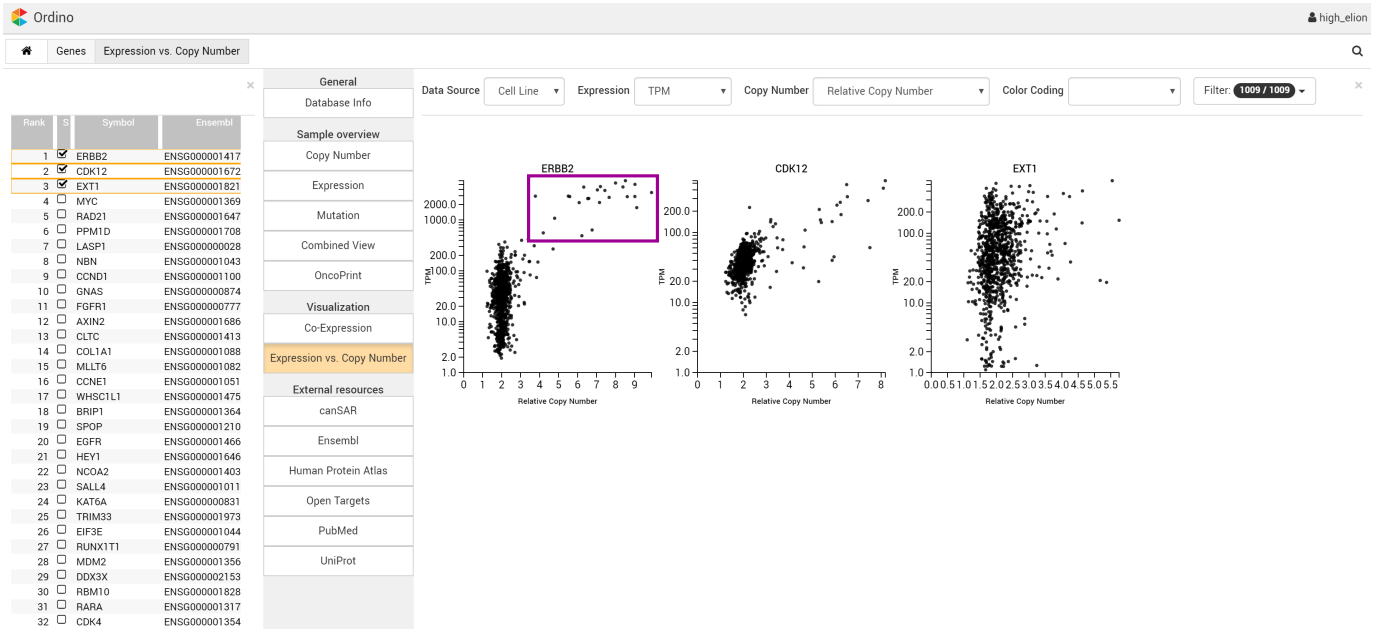

Fig. S 6. The user removes the filter to explore if the expression to copy number correlation is also present in other tumor types. The result is that all cell lines with a very high *ERBB2* copy number have high *ERBB2* expression.

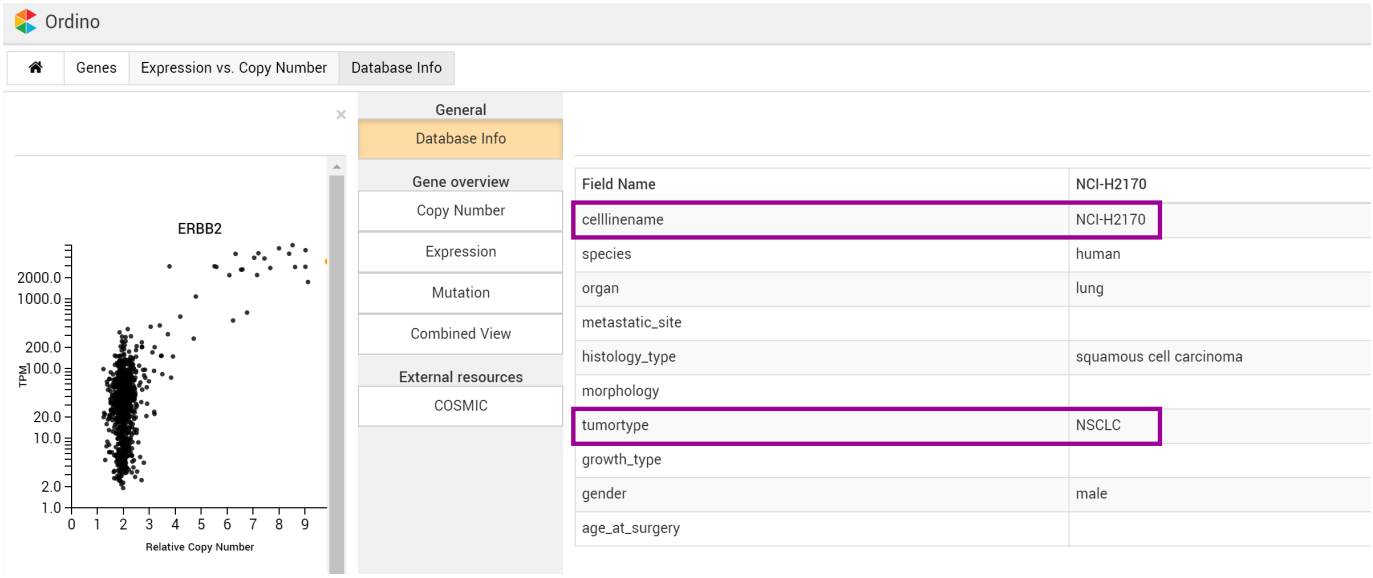

Fig. S 7. Selecting the most right point in the *ERBB2* scatterplot and opening the *Database Info* detail view, reveals that the cell line with the highest *ERBB2* amplification is *NCI-H2170* and belongs to *non-small-cell lung cancer* (NSCLC).

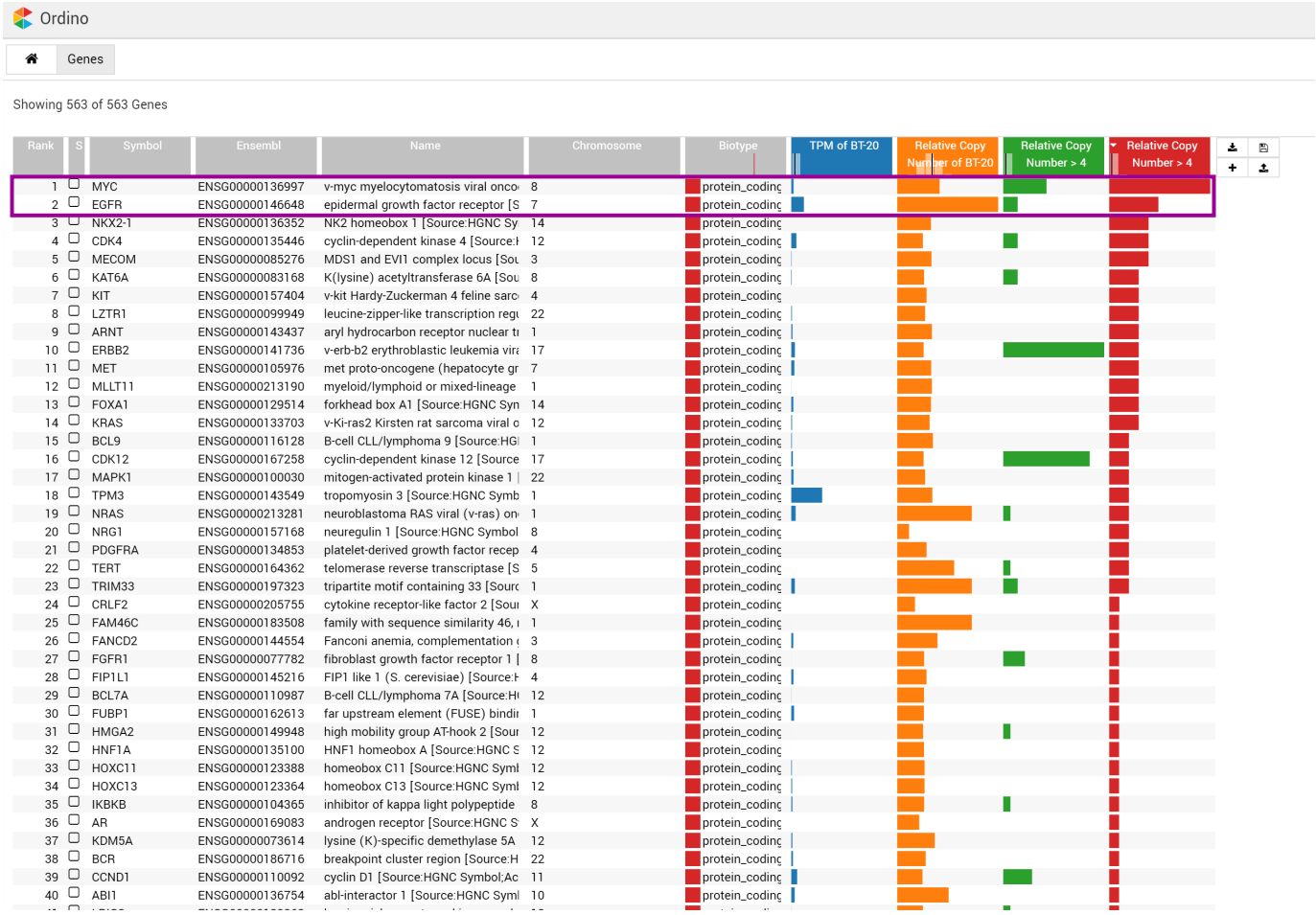

Fig. S 8. With the next step the user wants to investigate, which genes are typically amplified in *non-small-cell lung cancer* (NSCLC)? He switches back to the initial gene ranking, adds another cell line score column with the parameters tumor type *NSCLC*; data type *Relative Copy Number*; aggregation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

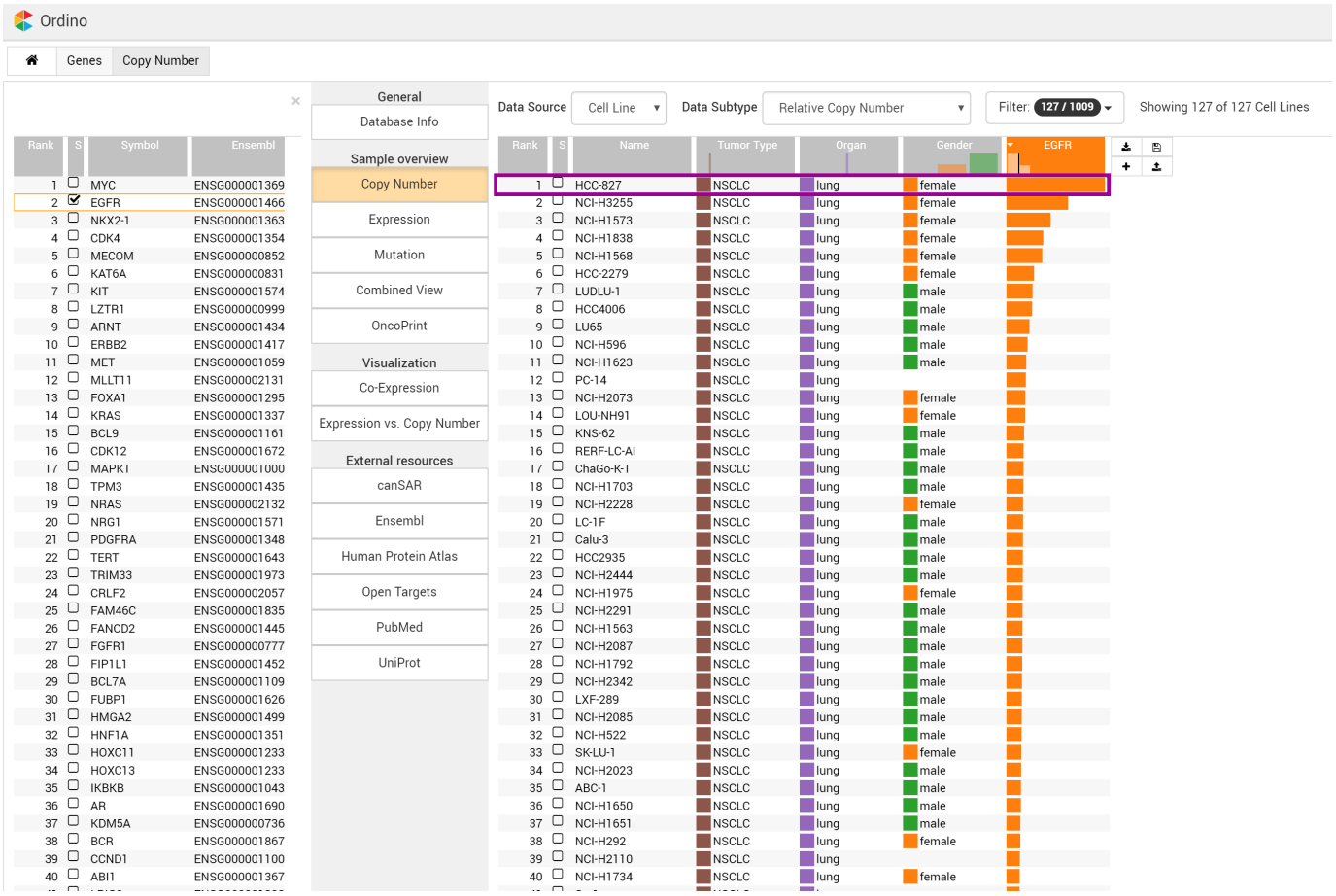

Fig. S 9. The user is still interested in *EGFR* and wants to know which *NSCLC* cell lines have an *EGFR* amplification? He selects the gene *EGFR* from the ranking and opens the *Copy Number* detail view. The copy number value for *EGFR* is added as additional column to the following cell line ranking.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

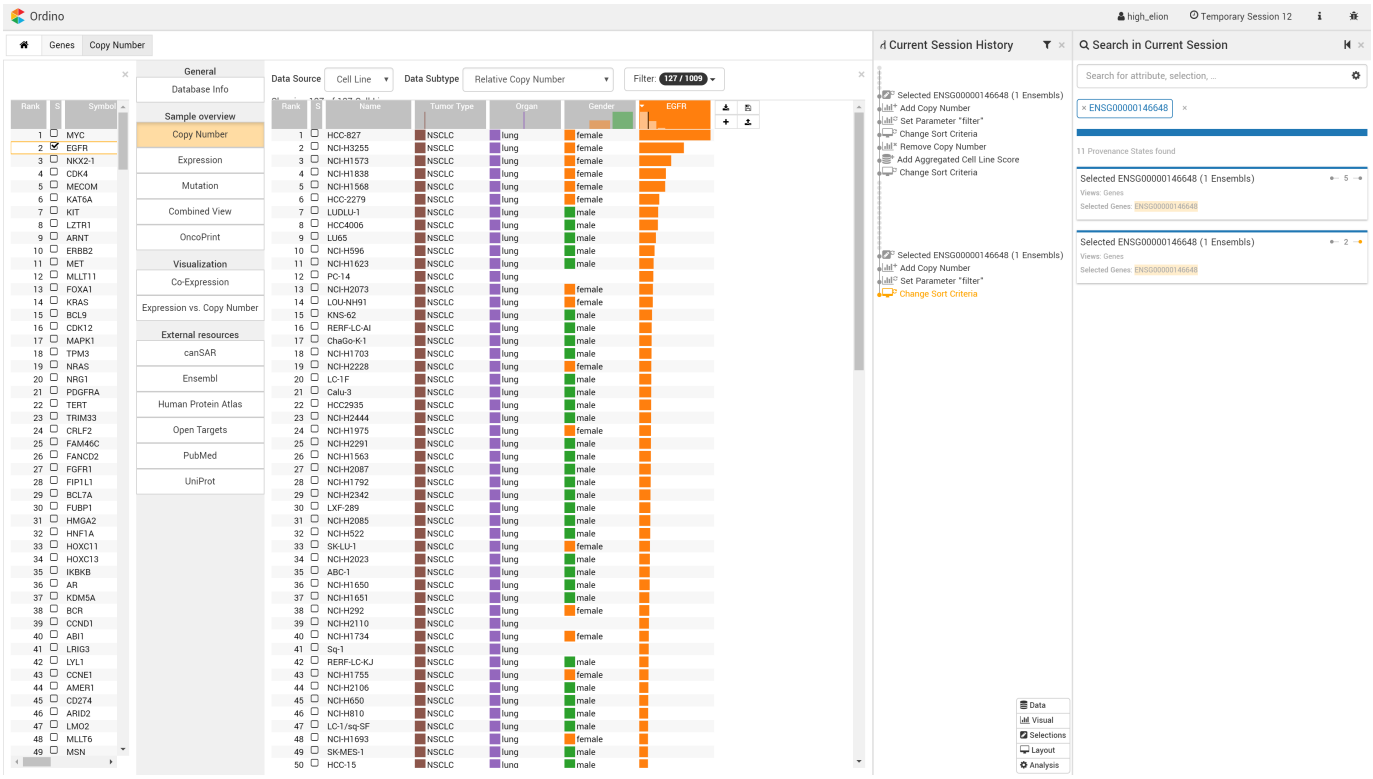

Fig. S 10. Filtering the search suggestions to properties of the active state and selecting the only gene *EGFR* with the identifier *ENSG00000146648* as search term results in two sequences where this gene was selected. When hovering over the search results the corresponding states in the provenance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

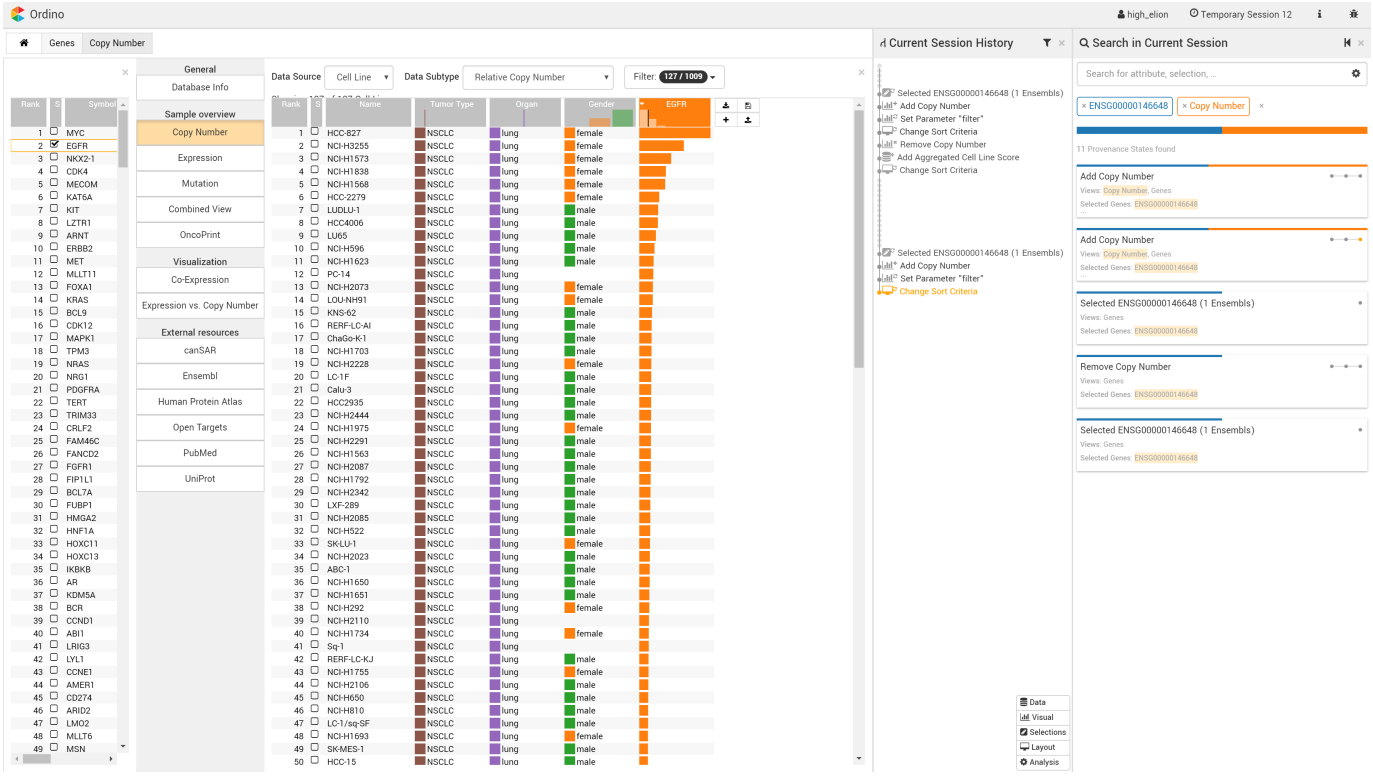

Fig. S 11. Refining the search with the *Copy Number* detail view results in five shorter sequences, where two of them are matching both search terms. The latter sequence contains the currently active state.

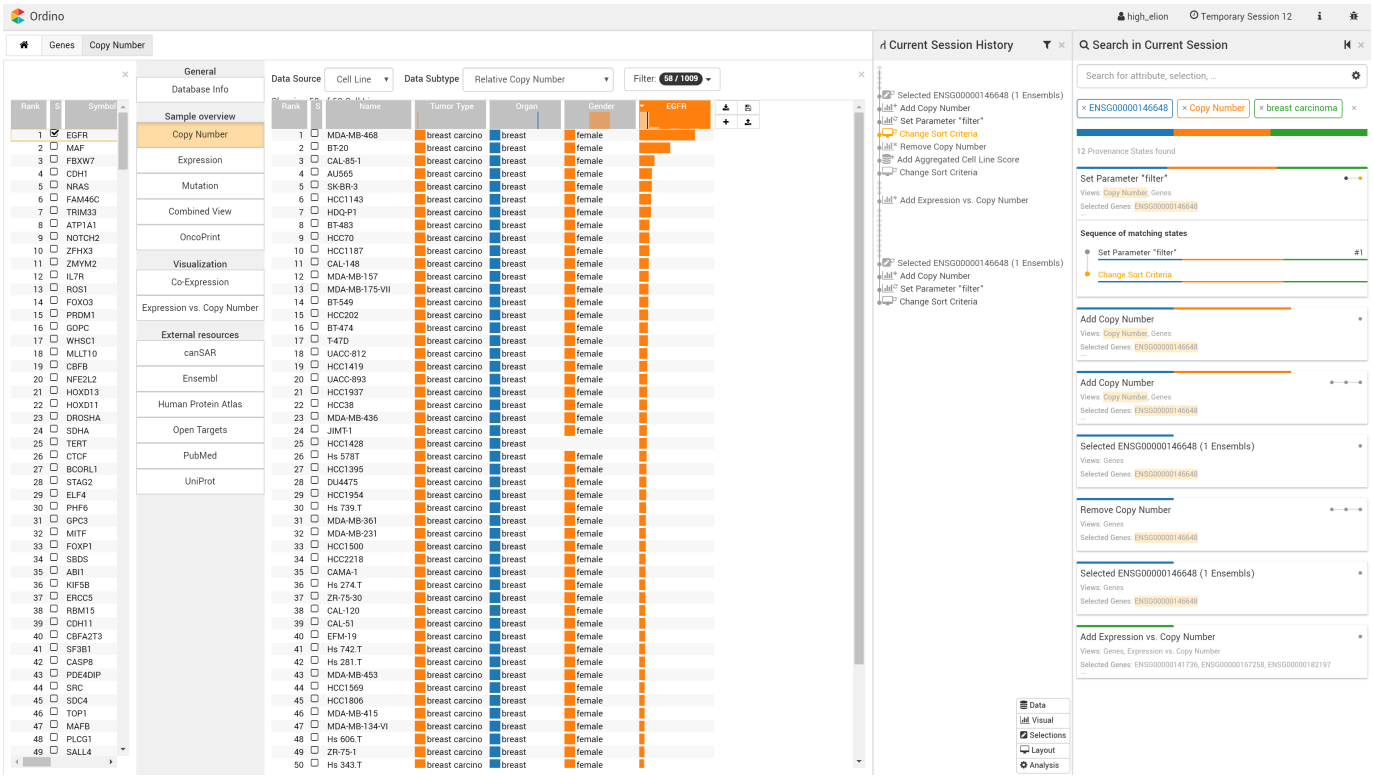

Fig. S 12. Refining the search results with *breast cancer* tumor type as third search term shows that only one sequence matches all search terms. Expanding the sequence and jumping to the last state shows that the *EGFR* copy number was assessed for *breast cancer* cell lines with *MDA-MB-468* as the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

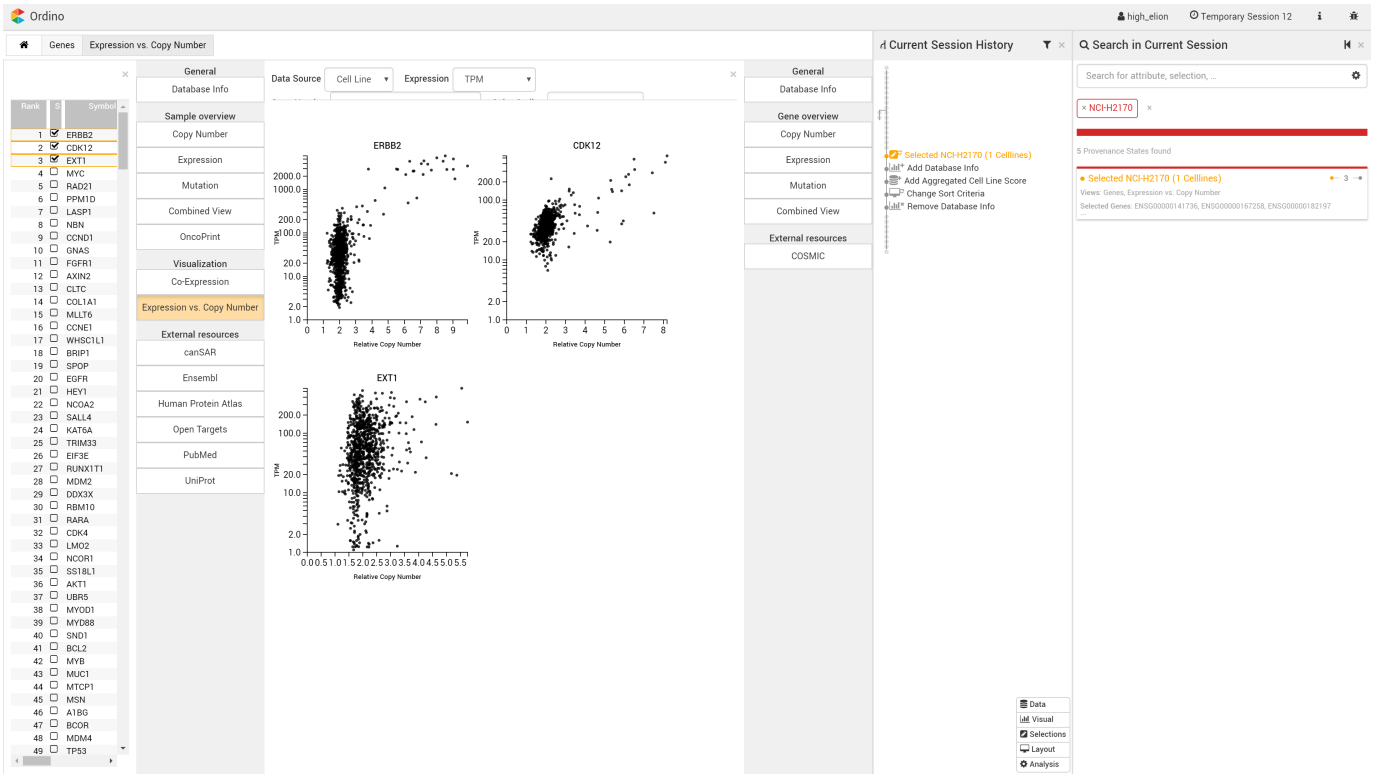

Fig. S 13. The user searches for the cell line *NCI-H2170* to recall if that cell line was used in the analysis. Indeed, the cell line was found in one state sequence and was selected in the *Expression vs. Copy Number* detail view.

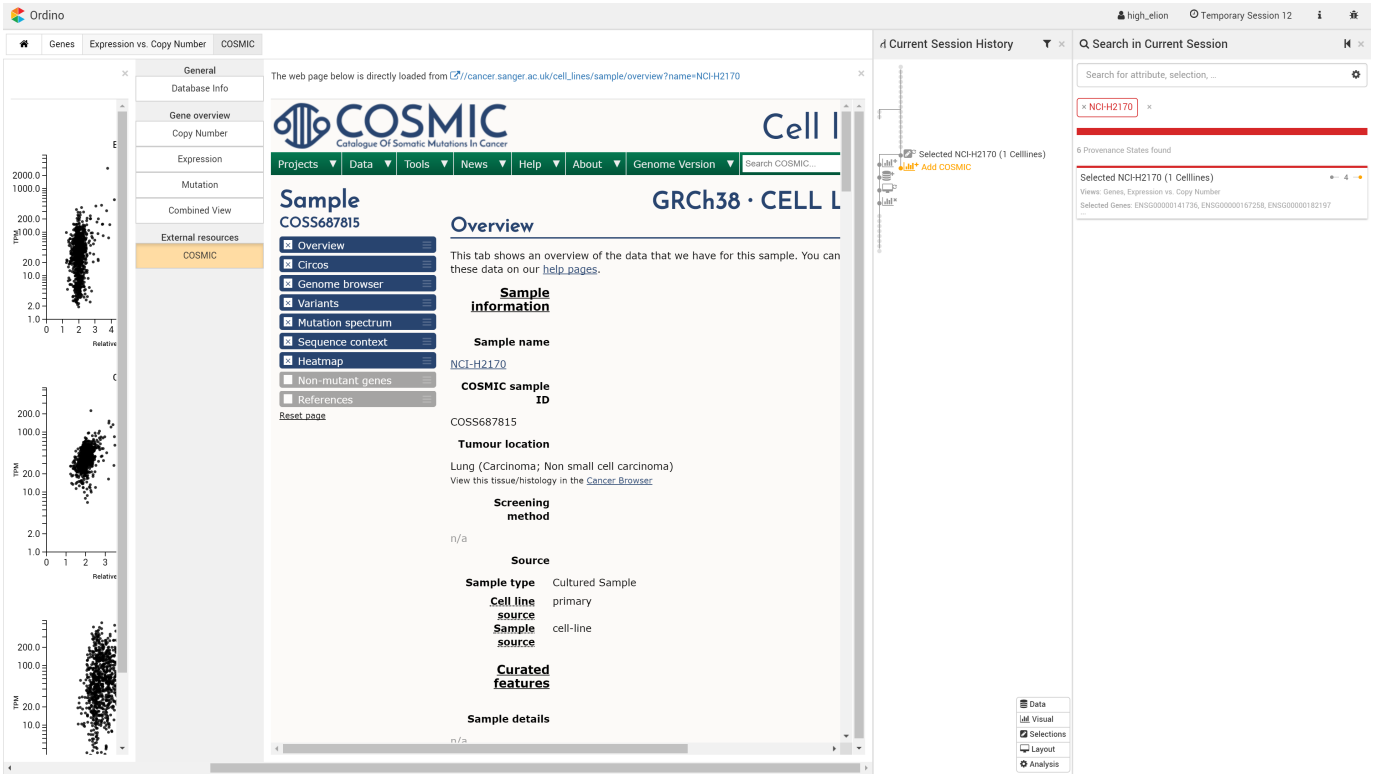

Fig. S 14. Following up with the analysis of cell line *NCI-H2170* the user opens the *COSMIC* detail view and browses the information that is available about this cell line in the COSMIC data base.