Classifying the Progression of Ductal Carcinoma from Single-Cell Sample Data: A Case Study
Supplementary Data

Daniele Catanzaro* Stanley E. Shackney† Alejandro A. Schäffer‡
Russell Schwartz§

January 12, 2015

Abstract

Ductal carcinoma in situ is a precursor lesion of invasive ductal carcinoma of the breast. Investigating its temporal progression could provide fundamental new insights for the development of more effective diagnoses and treatments. In this article we address this major issue by investigating the problem of reconstructing a plausible progression of the carcinoma from single-cell sample data of an affected individual. Specifically, by using a number of assumptions derived from the observation of cellular atypia occurring in ductal carcinoma, we design a possible predictive model based on the parsimony criterion. Preliminary experiments carried out on a population of 13 patients show that the corresponding predicted progressions are non-random and classifiable in subfamilies having specific evolutionary characteristics.

Keywords: network design, combinatorial optimization, hierarchical classification, computational biology, tumorigenesis, ductal carcinoma.

Results

Figures 1-13 show the predictions obtained by applying the model described in the main article on the 13 considered datasets. Similarly, Figures 15-27 show the predicted gene-driven correlated variation in the corresponding datasets, i.e., the overall number of times (expressed in percentage) that a change in the copy number of a specific gene in a taxon would cause a change in the copy number of the other genes in the immediate descendent taxon. As general trend, the predictions show that the progression of ductal carcinoma is characterized by a high tendency to loose copies of the TP53 gene ($17.79\% \pm 5.11\%$) and, more in general, of the suppressor genes ($36\% \pm 5.35\%$); moreover, in the considered datasets the suppressor genes are characterized by a high level of *spontaneous variation* (see Figure 4 in the main article), i.e., the tendency of a gene in a taxon to increase or decrease its copy number with respect to its ancestor provided that the copy numbers of the remaining genes in both taxa are unchanged. Specifically, CDH1 is the one showing in average the highest rate of spontaneous variation (16.53 ± 6.59), followed by TP53 (15.36 ± 5.18) and DBC2 (14.23 ± 6.51). Among the oncogenes, COX2 is the gene that shows the highest level of spontaneous variation (13.19 ± 6.02), followed by ZNF217 (10.75 ± 6.07), CCND1 (10.13 ± 4.69), HER2 (10.03 ± 3.33) and MYC (9.79 ± 3.87). Interestingly, ZNF217 is more prone to spontaneous variations in datasets DAT07 and DAT09 and such phenomenon seems to be correlated to the variation induced by CDH1. (see Figures 21 and 23).

The frequency of the doubling-loss phenomenon is in proportion less preponderant ($7\% \pm 3.99\%$) than the tendency to loose tumor suppressor genes and in general more localized in certain datasets (namely, regular datasets, see Table 1) than others. An accurate analysis of the results suggests that the predicted progressions are non-random and classifiable in subfamilies having specific evolutionary characteristics (see

*Louvain School of Management, Catholic University of Louvain, Mons, Belgium.
†Departments of Human Oncology and Human Genetics, Drexel University School of Medicine, Pittsburgh, PA.
‡Computational Biology Branch, NCBI, NIH, Bethesda, Maryland, United States of America.
§Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA.
Correspondence should be addressed to: russells@andrew.cmu.edu

Doubling-driven			Doubling Absent
Regular	Abnormal		
	TP53-driven	CDH1-driven	
DAT02	DAT06	DAT01	DAT05
DAT03	DAT08	DAT07	DAT10
DAT04		DAT09	
DAT11			
DAT12			
DAT13			

Table 1: A possible classification of the analyzed datasets.

Table 1). Specifically, we can distinguish between progressions showing a preponderance of doubling-loss phenomena ("doubling-driven" column in Table 1) and progressions showing a low or absent presence of the doubling-loss phenomenon ("Doubling Absent" column in Table 1). The first group is the largest and can be in turn subdivided in three main subgroups, namely: the *regular*, the *abnormal with TP53 predominance* and the *abnormal with CDH1 predominance*. The regular group is the largest subgroup and includes datasets DAT02, DAT03, DAT04, DAT11, DAT12 and DAT13. As general trend, the group shows a high spontaneous variation, usually affecting the tumor suppressor genes CDH1 and TP53 which in general tend to be lost. Moreover, the copy number of the genes in general do not tend to increase with respect to the root node as in the abnormal groups and usually do not exceed 8 copies. The doubling-loss phenomenon is more preponderant in the regular group than in others, it usually tends to affect (almost) all genes (see e.g., taxa $\langle 2.2.2.1.2 - 1.1.1\rangle$ and $\langle 4.4.4.2.4 - 2.2.2\rangle$, or $\langle 2.3.2.1.2 - 1.1.1\rangle$ and $\langle 4.6.4.2.4 - 1.2.2\rangle$ in Figure 2), and it can be considered as a possible source of the progression of the carcinoma, being located in several internal vertices of corresponding predictions.

The abnormal with TP53 predominance subgroup includes datasets DAT08 and DAT06. It is characterized by a very high spontaneous variation for TP53 with frequent lost either of the tumor suppressor gene CDH1 if DBC2 is affected by high spontaneous variations or, vice versa, of the tumor suppressor gene DBC2 if CDH1 is affected by high spontaneous variations. The copy number of the genes tends to increase more than in the regular subgroups but usually it does not exceed again 8 copies. The instance DAT06 shows multiple situations in which some or all of the tumor suppressor genes are lost and this phenomenon usually comes together with a simultaneous loss one or more oncogenes. A similar situation can be observed also in DAT08, although the loss of gene copies is less predominant. Interestingly, dataset DAT08 shows a very high level of variation of TP53 with copy numbers ranging from 0 to 5. This fact seems to suggest the presence of a strong pressure acting on this gene. Similarly, in both datasets COX2 and MYC show a high level of variation, in particular COX2 which tends to increase or double in abnormal way. The doubling-loss phenomenon is less preponderant than in the regular subgroup and it is "abnormal" in the sense that it usually does not interest all of the genes but just part of them (see e.g., taxa $\langle 4.2.2.2.3 - 2.2.2\rangle$ and $\langle 5.4.3.4.1 - 2.2.3\rangle$ in Figure 6 or taxa $\langle 2.2.2.2.2 - 1.1.2\rangle$ and $\langle 4.2.2.2.2 - 2.1.3\rangle$ in Figure 8). Also in this case, the doubling-loss phenomenon can be considered as a possible source of the progression of the carcinoma, being located in several internal vertices of corresponding predictions.

The abnormal with CDH1 predominance subgroup includes datasets DAT07, DAT01, and DAT09. It is characterized by a very high spontaneous variation of CDH1 with respect to TP53, high spontaneous variation of CCND1 and ZNF217, and a low tendency to loose tumor suppressor genes. This subgroup is the one that shows in absolute the highest increment of the copy number of the genes in the considered datasets, by containing taxa having gene copy number approaching 25 (see e.g., DAT07). The doubling-loss phenomenon is still present, although in an abnormal and less preponderant way than the previous two subgroups, hence it can hardly be considered as a possible source of the progression of the carcinoma. Instead, possibly COX2, MYC and ZNF217 could play an important role in the progression, being characterized by very high variations in copy numbers.

Finally, the doubling absent subgroup includes datasets DAT05 and DAT10. It is characterized by a very high spontaneous variation of CDH1, high spontaneous variation of COX2 and MYC, and a low tendency to loose tumor suppressor genes. This subgroup does not show a significative presence of doubling-loss phenomena, although some genes (namely, COX2 and MYC) tends to approach high number of gene copy

numbers with respected to the root, above all in DAT10. Specifically, in this dataset the progression of the carcinoma seems to be caused by iterated increments of COX2 over time which seems to affect the variation of the copy number of CDH1.

Acknowledgements

This research was supported in part by the Belgian National Fund for Scientific Research (FRS-FNRS) (D.C), U.S. National Institutes of Health grants 1R01CA140214 (R.S.) and 1R01AI076318 (R.S.) and the Intramural Research Program of the NIH, NLM (A.A.S.).

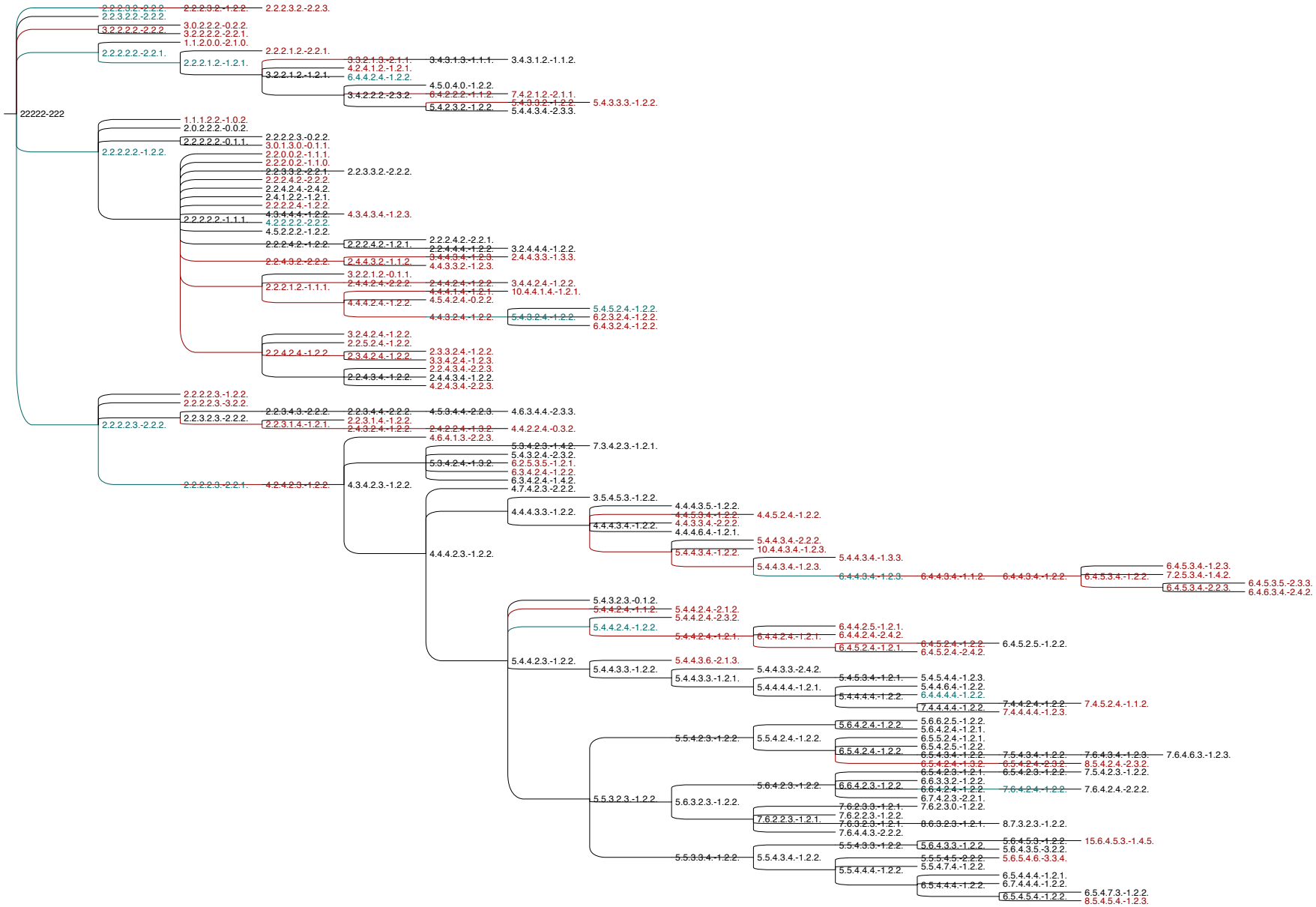

Figure 1: Predicted progression from DCIS to IDC for the dataset DAT01.

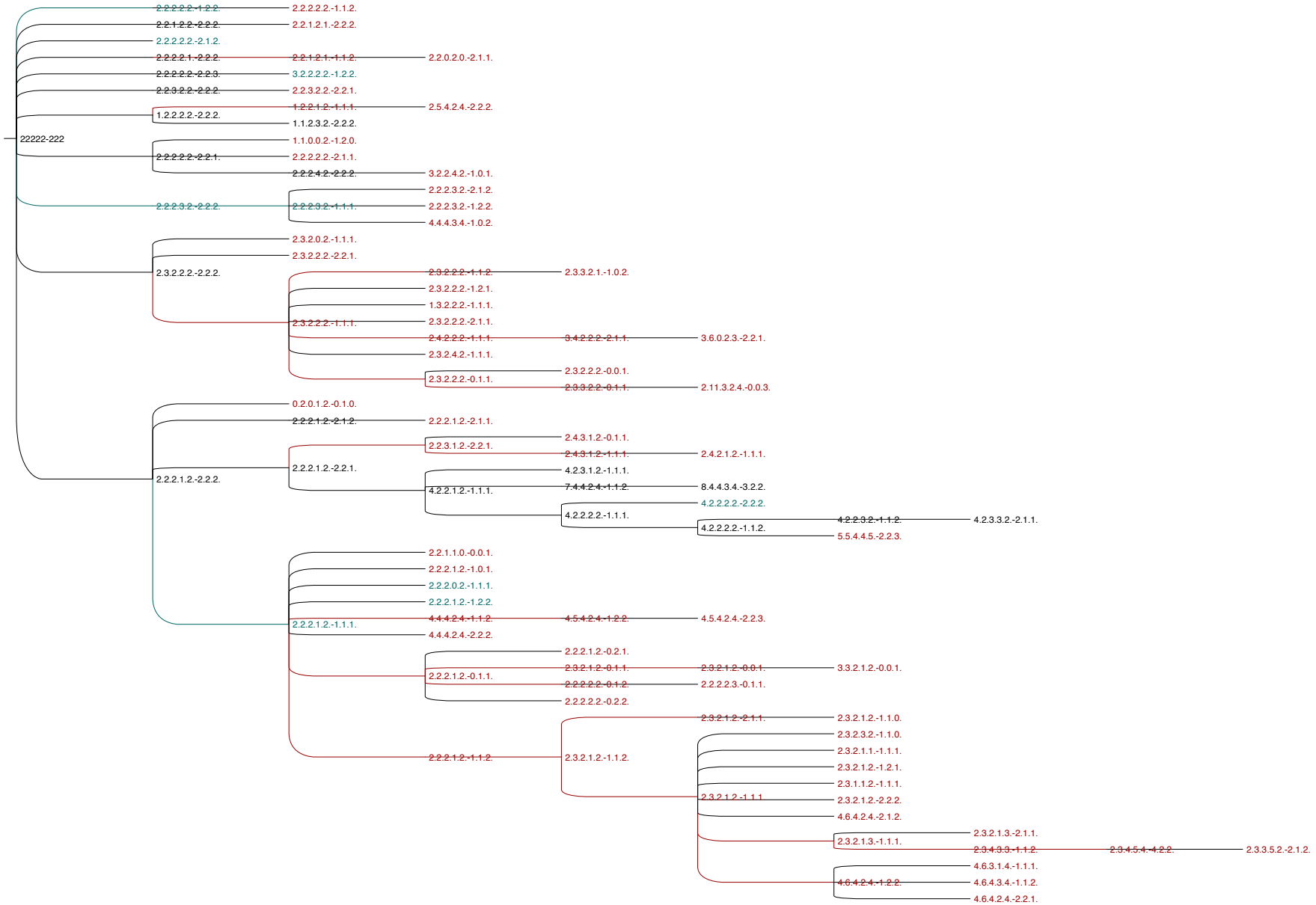

Figure 2: Predicted progression from DCIS to IDC for the dataset DAT02.

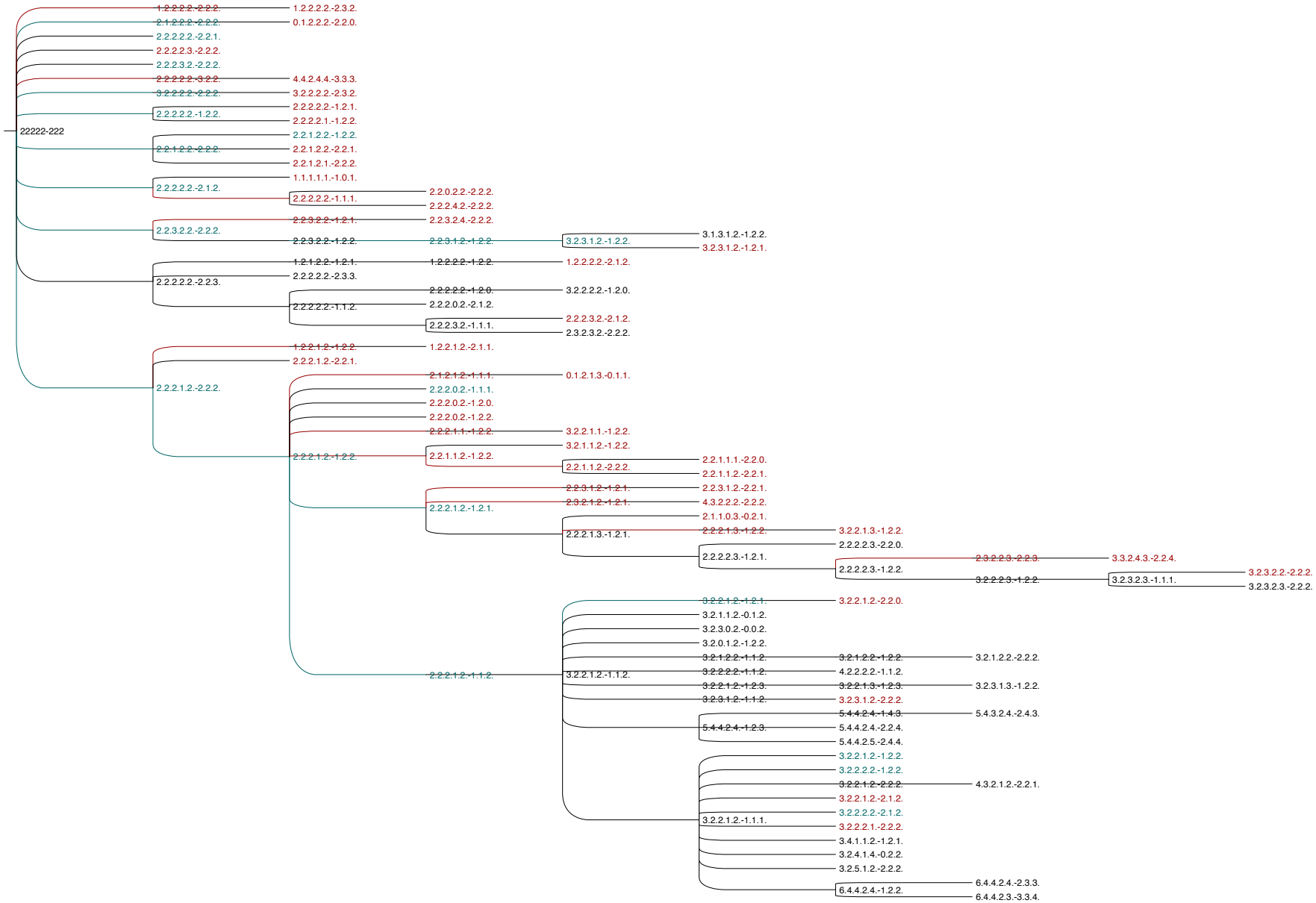

Figure 3: Predicted progression from DCIS to IDC for the dataset DAT03.

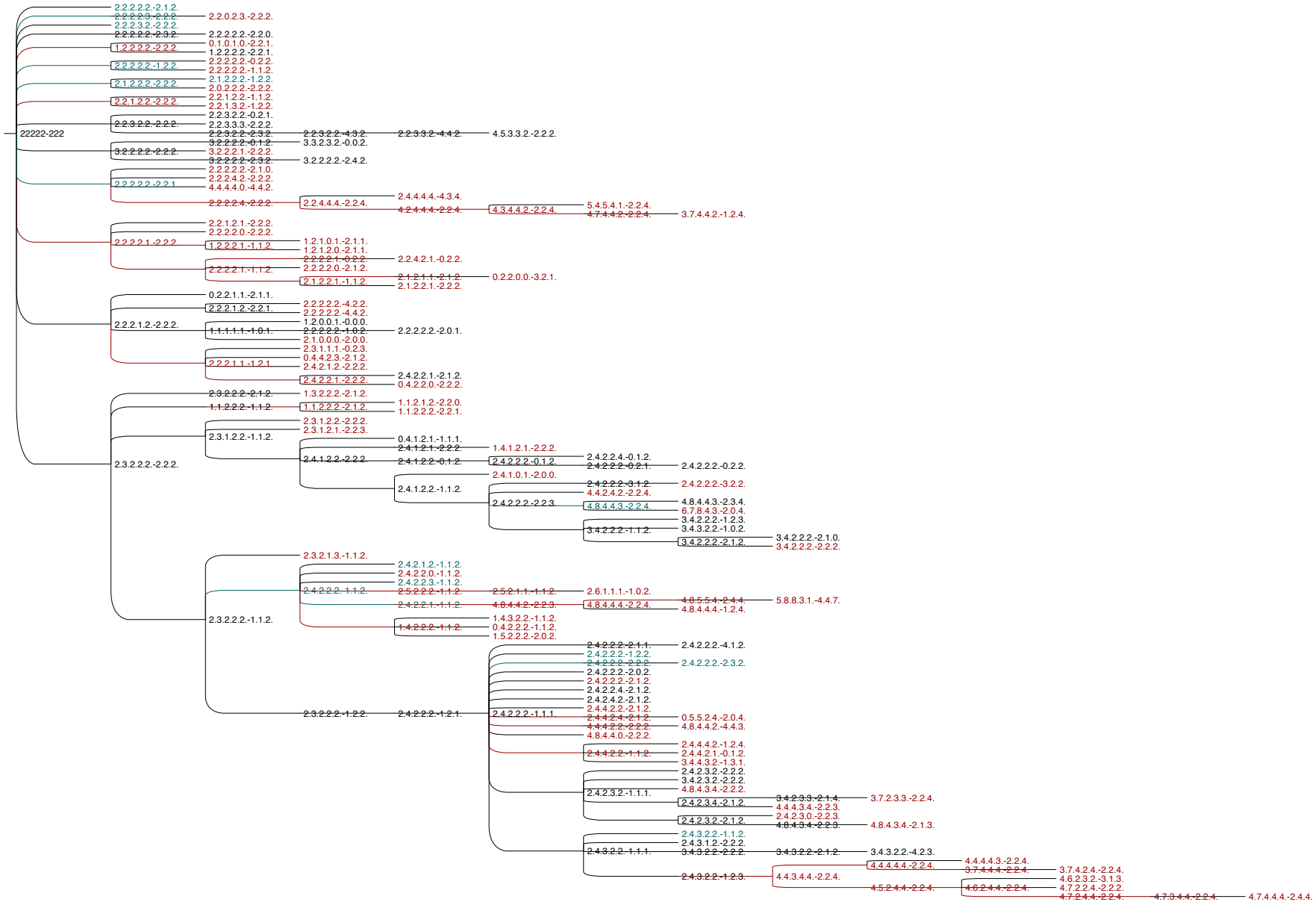

Figure 4: Predicted progression from DCIS to IDC for the dataset DAT04.

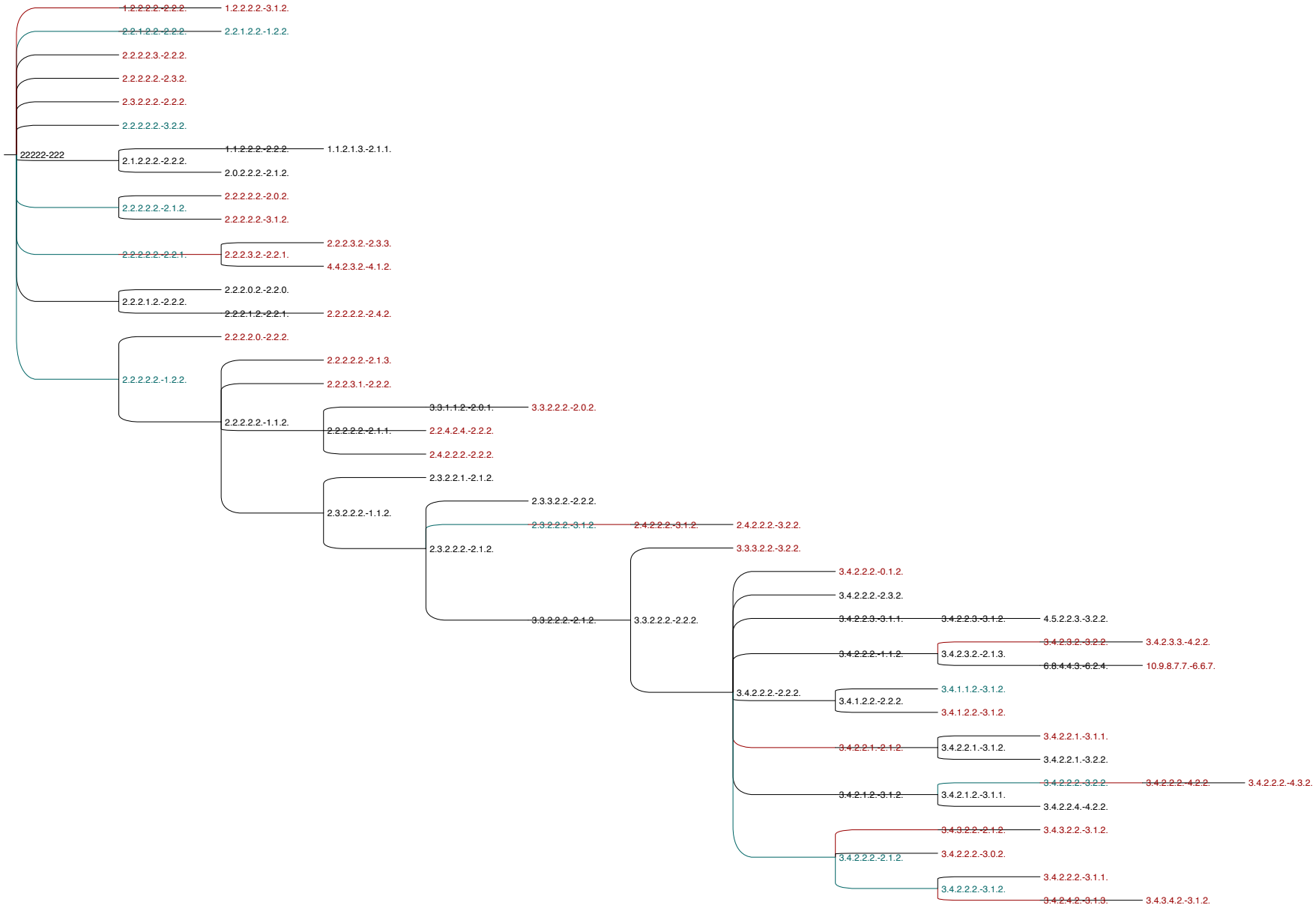

Figure 5: Predicted progression from DCIS to IDC for the dataset DAT05.

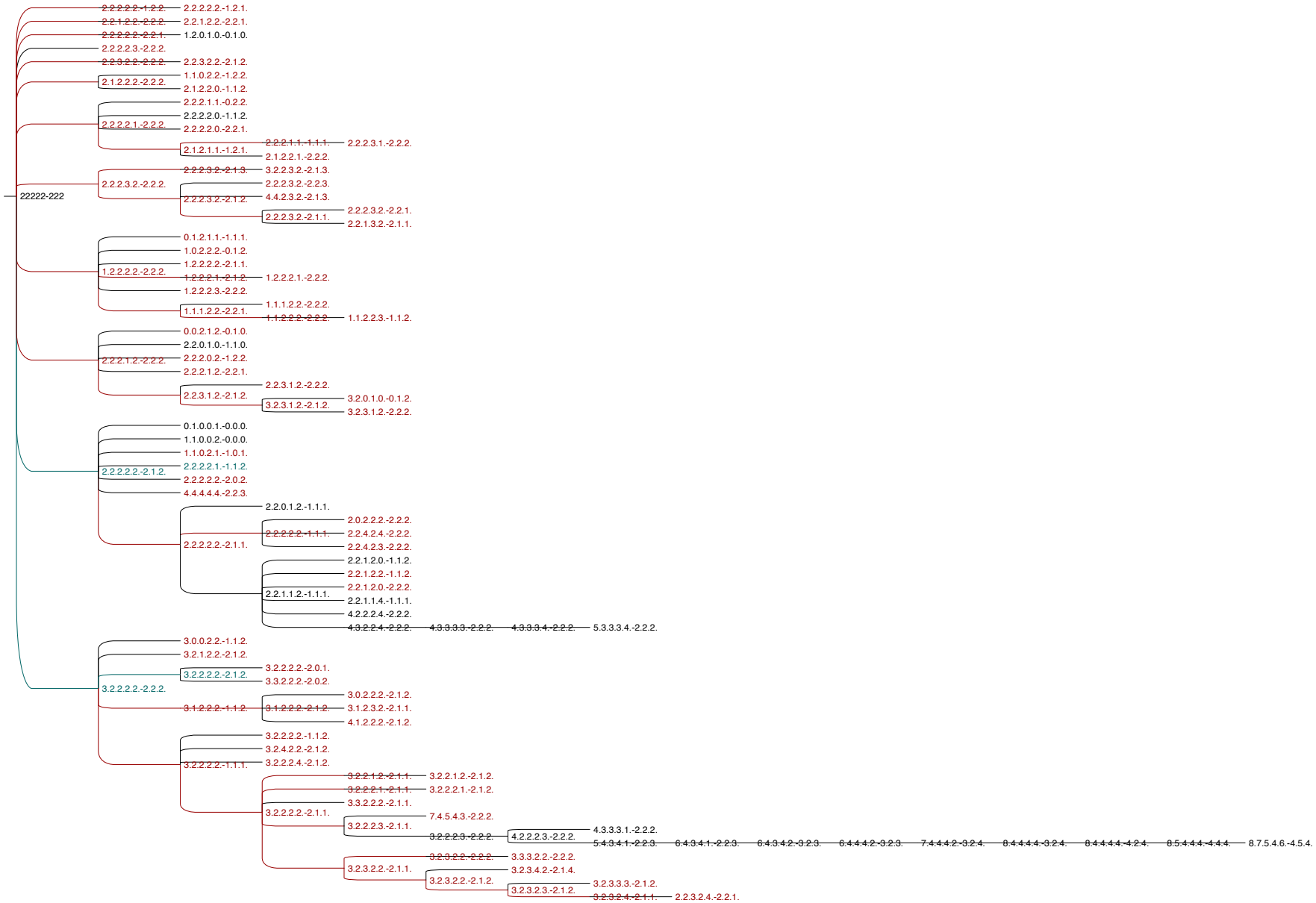

Figure 6: Predicted progression from DCIS to IDC for the dataset DAT06.

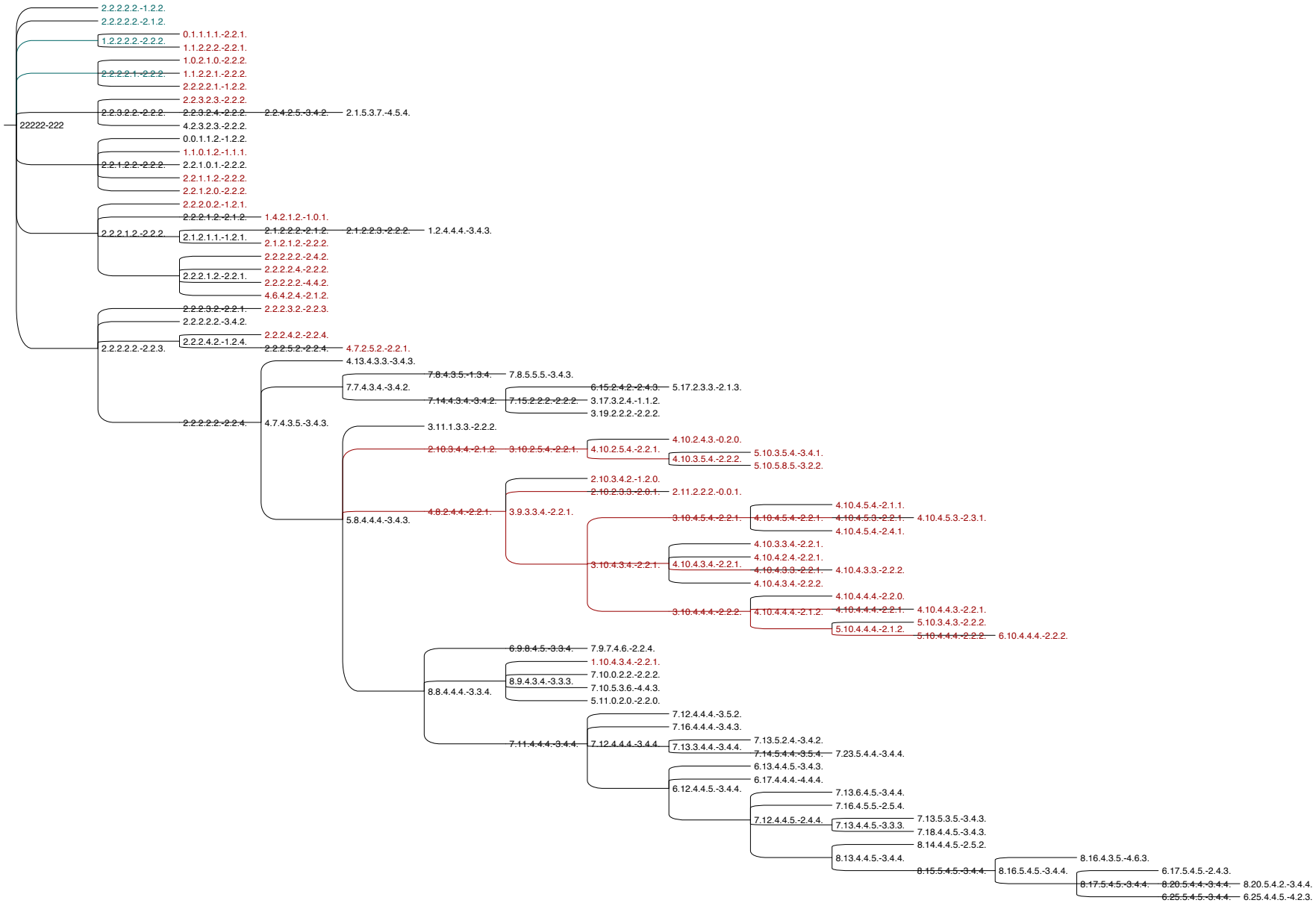

Figure 7: Predicted progression from DCIS to IDC for the dataset DAT07.

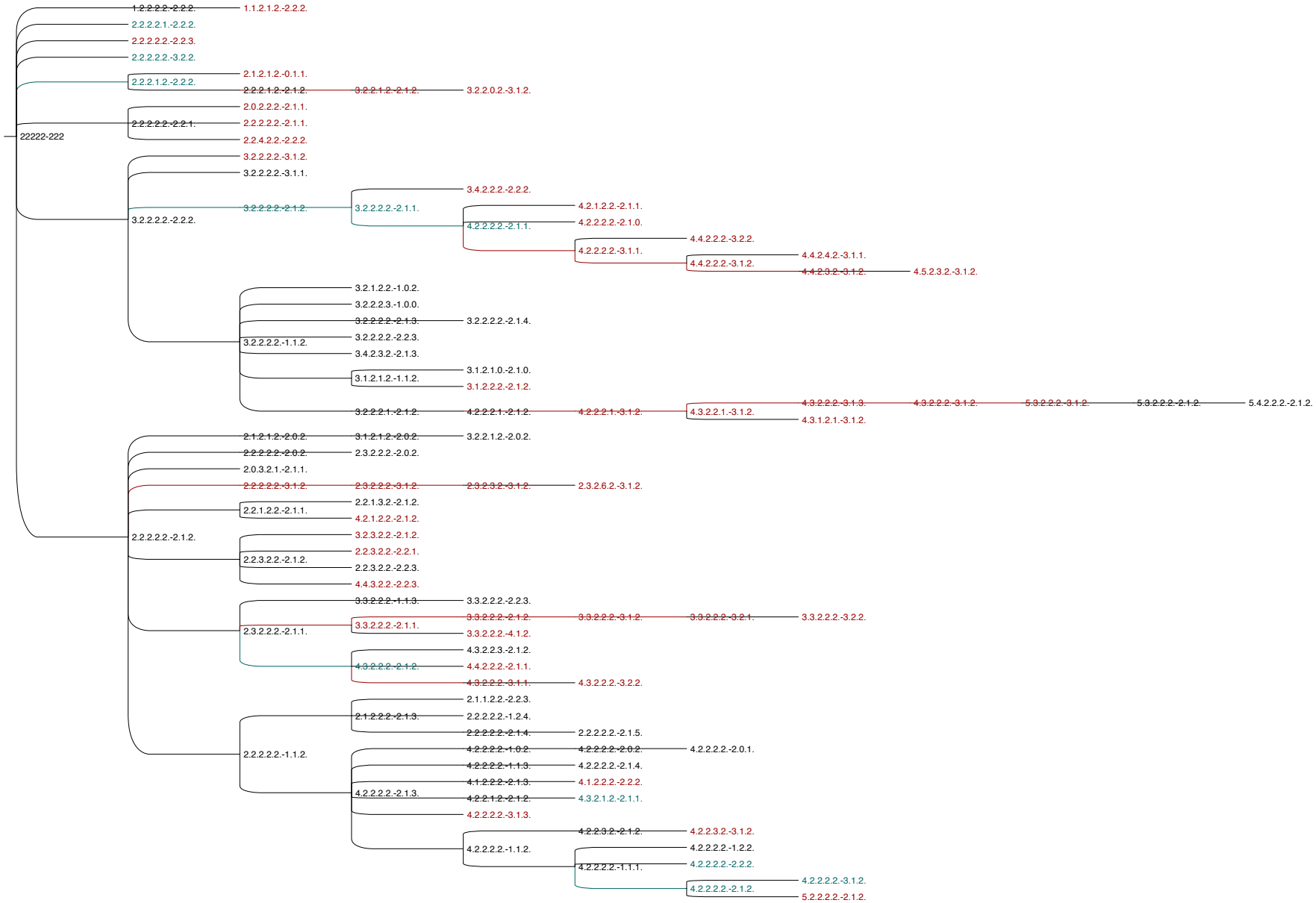

Figure 8: Predicted progression from DCIS to IDC for the dataset DAT08.

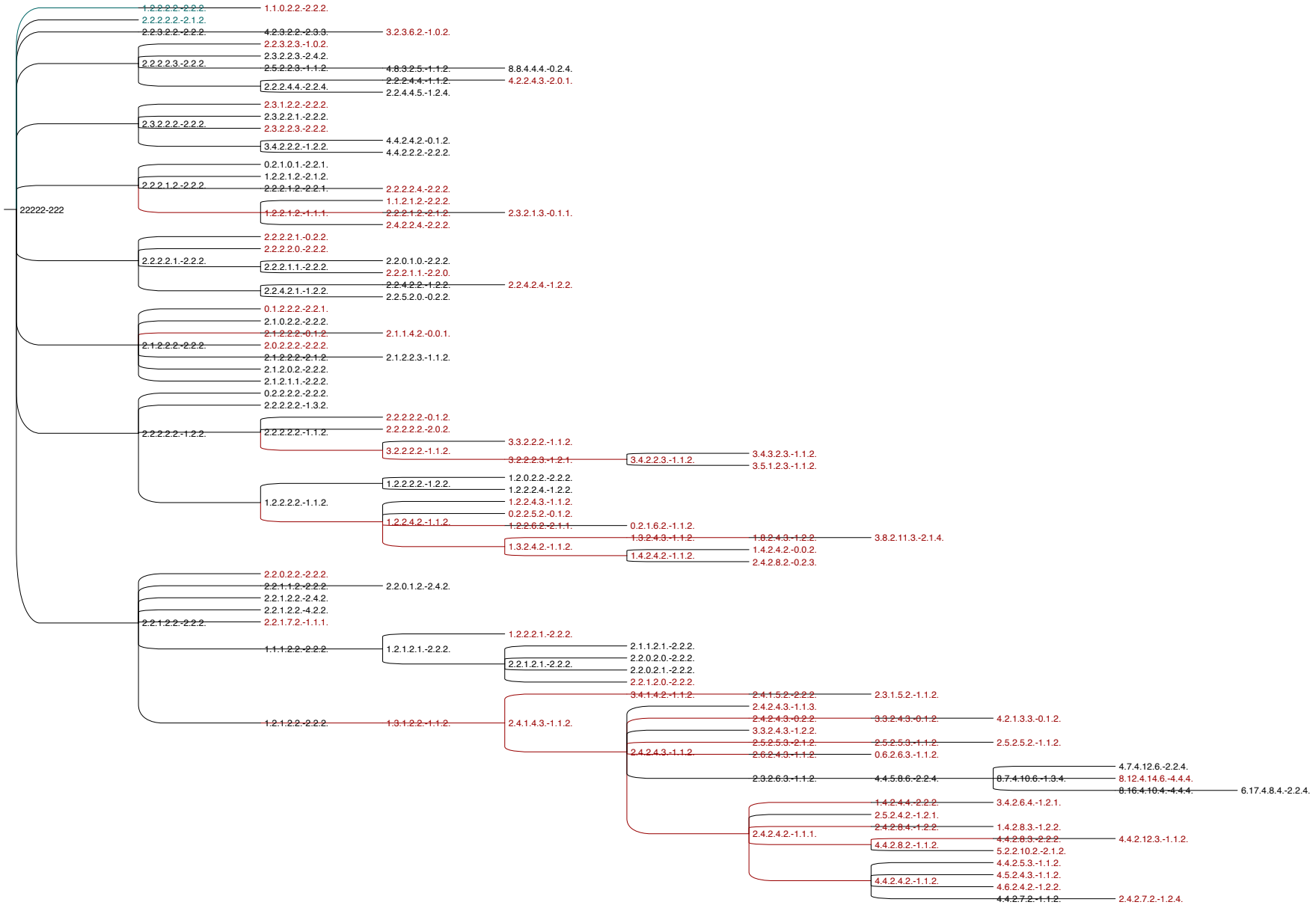

Figure 9: Predicted progression from DCIS to IDC for the dataset DAT09.

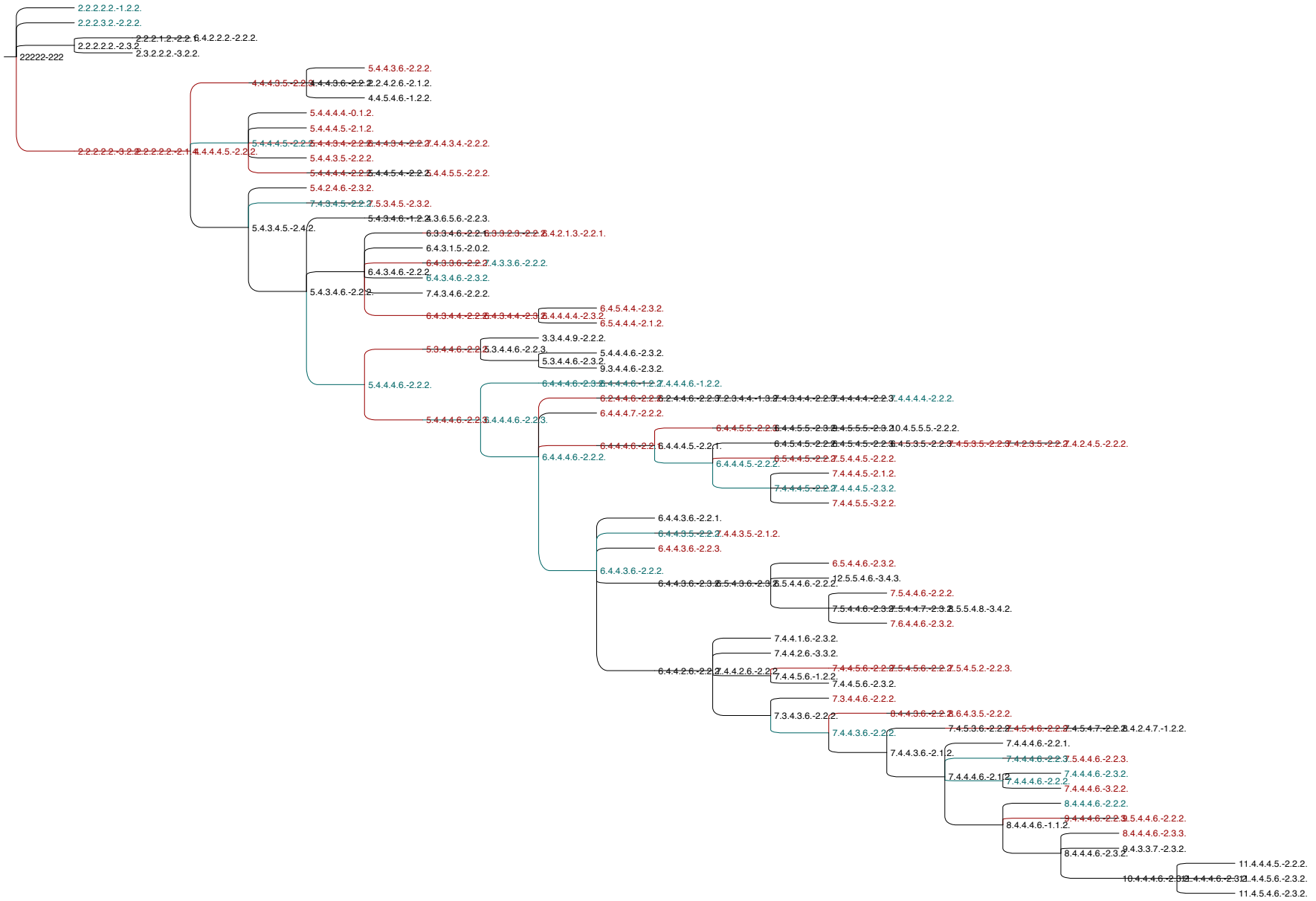

Figure 10: Predicted progression from DCIS to IDC for the dataset DAT10.

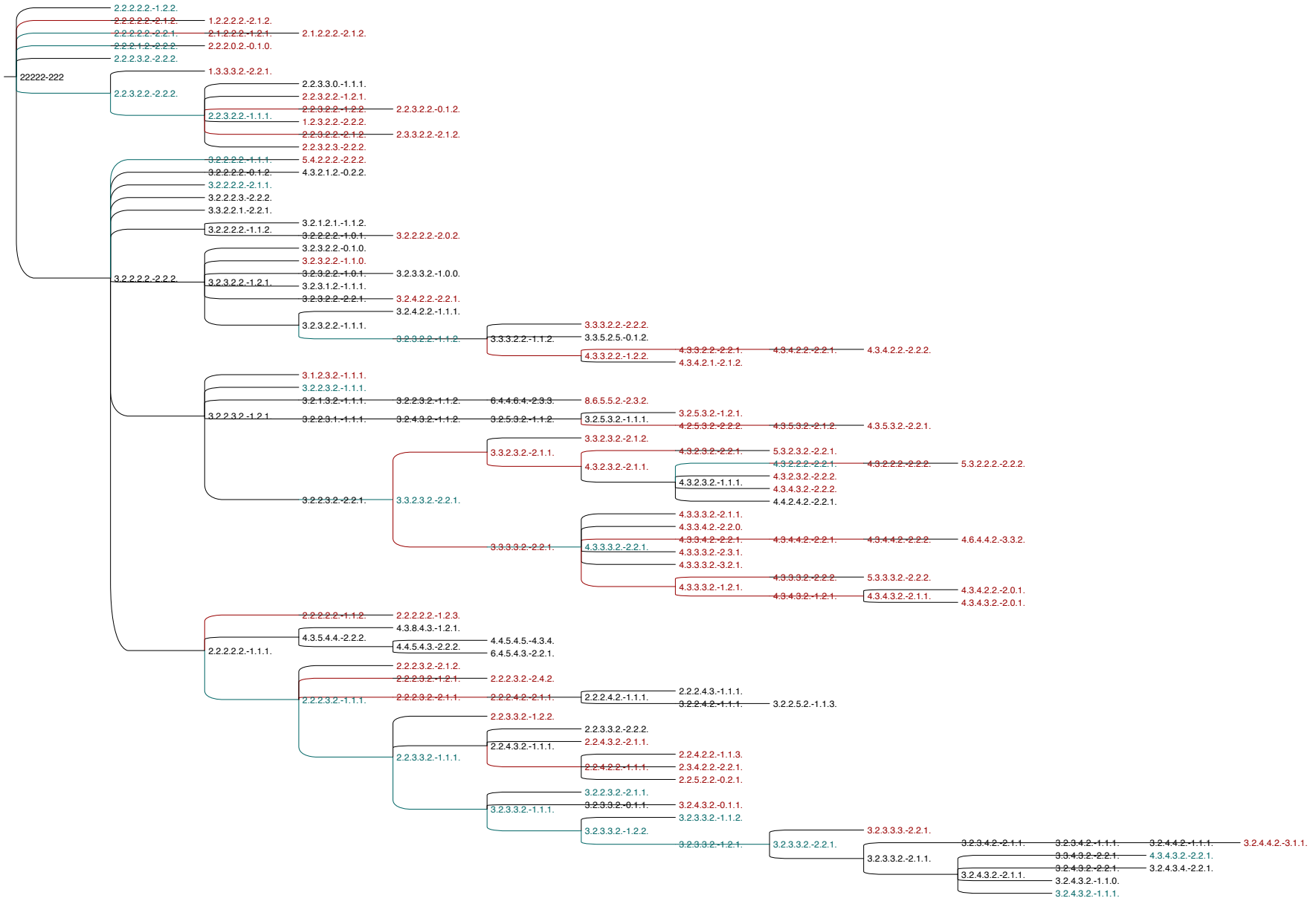

Figure 11: Predicted progression from DCIS to IDC for the dataset DAT11.

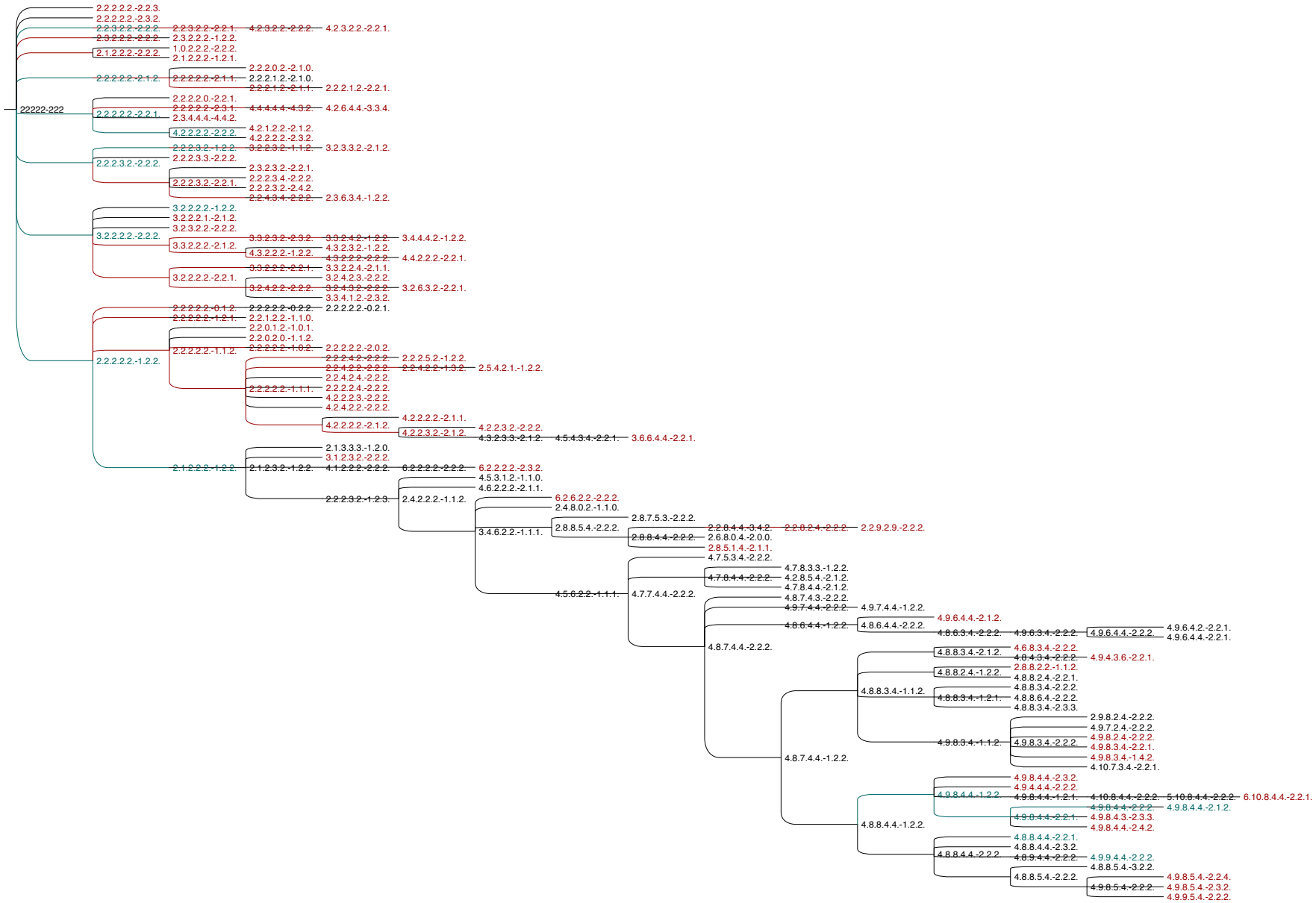

Figure 12: Predicted progression from DCIS to IDC for the dataset DAT12.

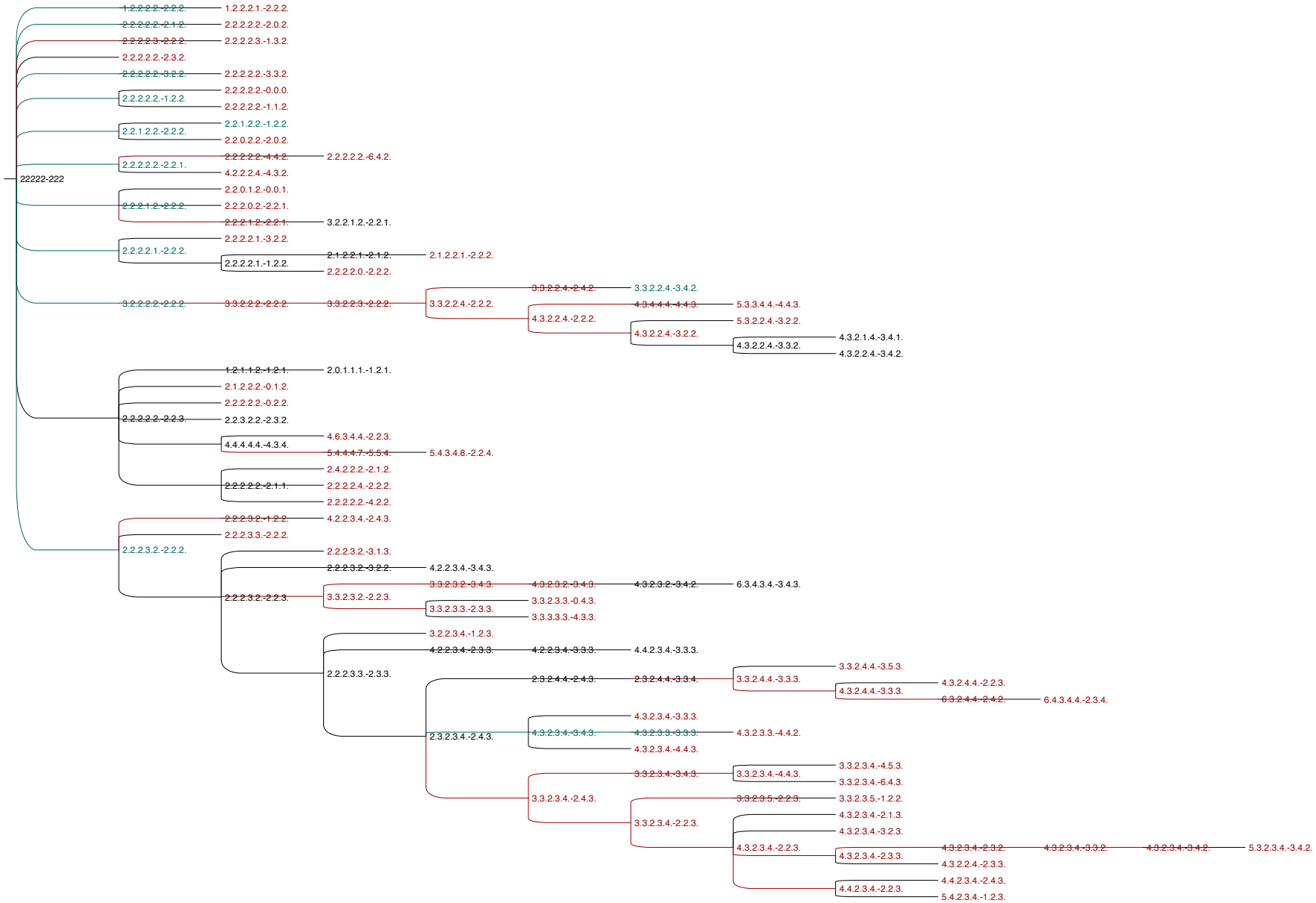

Figure 13: Predicted progression from DCIS to IDC for the dataset DAT13.

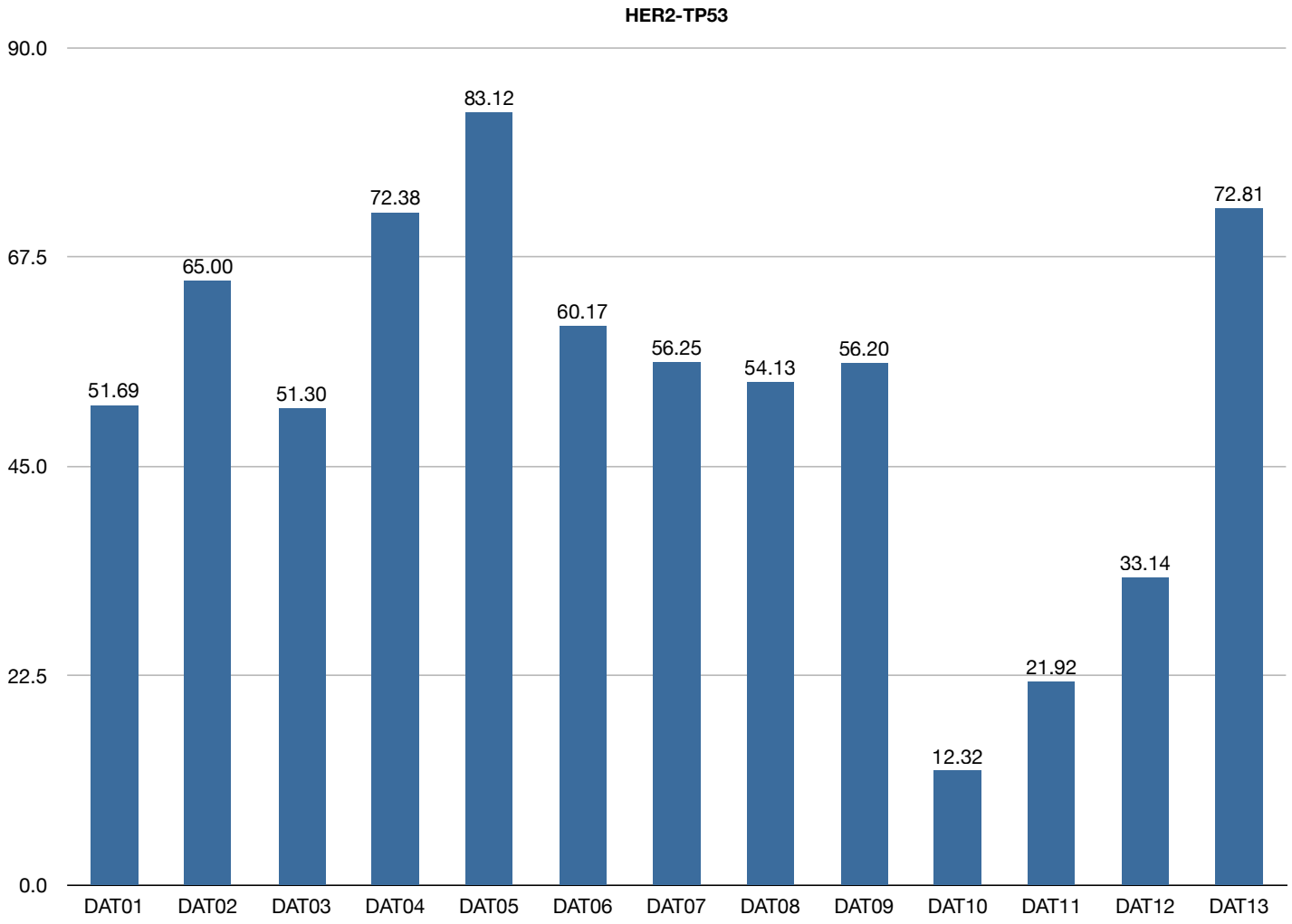

Figure 14: Predicted correlated variation (expressed in percentage) between HER2 and TP53 in the considered datasets.

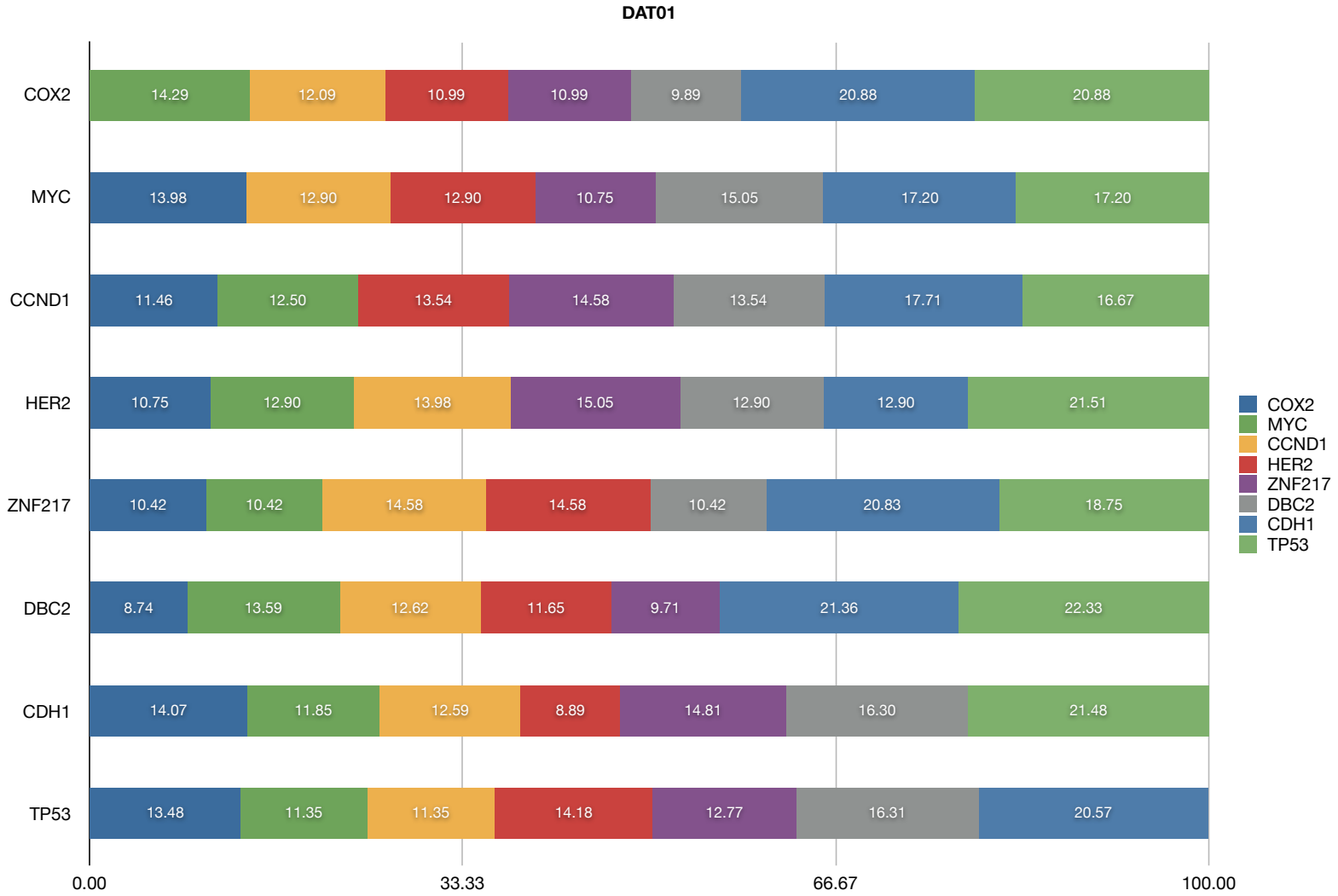

Figure 15: Predicted gene-driven correlated variation (expressed in percentage) in DAT01.

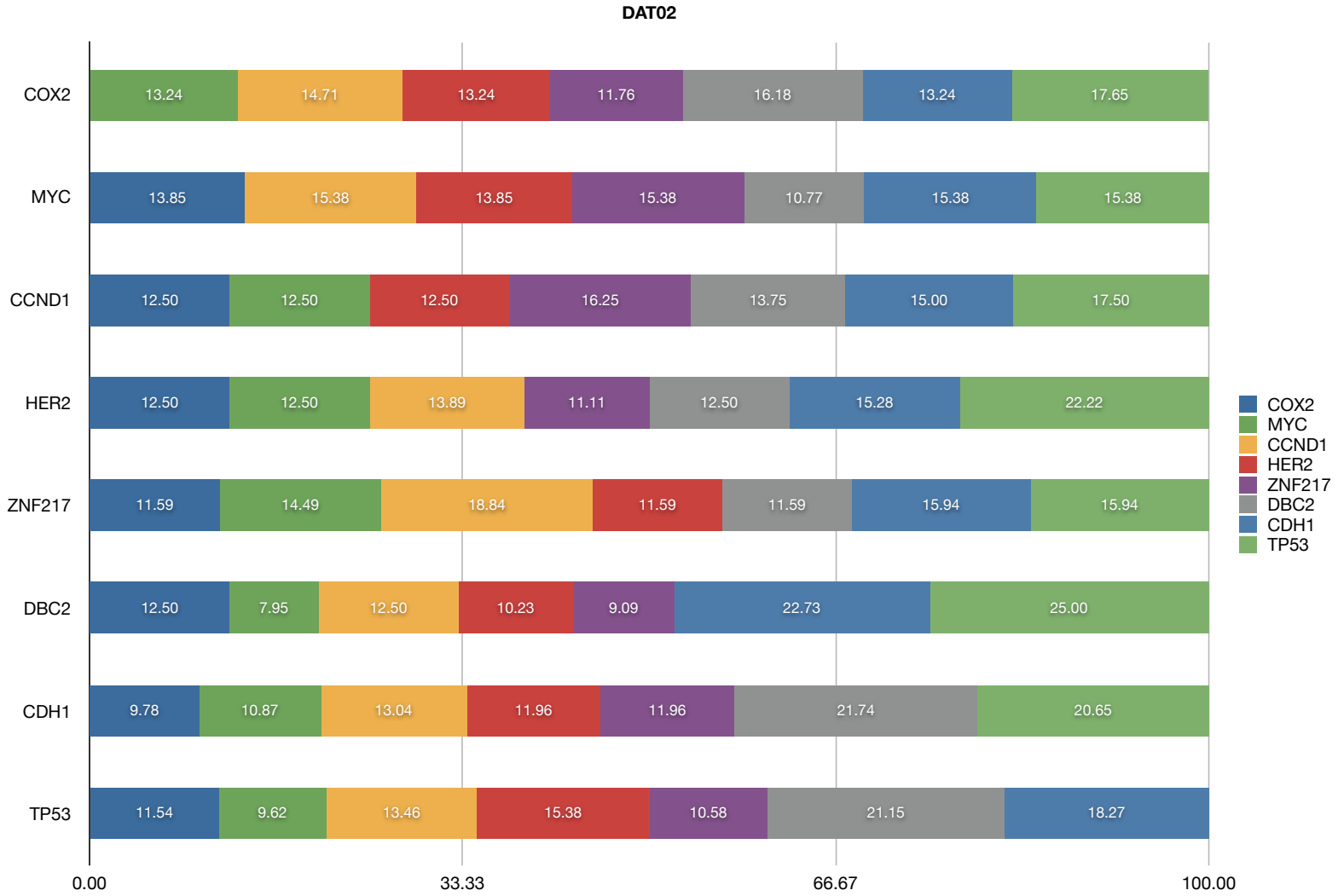

Figure 16: Predicted gene-driven correlated variation (expressed in percentage) in DAT02.

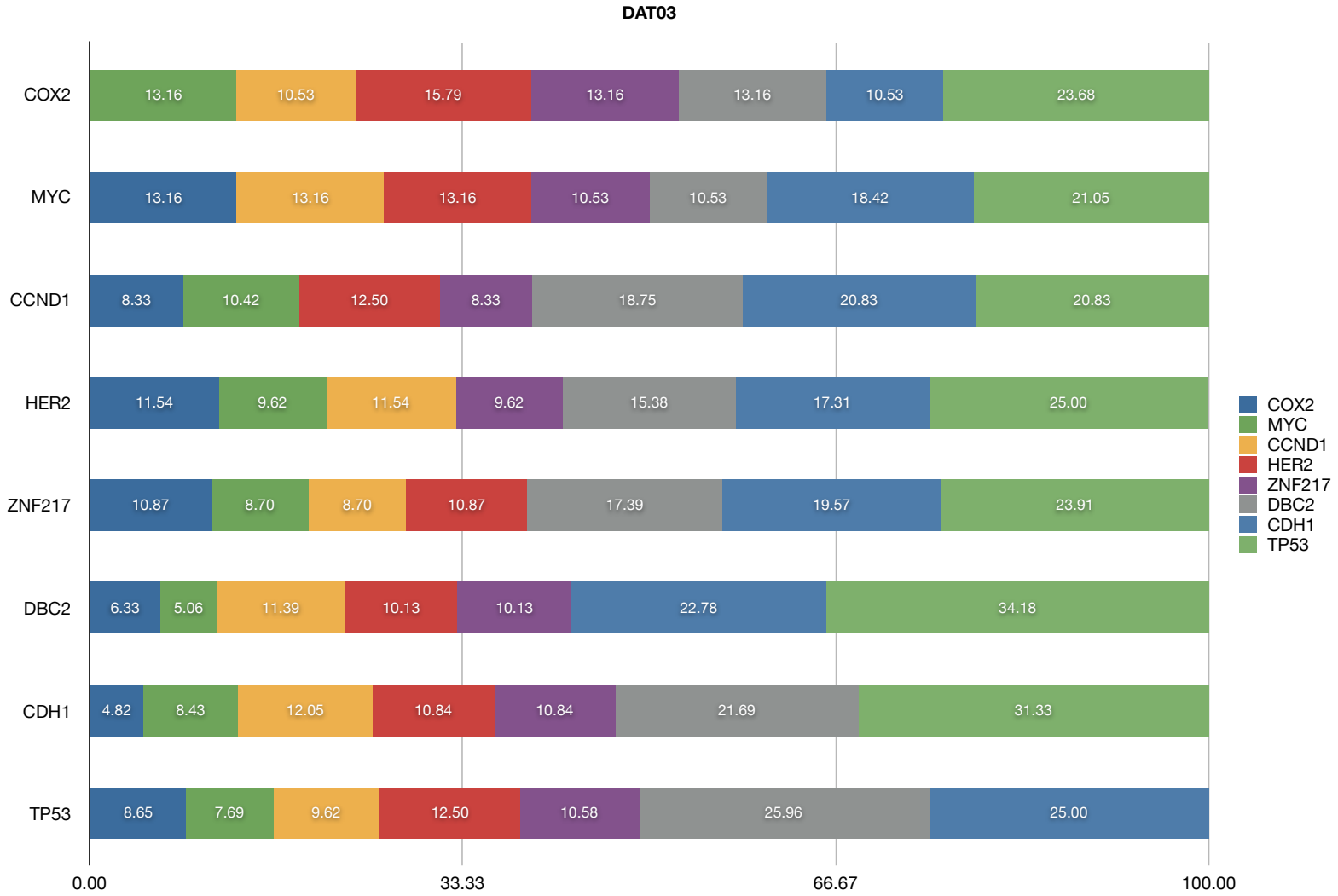

Figure 17: Predicted gene-driven correlated variation (expressed in percentage) in DAT03.

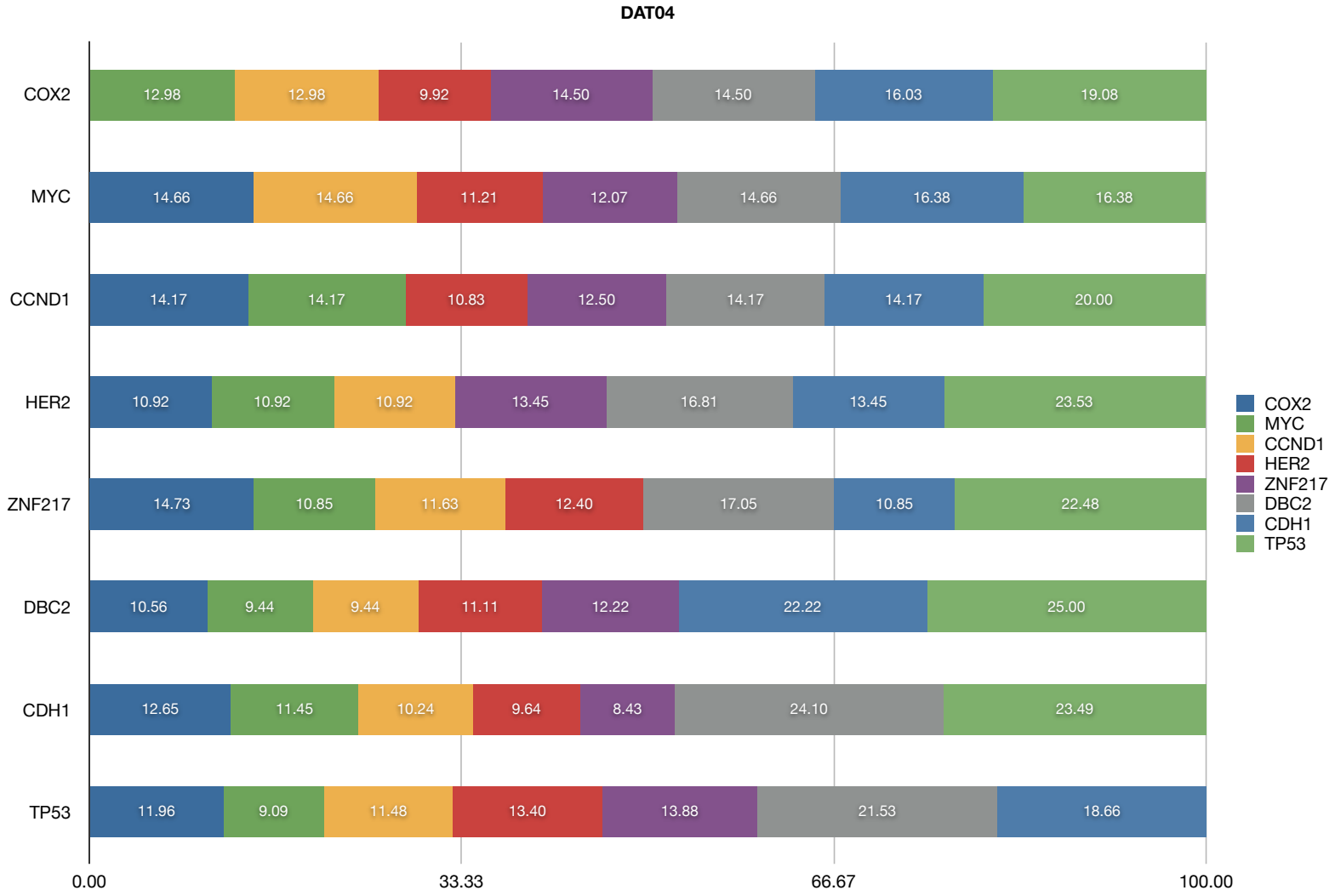

Figure 18: Predicted gene-driven correlated variation (expressed in percentage) in DAT04.

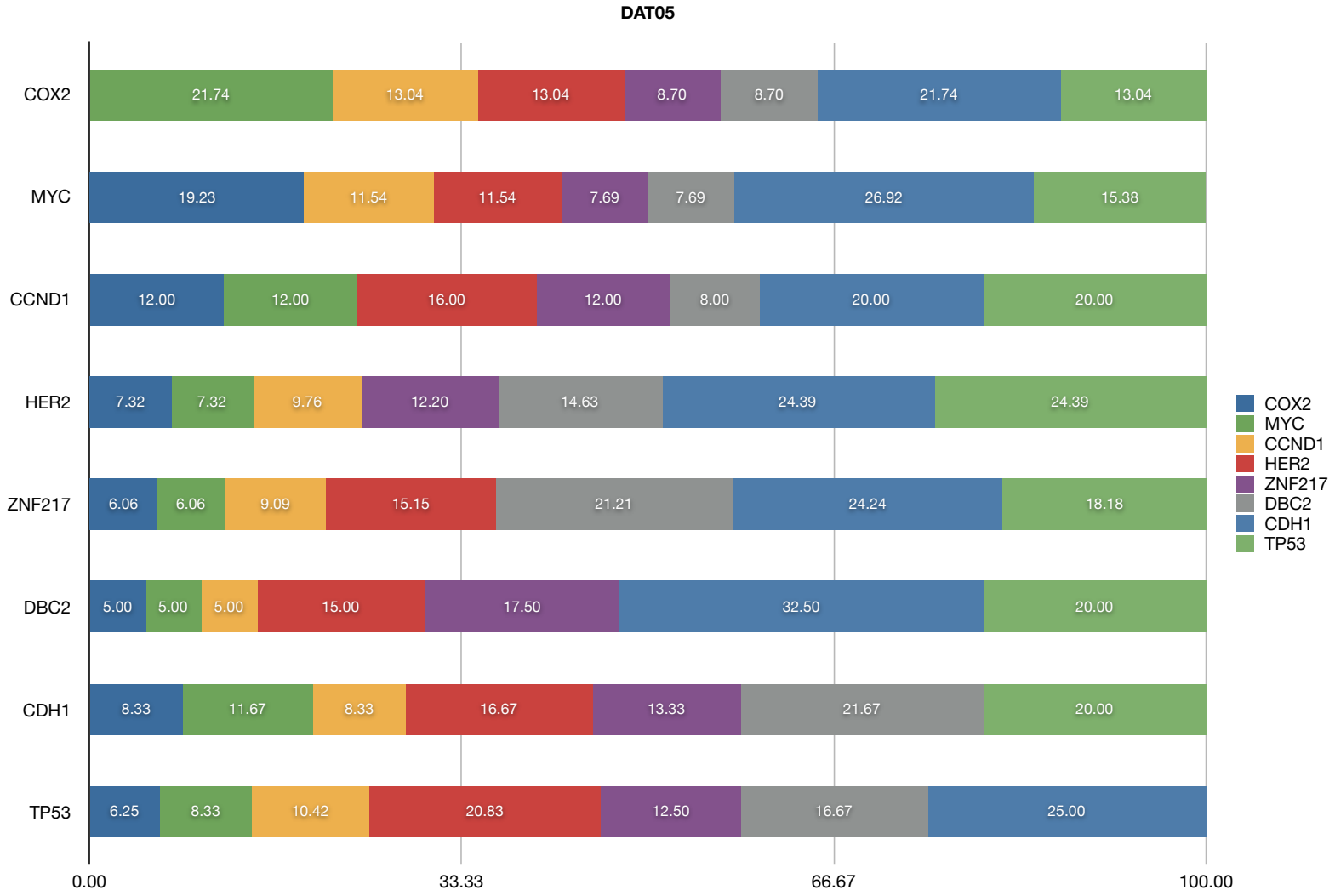

Figure 19: Predicted gene-driven correlated variation (expressed in percentage) in DAT05.

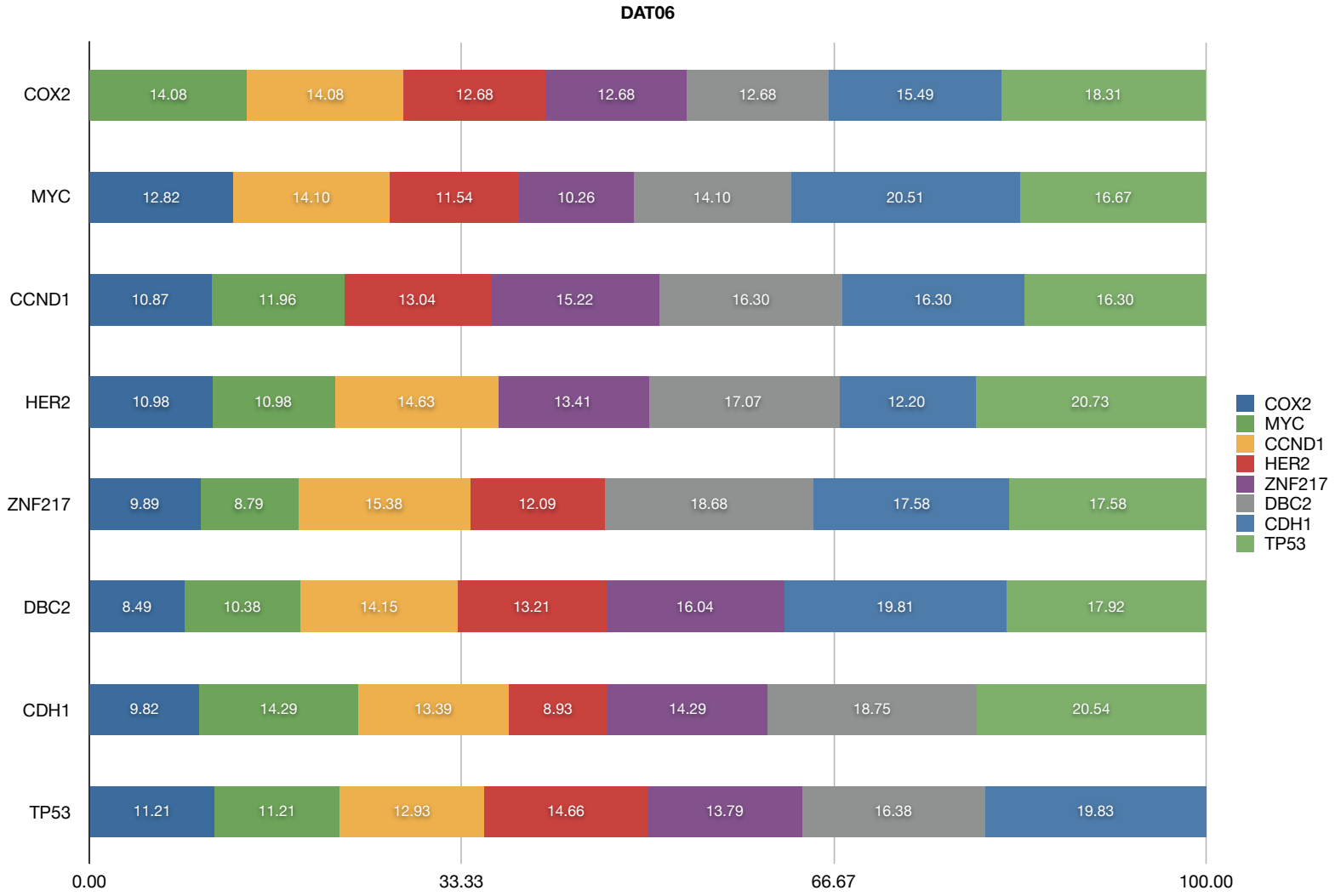

Figure 20: Predicted gene-driven correlated variation (expressed in percentage) in DAT06.

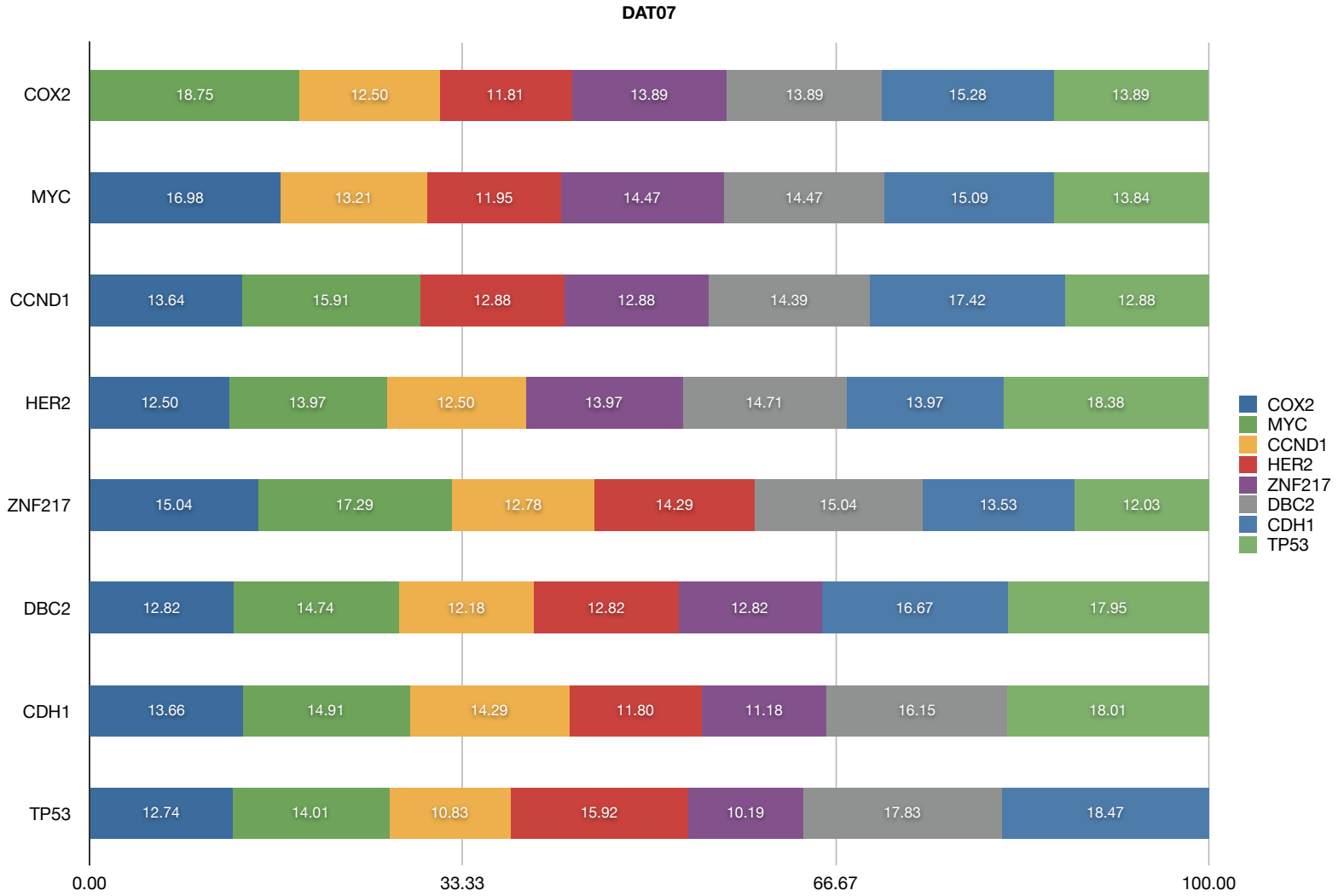

Figure 21: Predicted gene-driven correlated variation (expressed in percentage) in DAT07.

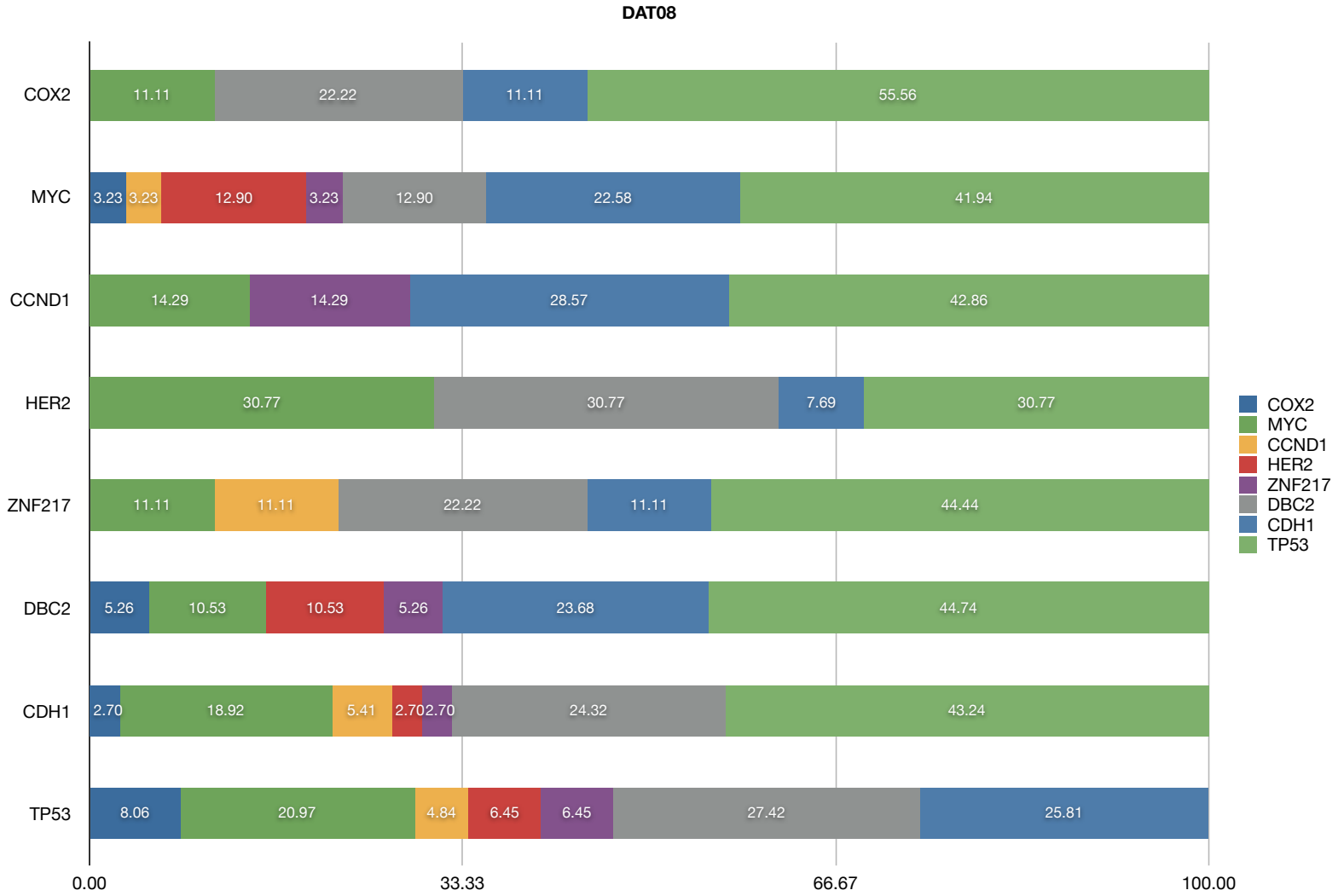

Figure 22: Predicted gene-driven correlated variation (expressed in percentage) in DAT08.

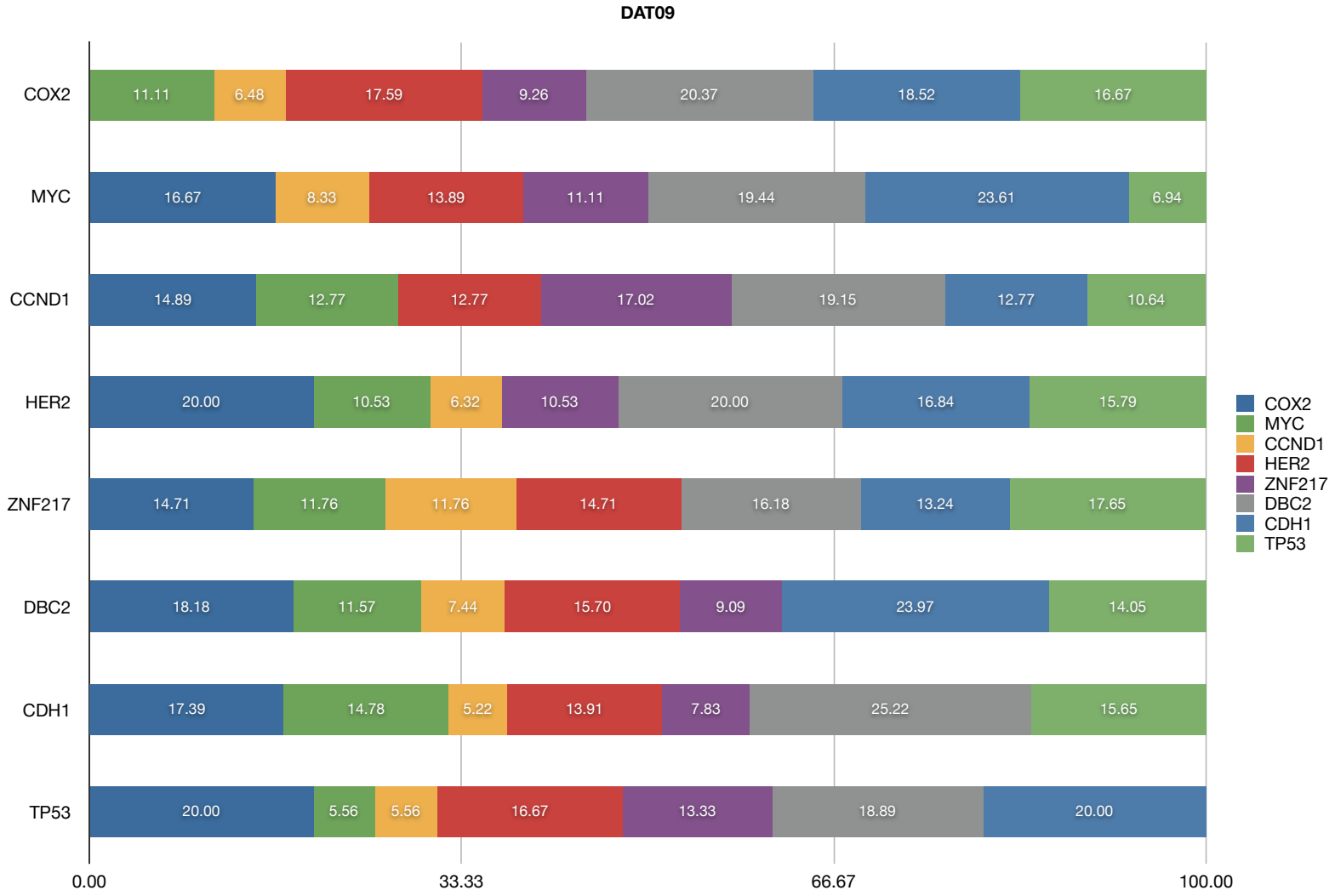

Figure 23: Predicted gene-driven correlated variation (expressed in percentage) in DAT09.

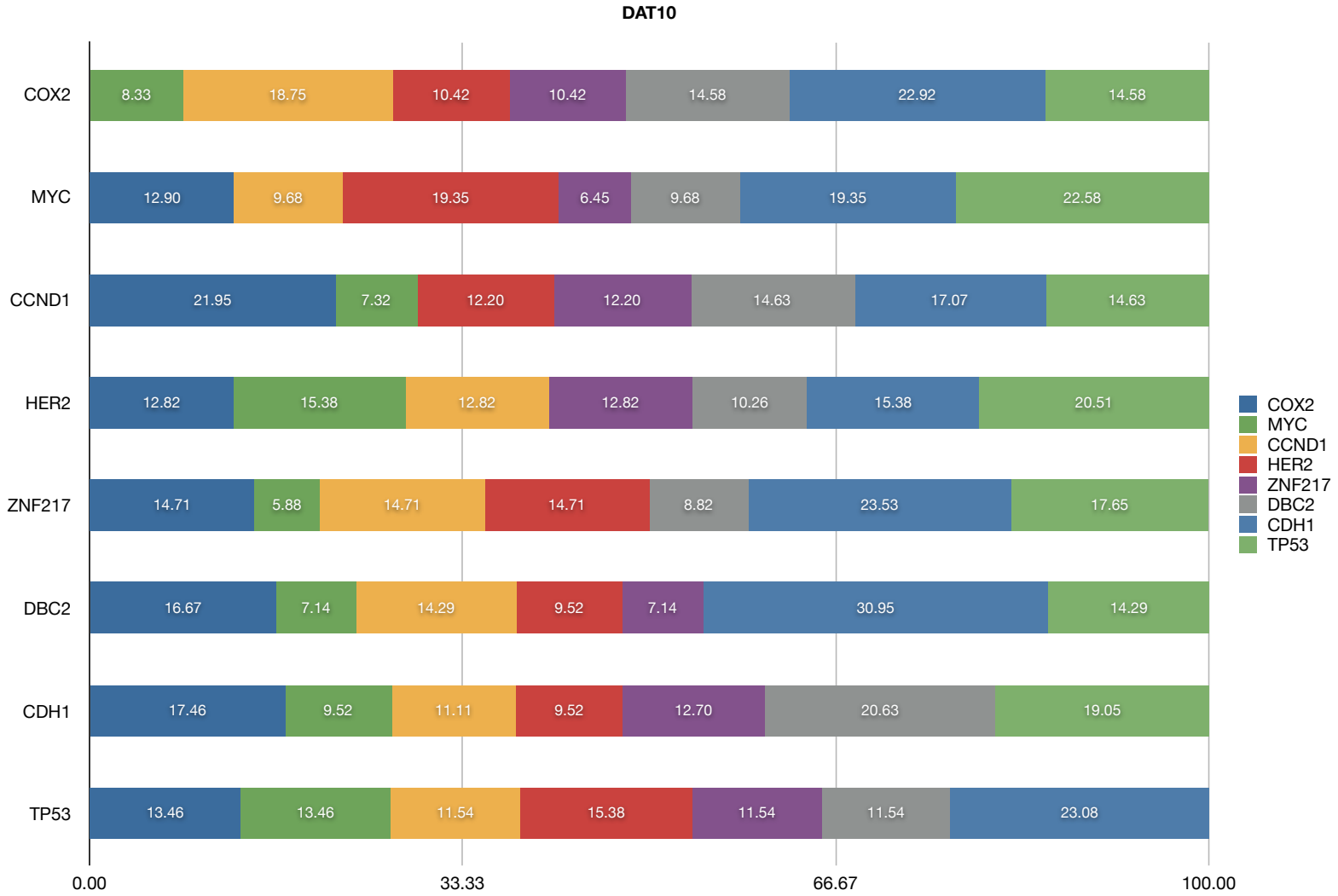

Figure 24: Predicted gene-driven correlated variation (expressed in percentage) in DAT10.

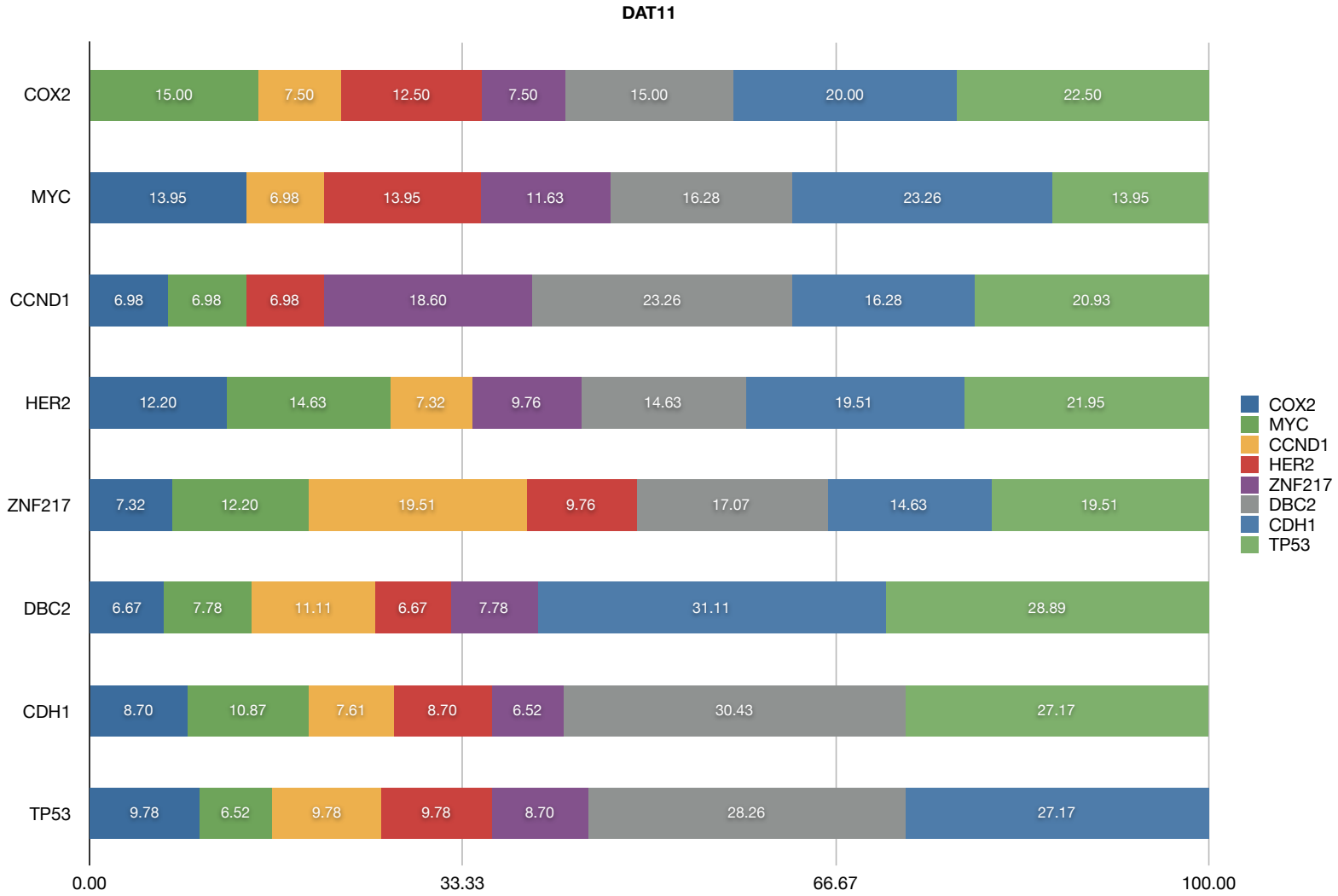

Figure 25: Predicted gene-driven correlated variation (expressed in percentage) in DAT11.

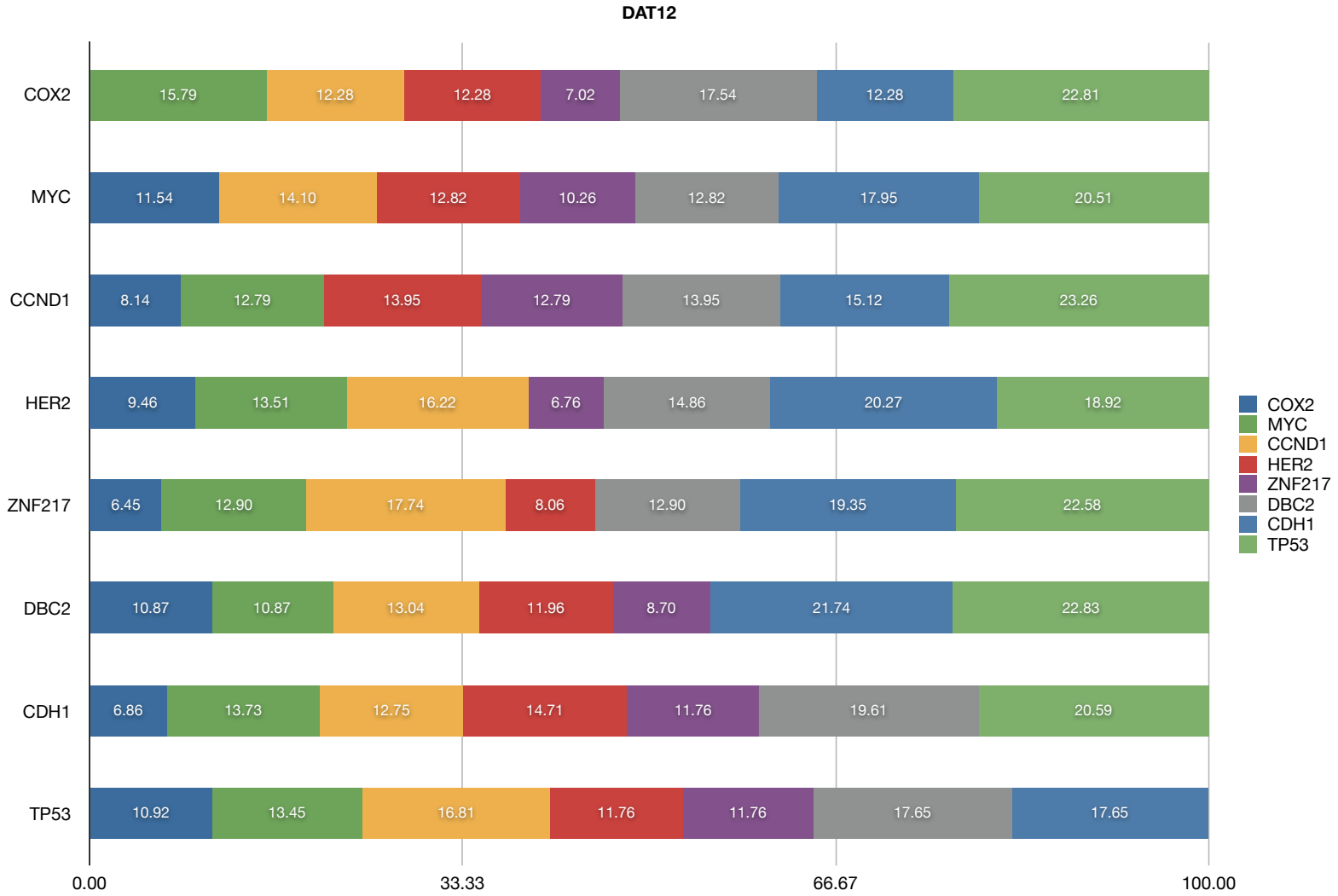

Figure 26: Predicted gene-driven correlated variation (expressed in percentage) in DAT12.

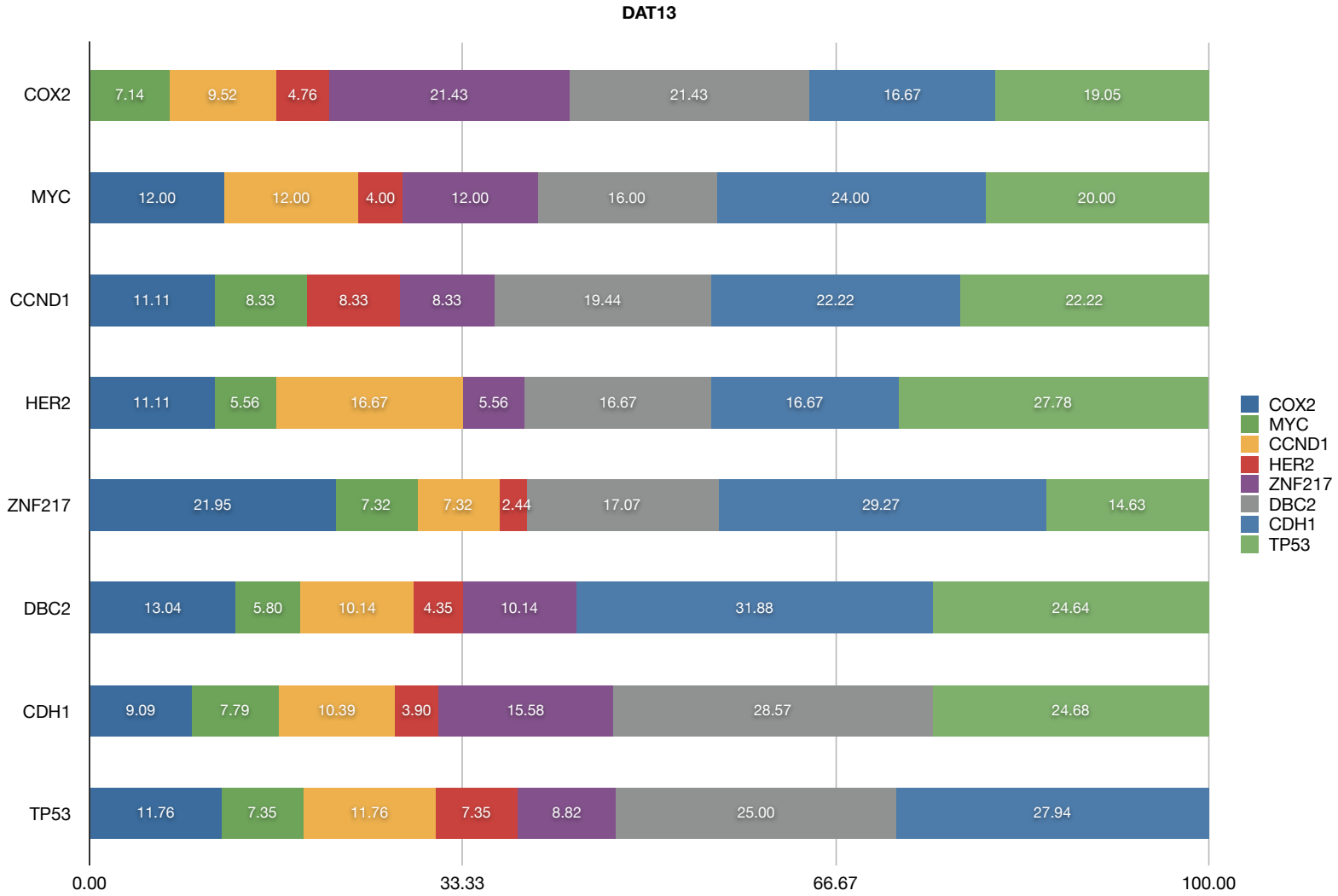

Figure 27: Predicted gene-driven correlated variation (expressed in percentage) in DAT13.