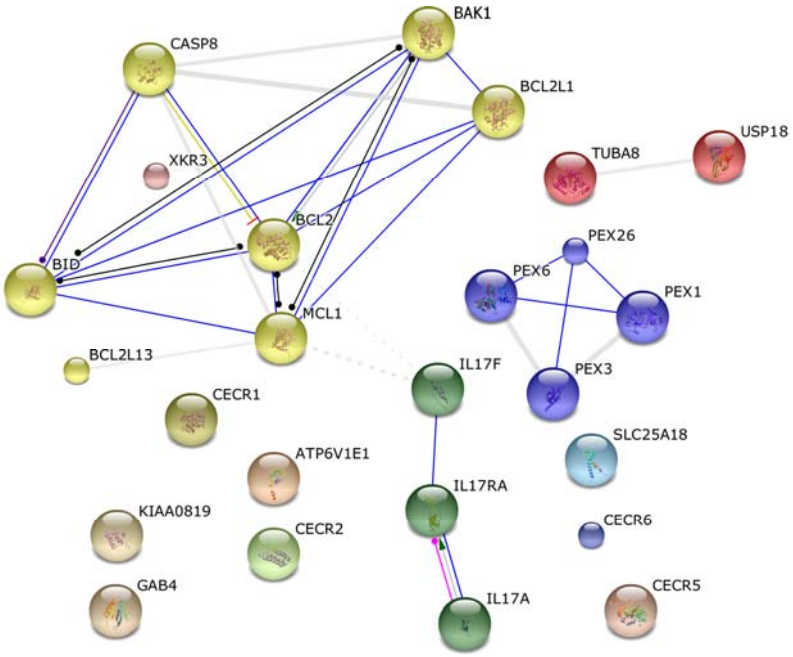
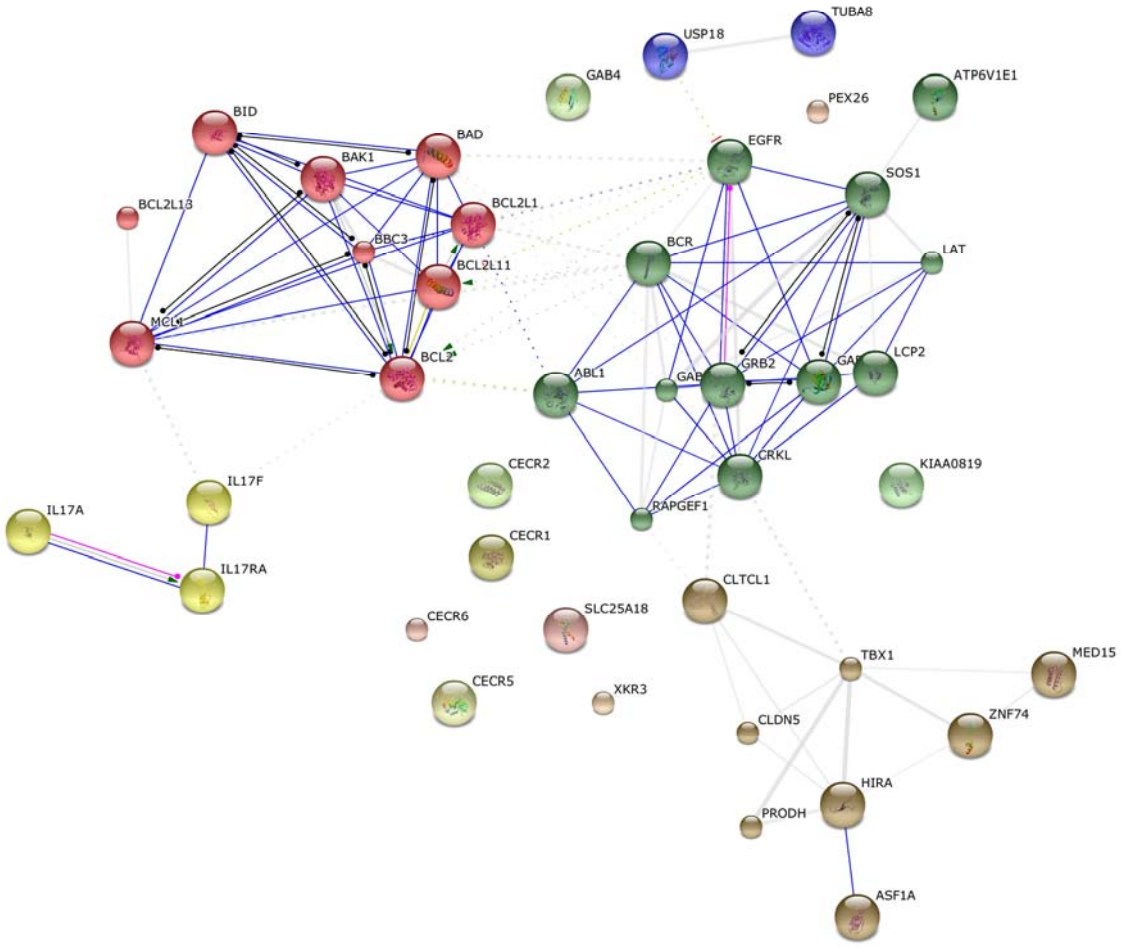
Supplementary figure and Network analysis



Supplementary Fig. 1 Protein-protein interaction network of the duplicated genes in the present case, *XKR3*, *GAB4*, *CECR7*, *IL17RA*, *CECR6*, *CECR5*, *CECR1*, *CECR2*, *SLC25A18*, *ATP6V1E1*, *BCL2L13*, *BID*, *MICAL3*, *PEX26*, *TUBA8* and *USP18* and proteins directly associated to the corresponding proteins as constructed with STRING 9.0. The BID, BCL2L13 and PEX26 proteins are found associated to functionally related proteins. Note that BID and BCL2L13 proteins were connected indirectly via MCL1 protein in the same cluster. Note: please consider that the image is a 2-Dimension projection of the network.



Supplementary Fig. 2. Protein-protein interaction network that includes the duplicated genes in the present case plus a set of genes randomly selected (*PRODH*, *CLTCL1*, *CLDN5*, *ZNF74*, *MED15*, *HIRA*, *TBX1*, *BCR*) related to other described distally duplicated regions in patients CES. Noticeably, USP18 (connected with TUBA8) protein is found functionally associated to EGFR (epidermal growth factor receptor) protein which in turn is found within the BCR protein cluster. Clusters of BCR and BID are functionally related via MCL1, BCL2L1-ABL1 proteins. Note: please consider that the image is a 2-Dimension projection of the network.

Network analysis

The aim to perform this analysis of the architecture of protein-protein interaction network was to search new insights regarding the underlying mechanisms to variability of CES phenotypes. That is, we believe that possible positive or negative interaction effect among certain genes of any duplication could account for the discrepancies related to correlation no between size and/or genetic content of duplication and the CES phenotypes. This notion is motivated by the observation that *MIL1* and *BID* have opposite functions into the same pathway (inhibition and activation, respectively) and that a probable antisense regulation among them was suggested [Kanehisa and Goto, 2000; Footz et al., 2001; cited in text].

We selected a set of 16 proteins corresponding to the list of duplicated genes and generated the network of protein-protein interactions associated to them. In addition, we included in the analysis eight distal genes randomly selected related to larger duplicated regions in other CES patients. To generate the network, we used the STRING 9.0 database [Jensen et al., 2009] to first retrieve the proteins directly associated to each of the 16 or 24 (16 + 8) selected proteins, and then to form the corresponding network (Supplementary figs. 1 and 2, respectively). The majority of the associations between proteins results from text mining, which assumes a functional similarity between connected proteins, whilst the remaining edges of the network represent either binding, activation, inhibition or co-expression associations between corresponding protein pairs (only the association above a 0.4 confidence were retained). To determine the clusters to which they belong, MCL algorithm (Markov Clustering) was used.

Various proteins were strongly clustered in two main clusters according their functional relation. Other two small clusters were observed under the same basis. Interestingly, cluster

of BID and BCL2L13 proteins is functionally associated found to BCR, ATP6V1E1 proteins cluster one, mainly via BCR, ABL1, EGFR and MCL1, BCL2, BCL2L11, BCL2L1 proteins (Supplementary fig. 2). The PRODH, CLTCL1, CLDN5, ZNF74, MED15, HIRA, TBX1 proteins were clustered in a same cluster which in turn was functionally associated to BCR cluster via EGFR, GRB2, RAPGEF1-CLTCL1 and TBX1-CRKL. Noticeably, USP18 (connected with TUBA8) was functionally associated to the BCR cluster via USP18-EGFR interaction and in turn, EGFR is also connected with BID cluster via BCL2L1 (Supplementary fig. 2). The IL17RA cluster was connected with that of BID and BCL2L13 via IL17F-MCL1.

Increasing the complexity level of the network (more than 40 nodes), we could see association between the BCR and TBX clusters via interactions between PRODH and TP53 proteins (network not showed). According to the used network level, CECR proteins had no interactions among themselves and neither with other proteins. It is due to the low complexity level of the network. Then, sub-networks for CECR1 and CECR2 were performed (data not showed). We observed that CECR1 is connected to BCL2L13 cluster via BCL2-TSPO. For CECR2, at level of network complexity (40 nodes), we did not observe any association with duplicated genes. However, interesting connections with gene clusters related to fibroblast growth factor, apoptosis, cytoskeleton and microtubule organization were observed; among them were the *MAP1S*, *UXT*, *LRPPRC*, *BUB1*, *BUB3*, *BUB1B* genes. A sub-network with both CECR1 and CECR2 proteins, showed an association between their clusters via chromatin remodeling genes such as *SMARCA1* and *TADA2A*.

Despite that protein-protein interaction network analysis is complex and represents just an approximation from biological phenomena, we can appreciate certain interesting connections among duplicated genes. Thereby, we speculate that such interactions could positive or negatively be modified (depending on the case) when varying gene dosage; consequently it could influence the phenotypes expressivity. However, it is still to be demonstrated and understood.

References

Jensen et al. Nucleic Acids Res. 2009, 37(Database issue):D412-6.