Background hypothesis. We hypothesized the following. **i)** Physical interactions between proteins is an indication of functional association. Proteins close to one another in a human Protein Protein Interaction (PPI) network are more likely to be functionally associated, than proteins separated by longer distances. Distance is defined as a combination of the number of interactions separating the proteins and the reliability of each connecting interaction. **ii)** Genes responsible for overlapping or identical phenotypes are likely to code for proteins that are related by participating in the same cellular pathway, cellular structure or molecular complex. **iii)** For a given phenotype, in this case Congenital Diaphragmatic Hernia (CDH), the causal gene can be identified amongst non-causal genes in a causal CNV by assessing the proximity of its protein product to proteins coded by genes from other causal CNVs. We expect the causal proteins to be significantly closer to each other in the ppi network than what would be expected by random. We have formalized this hypothesis into a model named CNVconnect. A particular application of CNVconnect, which was used in this work, employs phenotype associated "bait" genes rather than CNVs. Computationally, phenotype associated genes are equated to phenotype associated CNVs containing a single gene.
We describe below the generalized form of the CNVconnect algorithm.

Protein protein interaction data. We have previously described the generation of a human protein-protein interaction network, by integrating data from the major interaction databases (MINT, BIND, Intact, GRID, DIP, Reactome, KEGG, HPRD and more) and across organisms with strict orthology requirements. The network now has ~400.000 interactions between 12,500 human proteins. This network is called InWeb and is described in detail in Lage et al., 2007(Lage, Karlberg et al. 2007).

Determining pair wise path lengths between interacting proteins in InWeb. Path lengths between directly interacting proteins in InWeb, were determined by first assigning all pair wise interactions a probabilistic confidence score based on their reliability and reproducibility across publications in Medline. The score is described in detail and thoroughly benchmarked in Lage et al., 2007(Lage, Karlberg et al. 2007). The more reliable the interaction between proteins P1 and P2, the more likely we believe they are to be functionally connected. The score ranges from 0 to 1, where interactions scoring close to 1 are highly reliable. In this work, we transformed the confidence score into a path length by taking the negative log of the confidence score (see below), which ensures that protein pairs with connected by interactions of high confidence are connected by a short path length and vice versa (see **Figure 1a**).

$$PPL(P1, P2) = -\log(CS(P1,P2))$$

Determining pair wise path lengths between protein pairs that directly interact in InWeb | CS is the probabilistic confidence score between proteins P1 and P2. PPL(P1,P2) denotes the path length between proteins P1 and P2

Determining the shortest path lengths between all proteins in InWeb. An all paths shortest path (APSP) matrix was calculated between all proteins (that either directly interact, or indirectly interact through common interaction partners) using the pair wise path lengths (PPLs) and the the Floyd-Warshall algorithm for finding all path shortest paths in a weighted graph (see http://en.wikipedia.org/wiki/Floyd-Warshall_algorithm

and **Figure1b**). This matrix contains the shortest pathlength (SPL) between all proteins in InWeb.

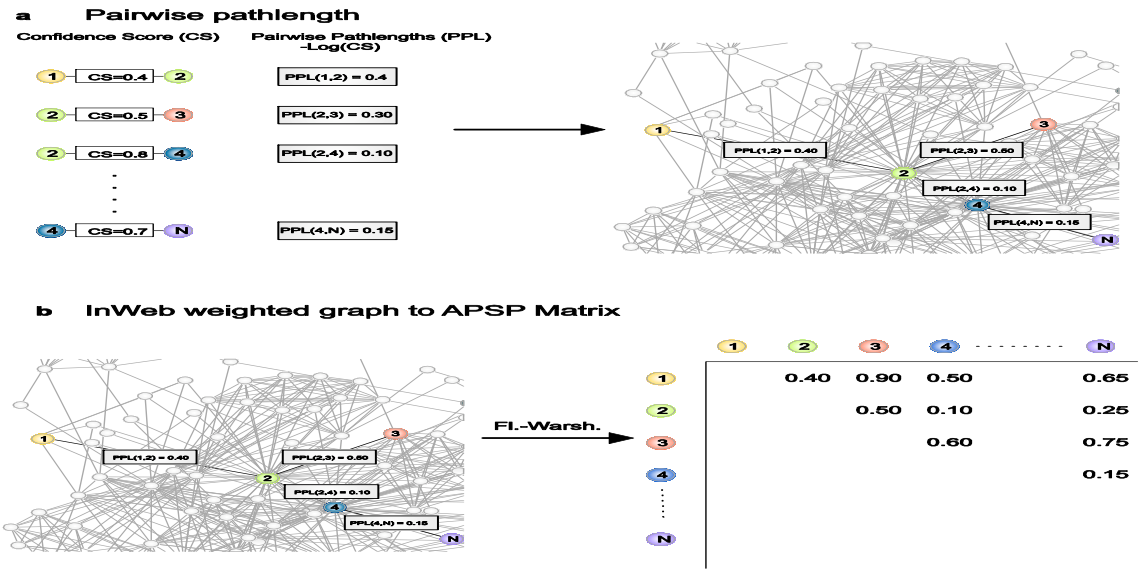
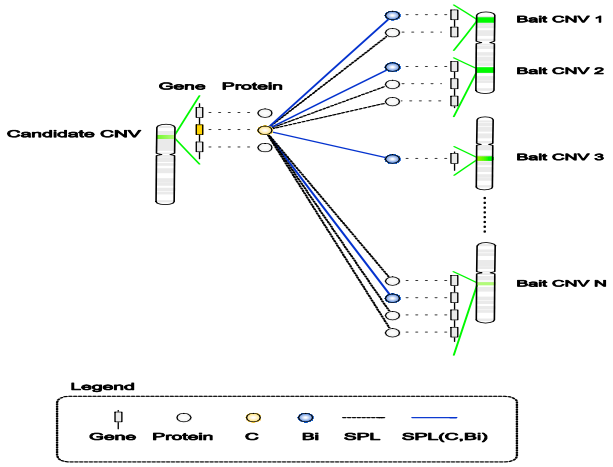


Figure 1 | Calculating an APSP matrix from InWeb. Pair wise confidence scores (CS) are turned into pair wise path lengths PPL by taking the negative log of the CS, as seen in **a**). Hereby InWeb is turned into a weighted graph where the weights are path lengths between the proteins indicating their likelihood to be functionally related. The Floyd-Warshall algorithm is then used to calculate an all paths shortest path (APSP) matrix of distances between all proteins in InWeb **b**). This matrix contains the shortest path length (SPL) between all proteins in InWeb.

a



b

$$\mathbf{AC_N} = (\sum_{i=1}^{N} SPL(C, Bi)) / N$$

Figure 2 | Determining the average pathlength between a candidate and the bait CNVs. For a given candidate gene its protein product (C), is determined. For C all bait CNVs are identified. For a given bait interval i, the proteins coded by genes in the interval are identified. Then the protein with the shortest pathlength (SPL) to the candidate (C) is identified amongst the bait proteins from i. This protein is called Bi, and the shortest pathlength from C to Bi (SPL(C,Bi)) is determined in the APSP matrix. This step is carried out for all N intervals **a**). The average pathlength from C to N intervals (AC_N) is then calculated in a straight forward manner, as see in **b**). This step is carried out for all genes in all CNV regions.

Determining the average path lengths connecting proteins in distinct causal CNVs. For each set of CNVs identified in a given study we determine the average pathlength from proteins coded by genes in the CNVs to all other proteins coded by genes in other CNVs. Pathlengths between proteins from the same CNV are not considered in this work. Specifically, the gene we are scoring is named the candidate gene and codes for a protein denoted the candidate protein (C). The CNV region of the candidate gene and protein is named the candidate CNV. All other regions are named bait CNVs. The genes in bait CNVs are called bait genes and their protein products are named bait proteins. A given candidate protein was scored by finding the average shortest path length to N bait intervals (AC_N), where 1<=N<=Q and Q is the amount of CNVs found in the study in question not including the candidate CNV. Under the hypothesis that only one gene per interval is responsible for the phenotype, the shortest path from a candidate protein C to a given bait interval i was determined by nearest bait protein (Bi) from interval i (see

Figure 2). The shortest path length C to Bi (SPL(C, Bi)) was determined from the APSP matrix calculated earlier.

Determining the significance of the average path between a candidate protein and the bait intervals. To determine the significance of the average pathlength of a given candidate protein to N bait intervals ($ProbAC_N$), we empirically estimated the probability of observing an equal or shorter path length in InWeb between C and a random set of CNVs modelled on the bait intervals fromt the data set in question (see below for details).

$ProbAC_N = Freq (AC_N > AC_{NRandom})$

The probability estimates were calculated from the frequency of randomized trials that result in a shorter average path length between a candidate protein and N random bait intervals. For example if the candidate 3 times of 1.0e5 got a shorter average pathlength to the N random CNVs than to N CNVs from the case data, **$ProbAC_N$** is 3/1.0e5 = 0.00003.

Making a random set of CNVs modeled on the case data.
If a CNV i from the case data has three genes i1, i2 and i3, these three genes were replaced by random genes i1Random, i2Random and i3Random. The random genes were chosen so their interaction profile matches the profile of the case genes. For example, i1Random was chosen so its interaction profile matches the profile of i1, meaning that if i1 has 10 interactions in InWeb, it was substituted by i1Random that had a similar amount of interactions. Finding random proteins with similar interaction properties as the case protein being substituted is important because some proteins have hundreds of biologically valid interactions, making them central in the network and thus close to a high proportion of the InWeb proteins compared to proteins that are not highly connected due to fewer biologically true interactions. Not taking this into account will give erroneous significance estimates. Once i1, i2 and i3 were replaced with i1Random, i2Random and i3Random, hereby creating a random interval iRandom, this replacement was carried out for the next interval in the case data until all CNVs had been replaced by random CNVs. This yields a set of random intervals with similar properties as the intervals in the original dataset. However, gene sets are not contiguous gene regions. Using this random data set the average path length of the candidate protein under consideration was calculated to all random bait intervals similarly to the method described above. After a pre-determined amount of randomizations (typically 1.0e5 to 1.0e7) Freq ($AC_N > AC_{NRandom}$) was calculated and $ProbAC_N$ determined.

Lage, K., E. O. Karlberg, et al. (2007). "A human phenome-interactome network of protein complexes implicated in genetic disorders." Nat Biotechnol **25**(3): 309-316.