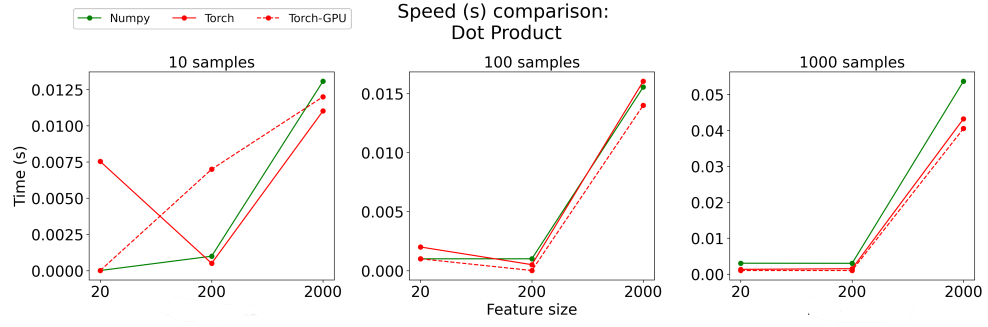


Supplementary Material

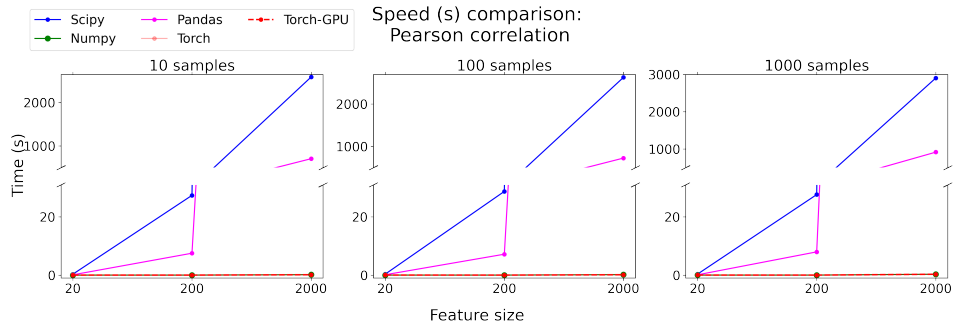
1 Personalized and Individual Specific Networks (ISNs)

1.1 Benchmarking correlation measures speed using different libraries

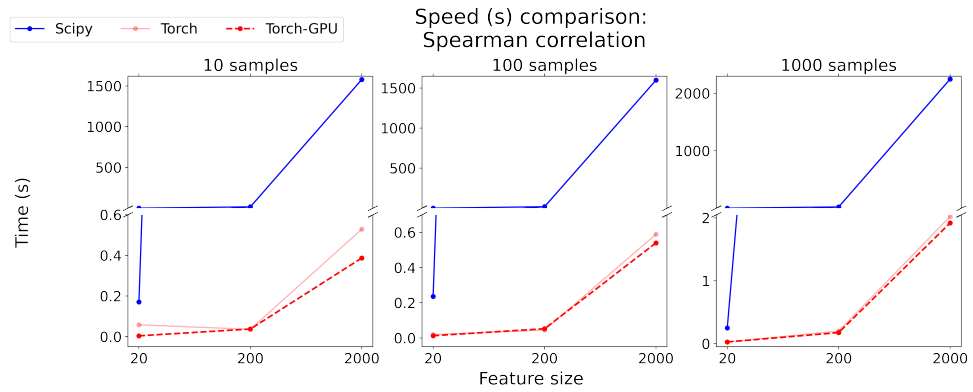
In this analysis (Figure S1), we compare the performance of different libraries over time using three types of correlation measures. The figure includes nine plots, each showing the performance results for one specific correlation measure. The horizontal axis shows the feature size used for evaluation, while the vertical axis shows the execution time in seconds. The correlation measures are Dot Product (Figure S1a), Pearson Correlation (Figure S1b), and Spearman Correlation (Figure S1c). These measures are arranged in rows, and different libraries are shown with different coloured lines in each plot. The plots help us see how efficient each library is when using the same correlation measures. The figure also shows the speed of the processing unit, with GPU computation shown by dotted lines and CPU computation shown by continuous lines. Notably, our library allows users to compute the ISNs with a custom metric. This plot shows the general trends observed with existing options (Numpy, Scipy, Pandas, PyTorch). Since our library is based on PyTorch, all the metrics we provide are Torch-based, resulting in the best overall performances. This figure effectively highlights the time-based performance of various libraries, giving a clear understanding of their efficiency in handling different correlation measures.



(a)



(b)



(c)

Figure S1: Performance comparison of metrics implemented in various Python libraries, showing the execution times across different sample sizes with varying feature sizes.

To conclude this section, we compared the fastest way to calculate Pearson Correlation in Python (using PyTorch) with the Incremental Pearson method. We did this comparison using the same number of features and samples on both CPU and GPU since both methods can run on these platforms. As shown in Figure S2, the Incremental Pearson method consistently remains the fastest computational metric.

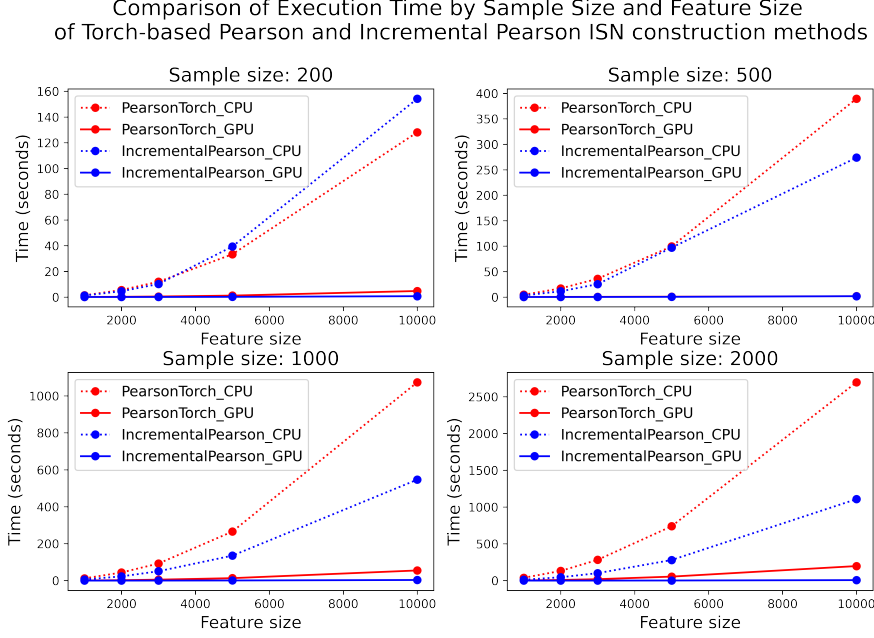


Figure S2: PyTorch-based Pearson correlation comparison with Incremental Pearson method on CPU and GPU.

1.2 Advantages of ISN-tractor for Computing Individual Specific Networks (ISNs)

Figure S3 shows the main features of ISN-tractor. These include the incorporation of pooling measures, utilization of Single Nucleotide Polymorphisms (SNPs) as input data, employment of both CPU and GPU processing, and integration of the Human Reference Interactome (HuRI) for accurate SNP-to-gene mapping. As a result, ISN-tractor produces biologically informed Integrated SNP Networks (ISNs) that capture the underlying biological mechanisms, offering a comprehensive understanding of the genomic landscape. This framework holds great promise for various applications. For instance, in disease-gene discovery, ISN-tractor can identify genetic variants consistently associated with specific diseases across different biological pathways. This

capability enables researchers to prioritise candidate genes for further experimental validation, potentially uncovering novel disease mechanisms.

In drug discovery, ISN-tractor facilitates the identification of drug targets by revealing key genes or pathways that are perturbed in disease contexts. By analyzing ISNs derived from patient-specific genomic data, researchers can pinpoint drug targets that are tailored to individual genetic profiles, thereby advancing personalized medicine approaches.

Furthermore, ISN-tractor supports functional genomics studies by elucidating how genetic variations influence molecular interactions within biological networks. This insight aids in understanding complex phenotypes and their underlying genetic basis, paving the way for targeted therapeutic interventions.

In summary, ISN-tractor not only provides a robust framework for constructing ISNs but also empowers researchers with valuable insights into disease mechanisms, drug targets, and functional genomics, thereby contributing to advancements in biomedical research and precision medicine.

Advantages of ISN-Tractor over lionessR

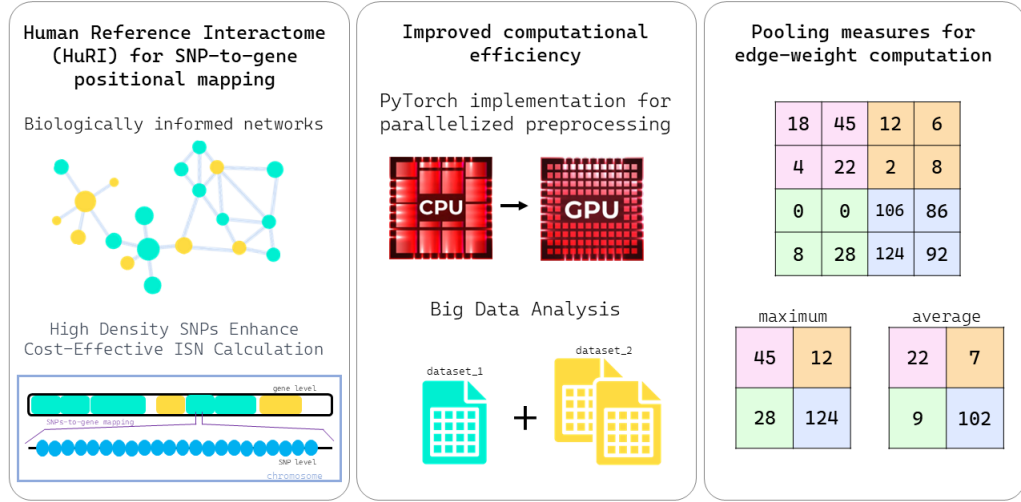


Figure S3: ISN-tractor: A Comprehensive Scheme

1.3 The vectorized Incremental Pearson Correlation algorithm

In section 2.2.1 of this article, we provided the incremental Pearson algorithm on one single edge in the ISN, summarized in Figure S4. Here, we present the final version of this algorithm, optimised to compute the Pearson correlations for all node pairs at once, by vectorizing the previous calculation. This is the version implemented in ISN-tractor because it is more suitable for parallelization.

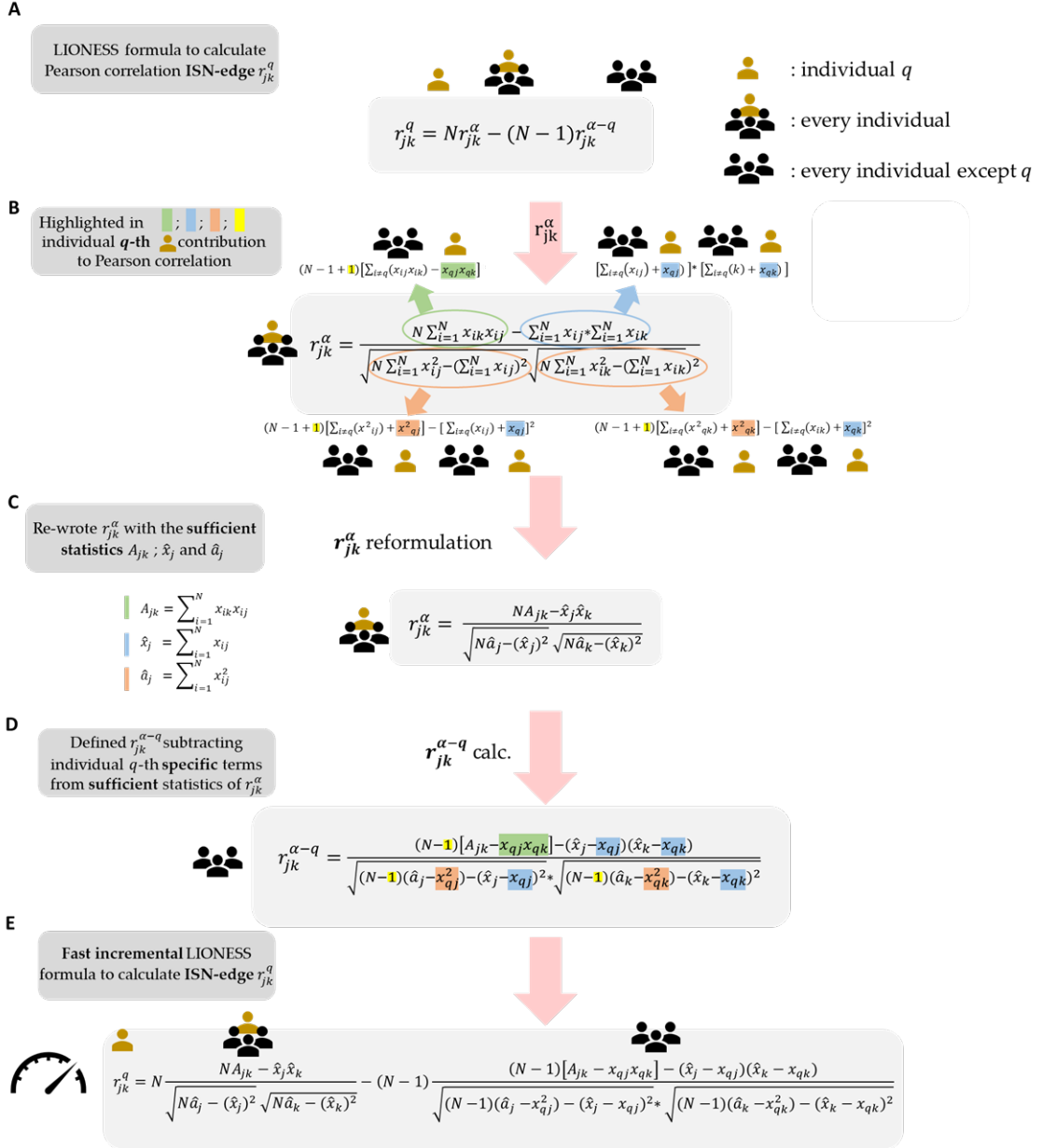


Figure S4: **A** shows the LIONESS ISN formula, for a single edge. **B** shows the canonical Pearson correlation formula, with highlighted the contribution of a generic individual q to the formula, for calculating one single edge between genes j and k . **C** shows the formula in B reorganised to focus on the sufficient statistics. **D** shows the formula when individual q is excluded from the pool. In detail, the q -th contribution highlighted in B is subtracted from the sufficient statistic shown in C. Finally, **E** shows the fast incremental formula to calculate the ISN with the LIONESS formula, equivalent to the one in A.

To explain the vectorized algorithm, we divide Eq. 1, calculating a single edge of the leave-one-out network, into three parts and propose a parallelizable implementation for each.

$$r_{jk}^{\alpha-q} = \frac{(N-1)(A_{jk} - x_{qj}x_{qk}) - (\hat{x}_j - x_{qj})(\hat{x}_k - x_{qk})}{\sqrt{(N-1)(\hat{a}_k - x_{qk}^2) - (\hat{x}_k - x_{qk})^2} \sqrt{(N-1)(\hat{a}_j - x_{qj}^2) - (\hat{x}_j - x_{qj})^2}} \quad (1)$$

1.3.1 First term

The first term of Eq. 1 is $(N-1)(A_{jk} - x_{qj}x_{qk})$. To derive a vectorized (easy to parallelize) version, we define: $M^q = (x_{q.}x_{q.})^T$, with M^q a matrix of dimension $p \times p$ in which element (j, k) is $x_{qj}x_{qk}$. Then, we rewrite the first term in a vectorized way as: $(N-1)(A_{p \times p} - M^q)$, a $p \times p$ matrix, with single element, for genes j and k , $(N-1)(A_{jk} - x_{qj}x_{qk})$.

1.3.2 Second term

To vectorize the second term of Eq. 1, $(\hat{x}_j - x_{qj})(\hat{x}_k - x_{qk})$, we define $\hat{x}^q = \hat{x} - x_{q.}$ of dimension $p \times 1$, with $x_{q.} = X_{q.}$ the q -th row of matrix X . Then, we rewrite the second term in a vectorized way as: $\hat{x}^q(\hat{x}^q)^T$, a $p \times p$ matrix, with single element, for genes j and k , $(\hat{x}_j - x_{qj})(\hat{x}_k - x_{qk})$. The key operation, the vector multiplication, can be fastly executed on both CPUs and GPUs.

1.3.3 Third term

The third and last term of Eq. 1 we consider is the denominator. In the denominator, there are two independent square roots, one for each gene involved in the target edge. Here, the formula does not need to know for which two genes j and k we are calculating the Pearson correlation; hence, we can compute and pre-store the denominator separately for each gene, and combine it for every two pairs of genes, reducing the complexity from $\Theta(p^2)$ to $\Theta(p)$. Taking into account $\hat{x}^q$ as defined in the previous section and for each single gene, the formula to calculate the denominator is:

$$Den_{p \times 1}^q = \sqrt{(N-1)(\hat{a}_{p \times 1} - x_{q.}^2) - (\hat{x}^q)^2}, \quad (2)$$

for each individual q . Thus, the full denominator becomes: $Den_{p \times 1}^q (Den_{p \times 1}^q)^T$, a $p \times p$ matrix, whose element for genes j and k is $\sqrt{(N-1)(\hat{a}_j - x_{qj}^2) - (\hat{x}_j - x_{qj})^2} * \sqrt{(N-1)(\hat{a}_k - x_{qk}^2) - (\hat{x}_k - x_{qk})^2}$

Combining together the three previous terms, the vectorized version of the **incremental** Pearson correlation formula is :

$$r_{p \times p}^{\alpha-q} = \frac{(n-1)(A_{p \times p} - M^q) - (\hat{x}^q \hat{x}^q)^T}{(Den_{p \times 1}^q * (Den_{p \times 1}^q)^T)} \quad (3)$$

2 Optimizing ISNs with underlying biological networks

The Human Reference Interactome (HuRI) is a map of human binary protein interactions that serves as a reference for understanding the complex network of protein interactions involved in biological processes and disease [1]. Utilizing HuRI enhances the accuracy of biological findings and deepens the understanding of human biology, enabling the development of new diagnostic and therapeutic strategies. HuRI has various applications in bioinformatics, including improved understanding of biological processes and disease mechanisms, identification of new drug targets, and creation of bioinformatics tools for protein interaction analysis. By using HuRI as a reference, results are based on a solid foundation of human biology knowledge, improving their reliability and validity. ISN-tractor brings a novel approach to ISN computation by integrating the HuRI for SNP-to-gene positional mapping, resulting in biologically informed ISNs that better reflect the underlying biology compared to prior methods that produce fully connected networks. The integration of the HuRI allows for the incorporation of biologically relevant information such as Protein-Protein Interactions (PPI) and gene expression data, providing a more nuanced view of the genomic landscape and facilitating the identification of biologically meaningful connections between genes and SNPs. This leads to a more biologically informed representation of the genome.

2.1 Strategies to aggregate data at the node level

This study represents the first attempt to compute ISNs using SNP array as input data. SNPs are cost-effective and have genetic marker properties that allow for the computation of Individual Specific edge weights, which capture the impact of genetic variations on disease susceptibility and treatment response. Despite their large size and low variability, SNPs are valuable because they enable the identification of genes or genomic regions associated with specific traits due to their common points of variation in DNA. However, compared to gene expression-based data, additional steps are needed for the calculation of ISNs from SNP data. After imputing missing values, SNP-to-gene positional mapping is performed using the GRCh38.p13 Human Genome Assembly (publicly available at https://www.ensembl.org/Homo_sapiens/Info/Index). To obtain reliable expression values for the resulting genes, we implement two pooling measures (maximum and average) to reduce noise and increase accuracy. The pooling method aggregates expression values across multiple SNPs that map to the same gene, providing a more precise representation of gene expression patterns associated with genetic variations captured by the SNPs. Finally, we use GGIs from the HuRI [1] for feature selection, which results in the computation of sparse biologically prompted networks that better represent gene interactions in the body. Guiding network construction with HuRI not only increases biological significance but also reduces the computational intensity by reducing edge calculations.

3 Individual Specific Networks (ISNs) Use Cases

3.1 Differentially expressed genes resulting from LIMMA analysis

In this study, we utilized the LIMMA (Linear Models for Microarray Data) algorithm, a statistical tool (<https://bioconductor.org/packages/release/bioc/html/limma.html>) for analyzing gene expression data, to identify differentially expressed gene-interactions (i.e., edges) between two populations in the HapMap dataset. We took inspiration from analogous work done for LionessR, showcasing the algorithm potential to find relevant edges that discriminate between groups. The HapMap dataset comprises 11 populations, and for this study, we selected the Maasai in Kinyawa, Kenya (abbr. MKK) and Yoruba in Ibadan, Nigeria (abbr. YRI) populations. These two populations were chosen due to their well-defined separation into distinct clusters and their balanced sample size. We used a threshold of 0.25, i.e., the difference between the average edge weight in the two groups, as a pre-filter of the edge results. Hence, following an analogous pipeline as in LionessR, we identify the significant (adj. p-value < 0.05) differential edges between MKK and YRI. The results were then visualized in a graph (Figure S5) with red edges indicating a stronger expression in MKK, and blue edges indicating a stronger expression in YRI. This analysis demonstrates the capability of Individual Specific Networks (ISNs) to detect differential expressions of edges within a network constructed with the same set of nodes. By using ISNs, this study sheds light on the nodes and edges that differentiate the two populations, MKK and YRI.

Both the genes in the S100A7-S100A7L2 interactions are protein-coding genes of the S100A7 family, associated with the homonymous protein. Those genes enable calcium ion binding, protein binding and zinc ion binding [2–5]. KRTAP10-7 Gene, i.e., Keratin Associated Protein 10-7, is the node with the highest degree and centrality in the LIMMA analysis. It is a Protein Coding gene. Diseases associated with KRTAP10-7 include Trochlear Nerve Disease and Nodular Basal Cell Carcinoma. Among its related pathways are Keratinization and Nervous system development.

It would be interesting to study the difference in the insurgence and development of related diseases in the Maasai and Yoruba populations, since the relations of said genes, and in general of three genes in the KRTAP10 family, are significantly different among the two populations. Overall, in this study, we highlight relevant differences among African population, remarking the need for more in-detail analysis.

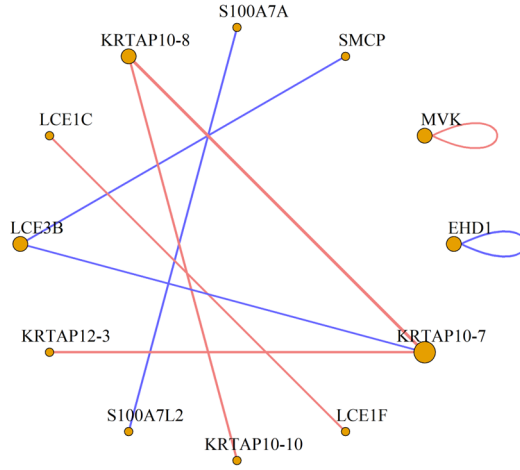


Figure S5: Identification of Differentially Expressed Gene-Interactions between Maasai and Yoruba Populations Using LIMMA Algorithm on HapMap Dataset

3.2 Multi-View Individual Specific Networks (ISNs) Analysis for BMI-Related Clustering in Facial Images and SNPs

In a separate project, ISNs were utilized for a multi-view dataset consisting of both facial images and SNPs. As a preprocessing step, the SNPs were mapped to genes and the facial images were segmented. Then, Principal Component Analysis (PCA) was performed on both the genes and facial segments to generate multi-view PC features. Only the features most related to Body Mass Index (BMI) were selected for constructing the ISNs, leading to the identification of five BMI-associated clusters and their representative features.

3.3 Analysis of TCGA data with ISN-tractor: graph representation

Figure S6 (Part 1 and 2) shows the graph representation of distinct patterns of edge importance within and across different cancer types from TCGA data.

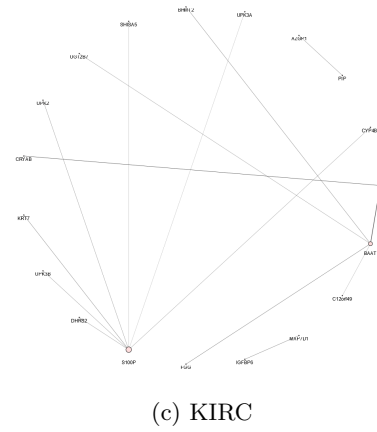
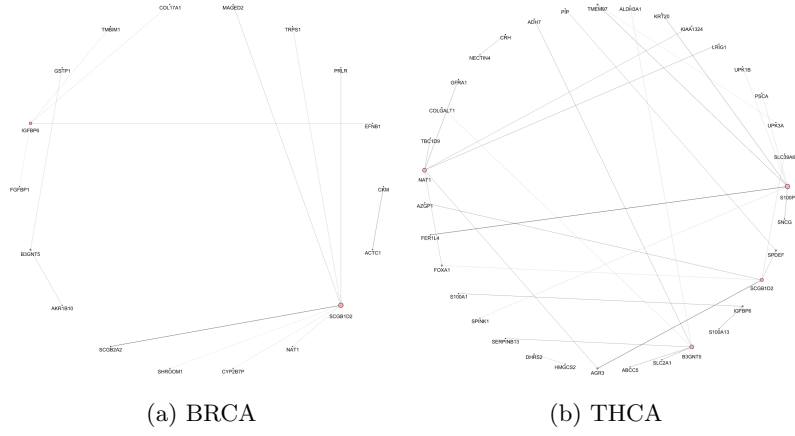


Figure S6: Graph representation of distinct patterns of edge importance within and across different cancer types from TCGA data (Part 1)

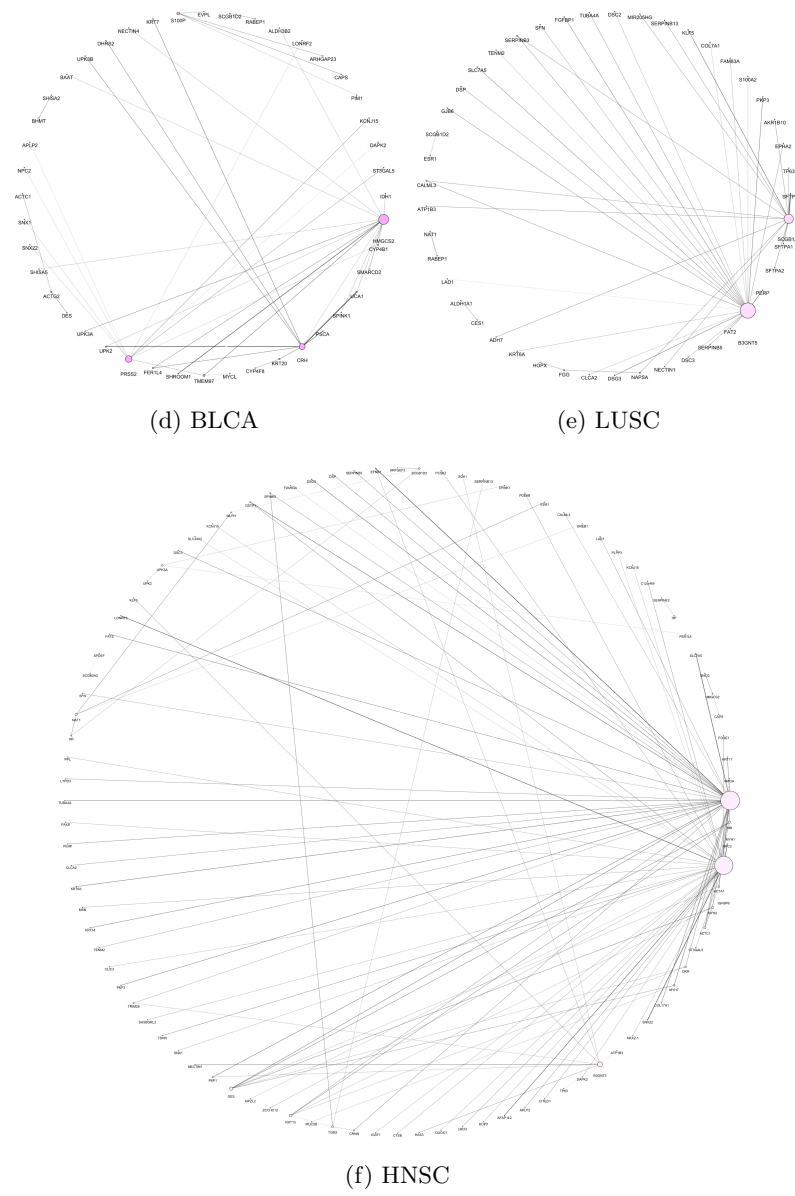


Figure S7: Graph representation of distinct patterns of edge importance within and across different cancer types from TCGA data (Part 2)

References

- [1] Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., *et al.*: A reference map of the human binary protein interactome. *Nature* **580**(7803), 402–408 (2020)
- [2] Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., Choi, D., Côté, A.G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M.F., Kishore, N., Knapp, J.J., Kovács, I.A., Lemmens, I., Mee, M.W., Mellor, J.C., Pollis, C., Pons, C., Richardson, A.D., Schlabach, S., Teeking, B., Yadav, A., Babor, M., Balcha, D., Basha, O., Bowman-Colin, C., Chin, S.F., Choi, S.G., Colabella, C., Coppin, G., D’Amata, C., Ridder, D., Rouck, S., Duran-Frigola, M., Ennajdaoui, H., Goebels, F., Goehring, L., Gopal, A., Haddad, G., Hatchi, E., Helmy, M., Jacob, Y., Kassa, Y., Landini, S., Li, R., Lieshout, N., MacWilliams, A., Markey, D., Paulson, J.N., Rangarajan, S., Rasla, J., Rayhan, A., Rolland, T., San-Miguel, A., Shen, Y., Sheykhkarimli, D., Sheynkman, G.M., Simonovsky, E., Taşan, M., Tejada, A., Tropepe, V., Twizere, J.C., Wang, Y., Weatheritt, R.J., Weile, J., Xia, Y., Yang, X., Yeger-Lotem, E., Zhong, Q., Aloy, P., Bader, G.D., Las Rivas, J., Gaudet, S., Hao, T., Rak, J., Tavernier, J., Hill, D.E., Vidal, M., Roth, F.P., Calderwood, M.A.: A reference map of the human binary protein interactome. *Nature* **Apr;580(7803):402-408** <https://doi.org/10.1038/s41586-020-2188-x>. . Epub 2020 Apr 8. PMID: 32296183; PMCID: PMC7169983.
- [3] Hoffmann, H.J., Olsen, E., Etzerodt, M., Madsen, P., Thøgersen, H.C., Kruse, T., Celis, J.E.: Psoriasin binds calcium and is upregulated by calcium to levels that resemble those observed in normal skin. *J Invest Dermatol* **Sep;103(3):370-5** <https://doi.org/10.1111/1523-1747.ep12395202>. . PMID: 8077703.
- [4] Emberley, E.D., Gietz, R.D., Campbell, J.D., HayGlass, K.T., Murphy, L.C., Watson, P.H.: Ranbpm interacts with psoriasin in vitro and their expression correlates with specific clinical features in vivo in breast cancer. *BMC Cancer* <https://doi.org/10.1186/1471-2407-2-28>. . PMID: 12421467; PMCID: PMC137593.
- [5] Brodersen, D.E., Nyborg, J., Kjeldgaard, M.: Zinc-binding site of an s100 protein revealed. two crystal structures of ca2+-bound human psoriasin (s100a7) in the zn2+-loaded and zn2+-free states. *Biochemistry* **9;38(6):1695-704** <https://doi.org/10.1021/bi982483d>. . PMID: 10026247.