**Large-scale Integration of Proteomics and Genomics Data to Discover Risk Proteins in Human Cancers**

Qing Li1,†, Zhishan Chen2,†, Jie Ping2, Wanqing Wen2, Xiang Shu3, Jun Yan4, Xiao-ou Shu2, Wei Zheng2, Quan Long1\* and Xingyi Guo2\*

**Affiliations:**

1 Department of Biochemistry & Molecular Biology, University of Calgary, Calgary, Canada

2 Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville TN 37203, USA.

3 Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

4 Physiology and Pharmacology, University of Calgary, Calgary, Canada

† Author names shared co-first authorship

\*Corresponding authors:

**Abstract**

Genomic-wide association studies (GWAS) have identified numerous genetic risk variants associated with various types of cancer. To enhance our understanding of the genetic basis of cancer susceptibility, it is crucial to conduct fine-mapping analyses within these GWAS-identified risk loci, which can uncover additional independent risk association signals and uncover potential causal variants and genes for cancer. While fine-mapping studies have been conducted for breast and colorectal cancers, several other cancer types have not been thoroughly investigated. Furthermore, while candidate target genes for the risk variants identified in GWAS have been extensively studied through expression quantitative trait locus (eQTL) analysis in human cancers, the investigation of target proteins for most of them remains unexplored.

In this study, we characterized 712 independent signals associated with cancer risk, through examining a comprehensive collection of previously reported risk variants and performing additional fine-mapping analyses using summary statistics data from European-ancestry populations from six cancer types: breast (N=247,173), ovary (N=63,347), prostate (N=140,306), colorectum (N=125,478), lung (N=85,716), and pancreas (N=21,536). To identify candidate proteins associated with these risk variants, we conducted a meta-analysis of plasma protein quantitative trait locus (pQTL) analysis results from two European-ancestry population studies (N=42,772; PMID: 35501419 and PMID: 34857953). By integrating the findings from these six cancer types, we discovered 259 protein at a Bonferroni corrected p-value threshold of < 0.05, corresponding to 207 unique proteins for 162 risk variants, including 34 proteins that were shared by at least two cancer types. Of them, associations of cancer risk with over 40% of these proteins was supported by additional evidence from functional genomic data in target cancer-related cells, eQTL analysis in relevant tissues, and colocalization analyses with GWAS risk signals. Enrichment analyses of these proteins highlighted the prominent involvement of well-established cancer signaling pathways, such as acute-phase responses, IL-6, Natural Killer cell wound healing, and STAT3 signaling pathways. Notably, a subset of proteins (7 out of 43 in breast cancer) identified in our study have previously been reported as cancer risk proteins in the existing literature. Therefore, our findings present a substantial number of potentially influential proteins that may impact cancer risks, by leveraging the strength of large-scale genome-wide association studies (GWAS) and protein quantitative trait loci (pQTL) data.

Our study identified a significant number of novel putative susceptibility proteins associated with cancer risk. By shedding light on the intricate pathways connecting risk genetic variants, target proteins, and signaling pathways related to cancer risk, our findings provide new insights and potential avenues for prevention and therapeutic interventions in these common cancers.