An overarching objective of our study is to leverage ENCODE data in order to provide novel insights and resources for cancer research. We aim integrate ENCODE and cancer genomic data to gain a more comprehensive understanding of the non-coding elements involved in oncogenesis, their associated linkages to protein-coding genes and the background mutation rates therein, and the global regulatory nature of TFs in the context of matched tumor-normal cell lines. The recent ENCODE data release provides a rich source of information for investigating questions both in basic biology and human disease. In large part, this wealth of information derives from the multiple genomic annotations provided across multiple cell lines. In addition to providing new opportunities, however, the very richness of this data provides considerable challenges in terms of data integration and organization. In addition to the complexity of this data resource, our analyses relies on an array methodologies, the details for which are difficult to include within the main text of this paper. As such, the purpose of this Supplementary document is to provide a clear and organized reference to support and explain the datasets, pipelines, and analyses associated with this study. In addition to supplementary text, supplementary figures and tables provide additional information not included in the main figures.

Our study is broadly organized into 4 main parts: a description of the assays, the construction of enhancer-target gene linkages, the workflow for variant prioritizing key genomic features associated with cancer, and concluding remarks. This supplement is presented in roughly a parallel fashion to the main text. The supplement is also connected to main text through the major results presented in the form of main text figures – captions associated with main text figures point to relevant sub-sections within the supplement. We have written our study in roughly a hierarchical fashion, and aim to present data and results (including predications) in an organized way. The main text lies at the top of this hierarchy, and synthesizes everything in a broad fashion. It refers to more detailed descriptions of our methods and datasets, as provided in the supplement.

Part 1 provides in-depth documentation of the ENCODE data we use, along with the subsidiary steps (including ENCODE data processing, enhancer and enhancer-target predictions, and extended gene definitions). Part 2 provides details on our recurrence analyses. Part 3 provides in-depth discussions and data regarding our TF network construction and analyses. Part 4 aims to expand on our expression aggregation analysis. Finally, Part 5 deals with the validation of prioritized SNVs.