# Experience on multi-omics and multi-scale data fusion to study mutation effect on key genes in the network modules

***Experience on data integration*.** We have extensive experience with the ENCODE (Djebali, Davis et al. 2012, Encode Project Consortium 2012, Gerstein, Kundaje et al. 2012) modENCODE (Gerstein, Lu et al. 2010, Gerstein, Rozowsky et al. 2014), 1,000 Genomes (Genomes Project, Abecasis et al. 2010, Khurana, Fu et al. 2013, Sudmant, Rausch et al. 2015), and PsychENCODE (Psych, Akbarian et al. 2015) consortia, in which we served a variety of leadership roles (i.e., co-lead of the analysis working group for ENCODE and modENCODE) (Gerstein, Lu et al. 2010, Encode Project Consortium 2012, Gerstein, Kundaje et al. 2012, Boyle, Araya et al. 2014). We also have extensive experience analyzing cancer genomes through our participation in TCGA and PCAWG consortia. We participated in the TCGA consortium studies of prostate (Cancer Genome Atlas Research 2015) and kidney (Cancer Genome Atlas Research, Linehan et al. 2016) cancers and recently conducted a detailed investigation of the non-coding mutations in TCGA kidney papillary cancer samples (Li, Shuch et al. 2017). We have also developed tools for somatic variant calling (Fang, Afshar et al. 2015). Currently, we are co-leading a PCAWG sub-group to investigate the impact of non-coding mutations.

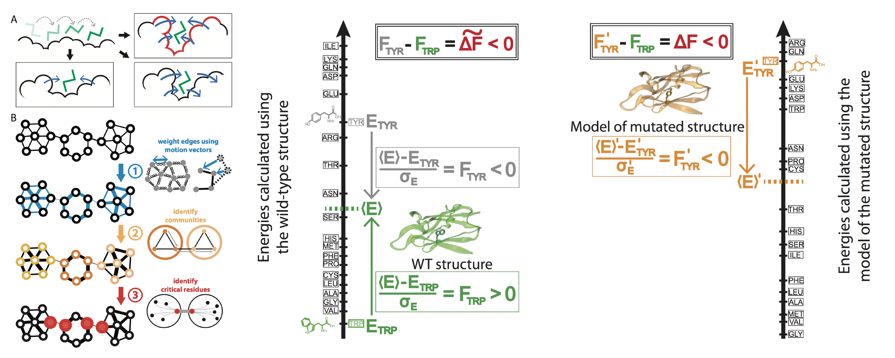
We have extensively analyzed patterns of variation in non-coding regions and their coding targets (Mu, Lu et al. 2011, Gerstein, Kundaje et al. 2012, Yip, Cheng et al. 2012). In recent projects (Khurana, Fu et al. 2013, Fu, Liu et al. 2014), we integrated multiple methods into a comprehensive prioritization pipeline called FunSeq. The pipeline identifies sensitive regions with annotations under high selective pressure, links non-coding mutations to their target genes, and prioritizes mutations based on network connectivity. It also identifies deleterious mutations in non-coding elements including transcription factor binding sites, enhancers, and regions corresponding to DNase I hypersensitive sites. Using integrated data from large-scale resources (including ENCODE and the 1,000 Genomes Project) with cancer genomics data, FunSeq can prioritize known TERT promoter driver mutations.

***Experience on statistical modelling*.** Our extensive experience with mining DNA and RNA data in cancer and other settings make us well prepared to interrogate GBM invasiveness.

Our lab has extensive experience in modelling genomic variations and their impact on regulatory networks, and specifically to detect driver mutations. A common method to search for driver mutations is to find genes or regions of the genome that are highly enriched for mutations. However, this search can be confounded by the fact that different regions of the genome have different mutation rates. We developed a computational framework called LARVA, which integrates mutations with a set of non-coding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle over-dispersion (Lochovsky, Zhang et al. 2015). Importantly, this method incorporates regional genomic features such as replication timing to better estimate local mutation rates and find mutational hotspots. Applying LARVA to 760 whole-genome tumor sequences shows that it identifies well-known non-coding drivers, such as mutations in the TERT promoter, in addition to uncovering new potential non-coding driver regions.

We have also extensively used TCGA RNA-seq data to develop and apply tools. For example, we developed a computational method called DREISS for analyzing the “Dynamics of gene expression driven by Regulatory networks, both External and Internal, based on State Space models” (Wang, He et al. 2016). DREISS employs dimensionality reduction to help identify canonical temporal dynamics (e.g., degradation, growth, and oscillation) representing the regulatory effects emanating from various subsystems. Another such tool that we developed, Loregic, computationally integrates gene expression and regulatory network data to characterize the cooperativity of regulatory factors (Li, Deng et al. 2015). These tools can be applied to sequencing data in cancer samples to identify crucial regulatory motifs.

***Experience on structural modelling*.** We have developed a number of tools to evaluate mutation deleteriousness in the context of protein structures and their interactions. Our comprehensive framework incorporates protein structure and dynamics for predicting allosteric residues both on the surface and in the interior (stress.molmovdb.org). As shown in **Fig. 1**, our STRESS tool identifies mutations that might affect allosteric hotspots in proteins, which can be key to protein function (Clarke, Sethi et al. 2016). This has been added to our suite of tools in MolMovDB, in which users visualize conformational changes (Echols, Milburn et al. 2003). Using RigidFinder, one may predict regions that remain relatively rigid upon such large-scale conformational changes (Abyzov, Bjornson et al. 2010). In addition, DynaSIN (which is also a part of the MolMovDB suite) catalogues instances in which protein conformational changes contribute favorably to protein interactions. Localized structural frustration identifies key functional regions (Kumar, Clarke et al. 2016) (**Fig. 1**). In addition, our Intensification tool searches for deleterious mutations particularly within repeated regions of proteins (Chen, Wang et al. 2017).



**Fig. 1** Schematic of STRESS & Frustration.

References:

Abyzov, A., et al. (2010). "RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes." Proteins **78**(2): 309-324.

Boyle, A. P., et al. (2014). "Comparative analysis of regulatory information and circuits across distant species." Nature **512**(7515): 453-456.

Cancer Genome Atlas Research, N. (2015). "The Molecular Taxonomy of Primary Prostate Cancer." Cell **163**(4): 1011-1025.

Cancer Genome Atlas Research, N., et al. (2016). "Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma." N Engl J Med **374**(2): 135-145.

Chen, J., et al. (2017). "Intensification: A Resource for Amplifying Population-Genetic Signals with Protein Repeats." J Mol Biol **429**(3): 435-445.

Clarke, D., et al. (2016). "Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation." Structure **24**(5): 826-837.

Djebali, S., et al. (2012). "Landscape of transcription in human cells." Nature **489**(7414): 101-108.

Echols, N., et al. (2003). "MolMovDB: analysis and visualization of conformational change and structural flexibility." Nucleic Acids Res **31**(1): 478-482.

Encode Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Fang, L. T., et al. (2015). "An ensemble approach to accurately detect somatic mutations using SomaticSeq." Genome Biol **16**: 197.

Fu, Y., et al. (2014). "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome Biol **15**(10): 480.

Genomes Project, C., et al. (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Gerstein, M. B., et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Gerstein, M. B., et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Gerstein, M. B., et al. (2014). "Comparative analysis of the transcriptome across distant species." Nature **512**(7515): 445-448.

Khurana, E., et al. (2013). "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science **342**(6154): 1235587.

Kumar, S., et al. (2016). "Localized structural frustration for evaluating the impact of sequence variants." Nucleic Acids Res **44**(21): 10062-10073.

Li, M. J., et al. (2015). "wKGGSeq: A Comprehensive Strategy-Based and Disease-Targeted Online Framework to Facilitate Exome Sequencing Studies of Inherited Disorders." Hum Mutat **36**(5): 496-503.

Li, S., et al. (2017). "Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations." PLoS Genet **13**(3): e1006685.

Lochovsky, L., et al. (2015). "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." Nucleic Acids Res **43**(17): 8123-8134.

Mu, X. J., et al. (2011). "Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project." Nucleic Acids Res **39**(16): 7058-7076.

Psych, E. C., et al. (2015). "The PsychENCODE project." Nat Neurosci **18**(12): 1707-1712.

Sudmant, P. H., et al. (2015). "An integrated map of structural variation in 2,504 human genomes." Nature **526**(7571): 75-81.

Wang, D., et al. (2016). "DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks." PLoS Comput Biol **12**(10): e1005146.

Yip, K. Y., et al. (2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors." Genome Biol **13**(9): R48.