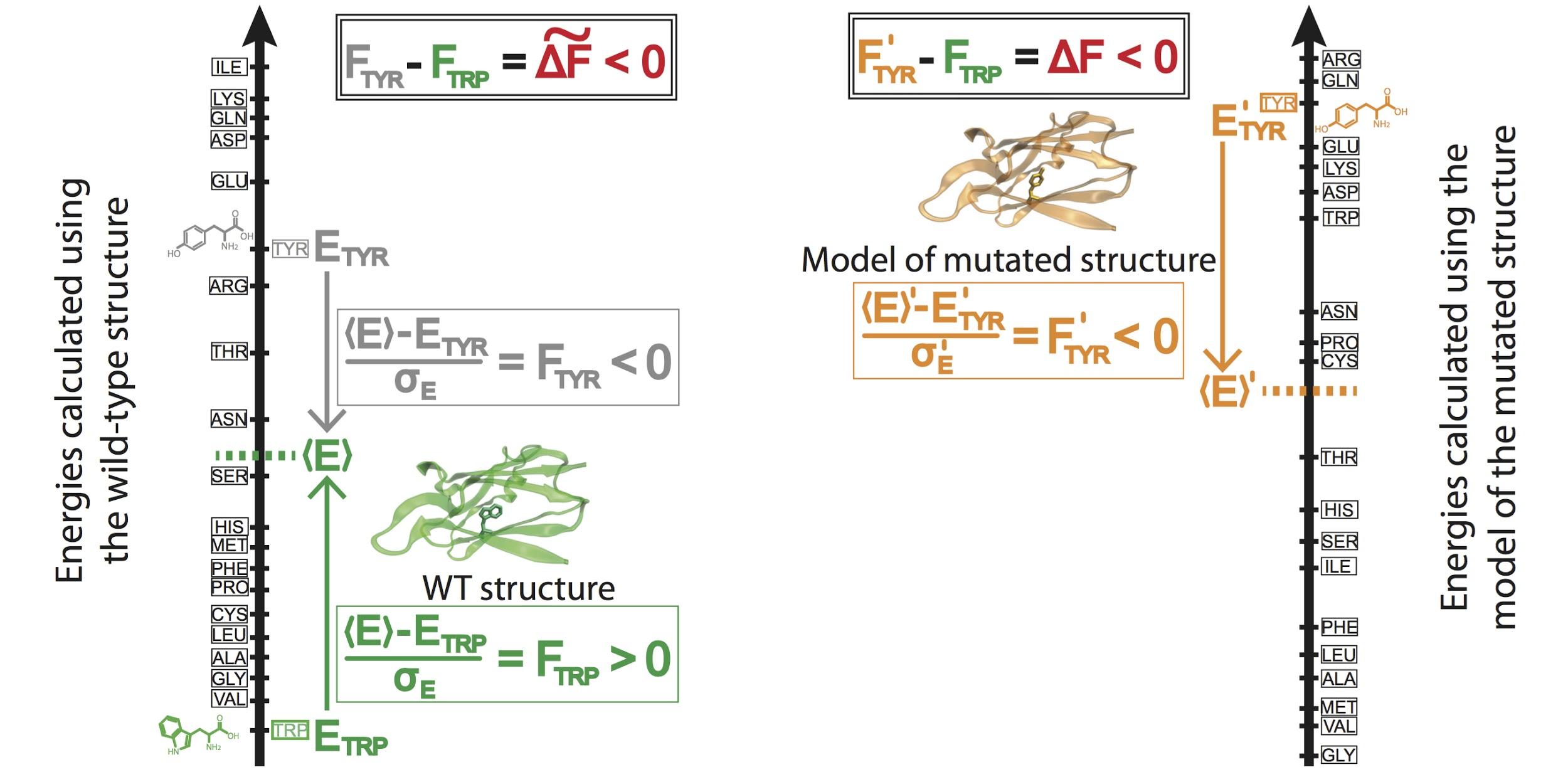
We have extensive experience in large-scale variant calling and interpretation through our active participation in the 1000 Genomes Consortium. In particular, we were involved in the analysis working group and the SV and functional interpretation subgroups, in which the majority of the variant calling tools were developed, deployed, and interpreted [21, 22, 23].

SVs are important contributors to human polymorphisms, have great functional impact, and are implicated in a number of diseases such as cancer. We have developed a number of SV-calling algorithms, including BreakSeq [28], CNVnator [29], AGE [30], and PEMer [31]. Furthermore, we have studied the SVs that originate from different mechanisms and may have potentially divergent functional impacts [32, 33].

Tools for somatic and germline burden tests:We have developed a number of software tools to annotate and understand the effects of variants within the coding regions of the human genome. We developed **VAT** to annotate coding variants; for example, VAT can determine them to be synonymous, non-synonymous, premature stop codons, or splice-site changes [35]. Once mapping the annotated variants to three-dimensional (3D) structures from the Protein Data Bank, we can study the effects of them in detail by measuring events associated with their LoF or in the contexts of allosteric regulation and local mechanistic perturbations.

In addition, we have developed **ALoFT**, a tool specifically tailored to annotate and predict the disease-causing potential of LoF events [36]. Short for “annotation of loss-of-function transcripts”, we have used ALoFT to successfully discriminate between LoF mutations that are deleterious in heterozygous states from those that may cause disease in the homozygous state. We analyzed somatic variants in more than 6,500 cancer exomes and demonstrated that variants predicted to be deleterious by ALoFT are enriched in canonical cancer driver genes [36].

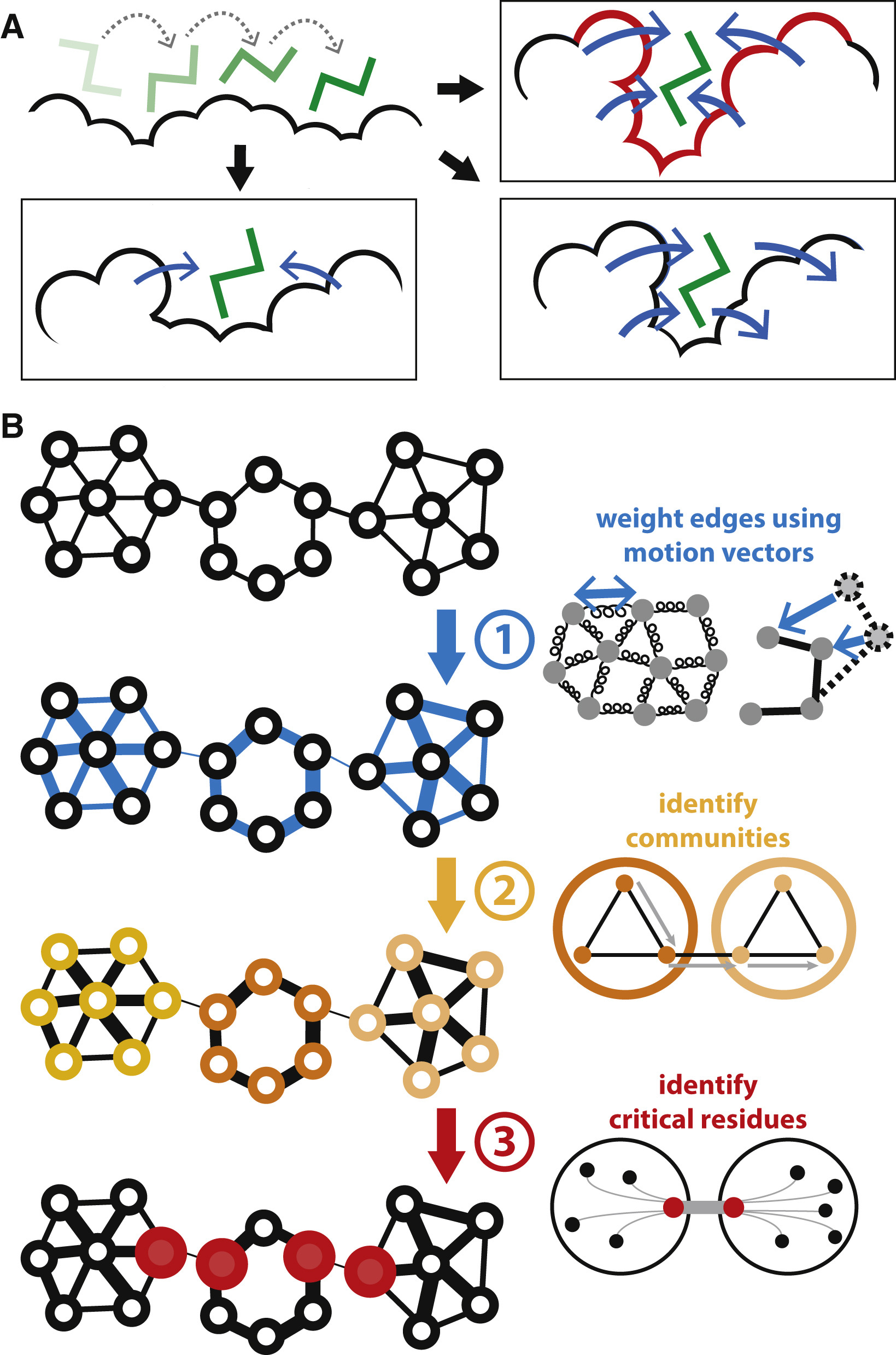
With respect to allosteric effects, we have developed the **STRESS** software tool [37]. STRESS (STRucturally-identified ESSential residues) employs models of large-scale protein conformational changes in order to predict key allosteric residues from both the protein surface (by finding essential pockets) as well as the interior (by identifying information-flow bottlenecks). Our reported results demonstrate that this software selects residues that are highly conserved over both long and short evolutionary timescales [37]. This software has also been used to help rationalize otherwise poorly understood (“cryptic”) disease-associated SNVs.

Figure 6: Prioritizing the effects of SNVs based on predicted allosteric residues at the surface (A) and within the interior (B).

Figure 5: Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

In addition to coding variants, we have developed a tool to prioritize non-coding variants in cancer called **FunSeq** (Fig. 7). In brief, FunSeq prioritizes variants based on network connectivity and their disruptiveness (e.g., by finding motif breakers), by identifying deleterious variants in many non-coding functional elements (including transcription factor binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites). In our published work using FunSeq [32], we integrated large-scale data from various resources, including ENCODE and the 1000 Genomes Project, with cancer genomics data. By comparing patterns of inherited polymorphisms from 1,092 humans with somatic variants, FunSeq identified candidate non-coding driver mutations.

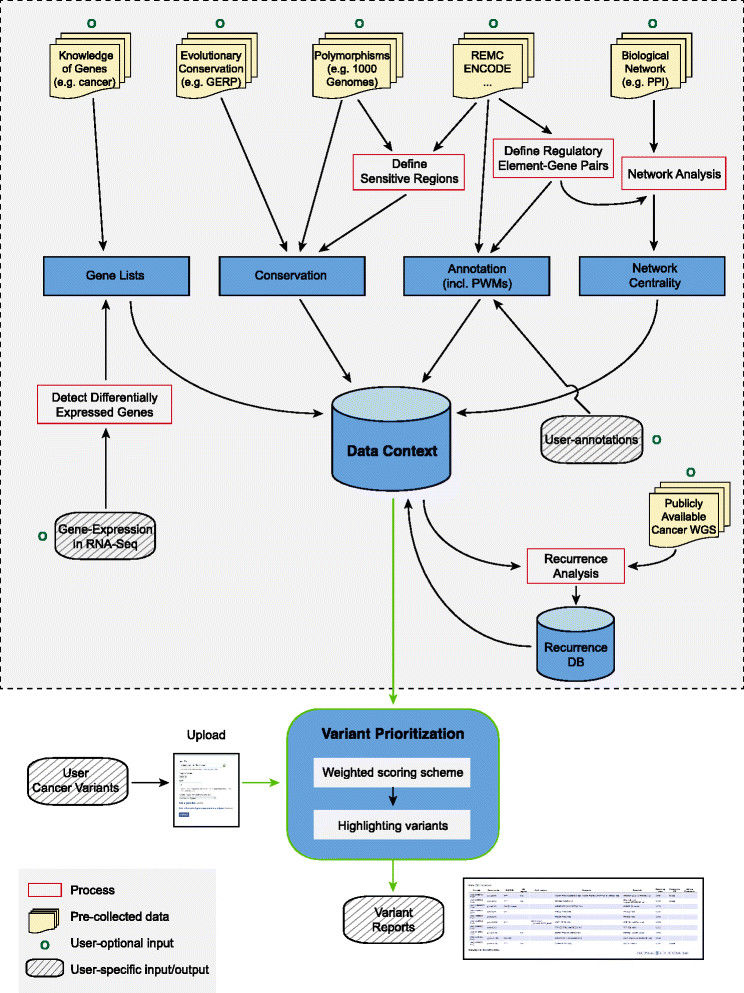
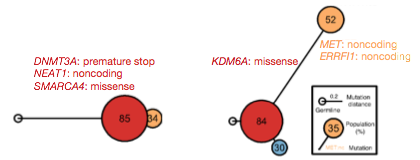
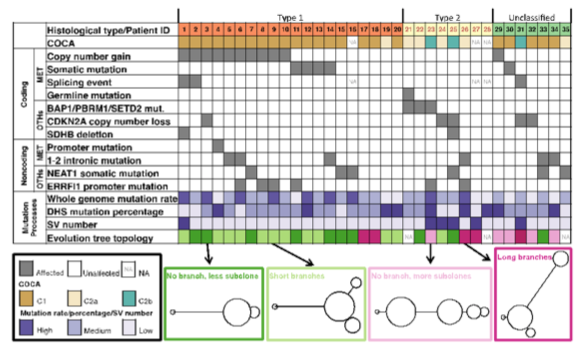
We have developed statistical methods for the analysis of non-coding regulatory regions. **LARVA** (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichment in non-coding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model [39]. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can be targeted to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of variants in 760 cancer whole genomes, spanning a number of cancer data portals and published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and non-coding cancer drivers, including the TERT and TP53 promoters [39]. Finally, we developed **MOAT** (Mutations Overburdening Annotations Tool), an alternative empirical mutation burden approach that evaluates mutation enrichment based upon permutations of the input data (in press). This tool supports both annotation-based and variant-based permutation.

Figure 7: The workflow of FunSeq

We have played key roles in TCGA investigations into prostate [40] and kidney [25] cancers. We participated in TCGA KICH (chromophobe RCC) project [41] and a following pan-subtype kidney analysis [42]. Our team analyzed the WGS data for the TCGA KIRP (pRCC), now published in The New England Journal of Medicine [25]. In recent work, we leveraged our expertise in non-coding regions in the first whole-genome analysis of pRCC samples [43]. Our work found significant genomic alterations beyond traditional known drivers of pRCC located within coding exons (Fig. 8). We hypothesize that these alterations may have non-canonical effects on known tumorigenic pathways (for example, *MET* in type 1 pRCC). We discovered genomic markers in *MET* and *NEAT1* that influence prognosis. Moreover, we constructed evolutionary trees using the abundant mutation information from WGS. The tree structure implies tumor evolution path and correlates with tumor subtypes (Fig. 9).This experience provided further practical knowledge of working with available RCC genomic datasets. Finally, our team has participated in two ongoing pan-RCC manuscripts by playing a central role in assessing the cluster-of-cluster assignments immunologic profile from gene and microRNA expression datasets.

Figure 9: Evolutionary trees help elucidate pRCC tumor development and complete molecular subtyping.



Together with other published results on RCC [44, 45, 46, 47, 48], we have assembled an extensive list of impactful and statistically significant regions of RCC genomes. Similarly, as part of the driver discovery subgroup in PCAWG, we participated in a comprehensive variant prioritization exercise to generate a catalog of driver elements in many cancer cohorts. Furthermore, we are currently leading the PCAWG group to investigate the impact of non-coding mutations on cancer development, progression, and prognosis. As part of this effort, we ran our FunSeq pipeline on each variant (~30 million total somatic mutations among 39 cancer subtypes) in PCAWG. In addition to identifying canonical driver mutations, we identified many high-impact mutations that can potentially influence cancer progression.

References

21. Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O., 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59-65.22. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061-73.23. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., McVean, G. A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.

28. Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., Gerstein, M. B. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28(1):47-55.29. Abyzov, A., Urban, A. E., Snyder, M., Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21(6):974-84.30. Abyzov, A., Gerstein, M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27(5):595-603.31. Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., Gerstein, M. B. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10(2):R23.32. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüş, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacArthur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., 1000 Genomes Project Consortium, Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C., Gerstein, M. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342(6154):1235587.33. Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., Hurles, M., Korbel, J. O., Lam, H. Y. K., Lee, C., Gerstein, M. B. (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* 6():7256.

35. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., Gerstein, M. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 28(17):2267-9.36. Balasubramanian, S., Fu, Y., Pawashe, M., McGillivray, P., Jin, M., Liu, J., Karczewski, K. J., MacArthur, D. G., Gerstein, M. (2017) Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Commun* 8(1):382.37. Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W. F., Chen, J., Gerstein, M. (2016) Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure* 24(5):826-837.38. Kumar, S., Clarke, D., Gerstein, M. (2016) Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res* 44(21):10062-10073.39. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., Gerstein, M. (2015) LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* 43(17):8123-34.40. Cancer Genome Atlas Research Network (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163(4):1011-25.41. Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., Buhay, C., Kang, H., Kim, S. C., Fahey, C. C., Hacker, K. E., Bhanot, G., Gordenin, D. A., Chu, A., Gunaratne, P. H., Biehl, M., Seth, S., Kaipparettu, B. A., Bristow, C. A., Donehower, L. A., Wallen, E. M., Smith, A. B., Tickoo, S. K., Tamboli, P., Reuter, V., Schmidt, L. S., Hsieh, J. J., Choueiri, T. K., Hakimi, A. A., The Cancer Genome Atlas Research Network, Chin, L., Meyerson, M., Kucherlapati, R., Park, W.-Y., Robertson, A. G., Laird, P. W., Henske, E. P., Kwiatkowski, D. J., Park, P. J., Morgan, M., Shuch, B., Muzny, D., Wheeler, D. A., Linehan, W. M., Gibbs, R. A., Rathmell, W. K., Creighton, C. J. (2014) The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* 26(3):319-330.42. Leont'ev, A. P., Minchenkova, B. I., Osipov, V. T. (1973) [Emitters for the apparatus for ultrasonic therapy]. *Med Tekh* 5(0):23-5.43. Bütler, R., Brunner, E. (1969) Inability of Ag-Anti-Ag complexes to fix complement and its significance. *Vox Sang* 17(5):462-4.44. Kovac, M., Navas, C., Horswell, S., Salm, M., Bardella, C., Rowan, A., Stares, M., Castro-Giner, F., Fisher, R., de Bruin, E. C., Kovacova, M., Gorman, M., Makino, S., Williams, J., Jaeger, E., Jones, A., Howarth, K., Larkin, J., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Stamp, G., O'Brien, T., Challacombe, B., Matthews, N., Phillimore, B., Begum, S., Rabinowitz, A., Varela, I., Chandra, A., Horsfield, C., Polson, A., Tran, M., Bhatt, R., Terracciano, L., Eppenberger-Castori, S., Protheroe, A., Maher, E., El Bahrawy, M., Fleming, S., Ratcliffe, P., Heinimann, K., Swanton, C., Tomlinson, I. (2015) Recurrent chromosomal gains and heterogeneous driver mutations characterise papillary renal cancer evolution. *Nat Commun* 6():6336.45. Durinck, S., Stawiski, E. W., Pavía-Jiménez, A., Modrusan, Z., Kapur, P., Jaiswal, B. S., Zhang, N., Toffessi-Tcheuyap, V., Nguyen, T. T., Pahuja, K. B., Chen, Y.-J., Saleem, S., Chaudhuri, S., Heldens, S., Jackson, M., Peña-Llopis, S., Guillory, J., Toy, K., Ha, C., Harris, C. J., Holloman, E., Hill, H. M., Stinson, J., Rivers, C. S., Janakiraman, V., Wang, W., Kinch, L. N., Grishin, N. V., Haverty, P. M., Chow, B., Gehring, J. S., Reeder, J., Pau, G., Wu, T. D., Margulis, V., Lotan, Y., Sagalowsky, A., Pedrosa, I., de Sauvage, F. J., Brugarolas, J., Seshagiri, S. (2015) Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. *Nat Genet* 47(1):13-21.46. Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-9.47. Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H., Nagata, Y., Yoshida, K., Kon, A., Suzuki, Y., Chiba, K., Tanaka, H., Niida, A., Fujimoto, A., Tsunoda, T., Morikawa, T., Maeda, D., Kume, H., Sugano, S., Fukayama, M., Aburatani, H., Sanada, M., Miyano, S., Homma, Y., Ogawa, S. (2013) Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* 45(8):860-7.48. Christinat, Y., Krek, W. (2015) Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. *Oncotarget* 6(12):10521-31.