1. We have experience in annotating noncoding regions of the genome, including both TF-binding sites and noncoding RNAs. We have experience in noncoding genome annotation, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of methods to define the binding peaks of TFs[1], prediction of a TF's target genes[2], and new machine learning techniques[3]. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers[4], which we have partially validated[5]. We also constructed linear and non-linear models that utilize TF binding and histone modification signals to accurately predict the transcriptional output of a gene in different cell types of several organisms including yeast, worm, fly, and human[6-10]. We have also constructed regulatory networks for human and model organisms[11,12], and completed many analyses on them[5,7,11,13-26].

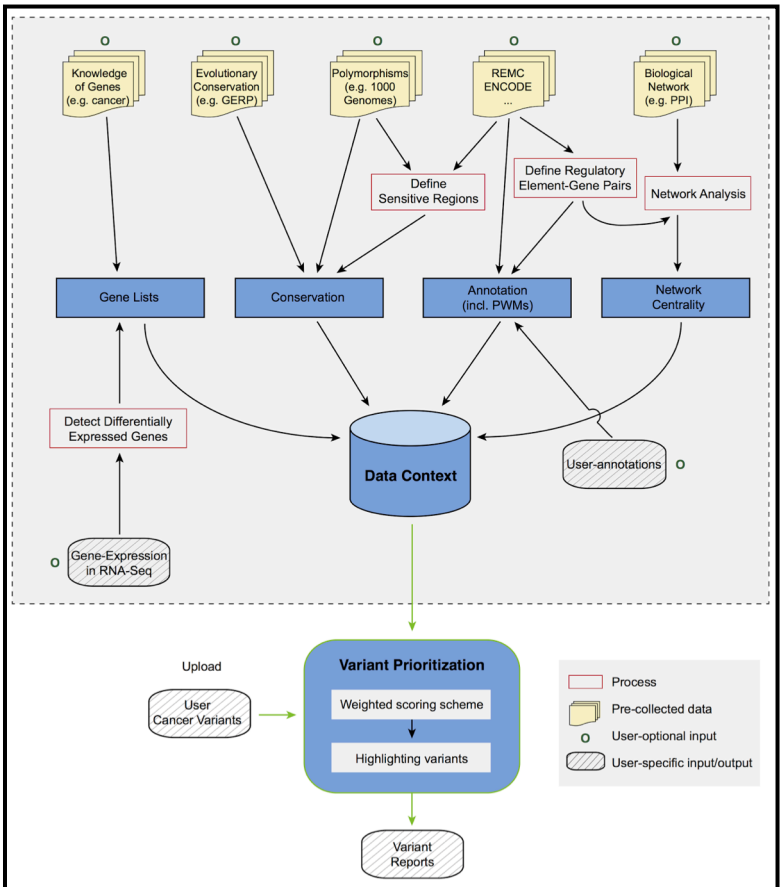

Fig. 1. Description of FunSeq workflow & data context.

2. We have experience in allelic analyses. A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly for transcription factors or DNA-binding proteins, and allele-specific expression (ASE)[27,28]. We have previously developed a tool, AlleleSeq,[24] for the detection of candidate variants associated with ASB and ASE. Using this we have generated comprehensive lists of allelic variants for ENCODE and 1000 Genomes and found that allelic variants are under differential selection from non-allelic ones[11,22,29]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression[11]. Furthermore, we have constructed a personal diploid genome and transcriptome of NA12878[30].

3. We have experience in relating annotation to variation: the FunSeq pipeline. We have extensively analyzed patterns of variation in noncoding regions, along with their coding targets[5,11,31]. In recent studies[22,32], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (**Fig. 1**). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each noncoding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many noncoding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations. Using FunSeq, we identified ~100 noncoding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples[22]. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

1. Rozowsky, J.*, et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66-75 (2009).
2. Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-3227 (2011).

3. Yip, K.Y. & Gerstein, M. Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics* **25**, 243-250 (2009).
4. Yip, K.Y., Alexander, R.P., Yan, K.-K. & Gerstein, M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One* **5**, e8121 (2010).
5. Yip, K.Y.*, et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48 (2012).
6. Cheng, C., Shou, C., Yip, K.Y. & Gerstein, M.B. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. *Genome Biol* **12**, R111 (2011).
7. Gerstein, M.B.*, et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787 (2010).
8. Cheng, C.*, et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15 (2011).
9. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* **40**, 553-568 (2012).
10. Cheng, C.*, et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* **22**, 1658-1667 (2012).
11. Gerstein, M.B.*, et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
12. Nègre, N.*, et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471**, 527-531 (2011).
13. Cheng, C.*, et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* **7**, e1002190 (2011).
14. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R.P. & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* **107**, 9186-9191 (2010).
15. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends Genet* **20**, 227-231 (2004).
16. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* **32**, 328-337 (2004).
17. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59 (2007).
18. Luscombe, N.M.*, et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-312 (2004).
19. Gianoulis, T.A.*, et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**, 1374-1379 (2009).
20. Yu, H., Paccanaro, A., Trifonov, V. & Gerstein, M. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**, 823-829 (2006).
21. Kim, P.M., Korbel, J.O. & Gerstein, M.B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-20279 (2007).
22. Khurana, E.*, et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
23. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886 (2013).
24. Rozowsky, J.*, et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).

25. Lörcher, U., Peters, J. & Kollath, J. [Changes in the lungs and pleura following chemoembolization of liver tumors with mitomycin-lipiodol]. *Rofo* **152**, 569-573 (1990).
26. Shou, C.*, et al.* Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol* **7**, e1001050 (2011).
27. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-538 (2010).
28. Birney, E., Lieb, J.D., Furey, T.S., Crawford, G.E. & Iyer, V.R. Allele-specific and heritable chromatin signatures in humans. *Hum Mol Genet* **19**, R204-209 (2010).
29. Djebali, S.*, et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
30. http://alleleseq.gersteinlab.org. Last accessed on 21st May 2015.
31. Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K. & Gerstein, M.B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058-7076 (2011).
32. Fu, Y.*, et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).