**Participation in Several Major National Biomedical Data Science Initiatives**

Prof. Mark Gerstein and his colleagues are involved in several large-scale national collaborations focused on aspects of genomics and data science. Many CBB graduate students are participating in these projects. These projects all have translational implications and provide a spectrum of translational bioinformatics research opportunities.

• **The 1000 Genomes Project** (http://1000genomes.org/) This is essentially NIH's marquee effort on personal genomics, the sequencing of individual people's genomes. The overall project aims to sequence thousands of individuals’ genomes to get a sense of their variability. Dr. Gerstein’s group developed an annotation pipeline that maps SNPs, indels and structural variations (SVs) on to protein coding genes. We also developed algorithms to identify indels and structural variations based on split-read, read-depth and paired-end mapping methods. Our methods were invaluable to the production phase of 1000 Genomes, which provides a comprehensive view of human variation based on the genomes of 2,500 individuals—a valuable resource for GWAS studies and other types of translational research. Furthermore, we have investigated the link between retroduplications and cell division using 1000 Genomes Project data. Our SV work has included the development of AGE (Alignment with Gap Excision) for finding the precise locations of SV breakpoints, and CNVnator, which offers a novel approach to finding copy number variants (CNVs) that might be missed by other methods. Finally, we have developed AlleleSeq, a computational pipeline for analyzing allele-specific expression and binding differences between maternally and paternally derived alleles. We are also involved in the 1KG SV trio project, a plan to sequence trios of individual from multiple families to very high coverage. The aim of this project is to identify structural variations with high confidence, compare the effectiveness of different sequencing technologies for this purpose, and compare sequences derived from different populations.

• **ENCODE** (http://www.genome.gov/10005107) As part of a multi-institutional collaboration, we are involved in annotating the human genome and developing

methods for analyzing large-scale genomic experiments. In particular, we are working extensively on pseudogene identification and annotation of the human genome in collaboration with the GENCODE team members (http://www.gencodegenes.org/). We are also elucidating transcription factor binding sites and chromatin structure based on ChIP-Seq experiments. To this end, we have developed Peak-Seq, an approach to identify peak regions in ChIP-seq data sets that correspond to sites of transcription factor binding (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2924752). We continue to refine this method as well as develop new methods for extensive human genome analyses. From a translational perspective, this project is providing annotation that enables genomic correlation with disease. Additional methods that we have developed include the development of ACT, a toolbox that facilitates many common operations on whole genome signal tracks, and IQSeq, which employs RNA-seq data towards measuring gene isoform abundance. Furthermore, we developed MUSIC, a tool for identifying enriched regions in ChIP-seq data. We were heavily involved in the ENCODE’s consortium’s rollout of a number of major human genome annotation papers in 2012 and 2014, with substantial contributions to the regulatory element annotation. In particular, we used ENCODE data to extensively study the architecture of the human regulatory network, and compared it to the regulatory networks of modENCODE model organisms *C. elegans* and *D. melanogaster*, both on a large-scale in terms of global statistics (e.g., the network diameter) and on a small-scale in terms of local network motifs..

• **Brainspan/psychENCODE** In collaboration with Prof. Nenad Sestan’s and Flora Vaccarino’s group at Yale, together with groups at USC, the Allen Brain Institute and elsewhere, we are analyzing large amounts of RNA-seq data to characterize the transcriptome of the human brain during development. The aim of this project is to create a comprehensive map of gene expression and to understand how the human brain changes throughout life. We have already developed RSEQtools, a suite of tools that performs common tasks on RNA-seq data such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. This project provides a reference atlas of gene expression in different regions of the brain that will provide valuable information to help interpret neurological function and dysfunction. More recently, this collaborative consortium has started work on PsychENCODE, a project aimed at understanding regulatory variants in the context of their functional connections to psychiatric disease. The project’s approach involves a comprehensive examination of the genome, transcriptome, epigenome, and proteome in relation to brain function.

* **Cancer Genomics** We worked on a number of cancer projects over the past year. One particular focus of ours was noncoding sequence variants. Our software projects included Funseq, a tool for functionally annotating regulatory variants in cancer genome sequences, and LARVA, a tool for detecting significant mutation burdens in noncoding elements in cancer whole genomes. As the number of cancer whole genome sequences continues to grow at an impressive rate, we anticipate these tools will be extremely useful in analyses to extract useful insights from this data. Furthermore, we are heavily involved in the Pan-Cancer Analysis Working Group (PCAWG): we are co-leaders of the PCAWG-2 group, and participate in the analyses of the PCAWG-3, 8, and 11 groups.
* **Privacy of Genomic Information & Data Science** Given the growing amount of available human genome data, a lot of which may be personally identifying, we investigated the privacy risks associated with the proliferation of personal genome data. We also looked into possible mechanisms for mitigating those risks. Specifically, we considered options for balancing the requirement of individual privacy with the need for sufficient data to perform useful analyses. The large scale of genome data also lends itself naturally to data science approaches, including machine learning, knowledgebase design, and modeling by simulation.
* **Proteomics & Networks** We aim to understand protein function by studying its structure and molecular motions, taking into consideration the permitted packing geometry. One particular focus of this work was to relate 3D protein structure to the topology of the protein-protein interaction network. Our network analyses have extended to regulatory, metabolic, and gene-expression networks. Identifying key hubs and bottlenecks in these systems is important for understanding how they operate. Specific examples include analyzing cooperative transcription factor binding, and comparing the phosphorylome to the regulome. We have also investigated the dynamics of networks, i.e., how their topology changes over time. In addition, we have identified changing hubs and systematic patterns of connectivity rewiring. One translational bioinformatics domain involves inferring regulatory networks in cancer (above). We have also correlated network hubs with gene essentiality, and consequently developed a number of tools to build and analyze networks derived from genes and also from literature citations