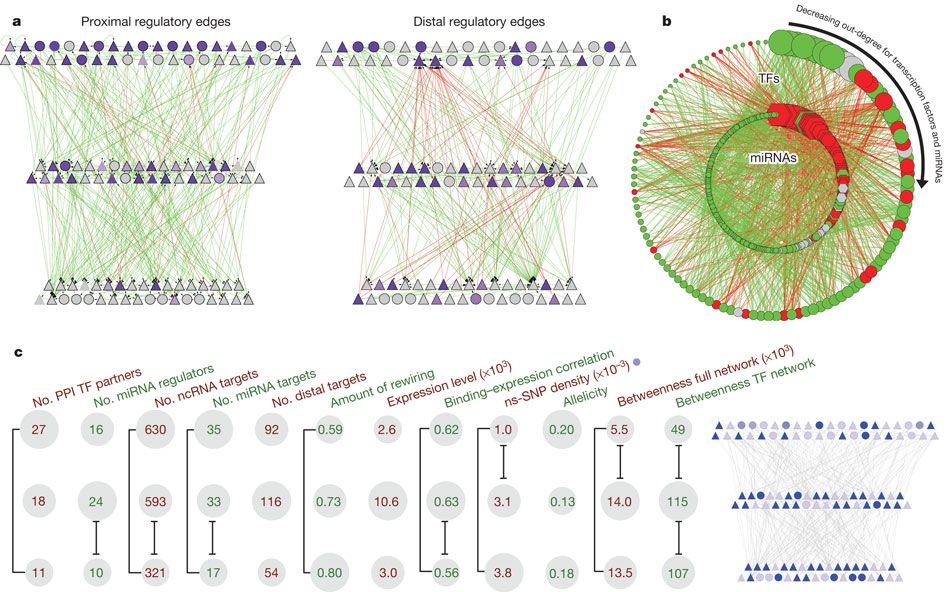
**Gerstein Lab Contributions in Gene Regulatory Networks, Citation Networks, and Personal Genome**

Analysis of biological network structure & dynamics

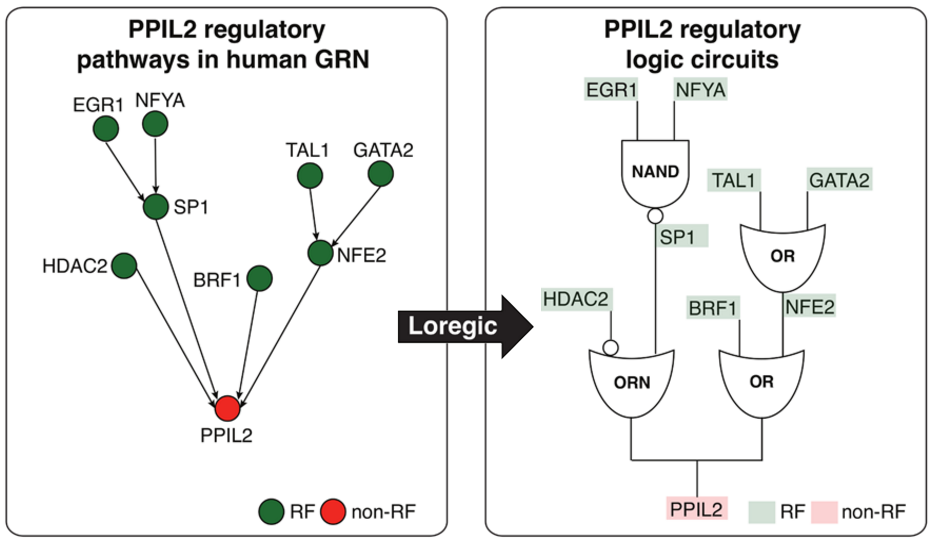
We have developed a number of approaches for constructing and studying biological networks that can be applied to analyze ENCODE4 datasets. We integrated multiple genomic datasets to construct gene regulatory networks consisting of various regulatory factors including transcription factors and micro-RNAs and their target genes {Gerstein:2012fq}{Boyle:2015bq}{Cheng:2011dx}. For constructed gene regulatory networks, we developed methods to construct and analyze human and model organism gene regulatory networks {Yan:2010fu}{Gerstein:2010bu}{Cheng:2011dx}{Negre:2011gg}{Gerstein:2012fq} using ENCODE and modENCODE datasets. We also analyzed hierarchical structures of gene regulatory networks and found that hierarchy rather than centrality ("hubiness") better reflects the importance of regulators {Gerstein:2012fq}{Yu:2006jg}{Bhardwaj:2010fr}{Bhardwaj:2010jj}{Bhardwaj:2010em}. We also developed a novel and general purpose algorithm to determine and measure the hierarchical structure of any type of gene regulatory network {Cheng:2015dl}. In addition, we integrated regulatory networks with gene expression to uncover different types of functional modules {Luscombe:2004ei}{Cheng:2009ks}{Yu:2003jd}{Qian:2003wh}. We also introduced several software tools for network analysis including Topnet {Yu:2004cv}, tYNA {Yip:2006kv}, and PubNet{Douglas:2005ky}.

In addition to the global features of regulatory networks such as hierarchy, we also analyzed their local topologies such as network motifs. For example, we analyzed feed-forward loops (FFLs), a common regulatory network motif in human, worm and fly gene regulatory networks {Gerstein:2012fq}{Gerstein:2010bu}{Boyle:2015bq}. The motif analyses will enable characterization of key regulatory mechanisms in each species and the comparison between species will enable us to observe how these mechanisms evolve.



**Figure Hierarchy** [**of the human regulatory network derived from ENCODE data**http://www.nature.com/nature/journal/v489/n7414/full/nature11245.html](http://www.nature.com/nature/journal/v489/n7414/full/nature11245.html)

**a**, Close-up representation of the transcription factor hierarchy. Nodes depict transcription factors. TFSSs are triangles, and non-TFSSs are circles. Left: proximal-edge hierarchy with downward pointing edges coloured in green and upward pointing edges coloured in red. The nodes are shaded according to their out-degree in the full network (as described in [Table 1](http://www.nature.com/nature/journal/v489/n7414/full/nature11245.html#t1)). Right: factors placed in the same proximal hierarchy but now with edges corresponding to distal regulation coloured green and red, and nodes re-coloured according to out-degree in the distal network. The distal edges do not follow the proximal-edge hierarchy. **b**, Close-up view of transcription-factor–miRNA regulation. The outer circle contains 119 transcription factors, whereas the inner circle contains miRNAs. Red edges correspond to miRNAs regulating transcription factors; green edges show transcription factors regulating miRNAs. Transcription factors and miRNAs each are arranged by their out-degree, beginning at the top (12:00) and decreasing in a clockwise order. Node sizes are proportional to out-degree. For transcription factors, the out-degree is as described in [Table 1](http://www.nature.com/nature/journal/v489/n7414/full/nature11245.html#t1); for miRNAs, it is according to the out-degree in this network. Red nodes are enriched for miRNA–transcription factor edges and green nodes are enriched for transcription factor–miRNA edges. Grey nodes have a balanced number of edges (within ±1).



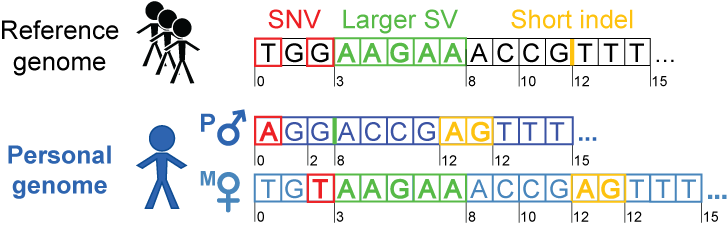
**Figure Depiction of two logic circuit regulatory pathways targeting PPIL2**.

Personal genomes and their use to develop allelic annotations

1.4.1 Personal genome construction

The alignment of assay reads is one of the main steps in processing functional genomic datasets. Conventionally, reads are aligned to the human reference genome. However, a systematic reference bias is introduced when reads are mapped to this haploid human reference sequence since reads that harbor an alternate allele are less likely to be aligned. In addition, reads can be improperly mapped to the reference genome in regions (or samples) with more genetic variation, especially when indels and larger structural variants are involved. This reduced mappability impairs estimation of read abundance and therefore compromises any downstream analyses.

For personal genome construction, we have developed a computational tool, *vcf2diploid* {Rozowsky:2011gx}. The tool integrates an individual’s genomic variation data (SNVs, indels, and SVs) into the reference genome. Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes. Chain files generated by the program can be used to account for coordinate offsets between the individual’s parental haplotypes and the original reference genomic sequence. The versatility to convert between reference and personal genome coordinates allows mapping of genomic annotated regions (e. g. gene or peak coordinates for RNA-seq and ChIP-seq, respectively) between the genomes using available tools, such as the UCSC LiftOver tool {Rhead:2010if}.



**Figure 1.4.1. Personal genome construction.** Each haplotype in the diploid personal genome is derived by incorporating phased or unphased variants (SNVs, indels and SVs) into the human reference genome. The coordinates can be mapped back to the human reference coordinates to facilitate comparisons with other reference-based resources, such as gene annotations from GENCODE.

We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878 (also known as GM12878). We have made this assembly available as a resource at alleleseq.gersteinlab.org and have been updating it as new versions of the human reference genome, genomic annotations, and NA12878 genetic variation data are released. Furthermore, the availability of a computational tool enables the construction of personal genomes in a high-throughput fashion, as demonstrated in a recent publication\cite{Chen et al. Nat. Commun., in press} where we built 382 personal genomes using the variant call sets from the 1000 Genomes Project.

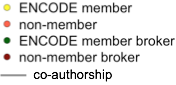
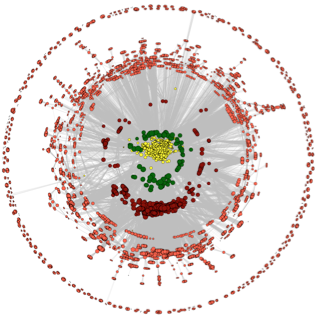
Allelic annotations

We also have extensive experience using personal genomes in analyses involving RNA-seq and ChIP-seq datasets. In particular, allele-specific (AS) analyses are very sensitive to mapping biases. Therefore in our pipeline *AlleleSeq*{Rozowsky:2011gx} the functional assay reads are aligned to a diploid personal genome, which alleviates the reference bias. We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project {Djebali:2012hc}{Gerstein:2012fq}{Khurana:2013em}{Sudmant:2015kz}.

More recently\cite{Chen et al. Nat. Commun., in press}, we annotated variants associated with allele-specific expression (ASE) and binding (ASB) in a large pool of individuals from the 1000 Genomes Project. For this analysis, we integrated matching functional datasets (955 RNA-seq and 165 ChIP-seq in total), which include ChIP-seq datasets from 14 lymphoblastoid cell lines in ENCODE{ENCODEProjectConsortium:2012gc}. We developed a standardized framework and statistical approaches to aggregate and uniformly process the datasets. Using the beta-binomial test, we first estimated overdispersion in the allelic ratios distribution for each dataset (filtering ones with a high overdispersion parameter) and then identified statistically significant AS events. We have also incorporated strategies to alleviate ambiguous mapping bias – bias occurring when reads originating from one of the two alleles map to multiple locations. Overall, we detected more than 6K and 63K SNVs associated with ASB and ASE, respectively. These results were made available as an online resource, AlleleDB (alleledb.gersteinlab.org) and serve as an allele-specific annotation for the 1000-Genome variant catalogue.

Analysis of ENCODE citation network:

The consortium publications on which the DAC has worked have been highly cited. The DAC has analyzed the patterns of dissemination of ENCODE publications to study how the outside scientific community benefits from ENCODE data. In particular, using publication data related to the ENCODE consortium {ENCODEProjectConsortium:2012gc}, we constructed co-authorship networks using temporal data from 2004 to 2014 (**Figure 2.4**). The networks show how the information from the consortium has diffused through authorship relationships. We found that the Consortium works as a community, whereas non-members collaborate on a smaller scale, and a few brokers initiate the connections between the consortium and non-members. Thus, large scientific consortia should set up formal outreach groups to communicate with outside researchers {Wang:2016gr}.



**Figure 2.4. The ENCODE co-authorship network.**