Hi-C data analysis: We have a long track record of generating and analyzing Hi-C libraries from adult dorsolateral prefrontal cortex (DLPFC)1. We also have extensive experience with Hi-C data processing through published studies on quantifying the reproducibility and quality of Hi-C libraries2,3 (**Fig. 2A**) and TAD calling (**Fig. 2B**)4. Our TAD caller uses a network-modularity-based approach and modified Louvain algorithm to identify TADs within Hi-C matrices at both low and high resolutions.

A screenshot of a cell phone

Description automatically generated**Figure 2.** Hi-C data analytic tools. (**A**) HiC-spector2 reproducibility score Q distribution across biological replicates, cell types, and pseudo replicates2. Q scores range from zero to one, where a score of one indicates that paired Hi-C matrices are highly reproducible from each other. Pseudo replicates were obtained by combining two replicates and then down sampling of paired-end reads. (**B**) Hi-C intra-chromosomal contact matrix of human embryonic stem cells (hESC, *top*)4 and its representative TADs called by MrTADFinder (golden yellow triangles) at different gamma resolution parameters (*bottom*). Red and blue triangles represent TADs in hESC and IMR90 cells called by Dixon et al., 20125.

Haplotype phasing: Our proposed workflow starts with designing a variant calling module that is followed by phasing and assembling the diploid genome (**Fig. 4A**). We have successfully performed SNV calling using an initial draft of our module. We show the success of variant calling with respect to the amount of coverage using various Hi-C data using the individual NA128786 (**Fig. 4B**). We developed an information metric that weights the captured rare variants more than the common variants, as common variants can be imputed using population genomics data. We showed that Hi-C performs as well as whole genome sequencing (WGS) in terms of capturing the correct SNVs as well as a low false-positive rate (**Fig. 4B**).

A picture containing screenshot

Description automatically generated

**Figure 4.** (**A**)Schematic of a pipeline that generates personal diploid genome using Hi-C reads as input. We will first map the reads to human reference genome. We will then call SNVs using the mapped reads. After filtering for the Hi-C specific biases from the called SNVs, we will phase the remaining SNVs either using Hi-C based methods or using population genetic based methods. Phased variants are then going to be used as input to our in-house vcf2diploid script with the reference genome to build haplotype-specific fasta files. These fasta files will then be used to phase the Hi-C interactions to haplotype resolved versions. (**B**) Pointwise mutual information (pmi) is a bit metric that measures the amount of correctly identified SNPs from an assay by weighing the variants by its frequency in the population. The pmi calculated for the individual NA12878 using the called variants from high resolution Hi-C data using 1,000 Genomes variants as the gold standard and SNP-chip as comparison. (**C**) Construction of a personal genome by vcf2diploid tool is made by incorporating personal variants into the reference genome. The output is the two (paternal and maternal) haplotypes of the personal genome. During the construction step, the reference genome is represented as an array of nucleotides with each cell representing a single base. Iteratively, the nucleotides in the array are being modified to reflect personal variations. Once all the variations have been applied, a personal haplotype is constructed by reading through the array.

We have also applied the initial version of our haplotype resolved Hi-C pipeline (**Fig. 5A**) to a Hi-C data of a normal gastrocnemius medialis tissue from a female donor hosted by the ENCODE consortium. Personal genome of this donor was also assembled by the consortium via Hi-C reads-guided phasing. We have successfully generated Hi-C contact matrices for two haplotypes (chr20 is shown in **Fig. 5C**). We then performed our statistical analysis and were able to identify statistically significant allelic Hi-C interactions (**Fig 5D**).

A close up of a map

Description automatically generated

**Figure 5.** (**A**)Schematic of personal diploid Hi-C map generation pipeline using personal genomes as reference. We will first map the reads to the diploid fasta files. We will then annotate the reads as haplotype 1 or 2 based on the alignment scores. We will pool the reads separately for each haplotype and run a traditional Hi-C contact matrix generation pipeline to generate maps for each haplotype.(**B**)Schematic of statistical significance analyses for allelic imbalance in chromatin interactions. We will first calculate the difference in interactions between different haplotypes. We will then generate two separate contact maps using the reference mapped Hi-C contact matrix and calculate the difference in interactions between these psuedo-haplotypes. This will be repeated for N times to generate an empirical null distribution. P-values are then calculated based on this null distribution.(**C**)We applied the pipeline to Hi-C data obtained from somatic tissues of an individual with known diploid personal genome to obtain haplotype specific Hi-C contact maps. **(D)** We analyzed the resulting contact maps in(C)to calculate the allele-specific chromatin interactions. Here we show the differential Hi-C interactions in different haplotypes as volcano plots with respect to the p-values we obtain from(B).We also show the difference between the randomly generated haplotype specific contact matrices and the real haplotype specific contact matrices.

Integrated analyses: We have considerable experience developing large-scale integrative and comparative analyses. This experience is gained from our participation in the ENCODE, modENCODE, 1000 Genomes, and KBase consortia. In particular, these consortia experiences have led to the development of numerous tools and frameworks that have been widely adapted by the genomics community: We co-developed the ENCODE ChIP-seq data processing pipeline7 such as PeakSeq8, a computational tool for identifying TF binding sites. PeakSeq is a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq. In addition, we have also developed MUSIC9 which enables multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at narrow and broad peak breadths. This tool is particularly applicable to studies of histone modifications and previously uncharacterized TFs, both of which may display both broad and punctate regions of enrichment. Another tool highly relevant to this proposal is AlleleSeq10, a computational tool for the detection of allelic sites, including those associated with gene expression and TF binding using RNA-Seq and ChIP-seq datasets. Using AlleleSeq, we have spearheaded and published allele-specific analyses as part of their work in several major consortia, including ENCODE and the 1000 Genomes Project. By constructing regulatory networks based on allele-specific binding and allele-specific expression of TFs and their target genes, we further revealed substantial coordination between allele-specific binding and expression (**Fig. 6**). Furthermore, AlleleSeq provides lists of detected allelic variants, and the personal diploid genome and transcriptome of NA12878.

A screenshot of a cell phone

Description automatically generated

**Figure 6.** Workflow for the uniform processing of functional genomics data from hundreds of individuals; including the assessment of allele-specific expression, binding events, and construction of AlleleDB11. AlleleSeq was recently updated and applied to 1,139 RNA-seq and ChIP-seq datasets for 382 cell lines in the 1000 Genomes Project. For each cell line, multiple RNA-seq and ChIP-seq datasets were harmonized and aggregated separately, and then each data type was uniformly reprocessed using the updated AlleleSeq. This allowed the annotation of the 1000 Genomes Project SNP catalog with allelic information. To house the results, we also constructed a database, AlleleDB.

With these tools, a comprehensive regulatory map of the adult cortex12 (**Fig. 7**) was built for the PsychENCODE consortium. In this regulatory map, promoter-interacting regions were highly enriched in promoter and enhancer states, indicating that physically interacting regions share similar epigenetic profiles (**Fig. 7B**). Importantly, we have shown that chromatin interactions mediate gene regulation: the number of enhancers that interact with a given promoter linearly correlates with the target gene expression level (**Fig. 7C**). Chromatin interactions also mediate gene co-regulation, such that genes that interact with higher frequency are co-expressed (**Fig. 7D**). Further, while genes that interact with promoters or enhancers are highly expressed, genes that interact with repressive marks and bivalent enhancers are lowly expressed (**Fig. 7E**). Taken together, these data support that **we are capable of obtaining and analyzing multi-omic datasets, which enables decoding the gene regulatory relationships in DS conditions**.

A close up of a map

Description automatically generated

**Figure 7.** Regulatory relationships in the adult cortex. (**A**) Gene regulatory networks generated by compiling Hi-C, TFs, and enhancers. (**B**) Regions that interact with transcription start sites (TSS) are enriched with other TSS and enhancers. (**C**) The number of assigned enhancers to promoters are correlated to the target gene expression level. (**D**) Physically interacting genes exhibit higher correlations in their expression level. (**E**) Genes that interact with enhancers or promoters are more highly expressed than genes that interact with bivalent enhancers or repressive marks.

**References**

1. [Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).](http://paperpile.com/b/Vq8ydx/UIJSB)
2. [Yan, K.-K., Yardımcı, G. G., Yan, C., Noble, W. S. & Gerstein, M. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199 (2017).](http://paperpile.com/b/Vq8ydx/u6kK5)
3. [Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* **20**, (2019).](http://paperpile.com/b/Vq8ydx/Dl5DI)
4. [Yan, K.-K., Lou, S. & Gerstein, M. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput. Biol.* **13**, (2017).](http://paperpile.com/b/Vq8ydx/hiBc0)
5. [Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).](http://paperpile.com/b/Vq8ydx/KUoy3)
6. [Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).](http://paperpile.com/b/Vq8ydx/wgX7e)
7. [Data Processing Pipelines – ENCODE.](http://paperpile.com/b/Vq8ydx/qPZQU) <https://www.encodeproject.org/pipelines/>[.](http://paperpile.com/b/Vq8ydx/qPZQU)
8. [Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75 (2009).](http://paperpile.com/b/Vq8ydx/e1FrK)
9. [Harmanci, A., Rozowsky, J. & Gerstein, M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.* **15**, 474 (2014).](http://paperpile.com/b/Vq8ydx/BGsGc)
10. [Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, (2011).](http://paperpile.com/b/Vq8ydx/E83nd)
11. [Chen, J. *et al* A Uniform Survey of Allele-Specific Binding and Expression over 1000-Genomes-Project Individuals. *Nature Comm.,* **7**(1), 1–13 (2016).](http://paperpile.com/b/cboIz4/xqWO)
12. [Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).](http://paperpile.com/b/Vq8ydx/UIJSB)