Analysis of functional genomics data. We have extensive experience in developing and maintaining pipelines for the quality assessment and processing of different types of functional genomics data. Additionally, we have been leading pipeline development efforts in a number of large-scale genomics consortia such as ENCODE, modENCODE, Gencode and the 1000 Genomes Project. In terms of genome annotation, we have pioneered the identification of non-coding transcription and novel transcribed elements both in the human species and in model organisms **[1-5]**: among these, *incRNA* predicts novel non-coding RNAs (ncRNAs) using known ncRNAs of various biotypes, and *FusionSeq* detects transcripts that arise due to trans-splicing or chromosomal translocations. Our group has also led efforts for the annotation and analysis of pseudogenes in the framework of the Gencode project. In collaboration with the UCSC and HAVANA teams, we have developed a variety of methods to identify pseudogenes **[6-9]**. These include *PseudoSeq* and *PseudoPipe*, which take as input all known protein sequences in the genome and use homology search to identify disabled copies of functional paralogs (referred to as pseudogene parents). As concerns transcriptome analysis, in the framework of the ENCODE and modENCODE projects we have curated pipelines for gene expression quantification that ensure uniform processing and comprehensive annotation of RNA-seq data, allowing direct comparison of gene expression patterns across multiple species **[10-11]**. Our pipeline *IQSeq* calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-Seq data **[12]**. To ensure the anonymization of confidential sequence information that can be potentially extracted from RNA-seq reads, we have developed the Mapped Read Format (MRF), a compact data summary format to store both short and long read alignments, as well as an accompanying suite of tools (*RSEQtools*) **[13]**. We have also significantly contributed to the analysis of extracellular small RNA-Seq experiments with our *exceRpt* pipeline **[14]**, which we developed in the framework of the NIH Extracellular RNA Communication Consortium. Besides transcriptomic data, we have created a number of tools for epigenome analysis. First, we developed PeakSeq **[15]**, a tool for the genome-wide identification of TF binding sites from ChIP-Seq data, which was extensively employed by the ENCODE consortium. Second, we developed MUSIC **[16]**, a peak caller that performs multiscale decomposition of ChIP-seq signal, which is applicable to studies of histone modifications enabling detection of broad and punctate regions of enrichment.

References:

[1] Bertone, Paul et al. “Global identification of human transcribed sequences with genome tiling arrays.” *Science* vol. 306,5705 (2004): 2242-6. doi:10.1126/science.1103388

[2] Clark, Michael B et al. “The reality of pervasive transcription.” *PLoS biology* vol. 9,7 (2011): e1000625; discussion e1001102. doi:10.1371/journal.pbio.1000625

[3] Lu, Zhi John et al. “Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data.” *Genome research* vol. 21,2 (2011): 276-85. doi:10.1101/gr.110189.110

[4] Rozowsky, Joel S et al. “The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.” *Genome research* vol. 17,6 (2007): 732-45. doi:10.1101/gr.5696007

[5] Sboner, Andrea et al. “FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.” *Genome biology* vol. 11,10 (2010): R104. doi:10.1186/gb-2010-11-10-r104

[6] Zhang, Zhaolei et al. “PseudoPipe: an automated pseudogene identification pipeline.” *Bioinformatics* vol. 22,12 (2006): 1437-9. doi:10.1093/bioinformatics/btl116

[7] Zheng, Deyou, and Mark B Gerstein. “A computational approach for identifying pseudogenes in the ENCODE regions.” *Genome biology* vol. 7 Suppl 1,Suppl 1 (2006): S13.1-10. doi:10.1186/gb-2006-7-s1-s13

[8] Pei, Baikang et al. “The GENCODE pseudogene resource.” *Genome biology* vol. 13,9 R51. 26 Sep. 2012, doi:10.1186/gb-2012-13-9-r51

[9] Sisu, Cristina et al. “Comparative analysis of pseudogenes across three phyla.” *Proceedings of the National Academy of Sciences* vol. 111,37 (2014): 13361-6. doi:10.1073/pnas.1407293111
[10] Djebali, Sarah et al. “Landscape of transcription in human cells.” *Nature* vol. 489,7414 (2012): 101-8. doi:10.1038/nature11233
[11] Gerstein, Mark B et al. “Comparative analysis of the transcriptome across distant species.” *Nature* vol. 512,7515 (2014): 445-8. doi:10.1038/nature13424
[12] Du, Jiang et al. “IQSeq: integrated isoform quantification analysis based on next-generation sequencing.” *PloS one* vol. 7,1 (2012): e29175. doi:10.1371/journal.pone.0029175
[13] Habegger, Lukas et al. “RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.” *Bioinformatics* vol. 27,2 (2011): 281-3. doi:10.1093/bioinformatics/btq643
[14] Rozowsky, Joel et al. “exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling.” *Cell systems* vol. 8,4 (2019): 352-357.e3. doi:10.1016/j.cels.2019.03.004
[15] Rozowsky, Joel et al. “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.” *Nature biotechnology* vol. 27,1 (2009): 66-75. doi:10.1038/nbt.1518
[16] Gerstein, Mark B et al. “Architecture of the human regulatory network derived from ENCODE data.” *Nature* vol. 489,7414 (2012): 91-100. doi:10.1038/nature11245

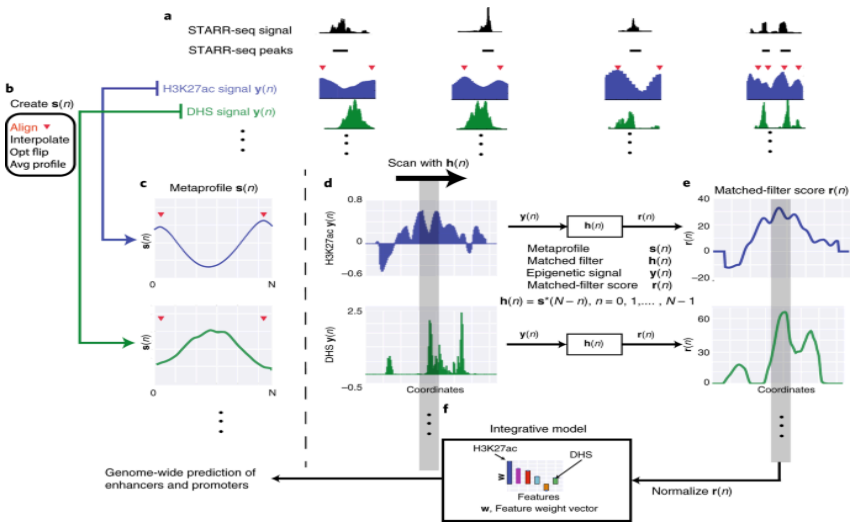

Figure 1: Flowchart of the matched-filter model. (**a**): We identified the double-peak pattern in the H3K27ac signal close to STARR-seq peaks. (**b**): We aggregated the H3K27ac signal around these regions after aligning the flanking maxima, using interpolation and smoothing on the H3K27ac signal. (**c**): The same operations were performed on other histone signals and DHS to create metaprofiles in other dependent epigenetic signals. (**d**): Matched filters were used to scan the histone and/or DHS datasets to identify the occurrence of the corresponding pattern in the genome. (**e**): The matched-filter scores are high in regions where the profile occurs (gray region shows an example), but are low when only noise is present in the data. (**f**): The individual matched-filter scores from different epigenetic datasets were combined using integrated model to predict active promoters and enhancers in a genome-wide fashion.

Measuring the regulatory potential of non-coding regions. Besides the analysis of functional genomics data, we have large experience analyzing data from massively parallel reporter assays. Analysis of data such as that obtained from STARR-seq assays brings in an additional level of complexity, since in this kind of experiments the coverage is typically non-uniform, overdispersed, and often confounded by sequencing biases such as GC content, or other factors like RNA secondary structure and thermodynamic stability. To overcome these limitations, we developed a negative binomial regression framework for uniformly processing STARR-seq data, *STARRPeaker*, which we used to generate comprehensive and unbiased catalogs of putative enhancers in various ENCODE cell lines **[1]**. We further integrated epigenomics and STARR-seq data to improve the prediction of enhancers across multiple species. To do so, we developed *matched-filter* (**Figure 1**) a framework that uses *Drosophila* STARR-seq peaks to create shape-matching filters based on meta-profiles of epigenetic features **[2]**. We integrated the resultant features with supervised machine-learning algorithms to predict enhancers in both *Drosophila* and mammals. Finally, we have also extensively contributed to the identification of cell-type specific enhancers with *DECODE*, a deep-learning framework that improves the annotation of enhancers by precise detection of their genomic boundaries **[3]**.

References:

[1] Lee, Donghoon et al. “STARRPeaker: uniform processing and accurate identification of STARR-seq active regions.” *Genome biology* vol. 21,1 298. 8 Dec. 2020, doi:10.1186/s13059-020-02194-x

[2] Sethi, Anurag et al. “Supervised enhancer prediction with epigenetic pattern recognition and targeted validation.” *Nature methods* vol. 17,8 (2020): 807-814. doi:10.1038/s41592-020-0907-8

[3] Chen, Zhanlin et al. “DECODE: a Deep-learning framework for Condensing enhancers and refining boundaries with large-scale functional assays.” *Bioinformatics* vol. 37,Suppl_1 (2021): i280-i288. doi:10.1093/bioinformatics/btab283

Identifying and interpreting genetic variants. We have pioneered the identification of genetic variants, in particular large structural variants (SVs), for the advancement of personalized genomics. We developed *Paired-End Mapper* (PEMer), a toolkit for the detection of SVs from paired-end sequencing data **[1]**, and *CNVnator*, a pipeline for the discovery and annotation of typical and atypical CNVs from family and population genome sequencing **[2]**. We have also led efforts for the identification of allele-specific variants. Our pipeline *AlleleSeq* integrates an individual's genomic variation data (SNVs, indels, and SVs) into the reference genome, phases information of heterozygous variants producing maternal and paternal haplotypes, and maps genomic loci that display imbalance in gene expression or chromatin binding between the two alleles (allele-specific events) **[3]**. We used *AlleleSeq* to construct the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878 **[3]**, and to build 382 personal genomes using the variant call sets from the 1000 Genomes Project **[4]**. Furthermore, using the extensive Roadmap dataset, we constructed a high-resolution map that reveals allelic imbalances in DNA methylation, histone marks, and transcription across 71 epigenomes from 36 distinct cell and tissue types from 13 donors **[5]**. We recently expanded this pipeline to call allele-specific genomic elements, such as genes or regulatory regions, giving rise to our updated tool, *AlleleSeq2*. We applied *AlleleSeq2* to the EN-TEx resource encompassing ~1.6K datasets from four donors (~30 tissues x 15 assays) and generated the largest catalog (>1M) of allele-specific loci available to date in the human genome **[6]**. We leveraged this catalog to develop a deep-learning transformer model that can predict the allele-specific activity based only on local nucleotide-sequence context (**Figure 2**) highlighting the importance of transcription-factor-binding motifs particularly sensitive to variants.

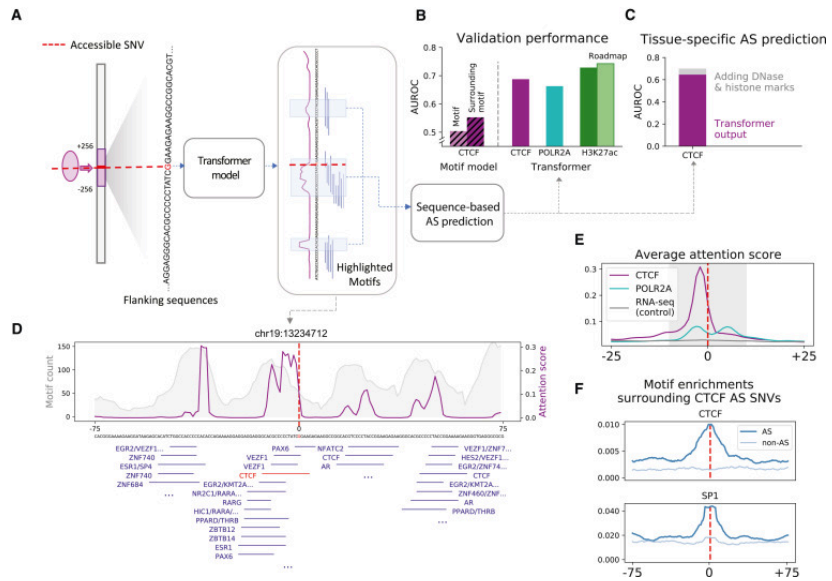

Figure 2: **Deep-learning model predicting AS activity from nucleotide sequence.** (**A**) Schematic of the sequence-based predictive model. (**B**) Average performance of models predicting AS activity. (**C**) Performance of a tissue-specific model for CTCF. (**D**) Attention patterns learned by the model. (**E**) Average attention pattern of sequence-based models for various assays. (**F**) Motif enrichment surrounding the AS CTCF SNV agrees with the average attention pattern in (**E**).

References:
[1] Korbel, Jan O et al. "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data." *Genome biology* vol. 10,2 R23. 23 Feb. 2009, doi:10.1186/gb-2009-10-2-r23
[2] Abyzov, Alexej et al. "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome research* vol. 21,6 (2011): 974-84. doi:10.1101/gr.114876.110
[3] Rozowsky, Joel et al. "AlleleSeq: analysis of allele-specific expression and binding in a network framework." *Molecular systems biology* vol. 7 522. 2 Aug. 2011, doi:10.1038/msb.2011.54

[4] Chen, Jieming et al. “A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.” *Nature communications* vol. 7 11101. 18 Apr. 2016, doi:10.1038/ncomms11101
[5] Onuchic, Vitor et al. “Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci.” *Science* vol. 361,6409 (2018): eaar3146. doi:10.1126/science.aar3146
[6] Rozowsky, Joel et al. “The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models.” *Cell* vol. 186,7 (2023): 1493-1511.e40. doi:10.1016/j.cell.2023.02.018

Analyzing biological networks to elucidate the effects of genomic variants. Reconstructed networks can help infer the direct and indirect effects of genomic variants. For example, we have used network properties such as centrality to evaluate the functional significance of genomic variants **[1]**. Genomic variants can also lead to disruptions of network connections. We developed *DiNeR* for identifying disruptions of TF co-regulation by variants and analyzing their consequences **[2]**. On a larger scale, some network perturbations may propagate to cause major network rewiring. We developed the *TopicNet* method to measure such rewiring in transcriptional regulatory networks **[3]**. We have also applied this idea to study network rewiring in cancer cells, as part of our efforts toward producing a general resource for cancer research based on ENCODE data **[4]**. In addition to studying individual networks, ultimately it is necessary to study multiple networks jointly to understand how they affect each other. Finally, we have recently developed a unified pre-trained language model, *MolLM*, to integrate biomedical text and improve 2D and 3D molecular representations **[5].**

References:

[1] Khurana, Ekta et al. “Integrative annotation of variants from 1092 humans: application to cancer genomics.” *Science* vol. 342,6154 (2013): 1235587. doi:10.1126/science.1235587

[2] Zhang, Jing et al. “DiNeR: a Differential graphical model for analysis of co-regulation Network Rewiring.” *BMC bioinformatics* vol. 21,1 281. 2 Jul. 2020, doi:10.1186/s12859-020-03605-3

[3] Lou, Shaoke et al. “TopicNet: a framework for measuring transcriptional regulatory network change.” *Bioinformatics* vol. 36,Suppl_1 (2020): i474-i481. doi:10.1093/bioinformatics/btaa403

[4] Zhang, Jing et al. “An integrative ENCODE resource for cancer genomics.” *Nature communications* vol. 11,1 3696. 29 Jul. 2020, doi:10.1038/s41467-020-14743-w

[5] Tang, Xiangru et al. “*MolLM*: A Unified Language Model for Integrating Biomedical Text with 2D and 3D Molecular Representations.” *bioRxiv* (2024), doi: https://doi.org/10.1101/2023.11.25.568656