Utilizing Hi-C sequencing and graph genome alignment to uncover structural variant effects in personal genomes

Isabella Wu, Matthew Jensen, Mark Gerstein
Department of Biomedical Informatics, Yale University, New Haven, CT

Introduction

Personal genome sequencing is a rapidly advancing field that is transforming the way that human health is understood and managed, providing unprecedented insights into the genetic predispositions, susceptibility to disease, and response to treatments of an individual, rather than a population. This innovative technology enables us to decode an individual's complete genome, encompassing all their unique genetic variants that cannot be captured by the reference genome. This allows us, for the first time, to investigate the effects and consequences of these genetic variants.

Structural Variants (SVs) are large genetic variants, including insertions or deletions, that play a crucial role in shaping gene regulation, functionality, and ultimately, the progression of diseases. However, despite their significant involvement in disease progression, determining the pathogenicity of SVs and their effects remains a challenge due to their complexity, which limits our understanding of their precise impacts. In recent months, however, the EN-TEx project has pioneered the systematic mapping of functional datasets for gene regulation onto personal genomes [1], introducing a groundbreaking approach to uncover the effects of SVs with unprecedented precision.

The primary objective of this project is to utilize Hi-C sequencing and graph genome alignment techniques to investigate and uncover the tissue-specific effects of SVs on regulatory signals and altered gene regulation through personal genome mapping. We seek to uncover novel effects and better understand the functional consequences of SVs in complex diseases within the context of personal genomes.

EN-TEx Diploid Personal Genomes

Previous studies exploring genetic variant effects have predominantly relied on the haploid reference genome for comparison, which represents an 'average' human genome that lacks individual-specific variants. This hinders accurate variant identification and can result in incomplete or inaccurate analysis of variants and their impact on gene regulation.

To overcome this limitation and bridge the gap between personal genomes and functional genomics, EN-TEx was created. Unlike previous efforts, EN-TEx sequenced diploid personal genomes, which more accurately captures an individual's genetic variants by accounting for the differences between maternal and paternal chromosomes, or haplotypes. Fig. 1 illustrates the diploid personal genomes. This approach enables the direct annotation of observed genetic differences to represent an individual's unique genetic variations, particularly those involved in regulatory mechanisms.

EN-TEx sequenced the personal genomes of four individuals (two female, two male), designated as ENC001 through ENC004. A total of 1,635 experiments, spanning 15 different assays targeting various epigenomic, transcriptomic, and proteomic aspects, were conducted from 30 tissues from each of the four individuals. This rich functional data diversity, when mapped directly onto personal genomes, allows us to annotate the effects of SVs on a wide range of processes, including chromatin structure, and DNA methylation.

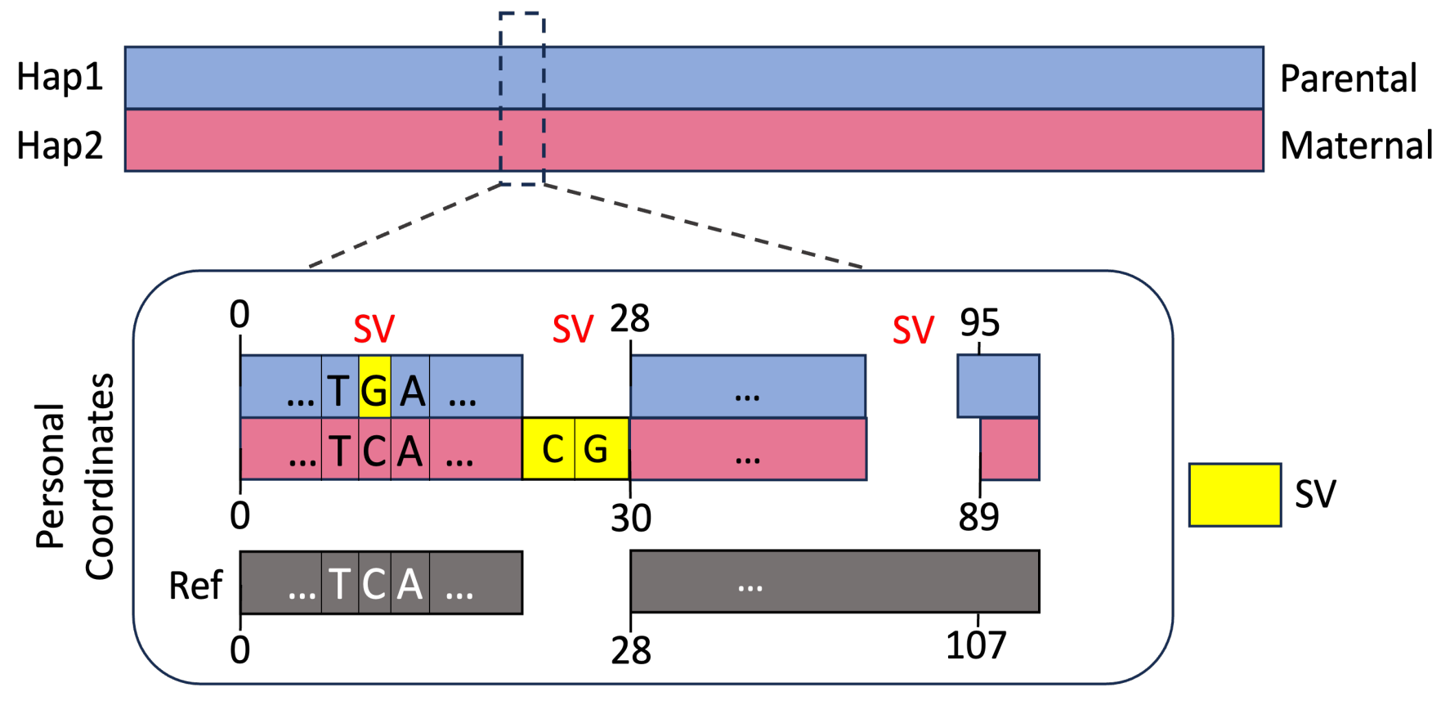

Figure 1. EN-TEx diploid personal genome. The maternal (red) and paternal (blue) haplotypes form a diploid genome that enables us to capture an individual's SVs, including insertions and deletions, by using each haplotype as a reference for the other. The reference genome is shown in grey for comparison.

Hi-C Pipeline Analysis

The 3D human genome is organized into topologically associated domains (TADs), structured regions containing regulatory loops that control interactions between regulatory sites and genes. To form a regulatory loop, DNA bends back on itself, and distant loci come into close spatial proximity with each other. SVs that disrupt these regulatory loops consequentially disrupt TADs and the 3D structure of DNA, leading to aberrant interactions between regulatory elements and genes. These disturbances can modify gene expression patterns and potentially affect disease onset and progression. We utilized Hi-C sequencing, a state-of-the-art technique that quantifies physical interactions within the genome, to identify disrupted regulatory loops and TADs. We designed a novel Hi-C analysis pipeline aligned to the personal genomes for the first time. This pipeline produces a comprehensive catalog of TADs disrupted by SVs, which facilitates the discovery of novel regulatory effects exerted by these variants. The development of this pipeline encompassed three main stages: (1) Data preprocessing with the *Juicer* pipeline [2], (2) Calling TADs with *Arrowhead* and loops with *HICCUPs*, and (3) Pinpointing disrupted regulatory regions. An overview of the pipeline is shown in Fig. 2.

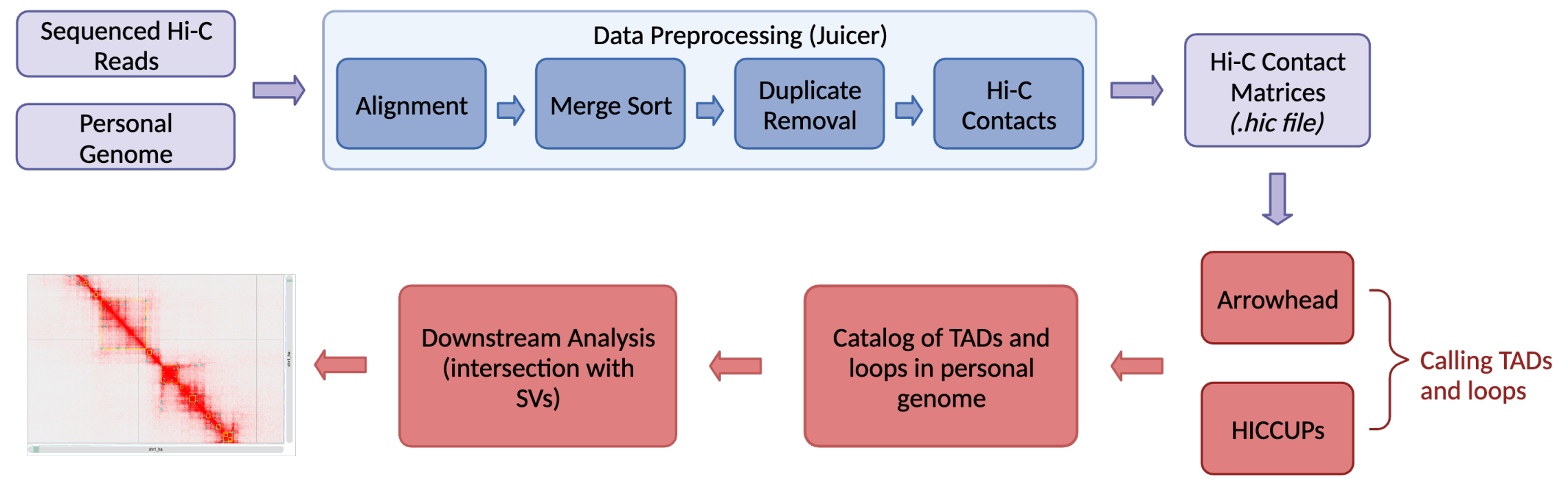
Sequenced Hi-C Reads
Personal Genome
Data Preprocessing (Juicer)
Alignment
Merge Sort
Duplicate Removal
Hi-C Contacts
Hi-C Contact Matrices (.hic file)
Arrowhead
HICCUPs
Calling TADs and loops
Catalog of TADs and loops in personal genome
Downstream Analysis (intersection with SVs)

Figure 2. HI-C methods pipeline. Hi-C reads aligned to the personal genome were preprocessed with *Juicer* into contact matrices. *Arrowhead* and *HICCUPs* were used to call TADs and loops, and downstream analysis was performed.

We first preprocessed the raw reads with the *Juicer* software pipeline. Sequenced Hi-C reads were aligned to the personal genomes using a BWA mem aligner. Subsequently, we filtered out unmapped, abnormal, and duplicate reads, as well as read pairs with low mapping quality to ensure that only high-quality Hi-C contacts are considered. The Hi-C contacts were then used to construct contact matrices for each sample. These matrices represent interactions between different genomic regions within the personal genome in the form of a compressed .hic file, thus capturing TADs and regulatory loops. Fig. 3 depicts the Hi-C contact matrix for individual ENC001, haplotype 1.

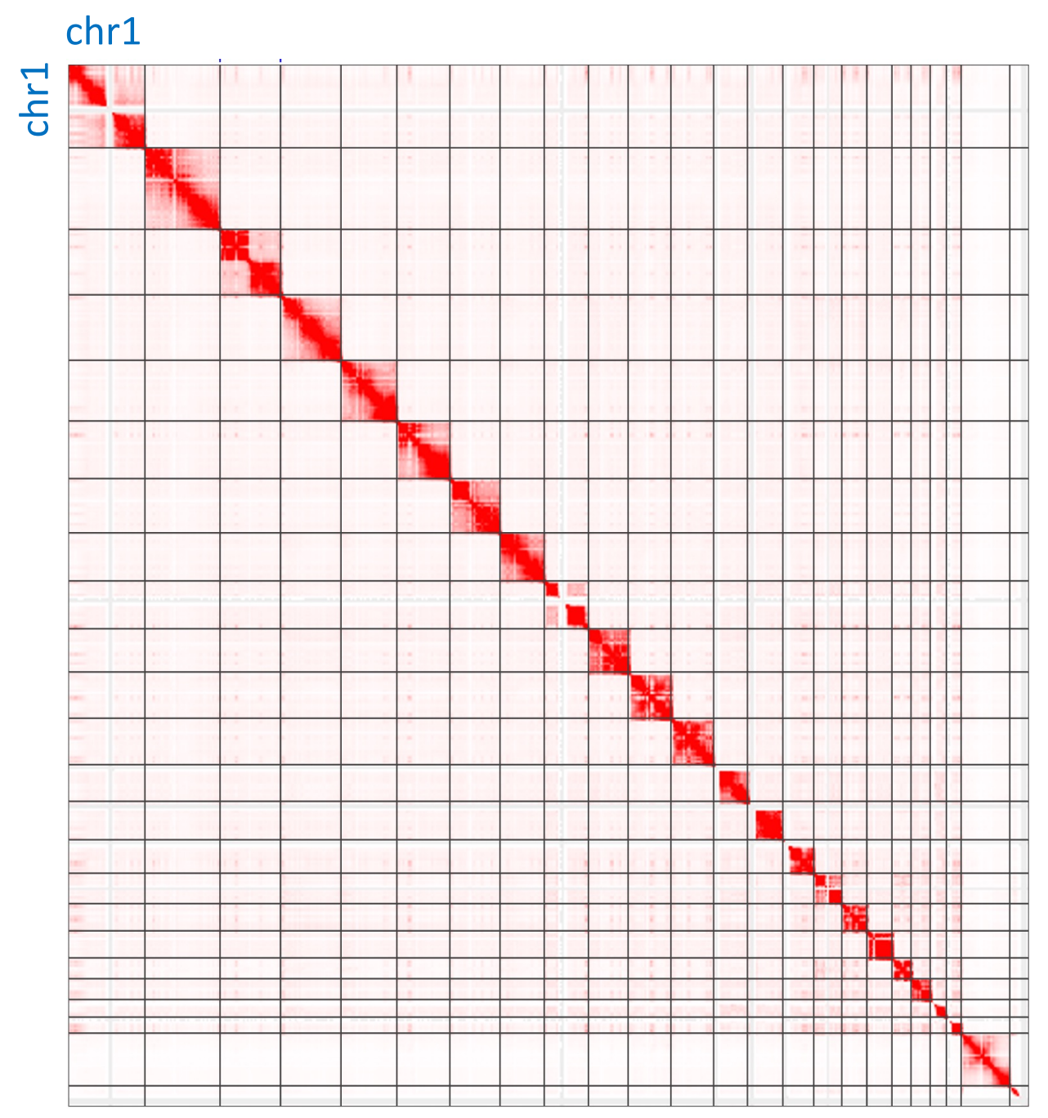

Figure 3. Hi-C contact matrix for individual ENC001 haplotype 1. Each square corresponds to a chromosome, with the intensity of color indicating the frequency of interactions between specific chromosomal regions, illustrating regulatory loops within the genomic structural organization.

We utilized the *Arrowhead* and *HICCUPs* algorithms to call TADs and loops. We then used *Bedtools* to intersect with SVs from the EN-TEx database to identify possible disruptions. We focused on identifying unique disruptions in one individual that are absent in the other individuals. This aids in detecting novel TADs and loops disrupted by SVs, which we term SV-TADs and SV-loops. We then investigated the effects of these disruptions on adjacent genes and potential changes in their expression.

Hi-C Results

This section will focus on results from the Hi-C pipeline for individual ENC001 only. For each haplotype, we focused on finding the number of: 1) total TADs, 2) SV-TADs, 3) SV-TADs within 50 kilobases (kb) proximity to a gene, which could suggest an impact on gene regulation, and 4) SV-TADs present in only the haplotype of interest and absent in any other individual, which could induce unique expression changes within that haplotype (Fig. 4). We repeated these analyses for loops (Fig. 5).

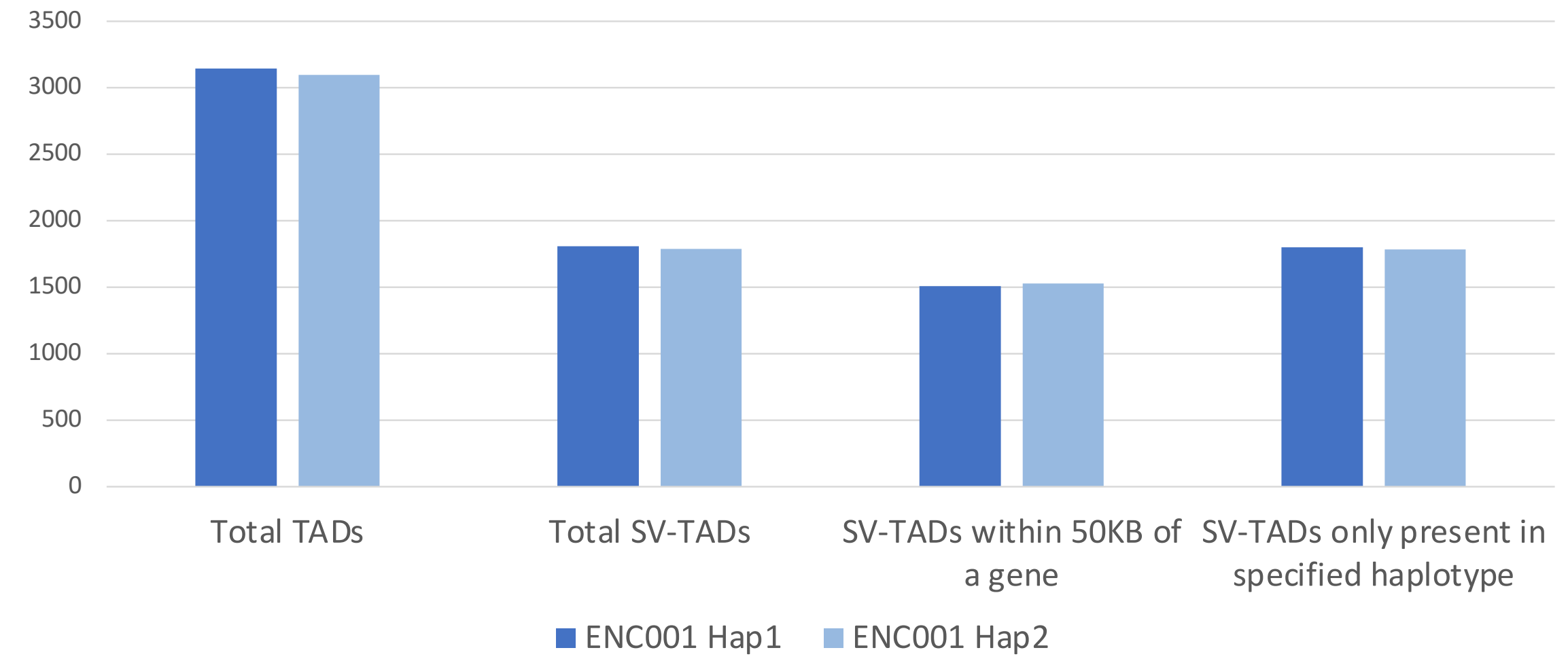

Figure 4. TAD and SV-TAD counts in ENC001 hap1 and hap2. The average TAD length is 256kb.

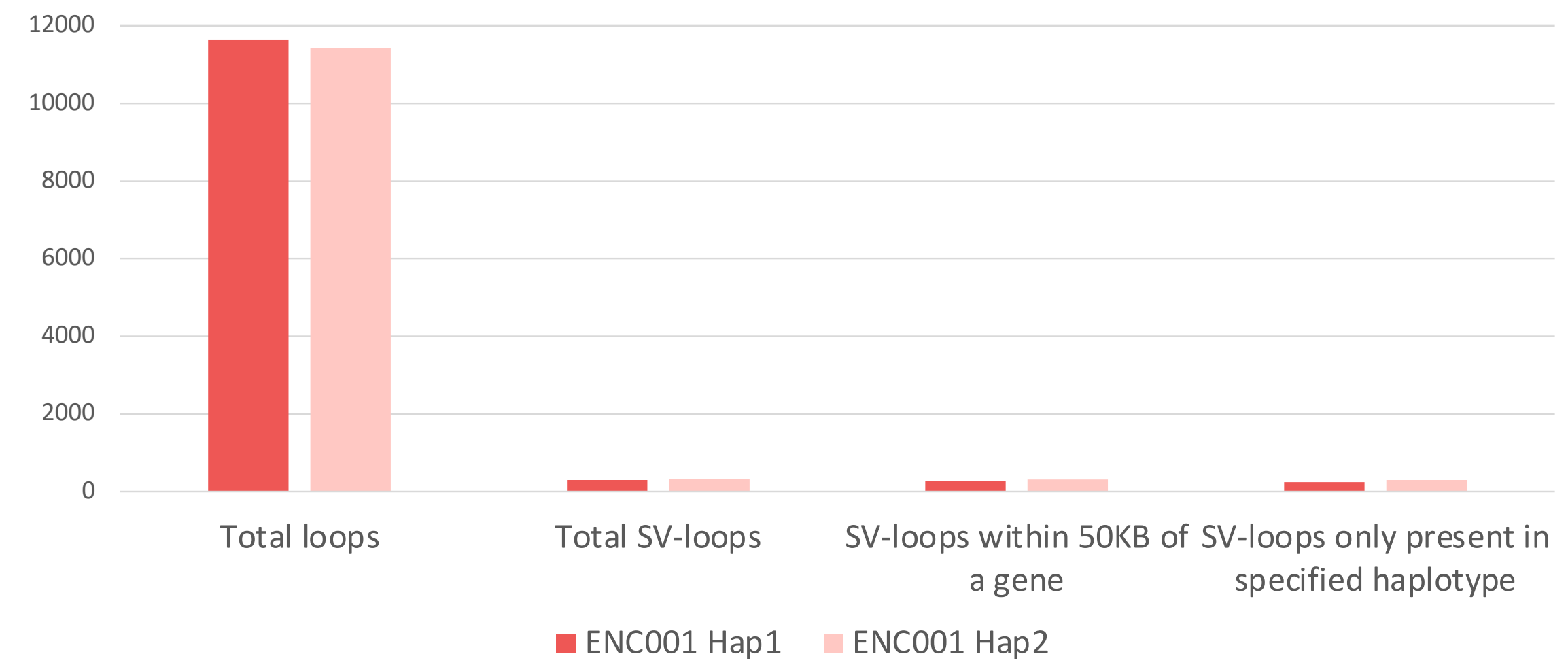

Figure 5. Loops and SV-loops counts in ENC001 hap1 and hap2. The average loop length is 11.8kb.

Fig. 6 shows two novel SV-TADs that differ between haplotype 1 and 2 due to insertion SVs.

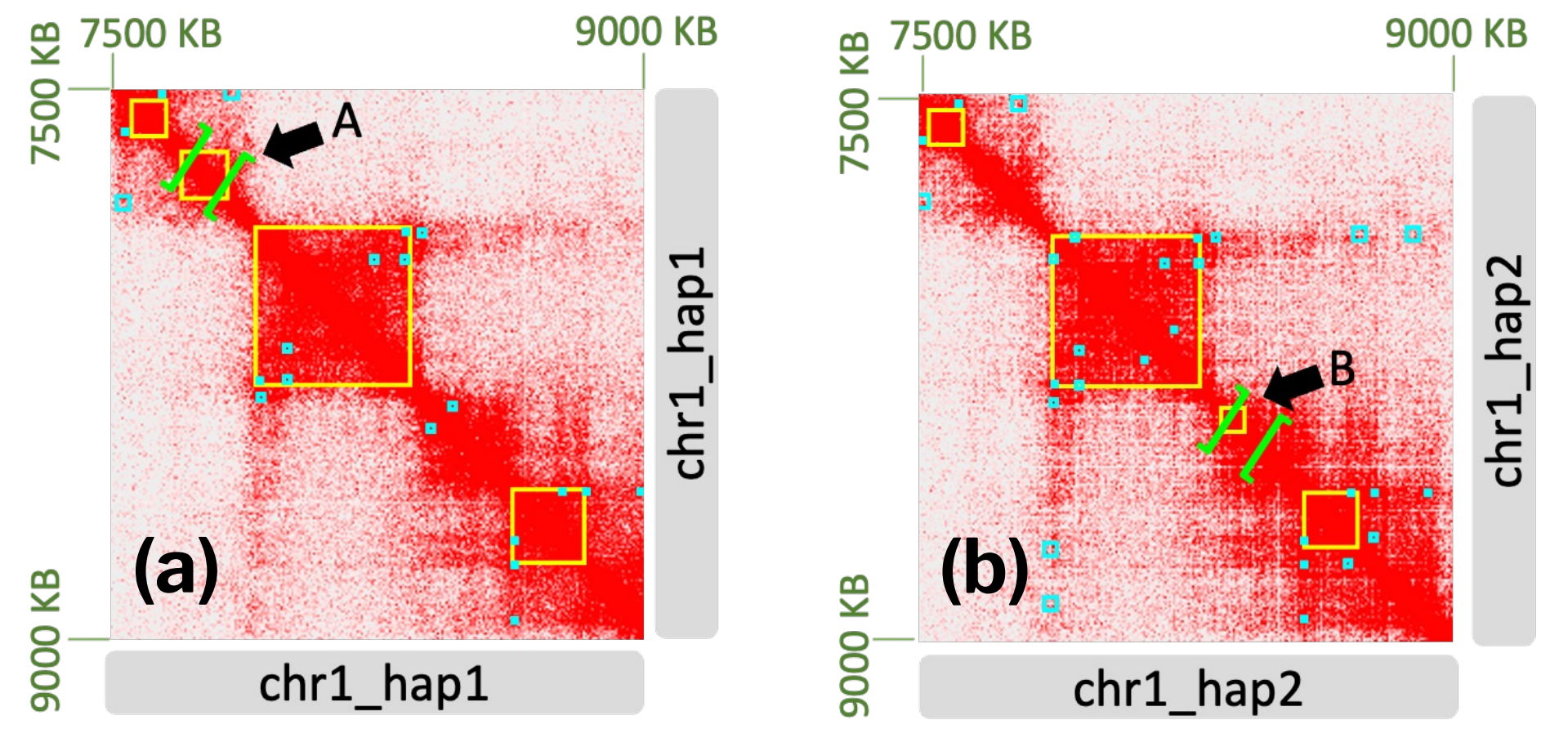

Figure 6. Yellow boxes denote TADs, cyan boxes denote loops, and green brackets denote an insertion SV. **(a)** An insertion SV creates a novel TAD (arrow A) in hap1 that is absent in hap2. **(b)** Conversely, an insertion SV in hap2 creates a novel TAD (arrow B) not found in hap1.

Additionally, we identified individual-specific SV-TADs located within 50kb of a gene that subsequently cause outlier expression of that gene, indicating an alteration in gene expression cause by the SV-TAD. An example of one of these identified SV-TADs is shown in Fig. 7.

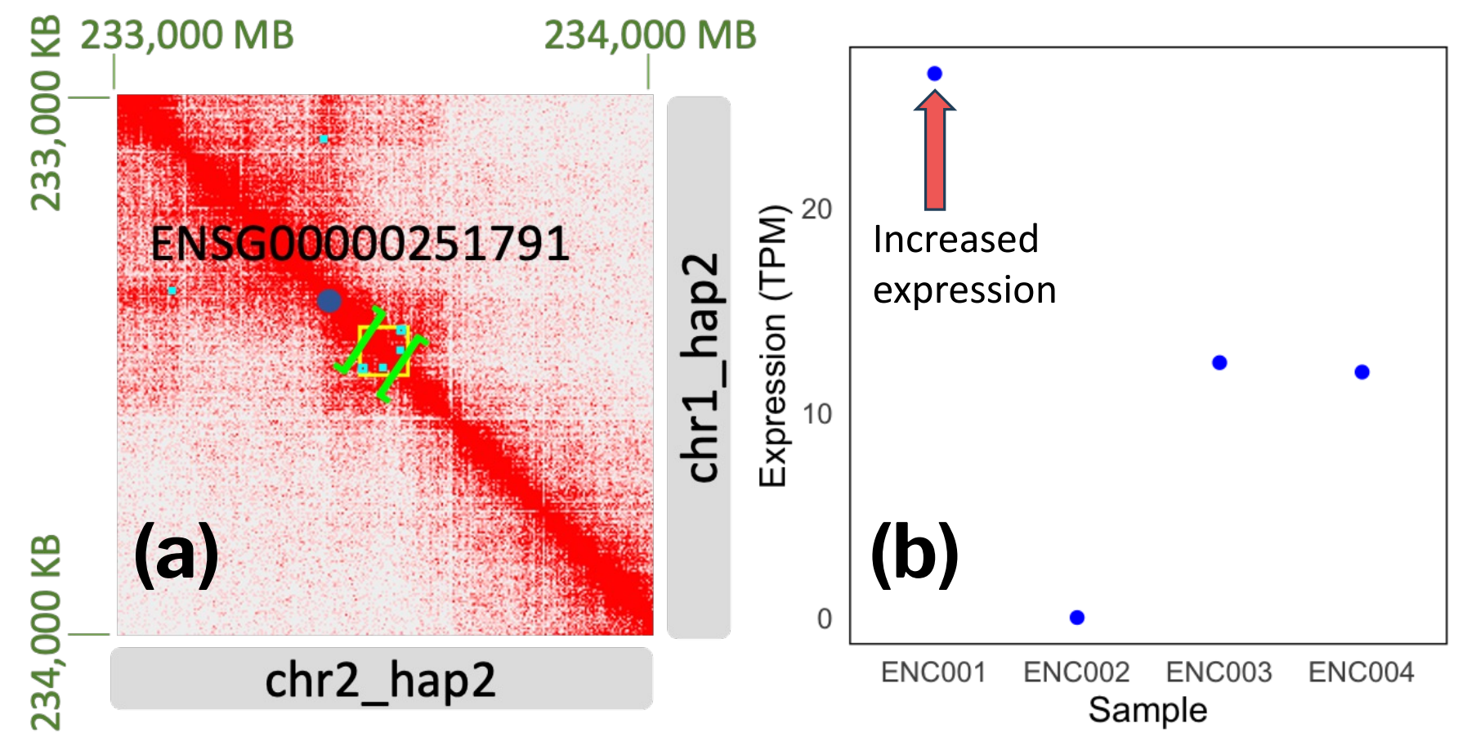

Figure 7. (a) A novel TAD in ENC001 that is unique compared to other individuals. The blue dot marks the gene ENSG00000251791 (SCARNA6). **(b)** SCARNA6 is a potential oncogene. The insertion SV leads to significantly increased expression in ENC001 compared to other individuals, which may explain disease effects in ENC001.

Graph Genome Alignment

While haplotype-specific personal genomes enable the discovery of regulatory elements within novel insertions and complex SVs, there is a need for improved alignment methods of functional genomic data to these genomes. Current techniques have limitations including missing novel regions and low mapping quality. These challenges underscore the need for the development of improved alignment methods to the personal genome.

In recent months, a new alignment approach known as graph genome alignment has emerged. Graph genome alignment provides a flexible and versatile way to capture SVs by representing genetic information as a graph, allowing for a more comprehensive and individualized understanding of an individual's genome. Within this graph-based framework, nodes denote specific genetic sequences or variants, and edges connect these nodes to depict their relationships (Fig. 8). This can help improve alignment to personal genomes to accurately identify novel disrupted regions.

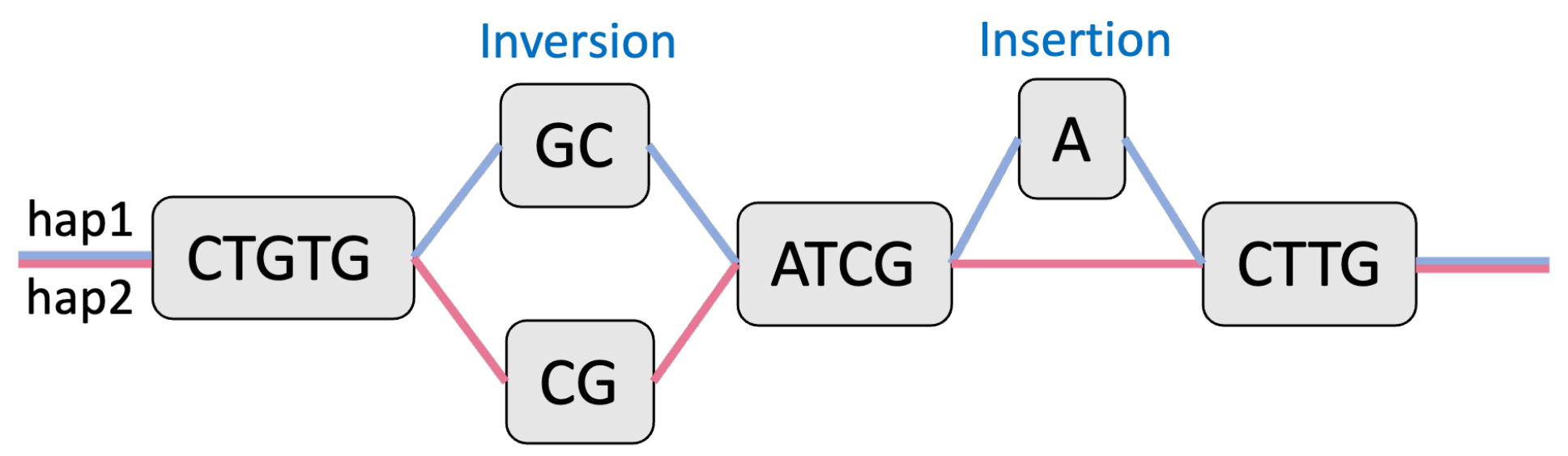

Figure 8. Simplified graph genome representation. Graph genomes can represent both haplotypes at once, so SVs can be captured by "bubbles" with diverging nodes. This figure depicts an inversion and insertion in haplotype 1.

We employed the *Minigraph-cactus* pipeline by Hickey et al. [3] to sequence graph genomes from the personal genome, and identified 131 alignment fragments. However, the final pipeline requires additional scaffolding of the personal genome onto the reference genome chromosomes, a task that the field is currently still addressing. As such, our graph genomes are still a work in progress.

Impact and Future Work

In this project, we showed that personal genomes can reveal unprecedented insights about the effects of SVs on our genome. We identified novel TADs and loops disrupted by SVs that lead to differences in gene expression and potential disease effects. To improve upon the alignment to the personal genome, we are hoping to complete our graph genome alignments to improve our mapping results and reveal novel disruptions by SVs.

This project contributes to the larger EN-TEx project, which aims to integrate many different types of gene regulatory data and systematically identify the effects of SVs. In the future, our findings from our Hi-C sequencing will be fed into a larger machine learning model, to predict the tissue-specific effects of newly discovered SVs. We will further refine and implement various graph genome alignment techniques to investigate improvement of performance.

Acknowledgements

I want to extend my greatest thank you to my PI Dr. Mark Gerstein and mentor Dr. Matthew Jensen for their support throughout my summer. I also want to extend my greatest thank you to Dr. Hogue and my fellow SRP Quant Class of 2024 for their endless encouragement, advice, and support.

Works Cited

[1] Beatrice Borsari ... Roderic Guigo Thomas R. Gingeras Mark Gerstein Joel Rozowsky, Jiahao Gao. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell, 186(7):1493–1511.e40, March 2023.
[2] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. Nat. Rev. Genet., 12(5):363–376, May 2011.
[3] Hickey, G., et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol (2023). https://doi.org/10.1038/s41587-023-01793-w