Life requires mistakes. If DNA replication machinery operated with perfect fidelity and every damaged nucleotide were faithfully repaired, the forces of selection, drift and gene flow would be completely powerless. Infrequently, the safeguards of genome integrity fail and germline mutations provide the raw material for both evolution and genetic disease. I've long been fascinated by the factors that influence germline and somatic mutation rates. I am keenly interested in discovering "mutator alleles," genetic variants that augment the mutation rate on haplotypes that carry them. Mutator alleles provide a powerful lens into the mutation process; they can reveal the limits of a cell's ability to reduce its mutation rate, and may represent novel drivers of heritable human cancer. Hundreds, if not thousands, of genes are involved in DNA repair, replication, and proofreading, but their collective effects on mutation rates and spectra remain poorly characterized.

Previous work
During my career, I've made important contributions to our knowledge of germline mutation. As a graduate student I precisely measured germline mutation rates in human families (Sasani et al., 2019, *eLife*), and observed high degrees of post-zygotic mosaicism in the developing embryo for the first time. Later, I discovered one of the first examples of a mammalian germline mutator allele (Sasani et al., 2022, *Nature*) by appling quantitative trait locus mapping to a population of recombinant inbred mice. More recently, I developed a novel statistical approach to map the first epistatic interaction between mutator alleles in a mammalian species (Sasani et al., 2024, *eLife*). And now, I've begun to explore the rates and patterns of tandem repeat mutation in the human genome, discovering recurrent *de novo* TR alleles in a single family for the first time (Sasani et al., 2024, *bioRxiv*).

Although the list of genes involved in DNA replication, proofreading, and repair runs long, we know very little about the genetic factors that influence mutation rate evolution. To develop a comprehensive understanding of mutation rate evolution and its effects on human health, we will need to apply new statistical and computational methods to the right datasets. I'm confident that my background and research experience make me well-suited to the task.

Using innovative computational approaches, my group will answer major questions about germline mutation — one of life's most fundamental phenotypes.

Theme 1: Machine learning to find the genomic signatures of mutators (and more)

Statistical approaches like QTL mapping require genotype and phenotype data from careful crosses of inbred model organisms. Large-effect mutator alleles are also far more likely to reach high population allele frequencies in recombinant inbred lines (RILs), since the effects of negative selection are attenuated by strict inbreeding. With these complications in mind, how can we identify regions of the genome that have experienced (or are currently experiencing) the effects of a germline mutator allele in natural populations? I'm confident that new deep learning (DL) approaches, which have already proven "unreasonably effective" for population genetic inference[1], can help.

Contrastive learning to uncover unusual mutation rate histories

Contrastive learning models are trained to embed data in multi-dimensional space so that similar pieces of data are closer together in that "latent" space, and dissimilar pieces of data are further apart (Figure 1). These embeddings are expected to capture salient characteristics of input data types (say, the features that distinguish images of dogs from images of cats), and enable downstream classification of new, unseen inputs. Inspired by recent applications of DL to population genetics[2], my group will develop contrastive learning approaches to embed images of haplotypes in genomic windows (Figure 1a). These models will be trained to minimize the distance between *positive pairs* of haplotype images (simulated from the same demographic model) and maximize the distance between *negative pairs* (simulated from two different demographic models). In the process, the models should learn the features that distinguish even subtly different demographic histories. Once trained, the model will be broadly useful for population genetics inference, and can be fine-tuned on images of regions with and without mutator alleles that have been simulated more rigorously with forwards-in-time methods, such as `SLiM`.

Theme 2: How do inherited genetic variants affect the mutation process in cancer?

[1] Flagel et al. (2019) *Mol. Biol. & Evolution*
[2] Wang et al. (2021) *Mol. Ecol. Resources*

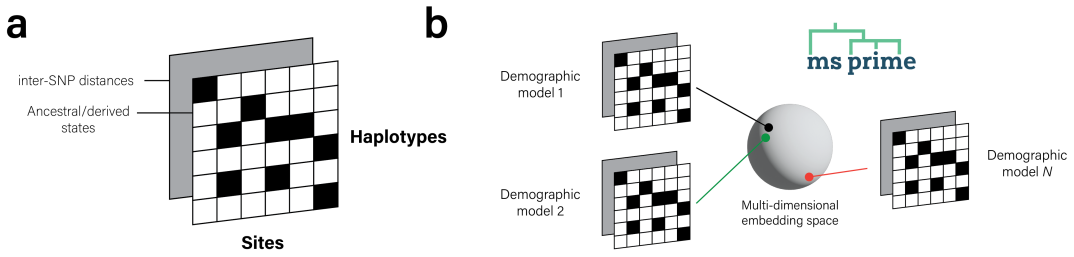

Figure 1: **(Deep) learning about the mutation process.** a) We can represent genomic regions as two-channel "images" in which rows correspond to haplotypes and columns to SNPs. The first channel encodes derived alleles as 1s and ancestral alleles as −1s, while the second channel encodes either absolute genomic positions or inter-SNP distances. b) Contrastive learning models seek to embed input images in multi-dimensional space such that an *anchor* is close to a image belonging to the same class and far apart from an image from a different class. We can simulate thousands of haplotype images from N neutral demographic histories and train a model to extract salient features from these images, thereby "learning" a useful embedding space we can use for downstream classification of regions containing mutators, adaptive admixture, or evidence of selection.

Cancer genomes present a remarkably well-powered system for mutator allele detection. In most sexually-reproducing populations, negative selection will efficiently remove large-effect mutator alleles and recombination will break up linkage between mutators and the excess deleterious mutations they cause. Cancer genomes do not recombine, steadily accumulating mutations that remain perfectly linked to a mutator under the drastically attenuated effects of negative selection.

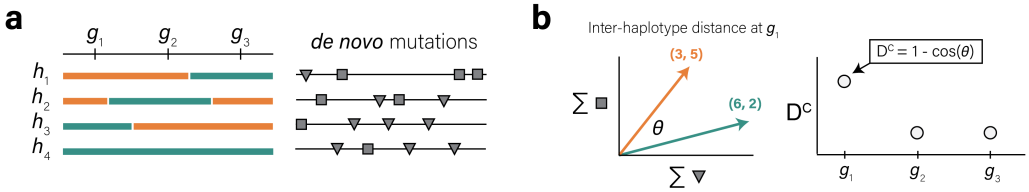

Figure 2: **The aggregate mutation spectrum distance approach.** a) A population of four haplotypes has been genotyped at three informative markers (g_1 through g_3); each haplotype also harbors unique de novo germline mutations. For simplicity in this toy example, de novo mutations are simply classified into two possible mutation types (squares or triangles). b) At each informative marker g_n, we calculate the total number of each mutation type observed on haplotypes that carry either parental allele (i.e., the aggregate mutation spectrum) using all genome-wide de novo mutations. We then calculate the cosine distance between the two aggregate mutation spectra, which we call the "aggregate mutation spectrum distance," and repeat this process for every informative marker.

We will first refine the *aggregate mutation spectrum distance* approach to test segregating genetic variants for their impacts on somatic mutation spectra, and implement approaches to account for the biased effects of genetic relatedness and population structure in mutator allele discovery. We will apply these approaches to large collections of paired tumor-normal sequencing data, enabling the discovery of inherited mutator alleles that exert even subtle influences on the cancer mutation spectrum. We will also characterize the effects of mutator alleles on tumor mutation spectra in a wide variety of tissues in order to understand why mutators active in one cancer subtype (e.g., colorectal) do not exert similar effects in others (e.g., breast).

We will also explore new methods to characterize the mutation process in human cancer. Contrastive PCA (and its nonlinear counterpart, the contrastive variational autoencoder) are two techniques for exploring features enriched in a target dataset with respect to a background dataset. For example, cPCA and cVAE can robustly extract features unique to admixed genomes with respect to the source populations that originally interbred. My research group will leverage these new computational methods to discover features of the mutation spectrum that are unique to tumor and germline-derived sequencing data. Dimensionality reduction techniques, such as non-negative matrix factorization, have proven incredibly useful for the detection of mutational "signatures" that are uniquely active in particular tissues and cancer subtypes. I anticipate that by combining contrastive approaches like cPCA and cVAE with techniques like NMF, we will be able to dissect the mutation process in healthy and diseased tissue

with even greater precision.

Theme 3: How much mutation rate variation are we missing?

Graph genome representations for mutation rate inference

When we want to detect genetic variants in a sample of interest, we typically align that sample's sequencing reads to a linear reference assembly. In humans, mice, and numerous other species, these linear reference genomes are typically derived from a single individual, meaning that the vast majority of the species' genetic variation is missing from the reference. Sequencing reads that harbor such variation often map poorly to the linear reference assembly and are ignored when calling DNA variants, thwarting our ability to discover new mutations in diverse, newly sequenced genomes. New computational methods can represent a collection of genomes as a large, interconnected graph, often termed a "pangenome." Pangenomes capture all of the genetic variation present in a collection of genomes and enable far more accurate sequence alignment and variant calling. My research group will develop new methods to infer mutation spectra from these pangenome representations in order to more precisely characterize mutation spectrum variation within diverse populations and investigate spatial variation in mutation rates and spectra.

Rates and patterns of tandem repeat and large structural variants

Large structural variants, including duplications, deletions, and inversions, as well as tandem repeats (TRs), affect many more base pairs of the human genome than single-nucleotide variants (SNVs). Now, Long-read sequencing technologies have also dramatically improved our ability to assemble eukaryotic genomes, culminating in the recent release of a complete "telomere-to-telomere" human genome. Hundreds of millions of base pairs in centromeric, duplicated, and repetitive regions of the genome — previously referred to as genomic "dark matter" — are now accessible to sequence analysis and can be incorporated into pangenome graph representations. We still know very little about the rates and patterns of large, complex, structural variants in the human genome, and whether inherited genetic variants can influence their developmental timing, spatial distribution, and overall frequencies. By leveraging new sequencing technologies, my group will explore the mutation process in structurally complex regions of the genome and the effects of inherited genetic variation on structural variation.