

ABSTRACT

Although certain subgroups of pediatric leukemia achieve promising outcomes, others still face dismal results marked by initial treatment failure and relapse. Genomic Studies have revealed that pediatric leukemia has distinct genetic profiles compared to adult diseases, characterized by a low mutation burden, diversity of fusion events, and a higher prevalence of germline predisposition. In this dissertation, I focus on developing computational methodologies to assess genetic risk at both the individual-case and population scales. In this dissertation, I developed computational methodologies to assess genetic risk at both the individual-case and population scales. In Chapter 2, I developed and validated a clonal evolution framework using patient-derived xenograft models, which enabled the lineage tracing of specific cancer cell populations in vivo. I applied this framework to longitudinal sequencing data from a relapsed ALL patient (multiple timepoints), demonstrating how clonal dynamics analyses can inform clinical decision-making. In Chapter 3, I created bamSliceR, an R/Bioconductor package for rapid cross-cohort variant analysis via BAM slicing. This tool efficiently retrieves and tallies reads spanning any locus across thousands of sequencing files, bypassing traditional variant-calling bottlenecks. bamSliceR enabled several large-scale investigations: for example, it rapidly uncovered rare H3K27M histone mutations in AML (a lesion prevalent in pediatric gliomas but found in ~0.3% of pediatric AML); identified novel in-frame insertions in the ENL (MLLT1) gene's YEATS domain in three AML patients; and re-diagnosed cryptic UBTF-ITD events by tallying soft-clipped reads across 452 AML transcriptomes. The package also facilitated our germline risk study by enabling efficient variant burden analysis, which revealed that ~5.5% of pediatric AML cases carry germline pathogenic variants in known leukemia predisposition genes. In Chapter 4, in parallel with bamSliceR, I developed a transcript-structure-aware approach to pinpoint mutant-bearing transcripts from short-read RNA-seq data. By testing for allele-stratified local exon usage, this method links genetic variants to specific isoforms, providing a mechanistic bridge from genotype to transcript structure- an advance that emulates capabilities of long-read sequencing within existing short-read data. In Chapter 5, I leveraged Oxford Nanopore long-read sequencing to resolve karyotype-scale structural variants that were missed by short-read methods. This is exemplified by decoding the precise junction of a FUS-ERG fusion from low-coverage nanopore data, which enabled the design of junction-specific siRNA therapeutics and rigorous validation of cell-line identity. Overall, this body of work advances our understanding of genetic risk in rare variants and rare subtypes, providing a critical bridge to the emerging long-read era of leukemia genomics