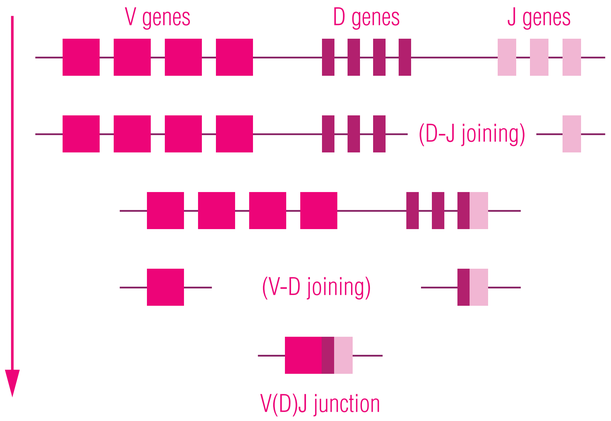
Computational Reconstruction of Human Immunoglobulin Sequences

Devang Thakkar, Sandeep S. Dave
Computational Biology and Bioinformatics
Duke University

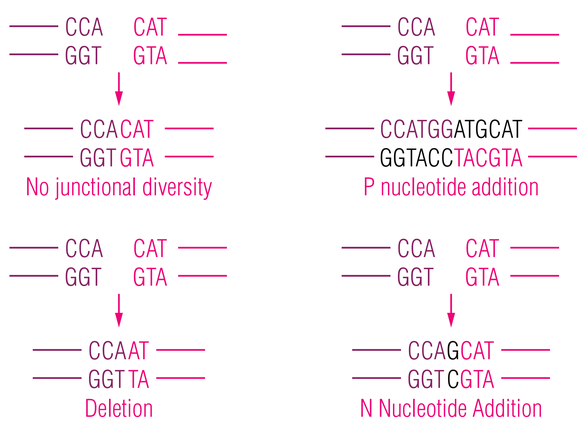
V(D)J Recombination

The antibody repertoire present in humans is generated as a result of two processes that take place in hematopoietic stem cells in the bone marrow - V(D)J recombination and somatic hypermutation. V(D)J recombination at the IGH (heavy), IGK (kappa), IGL (lambda) loci can lead to around 10^{11} combinations.



Junctional Diversity

The joining of the V, D, J regions is accompanied by a variety of junctional modifications. These alterations make it difficult to map confidently to the shorter D and J genes, which are only 15-35bp in length. This, combined with somatic hypermutation, often modifies the D and J genes beyond recognition.



Reconstruction of clonal sequences

Immunoglobulin reconstruction is a complicated problem primarily because of the high similarity in gene sequences. D and J genes are extremely short (15-35bp) in length, with each gene differing only slightly from the other. This, coupled with junctional modifications and somatic hypermutation, makes mapping reads to this region unfeasible. Mapping V genes, which can be up to 350bp in length is slightly more tractable. However, the V genes are extremely similar to each other as well. Consider the 100bp stretch for two V genes shown below.

1> CAGGTGCAGCTGGTGCAGTCTGGGGCTGAGGTGAAGAAGCCTGGGGCCTC
2> CAGGTCCAGCTTGTGCAGTCTGGGGCTGAGGTGAAGAAGCCTGGGGCCTC

1> AGTGAAGGTCTCCTGCAAGGCTTCTGGATACACCTTCACCGGCTACTATA
2> AGTGAAGGTTTCCTGCAAGGCTTCTGGATACACCTTCACTAGCTATGCTA

Burkitt Lymphoma is caused by the translocation of the MYC gene to either of the IGH, IGK, or IGL loci [1]. Under normal conditions, every B cell has a unique immunoglobulin sequence contributes to the antibody repertoire. However, in Burkitt Lymphoma a clonal B cell keeps replicating and thus the population has a immunoglobulin clonotype that is present in higher proportions than otherwise expected.

A normal B cell population

B cell population in Burkitt Lymphoma

Reconstruction Pipeline

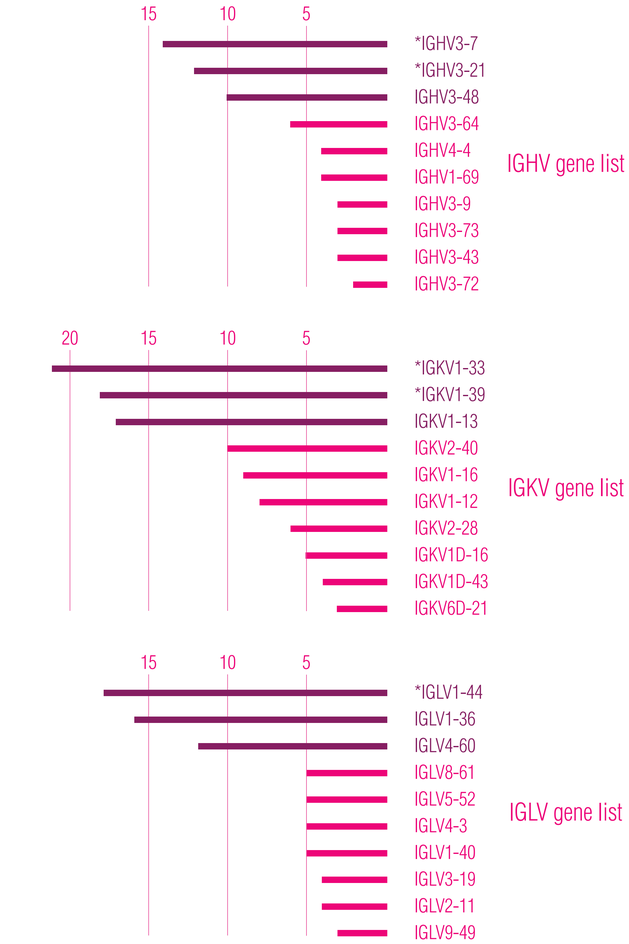
A majority of the methods for immunoglobulin sequence reconstruction use single cell data whereas we perform the same task using bulk sequencing data. IGBLAST, the BLAST equivalent for immunoglobulin mapping fails to work since it attempts to map V(D)J junctions.

Mapping V(D)J junctions is infeasible because of the short lengths of D and J genes, and the junctional diversity incorporated during recombination. In order to circumvent this problem, I created my own pipeline that identifies the most likely V genes present in the tumor clonotype sequence. My pipeline relies on hierarchical mapping of reads to contigs and contigs to genes to overcome the difficulty of mapping to this hypervariable locus.

- Select reads mapping to IG loci (samtools)
- Assemble reads into contigs (ABySS)
- Identify contigs that map to genes (LAST)
- Align reads to above contigs (BWA-MEM)
- Choose reads that map fully to other reads
- Assemble reads into contigs (ABySS)
- List genes with maximum contigs mapped

Top IG(H/K/L) genes

We processed 101 paired tumor-normal Burkitt Lymphoma samples on this pipeline and reported the top V genes for each of the H, K, L loci. Three genes at each loci stand out from the rest of the genes and are likely to be tumor clonotype genes - those marked with an asterisk were also identified by a recent study [2].



References

[1] Norris, Debra, and Jason Stone. "WHO classification of tumours of haematopoietic and lymphoid tissues." Geneva: WHO (2008): 22-23.

[2] Grande, Bruno M., et al. "Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma." Blood 133.12 (2019): 1313-1324.