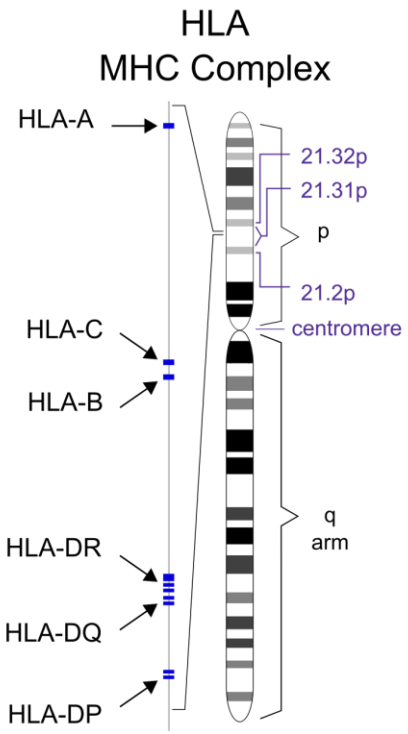
Haplotyping of MHC complex with nanopore reads

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

HLA

The HLA (human leukocyte antigens) complex is located on the short arm of chromosome 6. The HLA genes follow the principles of Mendelian genetics and the encoded antigens are co-dominantly expressed on the cell surface. In the absence of a recombination event, HLA genes are normally inherited *en bloc* from each parent due to their close proximity resulting in their close physical linkage. The HLA haplotype is a combination of linked HLA genes (HLA-A, -B,-C,-DR, -DQ,-DP) transmitted on a single parental chromosome. HLA antigens are expressed on the surface of many cells and play a major role in self-recognition, evoking the immune response to an antigenic stimulus, and to the orchestration of cellular and humoral immunity. HLA complex is known to be highly polygenic as it is composed of many genes, which can divide broadly into three categories: Class I, Class II, and Class III. Polymorphism is another feature of the HLA molecule. Polymorphism allows the presence of multiple variations of antigens or alleles. The HLA class I and class II antigens have the most highly polymorphic structural genes found in humans, which allows amino acids in any given HLA molecule to vary slightly from one person to the next. This variation generates distinct HLA types and also causes allograft rejection when tissues are transplanted.

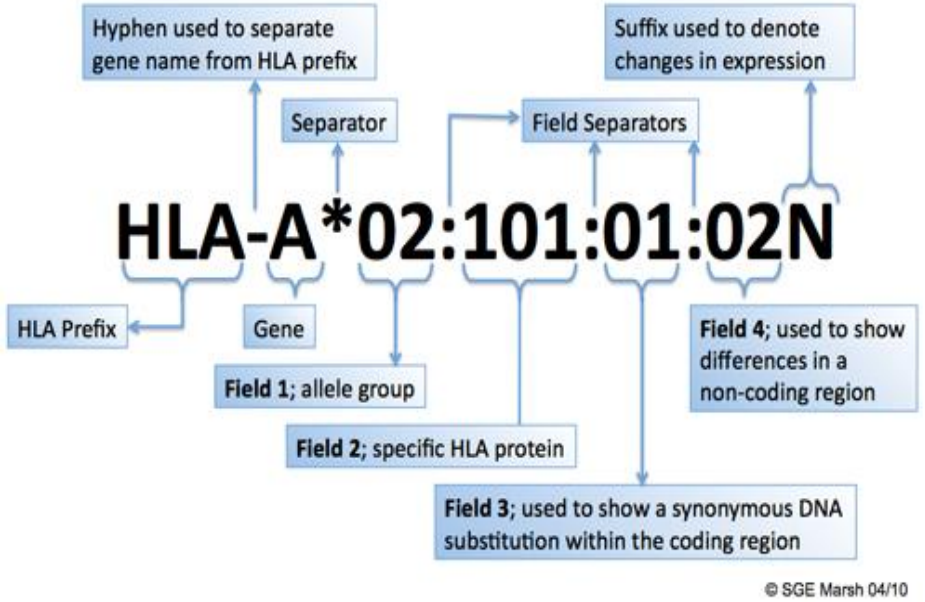


Tumer G, Simpson B, Roberts TK. StatPearls [Internet]. StatPearls Publishing; Treasure Island (FL): Sep 2, 2020. Genetics, Human Major Histocompatibility Complex (MHC) [PubMed]

Anthony noland Database

At Anthony Nolan, the Bioinformatics Team:

- Maintains an internationally recognised database of all known HLA gene variants. This resource is the world's leading authority on cataloguing, naming and publishing data on these genes.



https://www.anthonynolan.org/clinicians-and-researchers/scientists-and-researchers/hla-informatics-group
http://hla.alleles.org/nomenclature/naming.html

Data

NA12878: https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md

AshkenazimTrio: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/

data from IEVA --> enrichment of HLA-A, B, C

Minimap2

Minimap2 is a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database.

minimap2 -a -z 600,200 -x map-ont REFERENCE FASTQ > SAM

This options are optimized for MinION reads (see README_ONT-UL_GIAB_HG002.md)

https://github.com/lh3/minimap2

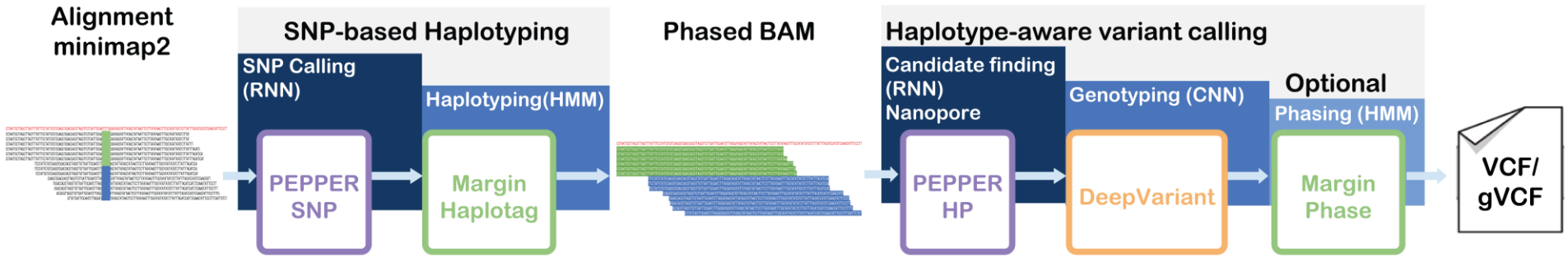
Samtools

Samtools is a set of utilities that manipulate alignments in the SAM (Sequence Alignment/Map), BAM, and CRAM formats. It converts between the formats, does sorting, merging and indexing, and can retrieve reads in any regions swiftly.

http://www.htslib.org/doc/samtools.html

Pepper-margin-deepvariant

PEPPER is a genome inference module based on recurrent neural networks that enables long-read variant calling and nanopore assembly polishing in the PEPPER-Margin-DeepVariant pipeline. This pipeline enables nanopore-based variant calling with DeepVariant.



Oxford Nanopore variant calling pipeline

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre
for Genomic
Regulation

De novo contig from haplotagged bam with canu

Canu is a fork of the Celera Assembler, designed for high-noise single-molecule sequencing (such as the PacBio RS II/Sequel or Oxford Nanopore MinION).

Canu is a hierarchical assembly pipeline which runs in four steps:

- Detect overlaps in high-noise sequences using MHAP
- Generate corrected sequence consensus
- Trim corrected sequences
- Assemble trimmed corrected sequences

https://github.com/marbl/canu

Analysis of the Blast results

Filtlong?