P16 - An end-to-end, fully automated bioinformatics pipeline for genotyping clinically relevant human platelet antigens

Raval R[1], Makwana V[1], Nevin C[1], Zhang R[1], Howell W[1], Browne T[1]
[1]NHS Blood and Transplant, Charcot Road, London NW2 5SX, UK

Introduction

Currently, clinically relevant human platelet antigen (HPA) genotyping (HPA-1 - 6, 9 and 15) in NHS Blood and Transplant (NHSBT) is performed by Sanger sequencing. This requires time-consuming, manual interpretation of the data, potentially causing transcription errors and erroneous genotype assignments. Long read sequencing technologies such as Oxford Nanopore sequencing (ONT) have the potential to revolutionise this service. However, the volume and complexity of the data requires specialist and highly technical bioinformatics expertise to output genotypes highlighting the need for a simple, end-to-end and scalable bioinformatics workflow with built in error handling that any scientist without bioinformatics expertise can run. The NHSBT Service Development Laboratory has developed a comprehensive HPA genotyping system using Oxford Nanopore Technologies Long read sequencing platform which also incorporates an easy to use, end-to-end bioinformatics workflow.

Methods

The pipeline was tested on data from 24 sequenced samples output from an Oxford Nanopore GridION sequencer using the R10.4.1 flowcells and basecalled using MinKNOW v5.9.12 and Dorado v.7.3.11.

Basecalled .fastq files and a sample sheet are saved in a standardised output directory by the user, serving as automated inputs to the pipeline. The sample sheet is a two column, tab separated file containing sample IDs and barcodes used to demultiplex the data during basecalling..

The pipeline is initiated using a single command via the terminal on a system running Ubuntu 20.04. The user is then be prompted to input a standardised run ID and then begin the analysis.

Figure 1 shows the pipeline modules. Each module has been optimised and parallelised to analyse multiple samples concurrently and speed up the computation time. The system used for testing the code utilised an Intel® Xeon® Gold 6238R with 112 threads, 512Gb of RAM and an nVidia A6000 with 48Gb of memory.

Summary statistics and variant calls are extracted from *.bam* and *.vcf* files to output genotyping results and summary statistics.

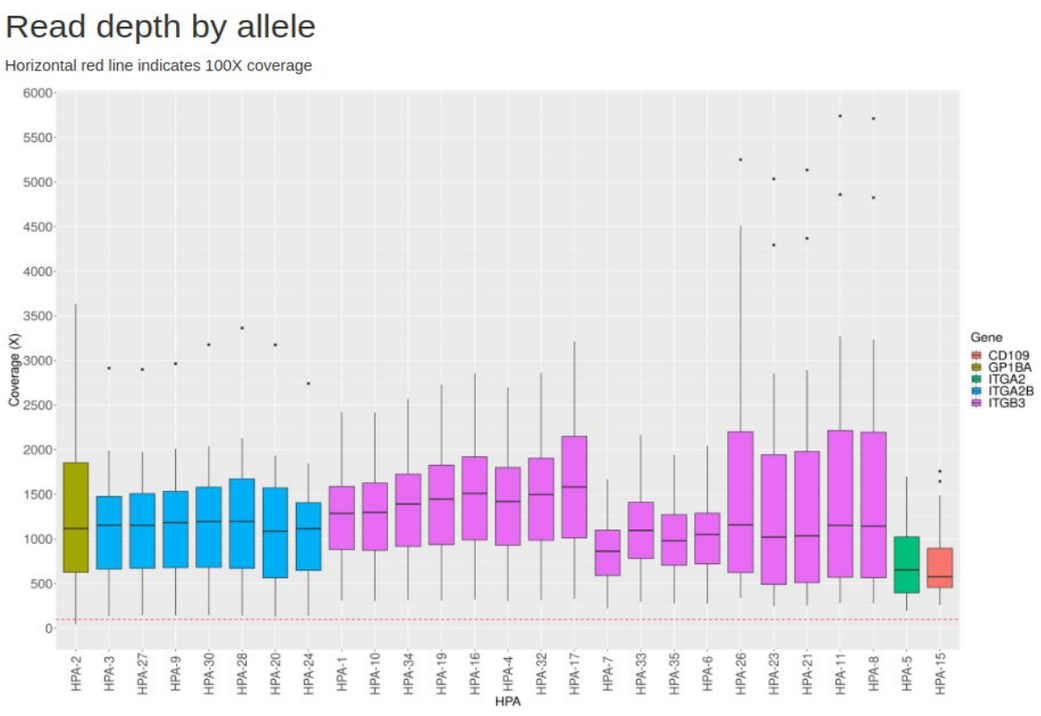

Figure 2. Depth of coverage plot included in interactive QC report.

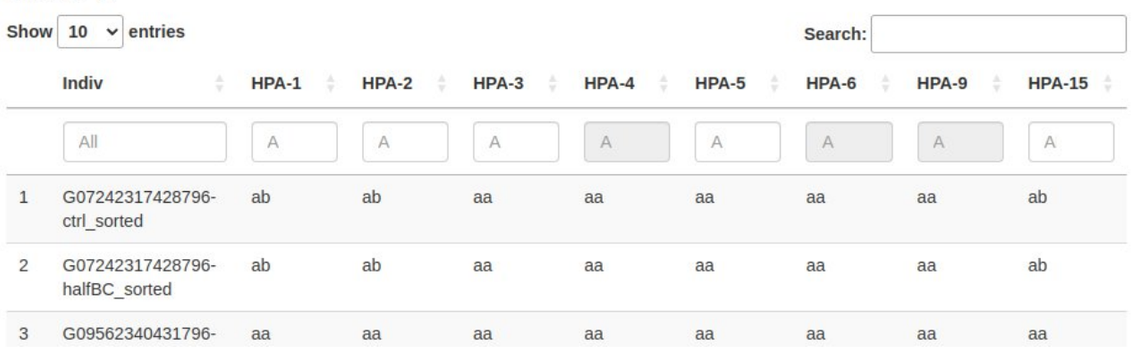

Figure 3. Clinical results are output as an interactive report that can be opened in any web browser.

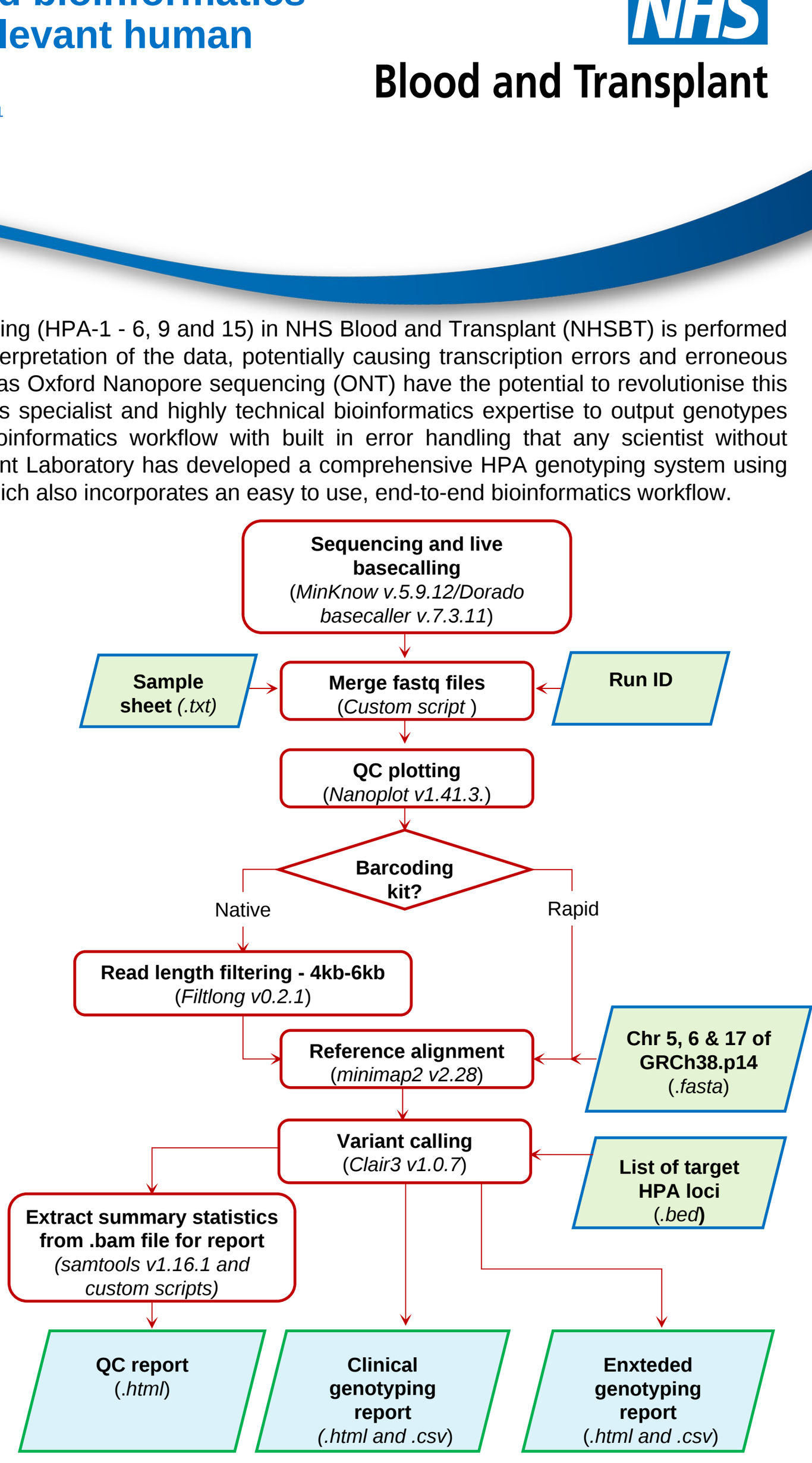

Figure 1. Bioinformatics pipeline for HPA genotyping.

Results

The analysis completed in 6 minutes for 24 samples, and variant calls HPA-1 - 6, 9 and 15 (currently clinically reported) as well as 19 extended genotypes also included on the enriched targets in this assay (Figure 2). All genotyping results are output as a *clear .html* report that can be opened on any web browser regardless of the operating system (Figure 3). These reports are interactive and allow the user to filter results by zygosity, sample ID, and/or HPA loci of interest. Two *.csv* files containing the HPA genotypes are also output so results can be uploaded to a LIMS system for clinical reporting.

A quality report is also output with results on read depth, read length distribution, number of reads aligned and alignment quality, as well as the allele balance of heterozygous and homozygous samples (Figure 2). A full audit trail is then produced showing the users ID, date, time, software versions, and any error logs and messages that need to be reviewed.

All genotypes showed 100% concordance with clinical results previously typed using current clinical methods.

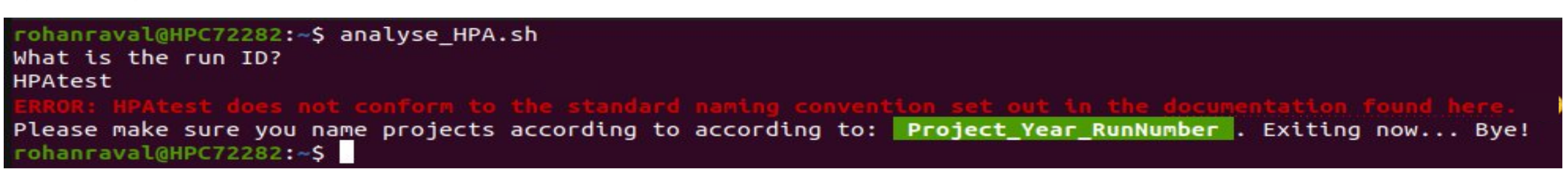

Figure 4. Simple and clear formatting for error handling to prompt troubleshooting. Underlined text contain hyperlinks that can be clicked on to open the documentation in a web browser for more detailed instructions.

Conclusions

This pipeline offers a user friendly approach to calling HPA genotypes derived from Nanopore sequencing for clinical use. By standardising the input and output, the entire pipeline has been automated to produce results that can be uploaded to a LIMS system for reporting results, and a simple QC report to aid Scientists in troubleshooting and clinical decision making. Furthermore, thorough error handling with embedded hyperlinks to documentation in the pipeline provides clear instructions to the user for simple troubleshooting of the pipeline, thus democratising access to bioinformatics techniques (Figure 4). The modularity of the pipeline also allows adaptability for genotyping other clinically relevant systems such as the Human Leukocyte Antigen or the Human Neutrophil Antigen. In future, the analysis will be automated further to remove the need for a Scientist to interact with a command line and conduct the analysis in the background following sequencing, thus allowing the analysis to fit into the heijunka of a high throughput clinical work flow.