ADDRESSING THE BIOINFORMATICS CHALLENGES OF HIGH THROUGHPUT HLA TYPING USING SMRT® DNA SEQUENCING

James Robinson [1, 2], Cristina Guijarro [1], Gayle Leen [1], Jeremy E Stein [1], Neema P Mayor [1,2], Katy Latham [1], J Alejandro Madrigal [1,2], Steven GE Marsh [1,2]

1 Anthony Nolan Research Institute, London, United Kingdom, **2** UCL Cancer Institute, London, United Kingdom

INTRODUCTION

The introduction of a novel approach to HLA typing, incorporating the use of Single Molecule Real Time (SMRT®) Sequencing on the Pacific Biosciences RS II platform raises a number of bioinformatics challenges. This platform enables the sequencing of fully phased long reads that are capable of spanning the full length of HLA class I and the majority of class II genes to produce allele resolution typing. Throughput rate is increased further by barcoding samples, allowing a high multiplex set-up to reduce the financial and environmental cost. The high throughput produces very large data-sets, known typically as “big data”, which creates challenges for conventional bioinformatics analysis and interpretation.

The {AT}toolset software was developed to meet the challenges of data analysis in a high-throughput clinical setting. A review of existing tools and applications available highlighted the need for a custom developed solution. The resulting software incorporates third-party, open source and software developed in-house, to provide a fully automated analysis pipeline. The software suite allows for monitoring and data collection from the Pacific Bioscience’s RS II sequencers, de-multiplexing of barcoded samples, consensus sequence generation, HLA typing assignment and lastly performs quality control checks on the final results.

HARDWARE

The HLA Informatics Group of the Anthony Nolan Research Institute have developed a bioinformatics pipeline and software for the analysis of RS II data that can be run on a singe platform.

- The application can be deployed on a single Ubuntu 14.04 LTS installation.
- The instance is maintained as a Virtual Machine (VM)
- At Anthony Nolan each RS II has a dedicated VM for data.

The final implementation does not require separate IT resources for the different analysis steps, *i.e.* assembly and typing. Anthony Nolan is currently utilising the Intel Xeon processors and HP Proliant DL560 Gen9 servers for analysis of Pacific Biosciences RS II data.

DATA PROCESSES

The following diagram presents an overview of the bioinformatics pipeline:

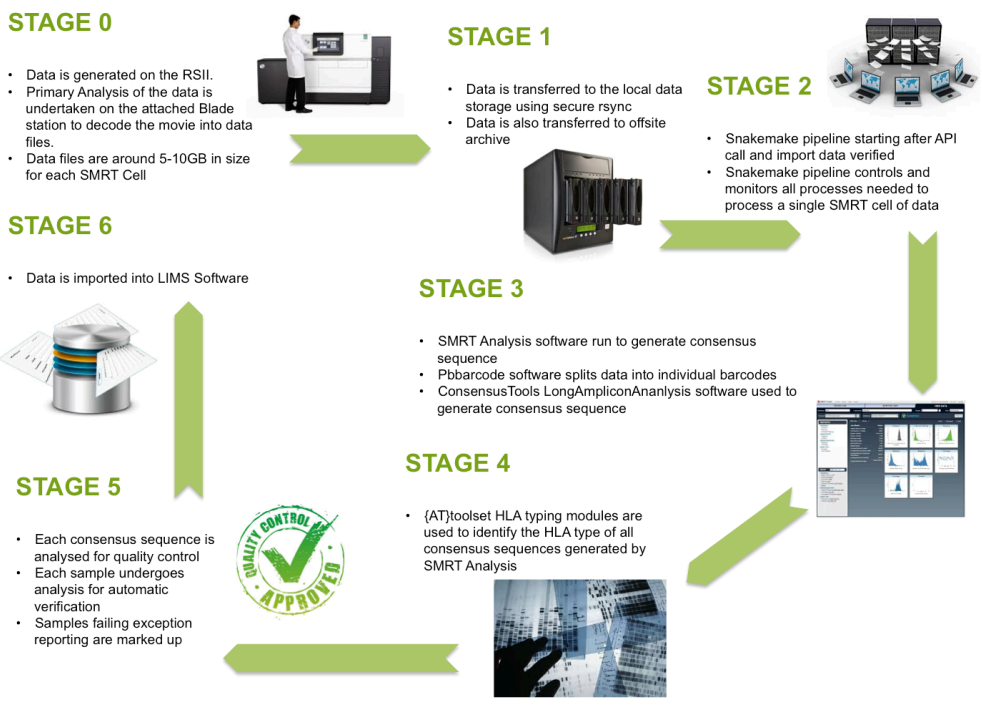

Figure 1: The {AT}toolset software has 7 key stages which are described above.

AUTOMATED BIOINFORMATICS ANALYSIS

The application uses Snakemake, a freely available Python workflow management system. {AT}toolset has been developed to be fully automated. The application will automatically monitor the analysis servers for new RS II data, and coordinate automated analysis. The {AT}toolset application will issue warnings for missing dependencies, and postpone analysis until these requirements are met.

DE-MULTIPLEXING & CONSENSUS SEQUENCES

The application utilises the Pacific Bioscience’s PBBarcode and ConsensusTools (LongAmpliconAnalysis) software to de-multiplex data, and to generate the consensus sequences for each allele. This software has been extensively tested and optimised for both speed and accuracy. The pipeline is able to run each SMRT cell using customised settings to improve the analysis. This allows the same process to be used for both HLA class I and class II typing protocols and also to handle variation in performance between SMRT cells.

HLA TYPING

The HLA typing is performed using software developed in-house. The pipeline uses a custom built HLA typing tool, to assess and match the consensus sequences against the latest version of the IPD-IMGT/HLA Database. The typing module is able to type sequences at the CDS and genomic level, identify off-target hits, as well as accurately describe any variants seen.

QUALITY CONTROL & VALIDATION

The analysis is subject to rigorous quality control testing, that is used to assess the results. These steps assess the quality of each sequence, as well as the results by gene and by sample. The data is also checked at the SMRT cell level. The final output files are designed to interact with either a Laboratory Information Management System or for manual analysis. The {AT}toolset application has been validated against other HLA typing software both freely and commercially available. Validation of the sequences generated from the RS II data has also been undertaken using reference cell lines with known sequence. Finally all automated exception reporting has been validated against analysis performed by registered clinical scientists.

HANDLING HIGH THROUGHPUT DATA

The software is configured to handle the high volume of data from multiple RS II platforms. Extensive work optimizing the analysis pipeline ensures that whilst each cell requires nearly 30 CPU hours to analyse, in the majority of runs the results are returned in less than two hours for each SMRT cell. The pipeline, from loading the RS II machine to the output of results, runs in the background without requiring any manual intervention thus reducing the opportunity for human error.

Analysis Metrics	Weekly	Monthly	Annually
SMRT Cell Runs	4	17	208
SMRT Cells	64	277	3328
Samples	1,536	6,656	79,782
Number of Sequences	17,792	77,099	925,184
Number of Bases	5,742,835,920	24,885,622,319	298,627,467,833
Size of Data (TB)	0.6	2.9	34.5
CPU Jobs Scheduled	9,728	42,154	505,856
CPU Hours	1,898	8,226	98,711

Table1: An estimate of the workload and compute resources handled by the {AT}toolset software

CONCLUSIONS

In summary {AT}toolset addresses the bioinformatics challenges of high-throughput HLA sequencing in a clinical setting and the full pipeline is currently been utilised at Anthony Nolan.

Reg charity no 803716/SC038827 978PA/0416

None of the authors have any conflict of interest to declare