Analysis and Visualisation Tools for NGS Sequence Data

Shay Leary¹, Don Cooper¹, Abha Chopra¹, Mark Watson¹ and Simon Mallal¹,²

¹The Institute for Immunology & Infectious Diseases, Murdoch University, WA, Australia, ²Vanderbilt University Medical Centre, Nashville, TN, USA

Background
- Data Analysis and Visualisation of Second and Third generation DNA sequence data is the next major roadblock to labs wanting to use this technology.
- While many tools already exist they are often difficult to use or require significant computing infrastructure to run or are too generic to generate the data visualisations required for publication.
- Here we describe several tools developed to facilitate NGS analysis and visualisation written in C# that run effectively under a windows environment on a standard windows desktop computer⁴.
- The Visual Genome Analysis Suite VGAS is freely available to collaborators for analysing data generated by IIID Murdoch.
- The tools are useful for both bulk and single cell type analysis within and across datasets.
- The VGAS software can run on as little as 4MB of RAM with no need for advanced graphics cards.
- The tools handle any aligned data from NGS platform in the form of Bam files or Multiple aligned files.

Genome variant analysis

Whole genome sequences can be visualised and the variants called and ranked by groups allowing comparisons within and across datasets. This has been successfully used to compare whole genome variants of H. Pylori to identify variants generated by different environmental stresses such as high salt (see below). An excellent tool for looking at genome variation microorganisms and could be adapted to identify Microbial resistance genes and mobile elements.

HIV Integration site analysis

The whole human genome can be visualised showing sites of HIV integration and the characteristics of the integration, Intronic, Exonic, productive and compare these between groups or longitudinally. While used for HIV integration the same display can be used to plot genomic variation in any organism longitudinally or cross sectional e.g. longitudinal changes during cancer progression.

Analysis of TCR

Both Single cell and bulk TCR analysis can be performed. Over 15 different fully customisable publication-ready analyses. High definition images can be produced to show TCR α/β frequencies and for single cells the α/β pairing frequencies. Samples used will dictate the analysis possible e.g. α/β pairing not possible from bulk samples. Productive and non-productive TCRs can also be identified. For single cells correlations with other phenotype or genotype data can be incorporated.

Conclusions
- VGAS is a desktop Windows PC tool that has no advanced graphic requirements.
- It supports multiple monitors for best results. 4K monitors are recommended but not required.
- Is freely available to collaborative research groups and is downloadable from our cloud resource.
- Highly configurable and produces publication ready images and data.

Acknowledgements: We would like to thank all staff at IIID for their support. http://iiid.murdoch.edu.au/

HLA Allele Caller

A proprietary HLA allele caller has been developed that can handle data from pooled samples (up to 384 samples per run) from any NGS platform. It filters noise and errors and produces HLA allele calls for HLA class I ABC, class II DQA, DQB1, DRB1 3,4,5, DPB1 using the latest IMGT database (1).** HLA analysis and reporting to ASHI accredited standards is handled by “HLA Analyse”. It can handle both Sanger and NGS data from the HLA Allele Caller software to facilitate data integrity and contamination checks and produce reports for both research and diagnostic use. Our current methods allow us to type from 1 sample to 384 in parallel to minimum 4 digit resolution for Class I ABC Class II DQA, DQB1, DRB1,3,4,5, DPB1.

Examples of TCR analysis plots

Integration Site analysis pipeline:

Integration site distribution.

Single Cell T-Cell receptor analysis pipeline.