Visual validation of NGS data features using Tablet

Iain Milne, Gordon Stephen, Micha Bayer, Linda Cardle, Paul D. Shaw, and David Marshall
Information and Computational Sciences Group, The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

tablet@hutton.ac.uk

The James Hutton Institute

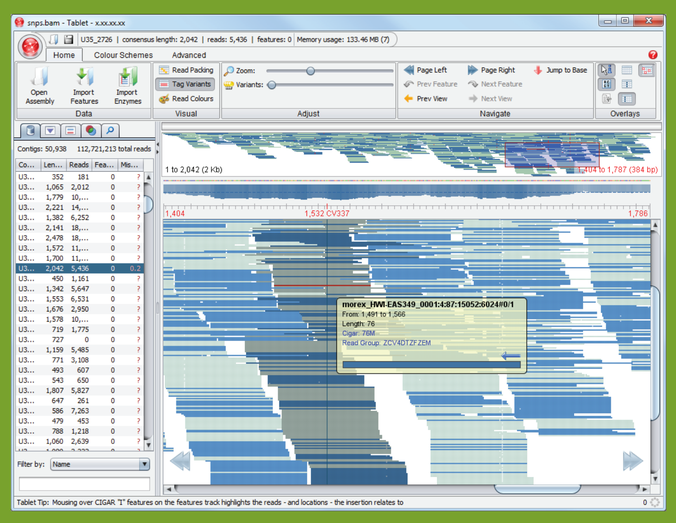
Tablet is written in Java with support for multiple desktop platforms (Windows, OS X and Linux) and is freely available in 32 and 64-bit formats from bioinf.hutton.ac.uk/tablet.

Visualization

The development of second and subsequent generation sequencing technologies introduces limitations on our ability to interrogate the results from analyses and identify patterns that reflect both major quality control issues, and biologically meaningful structures in datasets and complex data analysis outputs.

In our crop and pathogen genetics work at the James Hutton Institute, we find that we are increasingly using visualization to support and aid in the understanding of these datasets.

Tablet has been designed as a high-performance application for the visualization of NGS sequence mapping and assemblies. We have found it to be of particular value in tuning and choosing appropriate parameters for the components of our analysis pipelines.



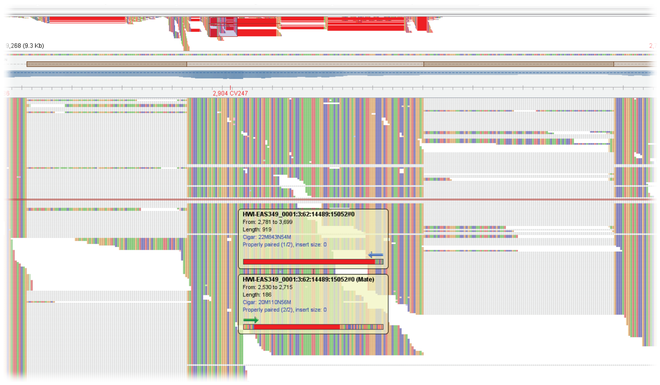
In particular, we use Tablet for the identification of mismapping and misassembly errors which have significant implications for the generation of false positive SNPs or erroneous splice junctions.

Acknowledgements
This work was supported by the Scottish Government (RERAD, Programme 1), the Scottish Funding Council and Scottish Enterprise through the Scottish Bioinformatics Research Network (SBRN) project.

We would also like to thank colleagues within the Cell & Molecular Sciences Group and Biomathematics & Statistics Scotland at The James Hutton Institute for their input to this project.

The Scottish Government · Scottish Enterprise · SBRN

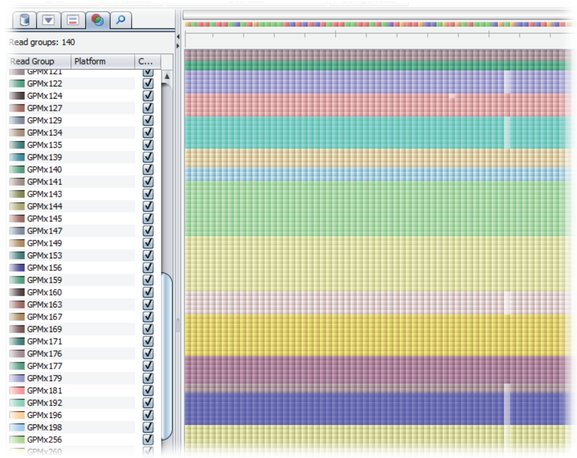
Alternative splicing



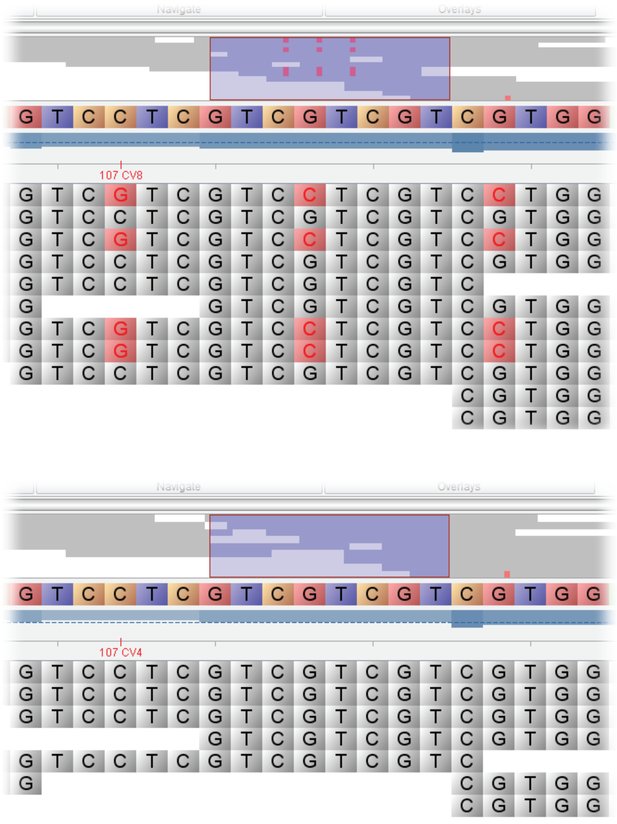
This example demonstrates the use of Tablet to visually confirm the evidence of alternative splice junctions. The image shows a set of barley RNA-Seq reads from a single tissue that have been mapped to a 9.3 Kb barley genomic contig containing two adjacent genes transcribed in opposite directions using the Bowtie/Tophat pipeline. The presence of overlapping CIGAR inserts confirms the alternative splice junctions, whereas simple read mapping in introns may be due to transcription from the opposite strand.

SNP validation

Here, genotyping-by-sequencing tags from barley have been mapped onto a reference sequence and visualized in Tablet using the read group colour scheme. Individual samples are clearly visible as coloured bands, and the variant highlighting functionality allows easy identification of those samples that have the alternate allele at the SNP location. This is helpful for the visual validation of spot samples of SNPs from a larger set.



Detection of mismapped short reads



Mapping parameters for short NGS reads are of crucial importance for the accuracy of downstream analyses such as single nucleotide polymorphism (SNP) discovery. A particularly critical parameter is the mapping mode - the strategy used by the mapping tool for handling reads that could potentially be mapped to more than one location, for example where closely related members of a gene family are involved.

This example features transcripts that have been *de novo* assembled from RNA-Seq reads which were subsequently mapped onto the assembled transcript contigs using the Bowtie mapping tool. In the upper image, all ambiguously mappable reads were mapped to all of their possible locations. This has resulted in cross-mapping of reads that belong to another, very similar transcript, and these are clearly visible in Tablet as a group of reads that feature several correlated variants. In this case, this would result in three false positives during SNP discovery.

The lower image shows the same transcript in a mapping of the same data, but this time applying a Bowtie switch that suppresses cross-mapping (--best --strata) by mapping multi-mappable reads only to a single location with the best fit, that is, the lowest number of mismatches. The reads that were mismapped originally are absent from this mapping.

Conclusions

- Visual validation of subsets of large data sets is an essential tool for next-generation sequencing data analysis.
- Tablet facilitates this task by placing visual emphasis on relevant features through a combination of colour schemes and feature layout.
- It also provides user-friendly management of large data sets through search functions and a multitude of navigation modes.