Artemis and ACT — browsing genomes and visualisation of next generation data

Tim Carver, Giles Velarde, Matthew Berriman, Julian Parkhill and Jacqueline A. McQuillan
Pathogen Genomics Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

ABSTRACT

The advent of next generation data has posed challenges for storing, distributing and visualising the vast quantity of data and information they contain. Artemis has traditionally been used as a genome browser and annotation tool and the Artemis Comparison Tool (ACT) is used to compare sequences and to highlight regions of similarity and differences. New functionality for both these tools is presented, including the ability to view next generation data in the context of the sequence, annotation and variation data. For example a window (BamView) has been incorporated to view read alignments and other types of data can be imported by reading them in as a user plot and displayed as graphs or as a heat map. In this way it provides the annotator with extra levels of information that can inform them about structural annotation. Additionally a new javascript version of the Artemis tool is presented. This has the advantage of delivering genome annotation straight to the community via their web-browser.

VIEWING READ ALIGNMENTS

- Artemis and ACT can visualise data stored in BAM (Binary Alignment/Map) files.
- BAM is a standard format for storing read alignments mapped to a reference genome. Several read alignment tools (e.g. SSAHA2[1] and BWA[2]) support BAM as an output file format.
- BAM files are first sorted and indexed using SAMTools[3]. They are then imported into Artemis or ACT which uses Picard[4] to rapidly access the read alignment information in the region being displayed.
- BAMView is available as a standalone application as well as being integrated into Artemis.

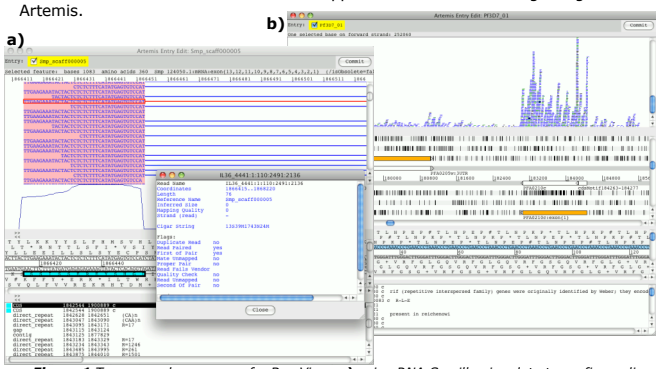


Figure 1 *Two example use cases for BamView,* ***a)*** *using RNA-Seq illumina data to confirm splice sites, the reads are shown split over exon boundaries and joined by a continuous line* ***b)*** *visualising expression of RNA-Seq illumna data obtained from the early ring stage in the life cycle of P. falciparum.*

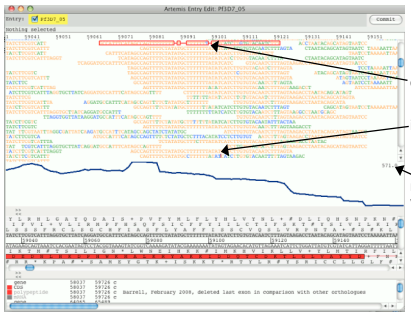


Figure 2. *Read alignments viewed at the nucleotide level along with a coverage plot generated from the alignments. The bases are colour-coded according to their quality scores: blue <10; green <20; orange <30; black >=30.*

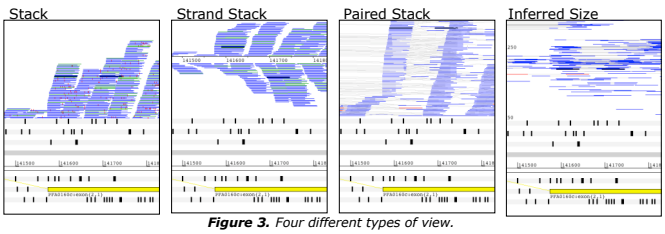


Figure 3. *Four different types of view.*

Features and Functionality Include:

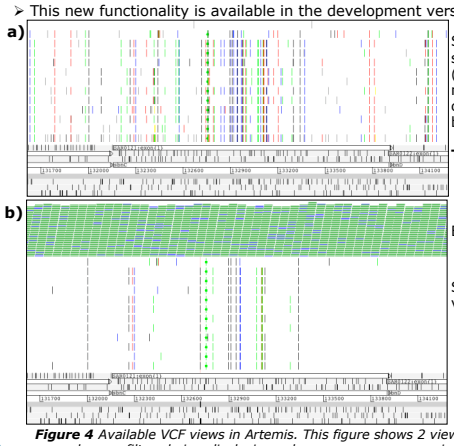
- Zoom to different levels, from broad view of coverage of the genome down to the nucleotide level.
- Different types of alignment views, see Figure 3.
- Highlight pairs or navigate to a reads mate.
- Detailed read information (*e.g.* flags, mapping quality) is available by right clicking over reads.
- Rapidly identify SNPs (see Figure 3a), insertions and deletions (see Figure 2) from the read alignments.
- Can filter reads by their score and/or their SAM flags (*e.g.* show only reads in a proper pair, show reads where the mate is unmapped)
- Auto-generation of coverage and SNP density plots from the BAM data.
- For multiple contigs, read positions are offset to match the position of the corresponding contig.

VARIANT CALL FORMAT (VCF)

- Functionality has been added to Artemis to read and visualise data in VCF[5] files.
- VCF is an emerging format to describe common types of sequence variation (*e.g.* SNPs, indels) relative to a reference genome. VCF files can be generated from BAM files by passing the output of a SAMtools pileup.
- The VCF files need to be compressed and indexed using bgzip and tabix[6] respectively.

```
bgzip file.vcf
tabix -p vcf file.vcf.gz
```

- This new functionality is available in the development version of Artemis.



Showing non-synonymous, synonymous, insertions (yellow), deletions (grey), multiple allele variant calls. The colour represents the base it is being changed to:

T black; G blue; A green; C red

BamView showing read alignment.

Showing only non-synonymous variant calls.

Figure 4 *Available VCF views in Artemis. This figure shows 2 views* ***a)*** *shows all variants and* ***b)*** *shows a filtered view displaying only non-synonymous variants in 12 strains of S.aureus*

WEB-ARTEMIS

- A new light weight version of Artemis for viewing and browsing genomes in a web-browser.
- Delivers sequence and annotation data straight to the community via their web browser. Users can browse and navigate genomes and search for and view their features/annotations of interest
- Web-Artemis is implemented in Javascript. It uses AJAX calls to **CRAWL** (Chado RESTful Access Web-service Layer) web-services to obtain the sequence and annotation data from the Pathogens Chado database which it then displays.
- An alpha version of Web-Artemis is accessible from GeneDB.

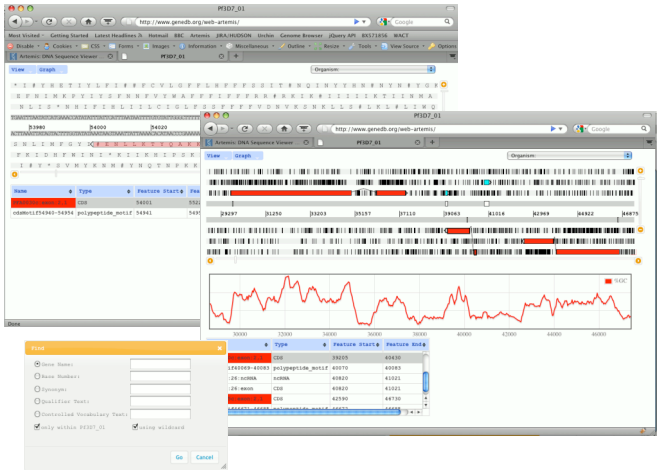


Figure 5 *The P.falciparum pathogen genome viewed in Web-Artemis. This shows a region of the genome and it's associated annotation. Users can zoom right down to the nucleotide level, display a GC plot of the region and display properties for selected features.*

Related Links:

[1]SSAHA2 http://www.sanger.ac.uk/resources/software/ssaha2/
[2]BWA http://bio-bwa.sourceforge.net/
[3]SAMtools http://samtools.sourceforge.net/
[4]Picard: http://picard.sourceforge.net/
[5]VCF: http://vcftools.sourceforge.net/specs.html
[6]tabix: http://samtools.sourceforge.net/tabix.shtml

Availability:

Artemis Home: http://www.sanger.ac.uk/resources/software/artemis/
Development Version: http://www.sanger.ac.uk/resources/software/artemis/#development
BamView: http://bamview.sourceforge.net/
Web-Artemis: http://www.genedb.org/web-artemis/

Contact:

artemis@sanger.ac.uk

Acknowledgements:

The Wellcome Trust funding of the Pathogen Genomics Group at the Wellcome Trust Sanger Institute [grant number WT085775/Z/08/Z].