The aim of projects, such as the UK10K (http://www.uk10k.org), is to construct detailed catalogues of genome variations and identify how genetic differences affect health and disease (Metzker, 2010). This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Artemis, traditionally a genome visualization and annotation tool, has recently been extended and enhanced with a wealth of new features and functionality to support the visualization and analysis of HTS datasets in the context of the genome sequence and annotation.

BamView was originally developed as a stand-alone application to visualize read alignments stored as binary versions of SAM (Sequence Alignment/Map) files and known as BAM files (Li et al., 2009). BamView was subsequently incorporated into Artemis, providing the functionality to view these read alignments in the context of the nucleotide sequence and genomic features. Reads can be displayed in one of five views, aimed at emphasizing different aspects of alignments: stacked, paired-stacked, strand stacked, inferred size and coverage view (Fig. 1). It is possible to zoom in and view the read alignments at the nucleotide level (Fig. 3b). Reads can be filtered by their mapping quality and/or by their SAM flags. Also the panel can be cloned, so that separate views of the same dataset at once to the user. This coupled with the wide range of read alignment and variant filters available allows users to easily customize and tailor their views for the particular experimental dataset they are interrogating. In addition, these new features are also available in the Artemis Comparison Tool, ACT (Carver et al., 2005), which means datasets can be compared across multiple genomes or assemblies. This comparison functionality is not available within the other visualization tools.

2 RESULTS

Artemis, traditionally a genome visualization and annotation tool, has recently been extended and enhanced with a wealth of new features and functionality to support the visualization and analysis of HTS datasets in the context of the genome sequence and annotation. In addition, functionality has been added to display the contents of Variant Call Format (VCF, Danecek et al., 2011) files. VCF files are compressed and indexed with ‘bgzip’ and ‘tabix’ (Li, 2011), respectively. Alternatively binary VCF (BCF) can be used, which are generated and indexed, using BCFtools (http://samtools.sourceforge.net/mpileup.shtml). Variants stored in these files are rendered in the VCF view allowing the user to visually inspect the variant calls against the reference sequence and annotation (Fig. 2). Artemis can access the BAM and VCF and the associated index files from a local file system or remotely over HTTP and FTP.

Furthermore, filtering mechanisms have been added to Artemis to allow users to filter variants in real time. The filter can be based on the type of variant (e.g. synonymous, non-synonymous, insertions) or on the properties present in the VCF file such as read depth and mapping quality. These variants can also be automatically annotated according to variant type, including synonymous SNPs, non-synonymous SNPs, insertion, deletion and multiple allele.
Fig. 2. A region of \textit{S.pneumoniae} antibiotic resistant pandemic clone PMEN1. This shows the variation data for 11 strains. The \textit{pspA} gene in the reference is a pseudogene as a result of a frame shift. The top VCF panel shows variant and non-variant sites without filtering. The VCF panel below this shows the same region but with filtering applied (i.e. minimum quality score of 60, a minimum read depth of 5). The magenta insertions in most strains show where this frame shift has been caused by a deletion in the reference strain (and one of the other strains). The gene has independently become a pseudogene in a second strain, because of an SNP leading to a premature stop codon (as indicated by the circle in the middle of the line), and a third strain has a recombination that spans this whole region (shown by the increased SNP density).

Artemis can be used for the initial data exploration of HTS datasets, cross checking of results and for small to medium size genomes it provides a limited set of analyses. This includes calculation of SNP density, read counts and the reads per kilobase per million mapped reads (RPKM, Mortazavi et al., 2008) for selected genes. The read count results can be saved in a tab delimited format file. These results can be used as input to DESeq (Anders and Huber, 2010), for example, for additional differential expression analysis.

Finally, to display other experimental datasets (e.g. microarray, RNA-Seq) it now has support for user-defined input types, which can be rendered graphically as either graphs or heat maps.

In the following sections, we describe these new features in more detail and outline how Artemis can be used in several different HTS experiments as a platform for interrogating, analysing, and interpreting these data.

2.1 Base pair variation

The high yield and low error rate of HTS technologies allows genome re-sequencing to a high depth of coverage. There are many tools available for mapping of short reads against a reference genome, most of which provide the option to output SAM or BAM format, which can now be opened in Artemis. In the simplest scenario, HTS genome re-sequencing allows correction of a reference sequence produced with older technologies, such as capillary sequencing. Figure 3 shows re-sequencing data of a bacterial genome. Red lines show the location of bases in mapped reads that differ from the sequence of the reference genome. Sequencing errors produce randomly distributed lines, while true SNPs between the sequenced DNA and the reference genome can be seen where lines align vertically through all of the reads mapping across a site. In this instance, the reference can be corrected/updated. There are tools to do this type of correction automatically e.g. iCORN (Otto et al., 2010a). However, Artemis can be used to visualize and confirm this error by inspecting the read and mapping qualities and BAM flags of the alignments.

For the analysis of genomic variation, sequence data from a set of isolates can be mapped against a reference genome to identify variations that are biologically significant. Genetic variation can be displayed in the VCF view of Artemis. Variant files are displayed in rows above the reference genomic annotation, with...
Table 1. Colour schemes for the variants in the VCF panel

1. Default colour scheme
   - Variant A: Green
   - Variant G: Blue
   - Variant T: Black
   - Variant C: Red
   - Multiple allele: Orange, with circle at the top
   - Insertion: Magenta
   - Deletion: Grey
   - Non-variant: Light Grey

2. Synonymous / non-synonymous
   - Synonymous SNP: Red
   - Non-synonymous SNP: Blue

3. Quality score
   - Variants are all on a red colour scale with those with a larger score being more intense.

lines representing the position of variations from the reference DNA sequence. By default, SNP variants are colour coded according to the variant base (Table 1), while indels are also highlighted. However, there are alternative colour schemes. Artemis also allows non-variant positions to be displayed, so that a lack of variation can be distinguished from a lack of data or coverage. Figure 2 shows how the Artemis can be used to visualize variation in 11 bacterial isolates mapped against the reference genome (Spm23F).

2.2 Population analysis

When very large numbers of isolates are sequenced, with the intention of identifying genetic variants at the population level, data from all of the samples or isolates can be loaded into Artemis and several types of analyses performed across the entire set. The VCF view supports real-time filtering of both variant and non-variant bases, enabling the user to adjust the parameters upon which base calls are made. For example, a user can choose to filter variants based on read depth or sequencing quality. Furthermore, by combining the BamView and VCF displays, it is possible to manually crosscheck base calls against the raw mapped reads. Filtered data can be exported in VCF format, or the reconstructed DNA sequences of variants can be exported in FASTA format for genes, regions selected in the Artemis annotation track or the entire reference. This output can then be used as input to multiple sequence alignment and phylogenetic tree construction tools. Using their set of filtered variants, users can also conduct further analyses within Artemis such as ranking genes based on SNP density.

2.3 Structural variation

The ability to visualize alignment of sequencing reads mapped against a reference also allows identification of structural variation. Deletions relative to the reference genome can show no mapping in the deleted region, but lack of mapping may also indicate lack of data. Where paired-read information is available, the insert sizes of read-pairs mapping across a deletion will be greater than expected. Figure 4a shows how a deletion in a subset of data from the Mouse Genomes Project (Keane et al., 2011) sequenced with Illumina technology can be clearly identified by visualizing read pairs in Artemis. One strain (green) shows continuous mapping across the entire region, which is also confirmed by the coverage plot. Coverage for the other strain drops to zero for a region of ~370 bases. For the same strain, paired-reads flanking the unmapped region exhibit an increased insertion length relative to the reference, indicated by their position on the y-axis. Together these two observations are strong evidence that this is a true deletion. Conversely, insertion locations relative to the reference are manifested by the presence of mapped reads with missing mate-pairs (i.e. reads sequenced from the opposite end of the same sequencing template). Upstream of the insertion only forward reads will be mapped, while downstream of
Artemis is implemented in Java and packaged for ease of installation and to explore conservation of synteny (Holden and others, 2008). Artemis allows RNAseq data to be viewed in the context of the genome annotation, providing useful guidance for correcting gene boundaries, splice-sites, or for identifying the location of non-coding RNA sequences. RNAseq data also provide valuable information about relative gene expression levels. Figure 5 shows RNAseq data from seven time points during the lifecycle of Plasmodium falciparum can be visualized simultaneously using the user graph heatmap view (Otto et al., 2010b). Artemis can also report RPKM values or the number of reads mapping against each gene.

3 IMPLEMENTATION AND PERFORMANCE

Artemis is implemented in Java and packaged for ease of installation on all major platforms (Windows, MacOSX and UNIX). It runs as a stand-alone application and requires a Java Runtime Environment (JRE) of at least 1.6. It uses the Picard library (http://picard.sourceforge.net) for accessing and manipulating BAM files. The latest version of Artemis available for download includes all the libraries needed to run the application.

ACT is an application for displaying pairwise comparisons of DNA sequences. It can be used to identify and analyse regions of similarity and difference between genomes and to explore conservation of synteny (Holden et al., 2009, Pain et al., 2008, Parkhill et al., 2001, Peacock et al., 2007). Since ACT is built from the components of Artemis it inherits all its features and functionality and so also supports the HTS formats.

With the ever-increasing size of the datasets produced by HTS experiments, effective management of resources is critical for visualizing and analysing such datasets. The original Artemis used a memory-based approach, where all the sequence and annotation data was loaded into memory. Although this type of approach has the advantage of being faster for visualization and navigation, it can be prohibitive for viewing large datasets as it is limited by the amount of memory available. To overcome this, a key focus of the new Artemis has been on developing it to handle larger datasets. To achieve this, Artemis now supports indexed file formats. In addition to support for indexed FASTA files, Artemis displays data contained in BAM, BCF and compressed and indexed VCF files. This has the advantage that the data can reside on disk and only the current visible region needs to be held in memory, thus reducing the amount of memory required and allowing Artemis to efficiently display large vertebrate genomes (Fig. 4).

Artemis has been optimized to minimize its memory requirements. However, with the implementation of the indexed FASTA reader in Artemis, the memory requirement is much less (Table 2). It is also faster at opening up as it does not read the entire sequence. There is still an increase in memory with an increase in sequence size because of a cache used to record stop codon positions. The caching mechanism has been further optimized to reduce the memory needed.

4 DISCUSSION

In this article, we have described Artemis, an integrated genome visualization and analysis platform for HTS datasets. We have presented several example datasets and discussed the role of Artemis in their visualization and interpretation. Further examples are available on the Artemis website (http://www.sanger.ac.uk/resources/software/artemis/ngs/).

Artemis has support for various standard file formats including BAM and VCF, in addition to user-defined input files. These formats can support data from many different types of sequencing platforms. Therefore, Artemis has the potential to visualize and investigate sequencing data from numerous sequencing technologies.

The new views available within Artemis are easily customizable by the user and can be viewed at different levels of resolution. Multiple BAMs and VCFs can be loaded simultaneously for comparison. This allows in-depth preliminary analysis and investigations of HTS datasets in the context of the sequence and annotation of a reference to reveal biologically relevant information.

Using indexed files (BAM, VCF, BCF and FASTA) means that Artemis can efficiently manage memory by only needing to load into memory the data that is currently being displayed in the interface. This improves the performance and enables large genomes to be viewed and interpreted along with large HTS datasets.

| Sequence size (Mb) | FASTA (Mb) | Indexed FASTA (Mb) | Indexed FASTA and optimized cache (Mb) |
|-------------------|-----------|-------------------|--------------------------------------|
| 100               | ~ 155     | ~ 75              | ~ 50                                 |
| 1000              | ~ 1200    | ~ 525             | ~ 280                                |
| 2000              | ~ 2300    | ~ 1024            | ~ 525                                |

Fig. 5. A region in Chromosome 2 of P. falciparum 3D7. Transcriptome data (in wiggle format) is plotted to show expression levels over time. Data are plotted for seven time intervals: 0, 6, 12, 18, 24, 30, 36 h, from the top of the plot to bottom.
In summary, the latest version of Artemis is a powerful tool that can be used in a variety of different sequencing-based experiments including genome re-sequencing, population scale variation detection and transcriptome sequencing.

ACKNOWLEDGEMENTS

We thank Kim Wong at the WTSI for providing feedback on the Mouse Genomes dataset.

Funding: Wellcome Trust through their funding of the Pathogen Genomics group at the Wellcome Trust Sanger Institute (grant number WT 076964).

Conflict of Interest: none declared.

REFERENCES

Aliothke,D. et al. (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073.
Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106.
Bergeron,M. et al. (2009) The genome of the blood fluke Schistosoma mansoni. Nature, 460, 352–358.
Burova,VM. et al. (2010) Comprehensive annotation of the transcriptome of the fungal pathogen Candida albicans using RNA-seq. Genome Res., 20, 1451–1458.
Carver,T.J. et al. (2005) ACT: the Artemis comparison tool. Bioinformatics, 21, 3422–3423.
Carver,T. et al. (2010) BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics, 26, 678–677.
Croucher,N.J. and Thomson,N.R. (2010) Studying bacterial transcriptomes using RNA-seq. Curr. Opin. Microbiol., 13, 419–424.
Croucher,N.J. et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. Science, 331, 430–434.
Daniel,C. et al. (2011) The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res., 21, 315–324.
Danzer,P. et al. (2011) The variant call format and VCF tools. Bioinformatics, 27, 2156–2158.
DePristo,M.A. et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet., 43, 491–498.
Flanc,M. et al. (2010) Variant: genome browser for high-throughput sequencing data. Bioinformatics, 26, 1938–1944.
Harris,S.R. et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science, 327, 469–474.
Holden,M.T.G. et al. (2009) The Genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol., 191, 261–277.
Hou,H. et al. (2010) MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. Nucleic Acids Res., 38, W732–W736.
Huang,W. and Mark,G.T. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res., 18, 1538–1543.
Kratz,T.M. et al. (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477, 289–294.
Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.
Li,H. (2011) Tabix: fast removal of sequence features from generic TAB-delimited files. Bioinformatics, 27, 718–719.
Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.
Li,H. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078–2079.
Lu,T. et al. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res., 20, 1238–1249.
Manske,H.M. and Kwaik,J.K. (2009) LookSeq a browser-based viewer for deep sequencing data. Genome Res., 19, 2125–2132.
Metzker,M.L. (2010) Sequencing technologies - the next generation. Nat. Rev. Genet., 11, 31–46.
Milne,J. et al. (2010) Tablet—next generation sequence assembly visualization. Bioinformatics, 26, 401–402.
Mortazavi,A. et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628.
Nicol,J.W. et al. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics, 25, 2730–2731.
Ott,T.D. et al. (2010a) Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics, 26, 1704–1707.
Ott,T.D. et al. (2010b) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Mol. Microbiol., 76, 12–24.
Pam,A. et al. (2008) The genome of the simian and human malaria parasites Plasmodium knowlesi. Nature, 455, 799–803.
Pasick,J. et al. (2011) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature, 413, 848–852.
Peacock,C.S. et al. (2007) Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat. Genet., 39, 839–847.
Robinson,J.T. et al. (2011) Integrative genomic viewers. Nat. Biotechnol., 29, 24–26.
Rutherford,K. et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics, 16, 944–945.
Shendure,J. and J.H. (2008) Next-generation DNA sequencing. Nat. Biotechnol., 26, 1153–1155.
Simpson,J.T. et al. (2009) Abyss: a parallel assembler for short read data. Genome Res., 19, 1117–1123.
Schmit,J.C. et al. (2007) HawkEye: an interactive visual analytics tool for genome assemblies. Genome Biol., 8, R34.
TrappeL.C. et al. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.
Walters,R.G. et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature, 463, 671–678.
Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821–829.