OXFORD

Genome analysis

gEVAL — a web-based browser for evaluating genome assemblies

William Chow[1], Kim Brugger[2], Mario Caccamo[3], Ian Sealy[1], James Torrance[1] and Kerstin Howe[1,*]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK, [2]East Anglian Medical Genetics Centre, Cambridge University Hospitals, NHS Foundation Trust, Cambridge CB2 0QQ, UK and [3]National Institute of Agricultural Botany, Cambridge CB3 0LE, UK

*To whom correspondence should be addressed.

Abstract

Motivation: For most research approaches, genome analyses are dependent on the existence of a high quality genome reference assembly. However, the local accuracy of an assembly remains difficult to assess and improve. The gEVAL browser allows the user to interrogate an assembly in any region of the genome by comparing it to different datasets and evaluating the concordance. These analyses include: a wide variety of sequence alignments, comparative analyses of multiple genome assemblies, and consistency with optical and other physical maps. gEVAL highlights allelic variations, regions of low complexity, abnormal coverage, and potential sequence and assembly errors, and offers strategies for improvement. Although gEVAL focuses primarily on sequence integrity, it can also display arbitrary annotation including from Ensembl or TrackHub sources. We provide gEVAL web sites for many human, mouse, zebrafish and chicken assemblies to support the Genome Reference Consortium, and gEVAL is also downloadable to enable its use for any organism and assembly.

Availability and Implementation: Web Browser: http://geval.sanger.ac.uk, Plugin: http://wchow.github.io/wtsi-geval-plugin.

Contact: kj2@sanger.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Reference genomes are the foundation for genomic biology. As more and more *de novo* sequencing projects are being conducted, and more draft genomes released, the continued challenge is to create sufficiently complete and correct assemblies that can be confidently used as references by the research community. Although it has been over a decade since it was announced that the Human genome was completed, this statement referred to the then achievable quality of sequence resolution. There remained errors and lack of sequence coverage for important regions, for example many areas encoding multi-gene families (IHGSC, 2004, Horton *et al.*, 2008). Since then, genome assemblies of varying quality have been produced for many more species and individuals, often resulting in annotation errors

with negative influence on their research use (Denton *et al.*, 2014). Groups such as the Genome Reference Consortium (GRC, genomereference.org) lead efforts to curate and resolve genome assembly issues, and maintain the best quality genome references possible for key research species (Church *et al.*, 2011).

Depending on the project, groups that tackle creating a genome reference may have at their disposal multiple alternative assemblies together with resources such as clone libraries, collections of short reads, cDNA sequences, RNAseq data, physical maps, optical maps or genetic markers. These datasets help in repairing and reorganizing genomic regions or to create a strategy to infuse new sequence into a draft assembly. In parallel, research biologists who observe discrepancies between their data and the genome reference would

Fig. 1. Region on GRCh38 Chromosome 11 with variation and missing sequence. **(A)** Purple clone end pair mappings indicate same end repeated, while red mappings indicate incorrect orientation of paired ends. **(B)** Two clone components are used to build this region of the assembly. The green box indicates a reliable overlap region (red would indicate high variation). **(C)** Orange indicates an incomplete transcript mapping. **(D)** Six Single molecule genome maps (orange/red) can be compared to *in silico* digest (purple). Red regions indicate discordance. In this case, a ~7.5 kb block variation is shared between three maps and the reference, whilst three other maps share two fragments. Furthermore, in the ~39 kb digest block, all six maps indicate a size of ~45–47 kb, giving evidence of missing sequence (~7–8 kb). **(E)** Comparative analysis between HuRef and YH2 assemblies, reveal this missing sequence (dotted box) as well as the region of variation (Supplementary Figure S1)

benefit from being able to assess evidence supporting the reference in the region.

Current public genome browsers include many of these types of datasets in their analysis repertoire, but usually focus on gene annotation (Cunningham *et al.*, 2014, Rosenbloom *et al.*, 2014). They also represent major assembly builds released every few years and neither reflect uncertainties in the sequence itself, nor more recent improvements. Here we introduce the gEVAL Browser project, a collection of software and frequently updated databases for key species that takes a tiling path as the backbone, conducts analyses using new sources of data and regularly releases the results in a web interface for users to evaluate sequence integrity and create strategies for sequence management.

2 Overview

The web based gEVAL browser and the underlying back end databases use the Ensembl project as a framework to build on (Birney *et al.*, 2004). Supporting the GRC, it features the current, previous and in-progress official reference genomes of human, mouse chicken and zebrafish while also hosting assemblies of these species not found in other public genome browsers including those curated by the Genome in a Bottle Consortium (Zook *et al.*, 2015) (Supplementary Table S1), and 16 lab and wild derived mouse strains from the international Mouse Genomes Project (Adams *et al.*, 2015). Other species present but less frequently updated include pig (*Sus scrofa*), rat (*Rattus norvegicus*) and three helminth species (*Echinococcus multilocularis*, *Strongyloides ratti*, *Schistosoma mansoni*) (Supplementary Table S1).

There are chromosomal, regional overview, detailed region and comparative browser views, in addition to lists of anomalies. Individual analysis tracks displaying a wide variety of glyphs can be selected interactively. Common tracks include clone library end sequences (Fig. 1A), cDNA alignments (Fig. 1C), genetic markers and self-alignments within the genome. Additional tracks available for some assemblies include issues in regions under review by the GRC, exact clone or sequence component placements and quality of

overlaps between neighbouring assembly components (Fig. 1B). The glyphs are coloured to indicate discrepancies and offer pop-up menus that reveal further information and navigational options. All tracks incorporate navigational links to access adjacent s regions, allowing the user to 'walk' along the genome from one feature/issue to the next. Custom external user-sourced tracks can also be added by attaching datasets stored in trackhubs or genome browser supported file formats (GFF, BED).

Comparative genome analysis is invaluable for revealing conserved regions between organisms. Uniquely in gEVAL, comparative analysis focuses on genomic alignments among the different assemblies available for the same species (Fig. 1E). This is useful in capturing sequence differences caused by both variation as well as misassembly, and aids the improvement of one assembly with components/guidance from another (Supplementary Figure S1).

Unique to gEVAL, it integrates general annotation with long range information produced from recent technologies such as single molecule genome/optical maps (Howe and Wood, 2015; Teague *et al.*, 2010; Supplementary Table S2). These maps prove invaluable for scaffolding assemblies but are also useful for capturing genome-wide structural variation (Mak *et al.*, 2015) (Fig. 1D). In addition to presenting evidence, gEVAL can suggest specific sequences from an aligned resource to fill reference gaps, or can suggest inversions or other rearrangements such as expansions or contractions of tandem repeats.

3 Conclusion

gEVAL is a web-based browser that allows easy detailed evaluation of genome assemblies through its tools and pre-computed analyses, and suggestion of fixes. It is a key tool used by the GRC for the improvement of the primary vertebrate reference genomes. It is widely used externally for evaluating detailed regions of these genomes in different assembly versions, and also by other reference assembly projects, e.g. the respective international genome consortiums of pig (Groenen *et al.*, 2012; Warr *et al.*, 2015), chicken (Schmid *et al.*, 2015) and mouse (Adams *et al.*, 2015). To enable this use the

gEVAL code can be downloaded (see Availability and Implementation Section) and installed alongside an Ensembl installation for any other genome assembly.

References

Adams,D.J. *et al*. (2015) The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mamm. Genome*, **26**, 403–412.

Birney,E. *et al*. (2004) An overview of Ensembl. *Genome Res*., **14**, 925–928.

Church,D.M. *et al*. (2011) Modernizing reference genome assemblies. *Plos Biol*., **9**, e1001091.

Cunningham,F. *et al*. (2014) Ensembl 2015. *Nucleic Acids Res*., **43**, D662–D669.

Denton,J.F. *et al*. (2014) Extensive error in the number of genes inferred from draft genome assemblies. *Plos Comput. Biol*., **10**, e1003998.

Groenen,M.A. *et al*. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, **491**, 393–398.

Horton,R. *et al*. (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics*, **60**, 1–18.

Howe,K., and Wood,J.M. (2015) Using optical mapping data for the improvement of vertebrate genome assemblies. *Gigascience*, **4**, 10.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Mak,A.C. *et al*. (2015) Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, **26**, 403–412.

Rosenbloom,K.R. *et al*. (2014) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*., **43**, D670–D681.

Schmid,M. *et al*. (2015) Third report on chicken genes and chromosomes 2015. *Cytogenet. Genome Res*., **145**, 78–179.

Teague,B. *et al*. (2010) High-resolution human genome structure by single molecule analysis. *Proc. Natl. Acad. Sci. USA*, **107**, 10848–10853.

Warr,A. *et al*. (2015) Identification of low-confidence regions in the pig reference genome (Sscrofa10.2). *Front. Genet*, **6**, 338.

Zook,J.M. *et al*. (2015) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *bioRxiv*, Preprint, submitted 23 December 2015. doi:10.1101/026468.