*D1018–D1024* *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 26 November 2008*
*doi:10.1093/nar/gkn939*

# The HuRef Browser: a web resource for individual human genomics

**Nelson Axelrod*, Yuan Lin, Pauline C. Ng, Timothy B. Stockwell, Jonathan Crabtree, Jiaqi Huang, Ewen Kirkness, Robert L. Strausberg, Marvin E. Frazier, J. Craig Venter, Saul Kravitz and Samuel Levy***

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received October 27, 2008; Revised November 4, 2008; Accepted November 5, 2008

## ABSTRACT

**The HuRef Genome Browser is a web application for the navigation and analysis of the previously published genome of a human individual, termed HuRef. The browser provides a comparative view between the NCBI human reference sequence and the HuRef assembly, and it enables the navigation of the HuRef genome in the context of HuRef, NCBI and Ensembl annotations. Single nucleotide polymorphisms, indels, inversions, structural and copy-number variations are shown in the context of existing functional annotations on either genome in the comparative view. Demonstrated here are some potential uses of the browser to enable a better understanding of individual human genetic variation. The browser provides full access to the underlying reads with sequence and quality information, the genome assembly and the evidence supporting the identification of DNA polymorphisms. The HuRef Browser is a unique and versatile tool for browsing genome assemblies and studying individual human sequence variation in a diploid context. The browser is available online at http://huref.jcvi.org.**

## INTRODUCTION

The first two drafts of human genome sequences were published by the Human Genome Sequencing consortium (1) and Celera Genomics (2) in 2001. These two genome assemblies were created as consensus sequences from a population of multiple individuals and DNA variations were identified in the form of single nucleotide polymorphisms (SNPs) (3). We sequenced the genome of a single individual and employed long range haplotype assembly to reconstruct a diploid human genome sequence, referred to as HuRef (4). This genome sequence assembly is unique in that it is constructed from the DNA of a human individual and the diploid phasing of the assembly has been determined in greater than 80% of the genome. The analysis of the diploid human genome produced a rich and complex picture of human sequence variation. We reported variations in the form of SNPs, block substitutions, indels, variable number tandem repeats (VNTRs), inversions and other complex variants. The detail and complexity of the datasets generated by genome sequencing and assembly projects necessitate visualization tools to provide composite displays that synthesize all of the underlying data.

There are two types of genome browsers most relevant to our project: assembly browsers and annotation browsers. There are a number of genome assembly browsers that are publicly available, including the NCBI Assembly Viewer, Hawkeye (5), TAMPA (6) and EagleView (7). These browsers are designed largely to enable users to detect, interrogate and correct errors of assembly, and are primarily used to refine the assembly process.

There are a number of genome annotation browsers available for the visualization and analysis of genome sequence annotations, including the UCSC Genome Browser (8), Ensembl Genome Browser (9), NCBI Sequence Viewer and GBrowse (10). These browsers provide many useful features for viewing genome sequences, annotations and variants. However, due to the highly coupled nature of the genome assembly and annotation processes, errors in assemblies are often propagated as errors in annotations. The accuracy of genome annotations, especially with regard to complex sequence variation, is nearly impossible to examine using annotation browsers. There are reported cases of incorrectly annotated genes or mischaracterized sequence variations that are due to errors in assembly (11).

During the analysis of the HuRef genome, we found that existing genome browsers were not designed to study complex, large-scale sequence variation and its

*To whom correspondence should be addressed. Tel: +1 301 795 7382; Fax: +1 301 294 3142; Email: naxelrod@jcvi.org
Correspondence may also be addressed to Samuel Levy. Tel: +1 301 795 7382; Fax: +1 301 294 3142; Email: slevy@jcvi.org

relation to the assembly and annotation of an individual human genome. To this end, we developed the HuRef genome browser. It is a unified, high-performance framework that integrates the functionality of annotation and assembly browsers, and incorporates assembly-to-assembly genome comparison views, haplotype views and multiple sequence alignment capabilities.

The HuRef genome browser is a unique visualization tool enabling the analysis of human sequence variation complexity within an individual and between an individual and a reference genome. The analysis of the HuRef genome determined that non-SNP sequence variation is responsible for 22% of all variant events, but accounts for 74% of all variant bases. Consequently, the annotation and assembly-to-assembly comparison (ATAC) views in the HuRef browser enable users to intuitively analyze indels and other types of large-scale sequence variants including inversions, transpositions, VNTRs and copy-number variations (CNVs). The assembly view provides the basis for judging the quality of the assembly, confirming structural variants, determining the possible presence of alternate alleles and interpreting the potential significance of a given variation in the context of sequence annotations. Finally, incorporating the multiple sequence alignment of the underlying assembly and the pairwise alignment of an individual to a reference genome helps to interrogate both heterozygous and homozygous sequence variation.

## METHODS

### The HuRef database

The HuRef database consists of approximately 32 million DNA reads sequenced using Sanger dideoxy chemistry, assembled into 4528 scaffolds and 4.1 million DNA variations identified by genome analysis (4). These variants include SNPs, block substitutions, short and large indels, inversions, rearrangements, and copy-number changes. The data sources comprising the assembly and annotation of the HuRef genome, including the HuRef genome assembly (version 6), HuRef sequence variants and the HuRef haplotype blocks have been previously described (4,12). Additionally, the consensus module of the Celera Assembler produces a multiple sequence alignment of the reference sequence (12,15) that has been stored in the HuRef database, and made accessible by the Sequence Alignment component.

The HuRef database contains external data sources, including the reference sequence of the human genome (NCBI version 36), the gene annotation of the NCBI human reference sequence (Ensembl version 41), dbSNP variants (build 126), OMIM Annotations (13), Gene Ontology (14) and HUGO Gene nomenclature. These external data sets are mapped to the NCBI and HuRef coordinate systems using the ATAC sequence alignment tool (15–17). Additionally, the HuRef database maintains a direct sequence annotation of the HuRef assembly by aligning genes in the RefSeq collection (18). There are approximately 1000 genes predicted by the RefSeq annotation that are uniquely annotated on the HuRef genome.

The HuRef SNPs and indels in coding regions were classified using our internal tool, SNP Classifier, according to their position on any given transcript such as in a promoter, splice site, UTR, exonic or intronic region, and the effects of these variants on translation such as causing a synonymous or non-synonymous substitution, frameshift or protein truncation. Tandem repeats were identified on the HuRef and NCBI reference sequences using the default parameters for Tandem Repeats Finder (19).

### The HuRef Browser software

The HuRef Genome Browser is an *n*-tier web application built on standards-based, open source technologies to facilitate software reuse and exchange of scientific information. The HuRef database was designed to solve the dual challenge of using open, standard and sensible biological data formats while delivering a high-performance web-based data resource. Our solution was to implement an optimized data warehouse schema to serve as the read-only data access layer of the browser application, built on top of the Chado (20) database schema which serves to manage the primary source of the data and the transactional data operations.

The Chado database schema enables us to use standard ontologies for the typing of annotations and relationships (14), which provides for the proper semantic encoding of complex biological information, and allows us to leverage existing open-source tools and software. The HuRef database uses the open-source PostgreSQL RDBMS in keeping to our commitment to standards-compliant, open-source technologies.

The data and logic tiers are implemented in object-oriented Perl, and the BioPerl (21) Graphic Modules are utilized in the presentation layer. We use the MultiPanel module from Sybil (http://sybil.sourceforge.net/) to generate images of the genome assembly-to-assembly mappings. We use JavaScript and the Dojo Toolkit (http://dojotoolkit.org/) to build a high-performance web application that provides the rich, user-experience of a thick-client application. The HuRef Browser is currently hosted on a single, 2-CPU Enterprise Red Hat Linux 5.0 server, running the Apache 2 HTTP server.

The HuRef Browser and database is available as an open-source project under the GNU General Public License (GPL) license at https://sourceforge.net/projects/huref.

## RESULTS

The HuRef Browser is a web application for the navigation and analysis of an individual human diploid genome. The browser provides a comparative view between the HuRef and NCBI human sequences enabling the examination of small-to-large scale structural differences between any two assemblies. Users can navigate the HuRef genome assembly and sequence variations, and compare it with the NCBI human assembly in the context of the NCBI (18) and Ensembl (9) gene annotations. The browser has tracks representing the haplotype blocks from which diploid genome sequence can be inferred and the

relation of variants to gene annotations. The Sequence Alignment view displays the underlying sequence reads of the HuRef assembly, and the haploid phasing of the assembly is shown in precise detail.

**Features of the HuRef Browser**

Users can search the HuRef Browser using: HUGO gene names; RefSeq, NCBI Gene, Ensembl or dbSNP accessions; HuRef supercontig or contig locations; read identifiers; or NCBI chromosome coordinates. Alternatively, users can navigate to a region by selecting a chromosome band of interest in the karyotype view. Users can navigate in the vicinity of any genomic region via the pan and zoom controls. The views are configurable by the user, and any tracks can be turned on and off by a drop-down menu and rearranged by drag and drop controls.

*Assembly-to-assembly comparison view*. We used the ATAC tool to align the HuRef and NCBI human genome consensus sequences, and the resulting pairwise alignments are shown in the ATAC view. This view is designed to enable the visualization of complex, large-scale sequence variations including indels, inversions, rearrangements, VNTRs and other structural variants (Figure 1). Ungapped alignments between the genome assemblies are linked by either blue (same orientation) or pink (opposite orientation) blocks.

*Assembly view*. The assembly view displays a contig or supercontig region of the HuRef assembly, and its underlying clones, reads and sequence annotations (Figure 2). A supercontig is a collection of ordered and oriented contigs, sometimes referred to as a scaffold. The assembly information is useful for the identification of structural variations such as large indels and inversions, to identify alternate alleles in the HuRef assembly, or to identify potential misassemblies. The assembly is displayed using: (i) a contig track which displays the tiling of contigs for any scaffold region, (ii) a histogram track of read coverage, (iii) a histogram track of clone coverage and (iv) tracks to display the individual clones and reads. The HuRef consensus has been annotated by the NCBI RefSeq group (18). The NCBI gene models and RefSeq mRNAs are displayed as tracks in the same context as the Assembly view, along with tracks for the various types of HuRef variants.

The underlying clone layout is useful to identify compressions and expansions, and assess the local accuracy of an assembly. The clone tracks are separately displayed and categorized according to their mate pair status: satisfied, stretched, compressed, externally mated, unknown mate and unassembled mate. Satisfied clones are defined by the furthest endpoints of the clone pair being placed within 3 SDs of the mean insert distance of its library. Stretched and compressed clones are cases where the clone ends are greater or less than the satisfied clone distance, respectively.

There are a number of patterns that are often associated with alternate alleles in the HuRef assembly that can be inspected using the browser. For example, a reduction in read coverage is often indicative of a heterozygous indel.

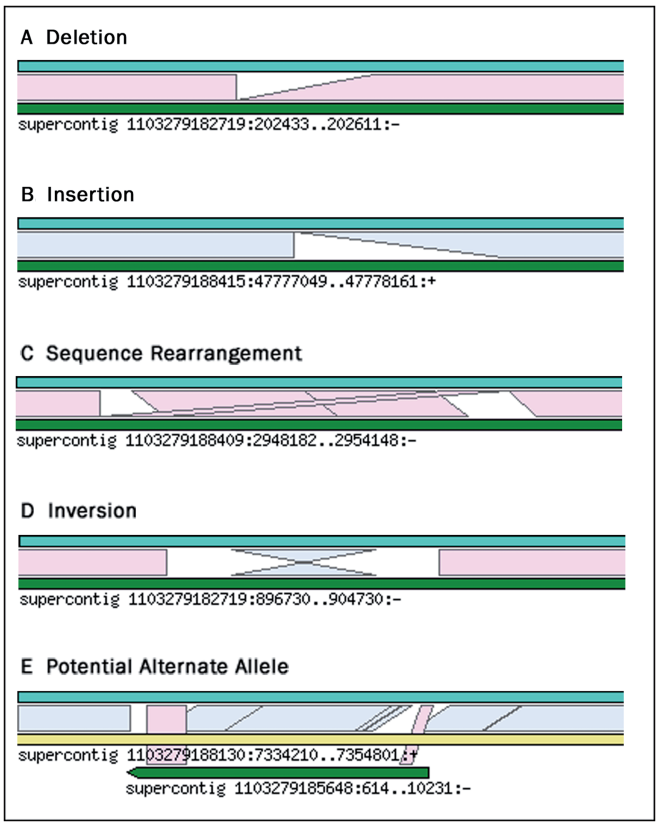


**Figure 1.** Representing complex sequence variation in the assembly comparison view. The ATAC view represents the mapping between two consensus sequences. In this example, the NCBI reference genome is represented as the teal rectangle on top, and the HuRef supercontig or contig that has the best match to the NCBI reference region is shown as the green and/or yellow rectangles at the bottom. Shaded lines represent the mapping between the two reference sequences, and the blue or pink fill color is use to denote the same or reverse orientation of the HuRef supercontig or contig with respect to the NCBI chromosome location. (**A**) Deletion of NCBI sequence with respect to the HuRef genome. (**B**) Insertion of HuRef supercontig sequence with respect to NCBI human reference. (**C**) Rearrangement of the HuRef supercontig with respect to the NCBI reference. (**D**) Inversion of HuRef with respect to NCBI reference. (**E**) HuRef assembly contains a ~10 kb scaffold that potentially represents an alternate haplotype to the larger supercontig that is the best match to this region on the NCBI reference.

A region with a number of externally placed clones that all share the same external contig assembly can be indicative of a large indel or structural variation. The assembly view enables the identification of these types of patterns, and to explore haplotype variability from single base pair SNPs to indels and large-scale sequence variations.

*Annotation view*. The annotation view displays Ensembl genes, dbSNP variants, HuRef variants and HuRef Haplotype Blocks as separate tracks (Figure 3). Differences in glyph color and shape are used to distinguish HuRef variants based upon their type and predicted functional impact (Supplementary Table 1). The annotation view is useful for understanding sequence variation within the context of genome annotation, to identify variants of interest, to review and compare the HuRef and

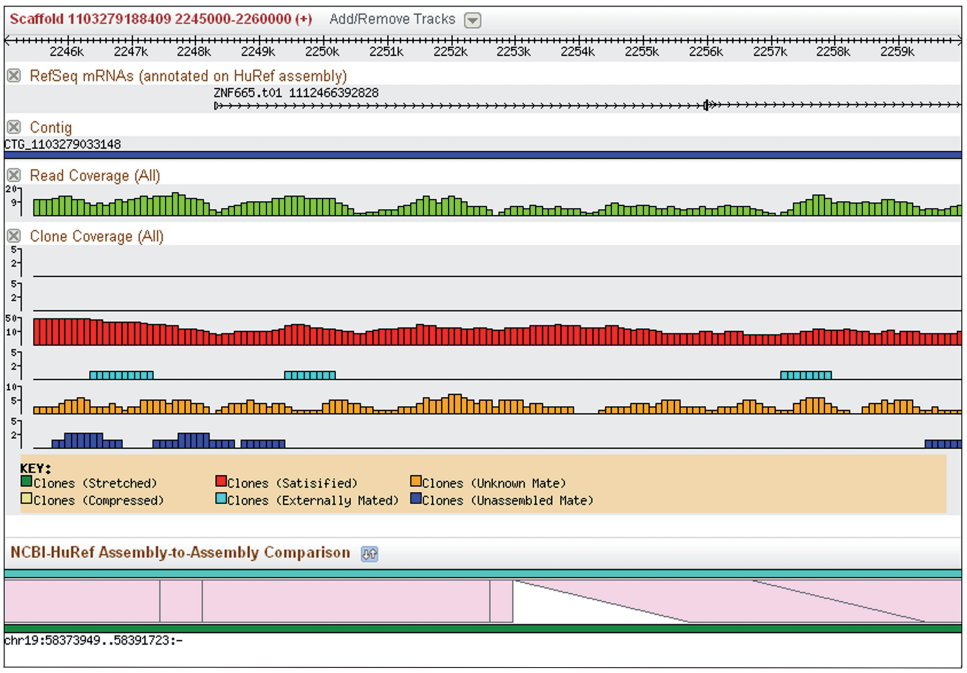


**Figure 2.** Assembly view showing a 3 kb Indel. This view represents the underlying structure of the HuRef assembly with tiers representing the contig layout, sequence read and clone coverage organized by their mate pair criteria. This example shows a 3 kb deletion on HuRef with respect to the NCBI reference genome. This indel ranges from 4–9× sequence read coverage and 10–25× clone coverage consisting of entirely well-placed clones. There are no stretched or compressed clones in this region which provides confidence in the quality of the assembly.

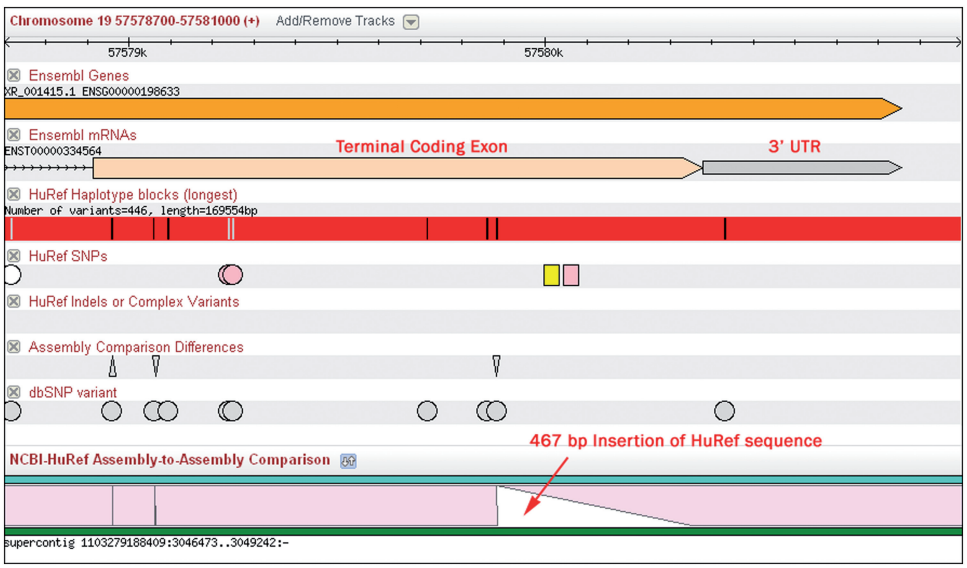


**Figure 3.** Annotation view. This view represents annotations on the reference sequence, including gene models, haplotype blocks, SNPs, indels, assembly comparison differences and complex sequence variations found on HuRef, as well as known variants from dbSNP. Glyph shape is used to distinguish between heterozygous and homozygous SNPs, and color is used to denote predicted impact on the encoded protein product. This example shows a 467 bp insertion on HuRef with respect to the NCBI reference sequence in the terminal coding exon of this Zinc-finger gene. The yellow and pink square glyphs in the HuRef SNPs tier represent heterozygous SNPs that introduce a synonymous and non-synonymous amino acid change, respectively. The pink circles represent homozygous SNPs, or differences between HuRef and the NCBI reference sequence, that introduce non-synonymous amino acid substitutions. See supplemental figure S1 for more detail on glyph shape and color.

dbSNP variants, and to inspect the Ensembl and NCBI gene annotations on either the HuRef or NCBI reference frame. Differences between the de novo annotation of HuRef and the mapped annotation on the NCBI human reference can be readily found using the HuRef Browser. The large-scale sequence variation found in the HuRef genome may lead to the identification of novel genes or alternate transcript forms (Supplementary Figure 2).

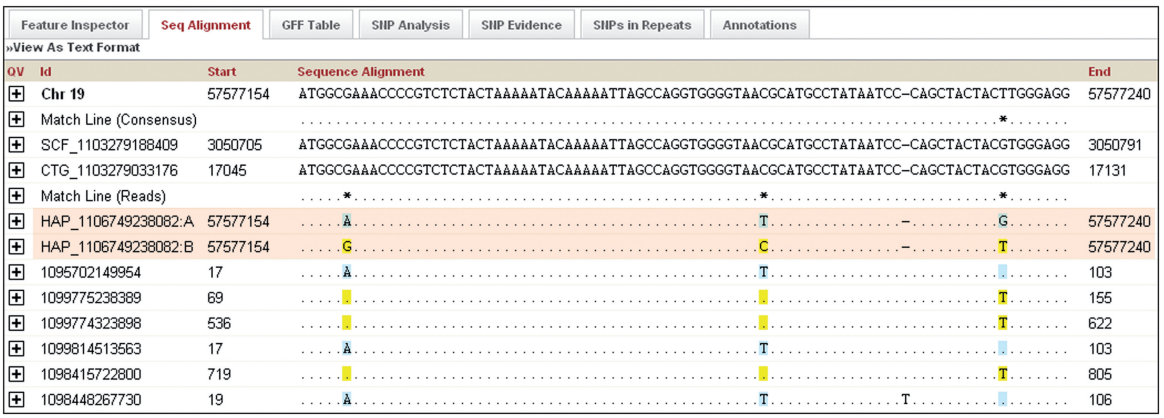


**Figure 4.** Sequence alignment view showing haplotype phasing. The sequence alignment of the 3′ intronic region of LOC400713, a Zinc finger gene on chromosome 19, shows three, closely-spaced SNP sites that have been phased to represent the diploid sequence of HuRef. This view includes the haploid sequences that overlap a given region, and color is used to highlight the variant columns and to show the phasing of variants. The haplotype assembly phases the three heterozygous SNPs in the haplotype block with the 'A' haplotype containing the A, T and G alleles, and the 'B' haplotype containing the G, C, T alleles for these three SNPs. Notice that the HuRef consensus is a mixed G, C, G allele and the correct phasing can only be observed by inspecting the haplotype sequences. The NCBI chromosome consensus is representative of the G, C and T haplotype.

*Browser components*. A selected feature or region of interest can be further analyzed using the components that are accessible at the bottom of the main browser window. The tabs in this portion of the browser page are: (i) *feature inspector*: this displays detailed information of any feature including coordinate locations, external database identifiers, links to related Web resources and the ability to export any feature in FASTA format (Supplementary Figure 3). (ii) *Seq alignment*: this displays the multiple sequence alignment of the NCBI consensus aligned to the sequence and associated quality values of the HuRef consensus and underlying sequence reads (Figure 4). The multiple sequence alignment can be examined to inspect variant calls and to determine the zygosity of the consensus at any position. (iii) *GFF table*: this displays the annotations visible in the Annotation View as a table that can be readily exported in GFF format (Supplementary Figure 4). (iv) *SNP analysis*: the SNP analysis component displays the predicted impact of the HuRef variants on functional sequence considering all overlapping transcript forms in a given region of interest (Supplementary Figure 5). (v) *SNP evidence*: this component provides the set of evidence criteria used for variant identification including read and clone placement, read coverage, read orientation to confirm a given variant in both sequencing directions, read and consensus quality values, and clone statistics (Supplementary Figure 6). (vi) *SNPs in repeats*: this component lists repeat features that overlap any annotated variants (Supplementary Figure 7). (vii) *Annotations*: this lists all external database references including OMIM and Gene Ontology links/assignments of genes and other annotations for a given region of interest. All of the datasets for these components can be exported in standard formats.

## Examples of using the browser

The HuRef Browser provides an easy way for users to identify variants of structural or functional interest in a genome, to check an individual's genotype, to examine the underlying evidence for any variant, to assess the impact of a variant on a particular protein product, or to verify genome assembly. The ability to provide accurate genotype calls whose underlying supporting evidence can be examined is paramount when providing a personal genomics profile to an individual. This is especially true for genotypes whose implication in genetic diseases has been established (11).

*Example 1. Analysis of a heterozygous SNP associated with night blindness*. *CACNA2D4* is a gene that encodes a regulatory subunit in the voltage-dependent calcium channel complex that mediates ion influx to the cell. A C-to-A transversion was identified (23) that introduces a premature stop codon (TAC → TAA) and truncates one-third of the open reading frame. This truncation has been shown to cause retinal cone dystrophy, an autosomal recessive condition. Upon navigating to this gene in the browser, users will notice a red round glyph in the HuRef SNPs track in the Annotation View representing a heterozygous variant that causes a frameshift or protein truncation according to our variant analysis (Supplementary Figure 1). The Sequence Alignment view of this G/T SNP shows ten reads that overlap the variant column, of which two support the alternate T allele, indicating that the HuRef donor is heterozygous for this variation and is a carrier of this recessive disorder. The predicted functional impact of the variant can be obtained by using the SNP Analysis component that confirms that the variant truncates the open reading frame of this protein-coding gene.

*Example 2. Analysis of a 14 kb inversion in a TNF receptor gene*. There is a 14 kb inversion in the HuRef donor on chromosome 1 from 2.47 to 2.49 MB (Supplementary Figure 2). This inversion fully spans the genic region of *TNFRSF14*, a member of the Tumor Necrosis Factor receptor superfamily. This gene plays a key role in

regulating the immune response to infection, and is involved in lymphocyte activation. The inversion is contained in a 2.3 MB contig with 23× and 27× clone coverage spanning the 5′ and 3′ breakpoints of the inversion, respectively, with a read in the inversion and its mate in the non-inverted region. This provides a significant degree of confidence in the inversion call. It is possible that this insertion may affect *TNFRSF14* regulation if this gene is *cis*-regulated by upstream or downstream sequences, although further analysis is required.

## DISCUSSION

In 2001, two versions of the human genome were published, both of which were comprised of a mosaic of an ethnically diverse population of individuals. Because of the composite nature of these genomes, the individual haplotype sequences were almost entirely lost. Subsequently, the genome sequences of Dr Craig Venter and Dr James Watson were published (4,22). This has signified a new era of human genomics and it is expected that thousands of personal human genomes will be publicly available in the near future. We can begin to unravel the unique traits and disease propensities that are encoded in an individual's double-stranded DNA on a genome-wide scale as this data becomes available.

In this article, we present the HuRef Genome Browser, a new web resource that provides open, public access to the de novo assembly and annotation of an individual diploid human genome (4). The HuRef Browser enables the analysis of the complete spectrum of sequence variation, from SNPs to indels, rearrangements, VNTRs and CNVs. It brings together the visual analytic tools of assembly, annotation and synteny viewers in an easy-to-use, integrated framework.

The underlying evidence, especially the assembly, which forms the basis of the subsequent annotation and analysis processes, is open and publicly available to the scientific community. This is especially important in the analysis of personal genomes because of the significant implications that some disease-associated genotypes may have to an individual donor. Due to the complexity of the human genome and limitations in technologies and algorithms, the cloning, sequencing, assembling and data analysis processes can lead to errors in the automated detection of sequence variation. Therefore, it is prudent to verify genotypes by a full examination of the sources of evidence. The HuRef Browser is designed to facilitate this kind of detailed examination by providing the variant evidence-related data, including the placement and status of reads and clones assembled into contigs and scaffolds, the sequence alignment, quality values and other measures that are associated with accurate variant detection methods. This information serves the scientific community to make the correct analysis of an individual genome based upon all of the available information.

It is now feasible to sequence individual human genomes on a large scale given the dramatic increase in the efficiencies of DNA sequencing. Researchers are generating individual human sequence data at an ever-increasing rate (4,22,24–26), motivating the need for effective, high-performance visualization tools to study sequence assembly and variation in a population of human individuals.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to thank Gennady Denisov for his help in the development of the sequence alignment capabilities. Brian Walenz for his valuable comments and contributions towards the assembly comparison displays. Kelvin Li for his support and the use of his variant classification software. We wish to express our gratitude to Stephen Scherer and Lars Feuk at The Centre for Applied Genomics at The Hospital for Sick Children, as well as Lisa Stubbs at the Lawrence Livermore Laboratory for their valuable suggestions to improve the user interface and information layout of the browser. Lastly, we would like to thank Tom Emmel and Hank Wu for their support in administering the IT system resources.

## FUNDING

This work was supported by the J. Craig Venter Institute. Funding for open access charge: J. Craig Venter Institute.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Sachidanandam,R., Weissman,D., Schmidt,S.C. and Kakol,J.M. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
4. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
5. Schatz,M.C., Phillipyy,A.M., Shneiderman,B. and Salzberg,S.L. (2007) Hawkeye: a visual analytics tool for genome assemblies. *Genome Biol.*, **8**, R34.
6. Dew,I.M., Walenz,B. and Sutton,G. (2005) A tool for analyzing mate pairs in assemblies (TAMPA). *J. Comput. Biol.*, **12**, 497–513.
7. Huang,W. and Marth,G.T. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.*, **18**, 1538–1543.
8. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
9. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
10. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The

generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
11. Ng,P.C., Levy,S., Huang,J., Stockwell,T.B., Walenz,B.P., Li,K, Axelrod,N., Busam,D.A., Strausberg,R.L. and Venter,J.C. (2008) Genetic Variation in an individual human exome. *PLoS Genet.*, **4**, e1000160.
12. Denisov,G., Walenz,B., Halpern,A.L., Axelrod,N., Levy,S. and Sutton,G. (2008) Consensus generation and variant detection by Celera Assembler. *Bioinformatics*, **24**, 1035–1040.
13. McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
15. Istrail,S., Sutton,G.G., Florea,L., Halpern,A.L., Mobarry,C.M., Lippert,R., Walenz,B., Shatkay,H., Dew,I., Miller,J.R. *et al.* (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA*, **101**, 1916–1921.
16. Lippert,R.A., Zhao,X., Florea,L., Mobarry,C. and Istrail,S (2005) Finding anchors for genomic sequence comparison. *J. Comput. Biol.*, **12**, 762–776.
17. Shatkay,H., Miller,J., Mobarry,C., Flanigan,M., Yooseph,S. and Sutton,G. (2004) ThurGood: evaluating assembly-to-assembly mapping. *J. Comput. Biol.*, **11**, 800–811.
18. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
19. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
20. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
21. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
22. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y. and Chen,L. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
23. Wycisk,K.A., Budde,B., Feil,S., Skosyrski,S., Buzzi,F., Neidhardt,J., Glaus,E., Nurnberg,P., Ruether,K. and Berger,W. (2006) Structural and functional abnormalities of retinal ribbon synapses due to Cacna2d4 mutation. *Invest. Ophthalmol. Vis. Sci.*, **47**, 3523–3530.
24. Pennisi,E. (2006) Genomics. On your mark. Get set. Sequence! *Science*, **314**, 760.
25. Church,G. (2005) The personal genome project. *Mol. Syst. Biol.*, **1**, 30.
26. Blow,N. (2007) Genomics: the personal side of genomics. *Nature*, **449**, 627–630.