Understanding genome browsing

Melissa S Cline & W James Kent

How can genome browsers help researchers to infer biological knowledge from data that might be misleading?

As genomic knowledge expands, new forms of data become available to help interpret genomic sequences. However, biological data can be noisy: living systems are complex and measurement technologies are rarely perfect. Two excellent approaches for reducing noise are data aggregation and visualization. When combined, multiple forms of evidence tend to be more accurate than a single source, as each distinct form reduces overall uncertainty[1]. The human mind is an outstanding data analysis tool. Although it absorbs textual data poorly, it can assimilate visual data in great detail[2], and can process it efficiently to identify common themes[3].

Genome browsers facilitate genomic analysis by presenting alignment, experimental and annotation data in the context of genomic DNA sequences. These include the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/), Ensembl (http://www.ensembl.org/), and National Center for Biotechnology Information (NCBI) Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/). They differ in their user interfaces, but address similar tasks, as described in **Supplementary Notes** online and reviewed elsewhere[4]. We focus here on the UCSC Genome Browser.

Figure 1 shows the display for a representative gene queried using the UCSC Genome Browser. The browser displays several tracks, or collections of data, some of which are hidden by default. The user controls which tracks are displayed by means of pull-down menus below the image. The track names are hyperlinked to pages that detail how the data were computed, outline any specific display conventions and may offer additional display options. Each track item within the browser is hyperlinked to a details page providing further information on that item, such as publications in PubMed and sequences in GenBank. The importance of studying these details cannot be overstated. Although genome browsers can simplify the task of generating hypotheses, the user must still evaluate the facts carefully to ensure that the hypotheses are likely to be valid.

Melissa S. Cline is in the Department of Molecular, Cell and Developmental Biology and W. James Kent is at the Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA.
e-mail: kent@soe.ucsc.edu

Gene structure and transcripts

Arguably the most important tracks are those that indicate the genes. No data indicate 'the genes' unambiguously. Genes are detected through experimental evidence (namely, observed transcription), and rare transcripts are often difficult to distinguish from measurement errors. To address this uncertainty, there are many gene and gene prediction tracks, each with its own evidence standards.

The high-confidence, low-coverage end of the spectrum contains tracks that derive gene structures from specific full-length transcripts (**Fig. 1a**, line 2). The track indicating genes[5] from the Mammalian Gene Collection (MGC) shows transcripts sequenced from selected high-quality clones. RefSeq Genes[6] shows expert-curated transcripts, along with some provisional transcripts awaiting curation.

For increased coverage, UCSC Genes[7] (**Fig. 1a**, line 1) and Ensembl Genes[8] (**Fig. 1a**, line 3) show predicted transcripts that are derived from mRNA, expressed sequence tag (EST) and protein-sequence alignments. Unlike the RefSeq and MGC transcripts, these transcripts do not always correspond to any single mRNA sequence, but represent composites of sets of similar aligned sequences with good overall evidence.

Aligned sequences offer the broadest but noisiest transcript data. The human mRNA (**Fig. 1a**, line 4) and spliced EST (**Fig. 1b**) tracks show GenBank sequences that align well to the genome. ESTs are short fragments obtained from a single sequencing pass, whereas mRNAs are obtained by high-quality sequencing of entire cDNAs. In general, ESTs describe more transcript isoforms, whereas mRNAs describe fewer isoforms but do so with greater accuracy. However, any aligned sequence is only as good as its underlying clone. If a clone is of poor quality, even the best sequencing protocols will yield misleading sequences. Thankfully, such sequences can often be identified—and disregarded—by following commonsense rules, such as those described below.

Interpreting aligned sequences

First, sequences that align with many errors should be trusted less, because they might not be bona fide products of the locus. Colored vertical lines indicate mismatches and insertions, and double horizontal lines indicate gaps. Sometimes, mismatches arise through normal genetic variation. Such cases can be identified by comparison against data from dbSNP[9] (**Fig. 1a**, line 10).

Second, one should not trust any variation evidenced from only one aligned sequence. For example, BE891408 (**Fig. 1b**, arrow vi) seems to suggest two novel exons, although no other alignment contains these exons. Furthermore, the details page of this EST indicates an older publication date. Together, these facts indicate that this EST should be disregarded.

Third, two or more questionable alignments support each other only if they were derived independently. Aligned sequences are often redundant, with multiple sequences derived from the same clone or from related clones in the same laboratory. Such cases are

not independent observations, but one observation recorded multiple times. The browser display is also redundant, as all MGC genes transcripts also appear under human mRNAs (**Fig. 1a**, arrows i and iv). This detail would be easy to miss, and could lead to misinterpretation of sequence-variation frequencies.

Fourth, one should be careful with alignments that suggest partial or erroneous cellular processing. This includes mRNAs that are not spliced or have retained introns (such as BC062326, **Fig. 1a**, arrow iii); mRNAs with premature stop codons, that fall well before the last splice site (such as AK023398, **Fig. 1a**, arrow ii); and run-on alignments that extend past the bounds of the loci (such as DA949381, **Fig. 1b**, arrow v). When such alignments are not supported by others, they probably indicate biological noise.

Finally, a short transcript does not imply a short transcribed region. Aligned sequences are often incomplete, especially in the untranslated regions (UTRs). Sequences are frequently cloned with incomplete UTRs for technical reasons, and sequencers often stop reading prematurely. Thus, variation in alignment lengths might not represent transcript variation; absence of evidence is not evidence of absence. Some tracks can indicate genuine variation: transcription factor binding site data can suggest alternative promoter usage, and the Poly(A) track[10,11] (**Fig. 1a**, line 5) can suggest alternative polyadenylation. For example, the polyA sites near the center of **Figure 1a** suggest that some of the shorter transcripts are actually complete isoforms.

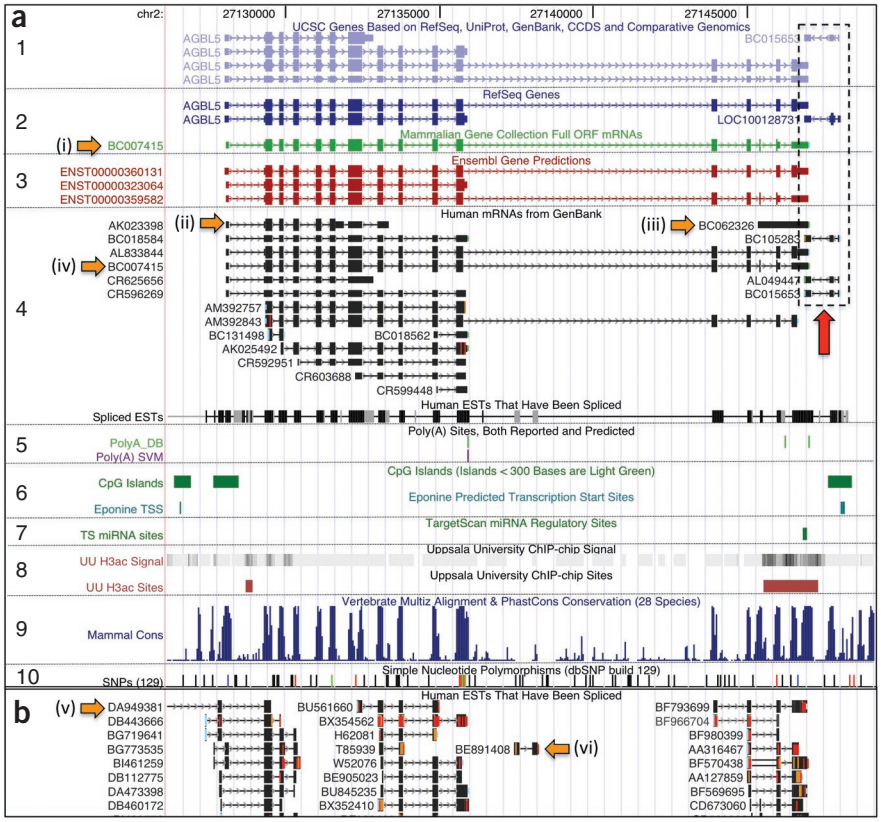

Figure 1 Illustrative screen shots from the current UCSC Genome Browser. (**a**) Selected tracks for the human *AGBL5* locus. 1. UCSC Genes[7]; 2. RefSeq Genes[6] and MGC[5] Genes; 3. Ensembl Genes[8]; 4. Human mRNAs and Spliced ESTs; 5. Poly(A)[10,11]; 6. CpG Islands[14] and Eponine TSS[15]; 7. TS miRNA sites; 8. Uppsala ChIP[16], 9. Conservation[12]; 10. SNPs (129)[9]. Most tracks are shown in pack display mode, with each item displayed separately. The CpG Islands, spliced ESTs, SNPs (129) and TS microRNA sites tracks are shown in dense mode, with all items condensed to a single display line. Darker portions of the EST track indicate regions of stronger evidence, which suggests greater likelihood that the regions are transcribed. In lines 1–4, each track item represents a transcript. Exons are shown as rectangles: taller rectangles indicate coding (CDS) segments, whereas shorter rectangles represent untranslated regions. Introns are shown as lines connecting exons, with arrowheads indicating the direction of transcription. Most transcripts shown are transcribed left to right, in the 5′ to 3′ direction on the sense strand. The dashed box, marked with the red arrow, indicates transcripts of the BC015653 locus on the antisense strand. The human mRNAs track is colored to show mRNA codons that are nonsynonymous to the genome. Orange arrows indicate (i,iv) an mRNA found in both the MGC genes and human mRNAs tracks (BC007415), (ii) an mRNA with a premature stop codon (AK023398) and (iii) an unspliced mRNA (BC062326). (**b**) Excerpt of the spliced ESTs track shown in pack mode, colored to indicate bases that differ from the genomic sequence. Orange arrows indicate (v) a run-on EST and (vi) an alignment consisting of two blocks that are not contained in any other aligned sequence.

Conservation and regulatory data

Figure 1a (line 9) shows genomic conservation, as inferred by MultiZ phylogenetic alignments of genomic sequences[12]. Overall, conservation is strongest in coding exons, weaker in UTRs and weakest in introns and intergenic regions. Strong conservation suggests functional importance, and highly conserved noncoding regions often contain regulatory signals.

TargetScan[13] (**Fig. 1a**, line 7) predicts microRNA binding sites in the highly conserved 3′ UTR. One might assume that this region is highly conserved to preserve these sites. Although this might be true, caution is warranted. TargetScan's track description page indicates that predictions are derived from MultiZ alignments: the predictions depend on conservation. This exemplifies the importance of investigating all of the details before drawing conclusions.

Figure 1a (lines 6 and 8) shows transcriptional start sites suggested by three separate lines of evidence: CpG islands[14], predicted transcription start sites[15] and experimentally determined acetylated histone H3 sites[16]. Each of these signals can be misleading: some genes have no CpG islands, transcription factor binding predictors often overpredict and histone measurement is noisy. However, in aggregate, such data can yield a strong, synergistic prediction.

Moving beyond visualization

After examining a locus, it is often valuable to save data in a text-based format for subsequent analysis. This can be done using the Table Browser[17], accessible through the 'Tables' link. It allows users to select a track, and extract the data from that track for a specific region (defaulting to the last region visualized), or genome-wide. For example, selecting the SNPs (build 129) track and position button allows users to extract a list of SNPs for the region last visualized.

Although genome browsers allow one to scan visually for loci with certain attributes, it can be easier to identify loci with those attributes and then evaluate them visually. This can be done with the Table Browser's filter and intersection functionality. Filtering

allows one to limit the track items according to data within the track. For example, one can obtain a list of predicted p53 binding sites by selecting the TFBS Conserved track and filtering for items with names matching "*p53*" (the 'describe table schema' button outlines the available fields). By intersecting the filtered track with the GIS ChIP-PET track[18], one can identify predicted p53 binding sites that are supported experimentally. For output, one can select a set of hyperlinks to the Genome Browser. Or, one can save the output as a custom track and further refine this track through additional filter and intersection actions. This allows users to build sophisticated queries to identify genomic regions sharing a combination of traits.

Conclusion

This primer describes a small subset of the analyses possible with genome browsers, but illustrates some basic principles. Virtually any genomic data can be erroneous, and one should be wary of data suggested by only a single observation. Nonetheless, the combination of multiple observations can suggest reliability, especially when the observations come from varying forms of evidence. Genome browsers facilitate such combination by presenting data visually, in a genomic context. Additional analysis scenarios are described under the recommended resources in **Supplementary Box 1** online.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Many people have made significant contributions to this manuscript. In particular, we wish to thank Mark Diekhans, Rachel Karchin, Donna Karolchik, Rachel Harte, Trey Lathe, Mary Mangan and Brooke Rhead for their insights. Bert Overduin and Deanna Church generously offered insights on how to perform similar analyses with the Ensembl and NCBI Map Viewer browsers, so that this paper might assist a broader community of users. This work was funded by the National Human Genome Research Institute (2 P41 HG002371-06 to UCSC Center for Genomic Science, 3 P41 HG002371-06S1 ENCODE supplement to UCSC Center for Genomic Science) and National Cancer Institute (contract no. N01-CO-12400 for Mammalian Gene Collection). M.S.C. is supported under National Institutes of Health GM-040478, and thanks Manuel Ares, Jr. for his mentorship and encouragement. We are grateful to the many researchers who have contributed tracks to the UCSC Genome Browser, as the value of any browser is determined by the underlying data.

COMPETING INTERESTS STATEMENT

The author declares competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/

1. Shafer, G.A. *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976).
2. Miller, G.A. *Psychol. Rev.* **63**, 81–97 (1956).
3. Bauer, M. & Johnson-Laird, P. *Psychol. Sci.* **4**, 372–378 (1993).
4. Furey, T.S. *Hum. Genomics* **2**, 266–270 (2006).
5. Gerhard, D.S. *et al. Genome Res.* **14**, 2121–2127 (2004).
6. Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **35**, D61–D65 (2007).
7. Hsu, F. *et al. Bioinformatics* **22**, 1036–1046 (2006).
8. Flicek, P. *et al. Nucleic Acids Res.* **36**, D707–D714 (2008).
9. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
10. Tian, B., Pan, Z. & Lee, J.Y. *Genome Res.* **17**, 156–165 (2007).
11. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. *Nucleic Acids Res.* **33**, 201–212 (2005).
12. Siepel, A. *et al. Genome Res.* **15**, 1034–1050 (2005).
13. Lewis, B.P. *et al. Cell* **115**, 787–798 (2003).
14. Gardiner-Garden, M. & Frommer, M. *J. Mol. Biol.* **196**, 261–282 (1987).
15. Down, T.A. & Hubbard, T.J. *Genome Res.* **12**, 458–461 (2002).
16. Rada-Iglesias, A. *et al. Genome Res.* **18**, 380–392 (2008).
17. Karolchik, D. *et al. Nucleic Acids Res.* **32**, D493–D496 (2004).
18. Ng, P. *et al. Nat. Methods* **2**, 105–111 (2005).