The Aldehyde Dehydrogenase Gene Superfamily Resource Center

*William Black and Vasilis Vasiliou**

Molecular Toxicology and Environmental Health Sciences Program, Department of Pharmaceutical Sciences, University of Colorado Denver, Aurora, CO 80045, USA
**Correspondence to:* E-mail: vasilis.vasiliou@ucdenver.edu

Date received (in revised form): 23 October 2009

Abstract

The website www.aldh.org is a publicly available database for nomenclature and functional and molecular sequence information for members of the aldehyde dehydrogenase (*ALDH*) gene superfamily for animals, plants, fungi and bacteria. The site has organised gene-specific records. It provides synopses of *ALDH* gene records, marries trivial terms to correct nomenclature and links global accession identifiers with source data. Server-side alignment software characterises the integrity of each sequence relative to the latest genomic assembly and provides identifier-specific detail reports, including a graphical presentation of the transcript's exon–intron structure, its size, coding sequence, genomic strand and locus. Also included are a summary of substrates, inhibitors and enzyme kinetics. The site provides reference lists and is designed to facilitate data mining by interested investigators.

Keywords: *Genomic database, aldehyde dehydrogenase, ALDH, nomenclature, gene superfamily*

Introduction

The completion of various genome projects and the growing trend towards high-throughput data production have created a significant knowledge base of molecular sequence data across a broad spectrum of species. This increase in available sequence information has led to a widening gap between the available raw sequence data and their functional analyses by molecular biological methods or other genetic approaches. As a consequence, the field of bioinformatics has rapidly developed as an essential aid for data analysis. A number of large-scale, gene-specific databases, including the National Center for Biotechnology Information (NCBI)'s Entrez Gene[1] and the European Bioinformatics Institute/Wellcome Trust Sanger Institute's Ensembl databases,[2] have developed to report and catalogue molecular sequence data. The intrinsic format of these databases in attempting to cover all genes for all species or to cover all genes for a given species (eg the mouse genome database[3]), however, has significant limitations. These include errors in sequence alignments due to a reliance on automated algorithms, poorly defined reference sequences and improper gene nomenclature. Other issues include lack of identification and/or categorisation of alternatively spliced transcriptional variants, as well as erroneous functional characterisations because generalised gene ontology entries do not distinguish the individual gene from other members of its gene superfamily. To address these limitations, we have developed a gene-specific database architecture and web-based scripting system which is tailored to report both the molecular sequence and functional data for all members of an individual gene superfamily across all species (Black and Vasiliou, manuscript in preparation). Using this software and relational database architecture, we have developed www.aldh.org, a publicly available

informational resource system for all members of the aldehyde dehydrogenase (ALDH) gene superfamily. The ALDH gene superfamily is an evolutionarily ancient group of genes spanning all the kingdoms and phyla known today. The ALDH website is designed to provide a comprehensive 'gold standard' dataset across a variety of species for the molecular and functional information pertaining to members of this superfamily.

Site design, hosting and software

The www.aldh.org website is designed, hosted and curated by the authors and is currently hosted at the University of Colorado's Anschutz campus in Denver, CO, USA. To accommodate various hosting platforms, the site's software infrastructure is capable of running on Microsoft Windows Server or Linux platforms running Internet Information Services 5 or higher (Microsoft) or Apache website hosting software. The site database operates on the open-source database software MySQL (version 5.0.51a) and content is dynamically generated via server-side scripting using the open-source script engine, PHP (version 5.2.9-2).

Organisation of the web database

From its launch in 1999, the ALDH gene superfamily website has grown exponentially and continues to do so as more genome projects are completed and become openly accessible. The www.aldh.org website provides comprehensive access to molecular, functional and bibliographic elements for each *ALDH* in the gene superfamily for human, mouse and rat. Additional superfamily data for other animal species, as well as for plants, bacteria and fungi, are regularly incorporated into the database and presented as completed.

The website's homepage welcomes users with regularly updated news and information about members of the superfamily, as well as any newly available site features. It displays a 'record status' summary table totalling the number of *ALDH* gene superfamily members within each species and quick links to each of the respective gene records within the database. The website's global navigation bar is located along the top of the homepage and allows visitors easily to access the 'ALDH overview' section, providing a general history and review of the *ALDH* gene superfamily and its nomenclature system. The '*ALDH* gene superfamily' link provides a complete tabular summary of all *ALDH* genes, with navigational links to all relevant records within the database. The '*ALDH* publications' section displays a comprehensive reference list for the *ALDH* gene superfamily, sorted by *ALDH* subfamily and gene. A 'Links' page contains data-mining sources and toolsets, and a 'Laboratory' page describes our personnel and their respective research interests. A local navigation system is situated on the left-hand side of each page within the www.aldh.org website, to enable users quickly to scroll their web browsers within the current page being viewed to pre-defined bookmarked subsections. Therefore, the global and local navigation systems provide users with a simple and uniform structure throughout the site, enabling ease of access to all database information.

The core functionality of the site is structured around the gene-specific records for each *ALDH* within the gene superfamily. All database information for each *ALDH* gene record is dynamically generated, organised and displayed to the user's web browser by the website's server-side scripting engine in a clear, concise and user-friendly approach.

Gene-specific records

Each gene-specific record is derived from a global template, utilising a series of curator-specified software data modules tailored to the *ALDH* gene superfamily. These modules perform pre-defined database queries and generate all of a record's text, tables and graphical representations for the end-user's web browser. At the time of writing, software data modules registered in this global template generate the 'Synopses', 'Trivial names', 'Global accession identifiers', 'Molecular features and cataloguing', 'Accession identifier details', 'Human polymorphisms', 'Enzyme kinetics', 'Tissue expression profiles' and 'Reference list' sections for each *ALDH* gene.

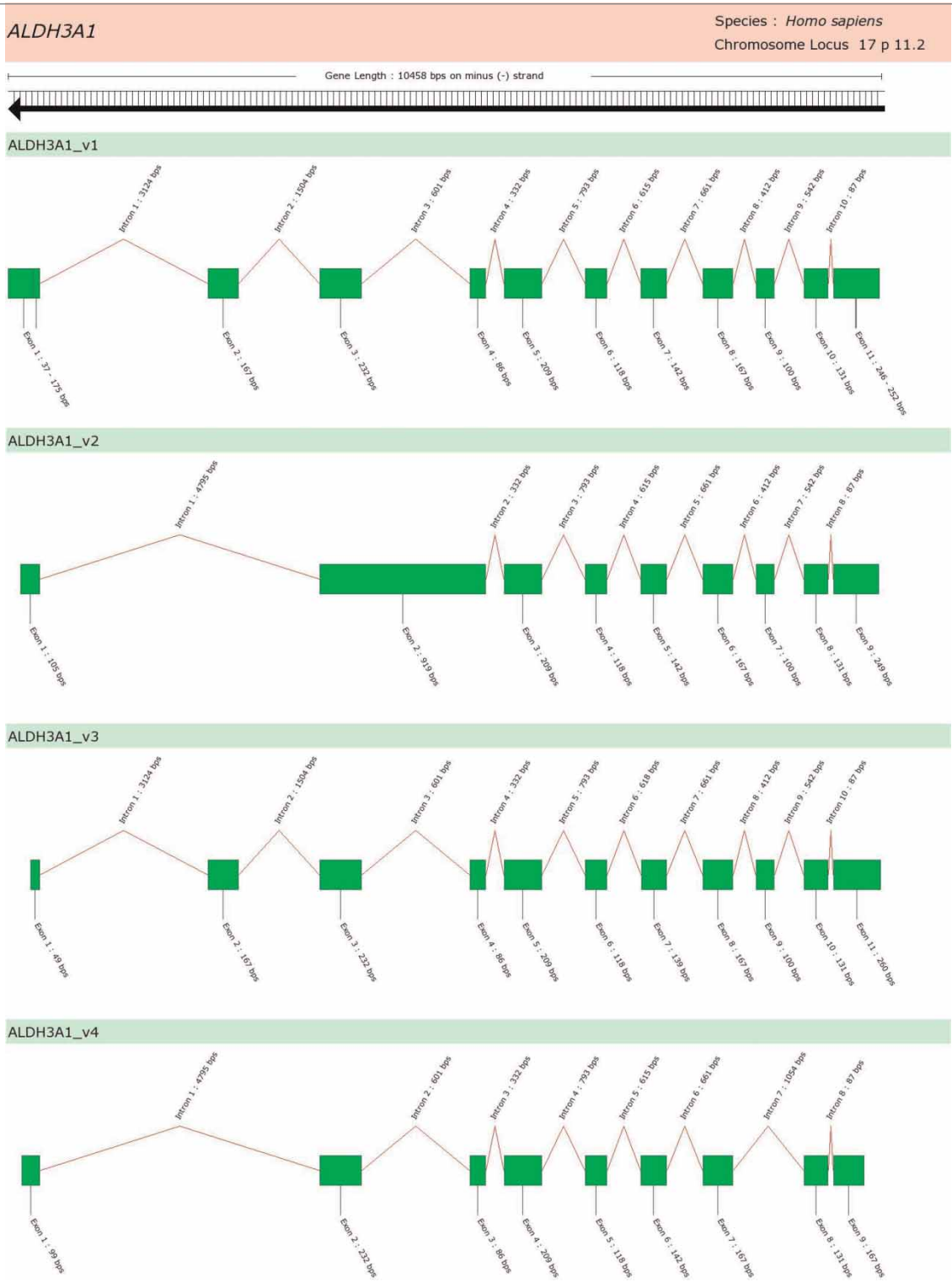

Figure 1. Alternatively spliced transcriptional variants for the *ALDH* gene superfamily member, *ALDH3A1*. The graphical representations for transcriptional variants for each gene superfamily member are dynamically generated based on sequences from each accession identifier and represent the splice variants for a particular gene and species, thereby allowing users quickly to identify similarities and differences between all available splice variants.

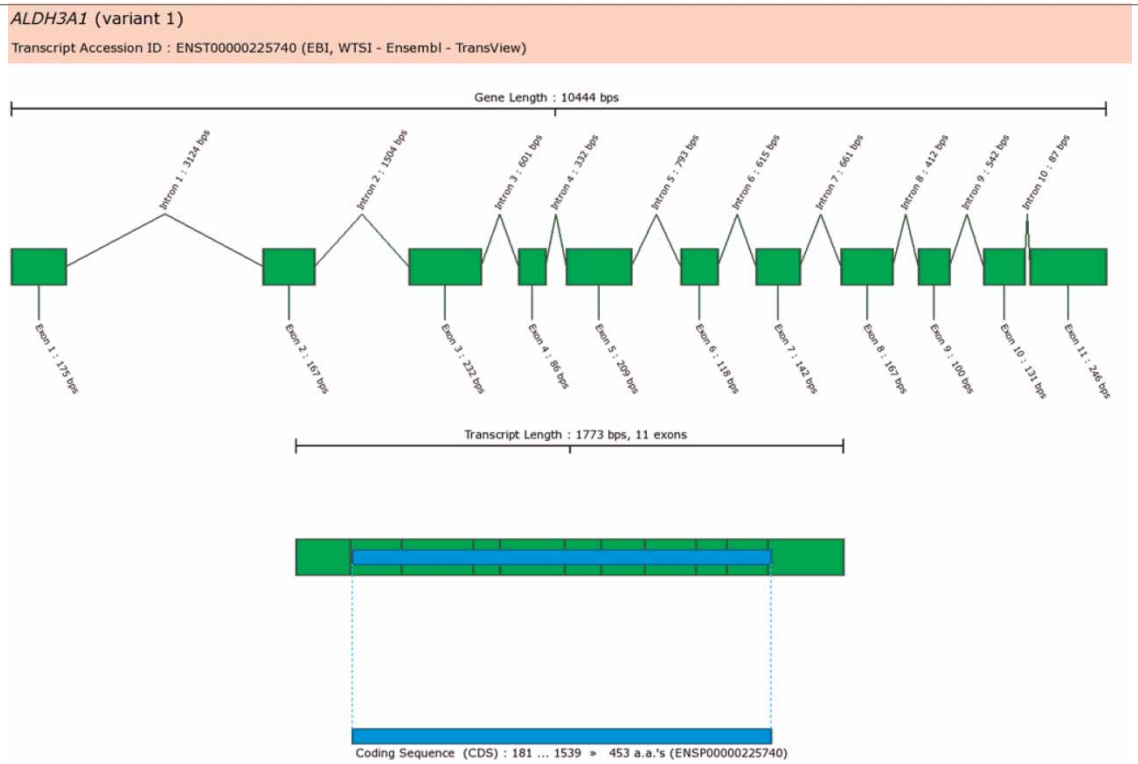

Figure 2. Graphical representation of the transcript sequence for individual accession identifiers. Each accession identifier's sequence within the database is analysed for inter- and intra-sequence structure. Data from this analysis is then parsed and graphically represented for each accession identifier in the 'Accession Identifier Detail Reports' section found under Molecular Features for each gene record. If present, any variations in the transcript sequence as a result of synonymous and non-synonymous polymorphisms relative to the genomic sequence are highlighted.

Synopses

ALDH gene records begin with a living review of the literature, prepared and regularly updated by the authors, describing available structural and functional data. All information and illustrations are referenced to the original publication using hyperlinked PubMed database identifiers (PMID). In addition, details for each reference are further described within the reference list section of the record.

Trivial names

Members of the *ALDH* gene superfamily have had a number of descriptive synonyms used in the literature over the past three decades. This section associates these synonyms with their corresponding nomenclature.

Global accession identifiers

Accession identifiers for a given *ALDH* gene, transcript or peptide sequence are numerous due to the multitude of disparate databases providing molecular sequence data. The 'Global accession identifiers' module tabulates these accessions into hyperlinked identifiers to all source databases. This enables users quickly to access all source data for an individual *ALDH* member.

Molecular features and cataloguing

The molecular sequence data available from source databases for most accession identifiers typically provide little perspective on the sequence's genomic alignment integrity or how the sequence compares with other transcripts for a particular gene. Publicly available alignment tools can detect if sequences for two or more transcripts are different at specific positions, but often cannot tell exactly where they differ relative to the gene's alternatively spliced transcription schema. This module uses the server-side alignment software to characterise the integrity of each

accession's sequence relative to the latest genomic assembly for its respective species. Specifically, all transcripts for each *ALDH* for a given species within the www.aldh.org database are co-aligned to their genomic assembly. These data are then used dynamically to generate a graphical representation of all transcripts for a particular gene and species, allowing the user quickly to identify the similarities and differences between alternatively spliced variants (Figure 1). The software module then categorises and tabulates all *ALDH* accession identifiers into their respective alternatively spliced transcriptional variant (ie *ALDH3A1_v1*, *ALDH3A1_v2*, etc) and briefly summarises size, isoelectric point, number of exons, FASTA sequence summaries and any sequence anomalies (single nucleotide polymorphisms [SNPs], insertions or deletions) relative to the genomic assembly for a each identifier.

Accession identifier detail reports

Most source databases provide limited or vague details about the accession identifier's sequence. Consequently, the www.aldh.org database uses the server-side scripting software to process each accession identifier's sequence for all available information. This processing generates an individual report of details for each accession identifier as a new webpage on the end-user's browser. The detail report for each identifier is found within the 'Molecular features' section by clicking on 'Click here for graphical and tabular details' for each accession identifier catalogued. The

Figure 3. Graphical representation of the peptide sequence for individual accession identifiers. Peptide translation for each transcript sequence is represented by a multicoloured graphic of the primary sequence based on amino acid content (Figure 3a). If present, any variations in amino acids as a result of synonymous and non-synonymous polymorphisms relative to the genomic sequence are highlighted. In addition, users are able to view the locations of various residues within the peptide sequence (eg cysteine residues; Figure 3b) by clicking on a specific residue in the amino acid content table of Figure 3a.

objective is to provide a succinct data analysis for each accession's sequence and begins with a tabular summary of the accession identifier's source information. Using the identifier's sequence, a graphical representation of the transcript's exon–intron structure, as well as its size, coding sequence, genomic strand and locus, is then generated (Figure 2). A multicoloured graphical representation of the transcript's corresponding peptide translation (Figure 3a) is generated and provides positional highlights for any synonymous or non-synonymous polymorphisms relative to the genomic assembly. A residue content summary for the peptide sequence, displayed next to the sequence image, facilitates easy review of the positional coordinates of any residues (eg cysteine residues; Figure 3b). Next, the 'Transcript sequence and structural features' section breaks the sequence into multiple linear representations, with all segments and their respective coordinates identified. These include 5′ and 3′ untranslated regions (5′ UTR, 3′ UTR), coding sequence (CDS), triplet codons, translations, exon segments, polyA signals and tails, polymorphisms, insertions and deletions. Additionally, hidden Markov models (HMM) for the ADHD peptide domain are being incorporated into this section, as well as the graphical representation of the peptide sequence further to strengthen the characterisation of the transcript and peptide sequence analysis.

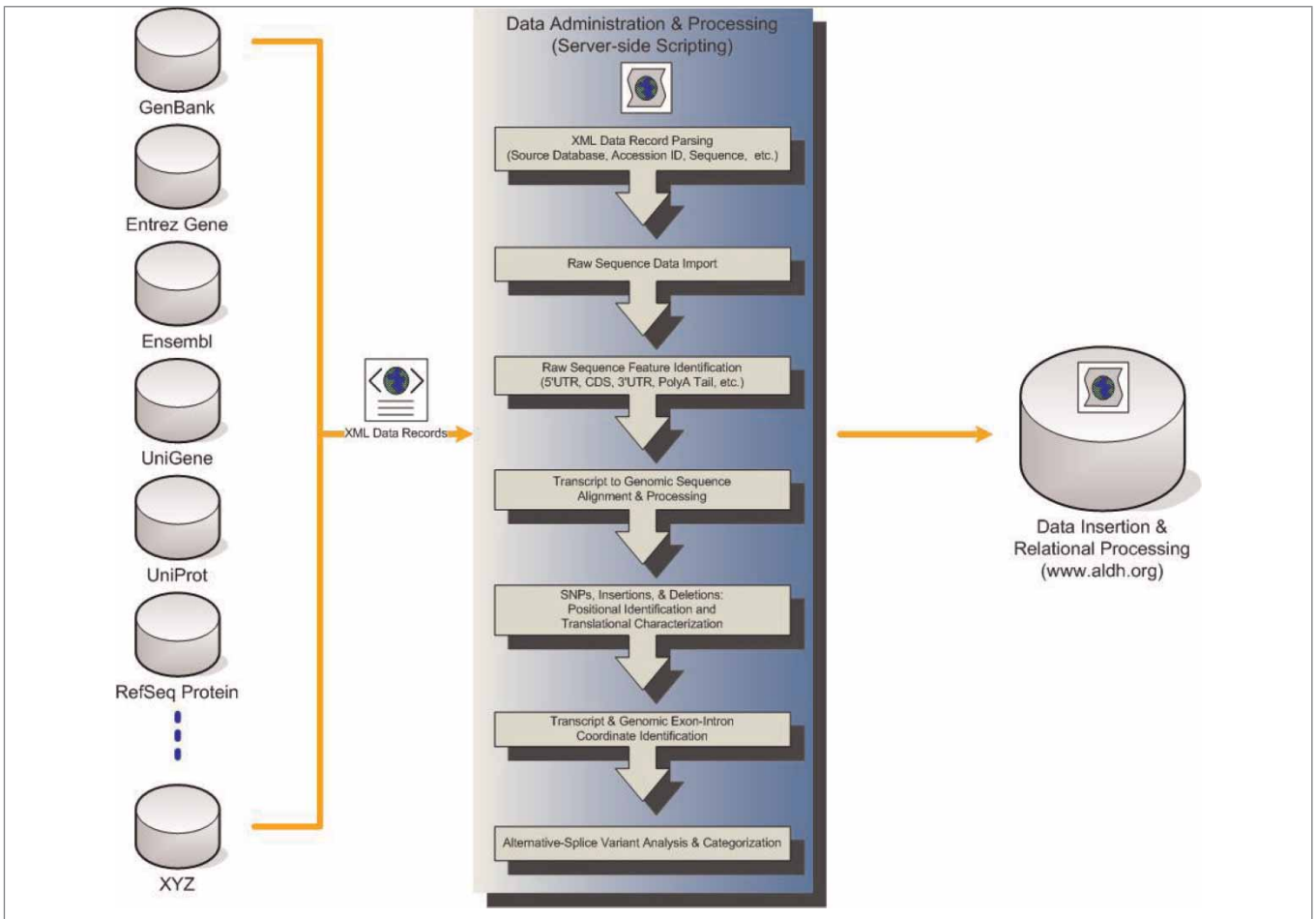

Figure 4. Data mining and processing for www.aldh.org. Site administrators collect specific data elements from major database repositories in the form of site-generated XML documents. These XML documents are then incorporated, parsed and inserted into the www.aldh.org database architecture by the server-side administrative script engines. The newly inserted data are then readily available for web server processing and display to the public.

Substrates, inhibitors and enzyme kinetics

Characterisation of *ALDH* gene superfamily members typically includes their enzymatic activity for an assortment of substrates and inhibitors using various test systems, isolation techniques, cofactors, tissues and species origin. This module provides a tabular summary of the literature for all reported kinetics values, sorted by species. Each entry includes the hyperlinked PubMed identifier to the original literature source for quick accessibility.

Reference list

References to the original sources for all data within the www.aldh.org database are an important priority for the *ALDH* gene superfamily curators. The reference list module provides a bibliography for all data reported within each gene-specific record, including those from the 'Synopsis and substrates', and 'Inhibitors and enzyme kinetics' modules, as well as additional references recommended by the curators. All references within this section are hyperlinked to the PubMed database via the PMID for ease of access.

Data mining and processing

All molecular sequence information, literature reviews, human polymorphisms and enzyme kinetics for www.aldh.org are developed from multiple publicly available databases. The data are either referenced to their source database records for each accession or to the literature abstracts via the PubMed identifier. All molecular data are handled using the administrative software modules, which pull pre-specified data from these public databases in the form of XML file records. The XML records are then parsed, analysed, processed and inserted into the ALDH web database through a series of functions using the administrative software modules (Figure 4). These data are then readily available for display on the website through the server-side scripting, as previously described.

Future directions

The www.aldh.org website is an ongoing project which enables the Vasiliou laboratory to maintain a living review of all *ALDH* genes, as well as provide detailed functional and molecular sequence analysis to the public. New *ALDH* genes and existing *ALDH* gene orthologues are continually being added as genome projects are completed. This site's identification and characterisation of the *ALDH* gene superfamily members will provide investigators with a degree of consistency in terms of HUGO-approved nomenclature[4] as they report their findings in the future. The server-side data modules producing the website are frequently updated for increased performance or additional data analysis, and new modules are designed and produced as an area of interest becomes apparent. At the time of writing, the authors are designing four additional gene record modules to address and characterise: (a) human polymorphisms from the NCBI dbSNP database; (b) subcellular localisation and tissue expression profiles; (c) upstream binding elements and promoters for each gene member; and (d) incorporation of the AmiGO gene ontology features. Lastly, it is our hope that users find this site helpful in their search for information and that investigators embarking on future gene superfamily initiatives utilise our site's structure and format in reporting such data to the web. The authors welcome all feedback regarding the website, as well as any ideas for new data modules that we have not yet addressed.

Acknowledgments

We thank our colleagues for valuable discussions and a careful reading of this manuscript. This work was supported, in part, by NIH grants EY17963 and AA017754.

References

1. Maglott, D., Ostell, J., Pruitt, K.D. *et al.* (2007), 'Entrez Gene: Gene-centered information at NCBI', *Nucleic Acids Res.* Vol. 35, pp. D26–D31.
2. Hubbard, T.J., Aken, B.L., Ayling, S. *et al.* (2009), 'Ensembl 2009', *Nucleic Acids Res.* Vol. 37, pp. D690–D697.
3. Bult, C.J., Eppig, J.T., Kadin, J.A. *et al.* (2008), 'The Mouse Genome Database (MGD): Mouse biology and model systems', *Nucleic Acids Res.* Vol. 36, pp. D724–D728.
4. Vasiliou, V., Bairoch, A., Tipton, K.F. *et al.* (1999), 'Eukaryotic aldehyde dehydrogenase (ALDH) genes: Human polymorphisms, and recommended nomenclature based on divergent evolution and chromosomal mapping', *Pharmacogenetics* Vol. 9, pp. 421–434.