BMC Bioinformatics

Software

Djinn Lite: a tool for customised gene transcript modelling, annotation-data enrichment and exploration

Erdahl T Teber[1,4,5], Edward Crawford[1], Kent B Bolton[2], Derek Van Dyk[3], Peter R Schofield[4,6], Vimal Kapoor[1,7] and W Bret Church*[1,4,5]

Address: [1]School of Medical Sciences, University of New South Wales NSW 2052, Australia, [2]EBM Pty Ltd, Level 6, 110 Sussex Street, Sydney, NSW 2000, Australia, [3]NSW Ministry for Science and Medical Research, GPO Box 5341, Sydney NSW 2001, Australia, [4]Neurobiology Division, Garvan Institute of Medical Research, Sydney NSW 2010, Australia, [5]Faculty of Pharmacy, University of Sydney NSW 2006, Australia, [6]Prince of Wales Medical Research Institute, Sydney NSW 2031, Australia and [7]Department of Medicine and Pharmacology, University of Western Australia, Crawley WA 6009, Australia

Email: Erdahl T Teber - e.teber@pharm.usyd.edu.au; Edward Crawford - e.crawford@unsw.edu.au; Kent B Bolton - kent@bolton.cc; Derek Van Dyk - derek.vandyk@msmr.nsw.gov.au; Peter R Schofield - p.schofield@unsw.edu.au; Vimal Kapoor - Vimal.Kapoor@uwa.edu.au; W Bret Church* - b.church@pharm.usyd.edu.au

* Corresponding author

Abstract

Background: There is an ever increasing rate of data made available on genetic variation, transcriptomes and proteomes. Similarly, a growing variety of bioinformatic programs are becoming available from many diverse sources, designed to identify a myriad of sequence patterns considered to have potential biological importance within inter-genic regions, genes, transcripts, and proteins. However, biologists require easy to use, uncomplicated tools to integrate this information, visualise and print gene annotations. Integrating this information usually requires considerable informatics skills, and comprehensive knowledge of the data format to make full use of this information. Tools are needed to explore gene model variants by allowing users the ability to create alternative transcript models using novel combinations of exons not necessarily represented in current database deposits of mRNA/cDNA sequences.

Results: Djinn Lite is designed to be an intuitive program for storing and visually exploring of custom annotations relating to a eukaryotic gene sequence and its modelled gene products. In particular, it is helpful in developing hypothesis regarding alternate splicing of transcripts by allowing the construction of model transcripts and inspection of their resulting translations. It facilitates the ability to view a gene and its gene products in one synchronised graphical view, allowing one to drill down into sequence related data. Colour highlighting of selected sequences and added annotations further supports exploration, visualisation of sequence regions and motifs known or predicted to be biologically significant.

Conclusion: Gene annotating remains an ongoing and challengingtask that will continue as gene structures, gene transcription repertoires, disease loci, protein products and their interactions become moreprecisely defined. Djinn Lite offers an accessible interface to help accumulate, enrich, and individualise sequence annotations relating to a gene, its transcripts and translations. The mechanism of transcript definition and creation, and subsequent navigation and exploration of features, are very intuitive and demand only a short learning curve. Ultimately, Djinn Lite can form the basis for providing valuable clues to plan new experiments, providing storage of sequences and annotations for dedication to customised projects. The application is appropriate for Windows 98-ME-2000-XP-2003 operating systems.

Background

There remains intense interest in the annotation of genomes, and work is continuing to be done to elucidate all of the human genes, its haplotypes [1,2] and its transcriptome [3,4]. It is estimated approximately 50% of the human transcriptome is not yet annotated [3]. According to the CCDS project, as at March, 2005, just over 13,000 genes can be reliably said to code for protein [5], which represents approximately half of the human genes [6]. Inconsistencies in gene annotations arise due to different human genome sequence database centers employing different methods for identifying the locations of genes and generating coding sequences [5], specifically from different computational methods and gene-finding programs. "Ab initio" gene finding programs detect genes by looking for distinct patterns that define where a gene begins and ends. Alternatively, comparative gene finding programs look for genes by comparing segments of sequence with those of known genes and proteins [6]. Current gene prediction algorithms also focus largely onpredicting coding regions and less on untranslated regions [7,8]. Gene finding computational methods alone are simply insufficient to generate accurate gene structures. Providing accurate gene annotations will take the coordinated efforts of experimentalists and computational biologists to learn from the inconsistencies between gene structures generated by manual curation and automated methods [9].

Also, limited numbers of tissue-specific EST and mRNA sequences deposited into public databases, as well as different cDNA construction protocols may miss tissue-specific transcript constructs [10]. Again, biological experiments may be necessary to confirm any transcript constructions, but programs which can suggest best-guesses and ranges of transcript options could be extremely valuable.

The need for manual gene curating is also necessitated due to errors and inaccuracies that may exist in the main sequence databases [11,12]. It is not uncommon for different sources to give different names to the same regions (e.g. Exons), particularly between literature and genome database sources (e.g. PTGS1/COX1 gene) [13-15]. These situations make manual curating all the more important, as biologists attempt to resolve inconsistencies.

Table 1: A sample of sources available on experimentally, computationally derived gene annotations and sequence pattern discovery programs

Annotation Type	Source	Reference
Alternative splicing	ASD http://www.ebi.ac.uk/asd/	[32]
	AltSplice and AltExtron – Computationally derived splicing events.	
	AEdb – splicing events manually generated from the literature.	
	Swiss-Prot/TrEMBL http://au.expasy.org/sprot/	[35-37]
Polymorphic sequence variations	dbSNP http://www.ncbi.nlm.nih.gov	[38]
Predicting RNA secondary structures	Mfold http://www.bioinfo.rpi.edu/applications/mfold/	[34]
Predicting protein motif/domains/structural similarities	SMART http://smart.embl-heidelberg.de/	[39]
Human gene mutations	OMIM http://www.ncbi.nlm.nih.gov/	[40]
	HGMD http://www.hgmd.cf.ac.uk	[41]
Predicting transcription factor binding sites	MAPPER http://bio.chip.org/mapper	[42]
Predicting exonic splicing enhancer sites	ESEFinder http://rulai.cshl.edu/tools/ESE	[43]
	RESCUE-ESE http://genes.mit.edu/burgelab/rescue-ese/	[44]
Predicting promoter elements	SIGNAL SCAN http://thr.cit.nih.gov/molbio/signal/	[45]
	FirstEF http://rulai.cshl.org/tools/FirstEF/	
Post-translational modifications	Swiss-Prot/TrEMBL http://au.expasy.org/sprot/	[35-37]
Predicting post-translational modifications	NetAcet – N-terminal acetylation in eukaryotic proteins.	[46]
	NetNGlyc – N-linked glycosylation sites in human proteins.	[47]
	NetOGlyc – O-GalNAc (mucin type) glycosylation sites in mammalian proteins.	[48]
	NetPhos – Serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins.	[49]
	ProP – Arginine and lysine propeptide cleavage sites in eukaryotic protein sequences.	
	YinOYang – O-(beta)-GlcNAc glycosylation and Yin-Yang sites (intracellular/nuclear proteins).	
	All above web server programs can be found at http://www.cbs.dtu.dk/services/	
Protein-protein interaction data	InterDom http://interdom.lit.org.sg/	[50]
	BIND http://bind.ca/	[51]
	DIP http://dip.doe-mbi.ucla.edu/	[52]
	MINT http://mint.bio.uniroma2.it/mint/	[53]

Some examples of web based sources providing free sequence related annotation data and "biologically significant" sequence pattern discovery programs. Most of these sources can provide specific annotations upon inputting of a gene/transcript/protein sequence, or the name of the gene/protein, along with the source organism. It is the thesis of our work that care needs to be taken to assess the quality of all sources of information. This is not intended to be an exhaustive list, but highlights sources used for the examples included in this paper.

In the past 30 years, genetic studies of multifactorial human diseases have identified approximately 50 genes and their allelic variants. However, it is estimated that there are likely to be hundreds of susceptibility loci that increase the risk for each common disease [16]. Therefore, annotating of genes remains ongoing and represents a challenging task that will be driven forward continually as the human gene structures and disease loci become more precisely defined. Current efforts are underway to develop a haplotype map of the human genome [1,2]. The haplotype map or "HapMap" aims to provide researchers with information to find genes and genetic variations that affect health and disease. Manual annotators will play a critical role cataloguing how different components interact and contribute to biological processes, diseases and physiological complexity. Thus, investigators require tools to be able to store, test and analyse the combinations of alleles experimentally observed to be inherited as units from DNA/polymorphism screenings.

It is estimated that 75% of alternative splicing events change the protein coding sequence [17,18]. It is thought that approximately a third to half of all human genes produce multiple transcript variants [19]. Alternative splicing can often produce protein isoforms with different domain compositions and motifs [20]. It will be vital to use tools to be able to model variant transcripts to test for splice-site plasticity and disease forming missplicing events [21].

Along with our understanding of alternate splicing events and post-translational modification it is likely that biologists with expertise in specific genes will continue to curate annotations and will benefit from "clue" providing visualisation tools to perform targeted experiments on regions of eukaryotic genomes. As sequences are accumulating at a great rate, biologists are also required to assimilate sequence related information from many diverse sources (Table 1). Moreover, putative or customised annotations are always required by individual laboratories, the implications of which can be further tested.

Comparisons
A number of sophisticated and powerful sequence annotation and visualisation tools are available including ARTEMIS [22], SeqVISTA [23], and Genotator [24,25]. These tools principally focus on features that are related to segments of nucleotide sequences or small genomes, translated amino acid sequences and their annotations. Many of the features that are necessary for visualising sequences such as ease of navigation, colour coding, and dynamic linking of macro level depictions with detailed sequences exist in these programs. They also contain additional integrated functionality which can be useful to experienced bioinformaticians, including exon predict-

ing, dbEST searches, protein secondary structure predictions and others [22-24].

Genome browsers on the other hand, are suited to large scale annotation and analysis of genomes and include UCSC's Genome Browser [26], ENSEMBL project viewer [27], NCBI Map Viewer and GeneViTo [28]. UCSC's Genome Browser can display requested portions of genomes by zooming in/out to any scale, together with many aligned annotation tracks, including known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, cross-species homologies, and tracks that have been deposited by others. Users can add and view their own custom tracks, however, this can require users to place annotations in formatted files before uploading into the browser.

NCBI Human Map viewer also has an additional function called Model Maker which is able to show the exons provided by GenBank, mRNAs, ESTs, and gene predictions. However, the numbering and alignments between transcript and genomic sequence, between transcript and protein are non-intuitive, requiring jumping between screens to obtain position number associations. Another key limitation of web-based visualisation and annotation tools is the available printing options. These are limited to printing only what is available on the page or images, and can make copying/pasting sequences and alignments cumbersome.

The number and type of annotations can vary and arranging annotations in a non-confusing manner is paramount for non-bioinformatic conversant biologists. Ultimately, excessive functionality, formatting of input files, genome wide analysis, and inflexible printing can be overwhelming for biologists whose key focus would be to judiciously conduct wet laboratory experiments on their gene of interest. Thus, unneeded complexities related to gene annotations need to be hidden from view, and software tools need to be less complicated in an effort to help in integrating, storing and visualising annotations as biologists gradually learn more about their gene of interest.

We have developed Djinn Lite for those users not requiring significant bioinformatics skills in customising gene/transcript annotations. The application is appropriate for Windows 98-ME-2000-XP-2003 operating systems.

Implementation
Input sequence, defining transcripts and coding regions
Djinn (pronounced jeen or jēn) Lite has been designed in keeping with a novice gene annotator's general workflow. This is facilitated by the use of "tabs" (analogous to dividers in a notebook or the labels in a file cabinet). A single application window contains six (6) tabs each corre-

sponding to a single page, these are: "View Sequences", "Nucleotide Regions", "Transcript Design", "Nucleotide Annotations", "Protein Annotations", "Sequence Reports", and "Graphical View" (Figure 1).

Djinn Lite invokes a wizard to allow for the input of a raw nucleotide sequence. The raw nucleotide sequence can be genomic, pre-mRNA, mRNA or partial or complete protein coding region (CDS) sequences. For example, raw nucleotide sequences can be obtained from any sequence database including NCBI Entrez Gene [14], Ensembl [15], GeneCards [29] and Celera [30] and then cut-and-paste into Djinn Lite's main sequence input text form as part of the initial wizard.

Djinn Lite uses the term "Nucleotide Regions" to describe core transcriptional regions, such as exons, 5' and 3' untranslated regions. Upon the input of a main sequence key regions may be defined by either providing a start and end nucleotide position or selecting the nucleotides using the mouse by click-and-dragging. Alternatively, regions may be assigned after carrying out a nucleotide string search match. Textual information describing the source of the information may be added to a "Reference" entry field pertaining to a defined "Nucleotide Region".

Transcripts can be generated by selecting from a list of previously defined transcriptional regions referred to as "Nucleotide Regions". When an mRNA transcript is created, a check box can be used to avoid translating of flanking regions such as 5' and 3' UTRs. The checkbox by default remains checked to allow the entire construct to be translated as in the case of a coding sequence (CDS).

Feature annotation and colour highlighting

Annotating and colour highlighting of particular nucleotide or translated protein sequence can be carried out by either conducting a sequence search or by providing a sequence start and end position. Adding annotations using the "Highlight Nucleotides" enables users to carry out searches against a list of previously defined transcripts, CDSs, including the main input sequence (e.g. genomic sequence). There are 15 colours to choose from for the highlighting of sequence regions and macro level depictions (a bar that displays colour annotations over graphical box representations of the gene, transcripts, and proteins). Thus, colour highlights can be overlaid on top of sequence regions which can include; 5' transcriptional control elements, promoters, translational control elements, start and stop codons, 3' polyadenylation signals, binding sites for transcription factors, splicing enhancer/silencing elements, polymorphic variants, SNPs, mutations, microRNA and small interfering RNA (siRNA) targets, RNA editing sequences, protein domains, motifs and protein binding regions, PCR covered regions, putative

regions or regions requiring experiment validation (Figure 1 and Table 1). Users have the discretion and flexibility to make any annotation of their choosing. Also, grey colour highlights are automatically generated by Djinn Lite where there are overlapping colour coded annotations along a sequence. Each annotation also enables the attachment of a textual description which appears in the legends in both the "Sequence Reports" and "Graphical View" pages.

Graphical representation and viewing sequences

Djinn Lite uses multiple rows or tracks [31] for handling the complexity of genomic sequence annotations enabling numerous annotations and incorporates a multiple dimensional data space (sequence, transcript regions, features). The "View Sequences" tab is the "working display" window showing the sequence(s) (base pairs) in their entirety viewable by horizontal scrolling bar.

The user can toggle between the macro view identified as "Graphical Overview" and the "View Sequence". The "Graphical Overview" provides the user with a high level picture representation of the gene, transcripts and protein.

The "Graphical Overview" was designed to be particularly useful for gaining an overview of the physical size of the gene map and its associated transcripts, in terms of the relative sizes of the introns, exons, and the density of features along these maps. The sizes of the transcripts and their associated proteins are scaled relative to each other. Thus, an inspection of key global differences between transcripts can provide clues to dissimilarities in transcriptional regions and protein domains or motifs.

Within the "Graphical Overview" there are two sections, the genomic view and the transcripts view. Within the genomic view boxes represent exons, narrow lines represent introns or non-genic regions, and below is an annotation ruler designed to display colour bars to assist in featured annotations alongside their corresponding relative locations along the genomic map. Djinn Lite is also able to depict "overlapping exons" and "overlapping untranslated regions", as can occur due to the plasticity of splice-site selection [21], as dark green coloured boxes. The transcripts view depicts all of the defined transcripts, where boxes represent untranslated, exons or coding regions. An annotation ruler displays colour bars to feature annotations alongside their corresponding relative location along the transcript. Also, the translated transcript is displayed as an outlined box and can be overlaid with colour code bars to correspond to annotations relative to the protein.

More critical inspection of features (regions and annotations) is facilitated by allowing for flexible viewing com-

Figure 1
Tab designed workflow of Djinn Lite. Prostaglandin-endoperoxide synthase (PTGS1) plays a major role in prostaglandin biosynthesis. Commonly referred to as Cyclooxygenase 1 (COX1), PTGS1 is a target for aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs), in particular, for reducing platelet aggregation (GenBank: NT_008470 Chromosome 9 genomic contig – nucleotides 32,450,650-32,481,650). This figure displays a section of the nucleotide annotations page showing a customised list of PTGS1 gene annotations, including predicted promoter elements, transcription factor binding regions, enhancer splicing elements and known SNPs.

binations via zooming and horizontal scroll bars which are located at the bottom of every genomic and transcript graphical representation. Clicking on any of the graphically displayed boxes (exon, untranslated regions) or narrow lines (non-genic or intronic regions) will invoke the "View Sequences" tab and call up in real-time the sequence corresponding to the selected region (Figure 2).

The user is able to maintain context between the "Graphical Overview" and "View Sequences", as the "View Sequences" tab, either displays transcript regions in context to its genomic DNA or a protein in context to its transcript. This allows the user to view sequence alignments and numbering of nucleotide and amino acid positions in an integrated context.

Previewing/printing/exporting sequences and picture representations
The user has two main options for the printing of annotations. These are previewing or printing of high level picture representations and the detailed sequences and annotations. The first option, the "Graphical View" tab, has the same functionality as the "Graphical Overview" window, but instead enables printing of high level picture representations of genomic, pre-mRNA, mRNA, partial or

complete protein coding region (CDS) sequences, together with the colour annotations and legends (Figure 3). The size and viewing pose of each picture object (gene, transcripts, CDSs) can be varied by clicking on any part of the object and then by zooming/scrolling. The current viewing pose can then be copied/pasted for reporting purposes.

The second option, the "Sequence Reports" tab, enables printing of sequences alongside aligned regions, amino acid sequences, with colour-codes overlaid on the sequences. This results in a table representation of the sequences, including a legend for the annotations at the end of the report. Printing is context specific, i.e. transcript regions are aligned in context to its genomic DNA or optionally a protein is aligned in context to its transcript. The width of the table can be modified, thus allowing users to have a large range of sequence lengths to aid their viewing requirements. It is then possible to copy/paste, which provides for report writing and ultimately for publication purposes.

Djinn Lite also allows for exporting of gene, transcript, CDS, transcriptional nucleotide regions (exons and untranslated regions) and protein sequences in FASTA for-
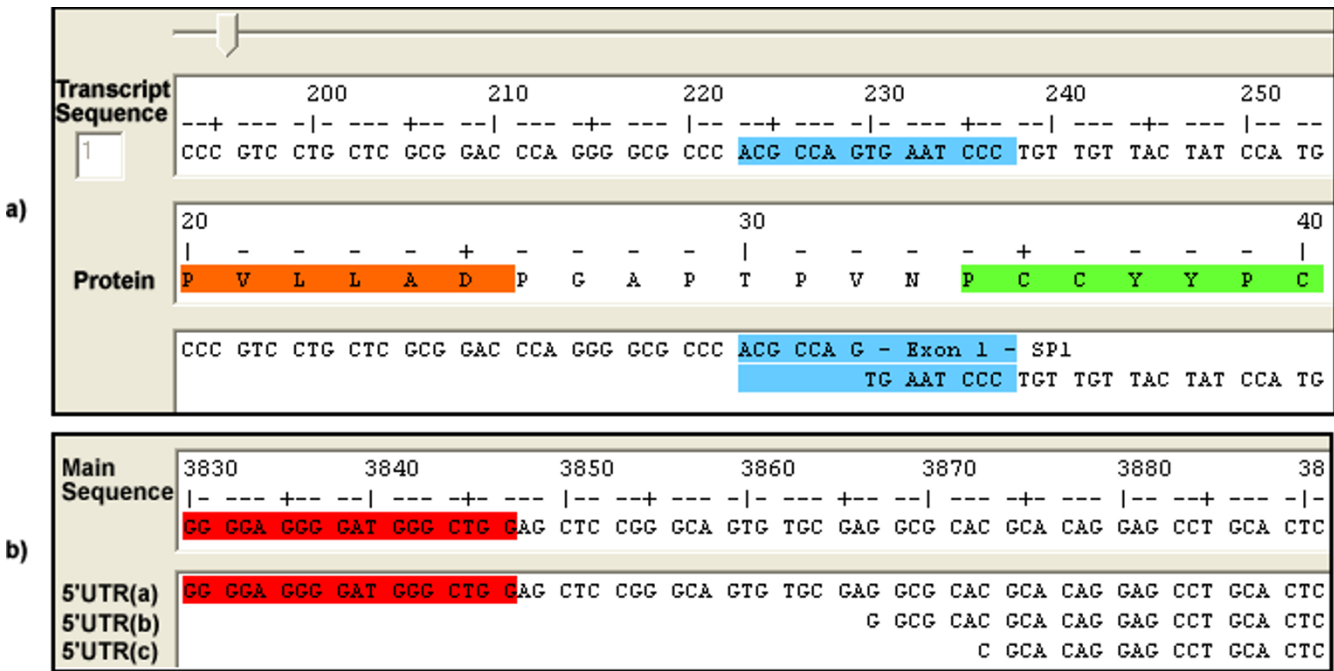
Figure 2
Viewing sequences and position numbering of associated regional segments. a) The Exon 1/Exon 2 splice junction of transcript 1 (SP1) alongside the translated sequence. Annotated regions have the colour highlights. Colours red, green, and aqua, respectively, refer to a putative signal peptide, an Epidermal Growth Factor domain, and the splice site junction between Exon 1/Exon2. b) The three varieties of 5'UTRs ([GenBank: NM_080591.1] and two computationally derived alternate splice variants from AltSplice [32]) aligned to the PTGS1 gene sequence (main sequence). Alternate 5' sequences can occur due to differential regulation of upstream promoters and splicing factors, giving rise to different transcription-start sites and splice donor/or acceptor sites [21, 33]. Mfold program [34] predicts a hair-pin RNA secondary structure within 5' UTR(a), represented in red. Secondary structures in the 5'UTR can modulate translation efficiency.

mat. All accumulated sequences, transcript models and annotations can be saved as a text formatted file. This text file can be easily copied into Microsoft Excel for further manipulation or analysis.

Software design limitations
Once a main sequence entry has been initially processed in Djinn Lite the nucleotide sequence state remains static and subsequent nucleotide changes (addition/deletions/substitutions) within a saved gene configuration are not allowed. Thus, Djinn Lite is not designed for automatic updating of downstream gene product sequences when alterations are made to the inputted nucleotide sequence. Real-time changing of nucleotides at the main sequence level would be a useful feature in helping to observe the effects of nucleotide changes at downstream levels, including changes to splicing, domains/motifs, and amino acid changes. This would force the program to respond to a multitude of subsequent effects, including changes to regions and transcript variants. This feature was not implemented in real time to prevent the user interface from becoming too complex, as it could poten-

tially yield multiple user notifications to highlight many of the subsequent downstream sequence alterations. However, a separate Djinn session and file can be set up to accommodate for different sequence variations of a gene or sequence entry. For example, Djinn Lite can be used for haplotype mapping, where each distinct haplotype (distinct set of polymorphic variations which are inherited as a unit) can be set up as a separate Djinn file.

Again, in an effort to maintain the simplicity of the Djinn Lite's user interface, some features which are biologically relevant to mRNA were not implemented. We believed that these features were not critical to the main emphasis and utility of Djinn Lite, which was the ease of use, uncomplicated transcript modeling, annotating and visualising. For example, thymine (T) is not replaced for uracil (U) when DNA is transformed to RNA. This aspect may be more important for programs that provide RNA secondary structural analysis, as uracil and thymine have different hybridization properties. This was not part of Djinn Lite's design scope. Djinn Lite avoids accommodating for addition of multiple adenosine nucleotides onto the 3' end of
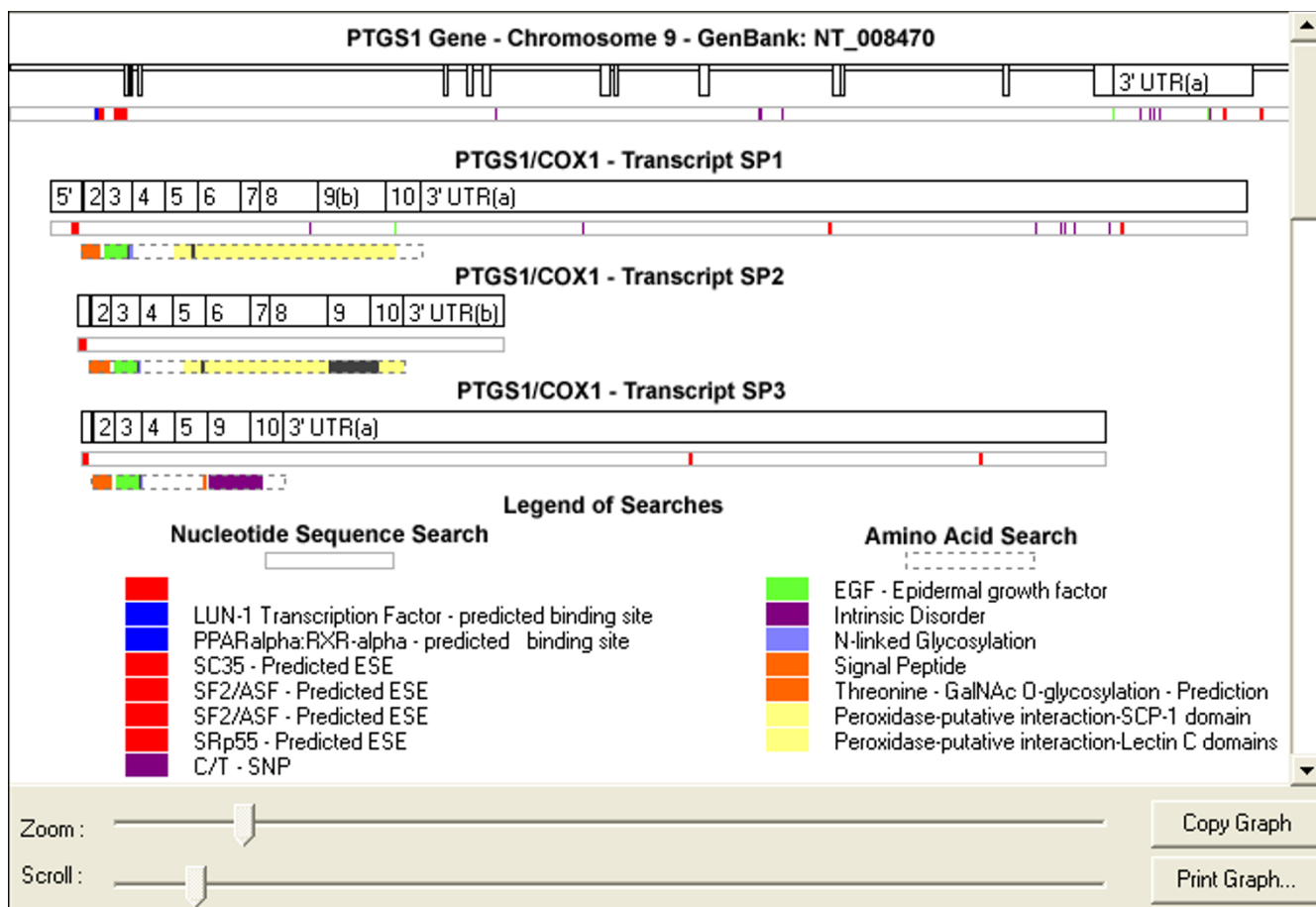
Figure 3
Graphical Representation. A representation of the PTGS1 gene (approx. 30 kb) and three (3) modelled gene products, transcript SP1 to SP3. The sizes of the transcripts and their associated proteins are scaled relative to each other, providing visual insights into the differences between transcripts, and clues to dissimilarities in transcriptional regions and protein domains or motifs. Boxes represent exons and untranslated regions and narrow lines represent introns or non-genic regions. An annotation ruler displays colour bars that feature annotations alongside their corresponding relative location along the transcript. Translations of transcripts are displayed as outlined boxes and are overlaid with colour code bars of annotations relative to the protein. Annotation legends are depicted at the bottom.

defined transcripts (polyadenylation) for subsequently defined and transformed downstream sequences. Likewise, RNA editing such as substitution or deletion or insertion editing are avoided. Also, Djinn Lite does not permit loose sequence alignments.

Conclusion

Djinn Lite represents a tool for the process of "annotation data enrichment", which involves the incremental gathering, qualifying and experimental verification of both putative and documented gene sequence annotations. Djinn Lite provides the ability to display annotations, easy to follow numbering of aligned sequences, creation of alternative transcript models using novel combinations of exons, as well as offer flexible printing options for annotated sequences and gene/transcript maps. The inter-

face is intuitive, requiring only a short learning curve, helping to quickly accumulate and individualise sequence information on genes and their flow on products, including sequence annotations relating to transcriptional/ translational regulation, post translational modifications and protein interactions. Djinn Lite can provide storage of gene annotations for personalised projects on particular genes of interest and therefore be the basis of valuable clues to plan new experiments so that the needs of biologists whose key concern is to judiciously plan and conduct experiments are met. Ultimately, extensive use of such a tool can help to improve the accuracy and comprehensiveness of genome wide annotations. Additionally, Djinn Lite may be a useful teaching aid to support the learning of undergraduate students on topics related to gene structure.

Availability and requirements

Project name: **Customised gene transcript modeling, annotating and exploring**

Project name home page: **http:// www.sbio.pharm.usyd.edu.au/DjinnLite**

Operating system: **Microsoft 98/ME/2000/XP/2003**

Programming language: **Visual Basic Version 6**

Other requirements: **None**

Licenses: **Executable is freeware**

Any restrictions to use by non-academics: **None**

Authors' contributions

ETT was responsible for the original concept and design, programming, testing and drafting of the manuscript; KBB programmed at the initial stage to create a viable program; EC was involved in design improvements of the user interface and graphics, a substantial overhaul of the data structures, and programming; DVD was involved in the design improvements; PRS provided experimentalist perspective on the biology of alternate splicing; VK provided experimentalist perspective on the biology of alternate splicing and testing; WBC was responsible for drafting of the manuscript, design improvements and substantial testing.

References

1. Foster MW, Sharp RR: **Beyond race: towards a whole-genome perspective on human populations and genetic variation.** *Nat Rev Genet* 2004, **5:**790-796.
2. **The International HapMap Project.** *Nature* 2003, **426:**789-796.
3. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.
4. Yeo G, Holste D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues.** *Genome Biol* 2004, **5:**R74.
5. **Consensus CDS project - European Bioinformatics Institute (EBI), National Center for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI), and the University of California, Santa Cruz (UCSC).** 2005 http://www.ncbi.nlm.nih.gov/projects/CCDS .
6. **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
7. Ashurst JL, Collins JE: **Gene annotation: prediction and testing.** *Annu Rev Genomics Hum Genet* 2003, **4:**69-88.
8. Dike S, Balija VS, Nascimento LU, Xuan Z, Ou J, Zutavern T, Palmer LE, Hannon G, Zhang MQ, McCombie WR: **The mouse genome: experimental examination of gene predictions and transcriptional start sites.** *Genome Res* 2004, **14:**2424-2429.
9. Pennisi E: **Bioinformatics. Gene counters struggle to get the right answer.** *Science* 2003, **301:**1040-1041.
10. Gupta S, Zink D, Korn B, Vingron M, Haas SA: **Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing.** *BMC Genomics* 2004, **5:**72.
11. Wesche PL, Gaffney DJ, Keightley PD: **DNA sequence error rates in Genbank records estimated using the mouse genome as a reference.** *DNA Seq* 2004, **15:**362-364.
12. Karlin S, Bergman A, Gentles AJ: **Genomics. Annotation of the Drosophila genome.** *Nature* 2001, **411:**259-260.
13. Hillarp A, Palmqvist B, Lethagen S, Villoutreix BO, Mattiasson I: **Mutations within the cyclooxygenase-1 gene in aspirin non-responders with recurrence of stroke.** *Thrombosis Research* 2003, **112:**275-283.
14. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2005, **33:**D54-8.
15. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33:**D447-53.
16. Wang WY, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns.** *Nat Rev Genet* 2005, **6:**109-118.
17. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome.** *Genome Res* 2003, **13:**1290-1300.
18. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
19. Lander ES, Linton LM, Birren B, C. N, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
20. Loraine AE, Helt GA: **Visualizing the genome: techniques for presenting human genome data and annotations.** *BMC Bioinformatics* 2002, **3:**19-26.
21. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing.** *Gene* 2005, **344:**1-20.
22. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
23. Hu Z, Frith M, Niu T, Weng Z: **SeqVISTA: a graphical tool for sequence feature visualization and comparison.** *BMC Bioinformatics* 2003, **4:**1.
24. Harris NL: **Annotating sequence data using Genotator.** *Mol Biotechnol* 2000, **16:**221-232.
25. Harris NL: **Genotator: a workbench for sequence annotation.** *Genome Res* 1997, **7:**754-762.
26. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler

D, Kent WJ: **The UCSC Genome Browser Database.** *Nucl Acids Res* 2003, **31**:51-54.
27. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T: **Ensembl 2004.** *Nucleic Acids Res* 2004, **32**:D468-70.
28. Vernikos GS, Gkogkas CG, Promponas VJ, Hamodrakas SJ: **GeneV-iTo: visualizing gene-product functional and structural features in genomic datasets.** *BMC Bioinformatics* 2003, **4**:53.
29. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D: **GeneCards 2002: towards a complete, object-oriented, human gene compendium.** *Bioinformatics* 2002, **18**:1542-1543.
30. Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, Mayberry F, Mural R, Nodell M, Yandell M, Zhang J, Thomas P: **The Celera Discovery System.** *Nucleic Acids Res* 2002, **30**:129-136.
31. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The Human Genome Browser at UCSC.** *Genome Res* 2002, **12**:996-1006.
32. Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Le Texier V, Muilu J: **ASD: the Alternative Splicing Database.** *Nucleic Acids Res* 2004, **32**:D64-9.
33. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biol* 2002, **3**:REVIEWS0004.
34. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31**:3406-3415.
35. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33**:D154-9.
36. Wu C, Nebert DW: **Update on genome completion and annotations: Protein Information Resource.** *Hum Genomics* 2004, **1**:229-233.
37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**:365-370.
38. Smigielski EM, Sirotkin K, Ward M, Sherry ST: **dbSNP: a database of single nucleotide polymorphisms.** *Nucl Acids Res* 2000, **28**:352-355.
39. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucleic Acids Res* 2004, **32**:D142-4.
40. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**:D514-7.
41. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN: **Human Gene Mutation Database (HGMD): 2003 update.** *Hum Mutat* 2003, **21**:577-581.
42. Marinescu VD, Kohane IS, Riva A: **MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes.** *BMC Bioinformatics* 2005, **6**:79.
43. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR: **ESEfinder: a web resource to identify exonic splicing enhancers.** *Nucl Acids Res* 2003, **31**:3568-3571.
44. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297**:1007-1013.
45. Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **5**:923-932.
46. Kiemer L, Bendtsen JD, Blom N: **NetAcet: prediction of N-terminal acetylation sites.** *Bioinformatics* 2005, **21**:1269-1270.
47. Julenius K, Molgaard A, Gupta R, Brunak S: **Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites.** *Glycobiology* 2005, **15**:153-164.
48. Blom N, Gammeltoft S, Brunak S: **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.** *J Mol Biol* 1999, **294**:1351-1362.
49. Duckert P, Brunak S, Blom N: **Prediction of proprotein convertase cleavage sites.** *Protein Eng Des Sel* 2004, **17**:107-112.
50. Ng SK, Zhang Z, Tan SH, Lin K: **InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes.** *Nucleic Acids Res* 2003, **31**:251-254.
51. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, Hogue CW: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33**:D418-24.
52. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**:D449-51.
53. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Letters* 2002, **513**:135-140.