

The Celera Genome Browser: Visualizing and Annotating the Human Genome



Russell Turner

*Johns Hopkins University
Applied Physics Laboratory*

James Baxendale

Synaptic Science, LLC

Peter Davies

*Howard Hughes
Medical Institute*

Patrick Dunn

*Northrop Grumman
IT Health Solutions*

Leslie Foster

*J. Craig Venter
Institute*

Todd Safford

*J. Craig Venter
Institute*

Scott Henderson

Celera

Genome Annotation

In order to interpret an organism's raw genomic sequence, the location of its biologically meaningful features such as genes must be annotated.

▶ Computational Annotation

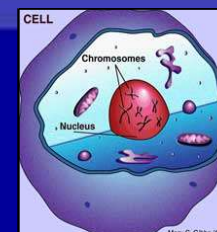
- Automated searches against sequence databases
- Gene prediction algorithms

▶ Manual Curation

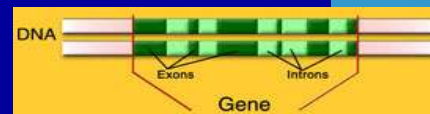
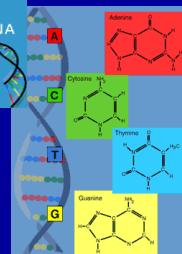
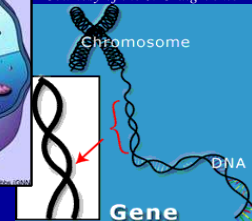
- Human expert biologists review evidence
- Determine definitive gene structure

▶ Interactive Visualization Tools

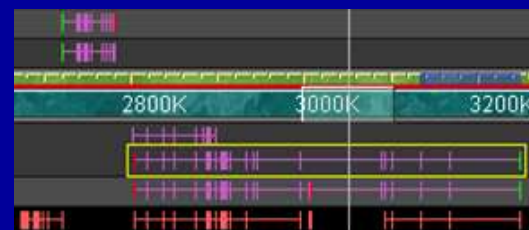
- Visual analysis of genomic data
- Creation and editing of gene annotations



Courtesy of the J. Craig Venter Institute



Frame +3	C	A	C	T	T	T	A	A	G	C	G	T	C	A	C	G	G	G	T	G	G	G	C
0																							
Frame +3				S	R				L	K				R	I								
52	C	T	T	C	G	C	G	C	T	G	A	G	C	G	G	A	T	A	G	C	G		
Frame +3	A	G		G	R				D	C	G	T	S										
104	G	C	A	G	G	T	G	G	A	G	A	G	A	C	T	G	C	G	G	T	A	C	C
Frame +3				R	K		S			G	Q	L											
156	C	G	A	G	G	A	A	G	A	G	C	G	G	C	A	G	T	T	A	A	C	C	
Frame +3				Q	C		V			W	R	I											
208	G	C	A	G	T	G	T	C	T	G	G	A	G	A	T	A	A	T	G	T			



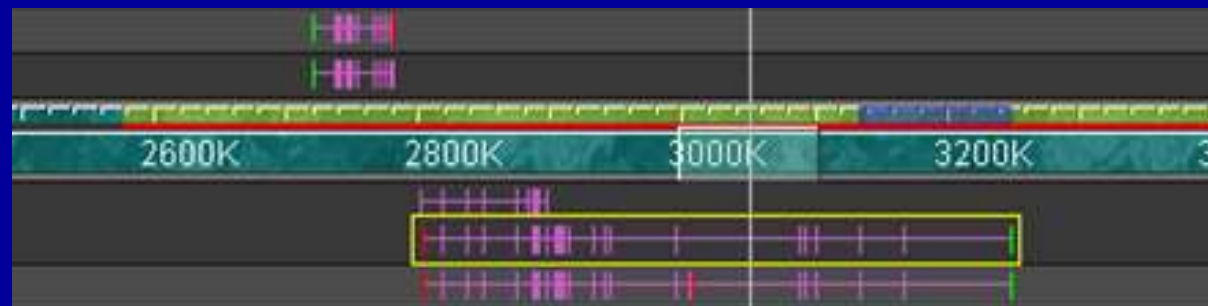
Genome Visualization

- ▶ Genomic Axis



Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations



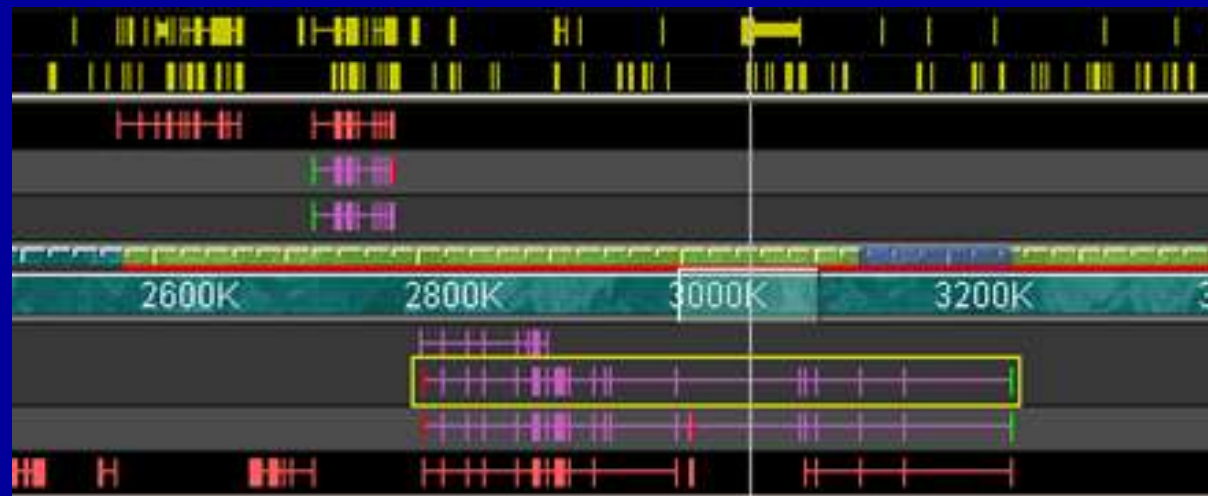
Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations
- ▶ Gene Predictions



Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations
- ▶ Gene Predictions
- ▶ Automated Searches



Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations
- ▶ Gene Predictions
- ▶ Automated Searches
- ▶ Polymorphisms



Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations
- ▶ Gene Predictions
- ▶ Automated Searches
- ▶ Polymorphisms
- ▶ Codons



Genome Visualization

- ▶ Genomic Axis
- ▶ Gene Annotations
- ▶ Gene Predictions
- ▶ Automated Searches
- ▶ Polymorphisms
- ▶ Codons
- ▶ Splice Sites



Sequence Visualization

► Consensus Sequence View

5488...	G	A	T	G	C	T	T	T	C	A	G	A	C	C	G	T	C	C	G	C	T	C	G	G	C	C	G	G	T	
5488...	C	C	A	G	T	C	G	A	T	C	A	G	T	C	G	C	A	C	C	C	G	A	C	A	G	T	T	G	C	T
5489...	C	G	T	G	G	T	A	A	G	C	C	T	G	T	T	G	C	G	C	A	T	C	T	G	T	T	A	G	C	C
5489...	C	G	G	T	G	C	T	G	G	G	A	A	A	G	G	G	C	A	A	G	C	A	G	G	G	T	A	G	A	T
5489...	G	G	C	C	C	G	G	C	A	C	C	A	A	T	G	A	T	C	A	A	G	A	C	A	T	C	G	A	C	A
5490...	A	T	G	G	T	C	A	C	G	T	G	G	G	C	T	G	G	A	A	G	C	A	G	A	T	T	C	G	C	

Sequence Visualization

► Consensus Sequence View

```

5488...GATGCTTTTCAGACCGTCCGCTCGGCCGGT
5488...CCAGTCGATCAGTCGCACCCGACAGTTGCT
5489...CGTGGTAAGCCTGTTGGCGCATCTGTTAGCC
5489...CGGTGCTGGGAAAGGGCAAGCAGGGTAGAT
5489...GGCCCGGCACCAATGATCAAGACATCGACA
5490...ATGGTCACTGTGGGCTGGAAAGCAGATTGCG
  
```

► Transcript Translate View

Frame +3	0	C	A	C	C	T	T	T	A	A	G	C	G	T	C	A	C	G	G	G	T	G	G	G	G	C
Frame +3																										
52		C	A	T	G	T	C	G	C	G	C	T	G	A	A	G	C	G	G	A	T	A	G	C	G	
Frame +3		A		G		G		R				D		C		G				T		S				
104		G	C	A	G	G	T	G	G	A	A	G	A	G	A	C	T	G	C	G	G	T	A	C	C	T
Frame +3																										
156		C	G	A	G	G	A	A	G	A	G	C	G	G	C	C	A	G	T	T	A	A	A	C	C	T
Frame +3		Q		C		V		W		R	I															
208		G	C	A	G	T	G	T	G	T	C	T	G	G	A	G	A	A	T	A	A	T	G	T	G	

Sequence Visualization

► Consensus Sequence View

```

5488...G A T G C T T T T C A G A C C G T C C G C T C G G C C G G T
5488...C C A G T C G A T C A G T C G C A C C C G A C A G T T G C T
5489...C G T G G T A A G C C T G T T G C G C A T C T G T T A G C C
5489...C G G T G C T G G G A A A G G G C A A G C A G G G T A G A T
5489...G G C C C G G C A C C A A T G A T C A A G A C A T C G A C A
5490...A T G G T C A C G T G G G C T G G A A G C A G A T T C G C
  
```

► Transcript Translate View

```

Frame +3
0      C A C C T T T A A G C G T C A C G G G T G G G G C
Frame +3      S R L K R I A
52     C A T G T C G C G C C T G A A G C G G A T A G C G
Frame +3      A G G R D C G T S
104    G C A G G T G G A A G A G A C T G C G G T A C C T
Frame +3      R K S G Q L N L
156    C G A G G A A G A G C G G C C A G T T A A A C C T
Frame +3      Q C V W R I N V
208    G C A G T G T G T C T G G A G A A T A A A T G T G
  
```

► Alignment View

```

                                     K K E L T Q I K Q K
                                     K K E L T Q I K Q K
S G Q R G S S K L K G D D L Q A I K K
S G Q R G S S E L K G D D L Q A I R R
gtgg ca tggggatcttc aat t gaaagg gat gaccttcaggccat aa aaggaa t gac cagat aaaac aaaa
gtgg ca tggggatcttc aat t gaaagg gat gaccttcaggccat aa aaggaa t gac cagat aaaac aaaa
GTGGGCAGTGGGGATCTTCCGAATTGAAAGGAGATGACCTTCAGGCCATAAGAAGGAATTGAACCCAGATAAAACAAAAA
5260960 5260970 5260980 5260990 5261000 5261010 5261020 5261030
  
```

The Celera Genome Browser

► Purpose

- To support Celera's scientists and customers in annotation of large eukaryotic genomes such as the human genome

► History

- Version 1 (1999) – Annotation of first assembly of fruitfly genome
- Version 2 (2000) – Annotation of Celera's first assembly of the human genome
- Subsequent versions supported Celera annotation projects including the mouse and rat genome assemblies
- Version 4.2 (2002) completed
- Open Source Release (2006)

Demo



- ▶ Comparative View of CFTR Region

Related Annotation Tools

- ▶ Apollo (Berkeley)
- ▶ Artemis (Sanger Institute)
- ▶ Early Work
 - BioPerl
 - BioTK
 - Neomorphic GeneViewer

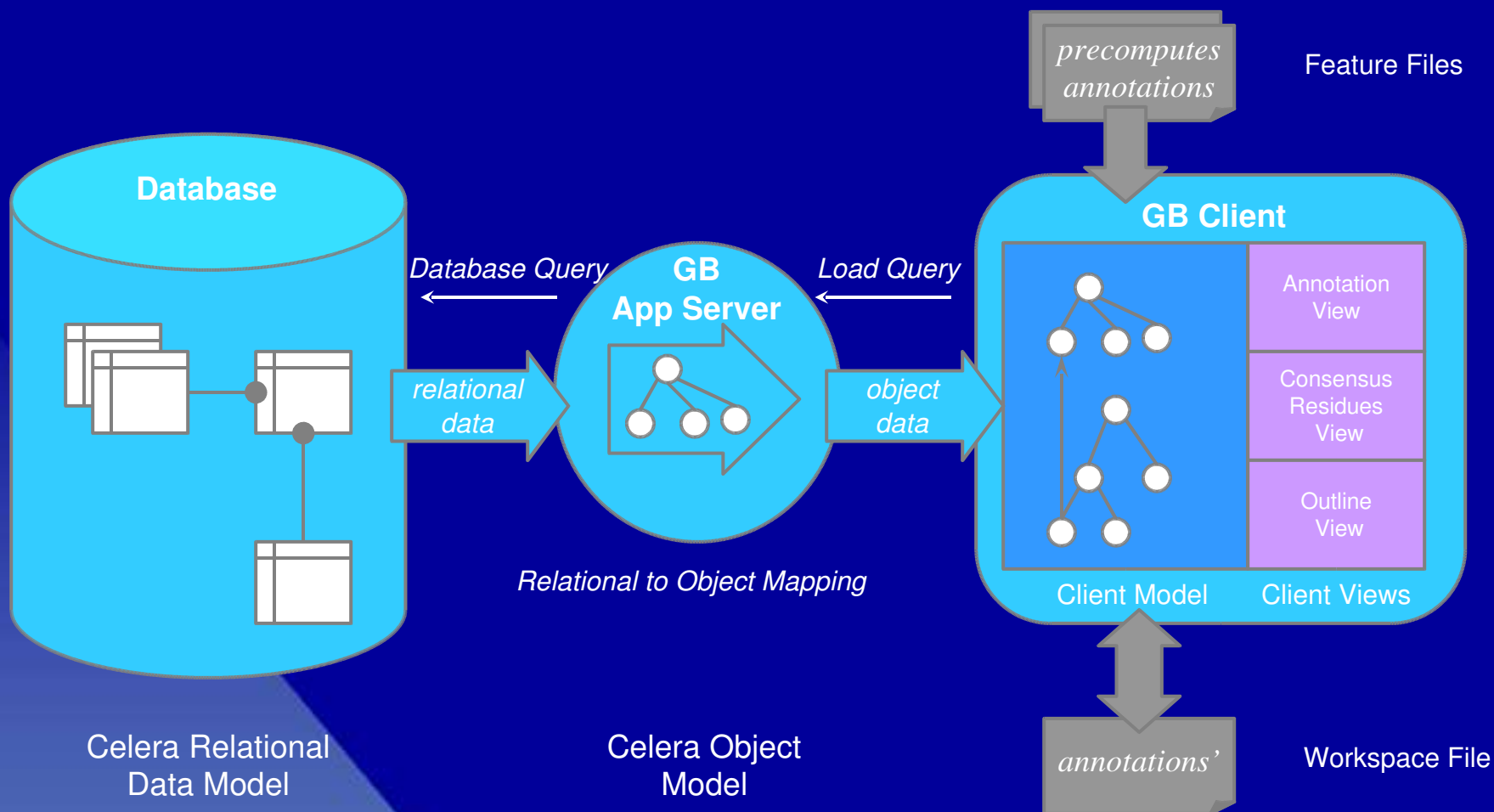
Notable Visualization Features

- ▶ Main View Tiered Feature Display
 - Overview with high-performance zoom
 - Dragable, dockable tiers
 - Full user preferences for tiers, features, colors, orders etc.
 - Color intensity mapping of arbitrary feature properties
 - Highlighting of curation evidence, exon edge/splice matching
 - Thick Intron lines for adjacent HSP's in subject space
 - Length of splice site glyph denotes score
 - SNP glyphs shape and color
- ▶ Subviews
 - Simultaneous micro-macro views with linked visual cues
 - Spliced Transcript View with start/stop codons and reading frames
 - Color-coded comparison of DNA and Protein sequences
 - Property view for details-on-demand
- ▶ Navigation
 - Bookmarks
 - Searching
 - Navigation launched from web browsers
 - Sequence Analysis on the fly

Notable Annotation Features

- ▶ Direct Manipulation Operations
 - “Drag and Drop” creation of transcripts, evidence
 - Dragging exon edges to resize
 - Infinite-level undo
 - Annotation logging
- ▶ Right-click popup actions
 - Deleting exons
 - Splitting/Merging transcripts
- ▶ Reloadable Curation “Workspace”

System Architecture



Availability

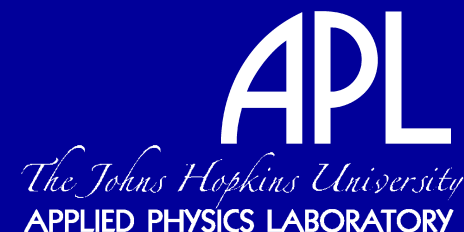
The Celera Genome Browser has been released under an LGPL license and will be available soon on Source Forge:

<http://sourceforge.net/projects/celeragb>

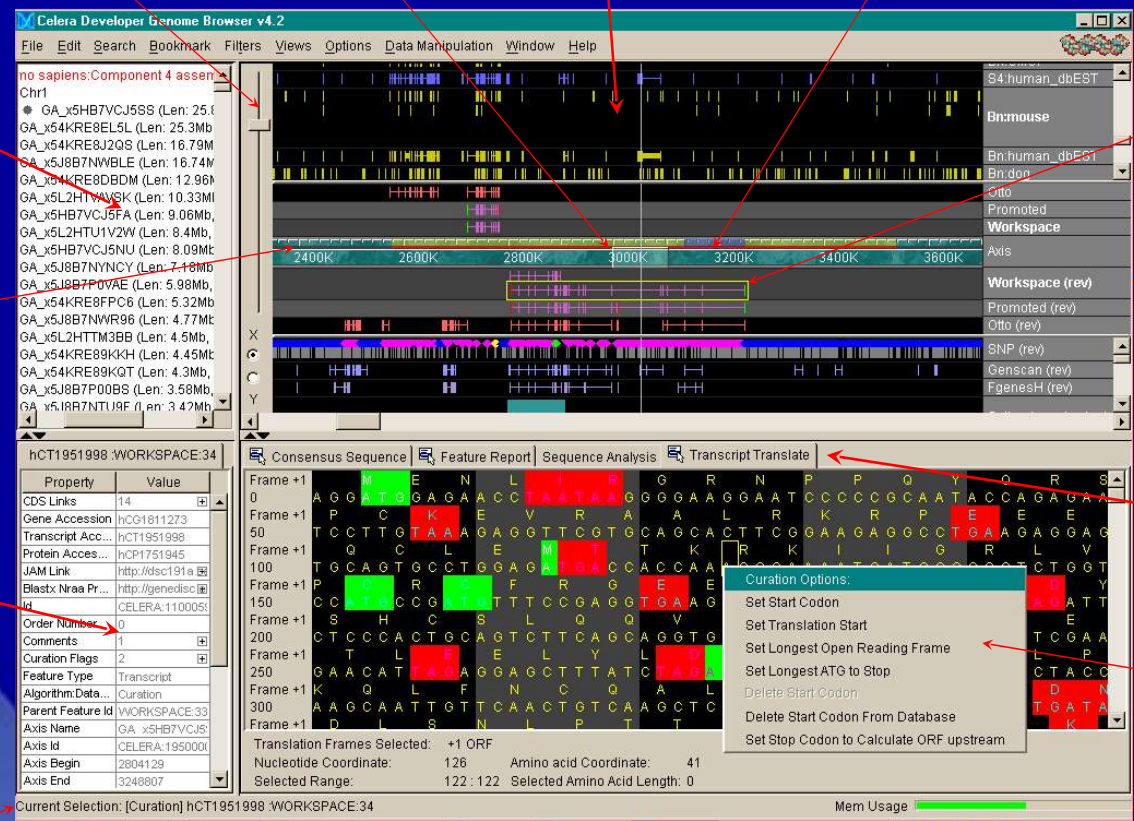
Acknowledgements

Thank you to others who helped create the Celera Genome Browser:

Mark Adams, Deepali Bhandari, Lou Blick, Rosane Charlab, Ken Graham, Michael Harris, Gregg Helt, Olivier Jojic, Jackie Keeling, Kennen Kellaris, Chinnappa Kodira, Mark Lewis, Anne Deslattes-Mays, Joe McDaniel, Natalia Milshina, John Otridge, Todd Pihl, Randall Ribaud, M'liz Robinson, Michael Simpson, Jay Shira, Marian Skupski, Deepak Srinivasan, Eric Sun, Jennifer Wortman, Rick Wise, David Wu, and Qing Zhang.



Graphical User Interface Overview



Zoom Slider

Zoom Focus Slider

Annotation View

Selected Region

Loading Indicator

Outline View

Genomic Axis

Property Inspector View

Subviews

Popup Curation Menu

Memory Usage Meter

Description Line

Selected Feature

Annotation Workspace Tier

Promoted Annotation Tier

Consensus Sequence

Feature Report

Sequence Analysis

Transcript Translate

Curation Options:

- Set Start Codon
- Set Translation Start
- Set Longest Open Reading Frame
- Set Longest ATG to Stop
- Delete Start Codon
- Delete Start Codon From Database
- Set Stop Codon to Calculate ORF upstream

hCT1951998 :WORKSPACE:34

Property	Value
CDS Links	14
Gene Accession	hCG1811273
Transcript Acc...	hCT1951998
Protein Acces...	hCP1751945
JAM Link	http://dsc191a
Blastx Nraa Pr...	http://genedisc
Id	CELERA:1100051
Order Number	0
Comments	1
Curation Flags	2
Feature Type	Transcript
Algorithm:Data...	Curation
Parent Feature Id	WORKSPACE:33
Axis Name	GA_x5HB7VCJS
Axis Id	CELERA:1950001
Axis Begin	2804129
Axis End	3248807

Current Selection: [Curation] hCT1951998 :WORKSPACE:34

Mem Usage