# Manual Annotation of Features in DNA sequences

J. Craig Venter Eukaryotic Genome Annotation and Analysis Course

# Overview

- What is manual annotation?
- Why do manual annotation?
- Tools for manual annotation

GBrowse

Apollo

Artemis

Critical evaluation of evidence

#### J. Craig Venter

## What is manual annotation?

- Automated gene structure annotations are often imperfect and can benefit from manual refinement.
- A manual annotator uses different tools to view and curate the gene structure.



# Why do manual annotation?

- Finding the genes is a primary goal of genome sequencing, and the gene set is the foundation for further studies. Accuracy is essential!
- Scientists can make informed decisions to correct faulty automated annotations.
- Model organisms have to be annotated to GOLD standards.

### Manual annotation tools

- Annotation tools allow you to visualize the location of features with respect to each other.
- They can also display evidence alignments and other information relevant to the process of annotation.
- Annotation editors allow the user to make changes in the annotation of a sequence and save these changes to a database.
- This talk will mainly focus on some open source annotation viewers and editors, however many commercial solutions are also available.

#### J. Craig Venter

#### Open source annotation tools/viewers

#### GBROWSE

#### Volvox Example Database

450 @ 640

800 🔘 1024

#### Showing 50 kbp from ctgA, positions 1 to 50,000

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position. Examples: cha.

[Hide banner] [Hide instructions] [Bookmark this view] [Link to an image of this view] [Help] Landmark or Region Scroll/Zoom ctgA:1..50000 Search Show 50 kbp 🔍 >>>verview of ctgf 204 30k 40k 10k 10k 20k 30k 40k -----Example Features £14 Data Source Volvox Example Database . Tracks [Hide] (External tracks 🕱 Example Features italicized) Image Width Key position Track Name Table Set Track Options... Update image

💿 Alphabetic 🔘 Varvir

### APOLLO









### J. Craig Venter

Between C Beneath



The Generic Genome Browser (also referred to as 'gbrowser') is a combination of database and interactive web page for manipulating and displaying annotations on genomes. It is similar to the genome browser used at UCSD and Ensemble

	Volvox Example Da	tabase						
Showing 50 kbp from ctgA, positions 1 to 50,000								
s on this afternoon	Instructions: Search using a seq the ruler. Use the Scroll/Zoom bu Examples: ctgA.	uence name, gene nam ittons to change magr	e, locus, or other land ification and position	dmark. The wildcard cl ).	naracter * is allowed. To ce	nter on a location, click		
	L andmark or Region	a) [Doownark cura vi	כאין נבוווג נס מו ווומ	ge of this view] [riei S	roll/Zoom:			
	ctgA:150000		Search Reset	Fip Sip	K K - Show 50 kbp	• 🗣 <mark> &gt; &gt;&gt;</mark>		
	0verview o (++++++	<b>f ctgA</b> + + + + + + + + + + + + + + + + + + +			+ + + + + + + + + + + + + + + + + + +	· · · · >		
	<+++++++++++++++++++++++++++++++++++++	 )k	20k		40k	******		
	Example Features	f08	f13 f0 f15 f14	2	f09 f03 f04	f01 f		
	Data Source							
	Volvox Example Database		-					
Craig Venter	Tracks [Hide] (External tracks	nple Features						
STITUTE	italicized) Image Width	Key position	Track Na	ame Table	Set Track Options	Update Image		

I N

### Important Features of GBrowse

It is mainly used to view and critically evaluate the various evidence

- Web based annotation Viewer
- Simultaneous bird's eye and detailed views of the genome.
- Scroll, zoom and center functions
- ✤ Attach arbitrary URLs to any annotation.
- Order and appearance of tracks are customizable by administrator and end-user.
- Search by annotation ID, name, or comment.
- Supports third party annotation using GFF format
- ✤ Settings persist across sessions.
- ✤ Allows DNA and GFF dumps.
- DNA sequence can be loaded into the database and displayed in a track along with the annotation.
- Annotation provided by the user can also be displayed in a track. The annotation data can be in a file on a remote server.

#### J. Craig Venter

### GBrowse needs:

### Data files

DNA and features files to load into the local database. e.g. gff files

Conf files

GBrowse configuration files for you to take and modify.

#### J. Craig Venter

## **GBrowse : An example**

٠

#### Showing 100 kbp from 8673, positions 80,001 to 180,000

#### □ Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: 8673:1..30000, NFIA\_000680.

[Hide banner] [Bookmark this] [Link to Image] [High	n-res Image] [Help] <mark>(Reset)</mark>		
⊟ <u>Search</u>			
Landmark or Region:	Reports & Analysis:		
8673:80001180000 Search	Download Decorated	FASTAFIle Configure Go	
	Scroll/Zoom: < <	Show 100 kbp 🔽 📑 2 꾠 🗹 Flip	
⊡ <u>Overview</u>			
Overview of 8673			
ÓM	0.5M	111 1.	5M
⊟ <u>Details</u>			
180k 170k	······································	130k 120k 110k	100k 90k 80k
TIGR working models	▶	HOH+ CD+ CD+ CD+	
EVM2 predictions 💶 🛛 🗧			
snap predictions			
twinscan predictions 💶 🛛 🕂			
GeneZilla predictions			
glinnerHMM predictions			
Iransposable elements			-
X gc			
6-frame translation			
pap vs AllGroup piga			
fungalDB_for_nfa1.pep	ו זה המשוע היה המשועה או היה היה היה היה היה היה היה היה היה הי	מסמו מתרכנינים בנו בנו בנו בנו או אלא הלהביר היו או אלא וויינט וויינט או היו אלא אלא איני או ארא א	
creat inginighting			opuale image
⊟ Tracks Tracks			
$\Box$ General $\Box$ All on $\Box$ All off			
☑ 6-frame translation	✓ fungalDB for nfa1.pep	🔽 nap vs. AllGroup.niaa	✓ Transposable elements
DNA/GC Content	GeneZilla predictions	snap predictions	✓ twinscan predictions
EVAD predictions	Viewworkhith a readictions	TICD working models	

# **Configure Tracks**

#### Settings for TIGR Neosartorya fischeri browser

	Undo Changes	Revert to Defaults	Refresh				Cancel Cha	inges and Return	Accept Change	s and Return	
Track Options											
he Show checkbox turns the track on and off. The Compact option forces the track to be condensed so that annotations will overlap. The Expand and Hyperexpand options turn on collision control using slower and aster layout algorithms. The Expand & label and Hyperexpand & label options force annotations to be labeled. If Auto is selected, the collision control and label options will be set automatically if space permits. To hange the track order use the Change Track Order popup menu to assign an annotation to a track. To limit the number of annotations of this type shown, change the value of the Limit menu.											
Track				Show		Format		Limit		Change Track Order	
Track 1	TIGR working mo	dels				Compact	•	No limit 💌			
Track 2	EVM2 predictions					Auto	•	No limit 💌		•	
Track 3	snap predictions			V		Expand	•	No limit 💌		•	
Track 4	twinscan predictions		V		Expand	•	No limit 💌		•		
Track 5	GeneZilla predictions		V		Auto	•	No limit 💌		•		
Track 6	glimmerHMM pre	dictions		V		Auto	•	No limit 💌		•	
Track 7	Transposable ele	ments		V		Auto	•	No limit 💌		•	
Track 8	DNA/GC Content			V		Auto	•	No limit 💌		•	
Track 9	6-frame translatio	n		V		Auto	•	No limit 💌		•	
Track 10	nap vs. AllGroup.niaa		V		Expand & Label	•	No limit 💌		•		
Track 11	fungaIDB_for_nfa	1.рер		V		Auto	•	No limit 💌		•	
	Undo Changes	Revert to Defaults	Refresh				Cancel Cha	inges and Return	Accept Change	s and Return	
External tracks it	alicized										

\*Overview track

\*\*Region track

### **Detailed information about a gene**

#### TIGR Aedes aegypti genome browser

Name:	24834.m01550					
Class:	mRNA					
Туре:	processed_transcript					
Source:	augustus					
Position:	supercont1.1:52583625274955	(+ strand)				
Length:	16594					
Parts:	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52583625258562 (+ strand)				
	Length:	201				
	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52586385258764 (+ strand)				
	Length:	127				
	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52588195259310 (+ strand)				
	Length:	492				
	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52595265259578 (+ strand)				
	Length:	53				
	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52596435259924 (+ strand)				
	Length:	282				
	Туре:	CDS				
	Source:	augustus				
	Position:	supercont1.1:52740295274955 (+ strand)				
	Length:	927				

>24834.m01550 class=mRNA position=supercontl.1:5258362..5274955 (+ strand) ATGGAGATTGGAGAAATGGCGATGTCCAATTTCGCATCGTGGTACCACGCGATGAATCCTGGCCATCTGACCACAGATGA  ${f T}{f G}{f A}{f A}{f G}{f G}{f A}{f A}{f A}{f G}{f G}{f G}{f A}{f A}{f T}{f C}{f G}{f G}{f A}{f G}{f C}{f G}{f A}{f G}{f T}{f G}{f G}{f$ GATAAACTGAATGGGATAAAAAATCACCTATAAAAGCGAGTGTCTAAGAAAGCGCCAGATCAGTTGTACAAAACGAAATT GGTTCATATCTTTTTCCGAATGCAGCGATTGCAAGCTCACACTGAGGAAGAAGACGAATTGAATAGCCTCGCGTTGATTG CTGGTGAATGTGTAAAGCTTCTCAACGTGTATTATTCACCCACATCCCACCTTTTAGAAGTGCGAGAGGCCGAATTGGCG AAGATCAAATGGGAGATCGACAGGTGAAAACGTGACTGAGCCGAATGGTAGGAATGTAGAAAATGATAGGATGAGTACCG GAGGTGAGCTATCTGAGAAGACGACGGATAGTGAGAAAGAGGGATTAGCTTCTGAGATCCGTCAGCTAACGGCGGTGGTG AGCCAACTACTGCAACGCATCACGGTACTTGAAGCTAATCAAAACGTTGCTCAACCGTTGAATCTTCATACAGGTACAGC GCTGGTTGAAACAGAGAAATGAGTCGTTAGATACGCTGAAAAATTCGGATGAAGACAAGCACGTCGGTAGTAAGCTGAAT ATATGACGGAATGGATAATGGCAGGCGGTTGAATGAATTCTTGAAAGAGGTGGAATTCAACGCCCGGTCAGAAGGATTTT 

# Clickable links

8 M						My NCBI
			Structure	SPMC -	Taxabaray	[Sign In] [Register]
Search Prote	in <b>T</b> for	o en onte	Go Clear		raxonomy	Biolog
	Limits	Preview/Index	History	Clipboard	Details	
Distant Conf						
Display   Gen		Warden and a second				
Range: from k	pegin to end Features: 🗹 🤇	DD + Refresh				
□1: <u>AAL35</u>	396. Reports Opie2a pol [Zea m[gi:170824	179]				BLink, Conserved Domains, Links
Features Se	equence					
LOCUS DEFINITION ACCESSION VERSION DBSOURCE KEYWORDS SOURCE ORGANISM REFERENCE AUTHORS TITLE JOURNAL	AAL35396 1048 aa Opie2a pol [Zea mays]. AAL35396 AAL35396.1 GI:17082479 accession <u>AF391808.2</u> Zea mays Zea mays Eukaryota; Viridiplantae; Strepto Spermatophyta; Magnoliophyta; Lil clade; Panicoideae; Andropogoneae 1 (residues 1 to 1048) Fu,H., Zheng,Z. and Dooner,H.K. Direct Submission Submitted (15-JUN-2001) Waksman I Frelinghuysen Rd, Piscataway, NJ	linear PLN 1 phyta; Embryophyta; Trac iopsida; Poales; Poaceae ; Zea. nstitute, Rutgers Univer 08854, USA	nov-2003 neophyta; ; PACCAD sity, 190			
REFERENCE	2 (residues 1 to 1048)	,				
TITLE	Direct Submission	ng,2. and booner,n.k.				
JOURNAL	Submitted (26-NOV-2001) Waksman I	nstitute, Rutgers Univer	sity, 190			
REMARK	Sequence update by submitter	00034, 0JA				
COMMENT	Method: conceptual translation su	pplied by author.				
FEATURES	Location/Qualifiers					
source	11048 /organism="Zea mays" /db_xref="taxon: <u>4577</u> " /chromosome="9" /man="95"					
Protei	<u>n</u> 11048					
Region	/product="Opie2a pol" 179342 /region_name="rve" /note="Integrase gore do	main, nfam00665″				
	/db_xref="CDD: <u>40748</u> "	main, pramotout				
CDS	11048					

### Load and compare annotations



## Repeat region information

#### TIGR Aedes aegypti genome browser

Name:	repeat_15051
Class:	Match
Туре:	match
Source:	repeat_regions
Position:	supercont1.1:52097775210002 (+ strand)
Length:	226
Note:	trf

>repeat\_15051 class=Match position=supercontl.l:5209777..5210002 (+ strand) TATCGATTTGAGAATCGATTGAGAAATATCGATTTGAGAATCAAGGGAGTAATTTTAGTTTGAGGATCGATGGAATGGTA TCGATTTGAGAATCTATTGAGAAATATTGATTTGAGGATTAATTGAGTAATATCGATTTGAGAATCAATTGAATAGTATC GATTTGCGAATCGATGGAATGGTATCGATTTGAGAATCGATGGAGTAATATCGATTTGAGGATCGA

Name:	repeat_15051
Class:	Match
Туре:	match
Source:	transposonPSI
Position:	supercont1.29:15403541540551 (+ strand)
Length:	198
Note:	LINE

>repeat\_15051 class=Match position=supercontl.29:1540354..1540551 (+ strand) GACATAGGCCAACATGTCAAAACATCTAAAAACATGAAGGCCCCAGGTTTCGACAGCATCCTGAACATGGAACTCAAGCA ATTAAGTCTCCAGTTCTATCTACATAGCTACTTCCGACTTAGCTACTTCCCCCTCGTCGTGGAAGTCAGCAAAAGTCATCC CCATCACAAAACCTGGGAAGGATCCACCAAAATCTGTT

# **Highlight Regions**



## **Highlight Features**



## **Display options**



### Links to Manatee

Rama and Linda will talk about Manatee tomorrow And Mathangi will demo manatee

Neosartorya fischeri	NFIA_000970
	Download sequence Show genomic region on 1099437636265
Gene Identification	
Gene Product Name: hexe	ose carrier protein
Locus Name: NFIA	_000970
Comment:	
Gene Ontology Class	ification   🕒
▶ None found	
Attributes	
Chromosome:	Unknown
Coordinates (5' - 3'):	129307 - 130996 on assembly <u>1099437636265</u>
Nucleotide length:	1635
Predicted protein length:	545
Predicted molecular weight:	0.00
Predicted pl:	0.00
Gene Structure	
400 	
Intron/Exon structure in nuc	leotide coordinates relative to the start of the gene. Click on the picture to display evidence supporting this gene structure.
Protein Features	
0 100	200 300 400 500
	NFIA_000970 PF00083: transporter, major facilitator superfamily

PF07690: transporter, major facilitator family

PR00171: PS50850:

#### HMM/PFAM Hits Accession Name Total Score Trusted cutoff Noise cutoff Total expect Curated transporter, major facilitator superfamily PF00083 386.0 -85.00 -85.10 6.7e-113 MFS transporter, sugar porter (SP) family 100.00 3.5e-101 TIGR00879 347.1 150.00

#### Open source annotation tools/viewers

#### GBROWSE

#### Volvox Example Database

#### Showing 50 kbp from ctgA, positions 1 to 50,000

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character " is allowed. To center on a location, click the ruler. Use the ScrollZoom buttons to change magnification and position. Examples - ctA



### APOLLO



#### ARTEMIS



### J. Craig Venter

### APOLLO http://www.fruitfly.org/annot/apollo/

# APOLLO Genome Annotation and Curation Tool Apollo was written by members of the FlyBase-Berkeley Informatics group (Suzanna Lewis, Mark Gibson & Nomi Harris), the Howard Hughes Medical Institute (John Richter) and the Sanger Institute/European Bioinformatics Institute (Steve Searle & Michele Clamp).

### J. Craig Venter

# **About Apollo**

- Apollo is a genome annotation viewer and editor
- It was developed as a collaboration between the <u>Berkeley Drosophila</u> <u>Genome Project</u> (part of the <u>FlyBase</u> consortium) and <u>The Sanger</u> <u>Institute</u> in Cambridge, UK.
- Apollo allows researchers to explore genomic annotations at many levels of detail, and to perform expert annotation curation, all in a graphical environment.
- The <u>Generic Model Organism Database (GMOD) project</u>, which aims to provide a complete ready-to-use toolkit for analyzing whole genomes, has adopted Apollo as its **annotation workbench**.
- Apollo is a Java application that can be downloaded and run on Windows, Mac OS X, or any Unix-type system (including Linux).

#### J. Craig Venter

# **Important features of Apollo**

- Apollo provides 'semantic zooming'
- You can drag and drop features from evidence panel.
- Simple interface for editing gene models (merging, splitting, deleting, adding exons/transcripts, set 5', 3' or calculate longest open reading frame)
- ✤ By default it can display both strands at the same time.
- Sequence-level adjustments are possible with Apollo's exon editor.
- ✤ Apollo allows you to tag results with a comment
- Apollo automatically creates alternatively spliced transcripts for a gene whenever the Open Reading Frames (ORFs) of transcripts overlap.

#### J. Craig Venter

## **Reading Data**

Apollo can transparently access data across the network from remote machines, as well as reading files that reside locally.

- Apollo can read and load data directly from a chado database
- Apollo can also read Chado XML or GAME XML files
- It can read Ensembl GFF files
- It can read Ensembl schema databases
- It can read GenBank or EMBL format

#### J. Craig Venter

# **Apollo – Main window**



# **Types Panel**

* Types			
Sort	Annotation	Show 🗾 Expand 🗹	ks <u>H</u> elp 
Sort	RepeatMasker	Show 🗾	
Sort	TRNAS	Show 🗹	
Label		Expand 🗌 Show 🔽	
Label	Copy_of_Annotation	Expand 🔲	
🛄 Sort 🔲 Label	Old_Working_Models	Show 🗾 Expand 🗌	eha3.transcript.111003.1
Sort 📃 Label	EVM	Show 💌 Expand 🗔	<u> </u>
Sort	EVM2	Show 💌	eha3.transcript.11064.1 eha3.transcript.11006.1
Sort	D/M alm	Expand Show 💌	
Label	Eveleni	Expand 🔄 Show 🔽	
Label	GeneZilla	Expand 🗌	
🛄 Sort 🛄 Label	glimmerHMM	Show 🗹 Expand 🗔	
Sort	phat	Show 💌 Expand 🛄	
Sort	snap	Show 🗾	
Sort	twincran1	Show 🗹	
C Label	twinstant	Expand 🗌 Show 💌	an a
Label	twinscan2	Expand 🛄	
🗌 Sort	AAT_AA AllGroup.NoTIGR.niaa	Show 🗾 Expand 🗹	eha3.gene.5415.1 eha3.transcript.11003.1
🔄 Sort	AAT_AA pfam_prots.db	Show 💌 Expand 💌	Id     Genomic Range     Genomic Length       eha3.exon.7072.1     7646-8968     1323
Sort	Start Codon	Show 💌	
Sort	Stop Codon	Show M	
Sort		Expand 🛄 Show 💌	
🗌 Label	Sequence selection	Expand 🗌	
Position	9710 Feature Action		

### Finding more about annotation and features

#### Right click on a gene

* eha3.assembly.1033.1				
Eile Edit View Tiers Analysis Bookmarks Annotation Window Links	s <u>H</u> elp			
				<u></u>
				30
twinscan2:236956-239607		twinscan2:242666-243532 twir	iscan2:243727-245037	t'
				1921
eha3.transcript.11100.1		eha3.transcript.10817.1 eha	3.transcript.10918.1	e
340000 340800 341600	242400	242200	Sequence	7454
, 240000 , 240800 , 241800	242900	, 245200 ,	Get info about this feature via Web	243
eha3.transcript.10857.1 eha3.transcript.11131.1			Exon detail editor	
			Delete selection	
			Merge transcripts	1000
			Split transcript Duplicate transcript	-
			Marga exons	
Genezilla.240029-2596/5 Genezilla.241659-240945			Move exon(s) to transcript	
			<u>S</u> plit exon	
			Create new annotation >	No.
			Calculate longest ORF	000000
twinscan2:241659-240945			Set as 3' end	
			Set both ends	
			not analyzed	
			Disown by mathangi Set eba3transcript.10918.1 completed	
			Change color of this feature type	-
Position			C <u>o</u> llapse tier	
Zoom x10 x2 x.5 x.1 Reset Zoom factor = 87.6656			Expand tier	4
Type Name Range Score		eha3.gene. <sup>5</sup>	5330. Close menu	
gene eha3.transcript.1  243727-245037  0.0	eha3.transcript.10918.1	Genomic R	ange Genomic Length	
	eha3.exon.6964.1	243727-245037	1311	
Position 244860 Feature eha3.transcript.10918.1 Action				

### Annotation info editor

#### View Tiers Analysis Bookmarks Annotation Window Links Help

eha3.gene.5330.1 Annotation Information			
eha3.gene.5364.1 📩 gene ena3.gene.5330.1			
eha3.gene.5585.1 Symbol	eha3.gene.5330.1	rtranscript eha3.transcript.10	1918.1
Feha3.gene.5460.1	eha3 gene 5330 1	Symbol	eha3 transcript 10918 1
Fehas.gene.5430.1	enas.gene.ssso.i	Concerne	endp.rrdibenpt.robro.r
F enas.gene.5343.1 Synonyms		Synunyms	
F enas.gene.5422.1		add delete	
h cho2 gono 5407.1	(	ass servere	
L obo2 gong 5501.1		la problemetic?	
L oba2 gone 5284.1		is problematic?	
L aha2 gana 5522 1 Type	gene 💌	Finished?	
Enabligene 5749.1		Author	mathangi
Heha3 gene 5330 1		Readthrough stop codon	
Evaluation of peptide	not analyzed 🔹 💌	No introns	
eha3.gene 5467.1		+ 1 translational frame shift	No
eha3.gene.5625.1 ID Valu	e DB Name	-1 translational frame shift	No
eha3.gene.5488.1	19. 19. 19. 19. 19. 19. 19. 19. 19. 19.	Missing start codon	No
eha3.gene.5445.1		Missing stop codon	No
⊢eha3.gene.5565.1		Unconventional start codon	No
⊢eha3.gene.5492.1		5	
Eeha3.gene.5432.1			
🗄 eha3.gene.5276.1 🛛 🚺 Comments and prope	rties		
eha3.gene.5591.1 eha3 transcript 1091	8.1 properties:		
Eeha3.gene.5480.1	ON: Tue Feb 06 10:35:59 EST 2007	1	
Eeha3.gene.5420.1		Edit eha3 ger	e 5330 1 comments
Eeha3.gene.5404.1		Eure enus.ger	
⊢eha3.gene.5599.1		1.5	
⊢eha3.gene.5513.1		(	
Eeha3.gene.5453.1		Edit eha3.tran	scrint.10918.1 com
eha3.gene.5405.1			
enas.gene.5280.1			
F eha3.gene.5524.1		-Genomic sequen	cing errors
Fenas.gene.5520.1		Scholine Sequen	
F enas.gene.5267.1			
F enas.gene.5334.1			
r enas.gene.5463.1			
L pho2 gono 5275 1			
l oho2 gone 5422 1			
Laho2 gene 5484 1			
b chap.gene.5704.1			
enay.gene.5520.1			

Follow selection

## **FIND Function**

# Exon detail editor

* Forward St	rand Exon Editor			11	
243649	G Y Q R F I D T R F S L L GGATACCAAAGAGTCTTTATT	K L G L N V AAACTAGGT	E E I K K K K L K K TAGAAGAAATTAAAAA	I K K D R T N I K R I E Q I NATAAAAAGGA <sup>S</sup> CGAACAAA <sup>S</sup> C	Q F CAGTTTTGATC
243730	M S R R P N S I L E D Q I V F S R T R AATTCTCGAAGACCAAATAGT	R R S <mark>K</mark> E E V I AGAAGAAGTA	Y F D Y E N T L I M K AATACTTTGATTATGA	I P S G K F N N F L L E N S I N S F W K I Q AATTCCTTCTGGAAAATTCAATA	E W A M N G Q M G M AATGAATGGGCA
243811	I T L Y N S T L H Y I I Q R ATTACATTATATAATTCAACO	D K D I I K I I GATAAAGATA	D S L X R L I H N E TTGATTCATTAAAACG	I A N E A A S A L R M K L L L NATTGCGAATGAAGCTGCTTCTC	Q L F V S C L C A G T T G T T T G T
243892	E T R N R S G K Q E T E V E GAAACAAGAAACAGAAGTGGA	T K K V	R G F F V L E D F L T R I F C TTAGAGGATTTTTTGT	F P Y I K S S S F L I L N L C ATTTCCTTATATTAAAATCTATGT	
243973	R E M E G R G E K W K E E V CGAGAAATGGAAGGAAGAGGT	I W G V F G E I	P K L K P F L N S Q TTCCTAAATTAAAGCC	S E M E R W A E V R W N D G L AAGTGAGATGGATGGGCTC	M L V M
244054	V I V D K Q G L L L I N K V GTTATTGTTGATAAACAAGGT	S A P R L P Q L TCTGCCCCAA	G K V S K K V K F P K AAGGTAAAGTTTCCAA	R E K Y R G M S K R K I S W I AAGAGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	S F N H L L I I F CATCTTTTAAT
244135	X H P T I L A N T L L Y L AAACACCCTACTATATTAGCT	T T P S Q L H F ACAACTCCAT	F S N K P H F Q I N L I F K T CATTTTCAAATAAACC	S N G F L N T S O M V F L I L F K W F S Y	V V V H
244216	P P Q T S K L   H L K H Q N   T S N I K T   CCACCTCAAACCTCAAAACTAAAACTAA	T K T S L K L M ACTAAAACTA(	T S I S À V H L F L L Y I Y F L GTACATCTATTTCTGC	PLGNNSS HVIILQ SIRTAGGTAATAATTCTTCAC	I A D P L I H S F CCATAGCTGAT
			1		
Transcript: el	ha3.transcript.10918.1 🔻	Translation len	ngth: 436	Go to next 5' annotation (eha3.gene.524	9.1)
Find seque	clear search hits			Go to next 3' annotation (eha3.gene.546)	7.1)
Show intro	ns in translation viewer		Follow external selection		30 20
J. (	Craig Venter				
IN	<b>S T I T U T E</b>	Sliding window			

## An example



#### Open source annotation tools/viewers

#### GBROWSE

#### Volvox Example Database

#### Showing 50 kbp from ctgA, positions 1 to 50,000

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character " is allowed. To center on a location, click the ruler. Use the ScrollZoom buttons to change magnification and position. Examples - ctA



### APOLLO



#### ARTEMIS



### J. Craig Venter

### Artemis

http://www.sanger.ac.uk/Software/Artemis/



- Developed by Sanger Institute, UK
- Reads in Genbank, EMBL and GFF3 files and sequence in FASTA format.
- Displays annotation derived from a feature table.
- ✤ Has a selection of tools such as a G/C graph.
- ✤ Allows the direct display of blast results.
- ✤ Artemis is Java based (cross platform), easy to run and install.
- Allows editing the annotation and export it in GenBank, EMBL, and GFF3 format

### J. Craig Venter

## Sequence Viewing



GC Plot

Sequence And Feature Overview

DNA View And 6 Frame Translation

Feature Table J. Craig Venter

# Annotating with Artemis



## Annotating with Artemis



# Comparing the viewers

	Artemis	Apollo	Genome Browser
Data Source	Flat file	Flat file or database	Flat file or database
Sharing	Not designed for sharing	Users can share databases	Convenient via the web, possible to share databases
Annotating	Designed for annotating sequence	Designed for annotating sequence	Users may add "private" annotation

#### J. Craig Venter

# **Critical evaluation of evidence**

Weighing the Evidence



Full-length cDNAs are the most reliable pieces of evidence, followed by EST, protein, and finally gene prediction evidence. Evidence of the highest reliability is used for structural annotation.

### J. Craig Venter

# full-length cDNA Evidence



# EST Evidence

When encountering EST evidence, one can examine information about the EST or the ESTs used to make a Tentative Consensus (TC) sequence at Gene Indices WebPages:

http://compbio.dfci.harvard.edu/tgi/

#### J. Craig Venter

INSTITUTE

#### **DFCI Cryptococcus Gene Index**

About CrGI Gene	Index	S 900 S
Development and Goals Release Summary Category Comparison	Background Information about CrGI display a statistical summary of all CrGI releases display estimated number of genes among all fungi releases	
Sequence Simila	rity Search	
BLAST	search TC sequences based on sequence similarity	0 29 9
Sequence Report	is	Balance 7.0 /Basambasa
Identifiers or Keywords TC Annotator	search TC reports using TC identifiers, GB accessions or keywords list all TC annotation	Release 7.0 (December
EST Annotator	list all EST annotation	Input Sequences
Libraries	search EST libraries by keywords or tissue origins	ESTs
CAT# Download	download EST and TC sequences originating from one library	ETs
Functional Annot	ation and Analysis	Output Sequences
Alternative Splice Forms EST Expression Gene Ontology	prediction of alternative splice variants compare EST expression between different libraries or tissues classification of TCs by GO vocabularies	TC sequences singleton ESTs singleton ETs
Metabolic Pathways	association of TCs with metabolic and signaling pathways	Total unique: 5674



7,2002

Input Sequen	ces
ESTs	17472
ETs	506
Output Sequer	ices
TC sequences	2384
singleton ESTs	3231
singleton ETs	59
Total unique: 5	i674

# **EST Evidence**



## Protein Evidence



# **Questions?**

