GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

12 October 2004

Lecture 13: Biological Interpretation

- Introduction
- Primary probe identifiers
  - Sigma-Genosys
  - Agilent
  - Affymetrix
  - IMAGE
- GenBank
- UniGene
- LocusLink
- Batch Resources

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Introduction

After analyzing a microarray experiment, you typically end up with a list of “interesting” genes. Today’s lecture deals with how to make biological sense of that list.

Keep in mind that the list may be quite long. For example, in our analysis of the prostate cancer data set, we found about 3500 genes that were differentially expressed (along with 3500 potential biomarkers, with an overlap of about 2500).

There are numerous databases (GenBank, UniGene, LocusLink, etc.) that include gene-related information. It can be difficult to keep track of exactly what kind of information each identifier describes. Each of the three kinds of microarrays has a different primary descriptor that tells you what bioloigcal material was placed on the array.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Primary Identifiers

synthesized oligonucleotide array：These are the Affymetrix arrays. The primary identifier is an Affymetrix probe set ID, which refers to the colletion of 25-mers that form a probe set.

spotted cDNA array：These are the most common glass microarrays. The primary identifier is an IMAGE clone id, which refers to the actual cDNA clone attached to a vector and propagated in bacteria

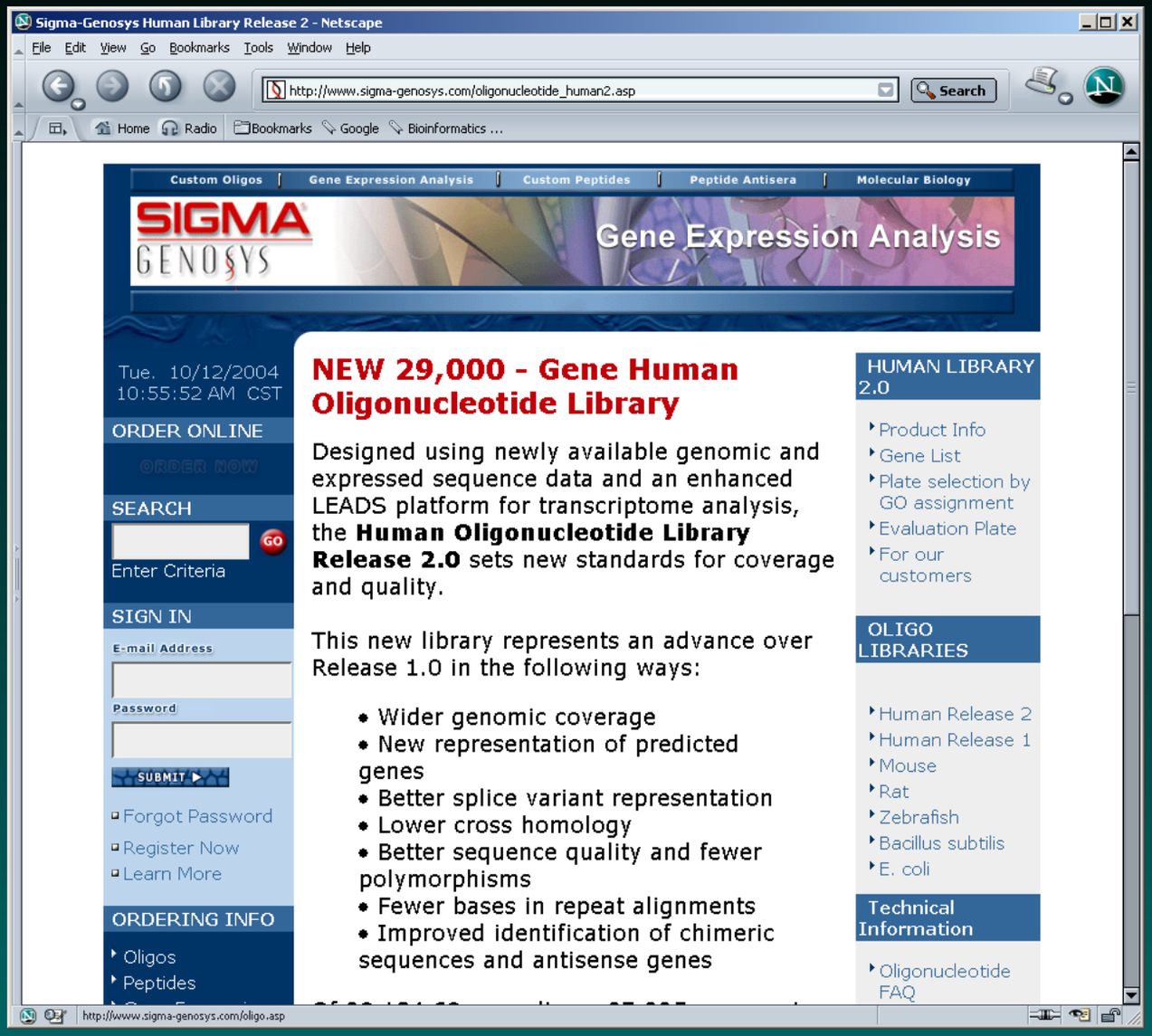
spotted oligonucleotide array：Newer glass arrays often spot commercially synthesized 60- or 70-mers instead of cDNA clones. The primary identifier is usually the commercial identifier that corresponds to the actual synthesized sequence.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

http://www.sigma-genosys.com

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Sigma-Genosys Human Oligo



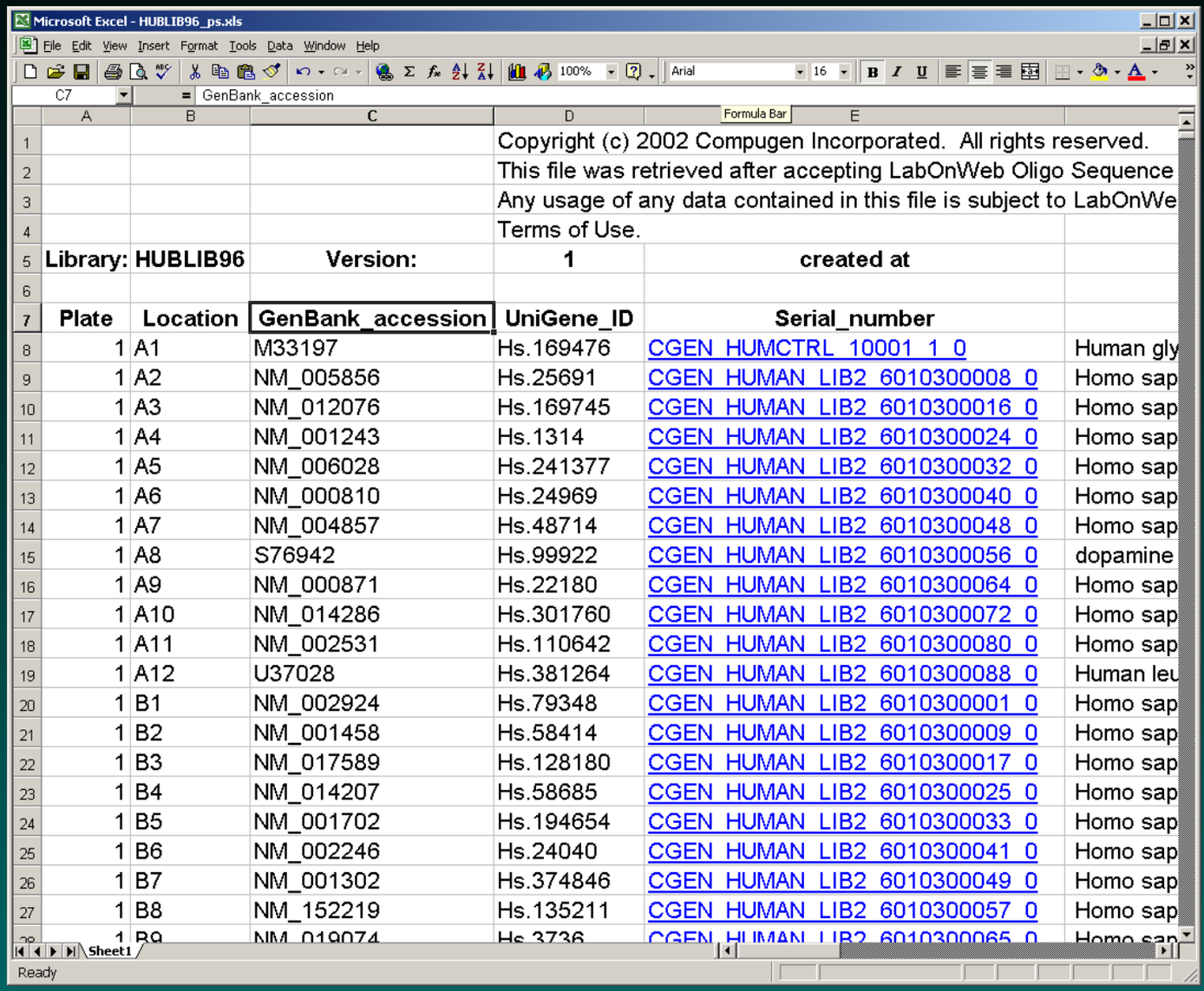
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Sigma-Genosys Product List

Catalog #	Description	Downloads	
		Gene list	Product info (PDF)
BACLIB96	B. Subtilis OligoLibrary: 96 well format, 43 plates (4,128 oligos)	Excel, 2 MB Zip, 262 KB	Info sheet Product profile
BACLIB384	B. Subtilis OligoLibrary: 384 well format, 11 plates (4,128 oligos)	Excel, 2 MB Zip, 285 KB	Info sheet Product profile
HUBLIB96	Human OligoLibrary Release 2.0: 96 well format, 304 plates (29,134 oligos)	Excel, 14 MB Zip, 2 MB	Info sheet Product profile

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Sigma-Genosys Oligo List



Microsoft Excel - HUBLIB96_ps.xls

File Edit View Insert Format Tools Data Window Help

C7 = GenBank_accession

Formula Bar

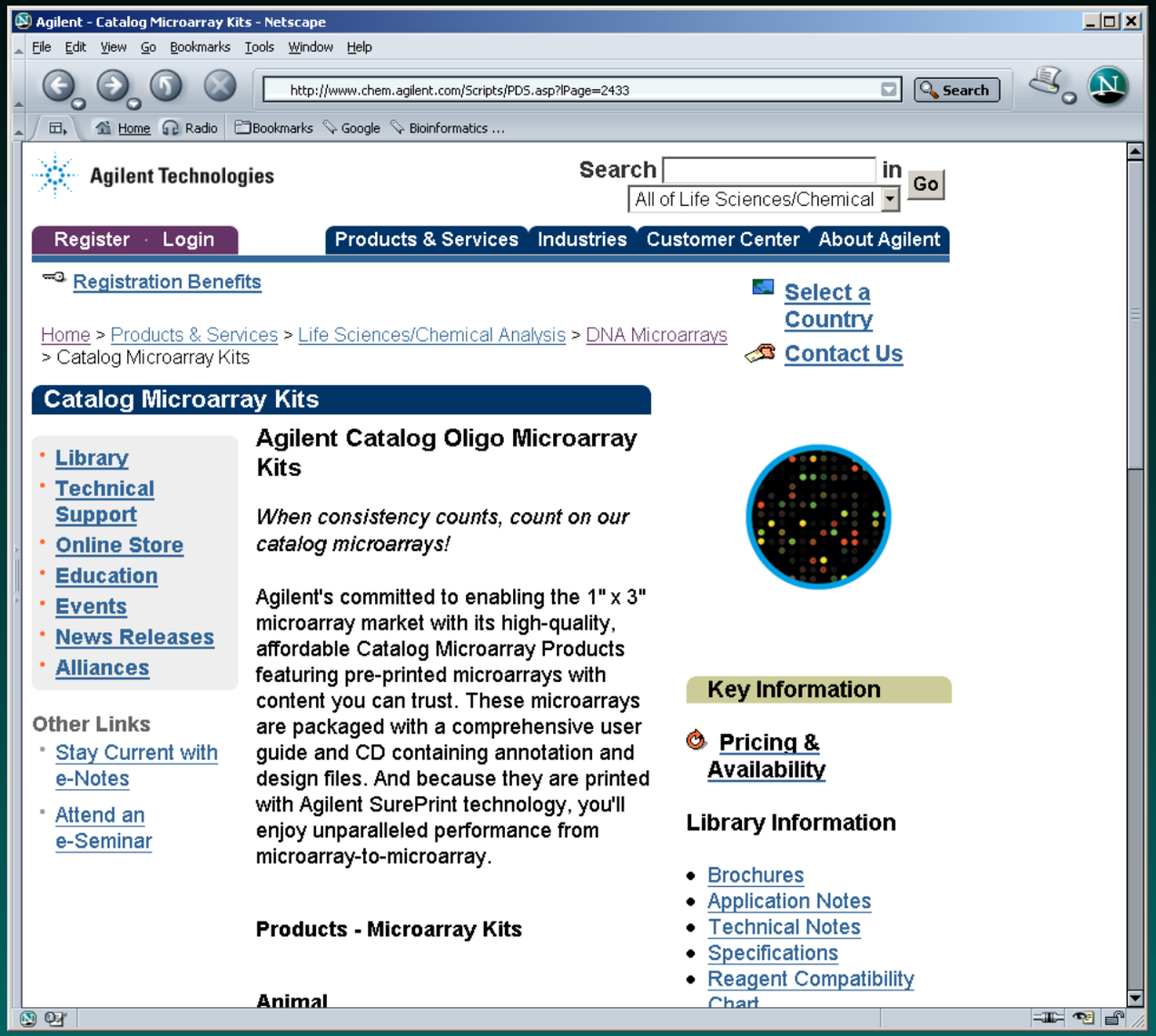
	A	B	C	D	E
1				Copyright (c) 2002 Compugen Incorporated. All rights reserved.	
2				This file was retrieved after accepting LabOnWeb Oligo Sequence	
3				Any usage of any data contained in this file is subject to LabOnWe	
4				Terms of Use.	
5	Library:	HUBLIB96	Version:	1	created at
6					
7	Plate	Location	GenBank_accession	UniGene_ID	Serial_number
8	1	A1	M33197	Hs.169476	CGEN_HUMCTRL_10001_1_0 Human gly
9	1	A2	NM_005856	Hs.25691	CGEN_HUMAN_LIB2_6010300008_0 Homo sap
10	1	A3	NM_012076	Hs.169745	CGEN_HUMAN_LIB2_6010300016_0 Homo sap
11	1	A4	NM_001243	Hs.1314	CGEN_HUMAN_LIB2_6010300024_0 Homo sap
12	1	A5	NM_006028	Hs.241377	CGEN_HUMAN_LIB2_6010300032_0 Homo sap
13	1	A6	NM_000810	Hs.24969	CGEN_HUMAN_LIB2_6010300040_0 Homo sap
14	1	A7	NM_004857	Hs.48714	CGEN_HUMAN_LIB2_6010300048_0 Homo sap
15	1	A8	S76942	Hs.99922	CGEN_HUMAN_LIB2_6010300056_0 dopamine
16	1	A9	NM_000871	Hs.22180	CGEN_HUMAN_LIB2_6010300064_0 Homo sap
17	1	A10	NM_014286	Hs.301760	CGEN_HUMAN_LIB2_6010300072_0 Homo sap
18	1	A11	NM_002531	Hs.110642	CGEN_HUMAN_LIB2_6010300080_0 Homo sap
19	1	A12	U37028	Hs.381264	CGEN_HUMAN_LIB2_6010300088_0 Human leu
20	1	B1	NM_002924	Hs.79348	CGEN_HUMAN_LIB2_6010300001_0 Homo sap
21	1	B2	NM_001458	Hs.58414	CGEN_HUMAN_LIB2_6010300009_0 Homo sap
22	1	B3	NM_017589	Hs.128180	CGEN_HUMAN_LIB2_6010300017_0 Homo sap
23	1	B4	NM_014207	Hs.58685	CGEN_HUMAN_LIB2_6010300025_0 Homo sap
24	1	B5	NM_001702	Hs.194654	CGEN_HUMAN_LIB2_6010300033_0 Homo sap
25	1	B6	NM_002246	Hs.24040	CGEN_HUMAN_LIB2_6010300041_0 Homo sap
26	1	B7	NM_001302	Hs.374846	CGEN_HUMAN_LIB2_6010300049_0 Homo sap
27	1	B8	NM_152219	Hs.135211	CGEN_HUMAN_LIB2_6010300057_0 Homo sap
28	1	B9	NM_019074	Hs.3736	CGEN_HUMAN_LIB2_6010300065_0 Homo sap

Sheet1

Ready

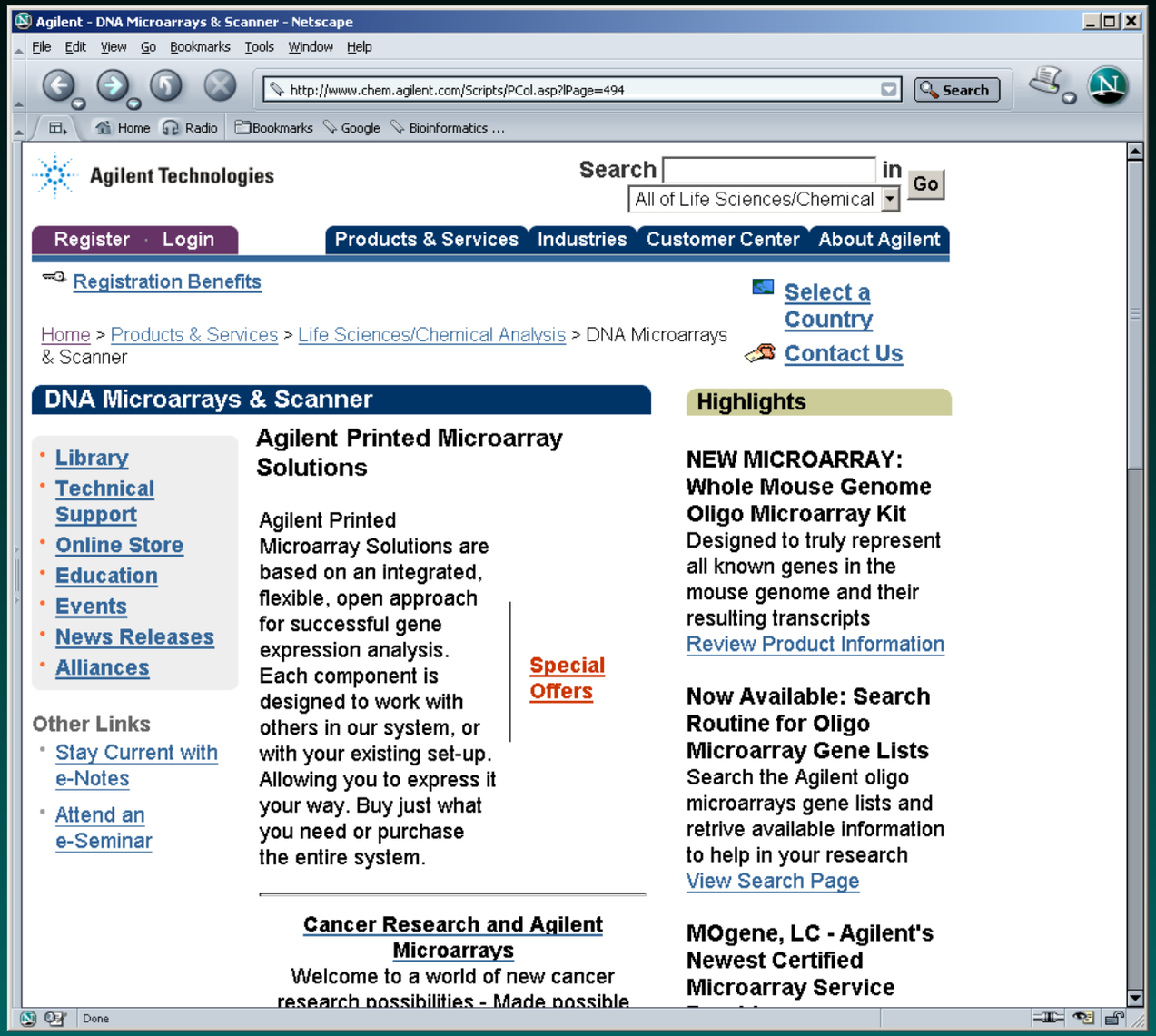
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

http://www.chem.agilent.com



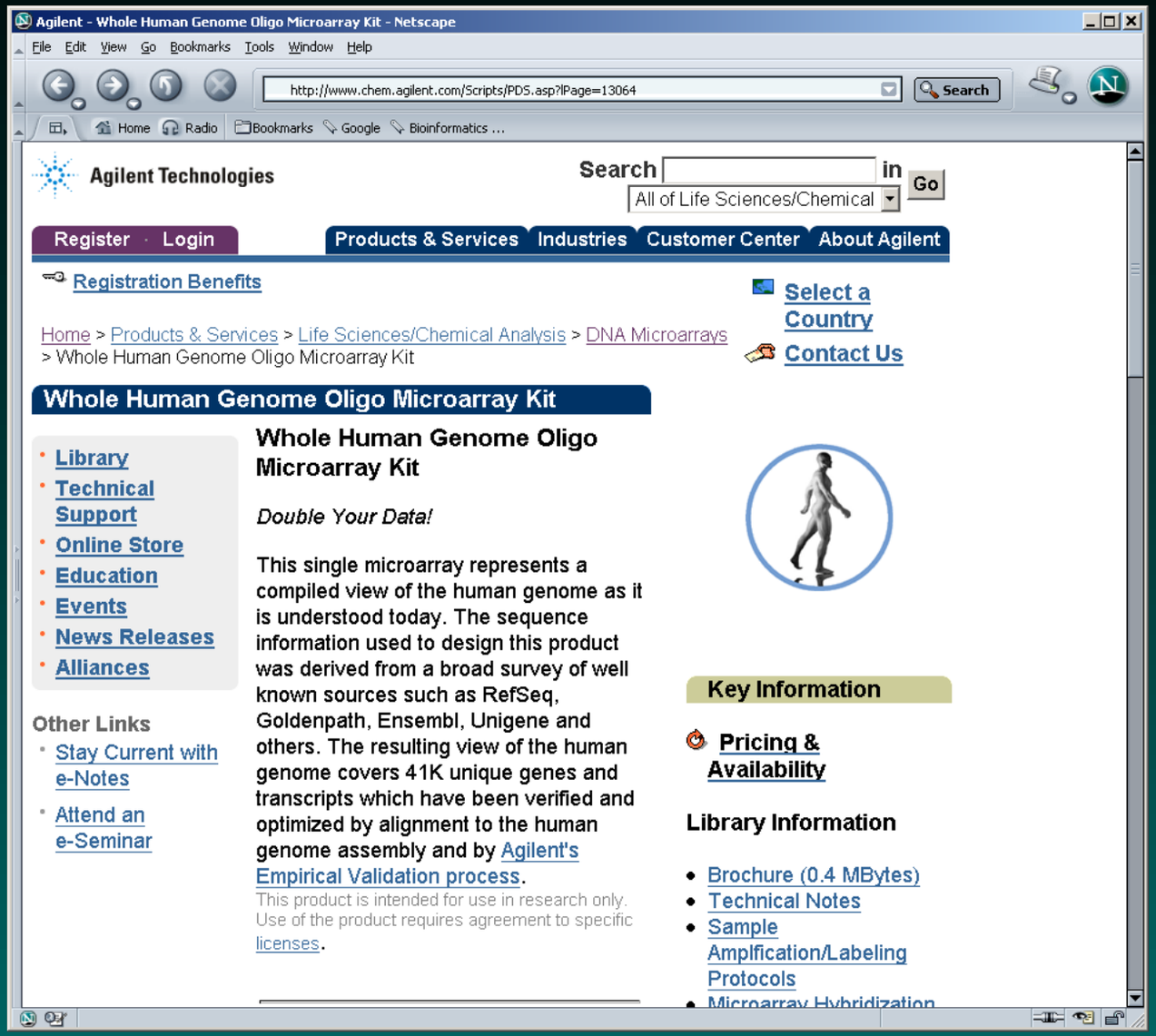
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Agilent Microarray



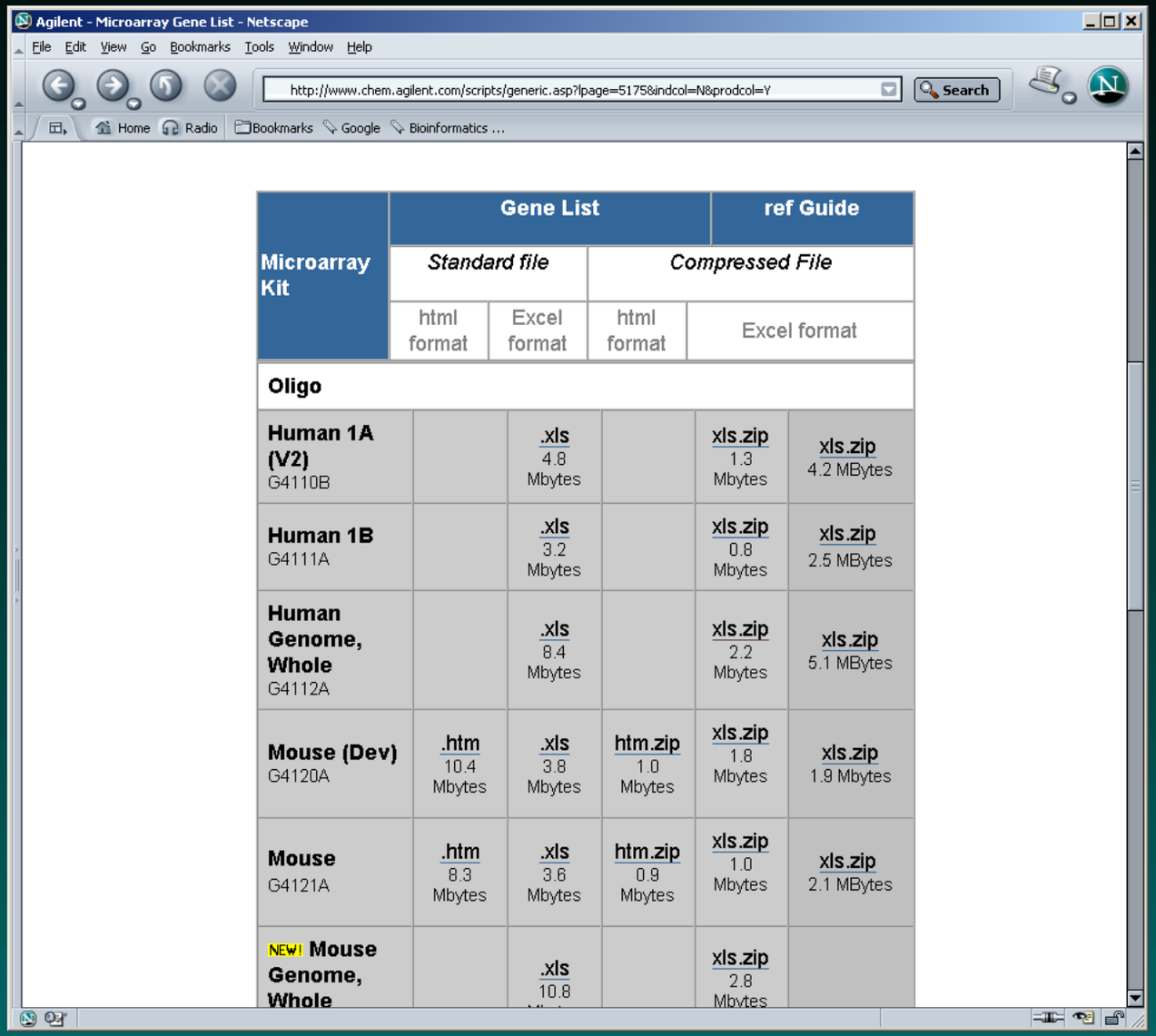
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Agilent Human Microarrays



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

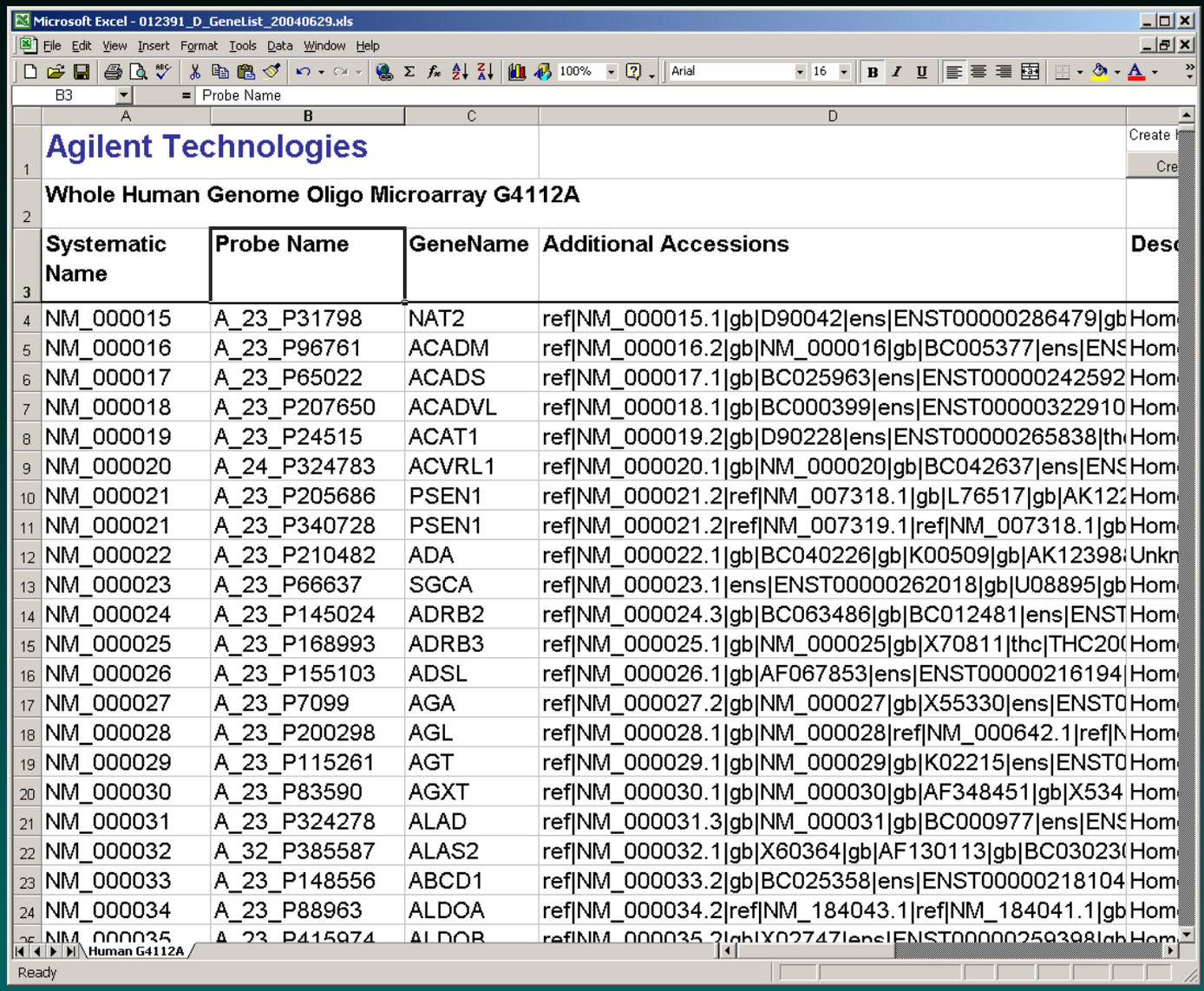
Agilent Product List



Microarray Kit	Gene List			ref Guide	
	Standard file		*Compressed File*		
	html format	Excel format	html format	Excel format	
Oligo					
Human 1A (V2) G4110B		.xls 4.8 Mbytes		xls.zip 1.3 Mbytes	xls.zip 4.2 MBytes
Human 1B G4111A		.xls 3.2 Mbytes		xls.zip 0.8 Mbytes	xls.zip 2.5 MBytes
Human Genome, Whole G4112A		.xls 8.4 Mbytes		xls.zip 2.2 Mbytes	xls.zip 5.1 MBytes
Mouse (Dev) G4120A	.htm 10.4 Mbytes	.xls 3.8 Mbytes	htm.zip 1.0 Mbytes	xls.zip 1.8 Mbytes	xls.zip 1.9 Mbytes
Mouse G4121A	.htm 8.3 Mbytes	.xls 3.6 Mbytes	htm.zip 0.9 Mbytes	xls.zip 1.0 Mbytes	xls.zip 2.1 MBytes
NEW! **Mouse Genome, Whole**		.xls 10.8		xls.zip 2.8 Mbytes	

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Agilent Oligo List



Microsoft Excel - 012391_D_GeneList_20040629.xls

File Edit View Insert Format Tools Data Window Help

B3 = Probe Name

Agilent Technologies

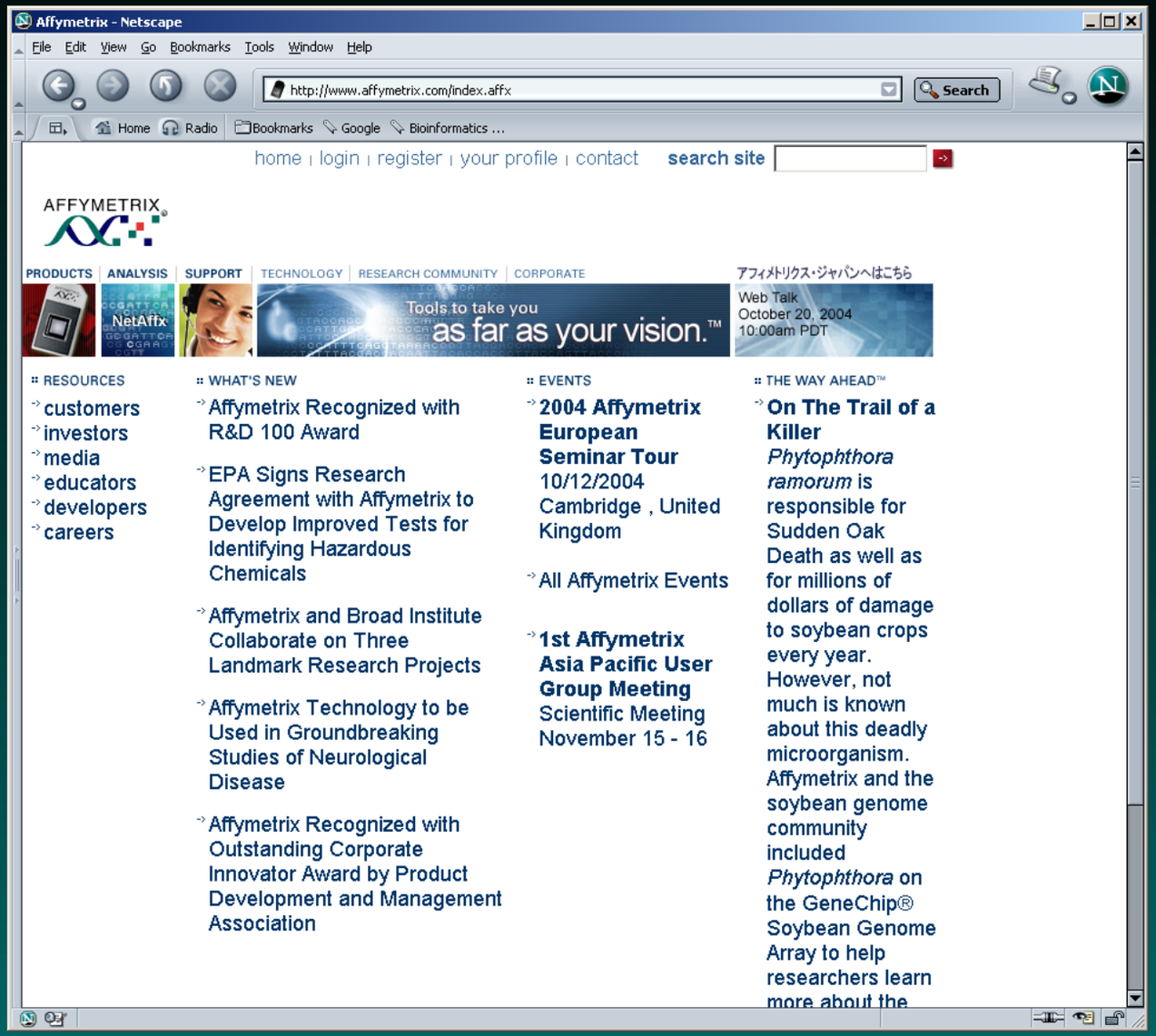
Whole Human Genome Oligo Microarray G4112A

Systematic Name	Probe Name	GeneName	Additional Accessions	Desc
NM_000015	A_23_P31798	NAT2	ref|NM_000015.1|gb|D90042|ens|ENST00000286479|gb	Hom
NM_000016	A_23_P96761	ACADM	ref|NM_000016.2|gb|NM_000016|gb|BC005377|ens|ENS	Hom
NM_000017	A_23_P65022	ACADS	ref|NM_000017.1|gb|BC025963|ens|ENST00000242592	Hom
NM_000018	A_23_P207650	ACADVL	ref|NM_000018.1|gb|BC000399|ens|ENST00000322910	Hom
NM_000019	A_23_P24515	ACAT1	ref|NM_000019.2|gb|D90228|ens|ENST00000265838|th	Hom
NM_000020	A_24_P324783	ACVRL1	ref|NM_000020.1|gb|NM_000020|gb|BC042637|ens|ENS	Hom
NM_000021	A_23_P205686	PSEN1	ref|NM_000021.2|ref|NM_007318.1|gb|L76517|gb|AK122	Hom
NM_000021	A_23_P340728	PSEN1	ref|NM_000021.2|ref|NM_007319.1|ref|NM_007318.1|gb	Hom
NM_000022	A_23_P210482	ADA	ref|NM_000022.1|gb|BC040226|gb|K00509|gb|AK12398	Unkn
NM_000023	A_23_P66637	SGCA	ref|NM_000023.1|ens|ENST00000262018|gb|U08895|gb	Hom
NM_000024	A_23_P145024	ADRB2	ref|NM_000024.3|gb|BC063486|gb|BC012481|ens|ENST	Hom
NM_000025	A_23_P168993	ADRB3	ref|NM_000025.1|gb|NM_000025|gb|X70811|thc|THC20(	Hom
NM_000026	A_23_P155103	ADSL	ref|NM_000026.1|gb|AF067853|ens|ENST00000216194|	Hom
NM_000027	A_23_P7099	AGA	ref|NM_000027.2|gb|NM_000027|gb|X55330|ens|ENST0	Hom
NM_000028	A_23_P200298	AGL	ref|NM_000028.1|gb|NM_000028|ref|NM_000642.1|ref|N	Hom
NM_000029	A_23_P115261	AGT	ref|NM_000029.1|gb|NM_000029|gb|K02215|ens|ENST0	Hom
NM_000030	A_23_P83590	AGXT	ref|NM_000030.1|gb|NM_000030|gb|AF348451|gb|X534	Hom
NM_000031	A_23_P324278	ALAD	ref|NM_000031.3|gb|NM_000031|gb|BC000977|ens|ENS	Hom
NM_000032	A_32_P385587	ALAS2	ref|NM_000032.1|gb|X60364|gb|AF130113|gb|BC03023(	Hom
NM_000033	A_23_P148556	ABCD1	ref|NM_000033.2|gb|BC025358|ens|ENST00000218104	Hom
NM_000034	A_23_P88963	ALDOA	ref|NM_000034.2|ref|NM_184043.1|ref|NM_184041.1|gb	Hom
NM_000035	A_23_P415974	ALDOB	ref|NM_000035.2|gb|X02747|ens|ENST00000259398|gb	Hom

Human G4112A

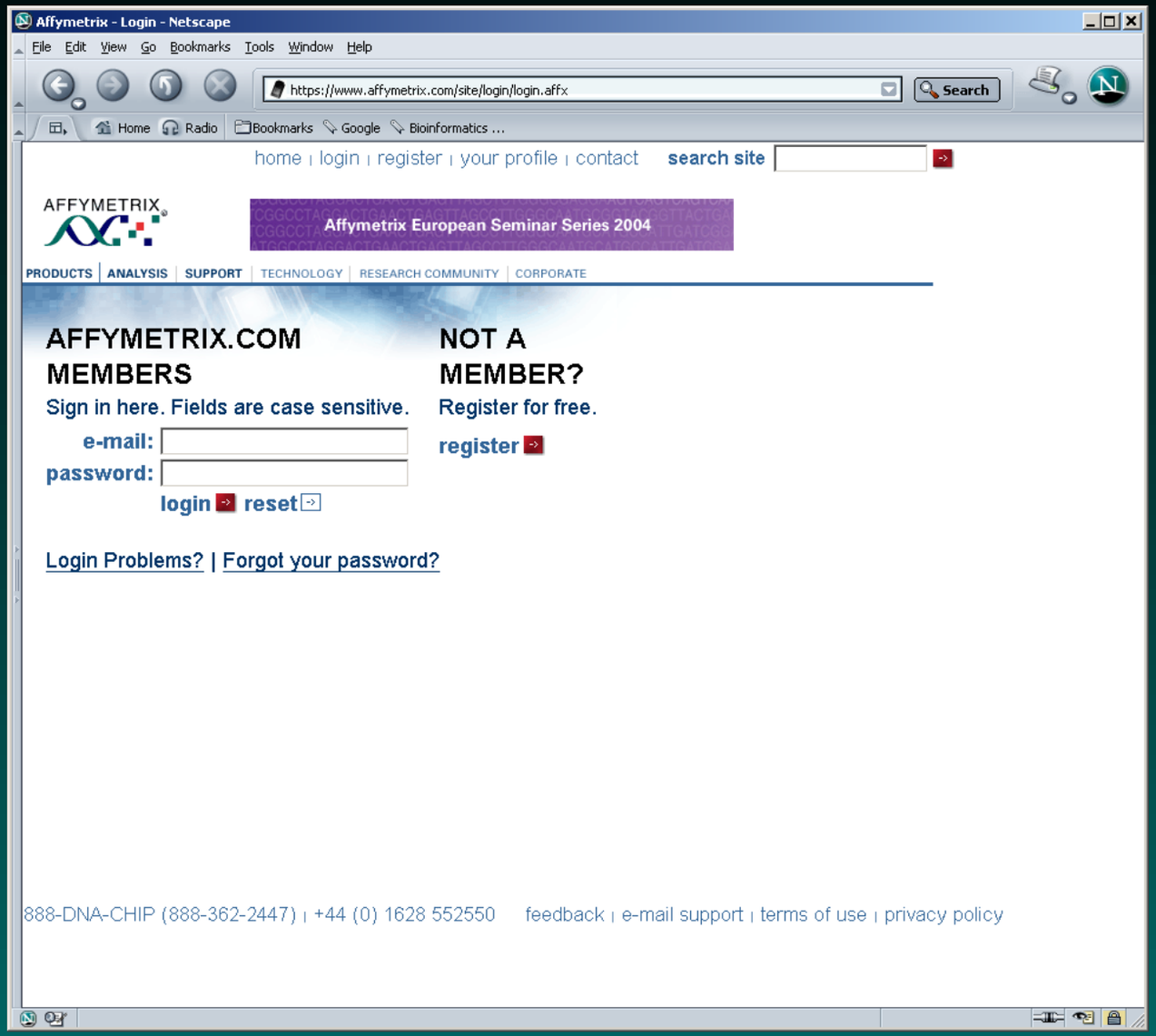
Ready

http://www.affymetrix.com



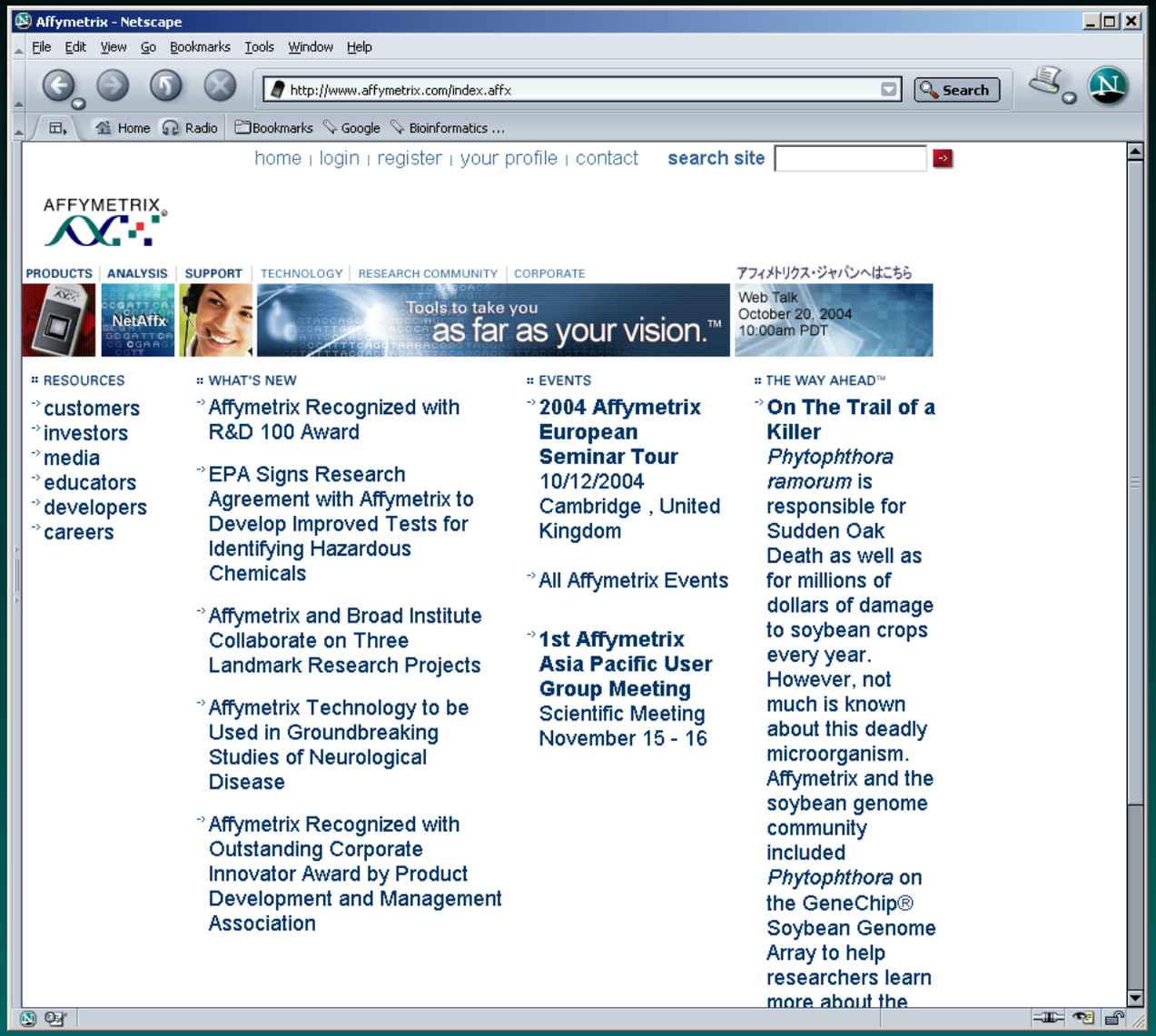
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix login page



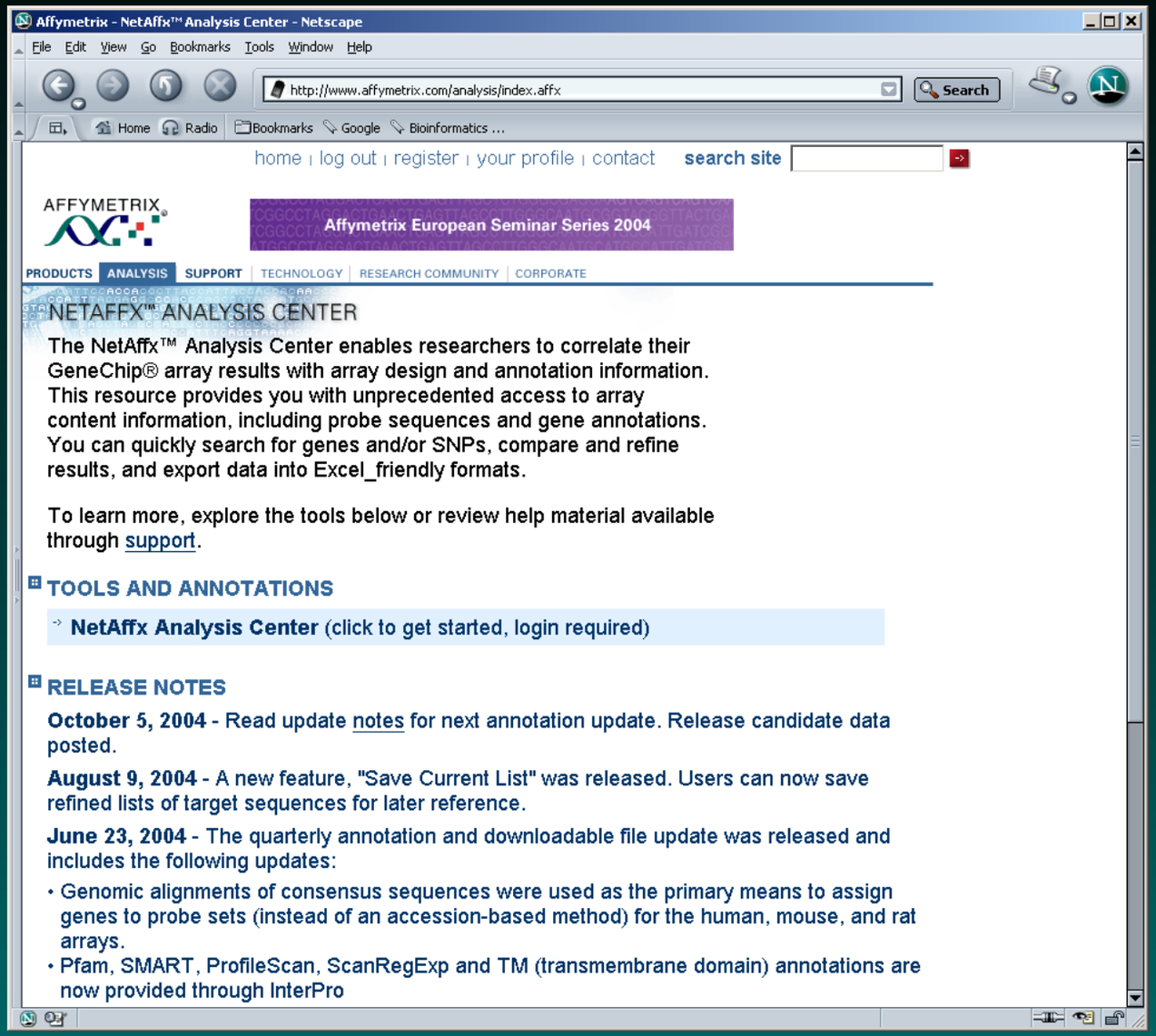
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Return to the home page



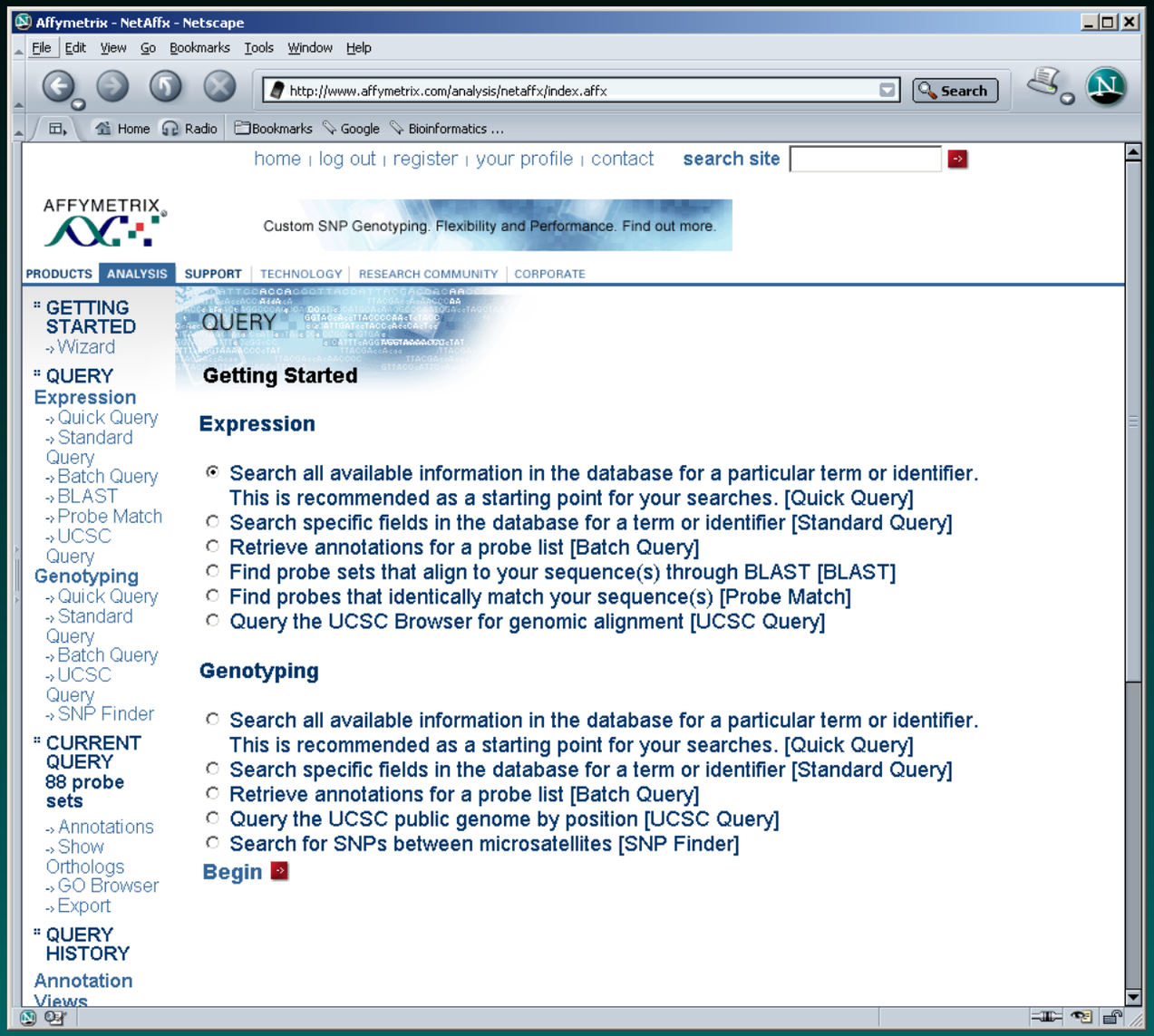
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Main page for NetAffx



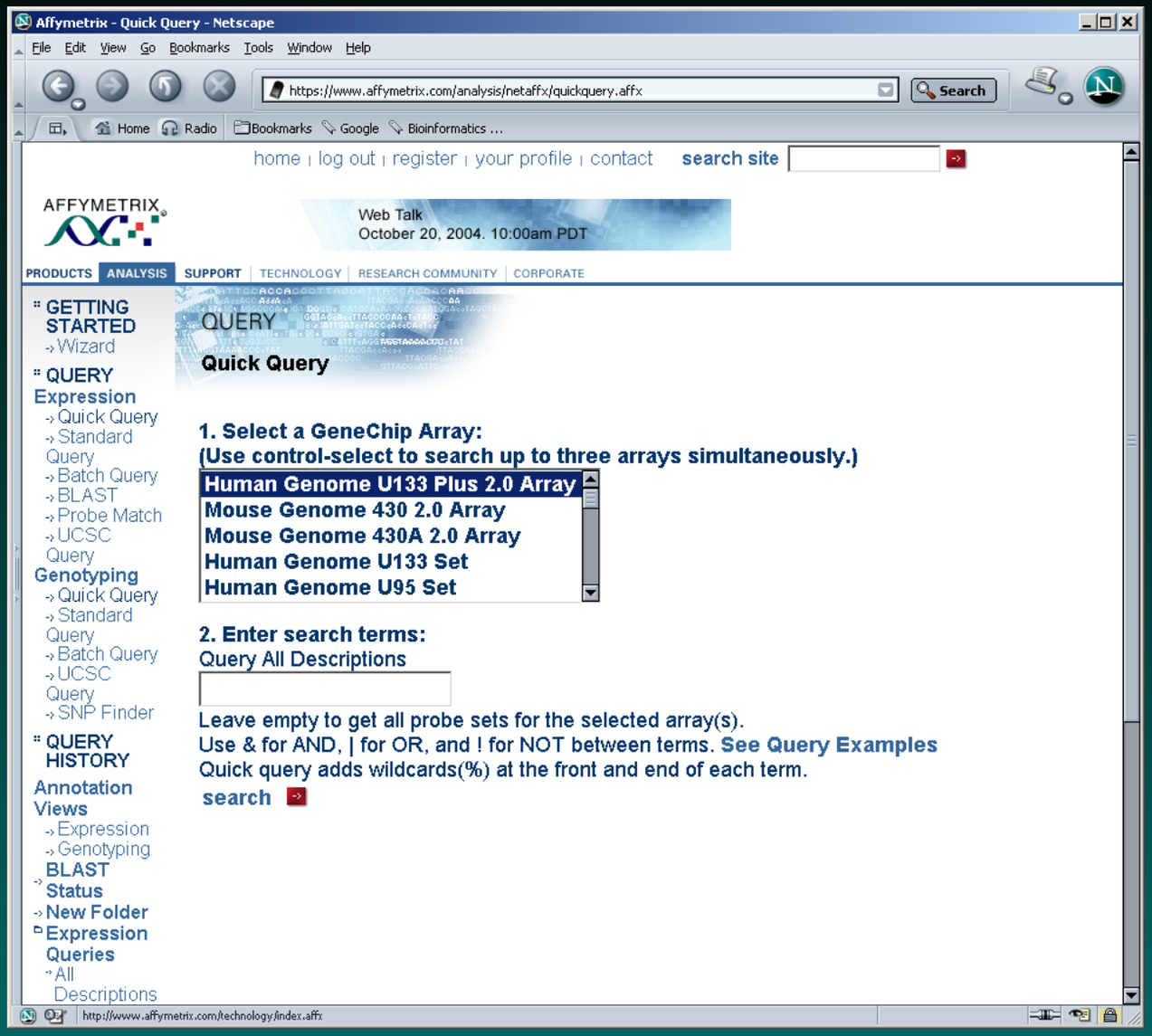
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Analysis Central



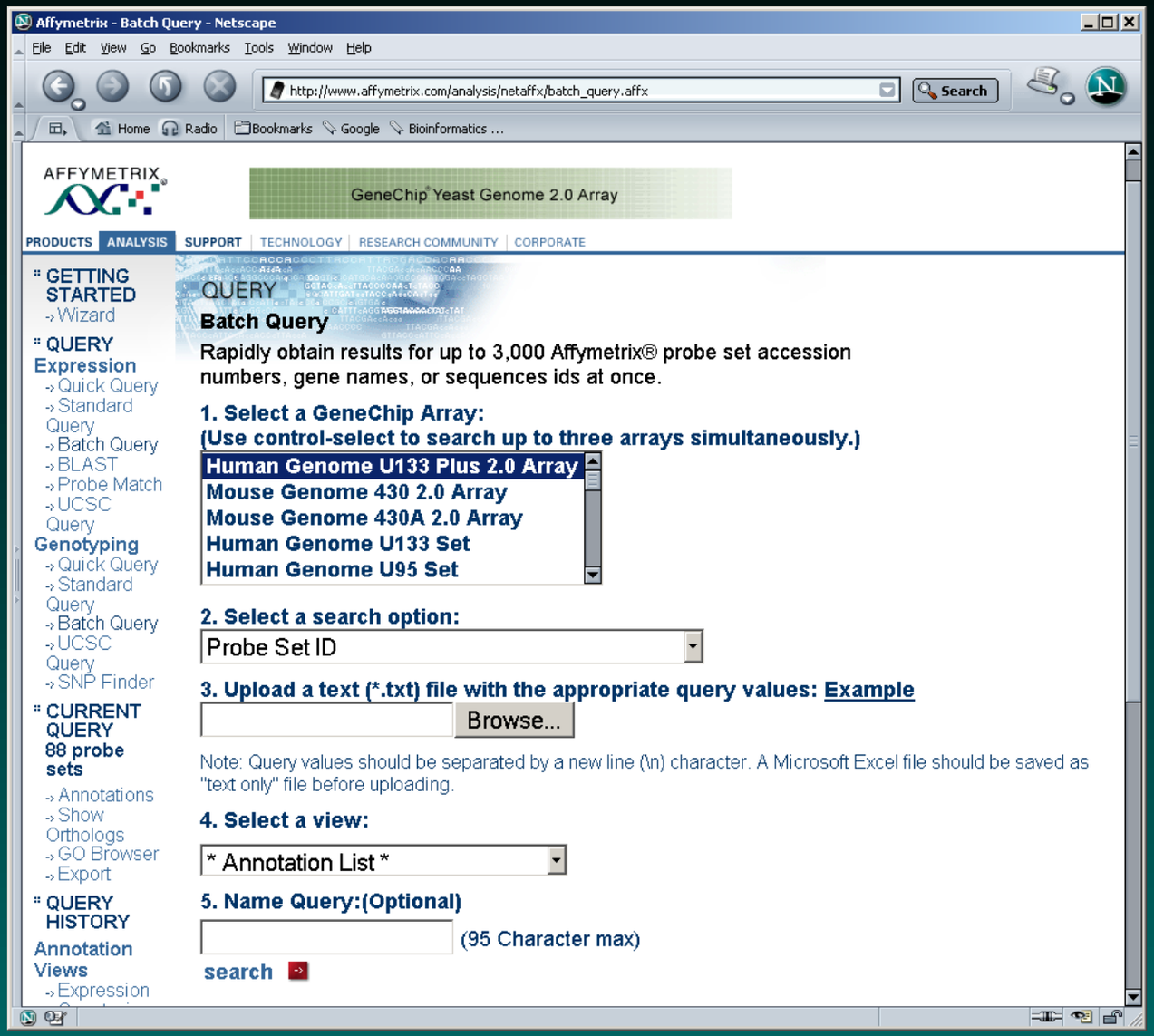
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Quick Search



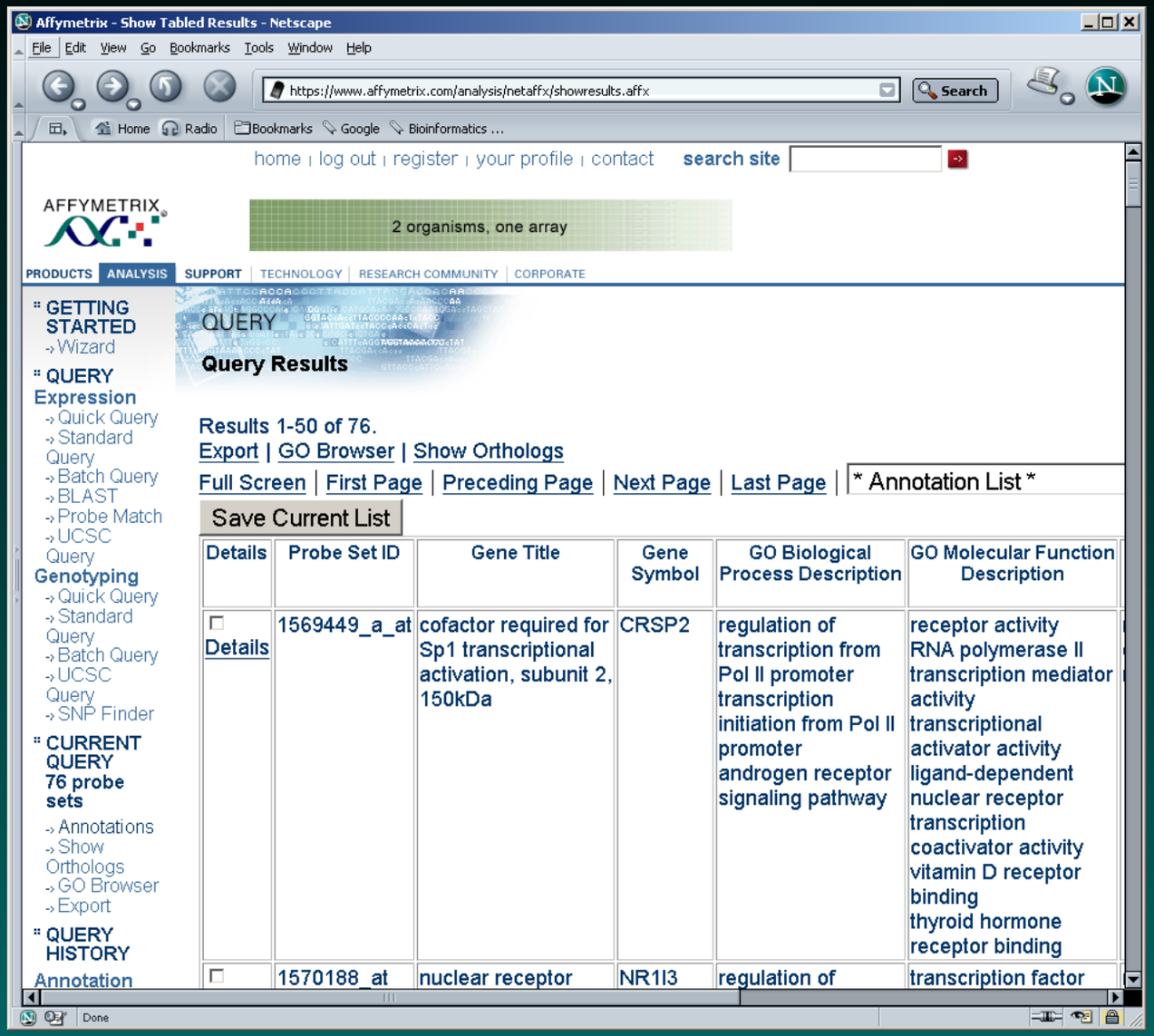
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Batch Search



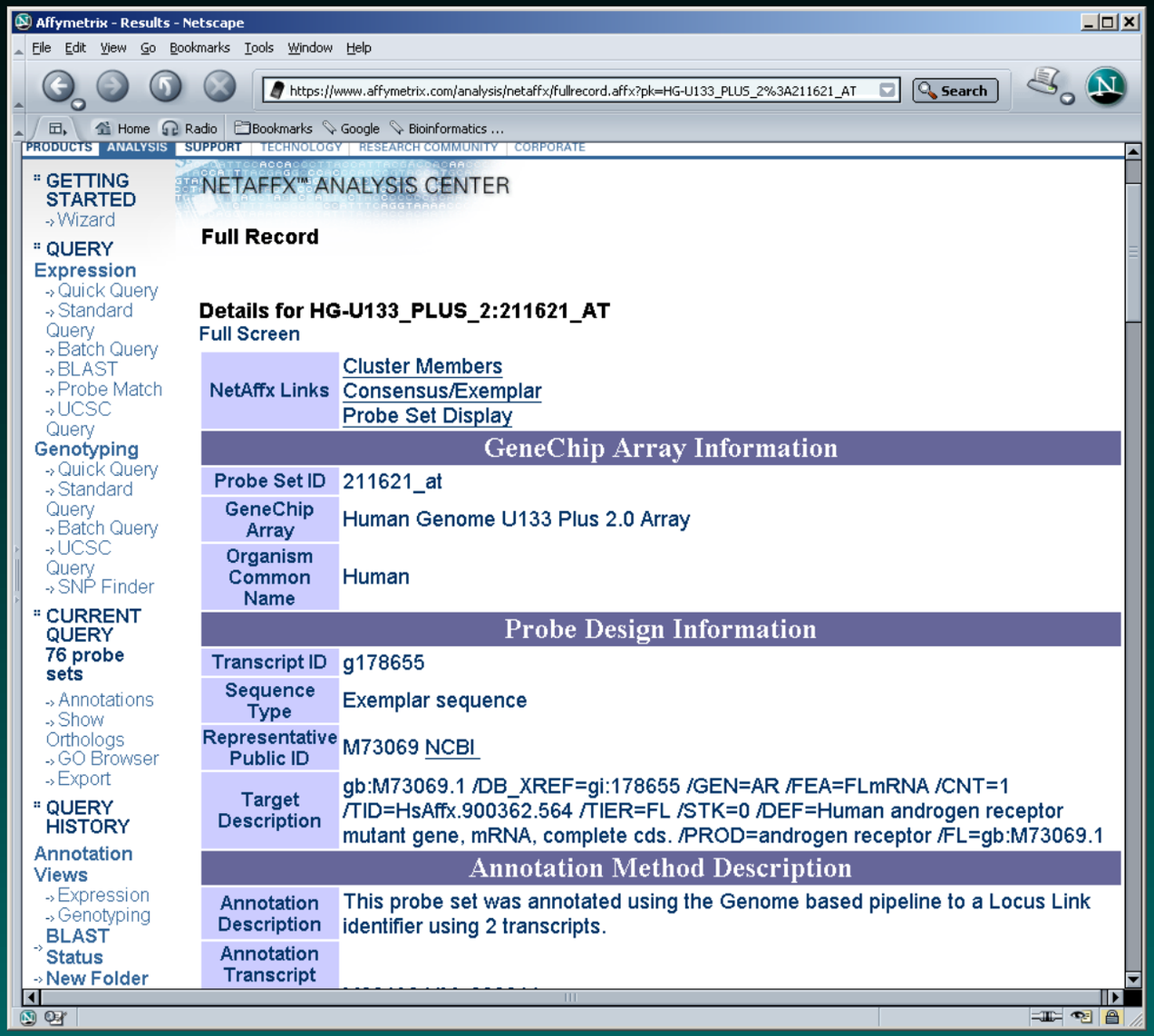
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Search Results



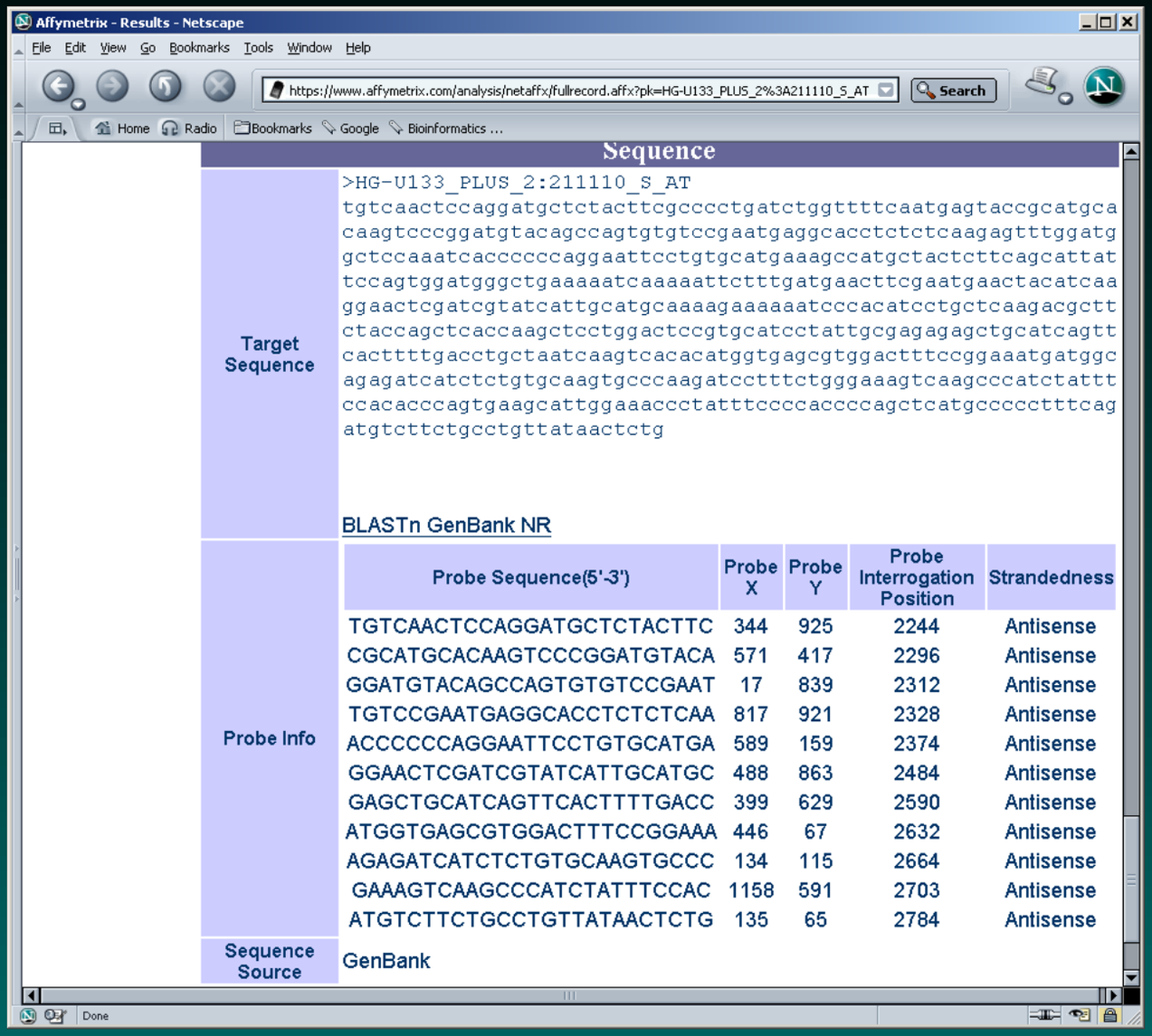
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Detailed Search Results



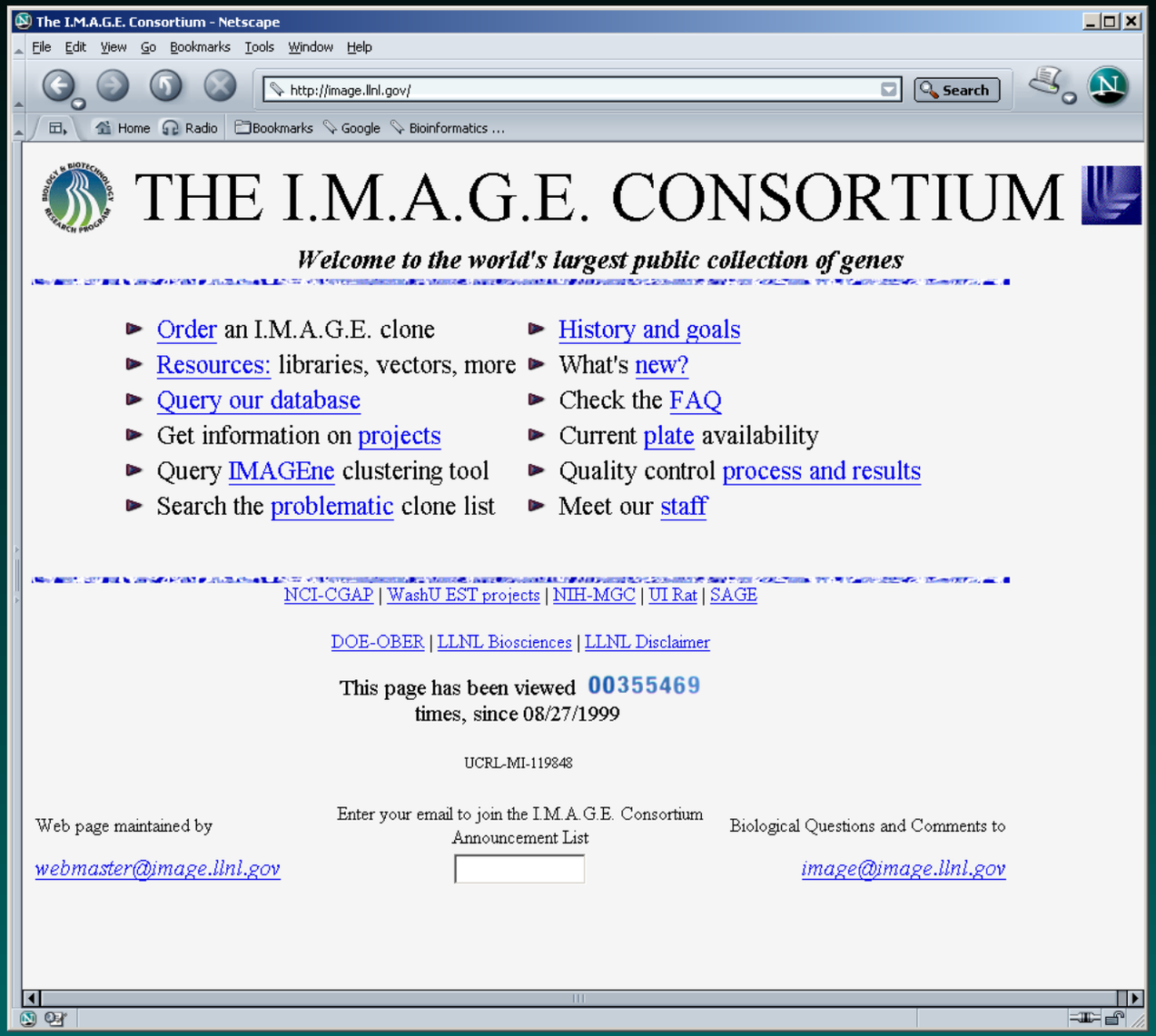
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Affymetrix Sequence Information



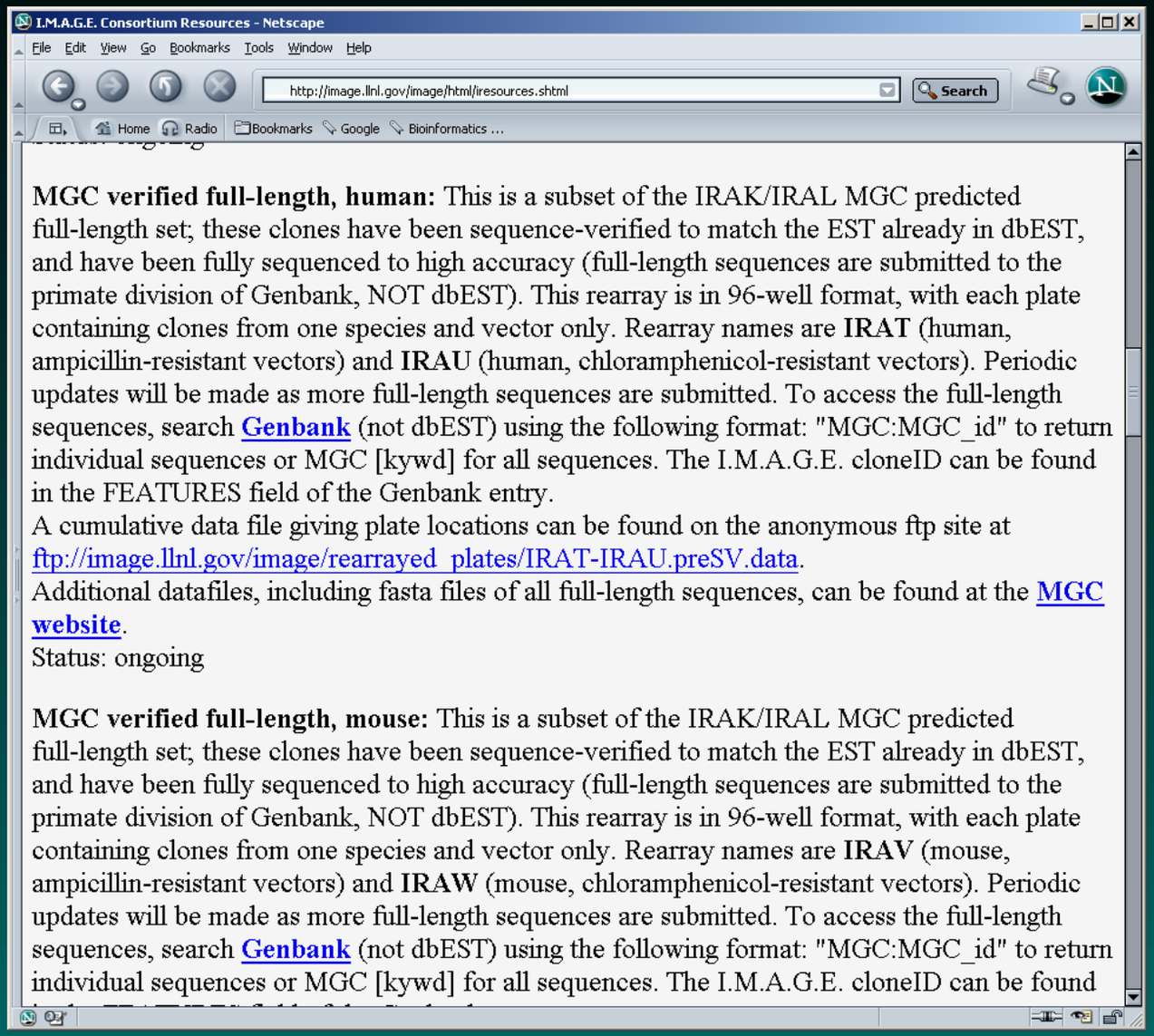
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

http://image.lnl.gov



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

IMAGE is integrated into GenBank



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GenBank: the common denominator

You may have noticed that all the commercial products map their propietary identifiers to GenBank. GenBank is the primary repository for sequence information.

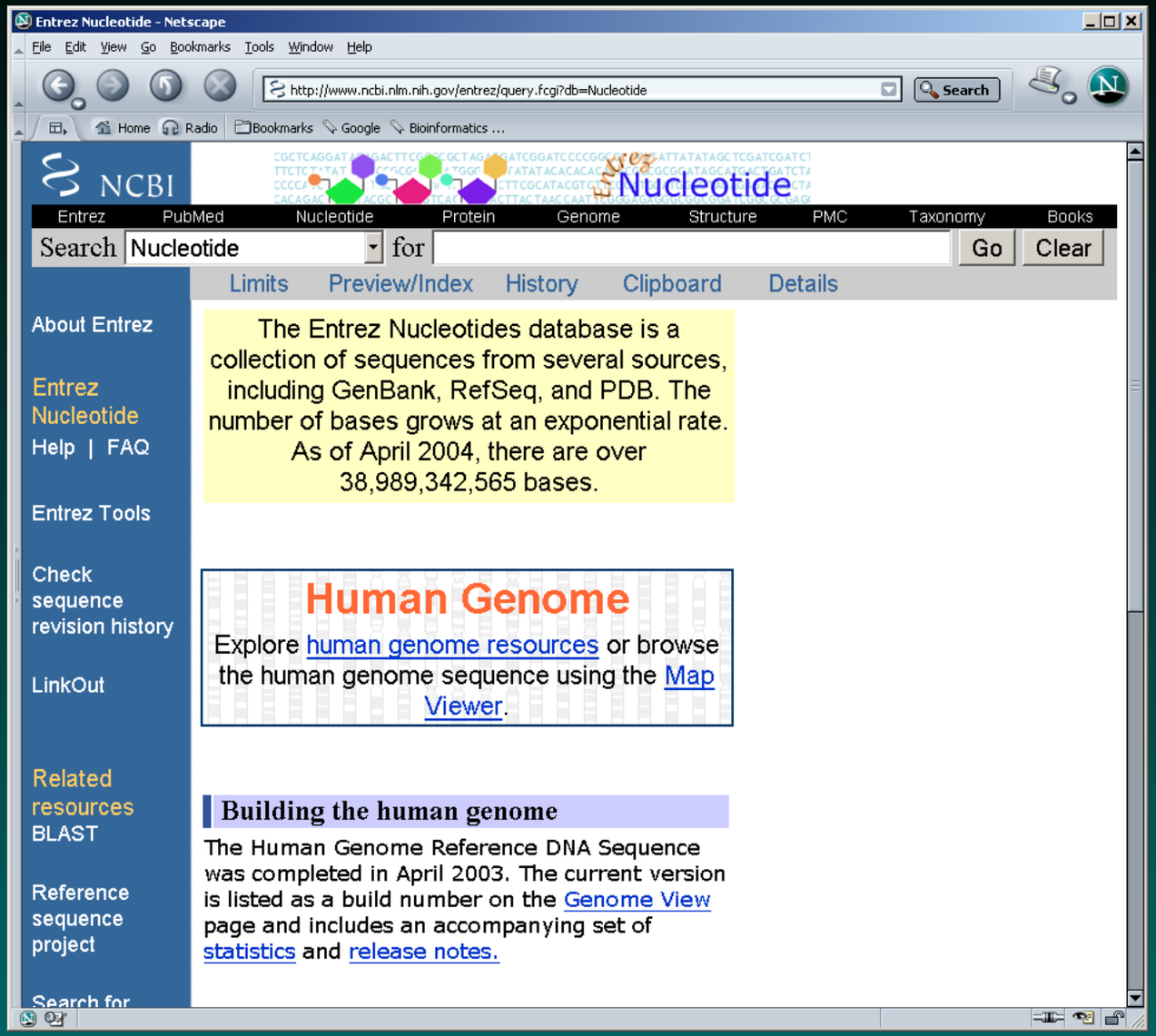
While the IMAGE ID refers to the actual clone, the corresponding GenBank entry (or entries) describes the actual mRNA sequence.

For Affymetrix, the GenBank entry describes the sequence from which the probe set was constructed.

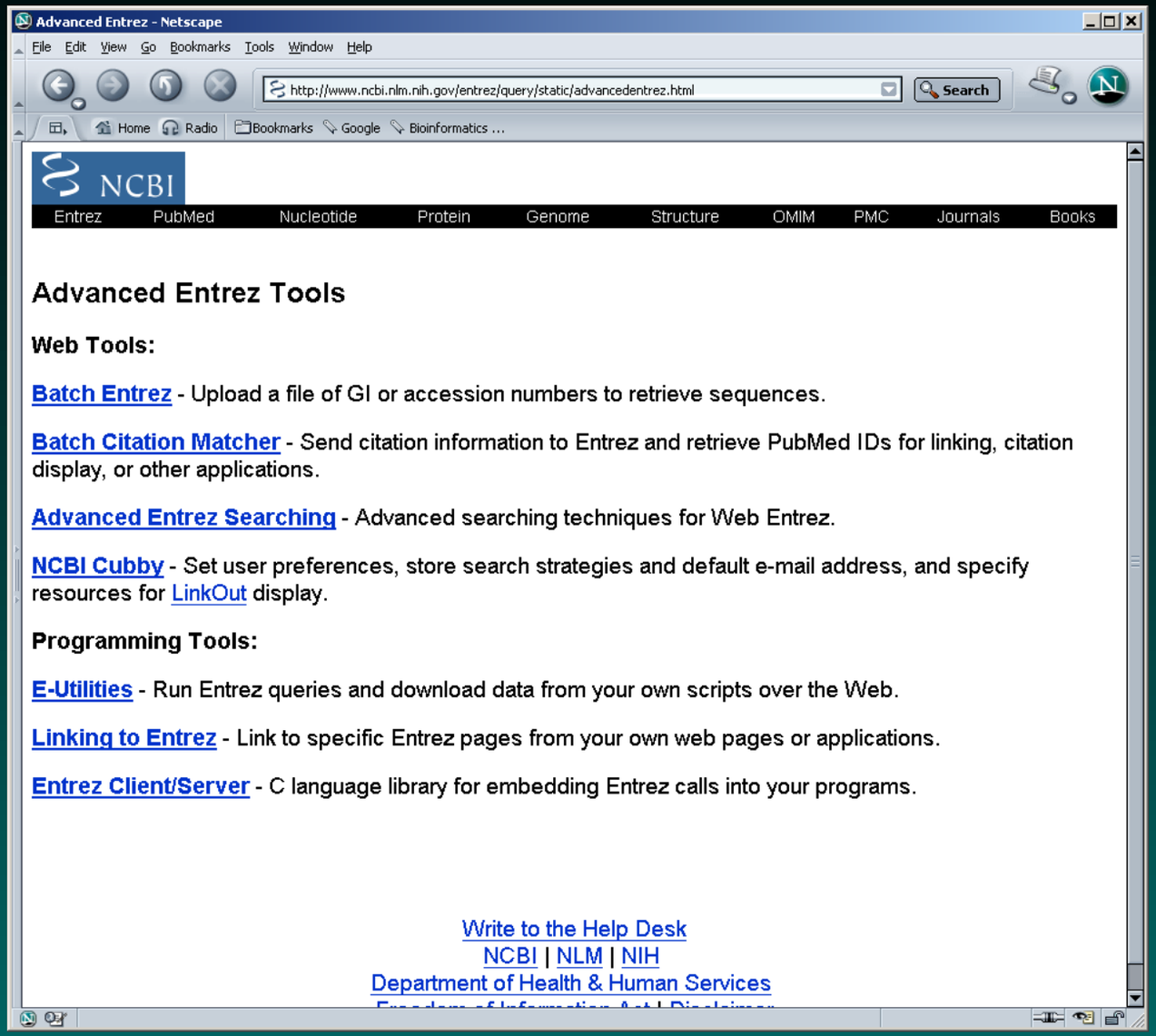
For Sigma-Genosys or Agilent, the GenBank entry describes the sequence from which the long oligo was selected.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Entrez Nucleotide

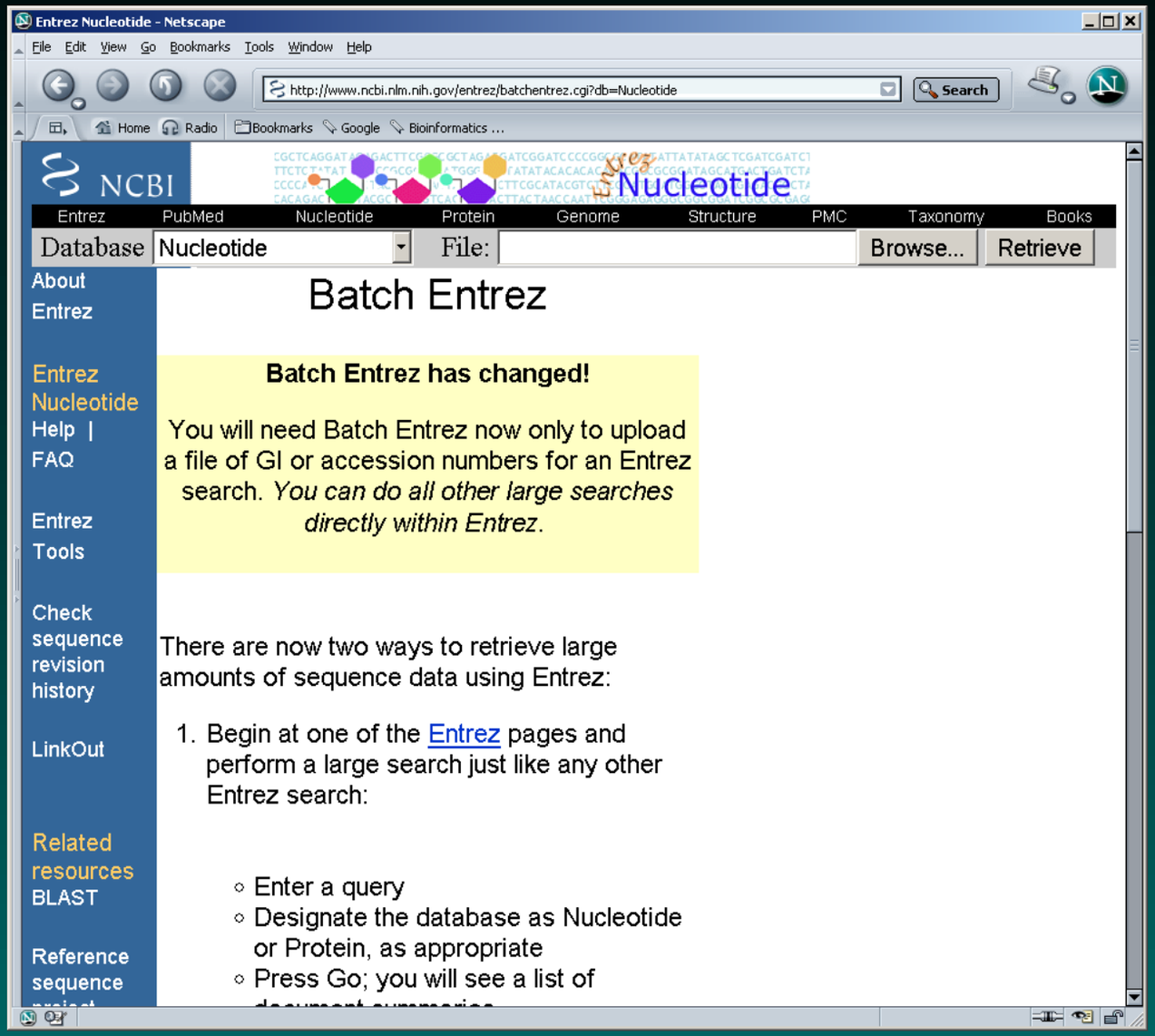


Entrez Tools



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Batch Entrez



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Getting a Gene List

```
> extreme <- abs(t.statistics)>6 & tr.stats > 20
> extreme.genes <- gene.info[extreme,]
> dim(extreme.genes)
[1] 1005    6
> extreme.genes[1:8,]
          Clone.ID Accession Gene.Symbol Cluster.ID
X24    IMAGE:34849    R20379        EEF2   Hs.75309
X27    IMAGE:45525    H08440         RFP  Hs.440382
X63   IMAGE:295831    N74602      CGI-26   Hs.24332
X82   IMAGE:143322    R74357                Hs.367688
X100  IMAGE:753381  AA406332      SEC23A  Hs.272927
X132  IMAGE:341083    W58562     C6orf56  Hs.102471
X205   IMAGE:78946    T61792        PDK4    Hs.8364
X225  IMAGE:586831  AA130866        TMLHE  Hs.133321
```

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

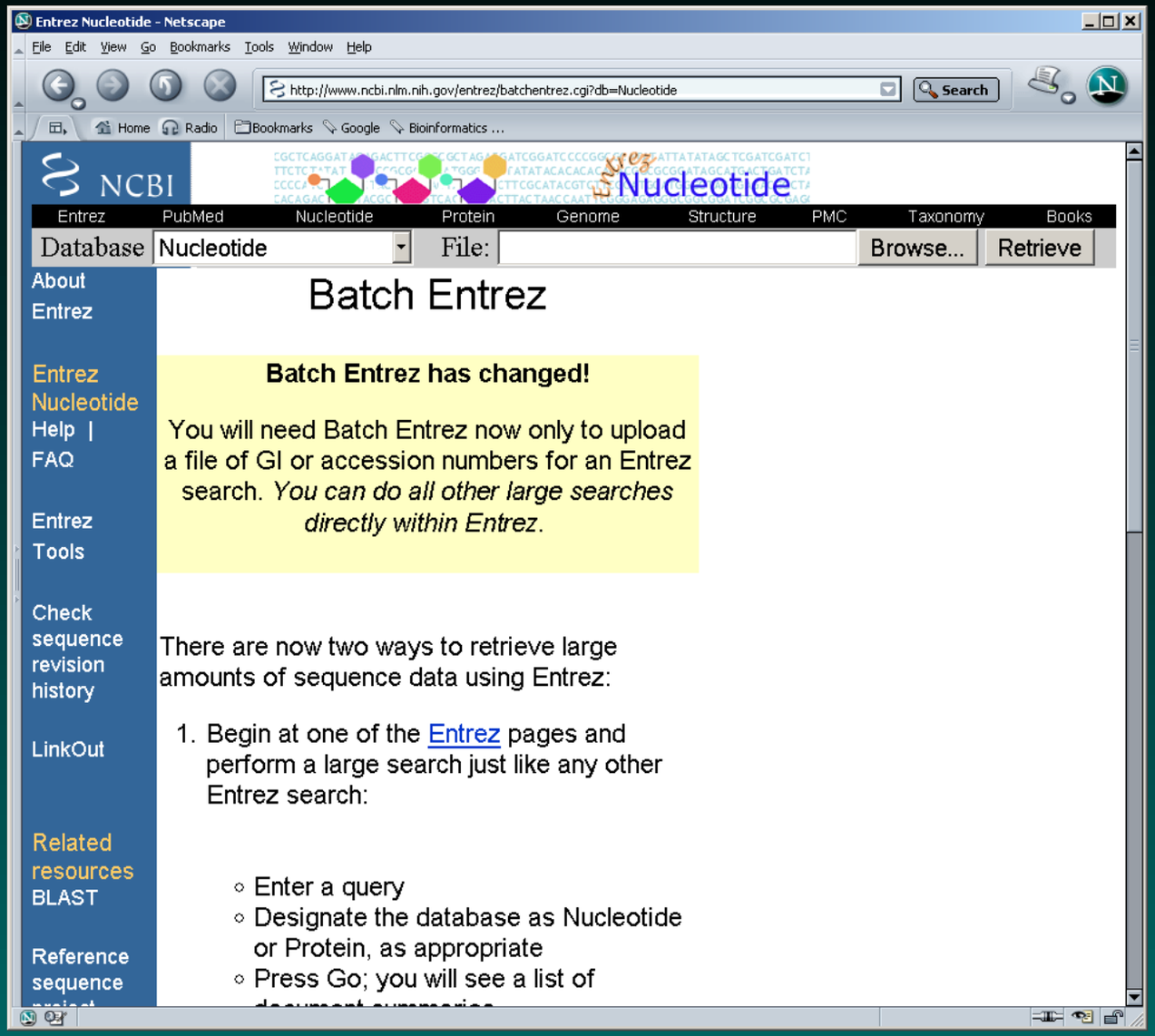
Just GenBank

```
> extreme.gb <- extreme.genes$Accession
> write.table(extreme.gb, 'extreme.txt',
+  quote=FALSE, row.names=FALSE, col.names=FALSE)
> write.table(extreme.genes, 'extreme-genes.txt',
+ quote=FALSE, row.names=FALSE, col.names=TRUE,
+ sep='\t')
```

We need the one-column list of GenBank identifiers for a batch search of GenBank. Other tools can use the more general list.

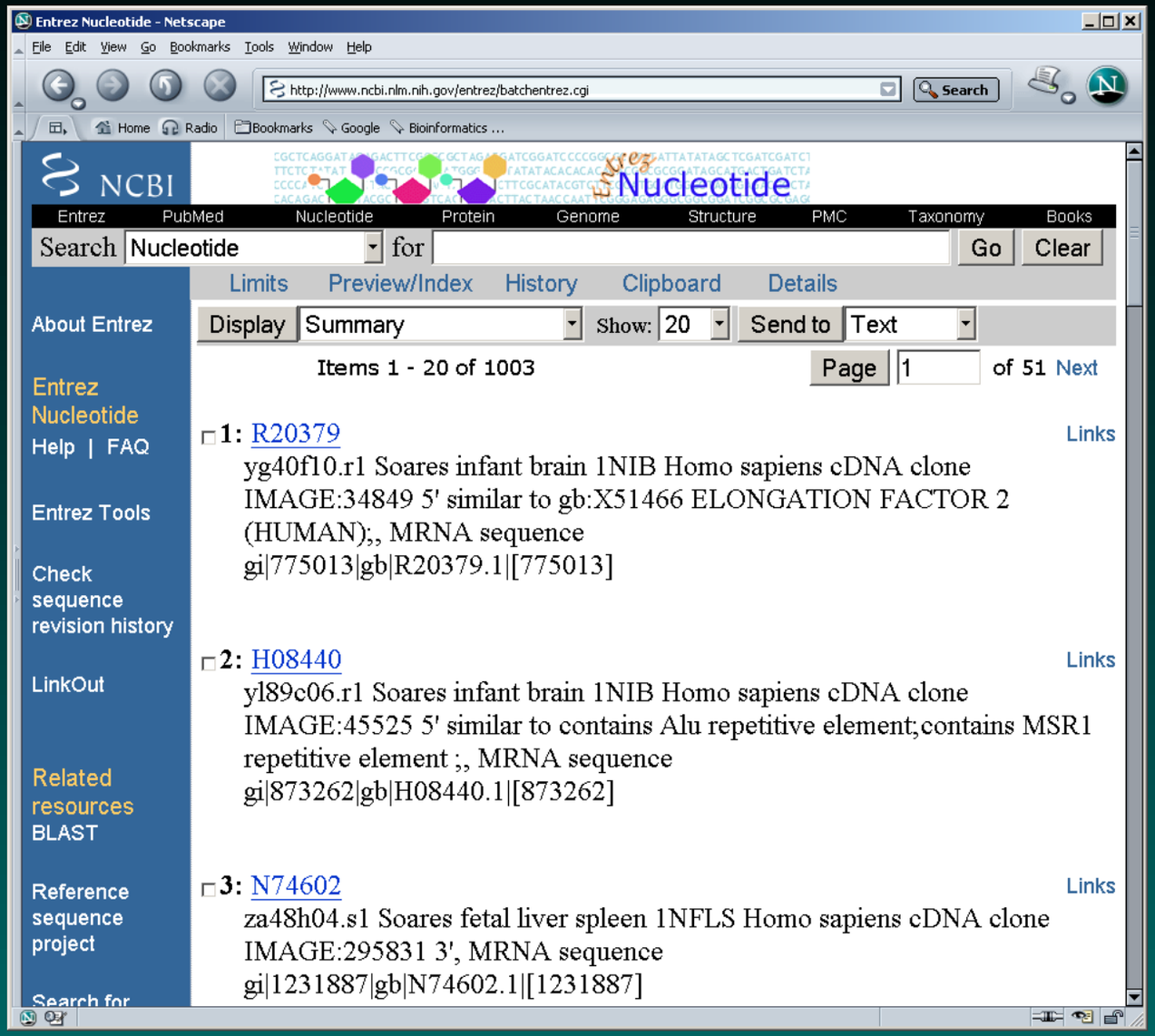
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Batch Entrez



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Entrez Results



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Entrez Results

Why did we only get 1003 results from a batch query with 1005 entries? Probably because two of the GenBank identifiers have been “retired” since the microarray was constructed. This typically happens when someone finds out there was a problem (often vector contamination) with the original sequence entry.

Note: To save the results, click the “Send to” button after first changing the destination to “File”.

We can use the results of the search to compare the IMAGE clone IDs supplied with the original array with the current record in GenBank. It is certainly possible that someone (in the distant past) typed one of the numbers incorrectly.

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene

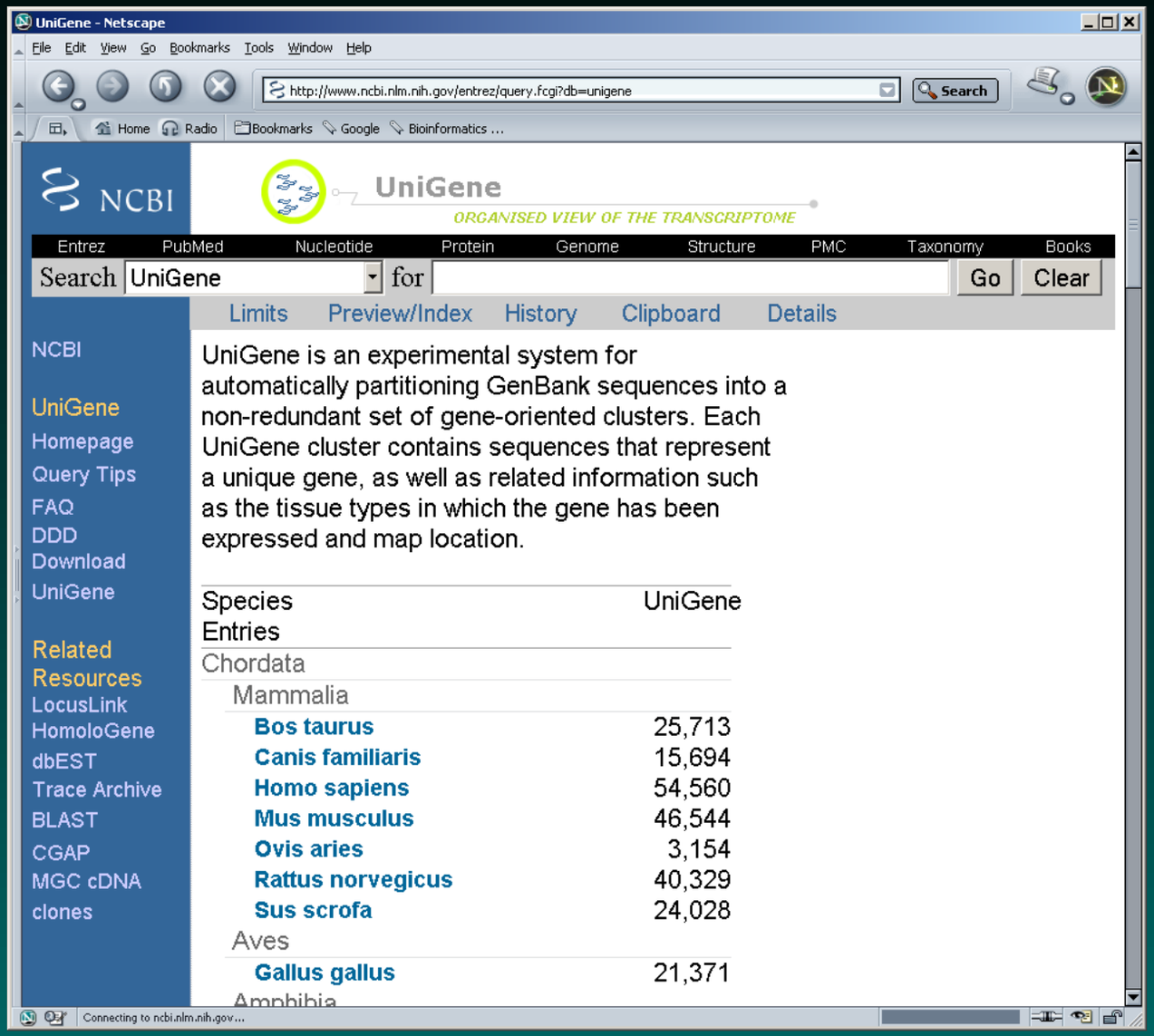
GenBank only refers to individual sequences. Because lots of people have sequenced lots of fragments of RNA and entered them into GenBank, a sequence entry is not the same thing as a gene.

UniGene is the NCBI's attempt to organize sequences into coherent clusters that should represent genes.

Critical Fact: UniGene changes regularly. The current sequence data is reclustered about once a month. The presence of new sequence information can change the clusters. As of two or three years ago, as many as 25% of the UniGene cluster assignments changed over the course of a year. One hopes that the rate of change is decreasing.

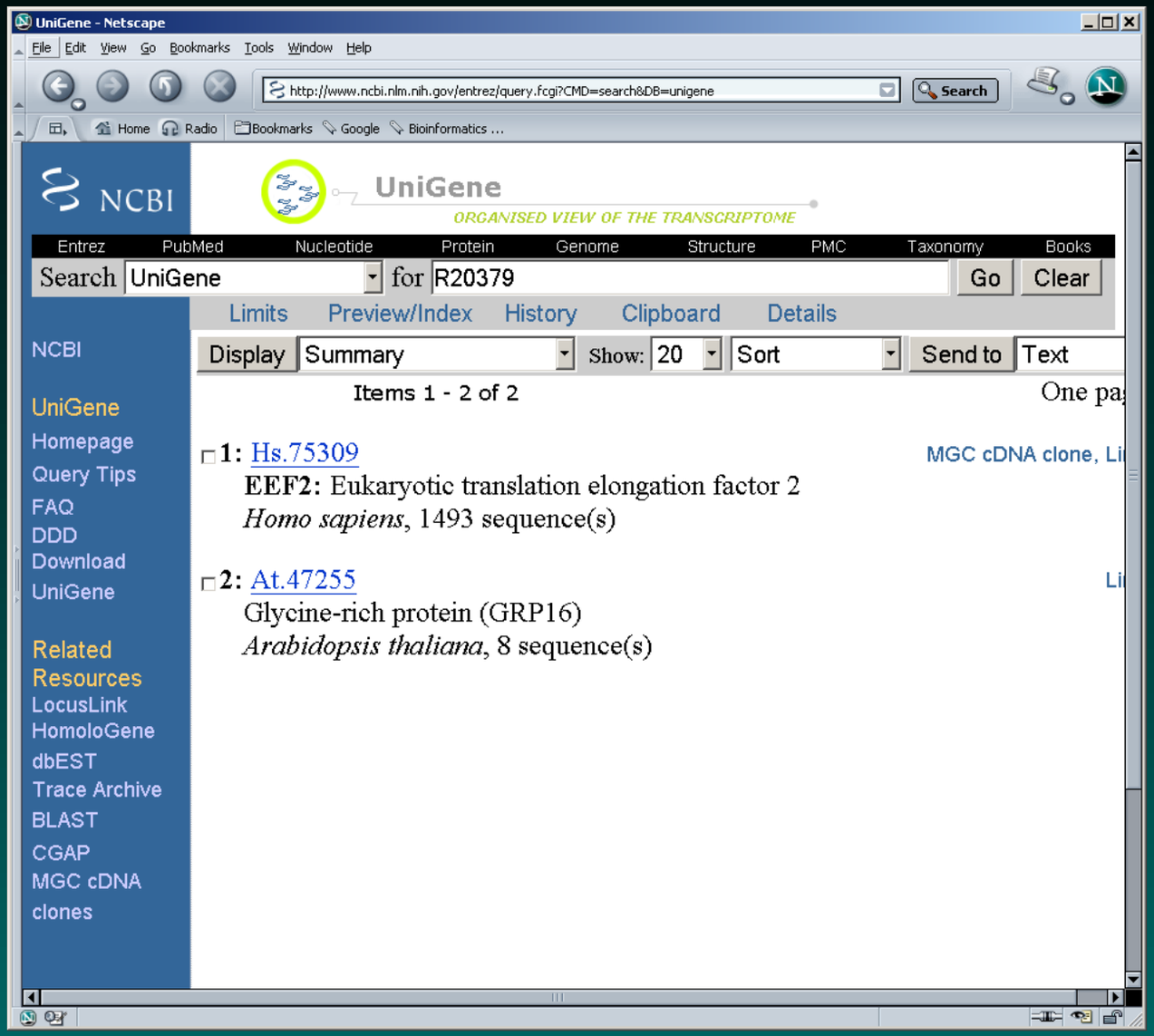
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Home Page



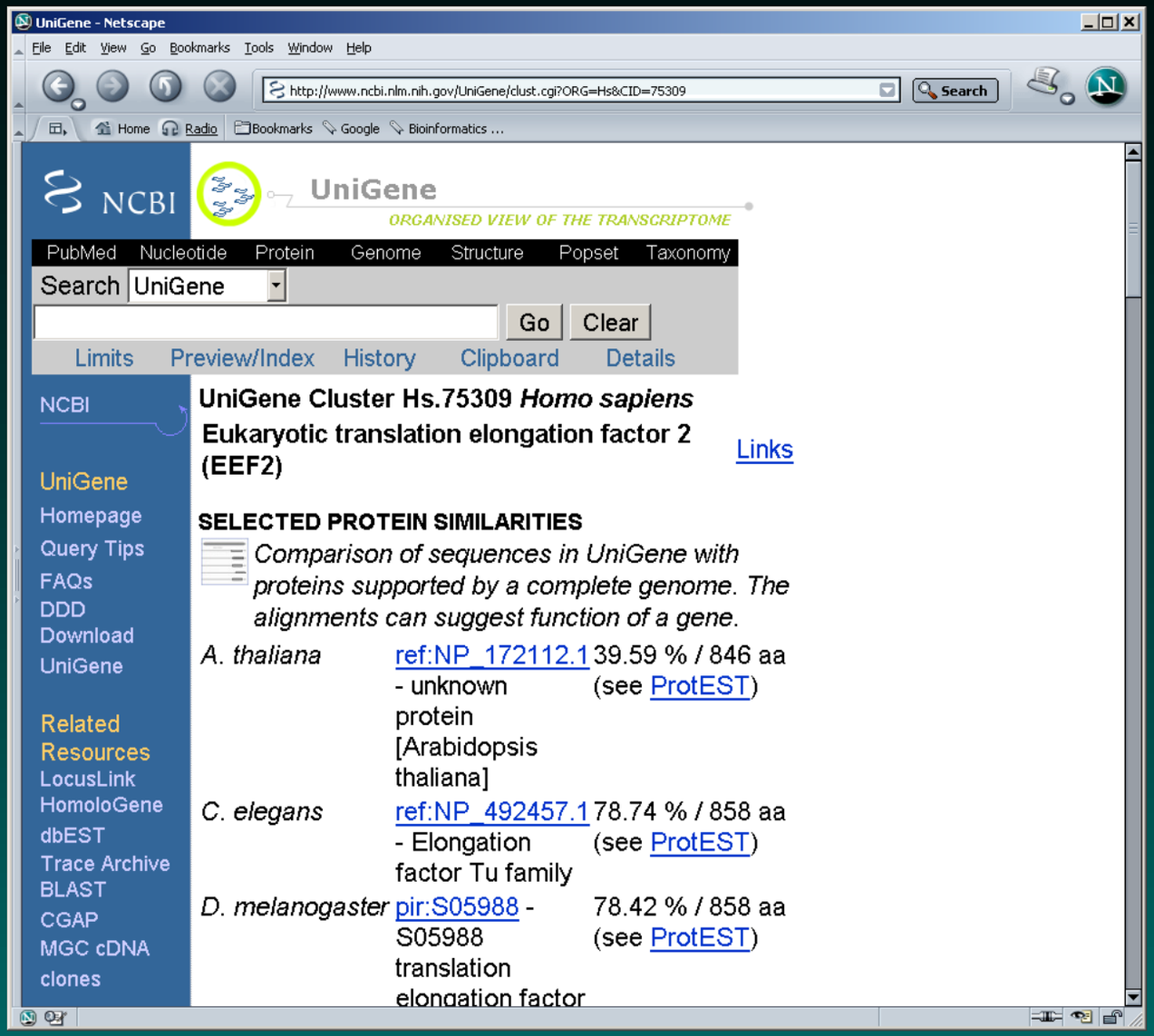
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Results



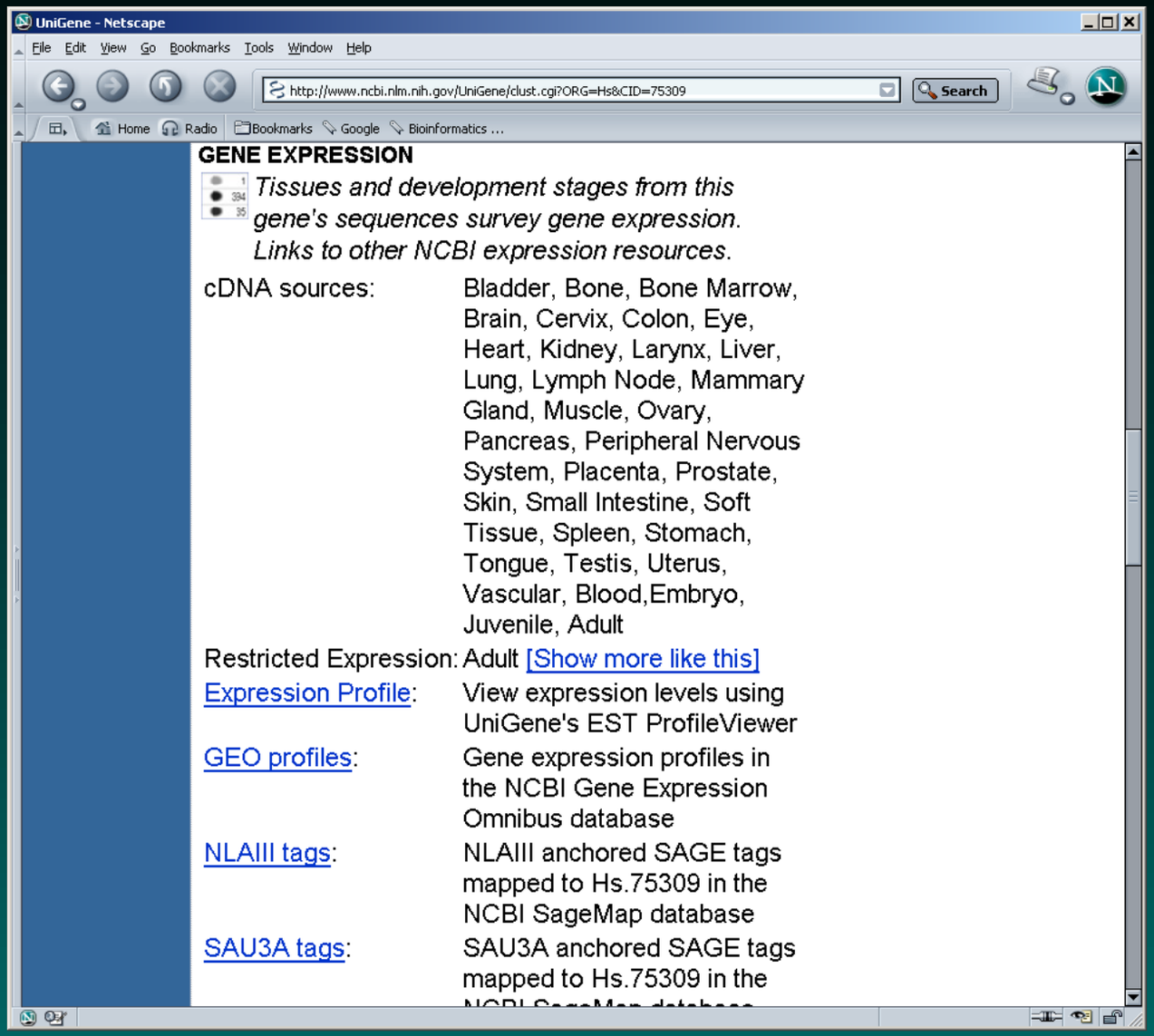
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Details



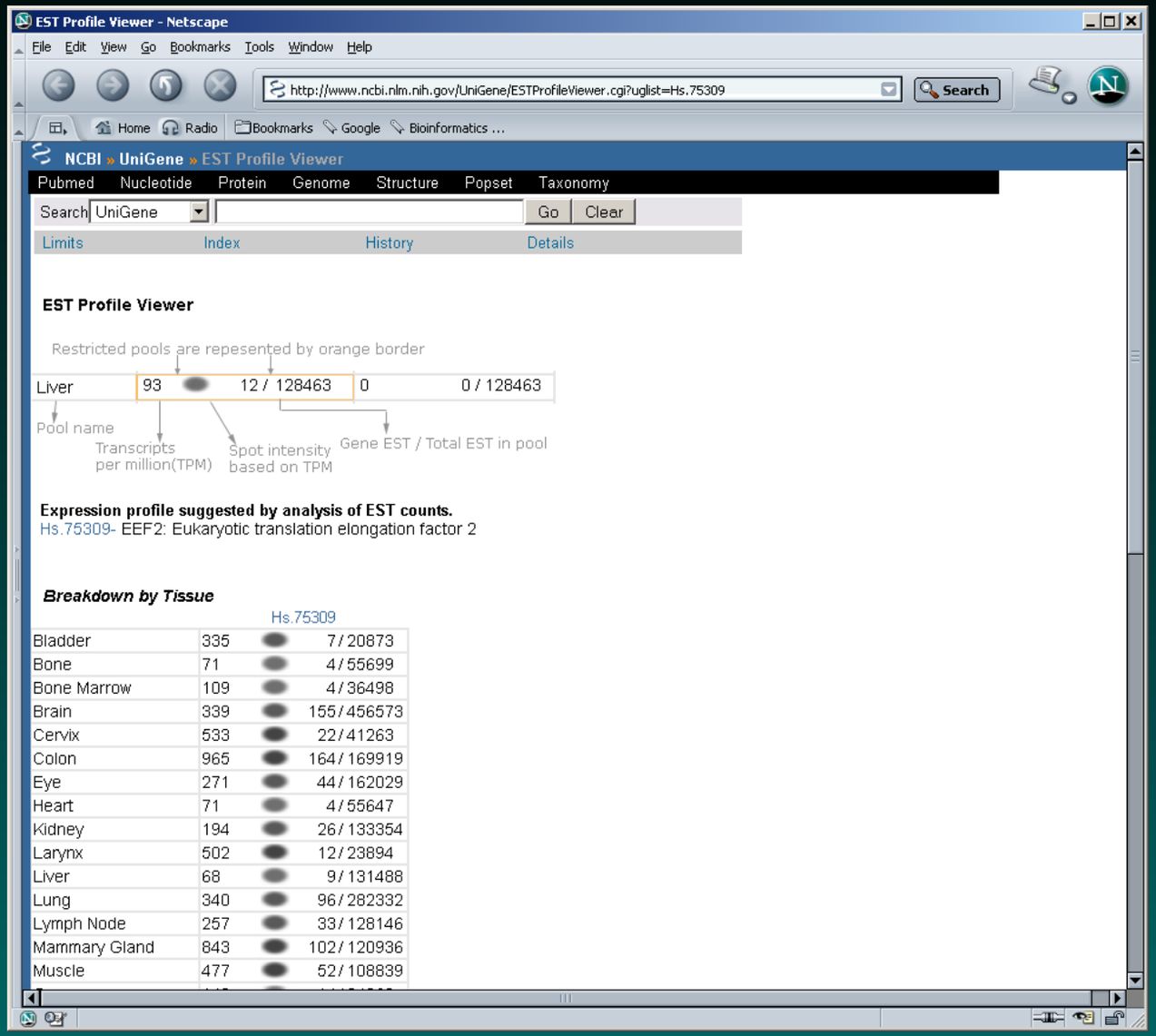
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Expression Information



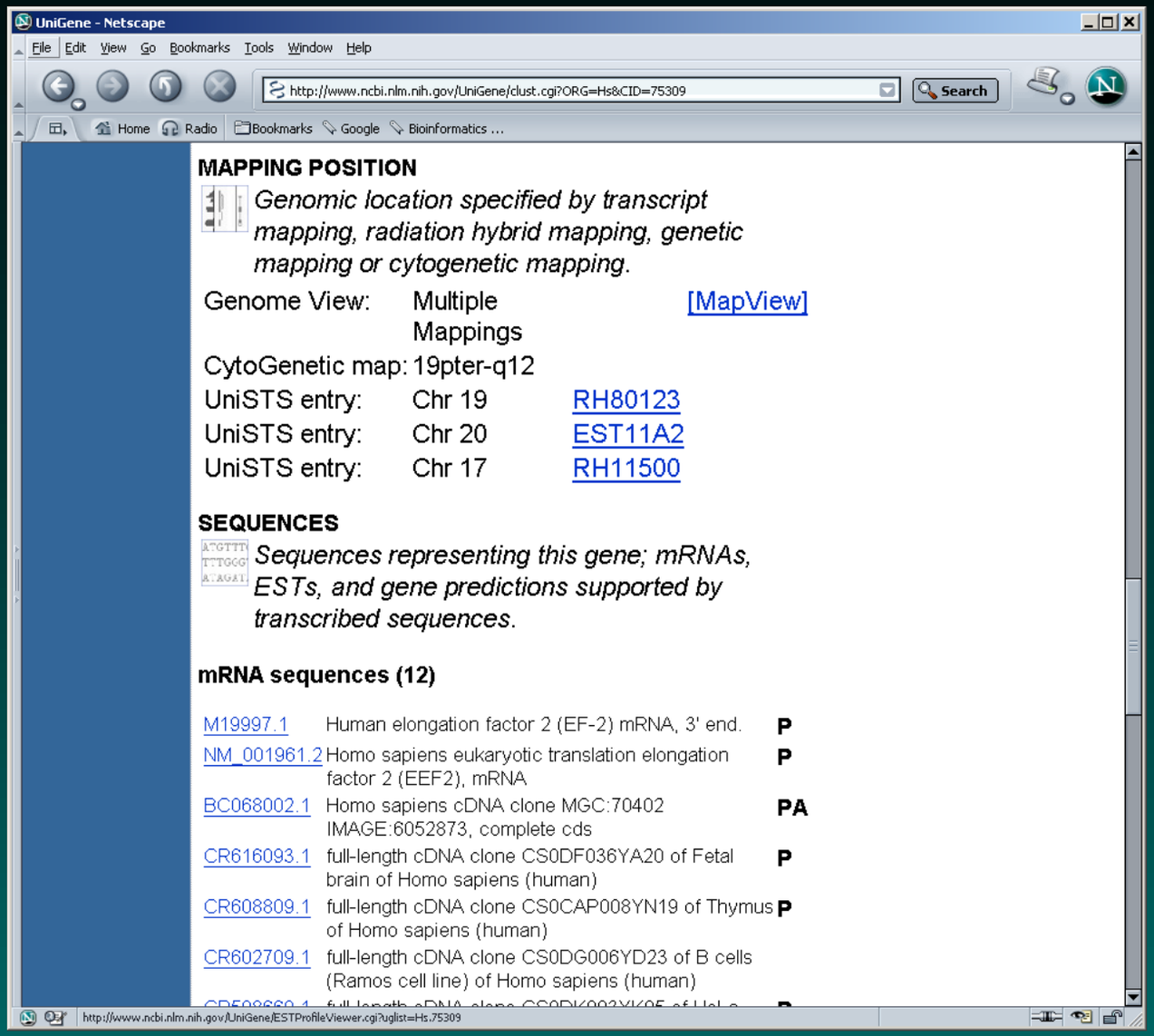
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Expression Profile



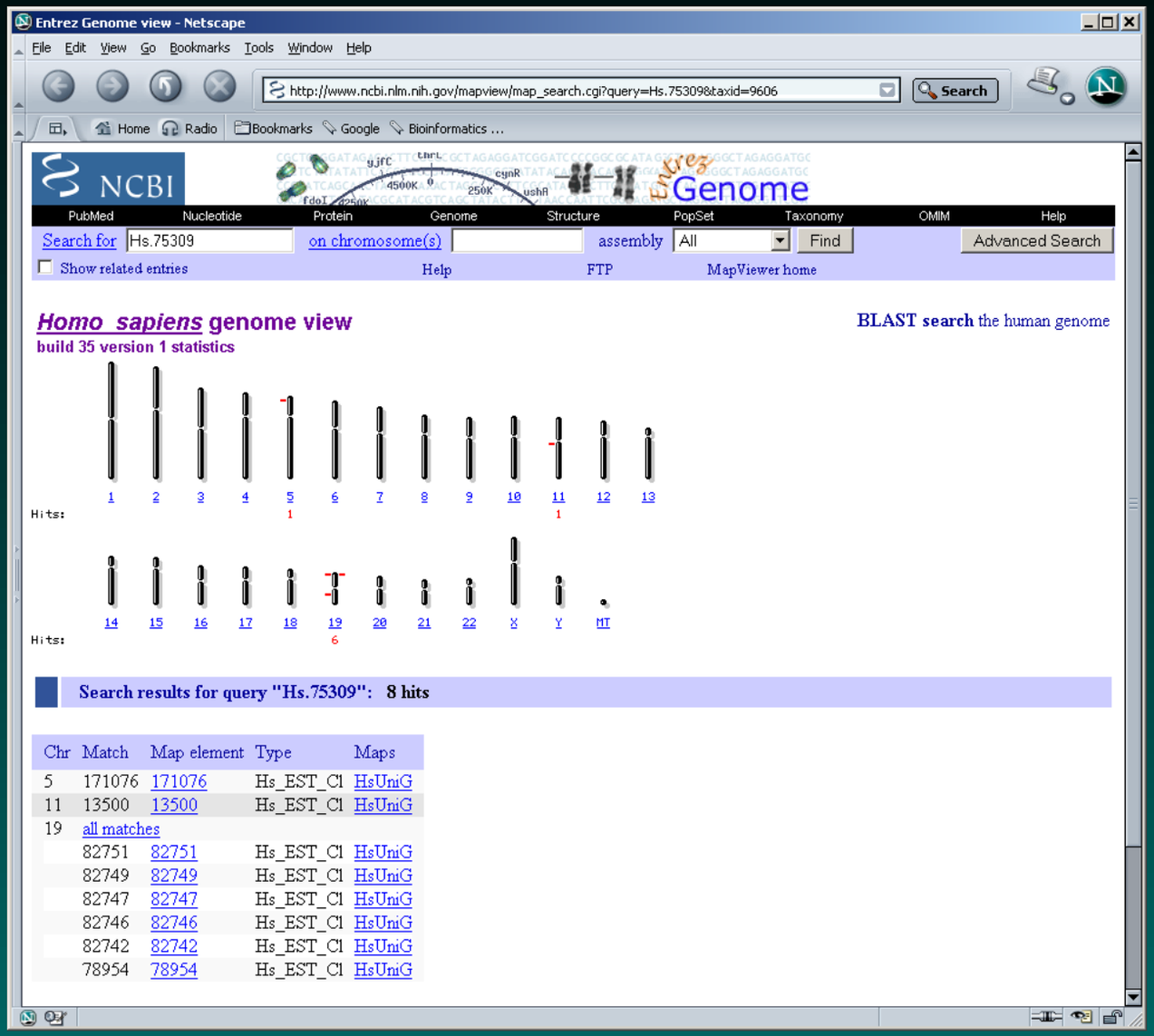
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Mapping Information



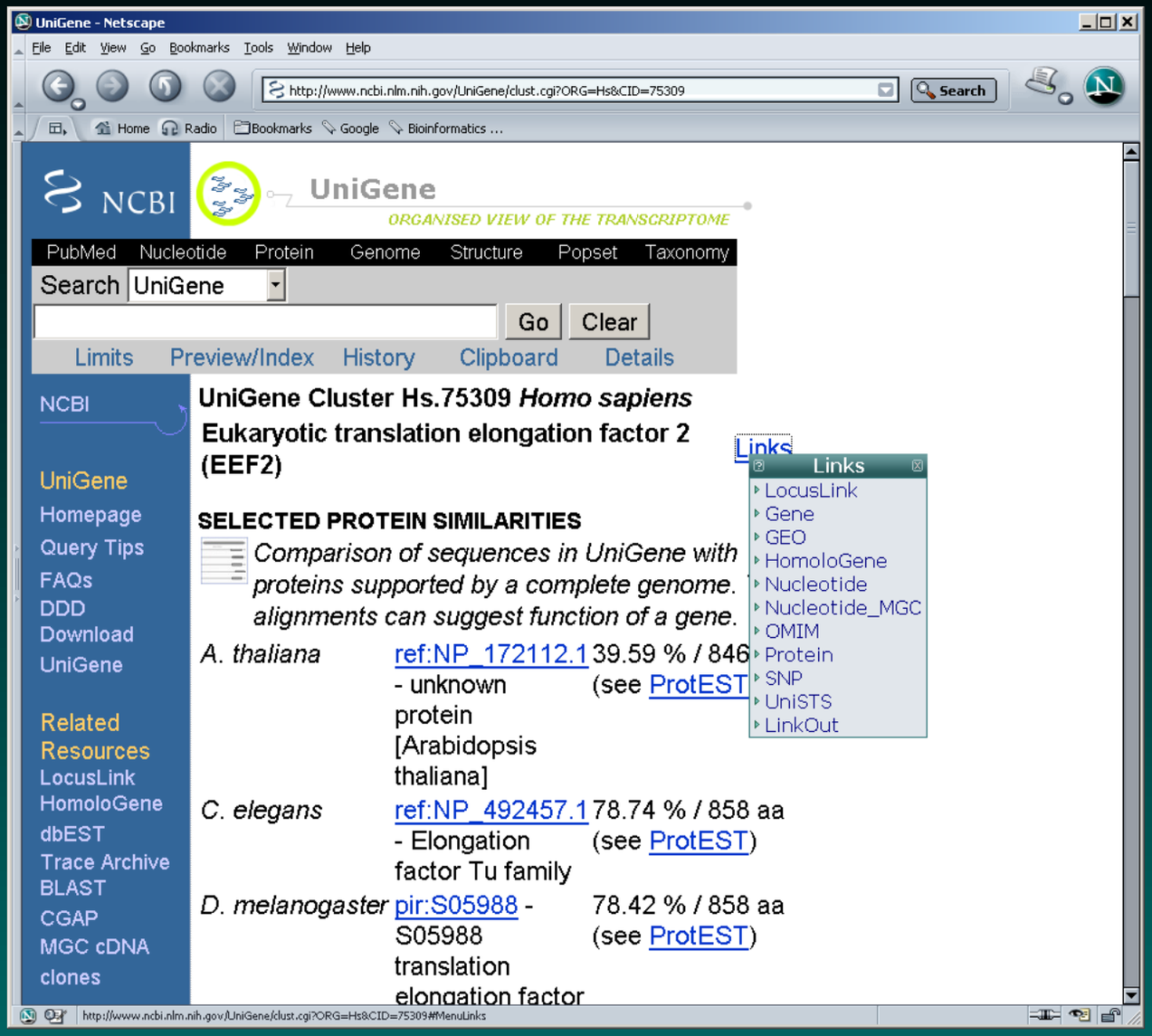
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Chromsome Mapping



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Links



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

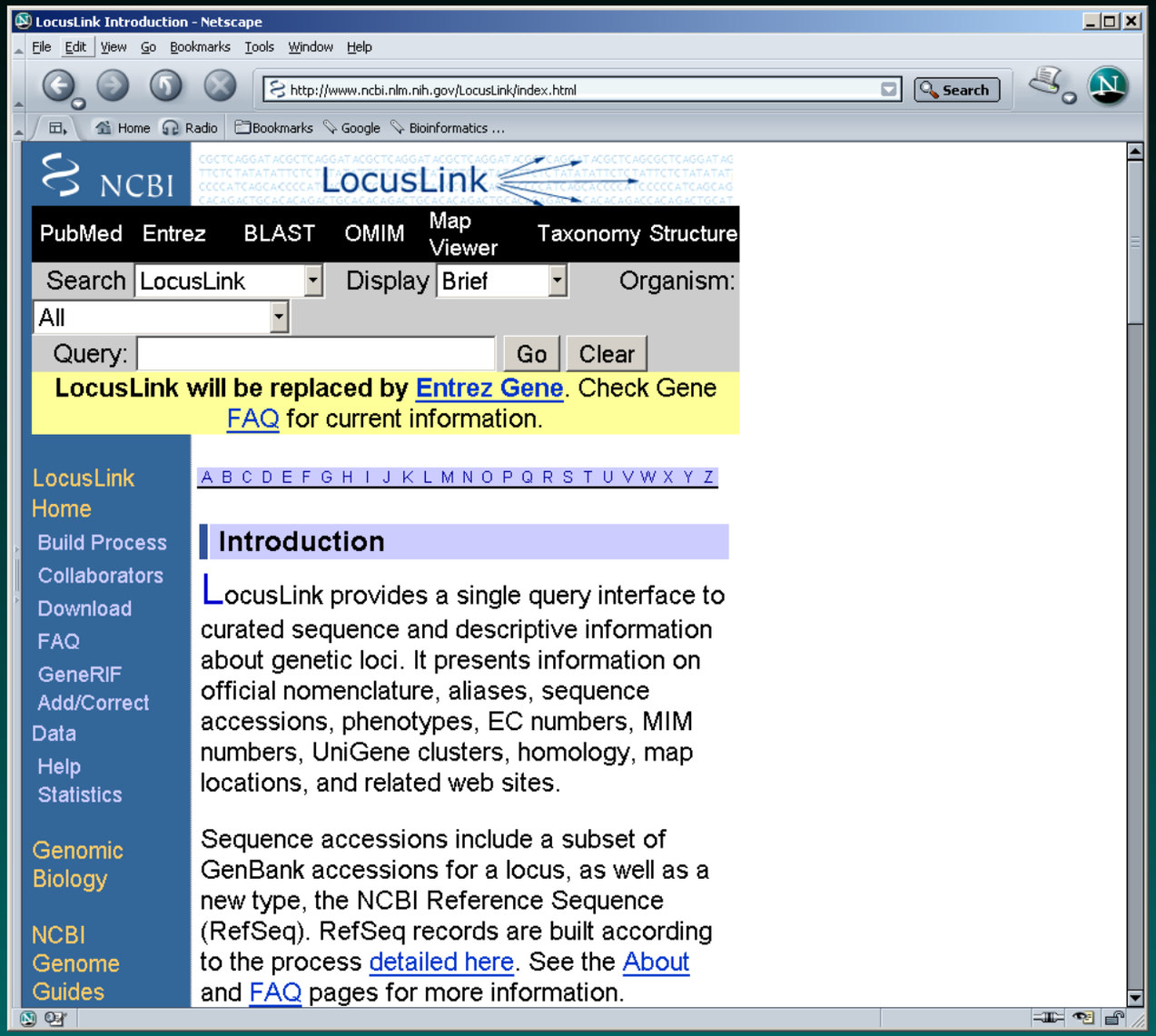
LocusLink

The first link out from UniGene is to LocusLink. LocuLik provides a single query interface to curated sequence information and descriptive information about genetic loci. This includes

- official nomenclature (symbol, name)
- aliases
- sequence accession numbers
- phenotypes
- MIM numbers
- UniGene clusters
- homology
- map locations

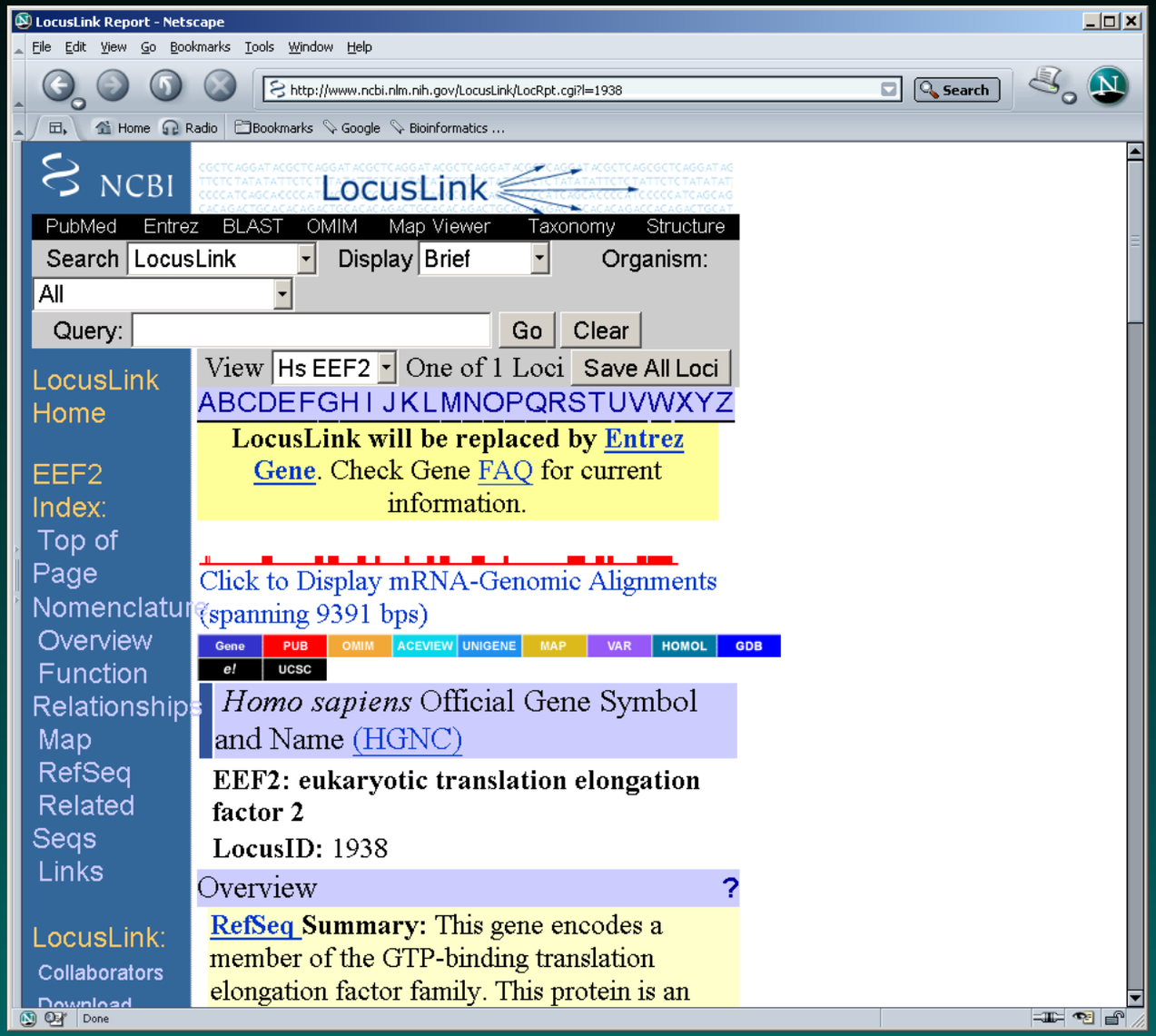
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

LocusLink Home Page



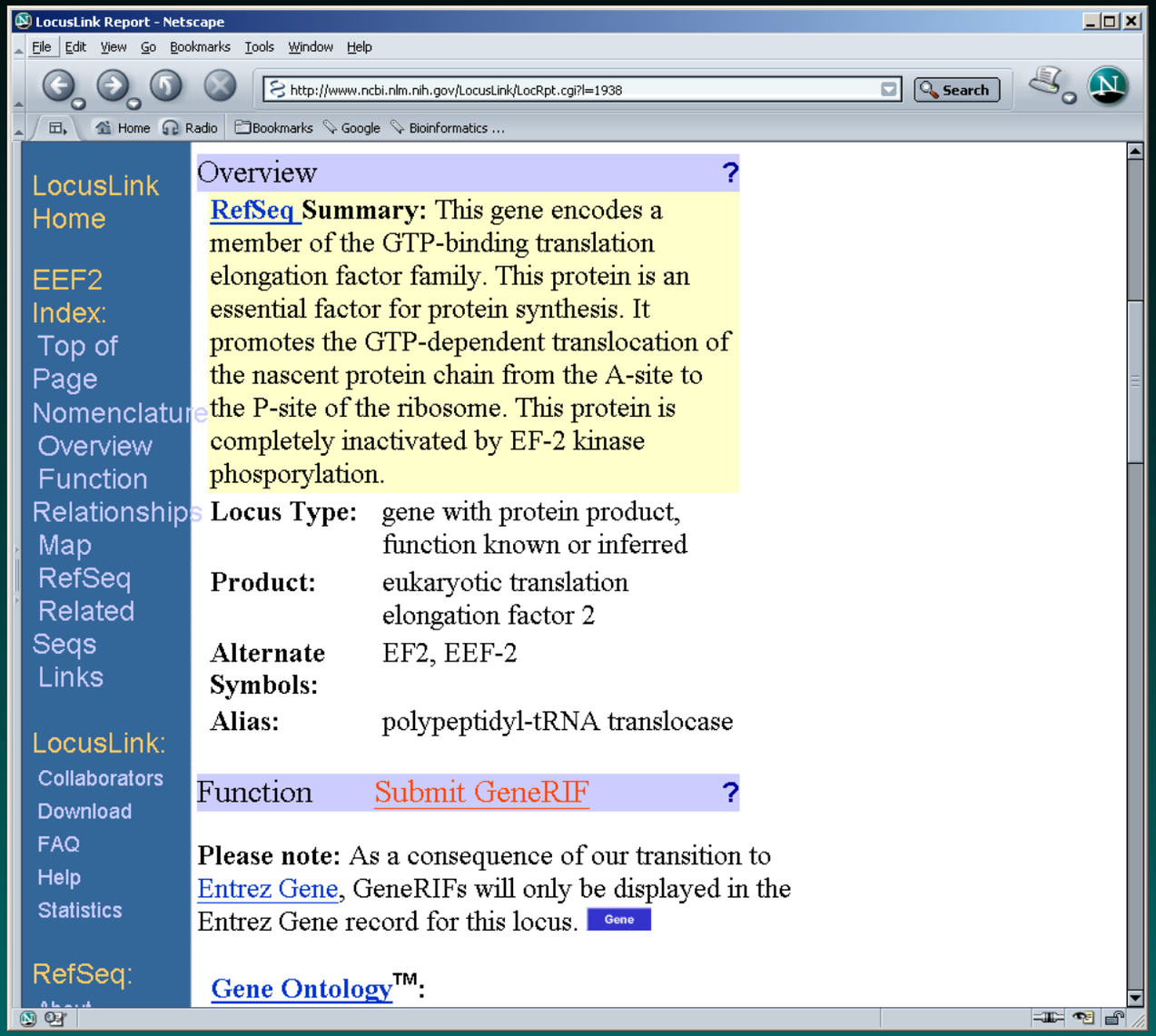
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

LocusLink Results



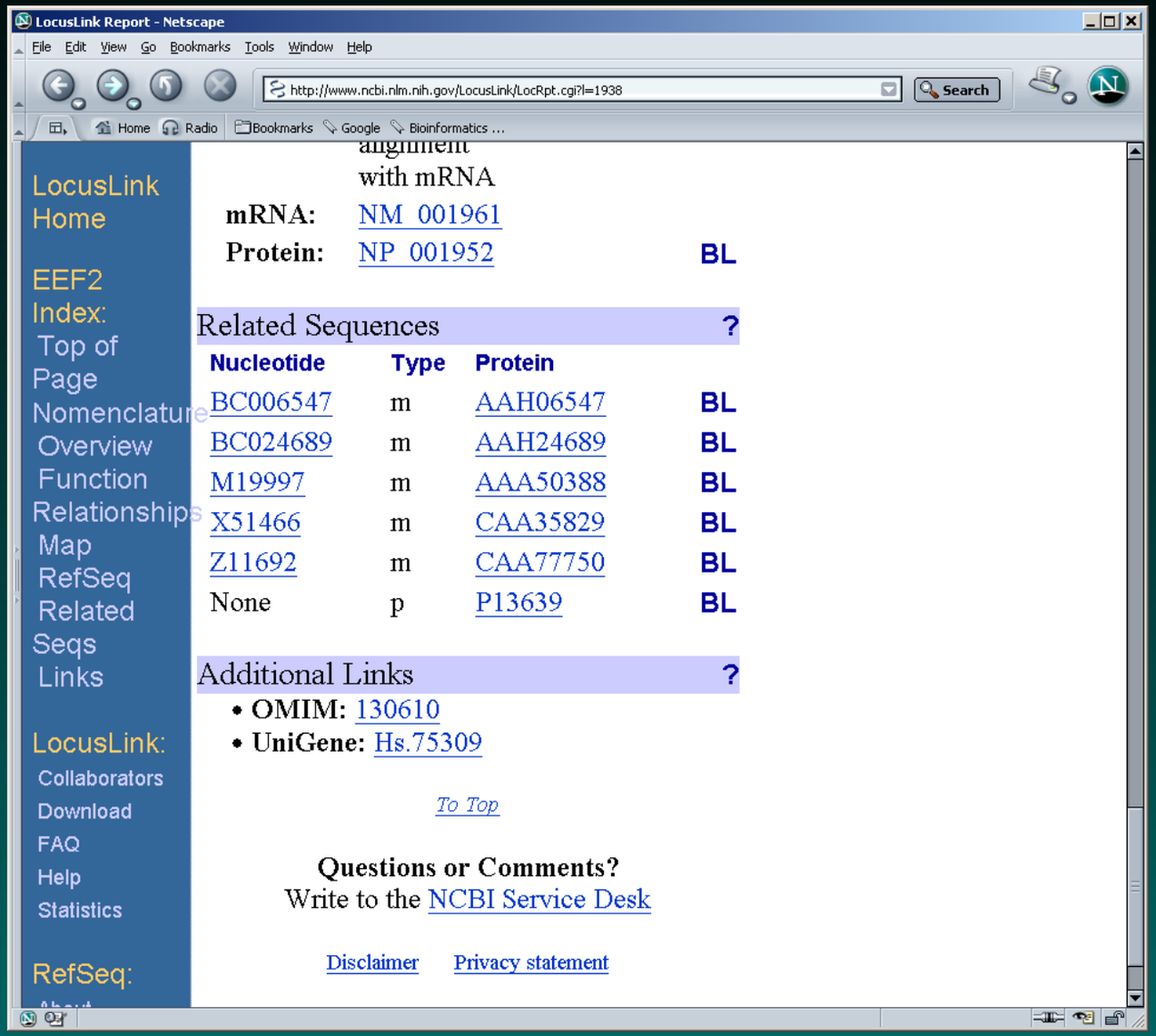
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

LocusLink Results



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

LocusLink Results



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

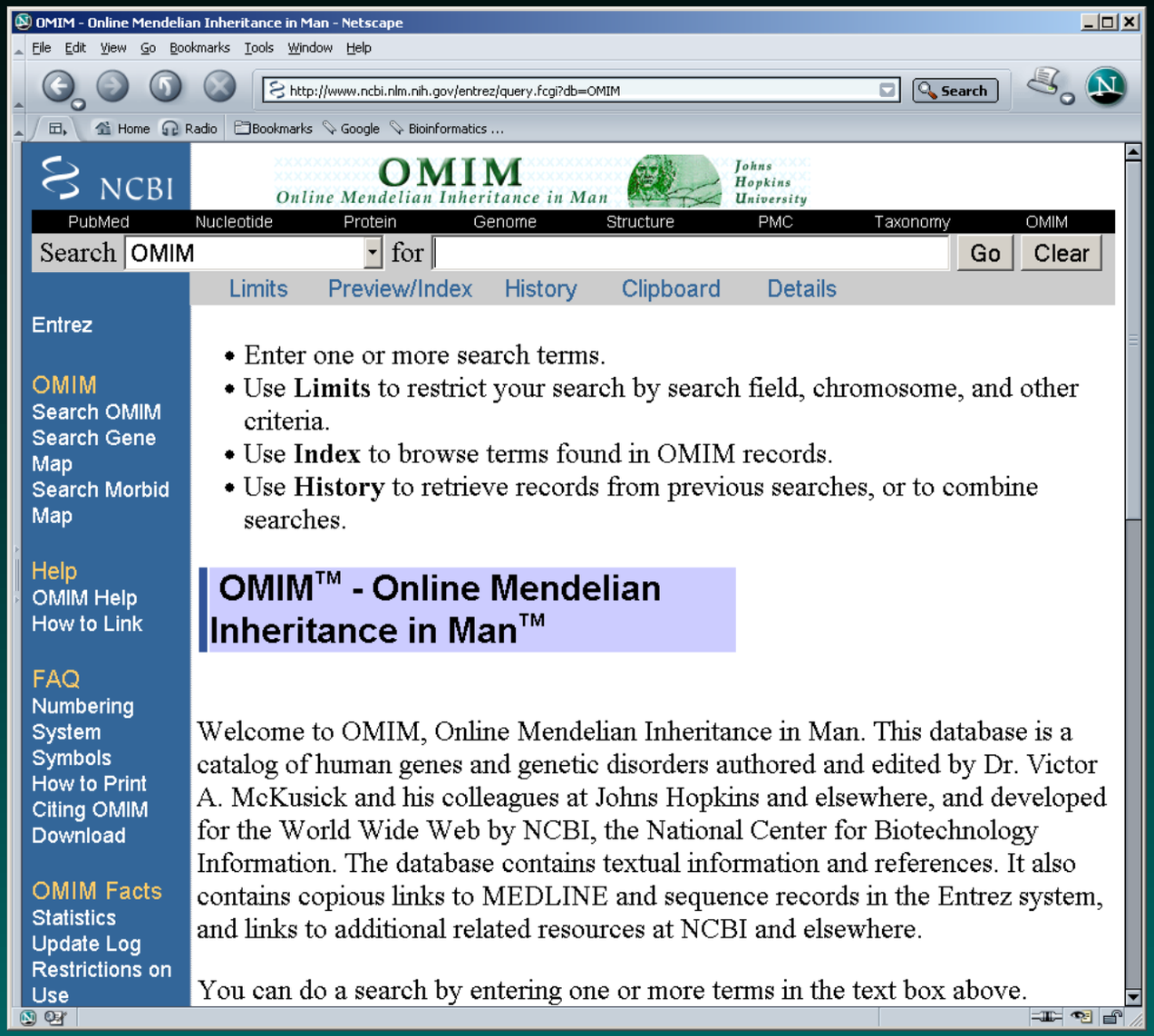
Online Mendelian Inheritance in Man

Both UniGene and LocusLink create automatic links to OMIM.

OMIM is a curated database of human genes and genetic disorders. It typically includes information about which diseases appear to be linked to specific genes, along with primary references that explain how the gene was sequenced and mapped to specific chromosomal regions.

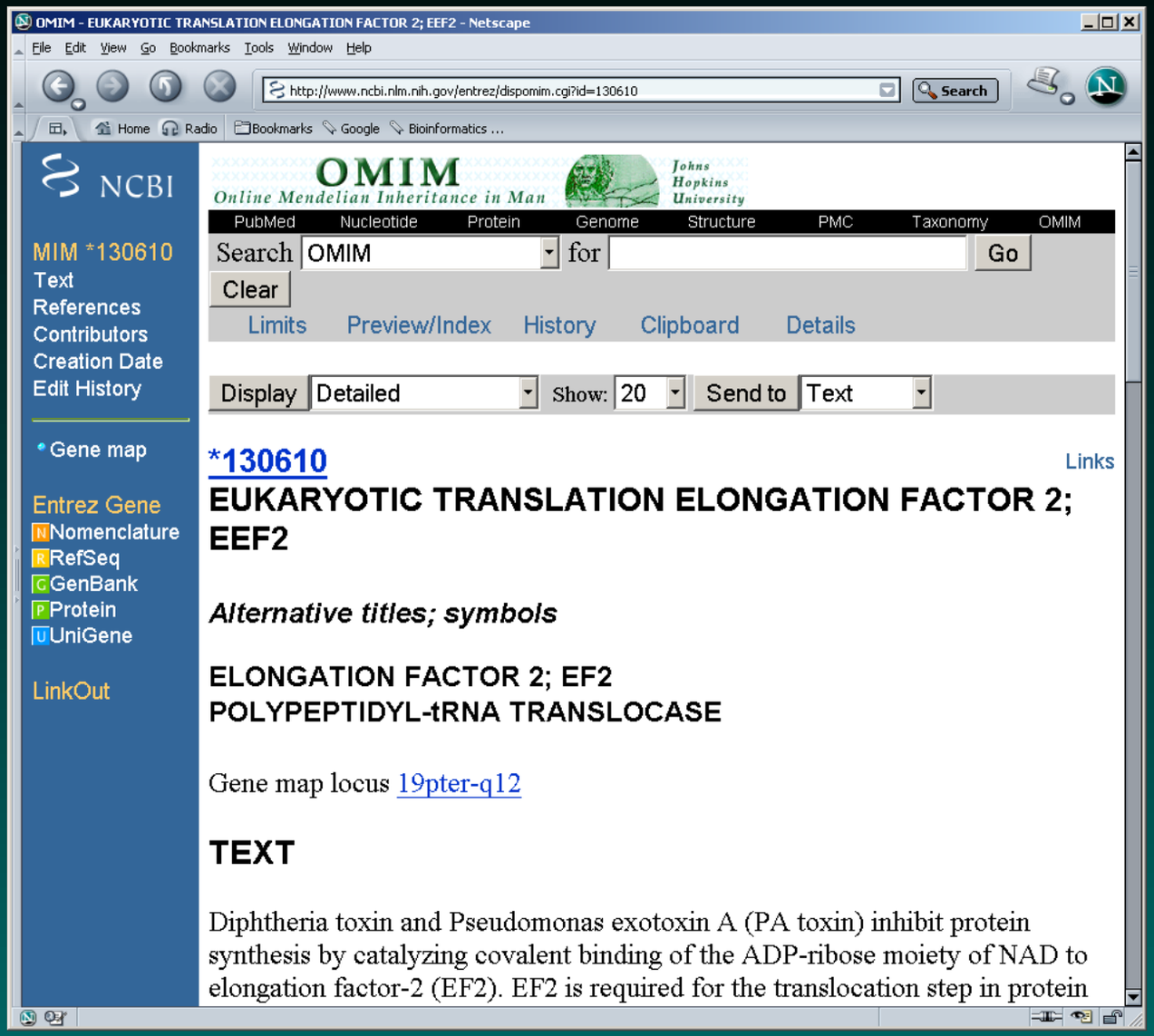
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

OMIM Main Page



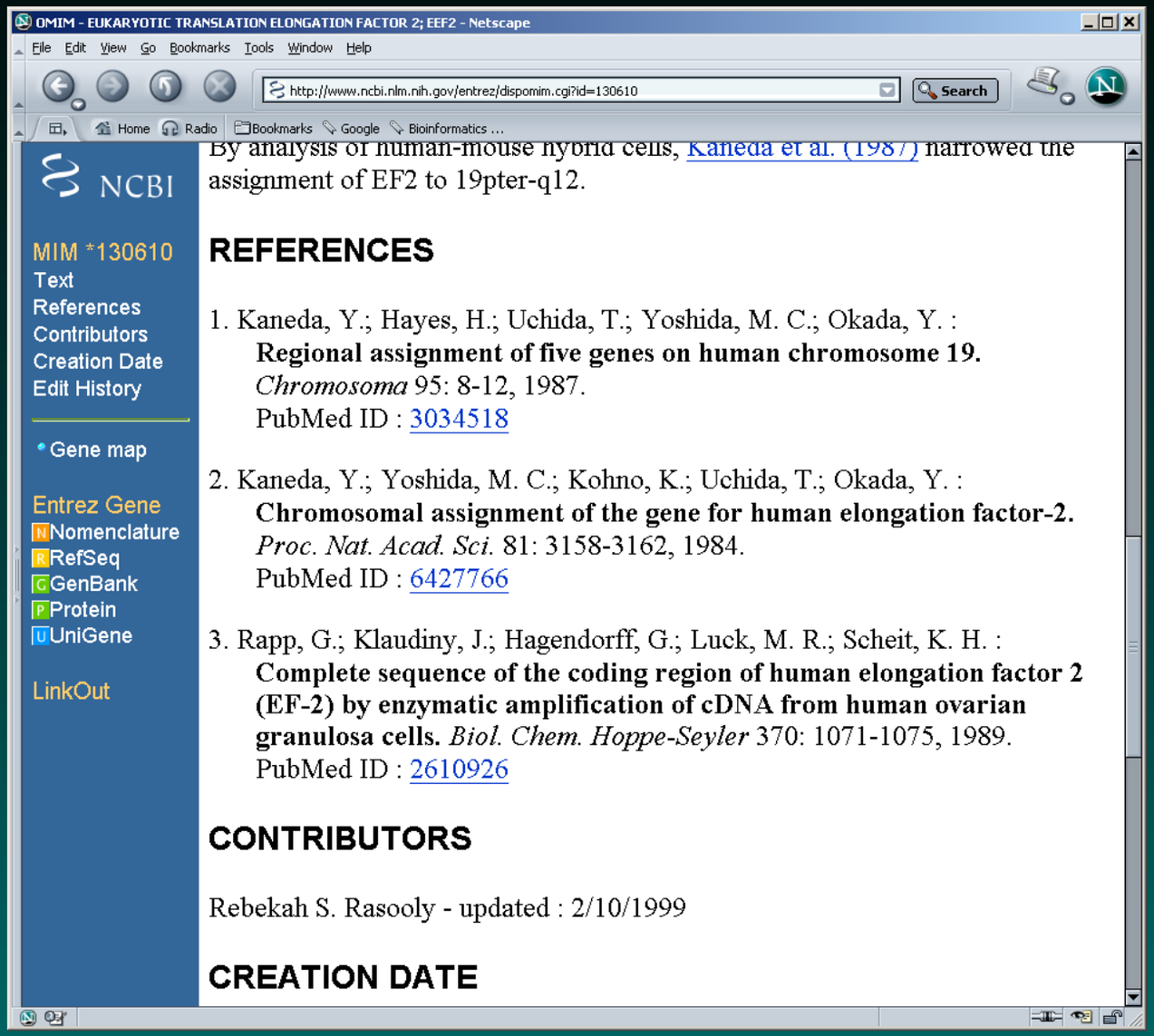
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

OMIM Results



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

OMIM References



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Batch Resources

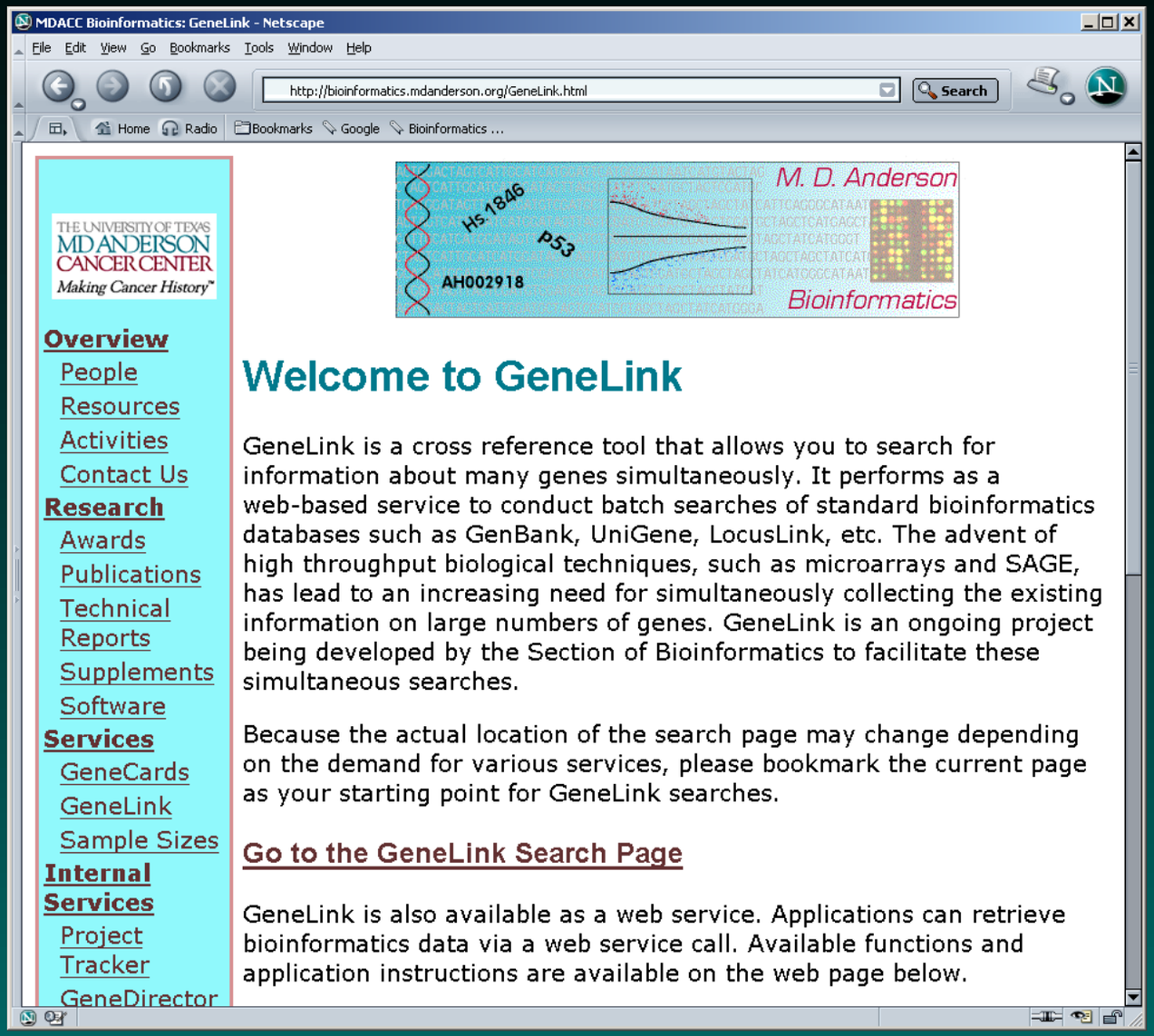
You may have noticed that only GenBank (Entrez-Nucleotide) provided a “batch” option that allowed us to search for an entire list of genes at once. Every other example we have presented works with one gene at a time. That’s probably not a good way to deal with 1000 genes.

M.D. Anderson maintains a service called GeneLink to deal with this problem. You can try it at `http://bioinformatics.mdanderson.org/GeneLink.html`

Stanford has a similar tool (that right now is faster) called Source. Check out `http://source.stanford.edu`.

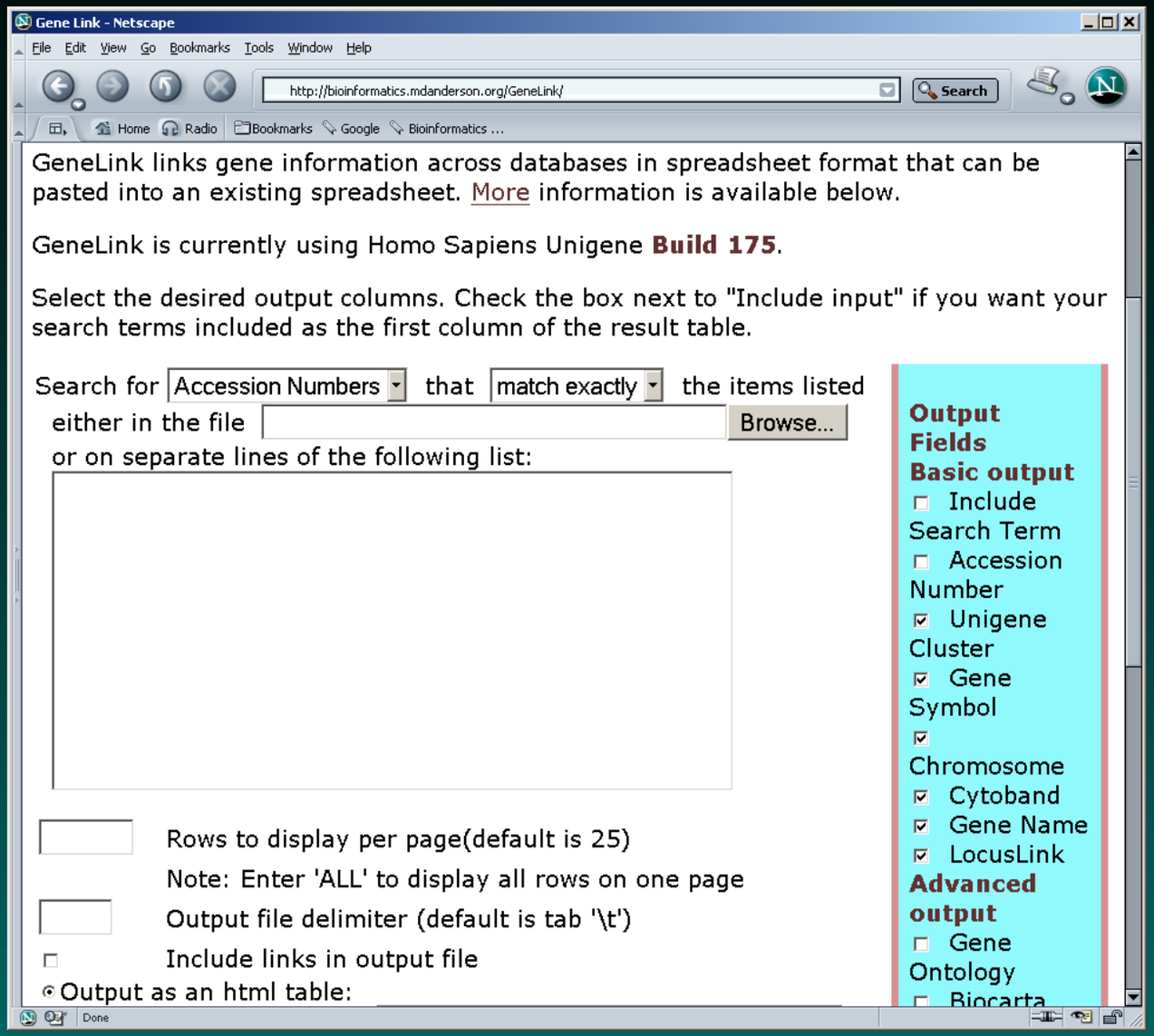
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GeneLink Description



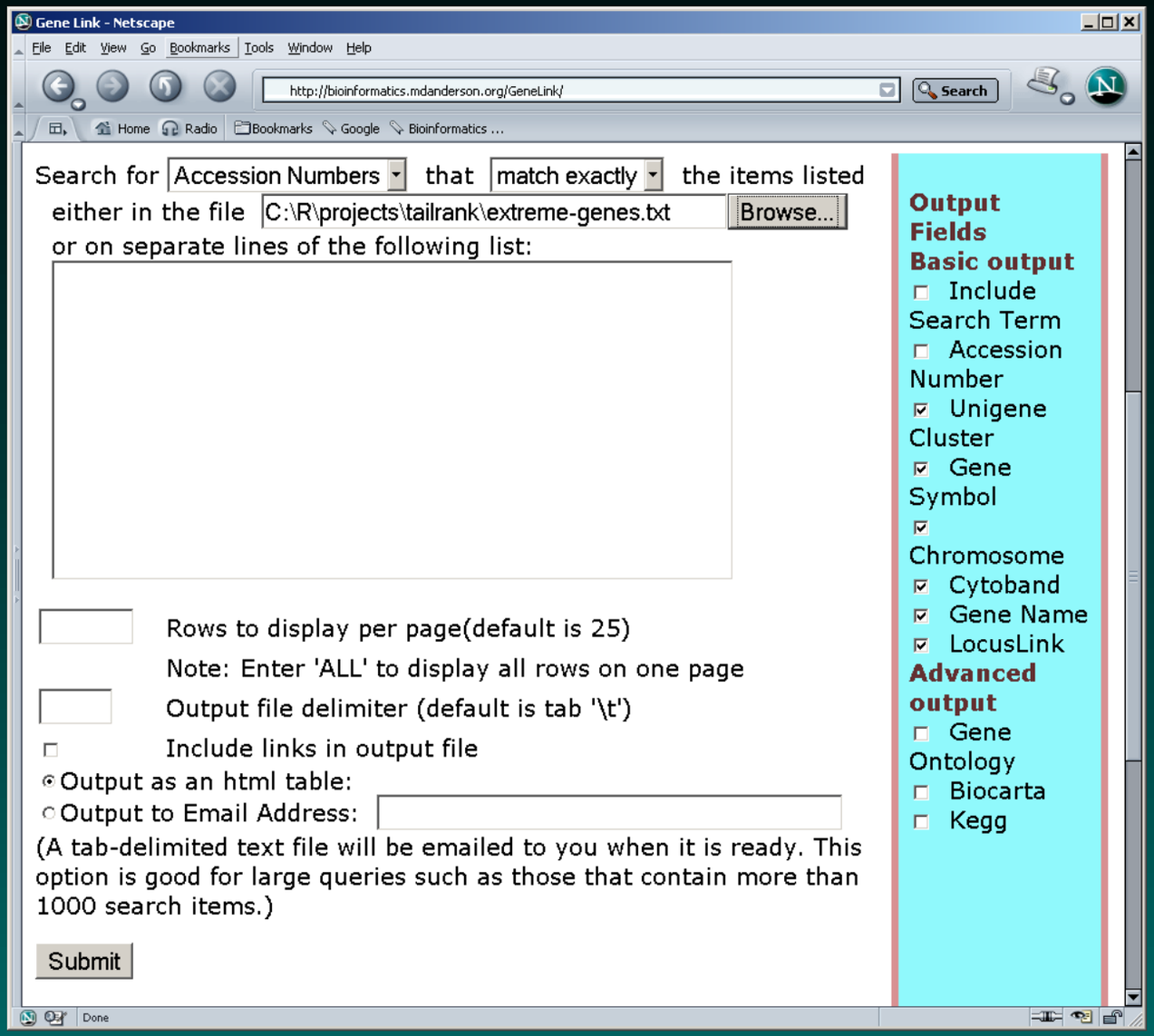
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GeneLink Search Page



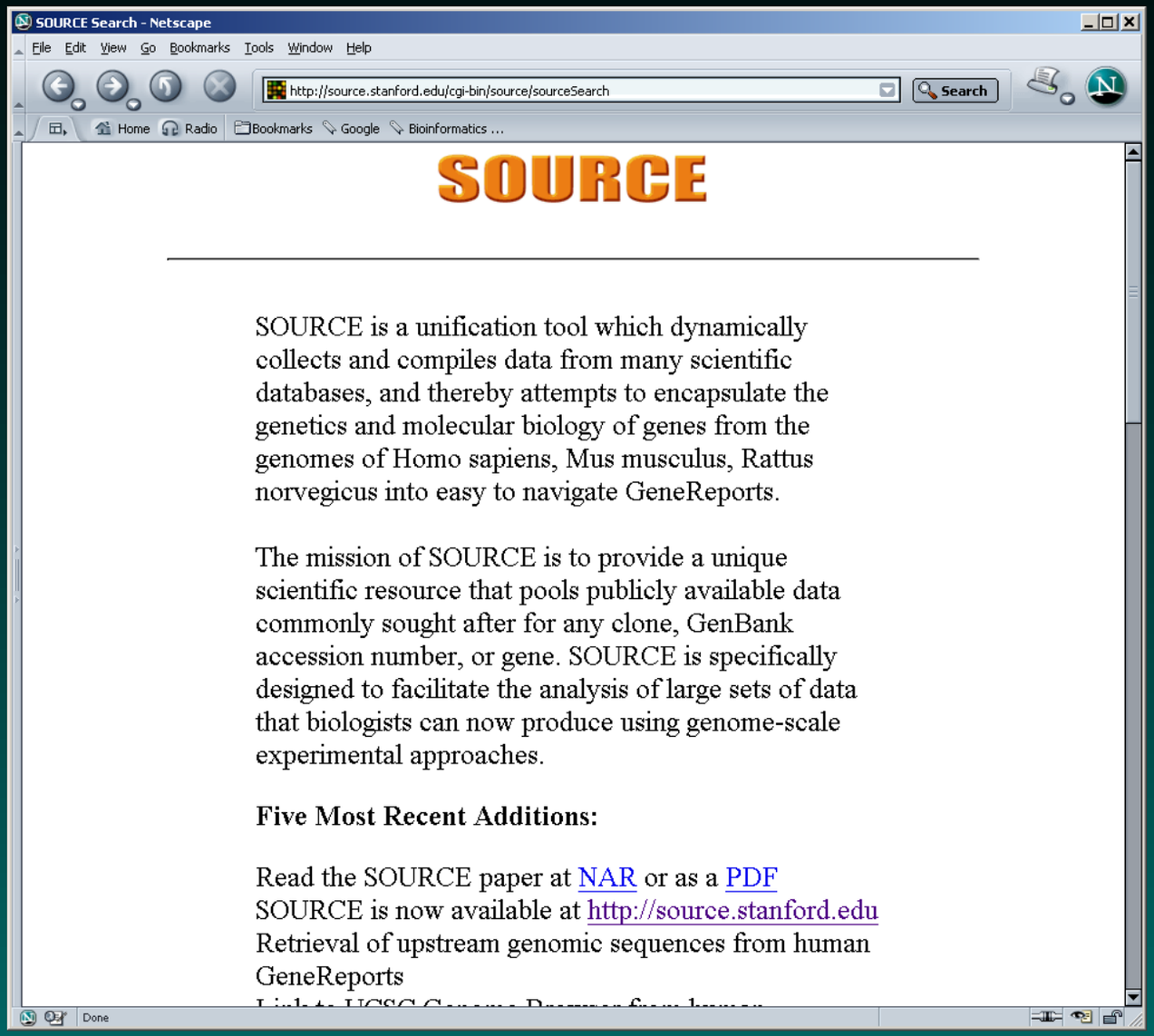
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GeneLink Search Page



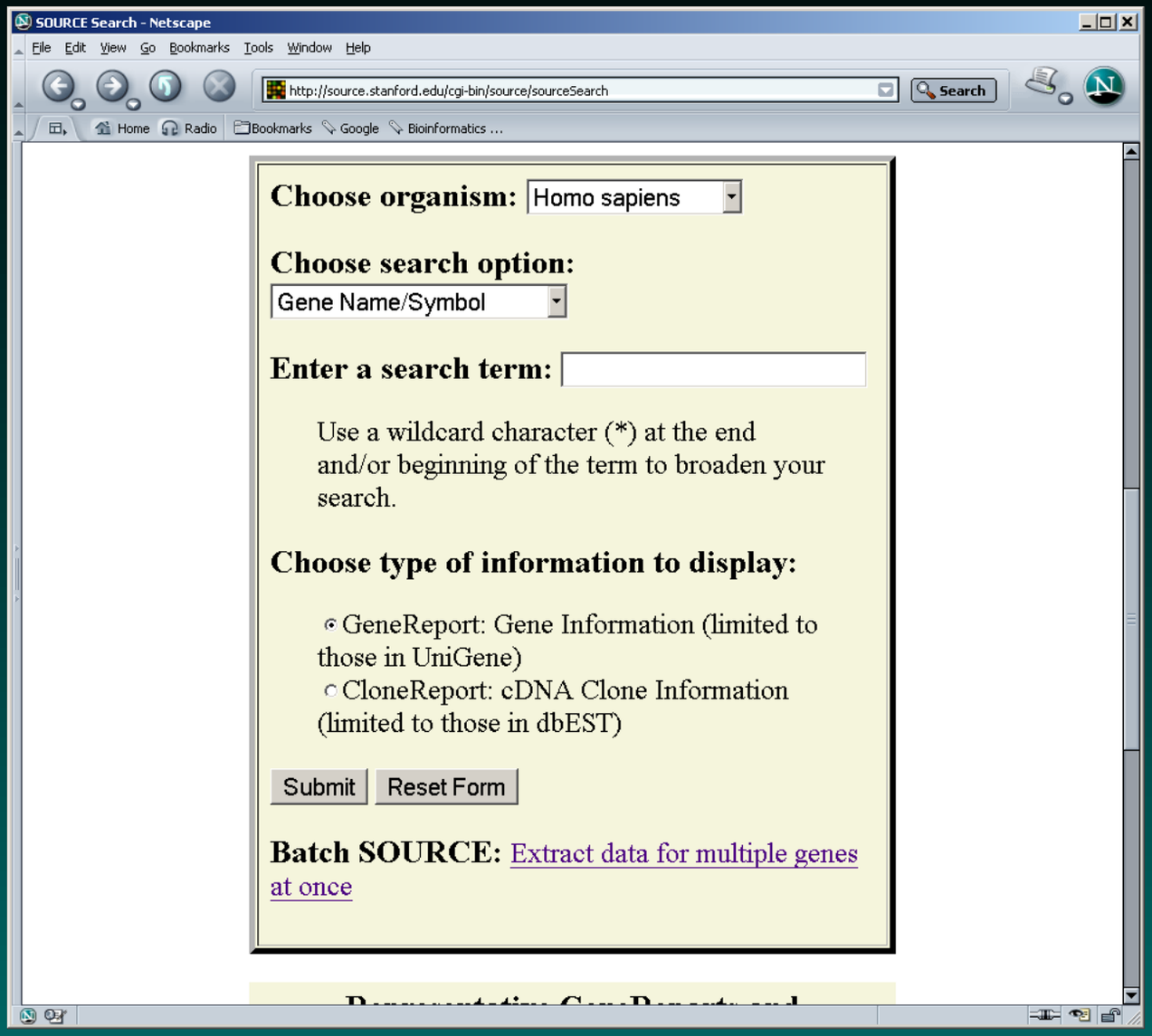
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Source Description



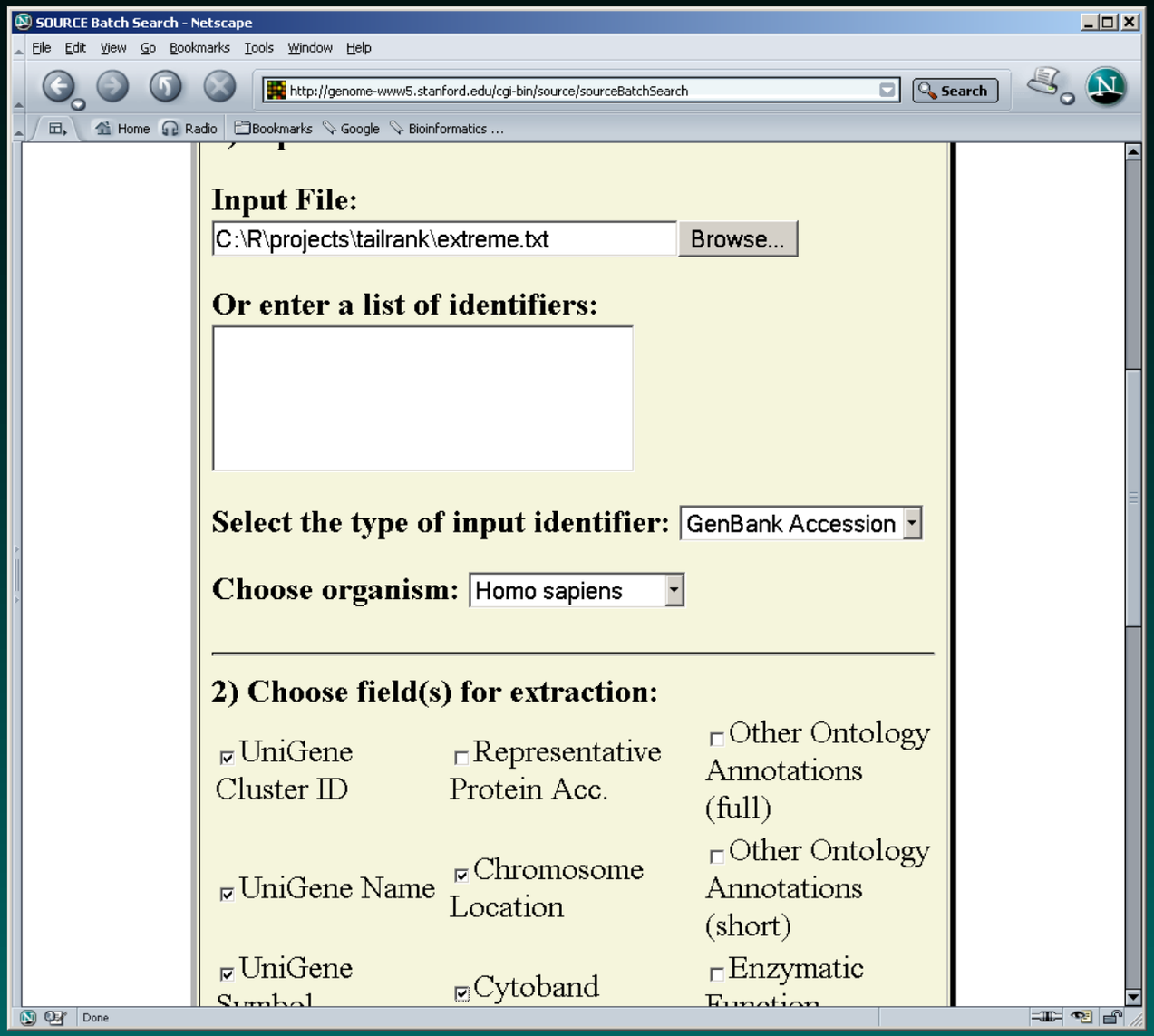
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Source Search Page



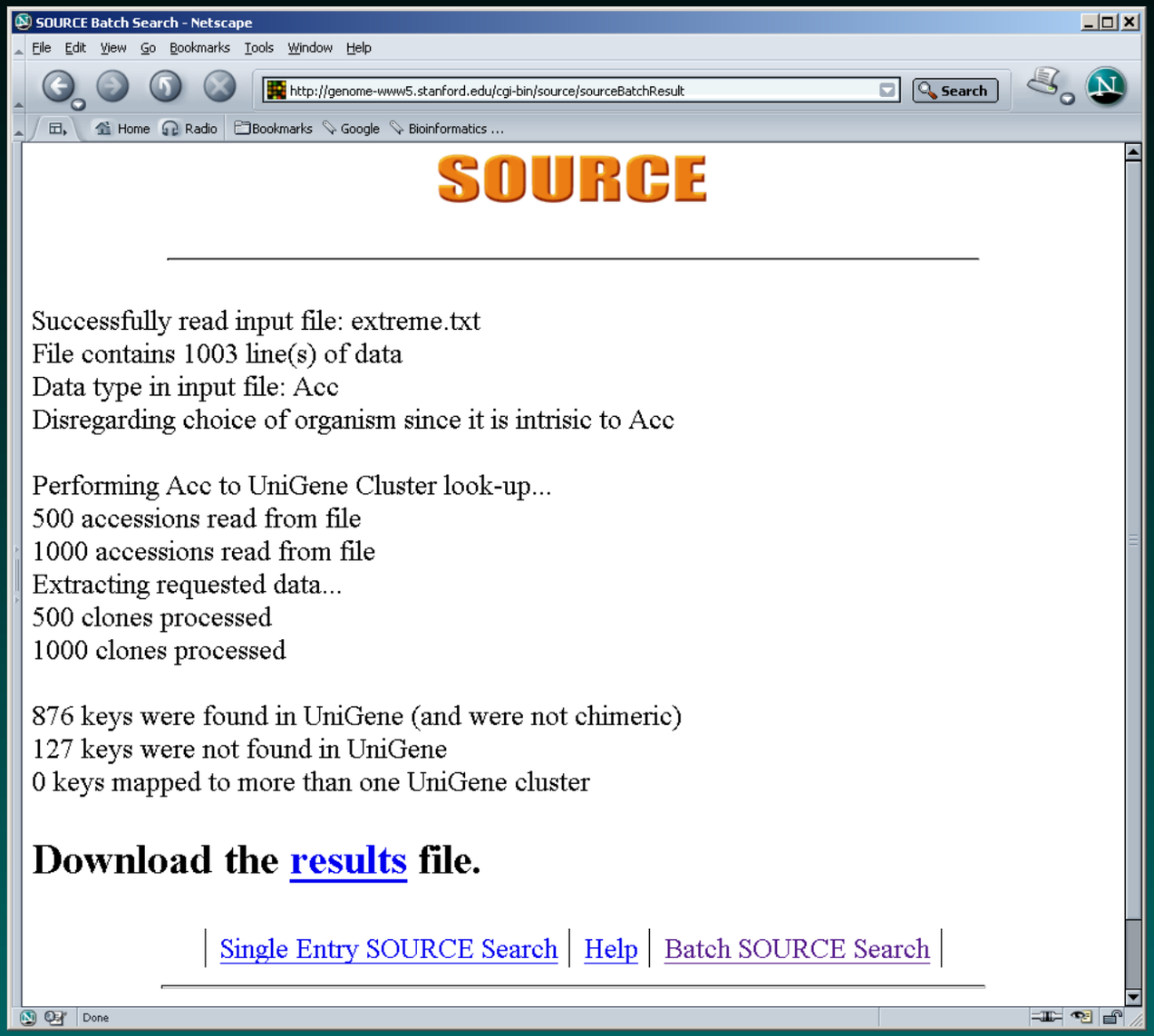
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Source Batch Search Page



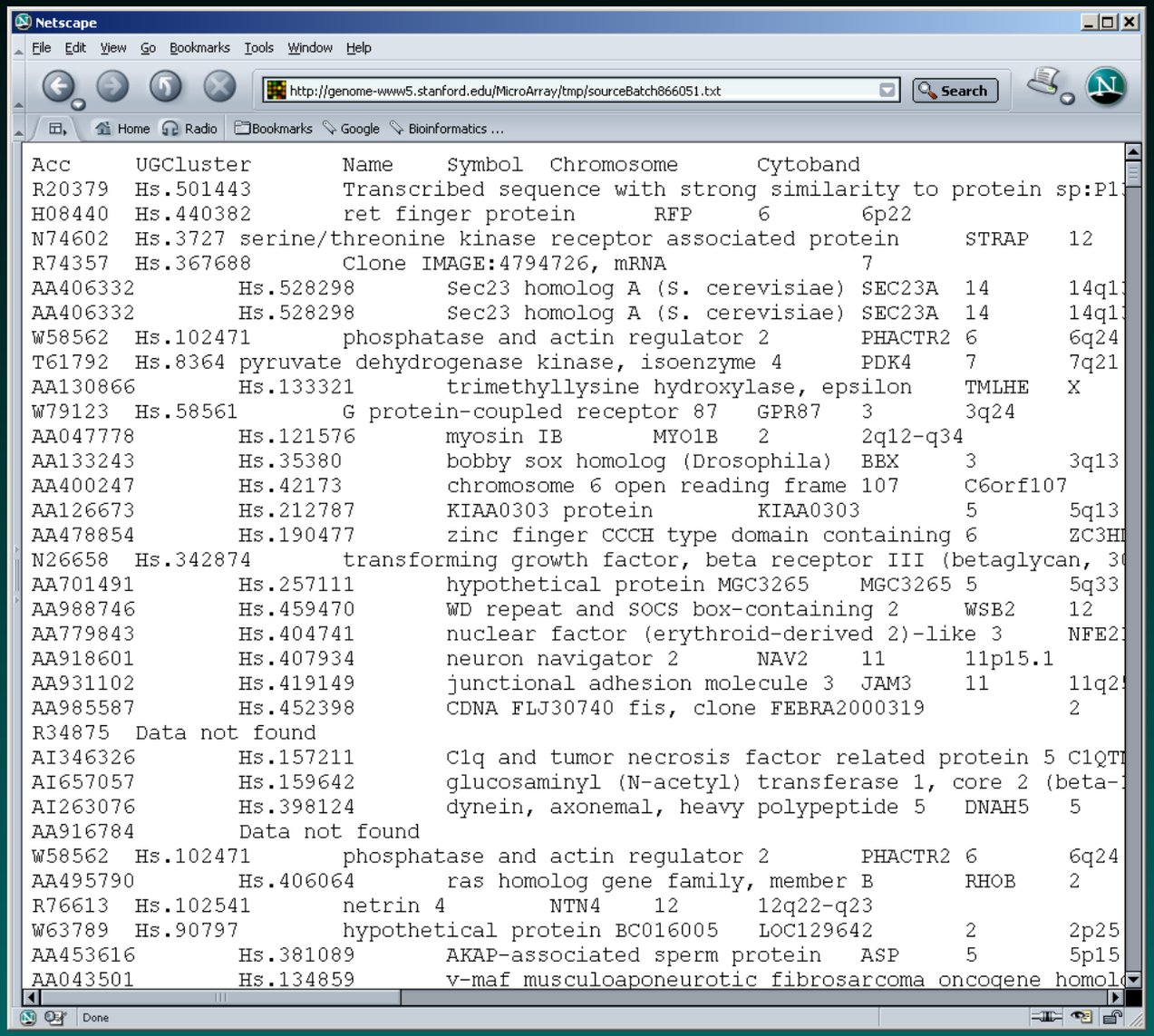
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Source Results



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Source Details



Acc	UGCluster	Name	Symbol	Chromosome	Cytoband
R20379	Hs.501443	Transcribed sequence with strong similarity to protein sp:P1			
H08440	Hs.440382	ret finger protein	RFP	6	6p22
N74602	Hs.3727	serine/threonine kinase receptor associated protein	STRAP	12	
R74357	Hs.367688	Clone IMAGE:4794726, mRNA		7	
AA406332	Hs.528298	Sec23 homolog A (S. cerevisiae)	SEC23A	14	14q1
AA406332	Hs.528298	Sec23 homolog A (S. cerevisiae)	SEC23A	14	14q1
W58562	Hs.102471	phosphatase and actin regulator 2	PHACTR2	6	6q24
T61792	Hs.8364	pyruvate dehydrogenase kinase, isoenzyme 4	PDK4	7	7q21
AA130866	Hs.133321	trimethyllysine hydroxylase, epsilon	TMLHE	X	
W79123	Hs.58561	G protein-coupled receptor 87	GPR87	3	3q24
AA047778	Hs.121576	myosin IB	MYO1B	2	2q12-q34
AA133243	Hs.35380	bobby sox homolog (Drosophila)	BBX	3	3q13
AA400247	Hs.42173	chromosome 6 open reading frame 107	C6orf107		
AA126673	Hs.212787	KIAA0303 protein	KIAA0303	5	5q13
AA478854	Hs.190477	zinc finger CCCH type domain containing 6			ZC3H
N26658	Hs.342874	transforming growth factor, beta receptor III (betaglycan, 3			
AA701491	Hs.257111	hypothetical protein MGC3265	MGC3265	5	5q33
AA988746	Hs.459470	WD repeat and SOCS box-containing 2	WSB2	12	
AA779843	Hs.404741	nuclear factor (erythroid-derived 2)-like 3			NFE2
AA918601	Hs.407934	neuron navigator 2	NAV2	11	11p15.1
AA931102	Hs.419149	junctional adhesion molecule 3	JAM3	11	11q2
AA985587	Hs.452398	CDNA FLJ30740 fis, clone FEBRA2000319			2
R34875	Data not found				
AI346326	Hs.157211	C1q and tumor necrosis factor related protein 5			C1QT
AI657057	Hs.159642	glucosaminyl (N-acetyl) transferase 1, core 2 (beta-			
AI263076	Hs.398124	dynein, axonemal, heavy polypeptide 5	DNAH5	5	
AA916784	Data not found				
W58562	Hs.102471	phosphatase and actin regulator 2	PHACTR2	6	6q24
AA495790	Hs.406064	ras homolog gene family, member B	RHOB	2	
R76613	Hs.102541	netrin 4	NTN4	12	12q22-q23
W63789	Hs.90797	hypothetical protein BC016005	LOC129642	2	2p25
AA453616	Hs.381089	AKAP-associated sperm protein	ASP	5	5p15
AA043501	Hs.134859	v-maf musculoaponeurotic fibrosarcoma oncogene homol			

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

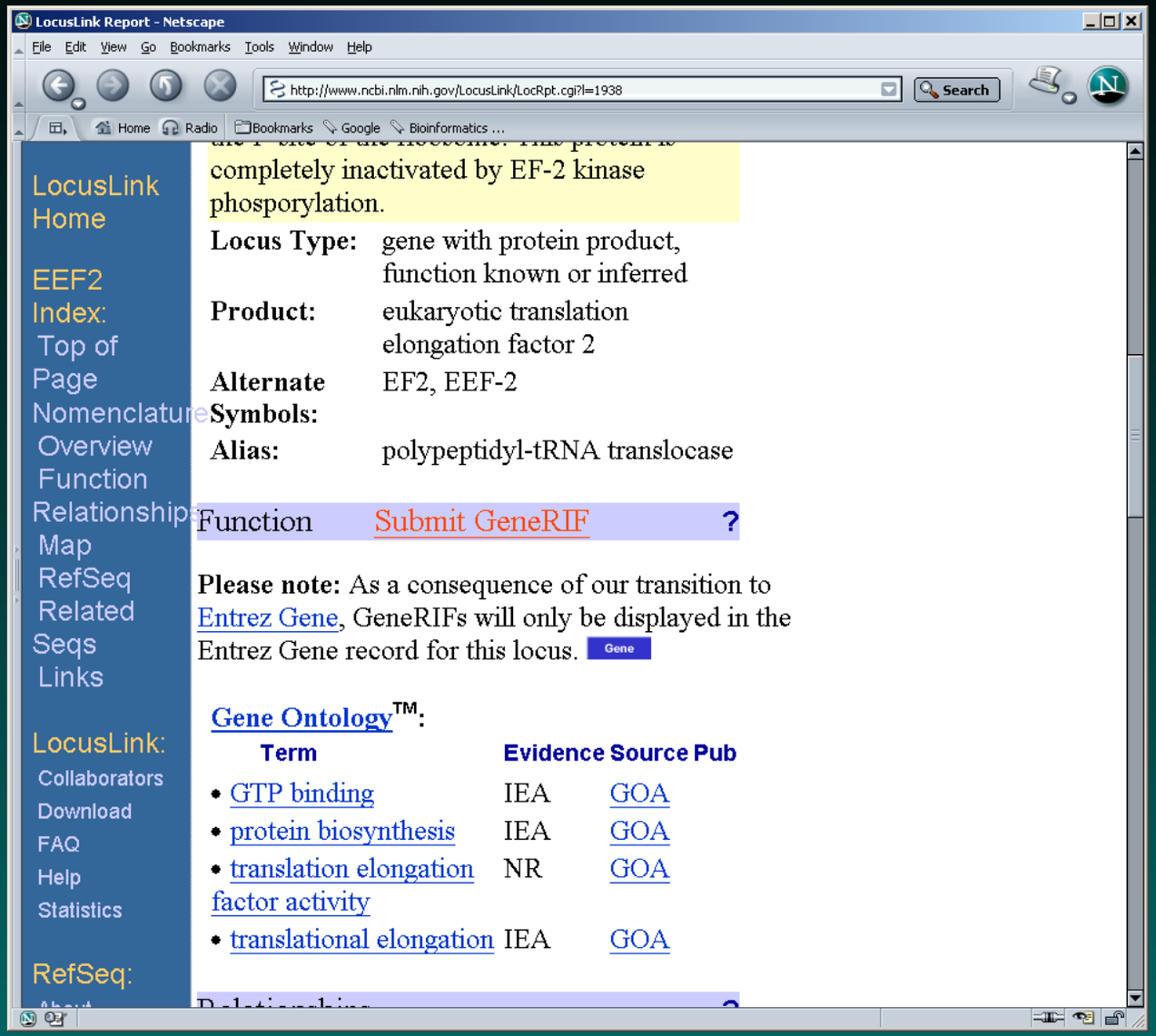
Reprise

- Primary identifiers describe the biological material used as probes
- GenBank accession numbers are a common denominator to start updating gene annotations
- UniGene clusters sequences that represent the same gene
- LocusLink contains curated sequence information about names, mappings, etc
- OMIM contains curated connections between genes and diseases, with references into the literature
- There are batch processing tools to go from a list of GenBank accession numbers through UniGene to LocusLink or OMIM.

Main Question: How do we learn about gene functions, networks, and pathways?

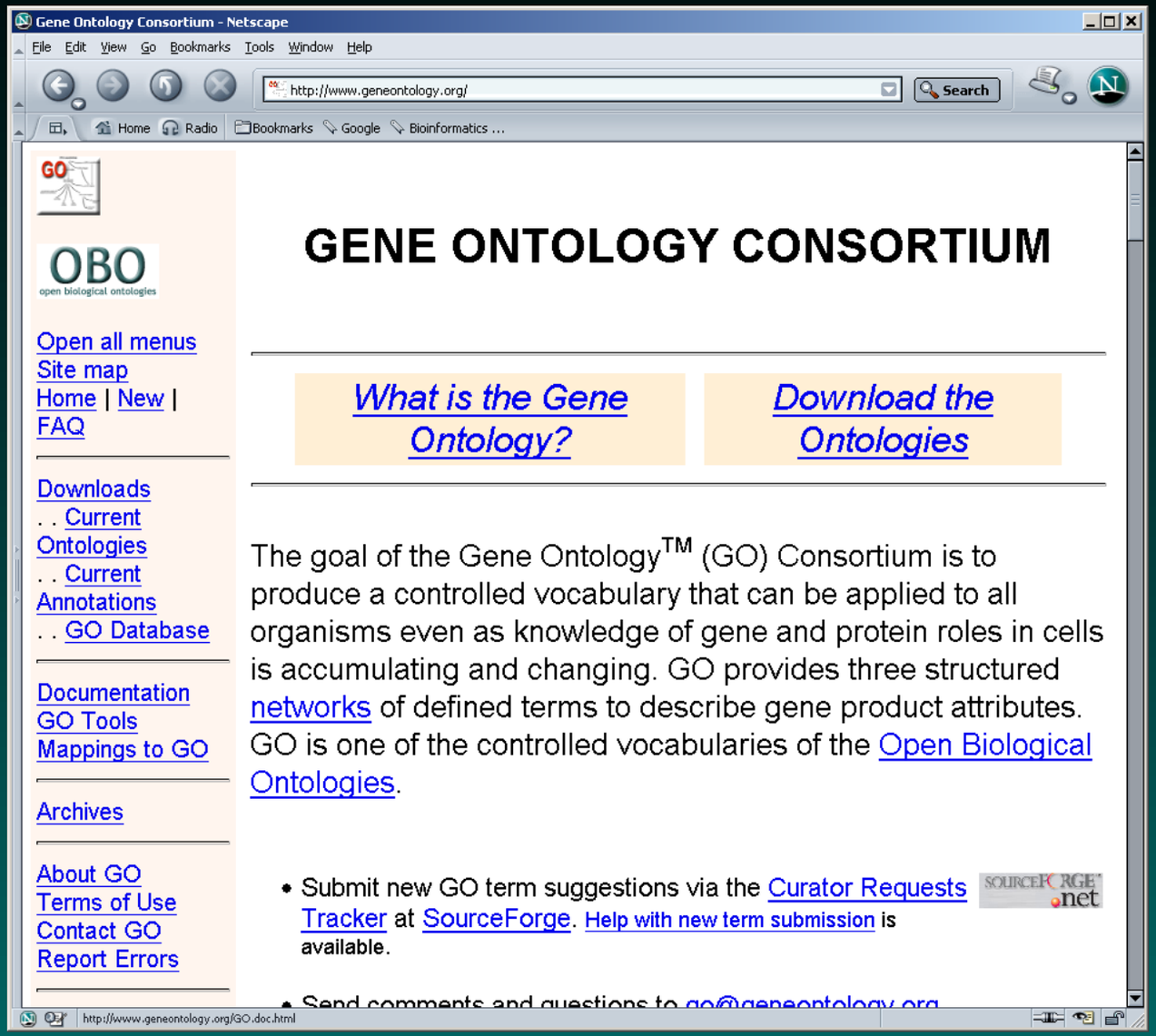
© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

Return to LocusLink



© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

http://www.geneontology.org



GeneOntology

GeneOntology uses controlled vocabularies to create a directed acyclic graph (a generalized tree) that describes the kinds of functions or properties that a gene might have.

The properties are divided into three categories:

1. Biological process (what)
2. Molecular function (how)
3. Cellular component (where)

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly