TRANSCRIPTOME 2002
Seattle, WA

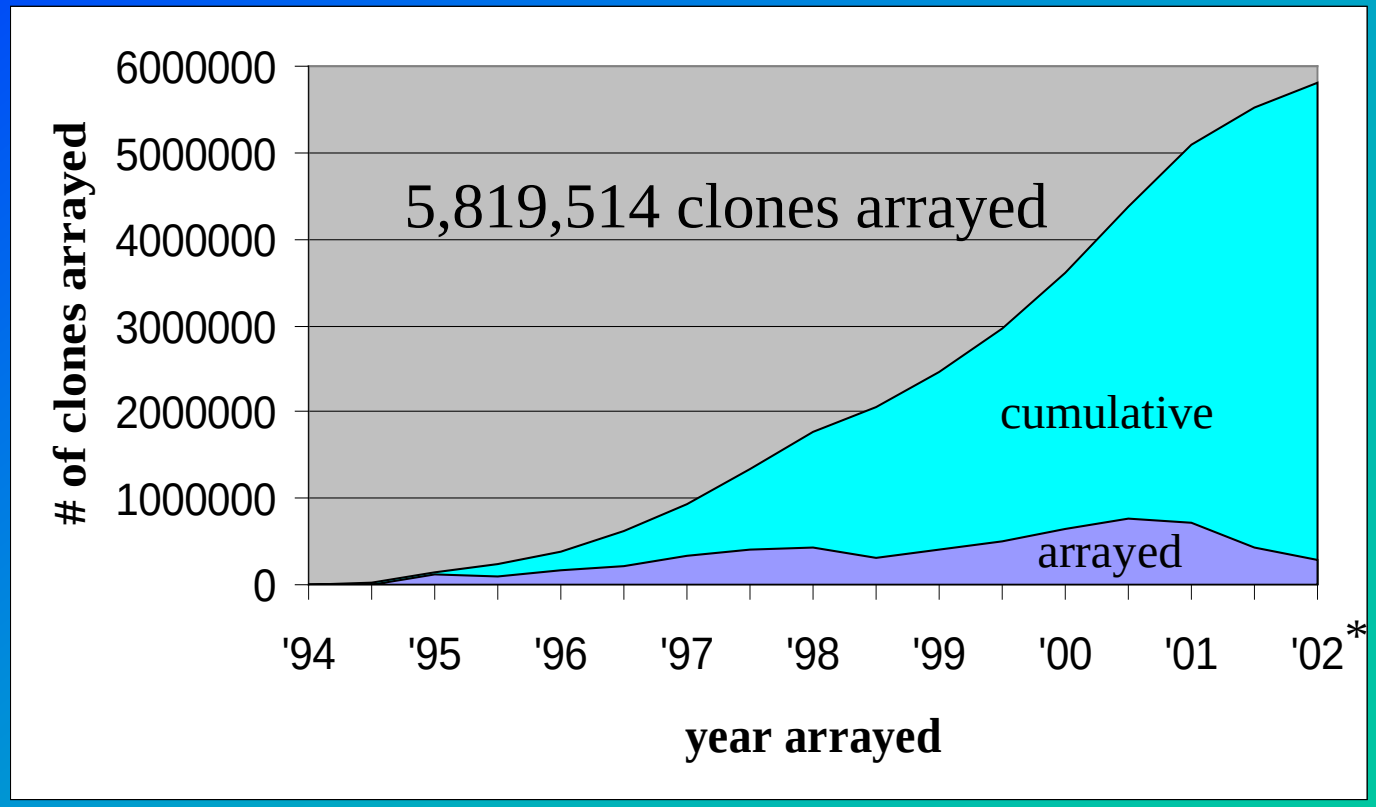
The Integrated Molecular Analysis of Genomes and their Expression Consortium's Data Mining Tools: Introducing the IQ

Peg Folta

Lawrence Livermore National Laboratory

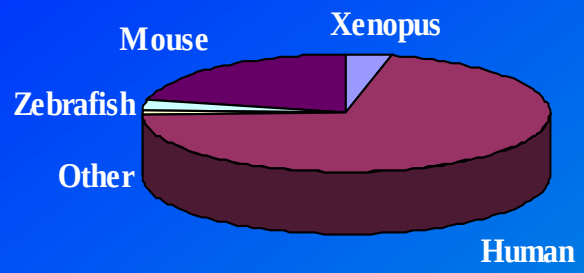
3/12/02

I.M.A.G.E. maintains world's largest publicly available cDNA collection

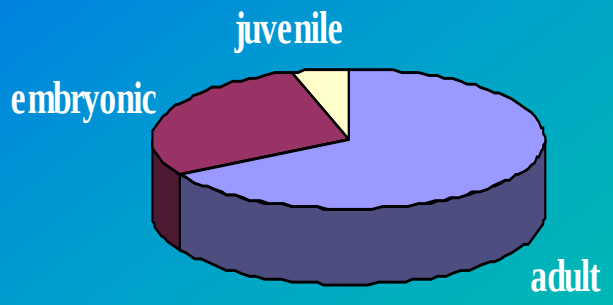


I.M.A.G.E. clones account for 64% of human ESTs in GenBank

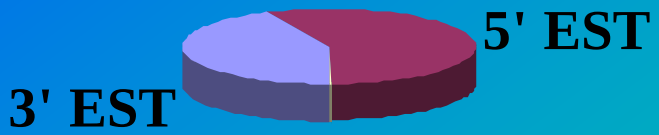
The I.M.A.G.E. collection has been shaped by projects (C-GAP, MGC…)



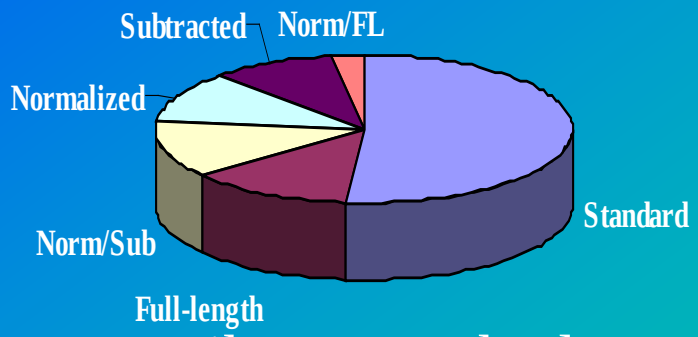
Species



Developmental state



Clone sequence



Library Method

Tissue

Informatics focus this year was on tools to characterize and query the collection.

- IMAGEne – mature clustering tool
- IMAGEne Tissue – allows searching of tissue type dominance in clusters
- IQ – Intelligent Query tool allows mining of I.M.A.G.E. data
- Library/plate query – allows selective searching of libraries and plates
- Problem report and query – allows users to report or query problems related to I.M.A.G.E. clones

Redesign of data management system

IMAGEne-Human Process

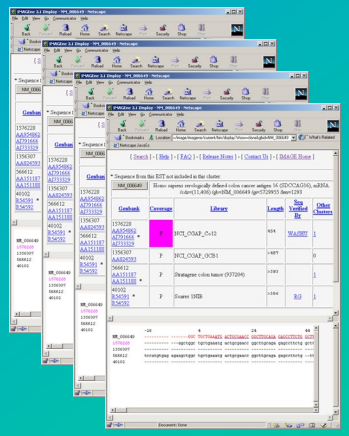
14,566
NCBI
Ref Seq

279,262
Lower quality
I.M.A.G.E
Sequences

2,289,020
Quality
I.M.A.G.E
sequences

>50 basepairs of contiguous, non-repeat sequence

IMAGEne



1,676,516
Sequences

623,294
Sequences

Remaining
Sequences

14,566
Known
Clusters

67,521
Candidate
Clusters
w/consensus

268,472
I.M.A.G.E.
Singletons

Initial query page, construct the query.

[Search] - [Help] - [FAQ] - [Release Notes] - [Contact Us] - [IMAGE Home]

All sequence from I.M.A.G.E. human cDNA clones have been used to create two types of gene clusters: Known Gene Clusters based on NCBI's Reference Genes, and Candidate Gene Clusters having no known gene association. TIGR clones are used to supplement the Candidate Gene clustering. Those clones whose sequence does not match any other cDNA are grouped as Singletons. Submit a query against either cluster set to obtain a ranked display of available I.M.A.G.E. clones aligned with the corresponding known human gene or consensus sequence. *Alternative methods to query against the Singleton database will be provided at a later date.*

IMAGEne Release 3.3 (data retrieved from Genbank 1-8-02)

14566 Known gene clusters					75805 Multi-Member Candidate gene clusters	284749 ESTs from 268472 Singleton Clones(no query yet)
8834 Full	1493 Predicted Full	348 Unknown	3252 Partial	639 Empty		

Search by: Gene Name/Keyword

Gene Name/Keyword
IMAGE Clone ID
GB Accession
Sequence
Cluster ID

Query

cancer

Minimum Blast2 Score 350

Show Alignments by Clone

Search Reset

Document: Done

Clusters matching query results, chose your cluster.

[Search] - [Help] - [FAQ] - [Release Notes] - [Contact Us] - [IMAGE Home]

Matches: 63

cluster id	description	fulls	predicted_fulls	unknowns	partials
NM_000059	Homo sapiens breast cancer 2, early onset (BRCA2), mRNA. /cds=(229,10485) /gb=NM_000059 /len=10987	0	0	5	19
NM_000249	Homo sapiens mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2) (MLH1) mRNA. /cds=(22,2292) /gb=NM_000249 /len=2484	2	1	46	47
NM_000251	Homo sapiens mutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1) (MSH2) mRNA. /cds=(69,2873) /gb=NM_000251 /len=3145	0	0	45	28
NM_000314	Homo sapiens phosphatase and tensin homolog (mutated in multiple advanced cancers 1) (PTEN) mRNA, and translated products. /cds=(1035,2246) /gb=NM_000314 /len=3160	2	0	42	72
NM_001327	Homo sapiens cancer/testis antigen (CTAG1), mRNA. /cds=(89,631) /gb=NM_001327 /len=806	6	0	4	6
NM_001635	Homo sapiens amphiphysin (Stiff-Mann syndrome with breast cancer 128kD autoantigen) (AMPH), mRNA. /cds=(75,2162) /gb=NM_001635 /len=3260	0	1	41	19
NM_002387	Homo sapiens mutated in colorectal cancers (MCC) mRNA. /cds=(221,2710) /gb=NM_002387 /len=4181	1	1	1	7
NM_003087	Homo sapiens synuclein, gamma (breast cancer-specific protein 1) (SNCG), mRNA. /cds=(49,432) /gb=NM_003087 /len=720	13	0	21	8
NM_003225	Homo sapiens trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in) (TFF1) mRNA. /cds=(41,295) /gb=NM_003225 /len=540	9	0	18	1
NM_003567	Homo sapiens breast cancer anti-estrogen resistance 3 (BCAR3) mRNA. /cds=(61,2538) /gb=NM_003567 /len=3004	0	0	29	33
NM_003627	Homo sapiens prostate cancer overexpressed gene 1 (POV1) mRNA, and translated products. /cds=(77,1756) /gb=NM_003627 /len=2326	1	1	34	20
	Homo sapiens Deleted in oral cancer-1 (DOC1) mRNA. /cds=(523,870)				

Document: Done

Display of cluster

NM_002387

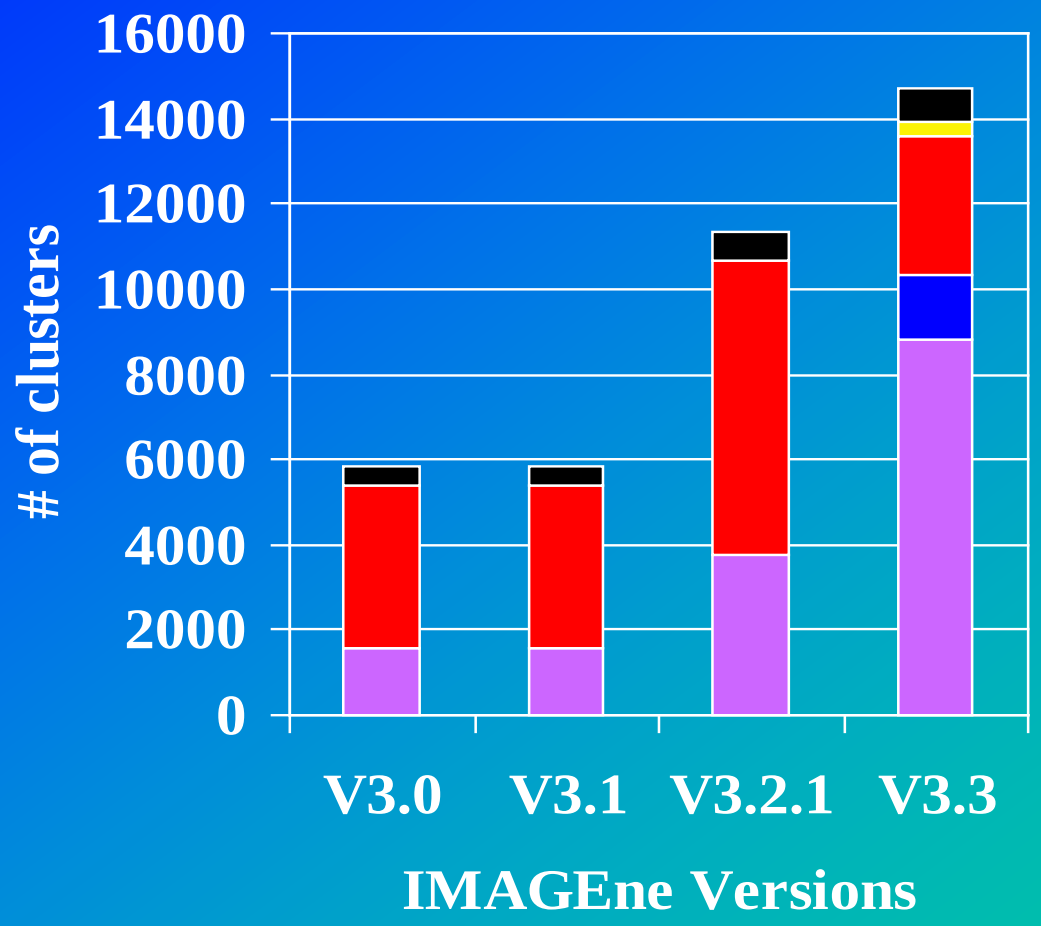
Homo sapiens mutated in colorectal cancers (MCC) mRNA. /cds=(221,2710) /gb=NM_002387 /len=4181

Genbank	Coverage	Library	Length	Seq Verified By	Other Clusters
4111706 BF203829 BC009279	F	NIH_MGC_17	3440	MGC	0
4131685 BF308292	PF	NIH_MGC_17	>607		0
3833189 BE730255	U	NIH_MGC_20	>812		0
3344331 BE259146	P	NIH_MGC_16	>805		0
5458442 BM011740	P	NIH_MGC_47	>650		0

```
            200                 220                 240                 260
NM_002387   CCCAGCTCTG AAAAGTGCAT CATGAATTCC GGAGTTGCCA TGAAATATGG AAACGACTCC TCGC
4111706     cccatgtctg aaaagtgcat catgaattcc ggagttgcca tgaaatatgg aaacgactcc tcgg
4131685     cccagctctg aaa-gtgcat catgaattcc ggagttggca tgaaatatgg aaacgactcc tcgg
3833189     gcaggccggg ggtatcagat ctgctaacgt gcgtcgaaca gagacgagag aacagagact tgat
3344331     ---------- ---------- ---------- ---------- ---------- ---------- ----
5458442     ---------- ---------- ---------- ---------- ---------- ---------- ----
3960921     ---------- ---------- ---------- ---------- ---------- ---------- ----
3356669     ---------- ---------- ---------- ---------- ---------- ---------- ----
4777009     ---------- ---------- ---------- ---------- ---------- ---------- ----
2695302     ---------- ---------- ---------- ---------- ---------- ---------- ----
263970      ---------- ---------- ---------- ---------- ---------- ---------- ----
```

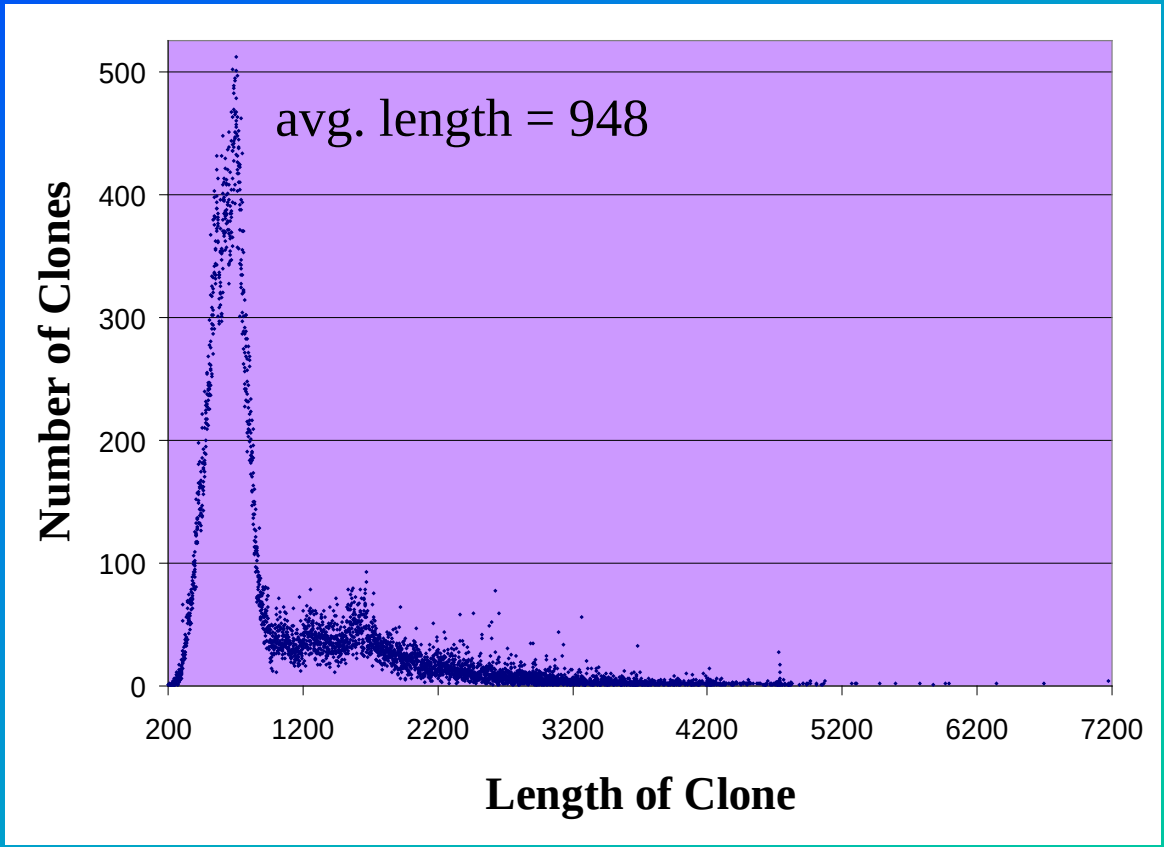
Applet Main running

Known gene clusters with full length I.M.A.G.E. clones have doubled in number.

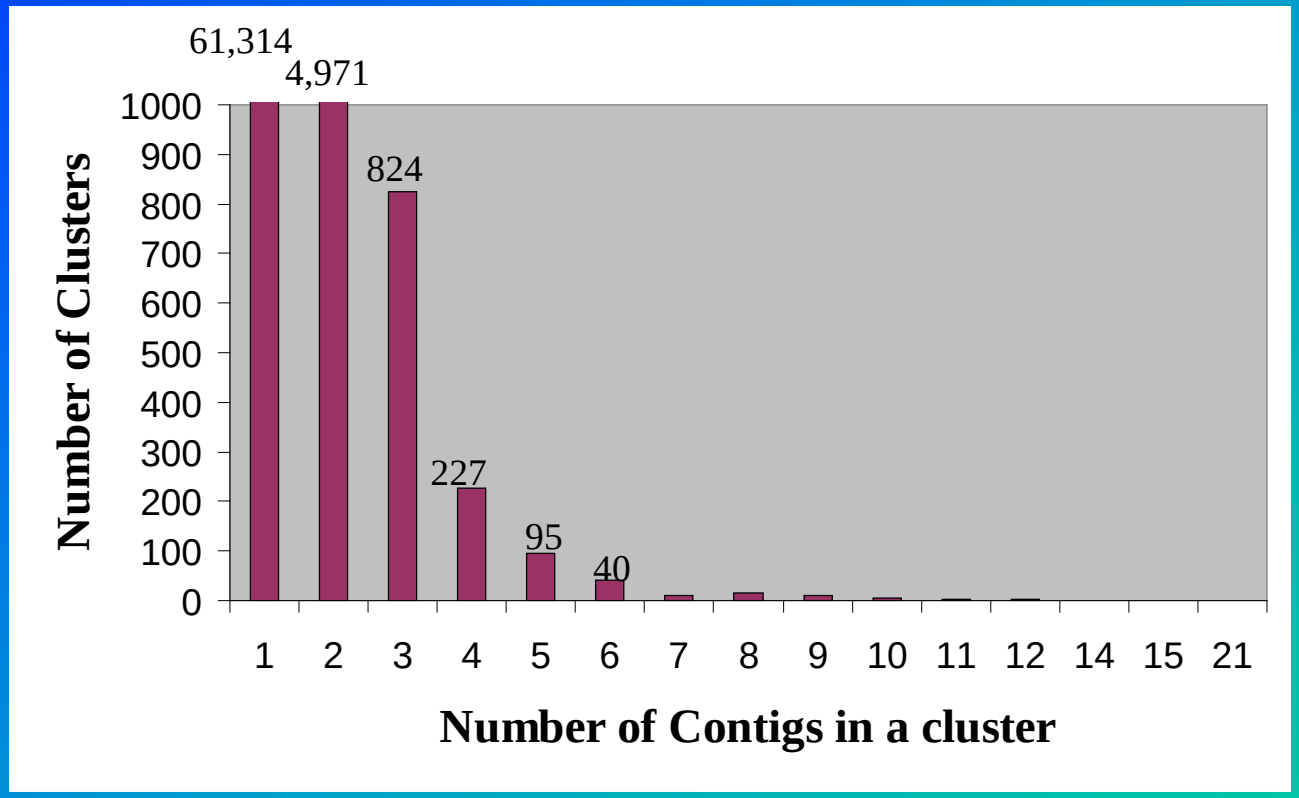


Cluster coverage	Avg. gene length
Empty	**1578**
Unknown	**3392**
Partial	**2763**
Predicted Full	**3380**
Full	**1896**

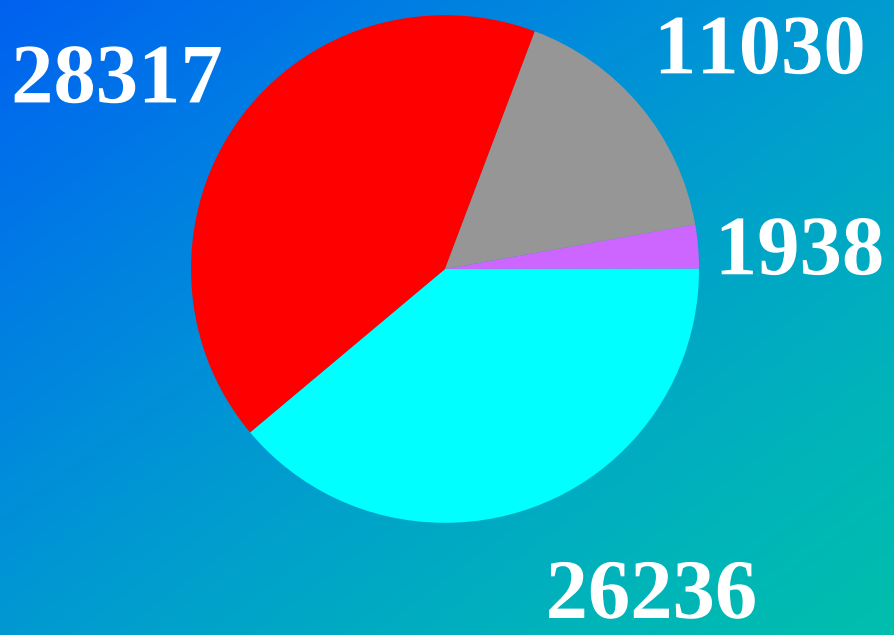
Known Gene Cluster distribution of full length clones



Candidate gene clusters consensus sequence and contigs are generated by CAP4

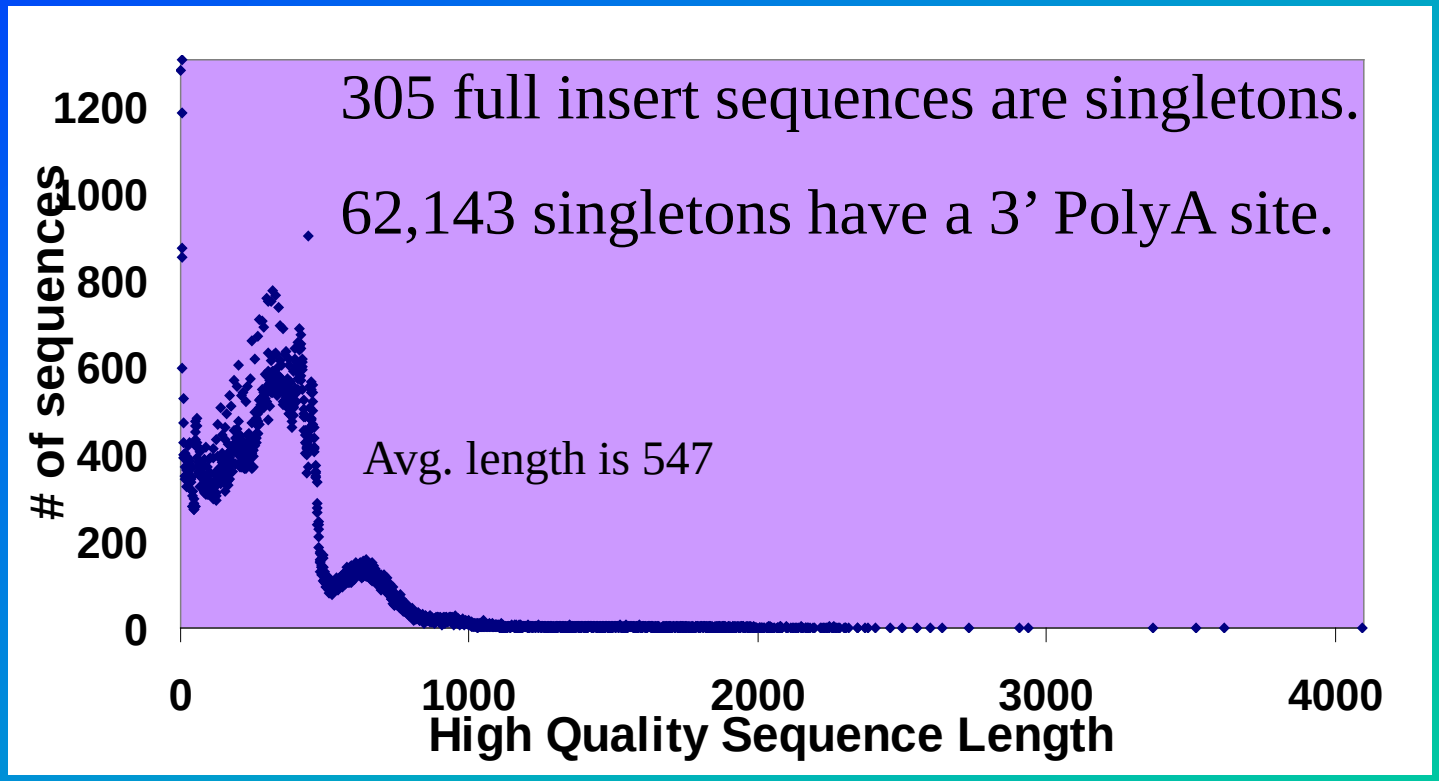


Candidate Gene cluster characteristics.

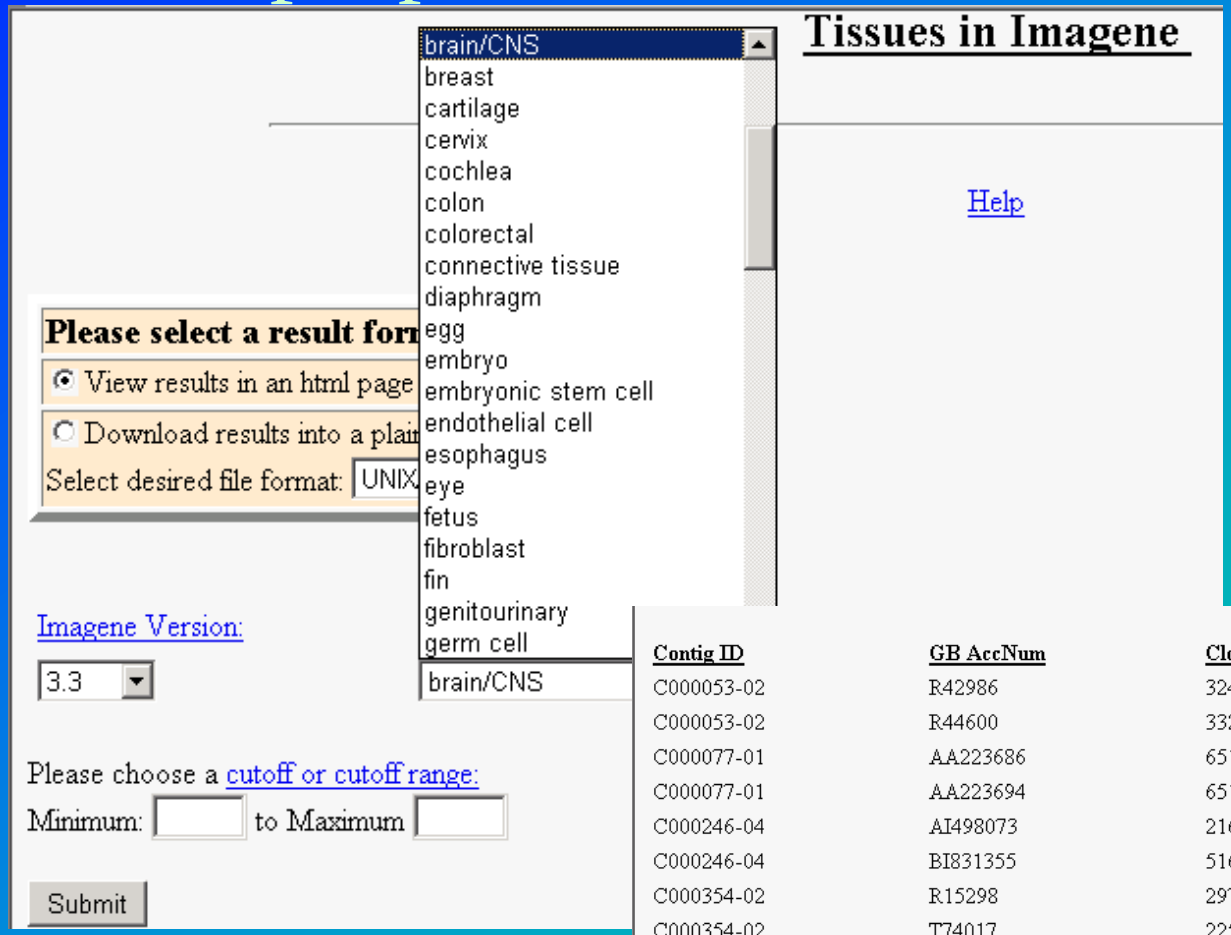


full insert | 3'&5' | 3' only | 5' only

Singleton: Wheat within the chaff



IMAGEne Tissue query allows searching for tissue proportions within clusters.



Contig ID	GB AccNum	Clone ID	Tissue	Ratio
C000053-02	R42986	32492	brain/CNS	1.0000
C000053-02	R44600	33222	brain/CNS	1.0000
C000077-01	AA223686	651002	brain/CNS	1.0000
C000077-01	AA223694	651003	brain/CNS	1.0000
C000246-04	AI498073	2163041	brain/CNS	1.0000
C000246-04	BI831355	5166226	brain/CNS	1.0000
C000354-02	R15298	29768	brain/CNS	1.0000
C000354-02	T74017	22595	brain/CNS	1.0000
C000522-02	BI915961	5241183	brain/CNS	1.0000
C000522-02	BI457868	5277635	brain/CNS	1.0000
C000581-03	BG910399	4938321	brain/CNS	1.0000

Introducing the Intelligent Query - IQ

- For a given category (currently clone and library) a user can specify a query based on key database attributes.
- The user can specify the fields returned.
- Various result format options (HTML, text)
- Initial version was rolled out last summer
- New functionality to be added this year (additional categories, etc.)

Specify a clone-based query.

Submit | Clear Form

Field Name	Data Input
Clone ID	Convert
Plate Number	Convert
I.M.A.G.E. Library ID	Convert
I.M.A.G.E. Library Name	Appel/Eisen 15-19 hr embryo Baker mouse embryo e6.5 Baker mouse embryo e7.5 Barstead HPL-RB1 Barstead HPL-RB2 Barstead HPL-RB3 Barstead HPL-RB4 Barstead HPL-RB4.1 Barstead HPL-RB5 Barstead HPL-RB6 Clear List
Row Position	none selected
Column Position	none selected
Genbank Accession Number	Convert
Genbank Identifier	Convert
Tissue Type	none selected
Is Reversed	yes no
Is Low Quality	yes no
Clone collection	none selected
Species	Xenopus laevis Xenopus tropicalis human mouse rat rhesus monkey zebrafish Clear List

Submit | Clear Form

Next specify what clone centric results will be provided and in what format.

Select the results you want to see back from the following list:

Check all | Clear all

Submit

- [x] Row Position
- [x] Column Position
- [x] Plate Number
- [ ] Clone collection
- [x] Clone ID
- [x] Sequence
- [ ] Repeat-masked Sequence
- [ ] Gb date created
- [ ] Date modified
- [x] Genbank Accession Number
- [ ] Genbank Identifier
- [x] Endedness
- [ ] Is Low Quality
- [ ] Is Reversed
- [ ] Seq length
- [ ] I.M.A.G.E. Library ID
- [ ] I.M.A.G.E. Library Name
- [ ] Tissue Count
- [x] Species
- [ ] Tissue Type
- [ ] Vector name

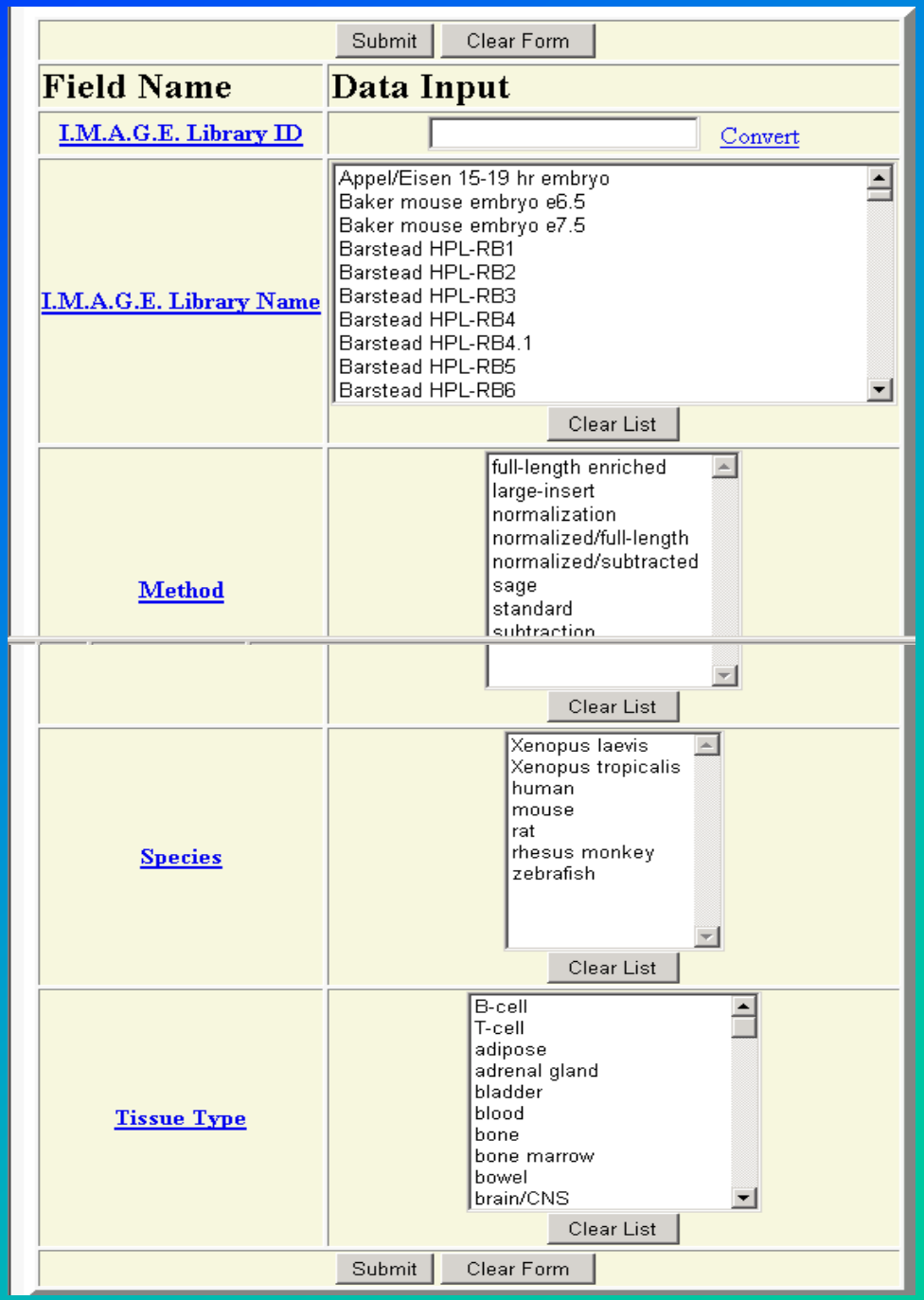
HTML version of clone-based query results.

I.M.A.G.E. Clone Query Results:

Your search returned **34** results! Here they are:

Result Number	PLATE	COLLECTION NAME	CLONE ID	SEQ	GB ACCNUM	ENDEDNESS	SPECIES	TISSUE TYPE
1	1651	LLAM	676104	gcgggttatctgtaaaggcctct aactttgtgaactgagtagcaag tagaagctaaggttacccacgaa ccccactttgcagttccccctcg tcctgctttgcagaatgaggatt ctcgccttgctgtccaccattac cttgggtggtcagttttgttgca	AA208983	5	mouse	liver
2	1652	LLAM	676488	ttcggcacgtagggaaaaaactg cattagattttcccattaaacct tggatccaggtggacatgcagaa ggaagttgtagtcaccgggatac aaacccaaggtgctaaacactac ctaaagtcctgctttaccacgga gttccaagtggcttacagctctg	AA208846	5	mouse	liver
3	1653	LLAM	676872	NULL	NULL	NULL	mouse	liver
4	1654	LLAM	677256	NULL	NULL	NULL	mouse	liver
5	1655	LLAM	677640	NULL	NULL	NULL	mouse	liver
6	1656	LLAM	678024	NULL	NULL	NULL	mouse	liver

Specify a library-based query.



Similarly specify what library centric results will be provided.

Check all | Clear all

Submit

- [ ] I.M.A.G.E. Library Name
- [ ] Tissue Count
- [ ] Species
- [ ] Strain
- [ ] Libr priming
- [ ] Vector digest
- [ ] Seq tag
- [ ] I.M.A.G.E. Library ID
- [ ] Method
- [ ] Host
- [ ] Libr provider
- [ ] Tissue Type
- [ ] Source age
- [ ] Source sex
- [ ] Source desc
- [ ] Source stage
- [ ] Tissue desc
- [ ] Vector name

HTML version of library-based query results.

Biology & Biotechnology Research PROGRAM

The I.M.A.G.E. Consortium

"Sharing resources to achieve a common goal - the discovery of all genes "

Begin new search | Begin new Library search

I.M.A.G.E. Library Query Results:

Your search returned **2** results! Here they are:

Result Number	LIBR NAME	SPECIES	LIBR PRIMING	METHOD	SOURCE AGE	SOURCE DESC	TISSUE DESC
1	NIH_MGC_87	human	oligo dT	full-length enriched	NULL	NULL	mammary adenocarcinoma, cell line
2	NIH_MGC_107	human	oligo dT	full-length enriched	NULL	NULL	adenocarcinoma, cell line

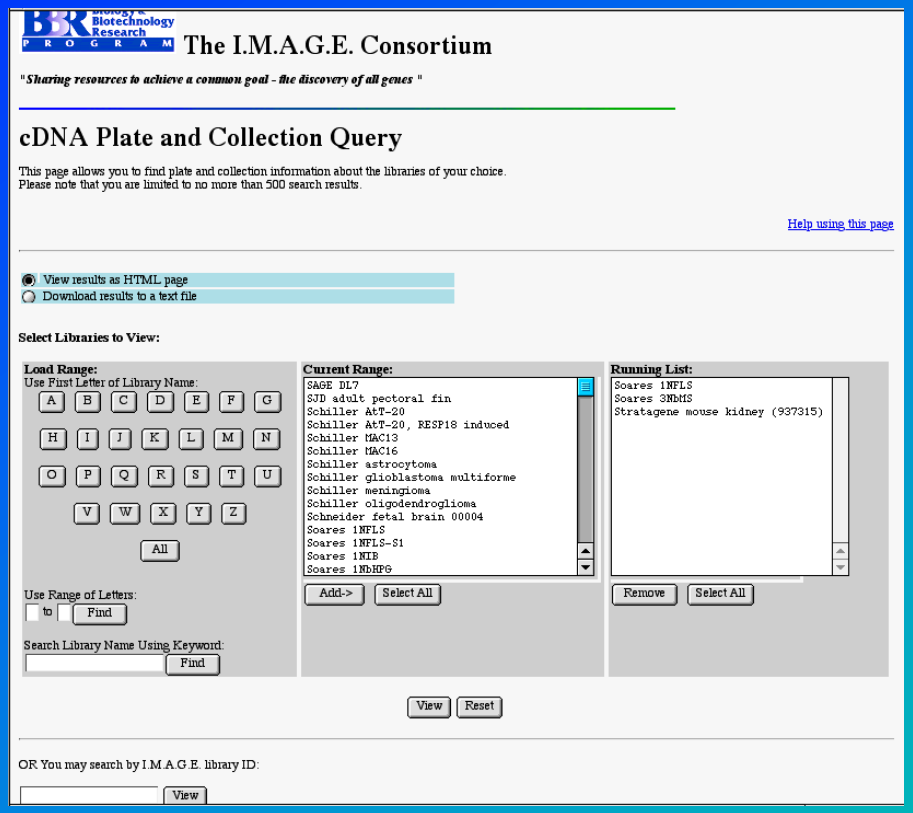
Begin new search | Begin new Library search

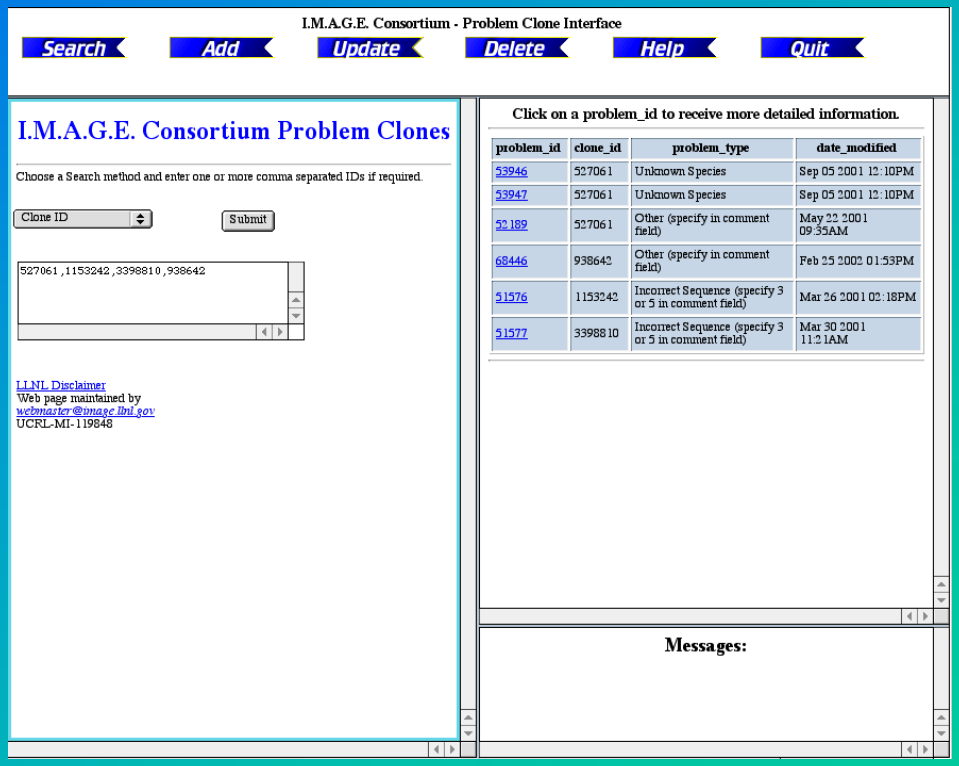
I.M.A.G.E. Consortium home page

BBRP home page

LLNL Programs, Projects, Centers and Consortia

Other tools to mine I.M.A.G.E. information





problem_id	clone_id	problem_type	date_modified
53946	527061	Unknown Species	Sep 05 2001 12:10PM
53947	527061	Unknown Species	Sep 05 2001 12:10PM
52189	527061	Other (specify in comment field)	May 22 2001 09:35AM
68446	938642	Other (specify in comment field)	Feb 25 2002 01:53PM
51576	1153242	Incorrect Sequence (specify 3 or 5 in comment field)	Mar 26 2001 02:18PM
51577	3398810	Incorrect Sequence (specify 3 or 5 in comment field)	Mar 30 2001 11:21AM

Query plates from libraries.

Query for reported problems.

Quality control for historical collection

Plates	Source	Well Error Rate
1-3705	Incyte	13
	LLNL Master	10
	Research Genetics	12
	Resource Center of HumanGenome Project	10
	ATTC	11
3,796-6000	Incyte	7
	LLNL Master	7
	Research Genetics	10
	Resource Center of Human Genome Project	12

QC on-going

	LLNL Replication		Master vs. GenBank	
Months	Well error rate	Plate Error Rate	Well error rate	Plate Error Rate
6/2000	1 (1,3)	0	7 (4,11)	2
10/2000	1 (0,3)	0	1 (0,3)	2
12/00	0 (0,2)	2	1 (0,3)	2
1/01	2 (1,4)	0	6 (4,11)	3
2/01	1 (0,3)	0	2 (1,5)	2
3/01	2 (1,5)	2	2 (1,5)	0
4/01	1 (0,3)	2	2 (1,4)	0
5/01	0 (0,1)	0	2 (1,5)	0
6/01	1 (0,3)	0	1 (0,4)	0
7/01	1 (0,4)	0	2 (1,6)	0
8/01	2 (1,3)	0	3 (2,6)	0

Ongoing QC results

R0 (LLNL Master Plate) compared to R3 (from R0, by LLNL, for LLNL)

Month	Number of Clones		Well Error (%)		Number of Plates		Plate Error (%)	Calc Date	Link to Details
	Compared	In Error	Rate	95% C.I.	Total	In Error			
6 / 2000	391	5	1	(0, 3)	57	0	0	Aug 15 2001	6 / 2000
10 / 2000	409	5	1	(0, 3)	56	0	0	Oct 30 2001	10 / 2000
11 / 2000	*	*	1	(*,*)	60	0	0	Feb 20 2001	11 / 2000
12 / 2000	407	2	0	(0, 2)	58	1	2	Aug 15 2001	12 / 2000
1 / 2001	344	7	2	(1, 4)	57	0	0	Aug 15 2001	1 / 2001
2 / 2001	385	5	1	(0, 3)	59	0	0	Aug 15 2001	2 / 2001
3 / 2001	323	8	2	(1, 5)	50	1	2	Aug 15 2001	3 / 2001
4 / 2001	372	5	1	(0, 3)	57	1	2	Aug 15 2001	4 / 2001
5 / 2001	345	0	0	(0, 1)	48	0	0	Oct 30 2001	5 / 2001
6 / 2001	359	3	1	(0, 3)	52	0	0	Nov 14 2001	6 / 2001
7 / 2001	304	4	1	(0, 4)	49	0	0	Nov 14 2001	7 / 2001
8 / 2001	389	6	2	(1, 3)	58	0	0	Jan 22 2002	8 / 2001
9 / 2001	154	3	2	(1, 6)	25	1	4	Jan 22 2002	9 / 2001

Error in replication @ LLNL

On-going
Comparing master to GenBa

Month	Number of Clones		Well Error (%)		Number of Plates		Plate Error (%)	Calc Date	Link to Details
	Compared	In Error	Rate	95% C.I.	Total	In Error			
6 / 2000	230	15	7	(4, 11)	40	1	2	Aug 15 2001	6 / 2000
10 / 2000	310	3	1	(0, 3)	50	1	2	Oct 30 2001	10 / 2000
11 / 2000	198	4	2	(1, 5)	32	0	0	Aug 15 2001	11 / 2000
12 / 2000	276	3	1	(0, 3)	46	1	2	Aug 15 2001	12 / 2000
1 / 2001	187	12	6	(4, 11)	36	1	3	Aug 15 2001	1 / 2001
2 / 2001	253	5	2	(1, 5)	45	1	2	Aug 15 2001	2 / 2001
3 / 2001	266	6	2	(1, 5)	46	0	0	Aug 15 2001	3 / 2001
4 / 2001	245	4	2	(1, 4)	43	0	0	Aug 15 2001	4 / 2001
5 / 2001	201	4	2	(1, 5)	34	0	0	Oct 30 2001	5 / 2001
6 / 2001	239	3	1	(0, 4)	41	0	0	Oct 30 2001	6 / 2001
7 / 2001	152	3	2	(1, 6)	27	0	0	Oct 31 2001	7 / 2001
8 / 2001	287	9	3	(2, 6)	46	0	0	Jan 30 2002	8 / 2001
9 / 2001	16	12	75	(47, 92)	38	34	89	Jan 22 2002	9 / 2001

Next for I.M.A.G.E. Informatics

- Extensive expansion of query tools and data access
- IMAGEne non-species specific
- Analysis of human cluster candidate genes and singletons
- Redo of web site, easier to navigate

MUCH influenced by public needs…..you have a say!

Acknowledgements

- LLNL
  - Christa Prange, I.M.A.G.E. PI
  - Tim Harsch, Amber Johnston, Julie Amundson

- Sponsors
  - DOE, Marv Stodolsky
  - NIH, Bob Strausberg

image.llnl.gov

This work was partially funded by the NIH and was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.