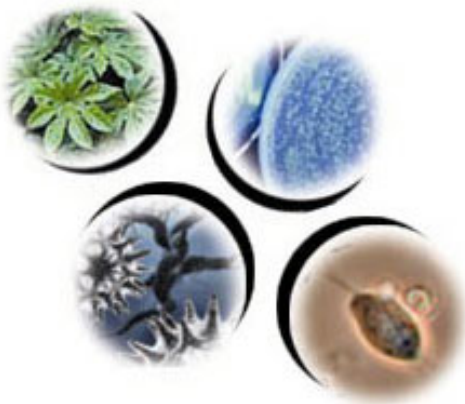
BioHealthBase

BioDefense Public Health Database



Bioinformatics Resource Center (BRC) for Biodefense and Emerging/Re-emerging Infectious Diseases

NIH Contract Number: HHSNS66200400041C ADB Contract Number: N01-AI-40041

Kick-Off Meeting - 13OCT2004

NATIONAL INSTITUTES OF HEALTH

NIAID

NORTHROP GRUMMAN

Information Technology

Partnership

BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC

- Purpose
  - Establishment and maintenance of a bioinformatics resource center
- Components
  - Web-based, curated, stable, relational databases
  - Collect, store and analyze
  - Genomic and emerging related data
  - Category A - C microorganisms and emerging/re-emerging infectious disease agents
  - Components include databases, analytical tools, user interfaces and training/support resources

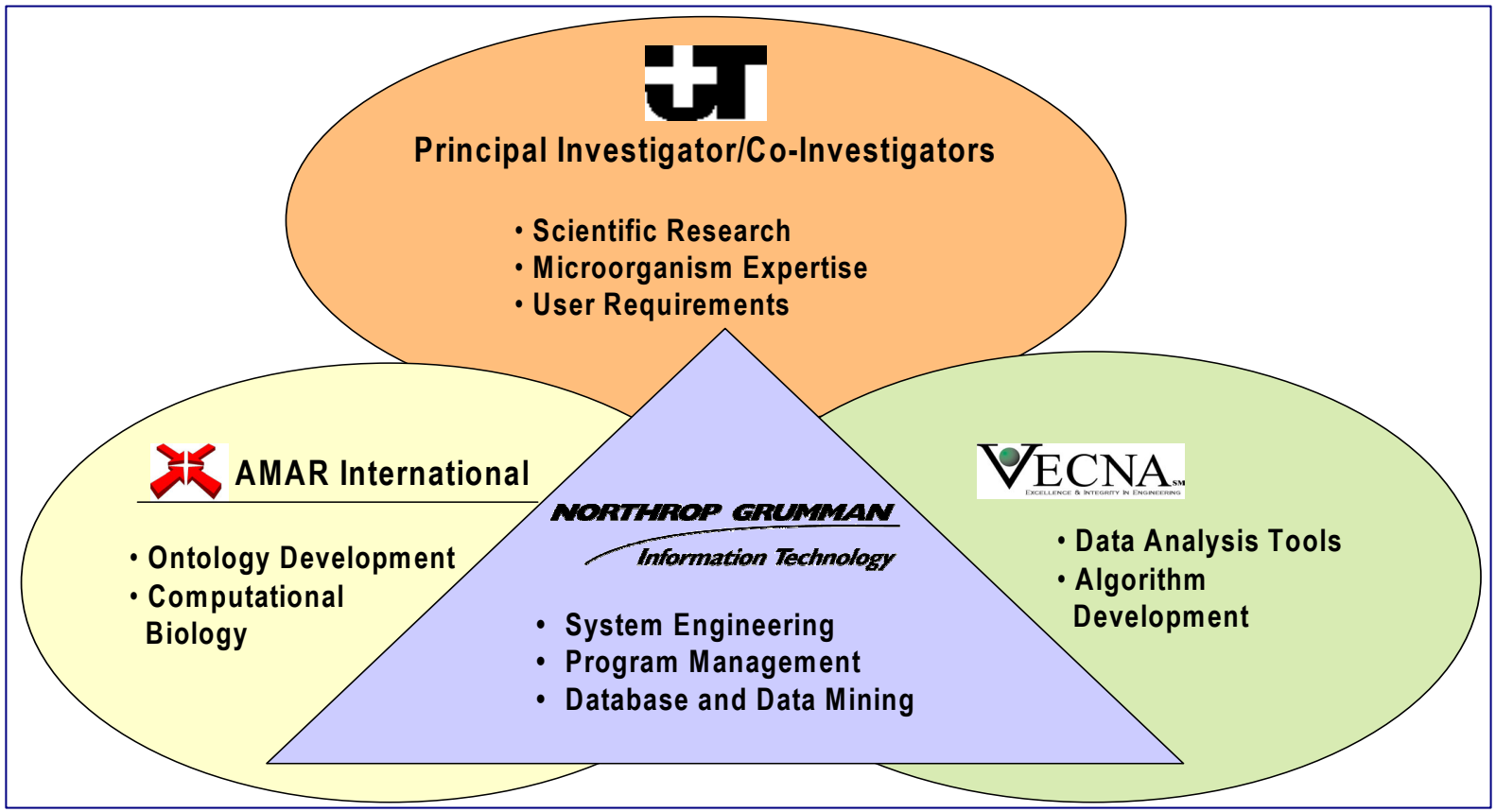
BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC Organisms

Influenza virus
Mycobacterium tuberculosis
Giardia lamblia
Entamoeba histolytica
Microsporidia
Ricinus communis

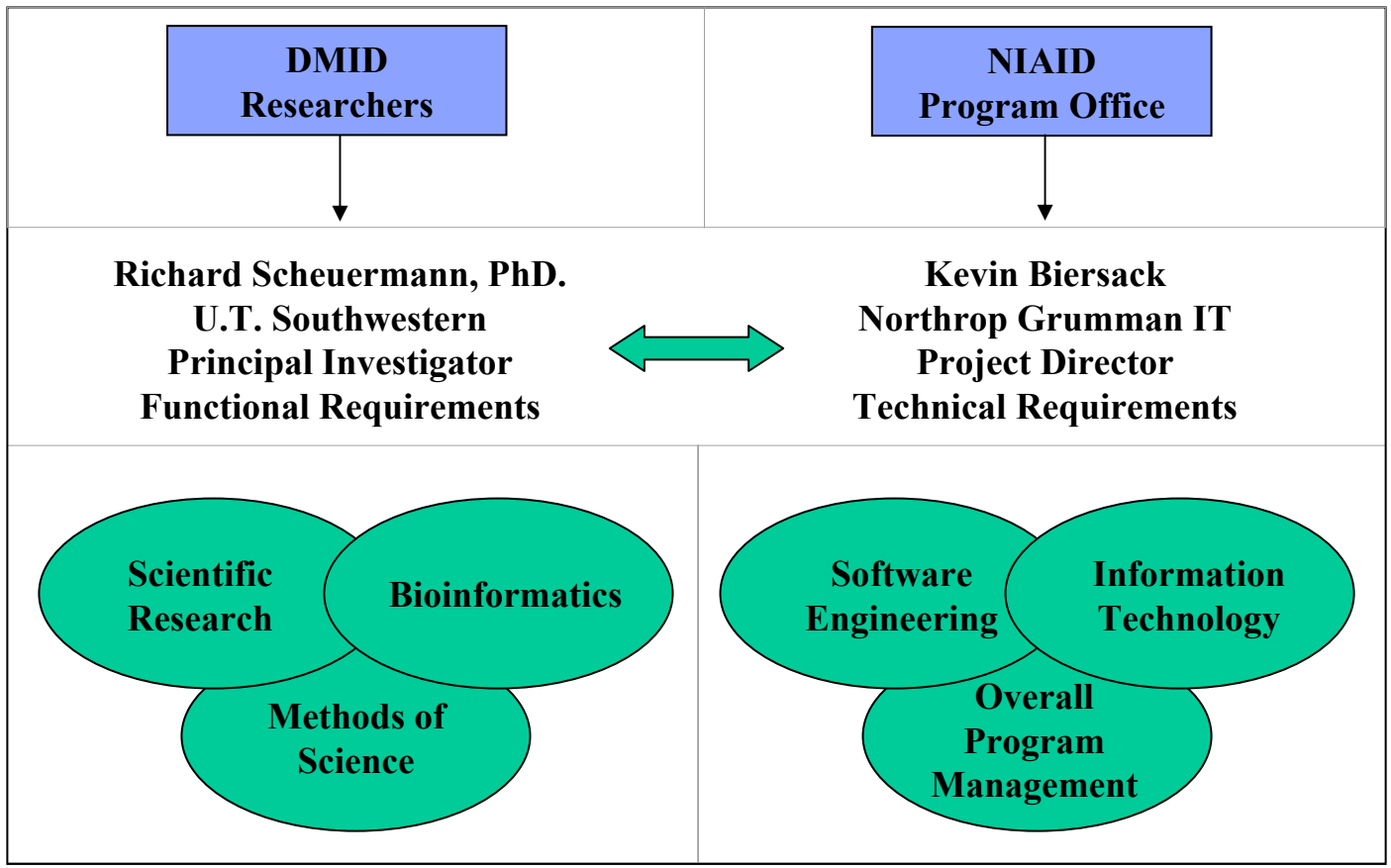
BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC Team



BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC Team Management



BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC Team Members

U.T. Southwestern Medical Center
Dr. Stephen Johnston, *M. tuberculosis* Co-PI
Dr. Ellen Vitetta, *R. communis* Co-PI
Dr. Feng Luo, Computational Biologist
Burke Squires, Database Analyst

Mt Sinai School of Medicine
Dr. Adolfo Garcia-Sastre, *Influenza* Co-PI

University of British Columbia
Dr. Patrick Keeling. *Microsporidia* Co-PI

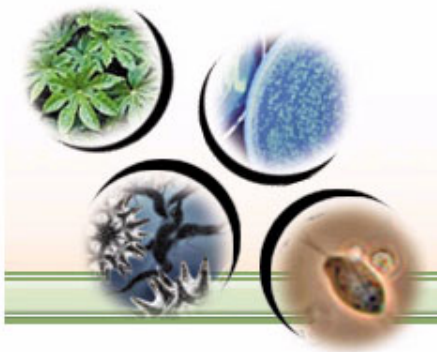
University of Virginia
Dr. Barbara Mann, *E. histolytica* Co-PI

Woods Hole Oceanographic Institute
Dr. Mitchell Sogin, *G. lamblia* Co-PI

Northrop Grumman IT
Juan Jose Cabrera, IS Technologist
Petra deBruin, IS Technologist
DeWayne Ejikeme, Test Engineer
Jason Lucas, Requirements/Config. Engineer
Johanan Lancaon, Software Architecture
Lyudmila Trebegova, Data Architect
Bryan Walters, Network Engineer

Vecna Technologies, Inc.
Carrie Stein, Product Manager
Dr. Rob Plovnick, Senior Software Engineer
Dr. Steve Niemczyk, Senior Systems Architect
Maki Duncan, Project Manager

AMAR International, Inc.
Ontology Development



BioHealthBase

BioDefense Public Health Database

Home | Microorganisms | Data Sources | Team | Schedule | Contact Us

The National Institute of Allergy and Infectious Disease (NIAID) Division of Microbiology and Infectious Diseases (DMID) recognizes the challenge posed by bioterrorism, the emergence of disease due to drug-resistent variants of etiologic organisms. DMID has envisioned a consortium of Bioinformatics Resource Centers (BRCs) for Biodefense and Emerging/Re-emerging Infectious Diseases that will provide information technology (IT) support for experimental studies of pathogenic organisms that could be used for biowarfare and bioterrorist activities, many of which also pose an ongoing threat to public health. The BRCs will provide both central repositories for a wide variety of scientific data on these pathogenic microorganisms and a platform for software tools that support investigator-driven data analysis. A description of the NIAID BRC program can be found at **www.niaid.nih.gov/dmid/genomes/brc/default.htm**.

BioHealthBase SYSTEM

The ***Bio***defense/Public ***Health*** Data***Base*** (***BioHealthBase***) BRC focuses on data about six priority pathogens to help fill in gaps in genomic and other data critical to scientific researchers. The six pathogens are Giardia lamblia (giardiasis) parasite; Mycobaterium tuberculosis (TB) bacteria; Influenza (flu) virus; Entamoeba histolytica (amoebiasis) parasite; eight genera of Microsporidia parasites; and Ricinus communis (castor bean) plant. The organisms were selected to represent a cross-section of the major groups of human pathogens, including bacteria, viruses and protozoan parasites. The majority also pose an ongoing threat to public health in the U.S. and abroad, especially in Third World countries.

Home | Microorganisms | Data Sources | Team |
Contact Us | Privacy Statement | Disclaimers | Accessibility

www.ngchealth.com/gov/nih/niaid/biohealthbase

BioHealthBase
BioDefense Public Health Database

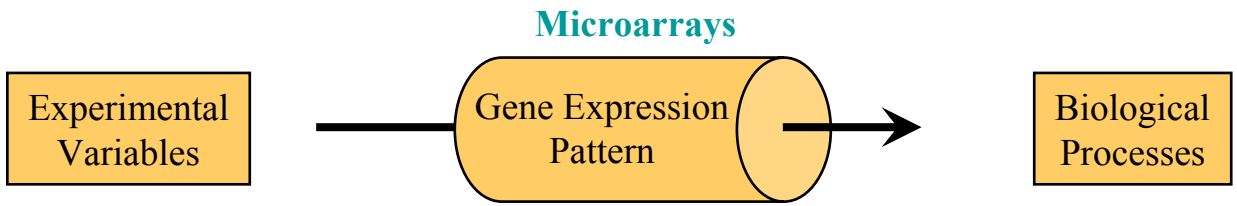
BioHealthBase BRC - Key features

- Beyond genomics
  - Support data regarding:
    - Functional genomics
    - Proteomics
    - ELISA
    - …. and beyond
  - Support data analysis
    - Analytical applications
    - Training
- Focus on host-pathogen interactions
  - Disease pathogenesis
  - Host sensitivity and resistance
  - Vaccinology
- Data integration
  - Supported by generic conceptual data model
  - Adding biological inference to experimental results

BioHealthBase
BioDefense Public Health Database

Data Integration - gene expression & functional gene annotation

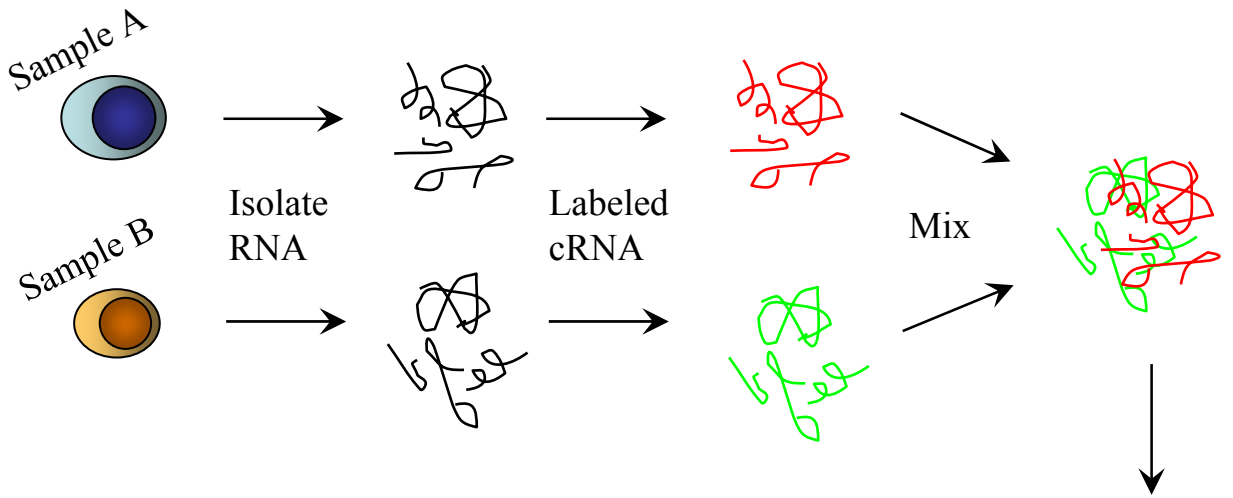
- Gene expression microarrays provide a tool for examining global gene expression patterns
- Approach is based on a comparison of expression patterns measured under different experimental conditions
  - Strain A vs Strain B
  - Condition A vs Condition B
- Goal is to use differences in gene expression patterns to understand the impact that the experimental variables have on cell physiology



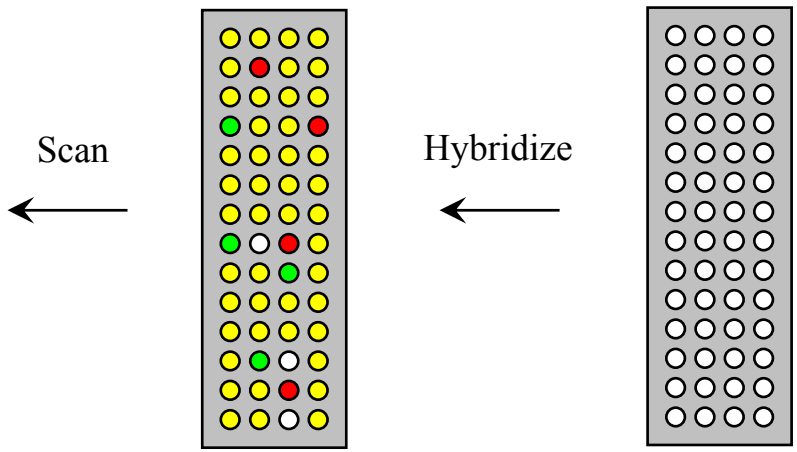
- Achieved through integration of microarray data with orthogonal data sets

BioHealthBase
BioDefense Public Health Database

Experimental Methodology



UNIQID	name	UNT0hr	UNT.5hr	UNT1hr	UNT2hr
2810011G06	homolog	0	0.553029	1.53488	NA
UI-M-AQ(	Mus musc	0	1.10078	0.791893	0.9969
H3122D05	CDK2 (cy	0	0.425552	0.119843	0.2989
H3122D05	CDK2 (cy	0	0.217208	0.014706	0.1438
H3122D05	CDK2 (cy	0	0.512776	-0.087039	0.322
H3077H03	cyclin G	0	0.177256	-0.158146	0.4515
H3122D05	CDK2 (cy	0	0.306455	0.107867	0.2766
2610510C05	cyclin B1,	0	0.011495	-0.370655	0.1666
2610016D16	cyclin D2	0	0.339981	-0.199669	0.0594
H3123E09	cyclin A2	0	0.295244	-0.175816	0.3272
1810028E17	cyclin-dep	0	1.12383	1.055665	0.5423
2810453O06	homolog	0	0.169857	-0.139112	0.0923
2900041I18	homolog	0	-0.030936	0.093679	0.1534
3100002O13	cyclin I	0	-0.049336	-0.29733	-0.0119
H3079A09	cyclin A2	0	-0.225899	0.664657	0.7023
5830435F03	cyclin-dep	0	NA	0.087183	-0.161
1700068G16	cyclin-dep	0	0.222088	-0.0987	0.2311
2310034I23	cyclin C	0	-0.030829	0.619777	0.8998
C030048F16	CYCLIN A	0	0.756419	-0.217605	0.1953
2610203O19	cyclin-dep	0	0.390739	0.041612	0.4092
2700001K05	homolog	0	-0.278647	0.593802	0.2563
5930414J15	cyclin-dep	0	0.664317	0.385508	0.6416
1110019B02	cyclin-dep	0	0.074843	0.149596	0.066

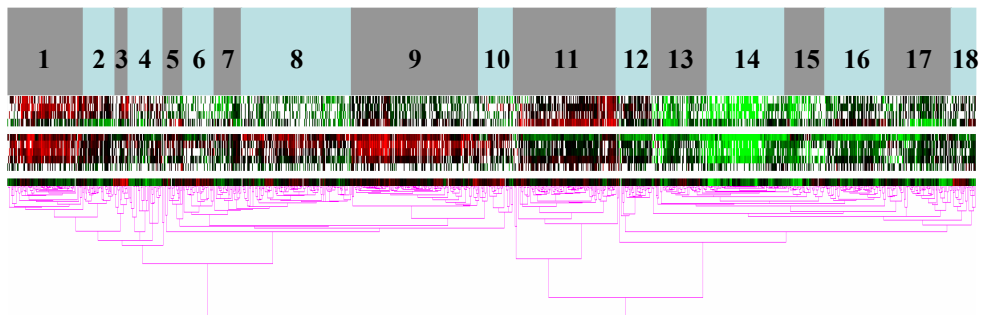


BioHealthBase
BioDefense Public Health Database

Data Processing

- Background subtraction - to reduce effects of non-specific intensity signals
- Data normalization - to allow pooling of data from different chips
- Statistical filtering - based on data precision
- Data clustering - distance measures to calculate the similarity between genes combined with clustering methods to group genes according to similar behaviors.

UNIQID	name	UNT0hr	UNT.5hr	UNT1hr	UNT2hr
2810011G06	omolog	0	0.553029	1.53485	NA
UI-M-AQ0	Mus musc	0	1.10078	0.791893	0.9969
H3122D05	CDK2 (cy	0	0.425552	0.119843	0.2989
H3122D05	CDK2 (cy	0	0.217208	0.014706	0.1438
H3122D05	CDK2 (cy	0	0.512776	-0.087039	0.322
H3077H02	cyclin G	0	0.177256	-0.158146	0.4515
H3122D05	CDK2 (cy	0	0.306455	0.107867	0.2766
2610510G05	cyclin B1,	0	0.011495	-0.370655	0.1666
2610016D16	cyclin D2	0	0.339981	-0.199669	0.0594
H3123E09	cyclin A2	0	0.295244	-0.175816	0.3272
1810028E17	cyclin-dep	0	1.12383	1.055665	0.5423
2810453O06	omolog	0	0.169857	-0.139112	0.0923
2900041I08	omolog	0	-0.030936	0.093679	0.1534
3100002O13	cyclin I	0	-0.049336	-0.29733	-0.0119
H3079A09	cyclin A2	0	-0.225899	0.664657	0.7023
5830435F03	cyclin-dep	0	NA	0.087183	-0.161
1700068G16	cyclin-dep	0	0.222088	-0.0987	0.2311
2310034I22	cyclin C	0	-0.030829	0.619777	0.8998
C030048F16	CYCLIN A	0	0.756419	-0.217605	0.1951
2610203O19	cyclin-dep	0	0.390739	0.041612	0.4092
2700001K05	omolog	0	-0.278647	0.593802	0.2563
5930414J15	cyclin-dep	0	0.664317	0.385508	0.6416
1110019B02	cyclin-dep	0	0.074843	0.149596	0.066



BioHealthBase
BioDefense Public Health Database

Biological inference from microarray data

- What does the expression pattern tell us about the underlying biological processes?
- We developed a method of gene cluster classification that combines gene ontology probe annotation with statistical evaluation of the significance of ontology co-clustering.
- CLASSIFI - Cluster assignment for biological inference.

BioHealthBase
BioDefense Public Health Database

CLASSIFI input

```
2210009M20      vanilloid receptor-like protein 1       1
9530047P18      unclassifiable  1
2610017G21      thymidylate synthase    1
4631402F24      unclassifiable  1
5730443M15      cathepsin H     1
2810449E13      cyclin B2       1
2700047H13      unclassifiable transcript       1
2700010I01      leukemia-associated gene        1
1600009N20      growth hormone releasing hormone        1
UI-M-AN1-afg-e-07-6-UI  "protein kinase C, alpha"       1
2900042M13      unclassifiable transcript       1
2610005C18      "gene rich cluster, C8 gene"    1
5430406J05      S100 calcium binding protein A1 1
2010306L21      diazepam binding inhibitor      1
5830417J06      24-dehydrocholesterol reductase 1
2010006I05      hypothetical protein    1
4933403O03      hypothetical protein    1
1700091B02      "cytochrome P450, 51"   1
4632419C16      unclassifiable transcript       1
2010002D23      triosephosphate isomerase       1
3110056A02      EST AA793972    1
2010109P17      triosephosphate isomerase       1
4931429A04      putative CUB domain containing protein  1
3110013K01      unclassifiable transcript       1
4933406L18      similar to SERINE PROTEASE INHIBITOR-2 RELATED SEQUENCE 1 (CONTRAPSIN-RELA
H3075G01        Cd63 antigen    1
H3080A11        Cd63 antigen    1
H3119E05        Cd63 antigen    1
2010003F10      "homolog to TRANSMEMBRANE 4 SUPERFAMILY, MEMBER 5 (TETRASPAN TRANSMEMBRANE
2210407G04      "nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibi
3021401C12      unclassifiable transcript       2
1600020H07      2-hydroxyphytanoyl-CoA lyase    2
1810037B02      homolog to TESTIS SPECIFIC PROTEIN.     2
2310061N23      homolog to ALPHA-INTERFERON INDUCIBLE PROTEIN (FRAGMENT).       2
2700043D05      "superoxide dismutase 2, mitochondrial" 2
2610001L11      "tubulin, beta 5"       2
2410011K15      hypothetical protein    2
1700112I01      "related to MR19,000 OUTER ARM DYNEIN LIGHT CHAIN."     2
2310009E04      "homolog to CDNA FLJ10986 FIS, CLONE PLACE1001869, WEAKLY SIMILAR TO L- R
2010321M18      similar to EPIDIDYMAL SECRETORY PROTEIN E1 PRECURSOR.   2
2010010H11      "ATPase, H+/K+ transporting, beta polypeptide, gastric specific"        2
1810020L16      aspartylglucosaminidase 2
3110024D01      homolog to HYPOTHETICAL 49.4 KDA PROTEIN.       2
```

BioHealthBase
BioDefense Public Health Database

CLASSIFI database extraction

2210009M20	vanilloid receptor-like protein 1	1	GO:0000267,GO:0003673,GO:0003674,GO:0004871,GO:0004872,GO:0005215,GO:0005216
9530047P18	unclassifiable	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
2610017G21	thymidylate synthase	1	GO:0003673,GO:0003674,GO:0003824,GO:0004799,GO:0006139,GO:0006220,GO:0006221
4631402F24	unclassifiable	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
5730443M15	cathepsin H	1	##
2.81E+19	cyclin B2	1	##
2700047H13	unclassifiable transcript	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
2700010I01	leukemia-associated gene	1	GO:0003673,GO:0006996,GO:0007010,GO:0007017,GO:0007028,GO:0007049,GO:0007154
1600009N20	growth hormone releasing hormone	1	GO:0003673,GO:0003674,GO:0004871,GO:0005102,GO:0005179,GO:0005180,GO:0005575
UI-M-AN1-afc	protein kinase C, alpha	1	##
2900042M13	unclassifiable transcript	1	GO:0000267,GO:0003673,GO:0003674,GO:0003824,GO:0004768,GO:0005575,GO:0005623
2610005C18	gene rich cluster, C8 gene	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
5430406J05	S100 calcium binding protein A1	1	GO:0003673,GO:0003674,GO:0005488,GO:0005509
2010306L21	diazepam binding inhibitor	1	GO:0000062,GO:0003673,GO:0003674,GO:0004871,GO:0005102,GO:0005488,GO:0005541
5830417J06	24-dehydrocholesterol reductase	1	GO:0003673,GO:0005575,GO:0006629,GO:0006694,GO:0006695,GO:0008150,GO:0008151
2010006I05	hypothetical protein	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
4933403O03	hypothetical protein	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
1700091B02	cytochrome P450, 51	1	##
4632419C16	unclassifiable transcript	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
2010002D23	triosephosphate isomerase	1	##
3110056A02	EST AA793972	1	GO:0003673,GO:0006411,GO:0006464,GO:0008150,GO:0008151,GO:0008152
2010109P17	triosephosphate isomerase	1	##
4931429A04	putative CUB domain containing protein	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
3110013K01	unclassifiable transcript	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
4933406L18	similar to SERINE PROTEASE INHIBITOR-2 RELA	1	GO:0003673,GO:0003674,GO:0004857,GO:0004866,GO:0004867,GO:0004868,GO:0006411
H3075G01	Cd63 antigen	1	GO:0000323,GO:0003673,GO:0005575,GO:0005622,GO:0005623,GO:0005737,GO:0005764
H3080A11	Cd63 antigen	1	GO:0000323,GO:0003673,GO:0005575,GO:0005622,GO:0005623,GO:0005737,GO:0005764
H3119E05	Cd63 antigen	1	GO:0000323,GO:0003673,GO:0005575,GO:0005622,GO:0005623,GO:0005737,GO:0005764
2210039G11	fatty acid binding protein 5, epidermal	1	GO:0003673,GO:0003674,GO:0005215,GO:0005468,GO:0005488,GO:0006629,GO:0006810
2410022L05	homolog to HYPOTHETICAL 11.8 KDA PROTEIN	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
2610025H03	fatty acid binding protein 5, epidermal	1	GO:0003673,GO:0003674,GO:0005215,GO:0005468,GO:0005488,GO:0006629,GO:0006810
5430413I02	differential display and activated by p53	1	GO:0003673,GO:0005575,GO:0005622,GO:0005623,GO:0005737,GO:0008150,GO:0008151
3300001H05	renin binding protein	1	GO:0003673,GO:0003674,GO:0003824,GO:0016853
2600013D04	unclassifiable transcript	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
1700022C02	hypothetical protein	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
1700108K13	hypothetical protein	1	GO:0000004,GO:0003673,GO:0003674,GO:0005554,GO:0005575,GO:0008150,GO:0008372
2700038N21	leukemia-associated gene	1	GO:0003673,GO:0006996,GO:0007010,GO:0007017,GO:0007028,GO:0007049,GO:0007154

BioHealthBase
BioDefense Public Health Database

CLASSIFI analysis

*** Clusterid: 1 ***

GO id	g	f	c	n	prob	GO type	GO name
GO:0016288	1188	21	93	11	6.98E-08	biological_process	cytokinesis
GO:0000279	1188	22	93	11	1.31E-07	biological_process	M phase
GO:0000280	1188	10	93	7	1.43E-06	biological_process	nuclear division
GO:0000087	1188	10	93	7	1.43E-06	biological_process	M phase of mitotic cell cycle
GO:0007067	1188	10	93	7	1.43E-06	biological_process	mitosis
GO:0007017	1188	10	93	7	1.43E-06	biological_process	microtubule-based process
GO:0007049	1188	57	93	15	1.21E-05	biological_process	cell cycle
GO:0000278	1188	15	93	7	5.53E-05	biological_process	mitotic cell cycle
GO:0007018	1188	5	93	4	1.66E-04	biological_process	microtubule-based movement
GO:0005814	1188	5	93	4	1.66E-04	cellular_component	centriole
GO:0008274	1188	5	93	4	1.66E-04	cellular_component	gamma-tubulin ring complex
GO:0006694	1188	3	93	3	4.66E-04	biological_process	steroid biosynthesis
GO:0006695	1188	3	93	3	4.66E-04	biological_process	cholesterol biosynthesis
GO:0005815	1188	6	93	4	4.69E-04	cellular_component	microtubule organizing center
GO:0005874	1188	6	93	4	4.69E-04	cellular_component	microtubule
GO:0005813	1188	6	93	4	4.69E-04	cellular_component	centrosome
GO:0003824	1188	178	93	26	5.63E-04	molecular_function	enzyme
GO:0006629	1188	41	93	10	7.80E-04	biological_process	lipid metabolism
GO:0016538	1188	7	93	4	1.03E-03	molecular_function	cyclin-dependent protein kinase, regula
GO:0004693	1188	7	93	4	1.03E-03	molecular_function	cyclin-dependent protein kinase
GO:0015630	1188	7	93	4	1.03E-03	cellular_component	microtubule cytoskeleton
GO:0019887	1188	7	93	4	1.03E-03	molecular_function	protein kinase regulator
GO:0007010	1188	23	93	7	1.25E-03	biological_process	cytoskeleton organization and biogenes
GO:0016125	1188	4	93	3	1.76E-03	biological_process	sterol metabolism
GO:0006731	1188	4	93	3	1.76E-03	biological_process	coenzymes and prosthetic group metab
GO:0008203	1188	4	93	3	1.76E-03	biological_process	cholesterol metabol
GO:0008202	1188	4	93	3	1.76E-03	biological_process	steroid metabolism

BioHealthBase
BioDefense Public Health Database

BioHealthBase BRC - Key features

- Beyond genomics
  - Support data regarding:
    - Functional genomics
    - Proteomics
    - ELISA
    - …. and beyond
  - Support data analysis
    - Analytical applications
    - Training
- Focus on host-pathogen interactions
  - Disease pathogenesis
  - Host sensitivity and resistance
  - Vaccinology
- Data integration
  - Supported by generic conceptual data model
  - Adding biological inference to experimental results

BioHealthBase
BioDefense Public Health Database

Structured Methodology for Information Technology

- Institute of Electrical and Electronics Engineers and the Electronic Industries Association (IEEE/EIA) 12207 Standard for Information Technology
- Carnegie Mellon Institute
- Software Engineering Institute (SEI) Capability Maturity Model Integrated (CMMI)
- Level 3, Defined Engineering Process
- Standard Procedures

Bioinformatics Integration Support Contract (BISC)

- NIAID's Division of Allergy, Immunology and Transplantation (DAIT)
- Awarded 30SEP2004: 2 yr base + 4 yr option
- **Imm**unology Database and Analysis **Port**al (***ImmPort***)
- *User groups:* Programs that conduct
  - basic scientific research of genetic correlates of immune disease
  - clinical trials to evaluate the safety, toxicity, and efficacy of immune disease therapies
  - studies of the underlying mechanisms of therapeutic agents
- *Purpose:* Provide advanced information technology support in the production, analysis, archiving, and exchange of scientific data by
  - conducting requirements assessment of bioinformatics needs
  - prototyping a system for the collection, storage, and query of data
  - designing, implementing and maintaining a data warehouse of genomic, proteomic and related data
  - developing specialized applications for the capture, storage, management, query, and analysis of these data
  - measuring performance and benefits of technical support activities