"It took millions of times more computing power to map the human genome than it did to land a man on the moon, and that’s only a fraction of what’s needed right now. The biggest challenge in biology is going to be computing."

Dr. J. Craig Venter
Chairman of the Board
The Institute for Genomic Research
Dec. 2001
Clusters: What and Why?

What?
- Collection of computers networked together to perform a particular application in parallel

Why?
- Scalability
- Cost-effective
Role of Your Cluster Vendor

Expertise with cluster management

- *Allow you to concentrate on the complexities of your models – not your computer system*

Expertise with turnkey solutions

- *Open Source distribution*
- *Proven compatibility with life sciences applications*

One-stop support
Trend to Clustered Supercomputers

Numbers Overwhelm Size
Clusters in the Life Sciences

- High throughput screening runs involving multiple, repetitive sequence comparisons on vast amounts of data.

- Data or query can be partitioned and dispatched to \( n \) nodes of a cluster to gain direct \( n \)-fold speedup and gain in throughput.

- Even the most compute-intensive fine-grain applications (e.g., molecular chemistry) can be deployed on a cluster equipped with a high-performance interconnect fabric.
Case Study—Business Case

Tularik, Incorporated
San Francisco, California

Situation
- Existing infrastructure would have taken 38 years to perform the 22 million genomic sequence comparisons

Solution
- Linux Networx cluster of 150 Pentium-3 nodes with ICEBox and Clusterworx management

Results
- Study completed in 34 days – 450x acceleration
“Cluster management tools from Linux Networx are setting the standard .... Without cluster management tools, we would be spending five times as long managing the cluster.... *ICE tools allow us to concentrate on finding new genes that cause disease and not worry about cluster management*”

- Gene Cutler
  Tularik, Inc.
Terascale Linux Clusters: Supercomputing Solutions for the Life Sciences

Dr. Padmanabhan Iyer, Linux Networx
Dr. Bruce Ling, Tularik
The Need For Computing

“IT took millions of times more computing power to map the human genome than it did to land a man on the moon, and that’s only a fraction of what’s needed right now. The biggest challenge in biology is going to be computing.”

Dr. J. Craig Venter
Chairman of the Board
The Institute for Genomic Research
Dec. 2001
Clusters: What and Why?

What?
- Collection of computers networked together to perform a particular application in parallel

Why?
- Scalability
- Cost-effective
Role of Your Cluster Vendor

Expertise with cluster management

- Allow you to concentrate on the complexities of your models – not your computer system

Expertise with turnkey solutions

- Open Source distribution
- Proven compatibility with life sciences applications

One-stop support
Trend to Clustered Supercomputers

Numbers Overwhelm Size
Clusters in the Life Sciences

• High throughput screening runs involving multiple, repetitive sequence comparisons on vast amounts of data.

• Data or query can be partitioned and dispatched to $n$ nodes of a cluster to gain direct $n$-fold speedup and gain in throughput.

• Even the most compute-intensive fine-grain applications (e.g., molecular chemistry) can be deployed on a cluster equipped with a high-performance interconnect fabric.
Case Study—Business Case

Tularik, Incorporated
San Francisco, California

Situation
• Existing infrastructure would have taken 38 years to perform the 22 million genomic sequence comparisons

Solution
• Linux Networx cluster of 150 Pentium-3 nodes with ICEBox and Clusterworx management

Results
• Study completed in 34 days – 450x acceleration
Case Study— Business Success

Tularik, Incorporated
San Francisco, California

“Cluster management tools from Linux Networx are setting the standard …. Without cluster management tools, we would be spending five times as long managing the cluster…. ICE tools allow us to concentrate on finding new genes that cause disease and not worry about cluster management”

- Gene Cutler
  Tularik, Inc.
Bruce Ling, Ph.D
Director, Bioinformatics
Discovery Platform – working around the clock

**Tularik, Inc.** in S. San Francisco, CA
- headquarter

**Tularik Genomics Division** in Cold Spring Harbor, NY
- Cancer gene discovery

**Tularik Pharmaceutical Company** in S. San Francisco, CA
- Metabolic disease drug discovery

**Tularik GmbH** in Regensburg, Germany
- Screening technology and HTS

**Cumbre Inc.** in Dallas, TX
- Anti-microbial drug discovery

**Tularik Ltd.** in Macclesfield, United Kingdom
- Computer-Aided Molecular Design (CAMD)
  - Virtual screening
Restoring Normal Gene Expression

Regulatory pathway

Tularik small molecule drug

Normal protein levels
Tularik’s Drug Discovery Pipeline

**Step 1:** Link disease to gene(s)
- Abnormal gene expression

**Step 2:** Elucidate regulatory pathway(s)
- Regulatory pathway

**Step 3:** Develop assays to identify leads
- Potential lead

**Step 4:** Optimize lead compounds
- Potency
- Specificity
- Oral bioavailability
- Drug/drug interactions

**Step 5:** Test preclinical candidates
- Efficacy models
- ADME
- Toxicity

**Step 6:** Perform clinical trials
- Safety
- Efficacy
<table>
<thead>
<tr>
<th></th>
<th>Drug Targets</th>
<th>Drug Leads</th>
<th>IND Candidates</th>
<th>Clinical Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Various tumor types</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Immunology</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMV - transplant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immune disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflammation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Metabolic disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>88</td>
<td>26</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Public Domain Genome Projects

Human Genome Sequencing

Sequencing Progress

Genome Watch

- Draft: 34.8%
- Finished: 63.0%
- Total: 97.8%

Assumptions and additional statistics...

- > 1000 kb
- 250 – 1000 kb
- < 250 kb

- draft sequence
- heterochromatin
Bioinformatics Data Mining

What is the function of this:

A. Sequence?

- High complexity
- Repetitive
- Very low complexity

B. Structure?

- β barrel
- Simple α bundle
- Complex multidomain

C. Motif?

- Calcium binding
- Nucleotide binding
- Membrane anchoring

Pattern Match
Data Mining Bottleneck

Issue

Computing power deficit
Start Small, Grow as Needed
High Throughput Computing

- Key: effective management and exploitation of all available computing resources
- Approach: using batch queuing systems
Benefits of Batch Queuing System

- Enable resource sharing across multiple platforms (No dedicated resource for particular project or personnel)

- Improve utilization of overall computing resources (CPU) from less than 20% to over 90%

- Increase visibility and availability of existing computing resources

- Scalability
  - Start with a small system, add more later
  - Revenue growth alignment
Linux Cluster Supporting Data Mining

single node
Old vs. New

- Time taken to BLAST raw mouse genomic sequence read against human genome database:

<table>
<thead>
<tr>
<th></th>
<th>Darwin (SGI Irix 6.5)</th>
<th>Linux Cluster (RedHat 7.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mouse sequence</td>
<td>1 minute</td>
<td>10 seconds</td>
</tr>
<tr>
<td>1000 mouse sequences</td>
<td>15 hours</td>
<td>3 minutes</td>
</tr>
<tr>
<td>All mouse genomic</td>
<td>38 years</td>
<td>34 days</td>
</tr>
<tr>
<td>sequence reads at NCBI (22 million reads)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note:
Darwin is a heavily used machine, so it is not a machine to machine comparison. However, it does accurately reflect the environment in which these computers are used.
Linux Cluster Dedicated to J2EE Discovery Architecture
Discovery Platform J2EE Application Cluster
Tularik in the news

Tularik Inc.
USA
Tularik Inc.

Monday August 5, 7:45 am Eastern Time

Press Release

SOURCE: Linux NetworX

Tularik Uses Linux NetworX Cluster Supercomputer For Genomic Research & Drug Development

Linux NetworX Evolocty(TM) Supercomputer Increases Performance by 78x

SALT LAKE CITY, Aug. 5 /PRNewswire/ -- Linux NetworX announced today that San Francisco-based Tularik Inc. (Nasdaq: TRLX - News) is using a Linux NetworX Evolocty(TM) cluster supercomputer to boost data mining performance by 78x for drug development.

Tularik, a biopharmaceutical company, relies on the Linux NetworX cluster to accelerate its drug discovery efforts by quickly identifying gene combinations behind diseases in the areas of cancer, immunology and metabolic disorders. Using the genomics information processed by the cluster, Tularik expects to develop pharmaceuticals that regulate gene expression.
On the sequencing of the human genome

Robert H. Waterston*, Erik S. Landerg, and John E. Sulston

*Genome Sequencing Center, Washington University, St. Louis, MO 63108; Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom.

On the sequencing and assembly of the human genome

Eugene W. Myers*, Granger G. Sutton, Hamilton O. Smith, Mark D. Adams, and J. Craig Venter

Celera Genomics, 48W. Gude Drive, Rockville, MD 20850
<table>
<thead>
<tr>
<th>Release date</th>
<th>Assembly HGSC (UCSC)</th>
<th>Assembly Celera</th>
<th>Curated genes HGSC (Ensembl)</th>
<th>Curated genes Celera</th>
</tr>
</thead>
<tbody>
<tr>
<td>05-2000</td>
<td></td>
<td>R18, R19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>06-2000</td>
<td>06-2000</td>
<td>R20, R21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>07-2000</td>
<td>07-2000</td>
<td>R22, R23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>08-2000</td>
<td></td>
<td>R24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09-2000</td>
<td>09-2000</td>
<td></td>
<td>E-0.8.0</td>
<td></td>
</tr>
<tr>
<td>10-2000</td>
<td>10-2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11-2000</td>
<td></td>
<td>R25e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-2000</td>
<td>12-2000</td>
<td></td>
<td>E-1.0.0</td>
<td>R25e</td>
</tr>
<tr>
<td>01-2001</td>
<td></td>
<td>R25h</td>
<td>R25h</td>
<td></td>
</tr>
<tr>
<td>04-2001</td>
<td>04-2001</td>
<td></td>
<td>E-1.1.0</td>
<td></td>
</tr>
<tr>
<td>07-2001</td>
<td></td>
<td>R26b</td>
<td>R26b</td>
<td></td>
</tr>
<tr>
<td>08-2001</td>
<td>08-2001</td>
<td></td>
<td>E-1.2.0</td>
<td></td>
</tr>
<tr>
<td>10-2001</td>
<td></td>
<td></td>
<td></td>
<td>R26d</td>
</tr>
<tr>
<td>11-2001</td>
<td></td>
<td></td>
<td></td>
<td>R26e</td>
</tr>
<tr>
<td>12-2001</td>
<td>12-2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01-2002</td>
<td></td>
<td>R26f</td>
<td>E-3.26</td>
<td>R26f, R26h</td>
</tr>
<tr>
<td>03-2002</td>
<td></td>
<td></td>
<td>E-4.28</td>
<td></td>
</tr>
<tr>
<td>04-2002</td>
<td>04-2002</td>
<td></td>
<td></td>
<td>R26j</td>
</tr>
<tr>
<td>05-2002</td>
<td></td>
<td>R26i</td>
<td>E-5.28</td>
<td></td>
</tr>
<tr>
<td>06-2002</td>
<td></td>
<td></td>
<td></td>
<td>R26k</td>
</tr>
</tbody>
</table>
Comparative Analysis of Human Genome Assemblies Reveals Genome-Level Differences

Shuyu Li,1 Jiayu Liao,2,* Gene Cutler,1 Timothy Hoey,1 John B. Hogenesch,2 Michael P. Cooke,2 Peter G. Schultz,2 and Xuefeng Bruce Ling1,*

1Tularik, Inc., Two Corporate Drive, South San Francisco, California 94080, USA
2The Genomic Institute of Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA

*To whom correspondence and reprint requests should be addressed. Fax: (858) 812-1502. E-mail: liae@gnf.org. Fax: (650) 825-7400. E-mail: xling@tularik.com.
Surfing the Human Genome

Databases of Genetic Code Are Moving to the Web

By LAWRENCE M. FISHER

SAN FRANCISCO -- Call it an end-of-the-century business case study.
Ontology Engineering

**GENE FUNCTION**

- **ROOT**
  - **STRUCTURAL PROTEINS**
  - **RECEPTORS**
  - **TRANSFERASES Level 2**
    - **ENZYME Level 1**
    - **HYDROLASES Level 2**

**TRANSFERASE signature domains**

**SWISS-PROT sequences**
- Human member A
- Human member B
- mouse
- chicken
- human

**ACCEPTED SEQUENCES**
- mouse

**REJECTED SEQUENCES**
- mouse
Methods

DIAN: A Novel Algorithm for Genome Ontological Classification

Yannick Pouliot, Jing Gao, Qiaojuan Jane Su, Guozhen Gordon Liu, and Xuefeng Bruce Ling$^{1,2}$
DIAN: A Novel Algorithm for Genome Ontological Classification

Yannick Poulitot, Jing Gao, Qiaojuan Jane Su, Guozhen Gordon Liu, and Xuefeng Bruce Ling

Methods

Computer-Aided Human Curation

Evolving Controlled vocabulary

Map unto ontology

Correct assignment

Incorrect assignment

Domain-based mapping

Vocabulary-based mapping

Assignment DB

DIAN Database
Use Ontology to Annotate MicroArray data

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOR154C</td>
<td>protein degradation</td>
</tr>
<tr>
<td>YBR62C</td>
<td>protein degradation</td>
</tr>
<tr>
<td>YPR158W</td>
<td>unknown</td>
</tr>
<tr>
<td>YGR145W</td>
<td>unknown</td>
</tr>
<tr>
<td>YOL092W</td>
<td>unknown</td>
</tr>
<tr>
<td>YDR171W</td>
<td>stress response</td>
</tr>
<tr>
<td>YKR011C</td>
<td>unknown</td>
</tr>
<tr>
<td>YGL087C</td>
<td>protein degradation</td>
</tr>
<tr>
<td>YHR023C</td>
<td>unknown</td>
</tr>
<tr>
<td>YHR010W</td>
<td>protein degradation</td>
</tr>
<tr>
<td>YKL213C</td>
<td>DDA1</td>
</tr>
<tr>
<td>YGR136W</td>
<td>unknown</td>
</tr>
<tr>
<td>YL078W</td>
<td>vesicle coat assembly</td>
</tr>
<tr>
<td>YOR077C</td>
<td>unknown</td>
</tr>
<tr>
<td>YNL027C</td>
<td>protein folding</td>
</tr>
<tr>
<td>YOR027W</td>
<td>protein folding</td>
</tr>
<tr>
<td>YLP246C</td>
<td>protein folding</td>
</tr>
<tr>
<td>YMR186W</td>
<td>protein folding</td>
</tr>
<tr>
<td>YER132W</td>
<td>protein folding</td>
</tr>
<tr>
<td>YDR286C</td>
<td>protein folding</td>
</tr>
<tr>
<td>YLR025W</td>
<td>protein folding</td>
</tr>
<tr>
<td>YLR217W</td>
<td>unknown</td>
</tr>
<tr>
<td>YLR218C</td>
<td>protein folding</td>
</tr>
<tr>
<td>YLR238C</td>
<td>protein folding</td>
</tr>
<tr>
<td>YDR214W</td>
<td>protein folding</td>
</tr>
</tbody>
</table>
Target Identification
Request

Patent & competitors’ information are badly needed for Tularik Genes of Interest
Introducing Commercial Patent Database In House
Automatic Pipeline Architecture enables
- Regular & automatic data analysis initiation
- Proper Update

Dynamically link to BLAST Application

Dynamically link to other intellect property sites
- Derwent.com
- Nerac.com
- Dialogweb.com

Synthesize information from various resources into the genomic relational database and notify relevant decision makers
Micro-Array Bioinformatics
Chip Design: Tularik Genes of Interest
Oligo Chip Design Algorithm Development

- **Software package: Oligo-Jungle**
  - 60 – 80 mer
  - Uniqueness in the genome
  - Optimal melting temperature
  - No secondary structure
  - Allow high throughput design
Algorithm Design Concerns

- Minimized distance from 3' end (<350 bases)
- Consideration of splice variants
  one probe designed to recognize all splice variants for a gene
- Minimized cross-homology
TMAX – Tularik Micro Array Explorer

Tularik Micro-array Data Storage and Analysis Software Suite
Welcome to the Tularik MicroArray EXplorer v2.0

Administration
- Add Experiments
- Batch Upload Application for Mac
- Batch Upload key
- Agilent to TMAX
- Delete Experiments
- Delete experiments from the database

Analysis
- Experiment Summarizer
  Quickly summarize data from microarray experiments
- Basic Query
  Use simple queries to download or view experiment data
- Complex Query
  Use complex queries to download or view experiment data
- ChipViewer
  View and analyze microarray images
- ChipPlotter
  Compare expression data from two microarray experiments
- ChipCluster
  Perform and view expression clustering of genes from microarray experiments
- GenomeScan
  View microarray copy number data based on chromosomal location

Extras
- Database schema diagram
- MRJ Libraries for using java on a Mac. Uncompress this file and put it in the Extensions Folder in the System Folder.
- Guide for running java applets on your computer
- Description of the database fields used by TMAX

Support
Tularik MAX was developed by Gene Cutler. Please email Gene for support.
TMAX Overview

- **TMAX** is Tularik’s in-house micro-array data storage and analysis software solution.

  - Designed for the needs of biologists studying gene expression or genomic amplification/deletion.

  - Flexible technology-independent design that handles data in in a variety of formats including Incyte, Affymetrix, Scanalyze, Genepix, Rosetta, Motorola, and simple spreadsheets.

  - TMAX is comprised of 12 front-end applications plus a database server and several administrator support tools.

  - Currently, TMAX contains about more than 14 million data points.
TMAX Overview

• Data Management -- Administration
  • Add Experiment
  • Delete Experiment
  • Edit Experiment Annotation
  • Combine Experiments
  • Export Experiments

• Data Analysis / Visualization
  • Experiment Summarizer
  • Basic Query
  • Complex Query
  • Chip-Viewer
  • Chip-Plotter
  • Chip-Cluster
  • Genome-Scan
TMAX Overview

- Administration
  - Add Experiment
    - Upload data, sequence, and image files into the TMAX database.
  - Delete Experiment
    - Remove unwanted experiments from the database.
  - Edit Experiment Annotation
    - Modify annotation for loaded experiments.
  - Combine Experiments
    - Generate synthetic data sets by combining replicate data sets.
  - Export Experiments
    - Export loaded experiments for use with other software.
TMAX Overview

- Analysis
  - Experiment Summarizer
    - Generate a quick summary of an experiment, including research notes; numbers of genes up-regulated, down-regulated and unchanged; top 5 up-regulated genes; top 5 down-regulated genes.
  - Basic Query
    - Retrieve data across multiple experiments based on signal intensity, fold change, quality score, gene name, and/or gene accession values.
  - Complex Query (functional group analysis)
    - Retrieve data across multiple experiments by specifying fold change criteria in individual conditions. Can do “fuzzy” matching.
TMAX Overview

- **Visualization analysis**
  - **Chip-Viewer**
    - Display fluorescence image of micro-array and pick spots to display corresponding data. Mainly for quality control.
  - **Chip-Plotter**
    - Display scatter plots of fold change values for a selected pair of experiments to quickly identify genes that are similarly or differentially modulated in different conditions.
  - **Chip-Cluster**
    - Perform clustering of data across multiple experiments so that patterns of change can be observed. Can choose between a variety of hierarchical and non-hierarchical algorithms.
  - **Genome-Scan**
    - Plot fold change data from multiple experiments along chromosomes using radiation-hybrid mapping data. Allows for quick identification of genomic regions that are amplified or deleted.
### TMAX Add Experiment

#### Experiment Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientist Name</td>
<td>Gene Cottler</td>
</tr>
<tr>
<td>Research Group</td>
<td>Testing</td>
</tr>
<tr>
<td>Microarray Source</td>
<td>Motorola, Human</td>
</tr>
<tr>
<td>Barcode</td>
<td>120938623</td>
</tr>
<tr>
<td>Fluor Reversal</td>
<td>No</td>
</tr>
<tr>
<td>Scanning Software</td>
<td>GenePix</td>
</tr>
<tr>
<td>Notes</td>
<td></td>
</tr>
</tbody>
</table>

#### Sample Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Experimental Sample</th>
<th>Reference Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T Number</td>
<td>T0801317</td>
<td></td>
</tr>
<tr>
<td>Compound Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compound Amount</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment Length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Series Index</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TMAX Basic Query -- Data Selection

#### Query

**11 experiments match your query**

<table>
<thead>
<tr>
<th>Experiment ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>LPS:180 min Whole Blood Cells</td>
</tr>
<tr>
<td>39</td>
<td>LPS:180 min Whole Blood Cells 2</td>
</tr>
<tr>
<td>36</td>
<td>LPS:60 min Whole Blood Cells</td>
</tr>
<tr>
<td>37</td>
<td>LPS:60 min Whole Blood Cells 2</td>
</tr>
<tr>
<td>28</td>
<td>T001117 Mono Mac 6</td>
</tr>
<tr>
<td>26</td>
<td>T00018 2 h Gm hypothalamus</td>
</tr>
<tr>
<td>34</td>
<td>T001158 6.6 h Gm hypothalamus</td>
</tr>
<tr>
<td>35</td>
<td>T001158 3.5 h Gm hypothalamus</td>
</tr>
<tr>
<td>33</td>
<td>T001158 2 h Gm hypothalamus</td>
</tr>
<tr>
<td>30</td>
<td>T001158 30 min Gm hypothalamus</td>
</tr>
<tr>
<td>31</td>
<td>T001158 30 min Gm hypothalamus</td>
</tr>
<tr>
<td>32</td>
<td>T001158 30 min Gm hypothalamus</td>
</tr>
</tbody>
</table>

#### Query Criteria

**Absolute Intensity**

Signal intensity of the experimental and/or reference sample. Uses the `Intensity_1_exp` and `Intensity_1_ref` database fields.

**Fold Change**

Change in gene expression or nucleotide level. Unchanged genes have a fold change of 1. Upregulated genes are greater than 1. Downregulated genes are less than 1. Choose `<>` (absolute) and a fold change greater than 1 to select both upregulated and downregulated genes.

Uses the `ratio_1` database fields.

**Sequence Name**

Selects sequences that match this text in their names.

Uses the `sequence_name` database field.

**Accession Number**

Selects sequences that match the supplied accession number(s). This match is always case insensitive. Use the upper input box for entering a single accession number. Use the lower input box to select a plain text file containing multiple accession numbers.

Uses the `sequence_accession` database field.

**Quality Score**

Quality Score is a measure for how reliable the fold change value is. It ranges from 0 to 100. Scores above 50 are moderately reliable, scores above 75 are highly reliable.

Uses the `quality` database field.

**Coefficient of Variance**

Measures the reproducibility of data in averaged ("synthetic") data sets. Smaller numbers are better.

Uses the `cv` database field.

**Exclude Controls**

Exclude control data from the query. This excludes sequences that come from the CONTROL database.

Uses the `sequence_database_name` database field.

#### Field Selection

**Spot Data**

- **Spot ID**
- **Array grid**

**Sequence Annotation**

- **Accession number**
- **Gene name**

**Experiment Annotation**

- **Experiment ID**
- **Experiment name**
### TMAX Basic Query

#### Basic Query: Query Results

- Results can also be downloaded as a spreadsheet for further processing.
Complex Query: Data Selection

- Select genes that have specific fold changes with different treatments
- Can do “fuzzy matching” (e.g. up-regulated in 3 out of 4 conditions)
TMAX Complex Query

<table>
<thead>
<tr>
<th>ID</th>
<th>Ratio</th>
<th>Ratio</th>
<th>Ratio</th>
<th>Ratio</th>
<th>Ratio</th>
<th>Ratio</th>
<th>Accession</th>
<th>Database</th>
<th>BLAST Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>T009156:2 k/z Hypothalamus</td>
<td>-1.37</td>
<td>-1.17</td>
<td>-1.05</td>
<td>-2.06</td>
<td>-2.24</td>
<td>11.97</td>
<td>NM_009152</td>
<td>NM_009152</td>
<td>BLAST NM_009152</td>
</tr>
<tr>
<td>T009156:2 k/z Hypothalamus</td>
<td>1.56</td>
<td>2.09</td>
<td>2.27</td>
<td>1.08</td>
<td>-2.83</td>
<td></td>
<td>NM_009156</td>
<td>NM_009156</td>
<td>BLAST NM_009156</td>
</tr>
<tr>
<td>T009156:2 k/z Hypothalamus</td>
<td>1.03</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
<td>NM_009157</td>
<td>NM_009157</td>
<td>BLAST NM_009157</td>
</tr>
</tbody>
</table>

Complex Query: Query Results
- Data can also be downloaded as a spreadsheet for further analysis
**Viewer Window:** Picking a spot gives information for that spot in the table view

**Table View:** Experimental data for selected spots can be sorted, columns rearranged.

**Query Window:** Find spots based on Gene Name or Accession

**Detail Window:** Additional spot details including gene chromosome position; links to Genbank, Unigene, BLAST search
TMAX Chip-Plotter

Plotter: Results Display
Cluster: Results Window

- Each row represents a single gene across several experimental conditions.
- The color of each box represents the fold change of one gene in one condition.
**View Window:** Each gene from each condition is plotted along the selected chromosome based on mapping data. Fold change is plotted on the Y-axis.

**Control Panel:** Selects the chromosome to plot, the mapping panel to display, the quality cutoff score for data display and the scale of the Y-axis

**Gene Table:** Displays additional data on the selected genes including experimental measurements, chromosome mapping information, and UniGene cluster information
Tularik’s lead discovery informatics

- ~80 to 90 current targets
- 50+ different HTS assays performed annually
- >950,000 compound screening library

- Proprietary & comprehensive screening database
- Information intensive lead discovery demands powerful informatics support
Types of Data, Links

- Biological results from HTS
- Equipment scheduling

- T-number database
- Chemical characterization
- Inventory

- Biological assays
- CYP, solubility/permeability, cytotoxicity
- PK

- Pharmacology
- Special PK experiments
IT&S for Research

Data Management
- Acquisition
- Reports
- Storage

Targets
- Identification → Validation

Compounds
- Acquisition → Inventory Processes

Assays
- HTS
- ADMET
- SAR
- MicroArray
Assays---Activities
Lead Discovery and Follow-up Experiments

HTS  SAR  ADMET  Microarrays

Assay Registration
On-line Data Management
Automated Data Analyses
Reporting (ISIS & Browsers)
Business Entities Modeling

- Resource Entities
- Reports Entities
- Security Entities
- Administration Entities
- Collaboration Entities
- Biology Entities
- Operation Entities
- Data Entities
- Chemistry Entities
“Discovery Platform for Drug Discovery Data management”

- Curve Fitting
- Raw Data
- Data Automatic Upload
- Biology Process Logic
- Configuration
- Data Interpretation e.g. “Screen Hits”
- Compound Priority
- Other Data
- Data QA
- Data Analysis
- Presentation Logic

Data Logic

Chemistry Process Logic
Dedicated Linux cluster Infrastructure

http://discovery.tularik.com

Corporate Database

1 mouse

1000 mouse sequences

All mouse genomic sequence reads at NCBI (22 million reads)

34 days

38 years

3 minutes

15 hours

1000 mouse sequences

10 seconds

1 minute

1 mouse sequence
Automated data flow
- http://discovery.tularik.com

Data Source

Live Data Capturing

CORPORATE DATABASE

Analysis

ISIS

Internet Explorer
Discovery Platform
– http://discovery.tularik.com

Discovery J2EE Architecture

J2EE
• Extensible
• Flexible
• Reusable
# Tularik Informatics

**Category:** HTS  **Program:** GPCR  **Month:** April  **Year:** 2002

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Robot:** CLIPR_1  **Assay Name:** TGR341_AGN_AEQ_S  **Run:** p1

[Search for Assay]